Н.С.Бахвалов, Н.П.Жидков, Г.М.Кобельков

ЧИСЛЕННЫЕ МЕТОДЫ

Данная книга представляет собой переработанный вариант учебного пособия «Численные методы» тех же авторов, вышедшего в 1987 году. Добавлен материал, относящийся к решению систем линейных уравнений с плохо обусловленными матрицами, решению задачи Коши для систем жестких обыкновенных дифференциальных уравнений, аппроксимации функций, методу сопряженных градиентов. Видоизменено изложение оптимального линейного итерационного процесса и рассмотрен многосеточный итерационный метод — один из наиболее применяемых в настоящее время методов решения сеточных краевых задач.

Содержание

Предисловие	7
Введение	8
1 Погрешность результата численного решения задачи	**17**
§ 1. Источники и классификация погрешности	17
§ 2. Запись чисел в ЭВМ	21
§ 3. Абсолютная и относительная погрешности. Формы записи данных	22
§ 4. О вычислительной погрешности	25
§ 5. Погрешность функции	27
§ 6. Обратная задача	32
2 Интерполяция и численное дифференцирование	**35**
§ 1 Постановка задачи приближения функций	36
§ 2. Интерполяционный многочлен Лагранжа	39
§ 3. Оценка остаточного члена интерполяционного многочлена Лагранжа	43
§ 4. Разделенные разности и их свойства	43
§ 5. Интерполяционная формула Ньютона с разделенными разностями	45
§ 6. Разделенные разности и интерполирование с кратными узлами	48
§ 7. Уравнения в конечных разностях	51
§ 8. Многочлены Чебышева	58
§ 9. Минимизация оценки остаточного члена интерполяционной формулы	62
§ 10. Конечные разности	65
§ 11. Интерполяционные формулы для таблиц с постоянным шагом	68
§ 12. Составление таблиц	71
§ 13. О погрешности округления при интерполяции	74
§ 14. Применения аппарата интерполирования. Обратная интерполяция	75
§ 15. Численное дифференцирование	76
§ 16. О вычислительной погрешности формул численного дифференцирования	83
§ 17. Рациональная интерполяция	84
3 Численное интегрирование	**86**
§1. Простейшие квадратурные формулы. Метод неопределенных коэффициентов	86
§ 2. Оценки погрешности квадратуры	89

§ 3. Квадратурные формулы Ньютона— Котеса 94
§ 4. Ортогональные многочлены 99
§ 5. Квадратурные формулы Гаусса 106
§ 6. Практическая оценка погрешности элементарных квадратурных формул 113
§ 7. Интегрирование быстро осциллирующих функций 116
§ 8. Повышение точности интегрирования за счет разбиения отрезка на равные части 119
§9. О постановках задач оптимизации 124
§ 10. Постановка задачи оптимизации квадратур 129
§11. Оптимизация распределения узлов квадратурной формулы 130
§ 12. Примеры оптимизации распределения узлов 137
§ 13. Главный член погрешности 140
§14. Правило Рунге практической оценки погрешности 144
§ 15. Уточнение результата интерполяцией более высокого порядка точности 148
§ 16. Вычисление интегралов в нерегулярном случае 150
§ 17. Принципы построения стандартных программ с автоматическим выбором шага 157
4 Приближение функций и смежные вопросы 164
§ 1. Наилучшие приближения в линейном нормированном пространстве 164
§ 2. Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении 166
§ 3. Тригонометрическая интерполяция. Дискретное преобразование Фурье 171
§ 4. Быстрое преобразование Фурье 175
§ 5. Наилучшее равномерное приближение 178
§ 6. Примеры наилучшего равномерного приближения 181
§ 7. О форме записи многочлена 187
§ 8. Интерполяция и приближение сплайнами 191
5 Многомерные задачи 201
§ 1. Метод неопределенных коэффициентов 202
§ 2. Метод наименьших квадратов и регуляризация. 203
§ 3. Примеры регуляризации 206
§ 4. Сведение многомерных задач к одномерным 212
§ 5. Интерполяция функций в треугольнике 220
§ 6. Оценка погрешности численного интегрирования на равномерной сетке 222
§ 7. Оценка снизу погрешности численного интегрирования 225
§ 8. Метод Монте-Карло 232
§ 9. Обсуждение правомерности использования недетерминированных методов решения задач 236
§ 10. Ускорение сходимости метода Монте-Карло 239
§ 11. О выборе метода решения задачи 243

6 Численные методы алгебры **250**
§ 1. Методы последовательного исключения неизвестных 253
§ 2. Метод отражений 262
§ 3. Метод простой итерации 265
§ 4. Особенности реализации метода простой итерации на ЭВМ 268
§ 5. δ^2 -процесс практической оценки погрешности и ускорения сходимости 271
§6. Оптимизация скорости сходимости итерационных процессов 275
§ 7. Метод Зейделя 285
§ 8. Метод наискорейшего градиентного спуска 290
§ 9. Метод сопряженных градиентов 294
§ 10. Итерационные методы с использованием спектрально-эквивалентных операторов 300
§11. Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация 304
§ 12. Проблема собственных значений 315
§ 13. Решение полной проблемы собственных значений при помощи QR-алгоритма 320
7 Решение систем нелинейных уравнений и задач оптимизации **324**
§ 1. Метод простой итерации и смежные вопросы 326
§ 2. Метод Ньютона решения нелинейных уравнений 330
§ 3. Методы спуска 336
§ 4. Другие методы сведения многомерных задач к задачам меньшей размерности 341
§ 5. Решение стационарных задач путем установления 345
§ 6. Как оптимизировать ? 352
8 Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений **360**
§ 1. Решение задачи Коши с помощью формулы Тейлора 361
§ 2. Методы Рунге— Кутта 363
§ 3. Методы с контролем погрешности на шаге 369
§ 4. Оценки погрешности одношаговых методов 371
§ 5. Конечно-разностные методы 376
§ 6. Метод неопределенных коэффициентов 379
§ 7. Исследование свойств конечно-разностных методов на модельных задачах 383
§ 8. Оценка погрешности конечно-разностных методов 388
§ 9. Особенности интегрирования систем уравнений 396
§ 10. Методы численного интегрирования уравнений второго порядка 409
§11. Оптимизация распределения узлов интегрирования 412
9 Численные методы решения краевых задач для обыкновенных дифференциальных уравнений **417**
§ 1. Простейшие методы решения краевой задачи для уравнений второго порядка 417

§ 2. Функция Грина сеточной краевой задачи 423
§ 3. Решение простейшей краевой сеточной задачи 428
§ 4. Замыкания вычислительных алгоритмов 436
§ 5. Обсуждение постановок краевых задач для линейных систем первого порядка 444
§ 6. Алгоритмы решения краевых задач для систем уравнений первого порядка 449
§ 7. Нелинейные краевые задачи 455
§ 8. Аппроксимации специального типа 461
§ 9. Конечно-разностные методы отыскания собственных значений 473
§ 10. Построение численных методов с помощью вариационных принципов 476
§ 11. Улучшение сходимости вариационных методов в нерегулярном случае 485
§ 12. Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения 488
10 Методы решения уравнений в частных производных 495
§ 1. Основные понятия теории метода сеток. 497
§ 2. Аппроксимация простейших гиперболических задач 505
§ 3. Принцип замороженных коэффициентов 521
§ 4. Численное решение нелинейных задач с разрывными решениями 524
§ 5. Разностные схемы для одномерного параболического уравнения 528
§ 6. Разностная аппроксимация эллиптических уравнений 543
§ 7. Решение параболических уравнений с несколькими пространственными переменными 566
§ 8. Методы решения сеточных эллиптических уравнений 580
11 Численные методы решения интегральных уравнений 599
§1. Решение интегральных уравнений методом замены интеграла квадратурной суммой 599
§ 2. Решение интегральных уравнений с помощью замены ядра на вырожденное 604
§ 3. Интегральные уравнения Фредгольма первого рода 608
Заключение 617
Список литературы 622
Предметный указатель 627

Предметный указатель

Алгоритм ненасыщаемый, 65
Аппроксимация, 379
дифференциального уравнения разностной схемой, 379
Большое число, 23
Валле-Пуссена теорема, 179
Вариация, 522
Верная цифра, 23
Весовая функция, 89
Восполнение, 559
Выделение весовой функции, 151
Гарантированные оценки погрешности на классе функций, 232
Главный член погрешности, 140
Дивергентность разностной схемы,

527
Дискретные коэффициенты Фурье, 173
Дифференцирование численное, 76
Замыкание алгоритма, 437
нерегулярное, 437
регулярное, 437
Значащие цифры, 23
Интегрирование
осциллирующих функций, 116
системы уравнений, 396
Интерполирование, 36
с кратными узлами, 49
Интерполяционная формула
Лагранжа, 41
Ньютона, 46
для интерполирования назад, 69
для интерполирования вперед, 69
Интерполяция, 36
квадратичная, 72
линейная, 71
тригонометрическая, 173
Итерационные методы построения многочлена наилучшего равномерного приближения с использованием спектрально эквивалентных операторов, 300
Квадратуры
Гаусса, 106
Грегори, 143
Лобатто, 111
Ньютона—Котеса, 94
Ромберга, 149
Симпсона, 88
Филона, 117
Эйлера, 142
обобщенные, 122
прямоугольников, 86
составные, 122
трапеций, 87
Количество арифметических операций, 41
Конечно-разностное уравнение, 51
Конечно-разностные
методы, 376
схемы, 376
Ленточная структура, 257
Линейная оценка погрешности, 28
Линейное разностное уравнение, 52
Мажорирующее разностное уравнение, 488
Матрица
Грама, 167
ортогонализации, 100
отражений, 262
Мера
обусловленности
матрицы, 305
системы, 305
погрешности аппроксимации, 504
Метод
Адамса, 378
Гаусса, 253
Зейделя, 290
Монте-Карло, 232
Ньютона решения нелинейных уравнений, 330
Ритца, 477
Рунге—Кутта, 363
Эйлера, 363
вариационно-разностный, 479
верхней релаксации, 290
вилки, 336
квадратного корня, 259
конечных элементов, 558
наименьших квадратов, 203
наискорейшего спуска, 290
неопределенных коэффициентов, 39, 202
оврагов, 342
оптимальный, 63
парабол, 336
покоординатного спуска, 288
пристрелки, 429
прогонки, 430
проекционно-разностный, 561
простой итерации, 265, 326
регуляризации, 205

релаксации, 289
сверхрелаксации, 290
секущих, 335
сопряженных градиентов, 294
спуска, 336
стрельбы, 449
суммарной аппроксимации, 579
циклической прогонки, 434
штрафа, 340
Методы
интерполяционные, 376
экстраполяционные, 376
Многочлен наилучшего равномерного приближения, 178
Многочлены
Лагерра, 104
Лежандра, 104
Чебышева, 58
второго рода, 104
первого рода, 104
Эрмита, 104
Якоби, 103
наименее уклоняющиеся от нуля, 60
Наилучшее равномерное приближение, 178
Недетерминированный метод, 242
Некорректные задачи, 608
Неравенство
Бесселя, 169
Чебышева, 233
ε-неравенство, 537
Неустойчивость, 35
Неявная схема, метод, 376, 530
Норма энергетическая, 555
Нормы
векторов и матриц, 250
эквивалентные, 266
Область
зависимости, 498
сходимости метода, 358
Обобщенное решение, 562
Обратная интерполяция, 76
Обратный ход метода
Гаусса, 254
прогонки, 431
Обусловленность
матрицы системы, 305
системы, 305
Однородные схемы, 469
Односторонние формулы численного дифференцирования, 80
Одношаговые методы численного интегрирования, 371
Оператор расщепляющийся, 573
Оптимальные квадратуры, 129
Оптимальный
линейный итерационный процесс, 279, 283
по порядку итерационный процесс, 42, 63
Оптимизация
методов, 63
оценки погрешности интерполирования, 63
распределения узлов интегрирования. 131
скорости сходимости итерационного процесса, 275
Ортогональная система, 99
Ортогональные многочлены, 101
Ортонормированная система элементов, 101
Остаточный член формулы Лагранжа, 43
Очень большое число. 23
Параметр регуляризации, 611
Переобуславливатель, 301
Планирование эксперимента, 37
Плохо обусловленные системы, 307
Повышение порядка точности разностной схемы, 419
Погрешность
абсолютная, 22
аппроксимации дифференциального уравнения разностной схемой, 504, 534
вычислительная, 17
квадратуры на классе функций, 129

математической модели, 17
метода, 17
на классе задач, 63
на шаге, 371
неустранимая, 17
относительная, 22
предельная, 27
Порядок
метода, 329
погрешности аппроксимации, 380
Почти собственные значения, 270
Преобразование Фурье
быстрое, 175
дискретное, 173
Принцип замороженных коэффициентов, 521
Проблема собственных значений
полная, 315
частичная, 315
δ^2-процесс ускорения сходимости, 273
Прямое произведение формул интегрирования, интерполирования, дифференцирования, 219
Прямой ход метода Гаусса, 254
прогонки, 431
Разности
вперед, 65
высшего порядка, 65
конечные. 65
назад, 65
разделенные, 43
центральные, 65
Разностная схема, 377
экономичная, 568
Ряд Фурье дискретный, конечный, 173
Сеточная функция Грина, 423
Сжатое отображение, 326
Симметризация системы уравнений, 276
Система жесткая, 398
Слой, 530
Согласованные нормы, 501
Спектральная эквивалентность, 300
Спектральный признак устойчивости, 506, 515
Сплайн, 191
интерполяционный, 197
локальный, 198
Стандартные программы, 47
Строго нормированное пространство, 165
Схема Эйткена, 47
Схемы, точные на решениях специального вида, 461
Сходимость, 501
Таблица разделенных разностей. 45
Теорема
Валле-Пуссена, 179
Чебышева, 179
Точки чебышевского альтернанса, 179
Триангуляции, 558
Тригонометрическая интерполяция, 173
Узел
внутренний, 544
граничный, 544
нерегулярный, 551
приграничный, 544
регулярный, 551
сетки, 543
Узлы интерполяции, 36
Уравнения в конечных разностях, 51
Условие
α, 386
сильной минимальности, 565
Устойчивость, 503
безусловная, 532
по начальным данным, 535
условная, 532
Формула
Абеля, 467
Ромберга, 149
Симпсона, 88
Филона, 117

прямоугольников, 86
с кратными узлами, 96
трапеций, 87
Формулы
Адамса, 363, 378
Грегори, 143
численного дифференцирования, 76
Функция сильно растущая, 23
Характеристическое уравнение, 55
разностной схемы, 384
Хорошо
обусловленная (поставленная) краевая задача, 448
обусловленные системы, 307
Шаблон, 507
Шаг таблицы, 65
Экстраполяция, 36
Элемент наилучшего приближения, 165
Элементарный треугольник, 558
Энергетическое
неравенство, 538
тождество, 537
Явная схема, 376, 529
Явный метод, 376, 529

Предисловие

Первый вариант этой книги вышел в свет около тридцати лет назад, когда численные методы и их приложения переживали бурный рост, а специалисты в области численных методов были весьма уважаемы в обществе.

«Вычислители» старшего поколения, многие из которых, как и Николай Петрович Жидков, уже ушли из жизни, внесли неоценимый вклад в развитие научного и промышленного потенциала нашей страны, в развитие ее обороноспособности.

В связи с общедоступностью вычислительной техники настоящее время характерно снижением интеллектуального уровня задач, требующих внимания математиков, в частности специалистов в области численных методов.

Однако мы сохранили общий теоретический настрой книги, исходя из следующих соображений.

Теория численных методов, однажды возникнув, развивается по своим внутренним законам так же, как и другие фундаментальные разделы математики.

Специалисты в области теории численных методов и практики их применения будут неминуемо востребованы в процессе развития промышленности и науки.

Снижение общей математической образованности и общедоступность вычислительной техники делают необходимым создание комплексов программ, допускающих их использование исследователями невысокой математической квалификации. Разработка таких комплексов невозможна без дальнейшего развития теории численных методов.

Авторы признательны О. Б. Арушаняну, С. Ф. Залеткину, В. И. Лебедеву и Е. В. Чижонкову за ряд полезных замечаний, а также Н. П. Валединской, и особенно С. Г. Кобелькову и В. М. Староверову за большую помощь в оформлении оригинал-макета книги.

Введение

Попытаемся определить место теории численных методов в системе других областей знаний и рассказать о проблемах, возникающих в связи с ее применением, прежде чем переходить к непосредственному ее изложению.

Математика как наука возникла в связи с необходимостью решения практических задач: измерений на местности, навигации и т. д. Вследствие этого математика была численной математикой, ее целью являлось получение решения в виде числа.

Численное решение прикладных задач всегда интересовало математиков. Крупнейшие представители прошлого сочетали в своих исследованиях изучение явлений природы, получение их математического описания, как иногда говорят, математической модели явления, и его исследование. Анализ усложненных моделей потребовал создания специальных, как правило численных или асимптотических методов решения задач. Названия некоторых из таких методов — методы Ньютона, Эйлера, Лобачевского, Гаусса, Чебышева, Эрмита, Крылова — свидетельствуют о том, что их разработкой занимались крупнейшие ученые своего времени.

Настоящее время характерно резким расширением приложений математики, во многом связанным с созданием и развитием средств вычислительной техники. В результате появления ЭВМ (электронно-вычислительных машин или, как часто говорят, компьютеров) с программным управлением менее чем за пятьдесят лет скорость выполнения арифметических операций возросла от 0,1 операции в секунду при ручном счете до 10^{12} операций на современных серийных ЭВМ, т. е. примерно в 10^{13} раз.

Рост возможностей в связи с созданием вычислительной техники носит качественный характер и иногда сравнивается с промышленной революцией, вызванной изобретением паровой машины. Уместно вспомнить, что в итоге промышленной революции и последующего на протяжении двух веков развития науки и техники скорость передвижения возросла от скорости пешехода 6 км/ч до скорости космонавта 30 000 км/ч, т. е. в 5 000 раз.

Распространенное мнение о всемогуществе современных ЭВМ часто порождает впечатление, что математики избавились почти от всех хлопот, связанных с численным решением задач, и разработка новых методов для их решения уже не столь существенна. В действительности дело обстоит иначе, поскольку потребности эволюции, как правило, ставят

перед наукой задачи, находящиеся на грани ее возможностей. Расширение возможностей приложения математики обусловило математизацию химии, экономики, биологии, геологии, географии, психологии, экологии, метеорологии, медицины, конкретных разделов техники и др. Суть математизации состоит в построении математических моделей процессов и явлений и в разработке методов их исследования.

В физике или механике, например, построение математических моделей для описания различных явлений и изучение этих моделей с целью объяснения старых или предсказания новых эффектов являются традиционными.

Однако в целом работа в этом направлении зачастую продвигалась относительно медленно, поскольку обычно не удавалось получить решение возникающих математических задач и приходилось ограничиваться рассмотрением простейших моделей. Применение ЭВМ и расширение математического образования резко увеличило возможности построения и исследования математических моделей. Все чаще результаты расчетов позволяют обнаруживать и предсказывать ранее никогда не наблюдавшиеся явления; это дает основания говорить о математическом эксперименте. В некоторых исследованиях доверие к результатам численных расчетов так велико, что при расхождении между результатами расчетов и экспериментов в первую очередь ищут погрешность в результатах экспериментов.

Современные успехи в решении таких, например, проблем как атомные и космические вряд ли были бы возможны без применения ЭВМ и численных методов.

Требование численного решения новых задач привело к появлению большого количества новых методов. Наряду с этим последние полвека происходило интенсивное теоретическое переосмысливание и старых методов, а также систематизация всех методов. Эти теоретические исследования оказывают большую помощь при решении конкретных задач и играют существенную роль в наблюдаемом сейчас широком распространении сферы приложений ЭВМ и математики вообще.

Как уже отмечалось, с помощью современных ЭВМ удалось успешно решить ряд важных научно-технических задач. У непосвященного человека может возникнуть превратное впечатление, что успехи в применении ЭВМ обусловлены только повышением их быстродействия. Реально дело обстоит иначе и сложнее.

Правильнее будет сказать, что достижения в области использования ЭВМ обусловлены сочетанием ряда существенных факторов, без пропорционального развития которых они были бы много скромнее:

1) увеличение быстродействия ЭВМ, расширение памяти, совершенствование структуры ЭВМ, неуклонное снижение стоимости арифметической операции и единицы памяти;

2) разработка программных средств общения с ЭВМ, включающая создание операционных систем, языков программирования, библиотек и па-

кетов стандартных программ, снижение требований (в случае персональных ЭВМ) к математической и программистской культуре;

3) рост понимания процессов и явлений науки, техники, природы и общества и создание их математических моделей;

4) совершенствование методов решения традиционных математических и прикладных задач и создание методов решения новых задач;

5) рост понимания возможностей применения ЭВМ среди широких слоев общества; распространение так называемой компьютерной грамотности; координация усилий специалистов разного профиля по использованию вычислительной техники.

Достижения, перечисленные в пп. 3), 5), позволяют ответить на вопрос, какие задачи следует решать с помощью ЭВМ, и организовать их решение, в пп. 2), 4) — как их решать, и в пп. 1), 2) — дают для этого технические и программные средства.

Просмотр методов решения сложных прикладных задач показывает, что, как правило, эффект, достигаемый за счет совершенствования численных методов, по порядку сравним с эффектом, достигаемым за счет повышения производительности ЭВМ. Трудно сформулировать критерий, по которому можно было бы оценивать эффект применения новых численных методов, и еще труднее дать его достоверную количественную оценку. Все же, если сказать, что эффект от применения новых численных методов (при измерении эффекта в логарифмической шкале) при решении прикладных естественнонаучных задач дает 40% общего эффекта, достигаемого за счет применения новой вычислительной техники и новых численных методов, то эта оценка не будет завышенной.

Рассмотрим пример, иллюстрирующий это утверждение. Решение дифференциальных уравнений в частных производных сводится к решению систем линейных алгебраических уравнений с матрицей, в каждой строке которой имеется 5–10 ненулевых элементов. Накануне появления ЭВМ такие системы уравнений решали в случае числа неизвестных порядка $10-10^2$; сейчас нередки случаи, когда решаются системы с числом неизвестных порядка 10^5–10^6. В гипотетическом случае решения этих задач на современных ЭВМ методами, известными тридцать лет назад, пришлось бы ограничиться системами уравнений с числом неизвестных порядка 10^3–10^4 (при тех же затратах времени ЭВМ). Конечность скорости распространения сигнала — 300 000 км/с — ставит уже сейчас существенное ограничение на возможный рост быстродействия однопроцессорных ЭВМ, поэтому значение дальнейшего развития теории численных методов трудно переоценить. В частности, становится все более актуальной проблема разработки численных методов и программных средств для многопроцессорных ЭВМ.

Быстрое проникновение математики во многие области знания, в частности, объясняется тем, что математические модели и методы их исследования применимы сразу ко многим явлениям, сходным по своей формальной структуре. Часто математическая модель, описывающая какое-

либо явление, появляется при изучении других явлений или при абстрактных математических построениях задолго до конкретного рассмотрения данного явления. В частности, и в теории численных методов, так же как в «чистой» математике, полезна разработка общих построений. Однако есть разница в подходе «чистого» и «прикладного» математика к решению какой-либо проблемы. На языке первого понятие «решить задачу» означает доказать существование решения и предложить процесс, сходящийся к решению. Сами по себе эти результаты полезны для прикладника, но, кроме этого, ему нужно, чтобы процесс получения приближения не требовал больших затрат, например времени или памяти ЭВМ. Ему важно не только то, что процесс сходится, но и то, как быстро он сходится. При численном решении задач возникают также новые вопросы, связанные с устойчивостью результата относительно возмущений исходных данных и округлений при вычислениях.

Наряду с теорией численных методов период бурного развития переживает и ряд других разделов математики, непосредственно обязанных ЭВМ своим возникновением. Применение численных методов и ЭВМ к решению естественнонаучных задач оказывает влияние и на традиционные разделы математики.

Математика возникла и развивается как часть естествознания, и долгое время ее развитие существенным образом определялось потребностями физики и механики. Требование математизации новых разделов науки неизбежно приводит к обратному влиянию этих разделов на развитие математики и должно существенно изменить лицо самой математики.

Развитие как теоретических, так и прикладных разделов математики в конечном счете определяется потребностями общества и его материальным вкладом в развитие науки, в частности в образование. Несколько десятилетий назад отношение вложений в науку к общим вложениям в народное хозяйство составляло доли процента. Сейчас в индустриально развитых странах это отношение настолько велико, что его дальнейший существенный рост невозможен. Поэтому происходит перераспределение вложений в различные направления науки. Это обуславливает еще один канал влияния прикладной стороны математики на развитие ее теоретических разделов. Прикладные исследования имеют непосредственную отдачу; это усиливает доверие общества к математике, расширяет понимание ее проблем и как следствие способствует увеличению вложения средств с целью ее развития.

При реальной работе в области приложений математики возникает большое количество осложнений самого различного, зачастую нематематического характера.

Хотя трудно надеяться, что какие-либо теоретические нравоучения могут заменить собственный опыт работы, попытаемся обратить внимание на некоторые вопросы общего характера, важные для работы в области приложений математики. Проводимая ниже систематизация этих вопросов является довольно случайной, условной; по-видимому, можно предло-

жить еще добрый десяток подобных классификаций, имеющих не меньшее право на существование.

1. Первостепенное значение имеет выбор направления исследования. Свобода выбора обычно довольно невелика, так как основные контуры направления исследования обычно задаются «извне».

При выборе направления исследования в пределах имеющихся возможностей полезно иметь в виду следующее «правило трех частей», по своему внешнему виду похожее на шутку. Проблемы делятся на: I — легкие, II — трудные, III — очень трудные. Проблемами I заниматься не стоит, они будут решены в ходе событий и без вашего вмешательства, проблемы III вряд ли удастся решить в настоящее время, поэтому стоит обратиться к проблемам II.

2. Нужно уметь сформулировать на языке математики конкретные задачи физики, механики, экономики, инженерные задачи и т. д., т. е. построить математическую модель рассматриваемого явления.

В теоретической науке исследователь, умеющий правильно формулировать, как говорят, ставить новые задачи, как правило ценится выше, чем исследователь, умеющий решать кем-то поставленные задачи. Еще более возрастает роль таких ученых в прикладной науке.

Начинающий работу математик часто жалуется на трудности контактов с представителями других наук, которые «даже» не могут сформулировать стоящих перед ними задач. Правильное формулирование задачи — это научная проблема, не менее сложная, чем само решение задачи, и не нужно надеяться, что кто-то другой целиком сделает это за вас. При постановке проблемы первостепенное внимание должно быть уделено выяснению цели исследования; принимаемая математическая модель явления не есть что-то однозначное, раз навсегда связанное с этим явлением, а зависит от цели исследования. Прежде чем выписывать дифференциальные уравнения, выбирать метод решения и обращаться к ЭВМ, стоит подумать, а не будут ли бесполезны все результаты вычислений? В то же время надо воспринимать как должное, что большая часть результатов вычислений будет выброшена сразу же после их получения. Дело в том, что производимая работа зачастую носит исследовательский характер и трудно заранее предсказать, что и в какой форме следует получить, на каком пути нужно искать численное решение задачи. Цель исследования и описание проблемы обычно уточняются в процессе контактов представителей конкретных наук или руководства организаций (заказчиков) и математиков (исследователей или исполнителей).

3. Успех в прикладной науке требует широкой математической подготовки, поскольку только такая подготовка может обеспечить приспособляемость к непрерывно меняющимся типам задач, предъявляемых к решению. Одной из причин необходимости изучения на первый взгляд «бесполезных» для практики разделов математики является достижение более

уверенного и более свободного владения «нужными» разделами математики.

При построении и анализе математических моделей привычка математика «докапываться до конца», подвергать все сомнению, обусловленная его строгим математическим образованием, часто не менее важна, чем интуиция и соображения здравого смысла. Типичное для человека с математическим образованием стремление к общности охвата различных явлений часто помогает выделить наиболее существенные черты явления и отбросить второстепенные.

4. Не следует думать, что совершенное знание математики, численных методов и навыки работы с ЭВМ позволяют сразу решить любую прикладную математическую задачу. Во многих случаях требуется «доводка» методов, приспособление их к решению конкретных задач. При этом типична обстановка, когда используются методы, применение которых теоретически не обосновано, или теоретические оценки погрешности численного метода неприемлемы для практического использования вследствие их громоздкости; при выборе метода решения задачи и анализе результатов приходится полагаться на опыт предшествующего решения задач, на интуицию и сравнение с экспериментом и при этом приходится отвечать за достоверность результата. Поэтому для успеха в работе необходимы развитое неформальное мышление, умение рассуждать по аналогии, дающие основания ручаться за достоверность результата там, где с позиций логики и математики, вообще говоря, ручаться нельзя.

В рассматриваемом вопросе есть и другая сторона. При численном решении конкретных трудных задач, возникающих в других областях знаний, математик действует как естествоиспытатель, полагаясь во многом лишь на опыт и «правдоподобные» рассуждения. Крайне желательно, чтобы такая эмпирическая работа подкреплялась теоретическими разработками методов, аккуратной проверкой качества методов на контрольных задачах с известным решением или частным сравнением с экспериментом. При длительном продвижении в каком-то направлении без такого подкрепления может теряться перспектива работы, уверенность в правильности получаемых результатов. Известное высказывание, что хороший теоретик может истолковать в желаемом направлении любые результаты как расчетов, так и эксперимента, содержит большую долю истины.

5. После завершения расчетов наступает этап использования результатов вычислений в практической деятельности, или, как часто говорят, этап внедрения результатов. Правильнее будет сказать, что подготовка к использованию результатов начинается уже с анализа постановки задачи и в процессе ее решения и, по существу, все моменты решения задачи и внедрения результатов неразрывно связаны между собой; в процессе формулирования задачи и ее решения заказчик и исполнитель взаимно уточняют постановку задачи и тем самым подготавливают почву

для приложения полученных результатов. Поскольку математика в сочетании с ЭВМ используется в самых разнообразных областях, то часто приходится иметь дело с заказчиками, не имеющими опыта применения ЭВМ. В процессе контакта с такими «начинающими» заказчиками особенно важно преодолеть их первоначальное недоверие к вторжению математики в их области исследования; результаты вычислений будут использоваться только тогда, когда заказчик осмыслит их со своих позиций и убедится в том, что их действительно можно и нужно использовать. При правильном подходе к взаимным контактам к концу процесса решения задачи «начинающий» заказчик приходит к пониманию, что ЭВМ и математика могут дать ему не все, но довольно много, а «начинающий» математик — к пониманию того, что он дает заказчику кое-что, но далеко не все нужное для реального решения задачи.

Большое значение имеет наглядность, доступность представления заказчику промежуточных и окончательных результатов исследования: таблицы, графики, вывод информации на экран: нельзя предполагать наличия или требовать от заказчика большего объема знаний, чем это требуется существом дела. Целесообразнее, чтобы биолог использовал свое умение дифференцировать для построения и исследования математической модели, а не для оценки погрешности метода численного интегрирования.

Математик должен принять во внимание образование и психологию людей, применяющих разработанные им методы и программы. Например, простейшая программа численного интегрирования, предназначенная для широкого круга нематематиков, использующих ЭВМ в своих конкретных исследованиях, должна быть рассчитана на человека, потолок математических знаний которого находится на интуитивном понимании того, что интеграл — это площадь. Чтобы не затруднять пользователя, в описании простейших программ даже ничего не говорится о точности результата. Предполагается, что пользователя удовлетворит невысокая точность результата, и программа реализуется, например, так, чтобы в большинстве случаев относительная погрешность результата не превосходила 1% (так называемая графическая точность).

6. Существенным моментом в прикладной работе является необходимость получения результатов в установленный срок. Заказчик, для которого проводятся исследования, расчеты, часто ограничен сроком завершения исследований и принятия решения на их основе. Если исследования не будут завершены к сроку, то решение все равно будет принято, но на основе более грубого, эмпирического или просто «волевого» подхода. Потерянное в таком случае доверие со стороны заказчика часто бывает невозможно восстановить.

В такой ситуации лучше найти по возможности удовлетворительное решение задачи, но в срок, чем получить полное решение задачи к тому времени, когда оно станет бесполезным. Поэтому, в частности, целесо-

образно начинать исследование новых задач с рассмотрения простейших моделей, применяя при численном решении испытанные методы.

7. Также существенным моментом в прикладной работе является то обстоятельство, что работа, как правило, проводится коллективом. Одна из причин этого состоит в том, что построение математической модели, выбор метода решения, непосредственное общение с ЭВМ и анализ результатов требуют различных знаний и квалификации. Другая причина кроется в упомянутой уже необходимости решения задачи в установленный срок. Это требование приводит к необходимости распараллеливания даже однотипной работы между большим числом исполнителей, например путем независимого написания различных блоков программы отдельными исполнителями. Параллельно могут идти отработка различных методов на модельных задачах, обсчет упрощенных моделей, подготовительная работа по написанию окончательной программы решения задачи.

Можно привести много реальных примеров неудачного решения больших вычислительных задач и работ по созданию программного обеспечения, вызванных следующей причиной. Распределение обязанностей между исполнителями не было в достаточной степени формализовано, т. е. не было выдано однозначного описания окончательного результата работы каждого исполнителя. В результате или основная доля времени уходила на непрерывное согласование отдельных частей работы, или после истечения существенного промежутка времени оказывалось, что эти части работы не стыкуются. Поэтому организаторские способности ученого, осуществляющего общее руководство решением задачи, зачастую не менее важны, чем его математические способности.

Приведенные выше рассуждения в определенной степени иллюстрируют специфику работы в области прикладной математики и показывают, что специалисты в этой области кроме широкой математической эрудиции должны обладать также другими важными свойствами человеческого интеллекта и характера.

Глава 1

Погрешность результата численного решения задачи



В этой главе объясняются источники возникновения погрешности решения задачи, даются основные правила задания приближенных величин и оценивается погрешность как простейших, так и более сложных функций от приближенно заданных величин.

В дальнейшем конкретные оценки этой главы по существу не используются, но сам разговор о них необходим для понимания реальной обстановки, в которой используются рассматриваемые в книге методы решения задач.

§ 1. Источники и классификация погрешности

Погрешность решения задачи обуславливается следующими причинами:

1) математическое описание задачи является неточным, в частности неточно заданы исходные данные описания;

2) применяемый для решения метод часто не является точным: получение точного решения возникающей математической задачи требует неограниченного или неприемлемо большого числа арифметических операций; поэтому вместо точного решения задачи приходится прибегать к приближенному;

3) при вводе данных в машину, при выполнении арифметических операций и при выводе данных производятся округления.

Погрешности, соответствующие этим причинам, называют:

1) *неустранимой погрешностью*,

2) *погрешностью метода*,

3) *вычислительной погрешностью*.

Часто неустранимую погрешность подразделяют на две части:

а) *неустранимой погрешностью* называют лишь погрешность, являющуюся следствием неточности задания числовых данных, входящих в математическое описание задачи;

б) погрешность, являющуюся следствием несоответствия математического описания задачи реальности, называют, соответственно, *погрешностью математической модели*.

Дадим иллюстрацию этих определений. Пусть у нас имеется маятник (рис. 1.1.1)[1)], начинающий движение в момент $t = t_0$. Требуется предсказать угол отклонения φ от вертикали в момент t_1.

Дифференциальное уравнение, описывающее колебание этого маятника, берется в виде

$$l\frac{d^2\varphi}{dt^2} + g\sin\varphi + \mu\frac{d\varphi}{dt} = 0, \tag{1}$$

где l — длина маятника, g — ускорение силы тяжести, μ — коэффициент трения.

Как только принимается такое описание задачи, решение уже приобретает неустранимую погрешность, в частности, потому, что реальное трение зависит от скорости не совсем линейно; другой источник неустранимой погрешности состоит в погрешностях определения l, g, μ, t_0 ,$\varphi(t_0)$, $\varphi'(t_0)$. Название этой погрешности — «неустранимая» — соответствует ее существу: она неконтролируема в процессе численного решения задачи и может уменьшиться только за счет более точного описания физической задачи и более точного определения параметров. Дифференциальное уравнение (1) не решается в явном виде; для его решения требуется применить какой-либо численный метод. Вследствие этой причины и возникает погрешность метода.

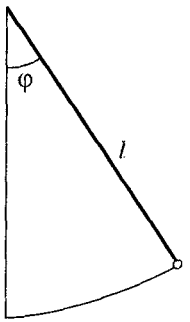


Рис. 1.1.1

Вычислительная погрешность может возникнуть, например, из-за конечности количества разрядов чисел, участвующих в вычислениях.

Введем формальные определения.

Пусть I — точное значение отыскиваемого параметра (в данном случае — реальный угол отклонения маятника φ в момент времени t_1), $\tilde{I}$ — значение этого параметра, соответствующее принятому математическому описанию (в данном случае — значение $\varphi(t_1)$ решения уравнения (1)), $\tilde{I}_h$ — решение задачи, получаемое при реализации численного метода в предположении отсутствия округлений, $\tilde{I}_h^*$ — приближение к решению задачи, получаемое при реальных вычислениях. Тогда

$$\begin{aligned} \rho_1 &= \tilde{I} - I && \text{— неустранимая погрешность,} \\ \rho_2 &= \tilde{I}_h - \tilde{I} && \text{— погрешность метода,} \\ \rho_3 &= \tilde{I}_h^* - \tilde{I}_h && \text{— вычислительная погрешность.} \end{aligned} \tag{2}$$

[1)] Тройная нумерация рисунков и формул указывает главу, параграф, номер формулы или рисунка; двойная, применяемая только для формул, — параграф и номер (в данной главе); одинарная, применяемая также только для формул, — только номер (в данном параграфе).

Полная погрешность $\rho_0 = \tilde{I}_h^* - I$, равная разности между реально получаемым и точным решениями задачи, удовлетворяет равенству

$$\rho_0 = \rho_1 + \rho_2 + \rho_3. \tag{3}$$

Во многих случаях под термином *погрешность* того или иного вида понимают не рассмотренные выше разности между приближениями, а некоторые меры близости между ними. Например, в скалярном случае полагают

$$\rho_0 = \left|\tilde{I}_h^* - I\right|, \ \rho_1 = \left|\tilde{I} - I\right|, \ \rho_2 = \left|\tilde{I}_h - \tilde{I}\right|, \ \rho_3 = \left|\tilde{I}_h^* - \tilde{I}_h\right|;$$

при таких обозначениях вместо (3) получаем

$$\rho_0 \leqslant \rho_1 + \rho_2 + \rho_3. \tag{4}$$

В других случаях решение I и приближения $\tilde{I}$, $\tilde{I}_h$, $\tilde{I}_h^*$ оказываются элементами некоторых функциональных пространств, часто различных. Например, $\tilde{I}$ может быть элементом пространства F непрерывных на $[0, 1]$ функций, а $\tilde{I}_h$ — элементом пространства F_h сеточных функций f_h, определенных в точках $x_n = nh$, $n = 0, 1, \ldots, h^{-1}$; h^{-1} — целое. Тогда в качестве меры погрешности вводят некоторую меру близости $\rho(z_1, z_2)$, где z_1 и z_2 могут быть элементами как одного, так и различных пространств. Требования на эту меру близости — возможность принять ее за естественную меру погрешности и выполнение неравенства треугольника

$$\rho(z_1, z_3) \leqslant \rho(z_1, z_2) + \rho(z_2, z_3) \tag{5}$$

при любых $z_1, z_2, z_3 \in F, F_h$. При этом не накладывается условие: если $\rho(z_1, z_2) = 0$, то $z_1 \equiv z_2$; таким образом, функция $\rho(z_1, z_2)$ не обязательно является расстоянием в некотором метрическом пространстве.

Например, можно положить

$$\rho(f_1, f_2) = \max_{0 \leqslant n \leqslant h^{-1}} |f_1(nh) - f_2(nh)|$$

независимо от того, каким пространствам принадлежат f_1 и f_2.

Может возникнуть такой вопрос по поводу проблемы исследования неустранимой погрешности: зачем изучать неустранимую погрешность решения задачи, если она «неустранима» ? По крайней мере такая точка зрения кажется оправданной, если математик получает для численного решения задачи уже готовые уравнения, не участвуя в обсуждении физической постановки задачи.

Это возражение нельзя признать разумным. Часто математик сам занимается исследованием постановки задачи, анализом и упрощением рассматриваемых уравнений. Поскольку все явления в природе взаимосвязаны, в принципе невозможно математически точно описать никакой реальный процесс, происходящий в природе. Однако анализ влияния различных факторов на погрешность решения может позволить получить

простейшее описание процесса с допустимой погрешностью. Обычно математик имеет представление о требуемой окончательной точности результата, и, исходя из этого, он может производить необходимые упрощения исходной задачи.

Если математик не участвует в обсуждении физической постановки задачи, то представление о величине неустранимой погрешности ему все равно необходимо по следующей причине. При решении большинства задач нет особого смысла применять метод решения задачи с погрешностью, существенно меньшей, чем величина неустранимой погрешности. Поэтому, имея представление о величине неустранимой погрешности, можно разумно сформулировать требования к точности результата численного решения задачи.

Непомерные требования заказчика к точности результата часто вызваны тем, что он имеет преувеличенные представления о возможностях ЭВМ и поэтому серьезно не продумывает, что все-таки ему нужно.

Такие требования часто снимаются в процессе обсуждения задачи на основе следующих соображений:

1) при более детальном подходе к изучению задачи в целом оказывается, что столь высокая точность и не нужна;

2а) математическая модель явления настолько груба, что требовать столь высокую точность бессмысленно;

2б) параметры модели не могут быть определены с высокой точностью;

3) заказчику нужен вообще не количественный, а качественный результат, например такого типа: будет ли работать данное устройство в заданном режиме или нет.

Разберем некоторые встретившиеся нам реальные задачи. К решению была предъявлена система интегральных уравнений с сильно осциллирующими ядрами с числом перемен у ядер порядка $\lambda^{-1} = N = 10^6$. Требовалось получить решение с относительной погрешностью (определение см. далее) порядка 10^{-6}. Эта система описывала режим работы некоторого оптического устройства. Решение такой системы интегральных уравнений непомерно сложно даже для современных ЭВМ, поэтому был предпринят ее подробный анализ. Оказалось, что относительные погрешности характеристик системы, обусловленные технологией изготовления устройства, являются величинами порядка 10^{-4}, поэтому нет смысла решать задачу со столь высокой точностью, как требовалось вначале. В результате требования к точности искомого решения были снижены до относительной погрешности 10^{-4}. Однако и такая точность все равно еще требовала непомерных затрат машинного времени.

Дальнейший анализ задачи показал, что по существу заказчика интересовал ответ только на один вопрос — будет ли данное устройство устойчиво функционировать или нет?

Естественно было предположить две возможности:

1) при малых значениях параметра $\lambda = 1/N$ решение плавно зависит от этого

параметра, поэтому при λ, меньших достаточно малого λ_0, система будет работать в одном режиме — или всегда устойчиво, или всегда неустойчиво;

2) при малых значениях λ решение существенно меняется при изменении этого параметра, и интервалы значений параметра λ, где режим работы устойчив, перемежаются с интервалами значений, где режим работы неустойчив.

В связи с этим были предприняты расчеты при довольно крупном значении $\lambda = 1/10$ с последующим уменьшением значений λ с тем, чтобы понять, какая из двух указанных выше возможностей реализуется.

Расчеты показывали, что при изменении λ в пределах от 10^{-1} до 10^{-2} имел место устойчивый режим работы устройства и был сделан вывод (подтвержденный потом экспериментально, после конструирования реального устройства) об устойчивости его работы при $\lambda = 10^{-6}$. В случае второй возможности, вследствие грубости изготовления устройства, вряд ли удалось бы вообще исследовать вопрос об устойчивости его работы при $\lambda = 10^{-6}$.

Другой, на первый взгляд выглядящий курьезным, но на самом деле весьма типичный пример реальной ситуации. Перед математиками была поставлена задача создания алгоритма и программы быстрого (менее чем за 1 с машинного времени) вычисления интегралов специального вида с относительной погрешностью 10^{-4}. Эта задача была ими успешно решена, т. е. был разработан метод вычисления таких интегралов и на его основе создана стандартная программа. В свою очередь исследователи, поставившие задачу, не скупясь на затраты машинного времени, для проверки качества предложенного математиками метода и надежности программы сами вычислили приближенно один из таких интегралов с относительной погрешностью, по их мнению, 10^{-6}. Но оказалось, что все попытки решить эту, так называемую *тестовую* задачу с погрешностью, лучшей, чем 10^{-2}, с помощью созданной математиками программы оканчивались неудачей. Возникло предположение о погрешности в самой тестовой задаче. Оказалось, что число π было взято равным 3,14, что вносило в тестовый пример *неустранимую* погрешность, которая, естественно, не могла быть устранена никакими усилиями математиков, создававших алгоритм и программу.

§ 2. Запись чисел в ЭВМ

Современные ЭВМ оперируют с числами, записанными в одной из приведенных ниже форм.

Первая форма записи — с фиксированной запятой: все числа в ЭВМ имеют модуль, меньший 1; число знаков после запятой фиксировано. Таким образом, машина оперирует с числами

$$x = \pm \sum_{k=1}^{t} \alpha_k q^{-k} = \pm(\alpha_1, \ldots, \alpha_t); \qquad (1)$$

здесь q — целое — основание системы счисления, $\alpha_1, \ldots, \alpha_t$ — целые в пределах $0 \leqslant \alpha_k < q$.

При операциях над числами x с $|x| < 1$ могут появляться числа y с $|y| \geqslant 1$, и тогда произойдет остановка работы ЭВМ («машинный останов» или «АВОСТ»). Чтобы избежать этого, производится масштабирование задачи — введение новых масштабов. Иногда заранее нельзя указать нужные масштабы; в других случаях введение очень больших масштабов с самого начала приведет к тому, что в исходных данных большое количество первых из α_i обратится в нуль и произойдет существенная потеря информации. Поэтому часто предусматривают изменение масштабов уже в процессе решения задачи.

Вторая форма записи, наиболее распространенная в ЭВМ, предназначенных для научных расчетов, — с плавающей запятой: машина оперирует с числами

$$x = \pm q^p \sum_{k=1}^{t} \alpha_k q^{-k} = \pm q^p (\alpha_1, \ldots, \alpha_t); \qquad (2)$$

порядок числа p удовлетворяет неравенству $|p| \leqslant p_0$.

Наиболее распространен случай двоичной системы счисления, когда $q = 2$.

При работе в режиме с плавающей запятой пользователь получает дополнительные удобства, так как не надо заботиться о масштабах; однако при этом происходит некоторое замедление работы ЭВМ.

§ 3. Абсолютная и относительная погрешности. Формы записи данных

Если a — точное значение некоторой величины, а a^* — известное приближение к нему, то *абсолютной погрешностью* приближенного значения a^* называют обычно некоторую величину $\Delta(a^*)$, про которую известно, что

$$|a^* - a| \leqslant \Delta(a^*).$$

Относительной погрешностью приближенного значения называют некоторую величину $\delta(a^*)$, про которую известно, что

$$\left|\frac{a^* - a}{a^*}\right| \leqslant \delta(a^*).$$

Относительную погрешность часто выражают в процентах.

Если a — известное число, например π, то иногда говорят об абсолютной $\Delta(a)$ и относительной $\delta(a)$ погрешностях задания этого числа: числа $\Delta(a)$ и $\delta(a)$ называют соответственно абсолютной и относительной погрешностью числа a, если про них известно, что $|a^* - a| \leqslant \Delta(a)$, $\left|\dfrac{a^* - a}{a}\right| \leqslant \delta(a)$.

Иногда в литературе абсолютной погрешностью называют величину $a^* - a$, а относительной — величину $\dfrac{a^* - a}{a^*}$; мы будем придерживаться исходных определений, и поэтому у нас всегда $0 \leqslant \Delta(a^*),\ \delta(a^*)$.

По ходу изложения материала будут употребляться выражения: *большое число, очень большое число, сильный рост функции.*

Чаще всего число x мы называем *большим*, если $|x| \gg 1$, но относительная погрешность результата решения задачи порядка $|x|2^{-t}$ является допустимой.

Если относительная погрешность порядка $|x|2^{-t}$ является недопустимо большой, то число x называем *очень большим.*

Выражение *функция сильно растет* чаще всего означает, что она возрастает в очень большое число раз.

Значащими цифрами числа называют все цифры в его записи, начиная с первой ненулевой слева.

Пример. У чисел $a^* = 0{,}0\underline{3045}$, $a^* = 0{,}0\underline{3045000}$ значащими цифрами являются подчеркнутые цифры. Число значащих цифр в первом случае равно 4, во втором — 7.

Значащую цифру называют *верной*, если абсолютная погрешность числа не превосходит единицы разряда, соответствующего этой цифре.

Примеры. $a^* = 0{,}0\underline{3045}$, $\Delta(a^*) = 0{,}000003$; $a^* = 0{,}0\underline{3045000}$, $\Delta(a^*) = 0{,}0000007$; подчеркнутые цифры являются верными.

Иногда уславливаются называть значащую цифру *верной*, если абсолютная погрешность не превосходит половины единиц разряда, соответствующих этой цифре.

Если все значащие цифры верные, то говорят, что число записано *со всеми верными цифрами.*

Пример. При $a^* = 0{,}0\underline{3045}$, $\Delta(a^*) = 0{,}000003$ число a^* записано со всеми верными цифрами.

Иногда употребляется термин *число верных цифр после запятой*: подсчитывается число цифр после запятой от первой цифры до последней верной цифры. В последнем примере это число равно 5.

Довольно часто информация о некоторой величине задается пределами ее измерения:

$$a_1 \leqslant a \leqslant a_2, \tag{1}$$

например

$$1{,}119 \leqslant a \leqslant 1{,}127.$$

Принято записывать эти пределы с одинаковым числом знаков после запятой; так как обычно достаточно грубого представления о погрешности, то в числах a_1, a_2 часто берут столько значащих десятичных цифр,

сколько нужно, чтобы разность $a_1 - a_2$ содержала одну-две значащие цифры.

Употребляемые далее оговорки «часто», «обычно», «принято» специально употребляются нами, чтобы не создавалось впечатления об обязательности каких-то стандартных форм задания информации о величине погрешности. Эти формы задания информации рассматриваются лишь потому, что они наиболее распространены, а следовательно, наиболее удобны при контактах.

Абсолютную или относительную погрешность обычно записывают в виде числа, содержащего одну или две значащих цифры.

Информацию о том, что a^* является приближенным значением числа a с абсолютной погрешностью $\Delta(a^*)$, иногда записывают в виде

$$a = a^* \pm \Delta(a^*); \tag{2}$$

числа a^* и $\Delta(a^*)$ принято записывать с одинаковым числом знаков после запятой. Например, записи

$$a = 1{,}123 \pm 0{,}004, \quad a = 1{,}123 \pm 4 \cdot 10^{-3}$$

относятся к общепринятым и означают, что

$$1{,}123 - 0{,}004 \leqslant a \leqslant 1{,}123 + 0{,}004.$$

Соответственно информацию о том, что a^* является приближенным значением числа a с относительной погрешностью $\delta(a^*)$, записывают в виде

$$a = a^*(1 \pm \delta(a^*)). \tag{3}$$

Например, записи

$$a = 1{,}123(1 \pm 0{,}003),\ a = 1{,}123(1 \pm 3 \cdot 10^{-3}),\ a = 1{,}123(1 \pm 0{,}3\%)$$

означают, что

$$(1 - 0{,}003)1{,}123 \leqslant a \leqslant (1 + 0{,}003)1{,}123.$$

При переходе от одной из форм записи к другой надо следить, чтобы пределы измерения, указываемые новой формой записи, были шире старых, иначе такой переход будет незаконным. Например, при переходе от (1) к (2) должны выполняться неравенства

$$a^* - \Delta(a^*) \leqslant a_1, \quad a_2 \leqslant a^* + \Delta(a^*),$$

при переходе от (2) к (3) — неравенства

$$a^*(1 - \delta(a^*)) \leqslant a^* - \Delta(a^*), \quad a^* + \Delta(a^*) \leqslant a^*(1 + \delta(a^*)),$$

при переходе от (3) к (2) должны выполняться противоположные неравенства (пределы всегда расширяются!).

Следует различать принятую нами выше формально математическую и обиходную терминологии в рассуждении о величине погрешности. Если

в постановке задачи говорится, что требуется найти решение с погрешностью 10^{-2}, то чаще всего не имеется в виду обязательность этого требования. Предполагается лишь, что погрешность имеет такой порядок. Если, например, решение будет найдено с погрешностью $2 \cdot 10^{-2}$, то такой результат, скорее всего, также удовлетворит заказчика.

§ 4. О вычислительной погрешности

Ограничение на порядки чисел в ЭВМ $|p| \leqslant p_0$ иногда приводит к прекращению вычислений; в других случаях относительно небольшая разрядность чисел в ЭВМ приводит к недопустимому искажению результата вычислительной погрешностью. Такие алгоритмы, где вследствие ограниченности p или малости t возникают подобные эффекты, называют «неустойчивыми».

Построение «устойчивых» алгоритмов, при использовании которых искажение окончательного результата вычислительной погрешностью находится в допустимых пределах, составляют существенную часть теории численных методов.

Рассмотрим пример, показывающий, что повышение точности иногда может быть достигнуто за счет несложного алгебраического преобразования.

Пусть отыскивается наименьший корень уравнения $y^2 - 140y + 1 = 0$. Для определенности условимся о следующих правилах округления. Вычисления производятся в десятичной системе счисления, причем в мантиссе числа после округлений удерживается 4 разряда. Имеем

$$y = 70 - \sqrt{4899}, \quad \sqrt{4899} = 69{,}992\ldots;$$

после округления получаем

$$\sqrt{4899} \approx 69{,}99, \quad y \approx 70 - 69{,}99 = 0{,}0\underline{1}.$$

То же самое значение y можно, «избавившись от иррациональности в числителе», представить в виде $y = 1/(70 + \sqrt{4899})$. Последовательно производя вычисления, получаем $\sqrt{4899} \approx 69{,}99$; $70 + 69{,}99 = 139{,}99$ и после округления

$$70 + 69{,}99 \approx 140{,}0.$$

Наконец,

$$1/140 = 0{,}00714285\ldots,$$

и после округления

$$y \approx 0{,}00\underline{7143}.$$

Производя вычисления с дополнительными разрядами, можно проверить, что в обоих случаях все подчеркнутые цифры результатов верные; однако во втором случае точность результата существенно выше. Дело в том, что в первом случае пришлось вычитать близкие большие числа;

так как эти числа были большие, то они были округлены с большой абсолютной погрешностью, в результате и ответ получился с большой абсолютной погрешностью. Здесь нам впервые встретилось явление *потери значащих цифр* (или *«пропадания» значащих цифр*), имеющее место при вычитании близких величин; это явление, например, довольно часто приводит к существенному искажению результата при решении систем линейных алгебраических уравнений.

Рассмотрим другой типичный пример, где порядок выполнения операций влияет на погрешность результата. На машине с плавающей запятой вычисляется значение суммы

$$S_{1\,000\,000} = \sum_{j=1}^{1\,000\,000} \frac{1}{j^2}.$$

Можно вычислять $S_{1\,000\,000}$ либо по рекуррентной формуле

$$S_n = S_{n-1} + \frac{1}{n^2}, \quad n = 1, \ldots, 1\,000\,000, \quad S_0 = 0,$$

либо по рекуррентной формуле

$$\sum_{n-1} = \sum_{n} + \frac{1}{n^2}, \quad n = 1\,000\,000, \ldots, 1, \quad \sum_{1\,000\,000} = 0, \quad \sum_{0} = S_{1\,000\,000}.$$

Оказывается, во втором случае суммарная вычислительная погрешность будет существенно меньше.

Дело заключается в следующем. В большинстве случаев сложение чисел в ЭВМ осуществляется по следующей схеме. Два числа a и b складываются абсолютно точно, а затем происходит отбрасывание последних знаков или округление результата с тем, чтобы осталось t или $t-1$ значащих цифр. В результате получается приближенное значение суммы $a+b$ с погрешностью, не превосходящей $2^{-t}|a+b|$, но в неблагоприятном случае большей, чем $2^{-1-t}|a+b|$. В первом случае у нас при каждом сложении значение суммы больше 1 и в принципе возможно получение погрешности около $10^6 \cdot 2^{-t}$. Во втором случае

$$\sum_{n} = O\left(\frac{1}{n}\right),$$

поэтому погрешность накапливается существенно медленнее. Можно показать, что что погрешность окончательного результата не превосходит $100 \cdot 2^{-t}$.

На конкретной ЭВМ было проведено вычисление $S_{1\,000\,000}$ по обоим алгоритмам и оказалось, что для первого алгоритма погрешность $\approx 2 \cdot 10^{-4}$, а для второго — $\approx 6 \cdot 10^{-8}$.

Заметим, что в настоящее время проблемы, возникающие в такого рода простейших задачах, обходятся за счет вычислений с двойной точностью.

§ 5. Погрешность функции

Довольно часто возникает следующая задача. Искомая величина y является функцией от параметров $a_1, \ldots, a_n$: $y = y(a_1, \ldots, a_n)$. Известна область G в пространстве переменных $a_1, \ldots, a_n$, которой принадлежат эти параметры. Требуется получить приближение к y и оценить его погрешность.

Если y^* — приближенное значение величины y, то *предельной абсолютной погрешностью* $A(y^*)$ называют наилучшую при имеющейся информации оценку погрешности величины y^*; согласно этому определению в данном случае

$$A(y^*) = \sup_{(a_1,\ldots,a_n)\in G} |y(a_1, \ldots, a_n) - y^*|; \tag{1}$$

предельной относительной погрешностью называют величину $\dfrac{A(y^*)}{|y^*|}$.

Задача 1. Доказать, что предельная абсолютная погрешность $A(y^*)$ минимальна при $y^* = (Y_1 + Y_2)/2$, где $Y_1 = \inf\limits_G y(a_1, \ldots, a_n)$, $Y_2 = \sup\limits_G y(a_1, \ldots, a_n)$.

Рассмотрим наиболее распространенный случай, когда область G — прямоугольник:

$$|a_j - a_j^*| \leqslant \Delta(a_j^*), \quad j = 1, \ldots, n,$$

и за приближенное значение принимается

$$y^* = y(a_1^*, \ldots, a_n^*).$$

Если y — непрерывно дифференцируемая функция своих аргументов, то, согласно формуле Лагранжа для функций многих переменных,

$$y(a_1, \ldots, a_n) - y^* = \sum_{j=1}^{n} b_j(\theta)(a_j - a_j^*), \tag{2}$$

где

$$b_j(\theta) = \frac{\partial y}{\partial a_j}\Big|_{(a_1^*+\theta(a_1-a_1^*),\ldots,a_n^*+\theta(a_n-a_n^*))}, \quad 0 \leqslant \theta \leqslant 1.$$

Отсюда следует оценка погрешности

$$|y(a_1, \ldots, a_n) - y^*| \leqslant A_0(y^*) = \sum_{j=1}^{n} B_j \Delta(a_j^*), \tag{3}$$

где

$$B_j = \sup_G \left|\frac{\partial y(a_1, \ldots, a_n)}{\partial a_j}\right|.$$

Положим

$$\rho = \sqrt{\sum_{j=1}^{n} \bigl(\Delta(a_j^*)\bigr)^2}.$$

Если производные $\dfrac{\partial y(a_1, \ldots, a_n)}{\partial a_j}$ непрерывны, то $b_j(\theta) = b_j(0) + o(1)$ и

$$B_j = \left| \frac{\partial y(a_1^*, \ldots, a_n^*)}{\partial a_j} \right| + o(1).$$

Здесь выражение $x = y + o(1)$ понимается в смысле: $x - y \to 0$ при $\rho \to 0$. Следовательно, $A_0(y^*) = A^0(y^*) + o(\rho)$, где

$$A^0(y^*) = \sum_{j=1}^{n} \left| \frac{\partial y(a_1^*, \ldots, a_n^*)}{\partial a_j} \right| \Delta(a_j^*).$$

При практической работе вместо оценки погрешности (3) обычно пользуются более простой, вообще говоря неверной, «оценкой»

$$|y(a_1, \ldots, a_n) - y^*| \leqslant A^0(y^*), \tag{4}$$

называемой *линейной оценкой погрешности.*

Задача 2. Доказать, что $A_0(y^*) - A(y^*) = o(\rho)$, $A^0(y^*) - A(y^*) = o(\rho)$.

Рассмотрим некоторые примеры определения величин $A(y^*)$, $A_0(y^*)$, $A^0(y^*)$ и произведем их сравнение.

1. $y = a^{10}$, $a^* = 1$, $\Delta(a^*) = 0{,}001$. Тогда

$$y^* = 1,\ y'_a = 10 \cdot a^9,\ b(0) = 10, \quad B = \sup_{|a-1| \leqslant 0{,}001} |10 \cdot a^9| = 10{,}09\ldots,$$

$$A(y^*) = \sup_{|a-1| \leqslant 0{,}001} |a^{10} - 1| = 1{,}001^{10} - 1 = 0{,}010045\ldots,$$

$$A_0(y^*) = B\Delta(a^*) = 0{,}01009\ldots, \quad A^0(y^*) = |b(0)|\,\Delta(a^*) = 0{,}01.$$

Здесь оценка погрешности через величину $A_0(y^*)$, предельно точная оценка (1) и линейная оценка (4) различаются несущественно.

2. $y = a^{10}$, $a^* = 1$, $\Delta(a^*) = 0{,}1$. Тогда

$$y^* = 1, \quad B = \sup_{|a-1| \leqslant 0{,}1} 10 \cdot a^9 = 10 \cdot 1{,}1^9 = 23,\ldots,$$

$$A(y^*) = \sup_{|a-1| \leqslant 0{,}1} |a^{10} - 1| = 1{,}1^{10} - 1 = 1{,}5\ldots,$$

$$A_0(y^*) = B\Delta(a^*) = 2{,}3\ldots, \quad A^0(y^*) = |b(0)|\,\Delta(a^*) = 1.$$

Здесь различие между этими оценками более заметно.

3. Проведем конкретную оценку погрешности в случае вычисления значений простейших функций. Пусть

$$y = \gamma_1 a_1 + \cdots + \gamma_n a_n,$$

где $\gamma_1, \ldots, \gamma_n$ принимают значения $+1$ или -1; пусть известны оценки $|a_j - a_j^*| \leqslant \Delta(a_j^*)$. В данном конкретном случае $b_j(\theta) = \gamma_j$, $|b_j(\theta)| = 1$, поэтому

$$A(y^*) = A_0(y^*) = A^0(y^*) = \Delta(a_1^*) + \cdots + \Delta(a_n^*).$$

Поскольку, по определению, погрешностью называют любую оценку для $y - y^*$, то это соотношение можно также записать в виде

$$\Delta(\pm a_1^* \pm \cdots \pm a_n^*) = \Delta(a_1^*) + \cdots + \Delta(a_n^*). \tag{5}$$

Это равенство иногда формулируется в виде правила:

предельная абсолютная погрешность суммы или разности равна сумме предельных погрешностей.

Если погрешности в величинах a_j^* зависимы, то оценка (5) часто может быть улучшена. Рассмотрим простейший пример: $a_1 = a$, $a_2 = 1 - a$; известно, что в обоих случаях a одно и то же; тогда независимо от погрешности в значении a сумма $a_1 + a_2$ равна 1 и погрешность суммы равна нулю.

Пусть теперь $y = a_1^{p_1} \ldots a_n^{p_n}$; тогда при всех j имеем $b_j(0) = p_j\Big((a_j^*)^{-1} \cdot y^*\Big)$ и

$$A(y^*) \approx A^0(y^*) = \sum_{j=1}^{n} |p_j|\, |a_j^*|^{-1} |y^*| \Delta(a_j^*).$$

После деления на $|y^*|$ получаем

$$\frac{A(y^*)}{|y^*|} \approx \frac{A^0(y^*)}{|y^*|} = \sum_{j=1}^{n} |p_j| \frac{\Delta(a_j^*)}{|a_j^*|} = \sum_{j=1}^{n} |p_j|\, \delta(a_j^*). \tag{6}$$

По отношению к частным случаям $y = a_1 \cdot a_2$ или $y = a_1 \cdot a_2^{-1}$ соотношение (6) иногда формулируют в виде правила:

предельная относительная погрешность произведения или частного приближенно равна сумме предельных относительных погрешностей.

4. Довольно часто возникает задача оценки погрешности функции, заданной неявно уравнением

$$F(y, a_1, \ldots, a_n) = 0. \tag{7}$$

Дифференцируя (7) по a_j, имеем

$$\frac{\partial F}{\partial y} \frac{\partial y}{\partial a_j} + \frac{\partial F}{\partial a_j} = 0,$$

откуда

$$\frac{\partial y}{\partial a_j} = -\left(\frac{\partial F}{\partial a_j}\right)\left(\frac{\partial F}{\partial y}\right)^{-1}. \tag{8}$$

При заданных $a_1^*, \ldots, a_n^*$ можно найти y^* как корень уравнения (7), а затем значения

$$b_j(0) = -\left(\frac{\partial F}{\partial a_j}\right)\left(\frac{\partial F}{\partial y}\right)^{-1}\Big|_{(y^*, a_1^*, \ldots, a_n^*)}. \tag{9}$$

С помощью этих величин можно получить «линейную оценку» погрешности (4).

Вследствие зависимости производных $\partial y/\partial a_j$ от самого значения y получение строгих оценок (1), (3) здесь довольно трудоемко.

Часто решение задачи зависит от приближенно задаваемых параметров настолько сложным образом, что получение или использование явных формул для производных по этим параметрам практически неприемлемо из-за своей громоздкости и трудоемкости. В такой ситуации для оценки этих производных целесообразно воспользоваться какими-либо приближенными формулами дифференцирования, например

$$f'(a) \approx \frac{f(a+\varepsilon) - f(a)}{\varepsilon}.$$

Так, производную решения дифференциального уравнения по начальному условию, в принципе, можно вычислить, интегрируя соответствующее уравнение в вариациях, решением которого является эта производная. Однако часто разумнее воспользоваться предыдущей формулой.

5. Рассмотрим один наиболее типичный частный случай из п. 4. Имеется приближение y^* к корню уравнения

$$f(y) = a;$$

требуется оценить его погрешность.

Вычислим величину

$$a^* = f(y^*).$$

При малых $y^* - y$ из равенства

$$f(y) - f(y^*) = a - a^*$$

следует, что

$$f'(y^*)(y - y^*) \approx a - a^*$$

и, таким образом,

$$y - y^* \approx \frac{a - a^*}{f'(y^*)} = \frac{a - f(y^*)}{f'(y^*)}.$$

В часто встречающемся случае $a = 0$ получаем

$$y - y^* \approx -\frac{f(y^*)}{f'(y^*)}.$$

6. Обратимся к оценке погрешности корней квадратного уравнения

$$F(y, a_1, a_2) = y^2 + a_1 y + a_2 = 0 \tag{10}$$

при заданных приближенных значениях коэффициентов a_1^*, a_2^* и их погрешностях $\Delta(a_1^*)$, $\Delta(a_2^*)$.

Пусть y^* — решение уравнения

$$y^{*2} + a_1^* y^* + a_2^* = 0.$$

Из формулы (9) имеем

$$b_1(0) = \frac{\partial y}{\partial a_1}\Big|_{(a_1^*, a_2^*)} = -\frac{y^*}{2y^* + a_1^*},$$

$$b_2(0) = \frac{\partial y}{\partial a_2}\Big|_{(a_1^*, a_2^*)} = -\frac{1}{2y^* + a_1^*}$$

и, следовательно,

$$A^0(y^*) = \frac{|y^*|\Delta(a_1^*) + \Delta(a_2^*)}{|2y^* + a_1^*|}. \tag{11}$$

Рассмотрим некоторую область $|a_1| \leqslant b_1$, $|a_2| \leqslant b_2$ изменения коэффициентов a_1, a_2. Из явного выражения корней

$$y = -\frac{a_1}{2} \pm \sqrt{\frac{a_1^2}{4} - a_2}$$

следует, что корни являются непрерывными функциями коэффициентов, поэтому

$$|y(a_1, a_2) - y(a_1^*, a_2^*)| \leqslant \omega(|a_1 - a_1^*|,\ |a_2 - a_2^*|);$$

при (a_1, a_2), (a_1^*, a_2^*) из этой области, $\omega(\lambda_1, \lambda_2) \to 0$ при λ_1, $\lambda_2 \to 0$. Правая часть в (11) стремится к ∞ при $2y^* + a_2^* \to 0$; поэтому «линейная оценка» погрешности при помощи формулы (4) может оказаться в некоторых случаях сильно завышенной по сравнению с точной оценкой погрешности (3). Дело в том, что ранее предполагалась непрерывная дифференцируемость $y(a_1, \ldots, a_n)$ по аргументам $(a_1, \ldots, a_n)$. Тогда погрешность y^* оказывалась величиной того же порядка, что и погрешности аргументов $\Delta(a_j^*)$. В случае, когда величина y^* определяется неявным образом, при некоторых значениях параметров она оказывается недифференцируемой функцией аргументов a_j^* и характер оценки меняется.

Пусть y^* является двукратным корнем уравнения (7) при $a_1 = a_1^*$, $a_2 = a_2^*$. Разложим левую часть (7) в ряд Тейлора в окрестности точки (y^*, a_1^*, a_2^*). Поскольку

$$F(y^*, a_1^*, a_2^*) = F_y(y^*, a_1^*, a_2^*) = 0$$

при y^* — двукратном корне уравнения (7), то уравнение (7) примет вид

$$d_{200}(y - y^*)^2 + d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + \cdots = 0,$$

где

$$d_{ijk} = \frac{F_{y^i a_1^j a_2^k}(y^*, a_1^*, a_2^*)}{i!j!k!},$$

а отброшенные члены имеют порядок $o(\rho)$. В случае уравнения (10) можно показать, что

$$y - y^* = \pm\sqrt{d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + o(\rho)}.$$

Таким образом, погрешность приближенного значения корня оказалась величиной порядка $O(\sqrt{\rho})$.

Задача 3. Показать, что в случае, когда уравнение имеет корень кратности k, погрешность корня имеет порядок $O(\sqrt[k]{\rho})$.

§ 6. Обратная задача

Часто приходится решать обратную задачу: с какой точностью надо задать значения аргументов $a_1^*, \ldots, a_n^*$ функции $y = y(a_1, \ldots, a_n)$, чтобы погрешность $y(a_1^*, \ldots, a_n^*)$ не превосходила заданной величины ε?

Пусть точки $(a_1, \ldots, a_n)$ и $(a_1^*, \ldots, a_n^*)$, соответствующие истинным и приближенным значениям параметров a_j, принадлежат некоторой выпуклой области G и $c_j = \sup\limits_G \left|\dfrac{\partial y}{\partial a_j}\right|$. Тогда имеем оценку погрешности

$$\left| y(a_1, \ldots, a_n) - y(a_1^*, \ldots, a_n^*) \right| \leqslant \sum_{j=1}^{n} c_j \Delta(a_j^*).$$

Любая совокупность $(\Delta(a_1^*), \ldots, \Delta(a_n^*))$ абсолютных погрешностей, удовлетворяющих неравенству

$$\sum_{j=1}^{n} c_j \Delta(a_j^*) \leqslant \varepsilon,$$

обеспечивает требуемую точность.

Если функция y зависит только от одного аргумента ($n = 1$), то имеем неравенство $c_1 \Delta(a_j^*) \leqslant \varepsilon$ и для достижения требуемой точности достаточно взять $\Delta(a_1^*) = \varepsilon / c_1$.

В случае $n > 1$ иногда рекомендуют отвести погрешности каждого аргумента равную долю, т. е. выбрать $\Delta(a_j^*)$ из условия $c_j \Delta(a_j^*) = \varepsilon / n$, т. е. $\Delta(a_j^*) = \varepsilon / (c_j n)$. В других случаях предлагают взять все оценки погрешностей равными, максимально возможными, т. е. положить

$$\Delta(a_1^*) = \cdots = \Delta(a_n^*) = \delta, \quad \text{где} \quad \delta = \varepsilon / (c_1 + \cdots + c_n).$$

В простейших случаях можно последовать этим рецептам, однако в более сложных случаях целесообразно подойти к вопросу о выборе верхних границ для допустимых погрешностей аргументов $\Delta(a_j^*)$ более аккуратно. Дело заключается в том, что достижение определенной точности в задании аргумента a_j может существенно зависеть от номера j. Тогда следует ввести в рассмотрение функцию стоимости $F(\Delta(a_1^*), \ldots, \Delta(a_n^*))$ затрат на задание точки $(a_1^*, \ldots, a_n^*)$ с заданными абсолютными погрешностями координат $\Delta(a_1^*), \ldots, \Delta(a_n^*)$ и найти минимум функции $F(x_1, \ldots, x_n)$ в области $c_1x_1 + \cdots + c_nx_n \leqslant \varepsilon$, $0 \leqslant x_1, \ldots, x_n$. Пусть он достигается в точке $x_1^0, \ldots, x_n^0$. Далее следует положить $\Delta(a_j^*) = x_j^0$, $j = 1, \ldots, n$.

В ряде типичных случаев функция $F(x_1, \ldots, x_n)$ имеет вид

$$F(x_1, \ldots, x_n) = \sum_{j=1}^{n} D_j x_j^{-d_j}, \quad D_j,\ d_j > 0, \quad j = 1, \ldots, n.$$

Ясно, что искомое минимальное значение функции $F(x_1, \ldots, x_n)$ достигается в некоторой точке $(x_1^0, x_2^0, \ldots, x_n^0)$ плоскости $c_1x_1 + \cdots + c_nx_n = \varepsilon$.

Рассмотрим случай $n = 2$. Составляем функцию Лагранжа

$$\Phi(x_1, x_2) = F(x_1, x_2) + \lambda(c_1x_1 + c_2x_2 - \varepsilon)$$

и, приравнивая нулю производные $\partial\Phi/\partial x_j$, получим систему уравнений

$$-d_1D_1x_1^{-d_1-1} + \lambda c_1 = 0, \quad -d_2D_2x_2^{-d_2-1} + \lambda c_2 = 0.$$

Отсюда

$$x_1 = \left(\frac{D_1d_1}{\lambda c_1}\right)^{1/(d_1+1)}, \quad x_2 = \left(\frac{D_2d_2}{\lambda c_2}\right)^{1/(d_2+1)}. \tag{1}$$

Подставляя x_j в равенство $c_1x_1 + c_2x_2 = \varepsilon$, получим уравнение относительно λ:

$$c_1\left(\frac{D_1d_1}{\lambda c_1}\right)^{1/(d_1+1)} + c_2\left(\frac{D_2d_2}{\lambda c_2}\right)^{1/(d_2+1)} = \varepsilon.$$

Видно, что $\lambda \to \infty$ при $\varepsilon \to 0$. Пусть $d_1 > d_2$. Тогда при больших λ главным членом в левой части является первый; поэтому имеем приближенное равенство

$$c_1\left(\frac{D_1d_1}{\lambda c_1}\right)^{1/(d_1+1)} \approx \varepsilon,$$

откуда следует

$$\lambda \approx \frac{D_1d_1}{c_1}\left(\frac{c_1}{\varepsilon}\right)^{d_1+1}.$$

Подставляя λ в (1), получим

$$\Delta(a_1^*) = x_1 \approx \frac{\varepsilon}{c_1},$$
$$\Delta(a_2^*) = x_2 \approx \left(\frac{D_2 d_2 c_1}{D_1 d_1 c_2}\right)^{1/(d_2+1)} \left(\frac{\varepsilon}{c_1}\right)^{(d_1+1)/(d_2+1)}. \qquad (2)$$

Поскольку $d_1 > d_2$, то в рассматриваемом примере стоимость задания первого аргумента при малых ε растет быстрее, чем стоимость задания второго аргумента. В соответствии с этим мы получили, что второй аргумент следует задавать с точностью более высокого порядка малости по ε, в то время как точность задания первого аргумента практически определяется из равенства $c_1\Delta(a_1^*) = \varepsilon$.

Различный характер зависимости функции стоимости от погрешностей задания аргументов может определяться многими факторами. Если, например, параметры a_j определяются численным решением некоторых вспомогательных задач, то слагаемые $D_j\Delta(a_j)^{-d_j}$ характеризуют различную трудоемкость решения этих задач. В других случаях этот характер может определяться сложностью получения экспериментальных данных или трудностью достижения нужной точности тех или иных параметров в реальной конструкции.

Литература

1. Березин И. С., Жидков Н. П. Методы вычислений. Т. 1. — М.: Наука, 1966.
2. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интерполирование и интегрирование. — Минск: Наука и техника, 1983.
3. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 1. — М.: Наука, 1976.

Глава 2

Интерполяция и численное дифференцирование

В этой главе излагаются наиболее широко используемые способы вычисления приближенных значений функции и ее производных в случае, когда известны значения функции в некоторых фиксированных точках. Множество этих точек иногда задается нам внешними обстоятельствами, а иногда мы можем выбрать их по своему усмотрению.

Такого рода задачи приближения и приближенного дифференцирования часто возникают как самостоятельные в ситуациях, некоторые из которых будут рассмотрены ниже. Кроме того, алгоритмы решения этих задач используются как вспомогательные при построении методов вычисления интегралов, решения дифференциальных и интегральных уравнений. Наличие большого количества методов вызвано историческим развитием теории и практики решения прикладных задач. Многие методы возникли как варианты предшествующих, отличаясь от них формой записи, изменением порядка вычислений, имевшими цель уменьшить влияние погрешности округлений при вычислениях.

В то же время развитие вычислительной техники и теории численных методов приводит к непрерывному пересмотру и некоторому сужению совокупности применяемых методов.

Некоторые методы вышли из употребления по следующим причинам. Произошло увеличение разрядности чисел в ЭВМ, и как следствие этого оказалось несущественным различие в вычислительной погрешности в зависимости от последовательности арифметических операций; в результате этого в практике вычислений постепенно закрепились простейшие по форме методы. С другой стороны, усложнение модели задачи и требование уменьшения погрешности метода, как правило, требуют и существенного роста числа выполняемых арифметических операций. Несмотря на повышение разрядности чисел в ЭВМ, для ряда методов это обстоятельство приводит к недопустимо большому значению вычислительной погрешности (так называемой *неустойчивости*). Поэтому при повышении требований к точности результата ряд методов также был забракован и изъят из вычислительной практики. Тем не менее сохраняется положение, когда для решения каждой конкретной задачи можно применить довольно много методов.

§ 1. Постановка задачи приближения функций

Иногда возникает следующий вопрос. Может быть, наличие большого количества различных способов приближения объясняется просто отсутствием научного подхода к постановке и решению проблемы; если бы такой подход был, то, может быть, удалось бы предложить один оптимальный способ приближения, пригодный во всех случаях? Такой вопрос возникает и при рассмотрении других разделов численного анализа. Сколь бы ни было заманчиво разработать единый подход к решению всех задач, следует все-таки признать, что многообразие методов вызывается существом дела — многообразием различных постановок проблемы. В частности, различные теоретические разделы теории приближений, например интерполяции, можно рассматривать как изучение абстрактных моделей некоторых реальных классов проблем.

1. Простейшая задача, приводящая к приближению функций, заключается в следующем. В дискретные моменты времени $x_1, \ldots, x_n$ наблюдаются значения функции $f(x)$; требуется восстановить ее значения при других x. Подобная задача возникает, в частности, в следующей обстановке. По ходу вычислений на ЭВМ приходится многократно вычислять одну и ту же сложную функцию в различных точках. Вместо ее непосредственного вычисления иногда целесообразно вычислить ее значение в отдельных выбираемых нами по своему усмотрению точках, а в других точках вычислять по каким-либо простым формулам, используя информацию об этих известных значениях.

Иногда из каких-то дополнительных соображений известно, что приближающую функцию целесообразно искать в виде

$$f(x) \approx g(x; a_1, \ldots, a_n).$$

Если параметры $a_1, \ldots, a_n$ определяются из условия совпадения $f(x)$ и приближающей функции в точках $x_1, \ldots, x_n$, так называемых *узлах интерполяции*,

$$g(x_j; a_1, \ldots, a_n) = f(x_j), \quad j = 1, \ldots, n,$$

то такой способ приближения называют *интерполяцией* или *интерполированием*.

2. Пусть y_1 — наименьшее из чисел x_i — узлов интерполяции, а y_2 — наибольшее из них. Если точка x, в которой вычисляется значение $f(x)$, лежит вне отрезка $[y_1, y_2]$, то наряду с термином *интерполяция* употребляют термин *экстраполяция*.

Например, известно поведение какой-либо переменной до данного момента времени и требуется высказать какое-то суждение о ее дальнейшем поведении. Это может быть температура, рост производства или потребления какого-либо продукта, рост народонаселения, урожайности и т. п. Задаются какими-то моментами времени, строят интерполирующую функцию и ее значение в какой-то будущий момент принимают за прогнозируемое (экстраполируемое) значение искомой величины.

Если узлы интерполяции выбраны далеко от момента времени, где приближается функция, то тем самым слабо используется существенная информация о поведении переменной в последнее время. Если они выбраны очень близко, то увеличивается роль погрешностей в используемой информации. Таким образом, вопрос о выборе узлов интерполяции и экстраполяции непрост, особенно в задачах, где значения исследуемой функции зависят от многих случайных или трудно учитываемых факторов. Сюда относятся задачи прогноза погоды, урожайности, медицины и т. д., в которых, как правило, требуется применять более сложные (в частности статистические) методы прогнозирования.

3. Наиболее часто используется рассматриваемая ниже интерполяция многочленами. Однако это не единственный возможный вид интерполяции. Иногда удобнее приближать функции тригонометрическими полиномами; в других задачах целесообразнее приближать многочленом не $f(x)$, а $\ln f(x)$, или приближать $f(x)$ не многочленом от x, а многочленом от $\ln x$.

Часто целесообразно использовать интерполяцию дробно-рациональными функциями

$$f(x) \approx g(x) = \frac{\sum\limits_{j=0}^{n} a_j x^j}{\sum\limits_{k=0}^{m} b_k x^k}.$$

4. В задачах *планирования экспериментов* в биологии, физике, химии, географии, медицине и других областях науки возникает следующая проблема. Известен вид хорошего приближения функции, например функция хорошо приближается многочленом второй степени. В то же время измеряемые значения функции содержат большие погрешности. Требуется получить наилучшее в определенной норме приближение при минимальном числе измерений значений функции.

5. Задача приближения появляется при составлении стандартных программ вычисления элементарных и специальных функций. Обычно такие функции обладают свойствами, позволяющими резко уменьшить объем вычислений.

Возникающая здесь проблема может быть сформулирована следующим образом. Рассматриваются все функции $g(x)$, программа вычисления которых требует некоторого фиксированного объема памяти ЭВМ, такие, что некоторая норма погрешности $\|f - g\|$ не превосходит ε. Среди всех таких функций нужно выбрать ту, вычисление которой требует минимальных затрат времени ЭВМ.

В зависимости от обстоятельств норма может выбираться по-разному. В большинстве случаев берется $\|f\| = \sup\limits_{[a,\,b]} |f|$, где $[a, b]$ — отрезок, на котором приближается функция.

Довольно часто требуется повышенная точность в отдельных точках. Например, одна из стандартных программ вычисления $\sin x$ обеспечивает малость погрешности в норме

$$\|f\| = \sup_{[0,\pi/2]} |p(x)f(x)|, \quad p(x) = \min\{10^{19}, x^{-1}\}.$$

Введение множителя $p(x)$ вызывается требованием малости относительной погрешности значений $\sin x$ при малых x.

Точно так же по-разному может толковаться требование минимальности затрат процессорного времени ЭВМ. Затраты, вообще говоря, могут зависеть от точки, в которой вычисляется значение функции. Обозначим их через $t(x)$. Если не имеется информации о том, с какой частотой вычисляются значения функции в тех или иных частях отрезка, то, например, можно в качестве общей меры затрат принять

$$T = \sup_x t(x).$$

Если такая информация есть, то можно принять

$$T = M(t(x)),$$

где $M(t(x))$ — математическое ожидание случайной величины $t(x)$.

6. При задании функции графиком или сложным аналитическим выражением возникают вариационные задачи других типов. Например, пусть решено разбить отрезок $[a, b]$ на l частей:

$$[a_{i-1}, a_i], \quad i = 1, \ldots, l, \quad a_0 = a, \quad a_l = b,$$

и на каждом отрезке $[a_{i-1}, a_i]$ приближать функцию многочленом степени n_i. Среди таких способов приближения отыскивается оптимальный в том или ином смысле. Чаще всего заранее накладывается требование $n_i \equiv n$, фиксируется число отрезков разбиения l и проводится оптимизация метода по $a_1, \ldots, a_{l-1}$.

В частном случае $l = 1$ возникает задача наилучшего приближения многочленами. Об этой задаче речь пойдет в гл. 4.

7. Вид приближающей функции существенно зависит от цели, с которой осуществляется приближение. Предположим, что с требуемой точностью функция может быть приближена многочленом десятой степени или выражением $a_1 \sin(\omega_1 x + \varphi_1) + a_2 \sin(\omega_2 x + \varphi_2)$. Если полученное приближение используется в теоретических исследованиях, для решения задачи на моделирующем устройстве или в технологическом процессе, то вторая форма записи может быть более удобной. Однако если значения функции вычисляются на ЭВМ, то вторая форма записи может потребовать при своей реализации большего числа арифметических операций.

Далее будет конкретно рассмотрена задача интерполирования многочленами. Ее выделение вызвано наличием непосредственных многочи-

сленных приложений, а также и следующими обстоятельствами: аппарат интерполирования многочленами является важнейшим аппаратом численного анализа; на его основе строится большинство методов решения других задач; его роль в численном анализе аналогична роли формулы Тейлора в классическом анализе.

Попутно будут рассмотрены некоторые другие вопросы общего характера, имеющие значение для других разделов численного анализа.

§ 2. Интерполяционный многочлен Лагранжа

Среди способов интерполирования наиболее распространен случай линейного интерполирования: приближение ищется в виде

$$g(x; a_1, \ldots, a_n) = \sum_{i=1}^{n} a_i \varphi_i(x),$$

где $\varphi_i(x)$ — фиксированные функции, значения коэффициентов a_i определяются из условия совпадения с приближаемой функцией в узлах интерполяции x_j:

$$f(x_j) = \sum_{i=1}^{n} a_i \varphi_i(x_j), \quad j = 1, \ldots, n. \tag{1}$$

Метод решения задачи, при котором коэффициенты a_i определяются непосредственным решением системы (1), называется *методом неопределенных коэффициентов.*

Как правило, в методе неопределенных коэффициентов число заданных условий равно числу свободных (неизвестных) параметров, подлежащих определению.

Наиболее изучен случай интерполирования многочленами

$$\sum_{i=1}^{n} a_i x^{i-1}. \tag{2}$$

Тогда

$$\varphi_i(x) = x^{i-1}, \quad i = 1, \ldots, n,$$

и система уравнений (1) имеет вид

$$\sum_{i=1}^{n} a_i x_j^{i-1} = f(x_j), \quad j = 1, \ldots, n. \tag{3}$$

Далее мы предполагаем, что все x_j различные. Определитель этой системы $\det[x_j^{i-1}]$ отличен от нуля (определитель Вандермонда). Следовательно, система (3) всегда имеет решение, и притом единственное. Та-

ким образом, доказано существование и единственность интерполяционного многочлена вида (2).

Непосредственное нахождение коэффициентов a_i с помощью решения этой системы уже при сравнительно небольших n, например, при $n = 20$, приводит к существенному искажению коэффициентов a_i вычислительной погрешностью. Кроме того, как мы увидим в гл. 4, уже сама запись многочлена в традиционной форме (2) часто приводит к большой вычислительной погрешности результата. При теоретических исследованиях, например при конструировании алгоритмов решения других задач, эти обстоятельства могут не играть роли. Однако при реальных вычислениях влияние вычислительной погрешности может быть недопустимо большим, и поэтому применяются другие виды интерполяционного многочлена и способы его записи.

Можно получить явные представления интерполяционного многочлена (2), не прибегая к непосредственному решению системы (3). Сразу же отметим, что в других случаях, например при интерполировании функций многих переменных, получение интерполяционного многочлена в явном виде затруднительно, и часто приходится прибегать к непосредственному решению системы уравнений типа (1).

Пусть δ_i^j есть символ Кронекера, определяемый соотношениями

$$\delta_i^j = \begin{cases} 1 & \text{при} \quad i = j, \\ 0 & \text{при} \quad i \neq j. \end{cases}$$

Задача интерполирования будет решена, если удастся построить многочлены $\Phi_i(x)$ степени не выше $n-1$ такие, что $\Phi_i(x_j) = \delta_i^j$ при $i, j = 1, \ldots, n$. Многочлен

$$g_n(x) = \sum_{i=1}^{n} f(x_i)\Phi_i(x)$$

будет искомым интерполяционным многочленом. В самом деле,

$$g_n(x_j) = \sum_{i=1}^{n} f(x_i)\Phi_i(x_j) = \sum_{i=1}^{n} f(x_i)\delta_i^j = f(x_j);$$

кроме того, $g_n(x)$ — многочлен степени $n-1$.

Поскольку $\Phi_i(x_j) = 0$ при $j \neq i$, то $\Phi_i(x)$ делится на $x - x_j$ при $j \neq i$. Таким образом, нам известны $n-1$ делителей многочлена степени $n-1$, поэтому

$$\Phi_i(x) = \text{const} \prod_{j \neq i} (x - x_j).$$

Из условия $\Phi_i(x_i) = 1$ получаем

$$\Phi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Интерполяционный многочлен (2), записанный в форме

$$g_n(x) \equiv L_n(x) = \sum_{i=1}^{n} f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}, \tag{4}$$

называют *интерполяционным многочленом Лагранжа.*

Существуют другие формы записи того же интерполяционного многочлена (2), например рассматриваемая далее *интерполяционная формула Ньютона* и ее варианты. При точных (без округлений) вычислениях значения, получаемые по различным интерполяционным формулам, совпадают. Наличие же округлений приводит к различию в получаемых по этим формулам значений интерполяционных многочленов. Запись многочлена в форме Лагранжа, как правило, приводит к меньшей величине вычислительной погрешности; запись же многочлена в форме Ньютона более наглядна и позволяет лучше проследить аналогию проводимых построений с основными построениями математического анализа. Кроме того, этим различным формам записи соответствует различное *количество арифметических операций* при вычислении с их помощью значений интерполяционного многочлена.

Мы употребили термин «количество арифметических операций». Поясним, что имеется в виду. Пусть рассматривается задача вычисления значения многочлена

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

в точке x. Вычисления можно проводить различными способами. Например, можно поступить следующим образом. Вычислить значение $a_1 x$ и сложить с a_0. Далее вычислить значение $a_2 x^2$ и сложить с полученным результатом и т. д. На j-м шаге, таким образом, вычисляется значение $a_j x^j$ и складывается с уже вычисленной суммой $a_0 + a_1 x + \cdots + a_{j-1} x^{j-1}$. Вычисление значения $a_j x^j$ требует j операций умножения. Таким образом, описанный выше алгоритм требует для вычисления значения многочлена $(1 + 2 + \cdots + n) = n(n+1)/2$ операций умножения и n операций сложения. Количество арифметических операций (действий) в данном случае будет равно $\Phi_1 = n(n+1)/2 + n$.

Ясно, что количество арифметических операций, необходимых для вычисления значения $P_n(x)$, может быть уменьшено. Например, можно последовательно вычислить и запомнить значения x^2, $x^3, \ldots, x^n$. Для этого потребуется $n-1$ операций умножения. Далее вычисляем величины $a_j x^j$ $(j = 1, \ldots, n)$. Это потребует n операций умножения. Складывая полученные значения (это требует n операций сложения), получаем $P_n(x)$. В этом случае $\Phi_2 = (2n-1) + n$ и уже при $n > 2$ имеет место неравенство $\Phi_2 < \Phi_1$.

Можно пойти еще дальше. Запишем $P_n(x)$ в виде

$$P_n = \Big(\cdots \big((a_n x + a_{n-1}) x + a_{n-2} \big) x + \cdots \Big) x + a_0.$$

Для вычисления значения во внутренних скобках $a_n x + a_{n-1}$ требуется одна операция умножения и одна операция сложения. Для вычисления значения в следующих скобках $(a_n x + a_{n-1})x + a_{n-2}$ требуется опять одна операция умножения и одна операция сложения, так как $a_n x + a_{n-1}$ уже вычислено, и т. д. Таким образом, вычисление $P_n(x)$ при помощи этого алгоритма потребует n операций умножения и n операций сложения, то есть $\Phi_3 = n + n$.

Ясно, что $\Phi_3 < \Phi_2 < \Phi_1$ при $n > 2$. Таким образом, вычисление $P_n(x)$ по последнему алгоритму потребует меньше арифметических операций и, соответственно, меньше времени ЭВМ. Количество арифметических операций, которое требуется для получения результата, является одной из важнейших характеристик метода, по которой происходит сравнение методов.

Иногда до начала вычислений не удается точно оценить требуемое количество арифметических операций, а удается оценить лишь *порядок количества арифметических операций* по отношению к какому-либо параметру. В рассматриваемом выше примере

$$\Phi_1 = O(n^2), \quad \Phi_2,\ \Phi_3 = O(n),$$

n — степень многочлена.

В последнем случае ($\Phi_2,\ \Phi_3 = O(n)$) говорят, что методы имеют одинаковый порядок количества арифметических операций.

В тех случаях, когда находится порядок количества арифметических операций, бывает важно найти постоянную в главном члене. Например,

$$\Phi_1 = \frac{1}{2}n^2 + o(n^2), \quad \Phi_2 = 3n + o(n), \quad \Phi_3 = 2n.$$

Как правило, метод, требующий меньшего количества арифметических операций, является более быстрым, и поэтому считается лучшим. Выбирая метод решения сложных задач, часто ограничиваются лишь сравнением порядков количества арифметических операций для различных методов.

Заметим, что значение многочлена $P_n(x)$ определяется параметрами $a_0, a_1, \ldots, a_n$ и величиной x. Поэтому в общем случае для вычисления $P_n(x)$ потребуется не менее n арифметических операций, т. е. мы имеем оценку снизу для количества арифметических операций. Таким образом, второй и третий методы вычисления $P_n(x)$ являются *оптимальными по порядку*, так как $\Phi_2,\ \Phi_3 = O(n)$ и для любого метода $\Phi \geqslant n$.

В связи с появлением многопроцессорных вычислительных комплексов может случиться так, что метод, требующий большего количества арифметических операций, будет быстрее другого метода с меньшим количеством арифметических операций. Поэтому для случая многопроцессорной ЭВМ нельзя оценивать качество метода только по количеству арифметических операций.

§ 3. Оценка остаточного члена интерполяционного многочлена Лагранжа

В предположении непрерывности $f^{(n)}(x)$ оценим разность между $f(x)$ и построенным интерполяционным многочленом $g_n(x)$. Положим

$$\varphi(t) = f(t) - g_n(t) - K\omega_n(t),$$

где $\omega_n(t) = (t - x_1)\cdots(t - x_n)$, а K выберем из условия $\varphi(x) = 0$, где x — точка, в которой оценивается погрешность. Из уравнения $\varphi(x) = 0$ получаем

$$K = \frac{f(x) - g_n(x)}{\omega_n(x)}.$$

При таком выборе K функция $\varphi(t)$ обращается в нуль в $(n+1)$-й точке $x_1, \ldots, x_n, x$. На основании теоремы Ролля ее производная $\varphi'(t)$ обращается в нуль по крайней мере в n точках. Применяя теорему Ролля к $\varphi'(t)$, получаем, что ее производная $\varphi''(t)$ обращается в нуль по крайней мере в $(n-1)$-й точке. Продолжая эти рассуждения дальше, получаем, что $\varphi^{(n)}(t)$ обращается в нуль по крайней мере в одной точке ζ, принадлежащей отрезку $[y_1, y_2]$, где

$$y_1 = \min\{x_1, \ldots, x_n, x\}, \quad y_2 = \max\{x_1, \ldots, x_n, x\}.$$

Поскольку

$$\varphi^{(n)}(t) = f^{(n)}(t) - Kn!,$$

из условия $\varphi^{(n)}(\zeta) = 0$ будем иметь

$$K = \frac{f^{(n)}(\zeta)}{n!}.$$

Следовательно, соотношение $\varphi(x) = 0$ можно переписать в виде

$$f(x) - g_n(x) = \frac{f^{(n)}(\zeta)\,\omega_n(x)}{n!}, \quad \zeta \in [y_1, y_2], \tag{1}$$

дающем представление остаточного члена.

§ 4. Разделенные разности и их свойства

Как будет видно далее, интерполяционный многочлен можно рассматривать как обобщение отрезка ряда Тейлора.

Обобщением понятия производной является понятие *разделенной разности*. Разделенные разности нулевого порядка $f(x_i)$ совпадают со значениями функции $f(x_i)$; разности первого порядка определяются равенством

$$f(x_i;\, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \tag{1}$$

разности второго порядка — равенством

$$f(x_i;\, x_j;\, x_k) = \frac{f(x_j;\, x_k) - f(x_i;\, x_j)}{x_k - x_i},$$

и, вообще, разности k-го порядка $f(x_1; \ldots; x_{k+1})$ определяются через разности $(k-1)$-го порядка по формуле

$$f(x_1; \ldots; x_{k+1}) = \frac{f(x_2; \ldots; x_{k+1}) - f(x_1; \ldots; x_k)}{x_{k+1} - x_1}. \tag{2}$$

Иногда вместо $f(x_1; \ldots; x_k)$ используют обозначения $(f)(x_1; \ldots; x_k)$ или $[x_1; \ldots; x_k]$.

Лемма. *Справедливо равенство*

$$f(x_1; \ldots; x_k) = \sum_{j=1}^{k} \frac{f(x_j)}{\prod\limits_{i \neq j} (x_j - x_i)}. \tag{3}$$

Доказательство будем проводить по индукции. При $k = 1$ это равенство превращается в равенство $f(x_1) = f(x_1)$, при $k = 2$ совпадает с (1). Пусть (3) доказано при $k < l$. Тогда

$$f(x_1; \ldots; x_{l+1}) = \frac{f(x_2; \ldots; x_{l+1}) - f(x_1; \ldots; x_l))}{x_{l+1} - x_1} =$$

$$= \frac{1}{x_{l+1} - x_1} \left(\sum_{j=2}^{l+1} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ 2 \leqslant i \leqslant l+1}} (x_j - x_i)} - \sum_{j=1}^{l} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l}} (x_j - x_i)} \right).$$

Если $j \neq 1,\ l+1$, то коэффициент при $f(x_j)$ в правой части есть

$$\frac{1}{x_{l+1} - x_1} \left(\frac{1}{\prod\limits_{\substack{i \neq j \\ 2 \leqslant i \leqslant l+1}} (x_j - x_i)} - \frac{1}{\prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l}} (x_j - x_i)} \right) =$$

$$= \frac{(x_j - x_1) - (x_j - x_{l+1})}{(x_{l+1} - x_1) \prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l+1}} (x_j - x_i)} = \frac{1}{\prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l+1}} (x_j - x_i)},$$

т. е. имеет требуемый вид; для $j = 1$ или $j = l + 1$ значение $f(x_j)$ входит только в одно слагаемое в правой части, и коэффициент при нем также имеет требуемый вид. Доказательство закончено.

Непосредственно из (3) вытекает ряд следствий.

1. При фиксированных $x_1, \ldots, x_k$ разделенная разность является линейным функционалом от функции f:

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_1; \ldots; x_k) = \alpha_1 f_1(x_1; \ldots; x_k) + \alpha_2 f_2(x_1; \ldots; x_k).$$

2. Разделенная разность есть симметрическая функция своих аргументов $x_1, \ldots, x_k$ (т. е. не меняется при любой их перестановке).

Если функция задана в точках $x_1, \ldots, x_n$, то таблицу

$$\begin{array}{lllll} f(x_1) & & & & \\ & f(x_1; x_2) & & & \\ f(x_2) & & f(x_1; x_2; x_3) & \ddots & \\ & f(x_2; x_3) & \vdots & \ddots & f(x_1; \ldots; x_n) \\ f(x_3) & \vdots & & \ddots & \\ \vdots & f(x_{n-1}, x_n) & \ddots & & \\ f(x_n) & & & & \end{array} \qquad (4)$$

называют *таблицей разделенных разностей.*

§ 5. Интерполяционная формула Ньютона с разделенными разностями

При помощи разделенных разностей можно получить другую форму записи интерполяционного многочлена (2.4).

Справедливо равенство

$$f(x) - L_n(x) = f(x) - \sum_{i=1}^{n} f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} =$$

$$= \prod_{i=1}^{n} (x - x_i) \left(\frac{f(x)}{\prod_{i=1}^{n} (x - x_i)} + \sum_{i=1}^{n} \frac{f(x_i)}{(x_i - x) \prod_{j \neq i} (x_i - x_j)} \right).$$

Сравнивая с (4.3), убеждаемся, что выражение в скобках есть $f(x; x_1; \ldots; x_n)$. Таким образом,

$$f(x) - L_n(x) = f(x; x_1; \ldots; x_n)\, \omega_n(x), \qquad (1)$$

где многочлен $\omega_n(x)$ определен в § 3.

Пусть $L_m(x)$ — интерполяционный многочлен Лагранжа с узлами интерполяции $x_1, \ldots, x_m$. Интерполяционный многочлен Лагранжа $L_n(x)$ можно представить в виде

$$L_n(x) = L_1(x) + \big(L_2(x) - L_1(x)\big) + \cdots + \big(L_n(x) - L_{n-1}(x)\big). \qquad (2)$$

Разность $L_m(x) - L_{m-1}(x)$ есть многочлен степени $m-1$, обращающийся в нуль в точках $x_1, \ldots, x_{m-1}$, поскольку $L_{m-1}(x_j) = L_m(x_j) = f(x_j)$ при $1 \leqslant j \leqslant m-1$. Следовательно,

$$L_m(x) - L_{m-1}(x) = A_{m-1}\,\omega_{m-1}(x), \quad \omega_{m-1}(x) = (x - x_1)\ldots(x - x_{m-1}),$$

где $A_{m-1} = \text{const}$. Полагая $x = x_m$, получим

$$f(x_m) - L_{m-1}(x_m) = A_{m-1}\omega_{m-1}(x_m).$$

С другой стороны, полагая в (1) $n = m-1$ и $x = x_m$, имеем

$$f(x_m) - L_{m-1}(x_m) = f(x_m;\, x_1;\, \ldots;\, x_{m-1})\,\omega_{m-1}(x_m).$$

Таким образом, $A_{m-1} = f(x_1;\, \ldots;\, x_m)$ и поэтому

$$L_m(x) - L_{m-1}(x) = f(x_1;\, \ldots;\, x_m)\,\omega_{m-1}(x).$$

Подставляя эти величины в (2), получим

$$L_n(x) = f(x_1) + f(x_1; x_2)(x - x_1) + \cdots + f(x_1; \ldots; x_n)(x - x_1)\ldots(x - x_{n-1}). \qquad (3)$$

Интерполяционный многочлен, записанный в такой форме, называется *интерполяционным многочленом Ньютона с разделенными разностями*. Из сравнения (1) с (3.1) следует важное равенство

$$f(x;\, x_1; \ldots;\, x_n) = \frac{f^{(n)}(\zeta)}{n!}, \quad y_1 \leqslant \zeta \leqslant y_2. \qquad (4)$$

В частности, если $f(x)$ — многочлен

$$P_l(x) = \sum_{j=0}^{l} b_j x^j$$

степени $l \leqslant n$, то на основании этой формулы имеем

$$P_l(x_0; \ldots;\, x_n) = \begin{cases} b_n & \text{при} \quad l = n, \\ 0 & \text{при} \quad l < n \end{cases}$$

при любых $x_0, \ldots, x_n$.

Предположим, что точки $x_1, \ldots, x_n$ пронумерованы в порядке возрастания: $x_1 < x_2 < \cdots < x_n$. Вследствие (4) имеем

$$m!\, f(x_k; \ldots;\, x_{k+m}) = f^{(m)}(\zeta_k), \quad \text{где } x_k \leqslant \zeta_k \leqslant x_{k+m}.$$

Поэтому величина

$$M_m = \sup_{1 \leqslant k \leqslant n-k} m!\, |f(x_k; \ldots;\, x_{k+m})|$$

может использоваться в качестве приближенной оценки для величины

$$M^{(m)} = \sup_{[x_1, x_n]} \left| f^{(m)}(x) \right|.$$

Задача 1. Доказать неравенство

$$|M_m - M^{(m)}| \leqslant M^{(m+1)} h_m,$$

где $h_m = \max\limits_{1 \leqslant k \leqslant n-m} (x_{k+m} - x_k)$.

Для упрощения вычислений интерполяционного многочлена иногда используется так называемая *схема Эйткена.*

Пусть $L_{(k,k+1,\ldots,l)}(x)$ — интерполяционный многочлен с узлами интерполяции $x_k, \ldots, x_l$, в частности $L_{(k)}(x) = f(x_k)$. Справедливо равенство

$$L_{(k,k+1,\ldots,l+1)}(x) = \frac{L_{(k+1,\ldots,l+1)}(x)\,(x - x_k) - L_{(k,\ldots,l)}(x)(x - x_{l+1})}{x_{l+1} - x_k}; \qquad (5)$$

действительно, правая часть является многочленом степени $l - k + 1$ и совпадает с $f(x)$ в точках $x_k, \ldots, x_{l+1}$. Схема Эйткена вычисления значения $L_{(1,\ldots,n)}(x)$ заключается в последовательном вычислении с помощью формулы (5) элементов таблицы значений интерполяционных многочленов

$$\begin{array}{lllll}
l_{(1)}(x) & & & & \\
 & l_{(1,2)}(x) & & & \\
l_{(2)}(x) & & l_{(1,2,3)}(x) & \ddots & \\
 & l_{(2,3)}(x) & \vdots & \ddots & l_{(1,2,\ldots,n)}(x) \\
l_{(3)}(x) & \vdots & \ddots & & \\
\vdots & l_{(n-1,n)}(x) & & & \\
l_{(n)}(x) & & & &
\end{array}$$

Эта схема положена в основу *стандартной программы* решения следующей задачи.

Дана таблица значений некоторой функции $f(x)$; требуется при каждом значении x вычислить значение $f(x)$ с заданной точностью ε или с наилучшей возможной точностью при имеющейся информации.

Трудно дать четкое определение термина *стандартная программа.* По установившейся традиции стандартной программой решения задач некоторого класса называют квалифицированно написанную программу, содержащую описание алгоритма решения задач данного класса. Решение конкретной задачи осуществляется подсоединением к стандартной программе информации об этой конкретной задаче. От стандартной программы также требуется, чтобы допускалось ее использование как элемента программы, предназначенной для решения более сложных задач.

Рассматриваемое ниже построение алгоритма решения задачи является довольно типичным для ситуации, возникающей в реальной практике.

Невозможно предложить обоснованный алгоритм решения поставленной задачи для всех функций, поскольку про функцию ничего не известно, кроме ее значений в заданных точках. Однако, предполагая функцию гладкой, мы выводим практический критерий оценки погрешности и, основываясь на нем, строим алгоритм решения задачи.

Пусть x фиксировано; перенумеруем узлы интерполяции в порядке возрастания $|x_i - x|$. Интерполяционные многочлены $L_{(1,\dots,m)}(x)$ будем обозначать, как обычно, $L_m(x)$.

Выше получено представление погрешности (1)

$$f(x) - L_m(x) = f(x; x_1; \dots; x_m)\,\omega_m(x),$$

а также равенство

$$L_{m+1}(x) - L_m(x) = f(x_1; \dots; x_{m+1})\,\omega_m(x). \tag{6}$$

Так как при малых $|x - x_k|$

$$f(x; x_1; \dots; x_m) \approx \frac{f^{(m)}(x)}{m!} \approx f(x_1; \dots; x_{m+1}),$$

то отсюда следует

$$f(x) - L_m(x) \approx L_{m+1}(x) - L_m(x). \tag{7}$$

Поэтому величину $\varepsilon_m = |L_{m+1}(x) - L_m(x)|$ можно рассматривать как приближенную оценку погрешности интерполяционной формулы $f(x) \approx L_m(x)$. Последовательно вычисляют значения $L_0(x)$, $L_1(x)$, ε_1, $L_2(x)$, $\varepsilon_2, \dots$; если при некотором m будет $\varepsilon_m \leqslant \varepsilon$, то вычисления прекращают и полагают

$$f(x) \approx L_m(x).$$

Если это неравенство не выполняется ни при каком m, то находят $\varepsilon_{m_0} = \min\limits_m \varepsilon_m$ и полагают $f(x) \approx L_{m_0}(x)$. Если этот минимум достигается при нескольких m, то среди них выбирают наименьшее. Если величины ε_m, начиная с некоторого m, имеют устойчивую тенденцию к увеличению, то с этого момента вычисление значений $L_m(x)$, ε_m прекращают.

§ 6. Разделенные разности и интерполирование с кратными узлами

Пусть требуется построить многочлен $g_s(x)$ степени $s - 1$, удовлетворяющий условиям

$$\begin{aligned} &g_s(x_1) = f(x_1), \quad \dots, \quad g_s^{(m_1-1)}(x_1) = f^{(m_1-1)}(x_1), \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ &g_s(x_n) = f(x_n), \quad \dots, \quad g_s^{(m_n-1)}(x_n) = f^{(m_n-1)}(x_n); \end{aligned} \tag{1}$$

здесь все x_i различные, $s = m_1 + \cdots + m_n$. Такой многочлен называют *интерполяционным многочленом с кратными узлами*, а числа $m_1, \ldots, m_n$ — *кратностями узлов* $x_1, \ldots, x_n$.

Интерполяционный многочлен $g_s(x)$ определяется единственным образом. В самом деле, предположим, что существуют два многочлена степени $s-1$, удовлетворяющих условиям (1). Тогда их разность $Q_s(x)$ удовлетворяет соотношениям

$$Q_s(x_1) = \cdots = Q_s^{(m_1-1)}(x_1) = 0, \ \ldots, \ Q_s(x_n) = \cdots = Q_s^{(m_1-1)}(x_n) = 0;$$

точки $x_1, \ldots, x_n$ являются нулями многочлена $Q_s(x)$ кратности $m_1, \ldots, m_n$ соответственно. Мы получили, что многочлен $Q_s(x) \not\equiv 0$ степени $s-1$ имеет s нулей. Следовательно, $Q_s(x) \equiv 0$.

Далее будем предполагать, что функция $f(x)$ непрерывно дифференцируема s раз. Существование интерполяционного многочлена $g_s(x)$, удовлетворяющего условиям (1), покажем, получив для него явное выражение.

Зададимся последовательностью совокупностей точек x_{ij}^{ε}, $0 < \varepsilon < \varepsilon_0$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, удовлетворяющих следующим условиям: при $0 < \varepsilon < \varepsilon_0$ все точки x_{ij}^{ε} различные, $x_{ij}^{\varepsilon} \to x_i$ при $\varepsilon \to 0$. В частности, можно положить $x_{ij}^{\varepsilon} = x_i + (j-1)\varepsilon$.

Построим интерполяционный многочлен $g_s^{\varepsilon}(x)$ степени $s-1$, совпадающий с $f(x)$ в точках x_{ij}^{ε}. Таблица разделенных разностей, соответствующих этому набору узлов, имеет вид

$$\begin{array}{ll}
f(x_{11}^{\varepsilon}) & \\
 & f(x_{11}^{\varepsilon}; x_{12}^{\varepsilon}) \\
f(x_{12}^{\varepsilon}) & \\
 & f(x_{12}^{\varepsilon}; x_{13}^{\varepsilon}) \\
f(x_{13}^{\varepsilon}) & \\
\vdots & \\
f(x_{1m_1}^{\varepsilon}) & \\
 & f(x_{1m_1}^{\varepsilon}; x_{21}^{\varepsilon}) \\
f(x_{21}^{\varepsilon}) & \\
\vdots & \\
f(x_{nm_n}^{\varepsilon}) &
\end{array}
\qquad
\begin{array}{l}
f(x_{11}^{\varepsilon}; x_{12}^{\varepsilon}; x_{13}^{\varepsilon}) \\
\vdots \\
f(x_{11}^{\varepsilon}; x_{12}^{\varepsilon}; \ldots; x_{nm_n}^{\varepsilon}).
\end{array} \tag{2}$$

Выпишем интерполяционную формулу Ньютона с разделенными разностями:

$$g_s^{\varepsilon}(x) = A_0^{\varepsilon} + A_1^{\varepsilon}(x - x_{11}^{\varepsilon}) + A_2^{\varepsilon}(x - x_{11}^{\varepsilon})(x - x_{12}^{\varepsilon}) + \cdots$$
$$\cdots + A_{s-1}^{\varepsilon}(x - x_{11}^{\varepsilon}) \ldots (x - x_{n,m_n-1}^{\varepsilon}),$$

где

$$A_0^{\varepsilon} = f(x_{11}^{\varepsilon}), \quad A_1^{\varepsilon} = f(x_{11}^{\varepsilon}; x_{12}^{\varepsilon}),$$
$$A_2^{\varepsilon} = f(x_{11}^{\varepsilon}; x_{12}^{\varepsilon}; x_{13}^{\varepsilon}), \ldots \quad , A_{s-1}^{\varepsilon} = f(x_{11}^{\varepsilon}; x_{12}^{\varepsilon}; \ldots; x_{n,m_n}^{\varepsilon}).$$

Выражая разделенные разности через производные, имеем

$$f(x_{il}^{\varepsilon};\dots;x_{im}^{\varepsilon})=\frac{f^{(m-l)}(\bar{x}_{il_m}^{\varepsilon})}{(m-l)!}.$$

Переходя к пределу при $\varepsilon\to 0$, получаем

$$\lim_{\varepsilon\to 0} f(x_{il}^{\varepsilon};\dots;x_{im}^{\varepsilon})=\frac{f^{(m-l)}(x_i)}{(m-l)!}. \tag{3}$$

Таким образом, из наших рассуждений следует, что все разделенные разности в таблице (2) вида $f(x_{il}^{\varepsilon};\dots;x_{im}^{\varepsilon})$ при $\varepsilon\to 0$ имеют пределы, которые естественно обозначать $f(\underbrace{x_i;\dots;x_i}_{m-l+1\text{ раз}})$. Из (3) следует, что

$$f(\underbrace{x_i;\dots;x_i}_{p+1\text{ раз}})=\frac{f^{(p)}(x_i)}{p!}. \tag{4}$$

Задача 1. Индукцией по порядку разности показать, что все разделенные разности, входящие в таблицу (2), имеют конечные пределы.

Если все элементы таблицы (2) имеют пределы, то на любом отрезке многочлены $g_s^{\varepsilon}(x)$ при $\varepsilon\to 0$ стремятся к некоторому многочлену

$$\begin{aligned} g_s(x)&=A_0+A_1(x-x_1)+A_2(x-x_1)^2+\cdots\\ &\quad\cdots+A_{s-1}(x-x_1)^{m_1}\dots(x-x_{n-1})^{m_{n-1}}(x-x_n)^{m_n-1}=\\ &=f(x_1)+f(x_1;\,x_1)(x-x_1)+f(x_1;\,x_1;\,x_1)(x-x_1)^2+\cdots; \end{aligned} \tag{5}$$

$$A_i=\lim_{\varepsilon\to 0}A_i^{\varepsilon}.$$

Многочлен $g_s(x)$ записывается в виде

$$g_s(x)=\sum_{i=1}^{m_1}\frac{f^{(i-1)}(x_1)}{(i-1)!}(x-x_1)^{i-1}+O((x-x_1)^{m_1}).$$

Отсюда вытекает, что он удовлетворяет условиям, заданным в точке x_1. Вследствие единственности интерполяционного многочлена многочлен $g_s^{\varepsilon}(x)$ не изменится при переобозначении $x_1=x_j$, $x_j=x_1$. Поэтому предельный многочлен будет удовлетворять заданным условиям в любой точке x_j. Следовательно, этот многочлен является искомым.

Задача 2. Доказать равенство

$$f(x)-g_s(x)=\frac{f^{(s)}(\zeta)}{s!}\omega_s(x),\quad \omega_s(x)=\prod_{i=1}^{n}(x-x_i)^{m_i},\quad y_1\leqslant\zeta\leqslant y_2, \tag{6}$$

где $y_1 = \min\{x, x_1, \ldots, x_n\}$, $y_2 = \max\{x, x_1, \ldots, x_n\}$.

Согласно (5.1) справедливо равенство

$$f(x) - g_s^\varepsilon(x) = f(x; x_{11}^\varepsilon; \ldots; x_{nm_n}^\varepsilon)\,\omega_s^\varepsilon(x),$$

где $\omega_s^\varepsilon(x) = \omega_s(x^\varepsilon)$. Переходя к пределу при $\varepsilon \to 0$, получим

$$f(x) - g_s(x) = f(x; x_1; \ldots; x_n)\,\omega_s(x).$$

Сравнивая это равенство с (6), имеем

$$f(x; x_1; \ldots; x_n) = \frac{f^{(s)}(\zeta)}{s!}.$$

Это соотношение остается в силе при предельном переходе $x \to x_j$, j — любое. Из этих соотношений следует, что формула (5.4)

$$f(x_1; \ldots; x_{N+1}) = \frac{f^{(N)}(\zeta)}{N!}$$

(переписанная в других обозначениях) справедлива и в случае, когда не все $x_1, \ldots, x_{N+1}$ — различные.

Мы доказали существование интерполяционного многочлена, удовлетворяющего условиям (1). Задачу интерполяции можно было бы поставить и таким образом.

Задана таблица чисел a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. Требуется построить многочлен $g_s(x)$ степени $s - 1$, удовлетворяющий условиям

$$g_s^{(j-1)}(x_i) = a_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i.$$

Эта задача равносильна исходной, поскольку всегда можно указать гладкую функцию $f(x)$ такую, что

$$f^{(j-1)}(x_i) = a_{ij}.$$

§ 7. Уравнения в конечных разностях

Конечно-разностными уравнениями называют уравнения относительно функций дискретного переменного. Такие уравнения, в частности, возникают при аппроксимации обыкновенных и многомерных дифференциальных уравнений.

Существует глубокая аналогия между непрерывными и дискретными случаями. В частности, справедливы разностные аналоги формул Грина; если в некоторой задаче применим метод Фурье, то в отношении соответствующей разностной задачи применим дискретный вариант метода Фурье. Практически каждому интегральному тождеству в теории дифференциальных уравнений

можно поставить в соответствие некоторый дискретный вариант. В руках квалифицированного математика методы решения конечно-разностных уравнений являются мощнейшим средством исследования чувствительности («устойчивости») алгоритмов к вычислительной погрешности. Если требуется исследовать алгоритм решения некоторой задачи, то подбирают близкую по структуре задачу (например, следуя *принципу замороженных коэффициентов* (см. гл. 10)), для которой решение соответствующей конечно-разностной задачи находится в явном виде. Анализируя алгоритм решения исходной задачи на примере этой конечно-разностной задачи, выносят предварительное суждение о его свойствах. Как правило, при практическом решении задач в большинстве случаев полученное на таком пути предварительное суждение дает правильное представление о свойствах алгоритма.

Непосредственно конечно-разностные уравнения потребуются нам в следующем параграфе при описании многочленов Чебышева. Ниже будет проведена аналогия между конечно-разностными уравнениями одного дискретного переменного и обыкновенными дифференциальными уравнениями.

Рассмотрим простейший случай одного линейного уравнения относительно неизвестной функции одного целочисленного аргумента

$$ly = \sum_{i=0}^{k} a_i(n)y(n+i) = f(n). \tag{1}$$

Это уравнение называется *линейным разностным уравнением* k-го порядка и является разностным аналогом линейного дифференциального уравнения k-го порядка

$$\bar{l}y = \sum_{i=0}^{k} b_i(x)y^{(i)}(x) = f(x). \tag{2}$$

Каждое из уравнений (1) и (2) имеет вид $Ly = h$, где L — линейный оператор. Уравнение $Ly = 0$ называют *однородным*; формулы

$$y = y(C_1, \ldots, C_l, n) \quad \text{или} \quad y = y(C_1, \ldots, C_l, x)$$

дают *общее решение* уравнения (1) или (2), если при подстановке значений параметров C_i можно получить любое решение рассматриваемого уравнения. Если v — частное решение неоднородного уравнения $Lv = h$, то разность $y - v$ является решением однородного уравнения $L(y - v) = h - h = 0$. Таким образом, общее решение неоднородного уравнения представимо в виде суммы частного решения неоднородного уравнения и общего решения однородного уравнения. Решения $y_1, \ldots, y_m$ однородного уравнения $Ly = 0$ называют *линейно зависимыми* в рассматриваемой области изменения независимого аргумента, если существуют $C_1, \ldots, C_m$, не все равные нулю, при которых $C_1y_1 + \cdots + C_my_m \equiv 0$. В противном случае эти решения называют *линейно независимыми*. Если функции y_i

являются решениями однородного уравнения $Ly = 0$, то любая функция $\sum_i C_i y_i$ также является решением этого уравнения, поскольку

$$L\left(\sum_i C_i y_i\right) = \sum_i C_i\, Ly_i = 0.$$

Проводимое далее параллельное рассмотрение уравнений (1), (2) подчеркивает общие черты этих уравнений и помогает найти путь исследования уравнения (1) по аналогии с уравнением (2).

Пусть для определенности уравнение (2) рассматривается в области $x \geqslant 0$, а уравнение (1) — в области $n \geqslant 0$.

Теорема.

Пусть $b_k(x) \neq 0$ *при* $x \geqslant 0$ *и все* $b_i(x)$ *непрерывны при* $x \geqslant 0$.

Пусть $a_k(n) \neq 0$ *при* $n \geqslant 0$.

Тогда общее решение однородного уравнения $Ly = 0$ *записывается в виде*

$$y = \sum_{i=1}^{k} C_i y_i,$$

где $y_1, \ldots, y_k$ — *линейно независимые решения уравнения* $Ly = 0$.

Доказательство.

Согласно теореме существования уравнение $ly = 0$ имеет решение при любых начальных условиях $y(0), \ldots, y^{(k-1)}(0)$.

Однородное уравнение $ly = 0$ можно представить в виде

$$y(n+k) = -\sum_{i=0}^{k-1} \frac{a_i(n)}{a_k(n)} y(n+i). \quad (3)$$

Если мы зададимся $y(0), \ldots, y(k-1)$, то из (3) сможем вычислить последовательно $y(k)$, $y(k+1), \ldots$. Таким образом, при любых $y(0), \ldots, y(k-1)$ уравнение $l(y) = 0$ имеет решение.

Согласно теореме единственности это решение единственно.

Это решение единственно, поскольку значения любого решения удовлетворяют уравнению (3), а из этого уравнения значения $y(k)$, $y(k+1), \ldots$ определяются однозначно.

Обозначим через $y_i(x)$ решения уравнения $\bar{l}y = 0$ при начальных условиях $y_i^{(j-1)} = \delta_i^j$, $i, j = 1, \ldots, k$.

Обозначим через $y_i(n)$ решения уравнения $ly = 0$ при начальных условиях $y_i(j-1) = \delta_i^j$, $i, j = 1, \ldots, k$.

Эти решения образуют линейно независимую систему. В самом деле, если

$$\sum_{i=1}^{k} C_i y_i(x) \equiv 0,$$

то при $j = 1, \ldots, k$ имеем

$$0 = \sum_{i=1}^{k} C_i y_i^{(j-1)}(0) = \sum_{i=1}^{k} C_i \delta_i^j = C_j.$$

$$\sum_{i=1}^{k} C_i y_i(n) \equiv 0,$$

то при $j = 1, \ldots, k$ имеем

$$0 = \sum_{i=1}^{k} C_i y_i(j-1) = \sum_{i=1}^{k} C_i \delta_i^j = C_j.$$

Следовательно, в случае

$$\sum_{i=1}^{k} C_i y_i \equiv 0$$

все C_i обязательно равны нулю, поэтому функции $y_1, \ldots, y_k$ линейно независимы.

Пусть $y(x)$ — какое-либо решение уравнения $\bar{l}y = 0$. Функция

$$z(x) = \sum_{j=1}^{k} y^{(j-1)}(0) y_j(x)$$

Пусть $y(n)$ — какое-либо решение уравнения $ly = 0$. Функция

$$z(n) = \sum_{j=1}^{k} y(j-1) y_j(n)$$

является решением этого уравнения при начальных условиях

$$y(0), \ldots, y^{(k-1)}(0).$$

$$y(0), \ldots, y(k-1).$$

Вследствие единственности решения уравнения

$\bar{l}y = 0$ имеем

$$y(x) = \sum_{j=1}^{k} y^{(j-1)}(0) y_j(x).$$

$ly = 0$ имеем

$$y(n) = \sum_{j=1}^{k} y(j-1) y_j(n).$$

Теорема доказана.

Далее в курсе дифференциальных уравнений устанавливается следующий факт. Если известны k линейно независимых решений однородного уравнения $\bar{l}y = 0$, то нахождение решения неоднородного уравнения (2) сводится к решению уравнений

$$dC_j/dx = g_j(x), \qquad (4)$$

где $g_j(x)$ — известные функции, т. е. к отысканию квадратур. Точно так же в случае, когда известны k линейно не независимых решений однородного урав-

нения $ly = 0$, нахождение решения неоднородного уравнения сводится к решению аналогичных (4) разностных уравнений

$$C_j(n+1) - C_j(n) = g_j(n),$$

где $g_j(n)$ — известные функции.

Перейдем к рассмотрению уравнений с постоянными коэффициентами

$$\bar{l}y = \sum_{i=0}^{k} b_i y^{(i)}(x) = f(x), \quad b_k \neq 0, \qquad (5)$$

$$ly = \sum_{i=0}^{k} a_i y(n+i) = f(n), \quad a_k \neq 0, \qquad (6)$$

и соответствующих однородных уравнений

$$\bar{l}y = \sum_{i=0}^{k} b_i y^{(i)}(x) = 0. \qquad (7)$$

$$ly = \sum_{i=0}^{k} a_i y(n+i) = 0. \qquad (8)$$

Займемся отысканием частных решений однородного уравнения.

Подставляя в (7) предполагаемый вид частного решения $\exp(\lambda x)$, получаем уравнение

$$\left(\sum_{i=0}^{k} b_i \lambda^i\right) \exp(\lambda x) = 0.$$

В случае уравнения (8) функцию $\exp(\lambda n)$ удобно записывать в виде μ^n, $\mu = \exp\lambda$. Подставляя ее в (8), получаем уравнение

$$\left(\sum_{i=0}^{k} a_i \mu^i\right) \mu^n = 0.$$

Таким образом, каждому корню уравнения

$$\sum_{i=0}^{k} b_i \lambda^i = 0, \qquad (9)$$

$$\sum_{i=0}^{k} a_i \mu^i = 0, \qquad (10)$$

называемому *характеристическим*, соответствует частное решение

$$\exp(\lambda x).$$

$$\mu^n.$$

Если все корни характеристического уравнения простые, мы получаем k различных решений. Покажем, что каждому s-кратному корню характеристического уравнения соответствуют s различных решений однородного уравнения

$$\exp(\lambda x),\ x\exp(\lambda x), \ldots, x^{s-1}\exp(\lambda x).$$

$$\mu^n,\ C_n^1\mu^{n-1}, \ldots, C_n^{s-1}\mu^{n-s+1}.$$

Пусть для определенности $\lambda_1 = \cdots = \lambda_s$, $\mu_1 = \cdots = \mu_s$. Разложим на множители характеристический многочлен

$$\sum_{i=0}^{k} b_i\lambda^i = b_k\prod_{j=1}^{k}(\lambda-\lambda_j). \qquad \sum_{i=0}^{k} a_i\mu^i = a_k\prod_{j=1}^{k}(\mu-\mu_j).$$

Зададимся действительным параметром $\varepsilon > 0$, $\varepsilon \to 0$.

Возьмем $\lambda_{j\varepsilon}$ такие, что:
а) $\lambda_{j\varepsilon}$ различны при $j = 1, \ldots, s$;
б) стремятся к λ_j при $\varepsilon \to 0$ для всех $j \leqslant k$.

Возьмем $\mu_{j\varepsilon}$ такие, что:
а) $\mu_{j\varepsilon}$ различны при $j = 1, \ldots, s$;
б) стремятся к μ_j при $\varepsilon \to 0$ для всех $j \leqslant k$.

Образуем характеристические уравнения, соответствующие этим корням:

$$0 = b_k\prod_{j=1}^{k}(\lambda-\lambda_{j\varepsilon}) = \sum_{i=0}^{k} b_{i\varepsilon}\lambda^i. \qquad 0 = a_k\prod_{j=1}^{k}(\mu-\mu_{j\varepsilon}) = \sum_{i=0}^{k} a_{i\varepsilon}\mu^i.$$

Ясно, что

$$b_{i\varepsilon} \to b_i \qquad a_{i\varepsilon} \to a_i \qquad (11)$$

при $\varepsilon \to 0$.

Этим характеристическим уравнениям соответствуют уравнения

$$\sum_{i=0}^{k} b_{i\varepsilon}y_\varepsilon^{(i)}(x) = 0. \qquad (12) \qquad \sum_{i=0}^{k} a_{i\varepsilon}y_\varepsilon(n+i) = 0. \qquad (13)$$

Пусть при $\varepsilon > 0$ мы можем указать решения

$y_\varepsilon(x)$ уравнения (12) такие, что при любом $x \geqslant 0$ существует предел

$$\lim_{\varepsilon\to 0} y_\varepsilon(x) = Y(x),$$

причем $y_\varepsilon(x)$ сходится к $Y(x)$ равномерно вместе со всеми производными до порядка k включительно на любом конечном отрезке $[x_1, x_2]$. Переходя к пределу в (12) с учетом (11), получаем, что предельная функция $Y(x)$ удовлетворяет уравнению (7).

$y_\varepsilon(n)$ уравнения (13) такие, что при любом $n \geqslant 0$ существует предел

$$\lim_{\varepsilon\to 0} y_\varepsilon(n) = Y(n).$$

Переходя к пределу в (13) с учетом (11), получаем, что предельная функция $Y(n)$ удовлетворяет уравнению (8).

Построим такие последовательности $y_\varepsilon(x)$ и $y_\varepsilon(n)$, которые будут сходиться к частным решениям (7), (8), соответствующим кратным корням.

При проведении этих построений удобно использовать разделенные разности. Рассмотрим сначала случай двукратного корня. Положим

$$y_{2\varepsilon}(x) = \exp(\lambda x)(\lambda_{1\varepsilon};\, \lambda_{2\varepsilon}) = \frac{\exp(\lambda_{2\varepsilon}x) - \exp(\lambda_{1\varepsilon}x)}{\lambda_{2\varepsilon} - \lambda_{1\varepsilon}}. \qquad y_{2\varepsilon}(n) = \mu^n(\mu_{1\varepsilon};\, \mu_{2\varepsilon}) = \frac{\mu_{2\varepsilon}^n - \mu_{1\varepsilon}^n}{\mu_{2\varepsilon} - \mu_{1\varepsilon}}.$$

Эти функции являются решениями соответственно уравнений (12), (13). Запишем их в виде

$$y_{2\varepsilon}(x) = x\exp(\lambda_{1\varepsilon}x) \times \frac{\exp((\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x) - 1}{(\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x}. \qquad y_{2\varepsilon}(n) = \mu_{2\varepsilon}^{n-1} + \mu_{2\varepsilon}^{n-2}\mu_{1\varepsilon} + \cdots + \mu_{1\varepsilon}^{n-1}. \tag{14}$$

Переходя к пределу при $\varepsilon \to 0$, получим

$$y_{2\varepsilon}(x) \to x\exp(\lambda_1 x). \qquad y_{2\varepsilon}(n) \to \mu_1^{n-1} + \cdots + \mu_1^{n-1} = n\mu_1^{n-1}.$$

В результате мы построили второе линейно независимое решение, соответствующее двукратному корню.

Случай корня более высокой кратности рассмотрим лишь для уравнения (1). Согласно (4.3) имеем

$$y_{q\varepsilon} = \mu^n(\mu_{1\varepsilon}; \ldots; \mu_{q\varepsilon}) = \sum_{j=1}^{q} \frac{\mu_{j\varepsilon}^n}{\prod\limits_{i\neq j}(\mu_{j\varepsilon} - \mu_{i\varepsilon})}.$$

Как линейная комбинация функций $\mu_{j\varepsilon}^n$, функция $y_{q\varepsilon}$ является решением уравнения (13). Аналогично (14) непосредственно устанавливается, что

$$y_{q\varepsilon} = \sum_{n_1+\cdots+n_q=n+1-q} \mu_{1\varepsilon}^{n_1}\mu_{2\varepsilon}^{n_2}\ldots\mu_{q\varepsilon}^{n_q}.$$

Общее число слагаемых равно C_n^{q-1}, поэтому

$$y_{q\varepsilon} \to Y_q(n) = C_n^{q-1}\mu^{n+1-q}.$$

Поскольку в случае s-кратного корня можно взять $q = 1, \ldots, s$, то получилось s частных решений

$$Y_1(n) = \mu^n,\ Y_2(n) = C_n^1\mu^{n-1}, \ldots,\ Y_s(n) = C_n^{s-1}\mu^{n+1-s}. \tag{15}$$

Задача 1. Доказать, что совокупность частных решений (15), соответствующих корням характеристического уравнения (10), образует фундаментальную систему (т. е. они линейно независимы и решение (8) может быть получено как линейная комбинация таких решений).

Задача 2. Пусть $P_{s-1}(n)$ — произвольный многочлен степени $s-1$. Доказать, что функция $P_{s-1}(n)\mu^n$ записывается как линейная комбинация функций (15)

$$P_{s-1}(n)\mu^n = \sum_{j=1}^{s} C_j Y_j(n).$$

Таким образом, вместо системы решений (15) можно взять систему решений

$$Y_1(n) = \mu_1^n, \quad Y_2(n) = n\mu_1^n, \quad \ldots, \quad Y_s(n) = n^{s-1}\mu_1^n.$$

Задача 3. Показать, что уравнение (16) имеет частное решение вида

$$\left(\sum_{j=s}^{s+m-1} d_j n^j\right)\sigma^n,$$

где d_j могут быть найдены методом неопределенных коэффициентов.

Рассмотрим теперь разностное уравнение

$$\sum_{i=0}^{k} a_i y(n+i) = \left(\sum_{j=0}^{m-1} C_j n^j\right)\sigma^n. \tag{16}$$

Пусть σ является корнем характеристического уравнения (10) кратности s; в частности, если σ не является корнем этого уравнения, то $s=0$.

§ 8. Многочлены Чебышева

Рассматриваемые ниже многочлены Чебышева играют фундаментальную роль в теории и практике использования численных методов. С их помощью решается значительная часть задач оптимизации свойств вычислительных алгоритмов. Запись многочленов в традиционной форме часто приводит к большому влиянию вычислительной погрешности, и в этих случаях их целесообразнее записывать в виде линейных комбинаций многочленов Чебышева.

Многочлены Чебышева $T_n(x)$, где $n \geqslant 0$, определяются соотношениями

$$\begin{aligned} &T_0(x) = 1, \quad T_1(x) = x, \\ &T_{n+1} = 2xT_n(x) - T_{n-1}(x) \quad \text{при} \quad n > 0. \end{aligned} \tag{1}$$

Пользуясь рекуррентной формулой (1), получаем, например,

$$T_2(x) = 2x^2 - 1, \qquad T_3(x) = 4x^3 - 3x,$$
$$T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \ldots .$$

Старший член многочлена $T_{n+1}(x)$ получается из старшего члена многочлена $T_n(x)$ умножением на $2x$, и, следовательно, старший член в $T_n(x)$ при $n > 0$ есть $2^{n-1}x^n$.

Все многочлены $T_{2n}(x)$ являются четными функциями, а $T_{2n+1}(x)$ — нечетными.

При $n = 0$ это утверждение верно. Предположив его справедливость при некотором n, мы получим, что $2xT_{2n+1}(x)$ — четная функция и, вследствие (1), $T_{2n+2}(x)$ — тоже четная функция. Тогда $2xT_{2n+2}(x)$ и $T_{2n+3}(x)$, вследствие (1), — нечетные функции.

При любом θ имеем

$$\cos((n+1)\theta) = 2\cos\theta\cos n\theta - \cos((n-1)\theta).$$

Полагая $\theta = \arccos x$, получим

$$\cos((n+1)\arccos x) = 2x\cos(n\arccos x) - \cos((n-1)\arccos x).$$

Функция $\cos(n\arccos x)$ удовлетворяет тому же разностному уравнению (1) по переменной n, что и $T_n(x)$. Начальные условия при $n = 0$ и $n = 1$ одни и те же:

$$\cos(0 \cdot \arccos x) = 1 = T_0(x), \quad \cos(1 \cdot \arccos x) = x = T_1(x);$$

поэтому при всех n

$$T_n(x) = \cos(n\arccos x). \tag{2}$$

Следовательно,

$$|T_n(x)| \leqslant 1 \quad \text{при} \quad |x| \leqslant 1. \tag{3}$$

Не нужно думать, что $|T_n(x)| \leqslant 1$ при всех вещественных x. Если $|x| \geqslant 1$, то $\arccos x$ не является действительным числом, а косинус такого числа больше 1.

Рекуррентное соотношение (1) является разностным уравнением; ему соответствует характеристическое уравнение

$$\mu^2 - 2\mu x + 1 = 0$$

с корнями

$$\mu_{1,2} = x \pm \sqrt{x^2 - 1}.$$

При $x \neq \pm 1$ корни простые, поэтому

$$T_n(x) = c_1(x)\mu_1^n + c_2(x)\mu_2^n.$$

Из начальных условий $T_0(x) = 1$, $T_1(x) = x$ получаем $c_1 = c_2 = 1/2$; таким образом,

$$T_n(x) = \frac{(x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n}{2}. \tag{4}$$

Задача 1. Проверить справедливость этой формулы при $x = 1$ и $x = -1$.

Из уравнения

$$T_n(x) = \cos(n \arccos x) = 0$$

получаем, что

$$x_m = \cos\left(\frac{\pi(2m-1)}{2n}\right), \quad m = 1, \ldots, n,$$

— нули $T_n(x)$. Вследствие (2), (3) точками экстремума $T_n(x)$ на $[-1,1]$ будут точки, где $|T_n(x)| = 1$. Решая это уравнение, получим

$$x_{(m)} = \cos\left(\frac{\pi m}{n}\right), \quad m = 0, \ldots, n,$$

причем

$$T_n(x_{(m)}) = \cos \pi m = (-1)^m.$$

Многочлены

$$\overline{T}_n(x) = 2^{1-n} T_n(x) = x^n + \cdots$$

называют *многочленами, наименее уклоняющимися от нуля*. Это определение объясняется следующим свойством.

Лемма. *Если $P_n(x)$ — многочлен степени n со старшим коэффициентом 1, то*

$$\max_{[-1,1]} |P_n(x)| \geqslant \max_{[-1,1]} |\overline{T}_n(x)| = 2^{1-n}. \tag{5}$$

Доказательство. Предположим противное. Многочлен $\overline{T}_n(x) - P_n(x)$ имеет степень $n-1$; в то же время

$$\operatorname{sign}(\overline{T}_n(x_{(m)}) - P_n(x_{(m)})) = \operatorname{sign}((-1)^m 2^{1-n} - P_n(x_{(m)})) = (-1)^m,$$

так как, согласно предположению, $|P_n(x_{(m)})| < 2^{1-n}$ при всех m. Таким образом, между каждыми двумя точками $x_{(m)}$, $x_{(m+1)}$ многочлен $\overline{T}_n(x) - P_n(x)$ меняет знак. Многочлен $\overline{T}_n(x) - P_n(x)$ степени $n-1$, отличный от нуля (поскольку он отличен от нуля в точках $x_{(m)}$), имеет n различных нулей. Мы пришли к противоречию.

Задача 2. Доказать более сильное утверждение: если

$$P_n(x) = x^n + \cdots \neq \overline{T}_n(x),$$

то

$$\max_{[-1,1]} |P_n(x)| > 2^{1-n}.$$

Линейной заменой переменных $x' = \frac{b+a}{2} + \frac{b-a}{2}x$ отрезок $[-1,1]$ можно перевести в заданный отрезок $[a,b]$. В многочлене $\overline{T}_n\left(\frac{2x-(b+a)}{b-a}\right)$ старший коэффициент равен $(2/(b-a))^n$. В соответствии с леммой можно утверждать, что многочлен

$$\overline{T}_n^{[a,b]}(x) = (b-a)^n\, 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

со старшим коэффициентом 1 является многочленом, наименее уклоняющимся от нуля на отрезке $[a,b]$. Это означает, что для любого многочлена $P_n(x)$ степени n со старшим коэффициентом 1 справедливо неравенство

$$\max_{[a,b]} |P_n(x)| \geqslant \max_{[a,b]} \left|\overline{T}_n^{[a,b]}(x)\right| = (b-a)^n\, 2^{1-2n}. \tag{6}$$

Нетрудно проверить, что нулями многочлена $\overline{T}_n^{[a,b]}(x)$ являются точки

$$x_m = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{\pi(2m-1)}{2n}\right), \quad m = 1, \ldots, n.$$

Многочлены $\widetilde{T}_0(x) = 1/\sqrt{2}$, $\widetilde{T}_n(x) = T_n(x)$ при $n \geqslant 1$ образуют на $[-1,1]$ ортонормированную систему с весом $2/(\pi\sqrt{1-x^2})$. Проверим свойство ортонормированности этих функций. После замены $x = \cos\theta$ имеем

$$\int_{-1}^{1} \frac{2T_n(x)T_m(x)}{\pi\sqrt{1-x^2}}\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} \cos n\theta \cos m\theta\, d\theta =$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \big(\cos((n-m)\theta) + \cos((n+m)\theta)\big)\, d\theta = \delta_m^n + \delta_m^{-n}.$$

Второе слагаемое обращается в нуль при $n, m \geqslant 0$, если $n^2 + m^2 \neq 0$. Отсюда вытекает требуемое равенство

$$\int_{-1}^{1} \frac{2\widetilde{T}_n(x)\widetilde{T}_m(x)}{\pi\sqrt{1-x^2}}\,dx = \delta_m^n.$$

Пусть многочлены Чебышева вычисляются по рекуррентной формуле (1). В процессе реальных вычислений вместо значений $T_n(x)$ получаем приближенные значения T_n^*, удовлетворяющие соотношениям

$$T_0^* = 1, \quad T_1 = x + \delta_1, \quad T_{n+1}^* = 2xT_n^* - T_{n-1}^* + \delta_{n+1},$$

где δ_k — погрешности, вызванные округлениями.

Задача 3. Получить представление погрешности

$$T_N^*(x) - T_N(x) = \sum_{k=1}^{N} \delta_k \frac{\sin((N+1-k)\arccos x)}{\sqrt{1-x^2}}.$$

Задача 4. Получить, используя решение задачи 2, оценку погрешности

$$|T_N^*(x) - T_N(x)| \leqslant \max_{1\leqslant k\leqslant N} |\delta_k| \cdot N \cdot \min\left\{N, \frac{1}{\sqrt{1-x^2}}\right\}. \tag{7}$$

Задача 5. Непосредственной проверкой убедиться, что в нулях многочленов Чебышева $x_m = \cos\left(\frac{\pi(2m-1)}{2N}\right)$ справедливо равенство

$$|T_N'(x_m)| = \frac{N}{\sqrt{1-x_m^2}}.$$

Отсюда заключаем, что при задании x в окрестности этих нулей с погрешностью δ погрешность вычисления значения $T_n(x)$ будет величиной, близкой к $\frac{n\delta}{\sqrt{1-x_m^2}}$; это означает, что оценка (7) не может быть существенно улучшена.

Задача 6. Пусть N — некоторое фиксированное число. Доказать, что векторы, образованные значениями многочленов $\overline{T}_n(x)$, $n < N$ в нулях $T_N(x)$, образуют некоторую ортогональную систему, а именно

$$\frac{2}{N}\sum_{j=1}^{N} \widetilde{T}_n\left(\frac{\pi(2j-1)}{2N}\right)\widetilde{T}_m\left(\frac{\pi(2j-1)}{2N}\right) = \delta_n^m \quad \text{при} \quad 0 \leqslant n,\ m < N-1. \tag{8}$$

§ 9. Минимизация оценки остаточного члена интерполяционной формулы

Пусть функция $f(x)$ приближается на $[a,b]$ с помощью интерполяционного многочлена степени $n-1$ с узлами интерполяции $x_1, \ldots, x_n \in [a,b]$. Согласно (3.1) имеем

$$f(x) - L_n(x) = \frac{f^{(n)}(\zeta)\omega_n(x)}{n!},$$

где $\zeta \in [a,b]$, если $x \in [a,b]$. Отсюда следует оценка погрешности интерполяции

$$\|f - L_n\| \leqslant \frac{\|f^{(n)}\| \cdot \|\omega_n\|}{n!}. \tag{1}$$

Здесь $\|\cdot\|$ есть обозначение равномерной нормы: $\|g(x)\| = \sup\limits_{[a,b]} |g(x)|$.

Займемся минимизацией правой части оценки (1) за счет выбора узлов $x_1, \ldots, x_n$. Многочлен $\omega_n(x) = (x-x_1)\ldots(x-x_n)$ имеет старший коэффициент 1, поэтому $\|\omega_n\| \geqslant (b-a)^n 2^{1-2n}$ согласно (8.6). Если взять в качестве узлов интерполяции

$$x_m = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{\pi(2m-1)}{2n}\right), \quad m = 1, \ldots, n, \tag{2}$$

то

$$\omega_n(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

и

$$\|\omega_n\| = (b-a)^n 2^{1-2n}.$$

Следовательно, при таком расположении узлов справедлива наилучшая из оценок, которая может быть получена как следствие оценки (1):

$$\|f - L_n\| \leqslant \frac{\|f^{(n)}\|(b-a)^n 2^{1-2n}}{n!}. \tag{3}$$

При получении оценки (1) максимум произведения заменен на произведение максимумов сомножителей. Поэтому может возникнуть надежда получить оценку погрешности, лучшую, чем (3). Однако это не так. Если $f(x) = a_n x^n + \cdots + a_0$ — многочлен степени n, то

$$f^{(n)}(\zeta) = a_n n! = \text{const},$$

поэтому неравенство (1) превращается в равенство; тогда, вследствие (8.6), при любых узлах интерполяции имеем

$$\|f - L_n\| \geqslant \frac{\|f^{(n)}\|(b-a)^n 2^{1-2n}}{n!} = |a_n|(b-a)^n 2^{1-2n}.$$

Как уже отмечалось, важной проблемой вычислительной математики является проблема оптимизации методов решения задач некоторого класса. Общая постановка ее такова.

Задается некоторый класс P решаемых задач p. Задается некоторое множество M методов решения. Пусть $e(p, m)$ — погрешность метода m при решении задачи p. Величину

$$e(P, m) = \sup_{p \in P} e(p, m)$$

называют *погрешностью метода на классе задач* P. Величину

$$e(P, M) = \inf_{m \in M} e(P, m)$$

называют *оптимальной оценкой погрешности методов из множества* M *на классе задач* P. Если существует метод $m \in M$, на котором эта оценка достигается, т. е. $e(P, M) = e(P, m)$, то такой метод называют *оптимальным*.

Полученное нами решение задачи об оптимизации узлов распределения интерполяционной формулы можно сформулировать в описанных выше терминах.

Пусть P — множество задач приближения функций, определенных на $[a, b]$ и удовлетворяющих условию $|f^{(n)}(x)| \leqslant A_n$. Пусть M — множество методов приближения, состоящих в том, что функция заменяется ее интерполяционным многочленом $L_n(x)$ по совокупности узлов $x_1, \ldots, x_n$;

таким образом, метод решения m определяется заданием узлов интерполяции $x_1, \ldots, x_n$. Наконец, пусть мера погрешности $e(p, m) = \|f - L_n\|$. Согласно (1) имеем

$$e(P, m) \leqslant \frac{A_n \|\omega_n\|}{n!}.$$

С другой стороны, для задачи приближения многочлена

$$P_{n+1}(x) = \frac{A_n}{n!} x^n + \cdots,$$

относящейся к рассматриваемому классу, имеем

$$e(p, m) = \frac{A_n \|\omega_n\|}{n!}.$$

Следовательно,

$$e(P, m) = \frac{A_n \|\omega_n\|}{n!}.$$

Как мы видели выше,

$$e(P, M) = \inf_m e(P, m) = \inf_{x_1, \ldots, x_n} \left(\frac{A_n \|\omega_n\|}{n!} \right) = \frac{A_n (b-a)^n 2^{1-2n}}{n!}.$$

Таким образом, способ интерполяции по узлам многочлена Чебышева (2) является оптимальным в рассматриваемом смысле.

В заключение произведем сравнение оценки (3) с оценкой погрешности разложения функции в ряд Тейлора. Согласно результатам § 6 отрезок ряда Тейлора

$$t_n(x) = \sum_{j=0}^{n-1} \frac{f^{(j)}((a+b)/2)}{j!} \left(x - \frac{a+b}{2} \right)^j$$

совпадает с интерполяционным многочленом Лагранжа при единственном, n-кратном узле интерполяции $(a+b)/2$. Поэтому, естественно, при оптимальном распределении узлов интерполяции (2) мы должны иметь лучшую оценку. В самом деле, оценка

$$\|f - t_n\| \leqslant \frac{\|f^{(n)}\| (b-a)^n 2^{-n}}{n!}$$

погрешности отрезка ряда Тейлора уступает оценке (3) в 2^{n-1} раз.

Приведем для сведения оценки погрешности интерполяции с узлами в нулях многочлена Чебышева. Для простоты возьмем случай, когда $[a, b] = [-1, 1]$.

Оценка 1. Если $f(x)$ удовлетворяет неравенству $\sup \left| f^{(m)}(x) \right| < \infty$, то справедливо соотношение $\|f(x) - L_n(x)\| = O(n^{-m} \ln n)$ при $n \to \infty$.

Оценка 2. Если функция $f(x)$ аналитична в каждой точке отрезка $[-1, 1]$, то $\|f(x) - L_n(x)\| = O(q^n)$, где $q < 1$.

Последнюю оценку можно конкретизировать. Пусть $f(z)$, $z = x + iy$ — функция, аналитическая в эллипсе на плоскости (x, y) с фокусами в точках -1, 1; тогда $\|f(x) - L_n(x)\| = O(c^{-n})$, где $c > 1$ — сумма полуосей этого эллипса.

Таким образом, при интерполяции по узлам многочлена Чебышева погрешность автоматически уменьшается, если алгоритм применяется к более гладкой функции. Такие алгоритмы называют *ненасыщаемыми*.

Если узлы интерполяции распределены существенно иначе, например равномерно, то даже для аналитической функции погрешность интерполяции может стремиться к бесконечности с ростом числа узлов. Например, для функции $f(x) = (1 + 25x^2)^{-1}$ имеет место соотношение

$$\|f(x) - L_n(x)\| \geqslant Aa^n, \quad A > 0, \quad a > 1.$$

Задача 1. Пользуясь формулой (8.8), показать, что интерполяционный многочлен с узлами в нулях многочленов Чебышева записывается в виде

$$L_n(x) = \sum_{j=0}^{n-1} a_j \widetilde{T}_j(x), \quad a_j = \frac{2}{n} \sum_{k=1}^{n} f(x_j) \widetilde{T}_j \left(\frac{\pi(2k-1)}{2n} \right).$$

Такая запись интерполяционного многочлена позволяет быстро и с малой чувствительностью по отношению к вычислительной погрешности вычислять его значения (см. § 4.8).

§ 10. Конечные разности

Пусть узлы таблицы x_i расположены на равных расстояниях: $x_i = x_0 + ih$, f_i — соответствующие значения функции; величину h называют *шагом таблицы*.

Разности $f_{i+1} - f_i$ называют *разностями первого порядка*. В зависимости от точки, к которой ее относят, эту величину обозначают: Δf_i — *разность вперед*, ∇f_{i+1} — *разность назад*, $\delta f_{i+1/2} = f^1_{i+1/2}$ — *центральная разность*. Таким образом,

$$f_{i+1} - f_i = \Delta f_i = \nabla f_{i+1} = \delta f_{i+1/2} = f^1_{i+1/2}. \tag{1}$$

Разности высшего порядка образуют при помощи рекуррентных соотношений

$$\Delta^m f_i = \Delta(\Delta^{m-1} f_i) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i,$$
$$\nabla^m f_i = \nabla(\nabla^{m-1} f_i) = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1},$$
$$\delta^m f_i = \delta(\delta^{m-1} f_i) = \delta^{m-1} f_{i+1/2} - \delta^{m-1} f_{i-1/2},$$
$$f^m_i = f^{m-1}_{i+1/2} - f^{m-1}_{i-1/2}.$$

Таблицу разностей обычно располагают в виде

x	f	f^1	f^2	f^3
x_0	f_0	$f^1_{1/2}$		
x_1	f_1	$f^1_{3/2}$	f^2_1	$f^3_{3/2}$
x_2	f_2	$f^1_{5/2}$	f^2_2	$f^3_{5/2}$
x_3	f_3	$f^1_{7/2}$	f^2_3	
x_4	f_4			

В некоторых интерполяционных формулах наряду с упомянутыми выше величинами используются средние арифметические двух последовательных величин одного и того же столбца:

$$\mu f_i^m = (f_{i+1/2}^m + f_{i-1/2}^m)/2 \text{при } m \text{ нечетном},$$
$$\mu f_{i+1/2}^m = (f_{i+1}^m + f_i^m)/2 \qquad \text{при } m \text{ четном}.$$

Лемма 1. *Разности m-го порядка выражаются через значения функции по формуле*

$$\Delta^m f_i = \sum_{j=0}^{m} (-1)^j C_m^j f_{i+m-j}, \tag{2}$$

где C_m^j — коэффициенты бинома Ньютона.

Доказательство проводим методом индукции. При $m = 1$ соотношение (2) выполняется согласно (1). Пусть оно доказано при $m = l$. Имеем

$$\Delta^{l+1} f_i = \Delta^l f_{i+1} - \Delta^l f_i = \sum_{j=0}^{l} (-1)^j C_l^j f_{i+1+l-j} - \sum_{j=0}^{l} (-1)^j C_l^j f_{i+l-j}.$$

Собирая коэффициенты при одинаковых f_k и пользуясь равенством

$$C_l^j + C_l^{j+1} = C_{l+1}^{j+1},$$

получим требуемое выражение для величины $\Delta^{l+1} f_i$. Лемма доказана.

Из (2) следует, что *оператор конечной разности* является линейным. В частности, из (2) имеем

$$\Delta^2 f_i = f_{i+2} - 2f_{i+1} + f_i,$$
$$\Delta^3 f_i = f_{i+3} - 3f_{i+2} + 3f_{i+1} - f_i,$$
$$\Delta^4 f_i = f_{i+4} - 4f_{i+3} + 6f_{i+2} - 4f_{i+1} + f_i.$$

Лемма 2. *При $x_i \equiv x_0 + ih$ справедливо равенство*

$$f(x_i; \ldots; x_{i+m}) = \frac{f^m_{i+m/2}}{h^m \cdot m!}. \tag{3}$$

Доказательство проводим по индукции. При $m = 1$ имеем

$$f(x_i;\, x_{i+1}) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{f^1_{i+1/2}}{h}.$$

Предположив, что соотношение (3) верно при всех $m \leqslant l$, имеем

$$f(x_i; \ldots; x_{i+l+1}) = \frac{f(x_{i+1}; \ldots; x_{i+l+1}) - f(x_i; \ldots; x_{i+l})}{x_{i+l+1} - x_i} =$$
$$= \frac{f^l_{i+1+l/2} - f^l_{i+l/2}}{h^l\, l!\, h(l+1)} = \frac{f^{l+1}_{i+(l+1)/2}}{h^{l+1}(l+1)!};$$

таким образом, (3) справедливо при $m = l + 1$. Лемма доказана.

Согласно (5.4) имеем $f(x_i; \ldots; x_{i+m}) = \dfrac{f^{(m)}(\zeta)}{m!}$, где $x_i \leqslant \zeta \leqslant x_{i+m}$. Сопоставляя это равенство с (3), получаем

$$\Delta^m f_i = \nabla^m f_{i+m} = \delta^m f_{i+m/2} = f^m_{i+m/2} = h^m f^{(m)}(\zeta). \tag{4}$$

Следствие. *Конечные разности n-го порядка от многочлена степени n постоянны, а разности любого более высокого порядка равны нулю.*

Рассмотрим влияние погрешности какого-либо значения f_i на конечные разности различных порядков; пусть вместо f_i стоит $f_i + \varepsilon$. Тогда имеем таблицу разностей

f_{i-2}			
	$f^1_{i-3/2}$		
f_{i-1}		$f^2_{i-1} + \varepsilon$	
	$f^1_{i-1/2} + \varepsilon$		
$f_i + \varepsilon$		$f^2_i - 2\varepsilon$	
	$f^1_{i+1/2} - \varepsilon$		
f_{i+1}		$f^2_{i+1} + \varepsilon$	
	$f^1_{i+3/2}$		
f_{i+2}			

Мы видим, что в соответствии с (2) на разности порядка m погрешность распространяется с коэффициентами $(-1)^j C^j_m$. Если функция достаточно гладкая, то ее разности не очень высокого порядка могут оказаться малыми. В то же время на их фоне величины $C^j_m \varepsilon$ будут выглядеть достаточно большими. Из наблюдений над таблицей разностей можно указать значение функции, содержащее погрешность, и исправить его.

Точно так же можно обнаруживать погрешности, имевшие место при составлении таблицы разностей. Пусть, например,

$$f^3_{1/2} = 1 \cdot 10^{-5}, \qquad f^3_{3/2} = 2 \cdot 10^{-5}, \qquad f^3_{5/2} = 12 \cdot 10^{-5},$$
$$f^3_{7/2} = -23 \cdot 10^{-5}, \quad f^3_{9/2} = 11 \cdot 10^{-5}, \quad f^3_{11/2} = 1 \cdot 10^{-5}.$$

Если бы какое-либо значение f_i содержало относительно большую погрешность ε, то в третьих разностях это обстоятельство проявилось бы в наличии величин вида ε, -3ε, 3ε, $-\varepsilon$. В рассматриваемом случае третьи разности практически равны нулю, за исключением $f^3_{5/2}$, $f^3_{7/2}$, $f^3_{9/2}$, которые примерно имеют вид ε, -2ε, ε, где $\varepsilon = 11 \cdot 10^{-5}$. Это наводит на мысль, что была допущена погрешность при вычислении значения $f^1_{7/2}$, которая и имела следствием эти возмущения в третьих разностях. Этот прием широко использовался при ручном счете для устранения случайных погрешностей расчетчика и на первом этапе использования ЭВМ, когда ЭВМ были малонадежны. Существовавшая на первом этапе использования ЭВМ общая рекомендация по устранению ненадежности заключалась в следующем. Задачу предлагалось решить два раза. В случае несовпадения результатов полагалось просчитать задачу повторно, пока результаты двух расчетов не совпадут.

Описанный выше метод исправления таблиц позволяет в такой ситуации уменьшить объем вычислений примерно вдвое.

Пусть требуется составить таблицу какой-либо гладкой функции, каждое вычисление которой обходится очень дорого. Вместо того, чтобы считать каждое значение дважды, просчитываем сразу всю таблицу, составляем (вручную или с помощью ЭВМ) таблицу разностей и выявляем значения, которые нужно исправить (или повторным расчетом, или описанным выше приемом исправления таблиц). В настоящее время описанный прием используется для выявления погрешностей в результатах измерений.

§ 11. Интерполяционные формулы для таблиц с постоянным шагом

Поскольку таблицы значений функций с постоянным шагом наиболее употребительны, приведем конкретные расчетные формулы для таких таблиц. Если узлы интерполирования выбираются вблизи точки x, где вычисляется значение функции, то промежуточная точка ζ в оценке остаточного члена (3.1) также находится вблизи точки x. Таким образом, величина $f^{(n)}(\zeta)$ изменяется не очень сильно при выборе узлов в окрестности точки x.

Следовательно, решающее влияние на значение погрешности оказывает величина

$$|\omega_n(x)| = \prod_{j=1}^{n} |x - x_j|,$$

т. е. произведение расстояний от точки x до узлов интерполирования. Величина $|\omega_n(x)|$ будет минимальной, если в качестве узлов интерполирования для нахождения $f(x)$ мы возьмем n узлов, ближайших к x. Для этого при четном $n = 2l$ следует взять по l узлов справа и слева от точки x. При нечетном $n = 2l + 1$ следует взять узел, ближайший к x, и по l узлов слева и справа от него. Если точка x находится вблизи одного из концов таблицы, то это правило несколько изменится.

При интерполировании в начале или конце таблицы принято записывать интерполяционный многочлен в виде так называемых *формул Ньютона* для интерполирования вперед или назад. Пусть $L_n(x)$ — интерполяционный многочлен Лагранжа по узлам $x_0, \ldots, x_{n-1}$. Согласно (5.3) имеем

$$L_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \cdots + f(x_0; \ldots; x_{n-1})(x - x_0)\ldots(x - x_{n-2}).$$

Произведем замену переменных $x = x_0 + ht$ и перейдем согласно (10.3) от разделенных разностей к конечным. Получим

$$L_n(x_0 + ht) = f_0 + f^1_{1/2}t + \cdots + f^{n-1}_{(n-1)/2}\frac{t(t-1)\ldots(t-(n-2))}{(n-1)!}. \qquad (1)$$

Остаточный член (3.1) представится в виде

$$f(x) - L_n(x) = f^{(n)}(\zeta)\,\frac{t(t-1)\ldots(t-(n-1))\,h^n}{n!}.$$

Формулу (1) называют *интерполяционной формулой Ньютона для интерполирования вперед*. Если мы произведем такую же замену переменных в интерполяционном многочлене $L_n(x)$ по узлам $x_0, x_{-1}, \ldots, x_{-(n-1)}$:

$$L_n(x) = f(x_0) + f(x_0; x_{-1})(x - x_0) + \cdots$$

$$\cdots + f(x_0; \ldots; x_{-(n-1)})(x - x_0)\ldots(x - x_{-(n-2)}),$$

то получим *интерполяционную формулу Ньютона для интерполирования назад*

$$L_n(x_0 + ht) = f_0 + f^1_{-1/2}t + \cdots + f^{n-1}_{-(n-1)/2}\,\frac{t(t+1)\ldots(t+(n-2))}{(n-1)!} \qquad (2)$$

с остаточным членом

$$f(x) - L_n(x) = f^{(n)}(\zeta)\frac{t(t+1)\ldots(t+(n-1))\,h^n}{n!}.$$

Эти формулы, в частности, используются при построении методов решения дифференциальных уравнений. Таблицы конечных разностей так же, как и таблицы разделенных разностей, используются для оценки производных функции. Если $f^{(n)}(x)$ непрерывна, то справедливо равенство $\lim\limits_{h\to 0}\sigma_n^h = M_n$; здесь $\sigma_n^h = \max\limits_{a\leqslant x_m\leqslant b-nh}\left|\dfrac{\Delta^n f_m}{h^n}\right|$; $M_n = \max\limits_{[a,b]}\left|f^{(n)}(x)\right|$. Поэтому при малых h можно принять $M_n \approx \sigma_n^h$.

Часто приобретает особо важное значение малость степени полиномов, приближающих функцию. Уменьшения степени таких полиномов без потери точности иногда можно достигнуть, образуя линейные комбинации интерполяционных полиномов. Рассмотрим простейший из таких способов приближения функций.

Требуется приблизить функцию на отрезке $[x_q, x_{q+1}]$ многочленом второй степени. Выпишем интерполяционную формулу Ньютона третьей степени по узлам x_{q-1}, x_q, x_{q+1}, x_{q+2}, взяв узлы в последовательности x_q, x_{q+1}, x_{q-1}, x_{q+2} и в последовательности x_{q+1}, x_q, x_{q+2}, x_{q-1}. Имеем

$$f(x) = P_3^1(x) + r^1(x); \tag{3}$$

$$\begin{aligned} P_3^1(x) = {} & f(x_q) + f(x_q; x_{q+1})(x - x_q) + f(x_q; x_{q+1}; x_{q-1})(x - x_q)(x - x_{q+1}) + \\ & + f(x_q; x_{q+1}; x_{q-1}; x_{q+2})(x - x_q)(x - x_{q+1})(x - x_{q-1}); \end{aligned} \tag{4}$$

$$r^1(x) = \frac{f^{(4)}(\zeta_1)}{4!}(x - x_{q-1})(x - x_q)(x - x_{q+1})(x - x_{q+2});$$

$$f(x) = P_3^2(x) + r^2(x);$$

$$\begin{aligned} P_3^2(x) = {} & f(x_{q+1}) + f(x_{q+1}; x_q)(x - x_{q+1}) + \\ & + f(x_{q+1}; x_q; x_{q+2})(x - x_{q+1})(x - x_q) + \\ & + f(x_{q+1}; x_q; x_{q+2}; x_{q-1})(x - x_{q+1})(x - x_q)(x - x_{q+2}); \end{aligned} \tag{5}$$

$$r^2(x) = \frac{f^{(4)}(\zeta_2)}{4!}(x - x_{q-1})(x - x_q)(x - x_{q+1})(x - x_{q+2}).$$

Поскольку интерполяционный многочлен третьей степени, совпадающий с функцией в четырех узлах, единствен, то

$$P_3^1(x) = P_3^2(x), \quad r^1(x) = r^2(x) \quad \text{и} \quad f^{(4)}(\zeta_1) = f^{(4)}(\zeta_2).$$

Образуем полусумму равенств (4), (5). Так как $P_3^1(x) = P_3^2(x)$, в левой части будет стоять многочлен $P_3^1(x)$; при вычислении правой части

образуем полусуммы от соответствующих слагаемых; введем обозначения: $x_{q+1/2} = x_q + h/2$, $\mu f^{2n}_{q+1/2} = (f^{2n}_q + f^{2n}_{q+1})/2$. Получим

$$P^1_3(x) = \mu f_{q+1/2} + f^1_{q+1/2} h^{-1}(x - x_{q+1/2}) + \frac{1}{2}\mu f^2_{q+1/2} h^{-2}(x - x_{q+1})(x - x_q) + \frac{1}{6} f^3_{q+1/2} h^{-3}(x - x_q)(x - x_{q+1})(x - x_{q+1/2}).$$

Обозначив первые три слагаемые в правой части последнего равенства через $B_2(x)$, соотношение (3) запишем в виде

$$f(x) = B_2(x) + R(x),$$

$$B_2(x) = \mu f_{q+1/2} + f^1_{q+1/2} h^{-1}(x - x_{q+1/2}) + \mu f^2_{q+1/2} h^{-2}(x - x_q)(x - x_{q+1}), \quad (6)$$

$$R(x) = f^3_{q+1/2} h^{-3}(x - x_q)(x - x_{q+1})(x - x_{q+1/2}) + \frac{f^{(4)}(\zeta_1)}{24}(x - x_{q-1})(x - x_q)(x - x_{q+1})(x - x_{q+2}). \quad (7)$$

Многочлен $B_2(x)$ называют *интерполяционным многочленом Бесселя*. Если подходить формально, то этот многочлен второй степени не является интерполяционным, поскольку он совпадает с $f(x)$ только в точках x_q, x_{q+1}.

В следующем параграфе будет видно, что использование многочлена Бесселя дает определенные преимущества по сравнению с непосредственным использованием интерполяционного многочлена второй степени.

§ 12. Составление таблиц

Рассмотрим следующую задачу. Требуется построить таблицу значений некоторой функции так, чтобы погрешность при интерполяции значений функции многочленом заданной степени m не превосходила ε. В этом случае говорят, что таблица допускает интерполяцию степени m (с погрешностью ε). Таблицы, выпускаемые для широкого круга пользователей, обычно составляются так, чтобы они допускали интерполяцию первой степени, иначе — *линейную интерполяцию*. Примером таких таблиц могут служить таблицы В.М. Брадиса, известные из школьного курса. В дальнейшем рассматриваем случай таблицы с постоянным шагом.

Для вычисления значения $f(x)$ при помощи такой таблицы берутся узлы x_q и x_{q+1} справа и слева от точки x: $x_q < x < x_{q+1}$ (они будут ближайшими к x); затем $f(x)$ заменяется интерполяционным многочленом первой степени по этим узлам (для удобства обозначаем $x = x_q + th$):

$$f(x) \approx L_2(x) = f_q + f^1_{q+1/2} t.$$

Погрешность этой формулы

$$f''(\zeta)\,h^2\frac{t(t-1)}{2},$$

где

$$x_q \leqslant \zeta \leqslant x_{q+1}.$$

Эта величина не превосходит ε, если

$$\max\left|f''(\zeta)\right| h^2 \max_{[0,1]}\left|\frac{t(t-1)}{2}\right| \leqslant \varepsilon;$$

$\max\limits_{[0,1]}\left|\frac{t(t-1)}{2}\right|$ достигается при $t=1/2$ и равен $1/8$. Таким образом, достаточно выполнения условия

$$\max\left|f''(\zeta)\right| h^2/8 \leqslant \varepsilon. \tag{1}$$

Пусть мы хотим составить или ввести в машину таблицу $\sin x$ на $[0,\,\pi/2]$ так, чтобы погрешность линейной интерполяции не превосходила $0{,}5\cdot 10^{-6}$. Поскольку $\max|(\sin x)''| \leqslant 1$, то из (1) вытекает требование на шаг таблицы

$$h^2/8 \leqslant 0{,}5\cdot 10^{-6} \quad \text{или} \quad h \leqslant 0{,}002.$$

Часто требование допустимости линейной интерполяции является слишком жестким и вместо него требуют допустимой *квадратичной интерполяции* (т. е. интерполяции многочленом второй степени). Простейшим случаем квадратичной интерполяции будет интерполяция многочленом Лагранжа по трем узлам. Пусть x_q — узел, ближайший к x, т. е. $|x-x_q| \leqslant h/2$. Имеем

$$f(x) \approx L_3(x) = f_q + f^1_{q+1/2}t + f^2_q\frac{t(t-1)}{2}. \tag{2}$$

Остаточный член этой формулы

$$f(x) - L_3(x) = f^{(3)}(\zeta)h^3\frac{t(t^2-1)}{3!}.$$

Чтобы таблица допускала квадратичную интерполяцию (2), достаточно выполнения условия

$$\max\left|f^{(3)}(\zeta)\right| h^3 \max_{|t|\leqslant 1/2}\left|\frac{t(t^2-1)}{6}\right| \leqslant \varepsilon.$$

Так как $\max\limits_{|t|\leqslant 1/2}\left|\frac{t(t^2-1)}{6}\right| = \frac{1}{16}$, то это требование на шаг перепишется в виде

$$\max\left|f^{(3)}(\zeta)\right| h^3/16 \leqslant \varepsilon. \tag{3}$$

В конкретном случае при $f(x)=\sin x$, $\varepsilon = 0{,}5\cdot 10^{-6}$ получаем $h \leqslant 0{,}02$.

Рассмотрим другой способ замены функции многочленом второй степени. Пусть $x \in (x_q, x_{q+1})$, $x = x_q + ht$; положим

$$f(x) \approx B_2(x) = \mu f_{q+1/2} + f^1_{q+1/2}(t - 1/2) + \mu f^2_{q+1/2}\frac{t(t-1)}{2}, \qquad (4)$$

т. е. заменим $f(x)$ многочленом Бесселя второй степени (11.6), выписанным по узлам x_{q-1}, x_q, x_{q+1}, x_{q+2}. Согласно (11.7) остаточный член (4) есть

$$f^3_{q+1/2}\frac{t(t-1)(t-1/2)}{6} + f^{(4)}(\zeta)h^4\frac{t(t^2-1)(t-2)}{24}.$$

Так как $f^3_{q+1/2} = h^3 f^{(3)}(\tilde{\zeta})$, то для допустимости интерполяции по формуле (4) достаточно выполнения соотношения

$$\max\left|f^{(3)}(\zeta)\right| h^3 \max_{0\leqslant t\leqslant 1}\left|\frac{t(t-1)(t-1/2)}{6}\right| +$$
$$+ \max\left|f^{(4)}(\zeta)\right| h^4 \max_{0\leqslant t\leqslant 1}\left|\frac{t(t^2-1)(t-2)}{24}\right| \leqslant \varepsilon.$$

Поскольку

$$\max_{0\leqslant t\leqslant 1}\left|\frac{t(t-1)(t-1/2)}{6}\right| = \frac{1}{72\sqrt{3}}, \quad \max_{0\leqslant t\leqslant 1}\left|\frac{t(t^2-1)(t-2)}{24}\right| = \frac{3}{128},$$

то интерполяция по формуле (4) допустима, если

$$\frac{\max\left|f^{(3)}(\zeta)\right| h^3}{72\sqrt{3}} + \frac{3\max\left|f^{(4)}(\zeta)\right|}{128}h^4 \leqslant \varepsilon. \qquad (5)$$

При малых h главной частью является первое слагаемое; оно меньше, чем левая часть (3), в $9\sqrt{3}/2 = 7{,}794\ldots$ раз. Следовательно, при малых h для выполнимости (5) можно взять шаг в $\approx \sqrt[3]{9\sqrt{3}/2} \approx 1{,}98$ раз больше, чем для выполнимости (3). В рассматриваемом примере условие (5) имеет вид

$$h^3/72\sqrt{3} + 3h^4/128 \leqslant 0{,}5 \cdot 10^{-6}.$$

Решая это неравенство, получим $h \leqslant h_0 = 0{,}038\ldots$. Заметим, что при ручном счете шаг $h = 0{,}038$ неудобен вследствие «некруглости» этого числа. Поэтому при составлении таблиц его заменили бы заведомо на меньшее, но «более круглое» число 0,03.

В многомерном случае иногда целесообразно дальнейшее увеличение степени используемого интерполяционного многочлена.

§ 13. О погрешности округления при интерполяции

Предположим, что выбран некоторый способ интерполяции. Выше мы получили некоторое представление о погрешности, являющейся следствием замены функции многочленом. Однако существует еще одна причина погрешности, в частности вследствие округления этих значений. Пусть требуется вычислить значение $f(x)$ по формуле

$$f(x) \approx \sum_{j=1}^{n} f_j P_j(x),$$

являющейся общим видом рассматриваемых нами интерполяционных формул. Поскольку реально заданы не f_j, а приближенные значения $f_j^* = f_j + \eta_j$, то в результате будет получено значение

$$f^*(x) = \sum_{j=1}^{n} f_j P_j(x) + \sum_{j=1}^{n} \eta_j P_j(x).$$

Если известны границы изменения значений η_j, то можно оценить верхнюю грань погрешности

$$\varepsilon = \sum_{j=1}^{n} \eta_j P_j(x).$$

Например, при условии $|\eta_j| \leqslant \eta$ имеем оценку

$$|\varepsilon| \leqslant \eta\Lambda(x), \quad \Lambda(x) = \sum_{j=1}^{n} |P_j(x)|.$$

Величина Λ может оказаться очень большой.

Задача 1. Пусть $f(x)$ интерполируется по узлам $x_j = -1 + 2\dfrac{j-1}{n-1}$, $j = 1, \ldots, n$. Показать, что $\max\limits_{[-1,\,1]} \Lambda(x) \geqslant \text{const} \cdot \dfrac{2^n}{\sqrt{n}}$.

Задача 2. Доказать, что если узлы интерполяции совпадают с нулями многочлена Чебышева, то $\max\limits_{[-1,\,1]} |\Lambda(x)| \leqslant \text{const} \cdot \ln n$.

Если мы вычисляем значение $f(x)$ при $x_0 < x < x_1$ интерполяцией по узлам x_0, x_1, то

$$P_0(x) = \frac{x_1 - x}{x_1 - x_0}, \quad P_1(x) = \frac{x - x_0}{x_1 - x_0}$$

и

$$\eta\Big(|P_0(x)| + |P_1(x)|\Big) = \eta.$$

Таким образом, при линейной интерполяции погрешность, являющаяся следствием округления значений функции, не превосходит погрешности этих значений.

Наличие большого числа формул интерполирования, применявшихся во времена ручного счета, отчасти объясняется именно поисками алгоритмов, порождающих минимальную вычислительную погрешность.

§ 14. Применения аппарата интерполирования. Обратная интерполяция

Употребление интерполяционных многочленов оказывается полезным при решении такой задачи.

Пусть требуется найти экстремум функции и точку экстремума. Составим таблицу значений функции с крупным шагом. Из рассмотрения этой таблицы можно увидеть место расположения экстремума. В предполагаемой области расположения экстремума приблизим функцию интерполяционным многочленом и найдем его точку экстремума P_1. В окрестности точки P_1 составим таблицу значений функции с более мелким шагом. Из рассмотрения этой таблицы можно уточнить расположение экстремума и т. д. Степень интерполяционного многочлена берется такой, чтобы точка экстремума определялась в явном виде. В одномерном случае берется интерполяционный многочлен Лагранжа или Бесселя второй степени или интерполяционный многочлен третьей степени. В многомерном случае, как правило, функция приближается многочленом второй степени.

На практике, начиная с окрестности приближения P_1 или следующего приближения P_2, уже не строят подробной таблицы значений функции, а ограничиваются минимальным числом точек в окрестности имеющегося приближения P_n, достаточным для построения интерполяционного многочлена.

Описанный способ является одним из наиболее употребительных при отыскании экстремума функции многих переменных.

В одномерном случае иногда после вычисления значения $f(P_n)$ не вычисляют дополнительно никаких новых значений функции, а проводят интерполяцию, используя это значение и ранее вычисленные значения.

Другой типичной задачей, где может быть применен аппарат интерполирования, является нахождение корня X уравнения $f(x) = d$.

Путь решения этой задачи тот же самый. Составляем таблицу значений функции; определяем по ней грубо, где находится корень уравнения, затем составляем таблицу с более мелким шагом и т. д.

Если вычисление функции относительно нетрудоемко, неразумно применять в процессе вычислений интерполяцию степени выше второй; в противном случае возникает задача нахождения корней многочленов, сама требующая достаточно большого числа арифметических операций.

Если вычисление функции трудоемко, может оказаться более выгодным пойти по пути увеличения степени интерполяционного многочлена.

В случае, когда в окрестности $y = d$ функция $g(y)$, обратная к $f(x)$, является достаточно гладкой, более эффективным может оказаться применение обратной интерполяции. *Обратной интерполяцией* называется следующий алгоритм. Пусть известны значения функции $y_i = f(x_i)$ при $i = 1, \ldots, n$. Эта информация эквивалентна тому, что известны значения $x_i = g(y_i)$ обратной функции. При условии допустимости интерполяции по переменной y можно заменить обратную функцию $g(y)$ интерполяционным многочленом $L_n(y)$, удовлетворяющим условиям

$$L_n(y_i) = x_i, \quad i = 1, \ldots, n,$$

и положить $X = g(d) \approx L_n(d)$. Такой способ особенно удобен, если нас интересуют значения решений уравнений при достаточно большом числе значений d или желательно получение явного выражения корня уравнения $f(x) = d$ в зависимости от параметра d. Если интерполяция по узлам $y_1, \ldots, y_n$ не обеспечивает нужной точности, полагаем $x_{n+1} = L_n(d)$, $y_{n+1} = f(x_{n+1})$. Далее, в зависимости от обстановки, целесообразно заменить $g(d)$ значением интерполяционного многочлена по всем узлам $y_1, \ldots, y_{n+1}$ или по некоторым из этих узлов, ближайшим к d.

§ 15. Численное дифференцирование

Простейшие формулы *численного дифференцирования* получаются в результате дифференцирования интерполяционных формул.

Пусть известны значения функции в точках $x_1, \ldots, x_n$ и требуется вычислить производную $f^{(k)}(x_0)$. Построим интерполяционный многочлен $L_n(x)$ и положим $f^{(k)}(x_0) \approx L_n^{(k)}(x_0)$. Точно так же мы можем заменять значения производных функций значениями производных других многочленов интерполяционного типа, например Бесселя.

Другой способ построения формул численного дифференцирования, приводящий к тем же формулам, — это метод неопределенных коэффициентов. Наиболее употребителен он в многомерном случае, когда не всегда просто выписывается интерполяционный многочлен. Коэффициенты c_i формулы численного дифференцирования

$$f^{(k)}(x) \approx \sum_{i=1}^{n} c_i f(x_i) \tag{1}$$

выбираются из условия, чтобы формула была точна для многочленов максимально высокой степени. Возьмем $f(x) = \sum_{j=0}^{n} a_j x^j$ и потребуем, что-

бы для такого многочлена соотношение (1) обратилось в равенство

$$\sum_{j=0}^{m} a_j (x^j)^{(k)} \Big|_{x_0} = \sum_{i=1}^{n} c_i \left(\sum_{j=0}^{m} a_j x_i^j \right).$$

Чтобы равенство выполнялось для любого многочлена степени m, необходимо и достаточно, чтобы коэффициенты при a_j в правой и левой частях были равны. Поскольку

$$(x^j)^{(k)} = j(j-1)\dots(j-k+1)x^{j-k},$$

то получаем линейную систему уравнений

$$j(j-1)\dots(j-k+1)x_0^{j-k} = \sum_{i=1}^{n} c_i x_i^j, \quad j = 0, \dots, m, \tag{2}$$

относительно неизвестных c_i. Если $m = n-1$, то число уравнений равно числу неизвестных. Определитель системы является определителем Вандермонда, поэтому отличен от нуля. Таким образом, всегда можно построить формулу численного дифференцирования с n узлами, точную для многочленов степени $n-1$.

При $m = n-1$ и определенном расположении узлов иногда оказывается, что равенство (2) выполнено и для $j = n$. Как правило, это будет в случае, когда узлы расположены симметрично относительно точки x_0.

В приведенных ниже задачах для простоты взято $x_0 = 0$. Пусть узлы x_i расположены симметрично относительно точки $x_0 = 0$, т. е. $x_1 = -x_n$, $x_2 = -x_{n-1}$ и т. д. Если n нечетно, $n = 2l+1$, то тогда $x_{l+1} = x_0 = 0$.

Задача 1. Пусть k четно (в частности, k может быть равно нулю и тогда речь идет об интерполировании). Доказать, что тогда $c_n = c_1$, $c_{n-1} = c_2$, и вообще $c_{n+1-k} = c_k$.

Доказать, что вследствие такого свойства симметрии формула (1) автоматически является точной для любой нечетной функции. В частности, при n нечетном формула (1) будет точна для x^n, поэтому она точна и для любого многочлена степени n (поскольку для любого многочлена степени $n-1$ она уже оказалась точной по построению).

Задача 2. Пусть k нечетно. Доказать, что тогда $c_n = -c_1$, $c_{n-1} = -c_2, \dots$, и вообще $c_{n+1-k} = -c_k$. Если n нечетно, $n = 2l+1$, то при $k = l+1$ имеем $c_{l+1} = -c_{l+1}$ и, следовательно, $c_{l+1} = 0$.

Доказать, что вследствие такого свойства симметрии формула (1) автоматически является точной для любой четной функции. В частности, при n четном она будет точна для x^n и поэтому будет точна для любого многочлена степени n.

Таким образом, при симметричном относительно x_0 расположении узлов, k четном, n нечетном или k нечетном, n четном формула (1) оказывается точной для многочленов на единицу большей степени.

Свойства симметрии формул численного дифференцирования используются для уменьшения числа уравнений, которые нужно решить при построении формулы.

Пусть требуется построить формулу численного дифференцирования

$$f'(0) \approx c_1 f(-h) + c_2 f(0) + c_3 f(h),$$

точную для многочленов второй степени. Система уравнений (2) в данном случае имеет вид

$$\begin{aligned} 0 &= c_1 + c_2 + c_3, \\ 1 &= c_1(-h) + c_3(h), \\ 0 &= c_1(-h)^2 + c_3(h)^2, \end{aligned}$$

и, решая ее, получаем $c_1 = -1/(2h)$, $c_2 = 0$, $c_3 = 1/(2h)$. Воспользуемся свойством симметрии и сразу возьмем формулу, для которой $c_3 = -c_1$, $c_2 = 0$. Тогда первое и третье уравнения выполнены автоматически, а второе приобретает вид $1 = 2c_3 h$, т. е. $c_3 = 1/(2h)$. Таким образом, $f'(0) \approx (f(h) - f(-h))/(2h)$.

Задача 3. Пусть все точки x_j удалены от точки x_0 на расстояние $O(h)$, h — малая величина. Показать, что при гладкой $f(x)$ приближенная формула численного дифференцирования (1) имеет порядок погрешности $O(h^m)$, где $m = l + 1 - k$, l — максимальная степень многочленов, для которых точна эта формула.

Построим приближенную формулу вычисления второй производной, использующую те же узлы:

$$f''(0) \approx \frac{c_1 f(-h) + c_2 f(0) + c_3 f(h)}{h^2}.$$

Из условий точности формулы для $1, x, x^2$ получаем систему уравнений

$$\begin{aligned} 0 &= c_1 + c_2 + c_3, \\ 0 &= \frac{c_1(-h) + c_3(h)}{h^2}, \\ 1 &= \frac{c_1 \frac{h^2}{2} + c_3 \frac{h^2}{2}}{h^2}. \end{aligned}$$

Решая эту систему, получим $c_1 = c_3 = 1$, $c_2 = -2$ и соответствующую приближенную формулу

$$f''(0) \approx \frac{f(h) - 2f(0) + f(-h)}{h^2}. \tag{3}$$

Мы можем не сомневаться в том, что получили правильную формулу. Выражение в правой части есть $f_0^2/h^2 = \Delta^2 f_{-1}/h^2$, и, согласно (10.4), оно равно значению $f''(\xi)$. У многочлена второй степени вторая производная постоянна, поэтому $f_0^2/h^2 = f''(\xi) \equiv f''(x)$ при любом x, в частности $f_0^2/h^2 = f''(0)$.

Построенная формула оказывается точной для любого многочлена третьей степени. Если подставим в левую и правую части (3) функцию $f(x) = x^3$, то в обеих частях получим нуль.

Оценим погрешность построенной выше приближенной формулы $f'(0) \approx (f(h) - f(-h))/(2h)$. В формуле Тейлора возьмем три члена разложения и остаточный член

$$f(h) = f(0) + hf'(0) + \frac{h^2}{2}f''(0) + \frac{h^3}{6}f'''(\xi_+), \quad 0 \leqslant \xi_+ \leqslant h,$$

$$f(-h) = f(0) - hf'(0) + \frac{h^2}{2}f''(0) - \frac{h^3}{6}f'''(\xi_-), \quad -h \leqslant \xi_- \leqslant 0.$$

Введем обозначение $R_k(f) = f^{(k)}(x_0) - \sum\limits_{i=1}^{n} c_i f(x_i)$. Имеем

$$\begin{aligned} R_1(f) &= f'(0) - \frac{f(h) - f(-h)}{2h} = \\ &= f'(0) - \left(f(0) + hf'(0) + \frac{h^2}{2}f''(0) + \frac{h^3}{6}f'''(\xi_+) - \right. \\ &\quad \left. - \left(f(0) - hf'(0) + \frac{h^2}{2}f''(0) - \frac{h^3}{6}f'''(\xi_-)\right)\right) / (2h) = \\ &= f'(0) - \left(f'(0) + \frac{h^2}{6}\left(\frac{f'''(\xi_+) + f'''(\xi_-)}{2}\right)\right) = \\ &= -\frac{h^2}{6}\alpha, \qquad \alpha = \frac{f'''(\xi_+) + f'''(\xi_-)}{2}. \end{aligned}$$

Значение α лежит между $f''(\xi_+)$ и $f''(\xi_-)$. Поэтому по теореме Ролля найдется ξ в пределах $[\xi_-, \xi_+]$ такое, что $\alpha = f'''(\xi)$. Таким образом, в итоге имеем

$$R_1(f) = -\frac{h^2}{6}f'''(\xi), \quad -h \leqslant \xi \leqslant h.$$

Рассмотрим приближенную формулу (3). Предположим сначала, что нам неизвестно, точна ли она для любого многочлена третьей степени. Беря в разложении Тейлора три члена

$$f(\pm h) = f(0) \pm hf'(0) + \frac{h^2}{2}f''(0) \pm \frac{h^3}{6}f'''(\xi_\pm),$$

получим

$$R_2(f) = f''(0) - \frac{f(h) - 2f(0) - f(-h)}{h^2} = f''(0) -$$
$$- \left[f(0) + hf'(0) + \frac{h^2}{2} f''(0) + \frac{h^3}{6} f'''(\xi_+) - 2f(0) + f(0) - hf'(0) + \right.$$
$$\left. + \frac{h^2}{2} f''(0) - \frac{h^3}{6} f'''(\xi_-) \right] h^{-2} = -\frac{h}{6}(f'''(\xi_+) - f'''(\xi_-)).$$

Если $f(x)$ — многочлен третьей степени, то $f'''(x) = \text{const}$, поэтому $R_2(f) \equiv 0$. Таким образом, из выражения для погрешности мы увидели, что формула (3) точна для всех многочленов третьей степени. По теореме Лагранжа

$$f'''(\xi_+) - f'''(\xi_-) = (\xi_+ - \xi_-) f^{(4)}(\bar{\xi});$$

$\bar{\xi} \in [\xi_-, \xi_+]$. В то же время $\xi_+ \in [0, h]$, $\xi_- \in [-h, 0]$, откуда следует $0 \leqslant \xi_+ - \xi_- \leqslant 2h$. Таким образом, $\xi_+ - \xi_- = \theta \cdot 2h$, где $0 \leqslant \theta \leqslant 1$, и

$$R_2(f) = -\theta \frac{h^2}{6} f^{(4)}(\bar{\xi}), \quad 0 \leqslant \theta \leqslant 1.$$

Если в разложении Тейлора взять четыре слагаемых

$$f(\pm h) = f(0) \pm hf'(0) + \frac{h^2}{2} f''(0) \pm \frac{h^3}{6} f'''(0) + \frac{h^4}{24} f^{(4)}(\xi_\pm),$$

то получим выражение для погрешности

$$R_2(f) = -\frac{h^2}{12} \left(\frac{f^{(4)}(\xi_+) + f^{(4)}(\xi_-)}{2} \right).$$

Рассуждая, как и при выводе оценки погрешности для первой производной, имеем

$$R_2(f) = -\frac{h^2}{12} f^{(4)}(\xi), \quad -h \leqslant \xi \leqslant h.$$

Приведем ряд формул численного дифференцирования функций, заданных на сетке с постоянным шагом $x_n = x_0 + nh$:

$$f'(x_0) \approx h^{-1} \sum_{j=1}^{n} \frac{(-1)^{j-1}}{j} f^j_{j/2}, R_1(f) = \frac{(-1)^n}{n+1} f^{(n+1)}(\xi) h^n;$$
$$f'(x_0) \approx h^{-1} \sum_{j=1}^{n} \frac{1}{j} f^j_{-j/2}, \qquad R_1(f) = \frac{1}{n+1} f^{(n+1)}(\xi) h^n. \tag{4}$$

Это так называемые односторонние формулы численного дифференцирования. В первой формуле (4) все узлы удовлетворяют условию $x_k \geqslant x_0$,

во второй $x_k \leqslant x_0$. Среди таких формул наиболее употребительны следующие:

$$n=1, \quad f'(x_0) \approx \frac{f^1_{1/2}}{h} = \frac{f(x_1)-f(x_0)}{h};$$

$$n=2, \quad f'(x_0) \approx \frac{1}{h}\left(f^1_{1/2} - \frac{1}{2}f^2_1\right) = \frac{-f(x_2)+4f(x_1)-3f(x_0)}{2h}$$

и

$$n=1, \quad f'(x_0) \approx \frac{f^1_{-1/2}}{h} = \frac{f(x_0)-f(x_{-1})}{h};$$

$$n=2, \quad f'(x_0) \approx \frac{1}{h}\left(f^1_{-1/2} + \frac{1}{2}f^2_{-1}\right) = \frac{3f(x_0)-4f(x_{-1})+f(x_{-2})}{2h}.$$

Такие приближения производных часто используются при решении дифференциальных уравнений для аппроксимации граничных условий. Приведем примеры симметричных формул:

$$f'\left(x_0+\frac{h}{2}\right) \approx h^{-1}\sum_{j=0}^{l-1} \frac{(-1)^j\left(\left(\frac{1}{2}\right)\cdot\left(\left(\frac{3}{2}\right)\cdots\left(\left(j-\frac{1}{2}\right)\right)\right)^2}{(2j+1)!} f^{2j+1}_{1/2};$$

$$R_1(f) = (-1)^l \frac{\left(\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\cdots\left(\left(l-\frac{1}{2}\right)\right)\right)^2}{(2l+1)!} f^{(2l+1)}(\xi)h^{2l}.$$

Наиболее употребительны следующие частные случаи:

$$l=1, \quad f'\left(x_0+\frac{h}{2}\right) \approx \frac{f^1_{1/2}}{h} = \frac{f(x_1)-f(x_0)}{h}$$

(уже рассмотренный нами выше в других обозначениях);

$$l=2, \quad f'\left(x_0+\frac{h}{2}\right) \approx \frac{1}{h}\left(f^1_{1/2} - \frac{1}{24}f^3_{1/2}\right) = \frac{-f(2h)+27f(h)-27f(0)+f(-h)}{24h}.$$

Формулы для второй производной записываются в виде

$$f''(x_0) \approx h^{-2}\sum_{j=1}^{l} \frac{2(-1)^{j-1}((j-1)!)^2}{(2j)!} f_0^{2j},$$

остаточный член

$$R_2(f) = \frac{2(-1)^l(l!)^2}{(2l+2)!} f^{(2l+2)}(\xi)h^{2l}.$$

Наиболее употребительны частные случаи:

$$l=1,\quad f''(x_0)\approx\frac{f_0^2}{h^2}=\frac{f(h)-2f(0)+f(-h)}{h^2};$$

$$l=2,\quad f''(x_0)\approx\frac{1}{h^2}\left(f_0^2-\frac{1}{12}f_0^4\right)=$$

$$=\frac{-f(2h)+16f(h)-30f(0)+16f(-h)-f(-2h)}{12h^2}.$$

Для высших производных простейшую грубую аппроксимацию можно получить, воспользовавшись (10.4). При $0\leqslant j\leqslant k$ имеем

$$R_k=f^{(k)}(x_j)-\frac{\Delta^k f_m}{h^k}=f^{(k)}(x_j)-f^{(k)}(\xi_{m,k})=O(h).$$

Наиболее употребительные частные случаи: односторонние формулы численного дифференцирования

$$f^{(k)}(0)\approx\frac{\Delta^k f_0}{h^k}=\frac{f_{k/2}^k}{h^k},\quad f^{(k)}(0)\approx\frac{\nabla^k f_0}{h^k}=\frac{f_{-k/2}^k}{h^k},$$

имеющие погрешность порядка $O(h)$, и симметричные формулы численного дифференцирования. При k четном

$$f^{(k)}(0)\approx f_0^k/h^k,$$

при k нечетном

$$f^{(k)}(0)\approx\frac{f_{1/2}^k+f_{-1/2}^k}{2h^k}.$$

Эти формулы имеют погрешность $O(h^2)$. При $k=1,\,2$ такими формулами как раз являются формулы, приведенные выше.

При выводе формул численного дифференцирования из приближенного равенства

$$f^{(k)}(x_0)\approx L_n^{(k)}(x_0)$$

оценку погрешности также можно получить, дифференцируя остаточный член в (5.1):

$$f^{(k)}(x_0)=L_n^{(k)}(x_0)+(f(x;x_1;\ldots;x_n)\omega_n(x))^{(k)}\Big|_{x_0}.$$

Для получения конкретной оценки надо воспользоваться правилом Лейбница и доказать равенство

$$(f(x;x_1;\ldots;x_n))^{(q)}=q!f(\underbrace{x;\ldots;x}_{q+1\text{ раз}};x_1;\ldots;x_n).$$

§ 16. О вычислительной погрешности формул численного дифференцирования

При решении одной задачи управления имела место следующая ситуация. Управление объектом выбиралось в зависимости от его скорости движения в данный момент; скорость вычислялась по простейшей формуле численного дифференцирования как отношение приращения координат к промежутку времени δt между двумя последовательными моментами измерения положения объекта.

Перед непосредственным конструированием системы было произведено подробное моделирование ее работы с помощью ЭВМ: координаты объекта брались со случайными погрешностями измерения и т. д.

Численные эксперименты показывали, что объект должен все время резко менять направление движения и требуемое управление движением нереализуемо. Однако уменьшение промежутка δt не приводило к улучшению дела. В данном конкретном примере проблема была решена путем увеличения промежутка δt в 100 раз по сравнению с предполагавшимся заранее. Попутно это привело к снижению стоимости управляющей системы.

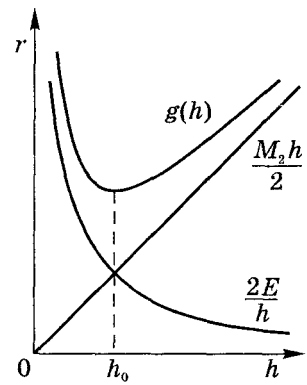


Рис. 2.16.1

Дело заключается в том, что часто уменьшение погрешности метода, в данном случае формулы численного дифференцирования, сопровождается ростом влияния погрешности исходных данных и вычислительной погрешности. Численное дифференцирование относится к таким задачам, где влияние этих погрешностей сказывается уже при умеренных значениях погрешности метода решения задачи.

Пусть для определенности значение $f'(x_0)$ определяется из соотношения

$$f'(x_0) \approx (f(x_1) - f(x_0))/h. \tag{1}$$

Согласно (15.4) остаточный член этой формулы имеет вид

$$r_1 = -f''(\zeta)h/2.$$

Пусть $|f''(\zeta)| \leqslant M_2$, тогда $|r_1| \leqslant M_2 h/2$.

Если значения функции $f(x_i)$ известны с некоторыми погрешностями $\varepsilon_i, |\varepsilon_i| \leqslant E$, то погрешность $f'(x_0)$ будет содержать дополнительное слагаемое

$$r_2 = -(\varepsilon_1 - \varepsilon_0)/h, \quad |r_2| \leqslant 2E/h.$$

Для простоты пренебрежем округлениями при реальном вычислении правой части (1). Тогда имеем оценку погрешности

$$|r| \leqslant |r_1| + |r_2| \leqslant g(h) = M_2 h/2 + 2E/h. \tag{2}$$

Для малости погрешности необходима малость h, но при уменьшении h растет второе слагаемое (рис. 2.16.1).

Из уравнения $g'(h) = 0$ получаем точку экстремума h_0 для $g(h)$:

$$h_0 = 2\sqrt{E/M_2},$$

затем значение

$$g(h_0) = 2\sqrt{M_2 E}.$$

Таким образом, ни при каком h нельзя гарантировать, что погрешность результата будет величиной порядка $o(\sqrt{E})$.

Погрешности ε_i возникают вследствие погрешностей в задаваемых значениях функций, например, если функция определяется из измерений или вычисляется по некоторой приближенной формуле. Поскольку эти значения округляются дополнительно при вводе в машину, то следует считать, что $E \geqslant \text{const} \cdot 2^{-t}$, где t — число разрядов. Таким образом, мы можем получить $f'(x_0)$ в лучшем случае с половиной верных разрядов.

В случае применения формул более высокого порядка точности положение несколько улучшается. Пусть производная $y^{(k)}(x)$ вычисляется по формуле

$$y^{(k)}(x) \approx \left(\sum_{j=1}^{n} c_j y_j\right) / h^k, \quad c_j = O(1), \tag{3}$$

с остаточным членом $O(h^l)$. Все рассмотренные выше формулы численного дифференцирования могут быть записаны в таком виде: в знаменателе — h^k, а в числителе — коэффициенты порядка $O(1)$. Погрешность, являющаяся следствием погрешностей в правой части (3), оценивается величиной $\text{const} \cdot E/h^k$. Таким образом, вместо (2) мы имеем

$$|r| \leqslant g_1(h) = A_1 h^l + A_2 E/h^k.$$

Минимум правой части достигается при h порядка $E^{1/(l+k)}$, при этом сама правая часть имеет порядок $E^{l/(l+k)}$. Таким образом, с ростом l порядок погрешности по отношению к E повышается; при этом значение шага, соответствующее минимуму оценки погрешности, становится все больше. Конечно, следует иметь в виду, что величины A_1 и A_2 могут расти с ростом l, поэтому увеличение l разумно лишь в определенных пределах.

Иногда складывается обстановка, когда повышение точности формул численного дифференцирования не приводит к требуемому результату. Тогда применяются методы предварительного сглаживания исследуемой функции. Одна группа методов базируется на идеях математической статистики. За счет обработки большого числа наблюдаемых значений функции уменьшается случайная погрешность в ее значениях. Другая группа методов, получающая распространение в последнее время, использует идеи регуляризации. О методах этой группы подробнее будет сказано в последующем.

§ 17. Рациональная интерполяция

В ряде случаев большую точность приближения можно достигнуть, используя рациональную интерполяцию. При заданных $f(x_1), \ldots, f(x_n)$ приближение к $f(x)$ ищется в виде

$$R(x) = \frac{a_0 + a_1 x + \cdots + a_p x^p}{b_0 + b_1 x + \cdots + b_q x^q}, \quad p + q + 1 = n.$$

Коэффициенты a_i, b_i находятся из совокупности соотношений $R(x_j) = f(x_j)$, $j = 1, \ldots, n$, которые можно записать в виде

$$\sum_{j=0}^{p} a_j x_i^j - f(x_i) \sum_{j=0}^{q} b_j x_i^j = 0, \quad i = 1, \ldots, n. \tag{1}$$

Уравнения (1) образуют систему n линейных алгебраических уравнений относительно $n + 1$ неизвестных.

Функция $R(x)$ может быть записана в явном виде в случаях, когда n нечетное и $p = q$, и когда n четное и $p - q = 1$.

Для этого следует вычислить так называемые обратные разделенные разности, определяемые условиями

$$f^-(x_l;\, x_k) = \frac{x_l - x_k}{f(x_l) - f(x_k)}$$

и рекуррентным соотношением

$$f^-(x_k; \ldots;\, x_l) = \frac{x_l - x_k}{f^-(x_{k+1}; \ldots;\, x_l) - f^-(x_k; \ldots;\, x_{l-1})}.$$

Интерполирующая рациональная функция записывается в виде цепной дроби

$$f(x) = f(x_1) + \cfrac{x - x_1}{f^-(x_1;\, x_2) + \cfrac{x - x_2}{f^-(x_1;\, x_2;\, x_3) + \cdots + \cfrac{x - x_{n-1}}{f^-(x_1; \ldots;\, x_n)}}}.$$

Использование рациональной интерполяции по подходящим образом выбранным узлам часто целесообразнее интерполяции многочленами в случае функций с нерегулярным характером поведения (резкое изменение или особенности производных в отдельных точках).

Литература

1. Бабенко К. И. Основы численного анализа. — М.: Наука, 1986.
2. Бахвалов Н. С. Численные методы. — М.: Наука, 1975.
3. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интерполирование и интегрирование. — Минск: Наука и техника, 1983.
4. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т.1. — М.: Наука, 1976.
5. Локуциевский О. В., Гавриков М. Б. Начала численного анализа. — М.: ТОО «Янус», 1995.

Глава 3

Численное интегрирование

Эта глава посвящена методам приближенного вычисления одномерных интегралов. Сначала строятся простейшие формулы для приближенного вычисления интегралов по отрезку. Такие формулы называют *квадратурными*. В многомерном случае (когда размерность интеграла больше единицы) формулы для приближенного вычисления интеграла называют *кубатурными*.

Изучается вопрос о повышении точности вычисления интегралов за счет повышения порядка точности квадратур (т. е. повышения степени полиномов, для которых квадратуры точны), за счет разбиения отрезка на части, за счет сведения интегралов от функций с «особенностями» к интегралам от более гладких функций.

На примере численного интегрирования иллюстрируются требования, предъявляемые к стандартным программам и алгоритмам, которые кладутся в их основу. Даются описания ряда стандартных программ численного интегрирования.

§ 1. Простейшие квадратурные формулы. Метод неопределенных коэффициентов

Простейшие квадратурные формулы можно получить из наглядных соображений. Пусть вычисляется интеграл

$$I = \int_a^b f(x)\,dx. \tag{1}$$

Если $f(x) \approx \text{const}$ на рассматриваемом отрезке $[a, b]$, то можно положить $I \approx (b-a)f(\zeta)$, ζ — произвольная точка на $[a, b]$. Естественно взять в качестве ζ среднюю точку отрезка; тогда получим *формулу прямоугольников*

$$I \approx (b-a)f\left(\frac{a+b}{2}\right).$$

Предположим, что функция $f(x)$ на $[a, b]$ близка к линейной; тогда естественно заменить интеграл площадью трапеции с высотой $(b-a)$ и основаниями $f(a)$ и $f(b)$ (рис. 3.1.1). Получим *формулу трапеций*

$$I \approx (b-a)\frac{f(a)+f(b)}{2}.$$

Если функция $f(x)$ близка к линейной, то из наглядных соображений видно, что формула прямоугольников также должна давать неплохой результат: дело в том, что $(b-a)f\left(\frac{a+b}{2}\right)$ есть площадь любой трапеции с высотой $b-a$ и средней линией $f\left(\frac{a+b}{2}\right)$; в частности, она равна площади трапеции, у которой одна из сторон лежит на касательной к графику функции в точке $\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right)$ (рис. 3.1.2).

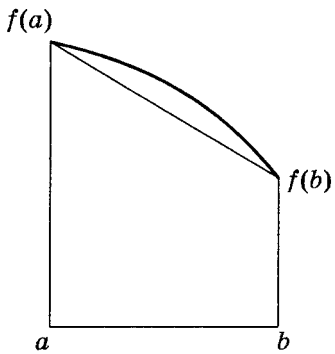


Рис. 3.1.1

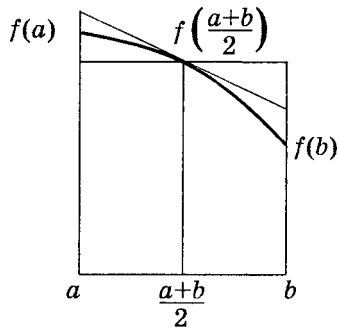


Рис. 3.1.2

Более сложные квадратурные формулы, так же как и формулы численного дифференцирования, строятся методом неопределенных коэффициентов или при помощи аппарата интерполирования.

Рассмотрим простейший пример построения квадратуры методом неопределенных коэффициентов. Строится квадратурная формула

$$\int_{-1}^{1} f(x)\,dx \approx S(f) = c_1 f(-1) + c_2 f(0) + c_3 f(1),$$

точная для многочленов наиболее высокой степени. Погрешность квадратуры

$$R(f) = \int_{-1}^{1} f(x)\,dx - S(f)$$

является линейным функционалом, и при $f = \sum_{j=0}^{m} a_j x^j$ имеем

$$R(f) = \sum_{j=0}^{m} a_j R(x^j).$$

Таким образом, нужно добиться выполнения равенств

$$R(1) = 0, \ldots, R(x^l) = 0$$

при возможно большем значении l. Получаем уравнения

$$\begin{aligned}
R(1) &= 2 - (c_1 + c_2 + c_3) = 0,\\
R(x) &= 0 - (-c_1 + c_3) = 0,\\
R(x^2) &= \frac{2}{3} - (c_1 + c_3) = 0,\\
R(x^3) &= 0 - (c_1 + c_3) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

Поскольку нужно определить 3 свободных параметра, то, вообще говоря, можно решить лишь первые три уравнения, из которых получаем

$$c_1 = c_3 = 1/3, \quad c_2 = 4/3.$$

В данном конкретном случае четвертое уравнение выполнено автоматически и мы получаем квадратуру, точную для многочленов третьей степени, называемую *формулой Симпсона.*

Вообще говоря, требуется вычислять интегралы не по отрезку $[-1, 1]$, а по произвольным отрезкам $[a, b]$. Переход к отрезку $[-1, 1]$ удобен тем, что для него арифметические выкладки, выполняемые при построении квадратуры, оказываются короче.

Иногда оказывается, что подынтегральная функция хорошо приближается не многочленами, а так называемыми обобщенными многочленами, т. е. линейными комбинациями вида $\sum_{j=0}^{m} b_j \varphi_j(x)$, где $\varphi_j(x)$ — какие-то конкретные линейно независимые функции. Тогда методом неопределенных коэффициентов строится квадратура, точная для функций такого вида.

Наиболее часто такие квадратуры используются в случае, когда $f(x)$ хорошо приближается функциями, представленными в виде произведения некоторой фиксированной функции $p(x)$ на многочлен, т. е. функциями вида

$$\sum_{j=0}^{m} a_j p(x) x^j.$$

В этом случае функцию $p(x)$ называют *весом* или *весовой функцией*, полагают $F(x) = f(x)/p(x)$ и исходный интеграл записывают в виде

$$\int_a^b F(x)p(x)\,dx. \tag{2}$$

Задача построения квадратуры

$$\int_a^b f(x)\,dx \approx \sum_{j=1}^{n} c_j f(x_j),$$

точной для всех функций вида $p(x)P_m(x)$, где $P_m(x)$ — многочлен степени m, заменяется задачей построения квадратуры

$$\int_a^b F(x)p(x)\,dx \approx \sum_{j=1}^{n} C_j F(x_j),$$

точной для всех многочленов степени m. В случае, когда все $p(x_j)$ отличны от нуля и бесконечности, эти задачи эквивалентны. В дальнейшем подынтегральную функцию в таких интегралах будем обозначать как $f(x)p(x)$.

Перейдем к оценке погрешности квадратурных формул.

§ 2. Оценки погрешности квадратуры

Пусть вычисляется интеграл (1.2). Если квадратура точна для многочленов $P_m(x)$ степени m, то

$$R(P_m) = I(P_m) - S(P_m) = 0,$$

поэтому

$$R(f) = R(f - P_m) + R(P_m) = R(f - P_m)$$

при любом многочлене $P_m(x)$ степени m. Оценивая в $R(g)$ каждое слагаемое, получим оценку

$$|R(g)| \leqslant \int_a^b |g(x)|\,|p(x)|\,dx + \sum_{j=1}^{n} |C_j|\,|g(x_j)| \leqslant V \sup_{[a,\,b]} |g(x)|,$$

где

$$V = \int_a^b |p(x)|\,dx + \sum_{j=1}^{n} |C_j|.$$

Поэтому

$$|R(f)| \leqslant |R(f - P_m)| \leqslant V\|f - P_m\|_C$$

при любом P_m — многочлене степени m; здесь

$$\|f - P_m\|_C = \sup_{[a,\,b]} |f(x) - P_m(x)|.$$

Взяв в правой части нижнюю грань по всем многочленам степени m, получим оценку

$$|R(f)| \leqslant V E_m(f), \tag{1}$$

где

$$E_m(f) = \inf_{P_m} \|f - P_m\|_C.$$

Построенные выше простейшие квадратурные формулы и ряд более сложных квадратур удовлетворяют условию $C_j \geqslant 0$, если $p(x) \geqslant 0$; в рассмотренных нами примерах $p(x) \equiv 1$.

Условие, что квадратура точна для многочлена нулевой степени, т. е. для функции $f \equiv 1$, имеет вид

$$I(1) = \int_a^b p(x)\,dx = S(1) = \sum_{j=1}^{n} C_j.$$

При $p(x) \geqslant 0$ и $C_j \geqslant 0$ имеем

$$\sum_{j=1}^{n} |C_j| = \sum_{j=1}^{n} C_j = \int_a^b p(x)\,dx, \tag{2}$$

поэтому $V = 2\int_a^b p(x)\,dx$. Обращаясь к (1), получим оценки

$$|R(f)| \leqslant 2\left(\int_a^b p(x)\,dx\right) E_m(f) \leqslant 2\left(\int_a^b p(x)\,dx\right) \|f - P_m\|_C,$$

где P_m — любой многочлен степени m. Если в качестве P_m взят интерполяционный многочлен по нулям многочлена Чебышева, то на основании (2.9.3) имеем

$$\|f - P_m\|_C \leqslant \frac{\left(\frac{b-a}{2}\right)^{m+1}}{2^m(m+1)!} \|f^{(m+1)}\|_C.$$

В конкретном случае для веса $p(x) \equiv 1$ и формул прямоугольников, трапеций, Симпсона, где все $C_j \geqslant 0$, имеем $V = 2(b-a)$ и

$$|R(f)| \leqslant \frac{(b-a)^{m+2}}{2^{2m}(m+1)!} \|f^{(m+1)}\|_C.$$

В частных случаях, например для формул прямоугольников и трапеций, где $m = 1$, отсюда имеем

$$|R(f)| \leqslant \frac{(b-a)^3}{8} \|f''\|_C;$$

для формулы Симпсона, где $m = 3$, имеем

$$|R(f)| \leqslant \frac{(b-a)^5}{1536} \|f^{(4)}\|_C.$$

Эти оценки одинаковы для всех квадратур, точных для многочленов какой-то определенной степени, например для формул трапеций и прямоугольников. Можно получить и более точные оценки погрешности этих квадратур.

Опишем универсальный способ получения наиболее точных оценок. В качестве $P_m(x)$ возьмем сумму первых $m+1$ членов разложения функции $f(x)$ по формуле Тейлора в какой-либо точке x_0 отрезка $[a, b]$. Для определенности возьмем $x_0 = a$ и рассмотрим случай, когда все $x_i \in [a, b]$. Пусть $P_m(x)$ — такая сумма, $r_m(x)$ — ее остаточный член:

$$f(x) = P_m(x) + r_m(x).$$

Имеем равенство

$$R(f) = R(r_m(x)) = I(r_m(x)) - \sum_{j=1}^{n} C_j r_m(x_j).$$

Остаточный член формулы Тейлора возьмем в интегральной форме:

$$r_m(x) = \int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t)\,dt.$$

В двукратном интеграле

$$I(r_m(x)) = \int_a^b p(x) \left(\int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t)dt \right) dx$$

сделаем интегрирование по t внешним, а по x — внутренним. Получим

$$I(r_m(x)) = \int_a^b K_m(t) f^{(m+1)}(t)\,dt,$$

где

$$K_m(t) = \int_t^b p(x) \frac{(x-t)^m}{m!}\,dx.$$

Таким образом, получим

$$R(f) = \int_a^b K_m(t) f^{(m+1)}(t)\,dt - \sum_{j=1}^{n} C_j \int_a^{x_j} \frac{(x_j - t)^m}{m!} f^{(m+1)}(t)\,dt.$$

Положим

$$(x_j - t)_+ = \begin{cases} x_j - t & \text{при } a \leqslant t \leqslant x_j, \\ 0 & \text{при остальных } t. \end{cases}$$

Используя это обозначение, представим погрешность $R(f)$ в виде

$$R(f) = \int_a^b Q(t) f^{(m+1)}(t)\,dt, \quad Q(t) = K_m(t) - \sum_{j=1}^{n} C_j \frac{(x_j - t)_+^m}{m!}. \tag{3}$$

Отсюда следует оценка погрешности

$$|R(f)| \leqslant \left(\int_a^b |Q(t)|\, dt \right) \|f^{(m+1)}\|_C. \tag{4}$$

Если $Q(t)$ не меняет знака на отрезке $[a, b]$, то по теореме о среднем из (3) получаем

$$|R(f)| = \left(\int_a^b Q(t)\, dt \right) f^{(m+1)}(\xi), \quad a \leqslant \xi \leqslant b. \tag{5}$$

Далее для простоты положим $p(x) \equiv 1$.

Задача 1. Предположим, что в качестве $P_m(x)$ берется сумма $(m+1)$-го члена разложения функции $f(x)$ в ряд Тейлора относительно произвольной точки $x_0 \neq a$. Доказать, что представление погрешности (3) при этом не изменится.

Задача 2. Пусть точка a фиксирована, $f^{(m+1)}(x)$ непрерывна в точке a. Доказать, что при $b \to a$, $C_i = O(b-a)$

$$\int_a^b Q(t)\, dt = O\left((b-a)^{m+2}\right),$$

$$R(f) = \left(\int_a^b Q(t)\, dt \right) f^{(m+1)}(a) + o\left((b-a)^{m+2}\right),$$

где $Q(t)$ определено в (3).

Рассмотрим для примера формулу трапеций. Тогда

$$\int_t^b (x-t)\, dt = \frac{(b-t)^2}{2},$$

$$Q(t) = \frac{(b-t)^2}{2} - \frac{(b-a)}{2}(b-t) = (a-t)(b-t) \leqslant 0;$$

подставляя $Q(t)$ в (5), получим

$$\int_a^b Q(t)\, dt = -\frac{(b-a)^3}{12},$$

т. е.

$$R(f) = -\frac{(b-a)^3}{12} f''(\xi_1).$$

В случае формулы прямоугольников

$$Q(t) = \frac{(b-t)^2}{2} - (b-a)\left(\frac{a+b}{2} - t\right) = \frac{(t-(a+b)/2)^2}{2} \geqslant 0;$$

подставляя $Q(t)$ в (5), получим

$$R(f) = \frac{(b-a)^3}{24} f''(\xi_2).$$

Часто на практике интересуются не оценкой погрешности (5), которая не поддается улучшению, а ее главным (при $(b-a)\to 0$) членом

$$f^{(m+1)}(a)\int_a^b Q(t)\,dt.$$

Если для некоторого m оказалось, что соответствующий интеграл $\int_a^b Q(t)\,dt$ равен нулю, то это значит, что квадратура точна для многочленов степени $m+1$. В этом случае надо увеличить m на 1 и провести аналогичные рассуждения для этого нового значения m.

Для вычисления главного члена погрешности можно поступить следующим образом. Представим $f(x)$ в виде суммы первых $(m+2)$-х членов разложения Тейлора относительно некоторой точки $x_0\in[a,b]$ и остаточного члена; при этом, объединив первые $m+1$ слагаемых в многочлен степени m, получим

$$f(x)=T_{x_0}^m(x)+\frac{(x-x_0)^{m+1}}{(m+1)!}f^{(m+1)}(x_0)+r_{m+1}(x),$$

где

$$T_{x_0}^m(x)=\sum_{i=0}^{m}\frac{(x-x_0)^i}{i!}f^{(i)}(x_0),\quad r_{m+1}(x)=o\left((x-x_0)^{m+1}\right).$$

Вследствие линейности функционала погрешности имеем равенство

$$R(f)=R(T_{x_0}^m(x))+\frac{f^{(m+1)}(x_0)}{(m+1)!}R\left((x-x_0)^{m+1}\right)+R(r_{m+1}(x)).$$

Первое слагаемое обращается в нуль, так как квадратура точна для многочленов степени m. Поскольку

$$|R(r_{m+1}(x))|\leqslant V\|r_{m+1}\|_C=o\left((b-a)^{m+2}\right)$$

(при условии, что $V<\infty$), то

$$\frac{f^{(m+1)}(x_0)}{(m+1)!}R\left((x-x_0)^{m+1}\right)$$

является главным членом погрешности $R(f)$. Для простоты выкладок при конкретном вычислении $R\left((x-x_0)^{m+1}\right)$ часто удобно произвести замену переменных

$$x=(a+b)/2+ht,\quad h=(b-a)/2,$$

и рассматривать разложение в ряд Тейлора относительно точки $t=0$.

Задача 3. Проверить, что

$$\frac{1}{(m+1)!}R\left((x-x_0)^{m+1}\right)=\int_a^b Q(t)\,dt.$$

Задача 4. Доказать, что $R\left((x-x_0)^{m+1}\right)$ не зависит от выбора x_0.

В частном случае для формулы трапеций имеем

$$R(x-a)^2 = \frac{(b-a)^3}{3} - \frac{1}{2}(b-a)^3 = -\frac{1}{6}(b-a)^3,$$

поэтому погрешность $R(f)$ с точностью до членов высшего порядка имеет вид

$$R(f) \sim -\frac{1}{12}(b-a)^3 f''(a).$$

§ 3. Квадратурные формулы Ньютона—Котеса

Рассмотренные далее квадратуры относятся к большой группе квадратурных формул, полученных с помощью интегрирования интерполяционного многочлена и объединенных под одним названием — квадратурные формулы Ньютона—Котеса. Зададимся некоторыми $d_1, \ldots, d_n \in [-1, 1]$ и построим интерполяционный многочлен $L_n(x)$ степени $n-1$, совпадающий с $f(x)$ в точках $x_j = \dfrac{a+b}{2} + \dfrac{b-a}{2} d_j$. Положим

$$\int_a^b f(x)p(x)\,dx \approx \int_a^b L_n(x)p(x)\,dx.$$

Имеем

$$R_n(f) = \int_a^b f(x)p(x)\,dx - \int_a^b L_n(x)p(x)\,dx = \int_a^b p(x)(f(x) - L_n(x))\,dx.$$

Разность $f(x) - L_n(x)$ оценим, воспользовавшись оценкой погрешности интерполяционного многочлена Лагранжа

$$|f(x) - L_n(x)| \leqslant \left(\max_{[a,b]} \left|f^{(n)}(x)\right|\right) \frac{|\omega_n(x)|}{n!},$$

где $\omega_n(x) = (x - x_1)\ldots(x - x_n)$. Отсюда

$$|R_n(f)| \leqslant \left(\max_{[a,b]} \left|f^{(n)}(x)\right|\right) \int_a^b \frac{|\omega_n(x)|\ |p(x)|}{n!}\,dx.$$

Произведем в последнем интеграле замену переменных, положив $x = X(t) = \dfrac{a+b}{2} + \dfrac{b-a}{2} t$. Тогда

$$\frac{1}{n!}\int_a^b \left|\,\omega_n(x)p(x)\,\right| dx = D(d_1, \ldots, d_n)\left(\frac{b-a}{2}\right)^{n+1},$$

где

$$D(d_1, \ldots, d_n) = \int_{-1}^{1} \frac{\left|\omega_n^0(t)p^0(t)\right|}{n!}\,dt,$$

$$\omega_n^0(t) = (t-d_1)\ldots(t-d_n), \quad p^0(t) = p\left(\frac{a+b}{2} + \frac{b-a}{2}t\right).$$

Таким образом, справедлива оценка

$$|R_n(f)| \leqslant D(d_1, \ldots, d_n)\left(\max_{[a\,b]}\left|f^{(n)}(x)\right|\right)\left(\frac{b-a}{2}\right)^{n+1}. \tag{1}$$

Пусть все d_j различны. Тогда

$$L_n(x) = \sum_{j=1}^{n} f(x_j)\prod_{i\neq j}\frac{x-x_i}{x_j-x_i}.$$

После замены переменных $x = X(t)$ получим

$$\int_a^b p(x)L_n(x)\,dx = \left(\frac{b-a}{2}\right)\sum_{j=1}^{n} D_j f(x_j), \tag{2}$$

где

$$D_j = \int_{-1}^{1} p^0(t)\prod_{i\neq j}\frac{t-d_i}{d_j-d_i}\,dt. \tag{3}$$

Таким образом, построенная квадратурная формула имеет вид

$$\int_a^b f(x)p(x)\,dx \approx \frac{b-a}{2}\sum_{j=1}^{n} D_j f\left(\frac{a+b}{2} + \frac{b-a}{2}d_j\right). \tag{4}$$

Как и при численном дифференцировании, можно обнаружить следующие обстоятельства: если задача имеет определенную симметрию, то метод с симметрией того же типа часто обладает дополнительными преимуществами.

Будем называть функцию *четной* относительно точки x_0, если $f(x-x_0) = f(x_0-x)$, и *нечетной*, если $f(x-x_0) = -f(x_0-x)$.

Можно показать, что для весовой функции $p(x)$, четной относительно середины отрезка $[a, b]$, и узлов x_j, расположенных симметрично середины отрезка, т. е. $d_j = -d_{n+1-j}$, коэффициенты квадратуры, соответствующие симметричным узлам, равны между собой:

$$D_j = D_{n+1-j}. \tag{5}$$

(Доказать!)

Такие «симметричные» квадратуры обладают следующим дополнительным свойством, которое, формально говоря, не предусматривалось при их построении. Они точны для любой функции, нечетной относительно середины отрезка $[a, b]$, т. е. удовлетворяющей условию $f\left(x - \frac{a+b}{2}\right) = -f\left(\frac{a+b}{2} - x\right)$. В самом деле, для таких функций

$\int_a^b p(x)f(x)\,dx = 0$ вследствие четности $p(x)$, а $\sum_{j=1}^{n} D_j f(x_j) = 0$ вследствие (5); поэтому и $R_n(f) = 0$. В частности, квадратуры будут точны для любого одночлена вида $\text{const}\cdot\left(x - \frac{a+b}{2}\right)^{2l+1}$. Свойство симметрии (5) помогает также при непосредственном построении формул методом неопределенных коэффициентов.

Рассмотрим теперь симметричную квадратуру, соответствующую нечетному n. Она точна для $f(x) = \text{const}\cdot\left(x - \frac{a+b}{2}\right)^{n}$ и, согласно построению, точна и для любого многочлена степени $n-1$. Следовательно, такая квадратура будет точна и для любого многочлена степени n. Таким образом, построенные квадратуры с $2q-1$ и $2q$ узлами, с симметричным расположением узлов оказываются точными для многочленов одинаковой степени $2q-1$ (для квадратур Гаусса (см. § 5) эта степень выше).

Чтобы получить уточненную оценку погрешности квадратур с нечетным числом узлов не через $f^{(n)}(x)$, а через $f^{(n+1)}(x)$, следует заменить подынтегральную функцию интерполяционным многочленом Лагранжа, имеющим точку $(a+b)/2$ двукратным узлом интерполирования. Ниже для случая $p(x) \equiv 1$ строится ряд элементарных квадратурных формул и дается оценка их погрешности; при $n=1$ и $n=3$ для симметричных формул производится оценка погрешности через $f^{(n+1)}(x)$.

1. Формула прямоугольников. Пусть $n=1$, $d_1 = 0$. Тогда

$$D = \int_{-1}^{1} |t|\,dt = 1, \quad D_1 = \int_{-1}^{1} 1\cdot dt = 2$$

и имеем квадратурную формулу

$$\int_a^b f(x)\,dx \approx (b-a)f\left(\frac{a+b}{2}\right) \tag{6}$$

с оценкой остаточного члена

$$R(f) = \max_{[a,b]} |f'(x)|\,\frac{(b-a)^2}{4}.$$

2. Формула прямоугольников как формула с кратным узлом. Пусть $n=2$, $d_1 = d_2 = 0$. Тогда

$$D = \int_{-1}^{1} \frac{t^2}{2}\,dt = \frac{1}{3},$$

$$L_2(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right),$$

$$\int_a^b L_2(x)\,dx = (b-a)f\left(\frac{a+b}{2}\right).$$

Таким образом, имеем ту же квадратурную формулу

$$\int_a^b f(x)\,dx \approx (b-a)f\left(\frac{a+b}{2}\right)$$

с оценкой остаточного члена

$$R(f) = \max_{[a,\,b]} \left|f''(x)\right| \frac{(b-a)^3}{24}.$$

В обоих случаях получилась одна и та же квадратурная формула, но с различной оценкой остаточного члена.

3. Формула трапеций. Пусть $n = 2$, $d_1 = -1$, $d_2 = 1$. Тогда

$$D = \int_{-1}^{1} \frac{|t^2-1|}{2}\,dt = \frac{2}{3}, \quad D_1 = \int_{-1}^{1} \frac{1-t}{2}\,dt = 1, \quad D_2 = \int_{-1}^{1} \frac{1+t}{2}\,dt = 1.$$

Получена *формула трапеций*

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{2}(f(a)+f(b)) \tag{7}$$

с оценкой остаточного члена

$$R(f) = \max_{[a,\,b]} \left|f''(x)\right| \frac{(b-a)^3}{12}.$$

4. Формула Симпсона. Пусть $n = 4$, $d_1 = -1$, $d_2 = d_4 = 0$, $d_3 = 1$. Тогда

$$D = \int_{-1}^{1} \frac{t^2(1-t^2)}{4!}\,dt = \frac{1}{90}.$$

Согласно формуле интерполирования с кратными узлами, можем написать, что

$$L_4(x) = L_3(x) + f\left(a;\, b;\, \frac{a+b}{2};\, \frac{a+b}{2}\right)(x-a)(x-b)\left(x-\frac{a+b}{2}\right),$$

где

$$L_3(x) = f(a) + f(a;\, b)(x-a) + f\left(a;\, b;\, \frac{a+b}{2}\right)(x-a)(x-b).$$

Второе слагаемое в выражении $L_4(x)$ является функцией, нечетной относительно середины отрезка $[a,\, b]$, поэтому

$$\int_a^b L_4(x)dx = \int_a^b L_3(x)\,dx.$$

Многочлен $L_3(x)$ является интерполяционным многочленом второй степени, соответствующим $d_1 = -1$, $d_2 = 0$, $d_3 = 1$ (рис. 3.3.1). Этим значениям d_1, d_2, d_3 соответствуют $D_1 = 1/3$,

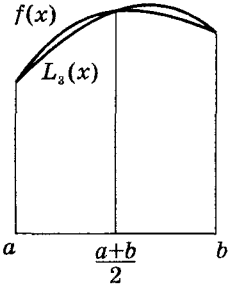


Рис. 3.3.1

$D_2 = 4/3$, $D_3 = 1/3$. В результате получаем *квадратурную формулу Симпсона*

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) \tag{8}$$

с оценкой остаточного члена

$$R(f) = \max_{[a,\,b]} \left|f^{(4)}(x)\right| \frac{(b-a)^5}{2880}.$$

Основной целью настоящей главы является рассмотрение способов вычисления интегралов от функций, заданных аналитическим выражением, и выработка принципов построения стандартных программ интегрирования таких функций. Естественно, что кроме этих задач в теории квадратурных формул имеются и другие задачи, например связанные с обработкой экспериментального материала.

Для примера обратим внимание на квадратурные формулы Чебышева, широко применявшиеся при подсчете водоизмещения судов. Постановка задачи, приведшая к построению этих квадратур, довольно близка к постановке задачи, возникающей при планировании экспериментов (см. гл. 2 § 1). Вычисляется интеграл $\int_{-1}^{1} f(x)\,dx$, причем известно, что функция $f(x)$ с приемлемой точностью может быть приближена многочленом степени q. Получение каждого значения $f(x)$, например путем измерений, обходится довольно дорого, и получаемые значения содержат довольно большие случайные погрешности. Предположим, что погрешности измерений независимы, имеют одинаковую дисперсию d и математическое ожидание, равное нулю. Тогда дисперсия приближенного значения $S_n(f)$, вычисляемого по квадратурной формуле

$$I(f) \approx S_n(f) = \sum_{j=1}^{n} c_j f(x_j),$$

равна $d\sum_{j=1}^{n} c_j^2$. Условие $I(f) = S_n(f)$ при $f = \text{const}$ имеет вид

$$\sum_{j=1}^{n} c_j = 2. \tag{9}$$

Как нетрудно проверить, минимум величины $d\sum_{j=1}^{n} c_j^2$ при условии (9) достигается при $c_1 = \cdots = c_n = 2/n$. Эти рассуждения привели к следующей постановке задачи: среди всех квадратур

$$I(f) \approx \frac{2}{n}\sum_{j=1}^{n} f(x_j),$$

точных для многочленов степени q, найти квадратуру, соответствующую наименьшему n. При $q = 0$ и $q = 1$ искомой будет квадратура прямоугольников $I(f) = 2f(0)$; при $q = 2$ и $q = 3$ — квадратура Гаусса (см. § 5).

§ 4. Ортогональные многочлены

Решения ряда задач математической физики часто исследуют, отыскивая их разложения по ортогональным функциям, в частности по ортогональным многочленам. Наиболее подробно изучены ортогональные системы функций одной переменной. Из ортогональных систем функций многих переменных рассматривают, как правило, лишь системы вида $\varphi^1_{n_1}(x_1) \times \cdots \times \varphi^s_{n_s}(x_s)$, где $\varphi^k_{n_k}(x_k)$ — некоторые ортогональные многочлены *одной* переменной.

Пусть H — пространство комплекснозначных функций, определенных на $[a, b]$, с ограниченным интегралом

$$\int_a^b |f(x)|^2 p(x)\,dx;$$

скалярное произведение задается равенством

$$(f, g) = \int_a^b f(x)\bar{g}(x)p(x)\,dx, \tag{1}$$

где $\bar{g}(x)$ — функция, комплексно сопряженная с $g(x)$; $p > 0$ почти всюду на $[a, b]$ и $\int_a^b p(x)dx < \infty$; функции, отличающиеся друг от друга на множестве меры 0, считаются равными.

Система $\Phi_n = \{\varphi_1, \ldots, \varphi_n\}$ ненулевых элементов из H называется *ортогональной*, если $(\varphi_i, \varphi_j) = 0$ при $i \neq j$. Система $\Phi_n = \{\varphi_1, \ldots, \varphi_n\}$ называется *линейно независимой*, если

$$\sum_{j=1}^n C_j\varphi_j = 0$$

только тогда, когда все $C_j = 0$.

Важным аппаратом многих исследований является ортогонализация заданной системы элементов гильбертова пространства.

Лемма 1. *Пусть в пространстве H задана линейно независимая система элементов $\Phi_n = \{\varphi_1, \ldots, \varphi_n\}$. Тогда можно построить ортогональную линейно независимую систему $\Psi_n = \{\psi_1, \ldots, \psi_n\}$ элементов вида*

$$\psi_j = \sum_{i=1}^j b_{ji}\varphi_i, \quad j = 1, \ldots, n, \tag{2}$$

где $b_{jj} = 1$.

Доказательство. Мы будем проводить построение такой системы методом индукции. При $n = 1$ имеем тривиальную систему $\psi_1 = \varphi_1$. Пусть требуемая система Ψ_n построена при некотором $n = k$; тогда элемент ψ_{k+1} отыскиваем в виде

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^k a_{ki}\psi_i. \tag{3}$$

Коэффициенты a_{ki} выбираем из условия ортогональности $(\psi_{k+1}, \psi_l) = 0$ при $l \leqslant k$. Вследствие ортогональности системы элементов Ψ_k последнее соотношение представится в виде

$$(\varphi_{k+1}, \psi_l) - a_{kl}(\psi_l, \psi_l) = 0,$$

откуда

$$a_{kl} = \frac{(\varphi_{k+1}, \psi_l)}{(\psi_l, \psi_l)};$$

следовательно, элемент

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^{k} \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)} \psi_i \tag{4}$$

будет ортогонален всем предшествующим. Подставляя в (3)

$$\psi_i = \sum_{q=1}^{i} b_{iq} \varphi_q$$

при $i \leqslant k$, получим требуемое соотношение. Лемма доказана.

Совокупность соотношений (2) при $j \leqslant n$ можно представить в виде

$$\Psi_n = B_n \Phi_n,$$

где

$$B_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ b_{21} & 1 & 0 & 0 & \cdots \\ b_{31} & b_{32} & 1 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{pmatrix},$$

и Ψ_n, Φ_n являются вектор-столбцами из соответствующих элементов. В то же время, перенося все ψ_i из правой части (4) в левую, получим

$$\Phi_n = A_n \Psi_n,$$

где

$$A_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ a_{21} & 1 & 0 & 0 & \cdots \\ a_{31} & a_{32} & 1 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{pmatrix}; \tag{5}$$

матрицу B_n иногда называют *матрицей ортогонализации*. Так как $\det B_n = 1$, то преобразование, задаваемое матрицей B_n, является невырожденным и переводит линейно независимую систему элементов Φ_n в линейно независимую систему Ψ_n.

В силу линейной независимости системы функций Φ_n отсюда следует, что $B_n = A_n^{-1}$.

При построении ортогональных многочленов в качестве элементов системы функций Φ_j берутся функции $1, x, \ldots, x^{j-1}$ и производится ортогонализация в пространстве со скалярным произведением (1) по описанной выше процедуре. Получаемые многочлены

$$\psi_j(x) = x^{j-1} + \sum_{i=1}^{j-1} b_{ji} x^{i-1} \tag{6}$$

называют *ортогональными многочленами, соответствующими весу* $p(x)$ *и отрезку* $[a, b]$. Иногда ортогональными многочленами, соответствующими весу $p(x)$, называют многочлены $g_j(x) = \alpha_j \psi_j(x)$, в которых величины α_j подбирают из каких-либо дополнительных соображений, например из условия $\|g_j\| = \sqrt{(g_j, g_j)} = 1$. Систему ортогональных элементов, удовлетворяющих такому условию, называют *ортонормированной.*

Мы уже имели дело с системой многочленов Чебышева

$$T_n(x) = 2^{n-1} x^n + \cdots,$$

ортогональных на отрезке $[-1, 1]$ с весом $2/\left(\pi\sqrt{1-x^2}\right)$. Как отмечалось, значения этих многочленов можно вычислять по рекуррентной формуле

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{7}$$

Вычисление значений ортогональных многочленов Чебышева при помощи формулы (7) более предпочтительно по сравнению с непосредственным вычислением их по явной формуле (6) по следующим причинам.

1. Вычисление по формуле (7) не требует хранения в памяти или вычисления коэффициентов b_{ji}.

2. Обычно требуется вычислять одновременно значения всех многочленов $\psi_1(x), \ldots, \psi_n(x)$ в одной и той же точке. При независимом вычислении значения каждого многочлена по формуле (6) вычисление значений всех многочленов потребует $\sim n^2$ арифметических операций. (Здесь и далее $a(n) \sim b(n)$ означает, что $a(n)$ и $b(n)$ одного порядка, т. е. $a(n) = O(b(n))$, $b(n) = O(a(n))$.)

При одновременном вычислении всех значений при помощи рекуррентного соотношения (7) потребуется $O(n)$ арифметических операций.

3. Значения $T_n(x)$, получаемые при непосредственном вычислении по формуле (6), могут содержать большую вычислительную погрешность.

Дело заключается в следующем: пусть $T_n(x)$ образуется как сумма слагаемых:

$$T_n(x) = \sum_{j=0}^{n} d_{nj} x^j. \tag{8}$$

При записи в машине эти слагаемые $d_{nj}x^j$ смогут приобрести абсолютную погрешность порядка $|d_{nj}x^j|2^{-t}$. Следствием этого может быть погрешность значения $T_n(x)$ порядка $D_n(x)2^{-t}$, где

$$D_n(x) = \sum_{j=0}^{n} |d_{nj}x^j|.$$

Оценим снизу $D_n(x)$. Из равенства

$$T_n(|x|\mathbf{i}) = \sum_{j=0}^{n} d_{nj}(|x|\mathbf{i})^j,$$

где $\mathbf{i}$ — мнимая единица, следует оценка

$$|T_n(|x|\mathbf{i})| \leqslant \sum_{j=0}^{n} |d_{nj}x^j| = D_n(x).$$

В то же время, согласно (2.8.4), при действительном x имеем

$$T_n(|x|\mathbf{i}) = \frac{\left(|x| + \sqrt{1+|x|^2}\right)^n + \left(|x| - \sqrt{1+|x|^2}\right)^n}{2}\,\mathbf{i}^n.$$

Так как

$$\left(|x| + \sqrt{1+|x|^2}\right)\left(|x| - \sqrt{1+|x|^2}\right) = -1,$$

то отсюда получаем

$$D_n(x) \geqslant |T_n(|x|\mathbf{i})| \geqslant \frac{\left(|x| + \sqrt{1+|x|^2}\right)^n - \left(|x| + \sqrt{1+|x|^2}\right)^{-n}}{2}.$$

Таким образом, при больших n и $x \neq 0$ при непосредственном использовании формулы (8) вычислительная погрешность может достигать значений порядка $\left(|x| + \sqrt{1+|x|^2}\right)^n 2^{-t}$.

Задача 1. Доказать равенство

$$D_n(x) = |T_n(|x|\mathbf{i})|.$$

Мы опять столкнулись здесь с явлением пропадания значащих цифр в вычислениях: $|T_n(x)| \leqslant 1$, но при вычислении значения $T_n(x)$ из (8) оно получается как сумма больших по модулю слагаемых переменного знака и поэтому приобретает большую погрешность.

В то же время можно показать, что при вычислении по рекуррентной формуле (7) погрешность $T_n(x)$ имеет порядок $\min\left\{n, \dfrac{1}{\sqrt{1-x^2}}\right\} \cdot O(n2^{-t})$. Из изложенного видна важность получения рекуррентных соотношений типа (7), связывающих значения ортогональных многочленов, соответствующих и другим весовым функциям $p(x)$.

Справедлива

Теорема (без доказательства). *Ортогональные многочлены*

$$\psi_j(x) = x^j + \sum_{i=0}^{j-1} b_{ji} x^{i-1}$$

связаны соотношениями

$$\psi_{j+1}(x) = (x + \alpha_j)\psi_j(x) - \beta_j \psi_{j-1}(x), \tag{9}$$

где $\beta_j > 0$.

При $Q_n(x) = \dfrac{\psi_n(x)}{\|\psi_n(x)\|}$ вместо (9) имеем

$$D_{n+1} Q_{n+1}(x) = (2x + G_n) Q_n(x) - D_n Q_{n-1}(x).$$

Если отрезок $[a, b]$ конечен, то известно, что $D_n \to 1$, $G_n \to 0$ при $n \to \infty$.

Приведем наиболее употребительные системы ортогональных многочленов, соответствующие различным весовым функциям.

1. Многочлены Якоби. Для отрезка $[-1, 1]$ и весовой функции $p(x) = (1-x)^\alpha (1+x)^\beta$, $\alpha, \beta > -1$, ортогональную систему образуют *многочлены Якоби*

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!} (1-x)^{-\alpha} (1+x)^{-\beta} \frac{d^n}{dx^n} \Big((1-x)^{\alpha+n} (1+x)^{\beta+n} \Big).$$

Имеют место соотношения

$$\left\| P_n^{(\alpha,\beta)}(x) \right\| = \left(\frac{2^{\alpha+\beta+1} \Gamma(\alpha+n+1)\Gamma(\beta+n+1)}{n!(\alpha+\beta+2n+1)\Gamma(\alpha+\beta+n+1)} \right)^{1/2},$$

$$\begin{aligned} (\alpha+\beta+2n)(\alpha+\beta+2n+1)(\alpha+\beta+2n+2) x P_n^{(\alpha,\beta)}(x) = \\ = 2(n+1)(\alpha+\beta+n+1)(\alpha+\beta+2n) P_{n+1}^{(\alpha,\beta)}(x) + \\ + (\beta^2 - \alpha^2)(\alpha+\beta+2n+1) P_n^{(\alpha,\beta)}(x) + \\ + 2(\alpha+n)(\beta+n)(\alpha+\beta+2n+2) P_{n-1}^{(\alpha,\beta)}(x); \end{aligned} \tag{10}$$

здесь Γ — гамма-функция Эйлера.

Многочлены Якоби удовлетворяют дифференциальному уравнению

$$L_{\alpha,\beta}\Big(P_n^{(\alpha,\beta)}(x) \Big) \equiv (1-x^2) \Big(P_n^{\alpha,\beta}(x) \Big)_{xx} +$$

$$+ \big((\beta - \alpha) - (\alpha + \beta + 2)x \big) \Big(P_n^{(\alpha,\beta)}(x) \Big)_x = -n(\alpha+\beta+n+1) P_n^{(\alpha,\beta)}(x).$$

Иначе говоря, они являются собственными функциями дифференциального оператора $L_{\alpha,\beta}$.

2. Многочлены Лежандра. Частным случаем многочленов Якоби при $\alpha = \beta = 0$, т. е. при весовой функции $p(x) \equiv 1$, являются *многочлены Лежандра*

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n$$

с нормой

$$\|L_n\| = \sqrt{2/(2n+1)},$$

удовлетворяющие рекуррентному соотношению

$$(n+1)L_{n+1}(x) - (2n+1)xL_n(x) + nL_{n-1}(x) = 0.$$

3. Многочлены Чебышева первого рода. При $\alpha = \beta = -1/2$, $p(x) = 1/\sqrt{1-x^2}$ многочлены Якоби после перенормировки превращаются в *многочлены Чебышева первого рода* $T_n(x)$.

4. Многочлены Чебышева второго рода. При $\alpha = \beta = -1/2$, $p(x) = \sqrt{1-x^2}$ многочлены Якоби после перенормировки превращаются в *многочлены Чебышева второго рода*

$$U_n(x) = (\sin((n+1)\arccos x))/\sqrt{1-x^2} = T'_{n+1}(x)/(n+1), \quad n = 0, 1, \ldots,$$

с нормой $\|U_n\| = \pi/2$. Рекуррентное соотношение для многочленов Чебышева второго рода такое же, как для многочленов Чебышева первого рода.

5. Многочлены Эрмита. При $(a, b) = (-\infty, \infty)$ и $p(x) = e^{-x^2}$ ортогональную систему образуют *многочлены Эрмита*

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}\left(e^{-x^2}\right)$$

с нормой

$$\|H_n\| = \sqrt{2^n \cdot n! \sqrt{\pi}},$$

удовлетворяющие рекуррентному соотношению

$$H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) = 0.$$

Многочлены Эрмита удовлетворяют дифференциальному уравнению

$$H''_n - 2xH'_n = -2nH_n.$$

6. Многочлены Лагерра. При $[a, b) = [0, \infty)$ и $p(x) = x^\alpha e^{-x}$, $\alpha > -1$, ортогональную систему образуют *многочлены Лагерра*

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} e^x \frac{d^n}{dx^n}\left(x^{\alpha+n} e^{-x}\right)$$

с нормой

$$\left\|L_n^{(\alpha)}\right\| = \sqrt{n!\Gamma(\alpha+n+1)}.$$

Для них справедливо рекуррентное соотношение

$$L_{n+1}^{(\alpha)}(x) - (x - \alpha - 2n - 1)L_n^{(\alpha)}(x) + n(\alpha+n)L_{n-1}^{(\alpha)}(x) = 0.$$

Многочлены Лагерра удовлетворяют дифференциальному уравнению

$$x\left(L_n^{(\alpha)}(x)\right)_{xx} + (\alpha + 1 - x)\left(L_n^{(\alpha)}(x)\right)_x = -nL_n^{(\alpha)}(x).$$

Рекуррентные соотношения для конкретных ортогональных многочленов, выписанные в пп. 2–5, имеют несколько иной вид, чем (9), поскольку соотношение (9) выписано для ортогональных многочленов, нормированных так, что их старший коэффициент равен 1.

Отметим ряд свойств ортогональных многочленов. Пусть $P_0(x), \ldots, P_n(x)$ — система ортогональных многочленов на отрезке $[a, b]$ вида

$$P_n(x) = x^n + \sum_{j=0}^{n-1} p_{nj}x^j.$$

Лемма. *Каждый многочлен $P_n(x)$ имеет ровно n различных нулей на открытом интервале (a, b).*

Доказательство. Предположим, что $P_n(x)$ имеет на (a, b) только $l < n$ нулей $x_1, \ldots, x_l$ нечетной кратности. Тогда многочлен

$$P_n(x)\prod_{j=1}^{l}(x - x_j)$$

не меняет знак на $[a, b]$, поэтому

$$\int_a^b P_n(x)\prod_{j=1}^{l}(x - x_j)p(x)\,dx \neq 0.$$

С другой стороны, этот интеграл равен нулю, поскольку $P_n(x)$ ортогонален всем многочленам меньшей степени. Получили противоречие.

Задача 2. Пусть $x_1^{(n)} < \cdots < x_n^{(n)}$ — нули $P_n(x)$. Тогда нули многочленов $P_{n-1}(x)$ и $P_n(x)$ перемежаются, т. е.

$$a < x_1^{(n)} < x_1^{(n-1)} < \cdots < x_{n-1}^{(n-1)} < x_n^{(n)} < b.$$

Это свойство ортогональных многочленов используется при составлении таблиц нулей ортогональных многочленов, являющихся узлами квадратур Гаусса.

Задача 3. Пусть вес является четной функцией относительно середины отрезка $[a, b]$ и, для определенности, $[a, b] = [-1, 1]$, т. е. $p(x) = p(-x)$. Доказать, что все многочлены $P_{2n}(x)$ четные, многочлены $P_{2n-1}(x)$ нечетные, т. е.

$$P_j(x) = (-1)^j P_j\left((-1)^j x\right) \quad \text{при всех} \quad j,$$

и рекуррентное соотношение (9) имеет вид

$$P_{j+1}(x) - xP_j(x) + \beta_{j+1}P_{j-1}(x) = 0.$$

При обработке результатов наблюдений возникает задача приближения функций, заданных на множестве точек $x_1, \ldots, x_N$ вещественной оси с помощью многочленов от переменной x. Эта задача часто решается с помощью ортогональных многочленов дискретной переменной. В теории таких многочленов установлены их свойства, аналогичные свойствам ортогональных многочленов непрерывной переменной; построены дискретные аналоги для всех рассмотренных выше типов ортогональных многочленов непрерывной переменной.

Отметим одно важное свойство распределения нулей ортогональных многочленов. Пусть $[a, b] = [-1, 1]$, вес $p(x)$ почти всюду положителен на $[-1, 1]$. Обозначим через $w_n(x_1, x_2)$ число нулей многочлена $P_n(x)$, принадлежащих отрезку $[x_1, x_2]$. Тогда справедливо соотношение

$$\lim_{n\to\infty} \frac{w_n(x_1, x_2)}{n} = \int_{x_1}^{x_2} \frac{1}{\pi\sqrt{1-x^2}}\, dx.$$

Таким образом, нули ортогональных многочленов независимо от весовой функции $p(x)$ распределены асимптотически одинаково, с плотностью $1/(\pi\sqrt{1-x^2})$.

§ 5. Квадратурные формулы Гаусса

Из оценки (2.1) следует, что погрешность квадратуры оценивается через погрешность приближения функции многочленами. Функция приближается многочленами более высокой степени точнее, чем многочленами низшей степени:

$$E_0(f) \geqslant E_1(f) \geqslant \cdots$$

Поэтому есть основания обратить внимание на квадратуры, точные для многочленов по возможности более высокой степени.

Рассмотрим следующую оптимизационную задачу. При заданном числе узлов n построить квадратуру

$$I(f) = \int_a^b f(x)p(x)\, dx \approx S_n(f) = \frac{b-a}{2} \sum_{j=1}^{n} D_j f(x_j), \tag{1}$$

точную для многочленов наиболее высокой степени. Такие квадратуры называют *квадратурами Гаусса.*

Мы видели (§ 2), что квадратура (1) точна для многочленов степени m, если она точна для всех функций x^q, $q = 0, \ldots, m$. Следовательно, должны выполняться соотношения

$$R_n(x^q) = \int_a^b x^q p(x)\, dx - \frac{b-a}{2} \sum_{j=1}^{n} D_j x_j^q = 0, \quad q = 0, \ldots, m. \tag{2}$$

Получили систему из $(m+1)$-го уравнения относительно неизвестных $x_1, \ldots, x_n, D_1, \ldots, D_n$, где $x_1, \ldots, x_n$ — неизвестные узлы, а $D_1, \ldots, D_n$ — неизвестные коэффициенты квадратурной формулы (1).

При $m \leqslant 2n-1$ число уравнений не превосходит числа неизвестных, поэтому можно ожидать, что алгебраическая система (2) имеет решение. Можно попытаться построить квадратурные формулы, соответствующие значению $m = 2n-1$, решая эту систему, однако неясно, будут ли узлы квадратур, получаемые из (2), принадлежать отрезку $[a, b]$. В противном случае может оказаться, что функция $f(x)$ не определена в узлах интегрирования и употребление квадратуры невозможно.

Заметим, что в гл. 8 при построении конечно-разностных методов решения обыкновенных дифференциальных уравнений возникнут квадратуры с узлами вне отрезка $[a, b]$.

Займемся построением квадратур, соответствующих максимальному значению $m = 2n-1$.

Лемма 1. *Если $x_1, \ldots, x_n$ — узлы квадратуры (1), точной для всех многочленов степени $2n-1$, то*

$$\int_a^b \omega_n(x) P_{n-1}(x)\, dx = 0$$

при $\omega_n(x) = (x-x_1)\ldots(x-x_n)$ и $P_{n-1}(x)$ — произвольном многочлене степени не выше $n-1$.

Доказательство. Пусть $P_{n-1}(x)$ — некоторый многочлен степени не выше $n-1$. Вследствие условия леммы квадратура (1) точна для многочлена $Q_{2n-1}(x) = \omega_n(x)P_{n-1}(x)$ степени $2n-1$. Поэтому

$$\int_a^b \omega_n(x)P_{n-1}(x)p(x)\, dx = \int_a^b Q_{2n-1}(x)p(x)\, dx = \\ = \frac{b-a}{2}\sum_{j=1}^{n} D_j Q_{2n-1}(x_j) = 0.$$

Последнее соотношение вытекает из равенства $Q_{2n-1}(x_j) = 0$ при всех j. Лемма 1 доказана.

Далее предполагается, что $p(x) > 0$ почти всюду на $[a, b]$.

Из результатов § 4 вытекает единственность многочлена $\psi_n(x)$, $\psi_n(x) = x^n + \cdots$, ортогонального всем многочленам низшей степени, если скалярное произведение задано соотношением

$$(f, g) = \int_a^b f(x)\bar{g}(x)p(x)\, dx.$$

Поэтому $\psi_n(x) = \omega_n(x)$ и узлы отыскиваемой квадратуры должны быть нулями $\psi_n(x)$. Согласно результатам § 4 многочлен $\psi_n(x)$ на (a, b) имеет n различных нулей.

Лемма 2. *Пусть $x_1, \ldots, x_n$ — нули ортогонального многочлена $\psi_n(x)$ степени n и (1) — квадратура, точная для многочленов степени $n-1$. Тогда квадратура (1) точна для многочленов степени $2n-1$.*

Доказательство. Произвольный многочлен $Q_{2n-1}(x)$ степени $2n-1$ представим в виде

$$Q_{2n-1}(x) = \psi_n(x)g_{n-1}(x) + r_{n-1}(x),$$

где g_{n-1} и r_{n-1} — многочлены степени $n-1$. Имеем

$$R_n(Q_{2n-1}) = R_n(\psi_n g_{n-1}) + R_n(r_{n-1}) = R_n(\psi_n g_{n-1}),$$

так как $R_n(r_{n-1}) = 0$ по условию леммы. Далее,

$$R_n(\psi_n g_{n-1}) = \int_a^b \psi_n(x)g_{n-1}(x)p(x)\,dx - \frac{b-a}{2}\sum_{j=1}^{n} D_j\psi_n(x_j)g_{n-1}(x_j) = 0,$$

поскольку $\displaystyle\int_a^b \psi_n(x)g_{n-1}(x)p(x)\,dx = 0$ вследствие свойства ортогональности многочлена $\psi_n(x)$ многочленам низшей степени, а все $\psi_n(x_j) = 0$ по предположению леммы. Следовательно, $R_n(Q_{2n-1}) = 0$. Лемма 2 доказана.

Теперь можно построить требуемую квадратурную формулу. Для этого зададимся узлами интерполяции $x_1, \ldots, x_n$, в которых $\psi_n(x_j) = 0$, и построим (например, следуя построениям § 3) квадратуру, точную для многочленов степени $n-1$. В итоге получим требуемую квадратуру

$$\int_a^b f(x)p(x)\,dx \approx \frac{b-a}{2}\sum_{j=1}^{n} D_j f(x_j), \tag{3}$$

точную для многочленов степени $2n-1$.

Если почти всюду $p(x) > 0$, то не существует квадратуры, точной для всех многочленов степени $2n$. В самом деле, возьмем $Q_{2n}(x) = (x - x_1)^2 \ldots (x - x_n)^2$; тогда левая часть (1)

$$\int_a^b \big((x-x_1)\ldots(x-x_n)\big)^2 p(x)\,dx > 0,$$

а правая равна 0.

Лемма 3. *Коэффициенты D_j положительны.*

Доказательство. Функция $\left(\frac{\psi_n(x)}{x-x_l}\right)^2$ является многочленом степени $2n-2$, обращающимся в нуль во всех точках $x_j \neq x_l$. Квадратура (3) будет точна для этой функции, поэтому

$$\int_a^b \left(\frac{\psi_n(x)}{x-x_l}\right)^2 p(x)\,dx = \frac{b-a}{2} D_l \left(\frac{\psi_n(x)}{x-x_l}\right)^2 \Bigg|_{x_l}.$$

Раскрывая выражение $\psi_n(x)/(x-x_l)$, получим

$$D_l = \frac{2}{b-a}\int_a^b \left(\prod_{j\neq l} \frac{x-x_j}{x_l-x_j}\right)^2 p(x)\,dx > 0.$$

Лемма доказана.

Поскольку все $D_j > 0$, то, воспользовавшись (2.1) и (2.2), имеем

$$|R_n(f)| \leqslant 2\left(\int_a^b p(x)\,dx\right) E_{2n-1}(f). \tag{4}$$

Можно получить также оценку погрешности квадратур Гаусса через $f^{(2n)}(x)$. Эта оценка имеет вид

$$R_n(f) = f^{(2n)}(\bar{\zeta}) \int_a^b \frac{\psi_n^2(x)}{(2n)!} p(x)\,dx. \tag{5}$$

Для практического применения формул Гаусса необходимо иметь в распоряжении узлы и коэффициенты этих квадратур. Можно показать, что для случая $p(x)$ — четной относительно точки $(a+b)/2$, нули ортогональных многочленов, т. е. узлы квадратур Гаусса, расположены симметрично относительно середины отрезка $[a, b]$. Вследствие (3.5) коэффициенты квадратуры Гаусса (3) будут удовлетворять условию четности $D_j = D_{n+1-j}$. Это обстоятельство наполовину уменьшает объем таблиц для формул Гаусса.

Если $p(x) \equiv 1$, то коэффициенты D_j и числа $d_j = \dfrac{2x_j - (a+b)}{b-a}$ не зависят от отрезка $[a, b]$. В самом деле, если многочлен $\psi_n(x) = (x-x_1)\ldots(x-x_n)$ принадлежит системе многочленов, ортогональных с весом 1 на $[a, b]$, то многочлен $(t-d_1)\ldots(t-d_n)$ принадлежит системе многочленов, ортогональных с весом 1 на $[-1, 1]$. Поэтому он сам, его нули, а согласно (3.3) и коэффициенты D_j определяются однозначно, независимо от исходного отрезка $[a, b]$.

Приведем для сведения параметры квадратур Гаусса для отрезка $[-1, 1]$ при $p(x) \equiv 1$. В этом случае остаточный член $R(f)$ для квадратурной формулы (3) есть

$$R(f) = f^{(2n)}(\zeta)\frac{2^{2n+1}(n!)^4}{(2n!)^3(2n+1)}.$$

Вследствие свойства симметрии мы указываем лишь неотрицательные d_j и коэффициенты при них (табл. 1).

Таблица 1

	d_1 D_1	d_2 D_2	d_3 D_3
1	0,0000000000 2,0000000000		
2	0,5773502692 1,0000000000		
3	0,0000000000 0,8888888888	0,7745966692 0,5555555556	
4	0,8611363115 0,3478548451	0,3399810436 0,6521451549	
5	0,0000000000 0,5688888888	0,9061798459 0,4786286705	0,5384693101 0,2369268851
6	0,9324695142 0,1713244924	0,6612093864 0,3607615730	0,2386191861 0,4679139346

В настоящее время составлены таблицы узлов и весов квадратур Гаусса по крайней мере до $n = 4096$ с 20 десятичными знаками. Вследствие их большого объема, начиная с некоторого n_0, их публикуют лишь для $n = 2^k$, $n = 3 \cdot 2^k$.

Иногда целесообразно видоизменить идею Гаусса построения квадратур, точных на многочленах максимально высокой степени. Например, пусть требуется вычислить $\int_0^1 f(x)\,dx$, а значение $f(a)$ вычисляется существенно быстрее, чем значения в других точках отрезка $[0, 1]$ (или почему-либо заранее известно). Тогда имеет смысл построить квадратуру

$$\int_0^1 f(x)\,dx \approx \sum_{j=0}^{n} D_j f(d_j), \quad d_0 = a,$$

точную для многочленов степени $2n$. Если требуется вычислить $\int_{-1}^{1} f(x)\,dx$, а значения $f(1)$ и $f(-1)$ вычисляются существенно быстрее, чем значения во внутренних точках отрезка $[-1, 1]$, то имеет смысл построить квадратуру

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{j=0}^{n} l_j f(d_j), \quad d_0 = -1, \quad d_n = 1, \tag{6}$$

точную для многочленов степени $2n-1$; в последнем случае оказывается, что $d_j = d_{n-j}$, $l_j = l_{n-j}$ при всех j.

Степень многочлена, для которого точна квадратура, определяется числом свободных параметров квадратуры. Квадратура (6) называется *квадратурой Лобатто* или *формулой Маркова*; при $n = 1$ она совпадает с формулой трапеций, при $n = 2$ — с формулой Симпсона.

Задача 1. Введением весовых функций и заменой переменных $x = \varphi(t)$ свести построение квадратур (6) к построению некоторых квадратур Гаусса.

Задача 2. Пусть $[a, b] = [-1, 1]$, $p(x) = \dfrac{1}{\sqrt{1-x^2}}$. Доказать, что соответствующей квадратурой Гаусса является

$$I(f) = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \approx S_n(f) = \frac{\pi}{n} \sum_{j=1}^{n} f(x_j),$$

где $x_j = \cos \dfrac{(2j-1)\pi}{2n}$ — нули многочлена Чебышева $T_n(x)$.

Указание. При проверке точности квадратуры для многочлена степени $2n-1$ представить многочлен в виде

$$\sum_{m=0}^{2n-1} a_m T_m(x)$$

и установить, что квадратура точна для $T_m(x)$ при $m < 2n$.

В настоящее время рассчитано много таблиц формул Гаусса и формул типа Лобатто, в частности, при

$$[a, b] = [-1, 1], \quad p(x) = 1,$$

а также в более общем случае при

$$[a, b] = [-1, 1], \quad p(x) = (1+x)^{\alpha}(1-x)^{\beta}, \quad \alpha, \beta > -1,$$

и при

$$[a, b] = [0, \infty), \quad p(x) = x^{\alpha} e^{-x}, \quad \alpha > -1.$$

Если подынтегральная функция интеграла

$$I(f) = \int_0^\omega f(x)\,dx$$

хорошо приближается тригонометрическими многочленами с периодом ω, то целесообразно применить квадратуру, являющуюся аналогом квадратуры Гаусса для этого случая вида

$$I(f) \approx S_N(f) = \frac{\omega}{N} \sum_{j=0}^{N-1} f\left(\frac{j\omega}{N}\right). \tag{7}$$

Имеем равенство

$$S_N\left(\exp\left\{2\pi m\mathbf{i}\,\frac{x}{\omega}\right\}\right) = \frac{\omega}{N} \sum_{j=0}^{N-1} \exp\left\{2\pi m\mathbf{i}\,\frac{j\omega}{N\omega}\right\} =$$

$$= \begin{cases} \dfrac{\omega}{N} \displaystyle\sum_{j=0}^{N-1} 1 = \omega & \text{при } \dfrac{m}{N} \text{ целом,} \\ \dfrac{\exp\{2\pi m\mathbf{i}\} - 1}{\exp\{2\pi m\mathbf{i}/N\} - 1} = 0 & \text{при } \dfrac{m}{N} \text{ не целом.} \end{cases}$$

В то же время

$$I\left(\exp\left\{2\pi m\mathbf{i}\,\frac{x}{\omega}\right\}\right) = \begin{cases} \omega & \text{при } m = 0, \\ 0 & \text{при } m \neq 0. \end{cases}$$

Следовательно, квадратура (7) точна для функции $\cos(2\pi mx/\omega)$ при $m = 0$ или при m/N не целом и для всех функций $\sin(2\pi mx/\omega)$. В результате этого оказывается, что квадратура точна для любого тригонометрического многочлена

$$t_N(x) = a_0 + \sum_{m=0}^{N-1} \left(a_m \cos\left(2\pi m\frac{x}{\omega}\right) + b_m \sin\left(2\pi m\frac{x}{\omega}\right)\right) + b_N \sin\left(2\pi N\frac{x}{\omega}\right); \tag{8}$$

следовательно,

$$R_N(f) = R_N(f - t_N) + R_N(t_N) = R_N(f - t_N).$$

Аналогично (4) получаем оценку

$$\left|R_N(f)\right| \leqslant 2\omega \inf_{t_N} \max_{[0,\omega]} \left|f(x) - t_N(x)\right|.$$

Нижняя грань берется по множеству всех многочленов вида (8).

Задача 3. Доказать, что не существует квадратур с N узлами, точных для всех тригонометрических многочленов степени N.

§ 6. Практическая оценка погрешности элементарных квадратурных формул

Выше получены ряд квадратурных формул и строгие оценки погрешности для них. Однако это не решает всех проблем задачи численного интегрирования. Важнейшей задачей вычислительной математики является создание алгоритмов и пакетов программ, обеспечивающих получение решения задач с заданной точностью при минимальном объеме затрат человеческого труда и работы машины. Практическое применение полученных выше оценок требует аналитических выкладок и поэтому достаточно большого объема работы исследователя; кроме того, эти оценки часто оказываются слишком завышенными. Поэтому при создании таких систем обычно отказываются от использования подобных оценок, зачастую жертвуя строгой гарантией малости погрешности приближенного решения.

Можно говорить, что задача от ее возникновения до получения результата проходит через некоторую систему, состоящую из людей, решающих задачу, и ЭВМ. На первоначальном этапе применения ЭВМ наиболее узким местом, тормозившим работу этой системы, являлось недостаточное количество ЭВМ. Поэтому применение аналитических методов решения или аналитическое проведение оценок погрешности было оправданным.

Однако теперь, с повсеместным распространением вычислительной техники и внедрением ее в различные сферы деятельности общества, обстановка меняется. Узким местом этой системы становятся длительность выбора математической модели, метода решения задачи, программирования и других этапов, предшествующих непосредственному решению задачи на ЭВМ. Прохождение этих этапов особенно замедляется в случае, когда решением задач на ЭВМ занимаются представители конкретных наук, например филологи, медики, экономисты, географы и т. п., мало знакомые с численными методами или программированием. Обучение их тонкостям теории численных методов может превратиться в самоцель, отвлекающую от решения основных задач их науки, и в конечном счете обойтись обществу довольно дорого. Поэтому в настоящее время важнейшей проблемой является создание систем решения задач с максимально простым обращением, предполагающих малую квалификацию пользователя в отношении численных методов и программирования. Например, естественно потребовать, чтобы к программе вычисления интеграла с заданной точностью мог обратиться исследователь, знающий, что такое интеграл, но не умеющий ни интегрировать, ни дифференцировать.

Конечно, в развитии многих областей знания и техники решающая роль математики состоит в создании математической модели явления, а потом уже в применении ЭВМ для ее исследования. При разработке модели от специалиста этой отрасли знания требуется определенная математическая культура, и наше высказывание не следует понимать как предложение полностью избавить его от математики.

Подоплекой проводимых здесь рассуждений является следующее известное рассуждение. Когда мы занимаемся решением каких-то задач, то нужно учитывать эффективность нашей работы не только по совокупным затратам на решение этих задач, но и принимать во внимание убыток, понесенный обществом в результате того, что нами не решены некоторые другие, возможно более важные задачи.

При практическом анализе погрешности численного интегрирования часто пользуются различными полуэмпирическими приемами. Наиболее распространенным из этих приемов является следующий. Производятся вычисления по двум квадратурным формулам

$$I(f)=\int_a^b f(x)\,dx\approx S^k(f)=\frac{b-a}{2}\sum_{j=1}^{N_k}D_j^k f\left(x_j^k\right),\quad k=1,2;$$

далее некоторая линейная комбинация $S(f)=S^1(f)+\theta(S^2(f)-S^1(f))$ этих квадратур принимается за приближенное значение интеграла, а величина $\rho=\left|S^2(f)-S^1(f)\right|$ — за меру погрешности приближенной формулы $I(f)\approx S(f)$. Довольно типичным является случай $\theta=0$.

Описанный выше подход нельзя считать полностью оправданным вследствие его неоднозначности.

Пусть, например, $S^1(f)$ — формула Симпсона:

$$\int_{-1}^1 f(x)\,dx\approx S^1(f)=\frac{f(-1)+4f(0)+f(1)}{3},$$

$S^2(f)$ — формула трапеций:

$$S^2(f)=f(-1)+f(1)$$

и $\theta=0$. Тогда в качестве меры погрешности выступает величина $\rho=\frac{2}{3}\left|f(1)-2f(0)+f(-1)\right|$. Если

$$S^2(f)=2f(0)$$

— формула прямоугольников, то соответствующее значение $\rho=\frac{1}{3}\left|f(1)-2f(0)+f(-1)\right|$. Таким образом, мы получили две различные эмпирические оценки погрешности одной и той же формулы Симпсона.

Попытаемся прояснить ситуацию. Выражение $S(f)$ является некоторой квадратурной суммой

$$S(f)=\frac{b-a}{2}\sum_{j=1}^{N}D_j f(x_j)\qquad(1)$$

по совокупности узлов x_j, принадлежащих объединению узлов, соответствующих квадратурам $S^1(f)$ и $S^2(f)$. В то же время

$$\rho=|l(f)|,$$

где

$$l(f)=\frac{b-a}{2}\sum_{j=1}^{N}B_j f(x_j).\qquad(2)$$

Возьмем произвольную линейную комбинацию вида (2) и положим

$$S^1(f) = S(f), \quad S^2(f) = S(f) - \theta l(f).$$

Тогда мы получим приближенное значение интеграла $I(f) \approx S(f)$ с оценкой погрешности $\rho = \theta\,|l(f)|$. Мы видим, что на таком пути можно получить неограниченное множество оценок погрешности одной и той же квадратуры (1).

Рассматриваемую задачу можно формулировать следующим образом. Приближенное значение интеграла вычисляется по формуле

$$I(f) \approx S(f) = \frac{b-a}{2} \sum_{j=1}^{N} D_j f(x_j). \tag{3}$$

Требуется построить выражение вида (2), дающее представление о погрешности квадратуры (3).

Предположим, что погрешность квадратуры (3) представляется в виде

$$R(f) = \overline{D}(b-a)^{m+1} f^{(m)}(\zeta). \tag{4}$$

Рассмотрим случай $m < N$. Тогда в качестве $l(f)$ можно взять величину

$$l(f) = \overline{D}(b-a)^{m+1} m! f(x_{i_1}; \ldots; x_{i_{m+1}}),$$

где $x_{i_1}, \ldots, x_{i_{m+1}}$ — различные узлы квадратуры (3). Разделенная разность может быть выражена через производную, поэтому имеем

$$l(f) = \bar{D}(b-a)^{m+1} f^{(m)}(\zeta), \quad a < \zeta < b.$$

Следовательно, при $a = \text{const}$, $f^{(m)}(a) \neq 0$, $(b-a) \to 0$ справедливо соотношение $R(f) \approx l(f)$ и величину $\rho = |l(f)|$ можно принять за меру погрешности.

Пусть, например, оценивается погрешность формулы трапеций

$$\int_a^b f(x)\,dx \approx S(f) = \frac{b-a}{2}\left(f(a) + 0 \cdot f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Согласно оценкам из § 3 имеем

$$R(f) = -\frac{(b-a)^3}{12} f''(\zeta).$$

Таким образом, мы можем принять за меру погрешности величину

$$\frac{b-a}{3}\left|f(b) - 2f\left(\frac{a+b}{2}\right) + f(a)\right|.$$

Иначе обстоит дело, когда $m \geqslant N$. Тогда нельзя получить никакого приближения к $f^{(m)}(\zeta)$ через величины $f(x_1), \ldots, f(x_N)$, и проблема получения эмпирической оценки погрешности в рассматриваемой выше постановке не может быть решена.

Например, мы не можем получить удовлетворительного представления об оценке погрешности формулы Симпсона через значения $f(a)$, $f\left(\frac{a+b}{2}\right)$, $f(b)$. Однако можно получить некоторую завышенную оценку погрешности.

Рассмотрим один подход к разрешению возникшей проблемы. Предположим, что нам удалось получить оценку погрешности вида

$$|R(f)| \leqslant (b-a)^N \bar{D} \max_{[a,b]} \left|f^{(N-1)}(x)\right| = \sigma.$$

Положим

$$\rho = (b-a)^N \bar{D}(N-1)! \left|f(x_1; \ldots; x_N)\right|. \tag{5}$$

При $a = \text{const}$, $f^{(N-1)}(a) \neq 0$, $(b-a) \to 0$ имеем $\sigma \sim \rho$. Таким образом, величину ρ можно принять за приближенную оценку погрешности формулы (3). Эта оценка будет сильно завышенной, поскольку при предположении $m \geqslant N$ имеем $R(f) = O\left((b-a)^{N+1}\right)$. Однако лучшей оценки погрешности формулы (3) по сравнению с оценкой через σ, по-видимому, нельзя предложить. В случае формулы Симпсона верна оценка (5) при $\bar{D} = \frac{1}{81}$ и, таким образом, за меру погрешности принимаем величину

$$\frac{4(b-a)}{81} \left|f(b) - 2f\left(\frac{a+b}{2}\right) + f(a)\right|. \tag{6}$$

В случае многомерных интегралов все практические способы оценки погрешности опираются на исходную, раскритикованную нами процедуру. Дело в том, что в многомерном случае погрешность оценивается через значения нескольких производных подынтегральной функции. Получение «обоснованных» оценок, подобных (6), для таких формул крайне затруднительно. Поэтому обращаются к исходной процедуре с последующей экспериментальной проверкой результатов ее применения.

§ 7. Интегрирование быстро осциллирующих функций

Пусть требуется вычислить интеграл

$$\int_a^b f(x) \exp\{\mathrm{i}\omega x\}\, dx,$$

где $\omega(b-a) \gg 1$, $f(x)$ — гладкая функция. Функции

$$\operatorname{Re}\left(f(x)\exp\{\mathrm{i}\omega x\}\right), \quad \operatorname{Im}\left(f(x)\exp\{\mathrm{i}\omega x\}\right)$$

имеют на рассматриваемом отрезке примерно $\omega(b-a)/\pi$ нулей. Поскольку многочлен степени n имеет не более n нулей, то такие функции

могут быть хорошо приближены многочленами степени n лишь при $n \gg \omega(b-a)/\pi$. Поэтому для непосредственного вычисления интегралов от таких функций потребуется применение квадратур, точных для многочленов высокой степени.

Более выгодным может оказаться путь рассмотрения функции $\exp\{\mathrm{i}\omega x\}$ как весовой.

Как и в § 1, зададимся узлами интерполирования

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j, \quad j = 1, \ldots, n$$

и заменим исходный интеграл на $\int_a^b L_n(x) \exp\{\mathrm{i}\omega x\}\, dx$, где $L_n(x)$ — интерполяционный многочлен с узлами x_j. Последний интеграл может быть вычислен в явном виде

$$\int_a^b L_n(x) \exp\{\mathrm{i}\omega x\}\, dx = S_n^\omega(f) = \frac{b-a}{2} \exp\left\{\mathrm{i}\omega \frac{b+a}{2}\right\} \sum_{j=1}^n D_j\left(\omega \frac{b-a}{2}\right) f(x_j),$$

где

$$D_j(p) = \int_{-1}^1 \left(\prod_{k \neq j} \frac{\xi - d_k}{d_j - d_k}\right) \exp\{\mathrm{i}p\xi\}\, d\xi. \tag{1}$$

Получилась квадратурная формула

$$\int_a^b f(x) \exp\{\mathrm{i}\omega x\}\, dx \approx S_n^\omega(f) \tag{2}$$

с остаточным членом

$$R_n(f) = \int_a^b (f(x) - L_n(x)) \exp\{\mathrm{i}\omega x\}\, dx.$$

В соответствии с (3.1)

$$R_n(f) \leqslant \int_a^b |f(x) - L_n(x)|\, dx \leqslant D(d_1, \ldots, d_n)\left(\max_{[a,b]} \left|f^{(n)}(x)\right|\right)\left(\frac{b-a}{2}\right)^{n+1}.$$

Вычисление интегралов такого рода является типичной задачей, встречающейся при разложении функций в ряды Фурье, при построении диаграмм направленности антенн и т. д.

В стандартных программах вычисления интегралов от быстро осциллирующих функций используются формулы (1), (2), соответствующие случаям: $n = 3$, $d_1 = -1$, $d_2 = 0$, $d_3 = 1$ (эту формулу называют *формулой Филона*) или $n = 5$, $d_1 = -1$, $d_2 = -0.5$, $d_3 = 0$, $d_4 = 0.5$, $d_5 = 1$.

Если формулы (1), (2) использовать для вычисления интегралов от функций, не являющихся быстро осциллирующими, то может возникнуть

следующая неприятность, которую мы проиллюстрируем для $n = 2$, $d_1 = -1$, $d_2 = 1$. В этом случае

$$D_1(p) = \int_{-1}^{1} \frac{1-\xi}{2} \exp\{\mathbf{i}p\xi\}\, d\xi = \frac{\sin p}{p} + \frac{p\cos p - \sin p}{p^2}\,\mathbf{i},$$

$$D_2(p) = \int_{-1}^{1} \frac{1+\xi}{2} \exp\{\mathbf{i}p\xi\}\, d\xi = \frac{\sin p}{p} - \frac{p\cos p - \sin p}{p^2}\,\mathbf{i}.$$

При $p \to 0$ имеем

$$\frac{p\cos p - \sin p}{p^2} = -\frac{p}{3} + O(p^3) \to 0, \qquad \frac{\sin p}{p} \to 1.$$

Таким образом, $D_1(p)$, $D_2(p) \to 1$ при $p \to 0$.

Пусть p — малое число. Функции $\sin p$ и $p\cos p$ вычисляются в машине с погрешностями $O(2^{-t})$ и $O(p2^{-t})$ соответственно. Вследствие этого коэффициенты $D_1(p)$, $D_2(p)$ приобретают погрешность $O(2^{-t}/p)$. При $n > 2$ оказывается, что погрешность коэффициентов $D_j(p)$, вычисляемых по формулам (1), может оказаться величиной порядка $2^{-t}/p^{n-1}$. Например, при $t = 30$, $n = 5$, $p = 0{,}01$ такая погрешность уже недопустима. Поэтому стандартные программы вычисления интегралов от быстро осциллирующих функций должны иметь специальный блок, предусматривающий изменение расчетных формул при малых p с тем, чтобы избежать существенного влияния вычислительной погрешности.

Если n не очень велико, например $n = 2$, то можно пойти по следующему пути: при $|p| > p_n$, где некоторое p_n подбираем экспериментально, вычисления производим по формулам (1), (2); если $|p| \leqslant p_n$, то вычисляем исходный интеграл по формуле трапеций (3.7), рассматривая всю функцию $f(x)\exp\{\mathbf{i}\omega x\}$ как подынтегральную. В рассматриваемом случае ($n = 2$, $d_1 = -1$, $d_2 = 1$) формула (3.7) приобретает вид

$$\int_a^b f(x)\exp\{\mathbf{i}\omega x\}\, dx \approx \frac{b-a}{2}[\exp\{\mathbf{i}\omega a\}f(a) + \exp\{\mathbf{i}\omega b\}f(b)]. \tag{3}$$

Формулы (1), (3) можно объединить в одну формулу

$$\int_a^b f(x)\exp\{\mathbf{i}\omega x\}\, dx = \frac{b-a}{2}\exp\left\{\mathbf{i}\omega\frac{a+b}{2}\right\}[D_1(p)f(a) + D_2(p)f(b)], \tag{4}$$

где

$$p = \omega\frac{b-a}{2},$$

$$D_{1,2}(p) = \begin{cases} \dfrac{\sin p}{p} \pm \dfrac{p\cos p - \sin p}{p^2}\,\mathbf{i} & \text{при} \quad |p| > p_2, \\ \exp\{\mp p\mathbf{i}\} & \text{при} \quad |p| \leqslant p_2. \end{cases} \tag{5}$$

В стандартных программах вычисления интегралов рассматриваемого типа применяются квадратуры вида (4), (5). Возникает вопрос: зачем усложнять

стандартную программу? Может быть, проще написать стандартную программу вычисления по формуле (2), стандартную программу вычисления по формуле трапеций и дать инструкцию: при больших $|p|$ обращаться к первой из стандартных программ, при малых $|p|$ — ко второй? Обсудим целесообразность такого подхода. Нашей целью является создание максимальных удобств пользователю ЭВМ. Если в описании правил использования стандартной программы написано слишком много, то пользователь может не понять того, что написано, и

1) обратиться к другой программе, худшей по качеству, но имеющей лучшее описание или более простое обращение;

2) воспользоваться программой неправильно и не получить результатов; например, в рассматриваемом случае, применив метод расчета по формулам (1), (2) при $p = 0$, он получит аварийную остановку ЭВМ;

3) воспользоваться программой неправильно и получить неверный результат, предполагая, что он верный; например, так случится, если он воспользуется формулами (1), (2) при $2^{-t}/p^{n-1}$ порядка 1.

Очевидно, что последний случай влечет за собой наиболее неприятные для пользователя последствия.

При выборе метода для стандартной программы надо стремиться к тому, чтобы описание программы, составляемой на базе этого метода, было по возможности простым и кратким. Надо иметь в виду, что всякое дополнительное высказывание может быть истолковано неправильно в ущерб для области применения программы.

Приведем один поучительный реальный пример. В описании одной из стандартных программ вычисления кратных интегралов было написано: «Применение настоящей программы нецелесообразно, если число узлов берется бóльшим 100 000». Спустя короткое время программа была практически изъята из употребления. Оказалось, что среди первых пользователей программы подавляющее большинство сразу задавались числом узлов 100 000, при этом на вычисление простого интеграла уходило слишком много машинного времени. Весть об этом распространилась повсюду, и вскоре к программе перестали обращаться. На самом деле при ее использовании большинство реальных интегралов вычислялось с приемлемой точностью при числе узлов порядка 1 000. Число 100 000 было указано лишь как ориентировочная верхняя граница значений, при которых вычислительная погрешность еще не оказывает катастрофического влияния на результат.

§ 8. Повышение точности интегрирования за счет разбиения отрезка на равные части

Из оценки (2.1) вытекает, что погрешность квадратуры оценивается через погрешность, с которой функция $f(x)$ может быть приближена многочленами степени m. Поэтому может показаться естественным добиваться увеличения точности за счет повышения степени многочленов, для которых эта квадратура точна. Однако такой путь содержит свои «подвод-

ные камни». При неудачном выборе узлов может оказаться, что величина $\sum_{j=1}^{n} |C_j|$ растет вместе с n. Тогда в оценке (2.1) величина V также растет вместе с n и может оказаться, что уменьшение $E_m(f)$ с ростом n не компенсирует увеличение V.

Например, для простейшего равномерного распределения узлов

$$d_j^{(n)} = -1 + \frac{2(j-1)}{n-1}, \quad j = 1, \ldots, n, \quad [a, b] = [-1, 1]$$

оказывается, что при $m = n - 1$ имеем $\log_2 V \sim n$. В результате этого, например, для аналитической функции $f(x) = (1 + 25x^2)^{-1}$

$$|R(f)| = \left| \int_{-1}^{1} f(x)\, dx - S_n(f) \right| \to \infty$$

при $n \to \infty$.

Рассмотрим случай, когда отрезок интегрирования есть $[-1, 1]$, и сформулируем общую теорему, указывающую на необходимость осторожного обращения с формулами, точными для многочленов очень высокой степени. Пусть при каждом n мы имеем квадратуру

$$\int_{-1}^{1} f(x)\, dx \approx \sum_{j=1}^{n} C_j^{(n)} f\left(d_j^{(n)}\right), \tag{1}$$

точную для многочленов степени $n-1$. Обозначим через $w_n(x_1, x_2)$ число узлов квадратуры (1), принадлежащих отрезку $[x_1, x_2]$.

Теорема (без доказательства). *Пусть существует отрезок* $[x_1, x_2] \in [-1, 1]$ *такой, что*

$$\frac{w_n(x_1, x_2)}{n} \not\to \int_{x_1}^{x_2} \frac{dx}{\pi\sqrt{1 - x^2}} \quad \textit{при} \quad n \to \infty.$$

Тогда можно указать $b \neq 0$ *и* a *такие, что для одной из аналитических функций*

$$\operatorname{Re}\left(\frac{1}{x - (a + b\mathbf{i})}\right) \quad \textit{или} \quad \operatorname{Im}\left(\frac{1}{x - (a + b\mathbf{i})}\right)$$

будет выполняться соотношение

$$\lim_{n\to\infty} |R_n(f)| = \infty.$$

Таким образом, узлы квадратур (1), точных для многочленов степени $n-1$, при больших n должны располагаться с такой же плотностью, как нули ортогональных многочленов, т. е. как и узлы квадратур Гаусса. Иначе такие квадратуры нельзя рассматривать как универсальные.

Перепишем оценку (5.4) погрешности формул Гаусса:

$$|R_n(f)| \leqslant 2\left(\int_a^b p(x)\,dx\right) E_{2n-1}(f). \tag{2}$$

Пусть подынтегральная функция $f(x)$ непрерывна. Тогда, согласно теореме Вейерштрасса, при любом $\varepsilon > 0$ найдется многочлен $P_q(x)$, для которого $\max\limits_{[a,\,b]} |f(x) - P_q(x)| \leqslant \varepsilon$, откуда следует, что $E_m(f) \to 0$ при $m \to \infty$. Таким образом, для любой непрерывной функции $f(x)$ погрешность формул Гаусса $R_n(f) \to 0$ при $n \to \infty$.

Задача 1. Пусть $f(x)$ — функция, интегрируемая по Риману. Доказать, что для формул Гаусса $R_n(f) \to 0$ при $n \to \infty$.

Из сказанного выше видно, что формулы Гаусса могли бы быть положены в основу универсальных программ вычисления интегралов с заданной точностью. При этом придется вводить в ЭВМ каким-либо образом узлы и веса этих квадратур.

Во многих случаях возникает задача вычисления интегралов, где подынтегральная функция или ее производные невысокого порядка имеют участки резкого изменения, например обращаются в бесконечность. Такие функции плохо приближаются многочленами сразу на всем отрезке интегрирования. Здесь часто оказывается более выгодным разбить исходный отрезок на части и на каждой части применять свою квадратурную формулу, Гаусса или какую-либо другую.

Пусть вычисляется интеграл $I = \int_A^B f(x)\,dx$. Разобьем отрезок $[A, B]$ на M частей $[a_{q-1}, a_q]$, где $a_0 = A$, $a_M = B$. Для вычисления интеграла по каждой из частей применим какую-либо квадратурную формулу из §§ 1, 3 вида

$$I_q = \int_{a_{q-1}}^{a_q} f(x)\,dx \approx s_q = \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{n} C_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \tag{3}$$

с оценкой остаточного члена

$$R_q(f) \leqslant D\Big(\max_{[a_{q-1},\,a_q]} \left|f^{(m)}(x)\right|\Big) \left(\frac{a_q - a_{q-1}}{2}\right)^{m+1}.$$

В результате интеграл по всему отрезку будет аппроксимирован суммой

$$S_M^n(f) = \sum_{q=1}^{M} s_q = \sum_{q=1}^{M} \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{n} C_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \tag{4}$$

с оценкой остаточного члена

$$\left|R_M^n(f)\right| = \left|I - S_M^n(f)\right| \leqslant D\sum_{q=1}^{M}\left(\max_{[a_{q-1},\,a_q]}\left|f^{(m)}(x)\right|\right)\left(\frac{a_q - a_{q-1}}{2}\right)^{m+1}. \qquad (5)$$

Выражение (4) часто называют *составной* или *обобщенной квадратурной формулой.*

Рассмотрим наиболее простой для исследования случай, когда отрезки разбиения имеют одинаковую длину $a_q - a_{q-1} \equiv H$. Тогда оценка погрешности (5) после замены $\max\limits_{[a_{q-1},\,a_q]}\left|f^{(m)}(x)\right|$ на величину $A_m = \max\limits_{[A,\,B]}\left|f^{(m)}(x)\right|$ приобретает вид

$$|R_M^n(f)| \leqslant \bar{D}A_m(B-A)H^m, \quad \text{где} \quad \bar{D} = 2^{-(m+1)}D, \qquad (6)$$

или

$$|R_M^n(f)| \leqslant DA_m\frac{(B-A)^{m+1}}{M^m}. \qquad (7)$$

Приведем конкретные квадратурные формулы и оценки погрешности для частных случаев формул (3).

1. Составная формула трапеций с постоянным шагом интегрирования. В этом случае при постоянном шаге $a_q - a_{q-1} \equiv H$ формула (3) приобретает вид

$$\int_{a_0}^{a_M} f(x)\,dx \approx H\left(\frac{f(a_0)}{2} + f(a_1) + \cdots + f(a_{M-1}) + \frac{f(a_M)}{2}\right),$$

а остаточный член оценивается следующим образом:

$$\left|R_M^2(f)\right| \leqslant A_2\frac{(a_M - a_0)H^2}{12} = A_2\frac{(a_M - a_0)^3}{12M^2}. \qquad (8)$$

2. Составная формула Симпсона с постоянным шагом интегрирования. При постоянном шаге $a_q - a_{q-1} \equiv H = 2h$ формула (3) приобретает вид

$$\int_{a_0}^{a_M} \approx H\left(\frac{1}{6}f(a_0) + \frac{4}{6}f(a_{1/2}) + \frac{1}{3}f(a_1) + \frac{4}{6}f(a_{3/2}) + \right.$$
$$\left. + \frac{1}{3}f(a_2) + \cdots + \frac{1}{3}f(a_{M-1}) + \frac{4}{6}f(a_{M-1/2}) + \frac{1}{6}f(a_M)\right),$$

где $a_j = a_0 + jH$. Для остаточного члена справедлива оценка

$$\left|R_M^4(f)\right| \leqslant \frac{A_4(a_M - a_0)H^4}{2880} = \frac{A_4(a_M - a_0)^5}{2880M^4} = \frac{A_4(a_M - a_0)}{180}h^4, \quad h = \frac{H}{2}.$$

Последняя запись оценки наиболее употребительна.

Мы получили формулы с порядком погрешности $O(N^{-m})$ по отношению к общему числу узлов интегрирования $N = Mn$ в предположении ограниченности $|f^{(m)}(x)|$. Заметим, что в случае формул трапеций и Симпсона общее число узлов N оказалось меньше, чем Mn, поскольку концы элементарных отрезков $[a_{q-1}, a_q]$ были узлами интегрирования и значения функции в этих концах использовались для вычисления интегралов по двум соседним элементарным отрезкам.

Задача 2. Пусть $\int_A^B \left|f^{(q)}(x)\right| dx < \infty$, $q \leqslant 2$; получить оценку погрешности формулы трапеций

$$\left|R_M^2(f)\right| \leqslant \gamma_q \left(\int_A^B \left|f^{(q)}(x)\right| dx\right) H^q, \tag{9}$$

где γ_q — абсолютная постоянная.

Задача 3. Пусть $\int_A^B \left|f^{(q)}(x)\right| dx < \infty$, $q \leqslant 4$; получить оценку погрешности формулы Симпсона

$$\left|R_M^4(f)\right| \leqslant \beta_q \left(\int_A^B \left|f^{(q)}(x)\right| dx\right) H^q,$$

где β_q — абсолютная постоянная.

Пусть, например, вычисляется

$$\int_0^1 (\sin x)^\alpha \, dx, \quad 0 < \alpha < 1.$$

Так как $((\sin x)^\alpha)'_x \to \infty$ при $x \to 0$, то мы не можем получить никаких оценок погрешности через $\max|f'(x)|$. В то же время функция $(\sin x)^\alpha$ монотонна на отрезке $[0, 1]$, поэтому

$$\int_0^1 \left|((\sin x)^\alpha)'\right| dx = (\sin x)^\alpha \Big|_0^1 = 1.$$

Следовательно, при использовании квадратуры (4), соответствующей $m = 2$ с постоянным шагом $a_q - a_{q-1} \equiv H$, согласно (9) имеем оценку погрешности $O(1/N)$.

В данном примере из оценки (6) малость погрешности не следовала; в то же время на основании (9) мы заключаем, что эта погрешность порядка $O(1/N)$.

Не следует думать, что в случае функций с малым числом ограниченных производных составные формулы численного интегрирования имеют лучший порядок сходимости по сравнению с формулами Гаусса. Предположим, что подынтегральная функция имеет q ограниченных производных. Тогда, применяя составную формулу (4), соответствующую $n = q$,

получим приближенное значение интеграла с погрешностью $O(N^{-q})$. С другой стороны, известно, что для такой функции $E_{2n-1}(f) = O(N^{-q})$. Поэтому из (5.4) следует оценка погрешности формулы Гаусса с N узлами

$$R_N(f) = O(N^{-q}).$$

Таким образом, порядок оценки в обоих случаях одинаков.

Обратим внимание еще на одно удобство использования формул Гаусса сразу по всему отрезку интегрирования. Не нужно оценивать число q_0 ограниченных производных подынтегральной функции и в соответствии с этим выбирать наиболее подходящую формулу численного интегрирования по отрезкам разбиения — при применении формул Гаусса порядок погрешности $O(N^{-q_0})$ обеспечивается автоматически. Конечно, не нужно думать, что формула, имеющая более высокий порядок скорости сходимости, при конкретном числе узлов всегда точнее формулы более низкого порядка скорости сходимости.

§ 9. О постановках задач оптимизации

Мы получили ряд формул численного интегрирования и могли бы получить еще большее количество таких формул. Возникает вопрос: можно ли получить лучшие по порядку оценки погрешности при тех же предположениях о подынтегральной функции или хотя бы улучшить константы в этих оценках? Изучение этого вопроса приводит к задаче построения квадратур с оптимальной оценкой погрешности, или, как говорят, оптимальных квадратур. В связи с этим возникает следующая проблема — чем больше способов решения задачи, тем труднее решиться выбрать какой-то из них, поскольку каждый способ может иметь свои положительные качества: простота программирования, малое время работы ЭВМ, малая загрузка памяти, простота оценки погрешности, применимость к широкому кругу задач. Таким образом, следует иметь в виду, что иногда излишняя информация о способах решения задач при большом их количестве может также и затормозить реальное решение задач. Поэтому необходима какая-то систематизация методов решения, их отбор. Естественно пытаться решить задачи наилучшим, оптимальным способом. Мы уже рассматривали некоторые модельные формулировки задач об оптимальных методах решения на примере вычисления значений функций; при этом возникают определенные математические постановки задач, требующих решения.

При рассмотрении каждой новой проблемы желательно получить ее наиболее подробное, наилучшее описание и затем решить возникшую задачу наилучшим образом. Однако обычно это не удается сделать и приходится довольствоваться меньшим: описать проблему наилучшим образом и решить ее удовлетворительно или описать проблему удовлетворительно и затем полностью решить возникшую задачу. При рассмотрении

проблемы оптимизации методов можно говорить о выборе между удовлетворительным решением проблемы оптимизации методов для классов задач, близких к реальным, или полным решением проблемы для эталонных математических классов, подобных рассматриваемому в следующем параграфе. Может вызвать недоумение высказывание об «удовлетворительном решении проблемы оптимизации методов» для каких-то классов — ведь задача оптимизации методов на классе сводится к вполне конкретной математической задаче, которую, по-видимому, можно решить окончательно, а не «удовлетворительно».

В принципе это высказывание верно, однако обычно полностью решить задачу в приемлемое для практики время не удается, так как время, необходимое для построения оптимального метода, обычно существенно превосходит время, в течение которого возникает новое, уточненное описание классов рассматриваемых задач. Также надо иметь в виду, что не всегда удается формализовать такое математическое описание класса реальных задач: имеется какое-то качественное представление о классе, имеются неплохие численные методы и интуитивно ясно, что в практическом плане оптимальность методов достигнута; в то же время нет даже четко формализованного описания класса решаемых задач. Например, реально требуется вычислить интегралы от кусочно-гладких функций, однако в течение долгого времени так и не удалось предложить описание множества таких функций, которое соответствовало бы реальной практике вычислений. Точно так же, например, при анализе физических моделей можно описать задачу с помощью сложной системы дифференциальных уравнений и решить эту систему с малой точностью или описать задачу более грубо с помощью простой системы и решить эту систему с большой точностью.

В разных случаях, в зависимости от конкретной обстановки, бывает целесообразен тот или иной подход к решению задачи. Представляется, что для прикладной науки часто более существенно не окончательное решение вариационных задач, а правильная их постановка. Обычно больший эффект дает удовлетворительное решение правильно поставленной в практическом смысле вариационной задачи, чем полное решение упрощенной задачи, не охватывающей все существенные характерные черты исходной проблемы.

Что такое оптимальный метод решения задачи?

Под оптимальным методом решения задачи можно понимать метод, требующий минимальной затраты машинного времени. Но это будет неправильная постановка, поскольку, в принципе, всякую задачу можно решить без применения машин, затратив на это очень большое количество времени.

Под оптимальным можно понимать метод, требующий минимальных затрат времени (или соответственно материальных затрат). При этом следует помнить, что время и затраты на поиск оптимального алгоритма уже входят в решение задачи.

Когда мы ставим вопрос о решении задачи оптимальным образом, мы не учитываем разных возможностей исследователей, решающих задачу. А ведь в зависимости от индивидуальных возможностей исследователя, наличия ЭВМ, библиотеки, лаборантов, сотрудников, обладающих опытом решения подобных задач, оптимальное решение, вообще говоря, будет различным.

Мы говорили об оптимальном решении каждой конкретной задачи. Такая постановка вопроса не совсем правильна также по следующей причине. На самом деле коллектив исследователей сталкивается с целой совокупностью задач. Мы можем поставить перед собой цель решить первую, вторую и т. д. задачи за кратчайшее время. Однако постановка вопроса в корне изменится, если мы зададимся целью решить с минимальными затратами целую совокупность задач. Мы можем решать первую, вторую и т. д. задачи оптимальными методами, но за это время наука уйдет вперед, и если мы не будем изучать новые методы, создавать новые алгоритмы и стандартные программы, рассчитанные на решение будущих задач, то в целом мы проиграем. При наличии большой серии задач, не требующей сверхсрочного решения, выгоднее заняться теоретическими исследованиями, создать новые алгоритмы, а затем уже решать эти задачи.

Нет однозначного ответа также на вопрос о том, кто должен заниматься оптимальным решением задач. Если мы имеем дело с единственной конкретной задачей, заведомо не требующей нового алгоритма и больших затрат машинного времени, ее лучше поручить выполнить работнику низкой квалификации. Увеличение числа задач и их сложности требует привлечения работников высокой квалификации, поскольку здесь их отдача будет наиболее полной. Конечно, важно изучить опыт решения как сложных, так и простых задач подобного типа. Для того чтобы задачу оптимизации методов можно было рассматривать как чисто математическую задачу, необходимо определить целевую функцию исследования, класс рассматриваемых задач и возможности исследователя.

Качество какого-то алгоритма на классе задач мы характеризуем его качеством на самой «плохой» для него задаче этого класса. Поэтому чем уже класс рассматриваемых задач, тем лучше качество алгоритмов на этом классе.

Казалось бы, следует построить оптимальные алгоритмы для решения возможно большего числа классов задач. Однако при излишней детализации слишком много усилий уйдет на отыскание оптимальных методов решения. В этом случае может не хватить сил и времени на реальное решение самих задач, разработку и создание стандартных программ. Конечно, эти замечания против детализации в меньшей мере относятся к случаю специализированных машин и устройств, где следует максимально сузить класс рассматриваемых задач.

При первом взгляде на проблему кажется, что в практике вычислений всегда имеют дело с конкретными задачами и никогда не рассматривают

класс задач. По этому поводу можно сказать следующее: формально при выборе метода решения задачи исследователь не относит ее к какому-либо классу; однако метод решения всегда выбирается в зависимости от некоторых типичных свойств задачи: тип дифференциального уравнения, наличие особенности у решения, число конечных производных и т. п. При выборе алгоритма исследователь учитывает эти свойства и тем самым вольно или невольно относит рассматриваемую задачу к классу задач, обладающих этими свойствами.

При анализе задач и разбиении их на классы часто возникает вопрос: какому классу задач уделить первоочередное внимание, в частности, на какой класс задач следует рассчитывать при составлении стандартной программы численного интегрирования? Пусть, например, при помощи этой программы примерно с одинаковой частотой будут вычисляться интегралы как от гладких, так и от не очень гладких функций. На какие функции нужно ориентироваться при выборе алгоритма для этой программы? Ответ на этот вопрос можно получить из следующих соображений. На гладких функциях алгоритм будет работать более эффективно, поэтому время, затрачиваемое на их решение, будет меньше времени, затрачиваемого на решение задач с негладкими функциями. 50% экономии времени при вычислении интегралов от гладких функций принесет меньшую выгоду, чем 50% экономии времени при вычислении интегралов от негладких функций. Следовательно, целесообразнее уделить внимание эффективности алгоритма в случае негладких функций. Конечно, вывод будет другим, если доля негладких функций будет мала.

При отыскании оптимального метода решения класса задач в случае, когда предполагается непосредственное внедрение алгоритма в вычислительную практику, обычно возникает следующая проблема: нам не удается сразу найти оптимальный метод решения задач рассматриваемого класса. После некоторого времени работы мы находим какой-то метод, иногда близкий к оптимальному, иногда просто лучший, чем ранее известные, а затем постепенно совершенствуем его. В какой момент надо остановиться в усовершенствовании метода и перейти к составлению стандартных программ решения этого класса задач? При ответе на этот вопрос надо учитывать следующие простые соображения. Если мы поспешим и быстро начнем внедрять только что полученный алгоритм, то, может быть, нам придется вскоре создавать новые алгоритмы из-за плохого качества этого алгоритма. Затягивание времени при составлении стандартных программ также нежелательно: при этом мы допустим большой перерасход машинного времени при решении задач по имеющимся стандартным программам.

Можно возразить, что создание совершенных стандартных программ приведет к большой экономии машинного времени в будущем. Однако при анализе этого возражения в свою очередь следует учесть, что машинное время становится все менее дефицитным и более дешевым. Кро-

ме того, чем раньше мы решим ту или иную задачу, тем бо́льшую реальную пользу получит общество в целом.

Нужно учитывать также важность быстрого внедрения вновь созданных алгоритмов для самой задачи отыскания оптимальных методов решения. Дело в том, что анализ результатов расчетов может указать на необходимость изменения класса рассматриваемых задач и открыть путь для новых теоретических исследований. Мы видим, что подход к проблеме оптимальности должен носить динамический характер: должны строиться оптимальные или близкие к ним методы для все новых классов задач, предъявляемых наукой и техникой, при изменении возможностей, предоставляемых ЭВМ.

При этом необходима как текущая работа, так и работа по доведению предложенных ранее постановок до окончательного решения. Для вычислительной математики, как и для всякой прикладной науки, характерна следующая обстановка. Обычно задачи новых типов предъявляются сначала в незначительном количестве и требуется срочное их решение любой ценой, не считаясь ни с какими затратами. На первом этапе применяется первый попавшийся приемлемый метод. Далее эти задачи поступают в большом количестве, производится более или менее удовлетворительная постановка задачи оптимизации методов и находится некоторое ее решение. Затем задача переводится на поток, т. е. решение задач этого типа производится при помощи пакетов соответствующих стандартных программ. Не следует думать, что на этом этапе полностью кончается исследование данного типа задач — чтобы создавать эффективные методы решения новых задач, нужно осмысливать те задачи, которые остались несколько позади, и проводить их теоретическое изучение. Иногда бывает целесообразно работу по быстрому решению первых поступивших задач серии и перспективную работу по созданию эффективных методов решения с самого начала организовать параллельно.

Соотношение между текущей и перспективной работой также является важнейшим фактором жизнедеятельности любой научной организации. В каждый момент времени организации предъявляются некоторые требования, выполнение которых необходимо для ее существования: самоокупаемость в случае хозрасчетных организаций, своевременная отчетность по годовому плану и т. п. Однако существуют и критические моменты времени, когда предъявляются повышенные требования к работе, например, необходимость своевременного решения новых классов задач. Поэтому при планировании работы должны предусматриваться какие-то теоретические разработки впрок, резервы людей, машинного времени.

Вернемся к вопросу об оптимизации методов.

Часто математик, создав оптимальный или близкий к нему метод решения задачи, сетует на то, что этот метод плохо внедряется в вычислительную практику. Ответ на этот вопрос может быть самым разным. Часто это происходит вследствие консерватизма практических работни-

ков, их желания работать старыми, привычными методами. Иногда это объясняется недостатками самого метода. Например, случается, что кроме (и даже вместо) оптимальности метода желательны и существенны простота метода и возможность надежного контроля точности получаемых результатов. Может случиться, что сам класс рассмотренных задач не совпадает с классом, к которому относится большинство задач, поступающих для решения.

Конечно, нужно учитывать также вопрос о том, насколько широк возможный круг решаемых задач рассматриваемого класса. Если сейчас и в перспективе ожидается решение небольшого числа задач рассматриваемого класса, разработка оптимальных алгоритмов и создание стандартных программ решения задач этого класса могут себя не оправдать. В то же время изучение задачи оптимизации методов на различных классах часто полезно тем, что при решении возникают новые методы, которые затем находят применение и при решении задач из других классов.

§ 10. Постановка задачи оптимизации квадратур

Рассуждения § 9 показывают важность изучения различных постановок проблемы оптимизации методов на классах функций.

Рассмотрим задачу вычисления интеграла

$$I(f) = \int_\Omega f(P)p(P)\,dP.$$

Область интегрирования Ω и весовая функция $p(P)$ предполагаются фиксированными. Класс рассматриваемых задач определим заданием класса F подынтегральных функций. *Погрешностью квадратуры*

$$I(f) \approx S_N(f) = \sum_{j=1}^{N} D_j f(P_j)$$

на классе F называют величину

$$R_N(F) = \sup_{f\in F} |R_N(f)|,$$

где, как обычно,

$$R_N(f) = I(f) - S_N(f).$$

Нижняя грань

$$W_N(F) = \inf_{D_j, P_j} R_N(F)$$

называется *оптимальной оценкой погрешности квадратур на рассматриваемом классе*. Если существует квадратура, для которой $R_N(F) = W_N(F)$, то такую квадратуру называют *оптимальной* или *наилучшей на рассматриваемом классе*.

В § 2 была получена оценка (2.4) погрешности квадратуры, точной для многочленов степени m, через $(m+1)$-ю производную функции. Эта оценка является неулучшаемой (см. задачу 2.3). Таким образом, для классов функций

$$F: \left|f^{(m+1)}(x)\right| \leqslant A \quad \text{при} \quad x \in [a, b]$$

величина $R_N(F)$ известна и задача построения оптимальной квадратуры сводится к нахождению коэффициентов и узлов, на которых достигается нижняя грань $R_N(F)$. Для ряда классов функций эту задачу удалось решить. Например, при $m = 0$ такой квадратурой является составная формула прямоугольников

$$\int_0^1 f(x)\,dx \approx \frac{1}{N}\sum_{j=1}^{N} f\left(\frac{j-1/2}{N}\right)$$

с оценкой погрешности

$$|R_N(f)| \leqslant A/(4N). \tag{1}$$

(Доказать!)

В настоящее время оптимальные квадратуры получены для небольшого набора классов функций, в основном одной переменной. Непосредственное значение этих квадратур для приложений невелико, однако это не значит, что не нужно заниматься построением таких квадратур.

Построение оптимальных квадратур и дальнейшее их развитие на случай большей гладкости и большего числа переменных оказались ценными не получением конкретных квадратурных формул, а выяснением качественной стороны вопроса: где какие методы лучше, на какую точность можно рассчитывать при использовании определенной информации о подынтегральной функции, какова плотность распределения узлов у «хорошей» квадратуры.

Пусть, например, при первоначальном анализе задачи мы решили воспользоваться информацией об ограниченности первой производной $|f'(x)| \leqslant A$; оценка погрешности (1) нас не устраивает, поскольку для достижения нужной точности ε требуется слишком большое число узлов: если $A/(4N) \leqslant \varepsilon$, то $N \geqslant A/(4\varepsilon)$. Оптимальность оценки (1) указывает на необходимость сужения класса рассматриваемых задач путем учета дополнительной информации о подынтегральной функции (ограниченность второй производной, тип особой точки подынтегральной функции, аналитичность и т. п.) или расширения множества используемых методов интегрирования.

§ 11. Оптимизация распределения узлов квадратурной формулы

Развитие методов численного интегрирования могло бы пойти по пути создания оптимальных методов на различных классах $C_r(A; [a, b])$ и создания программ на основе этих методов. Здесь и далее $C_r(A; [a, b])$ — класс функций с кусочно-непрерывной r-й производной, удовлетворяющей условию

$$\left|f^{(r)}(x)\right| \leqslant A \quad \text{при} \quad x \in [a, b].$$

Вскоре после начала рассмотрения задач оптимизации методов стало ясно, что уже известные на этих классах методы с оценкой погрешности (8.7) недалеки от оптимальных. Как увидим в § 7 гл. 5, за счет оптимизации квадратурной формулы оценка погрешности (8.7) на рассматриваемых классах не может быть улучшена по порядку. При этом также стало ясно, что некоторые методы, практически совпадающие с методами Эйлера и Грегори (см. § 13), являются асимптотически оптимальными по оценке главного члена погрешности. Имеется в виду следующее. Для этих методов на соответствующих классах функций были получены оценки погрешности

$$c_r/N^r + c_{r+1}/N^{r+1},$$

в то время как для оптимальных на этих классах методов оценка погрешности имеет вид

$$c_r/N^r + O(1/N^{r+1}).$$

Казалось бы, что поскольку есть почти оптимальные методы, то следующим этапом должен стать перевод всех программных комплексов интегрирования на использование этих методов.

Однако подобное утверждение нельзя рассматривать как бесспорное. Вследствие большого многообразия задач, требующих решения, при переходе к этапу внедрения методов в практику всегда следует проявлять известную осторожность. Нельзя полностью ручаться, что принятое нами описание классов этих задач наилучшим образом соответствует классам реальных задач. Например, следует признать, что классы C_r плохо описывают реальные задачи. Конечно, из-за привычки к традиционным методам внедрение новых методов обычно требует энергичных действий со стороны их приверженцев. В то же время надо иметь в виду, что старые методы прошли испытание временем и могут оказаться более пригодными для решения задач некоторых классов, поэтому отбрасывание старых методов должно производиться лишь после достаточного теоретического и практического анализа.

В случае рассматриваемой проблемы оптимизации методов интегрирования на практике еще до полного выяснения вопроса об оптимальности методов пришли к следующим заключениям.

При разбиении исходного отрезка интегрирования на одинаковые элементарные отрезки $a_q - a_{q-1} \equiv H$ мы получаем информацию о подынтегральной функции равномерно по всему отрезку интегрирования. В то же время подынтегральная функция может быть более гладкой на части отрезка интегрирования, поэтому там следует поместить относительно небольшое количество узлов, т. е. для практически оптимальных методов разбиение отрезка интегрирования на части должно быть приспособлено к специфике поведения подынтегральной функции.

Рассмотрим возможные постановки задач распределения узлов в зависимости от особенностей поведения производной подынтегральной функции. Эти постановки имеют много общего, однако для конкретных задач тот или иной подход иногда оказывается более удобным. Для простоты изложения мы будем проводить рассмотрение на примере формулы трапеций.

Пусть вычисляется интеграл

$$I(f) = \int_0^1 f(x)\,dx$$

и подынтегральная функция удовлетворяет условиям $|f''(x)| \leqslant A_l$ на отрезках $[B_{l-1}, B_l]$, $l = 1, \ldots, q$, где $0 = B_0 < B_1 < \cdots < B_q = 1$. Для вычисления интегралов

$$I_l(f) = \int_{B_{l-1}}^{B_l} f(x)\,dx$$

применим составную формулу трапеций с равными отрезками разбиения длины $H_l = b_l/N_l$, где $b_l = B_l - B_{l-1}$:

$$I_l(f) \approx S_l(f) = H_l \left(\frac{f(B_{l-1})}{2} + f(B_{l-1} + H_l) + \cdots + \frac{f(B_l)}{2} \right).$$

Из результатов § 3 следует, что остаточный член оценивается величиной $A_l b_l^3/(12N_l^2)$. Тогда суммарная погрешность при замене $I(f)$ суммой $\sum_{l=1}^{q} S_l(f)$ не превзойдет величины

$$\Phi = \sum_{l=1}^{q} \frac{A_l b_l^3}{12N_l^2}.$$

Поставим задачу: при заданном числе $N = N_1 + \cdots + N_q$ отрезков разбиения распорядиться выбором величин N_l так, чтобы суммарная оценка погрешности Φ была минимальной.

Найдем минимум Φ при условии $\Psi = N_1 + \cdots + N_q - N = 0$, не предполагая пока, что величины N_l целые. Приравнивая нулю производные функции Лагранжа $\Phi + \lambda\Psi$, получаем систему уравнений

$$\frac{\partial(\Phi + \lambda\Psi)}{\partial N_l} = -\frac{A_l b_l^3}{6N_l^3} + \lambda = 0.$$

Отсюда

$$N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}. \tag{1}$$

Из условия $\Psi = 0$ получим следующее уравнение относительно λ:

$$\sum_{l=1}^{q} b_l \sqrt[3]{\frac{A_l}{6\lambda}} = N.$$

Из этого уравнения определяем λ и затем из (1) находим N_l. Поскольку N_l должны быть целыми, то возьмем, например,

$$N_l = N_l^0 = \left[b_l \sqrt[3]{\frac{A_l}{6\lambda}} \right]. \tag{2}$$

Очевидно, что

$$N - q \leqslant N_1^0 + \cdots + N_q^0 \leqslant N.$$

Мы не нашли настоящего минимума Φ по множеству всевозможных целых $N_1, \ldots, N_q$, удовлетворяющих условию $N_1 + \cdots + N_q = N$, однако дальнейшее уточнение вряд ли разумно.

Обычно практический интерес представляет другая вариационная задача: найти минимальное значение $N = N_1 + \cdots + N_q$, при котором оценка погрешности Φ не превосходит заданного ε. Поскольку Φ монотонно убывает с ростом величины N_l, достаточно ограничиться случаем $\Phi = \varepsilon$. Возьмем функцию Лагранжа в виде

$$N_1 + \cdots + N_q + \lambda^{-1}(\Phi - \varepsilon).$$

Приравнивая нулю ее производные по N_l, $l = 1, 2, \ldots, q$, получим те же уравнения (1); подставляя значение $N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}$ в уравнение $\Phi = \varepsilon$, получим

$$\lambda^{2/3} \sum_{l=1}^{q} \left(\frac{A_l}{48} \right)^{1/3} b_l = \varepsilon.$$

Определяем λ, а затем соответствующие N_l. Соотношения для определения величин N_l, получившиеся при рассмотрении этих двух задач, имеют одинаковый вид. Поэтому для вывода качественных результатов об оптимальном распределении узлов достаточно было бы ограничиться рассмотрением одной из этих задач.

Нашей целью является разработка оптимальных методов решения и разработка на их основе систем программ решения типовых математических задач. Можно представить себе программу вычисления интеграла с заданной точностью, работающую по следующей схеме. Производится вычисление таблицы значений функции на некоторой сетке

$x_1, \ldots, x_n$. По этой таблице составляется таблица разделенных разностей $f(x_0; x_1; x_2), \ldots, f(x_{n-2}; x_{n-1}; x_n)$. Из рассмотрения этой таблицы делается вывод о наиболее целесообразном разбиении отрезка на части $[B_{l-1}, B_l]$ и значениях A_l, соответствующих этим частям. Затем, в соответствии с (1), выбираются N_l и производится интегрирование.

Большинство алгоритмов реально работающих стандартных программ базируется не на таком непосредственном использовании полученных соотношений, а на одном качественном выводе, являющемся следствием (1). Для этого перепишем равенство (1) в виде

$$\frac{1}{12} A_l (b_l/N_l)^3 = \frac{\lambda}{2}.$$

Левая часть этого выражения равна оценке погрешности по элементарному отрезку интегрирования длины b_l/N_l, на которые разбит отрезок $[B_{l-1}, B_l]$. Таким образом, *это соотношение означает, что при оптимальном распределении узлов интегрирования оценки погрешностей, приходящиеся на элементарные отрезки интегрирования, должны быть одинаковыми.*

Для получения этого вывода достаточно было ограничиться случаем $q = 2$. Это обстоятельство подчеркивает общее свойство качественных характеристик методов решения задач (не обязательно математических): *для их получения достаточно ограничиться рассмотрением простейших моделей, учитывающих основные стороны явления.*

Как правило, алгоритмы, основанные на качественных выводах о свойствах решения оптимизационной задачи, имеют *более широкую область применения*, чем алгоритмы, подобные вышеописанному, основывающиеся на количественных соотношениях. Описываемые далее программы вычисления интегралов, основывающиеся на этом качественном выводе, позволяют вычислять с высокой скоростью сходимости интегралы от функций с регулярными особенностями типа x^α, $\alpha > -1$.

Рассмотрим еще одну, близкую постановку задачи оптимизации распределения узлов интегрирования. Чтобы не утомлять читателя второстепенными деталями, мы не будем проводить подробных оценок членов высшего порядка в оценке погрешности.

Пусть отрезок интегрирования $[0, 1]$ разбит на части $[a_{q-1}, a_q]$, $q = 1, \ldots, N$, $a_0 = 0$, $a_N = 1$, и интеграл по каждой части вычисляется по формуле трапеций

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x)\,dx \approx s_q(f) = \frac{a_q - a_{q-1}}{2}\bigl(f(a_{q-1}) + f(a_q)\bigr).$$

Тогда интеграл по всему отрезку $[0, 1]$ вычисляется по формуле

$$I(f) = \sum_{i=1}^{q} I_q(f) \approx \sum_{q=1}^{N} s_q(f)$$

с оценкой остаточного члена

$$r = \sum_{q=1}^{N} \max_{[a_{q-1},\, a_q]} |f''(x)| \, \frac{(a_q - a_{q-1})^3}{12}. \tag{3}$$

Пусть известно, что $|f''(x)| \leqslant F(x)$ на $[0, 1]$, где $F(x)$ непрерывна, и пусть в качестве a_q взяты значения $\varphi(q/N)$ непрерывно дифференцируемой функции φ, удовлетворяющей условиям $\varphi(0) = 0$, $\varphi(1) = 1$. Поскольку

$$a_q - a_{q-1} = \varphi\left(\frac{q}{N}\right) - \varphi\left(\frac{q-1}{N}\right) = \varphi'\left(\frac{q}{N}\right)\frac{1}{N} + o\left(\frac{1}{N}\right) \quad \text{при} \quad N \to \infty,$$

то

$$\max_{[a_{q-1},\, a_q]} |f''(x)| \leqslant \max_{[a_{q-1},\, a_q]} F(x) = F(a_q) + o(1) = F\left(\varphi\left(\frac{q}{N}\right)\right) + o(1).$$

Из этих соотношений получаем

$$\max_{[a_{q-1},\, a_q]} |f''(x)| \, \frac{(a_q - a_{q-1})^3}{12} \leqslant \varepsilon_q = \left(\varphi'\left(\frac{q}{N}\right)\right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12N^3} + o\left(\frac{1}{N^3}\right).$$

Подставляя последние соотношения в (3), имеем

$$r \leqslant \bar{r} = \frac{1}{N^2}\left\{\sum_{q=1}^{N} \frac{1}{N}\left(\varphi'\left(\frac{q}{N}\right)\right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12}\right\} + o\left(\frac{1}{N^2}\right).$$

Выражение в фигурных скобках является квадратурной суммой Римана для интеграла

$$\int_0^1 \left(\varphi'(t)\right)^3 \frac{F(\varphi(t))}{12}\, dt$$

от непрерывной функции. Следовательно,

$$\bar{r} = \frac{1}{N^2}\left(\int_0^1 \left(\varphi'(t)\right)^3 \frac{F(\varphi(t))}{12}\, dt\right) + o\left(\frac{1}{N^2}\right). \tag{4}$$

Рассмотрим задачу минимизации первого, главного члена выражения (4). Для удобства решения уравнения Эйлера примем за независимую переменную функцию φ. Тогда коэффициент при $1/N^2$ в главном члене погрешности запишется в виде

$$\int_0^1 \left(t'(\varphi)\right)^{-2} \frac{F(\varphi)}{12}\, d\varphi.$$

Уравнение Эйлера для функции, минимизирующей функционал

$$\int_0^1 G(\varphi, t, t')\, d\varphi,$$

имеет вид

$$\frac{d}{d\varphi}\left(\frac{\partial G}{\partial t'}\right) - \frac{\partial G}{\partial t} = 0. \tag{5}$$

В рассматриваемом случае $\partial G/\partial t = 0$, поэтому из (5) имеем $\partial G/\partial t' = \text{const}$. Подставляя конкретное значение функции G, получим

$$(t'(\varphi))^{-3}\frac{F(\varphi)}{6} = \text{const}$$

или

$$F(\varphi)(\varphi'(t))^3 = C_1. \tag{6}$$

Общее решение этого уравнения зависит от C_1 и еще от некоторой постоянной C_2. Значения этих постоянных можно определить из граничных условий $\varphi(0) = 0$, $\varphi(1) = 1$. Решение рассмотренной вариационной постановки может практически использоваться различными способами. Например, в случае гладких функций $f(x)$ программы осуществляют численное интегрирование (6) на сетке с шагом, существенно большим $1/N$, и затем распределяют узлы в соответствии с полученным решением.

Из соотношения (6) можно сделать тот же вывод о равенстве оценок погрешностей на элементарных отрезках интегрирования при оптимальном распределении узлов. В самом деле, умножим (6) на $\frac{1}{12N^3}$, положим $t = \frac{q}{N}$ и заменим $F\left(\varphi\left(\frac{q}{N}\right)\right)$ и $\frac{1}{N}\varphi'\left(\frac{q}{N}\right)$ соответственно эквивалентными величинами $\max\limits_{[a_{q-1},\,a_q]} F(x)$, $a_q - a_{q-1}$. В результате получим

$$\max_{[a_{q-1},\,a_q]} F(x)\frac{(a_q - a_{q-1})^3}{12} \approx \frac{C_1}{12N^3}. \tag{7}$$

Другой из возможных путей практического использования решения уравнения (6) состоит в следующем. Пусть требуется вычислить большую серию интегралов с одинаковым характерным поведением подынтегральных функций. Выделим простейшую модельную функцию, для которой задача оптимизации узлов может быть решена в явном виде, и далее будем производить интегрирование с распределением узлов, соответствующим этой функции. Если характер изменения функций из рассматриваемой серии зависит от некоторого параметра, то этот параметр следует учесть при выборе модельной функции; естественно, что модельная функция не обязательно относится к рассматриваемому классу. Чем большее количество задач предъявляется для решения, тем более оправданными могут быть затраты, связанные с удачным выбором и рассмотрением модельной задачи.

§ 12. Примеры оптимизации распределения узлов

Рассмотрим примеры решения уравнения (11.6) для конкретных задач.

Пример 1. Пусть вычисляется серия интегралов

$$\int_0^1 f(b,\, x)\, dx,$$

где b — параметр серии, $f(b,\, x) = x^b g(b,\, x)$, $-1 < b < 2$, $g(b,\, x)$ — гладкая функция, $g(b,\, 0) \neq 0$. Если $b \neq 0,\, 1$, то вторая производная $f_{xx}(b,\, x)$ не ограничена в окрестности точки 0, поэтому при выборе модельной задачи следует учесть эту специфику поведения подынтегральной функции. В окрестности точки $x = 0$ мы имеем

$$f_{xx}(b,\, x) = b(b-1)x^{b-2}g(b,\, 0) + O(x^{b-1}).$$

Таким образом, в окрестности точки $x = 0$ вторая производная f_{xx} приблизительно пропорциональна второй производной функции $y = x^b$, поэтому функцию $y = x^b$ естественно рассматривать в качестве модельной. Примем за $F(x)$ величину $|b(b-1)|\, x^{b-2}$; тогда уравнение (11.6) запишется в виде

$$|b(b-1)|\, \varphi^{b-2} \left(\frac{d\varphi}{dt}\right)^3 = C_1,$$

отсюда

$$\left(\frac{3\sqrt[3]{|b(b-1)|}}{b+1}\right) \varphi^{\frac{b+1}{3}} = C_1 t + C_2.$$

Из условия $\varphi(0) = 0$ получаем, что $C_2 = 0$, а из условия $\varphi(1) = 1$ — что $\varphi^{\frac{b+1}{3}} = t$. Таким образом,

$$\varphi(t) = t^{\frac{3}{1+b}} \tag{1}$$

и для модельной задачи вычисления интеграла $\int_0^1 x^b\, dx$ оптимальным в рассматриваемом нами смысле является распределение узлов $a_q = \left(\frac{q}{N}\right)^{\frac{3}{1+b}}$.

Проведенные выше построения, вообще говоря, неприменимы к рассматриваемому случаю, поскольку при получении оценки (11.4) предполагалась ограниченность второй производной функции $F(x)$, не имеющая места для данной задачи. Однако можно обосновать применимость оценки (11.4) и в рассматриваемом случае.

Задача 1. Пусть для функции

$$f_b(x) = \begin{cases} 0 & \text{при} \quad x = 0, \\ x^b & \text{при} \quad x \in (0,\, 1], \end{cases}$$

где $-1 < b < 1$, по формуле трапеций с постоянным шагом $a_q - a_{q-1} = 1/N$ вычисляется

$$\int_0^1 f_b(x)\,dx. \tag{2}$$

Доказать, что суммарная погрешность удовлетворяет соотношению $R_N(f) \sim D_1(b)/\ N^{1+b}$, где $D_1(b) \neq 0$.

Задача 2. Интеграл (2) вычисляется по формуле трапеций с распределением узлов $a_q = \varphi(q/N)$, $\varphi(t) = t^{\frac{3}{1+b}}$, определяемым (1). Доказать, что суммарная погрешность удовлетворяет соотношению $R_N(f) \sim D_2(b)/N^2$.

Задача 3. Интеграл (2) вычисляется по формуле трапеций с распределением узлов $a_q = \varphi(q/N)$, $\varphi(t) = t^a$. Показать, что при $a > 2(b+1)$ суммарная погрешность $R_N(f) \sim D(a,\, b)/N^2$. Проверить, что $D(a,\, b) > D_2(b)$.

Сравнение результатов решения этих задач показывает, что перераспределение узлов в сторону большей их концентрации вблизи особенности, в частности оптимизация распределения узлов, приводит к увеличению порядка скорости сходимости.

Пример 2. Вычисляется серия интегралов

$$\int_0^1 x^b (\ln x) g(b,\, x)\,dx,$$

где $g(b,\, x)$ — гладкая функция, $g(b,\, 0) \neq 0$, b — параметр серии. Поскольку $\ln x$ имеет особенность в точке 0, то кажется естественным взять в качестве модельной функции $y = x^b \ln x$. Ее вторая производная имеет вид

$$y'' = x^{b-2}(b(b-1)\ln x + (2b-1)).$$

При $F(x) = \left|(x^b \ln x)''\right|$ уравнение (11.6) не решается в квадратурах, поэтому упростим задачу. При $x \to 0$ функция $\ln x$ растет медленнее, чем любая степенная функция $y = x^{-\varepsilon}$, $\varepsilon > 0$. Исходя из этого в уравнении (11.6) возьмем $F(x) = \mathrm{const} \cdot x^{b-2}$.

Пример 3. Вычисляется серия интегралов

$$\int_0^1 \exp\left\{-\frac{x}{b}\right\} g(b,\, x)\,dx, \tag{3}$$

где $g(b,\,x)$ — гладкая функция, $g(b,\,0) \neq 0$, b — параметр серии, который может принимать очень малые значения. При малых b подынтегральная функция и ее производные резко меняются в окрестности точки $x = 0$ за счет множителя $\exp\{-x/b\}$, поэтому имеет смысл произвести оптимизацию распределения узлов интегрирования на модельной задаче вычисления интеграла $\int_0^1 \exp\{-x/b\}\,dx$. Положим $F(x) = |(\exp\{-x/b\})''|$. Уравнение (11.6) приобретает вид

$$\frac{1}{b^2}\exp\left\{-\frac{\varphi}{b}\right\}\left(\frac{d\varphi}{dt}\right)^3 = C_1.$$

Отсюда

$$1 - \exp\left\{-\frac{\varphi}{3b}\right\} = C_3 t + C_4.$$

Из условия $\varphi(0) = 0$ следует, что $C_4 = 0$, а из условия $\varphi(1) = 1$ получаем

$$1 - \exp\left\{-\frac{\varphi}{3b}\right\} = \left(1 - \exp\left\{-\frac{1}{3b}\right\}\right) t,$$

откуда

$$\varphi(t) = 3b\,\ln\left[1 - \left(1 - \exp\left\{-\frac{1}{3b}\right\}\right) t\right]^{-1}.$$

Пример 4. Вычисляется серия интегралов

$$\int_0^1 \exp\{-x^2/b^2\} g(b,\,x)\,dx, \tag{4}$$

где $g(b,\,x)$ — гладкая функция, $g(b,\,0) \neq 0$, b — параметр серии, который может принимать очень малые значения. При малых b подынтегральная функция и ее производные резко меняются в окрестности точки $x = 0$ за счет множителя $\exp\{-x^2/b^2\}$. Поэтому в качестве модельной задачи возьмем задачу вычисления интеграла $\int_0^1 \exp\{-x^2/b^2\}\,dx$. Положим $F(x) = \left|(\exp\{-x^2/b^2\})''\right|$; тогда в качестве уравнения (11.6) получим уравнение

$$|1 - 2\varphi^2/b^2|\exp\{-\varphi^2/b^2\}(d\varphi/dt)^3 = C_1 b^2/2,$$

откуда

$$\int \sqrt[3]{|1 - 2\varphi^2/b^2|}\exp\{-\varphi^2/3b^2\}\,d\varphi = \int \sqrt[3]{C_1 b^2}/\sqrt[3]{2}\,dt.$$

Этот интеграл не вычисляется в явном виде, поэтому попытаемся произвести упрощения. Например, можно заменить $\sqrt[3]{|1 - 2\varphi^2/b^2|}$ на 1. При больших значениях φ/b, когда погрешность такой замены

большая, ее влияние не столь значительно из-за малого множителя $\exp\{-\varphi^2/(3b^2)\}$. После такого упрощения функция $\varphi(t)$ будет выражаться через функцию, обратную функции $\int_0^{\varphi} \exp\{-v^2\}\,dv$.

Задачи, подобные рассмотренным в примерах 3, 4, возникают довольно часто. Например, при расчетах диаграмм направленности антенн вычисляются серии интегралов $\int_0^1 \exp\{ibg(b,x)\}h(b,x)\,dx$ в широком диапазоне изменения b; функции $g(b,x)$, $h(b,x)$ являются довольно гладкими. При b не очень больших эти интегралы могут вычисляться с помощью простейших квадратурных формул. С ростом b производные подынтегральной функции растут, поэтому требуемое количество узлов интегрирования увеличивается. При очень больших b можно воспользоваться методом перевала или иными асимптотическими методами. Однако для «промежуточных» значений b оба эти метода будут плохи: первый — из-за трудоемкости, второй — из-за малой точности. Поэтому иногда применяют следующий метод: контур интегрирования преобразуется так, чтобы он проходил по линиям наискорейшего спуска функции $\exp\{ibg(b,x)\}$, как это делается при использовании метода перевала. Получаются интегралы от резко меняющихся функций, аналогичные рассмотренным в примерах 3, 4.

Из приведенных примеров видно, что оптимизация распределения узлов интегрирования на основе уравнения (11.6) требует достаточно высокой квалификации исследователя. Поэтому далее в § 17 будет рассмотрен вопрос о передаче этих функций ЭВМ.

§ 13. Главный член погрешности

Применение формул для оценок погрешности, подобных полученным в § 2, 3, требует достаточно высокой квалификации исследователя, например для получения требуемых оценок производных. При получении ряда из этих оценок, например оценок для составных формул трапеции и Симпсона, возможно существенное загрубление оценки, поскольку общая оценка погрешности равна сумме модулей оценок на отдельных отрезках.

Эти обстоятельства определили интерес к получению выражения для главного члена погрешности. По информации о величине главного члена погрешности можно полноценнее проводить сравнение методов.

Как будет видно далее, сам факт наличия главного члена у погрешности позволяет судить о реальной величине погрешности, не прибегая к теоретическим оценкам. Обратимся к составной квадратурной формуле трапеций вычисления интеграла $I(f) = \int_A^B f(x)\,dx$ с постоянным шагом

H. Для удобства обозначим $H = (B - A)/M$, $a_q = A + qH$, в частности $a_0 = A$, $a_M = B$. Имеем

$$I(f) \approx S_M(f) \equiv S_M^2(f) = H\left(\frac{f(a_0)}{2} + f(a_1) + \cdots + f(a_{M-1}) + \frac{f(a_M)}{2}\right).$$

Согласно § 3.2 справедливо равенство

$$\int_{a_{q-1}}^{a_q} f(x)\,dx = H\frac{f(a_{q-1}) + f(a_q)}{2} - \frac{f''(\zeta_q)H^3}{12}, \quad \zeta_q \in [a_{q-1}, a_q].$$

Просуммировав по q, получим

$$\int_{a_0}^{a_M} f(x)\,dx = S_M(f) + R(f), \qquad R(f) = -\sum_{q=1}^{M} f''(\zeta_q)\frac{H^3}{12}.$$

Величину погрешности $R(f)$ можно записать в виде

$$R(f) = -\frac{H^2}{12}i(f), \quad i(f) = \sum_{q=1}^{M} Hf''(\zeta_q).$$

Выражение в правой части есть квадратурная формула для интеграла $\int_{a_0}^{a_M} f''(x)\,dx$, поэтому при $H \to 0$ имеем

$$i(f) \to \int_{a_0}^{a_M} f''(x)\,dx.$$

Следовательно,

$$R(f) = -\frac{H^2}{12}\int_{a_0}^{a_M} f''(x)\,dx + R_1(f), \quad R_1(f) = o(H^2).$$

Задача 1. Пусть $|f^{(3)}(x)| \leqslant M_3$ на $[A, B]$. Показать, что в этом случае $|R_1(f)| \leqslant c_3 M_3 (B - A)H^3$.

Задача 2. Пусть $|f^{(4)}(x)| \leqslant M_4$ на отрезке $[A, B]$. Показать, что $|R_1(f)| \leqslant c_4 M_4 (B - A)H^4$.

Полученное соотношение для $R(f)$ может использоваться в различных целях. Например, его можно представить в виде

$$R(f) = -\frac{H^2}{12}(f'(B) - f'(A)) + o(H^2). \tag{1}$$

После вычисления $f'(B) - f'(A)$ получаем значение главного члена погрешности. Предположим, что достигнутая точность не является удовлетворительной. Запишем (1) в виде

$$I(f) = S_M^4(f) + o(H^2),$$

где

$$S_M^4(f) = S_M^2(f) - \frac{H^2}{12}(f'(B) - f'(A)).$$

Как следует из решения задачи 2, при $\left|f^{(4)}(x)\right| \leqslant M_4$ выражение $S_M^4(f)$ оказывается квадратурной суммой с погрешностью $O\big((B-A)H^4\big)$, т. е. такой же по порядку, как у формулы Симпсона.

Можно попытаться выделить главный член погрешности получившейся формулы. Имеем равенства

$$S_M^4(f) = \sum_{q=1}^{M} s_q^2(f),$$

$$s_q^2(f) = \frac{H}{2}(f(a_{q-1}) + f(a_q)) - \frac{H^2}{12}(f'(a_q) - f'(a_{q-1})).$$

Величину $s_q^2(f)$ будем рассматривать как приближенное значение интеграла

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x)\,dx.$$

Подставляя в разность $I_q(f) - s_q^2(f)$ представление $f(x)$ в виде отрезка ряда Тейлора, можно получить главный член погрешности на элементарном отрезке в виде $\dfrac{H^5}{720} f^{(4)}(\zeta_q^4)$ и т. д.

Продолжая процесс выделения главного члена погрешности, приходим к последовательности *квадратурных формул Эйлера*

$$I(f) \approx S_M^{2l}(f), \quad S_M^{2l}(f) = S_M^2(f) - \sum_{j=1}^{l-1} \gamma_{2j} H^{2j} \Big(f^{(2j-1)}(B) - f^{(2j-1)}(A)\Big)$$

с оценкой погрешности

$$I(f) - S_M^{2l}(f) = -\gamma_{2l} f^{(2l)}(\zeta^{2l})(B-A)H^{2l}. \tag{2}$$

Существует следующее соотношение, которому удовлетворяют числа γ_j:

$$\frac{x}{e^x - 1} = \sum_{j=0}^{\infty} \gamma_j x^j.$$

Обычно принято записывать числа γ_j в виде $B_j/j!$, где B_j — так называемые *числа Бернулли.*

Для сведения приведем несколько значений чисел γ_j:

$$\gamma_2 = \frac{1}{12}, \quad \gamma_4 = -\frac{1}{720}, \quad \gamma_6 = \frac{1}{30\,240}, \quad \gamma_8 = -\frac{1}{1\,209\,600},$$

$$\gamma_{10} = \frac{1}{47\,900\,160}.$$

Использование формул Эйлера неудобно, поскольку необходимо вычислять не только значения функции, но и значения ее производных.

Однако, если в выражении $S_M^{2p}(f)$ заменить производные $f^{(2k-1)}(A)$ и $f^{(2k-1)}(B)$ производными интерполяционных многочленов степени l соответственно с узлами $a_0, \dots, a_l$ и $a_N, \dots, a_{N-l}$, то при $l = 2p-3$ и $l = 2p-2$ после проведения промежуточных преобразований получаются *формулы численного интегрирования Грегори*

$$I(f) \approx G_M^l(f),$$

$$G_M^l(f) = S_M^2(f) - H\sum_{k=1}^{l} \beta_k \bigl(\nabla^k f(a_M) - (-1)^k \Delta^k f(a_0)\bigr),$$

где

$$\beta_k = (-1)^{k+1} \int_0^1 \frac{x(x-1)\dots(x-k)}{(k+1)!}\, dx.$$

В частности,

$$\beta_1 = \frac{1}{12}, \quad \beta_2 = \frac{1}{24}, \quad \beta_3 = \frac{19}{720}, \quad \beta_4 = \frac{3}{160}, \quad \beta_5 = \frac{863}{60\,480}, \quad \beta_6 = \frac{275}{24\,192}.$$

В случае подынтегральных функций с нерегулярным характером поведения, типа рассмотренных в § 11, применение формул Эйлера и Грегори неэффективно, поскольку производные высших порядков или не ограничены, или очень велики. Поэтому при непосредственном вычислении определенных интегралов эти формулы в настоящее время применяются редко. Однако они используются при интегрировании функций, заданных таблично, при вычислении неопределенных интегралов, при решении интегральных уравнений Вольтерра и других задачах, где существенно, чтобы значения подынтегральной функции вычислялись именно на равномерной сетке.

Задача 3. Доказать, что главный член погрешности квадратуры Грегори $I(f) \approx G_M^l(f)$ есть $\beta_{l+1} H^{l+2} (f^{(l+1)}(B) - (-1)^{l+1} f^{(l+1)}(A))$.

Задача 4. Пусть $\int_0^1 f(x)\, dx$ вычисляется по составной формуле трапеций с переменным шагом интегрирования: $a_q = \varphi(q/N)$, где φ — гладкая функция. Доказать, что главный член погрешности есть

$$-\frac{1}{12N^2} \int_0^1 f''(\varphi(t))(\varphi'(t))^3\, dt.$$

Указание. См. построения § 11.

§ 14. Правило Рунге практической оценки погрешности

Мы получили, что главный член погрешности формулы трапеций с постоянным шагом интегрирования равен

$$-\frac{1}{12}H^2(f'(B)-f'(A)).$$

В случае формул более высокого порядка точности можно получить представление главного члена погрешности квадратуры через производные высших порядков. Непосредственное использование этих выражений для оценки величины главного члена погрешности иногда неудобно, поскольку требует выполнения операции дифференцирования. В других задачах выражение главного члена погрешности может оказаться настолько сложным, что его вычисление требует дополнительного численного интегрирования. Поэтому в вычислительной практике применяется способ практической оценки погрешности, не использующий фактического выражения главного члена погрешности, а опирающийся лишь на факт существования такого главного члена. Для простейших задач типа численного интегрирования этот способ связывается с именем Рунге, в более сложных случаях — с именами Ричардсона и Филиппова. Этот способ основан на выделении главного члена погрешности по результатам расчетов с двумя различными шагами.

Рассмотрим простейший вариант применения этого правила. Осуществим приближенное вычисление интеграла $I(f)=\int_A^B f(x)\,dx$ с помощью формулы трапеций с постоянным шагом $H_1=(B-A)/M_1$ и $H_2=(B-A)/M_2$; $M_2=2M_1$, т. е. $H_2=H_1/2$. Согласно (13.1) имеем равенство

$$\begin{aligned} I(f)-S_{M_1}(f) &= -\frac{H_1^2}{12}\Big(f'(B)-f'(A)\Big)+o\big(H_1^2\big),\\ I(f)-S_{M_2}(f) &= -\frac{H_2^2}{12}\Big(f'(B)-f'(A)\Big)+o\big(H_2^2\big). \end{aligned} \tag{1}$$

Мы стремимся построить алгоритм вычисления главного члена погрешности, не использующий его конкретного выражения. Для этого запишем (1) в виде совокупности приближенных равенств

$$\begin{aligned} I(f)-S_{M_1}(f) &\approx CH_1^2,\\ I(f)-S_{M_2}(f) &\approx CH_2^2. \end{aligned} \tag{2}$$

Величины $S_{M_1}(f)$ и $S_{M_2}(f)$ определены в результате расчетов, поэтому мы имеем два приближенных равенства относительно двух неизвестных $I(f)$ и C. Вычитая второе равенство из первого, получим

$$S_{M_2}(f)-S_{M_1}(f)\approx CH_1^2-CH_2^2=3CH_2^2.$$

Таким образом,

$$CH_2^2 = \frac{1}{3}\Big(S_{M_2}(f) - S_{M_1}(f)\Big). \tag{3}$$

Подставляя приближенное выражение CH_2^2 в (2), получаем приближенное равенство

$$I(f) - S_{M_2}(f) \approx \frac{1}{3}\Big(S_{M_2}(f) - S_{M_1}(f)\Big). \tag{4}$$

Таким образом, величина $\frac{1}{3}\Big(S_{M_2}(f) - S_{M_1}(f)\Big)$ является главным членом погрешности приближенного значения интеграла $S_{M_2}(f)$. Перенося в (4) значение $S_{M_2}(f)$ в правую часть, получим формулу для более точного по порядку, чем $S_{M_2}(f)$, приближения к $I(f)$:

$$I(f) \approx S_{M_2}(f) + \frac{1}{3}\Big(S_{M_2}(f) - S_{M_1}(f)\Big). \tag{5}$$

Таким образом, описанный способ построения главного члена погрешности порождает некоторую квадратурную формулу более высокого порядка точности.

Задача 1. Доказать, что правая часть в (5) совпадает с составной квадратурной формулой Симпсона.

Информация о величине главного члена погрешности часто используется для приближенного определения минимального количества узлов, достаточного для достижения заданной точности. Из (3) находим, что

$$C \approx \frac{1}{3H_2^2}\Big(S_{M_2}(f) - S_{M_1}(f)\Big),$$

а затем выбираем шаг интегрирования из условия

$$\big|CH^2\big| \leqslant \varepsilon \quad \text{или} \quad \big|CH^2\big| \leqslant \varepsilon\Big|S_{M_2}(f)\Big|, \tag{6}$$

где ε — заданная абсолютная или относительная погрешность результата.

Выписанные выше соотношения (2)–(5) носят асимптотический характер, поэтому значение C, найденное из (3), будет достоверным (т. е. близким к истинному) лишь при достаточно малом H_2. В ответственных случаях после решения задачи с шагом H, удовлетворяющим условиям (6), для контроля над точностью решают задачу с шагом $2H$ и опять определяют главный член погрешности, соответствующий шагу H.

Описанный метод уточнения результата по итогам двух расчетов применим к методам *любого* порядка точности, причем не обязательно брать $M_2 = 2M_1$.

Задача 2. Имеется некоторый метод решения задачи с погрешностью

$$I(f) - S_M(f) \sim C/M^m.$$

Произведено вычисление интеграла с M_1 и $M_2 = \lambda M_1$ отрезками разбиения. Показать, что

$$I(f) - S_{M_2}(f) \sim \frac{S_{M_2}(f) - S_{M_1}(f)}{\lambda^m - 1};$$

здесь имеется в виду предельный переход при $M_2 \to \infty, \quad \lambda = \text{const}$.

Задача 3. Пусть

$$I(f) - S_M(f) = \frac{C}{M^m} + O\left(\frac{1}{M^{m+1}}\right).$$

Показать, что

$$I(f) - S_{M_2}(f) \sim \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1}$$

при условии, что $M_1,\ M_2 - M_1 \to \infty$.

Задача 4. Пусть

$$I(f) - S_M(f) = \frac{C}{M^m} + O\left(\frac{1}{M^{m+2}}\right).$$

Показать, что

$$I(f) - S_{M_2}(f) \sim \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1}$$

при условии, что $M_1 \to \infty,\ M_2 > M_1$.

В случаях, когда вычисляется большое количество интегралов с особенностями определенного вида, без серьезного теоретического анализа нельзя определить порядок сходимости метода на интегралах этого рода (из-за неограниченности производных мы не имеем права пользоваться результатом о существовании главного члена погрешности). В других же случаях порядок погрешности может быть известен, но неясно, каким он оказывается при реально используемых значениях M.

Рассмотрим вопрос о способах проверки выполнимости соотношения $R(f) \sim CM^{-m}$ при реально допустимых значениях M.

Можно постараться подобрать модельную задачу с известным ответом, близкую к рассматриваемой. Тогда после проведения расчета мы будем иметь в распоряжении приближенное значение S_m и погрешность $R_m = I - S_m$. Может случиться, что имеются какие-то предположения о характере поведения этой величины, например что

$$R_m \sim \text{const} \cdot M^{-m}. \tag{7}$$

В таком случае можно подсчитать для некоторой последовательности M_k значения $M_k^m R_{M_k}$ и посмотреть, стабилизируются ли эти величины с ростом M. Если нет предположения о характере поведения погрешности в данной задаче, то можно применить следующую методику.

Возьмем координатную плоскость $\ln M$, $\ln\left(\frac{1}{|R|}\right)$ (рис. 3.14.1), нанесем на нее точки

$$\left(\ln M_k,\ \ln\frac{1}{|R_{M_k}|}\right). \tag{8}$$

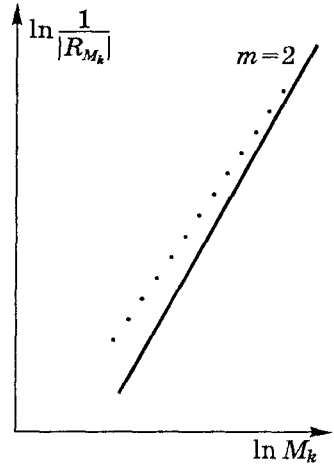


Рис. 3.14.1

Если эти точки расположены хаотически, то это означает, что числа M_k не настолько велики, чтобы в погрешности выделился главный член. Предположим, что асимптотическое неравенство (7) в данной области изменения параметра M выполняется с большой точностью. Из (7) следует, что

$$\ln\left|\frac{1}{R_m}\right| \sim m\ln M;$$

после дифференцирования имеем

$$\frac{d\ln\left|\frac{1}{R_M}\right|}{d\ln M} \sim m. \tag{9}$$

Заметим, что операция дифференцирования асимптотических равенств, вообще говоря, незаконна.

Согласно (9) в случае, когда (7) выполняется достаточно точно, точки $(\ln M_k,\ \ln(1/|R_{Mk}|))$, получаемые в результате эксперимента на ЭВМ, должны лежать на кривой, тангенс угла наклона которой стремится к m. Если угол наклона кривой меняется резко, то еще нет оснований применять правило Рунге.

Проверку справедливости предположения о характере поведения погрешности можно осуществлять и таким путем. Если справедливо равенство

$$R_M \sim c/M^m, \tag{10}$$

то

$$S_{M\lambda} - S_M \sim c(1-\lambda^{-m})/M^m. \tag{11}$$

С другой стороны, если (11) выполняется при $M \geqslant M_0$, то будет выполняться и (10). Поэтому вместо проверки практической выполнимости (10) можно производить проверку практической выполнимости (11), в частности, при помощи изучения графиков функций $g(M) = (S_{M\lambda} - S_m)M^m$ или расположения точек

$$\left(\ln M_k,\ \ln\frac{1}{|S_{M_k\lambda} - S_{M_k}|}\right). \tag{12}$$

Заметим, что возможности определения значения m и вообще проверки условия (10) путем численного эксперимента довольно ограничены. Например, случаи $R_m \sim \text{const}\cdot\dfrac{\ln M}{M}$ и $R_m \sim \text{const}\cdot\dfrac{1}{M}$ практически неразличимы при таком рассмотрении, потому что в обоих случаях $d\ln\left|\dfrac{1}{R_m}\right| \Big/ d\ln M \to 1$ при $M \to \infty$.

§ 15. Уточнение результата интерполяцией более высокого порядка точности

Если подынтегральная функция достаточно гладкая, то, как правило, погрешность квадратурной формулы может быть представлена в виде

$$I(f) - S_M(f) = R_M(f) = \sum_{k=1}^{l} D_k(f) M^{-i_k} + r(M), \tag{1}$$

где $i_1 < \ldots < i_l$, $r(M) = o(M^{-i_l})$. Обычно при гладкой подынтегральной функции имеем $i_2 - i_1 = \cdots = i_l - i_{l-1} = s$, где $s = 1$ или $s = 2$. Например, в предположении ограниченности $f^{(2m+2)}(x)$ погрешность формулы трапеций, согласно (13.2), представляется в виде

$$R_M(f) = \\ = -\sum_{k=1}^{m} \gamma_{2k} \left(\frac{B-A}{M}\right)^{2k} \left(f^{(2k-1)}(B) - f^{(2k-1)}(A)\right) + O\left(\left(\frac{B-A}{M}\right)^{2m+2}\right).$$

Предположим, что произведено вычисление $S_M(f)$ при значениях $M = M_0, \ldots, M_l$. Мы имеем равенства

$$I(f) = S_{M_j}(f) + \sum_{k=1}^{l} D_k(f) M_j^{-i_k} + r(M_j), \quad j = 0, \ldots, l.$$

Образуем линейную комбинацию этих соотношений с некоторыми коэффициентами c_j, потребовав, чтобы

$$\sum_{j=0}^{l} c_j = 1. \tag{2}$$

Получим соотношение

$$I(f) = \sum_{j=0}^{l} c_j S_{M_j}(f) + \sum_{k=1}^{l} D_k(f) \left(\sum_{j=0}^{l} c_j M_j^{-i_k}\right) + \sum_{j=0}^{l} c_j r(M_j).$$

Предположим, что выполняются равенства

$$\sum_{j=0}^{l} c_j M_j^{-i_k} = 0 \qquad \text{при} \qquad k = 1, \ldots, l, \tag{3}$$

тогда

$$I(f) = \sum_{j=0}^{l} c_j S_{M_j}(f) + \sum_{j=0}^{l} c_j r(M_j). \tag{4}$$

Если величиной $\sum_{j=0}^{l} c_j r(M_j)$ можно пренебречь, то

$$I(f) \approx \sum_{j=0}^{l} c_j S_{M_j}(f). \tag{5}$$

Система соотношений (2), (3) образует систему из $l+1$ линейных алгебраических уравнений с $l+1$ неизвестными, поэтому есть основания ожидать, что она имеет решение.

Аналогия между рассматриваемой задачей и задачей интерполяции позволяет найти c_j в явном виде. Перепишем (1) в виде

$$Q_l(M^{-l}) = S_M(f) - r(M), \quad \text{где} \quad Q_l(y) = I(f) - \sum_{k=1}^{l} D_k y^k. \tag{6}$$

Из соотношения (6) видно, что задача нахождения $I(f)$ может формулироваться следующим образом. Заданы значения многочлена $Q_l(y)$ при $y = M_0^{-1}, \ldots, M_l^{-1}$; требуется определить значение $Q_l(0) = I(f)$. Согласно интерполяционной формуле Лагранжа (гл. 2 § 2) имеем

$$Q_l(y) = \sum_{j=0}^{l} Q_l(y_j) \prod_{i \neq j} \frac{y - y_i}{y_j - y_i},$$

поэтому

$$Q_l(0) = \sum_{j=0}^{l} c_j Q_l(y_j), \qquad \text{где} \qquad c_j = \prod_{i \neq j} \frac{y_i}{y_i - y_j}, \quad y_j = M_j^{-1}, \tag{7}$$

и, следовательно,

$$I(f) = Q_l(0) = \sum_{j=0}^{l} c_j Q_l(y_j) = \sum_{j=0}^{l} c_j S_{M_j}(f) + \sum_{j=0}^{l} c_j r(M_j).$$

Мы получили соотношение (4) с выписанными явно значениями c_j.

Применение описанного метода, иногда называемого *методом Ромберга*, может быть полезным в следующей ситуации. Пусть мы задались какой-то квадратурой, вычислили на ЭВМ и выдали на печать значения $S_{M_0 2^i}(f)$, $i = 0, \ldots, l$, но оказалось, что нужной точности еще не достигли. Тогда можно попытаться получить приближение к интегралу, применив правило Ромберга по некоторой совокупности значений $S_{M_0 2^{s+1}}(f), \ldots, S_{M_0 2^{p+1}}(f)$.

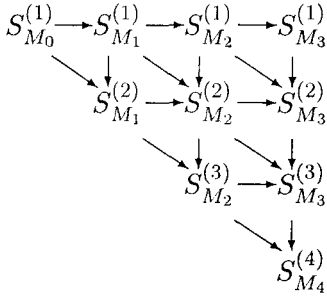
Иногда применяют следующую процедуру численного интегрирования. Задаются некоторым числом M_0 и последовательно вычисляют приближенные значения интеграла по формуле трапеций $I(f) \approx S^{(1)}_{M_k}(f)$ при M_k отрезках разбиения: $M_k = M_0 \cdot 2^k$. Удобнее всего вести вычисления по формуле

$$S^{(1)}_{M_k}(f) = \frac{1}{2} S^{(1)}_{M_{k-1}}(f) + \frac{B-A}{M_k} \sum_{j=0}^{M_{k-1}} f\left(A + \frac{2j-1}{M_k}(B-A)\right).$$

При каждом k после вычисления $S^{(1)}_{M_k}(f)$ последовательно вычисляют $S^{(1)}_{M_k}(f)$, ..., $S^{(k+1)}_{M_k}(f)$ по рекуррентной формуле

$$S^{(i)}_{M_k}(f) = S^{(i-1)}_{M_k}(f) + \frac{1}{2^i - 1}(S^{(i-1)}_{M_k}(f) - S^{(i-1)}_{M_{k-1}}(f)).$$

Таким образом, последовательность вычислений определяется схемой



Вычисления значений $S^{(i)}_{M_k}(f)$ обычно продолжают до тех пор, пока при некотором k не окажется, что $\min\limits_i |S^{(i)}_{M_k}(f) - S^{(i)}_{M_{k-1}}(f)| < \varepsilon$.

Как правило, метод Ромберга существенно уступает по эффективности квадратуре Гаусса и методам интегрирования с автоматическим выбором шага (см. § 17).

Задача 1. Показать, что $S^{(i)}_{M_k}(f)$ есть результат применения правила Ромберга к значениям $S^{(1)}_{M_k}(f), \ldots, S^{(1)}_{M_{k-i+1}}(f)$.

§ 16. Вычисление интегралов в нерегулярном случае

Существенную часть реально встречающихся подынтегральных функций составляют функции с особенностями, причем особенность может содержаться либо в функции, либо в ее производной, или функции, производные которых очень велики. Если такая нерегулярность подынтегральной

функции не вызвана колебательным характером ее поведения, то неплохой результат дают стандартные программы с автоматическим выбором шага, которые будут обсуждаться в § 17. В случае расчета малой серии интегралов с особенностями обращение к этим стандартным программам может оказаться наиболее целесообразным способом решения задачи. Для вычисления же большой серии интегралов с особенностями необходимо привлечь исследователей более высокой квалификации. Укажем ряд приемов, которые могут оказаться полезными при рассмотрении этих вопросов.

1. Выделение весовой функции. Пусть вычисляется интеграл $\int_a^b f(x)\,dx$, где пределы интегрирования a и b могут быть и бесконечными. Представим подынтегральную функцию в виде $f(x) = g(x)p(x)$, где $p(x)$ — достаточно простая, а $g(x)$ — гладкая функции. Далее применяем какой-либо из рассмотренных ранее способов вычисления интегралов с весом. Рассмотрим некоторые примеры.

Пусть вычисляется интеграл $\int_{-1}^{1} \frac{dx}{\sqrt{1-x^4}}$. Представим $f(x)$ в виде $\frac{1}{\sqrt{1+x^2}} \cdot \frac{1}{\sqrt{1-x^2}}$, где функция $\frac{1}{\sqrt{1+x^2}}$ является гладкой. Функцию $\frac{1}{\sqrt{1-x^2}}$ можно рассматривать как весовую. Этой весовой функции соответствует квадратура Мелера, часто называемая квадратурой Эрмита.

Пусть вычисляется $\int_0^1 f(x)dx$, причем $f(x)$ может быть представлена в виде $g(x)x^\alpha$, где $-1 < \alpha < 1$, $g(x)$ — гладкая функция, $g(0) \neq 0$. При вычислении интеграла по формуле трапеций с постоянным шагом $H = M^{-1}$ погрешность стремится к нулю медленнее, чем M^{-2}. Один из возможных способов вычисления интеграла — обращение к квадратурам Гаусса, соответствующим данной весовой функции.

2. Можно пойти по пути разбиения интеграла на части и вычисления интеграла по каждой части при помощи построений из § 7. Представим интеграл в виде

$$I = \sum_{q=1}^{M} I_q, \quad I_q = \int_{(q-1)H}^{qH} g(x)x^\alpha dx, \quad H = \frac{1}{M}.$$

Заменив функцию $g(x)$ на интерполяционный многочлен

$$P_{(q)}(x) = g\Big((q-1)H\Big) + \Big(x - (q-1)H\Big)\,\frac{g(qH) - g\big((q-1)H\big)}{H},$$

получим

$$I_q \approx \int_{(q-1)H}^{qH} P_{(q)}(x)x^\alpha dx =$$

$$= H^{\alpha+1}q^{\alpha+2}\left(\left(\frac{1-(1-1/q)^{\alpha+1}}{\alpha+1} - \frac{1-(1-1/q)^{\alpha+2}}{\alpha+2}\right)g(qH) + \right. \tag{1}$$

$$\left. + \left(\frac{1-(1-1/q)^{\alpha+2}}{\alpha+2} - \frac{1-1/q-(1-1/q)^{\alpha+2}}{\alpha+1}\right)g\big((q-1)H\big)\right).$$

Суммируя по q правые части в (1), получим квадратуру для вычисления исходного интеграла. В ряде случаев будет удобнее положить $g(qH) = (qH)^{-\alpha}f(qH)$ и, таким образом, получить квадратуру, имеющую вид

$$I \approx \sum_{q=0}^{M} D(q, H)f(qH). \tag{2}$$

Задача 1. Для квадратуры (2) получить оценку погрешности

$$\text{const} \cdot \max_{[0,1]} |g''(x)| M^{-2}.$$

Далее будут рассмотрены более простые по виду способы вычисления интегралов от функций с особенностями. Описанный выше способ аппроксимации интеграла по значениям функции на фиксированной, в частности, равномерной сетке обладает определенными преимуществами в случае, когда задача вычисления интеграла представляет часть более сложной задачи, например при решении интегральных уравнений путем сведения к решению системы линейных алгебраических уравнений. Иногда необходимая точность уже достигается при замене функции $g(x)$ на отрезках разбиения на постоянную. В этом случае полагаем

$$\int_{a_{q-1}}^{a_q} g(x)p(x)\,dx \approx g(\zeta_q)\int_{a_{q-1}}^{a_q} p(x)\,dx$$

и квадратура для вычисления исходного интеграла приобретает вид

$$\int_0^1 g(x)p(x)\,dx \approx \sum_{q=1}^{N} g(\zeta_q)\int_{a_{q-1}}^{a_q} p(x)\,dx. \tag{3}$$

Задача 2. Пусть вычисляется интеграл

$$I = \int_0^1 \frac{g(x)b}{b^2+x^2}\,dx, \quad \text{где} \quad b \text{ — малое число.}$$

Показать, что при использовании формулы трапеций с постоянным шагом $a_q - a_{q-1} \equiv H = M^{-1}$ погрешность оценивается через

$$\text{const} \cdot \min\left\{\frac{1}{Mb}, \frac{1}{(Mb)^2}\right\}. \tag{4}$$

Задача 3. При $a_q - a_{q-1} \equiv H$ квадратура (3) для этого интеграла имеет вид

$$I \approx \sum_{q=1}^{M} g(\zeta_q) \left(\operatorname{arctg}\left(\frac{qH}{b}\right) - \operatorname{arctg}\left(\frac{(q-1)H}{b}\right)\right),$$

где $(q-1)H \leqslant \zeta_q \leqslant qH$. Получить оценку погрешности

$$|R_M| \leqslant \text{const} \cdot \max_{[0,1]} \left|g'(x)\right| M^{-1}.$$

В рассматривавшихся выше случаях коэффициенты квадратур имеют вид

$$\int_{a_{q-1}}^{a_q} P_i(x) p(x)\, dx,$$

где $P_i(x)$ — некоторые многочлены, причем эти интегралы вычисляются в явном виде. Для ряда классов задач, где эти интегралы не вычисляются в явном виде, может оказаться разумным найти эти интегралы при помощи численного интегрирования. Эта дополнительная работа оправдывается, если получившиеся формулы используются многократно, например, при вычислении большой серии интегралов, при вычислении кратных интегралов как повторных (см. гл. 5), а также при решении интегральных уравнений.

3. Пусть теперь вычисляется

$$I_\omega(f) = \int_0^1 f(x) \exp(\mathrm{i}\omega x)\, dx,$$

где ω — большое число, $f(x)$ — достаточно гладкая функция. Будем рассматривать функцию $\exp\{\mathrm{i}\omega x\}$ как весовую. Представим интеграл в виде

$$I = \sum_{q=1}^{M} I_q, \quad I_q = \int_{(q-1)H}^{qH} f(x) \exp\{\mathrm{i}\omega x\}\, dx,$$

для вычисления интеграла I_q применим квадратуру типа (7.2).

4. В некоторых случаях подынтегральную функцию можно представить в виде $f(x) = G(x) + g(x)$, причем $\int_A^B G(x)dx$ берется в явном виде, а $g(x)$ — гладкая функция. Пусть вычисляется интеграл

$$I = \int_0^1 f(x)\, dx, \quad f(x) = \frac{\ln x}{1+x^2}. \tag{5}$$

Возьмем $G(x) = \ln x$. Тогда функция $g(x)$ имеет вид

$$g(x) = -\frac{x^2 \ln x}{1+x^2}.$$

Величина $\int_0^1 |g''(x)|\,dx$ будет конечна, и можно показать, что погрешность вычисления интеграла $\int_0^1 g(x)\,dx$ по формуле трапеций с постоянным шагом $a_q - a_{q-1} = M^{-1}$ имеет порядок $O(M^{-2})$. Чтобы погрешность формулы Симпсона имела порядок $O(M^{-4})$, следует взять $G(x) = (1-x^2)\ln x$.

В случае $f(x) = \left(x^2+b^2\right)^{-1} e^x$, b — малое число, целесообразно взять

$$G(x) = (\text{Выч}_{bi} f(z))(x-bi)^{-1} + (\text{Выч}_{-bi} f(z))(x+bi)^{-1}.$$

Достигаемое здесь расширение области аналитичности подынтегральной функции особенно эффективно при использовании формулы Гаусса.

5. Другим способом устранения особенности подынтегральной функции является замена переменной интегрирования. При замене переменных $x = \varphi(t)$, $\varphi(0) = 0$, $\varphi(1) = 1$ исходный интеграл $I = \int_0^1 f(x)\,dx$ преобразуется к виду

$$\int_0^1 g(t)\,dt, \tag{6}$$

где

$$g(t) = f(\varphi(t))\varphi'(t).$$

За счет множителя $\varphi'(t)$ происходит устранение особенностей подынтегральной функции в отдельных точках. Произведя в интеграле (5) замену переменной $x = t^k$, получим

$$I = \int_0^1 \frac{k^2 t^{k-1} \ln t}{1+t^{2k}}\,dt.$$

При $k > 2$ интеграл $\int_0^1 |g''(t)|\,dt$ конечен, поэтому погрешность формулы трапеций имеет порядок $O(M^{-2})$. С увеличением k растет порядок производных $g^{(n)}(t)$, для которых интеграл $\int_0^1 |g^{(n)}(t)|\,dt$ ограничен, поэтому можно применять квадратуры все более высокого порядка точности.

Если k очень большое, то производные функции $g(t)$ хотя и конечны, но очень большие, следовательно, должна соблюдаться определенная пропорция между величиной k и числом узлов N. Необходимость соблюдения осторожности при употреблении очень больших k видна хотя бы из следующего. Постоянный шаг $t_q - t_{q-1} = 1/M$ в интеграле (6) соответствует узлам интегрирования $a_q = \varphi(q/M) = (q/M)^k$ в исходном

интеграле, поэтому при больших k используется мало значений подынтегральной функции в правой части отрезка [0, 1] (рис. 3.16.1).

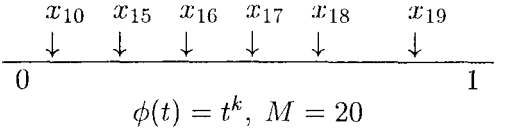


Рис. 3.16.1

6. Как мы уже видели в § 11, скорость сходимости при вычислении интегралов от функций с особенностями повышается также за счет распределения узлов интегрирования.

7. В некоторых случаях приходится идти по пути сочетания некоторых из описанных способов. Пусть вычисляется интеграл

$$\int_0^1 g(x)x^\alpha \exp\{\mathrm{i}\omega x\}\,dx,$$

где ω — большое число, $g(x)$ — гладкая функция, $g(0) \neq 0$, $|\alpha| < 1$. Наличие множителя $\exp\{\mathrm{i}\omega x\}$ требует выделения его как весового. Наличие множителя x^α требует принятия специальных мер для интегрирования в окрестности нуля. Замена переменных $x = \varphi(t)$ в данном случае является неприемлемой, поскольку для соответствующей весовой функции $\exp\{\mathrm{i}\omega x\}$ невозможно вычисление в явном виде коэффициентов квадратурных формул. Здесь целесообразнее разбить отрезок интегрирования на неравные части, соответствующие оптимальному распределению узлов при вычислении интеграла от функции x^α, и применить на каждой части интерполяционные квадратурные формулы (§ 3), соответствующие весовой функции $\exp\{\mathrm{i}\omega x\}$. В случае интегралов типа $\int_0^1 g(x)x^{-\alpha}\sin\omega x\,dx$ при $\alpha > 1$, $g(x)$ — гладкой функции, $g(0) \neq 0$, такой способ будет неприемлемым, поскольку в окрестности точки $x = 0$ неинтегрируемая функция $x^{-\alpha}$ не аппроксимируется многочленами. Здесь целесообразно разбить исходный интеграл на части $\int_0^\varepsilon$ и $\int_\varepsilon^1$, где $\varepsilon \sim 1/\omega$. Для вычисления второго интеграла разумно применить процедуру, которая описана выше. В первом интеграле функция $\sin\omega x$ не играет роли осциллирующего множителя, поскольку при таком выборе ε она имеет на $[0, \varepsilon]$ конечное число колебаний. Поэтому этот интеграл можно вычислять, например, распределив узлы интегрирования соответственно оптимальному распределению для функции $g(0)\omega x^{1-\alpha}$, аппроксимирующей подынтегральную при малых ωx.

8. Упомянем метод Ромберга. Погрешность формулы трапеций с постоянным шагом при вычислении интеграла $\int_0^1 g(x)x^\alpha dx$ для гладкой

функции $g(x)$, $g(0) \neq 0$, $-1 < \alpha < 1$, представляется в виде $D_1 N^{-1-\alpha} + D_2 N^{-2-\alpha} + \cdots$, и имеются основания для применения приемов, положенных в основу метода Ромберга.

9. Решение ряда задач сводится к вычислению сингулярных интегралов типа

$$I(a) = \int_A^B \frac{g(x)}{x-a}\,dx,$$

где $a \in (A, B)$, $g(a) \neq 0$. Интеграл понимается в смысле главного значения, т. е. как предел

$$\lim_{\varepsilon \to 0} \left(\int_A^{a-\varepsilon} \frac{g(x)}{x-a}\,dx + \int_{a+\varepsilon}^B \frac{g(x)}{x-a}\,dx \right).$$

Интеграл может быть записан как сумма интеграла по отрезку, симметричному относительно точки a, и интеграла от гладкой функции по оставшейся части. Для простоты предполагаем, что первый интеграл преобразован к виду $\int_{-b}^{b} \frac{g(x)}{x}dx$. Если функция $g(x)$ удовлетворяет условию Гельдера в точке $x = 0$, т. е. $|g(x) - g(0)| \leqslant A|x|^\alpha$, $\alpha > 0$, то последний интеграл равен несингулярному интегралу $\int_0^b \frac{g(x) - g(-x)}{x}\,dx$. В частности, если $g(x)$ — гладкая функция, то новая подынтегральная функция $\frac{g(x) - g(-x)}{x}$ также является гладкой.

В ряде случаев, например при решении интегральных уравнений с сингулярными ядрами, возникает следующая ситуация. Значения функции $g(x)$ задаются на некоторой фиксированной сетке. Исходя из информации об этих значениях, требуется вычислять значения интеграла $I(a)$ при различных a. Если $g(x)$ — достаточно гладкая функция, то здесь можно поступить следующим образом.

Разбиваем отрезок $[A, B]$ на части $[A_0, A_1], \ldots, [A_{M-1}, A_M]$, $A_0 = A$, $A_M = B$. На каждой из частей $[A_{q-1}, A_q]$ приближаем функцию $g(x)$ интерполяционным многочленом $L^q(x)$. При этом требуем, чтобы при всех q было выполнено условие

$$L^q(A_q) = L^{q+1}(A_q) = g(A_q).$$

Исходный интеграл заменяем суммой интегралов

$$i_q = \int_{A_{q-1}}^{A_q} \frac{L^q(x)}{x-a}dx.$$

Интегралы i_q вычисляем в явном виде. Если $a \in (A_{q-1}, A_q)$, то соответствующий интеграл i_q следует рассматривать как сингулярный. Если $a = A_q$, то

следует объединить интегралы i_q и i_{q+1} (расходящиеся) в один сингулярный интеграл

$$i_q + i_{q+1} = \int_{A_{q-1}}^{A_{q+1}} \frac{\bar{L}^q(x)}{x - A_q}\,dx,$$

$$\bar{L}^q(x) = \begin{cases} L^q(x) & \text{при } x < A_q, \\ L^{q+1}(x) & \text{при } x > A_q. \end{cases}$$

Получившиеся интегралы вычислим в явном виде.

Задача 5. Пусть отрезок интегрирования разбит на равные части длины H и на каждой части функция $g(x)$ аппроксимируется при помощи линейной интерполяции. Таким образом, исходный интеграл аппроксимируется суммой интегралов

$$\sum_{q=1}^{M} \int_{A+(q-1)H}^{A+qH} \frac{g((q-1)H) + (x - (q-1)H)\,\dfrac{g(qH) - g((q-1)H)}{H}}{x - a}\,dx,$$

где $H = (B - A)M^{-1}$. В предположении ограниченности $|g''(x)|$ получить оценку погрешности $O(M^{-2} \ln M)$.

Полезно указать на следующую практически важную деталь. Если решение задачи содержит какие-то неисследованные особенности, ухудшающие сходимость методов, то лучше сразу выделить простейшую модельную задачу, содержащую эти особенности, и провести выбор метода и проверку применимости различных асимптотических критериев на этой модельной задаче. Этот путь обычно приводит к более быстрому пониманию существа вопроса и избавляет от необходимости проведения многочисленных экспериментов на самой задаче. В частности, достигается экономия труда математика при программировании задачи, машинного времени и упрощается исследование за счет возможности построения более содержательных графиков поведения погрешности; здесь имеем возможность получить больше точек (14.8) или (14.12), поскольку для простой задачи их получение менее трудоемко.

§ 17. Принципы построения стандартных программ с автоматическим выбором шага

Как отмечалось в § 12, интегралы от функций с особенностями типа x^b хорошо вычисляются методами интегрирования с переменным шагом, если узлы интегрирования распределены оптимальным образом. По-видимому, столь же хорошо будут интегрироваться функции с особенностями других типов. Поэтому представляется заманчивым строить стандартные программы численного интегрирования так, чтобы для любой

функции распределение узлов являлось оптимальным или близким к нему.

В § 11 указана возможность распределения узлов, близкого к оптимальному, после исследования поведения подынтегральной функции на редкой сетке. Однако в случае резко меняющихся функций, например функций типа x^b, реализация этой возможности приводит к недостаточно удовлетворительным результатам.

Поэтому при разработке стандартных программ интегрирования приняты несколько другие процедуры распределения узлов интегрирования, обеспечивающие лучшее приближение к оптимальному распределению узлов для функций с особенностями. Рассмотрим некоторые из них.

Для вычисления интегралов по элементарным отрезкам разбиения $[a_{q-1}, a_q]$ выбираются: квадратурная формула

$$I_q(f) \approx \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{m} D_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \tag{1}$$

и мера погрешности

$$\rho_q(f) = \left| \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{m} B_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \right|. \tag{2}$$

Пусть вычисляется

$$\int_A^B f(x)dx, \quad A = a_0.$$

Первая процедура, которую естественно назвать *горизонтальной*, определяется заданием параметров β, $\sigma < 1$, h_0, и ε_0. Полагаем $\varepsilon_1 = \varepsilon_0\beta$. Предположим, что каким-то образом уже вычислено приближенное значение интеграла $\int_{a_0}^{a_q} f(x)\,dx$. Программа располагает в каждый момент времени некоторым значением h_q, с которым надо начинать считать оставшуюся часть интеграла. Вычисляем величину $\rho_q(f)$, соответствующую отрезку $[a_q, a_q + h_q]$. Если оказалось, что $\rho_q(f) \leqslant \varepsilon_0$, то вычисляем приближенное значение $\int_{a_q}^{a_q+h_q} f(x)\,dx$ по формуле (1) и полагаем $a_{q+1} = a_q + h_q$. Мы получили приближенное значение величины $\int_{a_0}^{a_{q+1}} f(x)\,dx$. В случае, когда $\varepsilon_1 < \rho_q(f)$, полагаем $h_{q+1} = h_q$, в противном случае полагаем $h_{q+1} = h_q/\sigma$. Мы готовы к следующему шагу. Если оказалось, что $\rho_q(f) > \varepsilon_0$, то принимаем σh_q за новое значение величины h_q и возвращаемся к исходной позиции: вычислено приближенное

значение интеграла $\int_{a_0}^{a_q} f(x)\,dx$ и задан шаг h_q. Начальные условия для применения процедуры:

$$q = 0, \quad \int_{a_0}^{a_q} f(x)\,dx = 0, \quad h_0.$$

Процедура должна также иметь блок окончания работы: если оказалось, что $a_q + h_q > B$, то следует положить $h_q = B - a_q$. Установилась практика брать $\sigma = 0{,}5$.

Другая процедура, которую можно назвать *вертикальной*, определяется заданием числа ε_0 и заключается в следующем. Пусть на каком-то шаге возникает необходимость вычисления интеграла по отрезку разбиения $[c, d]$, т. е. $\int_c^d f(x)\,dx$. Вычисляется величина $\rho(f)$, соответствующая этому отрезку. Если она оказалась меньше ε_0, то этот интеграл вычисляется по соответствующей формуле (1) и программа переходит к следующему справа отрезку разбиения. В противном случае отрезки $[c, (c+d)/2]$ и $[(c+d)/2, d]$ объявляются отрезками разбиения и программа сначала обращается к вычислению интеграла по левому из этих отрезков. В начале работы программа обращается к вычислению исходного интеграла $\int_A^B f(x)dx$.

Распределение узлов, осуществляемое этими процедурами, не является асимптотически оптимальным в смысле, определяемом в § 11, по следующим причинам: величина $\rho_q(f)$ не является главным членом погрешности квадратуры (1), а, как правило, есть некоторая грубая, завышенная по порядку оценка для него; отрезки для него могут принимать лишь довольно редкий ряд значений (при этом для первой процедуры они имеют вид $h_0\sigma^k$, а для второй — $(B - A)2^{-k}$).

Рассмотрим некоторые моменты, связанные с практическим использованием описанных выше процедур.

Чтобы в отдельных случаях сделать распределение узлов более близким к оптимальному, иногда величины $\rho_q(f)$, соответствующие отрезкам разбиения, сравнивают не с ε_0, а с $\varepsilon_0 h_q^\gamma$, где γ специально подбирается. На первоначальном этапе применения ЭВМ сложилась традиция производить сравнение величины $\rho_q(f)$ с величиной $\varepsilon_0 h_q$, т. е. брать $\gamma = 1$. Оказалось, что для резко меняющихся функций получающееся распределение узлов было далеко от оптимального, а в случае функций с разрывами программа не могла вычислить интеграл. В самом деле, пусть, например,

$$f(x) = \begin{cases} 0 & \text{при } x < c, \\ 1 & \text{при } x \geqslant c. \end{cases}$$

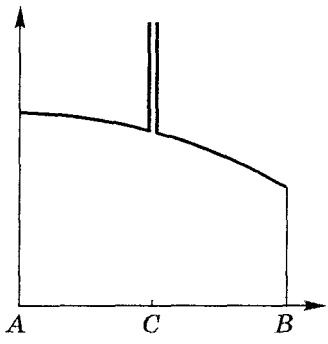


Рис. 3.17.1

Если при каком-то q величина $\rho_q(f)$ содержит одновременно значения f в узлах, где $f = 0$ и $f = 1$, то $\rho_q(f)$ имеет порядок h^q, поэтому нет оснований надеяться, что неравенство $|\rho_q(f)| \geqslant \varepsilon_0 h_q$ будет выполняться при малых h_q; программа будет осуществлять дробление шага до машинного нуля. Если функция меняется на малом участке очень резко, то шаг будет дробиться неоправданно сильно. Это обстоятельство было отмечено пользователями, и после теоретического анализа, в результате которого и была решена задача об оптимальном распределении узлов, было принято полагать $\gamma = 0$, если нет особых причин против этого.

В случае резко меняющихся функций следует иметь в виду, что программа может не заметить участка резкого изменения фупкции. Пусть $\rho(f) = 0$, когда $f(x) = P_s(x)$ — многочлен степени s, и пусть реальная подынтегральная функция есть

$$f(x) - P_s(x) + g(x),$$

где $g(x)$ практически отлична от нуля лишь на малом отрезке, например

$$g(x) = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left\{\frac{(x-C)^2}{2\varepsilon}\right\}$$

(рис. 3.17.1). Для определенности обратимся к первой из описанных выше процедур. Если отрезок разбиения $[a_{q-1}, a_q]$ удален от точки C, то $\rho(f) \approx \rho(P_s) = 0$ и программа будет увеличивать шаг. При подходе к точке C шаг может оказаться настолько большим, что окрестность, где функция $g(x)$ существенно больше 0, окажется заключенной между узлами. Тогда $\rho(f)$ будет близко к 0 и на отрезке, содержащем точку C. В результате в качестве значения $I(f)$ мы получим значение $I(P_s)$; погрешность этого приближенного значения близка к 1. Для контроля над точностью можно было бы попытаться произвести интегрирование с другим значением ε_0, однако с большой вероятностью получился бы тот же результат.

Не следует думать, что этот недостаток свойствен только методам интегрирования с автоматическим выбором шага. Если производить вычисление этого интеграла по формулам с постоянным шагом H, то при $H \gg \sqrt{\varepsilon}$ все равно может оказаться, что $g(x) \approx 0$ во всех узлах интегрирования, и мы получим приближенное значение $I(f) \approx I(P_s)$. Производя численное интегрирование при нескольких шагах и оценивая погрешность по правилу Рунге, можно прийти к неправильному выводу, что приближенное значение интеграла $I(f) \approx I(P_s)$. В целом можно сказать, что в случае использования алгоритмов интегрирования с автоматическим выбором шага возможность получения подобных неправильных выводов несколько больше.

При составлении стандартных программ всегда приходится балансировать между двумя крайностями: гарантией требуемой точности для любой подынте-

гральной функции или быстрым вычислением интеграла с нужной точностью для большинства предъявляемых к решению задач.

По-видимому, можно говорить о каком-то принципе неопределенности по отношению к методам высокой эффективности: дальнейшее повышение скорости работы должно сопровождаться уменьшением надежности. Чтобы избежать «проскакивания» областей резкого изменения функции, можно предусмотреть в программе наличие отрезков, где шаг интегрирования дробится принудительно. Например, в описание стандартной программы можно включить концы некоторого отрезка $[\alpha, \beta]$. Если отрезок разбиения $[a_{q-1}, a_q]$ пересекается с $[\alpha, \beta]$, то следующий отрезок выбирается по более сложному правилу. В качестве $[\alpha, \beta]$ следует задавать некоторый отрезок, где подынтегральная функция резко меняется.

Обратим внимание на ряд дополнительных моментов. Пусть для гладкой функции $f(x)$

$$\rho(f) = c_s \left| f^{(s)} \left(\frac{a_{q-1} + a_q}{2} \right) \right| (a_q - a_{q-1})^{s+1} + \varepsilon(a_{q-1}, a_q),$$
$$|\varepsilon(a_{q-1}, a_q)| \leqslant \bar{c}_s \max_{A,B} |f^{(s+1)}(x)| (a_q - a_{q-1})^{s+2}. \tag{3}$$

Для понимания механизма изменения шага предположим, что $f^{(s)}(x)$ — кусочно-постоянная функция. В области постоянства $f^{(s)}(x)$ имеем

$$\rho(f) = c_s |f^{(s)}(x)| (a_q - a_{q-1})^{s+1}.$$

Зададимся некоторым числом $\beta < 1$. Рассмотрим случай $\beta < \sigma^{s+1}$. Пусть для шага h_0, с которым мы приступаем к интегрированию в данной области постоянства производной, выполняется соотношение $c_s |f^{(s)}(x)| (h_0)^{s+1} \leqslant \varepsilon_1$. Тогда шаг будет увеличиваться каждый раз в σ^{-1} раз, пока не станет больше ε_1. Поскольку при шаге в σ^{-1} раз меньшем, выполнялось соотношение $\rho(f) \leqslant \varepsilon_1$, то интегрирование будет осуществляться с наименьшим шагом $h = h_0 \sigma^k$, для которого $c_s |f^{(s)}(x)| h^{s+1} > \varepsilon_1$. Если же для начального в этой области шага выполняется соотношение $c_s |f^{(s)}(x)| (h_0)^{s+1} > \varepsilon_0$, то шаг станет наибольшим из шагов $h = h_0 \sigma^k$, для которых выполняется условие $c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_0$. Если, как мы предположили, $\varepsilon_1 / \varepsilon_0 = \beta < \sigma^{s+1}$, то может оказаться, что в зависимости от поведения функции при меньших x в рассматриваемой области выбираются различные шаги. Мы видели, что шаг интегрирования желательно распределять так, чтобы оценка погрешности на всех отрезках разбиения интервала интегрирования была примерно одинаковой. Исходя из сказанного мы делаем вывод о желательности употребления $\beta \geqslant \sigma^{s+1}$.

Рассмотрим случай $\beta > \sigma^{s+1}$. Можно указать шаг $h = h_0^k$ такой, что

$$\varepsilon_0 \sigma^{s+1} < c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_0. \tag{4}$$

Если оказалось, что

$$\varepsilon_1 < c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_0, \tag{5}$$

то при интегрировании по рассматриваемому отрезку будет выбран именно такой шаг. Однако при

$$\varepsilon_0 \sigma^{s+1} < c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_1 = \varepsilon_0 \beta \tag{6}$$

программа выработает следующий шаг, равный h/σ. Для этого шага, вследствие (6), окажется, что $\varepsilon_0 < \rho(f)$, поэтому шаг h/σ будет признан непригодным. Программа возьмет шаг, равный h; поскольку для него выполняется условие (6), то интегрирование по текущему отрезку будет закончено и за исходный шаг для дальнейшего счета будет принят шаг h/σ. Таким образом, фактическое интегрирование происходит с шагом h и на каждом шаге делается попытка произвести интегрирование с шагом h/σ. Если вычисления с шагами h и h/σ независимы, то на каждом шаге объем работы будет удваиваться.

Известен следующий экспериментальный факт: если мы зададимся произвольным $\lambda > 1$ и из различных независимых источников получим некоторые числа y, то величины $\{\log_\lambda y\}$ будут асимптотически равномерно распределены на отрезке $[0, 1]$.

Задача 1. Считая, что $f^{(s)}(x)$ — числа, происходящие из независимых случайных источников, показать, что при $\beta > \sigma^{s+1}$ математическое ожидание затрат работ пропорционально $2 - \log_{\sigma^{s+1}} \beta$. (Отсюда следует целесообразность выбора $\beta = \sigma^{s+1}$.)

Наши рассуждения проведены в предположении, что $f^{(s)}(x) = \text{const}$, и поэтому требуют экспериментальной проверки. Исходя из результатов такой проверки на практике β обычно берется равным или несколько большим, чем σ^{s+1}, например, в одной из распространенных стандартных программ $s = 4$, $\sigma = 1/2$, $\beta = 1/10$.

Как и в случае интегрирования с постоянным шагом, возникает вопрос практической оценки погрешности. Обратимся ко второй процедуре, имеющей более простое описание. Как и ранее, проведем исследование на примере кусочно-постоянной производной $f^{(s)}(x)$. Шаг интегрирования в области постоянства $f^s(x)$ для этой процедуры определяется условием: это наибольший шаг h вида $(B-A)2^{-k}$, для которого выполнено условие

$$c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_0. \tag{7}$$

Очевидно, что для такого максимального шага справедливо соотношение

$$\varepsilon_0 2^{-(s+1)} < c_s |f^{(s)}(x)| h^{s+1}.$$

Произведем интегрирование с некоторым $\varepsilon_0' < \varepsilon_0$. В области, где

$$\varepsilon_0' < c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_0,$$

«старый» шаг интегрирования будет уже непригоден, поэтому произойдет дробление шага не менее чем вдвое. В области, где выполняется условие

$$\varepsilon_0 2^{-(s+1)} < c_s |f^{(s)}(x)| h^{s+1} \leqslant \varepsilon_0',$$

дробления шага не произойдет. Таким образом, при $\varepsilon_0 2^{-(s+1)} < \varepsilon_0'$ может случиться, что дробление шага интегрирования произойдет лишь на части отрезка интегрирования или его вообще не будет. Из сравнения

результатов расчетов с ε_0 и ε_0' мы не получим информации о величине погрешности интегрирования в той области, где дробления шага не произошло. При условии

$$\varepsilon_0' \leqslant \varepsilon_0 2^{-(s+1)} \tag{8}$$

шаг дробится всюду, однако отсюда не следует, что сравнение результатов расчетов с ε_0 и любым ε_0', удовлетворяющим (8), дает надежную гарантию точности результата.

Задача 2. Пусть $\varepsilon_0' \leqslant \varepsilon_0 2^{-(s+1)}$. Подобрать постоянные c_1, c_2, C такие, что при

$$f^{(s)}(x) = \begin{cases} c_1 & \text{на } [A, C], \\ c_2 & \text{на } [C, B] \end{cases}$$

приближенные значения $S_{\varepsilon_0}(f)$ и $S_{\varepsilon_0'}(f)$ интеграла, соответствующие ε_0 и ε_0', будут одинаковыми, а погрешность отлична от нуля.

Если $f^{(s)}(x)$ непрерывна, то при $\varepsilon_0' = \varepsilon_0 2^{-(s+1)}$ и при введении в алгоритм условия уменьшения максимального шага, пропорционального $\varepsilon_0^{1/(s+1)}$, можно предложить и обосновать аналог правила Рунге оценки погрешности

$$I(f) - S_{\varepsilon_0'}(f) \sim \frac{S_{\varepsilon_0'}(f) - S_{\varepsilon_0}(f)}{2^m - 1}, \quad \varepsilon_0 \to 0;$$

здесь $m+1$ — порядок главного (при $\max\limits_q (a_q - a_{q-1}) \to 0$) члена погрешности квадратуры (1).

Литература

1. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интерполирование и интегрирование. — Минск: Наука и техника, 1983.
2. Крылов В. И., Шульгина А. Т. Справочная книга по численному интегрированию. — М.: Наука, 1966.
3. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т.1. — М.: Наука, 1976.
4. Мысовских И. П. Интерполяционные кубатурные формулы. — М.: Наука, 1981.
5. Никифоров А. Ф., Суслов С. К., Уваров В. Б. Классические ортогональные полиномы дискретной переменной. — М.: Наука, 1985.
6. Никифоров А. Ф., Уваров В. Б. Специальные функции. — М.: Наука, 1979.
7. Никольский С. М. Квадратурные формулы. — М.: Наука, 1979.
8. Stroud A. H. and Secrest D. Gaussian Quadrature Formulas. — Englewood Cliffs, N. Y.: Prentice-Hall, 1966.

Глава 4

Приближение функций и смежные вопросы

Непрерывная функция не всегда может быть хорошо приближена интерполяционным многочленом Лагранжа. В частности, последовательность интерполяционных многочленов Лагранжа по равноотстоящим узлам не обязательно сходится к функции даже в том случае, если функция бесконечно дифференцируема. В тех случаях, когда сходимость имеет место, часто получение достаточно хорошего приближения требует использования полиномов высокой степени. В то же время, если для приближаемой функции удается подобрать подходящие узлы интерполяции, то степень интерполяционного многочлена, приближающего функцию с заданной точностью, может быть значительно снижена.

В ряде конкретных случаев целесообразно приближать функцию не путем интерполяции, а путем построения так называемого *наилучшего приближения*. Проблемы, связанные с построением наилучшего приближения, и будут рассмотрены в настоящей главе.

§ 1. Наилучшие приближения в линейном нормированном пространстве

Сформулируем задачу построения наилучшего приближения на абстрактном языке. Пусть имеется элемент f линейного нормированного пространства R. Требуется найти его наилучшее приближение линейной комбинацией $\sum_{j=1}^{n} c_j g_j$ данных линейно независимых элементов $g_1, \ldots, g_n \in R$. Это означает: найти элемент $\sum_{j=1}^{n} c_j^0 g_j$ такой, что

$$\left\| f - \sum_{j=1}^{n} c_j^0 g_j \right\| = \Delta = \inf_{c_1, \ldots, c_n} \left\| f - \sum_{j=1}^{n} c_j g_j \right\|.$$

По-другому это можно обозначить следующим образом:

$$\sum_{j=1}^{n} c_j^0 g_j = \arg \inf_{c_1, \ldots, c_n} \left\| f - \sum_{j=1}^{n} c_j g_j \right\|.$$

Если такой элемент существует, то он называется *элементом наилучшего приближения*.

Теорема. *Элемент наилучшего приближения существует.*

Доказательство. Вследствие соотношений (следствие из неравенства треугольника)

$$\left|\left\|f-\sum_{j=1}^{n}c_j^1g_j\right\|-\left\|f-\sum_{j=1}^{n}c_j^2g_j\right\|\right|\leqslant\left\|\sum_{j=1}^{n}\left(c_j^1-c_j^2\right)g_j\right\|\leqslant\sum_{j=1}^{n}|c_j^1-c_j^2|\,\|g_j\|$$

функция

$$F_f(c_1,\ldots,c_n)=\left\|f-\sum_{j=1}^{n}c_jg_j\right\|$$

является непрерывной функцией аргументов c_j при любом $f\in R$. Пусть $|\mathbf{c}|$ — евклидова норма вектора $\mathbf{c}=(c_1,\ldots,c_n)$. Функция $F_0(c_1,\ldots,c_n)=\|c_1g_1+\cdots+c_ng_n\|$ непрерывна на единичной сфере $|\mathbf{c}|=1$ и, следовательно, в некоторой ее точке $(\tilde{c}_1,\ldots,\tilde{c}_n)$ достигает своей нижней грани $\tilde{F}$ по сфере, причем $\tilde{F}\neq 0$, так как равенство $\tilde{F}=\|\tilde{c}_1g_1+\cdots+\tilde{c}_ng_n\|=0$ противоречит линейной независимости элементов $g_1,\ldots,g_n$. Для любого $\mathbf{c}=(c_1,\ldots,c_n)\neq(0,\ldots,0)$ справедлива оценка

$$\|c_1g_1+\cdots+c_ng_n\|=F_0(c_1,\ldots,c_n)=|\mathbf{c}|F_0\left(\frac{c_1}{|\mathbf{c}|},\ldots,\frac{c_n}{|\mathbf{c}|}\right)\geqslant\|\mathbf{c}\|\tilde{F}.$$

Пусть $\gamma>2\|f\|/\tilde{F}$. Функция $F_f(c_1,\ldots,c_n)$ непрерывна в шаре $|\mathbf{c}|\leqslant\gamma$; следовательно, в некоторой точке шара $(c_1^0,\ldots,c_n^0)$ она достигает своей нижней грани F^0 по шару. Имеем $F^0\leqslant F_f(0,\ldots,0)=\|f\|$. Вне этого шара выполняются соотношения

$$F_f(c_1,\ldots,c_n)\geqslant\|\,c_1g_1+\cdots+c_ng_n\,\|-\|f\|>$$
$$>\left(2\|f\|\,/\,\|\tilde{F}\|\right)\|\tilde{F}\|-\|f\|=\|f\|>F^0.$$

Таким образом, вне этого шара

$$F_f(c_1,\ldots,c_n)\geqslant F^0=F_f(c_1^0,\ldots,c_n^0)$$

при всех возможных $c_1,\ldots,c_n$. Теорема доказана.

Элементов наилучшего приближения, вообще говоря, может быть несколько.

Пространство R называется *строго нормированным*, если из условия

$$\|f+g\|=\|f\|+\|g\|,\quad\|f\|,\|g\|\neq 0$$

следует $f=\alpha g$, $\alpha>0$.

Задача 1. Доказать, что в случае строго нормированного пространства R элемент наилучшего приближения единствен.

Задача 2. Доказать, что пространство $L_p((0, 1), q(x))$, $q(x) \geqslant 0$ почти всюду, с нормой

$$\|f\|_p = \sqrt[p]{\int_0^1 |f(x)|^p q(x)\,dx}$$

строго нормированное при $1 < p < \infty$.

Рассмотреть отдельно простейший случай гильбертова пространства $p = 2$.

§ 2. Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении

Для гильбертова пространства элемент наилучшего приближения единствен (см. задачу 1.2) и проблема его нахождения формально сводится к решению системы линейных уравнений.

Наиболее простой способ получения этой системы следующий. По определению коэффициенты a_j элемента наилучшего приближения реализуют минимум выражения

$$\delta^2 = \left\|f - \sum_{j=1}^n a_j g_j\right\|^2 = \left(f - \sum_{j=1}^n a_j g_j,\ f - \sum_{j=1}^n a_j g_j\right).$$

Приравнивая нулю производные по $\operatorname{Re} a_i$ и $\operatorname{Im} a_i$, получаем искомую систему уравнений для определения a_i. Вследствие существования и единственности элемента наилучшего приближения, эта система имеет единственное решение.

Построим эту систему и исследуем вопрос о единственности ее решения несколько другим способом.

Для простоты изложения ограничимся случаем вещественных f и g_j.

Пусть $a_j = \alpha_j + \mathbf{i}\beta_j$, α_j, β_j — вещественные числа; имеем $f - \sum_{j=1}^n a_j g_j = \left(f - \sum_{j=1}^n \alpha_j g_j\right) - \mathbf{i}\sum_{j=1}^n \beta_j g_j$. Положим

$$\Phi(a_1, \ldots, a_n) = \left\|f - \sum_{j=1}^n a_j g_j\right\|^2.$$

Если f_1 и f_2 вещественны, то $\|f_1 + \mathbf{i}f_2\|^2 = (f_1 + \mathbf{i}f_2,\ f_1 + \mathbf{i}f_2) = (f_1, f_1) + \mathbf{i}(f_2, f_1) - \mathbf{i}(f_1, f_2) + (f_2, f_2) = \|f_1\|^2 + \|f_2\|^2$, поэтому

$$\left\|f - \sum_{j=1}^n a_j g_j\right\|^2 = \left\|f - \sum_{j=1}^n \alpha_j g_j\right\|^2 + \left\|\sum_{j=1}^n \beta_j g_j\right\|^2. \qquad (1)$$

Согласно (1) имеем равенство

$$\Phi(a_1, \ldots, a_n) = \Phi(\alpha_1, \ldots, \alpha_n) + \Big\|\sum_{j=1}^{n} \beta_j g_j\Big\|^2.$$

Отсюда следует, что

$$\inf_{a_1, \ldots, a_n} \Phi(a_1, \ldots, a_n) \geqslant \inf_{\alpha_1, \ldots, \alpha_n} \Phi(\alpha_1, \ldots, \alpha_n). \tag{2}$$

В то же время

$$\inf_{\alpha_1, \ldots, \alpha_n} \Phi(\alpha_1, \ldots, \alpha_n) \geqslant \inf_{a_1, \ldots, a_n} \Phi(a_1, \ldots, a_n), \tag{3}$$

поскольку в правой части берется нижняя грань по более широкому множеству всевозможных параметров $a_1, \ldots, a_n$, а в левой — по множеству вещественных параметров. Из (2) и (3) следует, что исходная задача сводится к нахождению

$$\inf_{\alpha_1, \ldots, \alpha_n} \Phi(\alpha_1, \ldots, \alpha_n).$$

В точке минимума должны выполняться условия $\partial\Phi/\partial\alpha_k = 0$. Имеем

$$\frac{\partial\Phi}{\partial\alpha_k} = \Big(-g_k, f - \sum_{j=1}^{n}\alpha_j g_j\Big) + \Big(f - \sum_{j=1}^{n}\alpha_j g_j, -g_k\Big) = -2\Big(f - \sum_{j=1}^{n}\alpha_j g_j, g_k\Big) = 0.$$

Отсюда получаем систему линейных уравнений относительно коэффициентов $\alpha_j = a_j$, соответствующих элементу наилучшего приближения

$$\sum_{j=1}^{n} a_j (g_j, g_k) = (f, g_k), \quad k = 1, \ldots, n. \tag{4}$$

Задача 1. Доказать, что коэффициенты a_j, соответствующие элементу наилучшего приближения, являются решением (4) и в том случае, когда f и g_j не обязательно вещественны.

Матрица $G_n = G(g_1, \ldots, g_n) = [(g_j, g_k)]$ называется *матрицей Грама* системы элементов $g_1, \ldots, g_n$. Поскольку $(g_j, g_k) = (g_k, g_j)$, то матрица Грама является эрмитовой.

Лемма. *Если элементы $g_1, \ldots, g_n$ линейно независимы, то матрица G_n положительно определена.*

Доказательство. Пусть $\mathbf{c} = (c_1, \ldots, c_n)^T$ — произвольный вектор с вещественными компонентами. Имеем равенство

$$\Big\|\sum_{j=1}^{n} c_j g_j\Big\| = \Big(\sum_{j=1}^{n} c_j g_j, \sum_{k=1}^{n} c_k g_k\Big) = \sum_{j,k=1}^{n} c_j c_k (g_j, g_k). \tag{5}$$

Последнее выражение совпадает с $(G_n\mathbf{c}, \mathbf{c})$, поэтому

$$(G_n\mathbf{c}, \mathbf{c}) = \Big\|\sum_{j=1}^{n} c_j g_j\Big\|^2 \geqslant 0.$$

Если элементы g_j линейно независимы, то $\Big\|\sum_{j=1}^{n} c_j g_j\Big\|^2 = 0$ только в том случае, когда все $c_j = 0$. Таким образом, $(G_n\mathbf{c}, \mathbf{c}) > 0$, если $\mathbf{c} \neq 0$, и согласно определению матрица G_n является положительно определенной. Поскольку матрица G_n положительно определена, то ее определитель отличен от нуля и, следовательно, система (4) имеет единственное решение.

Задача 2. Доказать неравенство

$$G(g_1, \ldots, g_{n+1}) \leqslant G(g_1, \ldots, g_n)(g_{n+1}, g_{n+1}).$$

При практическом приближении функций нужно проявлять осторожность при выборе системы функций g_j. Оказывается, что при неудачном выборе такой системы вычислительная погрешность коэффициентов a_j может достигать катастрофических размеров, и с добавлением новых функций g_n получаемое «наилучшее» приближение будет все хуже приближать заданную функцию.

Дело заключается в следующем. Матрица G_n при неудачном выборе системы функций g_j имеет большой разброс собственных значений, т. е. отношение максимального (по модулю) собственного значения к наименьшему (по модулю) велико; вычислительная погрешность при решении систем с такой матрицей возрастает по крайней мере пропорционально этому разбросу. Например, в случае отрезка $[-1, 1]$, веса $p(x) \equiv 1$ для системы функций $g_j = x^{j-1}$ разброс собственных значений матрицы G_n превосходит $a(\sqrt{2}+1)^{2n}/n^b$, где a, b — некоторые положительные постоянные. Более детальный анализ задачи показывает, что в качестве систем функций g_j целесообразно брать системы ортонормированных по отношению к некоторому, возможно другому, скалярному произведению, функций или в каком-то смысле близких к ним.

Если элементы g_k образуют ортонормированную систему $(g_k, g_j) = \delta_k^j$, то система (4) приобретает вид

$$a_j = (f, g_j). \qquad (6)$$

Тогда наилучшее приближение записывается в форме

$$g = \sum_{j=1}^{n} (f, g_j) g_j$$

и имеется следующее удобное представление для величины $\|f - g\|^2$:

$$\|f - g\|^2 = \Big(f - \sum_{j=1}^{n} a_j g_j,\ f - \sum_{j=1}^{n} a_j g_j\Big) = (f, f) - \sum_{j=1}^{n} |a_j|^2 = (f, f) - \sum_{j=1}^{n} |(f, g_j)|^2.$$

Поскольку $\|f-g\|^2 \geqslant 0$, то из равенства

$$\|f-g\|^2 = (f, f) - \sum_{j=1}^{n} |(f, g_j)|^2$$

следует, в частности, известное *неравенство Бесселя*

$$(f, f) \geqslant \sum_{j=1}^{n} |(f, g_j)|^2.$$

Если исходные элементы не образуют ортонормированной системы, то, вообще говоря, их можно ортогонализовать при помощи рассматривавшегося в гл. 3 алгоритма ортогонализации. Однако применение этого алгоритма довольно часто приводит к неудовлетворительным результатам. Например, при построении на отрезке $[-1, 1]$ ортонормированной с весом $p(x) > 0$ системы функций из указанной выше системы функций x^j будут получены некоторые ортогональные многочлены $P_j(x)$, сумма модулей коэффициентов у которых растет не медленнее чем $\left(\sqrt{2}+1\right)^n n^a$, где a определяется через $p(x)$. Если в дальнейшем значения многочленов вычисляются по явной формуле $P_j(x) = \sum_{k=0}^{l} a_{kj} x^k$, то из-за большой величины суммы модулей коэффициентов будет большая погрешность в значениях самих многочленов. Для устойчивого вычисления значений многочленов нужно применить какой-то иной алгоритм, например вычислять их по рекуррентным формулам или по явным формулам типа $T_n(x) = \cos(n \arccos x)$ в случае многочленов Чебышева. Некоторые более детальные сведения по этому вопросу будут приведены в § 8.

Пусть нам требуется приблизить функцию двух переменных $f(x, y)$ в некоторой области G на плоскости (x, y). Явный вид ортогональных функций известен только для простейших областей. Можно применить следующий прием. Возьмем некоторую область $\Omega \in G$, для которой известна ортогональная система функций, и продолжим f в области $\Omega \backslash G$. Далее будем приближать f в области G с помощью этой системы. Такой прием иногда неэффективен, поскольку не удается легко построить достаточно гладкое продолжение функции f в области $\Omega \backslash G$, а при недостаточно гладком продолжении приближение будет плохим. Следствием как этого обстоятельства, так и того, что приближение ищется в большей области, часто является неоправданное увеличение значения n, нужного для достижения заданной точности.

Другой возможный прием сведения к известной ортогональной системе функций состоит в следующем. Возьмем некоторую область Ω с известной ортогональной системой $\varphi_1(\zeta, \eta), \ldots, \varphi_n(\zeta, \eta)$. Пусть отображение $x = x(\zeta, \eta)$, $y = y(\zeta, \eta)$ переводит область Ω в G; $\zeta = \zeta(x, y)$, $\eta = \eta(x, y)$ — обратное отображение. Будем приближать функцию $f(x(\zeta, \eta), y(\zeta, \eta))$ в области Ω линейными комбинациями $\sum_{j=1}^{n} c_j \varphi_j(\zeta(x, y), \eta(x, y))$.

При приближении функций большого числа переменных не удается указать методов, пригодных во всех случаях. В каждой конкретной ситуации надо учитывать специфику задачи (вид функции, геометрию области и т. п.).

Рассмотрим для иллюстрации задачу другого рода, при решении которой используются проводимые выше построения. Пусть требуется построить интерполяционный многочлен степени $N-1$ с узлами интерполяции $x_1^N, \ldots, x_N^N$. Пусть эти узлы являются нулями $Q_N(x)$ степени N из ортонормированной системы многочленов $\{Q_n(x)\}$, соответствующей весу $p(x)$. Например, можно строить интерполяционные многочлены по нулям многочленов Чебышева, о которых шла речь в гл. 2. Будем отыскивать интерполяционный многочлен в виде линейной комбинации

$$P_{N-1}(x) = \sum_{j=0}^{N-1} a_j Q_j(x).$$

При $m, n < N$ многочлен $Q_m(x)Q_n(x)$ имеет степень не выше $2N-2$. Поэтому квадратура Гаусса с N узлами точна для этого многочлена:

$$\frac{b-a}{2} \sum_{q=1}^{N} D_q^N Q_m(x_q^N) Q_n(x_q^N) = \int_a^b Q_m(x)Q_n(x)p(x)\,dx = \delta_n^m. \tag{7}$$

Таким образом, векторы $\mathbf{Q}_m = \{Q_m(x_1^N), \ldots, Q_m(x_N^N)\}$ при $m < N$ образуют ортонормированную систему относительно скалярного произведения

$$(\mathbf{y}, \mathbf{z}) = \sum_{q=1}^{N} D_q^N y_q z_q \tag{8}$$

и вектор $\mathbf{f} = (f(x_1^N), \ldots, f(x_N^N))^T$ может быть разложен по этой системе векторов:

$$\mathbf{f} = \sum_{j=0}^{N-1} d_j \mathbf{Q}_j,$$

где $d_j = (\mathbf{f}, \mathbf{Q}_j)$. Многочлен

$$P_{N-1}(x) = \sum_{j=0}^{N-1} d_j Q_j(x) \tag{9}$$

будет искомым.

Таким образом, построение интерполяционного многочлена с узлами интерполяции, соответствующими корням ортогонального многочлена, сводится к вычислению коэффициентов разложения функции d_j по системе ортогональных многочленов. После этого искомый интерполяционный многочлен вычисляется по формуле (9). Этот алгоритм будет более устойчив по отношению к погрешностям округления по сравнению с непосредственным построением интерполяционного многочлена Лагранжа.

§ 3. Тригонометрическая интерполяция. Дискретное преобразование Фурье

Дискретное преобразование Фурье применяется при решении многих прикладных задач. К ним относятся тригонометрическая интерполяция, вычисление свертки функций, распознавание образов и многие другие. Дискретное преобразование Фурье стало особенно эффективным методом решения прикладных задач после создания быстрого преобразования Фурье (см. § 4).

Пусть $f(x)$ — периодическая функция с периодом 1 — разложена в ряд Фурье

$$f(x) = \sum_{q=-\infty}^{\infty} a_q \exp\{2\pi \mathbf{i} q x\}, \tag{1}$$

причем

$$\sum_{q=-\infty}^{\infty} |a_q| < \infty. \tag{2}$$

Здесь $\mathbf{i}$ — мнимая единица.

Рассмотрим значения этой функции на сетке из точек $x_l = l/N$, где l, N целые, N фиксировано, и обозначим $f(x_l) = f_l$. Если $q_2 - q_1 = kN$, где k целое, то $q_2 x_l - q_1 x_l = kNx_l = kl$, где kl целое. Следовательно,

$$\exp\{2\pi \mathbf{i} q_1 x\} = \exp\{2\pi \mathbf{i} q_2 x\} \tag{3}$$

в узлах сетки. Поэтому если функция $f(x)$ рассматривается лишь в узлах сетки x_l, то в соотношении (1) можно привести подобные члены

$$f_l = \sum_{q=0}^{N-1} A_q \exp\{2\pi \mathbf{i} q x_l\}, \tag{4}$$

где

$$A_q = \sum_{s=-\infty}^{\infty} a_{q+sN}. \tag{5}$$

Лемма. *При A_q, определяемых* (5), *соотношение* (4) *остается в силе, если пределы суммирования* $[0, N-1]$ *заменить на* $[m, N-1+m]$, *где m — любое целое.*

Доказательство. В самом деле, если $q' = q + kN$, то

$$A_{q'} = \sum_{m=-\infty}^{\infty} a_{q+kN+mN}.$$

Принимая $k + m$ за новую переменную суммирования m', получим

$$A_{q'} = \sum_{m'=-\infty}^{\infty} a_{q+m'N} = A_q.$$

Поскольку в узлах сетки $\exp\{2\pi\mathrm{i}q'x_l\} = \exp\{2\pi\mathrm{i}qx_l\}$ согласно (3), то в совокупности имеем

$$A_q \exp\{2\pi\mathrm{i}qx_l\} = A_{q'} \exp\{2\pi\mathrm{i}q'x_l\}.$$

Таким образом, при A_q, определяемом соотношением (5), функция $A_q \exp\{2\pi\mathrm{i}qx_l\}$ является периодической по q с периодом N и, следовательно, сумма

$$\sum_{q=m}^{N-1+m} A_q \exp\{2\pi\mathrm{i}qx_l\}$$

не зависит от m и совпадает с f_l. Лемма доказана.

Если с самого начала была задана функция, определенная только на сетке, то на этой сетке ее можно также представить в форме (1). Действительно, такую функцию можно продолжить на всю прямую, доопределив ее между узлами сетки путем линейной интерполяции. Для непрерывной кусочно-дифференцируемой функции выполняется (2), поэтому в точках сетки после приведения подобных членов получим (4).

Определим скалярное произведение для функций на сетке следующим образом:

$$(f, g) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \bar{g}_l.$$

(Множитель $1/N$ введен для согласованности получаемых соотношений с непрерывным случаем: если $f(x)$ и $g(x)$ — непрерывные функции на отрезке $[0, 1]$, то вследствие интегрируемости $f(x)g(x)$ по Риману

$$(f, g) \to \int_0^1 f(x)\bar{g}(x)\,dx$$

при $N \to \infty$.) Функции $g_q(x_l) = \exp\{2\pi\mathrm{i}qx_l\}$ при $0 \leqslant q < N$ образуют ортонормированную систему относительно введенного таким образом скалярного произведения. Действительно,

$$(g_q, g_j) = \frac{1}{N} \sum_{l=0}^{N-1} \exp\left\{2\pi\mathrm{i}\frac{q-j}{N}l\right\}.$$

При $q \neq j$, суммируя геометрическую прогрессию, имеем

$$(g_q, g_j) = \frac{1}{N} \frac{\exp\{2\pi\mathrm{i}(q-j)\} - 1}{\exp\left\{2\pi\mathrm{i}\dfrac{q-j}{N}\right\} - 1} = 0$$

(при $0 \leqslant q,\ j < N$, $q \neq j$ знаменатель отличен от 0). Поскольку $(g_q, g_q) = 1$, то в итоге имеем

$$(g_q, g_j) = \delta_q^j \quad \text{при} \quad 0 \leqslant q,\ j < N. \tag{6}$$

Умножая (4) скалярно на g_j, получим равенство

$$A_j = (f, g_j) = \frac{1}{N}\sum_{l=0}^{N-1} f_l \exp\{-2\pi \mathbf{i} j x_l\}. \tag{7}$$

Выражение в правой части образует квадратурную сумму для интеграла

$$\int_0^1 f(x)\exp\{-2\pi \mathbf{i} j x\}\,dx,$$

поэтому

$$A_j \to a_j = \int_0^1 f(x)\exp\{-2\pi \mathbf{i} j x\}\,dx$$

при $N \to \infty$ и фиксированном j.

Покажем, что соотношение

$$f(x) \approx \sum_{j=0}^{N-1} A_j \exp\{2\pi \mathbf{i} j x\} \tag{8}$$

в общем случае не имеет места. Пусть $f(x) = a_0 + a_{-1}\exp\{-2\pi \mathbf{i} x\}$. Из (4) получаем $A_0 = a_0$, $A_{N-1} = a_{-1}$, остальные $A_j = 0$. Таким образом, правая часть (8) есть $a_0 + a_{-1}\exp\{2\pi \mathbf{i}(N-1)x\}$. Она совпадает с $f(x)$ в точках x_l, но, как правило, далека от нее вне этих точек.

Воспользовавшись утверждением леммы, перепишем (4) в виде

$$f_l = \sum_{-N/2<q\leqslant N/2} A_q \exp\{2\pi \mathbf{i} q x_l\}. \tag{9}$$

Если $f(x)$ — достаточно гладкая функция, то величины $|a_j|$ с ростом j убывают быстро, поэтому $A_q \approx a_q$ при малых q. Кроме того, при гладкой $f(x)$ величины A_q и a_q малы при больших q.

Задача 1. Пусть $f(x)$ непрерывно дифференцируема. Доказать, что

$$\max_{[0,1]}\left| f(x) - \sum_{-N/2<q\leqslant N/2} A_q \exp\{2\pi \mathbf{i} q x\}\right| \to 0$$

при $N \to \infty$.

Напомним, что это приближенное равенство обращается в точное равенство в точках сетки. Способ аппроксимации функции

$$f(x) \approx \sum_{-N/2<q\leqslant N/2} A_q \exp\{2\pi \mathbf{i} q x\}$$

носит название *тригонометрической интерполяции*. Соотношение (9) называют *конечным* или *дискретным рядом Фурье*, а коэффициенты A_q — *дискретными коэффициентами Фурье*.

Игнорирование установленного нами факта о равенстве функций $\exp\{2\pi i q_1 x\}$ и $\exp\{2\pi i q_2 x\}$ в узлах сетки при $q_1 - q_2 = kN$ часто является источником получения неверных соотношений.

При решении одной инженерной задачи потребовалось определить первую собственную частоту колебаний конструкции. Было принято решение написать нестационарное уравнение, описывающее процесс колебаний, вывести на печать график и из рассмотрения графика определить частоту. Соответствующее уравнение, которое мы условно будем обозначать $x'' = F(x)$, решалось методом конечных разностей. Для контроля над надежностью результата производился повторный расчет с вдвое меньшим шагом. Графики кривых, полученных в результате расчетов, совпали с точностью до 10 %. Однако из сравнения с экспериментом оказалось, что полученная частота отличается от истинной в десятки раз. Причина недоразумения заключалась в том, что график решения строился с шагом $1/N$, существенно большим периода колебаний решения задачи. Решение было близко к функции $\text{const}\cdot\exp\{2\pi i q t\}$, где q/N близко к четному числу $2k$. Поэтому как на сетке с шагом $1/N$, так и на вдвое более мелкой с шагом $1/(2N)$ получался график одной и той же функции $\text{const}\cdot\exp\{2\pi i(q-2kN)t\}$.

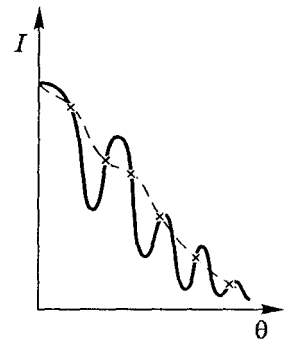


Рис. 4.3.1

В другом случае несоответствие со здравым смыслом возникло при расчете диаграммы направленности антенны. Предпринимавшиеся попытки найти ошибку в программе, методе решения или физическом описании задачи не приводили к положительному результату. Объяснение оказалось тем же: график сильно колеблющейся функции выдавался на очень редкой сетке. На рис. 4.3.1 сплошной кривой изображен реальный график сечения диаграммы направленности, пунктиром — график, который строился путем интерполяции полученных расчетных значений x и противоречил эксперименту.

Существует соответствие между задачей приближения функций линейными комбинациями многочленов Чебышева и тригонометрическими многочленами. Пусть на отрезке $[-1, 1]$ функция $f(x)$ приближается линейными комбинациями $\sum_{j=0}^{m-1} a_j T_j(x)$. Замена переменных $x = \cos t$ сводит исходную задачу к задаче приближения функции $f(\cos t)$ линейной комбинацией $\sum_{j=0}^{m-1} a_j T_j(\cos t) = \sum_{j=0}^{m-1} a_j \cos(jt)$.

Справедливо равенство

$$(f, g)_1 = \int_{-1}^{1} \frac{f(x)\bar{g}(x)}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} f(\cos\theta)\bar{g}(\cos\theta)\,d\theta.$$

Следовательно, задача наилучшего приближения $f(x)$ в норме, соответствующей скалярному произведению $(f, g)_1$, эквивалентна задаче приближения $f(\cos\theta)$ в норме, соответствующей скалярному произведению $(f, g)_2 = \int_0^\pi f(\cos\theta)\bar{g}(\cos\theta)\,d\theta$. Точно так же существует соответствие в случае задач интерполяции и наилучшего приближения в равномерной метрике. Задача интерполирования функции многочленом по узлам $x_j = \cos\left(\pi\dfrac{2j-1}{2m}\right)$ — нулям многочлена Чебышева $T_m(x)$ — после такой замены сводится к задаче интерполирования функции $f(\cos\theta)$ при помощи тригонометрического многочлена $\sum\limits_{j=0}^{m-1} a_j\cos(jt)$ по узлам $t_j = \pi\dfrac{2j-1}{2m}$, образующим равномерную сетку.

§ 4. Быстрое преобразование Фурье

Осуществление прямого и обратного дискретных преобразований Фурье

$$(f_0, \ldots, f_{N-1}) \Leftrightarrow (A_0, \ldots, A_{N-1})$$

является составной частью решения многих задач. Непосредственное осуществление этих преобразований по формулам (3.4), (3.7) требует $O(N^2)$ арифметических операций. Рассмотрим вопрос о возможности сокращения этого числа. Для определенности речь пойдет о вычислении коэффициентов A_q по заданным значениям функции. Идея построения алгоритмов *быстрого преобразования Фурье* опирается на то, что при составном N в слагаемых правой части (3.7) можно выделить группы, которые входят в выражения различных коэффициентов A_q. Вычисляя каждую группу только один раз, можно значительно сократить число операций.

Рассмотрим сначала случай $N = p_1p_2$; $p_1, p_2 \neq 1$. Представим q, j, лежащие в пределах $0 \leqslant q, j < N$, в виде $q = q_1 + p_1q_2$, $j = j_2 + p_2j_1$, где $0 \leqslant q_1, j_1 < p_1$, $0 \leqslant q_2, j_2 < p_2$. Имеем цепочку соотношений

$$A_q = A(q_1, q_2) = \frac{1}{N}\sum_{j=0}^{N-1} f_j \exp\left\{-2\pi\mathbf{i}\frac{qj}{N}\right\} =$$

$$= \frac{1}{N}\sum_{j_1=0}^{p_1-1}\sum_{j_2=0}^{p_2-1} f_{j_2+p_2j_1} \exp\left\{-2\pi\mathbf{i}\frac{(q_1+p_1q_2)(j_2+p_2j_1)}{p_1p_2}\right\}.$$

Из равенства

$$\frac{(q_1+p_1q_2)(j_2+p_2j_1)}{p_1p_2} = q_2j_1 + \frac{qj_2}{N} + \frac{j_1q_1}{p_1}$$

и предыдущего соотношения получим

$$A(q_1, q_2) = \frac{1}{p_2} \sum_{j_2=0}^{p_2-1} A^{(1)}(q_1, j_2) \exp\left\{-2\pi \mathbf{i} \frac{q j_2}{N}\right\},$$

где

$$A^{(1)}(q_1, j_2) = \frac{1}{p_1} \sum_{j_1=0}^{p_1-1} f_{j_2+p_2 j_1} \exp\left\{-2\pi \mathbf{i} \frac{j_1 q_1}{p_1}\right\}.$$

Непосредственное вычисление всех $A^{(1)}(q_1, j_2)$ требует $O\left(p_1^2 p_2\right)$ арифметических операций, а последующее вычисление $A(q_1, q_2)$ — еще $O\left(p_1 p_2^2\right)$ операций. Поэтому при $p_1, p_2 = O(\sqrt{N})$ общее число операций составит $O(N^{3/2})$. Точно так же при $N = p_1 \dots p_r$ строится алгоритм вычисления совокупности значений A_q, для которого общее число операций не превосходит $CN(p_1 + \dots + p_r)$, здесь C — постоянная, не зависящая от N. Выпишем соответствующие расчетные формулы для наиболее употребительного случая $p_1 = \dots = p_r = 2$. Представим числа q, j в виде

$$q = \sum_{k=1}^{r} q_k 2^{k-1}, \quad j = \sum_{m=1}^{r} j_{r+1-m} 2^{m-1},$$

где $q_k j_{r+1-m} = 0, 1$. Величину $qj2^{-r}$ представим в виде

$$qj2^{-r} = \sum_{m=1}^{r} q j_{r+1-m} 2^{m-1-r} = \sum_{m=1}^{r} \left(\sum_{k=1}^{r} q_k 2^{k+m-r-2} \right) j_{r+1-m} =$$

$$= \sum_{m=1}^{r} \left(\sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2} \right) j_{r+1-m} + s,$$

где s — целое, равное сумме всех слагаемых вида $q_k j_{r+1-m} 2^{k+m-r-2}$, у которых $k+m-r-2 \geqslant 0$. Очевидно, что $\exp\left\{-2\pi \mathbf{i} \frac{qj}{N}\right\} = \exp\left\{-2\pi \mathbf{i} \left(\frac{qj}{N} - s\right)\right\}$, поэтому

$$A(q_1, \dots, q_r) = A_q = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp\left\{-2\pi \mathbf{i} \frac{qj}{N}\right\} =$$

$$= \frac{1}{2} \sum_{j_r=0}^{1} \cdots \frac{1}{2} \sum_{j_1=0}^{1} f_{j_r+2j_{r-1}+\dots+2^{r-1}j_1} \exp\left\{-2\pi \mathbf{i} \sum_{m=1}^{r} \left(\sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2} \right) j_{r+1-m} \right\}.$$

После перегруппировки слагаемых имеем

$$A(q_1,\dots,q_r)=\frac{1}{2}\sum_{j_r=0}^{1}\exp\left\{-2\pi\mathbf{i}j_r2^{-r}\sum_{k=1}^{r}q_k2^{k-1}\right\}\times$$
$$\times\left(\frac{1}{2}\sum_{j_{r-1}=0}^{1}\exp\left\{-2\pi\mathbf{i}j_{r-1}2^{1-r}\sum_{k=1}^{r-1}q_k2^{k-1}\right\}\times\right.$$
$$\left.\times\cdots\times\left(\frac{1}{2}\sum_{j_1=0}^{1}\exp\{-2\pi\mathbf{i}j_12^{-1}q_1\}f_{j_r+2j_{r-1}+\cdots+2^{r-1}j_1}\right)\cdots\right).$$

Это соотношение можно записать в виде последовательности рекуррентных соотношений

$$A^{(m)}(q_1,\dots,q_m;\,j_{m+1},\dots,j_r)=$$
$$=\frac{1}{2}\sum_{j_m=0}^{1}\exp\left\{-2\pi j_m2^{-m}\sum_{k=1}^{m}q_k2^{k-1}\right\}A^{(m-1)}(q_1,\dots,q_{m-1};\,j_m,\dots,j_r),$$
$$m=1,\dots,r,$$

где

$$A^{(0)}(j_1,\dots,j_r)=f_{j_r+2j_{r-1}+\cdots+2^{r-1}j_1},$$
$$A^{(r)}(q_1,\dots,q_r)=A(q_1,\dots,q_r).$$

Переход от каждой совокупности $A^{(m-1)}$ к совокупности $A^{(m)}$ требует $O(N)$ арифметических и логических операций; всего таких шагов r, поэтому общее число операций имеет порядок $O(Nr)=O(N\log_2N)$.

Вычисление при помощи совокупностей $A^{(m)}$ дает меньшее накопление вычислительной погрешности по сравнению с формулами (3.7). Определенные удобства имеются также при вычислении экспонент, входящих в расчетные формулы. При вычислении величин $A^{(m)}$ используются значения $\exp\{-2\pi\mathbf{i}j2^{-m}\}$, $j=0,1,\dots,2^m-1$. В частности, при $m=1$ величина $\exp\{-\pi\mathbf{i}j\}$ принимает значения $+1$ или -1. Для вычисления значений $A^{(m+1)}$ потребуются еще значения $\exp\{-2\pi\mathbf{i}j2^{-(m+1)}\}$ при нечетных j, удовлетворяющих неравенству $0\leqslant j<2^{m+1}$. Их можно вычислить через уже вычисленные до этого величины, в частности, при помощи соотношений ($m\geqslant2$)

$$\exp\{-2\pi\mathbf{i}j\,2^{-(m+1)}\}=\frac{\exp\left\{-2\pi\mathbf{i}\,\frac{j+1}{2}2^{-m}\right\}+\exp\left\{-2\pi\mathbf{i}\,\frac{j-1}{2}2^{-m}\right\}}{2\cos(\pi2^{-m})},$$

где, в свою очередь,

$$\cos(\pi2^{-m})=\sqrt{\frac{1}{2}(1+\cos(\pi2^{1-m}))}$$

при $m\geqslant1$.

В ряде случаев удается еще уменьшить число операций. Один из таких случаев упоминался выше: дана вещественная функция $f(t) = g(\cos t)$, известная в точках $t_l = \pi(2l-1)/(2N)$; требуется найти коэффициенты интерполяционного многочлена

$$\sum_{j=0}^{N-1} A_j \cos jt.$$

Другой случай: при четном N заданы значения функции

$$\sum_{j=1}^{N/2-1} A_j \sin 2\pi jt$$

в точках $t = l/N$, $0 < l < N/2$; нужно определить коэффициенты A_j.

Задача 1. Найти коэффициенты c_j произведения двух многочленов

$$\left(\sum_{j=0}^{N-1} a_j x^j\right)\left(\sum_{j=0}^{N-1} b_j x^j\right) = \sum_{j=0}^{2N-2} c_j x^j.$$

Показать, что для их нахождения достаточно $O(N \log_2 N)$ операций.

§ 5. Наилучшее равномерное приближение

Если норма в линейном нормированном пространстве определяется не через скалярное произведение, то нахождение элемента наилучшего приближения существенно усложняется. Рассмотрим типичную задачу, встречающуюся, в частности, при составлении стандартных программ вычисления функций.

Пусть R — пространство ограниченных вещественных функций, определенных на отрезке $[a, b]$ вещественной оси, с нормой

$$\|f\| = \sup_{[a,b]} |f(x)|.$$

Ищется наилучшее приближение вида

$$Q_n(x) = \sum_{j=0}^{n} a_j x^j.$$

Согласно теореме из §1 существует элемент наилучшего приближения, т. е. многочлен $Q_n^0(x)$ такой, что

$$E_n(f) = \|f - Q_n^0\| \leqslant \|f - Q_n\|$$

при любом многочлене $Q_n(x)$ степени n. Такой многочлен $Q_n^0(x)$ называют *многочленом наилучшего равномерного приближения*. Далее будут установлены необходимые и достаточные условия того, чтобы многочлен

являлся многочленом наилучшего равномерного приближения для непрерывной функции.

Теорема Валле-Пуссена. *Пусть существуют* $n+2$ *точки* $x_0 < \cdots < x_{n+1}$ *отрезка* $[a, b]$ *такие, что*

$$\operatorname{sign}\bigl(f(x_i) - Q_n(x_i)\bigr)\,(-1)^i = \text{const},$$

т. е. при переходе от точки x_i *к точке* x_{i+1} *величина* $f(x) - Q_n(x)$ *меняет знак. Тогда*

$$E_n(f) \geqslant \mu = \min_{i=0,\ldots,n+1} \bigl|f(x_i) - Q_n(x_i)\bigr|. \tag{1}$$

Доказательство. В случае $\mu = 0$ утверждение теоремы очевидно. Пусть $\mu > 0$. Предположим противное, т. е. что для многочлена наилучшего приближения $Q_n^0(x)$

$$\|Q_n^0 - f\| = E_n(f) < \mu.$$

Имеем

$$\operatorname{sign}\Bigl(Q_n(x) - Q_n^0(x)\Bigr) = \operatorname{sign}\Bigl((Q_n(x) - f(x)) - (Q_n^0(x) - f(x))\Bigr).$$

В точках x_i первое слагаемое превосходит по модулю второе, поэтому $\operatorname{sign}\Bigl(Q_n(x_i) - Q_n^0(x_i)\Bigr) = \operatorname{sign}\Bigl(Q_n(x_i) - f(x_i)\Bigr)$. Следовательно, многочлен $Q_n(x) - Q_n^0(x)$ степени n меняет знак $n+1$ раз. Получили противоречие.

Теорема Чебышева. *Чтобы многочлен* $Q_n(x)$ *был многочленом наилучшего равномерного приближения непрерывной функции* $f(x)$, *необходимо и достаточно существования на* $[a, b]$ *по крайней мере* $n+2$ *точек* $x_0 < \cdots < x_{n+1}$ *таких, что*

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \,\|f - Q_n\|,$$

где $i = 0, \ldots, n+1$, $\alpha = 1$ *(или* $\alpha = -1$*) одновременно для всех* i.

Точки $x_0, \ldots, x_{n+1}$, удовлетворяющие условиям теоремы, принято называть *точками чебышевского альтернанса.*

Доказательство. Достаточность. Обозначим через L величину $\|f - Q_n\|$. Применяя (1), имеем $L = \mu \leqslant E_n(f)$, но $E_n(f) \leqslant \|f - Q_n\| = L$ вследствие определения величины $E_n(f)$. Следовательно, $E_n(f) = L$ и данный многочлен является многочленом наилучшего равномерного приближения.

Необходимость. Пусть данный многочлен $Q_n(x)$ является многочленом наилучшего равномерного приближения. Обозначим через y_1 нижнюю грань точек $x \in [a, b]$, в которых $|f(x) - Q_n(x)| = L$; из определения L следует существование такой точки. Вследствие непрерывности $f(x) - Q_n(x)$ имеем $|f(y_1) - Q_n(y_1)| = L$. Для определенности далее рассматриваем случай, когда $f(y_1) - Q_n(y_1) = +L$. Обозначим через y_2 нижнюю

грань всех точек $x \in (y_1, b]$, в которых $f(x) - Q_n(x) = -L$, последовательно через y_{k+1} обозначим нижнюю грань точек $x \in (y_k, b]$, в которых $f(x) - Q_n(x) = (-1)^k L, \ldots$ Вследствие непрерывности $f(x) - Q_n(x)$ при всех k имеем $f(y_{k+1}) - Q_n(y_{k+1}) = (-1)^k L$. Продолжаем этот процесс до значения $y_m = b$ или y_m такого, что $|f(x) - Q_n(x)| < L$ при $y_m < x \leqslant b$. Если $m \geqslant n+2$, то утверждение теоремы выполнено.

Предположим, что оказалось $m < n+2$. Вследствие непрерывности $f(x) - Q_n(x)$, при любом k $(1 < k \leqslant m)$ можно указать точку z_{k-1} такую, что $|f(x) - Q_n(x)| < L$ при $z_{k-1} \leqslant x < y_k$; положим $z_0 = a$, $z_m = b$. Согласно проведенным выше построениям, на отрезках $[z_{i-1}, z_i]$, $i = 1, \ldots, m$, имеются точки, в частности точки y_i, где $f(x) - Q_n(x) = (-1)^{i-1} L$, и нет точек, где $f(x) - Q_n(x) = (-1)^i L$. Положим

$$v(x) = \prod_{j=1}^{m-1} (z_j - x), \quad Q_n^d(x) = Q_n(x) + dv(x), \quad d > 0$$

и рассмотрим поведение разности

$$f(x) - Q_n^d(x) = f(x) - Q_n(x) - dv(x)$$

на отрезках $[z_{j-1}, z_j]$. Для примера обратимся к отрезку $[z_0, z_1]$. На $[z_0, z_1)$ имеем $v(x) > 0$, поэтому

$$f(x) - Q_n^d(x) \leqslant L - dv(x) < L.$$

Кроме того, на этом отрезке выполняется неравенство $f(x) - Q_n(x) > -L$; поэтому при достаточно малых d, например при

$$d < d_1 = \frac{\min\limits_{[z_0, z_1]} |f(x) - Q_n(x) + L|}{\max\limits_{[z_0, z_1]} |v(x)|}$$

на $[z_0, z_1)$ имеем $f(x) - Q_n(x) > -L$. В то же время

$$|f(z_1) - Q_n^d(z_1)| = |f(z_1) - Q_n(z_1)| < L.$$

Таким образом, $|f(x) - Q_n^d(x)| < L$ на этом отрезке при достаточно малом d. После проведения аналогичных рассуждений относительно остальных отрезков $[z_{i-1}, z_i]$ мы сможем указать малое d_0 такое, что на всех отрезках выполняется неравенство $|f(x) - Q_n^{d_0}(x)| < L$. Мы получили противоречие с предположением, что $Q_n(x)$ — многочлен наилучшего приближения, а $m < n+2$. Теорема доказана.

Теорема единственности. *Многочлен наилучшего равномерного приближения непрерывной функции единствен.*

Доказательство. Предположим, что существуют два многочлена степени n наилучшего равномерного приближения:

$$Q_n^1(x) \neq Q_n^2(x), \quad \|f - Q_n^1\| = \|f - Q_n^2\| = E_n(f).$$

Отсюда следует, что

$$\left\|f-\frac{Q_n^1+Q_n^2}{2}\right\|\leqslant\left\|\frac{f-Q_n^1}{2}\right\|+\left\|\frac{f-Q_n^2}{2}\right\|=E_n(f),$$

т. е. многочлен $\frac{1}{2}[Q_n^1(x)+Q_n^2(x)]$ также является многочленом наилучшего равномерного приближения. Пусть $x_0,\ldots,x_{n+1}$ — соответствующие этому многочлену точки чебышевского альтернанса; тогда

$$\left|\frac{1}{2}[Q_n^1(x_i)+Q_n^2(x_i)]-f(x_i)\right|=E_n(f),\quad i=0,\ldots,n+1,$$

или

$$|(Q_n^1(x_i)-f(x_i))+(Q_n^2(x_i)-f(x_i))|=2E_n(f).$$

Так как $|Q_n^k(x_i)-f(x_i)|\leqslant E_n(f)$, $k=1,2$, то последнее соотношение возможно лишь в том случае, когда

$$Q_n^1(x_i)-f(x_i)=Q_n^2(x_i)-f(x_i).$$

Мы получили, что два различных многочлена $Q_n^1(x)$ и $Q_n^2(x)$ степени n совпадают в точках $x_0,\ldots,x_{n+1}$, т. е. пришли к противоречию.

Задача 1. Функция $f(x)=\sin 100x$ приближается на отрезке $[0,\pi]$. Найти $Q_{90}(x)$.

§ 6. Примеры наилучшего равномерного приближения

1. Непрерывная на $[a,b]$ функция приближается многочленом нулевой степени. Пусть

$$\sup_{[a,b]}f(x)=f(x_1)=M,\quad\inf_{[a,b]}f(x)=f(x_2)=m.$$

Многочлен $Q_0(x)=(M+m)/2$ является многочленом наилучшего приближения, а x_1, x_2 — точками чебышевского альтернанса.

Задача 1. Доказать, что наилучшее приближение нулевой степени имеет вид $Q_0(x)=(M+m)/2$, если $f(x)$ не обязательно непрерывна.

2. Непрерывная, выпуклая на отрезке $[a,b]$ функция $f(x)$ приближается многочленом первой степени $Q_1(x)=a_0+a_1x$. Вследствие выпуклости $f(x)$ разность $f(x)-(a_0+a_1x)$ может иметь на интервале (a,b) только одну точку экстремума, поэтому точки a, b являются точками чебышевского альтернанса. Пусть d — третья точка чебышевского альтернанса. Согласно теореме Чебышева, имеем равенства

$$f(a)-(a_0+a_1a)=\alpha L,$$

$$f(d)-(a_0+a_1d)=-\alpha L,$$

$$f(b)-(a_0+a_1b)=\alpha L.$$

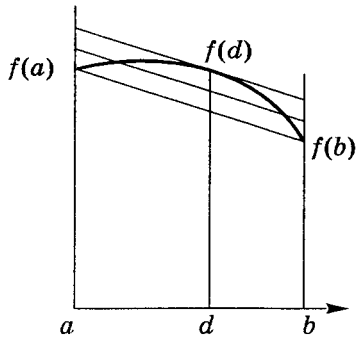


Рис. 4.6.1

Вычитая первое уравнение из третьего, получим $f(b)-f(a)=a_1(b-a)$. Отсюда находим $a_1=(f(b)-f(a))/(b-a)$. Для определения неизвестных d, L, a_0, a_1 и $\alpha=+1$ или $\alpha=-1$ получено всего три уравнения. Однако следует вспомнить, что точка d является точкой экстремума разности $f(x)-(a_0+a_1x)$. Если $f(x)$ — дифференцируемая функция, то для определения d имеем уравнение $f'(d)-a_1=0$. Теперь определяем a_0, например из уравнения, получающегося сложением первого и второго уравнений.

Геометрически эта процедура выглядит следующим образом (рис. 4.6.1). Проводим секущую через точки $(a,f(a))$, $(b,f(b))$. Для нее тангенс угла наклона равен a_1. Проводим параллельную ей касательную к кривой $y=f(x)$, а потом прямую, равноудаленную от секущей и касательной.

Задача 2. Построить пример функции и соответствующего многочлена первой степени наилучшего равномерного приближения на $[a,\,b]$ так, чтобы среди точек чебышевского альтернанса не было точек a и b.

Задача 3. Построить пример функции (естественно, не непрерывной), для которой многочлен наилучшего равномерного приближения не удовлетворяет условиям теоремы Чебышева.

Задача 4. Пусть $f(x)=|x|$, $[a,\,b]=[-1,\,5]$. Построить многочлен наилучшего равномерного приближения первой степени.

3. Функция $f(x)$, удовлетворяющая условию $f^{(n+1)}(x)\geqslant 0$, приближается на $[a,\,b]$ многочленом наилучшего равномерного приближения степени n; требуется оценить величину $E_n(f)$. В § 9 гл. 2 мы имели оценку погрешности интерполяции по узлам, являющимся нулями многочлена Чебышева:

$$x_k=\frac{a+b}{2}+\frac{b-a}{2}\cos\left(\frac{\pi(2k-1)}{2(n+1)}\right),$$

а именно,

$$|f(x)-P_n(x)|\leqslant\left(\max_{[a,\,b]}|f^{(n+1)}(x)|\right)\frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \tag{1}$$

Отсюда следует неравенство

$$E_n(f)\leqslant\left(\max_{[a,\,b]}|f^{(n+1)}(x)|\right)\frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}.$$

Пусть $Q_n(x)$ — многочлен наилучшего равномерного приближения. Поскольку вследствие теоремы Чебышева разность $f(x)-Q_n(x)$ меняет знак

при переходе от одной точки чебышевского альтернанса к другой, то она обращается в нуль в $(n+1)$-й точке $y_1, \dots, y_{n+1}$. Поэтому многочлен $Q_n(x)$ можно рассматривать как интерполяционный с узлами интерполяции $y_1, \dots, y_{n+1}$. Согласно (2.3.1) имеем представление для погрешности интерполирования следующего вида:

$$f(x) - Q_n(x) = f^{(n+1)}(\zeta)\,\frac{\omega_{n+1}(x)}{(n+1)!},$$

где $\omega_{n+1}(x) = (x - y_1)\dots(x - y_{n+1})$, $\zeta = \zeta(x) \in [a, b]$. Пусть

$$\max_{[a,\,b]} |\omega_{n+1}(x)| = |\omega_{n+1}(x_0)|.$$

Имеем

$$E_n(f) = \|f(x) - Q_n(x)\| \geqslant \Big|f(x_0) - Q_n(x_0)\Big| =$$

$$= \Big|f^{(n+1)}(\zeta(x_0))\Big|\,\frac{|\omega_{n+1}(x_0)|}{(n+1)!} \geqslant \Big(\min_{[a,\,b]} |f^{(n+1)}(x)|\Big)\Big(\max_{[a,\,b]} \frac{|\omega_{n+1}(x)|}{(n+1)!}\Big).$$

Согласно (2.8.6) выполняется неравенство

$$\max_{[a,\,b]} |\omega_{n+1}(x)| \geqslant (b-a)^{n+1}/2^{2n+1}.$$

Отсюда следует оценка

$$E_n(f) \geqslant \Big(\min_{[a,\,b]} |f^{(n+1)}(x)|\Big)\,\frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \tag{2}$$

Таким образом, если $f^{(n+1)}(x)$ сохраняет знак и меняется не очень сильно, то разность между погрешностью многочлена наилучшего равномерного приближения и интерполяционного многочлена по нулям многочленов Чебышева несущественна.

Задача 5. Доказать, что в случае, когда $f^{(n+1)}(x)$ сохраняет знак на отрезке $[a, b]$, чебышевский альтернанс содержит точки a и b.

4. Рассмотрим задачу нахождения многочлена наилучшего приближения степени n в случае, когда

$$f(x) = P_{n+1}(x) = a_0 + \dots + a_{n+1}x^{n+1}, \quad a_{n+1} \neq 0.$$

Тогда $f^{(n+1)}(x) = a_{n+1}(n+1)!$ и оценки сверху (1) и снизу (2) для $E_n(f)$ совпадают:

$$E_n(f) = |a_{n+1}|(b-a)^{n+1}2^{-2n-1}.$$

Таким образом, многочленом наилучшего приближения оказывается интерполяционный многочлен $Q_n(x)$ с узлами интерполяции

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{\pi(2k+1)}{2(n+1)}\right), \quad k = 1, \dots, n+1.$$

Можно получить другое представление этого многочлена наилучшего приближения, записав его в виде

$$Q_n(x) = P_{n+1}(x) - a_{n+1}T_{n+1}\left(\frac{2x-(a+b)}{b-a}\right)\frac{(b-a)^{n+1}}{2^{2n+1}}. \qquad (3)$$

Действительно, выражение в правой части является многочленом степени n, поскольку коэффициент при x^{n+1} равен нулю. Точки $x_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{\pi i}{n+1}$, $i = 0, \ldots, n$, образуют чебышевский альтернанс.

5. Пусть $[a, b] = [-1, 1]$ и $f(x)$ — нечетная функция относительно точки $\tilde{x} = 0$. Покажем, что здесь многочлен наилучшего приближения любой степени нечетен, т. е. записывается в виде суммы нечетных степеней x. Действительно, пусть $Q_n(x)$ — многочлен наилучшего приближения для $f(x)$. Имеем $|f(x) - Q_n(\tilde{x})| \leqslant E_n(f)$. После замены x на $-x$ и умножения выражения под знаком модуля на -1 получим

$$|-f(-x) - (-Q_n(-x))| \leqslant E_n(f),$$

иначе

$$|f(x) - (-Q_n(-x))| \leqslant E_n(f).$$

Следовательно, многочлен $-Q_n(-x)$ также является многочленом наилучшего равномерного приближения. По теореме единственности имеем $Q_n(x) = -Q_n(-x)$, что и требовалось доказать.

6. Пусть требуется приблизить функцию $f(x) = x^3$ на отрезке $[-1, 1]$ многочленом наилучшего приближения первой степени. Предшествующим результатом можно воспользоваться двояко. Один путь: поскольку искомый многочлен наилучшего приближения будет нечетным, то его достаточно отыскивать среди многочленов вида $Q_1(x) = \alpha_1 x$. Второй путь: поскольку многочлен наилучшего приближения второй степени для данной задачи оказывается многочленом первой степени, то исходная задача эквивалентна задаче построения многочлена наилучшего приближения второй степени. Последняя задача наилучшего приближения многочлена многочленом степени, на единицу меньшей, уже нами рассматривалась.

Задача 6. Пусть $f(x)$ — четная функция относительно середины отрезка приближения $[-1, 1]$: $f(x) = f(-x)$. Доказать, что многочлен наилучшего приближения $Q_n(x)$ четен.

Задача 7. Функцию $f(x) = \exp\{x^2\}$ приблизить на отрезке $[-1, 1]$ многочленом наилучшего приближения третьей степени.

Примечание. Из решения предыдущей задачи следует, что этот многочлен имеет вид $a_0 + a_2x^2$. Задача эквивалентна задаче наилучшего приближения функции $f_1(y) = e^y$ на отрезке $[0, 1]$ многочленом вида $a_0 + a_2y$.

7. Очень часто бывает, что многочлен наилучшего равномерного приближения точно найти не удается. В этих случаях ищется многочлен, близкий к

многочлену наилучшего приближения. Рассмотрим примеры такого рода. Для простоты рассматриваем случай приближения на отрезке $[-1, 1]$.

Разложим функцию $f(x)$ в ряд по ортогональной системе многочленов Чебышева:

$$f(x) \sim \sum_{j=0}^{\infty} d_j T_j(x).$$

Отрезок этого ряда

$$\sum_{j=0}^{n} d_j T_j(x)$$

невысокой степени часто обеспечивает неплохое равномерное приближение. Иногда бывает затруднительно вычислить явно коэффициенты d_j, но зато известно разложение Тейлора

$$f(x) = \sum_{j=0}^{\infty} a_j x^j,$$

сходящееся при $|x| \leqslant 1$. Тогда применяют следующий метод (называемый иногда *телескопическим*). Выбирают некоторое n такое, что погрешность формулы

$$f(x) \approx P_n(x) = \sum_{j=0}^{n} a_j x^j$$

является достаточно малой. Затем приближают многочлен $P_n(x)$ многочленом наилучшего равномерного приближения $P_{n-1}(x)$. Согласно формуле (3) имеем

$$P_{n-1}(x) = P_n(x) - a_n T_n(x) 2^{1-n}.$$

Поскольку $|T_n(x)| \leqslant 1$ на отрезке $[-1, 1]$, то

$$|P_{n-1}(x) - P_n(x)| \leqslant |a_n| 2^{1-n}.$$

Далее приближают многочлен $P_{n-1}(x)$ многочленом наилучшего равномерного приближения $P_{n-2}(x)$ и т. д. Понижение продолжается до тех пор, пока погрешность от таких последовательных аппроксимаций остается малой.

Рассматриваемый прием можно описать еще и следующим образом. Разложим многочлен $P_n(x)$ по многочленам Чебышева:

$$P_n(x) = \sum_{j=0}^{n} d_j T_j(x).$$

Введем обозначения $Q_m(x) = \sum_{j=0}^{m} d_j T_j(x)$ при $m \leqslant n$.

Всякий многочлен $Q_m(x)$ является многочленом наилучшего равномерного приближения степени m для многочлена $Q_{m+1}(x)$, при этом

$$E_m(Q_{m+1}) = \|Q_{m+1} - Q_m\| = |d_{m+1}|. \tag{4}$$

Это следует, например, из формулы (3) или непосредственно из теоремы Чебышева. Отсюда вытекает, что $Q_{n-1}(x) = P_{n-1}(x)$, $Q_{n-2}(x) = P_{n-2}(x)$ и т. д.

Таким образом, сущность описанного метода заключается в следующем. Исходная функция приближается отрезком ее ряда Тейлора $P_n(x)$. Затем многочлен $P_n(x)$ раскладывается на многочлены Чебышева и отбрасываются несколько последних членов разложения. Так как

$$|P_n(x) - P_m(x)| \leqslant \sum_{j=m+1}^{n} |d_j|,$$

то общая оценка погрешности такова:

$$\Big|f(x) - P_m(x)\Big| \leqslant \max\Big|f(x) - P_n(x)\Big| + \sum_{j=m+1}^{n} |d_j|.$$

Рассмотрим задачу приближения функции $f(x) = \operatorname{arctg} x$ на отрезке $[-\operatorname{tg}(\pi/8),\ \operatorname{tg}(\pi/8)]$ с точностью $0{,}5 \cdot 10^{-5}$. Для достижения такой точности при аппроксимации отрезком ряда Тейлора требуется положить

$$\operatorname{arctg} x \approx -\frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11}.$$

Приблизим полученный многочлен многочленом наилучшего приближения степени, на единицу меньшей. Повторяя данную процедуру три раза, получим многочлен

$$P_5(x) = 0{,}9999374x - 0{,}3303433x^3 + 0{,}1632823x^5,$$

также обеспечивающий требуемую точность. Отметим, что здесь из-за нечетности исходного многочлена показатель степени на каждом шаге уменьшался на 2.

Задача 8. Пусть функция $f(x)$ непрерывна на отрезке $[-1,\ 1]$:

$$Q_{n+1}(x) = a_{n+1}x^{n+1} + \cdots + a_0.$$

Доказать, что

$$E_n(f) \geqslant |a_{n+1}|2^{-n} - \|f - Q_{n+1}\|.$$

Задача 9. Пусть

$$f(x) = \sum_{j=0}^{\infty} d_j T_j(x).$$

Доказать, что

$$E_n(f) \geqslant |d_{n+1}| - \sum_{j=n+2}^{\infty} |d_j|.$$

Многочлены наилучшего равномерного приближения или близкие к ним используются как важный составной элемент в стандартных программах вычисления элементарных и специальных функций. Часто возникает ситуация, когда функция задается очень сложным явным выражением (например, в виде интеграла $f(x) = \int_G F(x, y)\,dy$), а по ходу решения конкретной задачи ее значение приходится вычислять в очень

многих точках. В этом случае часто полезно вместо непосредственного вычисления значений функции воспользоваться интерполяцией ее значений по таблице или приблизить функцию многочленом. Иногда для этой цели используют многочлены наилучшего приближения в норме L_2 или наилучшего равномерного приближения. Конечно, в каждом конкретном случае полезно посмотреть, оправдают ли себя затраты по построению приближающего многочлена.

§ 7. О форме записи многочлена

Одна из стандартных программ наилучшего равномерного приближения была отлажена в случае приближения функций многочленами невысокой степени и включена в пакет стандартных программ. Однако при практическом использовании программы для приближения функции многочленами высокой степени в ряде случаев оказывалось, что программа выдает приближение к функции, не обеспечивающее ожидаемой точности, или итерационный процесс не сходился, продолжаясь неограниченно долго, так что приходилось прекращать вычисления.

После практического и теоретического анализа возникшей ситуации удалось установить причину происходящего.

Для определенности будем говорить о приближении функции на отрезке $[-1, 1]$.

Было установлено, что в случае недостаточно гладкой функции коэффициенты многочлена, приближающего ее с высокой точностью, обязательно будут очень большими. За исключением редко встречающегося случая, когда эти коэффициенты записываются в ЭВМ без округлений, в этот многочлен вносится погрешность, и он плохо приближает рассматриваемую функцию. Если эти коэффициенты и записаны в машине без округлений, то все равно значения многочлена при $|x|$, близком к 1, будут находиться с большой вычислительной погрешностью.

Сформулируем соответствующие утверждения более строго. Предположим, что существует последовательность многочленов

$$P^m(x) = \sum_{j=0}^{n(m)} a_j^m x^j,$$

удовлетворяющих условию

$$|f(x) - P^m(x)| \leqslant 2^{-m} \quad \text{на} \quad [-1, 1] \tag{1}$$

($P^m(x)$ — не обязательно многочлены наилучшего равномерного приближения); предположим также, что коэффициенты этих многочленов растут не очень сильно:

$$l_m = \sum_{j=0}^{n(m)} |a_j^m| \leqslant M\, 2^{qm}, \quad q \geqslant 0. \tag{2}$$

Обозначим через G_φ открытую область, ограниченную дугами, из любой точки которых отрезок $[-1, 1]$ виден под углом φ. Известна

Теорема. *Пусть последовательность многочленов $P^m(x)$ удовлетворяет соотношениям (1), (2). Тогда функция $f(x)$ может быть аналитически продолжена в открытую область комплексного переменного G_φ, $\varphi = \pi(2q+1)/(2q+2)$.*

Примечание. Если вместо условия (1) имеем условие $|f(x) - P^m(x)| \leqslant 1/m^l$, то можно показать, что $\max\limits_{[-1+\delta,\,1-\delta]} |f^{(l)}(x)| < \infty$ при любом $\delta > 0$.

Какие практические следствия вытекают из этой теоремы? Если число a записывается в системе с плавающей запятой с t двоичными разрядами, то его погрешность может оказаться больше $|a|2^{-t-1}$; в худшем случае, когда погрешности коэффициентов a_j^m одного знака, погрешность значения

$$P^m(1) = \sum_{j=0}^{n(m)} a_j^m,$$

являющаяся следствием этих погрешностей, может превзойти величину

$$\left(\sum_{j=0}^{n(m)} |a_j^m|\right) 2^{-t-1} = l_m 2^{-t-1}.$$

Чтобы качественно представить характер влияния этой погрешности, будем считать, что число m, входящее в условия (1), (2), и число t велики. Пусть функция $f(z)$ аналитична в некоторой области G_{φ_0}, и не аналитична ни в какой области G_φ при $\varphi < \varphi_0$. Пусть q_0 выбирается из соотношения $q_0 = [2\varphi_0 - \pi]/2(\pi - \varphi_0)]$. Возьмем произвольное $q < q_0$. Если предположить, что для всех m выполняется неравенство $l_m \leqslant 2^{qm}$, то, согласно выше сформулированной теореме, $f(z)$ будет аналитична в соответствующей области G_φ при $\varphi < \varphi_0$ и получается противоречие с предположением о свойствах функции $f(z)$. Таким образом, будут встречаться сколь угодно большие m, для которых $l_m \geqslant 2^{qm}$. Следовательно, при неблагоприятном стечении обстоятельств погрешность от округления значений a_j^m может оказаться больше, чем $2^{qm}2^{-t-1}$. В то же время нельзя рассчитывать на оценку погрешности, лучшую, чем (1). Для суммарной погрешности, получающейся от замены $f(x)$ на $P^m(x)$ и от погрешностей в коэффициентах многочлена, мы не можем предложить оценки лучшей, чем $\varepsilon(m) = 2^{qm}2^{-t-1} + 2^{-m}$. Уравнение для точки экстремума $\bar{m}$ функции $\varepsilon(m)$ имеет вид

$$\varepsilon'(\bar{m}) = q \ln 2 \cdot 2^{q\bar{m}} 2^{-t-1} - \ln 2 \cdot 2^{-\bar{m}}.$$

Отсюда следует, что $2^{\bar{m}_l} = \left(\frac{2}{q}\right)^{1/(q+1)} 2^{t/(q+1)}$. Подставив это значение в выражение для $\varepsilon(\bar{m})$, получим

$$\varepsilon(\bar{m}) = \left(\frac{2}{q}\right)^{q/(q+1)} 2^{-t/(q+1)} 2^{-1} + \left(\frac{2}{q}\right)^{-1/(q+1)} 2^{-t/(q+1)} = \\ = \frac{q+1}{2}\left(\frac{2}{q}\right)^{q/(q+1)} 2^{-t/(q+1)}.$$

Таким образом, при вычислениях на ЭВМ с t разрядами нельзя надеяться на получение более точных приближений к значениям функции, чем с $t/(q+1)$ разрядами. Например, для функции $f(x) = (1+25x^2)^{-1}$ нижней гранью φ, при которой $f(z)$ аналитична в области G_φ, будет $\varphi_0 = \operatorname{arctg} 0{,}2$; соответствующее значение $q+1 \approx 3{,}2$. Таким образом, на ЭВМ с 60 двоичными разрядами мы в лучшем случае можем рассчитывать на получение приближений к значениям функции примерно с 20 верными двоичными разрядами.

Если производная $f^{(l)}(x)$ не существует в некоторой внутренней точке отрезка при не очень большом l, то согласно примечанию к теореме имеет место еще бóльшая потеря точности результата.

Можно привести следующий довод об исключительности подобной обстановки. Обычно приходится приближать целые функции или такие, особенности которых лежат далеко за единичным кругом. Тогда из проведенных выше рассмотрений столь удручающие выводы не следуют. Формально это верно, однако такого рода функции часто, являясь целыми аналитическими функциями, очень быстро возрастают при увеличении мнимой части у аргумента. Хотя для таких функций величины l_n остаются равномерно ограниченными по n, они могут принимать столь большие значения, что величина $|l_n|2^{-t-1}$ окажется существенно больше допустимой погрешности. Примером подобных функций является функция $\varphi(x) = \int_0^1 \exp\{\mathbf{i}\lambda t^2\}\,dt$ для больших λ.

Эти обстоятельства побуждают к отысканию иных простых форм записи и вычисления значений многочленов.

Обратимся к форме записи многочленов наилучшего приближения в виде суммы значений многочленов Чебышева

$$P_n(x) = \sum_{j=0}^{n} d_j T_j(x). \tag{3}$$

Имеем

$$\frac{2}{\pi}\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} 2 & \text{при } n = m = 0, \\ \delta_n^m & \text{при } n^2 + m^2 > 0, \end{cases}$$

поэтому

$$\frac{2}{\pi}\int_{-1}^{1} \frac{P_n^2(x)}{\sqrt{1-x^2}}dx = \frac{2}{\pi}\sum_{j=0}^{n}|d_j|^2 \int_{-1}^{1} \frac{T_j^2(x)}{\sqrt{1-x^2}}\,dx = 2d_0^2 + \sum_{j=1}^{n} d_j^2 \geqslant \sum_{j=0}^{n} d_j^2.$$

Воспользуемся неравенством Коши—Буняковского

$$\sum_{j=0}^{n}|d_j| = \sum_{j=0}^{n}1\cdot|d_j| \leqslant \sqrt{\Big(\sum_{j=0}^{n}1\Big)\cdot\Big(\sum_{j=0}^{n}d_j^2\Big)} = \sqrt{(n+1)\sum_{j=0}^{n}d_j^2}.$$

Отсюда заключаем, что

$$\sum_{j=0}^{n}|d_j| \leqslant \sqrt{\frac{2}{\pi}(n+1)\int_{-1}^{1}\frac{P_n^2(x)}{\sqrt{1-x^2}}\,dx}.$$

Если $\|P_n - f\|_C \to 0$, то

$$\frac{2}{\pi}\int_{-1}^{1}\frac{P_n^2(x)}{\sqrt{1-x^2}}\,dx \to \frac{2}{\pi}\int_{-1}^{1}\frac{f^2(x)}{\sqrt{1-x^2}}\,dx.$$

Таким образом, в этом случае величина $\sum\limits_{j=0}^{n}|d_j|$ растет не быстрее, чем $\sqrt{n}$. Поскольку $|T_n(x)| \leqslant 1$, то погрешность в значении многочлена, которая является следствием погрешностей в d_j, не превзойдет величины порядка $\Big(\sum\limits_{j=0}^{n}|d_j|\Big)2^{-t} = O(\sqrt{n}\,2^{-t})$. Такая оценка погрешности оказывается приемлемой при реальных вычислениях.

Конечно, не обязательно представлять многочлены, приближающие функцию, в виде линейной комбинации (3) многочленов Чебышева. В зависимости от конкретной обстановки иногда удобнее представлять многочлен как линейную комбинацию многочленов какой-либо другой ортогональной системы, например многочленов Лежандра. (Мы рассмотрели систему многочленов Чебышева вследствие наибольшей простоты рекуррентных соотношений для вычисления их значений.)

Дадим рекомендации по вычислению значений многочлена $P_n(x)$ на отрезке $[-1, 1]$ безотносительно к вопросу о наилучшем приближении функций.

Пусть в обычной форме этот многочлен записывается в виде $\sum\limits_{j=0}^{n}a_jx^j$ и погрешность порядка $\Big(\sum\limits_{j=0}^{n}|a_j|\Big)2^{-t}$ является допустимой. Тогда имеет смысл воспользоваться схемой Горнера

$$P_n(x) = a_0 + x(a_1 + x(\cdots + x(a_{n-1} + xa_n)\ldots)).$$

В противном случае этот многочлен целесообразно представлять в виде линейной комбинации ортогональных многочленов, например многочленов Чебышева:

$$P_n(x) = \sum_{j=0}^{n}d_jT_j(x).$$

Для уменьшения числа действий при вычислении значений $P_n(x)$ целесообразно поступить следующим образом. Полагаем

$$D_0 = d_0 - d_2/2, \quad D_j = (d_j - d_{j+2})/2, \quad j = 1, \ldots, n-2,$$
$$D_{n-1} = d_{n-1}/2, \qquad v_n = d_n/2.$$

При необходимости вычислить $P_n(x)$ находим $y = 2x$, $v_{n-1} = yv_n + D_{n-1}$, затем последовательно находим $v_{n-2}, \ldots, v_0 = P_n(x)$ по рекуррентной формуле $v_k = yv_{k+1} - v_{k+2} + D_k$. Известно, что вычислительная погрешность такого алгоритма есть

$$O\left(n \min\left\{n, \frac{1}{\sqrt{1-x^2}}\right\} \max_{[-1,1]} P_n(x)\, 2^{-t}\right).$$

Задача 2. Проверить, что действительно $v_0 = P_n(x)$.

§ 8. Интерполяция и приближение сплайнами

Для определенности будем говорить о приближении функции $f(x)$ на отрезке $[0, 1]$. Разобьем его на части $[x_0, x_1], \ldots, [x_{N-1}, x_N]$, $x_0 = 0$; $x_N = 1$, и обозначим это разбиение через Δ. Назовем *сплайном* $S^m_\Delta(f, x)$ *порядка* m функцию, являющуюся многочленом степени m на каждом из отрезков $[x_{n-1}, x_n]$, т. е.

$$S^m_\Delta(f, x) = P_{nm}(x) = a_{n0} + \cdots + a_{nm}x^m \quad \text{при} \quad x_{n-1} \leqslant x \leqslant x_n, \tag{1}$$

и удовлетворяющую условиям непрерывности производных до порядка $m-1$ в точках $x_1, \ldots, x_{N-1}$:

$$P^{(k)}_{nm}(x_n) = P^{(k)}_{n+1,m}(x_n) \quad \text{при} \quad k = 0, \ldots, m-1, \; n = 1, \ldots, N-1. \tag{2}$$

Всего имеется в распоряжении $Q = N(m+1)$ неизвестных коэффициентов a_{nm}, и соотношения (2) образуют систему из $(N-1)m$ линейных алгебраических уравнений. Другие уравнения для коэффициентов получаются из условия близости сплайна к приближаемой функции и из некоторых дополнительных условий.

Рассмотрим простейшую задачу приближения линейными сплайнами ($m = 1$). Тогда общее число Q свободных параметров равно $2N$. Поставим вопрос о построении сплайна $S^1_\Delta(f, x)$, совпадающего с функцией $f(x)$ в точках $x_0, \ldots, x_N$; получится система уравнений

$$P_{n1}(x_{n-1}) = f(x_{n-1}), \quad n = 1, \ldots, N,$$
$$P_{n1}(x_n) = f(x_n), \qquad n = 1, \ldots, N.$$

Эта система распадается на системы уравнений относительно коэффициентов отдельных многочленов

$$P_{n1}(x_{n-1}) = a_{n0} + a_{n1}x_{n-1} = f(x_{n-1}),$$
$$P_{n1}(x_n) = a_{n0} + a_{n1}x_n = f(x_n).$$

Отсюда находим

$$a_{n1} = (f(x_n) - f(x_{n-1}))/(x_n - x_{n-1}),$$
$$a_{n0} = f(x_{n-1}) - a_{n1}x_{n-1}.$$

Многочлен $P_{n1}(x)$ является многократно рассматривавшимся интерполяционным многочленом первой степени с узлами интерполяции x_{n-1}, x_n.

Широкое распространение сплайнов во многом вызвано тем, что они являются в определенном смысле наиболее гладкими функциями среди функций, принимающих заданные значения. Сплайны степени выше первой в случае гладкой $f(x)$ хорошо приближают не только саму функцию, но и ее производные.

Сплайны первой степени $S^1_\Delta(f, x)$ возникают из рассмотрения следующей вариационной задачи. Рассмотрим множество S_1 кусочно-дифференцируемых функций $s(x)$, удовлетворяющих условиям

$$s(x_n) = f(x_n), \quad n = 0, \ldots, N; \quad I_1(s) = \int_0^1 (s'(x))^2 dx < \infty.$$

Поставим задачу найти функцию $s_1(x) \in S_1$, реализующую

$$\inf_{s \in S_1} I_1(s).$$

Уравнение Эйлера для этого функционала имеет вид $s''(x) = 0$. Таким образом, $s_1(x)$ линейна на каждом из отрезков $[x_{n-1}, x_n]$ и, следовательно, $s_1(x) = S^1_\Delta(f, x)$.

Этот подход к приближению функции, при котором естественным образом возник сплайн $S^1_\Delta(f, x)$, допускает различные обобщения. Рассмотрим, например, множество S_2 непрерывных, непрерывно-дифференцируемых и дважды кусочно непрерывно-дифференцируемых функций $s(x)$, удовлетворяющих условиям

$$s(x_n) = f(x_n), \quad n = 0, \ldots, N; \quad I_2(s) = \int_0^1 (s''(x))^2 dx < \infty.$$

Ищется функция $s_2(x) \in S_2$, реализующая $\inf\limits_{s \in S_2} I_2(s)$. Решением этой задачи оказывается сплайн третьей степени, удовлетворяющий условиям $P''(x_0) = 0$, $P''(x_N) = 0$.

Справедливость этого утверждения непосредственно следует из соответствующих результатов вариационного исчисления. Однако для полноты изложения ниже будет приведено его обоснование.

Уравнением Эйлера для рассматриваемого функционала является уравнение

$$\frac{d^2}{dx^2}\frac{\partial}{\partial s''}(s'')^2 = 0,$$

т. е. $s^{(4)} = 0$. Поэтому естественно предположить, что экстремум реализуется на некоторой функции $s_2(x) \in S_2$, непрерывной вместе с первой производной, являющейся многочленом третьей степени $P_{n3}(x) = a_{n0} + a_{n1}x + a_{n2}x^2 + a_{n3}x^3$ на каждом из отрезков $[x_{n-1}, x_n]$. Пусть $\eta(x)$ — непрерывная, непрерывно-дифференцируемая и дважды кусочно непрерывно-дифференцируемая функция, удовлетворяющая условиям

$$\eta(x_0) = \cdots = \eta(x_N) = 0. \tag{3}$$

Положим

$$F(t) = \int_0^1 (s_2''(x) + t\eta''(x))^2 dx = \int_0^1 (s_2''(x))^2 dx + \\ + 2t\int_0^1 s_2''(x)\eta''(x)\,dx + t^2\int_0^1 (\eta''(x))^2\,dx.$$

Поскольку по предположению $F(t) \geqslant F(0)$, то

$$F'(0) = 2\int_0^1 s_2''(x)\eta''(x)\,dx = 0.$$

Представим величину $B(\eta) = F'(0)/2$ в виде $B = b_1 + \cdots + b_N$,

$$b_n = \int_{x_{n-1}}^{x_n} s_2''(x)\eta''(x)\,dx = \int_{x_{n-1}}^{x_n} P_{n3}''(x)\eta''(x)\,dx,$$

и произведем в выражении для b_n дважды интегрирование по частям. Получим

$$b_n = \int_{x_{n-1}}^{x_n} P_{n3}^{(4)}(x)\eta(x)\,dx + \left(P_{n3}''(x)\eta'(x) - P_{n3}'''(x)\eta(x)\right)\Big|_{x_{n-1}}^{x_n}.$$

Вследствие (3) и равенства $P_{n3}^{(4)}(x) = 0$ имеем

$$b_n = P_{n3}''(x)\eta'(x)\Big|_{x_{n-1}}^{x_n}.$$

После суммирования по n получим

$$B(\eta) = P_{N3}''(x_N)\eta'(x_N) + \sum_{n=1}^{N-1}(P_{n3}''(x_n) - P_{n+1,3}''(x_n))\eta'(x_n) - P_{13}''(x_0)\eta'(x_0) \tag{4}$$

для любой функции $\eta(x)$ рассматриваемого вида. При любом l можно подобрать функцию $\eta_l(x)$ такую, что $\eta_l'(x_l) = 1$, $\eta_l'(x_n) = 0$ при $n \neq l$ в пределах $0 \leqslant n \leqslant N$.

Заметим, что мы не имели права производить интегрирование по частям непосредственно в исходном интеграле $B(\eta)$, так как неясен вопрос о существовании производных у $s_2(x)$ более высокого порядка, чем первый, в точках $x_1, \ldots, x_N$.

Подставляя такие $\eta_l(x)$ в равенство

$$B(\eta_l(x)) = 0, \quad l = 0, \ldots, N,$$

получаем $N+1$ уравнение

$$P''_{13}(0) = P''_{N3}(1) = 0, \quad P''_{n3}(x_n) - P''_{n+1,3}(x_n) = 0, \quad n = 1, \ldots, N-1. \tag{5}$$

Условие, что минимизирующая функция принимает заданные значения в точках x_n, порождает $2N$ уравнений

$$\begin{aligned} &P_{n3}(x_n) = P_{n+1,3}(x_n) = f(x_n), \quad n = 1, \ldots, N-1, \\ &P_{13}(x_0) = f(x_0), \quad P_{N3}(x_N) = f(x_N), \end{aligned} \tag{6}$$

а условия непрерывности $s'_2(x)$ в точках x_n порождают уравнения

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n), \quad n = 1, \ldots, N-1. \tag{7}$$

В совокупности нами получено $4N$ уравнений (5)–(7) относительно $4N$ неизвестных коэффициентов многочленов $P_{n3}(x)$.

Исследуем вопрос о разрешимости и о практическом решении системы уравнений (5)–(7). Для удобства введем в рассмотрение величины $M_n = s''_2(x_n)$. Поскольку функция $s''_2(x)$ линейна на $[x_{n-1}, x_n]$, то

$$s''_2(x) = M_{n-1}\frac{x_n - x}{h_n} + M_n\frac{x - x_{n-1}}{h_n} \quad \text{на} \quad [x_{n-1}, x_n];$$

здесь $h_n = x_n - x_{n-1}$. Из этого соотношения и из условий

$$s_2(x_{n-1}) = f(x_{n-1}), \quad s_2(x_n) = f(x_n)$$

можно получить, что

$$\begin{aligned} s_2(x) = P_{n3}(x) = M_{n-1}\frac{(x_n - x)^3}{6h_n} + M_n\frac{(x - x_{n-1})^3}{6h_n} + \\ + \left(f(x_{n-1}) - \frac{M_{n-1}h_n^2}{6}\right)\frac{x_n - x}{h_n} + \left(f(x_n) - \frac{M_n h_n^2}{6}\right)\frac{x - x_{n-1}}{h_n} \end{aligned} \tag{8}$$

на $[x_{n-1}, x_n]$. Условия

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n), \quad n = 1, \ldots, N-1$$

порождают уравнения

$$\frac{h_n}{6}M_{n-1} + \frac{h_n + h_{n+1}}{3}M_n + \frac{h_{n+1}}{6}M_{n+1} = \frac{f(x_{n+1}) - f(x_n)}{h_{n+1}} - \frac{f(x_n) - f(x_{n-1})}{h_n}. \tag{9}$$

Кроме того, имеем условия

$$P''_{13}(x_0) = 0, \quad P''_{N3}(x_N) = 0,$$

иначе $M_0 = 0$, $M_N = 0$. После подстановки $M_0 = 0$ и $M_N = 0$ соответственно в первое и последнее уравнения (8) получим систему

$$C\mathbf{M} = \mathbf{d} \tag{10}$$

из $(N-1)$-го уравнения с $(N-1)$-м неизвестным:

$$\mathbf{M} = (M_1, \ldots, M_{N-1})^T, \quad \mathbf{d} = (d_1, \ldots, d_{N-1})^T.$$

Элементы c_{ij}, $i, j = 1, \ldots, N-1$, матрицы C согласно (9) задаются соотношениями

$$c_{ij} = \begin{cases} h_i/6 & \text{при} \quad j = i-1, \\ (h_i + h_{i+1})/3 & \text{при} \quad j = i, \\ h_{i+1}/6 & \text{при} \quad j = i+1, \\ 0 & \text{при} \quad |i-j| > 1, \end{cases} \tag{11}$$

а элементы d_i столбца $\mathbf{d}$ — соотношениями

$$d_i = \frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1})}{h_i}.$$

Эта система решается методом прогонки (см. гл. 9) примерно за $8N$ арифметических операций. После нахождения M_j по формуле (8) определяем многочлены $P_{n3}(x)$.

Покажем, что система уравнений (10) однозначно разрешима. Поскольку число уравнений равно числу неизвестных, то достаточно показать, что однородная система

$$C\mathbf{M} = \mathbf{0} \tag{12}$$

имеет только нулевое решение. Предположим противное, т. е. что существует ненулевое решение $\mathbf{M}^0 = (M_1^0, \ldots, M_{N-1}^0)^T$ системы (12). Пусть n_0 — значение n, при котором достигается $\max\limits_{1 \leqslant n < N} |M_n^0|$, т. е.

$$|M_{n_0}^0| = \max_{1 \leqslant n < N} |M_n^0| \neq 0.$$

В уравнении $\sum\limits_{j=1}^{N-1} c_{n_0 j} M_j^0 = 0$ перенесем в правую часть все слагаемые, кроме $c_{n_0 n_0} M_{n_0}^0$. Получим

$$c_{n_0 n_0} M_{n_0}^0 = \sum_{j \neq n_0} c_{n_0 j} M_j^0. \tag{13}$$

Из соотношений (11), определяющих c_{ij}, следует, что при любом i справедливы соотношения

$$\frac{1}{2}c_{ii} > \sum_{j\neq i}|c_{ij}| > 0,$$

поэтому имеем цепочку неравенств

$$|c_{n_0n_0}M^0_{n_0}| \leqslant \sum_{j\neq n_0}|c_{n_0j}|\,|M^0_j| \leqslant \frac{1}{2}|c_{n_0n_0}|\max_j|M^0_j| =$$
$$= \frac{1}{2}|c_{n_0n_0}|\cdot|M^0_{n_0}| = \frac{1}{2}|c_{n_0n_0}M^0_{n_0}|.$$

Вычтем из обеих частей результирующего неравенства

$$|c_{n_0n_0}M^0_{n_0}| \leqslant \frac{1}{2}|c_{n_0n_0}M^0_{n_0}|$$

выражение, стоящее в правой части. Получим $\frac{1}{2}|c_{n_0n_0}M^0_{n_0}| \leqslant 0$. Поскольку $c_{n_0n_0}\neq 0$, то $M^0_{n_0}=0$. Мы пришли к противоречию с предположением о существовании у системы (12) ненулевого решения.

Подведем итог проведенным построениям. Доказано существование решения системы (10). Сплайн третьей степени, определяемый соотношениями (8), будет искомым сплайном, удовлетворяющим условиям (5)–(7).

Лемма. *Полученный сплайн $s_2(x)$ реализует* $\inf\limits_{s\in S_2} I_2(s)$.

Доказательство. Пусть $s(x)$ — произвольная функция из S_2. Положим $\eta(x) = s(x) - s_2(x)$. Имеем

$$F(1) = \int_0^1 (s''(x))^2dx = \int_0^1 \big(s_2''(x)+\eta''(x)\big)^2dx =$$
$$= \int_0^1 (s_2''(x))^2dx + 2B(s(x)-s_2(x)) + \int_0^1 ((s(x)-s_2(x))'')^2dx. \tag{14}$$

Поскольку $s_2(x)$ удовлетворяет условиям (5), то в соотношении (14) имеем $B(s(x)-s_2(x)) = 0$. Таким образом, при любой функции $s(x)\in S_2$ справедливо соотношение

$$I_2(s) = \int_0^1 (s''(x))^2dx = \int_0^1 (s_2''(x))^2dx + \int_0^1 \Big((s(x)-s_2(x))''\Big)^2dx$$
$$= I_2(s_2) + \int_0^1 ((s(x)-s_2(x))'')^2dx.$$

Отсюда следует справедливость утверждения леммы.

Из последнего соотношения следует также, что сплайн $s_2(x)$ является единственной функцией из рассматриваемого класса, на которой $I_2(s)$ достигает своего наименьшего значения.

Задача 1. Доказать, что решение системы (10) удовлетворяет неравенству

$$\max_{1\leqslant n\leqslant N-1}|M_n|\leqslant\frac{3}{\min\limits_{1\leqslant n\leqslant N-1}h_n}\max_{1\leqslant n\leqslant N-1}|d_n|;$$

получить отсюда однозначную разрешимость системы (5)–(7).

При рассматриваемом подходе получаемый сплайн совпадает с $f(x)$ во всех узлах; такие сплайны называют *интерполяционными.*

Задача 2. Пусть точки x_n распределены равномерно: $x_{n+1}-x_n\equiv h$. Значение интерполяционного сплайна третьей степени $s_2(x)$ выражается через значения функции f_n некоторой формулой

$$s_2(x)=\sum_{n=0}^{N}C_n(x)f_n.$$

Получить оценку

$$|C_n(x)|\leqslant a_3e^{-b_3\dfrac{|x-nh|}{h}};\qquad(15)$$

a_3, $b_3>0$ — абсолютные постоянные, не зависящие от f, N, h.

Из оценки (15) следует, что значение сплайна $s_2(x)$ в точке x слабо зависит от значений f_n при большом $|(x-nh)/h|$.

Описанный выше способ построения сплайна третьей степени страдает следующим недостатком. Из соотношений (5) следует $s_2''(0)=s_2''(1)=0$, хотя, как правило, $f''(0)$, $f''(1)\neq 0$; поэтому точность приближенных формул $f^{(k)}(x)\approx s_2^{(k)}(x)$ вблизи границы ухудшается. Если значения $f''(0)$ и $f''(1)$ известны, то в (9) при $n=1$ и $n=N-1$ следует положить $M_0=f''(0)$ и $M_N=f''(1)$. Оказывается, что для повышения точности целесообразно задавать в точках x_0, x_N значения первой производной. Дифференцируя (8), имеем

$$s_2'(x_0)=-\frac{M_0h_1}{3}-\frac{M_1h_1}{6}+\frac{f(x_1)-f(x_0)}{h_1}.$$

В случае, если величина $f'(x_0)$ известна, полагаем правую часть равной $f'(x_0)$ и получаем дополнительное уравнение, связывающее M_0 и M_1. Обычно значение $f'(x_0)$ неизвестно, поэтому поступим следующим образом. Определим интерполяционный многочлен $Q_0(x)$ третьей степени, совпадающий с $f(x)$ в точках x_0, x_1, x_2, x_3. Величину $f'(x_0)$ заменим выражением $Q_0'(x_0)$. Окончательно получим

$$\frac{h_1}{3}M_0+\frac{h_1}{6}M_1=\frac{f(x_1)-f(x_0)}{h_1}-Q_0'(x_0).$$

Последняя формула может быть записана в виде

$$\frac{h_1}{3}M_0 + \frac{h_1}{6}M_1 = \frac{h_1}{3}Q_0''(x_0) + \frac{h_1}{6}Q_0''(x_1).$$

Аналогично построим интерполяционный многочлен третьей степени $Q_N(x)$, совпадающий с $f(x)$ в точках x_N, x_{N-1}, x_{N-2}, x_{N-3}.

Коэффициенты M_j будем находить из системы уравнений, состоящей из совокупности уравнений (9) при $n = 1, \ldots, N-1$, и уравнений

$$\begin{aligned} &\frac{h_1}{3}M_0 + \frac{h_1}{6}M_1 = \frac{h_1}{3}Q_0''(x_0) + \frac{h_1}{6}Q_0''(x_1), \\ &\frac{h_N}{6}M_{N-1} + \frac{h_N}{3}M_N = \frac{h_N}{3}Q_N''(x_N) + \frac{h_N}{6}Q_N''(x_{N-1}). \end{aligned} \tag{16}$$

Часто вместо Q_0 и Q_N лучше брать многочлены четвертой степени, совпадающие с $f(x)$ в пяти крайних точках.

Задача 3. Пусть $h_n \equiv h = 1/N$, $|f^{(4)}(x)| \leqslant A_4$. Показать, что для сплайна, определяемого системой соотношений (9), (16), выполнены оценки

$$\max_{[0,1]} \left| f^{(q)}(x) - s_2^{(q)}(x) \right| \leqslant \text{const} \cdot A_4 h^{4-q}, \quad q \leqslant 3.$$

Описанные выше сплайны часто неудобны из-за своей нелокальности: значение сплайна в точке x зависит от значений $f(x_n)$ во всех узлах. Если в процессе работы со сплайнами (а она часто проходит в диалоговом режиме с визуализацией результатов на экране) требуется исправить одно значение, приходится заново решать систему уравнений (9). Особенно эта процедура неприятна в случае приближения функций многих переменных многомерными сплайнами.

Чтобы избежать этого, используем так называемые *локальные* (*аппроксимационные*) *сплайны.*

Локальный сплайн первой степени совпадает с построенным выше сплайном $s_1(x)$.

Локальные сплайны более высоких степеней, как правило, не совпадают с $f(x)$ в узлах x_n. Однако это обстоятельство не носит принципиального характера. Все равно, как правило, значения $f(x_n)$ известны с некоторой погрешностью δ_n, т. е. нам заданы величины $f_n = f(x_n) + \delta_n$.

Построение локального сплайна третьей степени опишем на примере случая постоянного шага $h_n \equiv h = 1/N$. Для этого используется стандартный сплайн $B(x)$ третьей степени, определяемый соотношениями

$$B(x) = \begin{cases} \dfrac{2}{3} - x^2 + \dfrac{1}{2}|x|^3 & \text{при} \quad |x| \leqslant 1, \\ \dfrac{1}{6}(2 - |x|)^3 & \text{при} \quad 1 \leqslant |x| \leqslant 2, \\ 0 & \text{при} \quad 2 \leqslant |x|. \end{cases}$$

Локальные сплайны третьей степени $B_2^{(1)}$ и $B_2^{(2)}$ записывают в виде

$$B_2^{(i)}(x) = \sum_{n=-1}^{N+1} \alpha_n^{(i)} B\left(\frac{x-nh}{h}\right), \quad i = 1,2;$$

способ выбора $\alpha_n^{(i)}$ будет указан ниже.

Задача 4. Доказать, что при любых $\alpha_n^{(i)}$ функции $B_2^{(i)}$ являются сплайнами третьей степени, причем $B_2^{(i)} \equiv 0$ вне отрезка $[-3h, 1+3h]$.

При $i = 1$ доопределяют значения f_{-1} и f_{N+1} линейной интерполяцией по значениям f_0, f_1 и f_N, F_{N-1}, соответственно, т. е. берут $f_{-1} = 2f_0 - f_1$, $f_{N+1} = 2f_N - f_{N-1}$ и полагают $\alpha_n = f_n$ при $-1 \leqslant n \leqslant N+1$.

При $i = 2$ доопределяют f_{-2}, f_{-1} и f_{N+1}, f_{N+2} кубической интерполяцией по значениям f_0, f_1, f_2, f_3 и f_N, f_{N-1}, f_{N-2}, f_{N-3}, соответственно, и полагают

$$\alpha_n = (8f_n - f_{n+1} - f_{n-1})/6.$$

Конкретные формулы для вычисления величин f_{-1}, f_{-2}, f_{N+1}, f_{N+2} имеют вид

$$f_{-1} = 4f_0 - 6f_1 + 4f_2 - f_3,$$
$$f_{-2} = 10f_0 - 20f_1 + 15f_2 - 4f_3,$$
$$f_{N+1} = 4f_N - 6f_{N-1} + 4f_{N-2} - f_{N-3},$$
$$f_{N+2} = 10f_N - 20f_{N-1} + 15f_{N-2} - 4f_{N-3}.$$

Задача 5. Показать, что значение $B_2^{(1)}(x)$ зависит только от значений f_n в четырех ближайших к x точках x_n, а значения $B_2^{(2)}(x)$ — в шести.

В случае, когда погрешности δ_n велики, чаще используют сплайны $B_2^{(1)}(x)$, а в случае, когда малы — сплайны $B_2^{(2)}(x)$; дело в том, что сплайны $B_2^{(1)}(x)$ обладают несколько лучшими свойствами сглаживания погрешностей в значениях $f(x_n)$, но обеспечивают меньшую точность в случае $\delta_n \equiv 0$ и гладкой $f(x)$.

Задача 6. Показать, что

$$B_2^{(1)}(x_0) = f_0, \quad B_2^{(1)}(x_N) = f_N,$$
$$B_2^{(2)}(x_0) = f_0, \quad B_2^{(2)}(x_1) = f_1, \quad B_2^{(2)}(x_{N-1}) = f_{N-1}, \quad B_2^{(2)}(x_N) = f_N.$$

Совпадение сплайнов $B_2^{(1)}(x)$ и $B_2^{(2)}(x)$ с $f(x)$ в концевых точках существенно упрощает применение таких сплайнов в задачах машинной графики.

Задача 7. Пусть $|f^{(k)}(x)| \leqslant A_k$ при $k = 2, 3, 4$. Показать, что

$$\max_{[0,1]} \left| B_2^{(1)}(x))^{(k)} - f^{(k)}(x) \right| \leqslant \text{const} \cdot h^{2-k} \quad \text{при} \quad k \leqslant 1,$$

$$\max_{[0,1]} \left| B_2^{(2)}(x))^{(k)} - f^{(k)}(x) \right| \leqslant cA_4 h^{4-k} \quad \text{при} \quad k \leqslant 3.$$

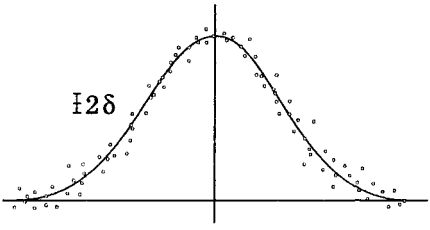


Рис. 4.8.1

Рассмотрим подход к построению сплайнов, основанный на идее регуляризации. Пусть известно, что погрешности δ_n — случайные величины с математическим ожиданием $M\delta_n = 0$ и дисперсией δ. Минимум $I_2(s)$ ищут при условии

$$\sum_{n=0}^{N} (s(x_n) - f_n)^2 \leqslant c(N+1)\delta^2; \quad (17)$$

постоянную c порядка 1 выбирают экспериментально.

Как правило, левая и правая части условия (17) равны в точке минимума $I_2(s)$. Поэтому обычно вместо исходной задачи рассматривают задачу безусловной минимизации по λ и s функционала

$$J_2(\lambda, s) = I_2(s) + \lambda\Big(\sum_{n=0}^{N} (s(x_n) - f_n)^2 - c(N+1)\delta^2\Big).$$

Рис. 4.8.1 соответствует некоторому конкретному случаю приближения при $N = 100$ и $c = 0{,}9$. При каждом фиксированном λ величина $J(\lambda) = \inf\limits_s J_2(\lambda, s)$ определяется с помощью решения системы линейных уравнений. Для нахождения $\inf\limits_\lambda J(\lambda)$ применяется какой-либо метод минимизации функций одной переменной (см. гл. 7).

Литература

1. Бабенко К. И. Основы численного анализа. — М.: Наука, 1986.
2. Бейкер Дж., Грейвс-Моррис П. Аппроксимации Паде. — М.: Мир, 1986.
3. Завьялов Ю. С., Квасов Б. И., Мирошниченко В.Л. Методы сплайн-функций. — М.: Наука, 1980.
4. Стечкин С. Б., Субботин Ю. Н. Сплайны в вычислительной математике. — М.: Наука, 1976.
5. Васильев Ф. П. Численные методы решения экстремальных задач. — М.: Наука, 1980.
6. Васильев Ф. П. Методы решения экстремальных задач. — М.: Наука, 1981.

Глава 5

Многомерные задачи



Вопросы приближения, интерполирования, численного интегрирования и дифференцирования функций одной переменной, как видно из предшествующего, разработаны достаточно подробно. В настоящее время на основе результатов теоретических исследований созданы довольно развитые системы стандартных программ решения одномерных задач. Значительная часть результатов теоретических исследований для одномерного случая может быть перенесена на случай функций двух и более переменных; однако при этом могут появляться практически недостаточно эффективные методы.

При теоретических построениях задачи разбивают на классы, например выделяют класс задач вычисления интегралов от функций с ограниченными производными, а затем проводят исследования, связанные с этими классами задач. Конечно, принимаемое описание не всегда (скорее даже редко) хорошо описывает класс реально встречающихся задач, однако скорости современных ЭВМ таковы, что, несмотря на грубость описания одномерных задач, удается решать большинство из них, пользуясь стандартными методами, разработанными в результате теоретических исследований.

Трудоемкость решения задач резко возрастает с ростом их размерности, и поэтому, как правило, не удается разработать стандартные методы решения широких классов многомерных задач со столь же высокой точностью, как в одномерном случае.

Несколько утешает следующее обстоятельство. Многомерные математические задачи обычно возникают из описания сложных процессов. Обычно уже эти описания являются довольно грубыми, и поэтому значительно реже предъявляется требование решения этих задач с такой же высокой точностью, как в одномерном случае. Например, требования к точности решения уравнений газовой динамики существенно ниже требований, предъявляемых к решению уравнений баллистики и небесной механики.

Не следует, однако, думать, что решение многомерных задач является почти безнадежным делом. Развитие теории методов решения многомерных задач и повышение скорости работы ЭВМ, несомненно, повлекут за

собой создание стандартных методов решения таких задач и, как следствие, снижение предъявляемых требований к квалификации пользователей. В настоящее время трудность многомерных задач требует, как правило, привлечения к их исследованию специалистов более высокой квалификации.

В настоящей главе рассмотрены вопросы интерполирования, численного интегрирования и дифференцирования функций в случае нескольких пространственных переменных.

§ 1. Метод неопределенных коэффициентов

Решение ряда многомерных задач часто сводится к решению следующих элементарных задач.

Пусть в некоторой области s-мерного пространства G заданы точки $P_1, \ldots, P_N$ и значения функции f в этих точках. Требуется:

1) получить приближение к значению функции $f(P)$;

2) получить приближение к значению некоторой производной Df функции в точке P;

3) вычислить интеграл

$$I(f) = \int_G f(P)p(P)\,dP,$$

где $p(P)$ — некоторая весовая функция.

Простейшим способом решения этих задач является многократно применявшийся нами в конкретных случаях *метод неопределенных коэффициентов*. Пусть из каких-то соображений известно, что функция $f(P)$ хорошо приближается линейными комбинациями вида

$$\sum_{j=1}^{N} B_j \omega_j(P).$$

Потребуем, чтобы такая линейная комбинация совпадала с $f(P)$ в заданных узлах, т. е. выполнялись равенства

$$\sum_{j=1}^{N} B_j \omega_j(P_q) = f(P_q), \quad q = 1, \ldots, N. \tag{1}$$

Предположим, что $\det \|\omega_j(P_q)\| \neq 0$. Тогда матрица $\|\omega_j(P_q)\|$ имеет обратную $A = \|a_{jq}\|$ и решение системы (1) записывается в виде

$$B_j = \sum_{q=1}^{N} a_{jq} f(P_q). \tag{2}$$

Функция

$$g(P) = \sum_{j=1}^{N} B_j \omega_j(P)$$

совпадает с функцией $f(P)$ в точках P_q. Подставляя в предыдущее соотношение B_j из (2), получим иное представление $g(P)$, а именно:

$$g(P) = \sum_{q=1}^{N} z_q(P) f(P_q), \qquad \text{где} \qquad z_q(P) = \sum_{j=1}^{N} a_{jq}\omega_j(P). \tag{3}$$

Такая форма записи интерполяционной функции является аналогом записи интерполяционного многочлена в форме Лагранжа. Как и в одномерном случае, можно надеяться, что при удачном выборе узлов P_q и функций $\omega_j(P)$ будет мала погрешность в приближенных равенствах:

$$f(P) \approx g(P) = \sum_{q=1}^{N} z_q(P) f(P_q), \tag{4}$$

$$Df(P) \approx Dg(P) = \sum_{q=1}^{N} Dz_q(P) f(P_q), \tag{5}$$

$$I(f) \approx I(g) = \sum_{q=1}^{N} C_q f(P_q), \tag{6}$$

где $C_q = I(z_q)$.

Как отмечалось в гл. 2 для одномерного случая, при неудачном выборе большого числа узлов интерполирования погрешность приближенного равенства (4) может оказаться катастрофически большой. Поскольку приближенные равенства (5), (6) являются следствием приближенного равенства (4), то может быть большой и погрешность в приближенных равенствах (5), (6). Поэтому применение соотношений (4)–(6) требует обоснования и выяснения условий законности использования.

§ 2. Метод наименьших квадратов и регуляризация

Как мы отмечали, применение описанного выше метода часто приводит к неудовлетворительным результатам.

Повышение качества приближения может достигаться различными способами. Рассмотрим первый из них, называемый *методом наименьших квадратов*. Перенумеруем функции $\omega_j(P)$ таким образом, чтобы меньшим значениям j соответствовали более гладкие функции. Приближение ищется в виде

$$g(P) = \sum_{j=1}^{n} B_j \omega_j(P),$$

где $n \ll N$. Параметры B_j определяются из условия

$$\min_{B_1,\dots,B_n} \Phi(g),$$

где, например,

$$\Phi(g) = \Phi(B_1, \dots, B_n) = \sum_{q=1}^{N} p_q (g(P_q) - f(P_q))^2 = \\ = \sum_{q=1}^{N} p_q \left(\sum_{j=1}^{n} B_j \omega_j(P_q) - f(P_q) \right)^2. \tag{1}$$

В основе метода наименьших квадратов лежит следующее соображение. Малость величины $\Phi(g)$ обеспечивает близость функций $g(P)$ и $f(P)$ в точках P_q; при $n \ll N$ функция $g(P)$ является линейной комбинацией относительно более гладких функций, поэтому у нее меньше возможностей отличаться от $f(P)$ вне узлов по сравнению со случаем $n = N$.

Числа $p_q > 0$, называемые весами, подбираются в зависимости от плотности распределения точек P_q. Если значения $f(P_q)$ содержат случайную погрешность, то p_q выбирают также в зависимости от дисперсии погрешностей измеряемых значений. Там, где точки P_q распределены плотнее, числа p_q берутся меньше; значениям $f(P_q)$ с большей дисперсией погрешности ставят в соответствие меньшие значения p_q. Такие рекомендации выглядят довольно неопределенными, поскольку нельзя предложить общего правила, пригодного для всех задач. Для конкретных классов задач принципы выбора p_q и $n = n(N)$ вырабатываются с учетом специфических свойств задач на основе статистических критериев и численного эксперимента.

Приравнивая нулю производные $\partial\Phi/\partial B_i$, получим систему линейных уравнений для определения B_j:

$$\frac{1}{2}\frac{\partial \Phi}{\partial B_i} = \sum_{j=1}^{n} d_{ij} B_j - d_i = 0, \tag{2}$$

$$d_{ji} = d_{ij} = \sum_{q=1}^{N} p_q \omega_i(P_q) \omega_j(P_q), \quad d_i = \sum_{q=1}^{N} p_q \omega_i(P_q) f(P_q).$$

Раскрывая скобки в (1), получим

$$\Phi(B_1, \dots, B_n) = \sum_{i,j=1}^{n} d_{ij} B_i B_j - 2 \sum_{j=1}^{n} d_j B_j + d_0,$$

где

$$d_0 = \sum_{q=1}^{N} p_q (f(P_q))^2.$$

Так как $\Phi(B_1, \ldots, B_n) \geqslant 0$, то симметричная матрица $D = \|d_{ij}\|$ неотрицательна. В связи с этим в ряде стандартных программ метода наименьших квадратов для решения системы уравнений (2) используется метод квадратного корня. Иногда целесообразно искать B_j, непосредственно минимизируя Φ каким-либо итерационным методом.

Выражая B_j из (2) через d_i, а затем через $f(P_q)$, получим

$$B_j = \sum_{q=1}^{N} a_{jq} f(P_q);$$

следовательно,

$$g(P) = \sum_{q=1}^{N} z_q(P) f(P_q), \tag{3}$$

где

$$z_q(P) = \sum_{i=1}^{n} a_{iq} \omega_i(P).$$

Воспользовавшись (3), можно получить также формулу численного дифференцирования $Df \approx Dg$ и квадратурную формулу $I(f) \approx I(g)$.

В основе *метода регуляризации* непосредственно лежат соображения о сглаживании аппроксимирующей функции. Наиболее распространенной формой метода регуляризации является следующая. Приближение отыскивается в виде

$$g(P) = \sum_{j=1}^{n} B_j \omega_j(P),$$

а коэффициенты B_j выбираются из условия минимума выражения

$$\Phi(\lambda, g) = \Phi(g) + \lambda \Psi(g), \quad \lambda > 0. \tag{4}$$

Функционал $\Psi(g)$ подбирается из следующего условия: если значение этого функционала невелико, то функция g обладает определенной гладкостью. Например, $\Psi(g)$ может быть некоторым приближением к интегралу $\int_G |\operatorname{grad} g(P)|^2 \, dP$. В приложениях часто используется случай, когда $n = N$, на котором мы далее и остановимся. Пусть минимум выражения $\Phi(\lambda, g)$ достигается при некоторых $B_1^\lambda, \ldots, B_N^\lambda$ и

$$g^\lambda(P) = \sum_{j=1}^{N} B_j^\lambda \omega_j(P).$$

Рассмотрим крайние случаи: $\lambda = 0$ и λ — очень большое число. Имеем равенство

$$\Phi(0, g) = \sum_{j=1}^{N} p_q \left(\sum_{j=1}^{N} B_j \omega_j(P_q) - f(P_q) \right)^2.$$

Если $\det\|\omega_j(P_q)\| \neq 0$, то система (1.1) имеет решение и на ее решении правая часть этого равенства обращается в нуль. В то же время выражение $\Phi(0, g)$ всегда неотрицательно. Таким образом, нижняя грань достигается на значениях B_j, являющихся решениями системы (1.1). Тогда $g^0(P)$ совпадает с интерполяционным многочленом с узлами интерполяции P_j. При больших λ в функционале $\Phi(\lambda, g)$ определяющим является второе слагаемое, нижняя грань которого достигается на гладкой функции. Следовательно, есть какие-то основания ожидать, что при промежуточных значениях λ функции $g^\lambda(P)$ будут гладкими и в то же время не очень сильно отличающимися от приближаемой функции в заданных узлах.

Наши рассуждения выглядят довольно расплывчато, однако при такой общей постановке задачи приближения функции без конкретного указания системы функций ω_j, распределения точек P_q и класса рассматриваемых функций вряд ли можно сказать что-либо определенное. Подобные грубые «физические» соображения часто помогают при конструировании новых методов решения задачи в условиях недостаточной информированности о самой задаче.

Ниже на конкретном примере будет объяснена сущность эффекта, достигаемого за счет применения регуляризации.

§ 3. Примеры регуляризации

Пусть $f(x)$ — действительная периодическая функция с периодом 1. Пусть известны значения f_q величин $f(x_q)$ при $x_q = qh$, $q = 0, \ldots, N-1$; $Nh = 1$. Погрешности приближенных значений $f_q - f(x_q) = d_q$ — независимые случайные величины с математическим ожиданием, равным нулю, и дисперсией d^2. Требуется получить таблицу приближенных значений производных $f'(x_q)$. Погрешность оценивается в норме, соответствующей скалярному произведению

$$(f, g) = \sum_{q=0}^{N-1} f_q \bar{g}_q h.$$

Далее $f(x)$ рассматривается как периодически продолженная функция.

Для нахождения производной используем простейшую формулу численного дифференцирования

$$f'(x_q) \approx Df_q = \frac{f_{q+1} - f_{q-1}}{2h}. \tag{1}$$

Предположим, что функция $f(x)$ достаточно гладкая и шаг h настолько мал, что при отсутствии погрешностей в значениях функции погрешность формулы численного дифференцирования

$$f'(x_q) \approx \frac{f(x_{q+1}) - f(x_{q-1})}{2h} \tag{2}$$

пренебрежимо мала. Тогда погрешность в значении производной представима в виде

$$R_q = \frac{f_{q+1} - f_{q-1}}{2h} - f'(x_q) = \left(\frac{f(x_{q+1}) - f(x_{q-1})}{2h} - f'(x_q)\right) + \\ + \frac{d_{q+1} - d_{q-1}}{2h} \approx \frac{d_{q+1} - d_{q-1}}{2h} = r_q.$$

Обозначая математическое ожидание знаком M, имеем равенство

$$M\|R_q\|^2 \approx M\|r_q\|^2 = M\left(\sum_{q=0}^{N-1} h\left(\frac{d_{q+1} - d_{q-1}}{2h}\right)^2\right) = \\ = \frac{N}{4}\sum_{q=0}^{N-1}\left(Md_{q+1}^2 - 2M(d_{q-1}d_{q+1}) + Md_{q-1}^2\right).$$

Соглсано указанным выше свойствам случайных величин d_q, имеем

$$Md_qd_p = d^2\delta_p^q. \tag{3}$$

Таким образом, $M\|R_q\|^2 = N^2d^2/2$, т. е. среднеквадратичное значение нормы погрешности равно $Nd/\sqrt{2}$. Мы видим, что погрешность приближенной формулы (1) возрастает с уменьшением $h = 1/N$; в то же время для малости погрешности в предположении отсутствия погрешностей в исходной информации (2) требуется достаточная малость h.

Чтобы выяснить пути уменьшения погрешности, проведем более детальное исследование. Пусть A_j^0 — дискретные коэффициенты Фурье (ДКФ) функции $f(x_q)$:

$$f(x_q) = \sum_{-N/2<j\leqslant N/2} A_j^0 \exp\{2\pi \mathbf{i} j x_q\},$$

а A_j — ДКФ функции f_q. Величины $a_j = A_j - A_j^0$ будут ДКФ функции $d_q = f_q - f(x_q)$. Имеем

$$d_q = \sum_{-N/2<j\leqslant N/2} a_j \exp\{2\pi \mathbf{i} j x_q\}, \tag{4}$$

$$a_j = h\sum_{q=0}^{N-1} d_q \exp\{-2\pi \mathbf{i} j x_q\}.$$

Чтобы вычислить ДКФ для r_q, применим оператор численного дифференцирования D к функции $\exp\{2\pi \mathbf{i} j x_q\}$:

$$D\exp\{2\pi \mathbf{i} j x_q\} = \frac{\exp\{2\pi \mathbf{i} j x_{q+1}\} - \exp\{2\pi \mathbf{i} j x_{q-1}\}}{2h} = \frac{\mathbf{i}\sin 2\pi jh}{h}\exp\{2\pi \mathbf{i} j x_q\}.$$

Применяя оператор D к обеим частям (4), получим

$$r_q = \sum_{-N/2<j\leqslant N/2} \frac{\mathbf{i}\sin 2\pi jh}{h} a_j \exp\{2\pi \mathbf{i} j x_q\}.$$

Из равенства Парсеваля

$$\|r_q\|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin(2\pi jh)}{h}\right)^2 |a_j|^2$$

следует, что

$$M\|r_q\|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin(2\pi jh)}{h}\right)^2 M|a_j|^2.$$

Имеем

$$|a_j|^2 = a_j\bar{a}_j = h^2 \sum_{q,\,p=0}^{N-1} d_q d_p \exp\{2\pi \mathbf{i} j(x_p - x_q)\};$$

мы заменили $\bar{d}_p$ на d_p, воспользовавшись тем, что d_p действительны. Таким образом,

$$M|a_j|^2 = h^2 \sum_{q,\,p=0}^{N-1} \exp\{2\pi \mathbf{i} j(x_p - x_q)\} M(d_q d_p) = h^2 \sum_{q,\,p=0}^{N-1} d^2 \delta_p^q = hd^2$$

и

$$M\|r_q\|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin(2\pi j/h)}{h}\right)^2 hd^2. \tag{5}$$

Возьмем некоторую функцию $\mu(j)$ и положим

$$f^\mu(x) = \sum_{-N/2<j\leqslant N/2} A_j^0 \mu(j) \exp\{2\pi \mathbf{i} jx\},$$

$$f_q^\mu = \sum_{-N/2<j\leqslant N/2} A_j \mu(j) \exp\{2\pi \mathbf{i} jx_q\}.$$

При заданных значениях f_q можно вычислить коэффициенты A_j, а следовательно, и величины f_q^μ. Пусть r_q^μ — составляющая погрешности формулы численного дифференцирования, являющаяся следствием погрешностей d_q:

$$r_q^\mu = Df_q^\mu - Df^\mu(x_q).$$

Аналогично (5) имеем

$$M\|r_q^\mu\|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin(2\pi jh)}{h}\right)^2 \mu^2(j) hd^2.$$

Если

$$Df(x_q) \approx Df^\mu(x_q), \qquad M\|r_q^\mu\|^2 \ll M\|r_q\|^2, \tag{6}$$

то имеет смысл положить $D(x_q) \approx Df_q^\mu$. Для выполнения первого из соотношений (6) существенна близость коэффициентов Фурье функций $f(x)$ и $f^\mu(x)$; для выполнения второго — малость $\mu(f)$ при больших j, т. е. определенная гладкость $f(x)$.

Попробуем прибегнуть к сглаживанию функции f_q при помощи метода регуляризации. Рассмотрим функционал

$$\Phi_1^h(\lambda, g) = h\sum_{q=0}^{N-1}(g_q - f_q)^2 + \lambda^2 h\sum_{q=0}^{N-1}\left(\frac{g_{q+1}-g_q}{h}\right)^2,$$

где $g_N = g_0$. Если g_q являются значениями $g(x_q)$ гладкой функции $g(x)$, то величина

$$h\sum_{q=0}^{N-1}\left(\frac{g_{q+1}-g_q}{h}\right)^2$$

стремится к интегралу $\int_0^1 (g'(x))^2\,dx$. Условие, что этот интеграл невелик, гарантирует определенную гладкость функции $g(x)$. Таким образом, есть какие-то основания считать, что функционал $\Phi_1^h(\lambda, g)$ удовлетворяет требованиям, накладываемым на функционалы метода регуляризации. Определим сеточную функцию g_q^λ из условия минимума функционала $\Phi_1^h(\lambda, g)$ и положим

$$f'(x_q) \approx (g_{q+1}^\lambda - g_{q-1}^\lambda)/(2h).$$

Посмотрим, что является аналогом такой регуляризации для случая функций непрерывного аргумента. Пусть

$$f(x) = \sum_{j=-\infty}^{\infty} A_j^0 \exp\{2\pi \mathbf{i} j x\}$$

и g_1^λ — функция, реализующая минимум функционала

$$\Phi_1(\lambda, g) = \int_0^1 (g(x) - f(x))^2\,dx + \lambda^2\int_0^1 (g'(x))^2\,dx.$$

Уравнение Эйлера для этого функционала записывается в виде

$$\lambda^2 g'' - (g - f) = 0.$$

Непосредственной проверкой убеждаемся, что функция

$$g_1^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (2\lambda\pi j)^2}\exp\{2\pi \mathbf{i} j x\}$$

является решением этого уравнения. Произведем сравнение $g_1^\lambda(x)$ и $f(x)$ для малых λ. Если $|2\lambda\pi j| \ll 1$, то коэффициенты Фурье этих функций $A_j^0/(1 + (2\lambda\pi j)^2)$ и A_j^0 близки между собой. Если $|2\lambda\pi j| \gg 1$, то коэффициенты Фурье функции $g_1^\lambda(x)$ много меньше коэффициентов Фурье функции $f(x)$. Таким образом, на языке техники регуляризация равносильна некоторой «фильтрации»: несущественно искажая гармоники с

малой частотой колебания, она сильно ослабляет гармоники с большой частотой. Если требуется еще меньше исказить амплитуды A_j^0 гармоник с малой частотой и лучше «отфильтровать» высокочастотные колебания, то можно рассмотреть функционал

$$\Phi_n(\lambda, g) = \int_0^1 (g(x) - f(x))^2\,dx + \lambda^{2n} \int_0^1 (g^{(n)}(x))^2\,dx.$$

Уравнение Эйлера для этого функционала имеет вид

$$\lambda^{2n} g^{(2n)} + (-1)^n (g - f) = 0,$$

отсюда

$$g_n^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (2\lambda\pi j)^{2n}} \exp\{2\pi \mathbf{i} j x\}.$$

Рассмотрим графики множителей $1/(1 + (2\lambda\pi j)^{2n})$; при $|j| < 1/(2\pi\lambda)$ и $n \to \infty$ эти множители стремятся к 1, и, таким образом, при больших n амплитуды соответствующих гармоник искажаются все меньше и меньше. В то же время при $|j| > 1/(2\pi\lambda)$ эти множители стремятся к 0 и, таким образом, соответствующие

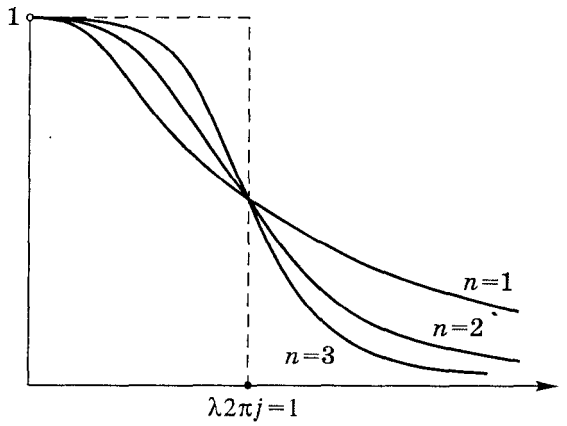


Рис. 5.3.1

гармоники умножаются на все меньшие множители (рис. 5.3.1). Таким образом, при $n \to \infty$ и $1/(2\pi\lambda)$ не целом

$$g_n^\lambda(x) \to \tilde{g}^\lambda(x) = \sum_{|j|<1/(2\pi\lambda)} A_j^0 \exp\{2\pi \mathbf{i} j x\}.$$

Вернемся к дискретному случаю. В выражении $\Phi_1^h(\lambda, g)$ величина g_q входит только в сумму слагаемых

$$h(g_q - f_q)^2 + \lambda^2 h \left(\left(\frac{g_{q+1} - g_q}{h} \right)^2 + \left(\frac{g_q - g_{q-1}}{h} \right)^2 \right).$$

Имеем систему уравнений относительно значений g_q в точке экстремума:

$$\frac{1}{2h}\frac{\partial \Phi_1^h}{\partial g_q} = (g_q - f_q) - \frac{\lambda^2}{h^2}(g_{q+1} - 2g_q + g_{q-1}) = 0. \tag{7}$$

Будем обозначать решение (7) через g_q^λ. Матрица системы уравнений (7) относительно неизвестных g_q совпадает с матрицей положительно определенной квадратичной формы

$$\sum_{q=0}^{N-1} g_q^2 + \frac{\lambda^2}{h^2}\sum_{q=0}^{N-1}(g_{q+1} - g_q)^2,$$

поэтому ее определитель отличен от нуля, и система (7) имеет, и притом единственное, решение.

Пусть

$$f_q = \sum_{-N/2<j\leqslant N/2} A_j \exp\{2\pi \mathbf{i} j x_q\}.$$

Будем отыскивать периодическую функцию g_q^λ в виде

$$g_q^\lambda = \sum_{-N/2<j\leqslant N/2} A_j^\lambda \exp\{2\pi \mathbf{i} j x_q\}.$$

Сначала вычислим оператор второй разности $\delta^2 f_q = f_{q+1} - 2f_q + f_{q-1}$ от функции $\exp\{2\pi \mathbf{i} j x_q\}$:

$$\begin{aligned}\delta^2 \exp\{2\pi \mathbf{i} j x_q\} &= \exp\{2\pi \mathbf{i} j x_{q+1}\} - 2\exp\{2\pi \mathbf{i} j x_q\} + \exp\{2\pi \mathbf{i} j x_{q-1}\} = \\ &= [\exp\{2\pi \mathbf{i} j h\} - 2 + \exp\{-2\pi \mathbf{i} j h\}]\exp\{2\pi \mathbf{i} j x_q\} = \\ &= \big(2\cos(2\pi j h) - 2\big)\exp\{2\pi \mathbf{i} j x_q\};\end{aligned}$$

отсюда получаем

$$\delta^2 \exp\{2\pi \mathbf{i} j x_q\} = -4\sin^2(\pi j h)\exp\{2\pi \mathbf{i} j x_q\}. \tag{8}$$

Подставим представление f_q и g_q в виде суммы Фурье в (7), преобразуем вторую разность $\delta^2 g_q$ с учетом (8), приведем подобные члены и получим

$$\sum_{-N/2<j\leqslant N/2}\left(A_j^\lambda - A_j + 4\lambda^2\left(\frac{\sin(\pi j h)}{h}\right)^2 A_j^\lambda\right)\exp\{2\pi \mathbf{i} j x_q\} = 0.$$

Это равенство будет удовлетворяться при

$$A_j^\lambda = A_j\left[1 + 4\lambda^2\left(\frac{\sin(\pi j h)}{h}\right)^2\right]^{-1}.$$

Таким образом,

$$g_q^\lambda = \sum_{-N/2<j\leqslant N/2} \frac{A_j}{1 + 4\lambda^2\left(\frac{\sin^2 \pi j h}{h}\right)^2}\exp\{2\pi \mathbf{i} j x_q\}.$$

Множитель $\left[1+4\lambda^2\left(\frac{\sin(\pi jh)}{h}\right)^2\right]^{-1}$ убывает с ростом j. Следовательно, регуляризация с использованием функционала $\Phi_1^h(\lambda, g)$ также приводит к ослаблению высших гармоник функции.

По аналогии с функционалами $\Phi_n(\lambda, g)$ можно осуществлять регуляризацию с помощью функционалов

$$\Phi_n^h(\lambda, g) = h\sum_{q=0}^{N-1}(g_q - f_q)^2 + \lambda^{2n}h\sum_{q=0}^{N-1}\left(\frac{\Delta^n g_q}{h^n}\right)^2.$$

Решение системы уравнений (7) и систем линейных уравнений, возникающих при минимизации функционалов $\Phi_n^h(\lambda, g)$, можно получить за $O(n^2N)$ арифметических операций методом прогонки решения периодической сеточной краевой задачи (см. гл. 9).

По сравнению с регуляризацией с помощью функционалов $\Phi_n^h(\lambda, g)$ при $n > 1$ часто бывает более удобна n-кратная регуляризация с помощью функционала $\Phi_1^h(\lambda, g)$.

§ 4. Сведение многомерных задач к одномерным

Выше рассматривались способы решения многомерных задач, не требующие дополнительной информации о распределении узлов $x_1, \ldots, x_N$, в которых известны значения функции. Такие способы применяются в случае, когда отсутствует возможность распоряжаться выбором узлов.

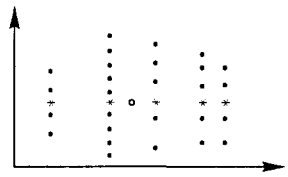


Рис. 5.4.1

Если рассматриваемая функция задана аналитически, то узлы можно выбирать по желанию. При удачном расположении узлов приближение, интерполирование, численное дифференцирование и интегрирование функций многих переменных могут быть сведены к последовательному осуществлению этих операций над функциями одной переменной. Рассмотрим случай расположения узлов, изображенных точками • на рис. 5.4.1. Здесь множество всех узлов Ω разбивается на подмножества узлов $\Omega_1, \ldots, \Omega_{m_1}$ (в данном случае $m_1 = 5$), лежащих соответственно на прямых $x_1 = x_1^1, \ldots, x_1 = x_1^{m_1}$.

Рассмотрим задачу вычисления значения некоторого оператора

$$\mathcal{D}f(P^0) = f^{(r_1, r_2)}(x_1^0, x_2^0).$$

Точка (x_1^0, x_2^0) на рис. 5.4.1 изображена символом ∘. Частный случай $r_1 = r_2 = 0$ соответствует задаче вычисления значения функции.

Возьмем какую-либо формулу вычисления производной $g^{(r_1)}(x_1^0)$ по значениям функции в узлах $x_1^{j_1}$:

$$g^{(r_1)}(x_1^0) \approx \sum_{j_1=1}^{m_1} A_{j_1} g(x_1^{j_1}); \tag{1}$$

при этом не имеется в виду, что используются значения функции g во всех точках $x_1^{j_1}$. Например, в случае численного дифференцирования речь может идти о простейшей формуле численного дифференцирования по ближайшим к x_1^0 двум узлам.

Подставляя $g(x_1) = f^{(0,r_2)}(x_1, x_2^0)$, получим

$$f^{(r_1,r_2)}(x_1^0, x_2^0) \approx \sum_{j_1=1}^{m_1} A_{j_1} f^{(0,r_2)}(x_1^{j_1}, x_2^0)$$

(точки $(x_1^{m_1}, x_2^0)$ изображены на рис. 5.4.1 символом $*$).

Задача численного дифференцирования функции двух переменных $f(x_1, x_2)$ свелась к задаче численного дифференцирования функций $f(x_1^{m_1}, x_2)$ одной переменной. Такая задача уже рассматривалась ранее.

Пусть $(x_1^m, x_2^{m_1}), \ldots, \left(x_1^m, x_2^{m,n(m)}\right)$ — узлы, образующие множество Ω_m. При каждом j_1 возьмем некоторую формулу вычисления производной $g^{(r_2)}(x_2^0)$ по значениям функции в узлах $x_2^{j_1 j_2}$:

$$g^{(r_2)}(x_2^0) \approx \sum_{j_2=1}^{n(j_1)} A_{j_1 j_2} g(x_2^{j_1 j_2}).$$

Подставляя сюда $g = f(x_1^{j_1}, x_2)$, аналогично (1) получим

$$f^{(0,r_2)}(x_1^{j_1}, x_2^0) \approx \sum_{j_2=1}^{n(j_1)} A_{j_1 j_2} f(x_1^{j_1}, x_2^{j_1 j_2}). \tag{2}$$

Воспользовавшись этими соотношениями, из (1) получаем

$$f^{(r_1,r_2)}(x_1^0, x_2^0) \approx \sum_{j_1=1}^{m_1} A_{j_1} \sum_{j_2=1}^{n(j_1)} A_{j_1 j_2} f(x_1^{j_1}, x_2^{j_1 j_2}).$$

Заметим, что формулу (2) нужно строить лишь для значений j_1 таких, что $A_{j_1} \neq 0$.

Наиболее часто узлы располагаются в узлах некоторой сетки, являющейся произведением одномерных сеток

$$\Omega = \Omega_1 \times \cdots \times \Omega_s, \tag{3}$$

иначе говоря, Ω состоит из точек $(x_1^{j_1}, \ldots, x_s^{j_s})$ при $j_1 = 1, \ldots, N_1, \ldots, j_s = 1, \ldots, N_s$. В случае расположения узлов в вершинах такой сетки та же формула численного дифференцирования получится, если поменять местами численное дифференцирование по переменным x_1 и x_2, т. е. сначала получить некоторую формулу

$$f^{(r_1, r_2)}(x_1^0, x_2^0) \approx \sum_{m_2} A_{m_2} f^{(r_1, 0)}(x_1^0, x_2^{m_2}),$$

а затем воспользоваться формулой

$$f^{(r_1, 0)}(x_1^0, x_2^{m_2}) \approx \sum_{m_1} A_{m_1 m_2} f(x_1^{m_1 m_2}, x_2^{m_2}).$$

Численное дифференцирование (интерполирование) функций большего числа переменных производится аналогично последовательным сведением к численному дифференцированию функций на единицу меньшего числа переменных.

При численном дифференцировании функций многих переменных нужно особенно следить за величиной отбрасываемых остаточных членов. Рассмотрим, например, задачу, где применение описанного выше приема последовательного численного дифференцирования может привести к получению неправильной формулы. Пусть значения функции $f(x_1, x_2)$ известны в узлах сетки $(m_1 h_1, m_2 h_2)$. Требуется вычислить значение $f_{x_1 x_2}(0, 0)$. Выпишем простейшую формулу численного дифференцирования функции одной переменной

$$f'(x) \approx \frac{f(x + h_1) - f(x)}{h_1}.$$

По формуле Тейлора

$$f(x + h) = f(x) + f'(x)h + f''(x + \theta h)\frac{h^2}{2}, \quad 0 \leqslant \theta \leqslant 1,$$

имеем

$$f'(x) = \frac{f(x + h) - f(x)}{h} - f''(x + \theta h)\frac{h}{2}.$$

Поэтому можно написать, что

$$f_{x_1 x_2}(0, 0) = \frac{f_{x_2}(h_1, 0) - f_{x_2}(0, 0)}{h_1} - f_{x_1^2 x_2}(\theta h_1, 0)\frac{h_1}{2}. \qquad (4)$$

В свою очередь, возьмем какие-либо аппроксимации производных $f_{x_2}(h_1, 0)$, $f_{x_2}(0, 0)$. Имеем равенства:

$$f_{x_2}(h_1, 0) = \frac{f(h_1, h_2) - f(h_1, 0)}{h_2} - f_{x_1 x_2}(h_1, \theta_1 h_2)\frac{h_2}{2}, \qquad (5)$$

$$f_{x_2}(0, 0) = \frac{f(0, 0) - f(0, -h_2)}{h_2} + f_{x_1 x_2}(0, \theta_2 h_2)\frac{h_2}{2}. \qquad (6)$$

После подстановки (5), (6) в (4) получим

$$f_{x_1x_2}(0,0)=\frac{f(h_1,h_2)-f(h_1,0)-f(0,0)+f(0,-h_2)}{h_1h_2}- \\ -f_{x_1^2x_2}(\theta h_1,0)\frac{h_1}{2}-\frac{1}{2}f_{x_2x_2}(0,\theta_1h_2)-\frac{1}{2}f_{x_2x_2}(h_1,\theta_2h_2). \tag{7}$$

Здесь при построении формулы численного дифференцирования одновременно учитывался остаточный член.

Соотношение (7) можно переписать в виде

$$f_{x_1x_2}(0,0)=\frac{f(h_1,h_2)-f(h_1,0)-f(0,0)+f(0,-h_2)}{h_1h_2}- \\ -f_{x_2x_2}(0,0)+O(h),\quad h=\max\{h_1,h_2\}.$$

Таким образом, если бы мы не обращали внимание на величину погрешности аппроксимации, то получили бы приближенную формулу с конечной погрешностью

$$f_{x_1x_2}(0,0)\approx\frac{f(h_1,h_2)-f(h_1,0)-f(0,0)+f(0,-h_2)}{h_1h_2},$$

в данном случае аппроксимирующую не требуемую производную, а выражение $f_{x_1x_2}(0,0)+f_{x_2x_2}(0,0)$.

Если в (5) и (6) использовать одни и те же формулы численного дифференцирования по переменной x_2, то вместо (7) получается формула с погрешностью, стремящейся к нулю при $h\to 0$. Например, вместо (6) воспользуемся равенством

$$f_{x_2}(0,0)=\frac{f(0,h_2)-f(0,0)}{h_2}-f_{x_2x_2}(0,\theta_3h_2)\frac{h_2}{2}. \tag{8}$$

После подстановки (5) и (8) в (4) получим

$$f_{x_1x_2}(0,0)=\frac{f(h_1,h_2)-f(h_1,0)-f(0,h_2)+f(0,0)}{h_1h_2}- \\ -f_{x_1^2x_2}(\theta h_1,0)\frac{h_1}{2}-\frac{1}{2}f_{x_2x_2}(h_1,\theta_1h_2)+\frac{1}{2}f_{x_2x_2}(0,\theta_3h_2).$$

Если производная $f_{x_2x_2}$ непрерывно дифференцируема в окрестности начала координат, то $\frac{1}{2}f_{x_2x_2}(0,\theta_3h_2)-\frac{1}{2}f_{x_2x_2}(h_1,\theta_1h_2)=O(h)$; поэтому

$$f_{x_1x_2}(0,0)=\frac{f(h_1,h_2)-f(h_1,0)-f(0,h_2)+f(0,0)}{h_1h_2}+O(h).$$

Мы получили, что формула численного дифференцирования

$$f_{x_1x_2}(0,0)\approx\frac{f(h_1,h_2)-f(h_1,0)-f(0,h_2)+f(0,0)}{h_1h_2} \tag{9}$$

имеет погрешность $O(h)$.

Покажем, как получить формулу (9) методом неопределенных коэффициентов. Зададимся видом формулы численного дифференцирования

$$f_{x_1x_2}(0,0) \approx l(f) = \frac{af(h_1,h_2)+bf(h_1,0)+cf(0,h_2)+df(0,0)}{h_1h_2}. \qquad (10)$$

Такой вид правой части выбран из *соображений размерности*. Пусть $[x]$ — обозначение размерности некоторой величины; например, если x — скорость, то $[x]$ = м/с. Производная $f_{x_1x_2}$ имеет размерность $[f]/([x_1]\cdot[x_2])$, функция f — размерность $[f]$, h_1 — размерность $[x_1]$, h_2 — размерность $[x_2]$. Таким образом, величины $f(P)/(h_1h_2)$ имеют ту же размерность, что и $f_{x_1x_2}$, поэтому есть основание ожидать, что в разумной формуле численного дифференцирования (10) коэффициенты a, b, c, d будут безразмерными величинами, не зависящими от шагов сетки.

Положим $R(f) = f_{x_1x_2}(0,0) - l(f)$. Выпишем разложение Тейлора для $f(x_1,x_2)$ относительно точки $(0,0)$ с точностью до членов второго порядка:

$$f(x_1,x_2) = P_2(x_1,x_2) + r(f),$$
$$P_2(x_1,x_2) = f(0,0) + x_1f_{x_1}(0,0) + x_2f_{x_2}(0,0) + \frac{1}{2}x_1^2f_{x_1x_1}(0,0) + x_1x_2f_{x_1x_2}(0,0) + \frac{1}{2}x_2^2f_{x_2x_2}(0,0).$$

В предположении, что f трижды непрерывно дифференцируема, можно показать справедливость соотношения $(r(f))_{x_1x_2}|_{(0,0)} = 0$ и $r(f) = O(h^3)$ в узлах сетки, входящих в выражение $l(f)$. Поэтому в предположении, что a, b, c, $d = O(1)$, имеем

$$R(f) = R(P_2) + R(r) = R(P_2) - O(h^3)/(h_1h_2).$$

Если $R(P_2) = 0$ и h_1, h_2 одного порядка, то $R(f) = O(h)$ и формула численного дифференцирования (10) имеет порядок $O(h)$. Выражение $R(f)$ является линейным функционалом от функции f, поэтому $R(P_2) = 0$ для любого многочлена P_2 второй степени, если

$$R(1) = R(x_1) = R(x_2) = R(x_1^2) = R(x_1x_2) = R(x_2^2) = 0.$$

Получаем систему уравнений

$$R(1) = -(a+b+c+d)/(h_1h_2) = 0,$$
$$R(x_1) = -(d+b)/h_2 = 0,$$
$$R(x_2) = -(d+c)/h_1 = 0,$$
$$R(x_1^2) = -(d+b)h_1/h_2 = 0,$$
$$R(x_2^2) = -(d+c)h_2/h_1 = 0,$$
$$R(x_1x_2) = 1-d = 0.$$

Эта система шести линейных уравнений с четырьмя неизвестными имеет решение $a = d = 1$, $b = c = -1$, соответствующее приближенной формуле (9).

Точно таким же образом строятся формулы численного интегрирования. Пусть вычисляется интеграл

$$I_0 = \int_G f(y_1, \ldots, y_s)\, dy_1 \ldots dy_s.$$

Его можно представить в виде

$$I_0 = \int_{G_0} I_1(y_1)\, dy_1, \quad I_1(y_1) = \int_{G_1(y_1)} I_2(y_1, y_2)\, dy_2,$$

$$I_2(y_1, y_2) = \int_{G_2(y_1, y_2)} I_2(y_1, y_2, y_3)\, dy_3, \ldots,$$

$$I_{s-1}(y_1, \ldots, y_{s-1}) = \int_{G_{s-1}(y_1, \ldots, y_{s-1})} I_s(y_1, \ldots, y_s)\, dy_s.$$

Здесь $I_s(y_1, \ldots, y_s) = f(y_1, \ldots, y_s)$, G_0 — проекция области G на ось y_1, $G_k(\bar{y}_1, \ldots, \bar{y}_k)$ — сечение множества G плоскостью $y_1 = \bar{y}_1, \ldots, y_k = \bar{y}_k$. Воспользуемся какой-либо квадратурной формулой для вычисления первого из этих интегралов

$$I_0 \approx \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}).$$

Задача вычисления исходного интеграла свелась к вычислению интегралов размерности, на единицу меньшей. Полагаем теперь

$$\begin{aligned} I_1(y_1^{m_1}) &\approx \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2}), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I_{s-1}(y_1^{m_1}, \ldots, y_{s-1}^{m_1 \ldots m_{s-1}}) &\approx \sum_{m_s=1}^{n_s(m_1, \ldots, m_{s-1})} D^{m_1 \ldots m_s} \times \\ &\qquad \times f(y_1^{m_1}, \ldots, y_s^{m_1 \ldots m_s}). \end{aligned} \tag{11}$$

В итоге получаем

$$I_0 \approx \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \ldots \ldots \sum_{m_s=1}^{n_s(m_1, \ldots, m_{s-1})} D^{m_1 \ldots m_s} f(y_1^{m_1}, \ldots, y_s^{m_1 \ldots m_s}). \tag{12}$$

В одномерном случае для всех операций численного анализа были получены оценки погрешности через производные рассматриваемой функции. Посмотрим, какие оценки погрешности являются следствием этих оценок для многомерных задач. Значение некоторого оператора от функции приближается значениями других операторов, причем погрешность этой замены можно оценить согласно формулам оценки погрешности одномерных формул. Однако эти новые значения сами уже содержат погрешности, поэтому в суммарную погрешность эти погрешности войдут с некоторыми множителями.

Обратимся к задаче интегрирования. Пусть

$$I_{k-1}(y_1^{m_1}, \ldots, y_{k-1}^{m_1 \ldots m_{k-1}}) = \\ = \sum_{m_k=1}^{n_k(m_1,\ldots,m_{k-1})} D^{m_1 \ldots m_k} I_k(y_1^{m_1}, \ldots, y_k^{m_1 \ldots m_k}) + E^{k-1}_{m_1 \ldots m_{k-1}}.$$

Последовательно подставляя в равенство

$$I_0 = \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) + E^0$$

выражения $I_1, I_2, \ldots$, получаем цепочку соотношений

$$I_0 = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) = \\ = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} \left(E^1_{m_1} + \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2}) \right) = \cdots \\ \cdots = R + \sum_{m_1=1}^{n_1} D^{m_1} \left(\sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \times \right. \\ \left. \times \left(\cdots \left(\sum_{m_s=1}^{n_s(m_1,\ldots,m_{s-1})} D^{m_1 \ldots m_s} f(y_1^{m_1}, \ldots, y_s^{m_1 \ldots m_s}) \right) \cdots \right) \right),$$

где

$$R = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} E^1_{m_1} + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} E^2_{m_1 m_2} + \cdots \\ \cdots + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \cdots \sum_{m_{s-1}=1}^{n_{s-1}(m_1,\ldots,m_{s-2})} D^{m_1 \ldots m_{s-1}} E^{s-1}_{m_1 \ldots m_{s-2}}. \quad (13)$$

Из последнего равенства видно, что погрешность аппроксимации может оказаться существенно больше, чем в одномерном случае, если коэффициенты $D^{m_1 \ldots m_k}$ — большие.

Формально сведение многомерной задачи к одномерной имеет одинаковый характер и в случае задачи численного дифференцирования (интерполирования), и в случае численного интегрирования. Однако между этими задачами есть такое существенное различие: задача численного дифференцирования (интерполирования) чаще ставится как задача нахождения оператора от функции по значениям на некоторой заданной совокупности узлов Ω. Для задачи интегрирования более типичной является возможность распоряжаться выбором узлов.

При осуществлении многомерных операций численного дифференцирования (интерполирования и интегрирования) функций, заданных на сет-

ке, являющейся произведением одномерных сеток, разумно использовать одинаковые формулы для аппроксимации промежуточных величин. Имеется в виду, что, например, формулы (11) должны иметь вид

$$I_{k-1}(y_1^{m_1}, \ldots, y_{k-1}^{m_{k-1}}) \approx \sum_{m_k=1}^{N_k} D_{k,N_k}^{m_k} I_k(y_1^{m_1}, \ldots, y_k^{m_k}), \tag{14}$$

т. е. $D_{k,N_k}^{m_k}$ и $y_k^{m_k}$ зависят только от m_k. Правая часть (12) тогда приобретает вид

$$\sum_{m_1=1}^{N_1} \ldots \sum_{m_s=1}^{N_s} D_{1,N_1}^{m_1} \ldots D_{s,N_s}^{m_s} f(y_1^{m_1}, \ldots, y_s^{m_s}).$$

При построении такой квадратуры мы неявно предполагаем, что область интегрирования — прямоугольный параллелепипед с ребрами, параллельными координатным осям.

Такие формулы численного интегрирования (интерполирования, дифференцирования) называют *прямым произведением* соответствующих одномерных формул численного интегрирования (интерполирования, дифференцирования).

В § 11 будет приведен более сложный пример прямого произведения квадратуры по отрезку на квадратуру по сфере. В случае применения таких аппроксимаций так же, как при вычислении производной $f_{x_1x_2}$ по формуле (9), оказывается, что некоторые составляющие погрешности аппроксимации компенсируются.

В случае задачи интегрирования возможна такая ситуация: при гладкой подынтегральной функции может оказаться, что промежуточные интегралы $I_k(y_1, \ldots, y_k)$ не обладают достаточной гладкостью.

Пусть вычисляется интеграл по единичному кругу

$$I = \int_{-1}^{1} \left(\int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} f(y_1, y_2)\, dy_2 \right) dy_1$$

при гладкой функции $f(y_1, y_2)$. Если $f(\pm 1, 0) \neq 0$, то функция

$$I_1(y_1) = \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} f(y_1, y_2)\, dy_2$$

имеет неограниченные производные в точках ± 1 и поэтому при численном интегрировании по переменной y_1 следует использовать специальные приемы вычисления интегралов от таких функций — переменный шаг интегрирования, в частности интегрирование с автоматическим выбором шага, выделение особенностей и т. д. Целесообразнее записать этот интеграл в виде

$$I = \int_0^1 r I_1(r)\, dr, \qquad I_1(r) = \int_0^{2\pi} f(r\cos\varphi, r\sin\varphi)\, d\varphi,$$

где все подынтегральные функции уже гладкие. Подынтегральная функция внутреннего интеграла периодическая, поэтому имеет смысл применять квадратуру (3.5.7)

$$\int_0^{2\pi} g(\varphi)d\varphi \approx \frac{2\pi}{n}\sum_{m=0}^{n-1} g\left(\frac{2\pi m}{n}\right).$$

Задача 1. Функция задана в узлах сетки $(m_1h_1,\, m_2h_2)$. Построить формулы с погрешностями аппроксимации $O(h_1^2+h_2^2)$, $O(h_1^4+h_2^4)$ для вычисления значения $f_{x_2}(h_1/2,\, 0)$.

§ 5. Интерполяция функций в треугольнике

При решении уравнений в частных производных вариационно-разностными методами возникает следующая задача. Имеется некоторый треугольник Δ, каждая сторона которого разбита на l равных частей, и через точки разбиения проведены прямые L_q, параллельные сторонам треугольника. Стороны треугольника также будем относить к множеству прямых L_q. Обозначим через Ω множество, состоящее из точек пересечения этих прямых, лежащих в замкнутом треугольнике Δ. (Таким образом, Ω включает также точки разбиения сторон треугольника и вершины треугольника.) Число таких точек равно $n = 1+2+\cdots+(l+1) = (l+1)(l+2)/2$. Будем обозначать их через $Q_1(x_1^1,\, x_2^1),\ldots,\, Q_n(x_1^n,\, x_2^n)$.

Ставится задача построения многочлена степени l

$$P(x_1,\, x_2) = \sum_{m_1+m_2\leqslant l} a_{m_1m_2}x_1^{m_1}x_2^{m_2},$$

принимающего в этих точках $Q_j(x_1^j,\, x_2^j)$ заданные значения

$$P(x_1^j,\, x_2^j) = f_j, \quad j = 1,\ldots,\, n. \tag{1}$$

Число неизвестных коэффициентов $a_{m_1m_2}$ также равно n, и, таким образом, соотношения (1) образуют систему n уравнений с n неизвестными. Если система (1) разрешима, то из нее могут быть найдены коэффициенты $a_{m_1m_2}$. Для их нахождения не обязательно прибегать к описанному выше варианту метода неопределенных коэффициентов, а можно выписать искомый многочлен $P(x)$ в явном виде.

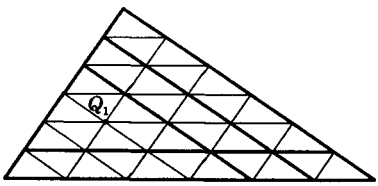


Рис. 5.5.1

Возьмем некоторую фиксированную точку Q_1. Можно показать, что среди прямых L_q имеется ровно l прямых, удовлетворяющих следующему условию. Существует не более одной вершины треугольника такой, что Q_1 и эта вершина лежат по одну сторону от такой прямой. При этом оказывается, что каждая точка из Ω, отличная от Q_1, лежит на одной из таких прямых. На рис. 5.5.1 эти прямые обозначены жирными линиями.

Пусть $L_{j,1}(x_1, x_2) = 0, \ldots, L_{j,l}(x_1, x_2) = 0$ — уравнения этих прямых. Функция

$$\varphi_j(x_1, x_2) = \prod_{i=1}^{l} \frac{L_{j,i}(x_1, x_2)}{L_{j,i}(x_1^j, x_2^j)}$$

является многочленом степени l, равна 1 в точке Q_1 и 0 в остальных точках Q_i. Поэтому многочлен степени l

$$P(x_1, x_2) = \sum_{j=1}^{n} f_j \varphi_j(x_1, x_2)$$

будет искомым.

В случае, когда $f_j = f(x_1^j, x_2^j)$ при всех j, многочлен $P(x_1, x_2)$ будет интерполяционным многочленом по отношению к функции $f(x_1, x_2)$.

Задача 1. Показать, что значения многочлена $P(x_1, x_2)$ на каждой из сторон треугольника зависят от значений f_j, соответствующих точкам Q_j этой стороны.

Задача 2. Пусть H — это длина максимальной из сторон треугольника Δ, $f_j = f(Q_j)$, f — некоторая гладкая функция,

$$M_{l+1} = \max_{r_1+r_2=l+1} \max_{\Delta} \left| f^{(r_1, r_2)}(x_1, x_2) \right|,$$

α — наименьший из углов треугольника. Получить оценку

$$\max_{\Delta} |f(x_1, x_2) - P(x_1, x_2)| \leqslant C(l, \alpha) M_{l+1} H^{l+1},$$

где C — постоянная, зависящая только от l и α.

Задача 3. При выполнении условия задачи 2 для $0 < r_1 + r_2 \leqslant l + 1$ получить оценку

$$\max_{\Delta} \left| f^{(r_1, r_2)}(x_1, x_2) - P^{(r_1, r_2)}(x_1, x_2) \right| \leqslant C(l, r_1, r_2, \alpha) M_{l+1} H^{l+1-r_1-r_2}.$$

Задача 4. Исследовать поведение постоянных $C(l, \alpha)$, $C(l, r_1, r_2, \alpha)$ при $\alpha \to 0$ в задачах 2 и 3 соответственно.

Описанный способ интерполяции широко используется при приближении функций двух переменных. Область G, где приближается функция, разбивается на треугольники с достаточно малой максимальной стороной. В каждом треугольнике функция приближается соответствующим интерполяционным многочленом $P(x_1, x_2)$ степени l. Если разбиение устроено так, что вершина одного треугольника не может быть внутренней точкой стороны другого треугольника, то полученная таким образом приближающая функция будет непрерывна в G (справедливость последнего утверждения следует из решения задачи 1).

В связи с неограниченным ростом постоянных $C(l, r_1, r_2, \alpha)$ при $\alpha \to 0$ (см. задачу 4) разумное разбиение области G не должно содержать треугольников Δ с очень малыми углами.

Задача 5. Построить аналог описанного выше способа интерполяции для задачи интерполяции функции в тетраэдре.

Замечание. В случае, когда две стороны треугольника или соответственно три ребра тетраэдра направлены вдоль координатных осей, интерполирующий многочлен может быть явно записан также с помощью аппарата разделенных разностей для функций многих переменных.

§ 6. Оценка погрешности численного интегрирования на равномерной сетке

Произведем конкретную оценку погрешности численного интегрирования в случае двукратного интеграла от дважды непрерывно дифференцируемой функции

$$I = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1 dx_2.$$

Запишем исходный интеграл как повторный:

$$I = \int_0^1 I_1(x_1)\, dx_1, \tag{1}$$

где

$$I_1(x_1) = \int_0^1 f(x_1, x_2)\, dx_2.$$

Обозначим

$$A_k = \max_{0 \leqslant x_1, x_2 \leqslant 1} \left| \frac{\partial^2 f}{\partial x_k^2} \right|.$$

Применим для вычисления интеграла (1) составную формулу трапеций с постоянным шагом разбиения $H_1 = 1/N_1$:

$$I \approx S_{N_1} = \frac{I_1(0) + I_1(1)}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{I_1(j_1/N_1)}{N_1}.$$

Поскольку

$$\frac{\partial^2 I_1(x_1)}{\partial x_1^2} = \int_0^1 \frac{\partial^2 f(x_1, x_2)}{\partial x_1^2}\, dx_2,$$

то

$$\left| \frac{\partial^2 I_1(x_1)}{\partial x_1^2} \right| \leqslant A_1.$$

Поэтому при $E^0 = I - S_{N_1}$, согласно оценке погрешности составной формулы трапеций (3.8.8), имеем

$$|E^0| \leqslant A_1/(12N_1^2). \tag{2}$$

Для вычисления интегралов $I_1(j_1/N_1)$ применим составную формулу трапеций; имеем

$$I_1\left(\frac{j_1}{N_1}\right) \approx S_{N_2}\left(\frac{j_1}{N_1}\right) = \frac{f(j_1/N_1,\,0) + f(j_1/N_1,\,1)}{2N_2} + \sum_{j_2=1}^{N_2-1} \frac{f(j_1/N_1,\,j_2/N_2)}{N_2}.$$

Для погрешности

$$E^1_{j_1} = I_1(j_1/N_1) - S_{N_2}(j_1/N_1)$$

согласно (3.8.8) имеем

$$|E^1_{j_1}| \leqslant A_2/(12N_2^2). \tag{3}$$

Подставляя $I_1(j_1/N_1) = E^1_{j_1} + S_{N_2}(j_1/N_1)$ в равенство $I = E^0 + S_{N_1}$, получим цепочку соотношений

$$I = E^0 + \frac{(S_{N_2}(0) + E^1_0) + (S_{N_2}(1) + E^1_{N_1})}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{S_{N_2}(j_1/N_1) + E^1_{j_1}}{N_1} = S_{N_1N_2} + R,$$

где

$$\begin{aligned} S_{N_1N_2} &= \frac{S_{N_2}(0) + S_{N_2}(1)}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{S_{N_2}(j_1/N_1)}{N_1}, \\ R &= E^0 + \frac{E^1_0 + E^1_{N_1}}{2N_1} + \sum_{j_1=1}^{N_1-1} \frac{E^1_{j_1}}{N_1}. \end{aligned} \tag{4}$$

Выражение $S_{N_1N_2}$ является квадратурной суммой, вычисляемой по значениям функции f в $(N_1+1)(N_2+1)$ точках $(j_1/N_1,\, j_2/N_2)$, R — погрешность численного интегрирования.

Из (4) следует оценка

$$|R| \leqslant |E^0| + \max_{0 \leqslant j_1 \leqslant N_1} |E^1_{j_1}|,$$

и вследствие (2), (3) имеем

$$|R| \leqslant A_1/(12N_1^2) + A_2/(12N_2^2).$$

При $N_1 = N_2 = n$ получаем, что

$$|R| \leqslant (A_1 + A_2)/(12n^2), \tag{5}$$

и по отношению к общему числу узлов интегрирования $N = (N_1 + 1) \times (N_2 + 1) = (n + 1)^2$ погрешность имеет порядок $O(N^{-1})$.

Предположим, что требуется гарантия, чтобы погрешность не превосходила ε. Для этого достаточно выполнения неравенства

$$A_1/(12N_1^2) + A_2/(12N_2^2) \leqslant \varepsilon. \tag{6}$$

Если A_1 и A_2 отличаются несущественно, то можно взять $N_1 = N_2 = n$ и исходя из оценки (5) определить минимальное n из условия

$$(A_1 + A_2)/(12n^2) \leqslant \varepsilon.$$

Если же A_1 и A_2 различаются существенно, то имеет смысл затратить время на минимизацию вычислительной работы, т. е. искать N_1 и N_2, минимизирующие $(N_1 + 1)(N_2 + 1)$ при условии

$$A_1/(12N_1^2) + A_2/(12N_2^2) \leqslant \varepsilon.$$

Обозначим через $C_{\mathbf{r}}(\mathbf{A}) = C_{r_1,\ldots,r_s}(A_1, \ldots, A_s)$ класс функций, у которых в рассматриваемой области определения производные $f_{x_k^{r_k-1}}$, $k = 1, \ldots, s$, непрерывны, а производные $f_{x_1^{r_1}\ldots x_s^{r_s}}(x_1, \ldots, x_s)$ кусочно-непрерывны и удовлетворяют условиям $|f_{x_k^{r_k}}(x_1, \ldots, x_s)| \leqslant A_k$.

Задача 1. Пусть для вычисления интеграла

$$\int_0^1 \ldots \int_0^1 f(x_1, \ldots, x_s)\, dx_1 \ldots dx_s$$

применяются составные формулы точности $O(N_k^{-r_k})$ по каждой оси, где N_k — число узлов, соответствующее направлению x_k. Получить оценку погрешности

$$O\left(\sum_1^k A_k N_k^{-r_k}\right). \tag{7}$$

Минимизируя оценку (7) при заданном общем числе узлов $N_1, \ldots, N_s = N$, получить оценку погрешности $O(N^{-r})$, $1/r = 1/r_1 + \cdots + 1/r_s$.

Рассмотрим частный случай: $r_1 = \cdots = r_s = r_0$, $A_1 = \cdots = A_s = A_0$. Тогда $1/r = 1/r_1 + \cdots + 1/r_s = s/r_0$. Таким образом, рассматриваемые квадратуры обеспечивают оценку погрешности $O(A_0 N^{-r_0/s})$. У реальных подынтегральных функций порядок ограниченных производных r часто оказывается не очень большим, поэтому при больших s, как показывает полученная оценка, скорость сходимости оказывается плохой.

Возникает вопрос об оптимальных квадратурах на классах многомерных функций. Так как ни для каких реальных классов функций такие квадратуры неизвестны, ограничимся оценкой снизу погрешности оптимальных квадратур.

§ 7. Оценка снизу погрешности численного интегрирования

Напомним постановку задачи оптимизации квадратурных формул на классе функций. Пусть приближенное значение интеграла вычисляется по формуле

$$I(f) = \int_G f(P)p(P)dP \approx S_N(f) = \sum_{j=1}^{N} D_j f(P_j).$$

Величина

$$R_N(f) = I(f) - S_N(f)$$

называется *погрешностью квадратуры*, величина

$$R_N(F) = \sup_{f \in F} |R_N(f)|$$

— *погрешностью квадратуры на классе функций* F, величина

$$W_N(F) = \inf_{D_j,\, P_j} R_N(F)$$

— *оптимальной оценкой погрешности квадратур на классе* F; квадратура (если такая существует), на которой эта нижняя грань достигается, называется *оптимальной*. Мы будем предполагать выполненным условие: существует некоторый куб $\Delta \in G$, в котором $p(P) \geqslant \gamma > 0$.

Теорема. $W_N(C_{\mathbf{r}}(\mathbf{A})) \geqslant d(\Delta, \mathbf{r}, \mathbf{A})\gamma N^{-r}$, где $d(\Delta, \mathbf{r}, \mathbf{A}) > 0$, $r = (r_1^{-1} + \cdots + r_s^{-1})^{-1}$.

Доказательство теоремы осуществляется следующим образом. Будет показано, что для любой совокупности узлов $P_1, \ldots, P_N$ можно построить функцию $f_{P_1,\ldots,P_N}(P)$ рассматриваемого класса $C_{\mathbf{r}}(\mathbf{A})$, обращающуюся в нуль во всех этих узлах и такую, что $I(f) \geqslant d(\Delta, \mathbf{r}, \mathbf{A})\gamma N^{-r}$. При этом постоянная $d > 0$ не зависит от точек $P_1, \ldots, P_N$. Тогда для любой квадратуры с этими узлами

$$R_N(C_{\mathbf{r}}(\mathbf{A})) \geqslant \left| I(f_{P_1,\ldots,P_N}) - \sum_{j=1}^{N} D_j f_{P_1,\ldots,P_N}(P_j) \right| =$$

$$= |I(f_{P_1,\ldots,P_N})| \geqslant d_1(\Delta, \mathbf{r}, \mathbf{A})\gamma N^{-r}.$$

Величина $R_N(C_{\mathbf{r}}(\mathbf{A}))$ оценена снизу постоянной, не зависящей от узлов квадратур P_j и весов D_j, поэтому и ее нижняя грань $W_N(C_{\mathbf{r}}(\mathbf{A}))$ по множеству всевозможных квадратур также оценивается снизу этой постоянной $d(\Delta, \mathbf{r}, \mathbf{A})\gamma N^{-r}$. Таким образом, доказательство теоремы сводится к построению соответствующей функции $f_{P_1,\ldots,P_N}(P)$ для каждой совокупности точек $P_1, \ldots, P_N$.

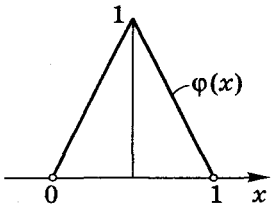


Рис. 5.7.1

Построение такой функции $f_{P_1,\dots,P_N}(P)$ и доказательство утверждения теоремы для простоты будут проведены для случая $\mathbf{r} = (1,\dots,1)$, $A_1 = \dots = A_s = A_0$, т. е. для класса непрерывных функций с кусочно-непрерывными производными $\partial f/\partial x_j$, ограниченными по модулю постоянной A_0.

Для простоты выкладок будем проводить построение для случая, когда Δ — единичный куб $0 \leqslant x_i \leqslant 1$, $i = 1,\dots,s$. Положим $n = [(2N)^{1/s}] + 1$ и разобьем куб Δ на n^s прямоугольных параллелепипедов $\Pi_{n_1,\dots,n_s}$: $(n_s - 1)/n \leqslant x_k \leqslant n_k/n$, $0 < n_k \leqslant n$, $k = 1,\dots,s$. Пусть (см. рис. 5.7.1)

$$\varphi(x) = \begin{cases} 0 & \text{при } x \leqslant 0 \text{ или } x \geqslant 1, \\ 1 - 2|x - 0{,}5| & \text{при } 0 \leqslant x \leqslant 1. \end{cases}$$

Из определения n следует, что $(2N)^{1/s} \leqslant n \leqslant 2(2N)^{1/s}$, поэтому

$$2N \leqslant n^s 2^{s+1} N. \tag{1}$$

Построим функцию $f_{P_1,\dots,P_N}(P) = f_0(P)$ следующим образом. В тех параллелепипедах $\Pi_{n_1,\dots,n_s}$, которые не содержат внутри ни одной из точек $P_1,\dots,P_s$, положим

$$f_0(P) = \frac{A_0}{2n} \prod_{k=1}^{s} \varphi(nx_k - (n_k - 1)). \tag{2}$$

Во всех остальных параллелипипедах положим $f_0(P) = 0$.

Функция $f_0(P)$ согласно своему определению непрерывна в каждом параллелепипеде $\Pi_{n_1,\dots,n_s}$ и обращается в нуль на его границе, поэтому она непрерывна в Δ. В точках, где функция $f_0(P)$ дифференцируема и отлична от 0, имеем

$$\frac{\partial f_0}{\partial x_1} = \frac{A_0}{2} \varphi'(nx_1 - (n_1 - 1)) \prod_{k=1}^{s} \varphi(nx_k - (n_k - 1)).$$

Поскольку $|\varphi| \leqslant 1$, $|\varphi'| \leqslant 2$, то $|\partial f_0/\partial x_1| \leqslant A_0$. Точно так же получаем оценки $|\partial f_0/\partial x_k| \leqslant A_0$, $k = 1, 2,\dots, s$. Согласно построению функции $f_0(P)$ производные $\partial f_0/\partial x_k$ могут иметь разрывы лишь в точках плоскостей $x_k = m/(2n)$, где m — целое; поэтому функция $f_0(P)$ принадлежит рассматриваемому классу. Из ее построения также следует, что она обращается в нуль со всеми производными во всех узлах $P_1,\dots,P_N$.

Оценим снизу значение $I(f_0)$. Воспользовавшись неравенством $p(P) \geqslant \gamma$, после замены переменных $x_k = (n_k - 1 + y_k)/n$ получим цепочку

соотношений

$$\int_{\Pi_{n_1,\dots,n_s}} p(x_1,\dots,x_s)\,\frac{A_0}{2n}\prod_{k=1}^{s}\varphi(nx_k-(n_k-1))\,dx_1\dots dx_s \geqslant$$

$$\geqslant \frac{A_0\gamma}{2n^{s+1}}\int_0^1\dots\int_0^1\prod_{k=1}^{s}\varphi(y_k)dy_1\dots dy_k = \frac{d_1A_0\gamma}{n^{s+1}};$$

здесь

$$d_1 = \frac{1}{2}\left(\int_0^1\varphi(y)\,dy\right)^s = \frac{1}{2^{s+1}}. \tag{3}$$

Каждая точка P_j может находиться внутри только одного из параллелепипедов $\Pi_{n_1,\dots,n_s}$. Следовательно, по крайней мере в n^s-N параллелепипедах $\Pi_{n_1,\dots,n_s}$ функция $f_0(P)$ отлична от тождественного нуля и определяется равенством (2). Согласно (1) имеем $N\leqslant n^s/2$, поэтому таких параллелепипедов не менее чем $n^s/2$. С учетом (3) получаем оценку

$$I(f_0)\geqslant\frac{n^s}{2}\,\frac{d_1A_0\gamma}{n^{s+1}} = \frac{d_1A_0\gamma}{2n}\geqslant\frac{d_1A_0\gamma}{2\cdot2\cdot(2N)^{1/s}} = \frac{d_2A_0\gamma}{N^{1/s}}, \tag{4}$$

где $d_2 = d_1/(4\cdot2^{1/s})$.

Таким образом, построенная функция принадлежит рассматриваемому классу и для нее выполняется неравенство (4) с постоянной d_2, не зависящей от узлов интегрирования, она обращается в нуль во всех узлах P_j. Следовательно, способ построения требуемой функции для любой совокупности узлов $P_1,\dots,P_N$ указан.

Выше производилась оценка снизу погрешности квадратур, т. е. формул интегрирования

$$I(f)\approx\sum_{j=1}^{N}D_jf(P_j),$$

где узлы интегрирования P_j и веса D_j не зависят от конкретной подынтегральной функции. Для многих задач практически более эффективными являются способы интегрирования, где последующие узлы интегрирования выбираются в зависимости от информации об уже вычисленных значениях функции, например способы интегрирования с автоматическим выбором шага.

Пусть имеется какой-то способ интегрирования, где информация о подынтегральной функции учитывается лишь в виде информации о ее значениях в отдельных точках. Этот способ определяется заданием первого узла интегрирования, правила, по которому отыскиваются следующие узлы, и способа вычисления приближенного значения интеграла. Таким образом, всякий такой способ укладывается в следующую схему: задаются некоторый узел P_1, функции

$$Q_q=\Phi_q(Q_1,\dots,Q_{q-1};\,y_1,\dots,y_{q-1}),\quad q=2,\dots,N,$$

определяющие выбор следующих узлов интегрирования в зависимости от ранее накопленной информации о подынтегральной функции, и функция

$$S_N(Q_1, \ldots, Q_N; y_1, \ldots, y_N).$$

Здесь Q_i — точки области G, y_j — числа. При приближенном вычислении конкретного интеграла последовательно вычисляются величины

$$f(P_1), \quad P_2 = \Phi_2\Big(P_1; f(P_1)\Big), \quad f(P_2),$$

$$P_3 = \Phi_3\Big(P_1, P_2; f(P_1), f(P_2)\Big), \quad f(P_3), \ldots, f(P_N)$$

и затем полагают

$$I(f) \approx S_N\Big(P_1, \ldots, P_N; f(P_1), \ldots, f(P_N)\Big).$$

Поскольку точка P_1 задается вместе с заданием функций Φ_j и S_N, зависимость этих функций от P_1 можно было бы опустить.

Так же, как в случае квадратурных формул, можно определить погрешность метода при вычислении данного интеграла

$$\tilde{R}_N(f) = I(f) - S\Big(P_1, \ldots, P_N; f(P_1), \ldots, f(P_N)\Big),$$

погрешность метода на классе

$$\tilde{R}_N(F) = \sup_{f \in F} \Big|\tilde{R}_N(f)\Big|$$

и оптимальную оценку погрешности на множестве всевозможных методов интегрирования

$$Z_N(F) = \inf_{P_1, \Phi_2, \ldots, \Phi_N, S_N} \tilde{R}_N(F).$$

Теорема (без доказательства). *Пусть класс функций F — выпуклый центрально симметричный компакт с центром симметрии $f \equiv 0$ и все функции этого класса равномерно ограничены. Тогда оптимальные оценки погрешности на множествах всевозможных квадратурных формул и всевозможных способов интегрирования совпадают:*

$$W_N(F) = Z_N(F).$$

Условие *выпуклости класса* F означает, что вместе с любыми $f_1, f_2 \in F$ этому классу принадлежат также все функции $f = \theta f_1 + (1-\theta) f_2$ при $0 < \theta < 1$. Условие, что класс является *центрально симметричным* с центром симметрии $f \equiv 0$, означает, что вместе со всякой функцией f этому классу принадлежит также функция $-f$.

В частности, все классы $C_r(A)$ удовлетворяют условиям теоремы и может показаться, что рассмотрение более широкого множества способов интегрирования не представляет интереса. Однако такой вывод нельзя признать правильным. Методы интегрирования с выбором узлов интегрирования в зависимости от полученной в процессе вычислений информации продемонстрировали на практике свою высокую эффективность

по сравнению с квадратурными формулами с заранее фиксированными узлами интегрирования. Поэтому более правильным будет вывод о том, что практически встречающиеся задачи более точно описываются некоторыми невыпуклыми классами функций. Например, типичным классом функций, встречающихся в приложениях, является класс кусочно-аналитических функций. Обратимся к одномерному случаю. В случае выпуклости класса функций полусумма двух функций класса также принадлежит этому классу. Полусумма функций, имеющих l точек, где нарушается аналитичность, может иметь $2l$ точек с нарушением аналитичности. Таким образом, класс аналитических функций, имеющих не более заданного числа l точек нарушения аналитичности, не является выпуклым. Если число точек нарушения аналитичности не ограничено сверху, то класс функций не является замкнутым: предел последовательности кусочно-аналитических функций с неограниченно растущим числом точек нарушения аналитичности может оказаться функцией, не являющейся кусочно-аналитической.

Выше производилось сравнение методов интегрирования по верхней грани погрешности на классе функций. Однако возможна следующая ситуация. Два метода имеют одинаковые погрешности на классе функций, в то же время на большинстве функций класса один из методов имеет меньшую погрешность. Ясно, что этот метод является более предпочтительным и сравнение методов по верхней грани погрешности на классе функций в данном случае не дает общей картины.

Из сказанного выше вытекает актуальность решения следующих задач.

Задача 1. Как правильно описать класс реально встречающихся функций?

Задача 2. Как правильно ввести меру в пространстве реально встречающихся подынтегральных функций? (Ни про одно из известных определений меры в пространстве функций нельзя сказать, что оно правильно описыает обстановку, характерную для приложений.)

Приведем примеры некоторых алгоритмов интегрирования с выбором узлов в зависимости от ранее полученной информации. Многомерный интеграл записывается как повторный

$$I(f) = \int_{a_0}^{b_0} I_1(x_1)\,dx_1,$$

$$I_1(x_1) = \int_{a_1(x_1)}^{b_1(x_1)} I_2(x_1,\, x_2)\,dx_2,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$I_{s-1}(x_1, \dots, x_{s-1}) = \int_{a_{s-1}(x_1,\dots,x_{s-1})}^{b_{s-1}(x_1,\dots,x_{s-1})} f(x_1, \dots x_s)\,dx_s,$$

и численное интегрирование по некоторым из переменных x_j производится посредством одномерных алгоритмов интегрирования с автоматическим выбором шага.

Следующий алгоритм вычисления многомерных интегралов имеет другую структуру. Среди известных алгоритмов интегрирования с автоматическим выбором узлов интегрирования этот алгоритм является наиболее эффективным по отношению к задаче вычисления интегралов от функций с особенностями функции или ее производных в изолированных точках.

Пусть вычисляется интеграл

$$\int_\Omega f(X)\,dX, \quad \Omega = [a_1 \leqslant x_1 \leqslant b_1, \ldots, a_s \leqslant x_s \leqslant b_s], \quad X = (x_1, \ldots, x_s).$$

Заменой переменных

$$x_i = 0{,}5(a_i + b_i) + 0{,}5(a_i - b_i)t_i, \quad i = 1, \ldots, s,$$

интеграл превращается в интеграл по кубу

$$\int_\Omega f(X)\,dX = \int_G g(t)\,dt,$$
$$G = [-1 \leqslant t_1 \leqslant 1, \ldots, -1 \leqslant t_s \leqslant 1], \quad t = (t_1, \ldots, t_s).$$

В основу метода положены кубатурные формулы

$$\int_G g(t)\,dt \approx Q_q^s(g), \quad q = 1, 2, 3.$$

Для вычисления интеграла $\int_G g(t)\,dt$ вычисляются $Q_1^s(g)$ и $Q_2^s(g)$ и проверяется условие

$$\left|Q_1^s(g) - Q_2^s(g)\right| \leqslant \varepsilon. \tag{5}$$

Здесь ε — некоторая условная мера погрешности. Если это условие выполнено, то за приближенное значение интеграла по G принимается значение, вычисленное по кубатурной формуле Q_3^s, обычно являющейся линейной комбинацией формул Q_1^s и Q_2^s. Если условие (5) не выполняется, то куб G разбивается на 2^s равных кубов и описанный алгоритм применяется к каждому из этих кубов. Процесс дробления продолжается до тех пор, пока условие (5) не будет выполнено. Если при делении шага h пополам наступит такой момент, когда h^s станет машинным нулем, то счет прекращается. Используемые в стандартных программах кубатурные формулы Q_q^s имеют следующий вид.

I. $s = 2$:

$$Q_1^2 = A_1 \sum_{|i|+|j|=1} g(i\alpha,\, j\alpha) + A_2 \sum_{|i|+|j|=1} g(i\beta,\, j\beta) + A_3 \sum_{|i|,\,|j|=1} g(i\gamma,\, j\gamma),$$

$$Q_2^2 = B_1 \sum_{|i|+|j|=1} g(i\alpha,\, j\alpha) + B_2 \sum_{|i|+|j|=1} g(i\beta,\, j\beta) + B_3 \sum_{|i|,\,|j|=1} g(i\gamma,\, j\gamma) + B_4 \sum_{|i|,\,|j|=1} g(i\nu,\, j\nu),$$

где

$$\alpha = 0{,}658149897623035910, \quad \beta = 0{,}549119831921783496,$$
$$\gamma = 0{,}894427190999915878, \quad \nu = 0{,}316227766016837933,$$
$$A_1 = 1{,}06136206790541224, \quad A_2 = -0{,}234973179016523356,$$
$$A_3 = 0{,}173611111111111111,$$
$$B_1 = 3{,}99942795838189963, \quad B_2 = -6{,}29803906949301074,$$
$$B_3 = 0{,}124007936507936507, \quad B_4 = 3{,}17460317460317460,$$
$$Q_3^2 = Q_2^2.$$

Формула Q_1^2 точна для всех многочленов степени не больше 5.
Формула Q_2^2 точна для всех многочленов степени не больше 7.

II. $s = 3$:

$$Q_1^3 = \frac{8}{225}\left(44g(0,\,0,\,0) + \frac{121}{8}\sum_{|i|,\,|j|,\,|k|=1} g\left(i\sqrt{\frac{5}{11}},\, j\sqrt{\frac{5}{11}},\, k\sqrt{\frac{5}{11}}\right) + 10\sum_{|i|+|j|+|k|=1} g(i,\,j,\,k)\right),$$

$$Q_2^3 = \frac{8}{1125}\left(-\frac{1552}{5}g(0,\,0,\,0) + \frac{1573}{40}\sum_{|i|,\,|j|,\,|k|=1} g\left(i\sqrt{\frac{5}{11}},\, j\sqrt{\frac{5}{11}},\, k\sqrt{\frac{5}{11}}\right) + \frac{784}{5}\sum_{|i|+|j|+|k|=1} g\left(i\sqrt{\frac{5}{14}},\, j\sqrt{\frac{5}{14}},\, k\sqrt{\frac{5}{14}}\right) + 15T_h\right),$$

$$T_h = \sum_{|i|,\,|j|=1} g(i,\,j,\,0) + \sum_{|i|,\,|k|=1} g(i,\,0,\,k) + \sum_{|j|,\,|k|=1} g(0,\,j,\,k),$$

$$Q_3^3 = \frac{4}{9}Q_1^3 + \frac{5}{9}Q_2^3.$$

Формулы Q_1^3 и Q_2^3 точны для всех многочленов степени не больше 5.
Формула Q_3^3 точна для всех многочленов степени не больше 7.

§ 8. Метод Монте-Карло

При построении квадратурных формул вместе с формулой, как правило, получалась оценка ее погрешности на некотором классе функций. Например, для одномерной формулы трапеций была получена оценка погрешности вида $\mathrm{const}\cdot A_2 N^{-2}$ на классе функций со второй производной, ограниченной по модулю постоянной A_2; здесь N — число узлов интегрирования. Такого рода оценки погрешности называют *гарантированными оценками погрешности на классе функций.* На основании этой оценки можно гарантировать, что погрешность приближенного значения интеграла не превосходит определенной величины для всех подынтегральных функций из рассматриваемого класса. Оценивая погрешность метода как оценку на классе функций, при оценке погрешности конкретного интеграла мы ориентируемся на величину, получающуюся в случае интегрирования «худшей» функции рассматриваемого класса. Для ряда классов функций эта оценка погрешности на классе настолько плоха, что не дает никакой надежды на получение приближенного значения интеграла с требуемой точностью. Например, согласно теоремам из § 7, не существует методов с оценкой погрешности на классе функций $C_{1,\dots,1}(A,\dots,A)$, лучшей, чем $dAN^{-1/s}$ (здесь s — кратность вычисляемого интеграла).

Предположим, что требуется гарантировать оценку погрешности, меньшую чем $0{,}01dA$. Тогда число узлов N должно удовлетворять неравенству $dAN^{-1/s} \leqslant 0{,}01dA$, т. е. должно выполняться неравенство $N \geqslant 100^s$. Поскольку вычисление каждого значения подынтегральной функции требует обычно большого числа арифметических операций, уже при $s = 6$ такое требование на число узлов оказывается практически невыполнимым.

Мы оказались в положении, когда на основе указанной выше оценки погрешности нет возможности вычислить значение интеграла с гарантированной оценкой погрешности $0{,}01dA$, поскольку это потребует непомерных затрат времени ЭВМ. Один из способов разрешения создавшегося противоречия состоит в более детальном описании классов подынтегральных функций. Другим выходом из создавшейся обстановки является отказ от получения строгой гарантированной оценки погрешности и получение оценки погрешности лишь с определенной степенью достоверности. В частности, при конструировании методов интегрирования в гл. 3 и в § 7 мы шли по пути отказа от строгой оценки погрешности: погрешность оценивалась через разность результатов вычисления приближенного значения интеграла при различных методах интегрирования.

Одним из методов приближенного вычисления значений интегралов, при котором погрешность оценивается не гарантированно, а лишь с некоторой степенью достоверности, является *метод Монте-Карло.*

Пусть требуется вычислить приближенное значение интеграла

$$I(f) = \int_G f(P)\,dP.$$

Для упрощения выкладок предполагаем, что $\mu(G)$ — мера области G равна 1. Как правило, это условие бывает выполнено, поскольку при практической реализации метода Монте-Карло область интегрирования обычно преобразуется в единичный куб. Предположим, что каким-то образом удалось получить N случайных попарно независимых точек $P_1, \ldots, P_N$, равномерно распределенных в G. Далее через $M(s)$ будем обозначать математическое ожидание случайной величины s, а через $D(s)$ — ее дисперсию. Случайные величины $s_j = f(P_j)$ попарно независимы и одинаково распределены, причем

$$M(s_j) = \int_G f(P)\,dP = I(f)$$

и

$$D(s_j) = M(s_j^2) - (M(s_j))^2 = D(f),$$

где

$$D(f) = I(f^2) - (I(f))^2.$$

Положим

$$S_N(f) = \frac{1}{N}\sum_{j=1}^{N} s_j.$$

Вследствие указанных свойств величин s_j имеем

$$M(S_N(f)) = \frac{1}{N}\sum_{j=1}^{N} M(s_j) = I(f),$$

$$D(S_N(f)) = \frac{1}{N^2}\sum_{j=1}^{N} D(s_j) = \frac{1}{N} D(f).$$

С вероятностью $1-\eta$ выполняется неравенство (*неравенство Чебышева*)

$$|S_N(f) - I(f)| \leqslant \sqrt{D(f)/(\eta N)}. \tag{1}$$

Полагая $\eta = 0{,}01$, получаем: с вероятностью 99% выполняется неравенство

$$|S_N(f) - I(f)| \leqslant 10\sqrt{D(f)/N}.$$

Еще лучшая оценка получается в предположении, что точки P_j не только попарно независимы, но и независимы в совокупности. Тогда, согласно центральной предельной теореме, случайная величина

$$(S_N(f) - I(f))/\sqrt{D(f)/N}$$

распределена асимптотически нормально с функцией распределения

$$\Phi(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Вероятность того, что случайная величина с такой функцией распределения не превосходит по модулю значение $y > 0$, асимптотически равна

$$p_0(y) = 1 - \frac{\sqrt{2}}{\sqrt{\pi}} \int_y^\infty \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Таким образом, при больших N с вероятностью, близкой к $p_0(y)$, выполняется неравенство

$$|S_N(f) - I(f)| \leqslant y\sqrt{D(f)/N}.$$

Полагая $y = 3$ и $y = 5$, получаем, что неравенства

$$|S_N(f) - I(f)| \leqslant 3\sqrt{D(f)/N} \quad \text{и} \quad |S_N(f) - I(f)| \leqslant 5\sqrt{D(f)/N}$$

выполняются соответственно с вероятностями 0,997 и 0,99999. Сформулированные выше утверждения называются иногда правилами «трех сигм» и «пяти сигм» соответственно.

В правой части всех этих оценок стоит неизвестная величина $D(f) = I(f^2) - (I(f))^2$, которую можно оценить на основании информации о вычисленных значениях $f(P_j)$.

Задача 1. Показать, что

$$D(f) = \frac{N}{N-1} M\Big(s_j(f) - S_N(f)\Big)^2. \tag{2}$$

Поскольку на основании закона больших чисел с большой вероятностью выполняется приближенное равенство

$$M\Big(s_j(f) - S_N(f)\Big)^2 \approx \frac{1}{N} \sum_{j=1}^N \Big(s_j(f) - S_N(f)\Big)^2,$$

то с большой вероятностью выполняется и приближенное равенство

$$D(f) \approx D^{(N)}(f),$$

где

$$D^{(N)}(f) = \frac{1}{N-1} \sum_{j=1}^N \Big(s_j(f) - S_N(f)\Big)^2.$$

Приведем схему применения метода Монте-Карло при $\mu(G) = 1$. Пусть $(1-\eta)$ — уровень надежности, с которым желательно получить приближенное значение интеграла, ε — заданная точность. Определяют y из равенства

$$\eta = \frac{\sqrt{2}}{\sqrt{\pi}} \int_y^\infty \exp\left\{-\frac{t^2}{2}\right\} dt.$$

Последовательно при $n = 1, \ldots$ получают случайные точки P_n и вычисляют величины $t_n(f)$, $S_n(f)$, $d_n(f)$, $D_n(f)$, пользуясь рекуррентными соотношениями

$$t_n(f) = t_{n-1}(f) + f(P_n), \quad S_n(f) = t_n(f)/n,$$

$$d_n(f) = d_{n-1}(f) + \frac{n}{n-1}\Big(f(P_n) - S_n(f)\Big)^2, \quad D_n(f) = d_n(f)/(n-1),$$

а также вычисляют значение

$$\lambda_n = y\sqrt{D_n(f)/n}.$$

Начальные условия рекурсии:

$$t_1(f) = S_1(f) = f(P_1), \quad d_1(f) = D_1(f) = 0.$$

Если оказалось, что $\lambda_n \leqslant \varepsilon$, то вычисления прекращаются; полагают $I(f) \approx S_N(f)$ и считают, что с вероятностью $1 - \eta$ выполняется неравенство $|I(f) - S_N(f)| \leqslant \varepsilon$.

Заметим, что в действительности неравенство $|I(f) - S_N(f)| \leqslant \varepsilon$ выполняется с вероятностью, несколько меньшей чем $1 - \eta$, хотя и близкой к ней.

Задача 2. Проверить, что $D_n(f) = D^{(n)}(f)$.

Для уменьшения погрешности метода Монте-Карло случайные узлы интегрирования берутся распределенными не равномерно, а с некоторой плотностью $p(P) \neq 1$, $\int_G p(P)\,dP = 1$. В этом случае полагают

$$I(f) \approx S_N(f) = \frac{1}{N}\sum_{j=1}^{N} f(P_j)/p(P_j).$$

Задача 3. Показать, что

$$M(S_N(f)) = I(f), \qquad D(f) = I(f^2/p) - (I(f))^2.$$

Убедиться в том, что переход к такому выбору случайных величин особенно целесообразен в случае $f(P)/p(P) \approx \text{const}$.

Часто в руководствах по использованию метода Монте-Карло говорится следующее. Метод Монте-Карло является универсальным методом вычисления интегралов высокой кратности. Порядок оценки скорости сходимости метода Монте-Карло есть $O(1/\sqrt{N})$ и не зависит от размерности интеграла, в то время как порядок гарантированных оценок скорости сходимости существенно ухудшается с ростом размерности. Для метода Монте-Карло при каждом n имеется эффективная оценка погрешности через величину $\sqrt{D_n(f)/n}$, и, таким образом, вычисления прекращаются именно в тот момент, когда достигнута требуемая точность.

Отнесемся осторожно к этому заявлению и рассмотрим различные соображения «за» и «против» метода Монте-Карло.

§ 9. Обсуждение правомерности использования недетерминированных методов решения задач

Часть пользователей предубеждена против метода Монте-Карло и отрицает правомерность его использования, поскольку малость погрешности метода обеспечивается лишь с некоторой вероятностью.

Выше мы уже обрисовали картину, складывающуюся при вычислении интегралов большой кратности, как почти полностью безнадежную в случае, если ставится цель получения приближенного значения интеграла с гарантированно малой оценкой погрешности. Такая обстановка и вызвала к жизни применение метода Монте-Карло.

Уже при вычислении однократных интегралов гарантия малости погрешности метода может быть получена только при использовании строгих теоретических оценок. Применение таких оценок требует высокой математической квалификации исследователя, затрат его умственного труда и не может быть перепоручена ЭВМ. Таким образом, ориентировка на методы вычисления интегралов с гарантированной оценкой погрешности противоречит общей тенденции использования ЭВМ.

Кроме того, при решении всякой задачи возможны ошибки в постановке задачи, в программе и т. д. В силу этих и ряда других причин редко можно дать стопроцентную гарантию малости погрешности результата расчета по отношению к реальной модели; некоторая вероятность ошибочности результатов вычислений имеется в любом случае. Все это подчеркивает, что полный отказ от метода Монте-Карло только из-за его вероятностной природы является неоправданным.

С другой стороны, при использовании метода Монте-Карло нужно учитывать следующие отрицательные эффекты.

Для применения метода Монте-Карло необходимо иметь в распоряжении последовательность независимых точек P_j с заданным законом распределения. Обычно пользователь располагает датчиками случайных или так называемых псевдослучайных чисел, которые выдают последовательности случайных чисел, равномерно распределенных на отрезке $[0, 1]$. При помощи преобразований таких случайных величин получаются случайные числа с заданным законом распределения. На первых ЭВМ датчиками случайных чисел были некоторые приборы, например использующие явление радиоактивного распада, которые выдавали последовательности случайных величин, иногда даже удовлетворяющие требованию независимости в совокупности. Однако такие приборы работают с малой скоростью, и поэтому с увеличением производительности ЭВМ от них отказались. Вместо датчиков случайных чисел используют датчики псевдослучайных чисел — некоторые программы, выдающие последовательности чисел, которые рекомендуется рассматривать как случайные. Использование датчиков псевдослучайных чисел явилось прогрессивным

шагом, позволившим широко применять вероятностные методы. Однако при использовании этих датчиков нужно всегда иметь в виду, какими свойствами обладают последовательности чисел, выдаваемые этими датчиками. Например, некоторые датчики псевдослучайных чисел вырабатывают последовательности чисел, которые можно рассматривать лишь как попарно независимые, а не как независимые в совокупности. В этом случае будет неправомерно пользоваться оценками погрешности, основанными на центральной предельной теореме.

Пусть методом Монте-Карло вычисляется интеграл

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s)\, dx_1 \dots dx_s.$$

Преположим, что в качестве узлов интегрирования мы хотим выбрать последовательность независимых равномерно распределенных точек единичного куба. Если датчик псевдослучайных чисел выдает последовательность $\xi_1, \dots$ чисел, равномерно распределенных на отрезке $[0, 1]$, то можно попытаться в качестве узлов интегрирования взять точки $(\xi_1, \dots, \xi_s)$, $(\xi_{s+1}, \dots, \xi_{2s}), \dots$.

Для законности применения неравенства Чебышева нужно выполнение предположения о независимости распределения любых точек P_j, т. е. независимости распределения совокупностей

$$(\xi_{(j-1)s+1}, \xi_{js}),\ (\xi_{(i-1)s+1}, \dots, \xi_{is}).$$

При увеличении s это условие накладывает все более жесткие требования на датчики псевдослучайных чисел. Известно много реальных примеров неудачного применения метода Монте-Карло в случае больших s, вызванных следующей причиной. При использовании метода и оценке погрешности делались допущения о тех или иных статистических свойствах псевдослучайных чисел, в то время как эти предположения на самом деле не выполнялись. В результате делался вывод о малости значения погрешности, который на самом деле не был справедлив.

Таким образом, опасность применения метода Монте-Карло заключается по большей части не в вероятностном характере оценки погрешности, а в том, что вероятностная оценка погрешности производится зачастую в предположении о свойствах датчиков случайных чисел, которые на самом деле не имеют места.

Задача 1. Пусть ξ_1, ξ_2 — случайные независимые величины, равномерно распределенные на отрезке $[0, 1]$. Пусть $\{y\}$ — дробная доля числа y, т. е. $\{y\} = y - [y]$, где $[y]$ — целая часть y. Положим $\xi_n = \{\xi_1 + (n-1)(\xi_2 - \xi_1)\}$. Показать, что точки ξ_n попарно независимы, равномерно распределены на $[0, 1]$ и, согласно построениям предыдущего параграфа,

$$I(f) = \int_0^1 f(x)\, dx \approx S_N(f) = \frac{1}{N} \sum_{j=1}^{N} f(\xi_j).$$

Задача 2. Пусть вычисляется двукратный интеграл

$$I(f) = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1 dx_2.$$

Показать, что при $P_n = (\xi_{2n-1}, \xi_{2n})$, где ξ_k из задачи 1, справедливо соотношение $M(S_N(f)) = I(f)$. Вычислить

$$\lim_{N\to\infty} D(S_N(f)) = d(f)$$

и убедиться, что, как правило, $d(f) \neq 0$ и поэтому $S_N(f)$ далеко от $I(f)$.

Задача 3. Пусть вычисляется трехкратный интеграл

$$I(f) = \int_0^1 \int_0^1 \int_0^1 f(x_1, x_2, x_3)\, dx_1 dx_2 dx_3.$$

При $P_n = (\xi_{3n-2}, \xi_{3n-1}, \xi_{3n})$ вычислить $M(S_n(f)) = i(f)$ и убедиться, что, как правило, $i(f) \neq I(f)$ и поэтому $S_N(f)$ далеко от $I(f)$.

Указание. Воспользоваться тем, что $(\xi_{3n} - 2\xi_{3n-1} + \xi_{3n-2})$ — целое в пределах $[-1, 1]$.

Пусть $\xi_1^k, \ldots, \xi_s^k$, $k = 1, \ldots, l$ — случайные независимые в совокупности равномерно распределенные на $[0, 1]$ случайные величины и $P_k^0 = (\xi_1^k, \ldots, \xi_s^k)$, $k = 1, \ldots, l$. Положим $\xi_j^n = \{\xi_j^1 + C_{n-1}^1 \Delta\xi_j^{1'} + \cdots + C_{n-1}^{l-1}\Delta^{l-1}\xi_j^1\}$; здесь $\{y\} = y - [y]$ — дробная часть числа y, $\Delta^q\xi_j^1 = \xi_j^{q+1} - C_q^1\xi_j^q + \cdots + (-1)^q C_q^q \xi_j^1$ — конечная разность q-го порядка, C_q^p — число сочетаний из q по p, равное нулю при $p > q$.

Пусть $P_n = (\xi_1^n, \ldots, \xi_s^n)$.

Задача 4. Проверить, что

$$P_n = P_n^0 \quad \text{при} \quad n = 1, \ldots, l.$$

Задача 5. Показать, что точки $P_1, \ldots, P_N$ равномерно распределены в единичном кубе и любые l из них независимы в совокупности.

К числу достоинств метода Монте-Карло относят независимость порядка оценки от размерности вычисляемого интеграла. Однако рассуждая только о порядке сходимости метода, можно не заметить следующую немаловажную деталь. Мы получали оценки погрешности приближенного значения интеграла вида

$$|S_N(f) - I(f)| \leqslant \text{const} \cdot \sqrt{D(f)/N}.$$

Типичным для практики является требование малости относительной погрешности приближенного значения интеграла, что в данном случае означает требование малости величины $\sqrt{D(f)}/(|I(f)|\sqrt{N})$. Статистика реально предъявляемых к вычислению интегралов показывает, что величина $\sqrt{D(f)}/|I(f)|$ имеет тенденцию к резкому росту с ростом размерности интегралов. В качестве иллюстрации приведем интеграл

$$I(f_s) = \int_0^1 \dots \int_0^1 \exp\{-32(x_1^2 + \dots + x_s^2)\}\, dx_1 \dots dx_s,$$

для которого $\sqrt{D(f_s)}/|I(f_s)| > 10^{s/2} - 1$.

Следовательно, практическая трудоемкость метода Монте-Карло существенно возрастает с ростом размерности интегралов (при одинаковой относительной погрешности). При действительном вычислении многократных интегралов методом Монте-Карло перед непосредственным применением метода зачастую с целью уменьшения величины $\sqrt{D(f)}/|I(f)|$ проводится довольно кропотливое исследование свойств подынтегральной функции, преобразование интегралов с помощью замен переменных и других приемов, требующие достаточно высокой квалификации исследователя.

§ 10. Ускорение сходимости метода Монте-Карло

Рассмотрим некоторые приемы повышения практической эффективности метода Монте-Карло.

1. Функция $f(P)$ представляется в виде

$$f(P) = F(P) + g(P),$$

где функция $F(P)$ интегрируется явно и содержит в себе все резко меняющиеся компоненты $f(P)$, а $g(P)$ — плавно меняющаяся функция с небольшой дисперсией $D(g)$. Иногда область интегрирования разбивается на малые подобласти и в каждой части в качестве $F(P)$ берут некоторый интерполяционный полином с узлами в этой подобласти.

2. Подходящий подбор плотности распределения узлов $p(P)$ (см. задачу 3 в § 8) также приводит к уменьшению дисперсии. Мы рассматривали случай, когда все узлы P_j имеют одинаковую функцию распределения $p(P)$. В ряде случаев оказывается целесообразным выбирать узлы интегрирования таким образом, чтобы каждый имел своею функцию распределения $p_j(P)$.

3. Следующий прием является частным случаем приемов 1 и 2. Исходный интеграл представляется в виде суммы интегралов

$$I(f) = \int_G f(P)\, dP = \sum_{l=1}^{n} I_l(f), I_l(f) = \int_{G_l} f(P)\, dP,$$

число узлов интегрирования N представляется в виде $N = N_1 + \cdots + N_n$, и каждый интеграл $I_l(f)$ вычисляется по методу Монте-Карло с N_l узлами интегрирования. Обратимся к случаю вычисления интеграла от функции, изображенной на рис. 5.10.1. При непосредственном вычислении исходного интеграла в случае $p(P) \equiv 1$ имеем $D(f) = 1/4$. Если $G_1 = [0, 1/2]$, $G_2 = [1/2, 1]$, то при любых $N_1, N_2 \neq 0$ оба интеграла $I_l(f)$, а следовательно, и исходный интеграл будут вычислены точно.

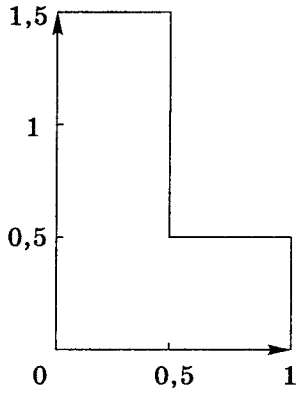


Рис. 5.10.1

Конечно, случай, когда исходную область интегрирования удается разбить на части, где подынтегральная функция постоянна, очень редкий. Однако, если все-таки удается разбить ее на части, где функция меняется мало, то можно получить существенное увеличение точности при том же объеме вычислений.

Разбиение области интегрирования на части с целью уменьшения дисперсии метода Монте-Карло широко используется, в частности, при обработке естественно-научной информации. Пусть требуется определить водосодержание снега в бассейне некоторой реки. При непосредственном применении метода Монте-Карло выбиралось бы с равномерной плотностью распределения несколько точек, где производилось бы измерение количества водосодержания на единице поверхности. Участки поверхности с однородными природными условиями (высота над уровнем моря, уровень облачности, залесненность, ориентация склонов гор, осадки, господствующее направление ветра) характеризуются примерно одинаковым водосодержанием. Поэтому удается добиться существенного повышения точности, разбивая бассейн на части с однородными условиями и применяя метод Монте-Карло для вычисления интегралов по этим частям.

В случае гладкой подынтегральной функции разбиение области интегрирования на части приводит к увеличению порядка скорости сходимости. Пусть вычисляется интеграл

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s)\, dx_1 \dots dx_s.$$

Положим $N = n^s$ и разобьем исходную область интегрирования на равные кубы $\Pi_{n_1,\dots,n_s}$: $(n_1 - 1)/n \leqslant x_1 \leqslant n_1/n, \dots, (n_s - 1)/n \leqslant x_s \leqslant n_s/n$. В каждом кубе выберем случайную точку $P_{n_1,\dots,n_s}$. Считаем, что ее плотность распределения — постоянная, равная n^s, и случайные точки в любых двух кубах выбираются независимо. Положим

$$\bar{S}_N(f) = \frac{1}{n^s} \sum_{n_1,\dots,n_s=1}^{n} f(P_{n_1,\dots,n_s}).$$

Задача 1. Доказать, что $D(\bar{S}_N(f)) \leqslant D(S_N(f)) = D(f)/N$.

Проведем оценку дисперсии в предположении, что функция $f(P)$ удовлетворяет условию Липшица с постоянной A по каждой из переменных. Справедливо равенство

$$D(\bar{S}_N(f)) = \sum_{n_1,\ldots,n_s=1} D\left(\frac{1}{n^s} f(P_{n_1,\ldots,n_s})\right) = \sum_{n_1,\ldots,n_s=1}^{n} \frac{1}{n^{2s}} D(f(P_{n_1,\ldots,n_s})). \quad (1)$$

Имеем равенство

$$M(f(P_{n_1,\ldots,n_s})) = \sigma_{n_1,\ldots,n_s} = \int_{\Pi_{n_1,\ldots,n_s}} n^s f(P)\, dP.$$

На основании теоремы о среднем имеем $\sigma_{n_1,\ldots,n_s} = f(\bar{P}_{n_1,\ldots,n_s})$, где $\bar{P}_{n_1,\ldots,n_s} \in \Pi_{n_1,\ldots,n_s}$, поэтому

$$M(f(P_{n_1,\ldots,n_s})) = f(\bar{P}_{n_1,\ldots,n_s}).$$

В то же время

$$\begin{aligned} f(x_1+\Delta_1,\ldots,x_s+\Delta_s) - f(x_1,\ldots,x_s) = \\ = \big(f(x_1+\Delta_1,\ldots,x_s+\Delta_s) - f(x_1+\Delta_1,\ldots,x_{s-1}+\Delta_{s-1},x_s)\big) + \\ + \big(f(x_1+\Delta_1,\ldots,x_{s-1}+\Delta_{s-1},x_s) - f(x_1+\Delta_1,\ldots,x_{s-2}+ \\ + \Delta_{s-2},x_{s-1},x_s)\big) + \cdots + \big(f(x_1+\Delta_1,x_2,\ldots,x_s) - f(x_1,\ldots,x_s)\big). \end{aligned}$$

Отсюда следует, что для функций рассматриваемого класса

$$\big|f(x_1+\Delta_1,\ldots,x_s+\Delta_s) - f(x_1,\ldots,x_s)\big| \leqslant A\Big(|\Delta_1| + \cdots + |\Delta_s|\Big).$$

Если точка $P_{n_1,\ldots,n_s}$ принадлежит $\Pi_{n_1,\ldots,n_s}$, то каждая из ее координат отличается от соответствующей координаты точки $\bar{P}_{n_1,\ldots,n_s}$ не более чем на n^{-1}. Поэтому из последнего неравенства следует оценка

$$\big|f(P_{n_1,\ldots,n_s}) - f(\bar{P}_{n_1,\ldots,n_s})\big| \leqslant Asn^{-1}$$

при $P_{n_1,\ldots,n_s} \in \Pi_{n_1,\ldots,n_s}$, и, следовательно,

$$|f(P_{n_1,\ldots,n_s}) - \sigma_{n_1,\ldots,n_s}| \leqslant Asn^{-1}. \quad (2)$$

Всякая случайная величина ξ удовлетворяет неравенству

$$|M(\xi)| \leqslant \sup|\xi|.$$

Следовательно,

$$\begin{aligned} D(f(P_{n_1,\ldots,n_s})) = M(f(P_{n_1,\ldots,n_s}) - \sigma)^2 \leqslant \\ \leqslant \sup_{P_{n_1,\ldots,n_s} \in \Pi_{n_1,\ldots,n_s}} \big(f(P_{n_1,\ldots,n_s}) - \sigma_{n_1,\ldots,n_s}\big)^2 \leqslant \left(\frac{As}{n}\right)^2. \end{aligned}$$

С помощью этой оценки заключаем, что правая часть равенства (1) не превосходит величины

$$n^s n^{-2s}(As/n)^2 = (As)^2/n^{s+2}.$$

Обозначив общее число узлов n^s через N, получаем оценку

$$D(\bar{S}_N(f)) \leqslant A^2 s^2/N^{1+2/s}. \tag{3}$$

Отсюда на основании неравенства Чебышева (8.1) заключаем, что с вероятностью $1-\eta$ выполняется неравенство

$$\left|\bar{S}_N(f) - I(f)\right| \leqslant AsN^{-1/s}/\sqrt{\eta N}.$$

Полученная оценка погрешности по порядку лучше, чем оценка погрешности $O(1/\sqrt{N})$ метода Монте-Карло.

Мы получили оценку погрешности по вероятности. Оказывается, что для рассматриваемого метода можно получить и гарантированную оценку погрешности. Умножая (2) на n^{-s}, получаем неравенство

$$\left|\frac{1}{n^s} f(P_{n_1,\ldots,n_s}) - \int_{\Pi_{n_1,\ldots,n_s}} f(P)\,dP\right| \leqslant As/n^{s+1}.$$

Величина $\bar{S}_N(f) - I(f)$ может быть представлена в виде суммы таких слагаемых:

$$\sum_{n_1,\ldots,n_s=1}^{n} \left(\frac{1}{n^s} f(P_{n_1,\ldots,n_s}) - \int_{\Pi_{n_1,\ldots,n_s}} f(P)\,dP\right).$$

Суммируя оценки для этих слагаемых, получим

$$\left|\bar{S}_N(f) - I(f)\right| \leqslant As/n = As/N^{1/s}.$$

Сопоставляя эту оценку с теоремой из § 7, заключаем, что рассматриваемый метод имеет гарантированную оценку погрешности, оптимальную по порядку на рассматриваемом классе функций.

Возникает вопрос, можно ли улучшить на этом классе оценку дисперсии (3). Метод Монте-Карло и другие способы интегрирования, подобные рассматриваемому, где приближенное значение интеграла зависит от некоторых случайных параметров, называют *недетерминированными*. Пусть $S_N(f)$ — приближенное значение интеграла $I(f)$, получаемое при применении некоторого недетерминированного способа вычисления.

Теорема (без доказательства). *Существуют* $d_1(\mathbf{r}, \mathbf{A})$, $d_2 > 0$, *удовлетворяющие следующему соотношению. Для любого способа вычисления интеграла, где информация о подынтегральной функции используется лишь как информация об ее значениях в* N *точках, найдется функция* $f \in C_{\mathbf{r}}(\mathbf{A})$, *для которой*

$$|S_N(f) - I(f)| > d_1(\mathbf{r}, \mathbf{A})/N^{r+1/2} \tag{4}$$

с вероятностью d_2, где $1/r = 1/r_1 + \cdots + 1/r_s$, а $S_N(f)$ — приближенное значение интеграла.

Следствием неравенства (4) является неравенство

$$\sqrt{D(S_N(f))} \geqslant d_3(\mathbf{r}, \mathbf{A})/N^{r+1/2}, \tag{5}$$

которое, в частности, означает, что оценка (3) не может быть улучшена по порядку.

Теорема (без доказательства). *Можно указать способы интегрирования, для которых имеет место гарантированная оценка погрешности*

$$|S_N(f) - I(f)| \leqslant d_4(\mathbf{r}, \mathbf{A})/N^r \tag{6}$$

и одновременно

$$M(S_N(f)) = I(f), \quad \sqrt{D(S_N(f))} \leqslant d_5(\mathbf{r}, \mathbf{A})/N^{r+1/2} \tag{7}$$

для всех $f \in C_{\mathbf{r}}(\mathbf{A})$.

Мы построили выше такой способ при $r_1 = \cdots = r_s = 1$. Идея построения таких способов состоит в разбиении исходной области интегрирования на малые части и вычислении интеграла по малой части при помощи некоторой «случайной квадратурной формулы».

Задача 2. Пусть $P_{n_1,\ldots,n_s}$ — случайная точка в кубе $\Pi_{n_1,\ldots,n_s}$ (при тех же условиях на распределение и независимость точек $P_{n_1,\ldots,n_s}$, как в задаче 1). Обозначим через $P^*_{n_1,\ldots,n_s}$ точку, симметричную $P_{n_1,\ldots,n_s}$ относительно центра куба $\Pi_{n_1,\ldots,n_s}$, и положим

$$s_{n_1,\ldots,n_s}(f) = \frac{1}{2n^s}(f(P_{n_1,\ldots,n_s}) + f(P^*_{n_1,\ldots,n_s})), \quad S_N(f) = \sum_{n_1,\ldots,n_s=1}^{n} s_{n_1,\ldots,n_s}(f).$$

Доказать, что для функции из класса $C_{2,\ldots,2}(\mathbf{A})$ выполнены одновременно оценки (6), (7). (Заметим, что здесь $r = 2/s$.)

§ 11. О выборе метода решения задачи

Обсудим схему решения одной задачи, где оказалось выгодным использование методов Монте-Карло, обсуждавшихся в предыдущем параграфе. Требовалось вычислить серию интегралов

$$I(\alpha_1, \alpha_2, \alpha_3) = \int_G f(P; \alpha_1, \alpha_2, \alpha_3)\, dP$$

при различных значениях параметров α_1, α_2, α_3. Область интегрирования G содержалась в единичном трехмерном кубе; при этом условие принадлежности точки к области G задавалось громоздкой системой неравенств. Было ясно, что нельзя сделать замену независимых переменных

так, чтобы область интегрирования G приобрела стандартный вид, где было бы возможным применение квадратурных формул высокого порядка точности. Поэтому пришлось вне области G продолжить f нулем и рассматривать исходную задачу как задачу вычисления интеграла

$$\bar{I}(\alpha_1, \alpha_2, \alpha_3) = \int_0^1 \int_0^1 \int_0^1 \bar{f}(P; \alpha_1, \alpha_2, \alpha_3)\, dx_1 dx_2 dx_3,$$

где

$$\bar{f}(P; \alpha_1, \alpha_2, \alpha_3) = \begin{cases} f(P; \alpha_1, \alpha_2, \alpha_3) & \text{при} \quad P \in G, \\ 0 & \text{при} \quad P \notin G. \end{cases}$$

Подынтегральная функция оказалась теперь разрывной. Для вычисления интеграла были применены формулы прямоугольников и Гаусса. С целью контроля над точностью проводились расчеты с различным числом узлов интегрирования. Оказалось, что результаты расчетов медленно устанавливаются (сильно меняются при изменении числа узлов), что указывает на малую точность получаемых приближенных значений. При непосредственном применении метода Монте-Карло установление получаемых приближенных значений было еще хуже. Поэтому было принято решение применить для вычисления описанные выше способы уменьшения дисперсии путем разбиения области на части. Были опробованы оба описанных выше способа. Область интегрирования $0 \leqslant x_1, x_2, x_3 \leqslant 1$ разбивалась на равные кубы с ребрами длиной $1/n$, и далее применялись оба метода, рассмотренные в предыдущем параграфе. Как и в случае других квадратур, исследовалось установление результатов вычислений. Оказалось, что оба метода могут обеспечить требуемую точность вычислений при приемлемых затратах машинного времени.

Во многих случаях вычисление интегралов высокой кратности или большой серии интегралов удается осуществить лишь за счет устранения повторения одинаковых вычислений. Решение рассматриваемой задачи представлялось сначала бесперспективным из-за большой трудоемкости проверки условия принадлежности узла P области G; нахождение же каждого значения $f(P; \alpha_1, \alpha_2, \alpha_3)$ требовало существенно меньшего объема вычислений. Так как все интегралы вычислялись по одной и той же области, то было принято решение применить следующий алгоритм. Куб $0 \leqslant x_1, x_2, x_3 \leqslant 1$ разбивался на кубики с длиной ребра $1/n$, в каждом кубике выбиралась случайная точка $P_{n_1,\ldots,n_s}$ и проверялось условие принадлежности точки $P_{n_1,\ldots,n_s}$ области G, координаты всех точек $P_{n_1,\ldots,n_s} \in G$ были записаны в памяти ЭВМ. Далее все интегралы серии заменяли суммами

$$\sum_{P_{n_1,\ldots,n_s} \in G} \frac{1}{n^3} f(P; \alpha_1, \alpha_2, \alpha_3)$$

с одними и теми же узлами. Отказ от проверки условия $P_{n_1,\ldots,n_s} \in G$ при вычислении каждого интеграла привел к снижению требований на затраты машинного времени примерно в 100 раз.

Другим фактором, позволившим резко снизить затраты машинного времени, оказался следующий. Подынтегральная функция имела вид

$$g(x_1, \alpha_1)h(x_1, x_2; \alpha_1, \alpha_2, \alpha_3),$$

где вычисление каждого значения функции h требовало относительно малого числа элементарных операций ЭВМ по сравнению с вычислением каждого значения функции g. Поэтому была применена следующая схема вычислений. Все интегралы $\bar{T}(\alpha_1, \alpha_2, \alpha_3)$, соответствующие одному и тому же значению параметра α_1, вычислялись одновременно; благодаря этому каждое значение $g(x_1, \alpha_1)$ вычислялось при расчете всей серии интегралов только один раз. Такая организация работы позволила довести затраты машинного времени при решении рассматриваемой задачи до приемлемых размеров. Анализ хода решения задачи, проведенный после окончания вычисления серии, показал ряд неиспользованных возможностей, которые предоставили бы дополнительные удобства как исполнителю, так и заказчику. С целью уменьшения затрат машинного времени заказчик старался уменьшить число M совокупностей значений параметров $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$, при которых требовалось вычислить значение интеграла. Наиболее трудоемкую часть при расчетах составляло вычисление значений функции $g(x_1, \alpha_1)$, поэтому общие затраты машинного времени были пропорциональны не числу M, а числу различных значений α_1^i. Таким образом, общие затраты времени могли бы быть снижены также за счет удачного выбора совокупности $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$.

В рассматриваемой задаче имелся еще один неиспользованный резерв повышения точности. Трудоемкость рассматриваемого метода пропорциональна произведению числа различных значений α на число различных значений координат x_1 узлов интегрирования. Узлы интегрирования выбирались случайно, и поэтому число различных значений x_1 было очень большим. Для уменьшения этого числа можно было бы пойти, например, по следующему пути: аппроксимировать исходный интеграл по формуле прямоугольников

$$I(\alpha_1, \alpha_2, \alpha_3) \approx \sum_{q=1}^{m} \frac{1}{m} \bar{I}\left(\alpha_1, \alpha_2, \alpha_3; \frac{q-1/2}{m}\right),$$

где

$$\bar{I}\left(\alpha_1, \alpha_2, \alpha_3; \frac{q-1/2}{m}\right) = \int_0^1 \int_0^1 f\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}, x_2, x_3\right) dx_2 dx_3,$$

и применять метод Монте-Карло с разбиением области интегрирования только для вычисления интегралов $\bar{I}(\alpha_1, \alpha_2, \alpha_3; x_1)$. Другой подход к решению задачи: разбить отрезок $[0, 1]$ на отрезки $[(q-1)/m, q/m]$, на каждом из них выбрать случайную точку ξ_q и положить

$$I(\alpha_1, \alpha_2, \alpha_3) \approx \frac{1}{m} \sum_{q=1}^{m} \bar{I}(\alpha_1, \alpha_2, \alpha_3; \xi_q).$$

Оба указанных выше метода имеют худшую скорость сходимости по сравнению с примененным методом, однако в этих методах вычисляется существенно меньше значений функции $g(x_1, \alpha_1)$.

Из последних рассуждений видно, что при вычислении больших серий интегралов (как, впрочем, и в случае больших серий других задач) часто большой эффект достигается не за счет повышения качества метода при решении каждой из задач серии, а за счет лучшей организации вычислений.

При применении процедуры, описанной в § 4, вычисление интеграла высокой кратности как бы сводилось к вычислению большого числа интегралов меньшей кратности. Поэтому при вычислении интегралов высокой кратности иногда можно использовать указанные выше резервы повышения эффективности в случае вычисления серий интегралов.

Обратим внимание на следующую опасность, возникающую при одновременном поучении результата сразу по всей серии задач.

Предъявляемая к решению совокупность задач соответствует какому-то реальному явлению. Может случиться, что ранее явление не «обсчитывалось» на ЭВМ и предъявляемая математическая модель не является удовлетворительной. Тогда в случае одновременного решения задач серии все результаты вычислений окажутся бесполезными. При последовательном решении задач можно уже после получения первых результатов обнаружить рассогласование математической модели с общей картиной явления. Такое рассогласование типично при решении новых задач, и обычно оно устраняется не сразу, а после большого числа пробных просчетов и совместного обсуждения модели заказчиком и исполнителем.

Исполнитель имеет особые основания быть заинтересованным в обсуждении постановки задачи. Если исходная постановка задачи окажется неразумной, ему придется потратить много времени на выбор нового алгоритма и написание новой программы. Участвуя в обсуждении, исполнитель может уяснить себе возможные варианты изменения постановок задачи и предусмотреть их при составлении программы решения задачи. В связи с этим при решении задач новых типов особенно важно, чтобы программа подразделялась на отдельные функциональные блоки, поддающиеся независимому изменению.

Часто столкновение исходных позиций заказчика (обсчет подробной модели) и исполнителя (минимальный объем работ) приводит к построению упрощенной модели явления, быстрый обсчет которой позволяет решить вопрос о целесообразности рассмотрения более сложной модели.

Рассмотрим другой пример, относящийся к организации работы по выбору метода интегрирования в случае кратных интегралов. Часто вычисление интегралов по сложной области сводят к вычислению интегралов по области, являющейся прямым произведением областей простейшего вида: отрезков, параллелепипедов, лучей, бесконечных прямых, кругов, сфер, шаров и т. д. Для таких стандартных областей имеется достаточное количество совершенных ме-

тодов интегрирования, и после такого преобразования области интегрирования можно применить процедуру, описанную в § 4, или какую-либо другую, аналогичную ей.

Пусть вычисляется интеграл

$$I(f) \equiv \int_{1 \leqslant x_1^2+x_2^2+x_3^2 \leqslant 4} f(x_1, x_2, x_3)\, dx_1 dx_2 dx_3.$$

Его удобно записать в виде

$$I(f) = \int_1^2 \int_{S_1} g(l, \omega)\, d\omega dl,$$

где S_l — единичная сфера, $d\omega$ — элемент ее площади,

$$g(l, \omega) = l^2 f(l\omega_1, l\omega_2, l\omega_3), \quad \omega = (\omega_1, \omega_2, \omega_3), \qquad \omega_1^2 + \omega_2^2 + \omega_3^2 = 1.$$

Предположим, что решено вычислять интеграл при помощи квадратур, являющихся прямым произведением квадратур по отрезку $[1, 2]$ и по сфере S_1.

Под *прямым произведением* квадратуры

$$\int_1^2 h(l)\, dl \approx \sum_{j=1}^{n} d_j h(l_j) \tag{1}$$

и квадратуры

$$\int_{S_1} p(\omega)\, d\omega \approx \sum_{q=1}^{m} k_q p(\omega_q) \tag{2}$$

здесь понимается квадратура

$$\int_1^2 \int_{S_1} g(l, \omega) d\omega dl \approx \sum_{j=1}^{n} \sum_{q=1}^{m} d_j k_q g(l_j, \omega_q).$$

Исследуем зависимость погрешности интегрирования от способа интегрирования и от числа узлов. Для исследовании поведения погрешности численного интегрирования по оси l выберем какую-то «базовую» квадратуру по единичной сфере:

$$\int_{S_1} p(\omega)\, d\omega \approx \sum_{q=1}^{m_0} k_q^0 p(\omega_q^0); \tag{3}$$

имеется в виду, что число узлов m_0 мало и в то же время выражения

$$G(l) = \int_{S_1} g(l, \omega)\, d\omega \quad \text{и} \quad G^0(l) = \sum_{q=1}^{m_0} k_q^0 g(l, \omega_q^0)$$

имеют одинаковый качественный характер поведения по l. Например, можно попытаться взять в качестве (3) квадратуру

$$\int_{S_1} p(\omega)\, d\omega \approx \frac{2\pi}{3} \sum p(\omega_1, \omega_2, \omega_3),$$

где суммирование производится по 6 точкам пересечения единичной сферы с координатными осями x_1, x_2, x_3.

Предположим, что для вычисления интеграла по оси l принято решение применить или формулу Гаусса, или формулу Симпсона. Последовательно применяя формулы Гаусса

$$\int_1^2 h(l)\,dl \approx \sum_{j=1}^{n} d_j^n h(l_j^n)$$

при $n = n_1, n_2, \ldots$ узлах, получаем некоторые величины

$$G_{n_i}^0 = \sum_{j=1}^{n_i} d_j^{n_i} G^0(l_j^{n_i}).$$

Точно так же при числе узлов $n = n_1', n_2', \ldots$ получаем приближения по формуле Симпсона $S_{n_i'}^0$.

Из рассмотрения поведения всех этих величин можно усмотреть значение предела I^0, к которому они стремятся. Далее, при каждом n среди всех приближений $G_{n_i}^0$ и $S_{n_i'}^0$ с $n_i, n_i' \leqslant n$ выберем приближение Γ_n, обеспечивающее лучшую точность, и введем функцию $\varphi_l(n) = |\Gamma_n - I^0|$ *погрешности численного интегрирования по оси* l.

Точно так же фиксируется некоторая базовая квадратура по переменной l (часто эта квадратура Гаусса с двумя узлами) и строится функция $\varphi_\omega(m)$ погрешности численного интегрирования по сфере ω. Предположим, что суммарная погрешность есть $R = \varphi_l(n) + \varphi_\omega(m)$. Если значения функции f вычисляются независимо, то трудоемкость метода пропорциональна nm. Минимизируя nm при заданном требовании на точность, чтобы $\varphi_l(n) + \varphi_\omega(m) \leqslant \varepsilon$, получаем искомые значения числа узлов n_0 и m_0. В зависимости от того, какой квадратуре — Гаусса или Симпсона — отвечает данное n, выбираем соответствующую квадратуру. В случае сомнений в правомерности использования этой методики остается возможность проверить правильность результата, проведя дополнительное интегрирование с несколько отличными от n_0 и m_0 значениями n и m.

Приведем типичный пример подобной организациии выбора способа интегрирования по каждой из осей в случае большой серии интегралов по единичному s-мерному кубу. По каждой из осей применяется формула Гаусса, при выборе числа ее узлов по каждой из осей в качестве базовых квадратур по остальным осям берутся квадратуры Гаусса с двумя узлами. Дробление числа узлов по каждой из рассматриваемых осей продолжается до тех пор, пока разность между двумя последующими приближениями не станет менее чем $\varepsilon \cdot s^{-1}$. После определения таким образом нужного числа узлов $n^{(i)}$, соответствующего каждой оси, производится вычисление интеграла с $n^{(i)}$ узлами по каждой из осей.

Из-за некоторой ненадежности описанного алгоритма обычно производится проверка правомерности его применения *в случае данной серии интегралов*: выбирается достаточно представительная подсерия интегралов и для нее результаты расчетов по данному алгоритму сравниваются с результатами расчетов при несколько измененных значениях числа узлов по осям.

Литература

1. Бахвалов Н. С. Об оптимальных оценках скорости сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций. // В кн.: Численные методы решения дифференциальных и интегральных уравнений и квадратурные формулы. — М.: Наука, 1964. С. 5–63.
2. Бахвалов Н. С. Численные методы. — М.: Наука, 1975.
3. Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. — М.: Наука, 1986.
4. Мысовских И. П. Интерполяционные кубатурные формулы. — М.: Наука, 1981.
5. Никифоров А. Ф., Суслов С. К., Уваров В. Б. Классические ортогональные полиномы дискретной переменной. — М.: Наука, 1985.
6. Никольский С. М. Квадратурные формулы. — М.: Наука, 1979.
7. Соболев С. Л. Введение в теорию кубатурных формул. — М.: Наука, 1974.

Глава 6

Численные методы алгебры



К численным методам алгебры традиционно относят численные методы решения систем линейных алгебраических уравнений, обращения матриц, вычисления определителей, нахождения собственных значений и собственных векторов матриц и нулей многочленов.

При формальном подходе решение этих задач не встречает затруднений: решение системы можно найти, раскрыв определители в формуле Крамера; для нахождения собственных значений матрицы достаточно выписать характеристическое уравнение и найти его корни. Однако эти рекомендации встречают возражения со многих сторон.

Так, при непосредственном раскрытии определителей решение системы с m неизвестными требует порядка $m!m$ арифметических операций; уже при $m = 30$ такое число операций недоступно для современных ЭВМ. При сколь-нибудь больших m применение методов с таким порядком числа операций будет невозможно и в обозримом будущем.

Другой причиной, по которой эти классические способы неприменимы даже при малых m, является сильное влияние на окончательный результат округлений при вычислениях. Уже при $m = 20$ при расчетах на современных ЭВМ типична аварийная остановка из-за переполнения порядка чисел. Даже если такая остановка не происходит, результат вычислений часто далек от истинного значения из-за влияния вычислительной погрешности. Точно так же обстоит дело при нахождении собственных значений матриц с использованием явного выражения характеристического многочлена.

Методы решения алгебраических задач разделяются на точные, итерационные и вероятностные. Классы задач, для решения которых обычно применяют методы этих групп, можно условно назвать соответственно классами задач с малым, средним и большим числом неизвестных. Изменение объема и структуры памяти ЭВМ, увеличение их быстродействия и развитие численных методов приводят к смещению границ применения методов в сторону систем более высоких порядков. В настоящее время точные методы обычно применяются для решения систем до порядка 10^4, итерационные — до порядка 10^7.

При изучении итерационных процессов нам понадобятся понятия норм вектора и матрицы. Напомним определения основных норм в простран-

ствах векторов и матриц. Если в пространстве векторов $\mathbf{x} = (x_1, \ldots, x_m)^T$ введена норма $\|\mathbf{x}\|$, то согласованной с ней нормой в пространстве матриц A называют норму

$$\|A\| = \sup_{\mathbf{x}\neq 0} \|A\mathbf{x}\|/\|\mathbf{x}\|. \tag{1}$$

Наиболее употребительны в пространстве векторов следующие нормы:

$$\|\mathbf{x}\|_\infty = \max_{1\leqslant j\leqslant m} |x_j|, \tag{2}$$

$$\|\mathbf{x}\|_1 = \sum_{j=1}^{m} |x_j|, \tag{3}$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^{m} |x_j|^2} = \sqrt{(\mathbf{x}, \mathbf{x})}, \tag{4}$$

а согласованными с ними нормами в пространстве матриц являются соответственно нормы

$$\|A\|_\infty = \max_{1\leqslant i\leqslant m} \Big(\sum_{j=1}^{m} |a_{ij}|\Big), \tag{5}$$

$$\|A\|_1 = \max_{1\leqslant j\leqslant m} \Big(\sum_{i=1}^{m} |a_{ij}|\Big), \tag{6}$$

$$\|A\|_2 = \sqrt{\max_{1\leqslant i\leqslant m} \lambda^i_{A^*A}}; \tag{7}$$

здесь и далее λ^i_D — собственные значения матрицы D.

Приведем вывод этих соотношений для вещественного случая. Поскольку, согласно (2),

$$\|A\mathbf{x}\|_\infty = \max_i \Big|\sum_j a_{ij}x_j\Big| \leqslant \max_i \Big(\sum_j |a_{ij}| \max_j |x_j|\Big) \leqslant \max_i \Big(\sum_j |a_{ij}|\Big) \max_j |x_j|,$$

то

$$\frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leqslant \max_i \Big(\sum_j |a_{ij}|\Big).$$

Пусть $\max\limits_i \Big(\sum\limits_j |a_{ij}|\Big)$ достигается при $i = l$; для вектора

$$\mathbf{x} = (\operatorname{sign}(a_{l1}), \ldots, \operatorname{sign}(a_{lm}))^T$$

имеем $\|\mathbf{x}\|_\infty = 1$,

$$\|A\mathbf{x}\|_\infty \geqslant \Big|\sum_j a_{lj}x_j\Big| = \sum_j |a_{lj}| = \Big(\max_i \sum_j |a_{ij}|\Big)\|\mathbf{x}\|_\infty.$$

Из этих соотношений следует (5).

Точно так же для нормы вектора, определяемой по формуле (3), имеем

$$\|A\mathbf{x}\|_1 = \sum_i \Big|\sum_j a_{ij}x_j\Big| \leqslant \Big(\max_j \sum_i |a_{ij}|\Big) \sum_j |x_j|,$$

т. е.

$$\frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leqslant \max_j \Big(\sum_i |a_{ij}|\Big).$$

Пусть $\max\limits_j \sum\limits_i |a_{ij}|$ достигается при $j = l$. Для вектора $\mathbf{x}$, у которого лишь одна компонента x_l отлична от нуля, имеем

$$\|A\mathbf{x}\|_1 = \sum_i \Big|\sum_i a_{ij}x_j\Big| = \sum_i |a_{il}|\,|x_l| = \Big(\sum_i |a_{il}|\Big)|x_l| =$$

$$= \Big(\max_j \sum_i |a_{ij}|\Big)\sum_j |x_j| = \Big(\max_j \sum_i |a_{ij}|\Big)\|\mathbf{x}\|_1;$$

отсюда следует (6).

Согласно определению $\|A\|_2$ и (4), имеем

$$\|A\|_2 = \sup_{\mathbf{x}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x}} \sqrt{\frac{(A\mathbf{x},\, A\mathbf{x})}{(\mathbf{x},\, \mathbf{x})}} = \sqrt{\sup_{\mathbf{x}} \frac{(A^TA\mathbf{x},\, \mathbf{x})}{(\mathbf{x},\, \mathbf{x})}}.$$

Матрица A^TA симметричная, поскольку $(A^TA)^T = A^T(A^T)^T = A^TA$.

Пусть матрица B симметричная, $\mathbf{e}_1, \ldots, \mathbf{e}_m$ — ортонормированная система ее собственных векторов, $\lambda_1, \ldots, \lambda_m$ — соответствующие собственные значения. Представим произвольный вектор $\mathbf{x}$ в виде $\sum\limits_{i=1}^{m} c_i\mathbf{e}_i$. Имеем

$$(B\mathbf{x},\, \mathbf{x}) = \Big(\sum_i \lambda_i c_i \mathbf{e}_i,\, \sum_i c_i\mathbf{e}_i\Big) = \sum_i \lambda_i |c_i|^2,$$

поэтому

$$(B\mathbf{x},\, \mathbf{x}) \leqslant \Big(\max_i \lambda_i\Big)\sum_i |c_i|^2 = (\max_i \lambda_i)(\mathbf{x},\, \mathbf{x}) \tag{8}$$

и

$$(B\mathbf{x},\, \mathbf{x}) \geqslant (\min_i \lambda_i)(\mathbf{x},\, \mathbf{x}). \tag{9}$$

В то же время $(B\mathbf{e}_i,\, \mathbf{e}_i)/(\mathbf{e}_i,\, \mathbf{e}_i) = \lambda_i$. Из этих соотношений следует, что

$$\sup_{\mathbf{x}} \frac{|(B\mathbf{x},\, \mathbf{x})|}{(\mathbf{x},\, \mathbf{x})} = \max_i |\lambda_i|. \tag{10}$$

Поскольку $(A^TA\mathbf{x},\, \mathbf{x}) = (A\mathbf{x},\, A\mathbf{x}) \geqslant 0$, то все $\lambda^i_{A^TA} \geqslant 0$. Полагая в (10) $B = A^TA$, получаем

$$\sup_{\mathbf{x}} \frac{|(A^TA\mathbf{x},\, \mathbf{x})|}{(\mathbf{x},\, \mathbf{x})} = \max_i |\lambda^i_{A^TA}| = \max_i \lambda^i_{A^TA}.$$

Из полученных соотношений следует (7).

Отметим важный частный случай.

Если A — симметричная матрица, то $\lambda^i_{A^TA} = \lambda^i_{A^2} = |\lambda^i_A|^2$, поэтому для нее

$$\|A\|_2 = \max |\lambda^i_A|. \tag{11}$$

Если $A\mathbf{x} = \lambda\mathbf{x}$, то $\|A\|\,\|\mathbf{x}\| \geqslant \|\lambda\mathbf{x}\| = |\lambda|\,\|\mathbf{x}\|$. Следовательно, модуль любого собственного значения матрицы A не больше любой ее нормы.

§ 1. Методы последовательного исключения неизвестных

Рассмотрим точные методы решения системы $A\mathbf{x} = \mathbf{b}$; здесь $A = [a_{ij}]$ — матрица размерности $m \times m$, $\det A \neq 0$,

$$\mathbf{b} = (a_{1,m+1}, \ldots, a_{m,m+1})^T.$$

Метод решения задачи относят к классу точных, если в предположении отсутствия округлений он дает точное решение задачи после конечного числа арифметических и логических операций. Если число ненулевых элементов матрицы системы имеет порядок m^2, то для большинства используемых в настоящее время точных методов решения таких систем требуемое число операций имеет порядок m^3. Поэтому для применимости точных методов необходимо, чтобы такой порядок числа операций был приемлем для данной ЭВМ; другие ограничения накладываются объемом и структурой памяти ЭВМ.

Оговорка об «используемых в настоящее время методах» имеет следующий смысл. Существуют методы решения таких систем с меньшим порядком числа операций, однако они не используются активно из-за сильной чувствительности результата к вычислительной погрешности.

Наиболее известным из точных методов решения систем линейных уравнений является *метод исключения Гаусса*. Рассмотрим одну из его возможных реализаций. В предположении, что $a_{11} \neq 0$, первое уравнение системы

$$\sum_{j=1}^{m} a_{ij}x_j = a_{i,m+1}, \quad i = 1, \ldots, m, \tag{1}$$

делим на коэффициент a_{11}, в результате получаем уравнение

$$x_1 + \sum_{j=2}^{m} a^1_{1j}x_j = a^1_{1,m+1}.$$

Затем из каждого из остальных уравнений вычитается первое уравнение, умноженное на соответствующий коэффициент a_{i1}. В результате эти уравнения преобразуются к виду

$$\sum_{j=2}^{m} a^1_{ij}x_j = a^1_{i,m+1}, \quad i = 2, \ldots, m.$$

Первое неизвестное оказалось исключенным из всех уравнений, кроме первого. Далее в предположении, что $a_{22}^1 \neq 0$, делим второе уравнение на коэффициент a_{22}^1 и исключаем неизвестное x_2 из всех уравнений, начиная со второго, и т. д. В результате последовательного исключения неизвестных система уравнений преобразуется в систему уравнений с треугольной матрицей

$$x_i + \sum_{j=i+1}^{m} a_{ij}^i x_j = a_{i,m+1}^i, \quad i = 1, \ldots, m. \tag{2}$$

Совокупность проведенных вычислений, в ходе которых исходная задача преобразовалась к виду (2), называется *прямым ходом метода Гаусса.*

Из m-го уравнения системы (2) определяем x_m, из $(m-1)$-го — x_{m-1} и т. д. до x_1. Совокупность таких вычислений называют *обратным ходом метода Гаусса.*

Нетрудно проверить, что реализация прямого хода метода Гаусса требует $N \sim 2m^3/3$ арифметических операций, а обратного — $N \sim m^2$ арифметических операций.

Исключение x_i происходит в результате следующих операций: 1) деления i-го уравнения на a_{ii}^{i-1}, 2) вычитания получающегося после такого деления i-го уравнения, умноженного на a_{ik}^{i-1}, из уравнений с номерами $k = i+1, \ldots, m$. Первая операция равносильна умножению системы уравнений слева на диагональную матрицу

$$C_i = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & (a_{ii}^{i-1})^{-1} & & & & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{pmatrix};$$

вторая операция равносильна умножению слева на матрицу

$$C_i' = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & -a_{i,i+1}^{i-1} & & & & \\ & & & \vdots & 1 & & & \\ & & & \vdots & & \ddots & & \\ & & & \vdots & & & 1 & \\ & & & -a_{i,m}^{i-1} & & & & 1 \end{pmatrix}.$$

Таким образом, система (2), получаемая в результате этих преобразований, запишется в виде

$$CA\mathbf{x} = C\mathbf{b}, \quad \text{где} \quad C = C_m \dots C_1' C_1.$$

Произведение левых (правых) треугольных матриц является левой (правой) треугольной матрицей, поэтому матрица C левая треугольная.

Из формулы для элементов обратной матрицы

$$(A^{-1})_{ij} = A_{ji}/\det A$$

следует, что матрица, обратная к левой (правой) треугольной, является левой (правой) треугольной. Следовательно, матрица $B = C^{-1}$ левая треугольная.

Введем обозначение $CA = D$. Согласно построению все $d_{ii} = 1$ и матрица D правая треугольная. Отсюда получаем представление матрицы A в виде произведения левой и правой треугольных матриц:

$$A = C^{-1}D = BD.$$

Равенство $A = BD$ вместе с условием $d_{ii} = 1$, $i = 1, \dots, m$, образует систему уравнений относительно элементов треугольных матриц B и D: $\sum\limits_{j=1}^{m} b_{ij}d_{jk} = a_{ik}$. Поскольку $b_{ij} = 0$ при $i < j$ и $d_{jk} = 0$ при $k < j$, эта система может быть записана в виде

$$\sum_{j=1}^{\min\{i,\,k\}} b_{ij}d_{jk} = a_{ik} \tag{3}$$

или, что то же самое,

$$\sum_{j=1}^{k} b_{ij}d_{jk} = a_{ik} \quad \text{при} \quad k \leqslant i,$$

$$\sum_{j=1}^{i} b_{ij}d_{jk} = a_{ik} \quad \text{при} \quad i < k.$$

Воспользовавшись условием, что все $d_{ii} = 1$, получаем систему рекуррентных соотношений для определения элементов b_{ij} и d_{ij}:

$$\begin{aligned} b_{ik} &= a_{ik} - \sum_{j=1}^{k-1} b_{ij}d_{jk} \text{ при} \quad k \leqslant i, \\ d_{ik} &= \frac{a_{ik} - \sum\limits_{j=1}^{i-1} b_{ij}d_{jk}}{b_{ii}} \text{при} \quad i < k. \end{aligned} \tag{4}$$

Вычисления проводятся последовательно для совокупностей $(i, k) = (1, 1)$, $\ldots$, $(1, m)$, $(2, 1), \ldots, (2, m), \ldots, (m, 1), \ldots, (m, m)$. Здесь и далее в случае, когда верхний предел суммирования меньше нижнего, считается, что вся сумма равна нулю.

Таким образом, вместо последовательных преобразований системы (1) к виду (2) можно непосредственно произвести вычисление матриц B и D с помощью формул (4). Эти вычисления можно осуществить, если только все элементы b_{ii} окажутся отличными от нуля. Пусть A_k, B_k, D_k — матрицы главных миноров k-го порядка матриц A, B, D. Согласно (3) $A_k = B_k D_k$. Поскольку $\det D_k = 1$, $\det B_k = b_{11} \ldots b_{kk}$, то $\det A_k = b_{11} \ldots b_{kk}$. Следовательно,

$$b_{kk} = \det A_k / \det A_{k-1}.$$

Итак, для осуществления вычислений по формулам (4) необходимо и достаточно выполнение условий

$$\det A_k \neq 0, \quad k = 1, \ldots, m. \tag{5}$$

В ряде случаев заранее известно, что условие (5) выполнено. Например, многие задачи математической физики сводятся к решению систем с положительно определенной матрицей A. Однако в общем случае этого заранее сказать нельзя. Возможен и такой случай: все $\det A_k \neq 0$, но среди величин b_{kk} есть очень малые и при делении на них будут получаться большие числа с большими абсолютными погрешностями. В результате этого решение сильно исказится.

Обозначим $C\mathbf{b} = \mathbf{d} = (d_{1,m+1}, \ldots, d_{m,m+1})^T$. Поскольку $C^{-1} = B$ и $CA = D$, то справедливы равенства $B\mathbf{d} = \mathbf{b}$, $D\mathbf{x} = \mathbf{d}$. Таким образом, после разложения матрицы исходной системы на произведение левой и правой треугольных матриц решение исходной системы сводится к последовательному решению двух систем $B\mathbf{d} = \mathbf{b}$, $D\mathbf{x} = \mathbf{d}$ с треугольными матрицами; это потребует $N \sim 2m^2$ арифметических операций.

Последовательность операций по разложению матрицы A на произведение треугольных матриц и по определению вектора $\mathbf{d}$ часто удобно объединить. Уравнения

$$\sum_{j=1}^{i} b_{ij} d_{j,m+1} = a_{i,m+1}$$

системы $B\mathbf{d} = \mathbf{b}$ можно записать в виде (3)

$$\sum_{j=1}^{\min(i,m+1)} b_{ij} d_{j,m+1} = a_{i,m+1}.$$

Следовательно, значения $d_{i,m+1}$ могут вычисляться одновременно с остальными значениями d_{ij} по формулам (4).

При решении практических задач часто возникает необходимость решения систем уравнений с матрицей, содержащей большое количество ну-

левых элементов. Обычно эти матрицы имеют так называемую *ленточную структуру*. Более точно, матрицу A называют $(2q+1)$-диагональной или имеющей *ленточную структуру*, если $a_{ij}=0$ при $|i-j|>q$. Число $2q+1$ называют *шириной ленты*. Оказывается, что при решении системы уравнений с ленточной матрицей методом Гаусса число арифметических операций и требуемый объем памяти ЭВМ могут быть существенно сокращены.

Задача 1. Исследовать характеристики метода Гаусса и метода решения системы с помощью разложения ленточной матрицы A на произведение левой и правой треугольных матриц. Показать, что для нахождения решения требуется $O(mq^2)$ арифметических операций (при $m, q \to \infty$). Найти главный член числа операций при условии $1 \ll q \ll m$.

Задача 2. Оценить объем загружаемой памяти ЭВМ в методе Гаусса для ленточных матриц.

При вычислениях без помощи ЭВМ велика вероятность случайных погрешностей. Для устранения таких погрешностей иногда вводят *контрольный столбец системы* $\mathbf{a}_{m+2}=(a_{1,\,m+2},\ldots,a_{m,\,m+2})^T$, состоящий из контрольных элементов уравнений системы

$$a_{i,\,m+2}=\sum_{j=1}^{m+1}a_{ij}.$$

При преобразовании уравнений над контрольными элементами производятся те же операции, что и над свободными членами уравнений. В результате этого контрольный элемент каждого нового уравнения должен равняться сумме коэффициентов этого уравнения. Большое расхождение между ними указывает на погрешности в вычислениях или на неустойчивость алгоритма вычислений по отношению к вычислительной погрешности.

К примеру, в случае приведения системы уравнений $A\mathbf{x}=\mathbf{b}$ к виду $D\mathbf{x}=\mathbf{d}$ с помощью формул (4) контрольный элемент $d_{i,\,m+2}$ каждого из уравнений системы $D\mathbf{x}=\mathbf{d}$ вычисляется по тем же формулам (4). После вычисления всех элементов d_{ij} при фиксированном i контроль осуществляется проверкой равенства

$$\sum_{j=i}^{m+1}d_{ij}=d_{i,\,m+2}.$$

Обратный ход метода Гаусса также сопровождается вычислением контрольных элементов строк системы.

Чтобы избежать катастрофического влияния вычислительной погрешности, применяют *метод Гаусса с выбором главного элемента*. Его отличие

от описанной выше схемы метода Гаусса состоит в следующем. Пусть по ходу исключения неизвестных получена система уравнений

$$x_i + \sum_{j=i+1}^{m} a_{ij}^i x_j = a_{i,m+1}^i, \quad i = 1, \ldots, k,$$

$$\sum_{j=k+1}^{m} a_{ij}^k x_j = a_{i,m+1}^k, \qquad i = k+1, \ldots, m.$$

Найдем l такое, что $|a_{k+1,l}^k| = \max\limits_j |a_{k+1,j}^k|$ и переобозначим $x_{k+1} = x_l$ и $x_l = x_{k+1}$; далее произведем исключение неизвестной x_{k+1} из всех уравнений, начиная с $(k+2)$-го. Такое переобозначение приводит к изменению порядка исключения неизвестных и во многих случаях существенно уменьшает чувствительность решения к погрешностям округления при вычислениях.

Часто требуется решить несколько систем уравнений $A\mathbf{x} = \mathbf{b}_q$, $q = 1, \ldots, p$, с одной и той же матрицей A. Удобно поступить следующим образом: введя обозначения

$$\mathbf{b}_q = (a_{1,m+q}, \ldots, a_{m,m+q})^T,$$

произведем вычисления по формулам (4), причем элементы d_{ik} вычислим при $i < k \leqslant m + p$. В результате будут получены p систем уравнений с треугольной матрицей, соответствующих исходной задаче

$$D\mathbf{x} = \mathbf{d}_q, \quad \mathbf{d}_q = (d_{1,m+q}, \ldots, d_{m,m+q})^T, \quad q = 1, \ldots, p.$$

Решаем эти системы каждую в отдельности. Оказывается, что общее число арифметических действий при решении p систем уравнений таким способом $N \sim 2m^3/3 + 2pm^2$.

Описанный выше прием иногда используется для того, чтобы без существенных дополнительных затрат получить суждение о погрешности решения, являющейся следствием погрешностей округления при вычислениях. Задаются вектором $\mathbf{z}$ с компонентами, имеющими по возможности тот же порядок и знак, что и компоненты искомого решения; часто из-за отсутствия достаточной информации берут $\mathbf{z} = (1, \ldots, 1)^T$. Вычисляется вектор $\mathbf{c} = A\mathbf{z}$, и наряду с исходной системой уравнений решается система $A\mathbf{z} = \mathbf{c}$.

Пусть $\mathbf{x}'$ и $\mathbf{z}'$ — реально получаемые решения этих систем. Суждение о погрешности $\mathbf{x}' - \mathbf{x}$ искомого решения можно получить, основываясь на гипотезе: относительные погрешности при решении методом исключения систем с одной и той же матрицей и различными правыми частями, которыми являются соответственно величины $\|\mathbf{x} - \mathbf{x}'\|/\|\mathbf{x}'\|$ и $\|\mathbf{z} - \mathbf{z}'\|/\|\mathbf{z}'\|$, отличаются не в очень большое число раз.

Другой прием для получения суждения о реальной величине погрешности, возникающей за счет округлений при вычислениях, состоит в *изменении масшта-*

бов, меняющем картину накопления вычислительной погрешности. Наряду с исходной системой тем же методом решается система

$$(\alpha A)\mathbf{x}' = \beta\mathbf{b}, \quad \text{где} \quad \alpha \text{ и } \beta \text{ — числа.}$$

При α и β, не являющихся целыми степенями двойки, сравнение векторов $\mathbf{x}$ и $\alpha\beta^{-1}\mathbf{x}'$ дает представление о величине вычислительной погрешности. Например, можно взять $\alpha = \sqrt{2}$, $\beta = \sqrt{3}$.

Изучение многих задач приводит к необходимости решения систем линейных уравнений с симметричной положительно определенной матрицей. Такие системы возникают, например, при решении дифференциальных уравнений методом конечных элементов или же конечно-разностными методами. В этих случаях матрица системы имеет также и ленточную структуру.

Для решения таких систем, а также систем уравнений более общего вида с эрмитовой не обязательно положительно определенной матрицей применяется *метод квадратного корня* (*метод Холецкого*). Матрица A представляется в виде

$$A = S^*DS, \tag{6}$$

где S — правая треугольная матрица, S^* — сопряженная с ней, т. е.

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots \\ 0 & s_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix},$$

причем все $s_{ii} > 0$, D — диагональная матрица с элементами d_{ii}, равными $+1$ или -1. Матричное равенство (6) образует систему уравнений

$$a_{ij} = \bar{s}_{1i}s_{1i}d_{11} + \dots + \bar{s}_{ij}s_{ij}d_{ii} \quad \text{при} \quad i \leqslant j.$$

Аналогичные уравнения при $i > j$ отброшены, так как уравнения, соответствующие парам (i, j) и (j, i), эквивалентны. Отсюда получаем рекуррентные формулы для определения элементов d_{ii} и s_{ij}:

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}\right), \quad s_{ii} = \sqrt{\left|a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}\right|},$$

$$s_{ij} = \frac{a_{ij} - \sum\limits_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk}}{\bar{s}_{ii}d_{ii}} \quad \text{при} \quad i < j.$$

Матрица S является правой треугольной, и, таким образом, после получения представления (6) решение исходной системы также сводится к последовательному решению двух систем с треугольными матрицами. Заметим, что в случае $A > 0$ все $d_{ii} = 1$ и $A = S^*S$.

Задача 3. Оценить число арифметических операций и загрузку памяти ЭВМ (при условии $a_{ij} = a_{ji}$ объем памяти, требуемый для запоминания матрицы A, уменьшается) при решении системы с вещественной положительно определеннной матрицей A методом квадратного корня.

Многие пакеты прикладных программ для решения краевых задач математической физики методом конечных элементов организованы по следующей схеме. После формирования матрицы системы A путем перестановки строк и столбцов (одновременно переставляются i-я и j-я строки и i-й и j-й столбцы) система преобразуется к виду с наименьшей шириной ленты. Далее применяется метод квадратного корня. При этом с целью уменьшения объема вычислений при решении системы $A\mathbf{x} = \mathbf{b}$ с другими правыми частями матрица S запоминается.

Замечание. Часто этот метод уступает по эффективности итерационным методам.

Задача 4. Оценить число арифметических операций и объем требуемой памяти метода квадратного корня в случае матриц ленточной структуры.

Если есть подозрение, что реально полученное решение $\mathbf{x}^1$ сильно искажено вычислительной погрешностью, то можно поступить следующим образом. Определим вектор $\mathbf{b}^1 = \mathbf{b} - A\mathbf{x}^1$. Погрешность $\mathbf{r}^1 = \mathbf{x} - \mathbf{x}^1$ удовлетворяет системе уравнений

$$A\mathbf{r}^1 = A\mathbf{x} - A\mathbf{x}^1 = \mathbf{b}^1. \tag{7}$$

Решая эту систему в условиях реальных округлений, получаем приближение $\mathbf{r}^{(1)}$ к $\mathbf{r}^1$. Полагаем $\mathbf{x}^2 = \mathbf{x}^1 + \mathbf{r}^{(1)}$. Если точность нового приближения представляется неудовлетворительной, то повторяем эту операцию. При решении системы (7) над компонентами правой части производятся те же линейные операции, что и над компонентами правой части при решении системы (1). Поэтому при вычислениях на ЭВМ с плавающей запятой естественно ожидать, что относительные погрешности решений этих систем будут одного порядка. Поскольку погрешности округлений обычно малы, то $\|\mathbf{b}^1\| \ll \|\mathbf{b}\|$; тогда $\|\mathbf{r}^1\| \ll \|\mathbf{x}^1\|$, и, как правило, решение (7) определится с существенно меньшей абсолютной погрешностью, чем решение системы (1). Таким образом, применение описанного приема приводит к повышению точности приближенного решения.

Особенно удобно применять этот прием, когда по ходу вычислений в памяти ЭВМ сохраняются матрицы B и D. Тогда для каждого уточнения требуется найти вектор $\mathbf{b}^k = \mathbf{b} - A\mathbf{x}^k$ и решить две системы с треугольными матрицами. Это потребует всего $N_1 \sim 4m^2$ арифметических операций, что составит малую долю от числа операций $N_0 \sim 2m^3/3$, требующихся для представления матрицы A в виде $A = BD$.

Идея описанного приема последовательного уточнения приближений к решению часто реализуется в такой форме. Пусть матрица B близка в каком-то

смысле к матрице A, но решение системы $B\mathbf{x} = \mathbf{c}$ требует существенно меньшего объема вычислений по сравнению с решением системы $A\mathbf{x} = \mathbf{b}$. Решение системы $B\mathbf{x} = \mathbf{b}$ принимаем в качестве первого приближения $\mathbf{x}^1$ к решению. Разность $\mathbf{x} - \mathbf{x}^1$ удовлетворяет системе уравнений

$$A(\mathbf{x} - \mathbf{x}') = \mathbf{b} - A\mathbf{x}^1.$$

Вместо решения этой системы находим решение системы

$$B\mathbf{r}^1 = \mathbf{b} - A\mathbf{x}^1$$

и полагаем $\mathbf{x}^2 = \mathbf{x}^1 + \mathbf{r}^1$. Таким образом, каждое приближение находится из предыдущего по формуле

$$\mathbf{x}^{n+1} = \mathbf{x}^n + B^{-1}(\mathbf{b} - A\mathbf{x}^n) = (E - B^{-1}A)\mathbf{x}^n + B^{-1}\mathbf{b}.$$

Если матрицы A и B достаточно близки, то матрица $E - B^{-1}A$ имеет малую норму и такой итерационный процесс быстро сходится (см. также § 10).

Значительно более редкой, чем задача решения системы уравнений, является задача обращения матриц. Для обратной матрицы $X = A^{-1}$ имеем равенство $AX = BDX = E$. Таким образом, для нахождения матрицы X достаточно последовательно решить две матричные системы $BY = E$, $DX = Y$. Нетрудно подсчитать, что при нахождении на таком пути матрицы A^{-1} общий объем вычислений составит $N_2 \sim 2m^3$ арифметических операций.

В случае необходимости уточнения приближения к обратной матрице могут производиться при помощи итерационного процесса $X_k = X_{k-1}(2E - AX_{k-1})$. Для исследования сходимости итерационного процесса рассмотрим матрицы $G_k = E - AX_k$. Имеем равенство

$$G_k = E - AX_k = E - AX_{k-1}(2E - AX_{k-1}) = (E - AX_{k-1})^2 = G_{k-1}^2.$$

Отсюда получаем цепочку равенств

$$G_k = G_{k-1}^2 = G_{k-2}^4 = \cdots = G_0^{2^k}.$$

Поскольку

$$A^{-1} - X_k = A^{-1}(E - AX_k) = A^{-1}G_k = A^{-1}G_0^{2^k},$$

то имеем оценку

$$\|A^{-1} - X_k\| \leqslant \|A^{-1}\| \cdot \|G_0\|^{2^k}.$$

Таким образом, при достаточно хорошем начальном приближении, т. е. если $\|E - AX_0\| \leqslant 1$, этот итерационный процесс сходится со скоростью более быстрой, чем геометрическая прогрессия.

§ 2. Метод отражений

В настоящее время разработано так много точных методов численного решения систем линейных алгебраических уравнений, что даже простое перечисление их затруднительно. Большинство этих методов, как и метод исключения Гаусса, основано на переходе от заданной системы $A\mathbf{x} = \mathbf{b}$ к новой системе $CA\mathbf{x} = C\mathbf{b}$ такой, что система $B\mathbf{x} = \mathbf{d}$, где $B = CA$ и $\mathbf{d} = C\mathbf{b}$, решается проще, чем исходная. При выборе подходящей матрицы C нужно учитывать по крайней мере следующие два фактора. Во-первых, ее вычисление не должно быть чересчур сложным и трудоемким. Во-вторых, умножение на матрицу C не должно в каком-то смысле портить матрицу A (мера обусловленности матрицы не должна меняться сильно (см. § 11)).

Этим условиям в определенной степени удовлетворяет описываемый ниже *метод отражений*. Среди методов, требующих для своей реализации числа операций $N \sim 2m^3/3$, этот метод в настоящее время рассматривается как один из наиболее устойчивых к вычислительной погрешности.

Рассмотрим случай вещественной матрицы A. Если $\mathbf{w}$ — некоторый вектор-столбец единичной длины, $(\mathbf{w}, \mathbf{w}) = 1$, то матрицу

$$U = E - 2\mathbf{w}\mathbf{w}^T$$

называют *матрицей отражений*. Под $\mathbf{w}\mathbf{w}^T$ здесь понимается матрица, являющаяся произведением вектора-столбца $\mathbf{w}$ на вектор-строку $\mathbf{w}^T$, т. е. $\mathbf{w}\mathbf{w}^T = (w_{ij})$, где $w_{ij} = w_i w_j$. Из определения следует, что $\mathbf{w}\mathbf{w}^T$ — симметричная матрица.

Непосредственной проверкой убеждаемся, что $U = U^T$ и

$$UU^T = (E - 2\mathbf{w}\mathbf{w}^T)(E - 2\mathbf{w}\mathbf{w}^T)^T = E - 2\mathbf{w}\mathbf{w}^T - 2\mathbf{w}\mathbf{w}^T + 4\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T = E;$$

здесь мы воспользовались тем, что

$$\mathbf{w}^T\mathbf{w} = (\mathbf{w}, \mathbf{w}) = 1. \tag{1}$$

Таким образом, матрица U — симметричная и ортогональная.

> Напомним один факт из алгебры. Пусть U и B — две матрицы порядка m, B — многочлен от U, $B = P_l(U)$. Тогда можно переупорядочить их собственные значения так, что $\lambda_j^B = P_l(\lambda_j^U)$ при $j = 1, \ldots, m$.

Поскольку U симметрична и $U^2 = UU^T = E$, а все собственные числа E равны 1, то все собственные числа матрицы U удовлетворяют условию $\lambda_U^2 = 1$, т. е. равны или $+1$ или -1.

Собственному значению -1 отвечает собственный вектор $\mathbf{w}$. В самом деле,

$$U\mathbf{w} = \mathbf{w} - 2\mathbf{w}\mathbf{w}^T\mathbf{w} = \mathbf{w} - 2\mathbf{w} = -\mathbf{w}. \tag{2}$$

Все векторы, ортогональные вектору $\mathbf{w}$, являются собственными. Им соответствует собственное значение, равное $+1$. Действительно, пусть $(\mathbf{v}, \mathbf{w}) = 0$. Тогда имеем

$$U\mathbf{v} = \mathbf{v} - 2\mathbf{w}\mathbf{w}^T\mathbf{v} = \mathbf{v} - 2\mathbf{w}(\mathbf{w}, \mathbf{v}) = \mathbf{v}. \tag{3}$$

Представим произвольный вектор $\mathbf{y}$ в виде $\mathbf{y} = \mathbf{z} + \mathbf{v}$, где $\mathbf{z} = \gamma\mathbf{w}$, $(\mathbf{v}, \mathbf{w}) = 0$. Для этого следует взять в качестве $\mathbf{z}$ проекцию вектора $\mathbf{y}$ на вектор $\mathbf{w}$, т. е. $\mathbf{z} = (\mathbf{y}, \mathbf{w})\mathbf{w}$, и $\mathbf{v} = \mathbf{y} - (\mathbf{y}, \mathbf{w})\mathbf{w}$. Вследствие (2) и (3) имеем $U\mathbf{y} = -\mathbf{z} + \mathbf{v}$. Таким образом, $U\mathbf{y}$ есть зеркальное отражение вектора $\mathbf{y}$ относительно гиперплоскости, ортогональной вектору $\mathbf{w}$.

Используя геометрические свойства матрицы отражений, нетрудно решить следующую задачу: подобрать вектор $\mathbf{w}$ в матрице отражений так, чтобы заданный вектор $\mathbf{y} \neq \mathbf{0}$ имел в результате преобразования $U\mathbf{y}$ матрицей отражения $U = E - 2\mathbf{w}\mathbf{w}^T$ направление заданного единичного вектора $\mathbf{e}$.

Так как U — ортогональная матрица, а при ортогональных преобразованиях длины векторов сохраняются, то мы должны иметь $U\mathbf{y} = \alpha\mathbf{e}$ или $U\mathbf{y} = -\alpha\mathbf{e}$, , где $\alpha = \sqrt{(\mathbf{y}, \mathbf{y})}$. Поэтому направление, перпендикулярное плоскости отражения, будет определяться либо вектором $\mathbf{y} - \alpha\mathbf{e}$, либо вектором $\mathbf{y} + \alpha\mathbf{e}$ (см. рис. 6.2.1).

Таким образом, векторы $\mathbf{w}_1 = \pm\rho_1^{-1}(\mathbf{y} - \alpha\mathbf{e})$ или $\mathbf{w}_2 = \pm\rho_2^{-1}(\mathbf{y} + \alpha\mathbf{e})$, где $\rho_1 = \sqrt{(\mathbf{y} - \alpha\mathbf{e}, \mathbf{y} - \alpha\mathbf{e})}$, $\rho_2 = \sqrt{(\mathbf{y} + \alpha\mathbf{e}, \mathbf{y} + \alpha\mathbf{e})}$, будут искомыми. Ясно, что данный процесс всегда осуществим. Если векторы $\mathbf{y}$ и $\mathbf{e}$ коллинеарны, а в этом случае либо ρ_1, либо ρ_2 будет равно нулю, то никаких отражений делать не надо.

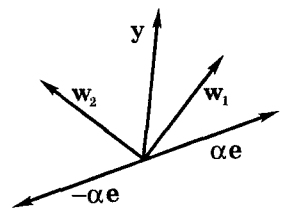


Рис. 6.2.1

Матрицы отражения нашли широкое применение при численном решении различных задач линейной алгебры (в частности, в рассматриваемой нами задаче приведения матрицы системы уравнений к треугольному виду).

Лемма. *Произвольная квадратная матрица может быть представлена в виде произведения ортогональной и верхней треугольной матриц.*

Доказательство. Пусть дана квадратная матрица порядка m. Будем приводить ее к правой треугольной матрице путем последовательного умножения слева на ортогональные матрицы. На первом шаге приведения рассмотрим в качестве вектора $\mathbf{y}$ из предыдущего рассуждения первый столбец матрицы A:

$$\mathbf{y}_1 = (a_{11}, \ldots, a_{m1})^T.$$

Если $a_{21} = a_{31} = \cdots = a_{m1} = 0$, то переходим к следующему шагу, положив $A^{(1)} = A$, $U_1 = E$ и введя обозначения $a_{ij}^{(1)} = a_{ij}$. В противном случае

умножаем матрицу A слева на матрицу отражения $U_1 = E_m - 2\mathbf{w}_1\mathbf{w}_1^T$, где $\mathbf{w}_1$ подбирается так, чтобы вектор $U_1\mathbf{y}_1$ был коллинеарен вектору $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$. Здесь и далее E_q — единичная матрица размерности q.

На этом первый шаг закончен, и на следующем шаге будем рассматривать матрицу $A^{(1)}$ с элементами $a_{ij}^{(1)}$, которая либо равна A, если имеет место первый случай, либо $A^{(1)} = U_1A$, если имеет место второй случай.

Пусть мы уже осуществили $l-1 > 0$ шагов и пришли к матрице $A^{(l-1)}$ с элементами $a_{ij}^{(l-1)}$ такими, что $a_{ij}^{(l-1)} = 0$ при $i > j$, $j = 1, 2, \ldots, l-1$. В пространстве R_{m-l+1} векторов размерности $m-l+1$ рассмотрим вектор

$$\mathbf{y}_l = (a_{l,l}^{(l-1)}, a_{l+1,l}^{(l-1)} \ldots, a_{m,l}^{(l-1)})^T.$$

Если $a_{l+1,l}^{(l-1)} = a_{l+2,l}^{(l-1)} = \cdots = 0$, то переходим к следующему шагу, полагая $A^{(l)} = A^{(l-1)}$, $U_l = E$. В противном случае строим матрицу отражения $V_l = E_{m-l+1} - 2\mathbf{w}_l\mathbf{w}_l^T$ (размеры матрицы V_l и вектора $\mathbf{w}_l$ равны $m-l+1$), переводящую вектор $\mathbf{y}_l$ в вектор, коллинеарный $\mathbf{e}_l = (1, 0, \ldots, 0)^T \in R_{m-l+1}$, и переходим к матрице

$$A^{(l)} = U_lA^{(l-1)};$$

здесь $U_l = \begin{pmatrix} E_{l-1} & 0 \\ 0 & V_l \end{pmatrix}$. Ясно, что процесс всегда осуществим, и после $(m-1)$-го шага мы приходим к матрице

$$A^{(m-1)} = U_{m-1}U_{m-2}\ldots U_1A,$$

имеющей правую треугольную форму.

Если обозначить $U_{m-1}U_{m-2}\ldots U_1 = U$, то из последнего равенства следует, что $A = U^TA^{(m-1)}$, где U^T — ортогональная, а $A^{(m-1)}$ — правая треугольная матрицы. Лемма доказана.

Вернемся к решению системы $A\mathbf{x} = \mathbf{b}$. С помощью указанных преобразований отражения последовательно приводим ее к эквивалентному виду

$$A^{(m-1)}\mathbf{x} = U\mathbf{b},$$

где $A^{(m-1)}$ — правая треугольная матрица. Если все диагональные элементы $A^{(m-1)}$ отличны от нуля, то последовательно находим $x_m, \ldots, x_1$. Если же хотя бы один из диагональных элементов равен нулю, то последняя система вырождена и в силу эквивалентности вырождена и исходная система.

Задача 1. Получить асимптотику числа операций метода отражений при $m \to \infty$.

Рассмотрим случай системы уравнений $A\mathbf{x} = \mathbf{b}$ с комплексными A и $\mathbf{b}$. Пусть

$$A = A_1 + iA_2, \quad \mathbf{b} = \mathbf{b}_1 + i\mathbf{b}_1, \quad \mathbf{x} = \mathbf{x}_1 + i\mathbf{x}_2.$$

Исходная система уравнений равносильна системе

$$C\mathbf{y} = \mathbf{d} \tag{4}$$

с вещественными C и $\mathbf{d}$:

$$C = \begin{pmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}.$$

Поэтому вместо непосредственного решения исходной задачи можно перейти к решению задачи (4) и применить для решения последней метод отражений.

Однако возможен и другой путь — непосредственное применение метода отражений к исходной системе $A\mathbf{x} = \mathbf{b}$. Здесь матрица отражения $U = E - 2\mathbf{w}\mathbf{w}^*$, $\mathbf{w}^* = (\bar{w}_1, \ldots, \bar{w}_m)^T$ будет унитарной с собственными значениями вида $\lambda_U = e^{i\varphi}$. (Через $\bar{z}$ обозначено комплексное число, сопряженное с z.)

Задача 2. *Перенести метод отражений на случай комплексных матриц.*

Задача 3. *Исследовать метод отражений в случае его применения для решения систем уравнений с ленточной матрицей.*

§ 3. Метод простой итерации

Простейшим итерационным методом решения систем линейных уравнений является *метод простой итерации.* Система уравнений

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

преобразуется к виду

$$\mathbf{x} = B\mathbf{x} + \mathbf{c}, \tag{2}$$

и ее решение находится как предел последовательности

$$\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}. \tag{3}$$

Всякая система

$$\mathbf{x} = \mathbf{x} - D(A\mathbf{x} - \mathbf{b}) \tag{4}$$

имеет вид (2) и при $\det D \neq 0$ эквивалентна системе (1). В то же время всякая система (2), эквивалентная (1), записывается в виде (4) с матрицей $D = (E - B)A^{-1}$.

Теорема (о достаточном условии сходимости метода простой итерации). *Если* $\|B\| < 1$, *то система уравнений* (2) *имеет единственное решение и*

итерационный процесс (3) *сходится к решению со скоростью геометрической прогрессии.*

Доказательство. Для всякого решения системы (2) имеет место $\|\mathbf{x}\| \leqslant \|B\|\,\|\mathbf{x}\| + \|\mathbf{c}\|$, поэтому справедливо неравенство $\|\mathbf{x}\|\,(1 - \|B\|) \leqslant \|\mathbf{c}\|$ или $\|\mathbf{x}\| \leqslant (1 - \|B\|)^{-1}\|\mathbf{c}\|$. Отсюда следует существование и единственность решения однородной системы $\mathbf{x} = B\mathbf{x}$, а следовательно, и системы (2). Пусть $\mathbf{X}$ — решение системы (2). Из (2) и (3) получаем уравнение относительно погрешности $\mathbf{r}^n = \mathbf{x}^n - \mathbf{X}$:

$$\mathbf{r}^{n+1} = B\mathbf{r}^n. \tag{5}$$

Из (5) получаем равенство

$$\mathbf{r}^n = B^n\mathbf{r}^0. \tag{6}$$

Отсюда следует, что $\|\mathbf{r}^n\| \leqslant \|B\|^n\|\mathbf{r}^0\| \to 0$. Теорема доказана.

Качество итерационного процесса удобно характеризовать скоростью убывания отношения погрешности после n итераций к начальной погрешности:

$$s_n = \sup_{\mathbf{x}^0 \neq \mathbf{X}} \frac{\|\mathbf{r}^n\|}{\|\mathbf{r}^0\|} = \sup_{\mathbf{r}^0 \neq 0} \frac{\|B^n\mathbf{r}^0\|}{\|\mathbf{r}^0\|} = \|B^n\|.$$

Можно гарантировать, что величина $s_n \leqslant \varepsilon$, если $\|B\|^n \leqslant \varepsilon$, т. е. при

$$n \geqslant n_\varepsilon = \ln(\varepsilon^{-1}) / \ln(\|B\|^{-1}). \tag{7}$$

Если существуют постоянные $\gamma_{\alpha\beta}$, $\gamma_{\beta\alpha}$ такие, что при $\mathbf{x} \neq 0$

$$\|\mathbf{x}\|_\beta / \|\mathbf{x}\|_\alpha \leqslant \gamma_{\alpha\beta}, \quad \|\mathbf{x}\|_\alpha / \|\mathbf{x}\|_\beta \leqslant \gamma_{\beta\alpha},$$

то нормы $\|\mathbf{x}\|_\alpha$ и $\|\mathbf{x}\|_\beta$ называются *эквивалентными*. Имеем

$$\|\mathbf{r}^n\|_\beta \leqslant \gamma_{\alpha\beta}\|\mathbf{r}^n\|_\alpha \leqslant \gamma_{\alpha\beta}\|B\|_\alpha^n\,\|\mathbf{r}^0\|_\alpha \leqslant \gamma_{\alpha\beta}\,\gamma_{\beta\alpha}\|B\|_\alpha^n\,\|\mathbf{r}^0\|_\beta.$$

Таким образом, если условие доказанной теоремы выполнено для нормы $\|\cdot\|_\alpha$, то утверждение справедливо относительно любой эквивалентной ей нормы.

Любые две нормы в конечномерном пространстве являются эквивалентными. В частности, нормы $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_\infty$, вычисляемые соответственно по формулам (2), (3), (4), приведенным во введении к настоящей главе, эквивалентны между собой вследствие справедливости цепочки неравенств

$$\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1 \leqslant m\|\mathbf{x}\|_\infty.$$

Лемма. *Пусть все собственные значения λ_i матрицы B лежат в круге $|\lambda| \leqslant q$, причем собственным значениям, по модулю равным q, соответствуют жордановы клетки размерности* 1. *Тогда существует матрица $\Lambda = D^{-1}BD$ с нормой $\|\Lambda\|_\infty \leqslant q$.*

Доказательство. Положим $\eta = q - \max\limits_{|\lambda_i|<q} |\lambda_i|$. Собственными значениями матрицы $\eta^{-1}B$ будут $\eta^{-1}\lambda_i$. Преобразуем матрицу $\eta^{-1}B$ к жордановой форме

$$D^{-1}(\eta^{-1}B)D = \begin{pmatrix} \eta^{-1}\lambda_1 & \alpha_{12} & 0 & \cdots \\ 0 & \eta^{-1}\lambda_2 & \alpha_2 & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{pmatrix},$$

где $\alpha_{i,i+1}$ принимают значения 0 или 1. После умножения на η получим

$$\Lambda = D^{-1}BD = \begin{pmatrix} \lambda_1 & \alpha_{12}\eta & 0 & \cdots \\ 0 & \lambda_2 & \alpha_{23}\eta & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{pmatrix}.$$

Если $|\lambda_i| = q$, то согласно условиям леммы, $\alpha_{i,i+1} = 0$. Отсюда следует, что $|\lambda_i| + |\alpha_{i,i+1}\eta| = q$. Если $|\lambda_i < q$, то

$$|\lambda_i| + |\alpha_{i,i+1}\eta| \leqslant \max_{|\lambda_i|<q} |\lambda_i| + \eta = q.$$

Таким образом, $\|\Lambda\|_\infty = \max\limits_i \left(|\lambda_i| + |\alpha_{i,i+1}\eta|\right) \leqslant q$.

Теорема (о необходимом и достаточном условии сходимости метода простой итерации). *Пусть система* (2) *имеет единственное решение. Итерационный процесс* (3) *сходится к решению системы* (2) *при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы* B *по модулю меньше* 1.

Доказательство. Достаточность. Возьмем произвольное q в пределах $\max\limits_i |\lambda_i| < q < 1$. Условие леммы выполнено по отношению к этому q, поэтому существует матрица D такая, что $\|\Lambda\|_\infty \leqslant q$ при $\Lambda = D^{-1}BD$. Поскольку $B = D\Lambda D^{-1}$, то

$$B^n = D\Lambda D^{-1}D\cdots D^{-1}D\Lambda D^{-1} = D\Lambda^n D^{-1}.$$

Поэтому

$$\|B^n\|_\infty \leqslant \|D\|_\infty \|D^{-1}\|_\infty q^n \to 0$$

и

$$\|\mathbf{x}^n - \mathbf{X}\|_\infty \leqslant \|D\|_\infty \|D^{-1}\|_\infty q^n \|\mathbf{x}^0 - \mathbf{X}\|_\infty \to 0 \tag{8}$$

при $n \to \infty$. Следовательно, и $\|\mathbf{x}^n - \mathbf{X}\|_1, \|\mathbf{x}^n - \mathbf{X}\|_2 \to 0$.

Если χ_i — координатные орты, $\mathbf{x} = (x_1, \ldots, x_m)^T$, то $\mathbf{x} = \sum\limits_i x_i\chi_i$. Пусть $\|\cdot\|$ — некоторая норма, тогда

$$\|\mathbf{x}\| \leqslant \sum_i |x_i|\,\|\chi_i\| \leqslant \|\mathbf{x}\|_\infty \sum_i \|\chi_i\|.$$

Поэтому при любой норме $\|\cdot\|$ имеем

$$\|\mathbf{x}^n - \mathbf{X}\| \leqslant \Big(\sum_i \|\chi_i\|\Big)\|D\|_\infty \|D^{-1}\|_\infty q^n \|\mathbf{x}^0 - \mathbf{X}\|_\infty \to 0. \tag{9}$$

Соотношения (8), (9) означают также, что любые нормы погрешности убывают быстрее любой геометрической прогрессии со знаменателем, большим $\max\limits_i |\lambda_i|$.

Необходимость. Пусть $|\lambda_i| \geqslant 1$ и $\mathbf{e}_1$ — соответствующий собственный вектор матрицы B. Тогда при начальном приближении $\mathbf{x}^0 = \mathbf{X} + c\mathbf{e}^1$, $c \neq 0$, имеем

$$\mathbf{r}^0 = c\mathbf{e}_1 \quad \text{и} \quad \mathbf{r}^n = \lambda_l^n c\mathbf{e}_1 \not\to 0 \quad \text{при} \quad n \to \infty.$$

Задача 1. Пусть все собственные значения матрицы B, за исключением простого $\lambda_1 = 1$, лежат внутри единичного круга и система (2) имеет решение $\mathbf{X}$. Решением системы будут также все $\mathbf{x} = \mathbf{X} + c\mathbf{e}_1$. Доказать, что итерационный процесс (3) сходится к одному из таких решений.

§ 4. Особенности реализации метода простой итерации на ЭВМ

Если все собственные значения матрицы B лежат внутри единичного круга, то может показаться, что не возникает никаких проблем относительно поведения метода в реальных условиях ограниченности порядков чисел в ЭВМ и присутствия округлений. В обоснование этого иногда приводят следующий довод: возмущения приближений в результате округлений равносильны возмущениям начальных условий итерационного процесса. Поскольку процесс сходящийся, «самоисправляющийся», эти возмущения в конце концов затухнут, и будет получено хорошее приближение к решению исходной задачи.

Однако при решении некоторых систем возникала следующая ситуация. Все собственные значения матрицы B лежали в круге $|\lambda| \leqslant 1/2$, а итерационный процесс останавливался после некоторого числа итераций из-за переполнения порядков чисел в ЭВМ. В других случаях такого переполнения не происходило, но векторы $\mathbf{x}^n$, получаемые при вычислениях, не сходились к решению. Последний случай особенно опасен по следующей причине. Можно необоснованно решить, что при условии $\max|\lambda_i| \leqslant 1/2$ какое-то определенное число итераций, например 100, заведомо достаточно для получения решения с требуемой точностью. Затем производим эти 100 итераций и рассматриваем полученный результат как требуемый. Поэтому наличие подобных явлений послужило толчком к более детальному исследованию итерационных процессов и формированию новых понятий в теории операторов.

Чтобы понять сущность явления, полезно построить пример, где это явление прослеживается в явном виде. В качестве модели выберем итерационный процесс, соответствующий двухдиагональной матрице

$$B_0 = \begin{pmatrix} \alpha & \beta & 0 & \cdots & 0 \\ 0 & \alpha & \beta & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & \alpha \end{pmatrix}.$$

При возведении матрицы B_0 в степень n, получается треугольная матрица

$$B_0^n = \left(b_{ij}^{(n)}\right) = \begin{pmatrix} \alpha^n & C_n^1\alpha^{n-1}\beta & C_n^2\alpha^{n-2}\beta^2 & \cdots \\ 0 & \alpha^n & C_n^1\alpha^{n-1}\beta & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{pmatrix}$$

с элементами $b_{ij}^{(n)} = C_n^{j-i}\alpha^{n-(j-i)}\beta^{j-i}$. Если $\mathbf{r}^0 = (0, \ldots, 0, 1)^T$, то

$$\mathbf{r}^n = B_0^n\mathbf{r}^0 = (b_{1m}^{(n)}, \ldots, b_{mm}^{(n)})^T, \quad \|\mathbf{r}^n\|_1 = \sum_{i=1}^{m} |b_{im}^{(n)}|.$$

При $n < m$ последнее выражение упрощается:

$$\|\mathbf{r}^n\|_1 = \sum_{i=1}^{m} C_n^{m-i}|\alpha|^{n-(m-i)}\,|\beta|^{m-i} =$$

$$= \sum_{k=0}^{m-1} C_n^k|\alpha|^{n-k}\,|\beta|^k = \sum_{k=0}^{n} C_n^k|\alpha|^{n-k}\,|\beta|^k = \Big(|\alpha| + |\beta|\Big)^n.$$

Рассмотрим случай $|\alpha| < 1$, $|\alpha| + |\beta| > 1$, $|\beta/(1-\alpha)| < 1$. Пусть $\mathbf{c} = \mathbf{c}^0 = (0, \ldots, 0, 1)^T$. Непосредственно проверяется, что при таком $\mathbf{c}$ решением рассматриваемой системы будет

$$\mathbf{X}^0 = \left(\frac{1}{1-\alpha}\left(\frac{\beta}{1-\alpha}\right)^{m-1}, \ldots, \frac{1}{1-\alpha}\right)^T.$$

Справедлива оценка

$$\|\mathbf{X}^0\|_1 \leqslant \omega,$$

где

$$\omega = \frac{1}{|1-\alpha|}\sum_{k=0}^{\infty}\left|\frac{\beta}{1-\alpha}\right|^k = \frac{1}{|1-\alpha|\left(1-\left|\frac{\beta}{1-\alpha}\right|\right)}.$$

При начальном приближении $\mathbf{x}^0 = \mathbf{X}^0 + \mathbf{c}^0$ имеем $\mathbf{r}^0 = \mathbf{c}^0$ и, согласно проводившимся выше построениям,

$$\|\mathbf{r}^n\|_1 = (|\alpha| + |\beta|)^n \qquad \text{для} \quad n < m.$$

Выберем m таким, чтобы число $\sigma = \left[(|\alpha| + |\beta|)^{m-1} - \omega\right]/m$ превосходило пределы, допустимые в ЭВМ. Из полученных ранее соотношений следует, что

$$\|\mathbf{x}^{m-1}\|_\infty \geqslant \|\mathbf{x}^{m-1}\|_1/m \geqslant \left(\|\mathbf{r}^{m-1}\|_1 - \|\mathbf{x}^0\|_1\right)/m \geqslant \sigma.$$

Поэтому построенный пример обладает следующими свойствами: норма начального приближения невелика, итерационный процесс сходится при отсутствии округлений и ограничения на порядки чисел в ЭВМ, но останавливается не позднее чем при $n = m - 1$ из-за недопустимо больших значений компонент приближений.

Обратимся к реальной ситуации, когда на каждом шаге вычислений происходят округления. Рассмотрим подробнее случай, когда переполнение не происходит. Вместо $\mathbf{x}^n$ получаются векторы $\mathbf{x}^{*n}$, связанные соотношениями

$$\mathbf{x}^{*n+1} = B\mathbf{x}^{*n} + \mathbf{c} + \rho^n,$$

где ρ^n — суммарное округление на шаге итерации.

Отсюда и из (3.2) получается уравнение относительно погрешности $\mathbf{r}^{*n} = \mathbf{x}^{*n} - \mathbf{X}$:

$$\mathbf{r}^{*n+1} = \rho^n + B\mathbf{r}^{*n}. \tag{1}$$

Выражая каждое $\mathbf{r}^{*n}$ через предыдущее, получаем

$$\begin{aligned}\mathbf{r}^{*n} &= \rho^{n-1} + B\mathbf{r}^{*n-1} = \rho^{n-1} + B(\rho^{n-2} + B\mathbf{r}^{*n-2}) = \\ &= \rho^{n-1} + B\rho^{n-2} + \cdots + B^{n-1}\rho^0 + B^n\mathbf{r}^0.\end{aligned} \tag{2}$$

Как мы видели, норма $\|B_0^n\|$ при $|\alpha| < 1$, $|\alpha|+|\beta| > 1$ имеет следующий характер поведения: при малых n она имеет тенденцию к возрастанию, при больших n стремится к нулю. (Можно показать, что максимальное значение $\varphi(B_0) = \max\limits_n \|B_0^n\|$ достигается при значении $n = n_0$ порядка m.) При таком характере поведения норм $\|B^n\|$ может возникнуть следующая ситуация. Величина $\max\limits_n \|\mathbf{x}^{*n}\|$ не настолько велика, чтобы происходило переполнение и остановка ЭВМ; в то же время $\varphi(B)2^{-t} \gg R$, где R — максимально допустимая погрешность решения. Поэтому, как правило, при $n > n_0$ среди слагаемых в правой части (2) присутствует слагаемое $B^{n_0}\rho^{n-1-n_0}$ с нормой, много большей, чем R. В результате установление приближений $\mathbf{x}^n$ с приемлемой точностью не происходит.

Подведем некоторый итог проведенных построений. Матрицы высокой размерности обладают свойствами, существенно отличными от свойств матриц малой размерности. Кроме собственных значений у таких матриц есть *почти собственные значения*, т. е. λ такие, что $\|A\mathbf{x} - \lambda\mathbf{x}\| \leqslant \varepsilon\|\mathbf{x}\|$ при $\|\mathbf{x}\| \neq 0$ и очень малом ε.

Например, в случае матрицы B_0 при любом λ_n, лежащем в круге $|\alpha - \lambda| < |\beta|$, можно построить вектор $\mathbf{x}_\lambda$ такой, что $\|B_0\mathbf{x}_\lambda - \lambda\mathbf{x}_\lambda\|_\infty \leqslant \varepsilon_\lambda\|\mathbf{x}_\lambda\|_\infty$, где $\varepsilon_\lambda = |\beta|\,|(\lambda - \alpha)/\beta|^m$.

Поведение степеней матрицы B^n при n порядка m определяется во многом такими «почти собственными векторами» $\mathbf{x}_\lambda$ и «почти собственными значениями» λ.

Задача 2. Построить «почти собственный вектор» $\mathbf{x}_\lambda$, соответствующий значению ε_λ, приведенному выше.

Суммарная вычислительная погрешность $\rho_n = \sum_{j=0}^{n-1} B^{n-1-j}\rho^j$ может оказаться большой не только из-за большой величины отдельных слагаемых, но и из-за того, что их много.

Пусть B — симметричная матрица и $\|B\|_2 = \max_i |\lambda_B^i| = \lambda_B^1 < 1$, $\mathbf{e}^1$ — соответствующий λ_B^1 нормированный собственный вектор. Предположим, что на каждом j-м шаге происходит округление $\rho^j = \rho\mathbf{e}^1$, где ρ порядка 2^{-t}. Имеем равенство

$$\rho_n = \rho\sum_{j=0}^{n-1}(\lambda_B^1)^j\mathbf{e}^1 = \rho\frac{1-(\lambda_B^1)^n}{1-\lambda_B^1}\mathbf{e}^1.$$

Поскольку число итераций берется таким, что $\|B^n\| \ll 1$, а $\|B^n\| = (\lambda_B^1)^n$, то можно считать, что $\|\rho_n\| \approx \rho/(1-\lambda_B^1)$. Таким образом, если λ_B^1 близко к 1, то суммарное влияние округлений на шагах интегрирования может оказаться довольно большим.

Покажем, что вычислительная погрешность такого порядка является неизбежной. Предположим, что вместо системы (3.2) решается система $\mathbf{x} = B\mathbf{x}+\mathbf{c}+\rho\mathbf{e}_1$. Разность $\bar{\mathbf{X}} - \mathbf{X}$ решений этих систем удовлетворяет соотношению $(\bar{\mathbf{X}} - \mathbf{X}) = B(\bar{\mathbf{X}}-X)+\rho\mathbf{e}_1$, отсюда $\bar{\mathbf{X}}-X = (E-B)^{-1}\rho\mathbf{e}_1 = (1-\lambda_B^1)^{-1}\rho\mathbf{e}_1$. Поэтому погрешность порядка $(1-\lambda_B^1)^{-1}\rho$ является неустранимой; возмущение приближений, создаваемое в ходе итераций, сравнимо с неустранимой погрешностью.

§ 5. δ^2-процесс практической оценки погрешности и ускорения сходимости

Рассмотрим вопрос об оценке погрешности приближенного решения системы уравнений. Если $\mathbf{X}^*$ — приближенное решение системы $A\mathbf{X} = \mathbf{b}$, а $\mathbf{X}$ — точное решение этой системы, то можно написать равенство

$$\|\mathbf{X}^* - \mathbf{X}\| = \|A^{-1}(A\mathbf{X}^* - \mathbf{b})\| \leqslant \|A^{-1}\|\,\|A\mathbf{X}^* - \mathbf{b}\|,$$

которое редко применяется из-за сложности оценки $\|A^{-1}\|$. Поэтому при практическом анализе погрешности приближений, получаемых итерационными методами, обычно вместо этой оценки используется рассматриваемая далее нестрогая, но более простая оценка погрешности, которая строится на основании дополнительной информации, получаемой в процессе вычислений.

Примем следующий *критерий разумности практической оценки погрешности*: $\mathbf{v}^n$ принимается за практическую погрешность приближения $\mathbf{x}^n$, стремящегося к $\mathbf{X}$ при $n \to \infty$, если

$$\|\mathbf{v}^n - (\mathbf{x}^n - \mathbf{X})\| \Big/ \|\mathbf{x}^n - \mathbf{X}\| \to 0 \quad \text{при} \quad n \to \infty. \tag{1}$$

Ясно, что тогда $\|\mathbf{v}^n\| \sim \|\mathbf{x}^n - \mathbf{X}\|$.

Рассмотрим метод простой итерации $\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}$. Для краткости изложения ограничимся случаем, когда матрица B простой структуры (т. е. ее жорданова форма диагональна и поэтому она обладает полной системой собственных векторов).

Пусть λ_i, $i = 1, \ldots, m$, — собственные значения матрицы B, занумерованные в порядке убывания $|\lambda_i|$, причем $1 > |\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant |\lambda_m|$, а $\mathbf{e}_i$, $\|\mathbf{e}_i\| = 1$, — соответствующие собственные векторы, образующие полную систему. Разложим вектор $\mathbf{r}^0$ по базису $\mathbf{e}_i$: $\mathbf{r}^0 = \sum c_i \mathbf{e}_i$. Тогда

$$\mathbf{r}^0 = \mathbf{x}^n - \mathbf{X} = B^n \mathbf{r}^0 = \sum c_i \lambda_i^n \mathbf{e}_i = c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n). \tag{2}$$

Здесь и далее выражение $\mathbf{x}^n = \mathbf{y}^n + O(\varepsilon_n)$ имеет следующий смысл:

$$\|\mathbf{x}^n - \mathbf{y}^n\| = O(\varepsilon_n) \quad \text{при} \quad n \to \infty.$$

Далее в этом параграфе $\|\mathbf{x}\|$ — это $\|\mathbf{x}\|_2$.

Укажем способ построения приближения к вектору $\mathbf{w}^n = c_1 \lambda_1^n \mathbf{e}_1$ на основании информации, получающейся в ходе вычислений. Согласно (2) имеем

$$\begin{aligned} \mathbf{x}^{n-2} - \mathbf{X} &= \mathbf{w}^n \lambda_1^{-2} + O\big(|\lambda_2|^n\big), \\ \mathbf{x}^{n-1} - \mathbf{X} &= \mathbf{w}^n \lambda_1^{-1} + O\big(|\lambda_2|^n\big), \\ \mathbf{x}^{n} - \mathbf{X} &= \mathbf{w}^n + O\big(|\lambda_2|^n\big). \end{aligned}$$

Вычитая друг из друга соседние соотношения, получим

$$\begin{aligned} \mathbf{x}^{n-1} - \mathbf{x}^{n-2} &= \mathbf{w}^n (1 - \lambda_1^{-1}) \lambda_1^{-1} + O\big(|\lambda_2|^n\big), \\ \mathbf{x}^{n} - \mathbf{x}^{n-1} &= \mathbf{w}^n (1 - \lambda_1^{-1}) + O\big(|\lambda_2|^n\big). \end{aligned} \tag{3}$$

Отсюда

$$\begin{aligned} (\mathbf{x}^n - \mathbf{x}^{n-1}, \mathbf{x}^n - \mathbf{x}^{n-1}) &= \|\mathbf{w}^n\|^2 |1 - \lambda_1^{-1}|^2 + O\big(\|\mathbf{w}^n\| \, |\lambda_2|^n\big), \\ (\mathbf{x}^{n-1} - \mathbf{x}^{n-2}, \mathbf{x}^n - \mathbf{x}^{n-1}) &= \|\mathbf{w}^n\|^2 |1 - \lambda_1^{-1}|^2 \lambda_1^{-1} + O\big(\|\mathbf{w}^n\| \, |\lambda_2|^n\big). \end{aligned} \tag{4}$$

Положим

$$\lambda_1^{(n)} = \frac{(\mathbf{x}^n - \mathbf{x}^{n-1}, \mathbf{x}^n - \mathbf{x}^{n-1})}{(\mathbf{x}^{n-1} - \mathbf{x}^{n-2}, \mathbf{x}^n - \mathbf{x}^{n-1})}.$$

Воспользуемся соотношениями (4) и в предположении $c_1 \neq 0$ поделим числитель и знаменатель в выражении для $\lambda_1^{(n)}$ на $\|\mathbf{w}^n\|^2 |1 - \lambda_1^{-1}|^2 \lambda_1^{-1}$; в результате получим

$$\lambda_1^{(n)} = \frac{\lambda_1 + O\left(\dfrac{|\lambda_2|^n}{\|\mathbf{w}^n\|}\right)}{1 + O\left(\dfrac{|\lambda_2|^n}{\|\mathbf{w}^n\|}\right)}.$$

Поскольку

$$\|\mathbf{w}^n\| = |c_1|\,|\lambda_1|^n, \tag{5}$$

то

$$\lambda_1^{(n)} = \lambda_1 + O\big(|\lambda_2/\lambda_1|^n\big). \tag{6}$$

Поделив второе из соотношений (3) на $1 - \left(\lambda_1^{(n)}\right)^{-1}$, получим

$$\frac{\mathbf{x}^n - \mathbf{x}^{n-1}}{1 - \left(\lambda_1^{(n)}\right)^{-1}} = \mathbf{w}^n \frac{1 - \lambda_1^{-1}}{1 - \left(\lambda_1^{(n)}\right)^{-1}} + O(|\lambda_2|^n) = \mathbf{w}^n + \mathbf{w}^n \frac{\lambda_1 - \lambda_1^{(n)}}{\lambda_1\left(\lambda_1^{(n)} - 1\right)} + O\big(|\lambda_2|^n\big).$$

Из (5), (6) следует $\left\|\mathbf{w}^n\left(\lambda_1 - \lambda_1^{(n)}\right)\right\| = O\big(|\lambda_2|^n\big)$; поэтому

$$\frac{\mathbf{x}^n - \mathbf{x}^{n-1}}{1 - \left(\lambda_1^{(n)}\right)^{-1}} = \mathbf{w}^n + O\big(|\lambda_2|^n\big).$$

Отсюда и из (2) получаем

$$\mathbf{x}^n - \mathbf{X} = \mathbf{v}^n + O\big(|\lambda_2|^n\big),$$

где $\mathbf{v}_n = (\mathbf{x}^n - \mathbf{x}^{n-1})/\big(1 - \left(\lambda_1^{(n)}\right)^{-1}\big)$. Заметим, что согласно (3), (6) $\|\mathbf{v}^n\| = |c_1|\,|\lambda_1|^n + O(|\lambda_2|^n)$. Из этих равенств вытекает, что $\mathbf{v}^n$ удовлетворяет критерию (1), и поэтому его можно принять за практическую погрешность приближения $\mathbf{x}^n$.

В случае $c_1 = \cdots = c_l = 0$, $c_{l+1} \neq 0$ проведенные рассуждения останутся в силе, если $|\lambda_{l+1}| > |\lambda_{l+2}|$. Во всех соотношениях следует заменить лишь λ_i, c_i, $\mathbf{e}_i$ при $i = 1, 2$ на λ_{l+i}, c_{l+i}, $\mathbf{e}_{l+i}$. Описанный способ получения оценки приближенного решения называется *δ^2-процессом*.

Если положить $\mathbf{y}^n = \mathbf{x}^n - \mathbf{v}^n$, то $\mathbf{y}^n - \mathbf{X} = O(|\lambda_2|^n)$, и поэтому $\mathbf{y}^n$, вообще говоря, является лучшим начальным условием для последующих итераций по сравнению с $\mathbf{x}^n$. Производя время от времени такие уточнения, иногда удается существенно уменьшить общее число итераций.

Для справедливости приближенного равенства

$$\mathbf{x}^n - \mathbf{X} \approx \mathbf{v}^n \tag{7}$$

необходимо, чтобы в правой части равенства

$$\mathbf{x}^n - \mathbf{X} = \sum_i c_i \lambda_i^n \mathbf{e}_i$$

одно из слагаемых преобладало над остальными. Если это так, то векторы $\mathbf{x}^n - \mathbf{x}^{n-1}$, $\mathbf{x}^{n-1} - \mathbf{x}^{n-2}$ приблизительно пропорциональны и

$$\mu_n = \frac{|(\mathbf{x}^{n-1} - \mathbf{x}^{n-2},\ \mathbf{x}^{n-1} - \mathbf{x}^n)|}{\|\mathbf{x}^{n-1} - \mathbf{x}^{n-2}\|\,\|\mathbf{x}^{n-1} - \mathbf{x}^n\|} \approx 1.$$

Таким образом, условие $\mu_1 \approx 1$ является необходимым для того, чтобы проводившиеся ранее построения были справедливы. Поэтому его можно принять за условие практической применимости (7).

Например, возможна следующая схема метода простой итерации с применением δ^2-процесса ускорения сходимости. Задаются некоторым η' в пределах $1 > \eta' > 0$ и малым $\eta > 0$. Если по ходу итераций оказалось, что $\mu_n \geqslant 1 - \eta$, то вычисляется $\mathbf{v}^n$ и вектор $\mathbf{y}^n$ принимается за начальное приближение для последующих итераций. Итерационный процесс прекращается, если $\mu_n \geqslant 1 - \eta'$ и $\|\mathbf{v}^n\| \leqslant \varepsilon$, где ε — требуемая точность.

Если η очень мало, то условие $\mu_n \geqslant 1 - \eta$ будет выполняться только после большого числа итераций, ускорение сходимости не будет иметь места. При большом η соотношения, положенные в основу наших построений, выполняются грубо, поэтому не исключено, что применение δ^2-процесса сходимости замедлит итерационный процесс. Картина итераций также осложняется наличием погрешности округлений, так что описанная выше схема требует практической отработки на большом числе примеров с целью выбора оптимальных η', η и указания нижней границы значений ε, при которых алгоритм применим. Если однородный итерационный процесс подвергается перестройке (в нашем случае при переходе от $\mathbf{x}^n$ к $\mathbf{y}^n$), то иногда полезно проверить, не ведет ли эта перестройка к ухудшению. В качестве критерия целесообразности перестройки можно взять некоторое соотношение, связывающее нормы невязок для $\mathbf{x}^n$, $\mathbf{y}^n$, например неравенство вида

$$\|(E - B)\mathbf{y}^n - \mathbf{c}\| \leqslant q\|(E - B)\mathbf{x}^n - \mathbf{c}\|.$$

Замечание о необходимости указания нижней грани значений ε вызывается следующим обстоятельством. Пусть для определенности $\lambda_1 > 0$. Уже при вычислении $\mathbf{x}^n$ по заданному $\mathbf{x}^{n-1}$ погрешности округления могут возмутить результат на величину $\delta\mathbf{x}^n$ с нормой порядка ρ. Следствием этого может явиться возмущение $\delta\mathbf{v}^n$, имеющее норму порядка $(1 - \lambda_1)^{-1}\rho$. Отсюда следует, что в случае $\varepsilon < (1 - \lambda_1)^{-1}\rho$ итерационный процесс может никогда не закончиться. Проведенные построения показывают, что при реализации метода возникает много таких моментов, разбор которых требует серьезной математической подготовки и проведения большой серии численных экспериментов. Поэтому, несмотря на «простоту» метода простой итерации, будет вполне оправданным создание стандартной программы этого метода.

§ 6. Оптимизация скорости сходимости итерационных процессов

Рассмотрим простейший итерационный способ решения системы уравнений $A\mathbf{x} = \mathbf{b}$:

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha(A\mathbf{x}^n - \mathbf{b}).$$

Мы видели, что скорость сходимости такого итерационного процесса существенно зависит от максимального модуля собственных значений матрицы $B = E - \alpha A$. Если $\lambda_1, \ldots, \lambda_n$ — собственные значения матрицы A, то $\max\limits_i |\lambda_i(B)| = \max\limits_i |1 - \alpha\lambda_i|$. Из рис. 6.6.1 видно, что при действительных собственных значениях различных знаков этот максимум больше 1 и итерационный процесс расходится.

Обратимся к часто встречающемуся случаю, когда все $\lambda_i > 0$. Значения λ_i бывают известны крайне редко, однако довольно типичен случай, когда известна оценка для этих чисел вида $0 < \mu \leqslant \lambda_i \leqslant M < \infty$ при всех i. Скорость сходимости итерационного процесса можно характеризовать величиной

$$\rho(\alpha) = \max_{\mu \leqslant \lambda \leqslant M} |1 - \alpha\lambda|.$$

Рассмотрим задачу минимизации $\rho(\alpha)$ за счет выбора α.

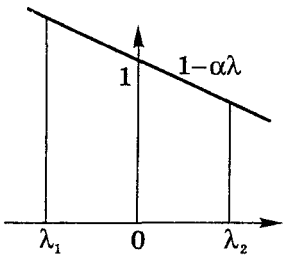


Рис. 6.6.1

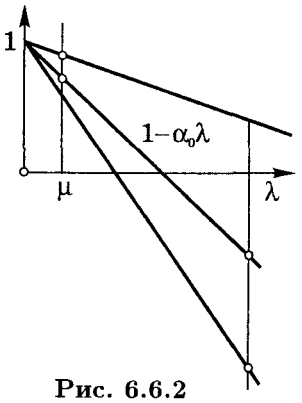


Рис. 6.6.2

Для нахождения $\min\limits_\alpha \rho(\alpha)$ удобно обратиться к геометрической картине (рис. 6.6.2). Ясно, что $\rho(\alpha) \geqslant 1$ при $\alpha \leqslant 0$. При $0 < \alpha \leqslant M^{-1}$ функция $1 - \alpha\lambda$ неотрицательна и монотонно убывает на отрезке $[\mu, M]$, поэтому $\rho(\alpha) = 1 - \alpha\mu$. При $M^{-1} < \alpha$ величина $1 - \alpha M$ отрицательна и модуль ее растет с ростом α. При некотором $\alpha = \alpha_0$ наступит момент, когда

$$1 - \alpha_0\mu = -(1 - \alpha_0 M), \tag{1}$$

и тогда $\rho(\alpha_0) = |1 - \alpha_0\mu|$. Если $\alpha < \alpha_0$, то $\rho(\alpha) = 1 - \alpha\mu > 1 - \alpha_0\mu = \rho(\alpha_0)$; если $\alpha_0 < \alpha$, то $\rho(\alpha) \geqslant |1 - \alpha M| = M\alpha - 1 \geqslant M\alpha_0 - 1 = \rho(\alpha_0)$.

Таким образом, значение $\alpha = \alpha_0$ является искомым. Решая уравнение (1) относительно α_0, получим $\alpha_0 = 2/(M + \mu)$. Отсюда

$$\rho(\alpha_0) = (M - \mu)/(M + \mu).$$

Задача 1. Доказать сходимость итерационного процесса при $\alpha = \|A\|^{-1}$.

На примере систем с матрицей $A > 0$ (здесь и далее неравенство $A > 0$ означает, что A — симметричная положительно определенная матрица) рассмотрим более формализованные постановки проблем оптимизации скорости сходимости итерационных процессов.

Если число ненулевых элементов матрицы много больше ее размерности, то операция умножения матрицы на вектор более трудоемка, чем умножение числа на вектор или сложение векторов. Поэтому при оценке трудоемкости итерационных процессов и оптимизации этих процессов далее за меру трудоемкости мы неявно принимаем число умножений матрицы A на вектор.

Всякая система $A\mathbf{x} = \mathbf{b}$ с $\det A \neq 0$, вообще говоря, может быть приведена (как говорят, *симметризована*) умножением обеих частей уравнения на матрицу A^T к системе с симметричной положительно определенной матрицей. В самом деле, система $A^TA\mathbf{x} = A^T\mathbf{b}$ эквивалентна исходной, матрица A^TA симметричная, так как $A^TA = (A^TA)^T$, и положительно определена, так как $(A^TA\mathbf{x}, \mathbf{x}) = \|A\mathbf{x}\|^2 > 0$ при $\mathbf{x} \neq 0$. По возможности стараются избегать симметризации, поскольку, как мы увидем далее, она часто приводит к ухудшению сходимости итерационных процессов.

Рассмотрим несколько более общий итерационный метод, чем метод простой итерации. А именно, в методе простой итерации

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau(A\mathbf{x}^k - \mathbf{b})$$

будем считать, что итерационный параметр τ может изменяться от шага к шагу. Тогда метод примет вид

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau_{k+1}(A\mathbf{x}^k - \mathbf{b}), \quad k = 0, 1, \ldots, \tag{2}$$

где $\mathbf{x}^0$ — некоторое начальное приближение.

Зададимся некоторым целым $n > 0$ и произведем n итераций по формуле (2). Согласно (2) погрешность $\mathbf{r}^k = \mathbf{x}^k - \mathbf{X}$ удовлетворяет соотношению

$$\mathbf{r}^{k+1} = \mathbf{r}^k - \tau_{k+1}A\mathbf{r}^k = (E - \tau_{k+1}A)\mathbf{r}^k. \tag{3}$$

Тогда через n шагов итерационного метода (2) погрешность $\mathbf{r}^n$ будет выражаться через погрешность начального приближения $\mathbf{r}^0$ следующим образом:

$$\mathbf{r}^n = (E - \tau_nA)\mathbf{r}^{n-1} = \cdots = (E - \tau_nA)\ldots(E - \tau_1A)\mathbf{r}^0, \tag{4}$$

где $\mathbf{r}^0 = \mathbf{x}^0 - \mathbf{X}$ — погрешность начального приближения.

Обозначим $Q_n(A) = (E - \tau_n A)\dots(E - \tau_1 A)$. Таким образом, оператор (матрица) $Q_n(A)$ связывает погрешности приближения на нулевом и n-м шагах итерационного процесса. Из (4) имеем

$$\|\mathbf{r}^n\|_2 \leqslant \|Q_n(A)\|_2 \, \|\mathbf{r}^0\|_2. \tag{5}$$

(Всюду далее на протяжении этого параграфа под знаком нормы $\|\cdot\|$ будем иметь в виду норму $\|\cdot\|_2$.)

Рассмотрим следующую оптимизационную задачу. Найти такие итерационные параметры $\tau_0, \dots, \tau_{n-1}$, чтобы норма $\|Q_n(A)\|$ была минимальной. Так как матрица A является симметричной, то матрица $Q_n(A)$ также будет симметричной. Отсюда следует, что если λ является собственным значением A, то $Q_n(\lambda)$ является собственным значением $Q_n(A)$. Таким образом,

$$\|Q_n(A)\| = \max_j |Q_n(\lambda_j)|, \tag{6}$$

где λ_j — собственные значения A. Предположим, что $\lambda_j \in [\mu, M]$, $\mu > 0$. Поскольку собственные значения в (6) неизвестны, а известен только интервал, которому они принадлежат, то задачу нахождения нормы оператора $Q_n(A)$ заменим задачей оценки нормы этого оператора при условии, что мы знаем отрезок, которому принадлежит спектр A, т. е. $\|Q_n(A)\| = \max\limits_{0 \leqslant \lambda \leqslant M} |Q_n(\lambda)|$. Заметим, что многочлен $Q_n(\lambda)$ имеет вид $Q_n(\lambda) = 1 + \cdots$. Введем класс K_n многочленов степени не выше n, равных единице в точке 0. Таким образом, мы можем переформулировать исходную задачу оптимизации следующим образом. На классе K_n требуется найти многочлен $Q_n^0(\lambda)$ такой, что

$$Q_n^0(A) = \arg \min_{Q_n \in K_n} \|Q_n(A)\| = \arg \min_{Q_n \in K_n} \max_{0 \leqslant \lambda \leqslant M} |Q_n(\lambda)|; \tag{7}$$

здесь arg, как обычно, означает аргумент, т. е. мы ищем многочлен $Q_n \in K_n$, для которого имеет место равенство $\min\limits_{Q_n \in K_n} \|Q_n(A)\| = \|Q_n^0(A)\|$.

На самом деле, вводя класс K_n, мы расширили класс многочленов, поскольку в исходной постановке задачи предполагалось, что на $[\mu, M]$ искомый многочлен должен иметь n корней. Тем не менее, как мы увидим далее, данное расширение класса не изменяет результат решения оптимизационной задачи.

Лемма. *Справедливо равенство*

$$Q_n^0(\lambda) = t_n^{-1} T_n\left(\frac{M + \mu - 2\lambda}{M - \mu}\right), \tag{8}$$

где T_n — многочлен Чебышева степени n, а $t_n = T_n\left(\frac{M+\mu}{M-\mu}\right)$.

Доказательство. Предположим, что утверждение леммы неверно, т. е. что существует многочлен $Q_n \in K_n$ с нормой, меньшей, чем у Q_n^0. Так

как $|T_n| \leqslant 1$ на $[-1, 1]$, то по предположению имеет место строгое неравенство

$$\max_{0\leqslant\lambda\leqslant M} |Q_n(\lambda)| < \max_{0\leqslant\lambda\leqslant M} |Q_n^0(\lambda)| = t_n^{-1}. \tag{9}$$

Рассмотрим многочлен $S_n(\lambda) = Q_n^0(\lambda) - Q_n(\lambda)$. Пусть

$$\lambda^j = \frac{M+m}{2} - \frac{M-m}{2}\cos\frac{\pi j}{n}, \quad j = 0, \ldots, n.$$

Из равенства $T_n(x) = \cos(n \arccos x)$ имеем

$$Q_n^0(\lambda^j) = (-1)^j t_n^{-1}.$$

Поскольку $\lambda^j \in [\mu, M]$, то, согласно (9),

$$|Q_n(\lambda^j)| < t_n^{-1}.$$

Отсюда следует, что $\operatorname{sign} S_n(\lambda^j) = \operatorname{sign} Q_n^0(\lambda^j)$.

Точки $\lambda^0, \ldots, \lambda^n$ расположены монотонно на отрезке $[\mu, M]$. Поскольку $S_n(\lambda)$ меняет знак при переходе от каждой из этих точек к следующей, то $S_n(\lambda)$ имеет n корней на $[\mu, M]$. Кроме того,

$$S_n(0) = Q_n^0(0) - Q_n(0) = 1 - 1 = 0.$$

Мы получили многочлен степени n, который имеет $n+1$ нуль; следовательно, $S_n(\lambda) \equiv 0$, $Q_n(\lambda) \equiv Q_n^0(\lambda)$,

$$\max_{[\mu, M]} |Q_n(\lambda)| = t_n^{-1}.$$

Мы пришли к противоречию с (9). Лемма доказана.

Заметим, что данный многочлен Q_n^0 решает также исходную оптимизационную задачу, так как по построению он имеет на отрезке $[\mu, M]$ n нулей.

Оценим скорость сходимости полученного метода. Воспользуемся явным представлением многочленов Чебышева

$$T_n(x) = (\lambda_1^n + \lambda_2^n)/2, \quad \lambda_{1,2} = x \pm \sqrt{x^2 - 1}.$$

При $x = \dfrac{M+\mu}{M-\mu}$ имеем

$$x \pm \sqrt{x^2-1} = \frac{M+\mu}{M-\mu} \pm \sqrt{\left(\frac{M+\mu}{M-\mu}\right)^2 - 1} = \frac{(\sqrt{M} \pm \sqrt{\mu})^2}{(\sqrt{M}+\sqrt{\mu})(\sqrt{M}-\sqrt{\mu})}. \tag{10}$$

Введем обозначение $\lambda_0 = (\sqrt{M}+\sqrt{m})/(\sqrt{M}-\sqrt{m})$. Из (10) имеем $\lambda_1 = \lambda_0$, $\lambda_2 = \lambda_0^{-1}$. Так как $\lambda_2 < 1$, то при больших n

$$t_n = T_n\left(\frac{M+\mu}{M-\mu}\right) \sim \frac{1}{2}\lambda_0^n.$$

Поскольку λ_1^n и λ_2^n одного знака, то

$$t_n \geqslant \lambda_1^n/2 = \lambda_0^n/2. \tag{11}$$

На основе приведенных построений можно предложить несколько типов итерационных процессов.

В одном случае задаются последовательностью значений $n_0 = 0 < n_1 < n_2 \cdots$, приближения $\mathbf{x}^{n_i}$ определяют по рекуррентной формуле

$$\mathbf{x}^{n_{i+1}} = \mathbf{x}^{n_i} - P^0_{q_i-1}(A)(A\mathbf{x}^{n_i} - \mathbf{b}), \tag{12}$$

где $q_i = n_{i+1} - n_i$ и $P^0_{q_i-1}(\lambda) = \lambda^{-1}\big(Q^0_{q_i}(\lambda) - 1\big)$. Имеем

$$\mathbf{r}^{n_{i+1}} = Q^0_{q_i}(A)\mathbf{r}^{n_i}, \quad \|\mathbf{r}^{n_{i+1}}\|_2 \leqslant \sigma_{q_i}\|\mathbf{r}^{n_i}\|_2,$$

где $\sigma_{q_i} = |t_{q_i}|^{-1}$, и в итоге

$$\|\mathbf{r}^{n_p}\|_2 \leqslant \sigma_{q_0} \ldots \sigma_{q_{p-1}}\|\mathbf{r}^0\|_2.$$

Рассмотрим случай $q_i \equiv k$, т. е. $n_i = ik$. Тогда, обозначив $\mathbf{x}^{n_i} = \mathbf{y}^i$, можно записать итерационный процесс (12) в виде

$$\mathbf{y}^{i+1} = \mathbf{y}^i - P^0_{k-1}(A)(A\mathbf{y}^i - \mathbf{b}).$$

Соответствующая оценка погрешности имеет вид

$$\|\mathbf{y}^i - \mathbf{X}\|_2 \leqslant (\sigma_k)^i\|\mathbf{y}^0 - \mathbf{X}\|_2.$$

Такой итерационный процесс называют *оптимальным* (по числу итераций) *линейным k-шаговым итерационным процессом*. В частном случае $k = 1$, согласно (5), выполняются соотношения

$$P^0_1(\lambda) = 1 - \lambda Q^0_0(\lambda) = \frac{\frac{M-\mu-2\lambda}{M-\mu}}{\frac{M+\mu}{M-\mu}} = 1 - \lambda\frac{2}{M+\mu},$$

$$t_1 = \frac{1}{\frac{M+\mu}{M-\mu}}, \quad \sigma_1 = \frac{1}{t_1} = \frac{M-\mu}{M+\mu}.$$

Таким образом, *оптимальный линейный одношаговый итерационный процесс* имеет вид

$$\mathbf{y}^{i+1} = \mathbf{y}^i - \frac{2}{M+\mu}(A\mathbf{y}^i - \mathbf{b}),$$

а погрешность оценивается следующим образом:

$$\|\mathbf{y}^i - \mathbf{X}\|_2 \leqslant \left(\frac{M-\mu}{M+\mu}\right)^i \|\mathbf{y}^0 - \mathbf{X}\|_2 \tag{13}$$

(этот метод мы уже построили выше в этом параграфе).

Можно проверить, что коэффициенты многочленов $P_{k-1}^0(A)$ быстро растут с ростом k, поэтому при больших k алгоритм вычисления $\mathbf{x}^n$, использующий информацию о значениях этих коэффициентов, сильно чувствителен к вычислительной погрешности. В связи с этим для вычисления $\mathbf{x}^k$ используют метод (2).

Поскольку из (4) следует, что

$$Q_k(\lambda) = (1 - \tau_k\lambda)\dots(1 - \tau_1\lambda) = Q_k^0(\lambda),$$

то τ_i являются величинами, обратными к корням многочлена $Q_k^0(\lambda)$. Но по доказанному выше $Q_k^0(\lambda) = t_k^{-1}T_k\left(\dfrac{M+\mu-2\lambda}{M-\mu}\right)$, т. е. корни этого многочлена равны

$$\lambda_j = \frac{M+\mu}{2} - \frac{M-\mu}{2}\cos\frac{(2j-1)\pi}{2k}, \quad j = 1,\dots,k. \tag{14}$$

Отсюда следует, что значения τ_i надо брать из совокупности

$$\left\{\frac{2}{M+\mu-(M-\mu)\cos\dfrac{(2j-1)\pi}{2k}}\right\}, \quad j = 1,\dots,k; \tag{15}$$

зафиксировав последовательность $\tau_1 = \lambda_{j_1}^{-1},\dots,\tau_k = \lambda_{j_k}^{-1}$, мы имеем алгоритм (2) для вычисления $\mathbf{y}^{i+1}$ по $\mathbf{y}^i$.

При больших k и произвольном выборе $j_1,\dots,j_k$ алгоритм вычисления по формулам (2), (15) также неустойчив к погрешностям округления. Так, например, если взять $\tau_i = \lambda_i^{-1}$, то согласно (3) уравнение для погрешности имеет вид $\mathbf{r}^i = (E - \tau_i A)\mathbf{r}^{i-1}$. При наличии округлений оно запишется в виде

$$\mathbf{r}^i = (E - \tau_i A)\mathbf{r}^{i-1} + \rho^{i-1}.$$

Последовательно выражая $\mathbf{r}^i$ через предыдущие, имеем равенство

$$\mathbf{r}^k = \prod_{i=1}^{k}(E - \tau_i A)\mathbf{r}^{(0)} + \sum_{i=0}^{k-1}\prod_{i+2\leqslant j\leqslant k}(E - \tau_j A)\rho^i.$$

Здесь к $\mathbf{r}^0$ применяется оператор $Q_k^0(A)$ с нормой $\sigma_k = |t^k|^{-1} < 1$, в то время как операторы, применяемые к ρ_i, могут быть с очень большими нормами.

При реальных вычислениях для обеспечения устойчивости алгоритма к округлениям осуществляют «перемешивание» чисел τ_i. Алгоритм перемешивания в случае $k = 2^l$ заключается в следующем. Последовательно, при $j = 1,\dots,l$ строится «наиболее перемешанная» перестановка чисел $1,\dots,2^j$. При $j = 1$ она состоит из двух чисел 2, 1. Пусть уже построена перестановка $(b_1^{j-1},\dots,b_{2^{j-1}}^{j-1})$; следующая перестановка берется в виде $(2^j + 1 - b_1^{j-1}, b_1^{j-1}, 2^j + 1 - b_2^{j-1}, b_2^{j-1},\dots)$. Например, при $l = 4$ эта перестановка имеет вид (11, 6, 14, 3, 10, 7, 15, 2, 12, 5, 13, 4, 9, 8, 16, 1). При таком алгоритме выбора

итерационных параметров норма оператора перехода всегда не будет превосходить 1. Задаваясь $k = 2^l$, строим таблицу чисел $b_1^l, \ldots, b_{2^l}^l$ и производим итерации (2) при значениях

$$\tau_i = 2 \Big/ \left(M + \mu - (M - \mu) \cos \frac{\pi(2b_i^l - 1)}{2k} \right). \tag{16}$$

Как уже отмечалось выше, k-шаговый оптимальный процесс обладает тем недостатком, что число итераций обязательно должно быть кратным k. В случае большого значения k (а это на практике имеет место, так как при этом улучшается скорость сходимости) это ведет к дополнительным затратам при решении системы.

Поставим задачу построения итерационного процесса, который при любом k дает такую же оценку для погрешности, как и k-шаговый оптимальный процесс. Исходя из такой постановки задачи, потребуем, чтобы при любом k вектор погрешности $\mathbf{r}^k$ удовлетворял уравнению

$$\mathbf{r}^k = t_k^{-1} T_k \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) \mathbf{r}^0. \tag{17}$$

Запишем соотношение (17) последовательно для $k = n - 1$, $k = n$, $k = n + 1$. Получим

$$\begin{aligned} \mathbf{r}^{n-1} &= t_{n-1}^{-1} T_{n-1} \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) \mathbf{r}^0, \\ \mathbf{r}^{n} &= t_{n}^{-1} T_{n} \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) \mathbf{r}^0, \\ \mathbf{r}^{n+1} &= t_{n+1}^{-1} T_{n+1} \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) \mathbf{r}^0. \end{aligned} \tag{18}$$

Многочлены Чебышева связаны рекуррентным соотношением

$$T_{n+1}(x) - 2T_n(x) + T_{n-1}(x) = 0. \tag{19}$$

Умножим первое уравнение (18) на t_{n-1}, третье — на t_{n+1}, а к обеим частям второго уравнения применим матрицу $-2t_n \dfrac{(M + \mu)E - 2A}{M - \mu}$. Складывая результаты, получим

$$\begin{aligned} t_{n-1}\mathbf{r}^{n-1} - 2t_n \frac{(M + \mu)E - 2A}{M - \mu} \mathbf{r}^n + t_{n+1}\mathbf{r}^{n+1} = \\ = \left\{ T_{n-1} \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) + T_{n+1} \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) - \right. \\ \left. - 2 \frac{(M + \mu)E - 2A}{M - \mu} T_n \left(\frac{(M + \mu)E - 2A}{M - \mu} \right) \right\} \mathbf{r}^0. \end{aligned} \tag{20}$$

Выражение в фигурных скобках в (20) равно нулю в силу (19). Таким образом, погрешности $\mathbf{r}^n$ искомого итерационного процесса должны удовлетворять трехчленному соотношению

$$t_{n-1}\mathbf{r}^{n-1} - 2t_n\frac{(M+\mu)E-2A}{M-\mu}\mathbf{r}^n + t_{n+1}\mathbf{r}^{n+1} = \mathbf{0}. \tag{21}$$

Так как $\mathbf{r}^n = \mathbf{x}^n - \mathbf{X}$, то подставляя это выражение в (21), получим

$$t_{n-1}\mathbf{x}^{n-1} - 2t_n\frac{(M+\mu)E-2A}{M-\mu}\mathbf{x}^n + t_{n+1}\mathbf{x}^{n+1} = \\ = \left(t_{n-1}E - 2t_n\frac{(M+\mu)E-2A}{M-\mu} + t_{n+1}E\right)\mathbf{X}. \tag{22}$$

Вследствие (19) и равенства $t_n = T_n\left(\frac{M+\mu}{M-\mu}\right)$ имеем

$$t_{n+1} - 2\left(\frac{M+\mu}{M-\mu}\right)t_n + t_{n-1} = 0, \tag{23}$$

и соотношение (22) может быть переписано в виде

$$t_{n+1}\mathbf{x}^{n+1} - 2\left(\frac{M+\mu}{M-\mu}\right)t_n\mathbf{x}^n + t_{n-1}\mathbf{x}^{n-1} = 4\frac{-t_n}{M-\mu}(A\mathbf{x}^n - \mathbf{b}). \tag{24}$$

Мы получили требуемое рекуррентное соотношение. Приведем его к более удобному виду. Вследствие (23) равенство (24) можно переписать в виде

$$t_{n+1}\mathbf{x}^{n+1} - (t_{n+1}+t_{n-1})\mathbf{x}^n + t_{n-1}\mathbf{x}^{n-1} = -\frac{2}{M+\mu}(t_{n+1}+t_{n-1})(A\mathbf{x}^n - \mathbf{b})$$

или

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \omega_n\omega_{n-1}(\mathbf{x}^n - \mathbf{x}^{n-1}) - \frac{2}{M+\mu}(1+\omega_n\omega_{n-1})(A\mathbf{x}^n - \mathbf{b}), \tag{25}$$

где $\omega_n = t_n/t_{n+1}$. Разделив обе части (23) на t_{n+1}, получим

$$1 - 2\frac{M+\mu}{M-\mu}\omega_n + \omega_n\omega_{n-1} = 0,$$

откуда

$$\omega_n = 1\Big/\left(2\frac{M+\mu}{M-\mu} - \omega_{n-1}\right) \quad \text{при} \quad n > 0, \quad \omega_0 = t_0/t_1. \tag{26}$$

Таким образом, можно рекуррентно вычислять величины ω_n из (26) и затем векторы $\mathbf{x}^{n+1}$ из (25). Для получения совокупности векторов $\mathbf{x}^1, \ldots, \mathbf{x}^n$ потребуется произвести n умножений матрицы на вектор, $O(n)$ умножений векторов на числа, сложений векторов и операций с числами. При этом для всех k, вследствие (5), (9), выполняется оценка

$$\|\mathbf{x}^k - \mathbf{X}\|_2 \leqslant t_k^{-1}\|\mathbf{x}^0 - \mathbf{X}\|_2. \tag{27}$$

Квадратное уравнение

$$\omega^2 - 2\frac{M+\mu}{M-\mu}\omega + 1 = 0$$

имеет два положительных корня

$$\omega = \frac{M+\mu}{M-\mu} \pm \sqrt{\left(\frac{M+\mu}{M-\mu}\right)^2 - 1}.$$

Наименьший из этих корней, равный

$$\frac{M+\mu}{M-\mu} - \sqrt{\left(\frac{M+\mu}{M-\mu}\right)^2 - 1} = \frac{\sqrt{M}-\sqrt{\mu}}{\sqrt{M}+\sqrt{\mu}} = \frac{1}{\lambda_0},$$

обозначим через ω.

Итерационный процесс (25), (26) называют *оптимальным* (по количеству итераций) *линейным итерационным процессом*. При реализации процесса (25), (26) после любых k применений матрицы A мы получаем оптимальный результат в смысле (6). Из сказанного видно, что по скорости сходимости итерационный процесс (25), (26) предпочтительнее, чем (12). Однако иногда от него отказываются в пользу (12) из-за соображений экономии памяти ЭВМ.

Получим более наглядную оценку скорости сходимости построенных итерационных процессов. Согласно (27) и (11) для оптимального итерационного процесса выполняется оценка

$$\|\mathbf{x}^n - \mathbf{X}\|_2 \leqslant 2\lambda_0^{-n}\|\mathbf{x}^0 - \mathbf{X}\|_2.$$

Норма погрешности $\mathbf{r}^n$ уменьшится по крайней мере в ε^{-1} раз, если $2\lambda_0^{-n} \leqslant \varepsilon$. Отсюда получаем оценку числа итераций, обеспечивающих получение решения с точностью ε:

$$n \geqslant n_1 = \log_{\lambda_0}(2/\varepsilon) = (\ln\lambda_0)^{-1}\ln(2/\varepsilon).$$

Для многих задач число M/μ оказывается очень большим. Поэтому при $M/\mu \to \infty$ имеем $\lambda_0 = 1 + 2\sqrt{\mu/M} + O(\mu/M)$. Таким образом,

$$\ln\lambda_0 \sim 2\sqrt{\mu/M}, \quad n_1 \sim 0{,}5\sqrt{M/\mu}\ln(2/\varepsilon).$$

Для сравнения рассмотрим оптимальный линейный одношаговый процесс, имеющий, согласно (13), оценку погрешности

$$\|\mathbf{x}^n - \mathbf{X}\|_2 \leqslant \left(\frac{M+\mu}{M-\mu}\right)^n \|\mathbf{x}^0 - \mathbf{X}\|_2;$$

отсюда получаем оценку числа итераций

$$n \geqslant n_2 = \left(\ln\frac{M+\mu}{M-\mu}\right)^{-1} \ln\frac{1}{\varepsilon}.$$

При $M/\mu \to \infty$ имеем

$$\ln\frac{M+\mu}{M-\mu} \sim \frac{2\mu}{M} \quad \text{и} \quad n_2 \sim \frac{1}{2}\frac{M}{\mu}\ln\frac{1}{\varepsilon}.$$

Таким образом, в этом сравнении оптимальный итерационный процесс дает выигрыш в числе итераций примерно в $\sqrt{M/\mu}$ раз.

Задача 2. Рассматривается итерационный процесс

$$\mathbf{x}^i = \mathbf{x}^{i-1} - \tau_{i-1}(A\mathbf{x}^{i-1} - \mathbf{b}), \quad i = 1, 2, \ldots .$$

Пусть p нечетное и при всех k совокупности $\tau_0, \ldots, \tau_{p^k-1}$ совпадают с совокупностями

$$\gamma_{i-1} = 2\left(M + \mu - (M - \mu)\cos\frac{\pi(2i-1)}{2p^k}\right)^{-1}, \quad i = 1, \ldots, p^k.$$

Проверить, что при всех $i = p^k$ приближения $\mathbf{x}^i$ совпадают с приближениями, получаемыми по оптимальному линейному итерационному процессу.

Задача 3. Пусть $\tau_0 = 2/(M + \mu)$, $\tau_1 = 1/M$, $\tau_2 = 1/\mu$ и совокупности $\tau_{2^{k+1}}, \ldots, \tau_{2^{k+1}}$ при каждом $k > 1$ совпадают с совокупностями величин

$$\gamma_i = 2\left(M + \mu - (M - \mu)\cos\frac{\pi(2i-1)}{2^{k+1}}\right)^{-1}, \quad i = 1, \ldots, 2^k.$$

Показать, что для такого итерационного процесса при всех $n = 2^{k+1}$ справедлива оценка

$$\|\mathbf{x}^n - \mathbf{X}\|_2 \leqslant \frac{4}{(\lambda_0^{n-1} - \lambda_0^{1-n})(\lambda_0 - \lambda_0^{-1})}\|\mathbf{x}^0 - \mathbf{X}\|_2,$$

т. е.

$$\|\mathbf{x}^n - \mathbf{X}\|_2 = O(\lambda_0^{-n})\|\mathbf{x}^0 - \mathbf{X}\|_2.$$

Рассмотрим типичную задачу математической физики, сводящуюся к решению системы уравнений с большим отношением M/μ. Пусть в квадрате $0 \leqslant x_1, x_2 \leqslant 1$ решается уравнение Пуассона $-\Delta u = f$ при нулевых условиях на границе. Зададимся сеткой с шагами $h = 1/l$ и напишем систему уравнений, аппроксимирующих дифференциальную задачу:

$$-\frac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2} - \frac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} = f_{mn} \tag{28}$$

при $0 < m, n < l$;

$$u_{mn} = 0, \quad \text{если} \quad m(l-m)(l-n)n = 0.$$

Матрица этой системы является положительно определенной, и для нее

$$\mu = \min\lambda_i = 8l^2\sin\left(\frac{\pi h}{2}\right) \sim 2\pi^2,$$

$$M = \max\lambda_i = 8l^2\cos\left(\frac{\pi h}{2}\right) \sim 8l^2,$$

т. е. $\sqrt{M/\mu} \sim 2l/\pi$. Например, при шаге $l = 30$ выигрыш в числе итераций примерно в 20 раз.

В начале этого параграфа было упомянуто, что при симметризации системы ее свойства могут ухудшаться. Действительно, пусть этот процесс применяется к уже симметричной матрице A, т. е. переходим от системы $A\mathbf{x} = \mathbf{b}$ к системе $A^2\mathbf{x} = A\mathbf{b}$. Если у старой системы отношение максимального и минимального собственных значений равнялось M/μ, то у новой оно будет $(M/\mu)^2$ и скорость сходимости итерационных процессов будет меньше.

Примечание. При $n \to \infty$ имеем

$$\omega_n = \frac{T_n\left(\dfrac{M+\mu}{M-\mu}\right)}{T_{n+1}\left(\dfrac{M+\mu}{M-\mu}\right)} \to \omega = \frac{\sqrt{M}-\sqrt{\mu}}{\sqrt{M}+\sqrt{\mu}}.$$

Таким образом, при больших n итерационная формула (25) близка к формуле

$$\mathbf{z}^{n+1} = \mathbf{z}^n + \omega^2(\mathbf{z}^n - \mathbf{z}^{n-1}) - \frac{2}{M+\mu}(1+\omega^2)(A\mathbf{z}^n - \mathbf{b}). \tag{29}$$

Если при $n > 1$ итерации будут производиться по этой формуле, то при условии $\mathbf{z}^0 = \mathbf{x}^0$, $\mathbf{z}^1 = \mathbf{y}^1$ этот итерационный процесс требует примерно столько же итераций, сколько итерационный процесс (25).

Задача 4. Для итерационного процесса (29) получить оценку погрешности

$$\frac{\|\mathbf{r}^{n+1}\|_2}{\|\mathbf{r}^0\|_2} \leqslant \frac{2+(n-1)(1-\lambda_0^{-2})}{1+\lambda_0^{-2}}\lambda_0^{-n}. \tag{30}$$

Указание. Представить погрешность в виде $\mathbf{r}^n = \sum_{k=1}^{m} z_k^m \mathbf{e}_k$, где $\{\mathbf{e}_k\}$ — полная ортогональная система собственных векторов матрицы A. Подстановкой в (29) получить разностное уравнение, связывающее $\mathbf{z}_k^{n-1}$, $\mathbf{z}_k^n$, $\mathbf{z}_k^{n+1}$. Получить явное выражение для $\mathbf{z}_k^n$ и с его помощью получить требуемую оценку (30).

Задача 5. Показать, что оценка (30) не может быть улучшена на множестве матриц, удовлетворяющих условию $S(A) \subset [\mu, M]$.

§ 7. Метод Зейделя

Пусть решается система уравнений $A\mathbf{x} = \mathbf{b}$, все диагональные элементы которой ненулевые. В итерационном методе Зейделя последовательно уточняются компоненты решения, причем k-я компонента находится из k-го уравнения. Именно, если $\mathbf{x}^n = (x_1^n, \ldots, x_m^n)^T$, то следующее прибли-

жение определяется из системы соотношений

$$\begin{aligned} &a_{11}x_1^{n+1} + a_{12}x_2^n + \cdots + a_{1m}x_m^n = b_1,\\ &a_{21}x_1^{n+1} + a_{22}x_2^{n+1} + a_{23}x_3^n + \cdots + a_{2m}x_m^n = b_2,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ &a_{m1}x_1^{n+1} + a_{m2}x_2^{n+1} + \cdots + a_{mm}x_m^{n+1} = b_m. \end{aligned} \tag{1}$$

Систему (1) можно представить в виде

$$B\mathbf{x}^{n+1} + C\mathbf{x}^n = \mathbf{b}, \tag{2}$$

где

$$B = \begin{pmatrix} a_{11} & 0 & 0 & \dots & 0\\ a_{21} & a_{22} & 0 & \dots & 0\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mm} \end{pmatrix}, \qquad C = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1m}\\ 0 & 0 & a_{23} & \cdots & a_{2m}\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Отсюда получаем

$$\mathbf{x}^{n+1} = -B^{-1}C\mathbf{x}^n + B^{-1}\mathbf{b}. \tag{3}$$

Таким образом, метод Зейделя эквивалентен некоторому методу простой итерации; поэтому для его сходимости при любом начальном приближении необходимо и достаточно, чтобы все собственные значения матрицы $B^{-1}C$ по модулю были меньше 1. Вследствие равенства

$$\det(-B^{-1}C - \lambda E) = \det(-B^{-1})\det(C + B\lambda)$$

собственные значения матрицы $(-B^{-1}C)$ являются корнями уравнения $\det(C + B\lambda) = 0$.

Таким образом, необходимое и достаточное условие сходимости метода Зейделя можно сформулировать следующим образом: все корни уравнения

$$\det\begin{pmatrix} a_{11}\lambda & a_{12} & a_{13} & \cdots & a_{1m}\\ a_{21}\lambda & a_{22}\lambda & a_{23} & \cdots & a_{2m}\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ a_{m1}\lambda & a_{m2}\lambda & a_{m3}\lambda & \cdots & a_{mm}\lambda \end{pmatrix} = 0 \tag{4}$$

должны быть по модулю меньше 1.

Часто можно предложить более удобные для применения достаточные условия сходимости метода Зейделя.

Задача 1. Пусть при всех i

$$\sum_{j\neq i} |a_{ij}| \leqslant q|a_{ii}|, \quad q < 1.$$

Получить оценку

$$\|\mathbf{x}^n - \mathbf{X}\|_\infty \leqslant \cdots \leqslant q^n\|\mathbf{x}^0 - \mathbf{X}\|_\infty.$$

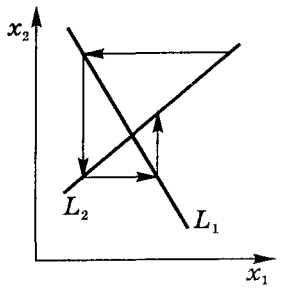


Рис. 6.7.1

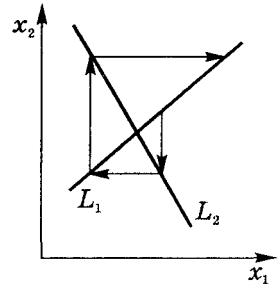


Рис. 6.7.2

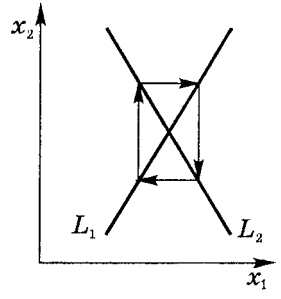


Рис. 6.7.3

Проблема решения систем линейных уравнений является модельной относительно более сложных задач решения систем нелинейных уравнений и минимизации функций многих переменных. Для перенесения метода на более сложные задачи важно понять его наиболее «грубые» качественные свойства, обеспечивающие сходимость. С этой целью наиболее желательно получить геометрическую интерпретацию метода. Обозначим через L_i плоскость $\sum_{j=1}^{m} a_{ij}x_j - b_i = 0$. При получении приближения $(x_1^{n+1}, \ldots, x_i^{n+1}, x_{i+1}^n, \ldots, x_m^n)$ из приближения $(x_1^{n+1}, \ldots, x_{i-1}^{n+1}, x_i^{n+1}, \ldots, x_m^n)$ происходит перемещение приближения параллельно оси x_i до пересечения с плоскостью L_i. Таким образом, геометрически метод Зейделя состоит в циклическом перемещении точек, соответствующих последовательно получаемым приближениям, параллельно координатным осям x_i до пересечения с плоскостями L_i. Рис. 6.7.1–6.7.3 иллюстрируют при $m = 2$ случаи, когда метод Зейделя сходится, расходится, имеет цикл (как говорят, «зацикливается»). Сравнение первых двух рисунков показывает, что сходимость метода Зейделя может изменить характер при перестановке уравнений.

Особенно интересная геометрическая картина возникает в случае, когда матрица A симметричная.

Теорема. *Пусть A — вещественная симметричная положительно определенная матрица. Тогда метод Зейделя сходится.*

Доказательство. При симметричной A имеем

$$F(\mathbf{y}) = (A(\mathbf{y}-\mathbf{X}), \mathbf{y}-\mathbf{X}) - (A\mathbf{X}, \mathbf{X}) = (A\mathbf{y}, \mathbf{y}) - 2(A\mathbf{X}, \mathbf{y}) = (A\mathbf{y}, \mathbf{y}) - 2(\mathbf{b}, \mathbf{y}).$$

Если $A > 0$, то $(A(\mathbf{y}-\mathbf{X}), \mathbf{y}-\mathbf{X}) > 0$ при $\mathbf{y} \neq \mathbf{X}$, поэтому функция $F(\mathbf{y})$ имеет минимум, и притом единственный, при $\mathbf{y} = \mathbf{X}$. Таким образом, за-

дача отыскания решения системы $A\mathbf{x} = \mathbf{b}$ оказалась равносильной задаче отыскания единственного минимума функции $F(\mathbf{y})$.

Одним из методов минимизации функции многих переменных является *метод покоординатного спуска.*

Пусть имеется приближение $(x_1^0, \ldots, x_m^0)$ к точке экстремума функции $F(x_1, \ldots, x_m)$. Рассмотрим функцию $F(x_1, x_2^0, \ldots, x_m^0)$ как функцию переменной x_1 и найдем точку x_1^1 ее минимума. Затем, исходя из приближения $(x_1^1, x_2^0, \ldots, x_m^0)$, путем минимизации функции $F(x_1^1, x_2, x_3^0, \ldots, x_m^0)$ по переменной x_2 находим следующее приближение $(x_1^1, x_2^1, x_3^0, \ldots, x_m^0)$. Процесс циклически повторяется. При уточнении компоненты x_k происходит смещение по прямой, параллельной оси x_k, до точки с наименьшим на этой прямой значением $F(\mathbf{x}) = c$. Ясно, что эта точка будет точкой касания рассматриваемой прямой и линии уровня $F(\mathbf{x}) = c$. Поэтому в двумерном случае картина приближений выглядит как на рис. 6.7.4.

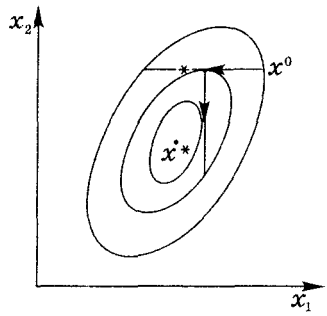


Рис. 6.7.4

Применим метод покоординатного спуска для отыскания экстремума функции $F(\mathbf{y})$. Обозначим $F(\mathbf{x}) + (A\mathbf{X}, \mathbf{X}) = (A(\mathbf{x} - \mathbf{X}), \mathbf{x} - \mathbf{X})$ через $F_0(\mathbf{x})$. При минимизации по переменной x_k происходит перемещение параллельно оси x_k до точки, где $F'_{x_k} = 0$. Следовательно, новое значение x_k определяется из того же уравнения

$$F'_{x_k} = 2\left(\sum_{j=1}^{m} a_{kj}x_j - b_k\right) = 0, \tag{5}$$

что и в случае метода Зейделя. Таким образом, приближения покоординатного спуска минимизации функции $F(\mathbf{y})$ и метода Зейделя решения исходной системы совпадают.

Если $\mathbf{x}^n \neq \mathbf{X}$, то хотя бы одно из уравнений системы не удовлетворяется и соответствующее значение $F'_{x_k}(\mathbf{x}^n) \neq 0$. Выберем среди таких k наименьшее. Тогда при уточнении компонент $x_1, \ldots, x_{k-1}$ мы остаемся в точке $\mathbf{x}^n$, а при уточнении компоненты x_k происходит смещение в сторону меньших значений $F(\mathbf{x})$; при уточнении остальных компонент значение $F(\mathbf{x})$ не возрастает. Таким образом,

$$F(\mathbf{x}^{n+1}) < F(\mathbf{x}^n), \quad F_0(\mathbf{x}^{n+1}) < F_0(\mathbf{x}^n);$$

поэтому

$$F_0(\mathbf{x}^{n+1})/F_0(\mathbf{x}^n) < 1 \tag{6}$$

при $\mathbf{x}^n \neq \mathbf{X}$. Вследствие (3) имеем равенство $\mathbf{r}^{n+1} = -B^{-1}C\mathbf{r}^n$, где $\mathbf{r}^n = \mathbf{x}^n - \mathbf{X}$. Соотношение (6) можно переписать в виде

$$\varphi(\mathbf{r}^n) = \frac{(AB^{-1}C\mathbf{r}^n,\ B^{-1}C\mathbf{r}^n)}{(A\mathbf{r}^n,\ \mathbf{r}^n)} < 1 \tag{7}$$

при $\mathbf{r}^n \neq 0$. На сфере $\|\mathbf{r}^n\|_2 = 1$ величина $\varphi(\mathbf{r}^n)$ непрерывна, поэтому она достигает своего наибольшего значения φ_0. Так как $A > 0$, то всегда $\varphi(\mathbf{r}^n) > 0$ и поэтому $\varphi_0 > 0$. Положим $\sqrt{\varphi_0} = \lambda$. Вследствие (7) имеем $\lambda^2 < 1$. Очевидно, $\varphi(c\mathbf{r}^n) = \varphi(\mathbf{r}^n)$ при любом $c \neq 0$, поэтому $\varphi(\mathbf{r}^n) = \varphi(\mathbf{r}^n/\|\mathbf{r}^n\|_2) \leqslant \lambda^2$ при любом $\mathbf{r}^n$. Отсюда получаем неравенство

$$F_0(\mathbf{x}^{n+1})/F_0(\mathbf{x}^n) \leqslant \lambda^2, \tag{8}$$

и, таким образом,

$$F_0(\mathbf{x}^n) \leqslant \lambda^{2n} F_0(\mathbf{x}^0).$$

Из (3.8), (3.9) следуют неравенства

$$\min \lambda_A^i \, \|\mathbf{y} - \mathbf{X}\|_2^2 \leqslant F_0(\mathbf{y}) \leqslant \max \lambda_A^i \, \|\mathbf{y} - \mathbf{X}\|_2^2.$$

Отсюда получаем оценку скорости сходимости

$$\|\mathbf{x}^n - \mathbf{X}\|_2 \leqslant \sqrt{\frac{F_0(\mathbf{x}^n)}{\min \lambda_A^i}} \leqslant \lambda^n \sqrt{\frac{F_0(\mathbf{x}^0)}{\min \lambda_A^i}} \leqslant \lambda^n \sqrt{\frac{\max \lambda_A^i}{\min \lambda_A^i}} \, \|\mathbf{x}^0 - \mathbf{X}\|_2. \tag{9}$$

Теорема доказана.

Из рис. 6.7.5 можно усмотреть, что метод Зейделя сходится быстрее, если направление осей эллипсоидов близко к направлению координатных осей, т. е. матрица A близка к диагональной.

В случае, изображенном на рис. 6.7.4, последовательные приближения смещаются все время монотонно влево и вниз. Такая картина — монотонное смещение отдельных компонент все время в одном и том же направлении — характерна для ряда классов матриц. При этом довольно часто монотонное смещение наблюдается именно у компонент решения, скорость сходимости которых наиболее плохая. В этих случаях для ускорения сходимости прибегают к *методу релаксации*, который заключается в следующем. После уточнения каждой координаты по методу Зейделя производится смещение в том же направлении на p-ю часть этого смещения. Таким обра-

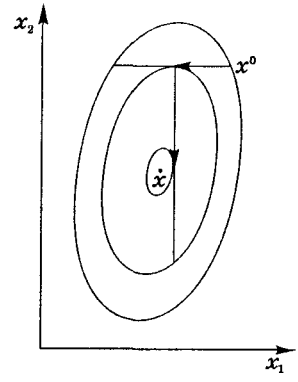


Рис. 6.7.5

зом, приближения отыскиваются из соотношения

$$(B-D)\mathbf{x}^{n+1}+D\left(\frac{\mathbf{x}^{n+1}+p\mathbf{x}^n}{1+p}\right)+C\mathbf{x}^n=\mathbf{b}, \tag{10}$$

D — диагональная матрица с элементами a_{ii} по диагонали. Как показала практика вычислений, при $A>0$ целесообразно брать показатель релаксации p в пределах $-1<p<1$; в случае $0<p<1$ метод релаксации обычно называют *методом сверхрелаксации* (или *методом верхней релаксации*). На рис. 6.7.4 изображены символами $\circ$ приближения метода Зейделя, $*$ — метода верхней релаксации при $p=1/4$. Например, для уменьшения погрешности в ε^{-1} раз при решении системы (6.28) методом Зейделя требуется порядка $ch^{-2}\ln(\varepsilon^{-1})$ итераций. Если применить метод верхней релаксации при $p=1-\omega h$, $\omega>0$, то потребуется порядка $c(\omega)h^{-1}\ln(\varepsilon^{-1})$ итераций. Подбор параметра ω в этой задаче требует детального рассмотрения. В частности, во многих случаях оптимальное значение параметра релаксации p определяется экспериментально. Иногда параметр релаксации p выбирают зависящим от n и i.

Для случая $A>0$ еще раз обратимся к геометрической картине (см. рис. 6.7.4). После уточнения компоненты x_1 по методу релаксации при $-1<p<1$ мы попадаем в точку, лежащую внутри или на границе эллипсоида $F(\mathbf{x})=F(\mathbf{x}^0)$. Рассуждая так же, как при обосновании сходимости метода Зейделя, заключаем, что при $-1<p<1$ всегда $F_0(\mathbf{x}^{n+1})<F_0(\mathbf{x}^n)$, если $\mathbf{x}^n\neq\mathbf{X}$.

Задача 2. Доказать, что при условии $|p_n|\leqslant q<1$ метод релаксации сходится со скоростью геометрической прогрессии.

§ 8. Метод наискорейшего градиентного спуска

Распространенным методом минимизации функций большого числа переменных является *метод градиентного спуска.* Последующее приближение получается из предыдущего смещением в направлении, противоположном градиенту функции $F(\mathbf{x})$. Каждое следующее приближение ищется в виде

$$\mathbf{x}^{n+1}=\mathbf{x}^n-\delta_n\operatorname{grad}F(\mathbf{x}^n). \tag{1}$$

Приведенное описание не определяет алгоритм однозначно, поскольку ничего не сказано о выборе параметра δ_n. Например, его можно определять из условия минимума величины

$$F\left(\mathbf{x}^n-\delta_n\operatorname{grad}F(\mathbf{x}^n)\right). \tag{2}$$

В этом случае рассматриваемый метод называют *методом наискорейшего градиентного спуска* или просто *методом наискорейшего спуска.*

Для функции $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$, соответствующей системе линейных уравнений с матрицей $A = A^T > 0$, задача нахождения минимума решается в явном виде. В этом конкретном случае

$$\operatorname{grad} F = 2(A\mathbf{x} - \mathbf{b})$$

и

$$\mathbf{x}^{n+1} = \mathbf{x}^n - 2\delta_n(A\mathbf{x}^n - \mathbf{b}).$$

Обозначим $2\delta_n$ через Δ_n, т. е. положим

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n(A\mathbf{x}^n - \mathbf{b}). \tag{3}$$

Пусть $\varphi(\Delta_n) = F(\mathbf{x}^{n+1})$. Вспоминая, что $A = A^T$, вычислим $\varphi'(\Delta_n)$. Имеем

$$\varphi(\Delta_n) = F(\mathbf{x}^n) - 2\Delta_n(A\mathbf{x}^n - \mathbf{b},\ A\mathbf{x}^n - \mathbf{b}) + (A(A\mathbf{x}^n - \mathbf{b}),\ A\mathbf{x}^n - \mathbf{b})\Delta_n - \\ - (A\mathbf{x}^n - \mathbf{b},\ A\mathbf{x}^n - \mathbf{b})) = 0,$$

откуда

$$\Delta_n = \frac{(A\mathbf{x}^n - \mathbf{b},\ A\mathbf{x}^n - \mathbf{b})}{(A(A\mathbf{x}^n - \mathbf{b}),\ A\mathbf{x}^n - \mathbf{b})}. \tag{4}$$

На рис. 6.8.1 изображены последовательные приближения метода наискорейшего спуска и линии уровня функции F. Итерационный процесс (3), (4) называют *методом наискорейшего спуска* решения рассматриваемой системы линейных уравнений.

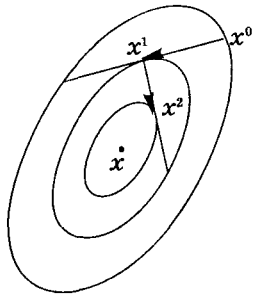


Рис. 6.8.1

Пусть собственные значения матрицы A расположены на $[\mu, M]$, т. е. $S_A \subset [\mu, M]$.

Теорема. *Приближения метода наискорейшего спуска удовлетворяют соотношению*

$$F_0(\mathbf{x}^n) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^{2n} F_0(\mathbf{x}^0), \quad F_0(\mathbf{x}) = (A(\mathbf{x} - \mathbf{X}),\ \mathbf{x} - \mathbf{X}). \tag{5}$$

Доказательство. При $\mathbf{y}^n = \mathbf{x}^n$ произведем одну итерацию оптимального одношагового итерационного процесса

$$\mathbf{y}^{n+1} = \mathbf{y}^n - \frac{2}{M+\mu}(A\mathbf{y}^n - \mathbf{b}). \tag{6}$$

Погрешности итераций $\mathbf{r}^n = \mathbf{y}^n - \mathbf{X}$ связаны соотношением

$$\mathbf{r}^{n+1} = \left(E - \frac{2}{M+\mu}A\right)\mathbf{r}^n.$$

Пусть $\mathbf{e}_1, \ldots$ — ортонормированная система собственных векторов матрицы A: $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$, $(\mathbf{e}_i, \mathbf{e}_j) = \delta_i^j$. Поскольку $\mu \leqslant \lambda_i \leqslant M$, то при всех i выполняются соотношения

$$-\frac{M-\mu}{M+\mu} \leqslant 1 - \frac{2}{M+\mu}\lambda_i \leqslant \frac{M-\mu}{M+\mu}$$

и, таким образом,

$$\left|1 - \frac{2}{M+\mu}\lambda_i\right| \leqslant \frac{M-\mu}{M+\mu}. \tag{7}$$

Пусть $\mathbf{r}^n = \sum c_i \mathbf{e}_i$. Справедливы соотношения

$$(A\mathbf{r}^n, \mathbf{r}^n) = \left(\sum c_i \lambda_i \mathbf{e}_i, \sum c_i \mathbf{e}_i\right) = \sum \lambda_i c_i^2,$$

$$\mathbf{r}^{n+1} = \sum b_i \mathbf{e}_i, \qquad \text{где} \qquad b_i = \left(1 - \frac{2\lambda_i}{M+\mu}\right) c_i,$$

$$(A\mathbf{r}^{n+1}, \mathbf{r}^{n+1}) \leqslant \sum \lambda_i b_i^2 = \sum \lambda_i \left(1 - \frac{2\lambda_i}{M+\mu}\right)^2 c_i^2.$$

С учетом (7) получаем

$$(A\mathbf{r}^{n+1}, \mathbf{r}^{n+1}) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^2 (A\mathbf{r}^n, \mathbf{r}^n).$$

Поскольку $F_0(\mathbf{y}^n) = (A\mathbf{r}^n, \mathbf{r}^n)$, то это означает, что

$$F_0(\mathbf{y}^{n+1}) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(\mathbf{y}^n) = \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(\mathbf{x}^n).$$

Приближение $\mathbf{y}^{n+1}$ можно записать в виде (1)

$$\mathbf{y}^{n+1} = \mathbf{x}^n - \alpha \operatorname{grad} F(\mathbf{x}^n), \quad \alpha = (M+\mu)^{-1}.$$

Так как на $\mathbf{x}^{n+1}$ достигается минимум $F(\mathbf{x})$ среди всех приближений вида (1), то $F(\mathbf{x}^{n+1}) \leqslant F(\mathbf{y}^{n+1})$. Отсюда следует оценка

$$F_0(\mathbf{x}^{n+1}) \leqslant F_0(\mathbf{y}^{n+1}) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(\mathbf{x}^n),$$

а поэтому и справедливость утверждения теоремы. Аналогично (7.9) можно получить неравенство

$$\|\mathbf{x}^n - \mathbf{X}\|_2 \leqslant \left(\frac{M-\mu}{M+\mu}\right)^n \sqrt{\frac{M}{\mu}}\, \|\mathbf{x}^0 - \mathbf{X}\|_2.$$

Хотя на каждом шаге метода наискорейшего спуска уменьшение величины $F_0(\mathbf{x})$ заведомо не меньше, чем у итерационного процесса (6),

мы получили примерно одинаковые оценки скорости сходимости. Однако есть принципиальное различие в этих методах. *Для написания итерационного процесса* (6) *требуется информация о границах спектра* μ, M. *В случае метода* (3), (4), *такой информации не требуется.*

Отметим также то важное обстоятельство, что метод наискорейшего спуска является *нелинейным* итерационным методом; параметры метода на каждом шаге выбираются в зависимости от полученного приближения.

У метода наискорейшего спуска (3), (4), однако, есть следующий недостаток по сравнению с простейшим процессом (6). При нахождении каждого следующего приближения он требует не одной, а двух трудоемких операций умножения матрицы на вектор.

Двукратного умножения матрицы на вектор при каждой итерации можно избежать следующим образом. Обозначим $\mathbf{w}^n = A\mathbf{x}^n - \mathbf{b}$ и перепишем (3) в виде

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n \mathbf{w}^n. \tag{8}$$

Вектор $\mathbf{w}^n$ называется *вектором невязки.* Умножая (8) слева на A и вычитая $\mathbf{b}$, получим

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta_n A\mathbf{w}^n. \tag{9}$$

Формулу (4) для определения Δ_n можно записать в виде

$$\Delta_n = \frac{(\mathbf{w}^n, \mathbf{w}^n)}{(A\mathbf{w}^n, \mathbf{w}^n)}. \tag{10}$$

В процессе итерации запоминаются векторы $\mathbf{x}^n$, $\mathbf{w}^n$ и на каждом шаге последовательно вычисляются $A\mathbf{w}^n$, Δ_n, $\mathbf{x}^{n+1}$, $\mathbf{w}^{n+1}$. В исходном методе наискорейшего спуска (3), (4) погрешность на шаге итерации равносильна возмущению начального приближения и, поскольку процесс сходящийся, ее влияние должно иметь тенденцию к затуханию.

В итерационном процессе (8)–(10) накопление вычислительной погрешности носит более сложный характер.

Задача 1. Получить оценку скорости сходимости метода наискорейшего спуска

$$\|\mathbf{x}^n - \mathbf{X}\|_2 \leqslant (1 - \mu/M)^n \|\mathbf{x}^0 - \mathbf{X}\|_2.$$

Реальный выбор итерационного процесса должен производиться с учетом имеющейся информации о границе спектра, объеме и структуре памяти ЭВМ. Например, при решении сеточных уравнений, аппроксимирующих дифференциальные уравнения в частных производных, иногда идут по следующему пути. Рассматривая задачу на более крупной сетке, проводят вспомогательную работу по возможно более точному определению значений μ и M, соответствующих более мелкой сетке, а затем применяют оптимальный линейный итерационный процесс.

Обратим внимание на интересное обстоятельство. Из геометрической картины итераций метода Зейделя видно, что скорость сходимости метода не меняется при умножении уравнений системы на множители и изменении масштабов по координатным осям, равносильном замене $x_i = k_i y_i$.

Иначе обстоит дело в случае метода наискорейшего спуска. Пусть, например, $A = E$ — единичная матрица. Тогда

$$F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x}) = \sum_{i=1}^{m} x_i^2 - 2\sum_{i=1}^{m} b_i x_i$$

и метод наискорейшего спуска сходится за одну итерацию (доказать!). Произведем замену масштабов $x_i = k_i y_i$, $k_i > 0$. Матрица системы A в данном случае будет диагональной с элементами на диагонали, равными k_i. Тогда минимизируется функционал

$$\bar{F}(\mathbf{y}) = (A\mathbf{y}, \mathbf{y}) - 2(\mathbf{b}, \mathbf{y}) = \sum_{i=1}^{m} k_i y_i^2 - 2\sum_{i=1}^{m} b_i y_i.$$

При большом разбросе k_i линиями уровня функции $\bar{F}$ будут сильно вытянутые эллипсоиды и скорость сходимости метода наискорейшего спуска будет очень медленной.

§ 9. Метод сопряженных градиентов

Метод сопряженных градиентов предназначен для решения систем линейных алгебраических уравнений

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

с симметричной положительно определенной матрицей A.

Предположим, что мы имеем некоторое начальное приближение $\mathbf{x}^0$. Обозначим через $\mathbf{r}^0 = A(\mathbf{x}^0 - \mathbf{X})$ невязку начального приближения; здесь $\mathbf{X}$ — точное решение системы (1). Поставим следующую задачу — построить многочлен $P_n(\lambda)$ степени n, удовлетворяющий условию $P_n(0) = 1$, такой, что значение функционала $F(\mathbf{x}^n) = (A\mathbf{x}^n, \mathbf{x}^n) - 2(\mathbf{b}, \mathbf{x}^n)$, где $\mathbf{x}^n$ находится из соотношения

$$\mathbf{r}^n = P_n(A)\mathbf{r}^0, \tag{2}$$

будет минимальным.

Так как

$$F(\mathbf{y}) = (A\mathbf{y}, \mathbf{y}) - 2(\mathbf{b}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\|_A^2 - \|\mathbf{X}\|_A^2 = \|A\mathbf{y} - \mathbf{b}\|_{A^{-1}}^2 - \|\mathbf{X}\|_A^2,$$

то данная задача может быть переформулирована следующим образом. Требуется найти $P_n(\lambda)$, $P_n(0) = 1$ такой, что норма невязки $\|\mathbf{r}^n\|_{A^{-1}}$ будет минимальной.

Покажем, что эта задача всегда разрешима единственным образом. Пусть

$$P_n(\lambda) = \sum_{k=0}^{n} c_k^{(n)} \lambda^k, \quad c_0^{(n)} = 1 \qquad \text{и} \qquad \mathbf{r}^0 = \sum_{i=1}^{q} r_i \mathbf{e}_i, \tag{3}$$

где $r_i \neq 0$, $i = 1, \ldots, q$, а $\mathbf{e}_i$ — собственные векторы A, соответствующие *различным* собственным значениям A. Представление (3) всегда возможно, так как матрица A симметрична и невырождена. Вектор $\mathbf{r}^n$ имеет вид

$$\mathbf{r}^n = P_n(A)\mathbf{r}^0 = \left(\sum_{k=0}^{n} c_k^{(n)} A^k \right) \mathbf{r}^0 = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k^{(n)} A^k \right) \mathbf{e}_i = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k^{(n)} \lambda_i^k \right) \mathbf{e}_i$$

и

$$\|\mathbf{r}^n\|_{A^{-1}}^2 = (A^{-1}\mathbf{r}^n, \mathbf{r}^n) = \sum_{i=1}^{q} r_i^2 \left(\sum_{k=0}^{n} c_k^{(n)} \lambda_i^{k-1} \right) \left(\sum_{j=0}^{n} c_j^{(n)} \lambda_i^j \right) =$$

$$= \sum_{k,j=0}^{n} c_k^{(n)} c_j^{(n)} \left(\sum_{i=1}^{q} \lambda_i^{k+j-1} r_i^2 \right).$$

Мы ищем минимум этого выражения относительно коэффициентов $c_l^{(n)}$. Приравнивая частные производные нулю, получим

$$\frac{1}{2} \frac{\partial}{\partial c_l^{(n)}} \|\mathbf{r}^n\|_{A^{-1}}^2 = 2 \sum_{j=0}^{n} c_j^{(n)} \left(\sum_{i=0}^{q} \lambda_i^{j+l-1} r_i^2 \right) =$$

$$= 2 \sum_{i=0}^{q} \left(\sum_{j=0}^{n} c_j^{(n)} \lambda_i^j r_i \lambda_i^{l-1} r_i \right) = 2(\mathbf{r}^n, A^{l-1}\mathbf{r}^0) = 0, \quad l = 1, \ldots, n.$$

Таким образом, в точке минимума должны выполняться равенства

$$(\mathbf{r}^n, A^l \mathbf{r}^0) = 0, \quad l = 0, \ldots, n-1. \tag{4}$$

Пусть $n \neq q - 1$. Из курса линейной алгебры известно, что векторы $\mathbf{r}^0$, $A\mathbf{r}^0, \ldots,$ $A^{q-1}\mathbf{r}^0$ образуют линейно независимую систему (пространство Крылова). Действительно, допустим противное. Тогда существуют постоянные $c_0, \ldots, c_{q-1}$, не равные нулю одновременно, такие, что $\sum_{i=0}^{q-1} c_i A^i \mathbf{r}^0 = \mathbf{0}$. Подставляя вместо $\mathbf{r}^0$ его разложение из (3), получим

$$\sum_{i=0}^{q-1} c_i A^i \mathbf{r}^0 = \sum_{i=0}^{q-1} c_i A^i \sum_{j=1}^{q} r_j \mathbf{e}_j = \sum_{i=0}^{q-1} c_i \sum_{j=1}^{q} r_j \lambda_j^i \mathbf{e}_j = \sum_{j=1}^{q} r_j \mathbf{e}_j \left(\sum_{i=0}^{q-1} c_i \lambda_j^i \right) = \mathbf{0}.$$

Таким образом, так как $r_j \neq 0$, то должны выполняться равенства

$$\sum_{i=0}^{q-1} \lambda_j^i c_i = 0, \quad j = 1, \ldots, q, \tag{5}$$

т. е. коэффициенты c_i должны удовлетворять системе уравнений (5). Определитель матрицы системы (5) совпадает с определителем Вандермонда и отличен от нуля, поскольку все λ_j различны по предположению. Отсюда следует, что равенства (5) могут выполняться только при $c_0 = \cdots = c_{q-1} = 0$. Таким образом, векторы $\mathbf{r}^0$, $A\mathbf{r}^0, \ldots, A^{q-1}\mathbf{r}^0$ действительно образуют линейно независимую систему.

Многочлен $P_n(\lambda)$ имеет n неизвестных коэффициентов ($P_n(0) = 1$); так как система линейных алгебраических уравнений (4) невырождена, то коэффициенты $c_k^{(n)}$, $k = 1, \ldots, n$ находятся из нее однозначно. Это означает, что поставленная задача всегда имеет единственное решение.

После нахождения коэффициентов $c_k^{(n)}$, $k = 1, \ldots, n$, значение $\mathbf{x}^n$ из (2) находится следующим образом. Имеем

$$A^{-1}\mathbf{r}^n = \mathbf{x}^n - \mathbf{X} = P_n(A)(\mathbf{x}^0 - \mathbf{X}) = \sum_{k=0}^{n} c_k^{(n)} A^k(\mathbf{x}^0 - \mathbf{X}) =$$

$$= \sum_{k=1}^{n} c_k^{(n)} A^k(\mathbf{x}^0 - \mathbf{X}) + \mathbf{x}^0 - \mathbf{X} = \mathbf{x}^0 - \mathbf{X} + \sum_{k=1}^{n} c_k^{(n)} A^{k-1}(A\mathbf{x}^0 - \mathbf{b}),$$

откуда следует

$$\mathbf{x}^n = \mathbf{x}^0 + \sum_{k=1}^{n} c_k^{(n)} A^{k-1}(A\mathbf{x}^0 - \mathbf{b}).$$

Обозначим через L_k линейную оболочку векторов $\mathbf{r}^0$, $A\mathbf{r}^0, \ldots, A^k\mathbf{r}^0$. Из построения видно, что

$$\mathbf{r}^j = P_j(A)\mathbf{r}^0 \in L_k, \quad j \leqslant k.$$

Покажем, что $\mathbf{r}^0$, $\mathbf{r}^1, \ldots, \mathbf{r}^n$ образуют базис в L_n. Доказательство проведем индукцией по n. Для $n = 0$ утверждение очевидно. Предположим, что для $n = k$ векторы $\mathbf{r}^0$, $\mathbf{r}^1, \ldots, \mathbf{r}^k$ образуют базис в L_k (и, как следствие, являются линейно независимыми). Покажем, что система $\mathbf{r}^0$, $\mathbf{r}^1, \ldots, \mathbf{r}^{k+1}$ также линейно независима. Предполагая противное, получим

$$\mathbf{r}^{k+1} = \sum_{i=0}^{k} \gamma_i \mathbf{r}^i, \quad \text{где} \quad \sum_{i=0}^{k} \gamma_i^2 \neq 0.$$

Тогда $\mathbf{r}^{k+1} \in L_k$. С другой стороны, в силу (4), вектор $\mathbf{r}^{k+1}$ ортогонален векторам $\mathbf{r}^0$, $A\mathbf{r}^0, \ldots, A^k\mathbf{r}^0$, образующим базис в L_k, а значит и всем векторам из L_k. Так как по предположению $\mathbf{r}^{k+1} \in L_k$, то это возможно

лишь при $\mathbf{r}^{k+1} = \mathbf{0}$, что противоречит исходной посылке о линейной зависимости $\mathbf{r}^0, \mathbf{r}^1, \ldots, \mathbf{r}^{k+1}$. Полученное противоречие доказывает линейную независимость $\mathbf{r}^0, \mathbf{r}^1, \ldots, \mathbf{r}^{k+1}$.

Так как $\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{n-1}\mathbf{r}^0$ образуют базис в L_{n-1}, то соотношения (4) означают, что вектор $\mathbf{r}^n$ ортогонален всему подпространству L_{n-1}, что в свою очередь может быть (по доказанному выше) записано в виде

$$(\mathbf{r}^n, \mathbf{r}^l) = 0, \quad l = 0, \ldots, n-1. \tag{6}$$

Покажем, что векторы $\mathbf{r}^0, \mathbf{r}^1, \ldots, \mathbf{r}^{n-1}, A\mathbf{r}^{n-1}$ также образуют базис в L_n. Вектор $\mathbf{r}^{n-1}$ по доказанному имеет вид

$$\mathbf{r}^{n-1} = \sum_{k=0}^{n-1} c_k^{(n-1)} A^k \mathbf{r}^0, \quad c_0^{(n-1)} = 1,$$

и этот вектор ортогонален векторам $\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{n-2}\mathbf{r}^0$. Коэффициент $c_{n-1}^{(n-1)}$ не равен нулю. Действительно, предполагая $c_{n-1}^{(n-1)} = 0$, получим

$$\mathbf{r}^{n-1} = \sum_{k=0}^{n-1} c_k^{(n-1)} A^k \mathbf{r}^0 = \sum_{k=0}^{n-2} c_k^{(n-1)} A^k \mathbf{r}^0.$$

Тогда по построению $\sum\limits_{k=0}^{n-2} c_k^{(n-1)} A^k = P_{n-2}(A)$, откуда $c_k^{(n-1)} = c_k^{(n-2)}$, $k = 0, \ldots, n-2$, и $\mathbf{r}^{n-1} = \mathbf{r}^{n-2}$, что противоречит ортогональности векторов $\mathbf{r}^{n-1}$ и $\mathbf{r}^{n-2}$.

Предположим, что векторы $\mathbf{r}^0, \mathbf{r}^1, \ldots, \mathbf{r}^{n-1}, A\mathbf{r}^{n-1}$ не образуют линейно независимую систему. Тогда $A\mathbf{r}^{n-1} \in L_{n-1}$. С другой стороны,

$$A\mathbf{r}^{n-1} = \sum_{k=0}^{n-1} c_k^{(n-1)} A^{k+1} \mathbf{r}^0 = \sum_{k=0}^{n-2} c_k^{(n-1)} A^{k+1} \mathbf{r}^0 + c_{n-1}^{(n-1)} A^n \mathbf{r}^0.$$

Но по доказанному выше векторы $\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^n\mathbf{r}^0$ линейно независимы при $n \leqslant q-1$ и $c_{n-1}^{(n-1)} \neq 0$, т. е. $A\mathbf{r}^{n-1} \notin L_{n-1}$.

Полученное противоречие показывает, что векторы $1\mathbf{r}^0, \mathbf{r}^1, \ldots, \mathbf{r}^{n-1}, A\mathbf{r}^{n-1}$ действительно образуют базис в L_n. Тогда вектор $\mathbf{r}^n \in L_n$ может быть разложен по этой системе единственным образом

$$\mathbf{r}^n = \sum_{k=0}^{n-1} \gamma_k \mathbf{r}^k + \gamma_n A\mathbf{r}^{n-1}. \tag{7}$$

Так как по построению $\mathbf{r}^n$ ортогонален $\mathbf{r}^j$ при $j = 0, \ldots, n-1$ и $(A\mathbf{r}^{n-1}, \mathbf{r}^j) = (\mathbf{r}^{n-1}, A\mathbf{r}^j) = 0$ при $j = 0, \ldots, n-3$, то из (7) следует, что $\gamma_k = 0$, $k = 0, \ldots, n-3$. Тогда (7) имеет вид

$$\mathbf{r}^n = \gamma_{n-1}\mathbf{r}^{n-1} + \gamma_{n-2}\mathbf{r}^{n-2} + \gamma_n A\mathbf{r}^{n-1}. \tag{8}$$

Из разложений

$$\mathbf{r}^j = \mathbf{r}^0 + \sum_{k=1}^{j} c_k^{(j)} A^k \mathbf{r}^0, \quad j = n-2,\ n-1, \qquad A\mathbf{r}^{n-1} = \sum_{k=0}^{n} p_k A^k \mathbf{r}^0$$

и условия $(A\mathbf{r}^{n-1}, \mathbf{r}^0) = 0$, $n \geqslant 2$, получаем $p_0 = 0$. Тогда из (8) имеем

$$\mathbf{r}^n = (\gamma_{n-1} + \gamma_{n-2})\mathbf{r}^0 + \sum_{k=1}^{n} c_k^{(n)} A^k \mathbf{r}^0 = P_n(A)\mathbf{r}^0.$$

Но $P_n(0) = 1$, откуда $\gamma_{n-1} + \gamma_{n-2} = 1$, и уравнение (8) может быть переписано в виде

$$\mathbf{r}^n = \gamma_{n-1}\mathbf{r}^{n-1} + (1 - \gamma_{n-1})\mathbf{r}^{n-2} + \gamma_n A\mathbf{r}^{n-1}.$$

Вводя обозначения $\gamma_{n-1} - 1 = \alpha_{n-1}$, $\gamma_n = \beta_{n-1}$, получим окончательное соотношение, связывающее невязки на трех соседних слоях

$$\mathbf{r}^n = \mathbf{r}^{n-1} + \alpha_{n-1}(\mathbf{r}^{n-1} - \mathbf{r}^{n-2}) + \beta_{n-1} A\mathbf{r}^{n-1}. \tag{9}$$

Подставляя в (9) вместо $\mathbf{r}^j$ его выражение $A\mathbf{x}^j - \mathbf{b}$ и применяя к обеим частям равенства оператор A^{-1}, получим

$$\mathbf{x}^n = \mathbf{x}^{n-1} + \alpha_{n-1}(\mathbf{x}^{n-1} - \mathbf{x}^{n-2}) + \beta_{n-1}(A\mathbf{x}^{n-1} - \mathbf{b}). \tag{10}$$

По доказанному выше, метод (10) эквивалентен исходному методу (2), который определен однозначно. Отсюда следует, что коэффициенты α_{n-1}, β_{n-1} находятся из условий (6) ортогональности невязок единственным образом; система уравнений для определения этих коэффициентов имеет вид

$$\begin{aligned} &(1 + \alpha_{n-1})\|\mathbf{r}^{n-1}\|^2 + \beta_{n-1}\|\mathbf{r}^{n-1}\|_A^2 = 0, \\ &-\alpha_{n-1}\|\mathbf{r}^{n-2}\|^2 + \beta_{n-1}(A\mathbf{r}^{n-1}, \mathbf{r}^{n-2}) = 0. \end{aligned} \tag{11}$$

На первом шаге, когда известно значение $\mathbf{x}^0$ и надо найти $\mathbf{x}^1$ из условия минимума функционала $F(\mathbf{x}^1)$, получаем формулу метода наискорейшего градиентного спуска

$$\mathbf{r}^1 = \mathbf{r}^0 - \frac{\|\mathbf{r}^0\|^2}{\|\mathbf{r}^0\|_A^2} A\mathbf{r}^0 \tag{12}$$

или, что то же самое,

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{\|\mathbf{r}^0\|^2}{\|\mathbf{r}^0\|_A^2}(A\mathbf{x}^0 - \mathbf{b}), \tag{13}$$

т. е. $\alpha_0 = 0$, $\beta_0 = -\dfrac{\|\mathbf{r}^0\|^2}{\|\mathbf{r}^0\|_A^2}$.

Покажем конечность итерационного процесса (10), т. е. что за конечное число итераций при отсутствии ошибок округления мы получим точное решение исходной системы (1). Пусть, как и ранее, $\mathbf{r}^0 = \sum_{i=1}^{q} r_i \mathbf{e}_i$, где $\mathbf{e}_i$ — собственные векторы A, отвечающие различным собственным значениям. Обозначим через L_{q-1} линейную оболочку векторов $\mathbf{e}_1, \ldots, \mathbf{e}_q$. Так как все $\mathbf{r}^k$, $k = 0, \ldots, q-1$, находятся из соотношений (9), (12), то $\mathbf{r}^k \in L_{q-1}$ и образуют в нем ортогональный базис. Вектор $\mathbf{r}^q \in L_{q-1}$ и по доказанному выше ортогонален векторам $\mathbf{r}^0, \ldots, \mathbf{r}^{q-1}$. Это возможно только в случае $\mathbf{r}^q = \mathbf{0}$.

Таким образом, производя q итераций по формулам (13), (9), мы получим точное решение системы уравнений (1) при условии отсутствия ошибок округления.

Оценим скорость сходимости метода. Для этого применим прием, который уже употреблялся при оценке скорости сходимости метода наискорейшего градиентного спуска. Пусть $\dot{\mathbf{y}}^0 = \mathbf{x}^0$ и $\mathbf{y}^j$ — приближения, получаемые при решении уравнения (1) линейным оптимальным процессом. Тогда по доказанному в § 6 погрешности $\mathbf{w}^n = \mathbf{y}^n - \mathbf{X}$ удовлетворяют соотношению

$$\mathbf{w}^n = Q_n(A)\mathbf{w}^0, \quad Q_n(0) = 1, \tag{14}$$

и имеет место оценка

$$\|Q_n(A)\| \leqslant \frac{2}{\lambda_0^n + \lambda_0^{-n}}, \quad \text{где} \quad \lambda_0 = \frac{\sqrt{M} + \sqrt{\mu}}{\sqrt{M} - \sqrt{\mu}}.$$

Отсюда, в частности, следует, что

$$\|\mathbf{w}^n\| \leqslant 2 \left(\frac{\sqrt{M} + \sqrt{\mu}}{\sqrt{M} - \sqrt{\mu}} \right)^n \|\mathbf{w}^0\|.$$

Получим оценку скорости сходимости линейного оптимального процесса в других нормах. Пусть $\mathbf{w}^0 = \sum_{i=1}^{q} w_i \mathbf{e}_i$. (Нетрудно видеть, что $w_i = r_i/\lambda_i$.) Тогда

$$\mathbf{w}^n = Q_n(A)\mathbf{w}^0 = \sum_{i=1}^{q} w_i Q_n(A)\mathbf{e}_i = \sum_{i=1}^{q} w_i Q_n(\lambda_i)\mathbf{e}_i$$

и

$$\|\mathbf{w}^n\|_A^2 = (A\mathbf{w}^n, \mathbf{w}^n) = \sum_{i=1}^{q} w_i^2 \lambda_i Q_n^2(\lambda_i) \leqslant \max_{\lambda \in [\mu, M]} Q_n^2(\lambda) \sum_{i=1}^{q} \lambda_i w_i^2 =$$

$$= \|Q_n(A)\|^2 \|\mathbf{w}^0\|_A^2 \leqslant \left(\frac{2}{\lambda_0^n + \lambda_0^{-n}} \right)^2 \|\mathbf{w}^0\|_A^2,$$

т. е.

$$\|\mathbf{w}^n\|_A \leqslant \frac{2}{\lambda_0^n + \lambda_0^{-n}} \|\mathbf{w}^0\|_A. \tag{15}$$

Так как

$$\|\mathbf{w}^n\|_A^2 = \Big(A(\mathbf{y}^n - \mathbf{X}),\, \mathbf{y}^n - \mathbf{X}\Big) = \Big(A^{-1}(A\mathbf{y}^n - \mathbf{b}),\, A\mathbf{y}^n - \mathbf{b}\Big) = \|\mathbf{z}^n\|_{A^{-1}}^2,$$

где $\mathbf{z}^n = A\mathbf{w}^n$, то из (15) следует оценка

$$\|\mathbf{z}^n\|_{A^{-1}} \leqslant \frac{2}{\lambda_0^n + \lambda_0^{-n}} \|\mathbf{z}^0\|_{A^{-1}}. \tag{16}$$

Применяя к обеим частям (14) матрицу A, получим уравнение, связывающее невязки на n-м и нулевом шагах линейного оптимального процесса

$$\mathbf{z}^n = Q_n(A)\mathbf{z}^0 = Q_n(A)\mathbf{r}^0. \tag{17}$$

По построению вектор $\mathbf{r}^n$ минимизирует функционал $\|\mathbf{y}\|_{A^{-1}}$ среди векторов вида (2) или, что то же самое, $\mathbf{x}^n$ минимизирует $F(\mathbf{y})$ среди векторов $\mathbf{y}$ таких, что $\mathbf{r} = \mathbf{y} - \mathbf{X}$ имеет вид (2). Тогда $F(\mathbf{x}^n) \leqslant F(\mathbf{y}^n)$, откуда

$$\|\mathbf{r}^n\|_{A^{-1}} \leqslant \|\mathbf{z}^n\|_{A^{-1}} \leqslant \frac{2}{\lambda_0^n + \lambda_0^{-n}} \|\mathbf{z}^0\|_{A^{-1}} = \frac{2}{\lambda_0^n + \lambda_0^{-n}} \|\mathbf{r}^0\|_{A^{-1}}.$$

Но так как $\|\mathbf{r}^n\|_{A^{-1}} = \|\mathbf{x}^n - \mathbf{X}\|_A$, то из последнего соотношения получаем оценку скорости сходимости метода сопряженных градиентов

$$\|\mathbf{x}^n - \mathbf{X}\|_A \leqslant \frac{2}{\lambda_0^n + \lambda_0^{-n}} \|\mathbf{x}^0 - \mathbf{X}\|_A. \tag{18}$$

Замечание. Мы не обсуждали вопрос численной устойчивости нахождения параметров α_n, β_n, поскольку он требует специального рассмотрения.

§ 10. Итерационные методы с использованием спектрально-эквивалентных операторов

Кроме методов простой итерации вида

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha(A\mathbf{x}^n - \mathbf{b}) \tag{1}$$

часто применяются итерационные методы

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n - \alpha(A\mathbf{x}^n - \mathbf{b}), \tag{2}$$

где матрица $B \neq E$ такова, что система уравнений $B\mathbf{y} = \mathbf{c}$ легко может быть решена. Если соотношение (2) умножить на B^{-1}, то получится

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha B^{-1}(A\mathbf{x}^n - \mathbf{b}).$$

Таким образом, итерационный процесс (2) равносилен методу простой итерации с матрицей $E - \alpha B^{-1}A$.

Рассмотрим случай, когда $A > 0$ и отношение максимального M и минимального μ из собственных значений матрицы A велико. Тогда рассматривавшиеся ранее итерационные процессы сходятся медленно. Пусть $B > 0$ и

$$M_1 = \sup_{\mathbf{x}} \frac{(A\mathbf{x}, \mathbf{x})}{(B\mathbf{x}, \mathbf{x})}, \quad \mu_1 = \inf_{\mathbf{x}} \frac{(A\mathbf{x}, \mathbf{x})}{(B\mathbf{x}, \mathbf{x})}. \tag{3}$$

Предположим, что система уравнений с матрицей B легко решается и $M_1/\mu_1 \ll M/\mu$.

В случае, когда отношение M_1/μ_1 не очень велико, итерационные методы типа рассматриваемых в настоящем параграфе принято условно называть *итерационными методами с использованием спектрально-эквивалентных операторов* (Л. В. Канторович, Е. Г. Дьяконов). В настоящее время эти итерационные методы называют *методами с переобуславливанием*, а матрицу (оператор) B *переобуславливателем*. Покажем, что при удачном подборе матрицы B метод итераций (2) обладает лучшей сходимостью по сравнению с простейшим методом (1).

Точное решение $\mathbf{X}$ удовлетворяет равенству

$$B\mathbf{X} = B\mathbf{X} - \alpha(A\mathbf{x} - \mathbf{b});$$

вычитая его из (2), получим уравнение относительно погрешности

$$B\mathbf{r}^{n+1} = B\mathbf{r}^n - \alpha A\mathbf{r}^n. \tag{4}$$

Приведем матрицу B при помощи ортогонального преобразования к диагональному виду. Пусть $B = U^T \Lambda U$, где

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix},$$

U — ортогональная матрица. Заметим, что все $\lambda_j > 0$.

Через $\sqrt{B}$ принято обозначать матрицу вида

$$U^T \begin{pmatrix} +\sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & +\sqrt{\lambda_m} \end{pmatrix} U.$$

Очевидно, $\sqrt{B} > 0$, $(\sqrt{B})^T = \sqrt{B}$, $\sqrt{B}\sqrt{B} = B$. Умножим обе части уравнения (4) слева на $(\sqrt{B})^{-1}$ и положим $\sqrt{B}\mathbf{r}^n = \mathbf{v}^n$. Получим равенство

$$\mathbf{v}^{n+1} = \mathbf{v}^n - \alpha C\mathbf{v}^n, \tag{5}$$

где $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$.

Вследствие соотношений $A = A^T$, $(\sqrt{B}) = (\sqrt{B})^T$, матрица C симметрична. Рассмотрим выражение

$$w(x) = (C\mathbf{x}, \mathbf{x})/(\mathbf{x}, \mathbf{x}).$$

Положив $(\sqrt{B})^{-1}\mathbf{x} = \mathbf{y}$, можем написать равенство

$$w(x) = w(\sqrt{B}\mathbf{y}) = (A\mathbf{y}, \mathbf{y})/(B\mathbf{y}, \mathbf{y}).$$

Согласно (3) имеем $w(x) \in [\mu_1, M_1]$, поэтому все собственные значения матрицы C также принадлежат отрезку $[\mu_1, M_1]$.

При $\alpha = 2/(M_1 + \mu_1)$ собственные значения матрицы $E - \alpha C$ по модулю не превосходят величины $(M_1 - \mu_1)/(M_1 + \mu_1)$. Поэтому $\|E - \alpha C\|_2 \leqslant (M_1 - \mu_1)/(M_1 + \mu_1)$ и аналогично (6.13) имеем

$$\|\mathbf{v}^n\|_2 \leqslant \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \|\mathbf{v}^0\|_2. \tag{6}$$

Напомним, что

$$\|\mathbf{v}^n\|_2 = \sqrt{(\sqrt{B}\mathbf{r}^n, \sqrt{B}\mathbf{r}^n)} = \sqrt{(B\mathbf{r}^n, \mathbf{r}^n)}.$$

Если $0 < \mu_2 \leqslant (B\mathbf{y}, \mathbf{y})/(\mathbf{y}, \mathbf{y}) \leqslant M_2$ при всех $\mathbf{y}$, то из (6) следует оценка

$$\|\mathbf{r}^n\|_2 \leqslant \frac{\|\mathbf{v}^n\|_2}{\sqrt{\mu_2}} \leqslant \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \frac{\|\mathbf{v}^0\|_2}{\sqrt{\mu_2}} \leqslant \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \sqrt{\frac{M_2}{\mu_2}}\, \|\mathbf{r}^0\|_2.$$

Так как функция

$$\frac{M - \mu}{M + \mu} = \frac{1 - (\mu/M)}{1 + (\mu/M)}$$

монотонно убывает с уменьшением M/μ, то скорость сходимости нового итерационного процесса быстрее, чем у (1).

При окончательном решении вопроса о переходе от итераций по формулам (1) *к итерациям по формулам* (2) *следует учесть количество арифметических операций, выполняемых во время этих итераций. Если итерации по формулам* (2) *требуют существенно большего количества операций, то такой переход может оказаться нецелесообразным.*

Задача 1. Пусть $T(A)$ — количество арифметических операций, выполняемых во время одной итерации по формуле (1), $T(B)$ — количество арифметических операций, выполняемых во время одной итерации по формуле (2). Привести соображения в пользу того, что переход к итерациям по формуле (2) целесообразен, если

$$T(B)\left(\ln\frac{M_1 + \mu_1}{M_1 - \mu_1}\right)^{-1} < T(A)\left(\ln\frac{M + \mu}{M - \mu}\right)^{-1}.$$

По аналогии с тем, как по итерационному методу (1) был построен метод (2), можно построить аналог любого из рассматривавшихся в §§ 8–10 методов.

Пусть имеется итерационный процесс, где погрешности последующих приближений связаны равенством

$$\mathbf{r}^{n+1} = \left(E - \sum_{i=1}^{k} \alpha_i A^i\right) \mathbf{r}^n.$$

Напишем равенство

$$\mathbf{v}^{n+1} = \left(E - \sum_{i=1}^{k} \alpha_i C^i\right) \mathbf{v}^n, \tag{7}$$

где $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$, положим $\mathbf{v}^n = \sqrt{B}\mathbf{r}^n$ и умножим (7) слева на $\sqrt{B}$. Получим уравнение

$$B\mathbf{r}^{n+1} = B\mathbf{r}^n - \sum_{i=1}^{k} \alpha_i (AB^{-1})^{i-1} A\mathbf{r}^n.$$

Такое уравнение для погрешности соответствует итерационному процессу

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n - \sum_{i=1}^{k} \alpha_i (AB^{-1})^{i-1} (A\mathbf{x}^n - \mathbf{b}). \tag{8}$$

Если произвести оптимизацию параметров α_i так, чтобы оценка отношения норм $\|\mathbf{v}^{n+1}\|_2 / \|\mathbf{v}^n\|_2$ была наилучшей, то так же, как в § 6, получится наилучший итерационный метод вида (8).

Аналогом оптимального линейного итерационного метода (6.25), (6.26) будет итерационный процесс

$$B\mathbf{y}^{n+1} = B\mathbf{y}^n + \bar{\omega}_n \bar{\omega}_{n-1} B(\mathbf{y}^n - \mathbf{y}^{n-1}) - \frac{2}{M_1 + \mu_1}(1 + \bar{\omega}_n \bar{\omega}_{n-1})(A\mathbf{y}^n - \mathbf{b}),$$

где $\bar{\omega}_n$ строятся по M_1 и μ_1 так же, как ω_n по M и μ (см. (6.26)).

Объединением метода (2) с методом наискорейшего спуска является следующий метод. Приближение $\mathbf{x}^{n+1}$ ищется из соотношения

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n - \alpha_n (A\mathbf{x}^n - \mathbf{b}). \tag{9}$$

При этом коэффициент α_n определятся из условия минимума функционала

$$F(\mathbf{x}^{n+1}) = (A\mathbf{x}^{n+1}, \mathbf{x}^{n+1}) - 2(\mathbf{b}, \mathbf{x}^{n+1}).$$

Из (9) получаем

$$B\frac{d\mathbf{x}^{n+1}}{d\alpha_n} = -(A\mathbf{x}^n - \mathbf{b}).$$

Поскольку

$$\frac{dF(\mathbf{x}^{n+1})}{d\alpha_n} = 2\left(A\mathbf{x}^{n+1} - \mathbf{b}, \frac{d\mathbf{x}^{n+1}}{d\alpha_n}\right) =$$

$$= 2(A(\mathbf{x}^n - \alpha_n B^{-1}(A\mathbf{x}^n - \mathbf{b})) - \mathbf{b}, -B^{-1}(A\mathbf{x}^n - \mathbf{b})),$$

то условие $dF(\mathbf{x}^{n+1}/d\alpha_n = 0$ является линейным уравнением относительно α_n. Аналог итерационного процесса (9.10) имеет вид

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n + \alpha_n(B\mathbf{x}^n - B\mathbf{x}^{n-1}) + \beta_n(A\mathbf{x}^n - \mathbf{b}).$$

Методы, рассмотренные в этом параграфе, получили широкое применение при решении систем уравнений $A\mathbf{x} = \mathbf{b}$ с большим разбросом собственных значений матрицы $A > 0$.

§ 11. Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация

Предположим, что матрица и правая часть системы заданы неточно и вместо предъявленной к решению системы

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

в действительности должна была решаться некоторая система

$$A_1\mathbf{x} = \mathbf{b}_1, \quad A_1 = A + \Delta, \quad \mathbf{b}_1 = \mathbf{b} + \eta. \tag{2}$$

Пусть известны оценки $\|\Delta\|$ и $\|\eta\|$. Займемся оценкой погрешности решения.

Сначала выделим главный член погрешности. Будем обозначать решения (1) и (2) через $\mathbf{X}$ и $\mathbf{X}^*$ и разность $\mathbf{X}^* - \mathbf{X}$ — через $\mathbf{r}$. Подставив выражения A_1, $\mathbf{b}_1$ и $\mathbf{X}^*$ в (2), будем иметь

$$(A + \Delta)(\mathbf{X} + \mathbf{r}) = \mathbf{b} + \eta.$$

Вычитая из этого равенства (1), получим

$$A\mathbf{r} + \Delta\mathbf{X} + \Delta\mathbf{r} = \eta,$$

откуда

$$A\mathbf{r} = \eta - \Delta\mathbf{X} - \Delta\mathbf{r}$$

и

$$\mathbf{r} = A^{-1}(\eta - \Delta\mathbf{X} - \Delta\mathbf{r}). \tag{3}$$

Если $\|\Delta\|$ и $\|\eta\|$ малы, то следует ожидать и малости $\|\mathbf{r}\|$. Тогда слагаемое $\Delta\mathbf{r}$ имеет более высокий порядок малости; отбрасывая это слагаемое, получаем

$$\mathbf{r} \approx A^{-1}(\eta - \Delta\mathbf{X}).$$

Отсюда следует оценка погрешности

$$\|\mathbf{r}\| \leqslant \sigma \approx \|A^{-1}\| \left(\|\eta\| + \|\Delta\| \, \|\mathbf{X}\|\right). \tag{4}$$

Строгая оценка погрешности получается следующим образом. Вследствие (3) выполняется неравенство

$$\|\mathbf{r}\| \leqslant \|A^{-1}\| \, \|\eta\| + \|A^{-1}\| \, \|\Delta\| \, \|\mathbf{X}\| + \|A^{-1}\| \, \|\Delta\| \, \|\mathbf{r}\|.$$

Предположим, что $\|A^{-1}\|\|\Delta\| < 1$. Перенеся последнее слагаемое в левую часть и поделив неравенство на коэффициент при $\|\mathbf{r}\|$, получим оценку

$$\|\mathbf{r}\| \leqslant \frac{\|A^{-1}\|\left(\|\eta\| + \|\Delta\|\,\|\mathbf{X}\|\right)}{1 - \|A^{-1}\|\,\|\Delta\|}. \tag{5}$$

Довольно распространен случай, когда погрешность матрицы системы существенно меньше погрешности правой части. В качестве модели этой ситуации будем рассматривать случай точного задания матрицы системы. Тогда, полагая в (5) $\Delta = 0$, имеем

$$\|\mathbf{r}\| \leqslant \|A^{-1}\|\,\|\eta\|.$$

Для качественной характеристики связи между погрешностями правой части и решения вводятся понятия *обусловленности системы* и *обусловленности матрицы системы*. Абсолютные погрешности правой части и решения системы зависят от масштабов, которыми измеряются коэффициенты системы. Поэтому удобнее характеризовать свойства системы через связь между относительными погрешностями правой части и решения.

Соответственно этому в качестве *меры обусловленности системы* принимается число

$$\tau = \sup_{\eta}\left(\frac{\|\mathbf{r}\|}{\|\mathbf{X}\|} : \frac{\|\eta\|}{\|\mathbf{b}\|}\right) = \frac{\|\mathbf{b}\|}{\|\mathbf{X}\|}\sup_{\eta}\frac{\|\mathbf{r}\|}{\|\eta\|}.$$

Отсюда получаем оценку относительной погрешности решения через меру обусловленности системы и относительную погрешность правой части:

$$\frac{\|\mathbf{r}\|}{\|\mathbf{X}\|} \leqslant \tau\frac{\|\eta\|}{\|\mathbf{b}\|}. \tag{6}$$

Так как $\mathbf{r} = A^{-1}\eta$, то

$$\sup_{\eta}\frac{\|\mathbf{r}\|}{\|\eta\|} = \|A^{-1}\|$$

и

$$\tau = \frac{\|\mathbf{b}\|}{\|\mathbf{X}\|}\|A^{-1}\|.$$

Иногда удобнее иметь более грубую характеристику свойств системы только через свойства матрицы A. Эту характеристику $\nu(A) = \sup_{\mathbf{b}}\tau$ называют *мерой* (или *числом*) *обусловленности матрицы* A. Согласно этому определению и (6), имеем оценку

$$\frac{\|\mathbf{r}\|}{\|\mathbf{X}\|} \leqslant \nu(A)\frac{\|\eta\|}{\|\mathbf{b}\|},$$

связывающую относительные погрешности правой части и решения только через свойства матрицы системы. Так как

$$\sup_{\mathbf{b}} \frac{\|\mathbf{b}\|}{\|\mathbf{X}\|} = \sup_{\mathbf{x}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\|,$$

то

$$\nu(A) = \|A\|\,\|A^{-1}\|.$$

Поскольку любая норма матрицы не меньше ее наибольшего по модулю собственного значения, то $\|A\| \geqslant \max|\lambda_A|$; поскольку собственные значения матриц A и A^{-1} взаимно обратны, то

$$\|A^{-1}\| \geqslant \max \frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}.$$

Таким образом,

$$\nu(A) \geqslant \max|\lambda_A| \Big/ \min|\lambda_A| \geqslant 1.$$

В частности, при $A = A^T$ имеем $\|A\|_2 = \max|\lambda_A|$ и

$$\|A^{-1}\|_2 = \max \frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}.$$

Следовательно, в случае нормы $\|\cdot\|_2$

$$\nu(A) = \max|\lambda_A| \Big/ \min|\lambda_A|.$$

Рассмотрим вопрос о погрешности решения вследствие округления в ЭВМ правой части. Пусть, как обычно, t — двоичная разрядность чисел в ЭВМ. Каждый элемент b_i правой части округляется с относительной погрешностью $O(2^{-t})$, т. е. с абсолютной погрешностью, равной $O(|b_i|2^{-t})$, поэтому

$$\|\eta\| = O\Big(\|\mathbf{b}\|2^{-t}\Big) \qquad \text{и} \qquad \|\eta\| \Big/ \|\mathbf{b}\| = O(2^{-t}).$$

Следовательно,

$$\|\mathbf{r}\| \Big/ \|\mathbf{X}\| \leqslant \nu(A)\, O(2^{-t}).$$

При практической работе вопрос о строгой оценке погрешности полученного приближенного решения системы линейных уравнений с помощью полученных неравенств или каким-либо иным способом возникает редко. Однако информация о порядке погрешности решения часто полезна для получения качественных выводов о том, с какой точностью разумно решать задачу. Соотношения (4), (5) оценивают сверху погрешность решения, являющуюся следствием погрешности исходных данных. Из равенства (3) видно, что оценки (4), (5) довольно точны, поэтому обычно не имеет смысла стремиться получать решение задачи с погрешностью, существенно меньшей чем σ.

Системы уравнений и матрицы с большими значениями мер обусловленности принято называть *плохо обусловленными*, а с малыми — *хорошо обусловленными*. Если правая часть (4), оценивающая погрешность решения через погрешность исходных данных, или оценка вычислительной погрешности недопустимо велики, то полезно принять во внимание какую-то дополнительную информацию о решении рассматриваемой задачи. Подход к решению такой задачи должен быть таким же, как в случае *некорректных задач*.

Рассмотрим простейший случай, когда A — симметричная матрица. Пусть $\lambda_1, \ldots, \lambda_m \neq 0$ — ее собственные значения, упорядоченные в порядке убывания $|\lambda_i|$; соответствующую ортонормированную систему собственных векторов обозначим через $\mathbf{e}_1, \ldots, \mathbf{e}_m$. Решением системы $A\mathbf{x} = \mathbf{b}$ является вектор

$$\mathbf{X} = \sum_{i=1}^{m} c_i \mathbf{e}_i, \quad c_i = \frac{(\mathbf{b}, \mathbf{e}_i)}{\lambda_i}.$$

При реально заданной правой части $\tilde{\mathbf{b}} = \mathbf{b} + \eta$ решением будет

$$\tilde{\mathbf{X}} = \sum_{i=1}^{m} \frac{(\mathbf{b}, \mathbf{e}_i) + (\eta, \mathbf{e}_i)}{\lambda_i} \mathbf{e}_i.$$

Коэффициент $(\eta, \mathbf{e}_i)/\lambda_i$ может оказаться очень большим при малых λ_i, что ведет к сильному искажению решения. Иногда заранее известно, что в разложении искомого решения задачи $\sum_{i=1}^{m} c_i \mathbf{e}_i$ коэффициенты c_i, соответствующие малым по модулю λ_i, малы. В этом случае следует принять какие-то меры с тем, чтобы «отфильтровать» эти составляющие решения.

При небольших m для решения этой задачи иногда применяется следующий способ: задаются некоторым $q > 0$, находят все λ_i и $\mathbf{e}_i$ при $i \leqslant q$ и полагают

$$\mathbf{X} \approx \sum_{i=1}^{q} \frac{(\tilde{\mathbf{b}}, \mathbf{e}_i)}{\lambda_i} \mathbf{e}_i;$$

q следует подобрать исходя из дополнительной информации о задаче.

Другие два способа проиллюстрируем на примере матриц, где все $\lambda_i > 0$.

Первый способ. Задавая некоторое $\alpha > 0$, находим решение $\mathbf{x}^\alpha$ системы

$$(\alpha E + A)\mathbf{x}^\alpha = \tilde{\mathbf{b}}.$$

Оно записывается в виде

$$\mathbf{x}_\alpha = \sum_{i=1}^{m} \frac{(\tilde{\mathbf{b}}, \mathbf{e}_i)}{\lambda_i + \alpha} \mathbf{e}_i.$$

Так как

$$\frac{1}{\lambda_i} - \frac{1}{\lambda_i + \alpha} = \frac{\alpha}{\lambda_i(\lambda_i + \alpha)},$$

то наличие малого параметра α несущественно изменит слагаемые, соответствующие большим λ_i. В то же время при $\lambda_i \ll \alpha$ имеем

$$\left|\frac{(\tilde{\mathbf{b}}, \mathbf{e}_i)}{\lambda_i + \alpha}\right| \ll \left|\frac{(\tilde{\mathbf{b}}, \mathbf{e}_i)}{\lambda_i}\right|.$$

Это означает, что введение параметра α приводит к существенному уменьшению роли слагаемых, соответствуюших малым λ_i. Подбор оптимального значения α обычно осуществляют экспериментально, сравнивая результаты расчетов при различных α.

Второй способ заключается в следующем. Будем решать систему уравнений каким-либо итерационным способом. Рассмотрим случай итераций по формуле

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha(A\mathbf{x}^n - \mathbf{b}) \tag{7}$$

при некотором начальном приближении $\mathbf{x}^0$. Пусть

$$\tilde{\mathbf{b}} = \sum_{i=1}^{m} \beta_i \mathbf{e}_i, \quad \mathbf{x}^n = \sum_{i=1}^{m} z_i^n \mathbf{e}_i.$$

Подставим эти выражения в (7) и, приравнивая коэффициенты при $\mathbf{e}_i$, получим соотношения

$$z_i^{n+1} = \alpha\beta_i + (1 - \alpha\lambda_i) z_i^n.$$

Последовательно выражая каждое z_i^k через предыдущие, имеем

$$z_i^n = \alpha\beta_i + (1 - \alpha\lambda_i) z_i^{n-1} = \alpha\beta_i + (1 - \alpha\lambda_i)(\alpha\beta_i + (1 - \alpha\lambda_i) z_i^{n-2}) = \cdots$$
$$= \alpha\beta_i \sum_{k=0}^{n-1} (1 - \alpha\lambda_i)^k + (1 - \alpha\lambda_i)^n z_i^0.$$

Если $|1 - \alpha\lambda_i| < 1$, то при $n \to \infty$

$$\sum_{k=0}^{n-1} (1 - \alpha\lambda_i)^k \to \sum_{k=0}^{\infty} (1 - \alpha\lambda_i)^k = \frac{1}{\alpha\lambda_i}; \quad (1 - \alpha\lambda_i)^n \to 0,$$

поэтому $z_i^n \to \beta_i/\lambda_i$. Пусть $\alpha \approx (\max \lambda_i)^{-1}$, т. е. относительно мало. При больших значениях λ_i величина $(1 - \alpha\lambda_i)^n$ быстро стремится к нулю с ростом n и z_i^n близко к своему предельному значению β_i/λ_i. В то же время иногда удается подобрать начальное приближение, для которого величины z_i^0 относительно малы при малых λ_i. Тогда при небольшом n коэффициенты z_i^n, соответствующие таким λ_i, еще не будут недопустимо большими и получаемое приближение может оказаться приемлемым.

В других случаях решение задачи находят, минимизируя некоторый функционал, близкий к $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$, например функционал $F(\mathbf{x}) + \alpha(\mathbf{x}, \mathbf{x})$ с малым $\alpha > 0$.

Успешность применения описанных приемов в случае несимметричных матриц A в существенной степени зависит от структуры жордановой формы и от ряда свойств матрицы. Здесь часто решение находят, минимизируя функционал

$$(A\mathbf{x} - \mathbf{b},\ A\mathbf{x} - \mathbf{b}) + \alpha(\mathbf{x}, \mathbf{x})$$

при малых $\alpha > 0$; значение α опять-таки подбирается экспериментально из сравнения результатов расчетов при различных α.

Другая группа методов основана на представлении матрицы системы A в виде

$$A = G\Lambda P,$$

где G и P — ортогональные матрицы, а Λ — двухдиагональная матрица, у которой могут быть отличными от нуля элементы λ_{ij} при $j = i$ и $j = i + 1$.

Большинство из описанных методов решения систем уравнений с плохо обусловленной матрицей относится к *методам регуляризации.*

Задача 1. Пусть $A^{(m)} = [a_{ij}^m]$ — матрица размерности $m \times m$ с элементами

$$a_{ij}^m = \begin{cases} p & \text{при} \quad j = i, \\ q & \text{при} \quad j = i + 1, \\ 0 & \text{при} \quad j \neq i, i + 1. \end{cases}$$

1. Вычислить матрицу $(A^{(m)})^{-1}$ и доказать утверждение: при $|q| < |p|$ матрицы $A^{(m)}$ в некотором смысле хорошо обусловлены, а при $|q| > |p|$ и m больших плохо обусловлены.

2. Выписать явно решение системы $A^{(m)}\mathbf{x} = \mathbf{b}$ через правую часть.

3. Выписать явно через правую часть $\mathbf{b}$ вектор $\mathbf{x}_\alpha$, минимизирующий функционал

$$(A^{(m)}\mathbf{x} - \mathbf{b},\ A^{(m)}\mathbf{x} - \mathbf{b}) + \alpha(\mathbf{x}, \mathbf{x}).$$

4. Попытаться качественно описать эффект, достигаемый за счет применения такой регуляризации.

Объясним еще одну причину, по которой стараются избегать симметризации матриц, предложенной в § 6. Сначала посмотрим, что происходит, когда операция симметризации применяется формально в случае симметричной матрицы. Тогда $A = A^T$ и $A^TA = A^2$. При возведении матрицы в квадрат собственные значения возводятся в квадрат, поэтому при $\|A\| = \|A\|_2$ имеем

$$\nu(A^2) = \frac{\max|\lambda_{A^2}|}{\min|\lambda_{A^2}|} = \frac{(\max|\lambda_A|)^2}{(\min|\lambda_A|)^2} = (\nu(A))^2.$$

Если $\nu(A) = 1$, то $\max|\lambda_A| = \min|\lambda_A|$ и все собственные значения матрицы A равны между собой по модулю, т. е. $A = \text{const} \cdot E$. За исключением этого частного случая $\nu(A) > 1$, поэтому $\nu(A^2) > \nu(A)$. Таким образом, в случае симметричной матрицы применение симметризации увеличивает число обусловленности. Число обусловленности является непрерывной функцией от матрицы, поэтому для матриц, близких к симметричным, применение симметризации также увеличивает число обусловленности матриц. Таким образом, имеется непустой класс несимметричных матриц, по отношению к которым алгоритм симметризации увеличивает число обусловленности.

Задача 2. Существуют ли несимметричные матрицы, для которых $\nu(A^2) = (\nu(A))^2$?

Рассмотрим еще один метод решения плохо обусловленных систем линейных алгебраических уравнений. Пусть

$$A\mathbf{x} = \mathbf{b}. \tag{8}$$

Относительно A будем считать, что в спектре матрицы A^*A есть как собственные числа λ_j порядка 1, так и собственные числа, близкие (или даже равные) к нулю. Это как раз и означает, что матрица A плохо обусловлена.

Заметим, что в силу наших предположений относительно собственных значений матрицы A^*A, часть из них может быть равна нулю. Таким образом, уравнение (8), вообще говоря, может не иметь решения в классическом смысле.

Назовем решением $\mathbf{X}$ уравнения (8) вектор, который минимизирует функционал невязки, а именно,

$$\mathbf{X} = \arg\min_{\mathbf{y}} \|A\mathbf{y} - \mathbf{b}\|; \tag{9}$$

здесь и далее в этом параграфе под нормой мы будем понимать евклидову норму вектора. Выписывая уравнение Эйлера для функционала $\Phi(\mathbf{y}) = \|A\mathbf{y} - \mathbf{b}\|^2$, мы получим

$$A^*A\mathbf{x} = A^*\mathbf{b}. \tag{10}$$

Уравнение (10), в отличие от (8), всегда имеет решение. Действительно, непосредственной проверкой убеждаемся, что $\ker A^*A = \ker A$. Необходимым и достаточным условием существования решения линейной системы уравнений (10) является ортогональность правой части ядру матрицы системы, т. е. вектор $A^*\mathbf{b}$ должен быть ортогонален ядру $\ker A^*A$, которое, как мы отметили выше, совпадает с ядром $\ker A$. Но из вида правой части видно, что она действительно ортогональна ядру A. Таким образом, система (10) всегда имеет решение. В общем случае таких решений может быть несколько.

Описываемый ниже метод заключается в минимизации функционала $\Phi(\mathbf{y})$ *методом оптимального покоординатного спуска*; на каждом шаге выбирается координата, спуск по которой будет оптимальным в смысле минимизации $\Phi(\mathbf{y})$. В качестве координатных (базисных) векторов можно выбрать любую ортонормированную систему.

Пусть $\mathbf{w}_1, \ldots, \mathbf{w}_q$ — ортонормированная система векторов в R_m (не обязательно базис) и $A\mathbf{w}_j \neq 0$ по крайней мере для одного из векторов. Обозначим через W линейную оболочку векторов $\mathbf{w}_1, \ldots, \mathbf{w}_q$. Будем искать вектор, минимизирующий функционал невязки $\Phi(\mathbf{y})$ на подпространстве W. Для этого рассмотрим следующий итерационный метод. Положим

$$\mathbf{x}^0 = \mathbf{0}.$$

Если приближение $\mathbf{x}^k$ уже найдено, то следующее приближение $\mathbf{x}^{k+1}$ будем искать в виде $\mathbf{x}^{k+1} = \mathbf{x}^k + C_k\mathbf{w}_{j_k}$, $C_k = \text{const}$, где

$$j_k = \arg\min_j \left(\min_{C_k} \Phi(\mathbf{x}^{k+1})\right). \tag{11}$$

Наряду с приближениями $\mathbf{x}^k$ введем невязки

$$\xi^k = A\mathbf{x}^k - \mathbf{b}. \tag{12}$$

Выпишем условия минимума функционала $\Phi(\mathbf{x}^{k+1})$ по C_k. Имеем

$$\Phi(\mathbf{x}^{k+1}) = \|C_kA\mathbf{w}_j + A\mathbf{x}^k - \mathbf{b}\|^2 = C_k^2\|A\mathbf{w}_j\|^2 + 2C_k(A\mathbf{w}_j, \xi^k) + \|\xi^k\|^2. \tag{13}$$

Заметим, что при поиске минимума $\Phi(\mathbf{x}^{k+1})$ достаточно рассматривать только те $\mathbf{w}_j$, для которых $\|A\mathbf{w}_j\| \neq 0$, так как в противном случае значение функционала не меняется. Функция $\Phi(\mathbf{x}^{k+1})$, как функция переменной C_k, является многочленом второй степени, причем коэффициент при C_k^2 положителен в силу замечания выше. Отсюда следует, что минимум $\Phi(\mathbf{x}^{k+1})$ по C_k при фиксированном j существует и единствен, если $\|A\mathbf{w}_j\| \neq 0$. Таким образом, из (13) следует, что C_k удовлетворяет уравнению

$$\frac{\partial\Phi(\mathbf{x}^{k+1})}{\partial C_k} = 2C_k\|A\mathbf{w}_j\|^2 + 2(A\mathbf{w}_j, \xi^k) = 0,$$

откуда

$$C_k = -\frac{(A\mathbf{w}_j, \xi^k)}{\|A\mathbf{w}_j\|^2}.$$

При таком выборе C_k

$$\Phi(\mathbf{x}^{k+1}) = \|\xi^{k+1}\|^2 = \|\xi^k\|^2 - \frac{(A\mathbf{w}_j, \xi^k)^2}{\|A\mathbf{w}_j\|^2}$$

и

$$j_k = \arg\max_j \frac{\left|(A\mathbf{w}_j, \xi^k)\right|}{\|A\mathbf{w}_j\|}, \quad \mathbf{x}^{k+1} = \mathbf{x}^k + C_k\mathbf{w}_{j_k}.$$

Суммируя вышесказанное, получаем следующий алгоритм:

1) Вычисляем векторы $A\mathbf{w}_j$, $j = 1, \ldots, q$, и их нормы $\|A\mathbf{w}_j\|$; в дальнейшем рассматриваем только те векторы $\mathbf{w}_j$, для которых $A\mathbf{w}_j \neq 0$. Не уменьшая общности, будем считать, что число таких векторов равно q.

2) Выбираем $\mathbf{x}^0$ на основе априорной информации; в частности, можно взять $\mathbf{x}^0 = \mathbf{0}$.

3) Если $\mathbf{x}^k$ найден, то j_k и C_k вычисляем по формулам

$$j_k = \arg\max_j \frac{|(A\mathbf{w}_j,\, \xi^k)|}{\|A\mathbf{w}_j\|}, \qquad C_k = -\frac{(A\mathbf{w}_{j_k},\, \xi^k)}{\|A\mathbf{w}_{j_k}\|^2}. \tag{14}$$

4) Следующее приближение $\mathbf{x}^{k+1}$ вычисляем по формуле

$$\mathbf{x}^{k+1} = \mathbf{x}^k + C_k \mathbf{w}_{j_k}. \tag{15}$$

Найдем трудоемкость метода. Для этого оценим число арифметических операций на шаге. Прежде всего заметим, что предварительные операции (этап 1) требуют в общем случае $O(m^2 q)$ операций.

Этап 3 итерационного метода требует $O(mq)$, а этап 4 — $O(m)$ арифметических операций.

Таким образом, общая трудоемкость метода составляет $O(m^2 + mql)$ арифметических операций, где l — число шагов итерационного процесса.

Отметим также, что j_k из (14) находится в общем случае неоднозначно (таких индексов может быть несколько). В этом случае в качестве j_k можно брать, например, наименьший.

Исследуем сходимость итерационного метода. Имеет место

Лемма. *Пусть $\mathbf{g}_1, \ldots, \mathbf{g}_l$ — произвольный набор линейно независимых единичных векторов из R_m и L — линейная оболочка этих векторов. Тогда существует γ, $0 < \gamma < 1$, такое, что для любого $\mathbf{x} \in L$ справедливо неравенство*

$$\|\mathbf{x} - (\mathbf{x},\, \mathbf{g}_k)\mathbf{g}_k\| \leqslant \gamma \|\mathbf{x}\|, \qquad k = \arg\max_j |(\mathbf{x},\, \mathbf{g}_j)|.$$

Доказательство. Положим

$$\psi(\mathbf{x}) = \big\|\mathbf{x} - (\mathbf{x},\, \mathbf{g}_k)\mathbf{g}_k\big\|^2, \qquad k = \arg\max_j \big|(\mathbf{x},\, \mathbf{g}_j)\big|.$$

Покажем, что функционал $\psi(\mathbf{x})$ непрерывен. Для этого достаточно показать непрерывность функционала $(\mathbf{x},\, \mathbf{g}_k)$, где k определено выше. Рассмотрим разность $\big|(\mathbf{x},\, \mathbf{g}_k)\big| - \big|(\mathbf{y},\, \mathbf{g}_i)\big|$, где индекс k определен выше, а i — индекс, определяемый аналогичным образом для $\mathbf{y}$.

Пусть для определенности $|(\mathbf{x}, \mathbf{g}_k)| \geqslant |(\mathbf{y}, \mathbf{g}_i)|$. Тогда справедлива цепочка неравенств

$$\big||(\mathbf{x}, \mathbf{g}_k)| - |(\mathbf{y}, \mathbf{g}_i)|\big| \leqslant \big||(\mathbf{x}, \mathbf{g}_k)| - |(\mathbf{y}, \mathbf{g}_k)|\big| \leqslant$$
$$\leqslant |(\mathbf{x}-\mathbf{y}, \mathbf{g}_k)| \leqslant \max_j |(\mathbf{x}-\mathbf{y}, \mathbf{g}_j)|,$$

откуда и следует непрерывность рассматриваемого функционала.

Предположим, что утверждение леммы неверно. Тогда существует последовательность $\{\mathbf{x}_i\}$ такая, что $\|\mathbf{x}_i\| = 1$ и $\psi(\mathbf{x}_i) \geqslant 1 - \varepsilon_i$, где $\varepsilon_i \to 0$ при $i \to \infty$. Так как в конечномерном пространстве сфера $S = \{\mathbf{x} : \|\mathbf{x}\| = 1\}$ компактна, то существует сходящаяся подпоследовательность. Для простоты изложения предположим, что сходится сама последовательность

$$\mathbf{x}^* = \lim_{i\to\infty} \mathbf{x}_i.$$

В силу непрерывности функционала ψ имеем $\psi(\mathbf{x}^*) = 1$ и $\|\mathbf{x}^*\| = 1$. Следовательно, при $k = \arg\max\limits_j |(\mathbf{x}^*, \mathbf{g}_j)|$ имеем

$$\psi(\mathbf{x}^*) = \|\mathbf{x}^* - (\mathbf{x}^*, \mathbf{g}_k)\mathbf{g}_k\|^2 = \|\mathbf{x}^*\|^2 - 2(\mathbf{x}^*, \mathbf{g}_k)^2 + (\mathbf{x}^*, \mathbf{g}_k)^2\|\mathbf{g}_k\|^2 =$$
$$= \|\mathbf{x}^*\|^2 - (\mathbf{x}^*, \mathbf{g}_k)^2 = 1.$$

Отсюда следует, что $(\mathbf{x}^*, \mathbf{g}_k) = 0$. Так как $|(\mathbf{x}^*, \mathbf{g}_k)| \geqslant |(\mathbf{x}^*, \mathbf{g}_j)|$ для любого $j = 1, \ldots, q$, то $(\mathbf{x}^*, \mathbf{g}_j) = 0$ при всех $j = 1, \ldots, q$. Так как $\mathbf{x}^*$ принадлежит линейной оболочке векторов $\mathbf{g}_1, \ldots, \mathbf{g}_l$, то последнее равенство может выполняться лишь при $\mathbf{x}^* = \mathbf{0}$, что противоречит условию $\|\mathbf{x}^*\| = 1$. Лемма доказана.

Теорема. *Последовательность приближений $\mathbf{x}^k$, получаемая в ходе итерационного метода (14), (15), является фундаментальной и сходится к некоторому вектору, минимизирующему функционал невязки $\psi(\mathbf{x})$ на подпространстве W, со скоростью геометрической прогрессии. А именно, существует $q < 1$ такое, что*

$$\|\mathbf{x}^k - \mathbf{x}^\infty\| \leqslant Cq^k, \quad \mathbf{x}^\infty = \lim_{k\to\infty} \mathbf{x}^k.$$

Постоянная q при этом зависит от выбора базиса $\{\mathbf{w}_j\}$ и оператора A.

Доказательство. Так как $A\mathbf{w}_{j_k} \neq 0$, то существует постоянная $\delta > 0$ такая, что

$$\|A\mathbf{w}_{j_k}\| \geqslant \delta \quad \forall k. \tag{16}$$

Из (14), (15) следует, что невязки $\xi^k = A\mathbf{x}^k - \mathbf{b}$ удовлетворяют соотношению

$$\xi^{k+1} = \xi^k - \frac{(A\mathbf{w}_{j_k}, \xi^k)}{\|A\mathbf{w}_{j_k}\|^2} A\mathbf{w}_{j_k}. \tag{17}$$

Положим $\mathbf{g}_i = A\mathbf{w}_i/\|A\mathbf{w}_i\|$. Тогда (17) примет вид

$$\xi^{k+1} = \xi^k - \left(\xi^k, \mathbf{g}_{j_k}\right)\mathbf{g}_{j_k}.$$

Поскольку C_k выбиралось из условия минимума $\|\xi^{k+1}\|$, то

$$j_k = \arg\max_j \left|(\xi^k, \mathbf{g}_j)\right|$$

и мы находимся в условиях предыдущей леммы. Тогда из условий леммы следует оценка

$$\|\xi^{k+1}\| \leqslant \gamma\|\xi^k\|, \quad \gamma < 1. \tag{18}$$

Из (14) и (15) имеем

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{(\xi^k, A\mathbf{w}_{j_k})}{\|A\mathbf{w}_{j_k}\|^2}\mathbf{w}_{j_k},$$

откуда, учитывая (16), получаем оценку

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \leqslant \|\xi^{(k)}\|/\|A\mathbf{w}_{j_k}\| \leqslant \|\xi^k\|/\delta.$$

Применяя к полученному неравенству оценку (18), имеем

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \leqslant \gamma^k\|\mathbf{b}\|/\delta,$$

откуда следует цепочка соотношений

$$\|\mathbf{x}^{k+p} - \mathbf{x}^k\| \leqslant \sum_{i=k}^{p+k}\|\mathbf{x}^{i+1} - \mathbf{x}^i\| \leqslant \sum_{i=0}^{p}\gamma^{k+i}\frac{\|\mathbf{b}\|}{\delta} \leqslant \frac{\gamma^k\|\mathbf{r}^0\|}{(1-\gamma)\delta} \quad \forall p \in \mathbf{N}.$$

Таким образом, последовательность $\{\mathbf{x}^k\}$ является фундаментальной и имеет предел $\mathbf{x}^\infty$. В силу построения, $\mathbf{x}^\infty$ минимизирует функционал $\Phi(\mathbf{y})$ на подпространстве W. Теорема доказана.

Описанный выше метод решения систем уравнений с плохо обусловленными матрицами особенно эффективен в случае, когда априорная информация представлена в виде каких-либо сведений о структурных особенностях искомого решения; например, когда известны базисные функции $\mathbf{w}_j$, и решение представимо в виде разложения по небольшому числу этих функций. Такая ситуация часто имеет место в задачах цифровой обработки сигналов.

Особенно эффективен данный метод в случае, когда q достаточно мало. Другими словами, для эффективного применения данного метода надо иметь разумную параметризацию исходной задачи. Довольно часто это можно сделать на основе априорной информации о решении. Например, если известно, что решение представляет собой некоторый колебательный процесс с небольшим числом гармоник, номера которых, вообще говоря, неизвестны.

Изложенный выше метод может быть обоснован также и в случае, когда спуск осуществляется не по одномерным подпространствам, соответствующим координатным осям в базисе $\{\mathbf{w}_j\}$, а по гиперплоскостям. При этом, естественно, скорость сходимости обычно бывает выше, однако число арифметических операций на шаге процесса возрастает.

§ 12. Проблема собственных значений

В различных случаях возникают разные требования к информации о собственных значениях и собственных векторах матриц, и это порождает многообразие проблем и приемов решения этой задачи.

1. Для решения ряда задач механики, физики, химии требуется получение всех собственных значений, а иногда и всех собственных векторов некоторых матриц. Эту задачу называют *полной проблемой собственных значений.*

2. В ряде случаев требуется найти лишь максимальное или минимальное по модулю собственное значение матрицы. Проблемы подобного сорта возникают, в частности, при решении некоторых задач ядерной физики. Здесь приходится решать задачи, эквивалентные задачам отыскания собственных значений матриц размерности порядка 10^3–10^6 или даже существенно большей. При малых размерностях матриц для решения этих задач чаще применяют итерационные методы, при больших — вероятностные.

3. При исследовании колебательных процессов иногда требуется найти два максимальных по модулю собственных значения матрицы, причем меньшее из них обычно достаточно определить с меньшей точностью.

4. Там же возникает задача отыскания собственного значения матрицы, наиболее близкого к заданному числу λ^0, или отыскания расстояния от заданного числа λ^0 до спектра матрицы.

Формально рассуждая, можно было бы сказать, что эти задачи, называемые *частичными проблемами собственных значений*, являются частным случаем общей проблемы собственных значений и достаточно ограничиться рассмотрением методов решения общей проблемы. Однако такой подход приведет к неоправданно большому объему вычислений. При обсуждении постановок конкретных задач, связанных с отысканием собственных значений матриц, зачастую значительные усилия тратятся именно на установление минимального объема информации о спектре матрицы, которым можно ограничиться.

Решение задач 2–4 обычно сводят к отысканию максимального по модулю собственного значения некоторой матрицы $B = g(A)$ такой, что это

собственное значение соответствует отыскиваемому собственному значению матрицы A.

Рассмотрим случай, когда все собственные значения матрицы A вещественны. Если требуется найти максимальное или минимальное значение матрицы A, то следует взять $g(A) = A + cE$. Очевидно, что при достаточно больших положительных (отрицательных) значениях c максимальному по модулю собственному значению матрицы $A + cE$ соответствует максимальное (минимальное) собственное значение матрицы A. В случае задачи 4 при некотором c максимальное по модулю собственное значение матрицы $E - c(A - \lambda^0 E)^2$ соответствует отыскиваемому собственному значению матрицы A. Иногда в качестве такой матрицы $g(A)$ может использоваться матрица $(A - \lambda^0 E)^{-1}$. При этом матрица $(A - \lambda^0 E)^{-1}$ не выписывается в явном виде, а необходимые по ходу вычислений векторы $(A - \lambda^0 E)^{-1}\mathbf{y}$ находят, решая систему уравнений $(A - \lambda^0 E)\mathbf{x} = \mathbf{y}$.

Рассмотрим типичную задачу отыскания двух максимальных по модулю собственных значений матрицы A. Для простоты предполагаем наличие полной системы собственных векторов $\mathbf{e}_j$:

$$A\mathbf{e}_j = \lambda_j \mathbf{e}_j, \qquad |\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \cdots \geqslant |\lambda_m|.$$

Зададимся некоторым вектором $\mathbf{x}^0$ и будем последовательно вычислять векторы $\mathbf{x}^{n+1} = A\mathbf{x}^n$. Представим $\mathbf{x}^0$ в виде $\mathbf{x}^0 = \sum_{i=1}^{m} c_i \mathbf{e}_i$; имеем

$$\mathbf{x}^n = \sum_{i=1}^{m} c_i A^n \mathbf{e}_i = \sum_{i=1}^{m} c_i \lambda_i^n \mathbf{e}_i.$$

Отсюда следуют соотношения

$$\mathbf{x}^n = c_1\lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n),$$

$$(\mathbf{x}^n, \mathbf{x}^n) = (c_1\lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n), c_1\lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n)) = |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n),$$

$$(\mathbf{x}^{n+1}, \mathbf{x}^n) = (c_1\lambda_1^{n+1} \mathbf{e}_1 + O(|\lambda_2|^{n+1}), c_1\lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n)) =$$

$$= \lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n).$$

Положим

$$\lambda_1^{(n)} = (\mathbf{x}^{n+1}, \mathbf{x}^n)/(\mathbf{x}^n, \mathbf{x}^n).$$

Из последних соотношений при $c_1 \neq 0$ получаем

$$\lambda_1^{(n)} = \frac{\lambda_1 c_1^2 \lambda_1^{2n} + O(|\lambda_1|^n |\lambda_2|^n)}{c_1^2 \lambda_1^{2n} + O(|\lambda_1|^n |\lambda_2|^n)} = \frac{\lambda_1 \left(1 + O\left(\frac{1}{c_1^2}\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)\right)}{1 + O\left(\frac{1}{c_1^2}\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right). \tag{1}$$

Задача 1. Доказать, что в случае симметричной матрицы $\lambda_1^{(n)} = \lambda_1 + O(|\lambda_2/\lambda_1|^{2n})$.

Кроме (1) имеем

$$\|\mathbf{x}^n\| = |c_1\lambda_1^n| + O\Big(|\lambda_2|^n\Big),$$

$$\mathbf{e}_1^{(n)} = \frac{\mathbf{x}^n}{\|\mathbf{x}^n\|} = \frac{\sum_{i=1}^{m} c_i\lambda_i^n\mathbf{e}_i}{|c_1|\,|\lambda_1|^n + O(|\lambda_2|^n)} =$$

$$= \frac{\frac{c_1\lambda_1^n}{|c_1\lambda_1^n|}\mathbf{e}_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}{1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = e^{i\varphi_n}\mathbf{e}_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right);$$

здесь $\varphi_n = \arg\{c_1\lambda_1^n\}$. Таким образом, в ходе этого итерационного процесса также получаем собственный вектор, соответствующий λ_1.

Может случиться, что у матрицы A имеются два максимальных по модулю собственных значения $\lambda_1 \neq \lambda_2$, $|\lambda_1| = |\lambda_2| > |\lambda_3| \geqslant \ldots$ В этом случае величина $\lambda_1^{(n)}$ будет устанавливаться только в частном случае, когда c_1 или c_2 равно нулю. Если заранее известно, что таких собственных значений два, то их и соответствующие им собственные векторы можно также определить, анализируя поведение $\lambda_1^{(n)}$ и $\mathbf{e}_1^{(n)}$.

Рассмотрим типичный случай, когда A, λ_1, λ_2, $\mathbf{x}^0$ вещественные, $\lambda_1 > 0$, $\lambda_1 = -\lambda_2$. Тогда

$$\mathbf{x}^n = c_1\lambda_1^n\mathbf{e}_1 + c_2(-\lambda_1)^n\mathbf{e}_2 + O(|\lambda_3|^n),$$

$$\mathbf{x}^{n+2} = c_1\lambda_1^{n+2}\mathbf{e}_1 + c_2(-\lambda_1)^{n+2}\mathbf{e}_2 + O(|\lambda_3|^n) =$$

$$= \lambda_1^2(c_1\lambda_1^n\mathbf{e}_1 + c_2(-\lambda_1)^n\mathbf{e}_2) + O(|\lambda_3|^n).$$

Отсюда получаем, что

$$\bar{\lambda}^{(n)} = (\mathbf{x}^{n+2}, \mathbf{x}^n)/(\mathbf{x}^n, \mathbf{x}^n) = \lambda_1^2 + O(|\lambda_3/\lambda_1|^n).$$

При $\bar{\lambda}^{(n)} > 0$ полагаем $\lambda_{1,2}^{(n)} = \pm\sqrt{\bar{\lambda}^{(n)}}$. Имеем

$$\mathbf{z}_1^{n+1} = \mathbf{x}^{n+1} + \lambda_1^{(n)}\mathbf{x}^{(n)} = 2c_1\lambda_1^{n+1}\mathbf{e}_1 + O(|\lambda_3|^n),$$

поэтому

$$\mathbf{e}_1^{(n)} = \mathbf{z}_1^{n+1}/\|\mathbf{z}_1^{n+1}\| = \mathbf{e}_1 + O(|\lambda_3/\lambda_1|^n).$$

Точно так же

$$\mathbf{z}_2^{n+1} = \mathbf{x}^{n+1} + \lambda_2^{(n)}\mathbf{x}^{(n)} = 2c_2\lambda_2^{n+1}\mathbf{e}_2 + O(|\lambda_3|^n)$$

и

$$\mathbf{e}_2^{(n)} = \mathbf{z}_2^{n+1}/\|\mathbf{z}_2^{n+1}\| = \mathbf{e}_2 + O(|\lambda_3/\lambda_1|^n).$$

Если λ_i — собственные значения матрицы A, то у сопряженной матрицы A^* собственным значением будет $\bar{\lambda}_i$; при этом если $A\mathbf{e}_i = \lambda_i\mathbf{e}_i$,

$A^*\mathbf{g}_j = \bar{\lambda}_j \mathbf{g}_j$, $\lambda_i \neq \lambda_j$, то $(\mathbf{e}_i, \mathbf{g}_j) = 0$. Поэтому при $|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots$ для нахождения собственного значения λ_2 можно поступать так. Получив приближение $\tilde{\mathbf{e}}_1 \approx \mathbf{e}_1$, аналогично определяем приближение $\tilde{\mathbf{g}}_1$ к вектору $\mathbf{g}_1$ и нормируем его условием $(\tilde{\mathbf{e}}_1, \tilde{\mathbf{g}}_1) = 1$. Далее ведем итерации по формулам $\mathbf{y}^{n+1} = A\mathbf{y}^n$; для исключения компоненты, пропорциональной $\mathbf{e}_1$, время от времени векторы $\mathbf{y}^n$ ортогонализируем по отношению к вектору $\tilde{\mathbf{g}}_1$, т. е. начальным для последующих итераций вместо $\mathbf{y}^n$ берем вектор

$$\tilde{\mathbf{y}}_n = \mathbf{y}_n - (\mathbf{y}_n, \tilde{\mathbf{g}}_1)\tilde{\mathbf{e}}_1.$$

Естественно, что сходимость итерационного процесса лучше, если в начальном приближении $\mathbf{y}^0 = \sum_{i=1}^{m} d_i \mathbf{e}_i$ слагаемое $d_2\mathbf{e}_2$ преобладает над остальными слагаемыми. Возьмем из описанного выше итерационного процесса отыскания λ_1 и $\mathbf{e}_1$ некоторое приближение $\mathbf{x}^l$. Вектор $\mathbf{y}^0 = \mathbf{x}^l - (\mathbf{x}^l, \tilde{\mathbf{g}}_1)\tilde{\mathbf{e}}_1$ примем за начальное приближение. Не следует брать l слишком малым, иначе компоненты $c_j\lambda_j^l\mathbf{e}_j$ при $j > 2$ не будут малы по сравнению с $c_2\lambda_2^l\mathbf{e}_2$; в то же время не следует брать l очень большим, поскольку в этом случае компонента $c_2\lambda_2^l\mathbf{e}_2$ будет малой по сравнению с вычислительной погрешностью.

Часто можно встретить следующее высказывание: если $c_1 = 0$, то описанный итерационный процесс, казалось бы, не должен давать приближения, сходящиеся к максимальному по модулю собственному значению; однако в действительности из-за присутствия округлений в процессе итераций может появиться компонента, пропорциональная $\mathbf{e}_1$, и вследствие этого требуемый результат все равно будет получен.

Реально при использовании современных ЭВМ с большой разрядностью может случиться, что после некоторого числа итераций влияние вычислительной погрешности еще не будет существенно, в то время как величина $\sum_{i=3}^{m} c_i\lambda_i^n\mathbf{e}_i$ будет мала по сравнению с $c_2\lambda_2^n\mathbf{e}_2$. Тогда $A\mathbf{x}^n \approx \text{const}\,\mathbf{x}^n$ и можно сделать неверный вывод, что искомое первое собственное значение найдено. Поэтому в тех случаях, когда нет уверенности в правильности найденного собственного значения, следует провести еще один или несколько расчетов с другими значениями $\mathbf{y}^0$.

В ряде случаев появление составляющей, пропорциональной $\mathbf{e}_1$, не является неизбежным даже в случае присутствия округлений. При решении проблемы собственных значений для дифференциальных и интегральных операторов иногда возникают матрицы A со специфическими свойствами. Например, часто встречается случай, когда при всех i, j выполняется равенство

$$a_{ij} = a_{m+1-i,\,m+1-j}. \tag{2}$$

Для определенности рассмотрим случай четного m. Назовем четными векторы $\mathbf{x}$, компоненты которых связаны равенством $x_i = x_{m+1-i}$, $i = 1, \ldots, m$, и нечетными — равенством $x_i = -x_{m+1-i}$. При условии (2) вектор $A\mathbf{x}$ четен, если $\mathbf{x}$ четен, и нечетен, если $\mathbf{x}$ нечетен. Поэтому подпространства четных и нечетных

векторов являются собственными для оператора A. Следовательно, существует полная система собственных векторов, принадлежащих этим подпространствам, т. е. являющихся или четными, или нечетными. Это обстоятельство может быть существенно использовано: если вектор $\mathbf{x}^0$ четен или нечетен, то этим же свойством обладают все векторы $\mathbf{x}^n$, поэтому при отыскании каждого последующего вектора $\mathbf{x}^n$ следует ограничиться определением первой половины его компонент. Кроме этого, можно объединить коэффициенты, соответствующие компонентам x_i и x_{m+1-i}. Например, при четном $\mathbf{x}^0$ вычисления компонент $\mathbf{x}^n$ можно производить по формулам

$$x_i^{n+1} = \sum_{j=1}^{m/2} (a_{ij} + a_{i,m+1-j}) x_i^n.$$

В случае итераций такого вида мы не выходим за пределы подпространств четных или нечетных векторов соответственно. Поэтому если $\mathbf{x}^0$ и $\mathbf{e}_1$ принадлежат различным подпространствам, составляющая, пропорциональная $\mathbf{e}_1$, так и не появится.

Кроме непосредственного уменьшения вычислений при отыскании каждого вектора $\mathbf{x}^n$, использование свойства (2) полезно также по следующей причине. Довольно типичен случай, когда собственные векторы с небольшими нечетными номерами являются четными, а с небольшими четными номерами — нечетными; предположим, что $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots$ Тогда при $\mathbf{x}^0$ четном всегда

$$\mathbf{x}^n = c_1 \lambda_1^n \mathbf{e}_1 + c_3 \lambda_3^n \mathbf{e}_3 + \cdots = c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_3|^n),$$

следовательно,

$$\lambda_1^{(n)} - \lambda_1 = O(|\lambda_3/\lambda_1|^n) \quad \text{при} \quad \lambda_1^{(n)} = (\mathbf{x}^{n+1}, \mathbf{x}^n)/(\mathbf{x}^n, \mathbf{x}^n).$$

Соответственно при $\mathbf{x}^0$ нечетном имеем

$$\mathbf{x}^n = c_2 \lambda_2^n \mathbf{e}_2 + O(|\lambda_4|^n)$$

и

$$\lambda_2^{(n)} - \lambda_2 = O(|\lambda_4/\lambda_2|^n) \quad \text{при} \quad \lambda_2^{(n)} = (\mathbf{x}^{n+1}, \mathbf{x}^n)/(\mathbf{x}^n, \mathbf{x}^n).$$

При этом не возникает никакой проблемы подавления составляющей, пропорциональной $\mathbf{e}_1$.

Если $|\lambda_1| > 1$, то $\|\mathbf{x}^n\| \to \infty$ при $n \to \infty$, поэтому при достаточно большом n произойдет переполнение разрядности чисел и остановка ЭВМ. Если $|\lambda_1| < 1$, то $\|\mathbf{x}^n\| \to 0$, и вследствие конечности порядков чисел в машине может случиться, что, начиная с некоторого n, $\mathbf{x}^n \equiv \mathbf{0}$. Чтобы избежать этих явлений, полезно время от времени нормировать вектор $\mathbf{x}^n$, чтобы $\|\mathbf{x}^n\| = 1$.

Для практической оценки погрешности и ускорения сходимости итерационных процессов может быть применен δ^2-процесс и другие приемы, аналогичные методам ускорения сходимости при решении систем линейных уравнений. Например, могут применяться итерации вида $\mathbf{x}^{n+1} =$

$g_k(A)\mathbf{x}^n$ со специальным подбором в зависимости от известной информации о спектре матрицы A многочлена $g_k(A)$.

Поскольку

$$\max \lambda_A^i = \sup_{\mathbf{x}} \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}, \quad \min \lambda_A^i = \inf_{\mathbf{x}} \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$$

при $A = A^T$, то некоторые приемы отыскания максимального и минимального собственных значений матрицы A основываются на идее отыскания стационарных точек функционала $\Phi(\mathbf{x}) = (A\mathbf{x}, \mathbf{x})/(\mathbf{x}, \mathbf{x})$.

Задача 2. Пусть $\lambda_1 \approx 5$, $1 \leqslant \lambda_i \leqslant 3$ при $i = 2, \ldots, m$. Построить итерационный процесс вида $\mathbf{x}^{n+1} = (A + cE)\mathbf{x}^n$ для получения λ_1 с наилучшей при данной информации скоростью сходимости.

Сделать то же самое, если $\lambda_1 \approx 1$, $2 \leqslant \lambda_i \leqslant 3$ при $i = 2, \ldots, m$.

§ 13. Решение полной проблемы собственных значений при помощи QR-алгоритма

Существует ряд тщательно отработанных алгоритмов и программ решения полной проблемы собственных значений. Поэтому в случае, когда возникает такая проблема, в первую очередь рекомендуется использовать стандартные программы решения таких задач. Наиболее совершенные из них основаны на различных модификациях QR-алгоритма, общая схема которого приводится ниже.

Пусть A — произвольная вещественная матрица. Согласно лемме из § 2 ее можно представить в виде $A = U^T A_{n-1}$, где U — ортогональная, а A_{n-1} — правая треугольная матрицы. Запишем это равенство в виде

$$A = Q_1 R_1, \tag{1}$$

где Q_1 — ортогональная, R_1 — правая треугольная матрицы. Из (1) имеем $R_1 = Q_1^{-1} A$, поэтому матрица $A_1 = R_1 Q_1 = Q_1^{-1} A Q_1$ подобна матрице A.

Построим последовательность матриц A_n по следующему правилу. Матрицу A_n разлагаем на произведение ортогональной и правой треугольной матриц в виде $A_n = Q_{n+1} R_{n+1}$ и полагаем $A_{n+1} = R_{n+1} Q_{n+1}$. Поскольку $A_{n+1} = Q_{n+1}^{-1} A_n Q_{n+1}$, то все матрицы A_n подобны между собой и подобны исходной матрице A.

Представим исходную матрицу A в виде $A = Q\Lambda Q^{-1}$, где Λ — правая каноническая форма Жордана, т. е. $\lambda_{ij} = 0$ при $j < i$ и $j > i+1$, $\lambda_{ii} = \lambda_i$ — собственные значения матрицы A, $\lambda_{i,i+1}$ равны нулю или 1. Всегда можно подобрать матрицу Q так, чтобы диагональные элементы матрицы Λ были упорядочены в порядке невозрастания модуля

$$|\lambda_1| = \cdots = |\lambda_{l_1}| > |\lambda_{l_1+1}| = \cdots = |\lambda_{l_2}| > \cdots > |\lambda_{l_{s-1}+1}| = \cdots = |\lambda_{l_s}|.$$

Теорема (без доказательства). *Пусть в разложении матрицы A все диагональные миноры матрицы Q не вырождены. Тогда последовательность*

матриц A_n при $n \to \infty$ сходится по форме к клеточному правому треугольному виду $\widehat{A}$.

Имеется в виду, что после некоторой перестановки строк и одновременно такой же перестановки столбцов матрицы A будут выполняться соотношения: если $l_k < i \leqslant l_{k+1}$, $j < i$ или $l_{k+1} < j$, $k = 1, \ldots, s$, то $a_{ij}^{(n)} \to 0$.

При реализации описанного алгоритма построения матриц A_n на практике мы увидим, что некоторые элементы матриц A_n оказываются малыми. Приравняв их нулю, мы получим клеточную правую треугольную матрицу. Характеристический многочлен этой матрицы равен произведению характеристических многочленов ее диагональных клеток.

Если требуется найти не только собственные значения матрицы A, но и ее собственные и присоединенные векторы, то в процессе построения последовательности матриц A_n следует запоминать ортогональные матрицы $P_n = Q_1 \ldots Q_n$, вычисляемые по рекуррентной формуле $P_{n+1} = P_n Q_{n+1}$.

Задача 1. Доказать, что каждый шаг QR-алгоритма требует $N \sim 10m^3/3$ арифметических операций.

На практике прибегают к различным способам ускорения сходимости. Один из этих способов заключается в следующем. Матрица A предварительно преобразуется в эквивалентную ей правую почти треугольную матрицу.

Матрица A называется *правой почти треугольной*, если $a_{ij} = 0$ при $j < i - 1$.

Алгоритм преобразования матрицы A в правую почти треугольную матрицу заключается в последовательном построении матриц A_l таких, что первые l столбцов матрицы A_l имеют вид первых l столбцов правой почти треугольной матрицы, т. е. $a_{ij} = 0$, если $j < i-1$ и $j \leqslant l$. По элементам $(l+1)$-го столбца матрицы A_l построим матрицу отражения U_{l+1} (см. § 2) так, чтобы в матрице $B = U_{l+1}A_l$ элементы $b_{1,l+1}, \ldots, b_{l,l+1}$ были те же, что у матрицы A_l, а элементы $b_{l+3,l+1}, \ldots, b_{m,l+1}$ были нулевыми. Положим $A_{l+1} = U_{l+1}A_lU_{l+1}$. Умножение справа на матрицу U_{l+1}^T не меняет первых $l+1$ столбцов матрицы B, поэтому матрица A_{l+1} является матрицей требуемого вида. После получения правой почти треугольной матрицы A_{m-1} применяют QR-алгоритм в его первоначальной форме.

Задача 2. Доказать, что в этом случае каждый шаг QR-алгоритма требует $N \sim 6m^2$ арифметических операций.

Для еще большего ускорения сходимости применяется вариант QR-алгоритма со сдвигом. А именно, строится последовательность ортого-

нальных матриц Q_l и правых треугольных матриц R_l по рекуррентным формулам

$$A - \nu_1 E = Q_1 R_1, \quad A_1 = R_1 Q_1 + \nu_1 E,$$
$$\dots$$
$$A_{l-1} - \nu_l E = Q_l R_l, \quad A_l = R_l Q_l + \nu_l E.$$

Матрицы A_l подобны матрице A; за счет введения «сдвигов» ν_l удается добиться ускорения сходимости. Вопрос о наиболее целесообразном выборе параметров ν_l мы рассматривать не будем.

В этой главе было, в частности, приведено большое количество методов решения линейных систем уравнений. Какой же из этих методов все-таки стоит выбрать, решая задачу?

Если порядок системы небольшой и по затратам машинного времени число арифметических операций порядка m^3, где m — порядок системы, является приемлемым, то проще всего обратиться к стандартным программам метода отражений (число арифметических действий $N \approx 2m^3/3$) или метода вращений (число арифметических действий $N \approx 4m^3/3$, но меньше накопление вычислительной погрешности).

Конечно, при этом надо иметь в виду, что точные методы типа методов отражений или вращений для матрицы общего вида требуют одновременного хранения в памяти ЭВМ порядка m^2 чисел. Если такое количество чисел не помещается в оперативной памяти ЭВМ, а обмен информацией между оперативной и внешней памятью происходит недостаточно быстро или из-за структуры программы, или из-за возможностей ЭВМ, то применение этих программ может оказаться нецелесообразным.

В случае, когда применение этих методов нецелесообразно, имеет смысл проанализировать возможности применения простейших по своей структуре итерационных методов: простой итерации, Зейделя, сверхрелаксации, наискорейшего спуска. Если решается отдельная задача, то вследствие простоты соответствующих программ применение этих методов может быть вполне целесообразным. Если применение этих методов требует больших затрат машинного времени, то следует проанализировать возможности применения более сложных по своей структуре методов: оптимального линейного итерационного процесса, метода с использованием корней многочлена Чебышева, метода сопряженных градиентов, итерационных методов, использующих спектрально эквивалентные операторы.

Если размерность задачи столь велика, что само решение задачи, т. е. вектор $\mathbf{X}$, не помещается в оперативной памяти ЭВМ, то иногда применяются вероятностные методы решения систем линейных уравнений, которые остались вне нашего рассмотрения.

Литература

1. Абрамов А.А. О численном решении некоторых алгебраических задач, возникающих в теории устойчивости. // ЖВМиМФ — 1984. **24**, N 3. С. 339–347.
2. Бахвалов Н.С. Численные методы. — М.: Наука, 1975.
3. Воеводин В.В. Численные методы алгебры. Теория и алгоритмы. — М.: Наука, 1966.
4. Воеводин В.В. Вычислительные основы линейной алгебры. — М.: Наука, 1977.
5. Воеводин В.В., Кузнецов Ю.А. Матрицы и вычисления. — М.: Наука, 1984.
6. Годунов С.К. Решение систем линейных уравнений. — Новосибирск: Наука, 1980.
7. Годунов С.К. Современные аспекты линейной алгебры. — Новосибирск: Научная книга, 1997.
8. Джордж А., Лю Д. Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.
9. Дьяконов Е.Г. О построении итерационных методов на основе использования операторов, эквивалентных по спектру // ЖВМ и МФ. — 1966. — **6**, N 1. — С. 12–34.
10. Икрамов Х.Д. Численное решение матричных уравнений. — М.: Наука, 1984.
11. Канторович Л.В. Функциональный анализ и прикладная математика. // УМН — 1948. **3** N 6 (28). С. 89–185.
12. Крылов В.И., Бобков В.В., Монастырный П.И. Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника, 1982.
13. Марчук Г.И., Лебедев В.И. Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1981.
14. Ортега Д. Введение в параллельные и векторные методы решения линейных систем. — М.: Мир, 1991.
15. Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983.
16. Поспелов В.В. Метод оптимального спуска по базису для решения вырожденных систем линейных алгебраических уравнений. // ЖВМиМФ — 1991. **31**, N 7. С. 961–969.
17. Фаддеев Л.К., Фаддеева В.Н. Вычислительные методы линейной алгебры. — М.: Физматгиз, 1963.
18. Форсайт Дж. и др. Машинные методы математических вычислений. — М.: Мир, 1980.

Глава 7

Решение систем нелинейных уравнений и задач оптимизации

Решение задач оптимизации, будь то оптимизация производственных или экономических процессов, оптимизация конструкций или оптимизация численных алгоритмов, сводится в математической формулировке исследуемой задачи к отысканию экстремума функционалов. В наиболее типичных случаях возникает задача минимизации функции большого числа переменных в области Ω, задаваемой большим числом ограничений типа неравенств или равенств: ищется

$$\inf_{x_1,\ldots,x_m} \Phi(x_1, \ldots, x_m)$$

при условиях

$$\varphi_i(x_1, \ldots, x_m) \geqslant 0, \quad i = 1, \ldots, l;$$
$$\psi_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, q.$$

Задача минимизации функций большого числа переменных возникает также в случае применения вариационных методов к решению задач математической физики и в других разделах прикладной математики.

Системы уравнений

$$f_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, m,$$

которые мы будем также обозначать

$$\mathbf{F}(\mathbf{x}) = 0,$$

возникают в случае многих задач указанных выше типов; например, в гл. 10 будет идти речь о подобных системах, возникающих при решении краевых задач.

Задачи минимизации функции и решения системы уравнений сводятся друг к другу. Если $\Psi(y_1, \ldots, y_m) > 0$ при

$$(y_1, \ldots, y_m) \neq (0, \ldots, 0), \quad \Psi(0, \ldots, 0) = 0,$$

то решение системы $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ равносильно минимизации функции

$$\Psi\Big(f_1(x_1, \ldots, x_m), \ldots, f_m(x_1, \ldots, x_m)\Big).$$

С другой стороны, пусть $\inf\limits_{G} \Phi(x_1, \ldots, x_m)$ достигается в точке X, внутренней по отношению к G, и функция Φ дифференцируема в этой точке. Тогда точка минимума является решением системы уравнений

$$\Phi'_{x_i} = 0, \quad i = 1, \ldots, m.$$

Возможность сведения одной из этих задач к другой не означает, что достаточно ограничиться рассмотрением только одной из них; родство этих задач скорее подчеркивает, что они одинаково трудны. О трудности этих задач свидетельствует то, что не существует универсальных алгоритмов решения, практически пригодных уже не при очень больших m; отсутствие таких алгоритмов вызвано существом дела.

В то же время и при больших m существуют эффективные алгоритмы решения задач, обладающих определенной внутренней структурой (в частности задач, возникающих при решении краевых задач математической физики вариационными методами).

При решении задач каждого типичного в приложениях класса приходится заниматься теоретической и экспериментальной «доводкой» методов применительно к этому классу задач. Сведение задачи минимизации функции к системе нелинейных уравнений или наоборот производится на практике с целью снижения трудоемкости решения.

Например, при решении систем нелинейных уравнений иногда поступают следующим образом. Строится функционал, минимум которого достигается на решении системы. Затем, задавшись начальным приближением к точке минимума, проводят итерации каким-либо из методов спуска (см. § 3) и таким путем получают удовлетворительное приближение к решению системы. Исходя из этого приближения производят уточнения при помощи какого-либо итерационного метода, специфического для задачи решения системы уравнений, например метода Ньютона (см. § 2).

Поясним причины, вызывающие такое комбинированное применение методов. Назовем *областью сходимости метода* множества начальных условий, при которых итерации по данному методу сходятся к решению задачи. Применение методов спуска на первоначальном этапе вызвано тем, что обычно они имеют более широкую область сходимости, чем методы, специфические для задачи решения системы уравнений. В то же время последние методы обычно обладают лучшей скоростью сходимости при наличии достаточно хорошего начального приближения; это и обуславливает их применение на заключительном этапе итераций.

На примере решения системы линейных уравнений было также видно, что сведение этой задачи к задаче нахождения минимума функционала приводит к конструированию новых методов решения исходной задачи.

§ 1. Метод простой итерации и смежные вопросы

Так же как в случае линейных уравнений, начнем изучение итерационных методов с *метода простой итерации.*

Этот метод состоит в следующем: система уравнений преобразуется к виду

$$\mathbf{x} = \mathbf{g}(\mathbf{x}), \tag{1}$$

иначе,

$$x_i = g_i(x_1, \ldots, x_m), \quad i = 1, \ldots, m,$$

и итерации проводятся по формуле

$$\mathbf{x}^{n+1} = \mathbf{g}(\mathbf{x}^n), \tag{2}$$

иначе,

$$x_i^{n+1} = g_i(x_1^n, \ldots, x_m^n), \quad i = 1, \ldots, m.$$

Подойдем к изучению этого метода с более общих позиций. Пусть H — полное метрическое пространство, а оператор $\mathbf{y} = \mathbf{g}(\mathbf{x})$ отображает H в себя. Рассмотрим итерационный процесс

$$\mathbf{x}^{n+1} = \mathbf{g}(\mathbf{x}^n) \tag{3}$$

решения уравнения

$$\mathbf{x} = \mathbf{g}(\mathbf{x}). \tag{4}$$

Если при некотором $q < 1$ отображение $\mathbf{y} = \mathbf{g}(\mathbf{x})$ удовлетворяет условию

$$\rho(\mathbf{g}(\mathbf{x}_1), \mathbf{g}(\mathbf{x}_2)) \leqslant q\rho(\mathbf{x}_1, \mathbf{x}_2) \tag{5}$$

при всех $\mathbf{x}_1, \mathbf{x}_2$, то такое отображение называют *сжимающим.*

Теорема. *Если отображение* $\mathbf{y} = \mathbf{g}(\mathbf{x})$ *сжимающее, то уравнение* $\mathbf{x} = \mathbf{g}(\mathbf{x})$ *имеет единственное решение* $\mathbf{X}$ *и*

$$\rho(\mathbf{X}, \mathbf{x}^n) \leqslant \frac{q^n a}{1-q};$$

здесь $a = \rho(\mathbf{x}^1, \mathbf{x}^0)$, $\rho(\mathbf{x}, \mathbf{y})$ *— расстояние между* $\mathbf{x}$ *и* $\mathbf{y}$.

Доказательство. Согласно (5) имеем

$$\rho(\mathbf{x}^{n+1}, \mathbf{x}^n) = \rho(\mathbf{g}(\mathbf{x}^n), \mathbf{g}(\mathbf{x}^{n-1})) \leqslant q\rho(\mathbf{x}^n, \mathbf{x}^{n-1}),$$

поэтому $\rho(\mathbf{x}^{n+1}, \mathbf{x}^n) \leqslant q^n\rho(\mathbf{x}^1, \mathbf{x}^0) = q^n a$. При $l > n$ имеем цепочку неравенств

$$\begin{aligned}\rho(\mathbf{x}^l, \mathbf{x}^n) &\leqslant \rho(\mathbf{x}^l, \mathbf{x}^{l-1}) + \cdots + \rho(\mathbf{x}^{n+1}, \mathbf{x}^n) \leqslant \\ &\leqslant q^{l-1}a + \cdots + q^n a \leqslant q^n a \sum_{i=0}^{\infty} q^i = \frac{q^n a}{1-q}.\end{aligned} \tag{6}$$

Согласно критерию Коши последовательность $\mathbf{x}^n$ имеет некоторый предел $\mathbf{X}$. Переходя к пределу в (6) при $l \to \infty$, получаем

$$\rho(\mathbf{X}, \mathbf{x}^n) \leqslant \frac{q^n a}{1-q}.$$

Справедлива цепочка соотношений

$$\rho(\mathbf{X}, \mathbf{g}(\mathbf{X})) \leqslant \rho(\mathbf{X}, \mathbf{x}^{n+1}) + \rho\Big(\mathbf{x}^{n+1}, \mathbf{g}(\mathbf{X})\Big) = \rho(\mathbf{X}, \mathbf{x}^{n+1}) + \rho\Big(\mathbf{g}(\mathbf{x}^n), \mathbf{g}(\mathbf{X})\Big)$$
$$\leqslant \rho(\mathbf{X}, \mathbf{x}^{n+1}) + q\rho(\mathbf{x}^n, \mathbf{X}) \leqslant 2\frac{q^{n+1}a}{1-q}.$$

Поскольку n произвольное, то $\rho(\mathbf{X}, \mathbf{g}(\mathbf{X})) = 0$, и, следовательно, $\mathbf{X} = \mathbf{g}(\mathbf{X})$. Предположим, что уравнение (4) имеет два решения $\mathbf{X}_1$ и $\mathbf{X}_2$. Тогда $\rho(\mathbf{X}_1, \mathbf{X}_2) = \rho\big(g(\mathbf{X}_1), g(\mathbf{X}_2)\big) \leqslant q\rho(\mathbf{X}_1, \mathbf{X}_2) < \rho(\mathbf{X}_1, \mathbf{X}_2)$. Мы пришли к противоречию. Теорема доказана.

Замечание. При $n = 0$ из (6) следует, что $\rho(\mathbf{x}^l, \mathbf{x}^0) \leqslant \dfrac{a}{1-q}$. Таким образом, все приближения принадлежат области

$$\Omega(\mathbf{x}^0, h) : \ \rho(\mathbf{x}, \mathbf{x}^0) \leqslant h, \quad h = a/(1-q).$$

При доказательстве теоремы отображение $\mathbf{g}(\mathbf{x})$ применяется лишь к элементам множества $\Omega(\mathbf{x}^0, h)$ и условие сжимаемости применяется лишь относительно пары элементов из $\Omega(\mathbf{x}^0, h)$. Поэтому в формулировке теоремы достаточно предполагать лишь, что отображение $\mathbf{g}(\mathbf{x})$ определено на элементах из $\Omega(\mathbf{x}^0, h)$ и удовлетворяет условию (5) при $\mathbf{x}_1, \mathbf{x}_2 \in \Omega(\mathbf{x}^0, h)$.

Если решается одно скалярное уравнение, то метод простой итерации имеет простую геометрическую интерпретацию. Построим на плоскости (x, y) графики $y = g(x)$ и $y = x$. Точки пересечения этих линий соответствуют искомому решению. Если на чертеже имеется точка $(x^n, x^{n+1}) = (x^n, g(x^n))$, то, проведя через нее прямую $y = x^{n+1}$ до пересечения с прямой $y = x$, а затем прямую $x = x^{n+1}$ до пересечения с кривой $y = g(x)$, мы получим точку (x^{n+1}, x^{n+2}). На рис. 7.1.1 изображено поведение последовательных приближений в случаях: *а)* $0 < g'(x) < 1$, *б)* $-1 < g'(x) < 0$, *в)* $1 < g'(x)$, *г)* $g'(x) < -1$. Монотонное поведение x^n при $g'(x) > 0$ и колебательное при $g'(x) < 0$ нетрудно усмотреть также из соотношения

$$x^{n+1} - X = g(x^n) - g(X) \sim g'(X)(x^n - X).$$

В случае системы нелинейных уравнений $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ аналогом метода Зейделя является итерационный процесс, где компоненты приближений определяются из соотношений

$$\begin{aligned} &f_1(x_1^{n+1}, x_2^n, \ldots, x_m^n) = 0, \\ &f_2(x_1^{n+1}, x_2^{n+1}, \ldots, x_m^n) = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &f_m(x_1^{n+1}, x_2^{n+1}, \ldots, x_m^{n+1}) = 0. \end{aligned} \tag{7}$$

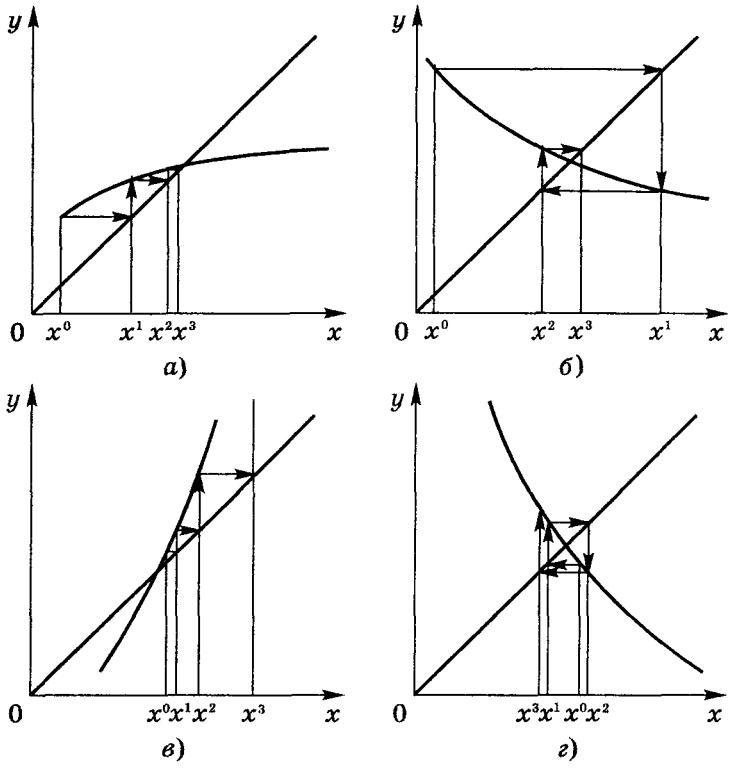


Рис. 7.1.1

Нахождение каждого нового значения x_i^{n+1} требует решения, вообще говоря, нелинейного уравнения

$$f_i(x_1^{n+1}, \ldots, x_{i-1}^{n+1}, x_i^{n+1}, x_{i+1}^n, \ldots, x_m^n) = 0$$

с одним неизвестным.

Промежуточное место между итерационными методами (2) и (7) занимает метод, где компоненты приближений определяются из соотношений

$$\begin{aligned} x_1^{n+1} &= g_1(x_1^n, \ldots, x_m^n), \\ x_2^{n+1} &= g_2(x_1^{n+1}, x_2^n, \ldots, x_m^n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_m^{n+1} &= g_m(x_1^{n+1}, \ldots, x_{m-1}^{n+1}, x_m^n). \end{aligned} \tag{8}$$

Методы (7) и (8) особенно широко использовались в различных моделирующих устройствах, так как они требуют малого объема памяти и просты в реализации.

В достаточно малой окрестности решения $\mathbf{X}$ системы для приближений методом простой итерации имеем

$$\mathbf{x}^{n+1} - \mathbf{X} = \mathbf{g}(\mathbf{x}^n) - \mathbf{g}(\mathbf{X}) \approx B(\mathbf{x}^n - \mathbf{X}), \tag{9}$$

где

$$B = \left[\frac{\partial g_i}{\partial x_j}\right]\Big|_{\mathbf{X}}.$$

Таким образом, при приближениях, находящихся в малой окрестности решения, погрешности приближений итерационного процесса (2) (а также и процессов (7) и (8)) подчиняются примерно тем же законам, что и погрешности итерационных методов решения систем линейных уравнений. Наличие соотношения (9) позволяет производить ускорение сходимости итерационных процессов.

Рассмотрим случай $m = 1$ и построим аналог δ^2-процесса. При имеющемся приближении x^n обозначим $x^{n1} = g(x^n)$, $x^{n2} = g(x^{n1})$. Согласно (9)

$$x^{n1} - X \approx g'(X)(x^n - X),$$
$$x^{n2} - X \approx g'(X)(x^{n1} - X).$$

Из этих соотношений получаем

$$g'(X) \approx \frac{x^{n2} - x^{n1}}{x^{n1} - x^n},$$

$$X \approx \frac{x^{n2} - g'(X)x^{n1}}{1 - g'(X)} \approx \frac{x^{n2} - \dfrac{x^{n2} - x^{n1}}{x^{n1} - x^n}x^{n1}}{1 - \dfrac{x^{n2} - x^{n1}}{x^{n1} - x^n}} = \frac{x^{n2}x^n - (x^{n1})^2}{x^{n2} - 2x^{n1} + x^n}.$$

За следующее после x^n приближение примем

$$x^{n+1} = \frac{x^{n2}x^n - (x^{n1})^2}{x^{n2} - 2x^{n1} + x^n} = \frac{x^n g(g(x^n)) - (g(x^n))^2}{g(g(x^n)) - 2g(x^n) + x^n}. \tag{10}$$

Для характеристики методов решения уравнений вводится понятие *порядка метода*. Говорят, что метод имеет k-й порядок, если существуют $c_1 > 0$, $c_2 < \infty$ такие, что

$$\rho(\mathbf{x}^{n+1}, \mathbf{X}) \leqslant c_2(\rho(\mathbf{x}^n, \mathbf{X}))^k$$

при условии $\rho(\mathbf{x}^n, \mathbf{X}) \leqslant c_1$. Чем больше k, тем быстрее сходится процесс итераций при малых значениях $\rho(\mathbf{x}^n, \mathbf{X})$, но каждая итерация метода при этом более трудоемка. В связи с этим в вычислительной практике наиболее распространены методы первого и второго порядков (например, метод, определяемый формулой (10), или метод Ньютона, рассматриваемый в следующем параграфе).

Примечание. Иногда в литературе встречается другое, на наш взгляд неразумное, определение порядка метода: говорят, что метод решения системы уравнений $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ имеет порядок k, если при его реализации вычисляются производные функций f_i до порядка $k - 1$ включительно.

В § 6.10 были рассмотрены итерационные методы решения линейных систем с помощью спектрально-эквивалентных операторов. Аналогичные методы применяются и для решения нелинейных систем. Выбирается оператор $\mathbf{G}(\mathbf{x})$ такой, что $\mathbf{x} = \mathbf{0}$ является единственным решением уравнения $\mathbf{G}(\mathbf{x}) = \mathbf{0}$. Приближения $\mathbf{x}^{n+1}$ к решению системы $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ определяются из соотношения

$$\mathbf{G}(\mathbf{x}^{n+1} - \mathbf{x}^n) = \mathbf{F}(\mathbf{x}^n). \tag{11}$$

Наиболее распространен случай, когда $\mathbf{G}$ — линейный оператор. В ряде случаев оператор $\mathbf{G}$ выбирается зависящим от n, а также от приближения $\mathbf{x}^n$. Тогда в схему (11) укладывается также рассматриваемый ниже метод Ньютона решения нелинейных уравнений.

§ 2. Метод Ньютона решения нелинейных уравнений

Если известно достаточно хорошее начальное приближение к решению системы уравнений

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}, \tag{1}$$

то эффективным методом повышения точности является *метод Ньютона.*

Идея метода Ньютона заключается в том, что в окрестности имеющегося приближения $\mathbf{x}^n$ задача заменяется некоторой вспомогательной линейной задачей.

Последняя задача выбирается так, чтобы погрешность замены имела более высокий порядок малости, чем первый (в определяемом далее смысле), в окрестности имеющегося приближения. За следующее приближение принимается решение этой вспомогательной задачи.

Рассмотрим случай скалярного уравнения $f(x) = 0$. В качестве такой вспомогательной задачи естественно взять линейную задачу

$$f(x_n) + f'(x_n)(x - x_n) = 0.$$

Ее решение $x = x_n - f(x_n)/f'(x_n)$ принимается за следующее приближение x_{n+1} к решению исходного уравнения, т. е. итерации ведутся по формуле

$$x_{n+1} = x_n - f(x_n)/f'(x_n).$$

Рассмотрим более общий случай — решение нелинейного функционального уравнения.

Пусть $\mathbf{F}(\mathbf{x})$ — оператор, отображающий линейное нормированное пространство H на линейное нормированное пространство Y, может быть и совпадающее с H. Нормы в этих пространствах соответственно обознача-

ем $\|\cdot\|_H$ и $\|\cdot\|_Y$. Линейный оператор $\mathbf{P}$, действующий из пространства H в пространство Y, назовем производной оператора $\mathbf{F}(\mathbf{x})$ в точке $\mathbf{x}$, если

$$\|\mathbf{F}(\mathbf{x}+\eta)-\mathbf{F}(\mathbf{x})-\mathbf{P}\eta\|_Y = o\Big(\|\eta\|_H\Big) \tag{2}$$

при $\|\eta\|_H \to 0$.

В дальнейшем будем обозначать такой оператор $\mathbf{P}$ через $\mathbf{F}'(\mathbf{x})$. Пусть, например,

$$\mathbf{x} = (x_1, \ldots, x_m)^T, \quad \mathbf{F} = (f_1, \ldots, f_m)^T.$$

Если функции f_i непрерывно дифференцируемы в окрестности данной точки $\mathbf{x}$, то

$$f_i(x_1+\eta_1, \ldots, x_m+\eta_m) = f_i(x_1, \ldots, x_m) + \sum_{i=1}^{m} \frac{\partial f_i(x_1, \ldots, x_m)}{\partial x_j}\eta_j + o\Big(\|\eta\|\Big).$$

Совокупность этих соотношений можно переписать в виде (2), если за $\mathbf{P}$ принять оператор умножения слева на матрицу

$$\mathbf{F}'(\mathbf{x}) = \left[\frac{\partial f_i}{\partial x_j}\right].$$

В простейшем случае $m = 1$ оператор $\mathbf{P}$ превращается в оператор умножения на производную f'_x.

Пусть $\mathbf{X}$ — решение уравнения $\mathbf{F}(\mathbf{X}) = \mathbf{0}$, $\mathbf{x}^n$ — некоторое приближение к $\mathbf{X}$. В предположении существования производной $\mathbf{F}'$, согласно (2), имеем

$$\big\|\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{x}^n) - \mathbf{F}'(\mathbf{x}^n)(\mathbf{X}-\mathbf{x}^n)\big\|_Y = o\Big(\|\mathbf{X}-\mathbf{x}^n\|_H\Big). \tag{3}$$

Если величина $\|\mathbf{X}-\mathbf{x}^n\|_H$ мала, то можно написать приближенное равенство

$$\mathbf{F}(\mathbf{x}^n) + \mathbf{F}'(\mathbf{x}^n)(\mathbf{X}-\mathbf{x}^n) \approx \mathbf{F}(\mathbf{X}).$$

Поскольку $\mathbf{F}(\mathbf{X}) = \mathbf{0}$, то

$$\mathbf{F}(\mathbf{x}^n) + \mathbf{F}'(\mathbf{x}^n)(\mathbf{X}-\mathbf{x}^n) \approx \mathbf{0}.$$

Возьмем в качестве следующего приближения $\mathbf{x}^{n+1}$ решение уравнения

$$\mathbf{F}(\mathbf{x}^n) + \mathbf{F}'(\mathbf{x}^n)(\mathbf{x}^{n+1}-\mathbf{x}^n) = \mathbf{0},$$

если такое решение существует. Между прочим, последнее уравнение имеет вид (1.11). В предположении, что оператор $\mathbf{F}'$ обратим, это решение можно записать в виде

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Big(\mathbf{F}'(\mathbf{x}^n)\Big)^{-1}\mathbf{F}(\mathbf{x}^n). \tag{4}$$

Такой итерационный процесс называют *методом Ньютона*.

Пусть $\Omega_a = \{\mathbf{x} : \|\mathbf{x} - \mathbf{X}\|_H < a\}$. Пусть при некоторых a, a_1, a_2, $0 < a$, $0 \leqslant a_1, a_2 < \infty$, выполнены условия:

$$\left\|(\mathbf{F}'(\mathbf{X}))^{-1}\right\|_Y \leqslant a_1 \quad \text{при} \quad \mathbf{x} \in \Omega_a; \tag{5}$$

$$\left\|\mathbf{F}(\mathbf{u}_1) - \mathbf{F}(\mathbf{u}_2) - \mathbf{F}'(\mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)\right\|_Y \leqslant a_2\|\mathbf{u}_2 - \mathbf{u}_1\|_H^2 \tag{6}$$

при $\mathbf{u}_1, \mathbf{u}_2 \in \Omega_a$. Обозначим $c = a_1 a_2$, $b = \min\left\{a, c^{-1}\right\}$.

Теорема (о сходимости метода Ньютона). *При условиях* (5), (6) *и* $\mathbf{x}^0 \in \Omega_b$ *итерационный процесс Ньютона* (4) *сходится с оценкой погрешности*

$$\|\mathbf{x}^n - \mathbf{X}\|_H \leqslant c^{-1}\left(c\|\mathbf{x}^0 - \mathbf{X}\|_H\right)^{2^n}. \tag{7}$$

Примечание. Если в рассматривавшемся выше примере в некоторой окрестности решения функции f_i имеют ограниченные вторые производные, то, согласно формуле Тейлора, имеем

$$f_i(\mathbf{y}) = f_i(\mathbf{x}) + \sum_{j=1}^{n} \frac{\partial f_i(x_1, \ldots, x_n)}{\partial x_j}(y_j - x_i) + O\left(\|\mathbf{y} - \mathbf{x}\|^2\right),$$

и, таким образом, условие (2) выполнено.

Доказательство. Пусть $\mathbf{x}^0 \in \Omega_b$. Индукцией по n докажем, что все $\mathbf{x}^n \in \Omega_b$. Пусть это утверждение доказано при некотором n; так как $b \leqslant a$, то тогда $\mathbf{x}^n \in \Omega_a$. Подставив в (6) $\mathbf{u}_1 = \mathbf{X}$ и $\mathbf{u}_2 = \mathbf{x}^n$, получим

$$\left\|\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{x}^n) - \mathbf{F}'(\mathbf{x}^n)(\mathbf{X} - \mathbf{x}^n)\right\|_Y \leqslant a_2\|\mathbf{x}^n - \mathbf{X}\|_H^2.$$

Поскольку $\mathbf{F}(\mathbf{x}^n) = -\mathbf{F}'(\mathbf{x}^n)(\mathbf{x}^{n+1} - \mathbf{x}^n)$, а $\mathbf{F}(\mathbf{X}) = \mathbf{0}$, то это соотношение может быть переписано в виде

$$\left\|\mathbf{F}'(\mathbf{x}^n)(\mathbf{x}^{n+1} - \mathbf{X})\right\|_Y \leqslant a_2\|\mathbf{x}^n - \mathbf{X}\|_H^2.$$

Воспользовавшись (5), получаем неравенство

$$\|\mathbf{x}^{n+1} - \mathbf{X}\|_H \leqslant c\|\mathbf{x}^n - \mathbf{X}\|_H^2. \tag{8}$$

Отсюда следует, что

$$\|\mathbf{x}^{n+1} - \mathbf{X}\|_H < cb^2 = (cb)b \leqslant b,$$

поэтому $\mathbf{x}^{n+1}$ также принадлежит Ω_b. Таким образом, при $\mathbf{x}^0 \in \Omega_b$ все $\mathbf{x}^n$ принадлежат Ω_b и, следовательно, для них выполняется (8).

Пусть $q_n = c\|\mathbf{x}^n - \mathbf{X}\|_H$. После умножения на c неравенство (8) запишется в виде $q_{n+1} \leqslant q_n^2$. Индукцией по n докажем справедливость неравенства

$$q_n \leqslant q_0^{2^n}.$$

При $n = 0$ оно очевидно. Предположив его верным при $n = k$, получаем

$$q_{k+1} \leqslant q_k^2 \leqslant (q_0^{2^k})^2 = q_0^{2^{k+1}}.$$

Таким образом, $q_n \leqslant q_0^{2^n}$ при всех n. Это означает, что

$$c\|\mathbf{x}^n - \mathbf{X}\|_H \leqslant \left(c\|\mathbf{x}^0 - \mathbf{X}\|_H\right)^{2^n}.$$

Отсюда следует (7). Согласно определению c и b,

$$c\|\mathbf{x}^0 - \mathbf{X}\|_H < cb \leqslant 1,$$

и поэтому $\mathbf{x}^n \to \mathbf{X}$. Теорема доказана.

Обращение оператора $\mathbf{F}'(\mathbf{x}^n)$ зачастую оказывается более трудоемкой операцией, чем вычисление значения $\mathbf{F}(\mathbf{x}^n)$. Поэтому метод Ньютона часто модифицируется следующим образом. По ходу вычислений выбирают или заранее задаются некоторой возрастающей последовательностью чисел $n_0 = 0, n_1, n_2 \ldots$. При $n_k \leqslant n < n_{k+1}$ итерации производят по формуле

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \left(\mathbf{F}'(\mathbf{x}^{n_k})\right)^{-1}\mathbf{F}(\mathbf{x}^n).$$

Увеличение числа итераций, сопровождающее такую модификацию, компенсируется большей «дешевизной» одного шага итерации. Выбор последовательности $\{n_k\}$ нужно производить с обоюдным учетом этих факторов.

Рассмотрим геометрическую интерпретацию метода Ньютона в случае решения скалярного уравнения $f(x) = 0$, когда расчетная формула (4) приобретает вид

$$x^{n+1} = x^n - f(x^n)/f'(x^n). \tag{9}$$

Для получения x^{n+1} геометрически надо найти абсциссу точки пересечения с осью x касательной к кривой $y = f(x)$ в точке $(x^n, f(x^n))$ (рис. 7.2.1). Уже в случае, когда $f(x)$ — многочлен третьей степени, может случиться, что последовательность $\{x_n\}$ не сходится к корню при плохом начальном приближении.

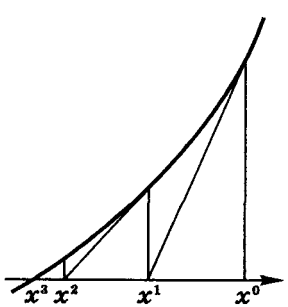


Рис. 7.2.1

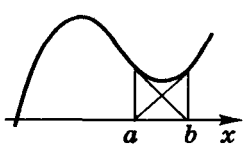


Рис. 7.2.2

Например, в случае, изображенном на рис. 7.2.2, все четные приближения совпадают с a, а нечетные — с b; метод, как говорят, «зациклился». Для более сложных задач реальное поведение приближений x^n при плохом начальном приближении становится существенно более запутанным и трудно поддающимся анализу.

Сравним асимптотическую скорость сходимости методов Ньютона и простой итерации. Для последнего мы имели оценку погрешности

$$\|\mathbf{x}^n - \mathbf{X}\| \leqslant q^n \|\mathbf{x}^0 - \mathbf{X}\|, \quad q < 1.$$

Чтобы погрешность стала меньше ε, согласно этой оценке достаточно взять

$$n \geqslant \log_{q^{-1}} \frac{\|\mathbf{x}^0 - \mathbf{X}\|}{\varepsilon} \sim \log_{q^{-1}} \frac{1}{\varepsilon}.$$

В случае метода Ньютона правая часть (7) будет меньше ε, если

$$n \geqslant -\log_2 \frac{\log_2(c\|\mathbf{x}^0 - \mathbf{X}\|)}{\log_2(c\varepsilon)} \sim \log_2 \log_2 \frac{1}{\varepsilon}. \tag{10}$$

Таким образом, асимптотически, при $\varepsilon \to 0$, метод Ньютона требует меньшего числа итераций.

Задача 1. Доказать, что для метода k-го порядка, $k > 1$, при наличии достаточно хорошего начального приближения число итераций, требуемое для достижения точности ε, будет $n \sim \log\log \varepsilon^{-1} / \log k$.

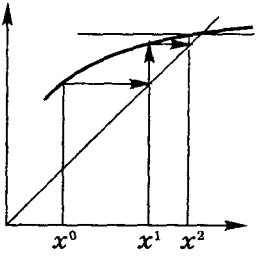


Рис. 7.2.3

Обратим внимание, что метод Ньютона, записанный в форме (4), сам является разновидностью метода простой итерации. В случае скалярного уравнения $f(x) = 0$ хорошо видна еще одна особенность метода Ньютона. Производная правой части (9) $g(x) = x - f(x)/f'(x)$ по x равна $f(x)f''(x)/(f'(x))^2$. Таким образом, $g'(X) = 0$, если $f'(X) \neq 0$, и рис. 7.1.1 в этом случае приобретает следующий вид (рис. 7.2.3)

Метод Ньютона оказывается удобным способом извлечения корней целой степени. Задача извлечения корня $\sqrt[p]{a}$, p — целое число, равносильна задаче решения уравнений $x^p - a = 0$. Расчетная формула метода Ньютона в этом случае приобретает вид

$$x_{n+1} = \frac{p-1}{p} x_n + \frac{a}{p x_n^{p-1}}.$$

Задача 2. Рассматривается алгоритм вычисления $\sqrt{a}$ при $1 \leqslant a \leqslant 4$, x_0 полагается равным значению многочлена наилучшего равномерного при-

ближения для $\sqrt{a}$ на $[1, 4]$: $x_0 = p_1(a) = \frac{17}{24} + \frac{a}{3}$. Убедиться в справедливости неравенства $|x_4 - \sqrt{a}| \leqslant 0{,}5 \cdot 10^{-25}$.

В случае решения одного скалярного уравнения $f(x) = 0$ наряду с методом Ньютона употребителен *метод секущих.*

Простейший вариант этого метода заключается в следующем. В процессе итераций фиксируется некоторая точка x^0. Приближение x^{n+1} находится как абсцисса точки пересечения прямой, проходящей через точки $(x_0, f(x^0))$ и $(x^n, f(x^n))$, с осью x (рис. 7.2.4).

Более эффективен способ, где за x^{n+1} принимается абсцисса точки пересечения с осью x прямой, проходящей через точки $(x^{n-1}, f(x^{n-1}))$ и $(x^n, f(x^n))$ (рис. 7.2.5). Уравнение этой прямой

$$y_n(x) = f(x^n) + (x - x^n)\frac{f(x^n) - f(x^{n-1})}{x^n - x^{n-1}}.$$

Из условия $y_n(x^{n+1}) = 0$ получаем

$$x^{n+1} = x^n - \frac{f(x^n)(x^n - x^{n-1})}{f(x^n) - f(x^{n-1})}. \tag{11}$$

Вычисления прекращают, когда одна из величин $|x^{n+1} - x^n|$ или $|f(x^{n+1}) - f(x^n)|$ становится меньше некоторого заранее заданного малого $\delta > 0$. Для достижения точности ε этим методом, как и в случае метода Ньютона, при достаточно хороших начальных приближениях требуется $O(\ln\ln(1/\varepsilon))$ итераций.

При решении системы m уравнений

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}$$

одним из возможных обобщений метода секущих является следующий метод. Пусть определены приближения $\mathbf{x}^{n-m}, \ldots, \mathbf{x}^n$ и известны значения

$$f_i(\mathbf{x}^{n-m}), \ldots, f_i(\mathbf{x}^n).$$

Пусть $y = L_i(\mathbf{x})$ — уравнение плоскости, проходящей через точки

$$\big(\mathbf{x}^{n-m}, f_i(\mathbf{x}^{n-m})\big), \ldots, \big(\mathbf{x}^n, f_i(\mathbf{x}^n)\big);$$

за следующее приближение $\mathbf{x}^{n+1}$ принимаем решение системы уравнений

$$L_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m.$$

При больших n эти плоскости становятся практически параллельными, поэтому для $m > 1$ этот метод применяется редко, обычно в случае, когда можно ограничиться невысокой точностью.

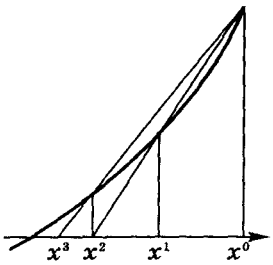


Рис. 7.2.4

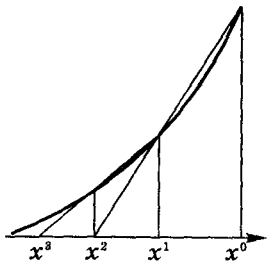


Рис. 7.2.5

В последнее время появились более совершенные обобщения метода секущих.

Дело в том, что для этого метода при $n \to \infty$ характерно «сплющивание» m-мерного тетраэдра с вершинами в точках $\mathbf{x}^{n-m}, \ldots, \mathbf{x}^n$. Следствием этого является быстрое ухудшение обусловленности системы уравнений $L_i(\mathbf{x}) = 0$. В результате алгоритм вычисления становится неустойчивым к вычислительной погрешности и часто перестает сходиться.

Кроме описанных выше, существует большое число других методов подобного типа, где в окрестности корня функция $f(x)$ приближается некоторой функцией $g(x)$, для которой уравнение $g(x) = 0$ решается в явном виде. Однако для применения всех этих методов необходимо достаточно хорошее приближение к решению. Иногда для его определения используется *метод вилки*. Определяют a_0, b_0 такие, что $f(a_0)f(b_0) < 0$; выбирают каким-либо образом точку $c_0 \in (a_0, b_0)$, например берут $c_0 = (a_0 + b_0)/2$ или за c_0 берут точку пересечения секущей, проходящей через точки $(a_0, f(a_0))$, $(b_0, f(b_0))$, с осью x. После вычисления $f(c_0)$ за $[a_1, b_1]$ принимают тот из отрезков $[a_0, c_0]$, $[c_0, b_0]$, на концах которого $f(x)$ принимает противоположные знаки, и т. д.

Важной задачей является разработка эффективных методов решения уравнений отдельных типичных классов. Для нахождения корней многочлена $P(z) = a_0 z^m + \cdots + a_m$ как с действительными, так и с комплексными коэффициентами таким методом является *метод парабол*. При заданных приближениях к корню z_{n-2}, z_{n-1}, z_n приближение z_{n+1} определяется следующим образом. Строится интерполяционный многочлен второй степени, совпадающий с $P(z)$ в точках z_{n-2}, z_{n-1}, z_n. За z_{n+1} принимается корень этого многочлена, наиболее близкий к z_n. В стандартных программах метода парабол эта схема подвергнута некоторой модификации.

§ 3. Методы спуска

Для решения задачи минимизации функционала наиболее часто применяются *методы спуска*. При заданном приближении определяется какое-либо направление, в котором функционал убывает, и производится перемещение приближения в этом направлении. Если величина перемещения

взята не очень большой, то значение функционала обязательно уменьшится.

Рассмотрим примеры методов спуска.

При исследовании сходимости метода Зейделя в случае системы уравнений $A\mathbf{x} = \mathbf{b}$ при $A > 0$ мы описали циклический метод покоординатного спуска минимизации функции $\Phi(x_1, \ldots, x_m)$: при заданном приближении $\Phi(x_1^0, \ldots, x_m^0)$ отыскивается значение $x_1 = x_1^1$, при котором достигается $\inf\limits_{x_1} \Phi(x_1, x_2^0, \ldots, x_m^0)$, затем отыскивается значение $x_2 = x_2^1$, при котором достигается $\inf\limits_{x_2} \Phi(x_1^1, x_2, x_3^0, \ldots, x_m^0)$, и т. д. Процесс циклически повторяется.

Обозначим через $P_k\mathbf{x}$ приближение, получаемое при спуске из $\mathbf{x}$ по координате x_k. Присваивая приближению, получающемуся при спуске по очередной координате, следующий номер, можно записать приближения метода циклического покоординатного спуска в виде

$$\mathbf{x}^1 = P_1\mathbf{x}^0, \qquad \mathbf{x}^2 = P_2P_1\mathbf{x}^0, \ldots,$$
$$\mathbf{x}^m = P_m \ldots P_1\mathbf{x}^0, \quad \mathbf{x}^{m+1} = P_1P_m \ldots P_1\mathbf{x}^0, \ldots$$

При практической реализации этого метода возникает проблема минимизации функций одной переменной. Рассмотрим отдельно задачу минимизации функций одной переменной $\Phi(x)$ при начальном приближении к точке минимума $x = x^0$. Так как эта задача обычно не может быть решена точно, то часто поступают следующим образом: берут некоторые значения $\bar{x}^0$, $\bar{\bar{x}}^0$, и строят параболу $y = Q_2(x)$, удовлетворяющую условиям

$$Q_2(x^0) = \Phi(x^0), \quad Q_2(\bar{x}^0) = \Phi(\bar{x}^0), \quad Q_2(\bar{\bar{x}}^0) = \Phi(\bar{\bar{x}}^0).$$

Абсциссу x точки минимума $Q_2(x)$ принимают за следующее приближение x^1. Уже в одномерном случае можно построить пример, когда последовательность точек, получаемых по описываемому методу, не обязательно сходится к искомой точке экстремума функции $\Phi(x)$.

Даже если на каждом шаге отыскивается абсолютный экстремум функции $\Phi(x_1, \ldots, x_m)$ по соответствующей координате, то уже при $m = 2$ может случиться, что итерационный процесс сходится не к искомой точке абсолютного экстремума, а к некоторой точке локального экстремума. На рис. 7.3.1 изображены линии уровня такой функции и получаемые приближения. Спуск в циклическом порядке необязателен. Если из рассмотрения проводившихся ранее вычислений видно, что спуск по каким-либо координатам обеспе-

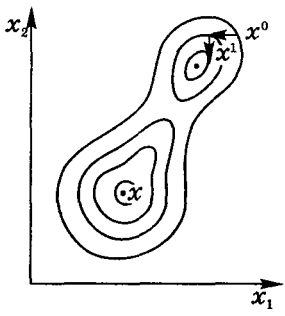


Рис. 7.3.1

чивает наибольшее убывание $\Phi(x)$, то иногда целесообразен более частый спуск по этим координатам.

В других случаях при каждом n после получения приближения x^n выбирается некоторая совокупность координат

$$x_{i(n,1)}, \ldots, x_{i(n,q(n))},$$

производятся независимые спуски по координатам этой группы исходя из приближения $\mathbf{x}^n$, т. е. находят точки $P_{i(n,k)}\mathbf{x}^n$. Далее вычисляется

$$\min_{1 \leqslant k \leqslant q(n)} \Phi(P_{i(n,k)}\mathbf{x}^n),$$

и соответствующая минимуму точка $P_{i(n,k)}\mathbf{x}^n$ принимается за $\mathbf{x}^{n+1}$.

Иногда номер очередной координаты, по которой осуществляется спуск, выбирается недетерминированно. В этом случае говорят *о случайном покоординатном спуске.*

Другой вариант метода спуска — *метод наискорейшего (градиентного) спуска.* Следующее приближение отыскивается в виде

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \delta_n \operatorname{grad} \Phi(\mathbf{x}^n)$$

(рис. 7.3.2). Значение δ_n определяется из условия

$$\min_{\delta_n} \Phi(\mathbf{x}^n - \delta_n \operatorname{grad} \Phi(\mathbf{x}^n)),$$

т. е. этот алгоритм опять состоит в последовательной минимизации функции одной переменной δ_n.

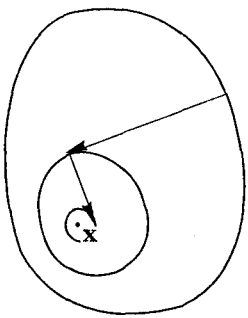


Рис. 7.3.2

Как и в методе покоординатного спуска, в методе наискорейшего спуска нет необходимости полного решения вспомогательной задачи минимизации функции одной переменной. В окрестности точки своего минимума эта функция меняется мало, и тщательное нахождение ее точки минимума не приводит к существенному эффекту. В случае метода наискорейшего спуска вопрос об объеме вычислений при минимизации вспомогательных функций одной переменной должен решаться также с учетом относительной трудоемкости вычисления значений функции $\Phi(x)$ и ее градиента.

Для иллюстрации решения вопроса о выборе метода рассмотрим следующую типичную задачу: решается нелинейная краевая задача для системы обыкновенных дифференциальных уравнений. В гл. 9 будет показано, что эта задача сводится к решению нелинейной системы уравнений $\mathbf{F}(\mathbf{x}) = \mathbf{0}$:

$$f_i(\mathbf{x}) = f_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, m,$$

обладающей следующими свойствами.

Количество операций при отыскании одного значения $f_i(\mathbf{x})$ и при одновременном отыскании всех значений $f_i(\mathbf{x})$, $i = 1, \ldots, m$, в той же точке одинаково; обозначим его через A. Количество операций при непосредственном отыскании значений $\partial f_i(\mathbf{x})/\partial x_j$ и при одновременном отыскании всех значений $\partial f_i(\mathbf{x})/\partial x_j$, $i = 1, \ldots, m$, в той же точке одинаково. Обозначим его через B; обычно $B \gg A$.

Тогда при решении задачи методом Ньютона целесообразно вычислять производные $\partial f_i(\mathbf{x})/\partial x_j$, $j = 1, \ldots, m$, пользуясь приближенной формулой

$$\frac{\partial f_i(x_1, \ldots, x_m)}{\partial x_j} \approx \approx \frac{f_i(x_1, \ldots, x_{j-1}, x_j + \Delta, x_{j+1}, \ldots, x_m) - f_i(x_1, \ldots, x_j, \ldots, x_m)}{\Delta}. \tag{1}$$

Область сходимости метода Ньютона обычно невелика, поэтому по крайней мере на начальном этапе итераций целесообразно свести решение этой задачи к минимизации некоторого функционала и применить какой-либо из методов спуска. Рассмотрим простейший случай функционала

$$\Phi(\mathbf{x}) = \sum_{i=1}^{m} (\lambda_i f_i(\mathbf{x}))^2;$$

множители $\lambda_i = \text{const} \neq 0$, называемые *масштабными*, подбираются из условий конкретной задачи.

Пусть $\mathbf{x}^n = (x_1^n, \ldots, x_m^n)$ — полученное приближение и решено сделать спуск в направлении $\Delta = (\Delta_1, \ldots, \Delta_m)$, $\|\Delta\| = 1$. Вычисляют приближенные значения производных в этом направлении:

$$l_i = (f_i(\mathbf{x}^n + \varepsilon\Delta) - f_i(\mathbf{x}^n))/\varepsilon. \tag{2}$$

На прямой $\mathbf{x} = \mathbf{x}^n + t\Delta$ имеем

$$f_i(\mathbf{x}) \approx f_i(\mathbf{x}^n) + tl_i,$$

поэтому следующее приближение $\mathbf{x}^{n+1} = \mathbf{x}^n + t\Delta$ определяют из условия

$$\min_i \sum_{i=1}^{m} \Big(\lambda_i (f_i(\mathbf{x}^n) + tl_i)\Big)^2.$$

Отметим, что в данном случае существен вопрос о разумном выборе ε в формуле (2). (По этому поводу см. § 2.16.)

Задача минимизации функции при наличии ограничений, так называемая *задача условной минимизации*, формулируется следующим образом. Ищется величина

$$A = \inf_{x_1, \ldots, x_m} \Phi(x_1, \ldots, x_m) \tag{3}$$

при условиях

$$\varphi_i(x_1, \ldots, x_m) \geqslant 0, \quad i = 1, \ldots, l, \tag{4}$$

$$\psi_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, q. \tag{5}$$

При решении этой задачи возникают дополнительные трудности по сравнению с решением задачи отыскания безусловного минимума, в которой ищется

$$\inf_{(x_1, \ldots, x_m) \in \mathbf{R}_m} \Phi(x_1, \ldots, x_m)$$

— нижняя грань $\Phi(x_1, \ldots, x_m)$ по всему пространству $\mathbf{R}_m$.

Непосредственное использование многих из описанных выше методов становится невозможным, и возникает необходимость в их модификации.

В то же время задача минимизации функции при ограничениях типа (4), (5) является весьма актуальной для приложений. Например, существует целый раздел математики — *линейное программирование*, — занимающийся решением задачи (3)–(5) в случае, когда Φ, φ_i, ψ_i — линейные функции аргументов x_j.

Среди других методов, связанных с решением задачи (3)–(5), упомянем *метод штрафа*. Строится последовательность функций $\Phi_\lambda(x_1, \ldots, x_m)$, удовлетворяющая следующим условиям:

1) $\Phi_\lambda(x_1, \ldots, x_m)$ определена при всех $(x_1, \ldots, x_m)$;

2) $\inf\limits_{\mathbf{R}_m} \Phi_\lambda(x_1, \ldots, x_m) = A_\lambda \to A$ при $\lambda \to \infty$;

3) если существуют точки $(\bar{x}_1, \ldots, \bar{x}_m)$ и $(x_1^\lambda, \ldots, x_m^\lambda)$ такие, что

$$\Phi(\bar{x}_1, \ldots, \bar{x}_m) = A, \quad \Phi_\lambda(x_1^\lambda, \ldots, x_m^\lambda) = A_\lambda,$$

то $(x_1^\lambda, \ldots, x_m^\lambda) \to (\bar{x}_1, \ldots, \bar{x}_m)$ при $\lambda \to \infty$.

Вместо решения исходной задачи (3)–(5) решается задача отыскания минимума $\inf\limits_{\mathbf{R}_m} \Phi_\lambda(x_1, \ldots, x_m)$ при достаточно больших λ.

Часто вместо первого условия на функцию Φ_λ требуют выполнения более слабого условия. Функция Φ_λ определена в точках некоторой односвязной области G_λ, $\Phi_\lambda(x_1, \ldots, x_m) \to \infty$ при приближении точки $(x_1, \ldots, x_m)$ к границе области или при $x_1^2 + \cdots + x_m^2 \to \infty$ (в случае, когда область G_λ неограниченная).

В отдельных случаях условие 3) также несколько модифицируется.

Приведем пример построения такой функции Φ_λ. Для этого соотношения (4), (5) записываются в таком виде, что φ_i, ψ_i будут определены при всех $(x_1, \ldots, x_m)$ (или при всех $x_1, \ldots, x_m$ из G_λ).

Вводится некоторая невозрастающая функция $h(t)$, определенная при $-\infty < t < \infty$, такая, что $\lim\limits_{t \to -\infty} h(t) > 0$, $\lim\limits_{t \to +\infty} h(t) = 0$. Например, можно взять

$$h(t) = \frac{1}{2} - \frac{1}{\pi} \operatorname{arctg} t.$$

В качестве $\Phi_\lambda(x_1, \ldots, x_m)$ берется функция

$$\Phi(x_1, \ldots, x_s) + \lambda \sum_{i=1}^{m} \psi_i^2(x_1, \ldots, x_m) + \lambda \sum_{i=1}^{q} h(\lambda \varphi_i(x_1, \ldots, x_m)).$$

Наличие слагаемого $\lambda \psi_i^2(x_1, \ldots, x_m)$ заставляет смещаться точку экстремума $(x_1^\lambda, \ldots, x_m^\lambda)$ в область, где $\psi_i(x_1, \ldots, x_m) = 0$; в то время как наличие слагаемого $\lambda h(\lambda \varphi_i(x_1, \ldots, x_m))$ заставляет смещаться точку экстремума $(x_1^\lambda, \ldots, x_m^\lambda)$ в область, где $\varphi_i(x_1, \ldots, x_m) \geqslant 0$.

Метод штрафа обладает следующим недостатком. Оказывается, что при больших λ структура линий уровня Φ_λ, как правило, такова, что сходимость методов минимизации существенно замедляется. Искусство применения метода штрафа при решении конкретных задач состоит в удачном выборе функции Φ_λ такой, что при заданной близости значений нижней грани $|A - A_\lambda| \leqslant \varepsilon$ замедление скорости сходимости применяемого итерационного метода будет минимальным.

В связи с отмеченными недостатками метода штрафа разработано большое число других методов решения задачи условной минимизации (3)–(5).

§ 4. Другие методы сведения многомерных задач к задачам меньшей размерности

Иногда полезно рассмотреть следующую формальную процедуру сведения многомерных задач к одномерным.

Пусть ищется минимум A функции $\Phi(x_1, \ldots, x_m)$ в области

$$\begin{aligned} &\varphi_i(x_1, \ldots, x_m) \geqslant 0, \; i = 1, \ldots, l, \\ &\psi_j(x_1, \ldots, x_m) = 0, \; j = 1, \ldots, q. \end{aligned} \tag{1}$$

Можно написать равенство

$$A = \min_{x_1, \ldots, x_m} \Phi(x_1, \ldots, x_m) = \min_{x_m} \Phi_m(x_m),$$

где

$$\Phi_m(x_m) = \min_{x_1, \ldots, x_{m-1}} \Phi(x_1, \ldots, x_m);$$

минимум каждый раз берется по области определения минимизируемой функции в соответствии с условиями (1). Таким образом, исходная задача минимизации функции m переменных свелась к минимизации функции одной переменной, каждое значение которой определяется минимизацией функции $(m-1)$-й переменной. В свою очередь минимизацию функции $\Phi_m(x_m)$ сведем к минимизации фунции одной переменной, каждое

значение которой определяется минимизацией функции от $(m-2)$-х переменных, и т. д. Получим цепочку соотношений

$$A = \min_{x_m} \Phi_m(x_m),$$
$$\Phi_m(x_m) = \min_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m),$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$\Phi_3(x_3, \dots, x_m) = \min_{x_2} \Phi_2(x_2, \dots, x_m),$$
$$\Phi_2(x_2, \dots, x_m) = \min_{x_1} \Phi_1(x_1, \dots, x_m).$$

Кажущейся простоте метода сопутствует его большая трудоемкость. Предположим, что каждая минимизация функций одной переменной потребует вычисления s значений минимизируемой функции. Тогда минимизация $\Phi_m(x_m)$ требует нахождения $\min\limits_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m)$ при s значениях параметра x_m, т. е. s^2 вычислений значений функции $\Phi_{m-1}(x_{m-1}, x_m)$. Это в свою очередь потребует вычисления s^2 значений $\Phi_{m-2}(x_{m-2}, x_{m-1}, x_m)$ и т. д. В конечном счете потребуется вычисление s^m значений функции Φ. Уже при умеренных s и m, например $s = 10$, $m = 10$, такой объем вычислений окажется недопустимо большим. Однако при малых m некоторые модификации рассматриваемой идеи оказываются полезными. Например, возможен такой вариант. Задаются начальным приближением $(x_1^0, \dots, x_m^0)$. Реализуют указанный алгоритм при не очень большом значении s, например $s = 3$ или $s = 4$. При этом значения функции вычисляются в точках некоторого параллелепипеда $|x_i - x_i^0| \leqslant \Delta_i^0$. Получаемую точку минимума $(x_1^1, \dots, x_m^1)$ принимают за следующее приближение; приближение $(x_1^2, \dots, x_m^2)$ находят аналогичным образом, но значения функции Φ вычисляются в точках параллелепипеда $|x_i - x_i^1| \leqslant \Delta_i^1$ и т. д.

Рассмотрим еще один метод, общая структура которого похожа на структуру описанного выше.

Если линии уровня функции Φ похожи на сферы, то при применении методов спуска происходит быстрое смещение в направлении минимума (рис. 7.4.1). Однако практически более типичен случай, когда эти линии похожи на эллипсоиды с большим разбросом полуосей. Тогда при движении по градиенту смещение в направлении точки минимума будет довольно медленным (рис. 7.4.2).

Предположим, что оси этих «эллипсоидов» естественным образом разбиваются на две группы: первая группа состоит из m_1 осей одного порядка и относительно малых, вторая группа состоит из m_2 осей одного порядка и относительно больших.

При решении некоторых задач такого рода хорошо зарекомендовал себя следующий метод, называемый *методом оврагов*. Задаются какими-то приближениями $\mathbf{x}^0$ и $\mathbf{x}^1$ и производят несколько шагов метода спуска, исходя из каждого из этих приближений. Будут получены прибли-

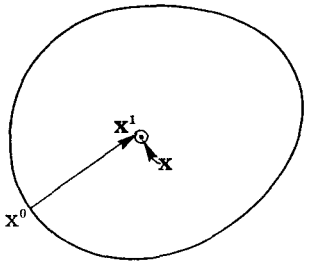


Рис. 7.4.1

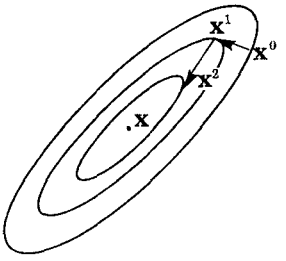


Рис. 7.4.2

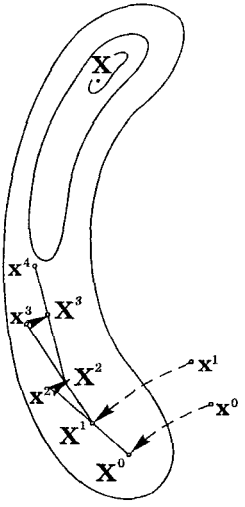


Рис. 7.4.3

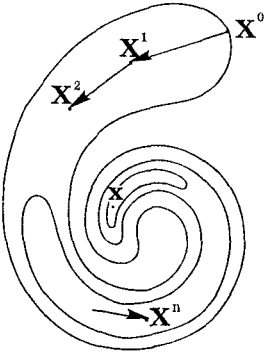


Рис. 7.4.4

жения $\mathbf{X}^0$ и $\mathbf{X}^1$. Из рис. 7.4.3 видно, что эти приближения будут лежать в многообразии, расположенном в окрестности осей второй группы. Процесс итераций состоит в получении последовательных приближений $\mathbf{X}^0, \mathbf{X}^1, \ldots,$ лежащих в окрестности этого многообразия. В случае $m_2 = 1$ приближение $\mathbf{X}^{l+1}$ отыскивается следующим образом. Проведем через $\mathbf{X}^{l-1}$ и $\mathbf{X}^l$ прямую и найдем приближение $\mathbf{x}^{l+1}$ к точке минимума $\Phi(x)$ на этой прямой. Таким образом, это приближение ищется в виде

$$\mathbf{x}^{l+1} = \mathbf{X}^l + \alpha(\mathbf{X}^l - \mathbf{X}^{l-1}).$$

Далее проводим несколько итераций исходя из $\mathbf{x}^{l+1}$ и получаем приближение $\mathbf{X}^{l+1}$, также лежащее в овраге. В случае $m_2 > 1$ приближение $\mathbf{x}^{l+1}$ иногда удобно отыскивать в виде

$$\mathbf{x}^{l+1} = \mathbf{X}^l + \alpha(\mathbf{X}^l - \mathbf{X}^{l-1}) + \beta \operatorname{grad} \Phi(\mathbf{X}^l).$$

Из рис. 7.4.4 видно, что описанный способ оказывается эффективным и в ряде случаев, когда линии уровня функции $\Phi(\mathbf{x})$ имеют более сложную структуру.

Решение системы уравнений

$$f_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, m, \tag{2}$$

также формально сводится к последовательному решению уравнений с одним неизвестным. Рассмотрим систему уравнений

$$f_i(x_1, \ldots, x_m) = 0, \quad i = 2, \ldots, m, \tag{3}$$

относительно неизвестных $x_2, \ldots, x_m$. Пусть $x_2(x_1), \ldots, x_m(x_1)$ — ее решение. Подставляя выражения $x_2(x_1), \ldots, x_m(x_1)$ в первое из уравнений (2), получим уравнение

$$F_1(x_1) = f_1\Big(x_1, x_2(x_1), \ldots, x_m(x_1)\Big) = 0 \tag{4}$$

относительно одной неизвестной x_1.

Отыскание значений $x_2(x_1), \ldots, x_m(x_1)$ и решение уравнения (4) можно проводить численно. Выбирается какой-то метод решения (4) по значениям функции $F(x_1)$; при каждом требуемом значении x_1 в результате решения (3) получаются значения $x_2(x_1), \ldots, x_m(x_1)$, которые подставляются затем в правую часть (4). Для решения системы уравнений (3) при каждом значении x_1 применим тот же прием.

Пусть $x_3(x_1, x_2), \ldots, x_m(x_1, x_2)$ — решение системы уравнений

$$f_i(x_1, \ldots, x_m) = 0, \quad i = 3, \ldots, m, \tag{5}$$

относительно неизвестных $x_3, \ldots, x_m$. Подставляя $x_3(x_1, x_2), \ldots, x_m(x_1, x_2)$ во второе уравнение системы, получим уравнение

$$F_2(x_1, x_2) = f_2\Big(x_1, x_2, x_3(x_1, x_2), \ldots, x_m(x_1, x_2)\Big) = 0.$$

При каждом x_1 это уравнение может быть разрешено относительно x_2. Его решение $x_2(x_1)$, а также $x_3(x_1, x_2(x_1)), \ldots, x_m(x_1, x_2(x_1))$ образуют решение системы (3). Систему (5) при каждых x_1, x_2 опять решаем, сводя к системе, где число неизвестных на единицу меньше, и т. д.

Если для решения каждого вспомогательного уравнения с одной неизвестной потребуется s вычислений функций, то суммарно этот алгоритм потребует порядка s^m вычислений правых частей уравнений системы.

По поводу реального применения этого алгоритма можно сказать все то же, что и по поводу применения описанного в начале параграфа метода минимизации.

Задача 1. Рассмотреть, во что переходит описанный метод в случае, когда система уравнений (2) линейная.

§ 5. Решение стационарных задач путем установления

Распространенным методом решения стационарных задач является *метод установления*. В этом случае для решения стационарной задачи строится нестационарный процесс, решение которого с течением времени оказывается независимым от него и устанавливается к решению исходной стационарной задачи. Рассмотрим систему дифференциальных уравнений

$$\frac{d\mathbf{x}}{dt} + \operatorname{grad}\Phi(\mathbf{x}) = 0. \tag{1}$$

Вектор $d\mathbf{x}/dt$ пропорционален градиенту функции $\Phi(\mathbf{x})$, т. е. ортогонален ее линиям уровня и направлен в сторону убывания значений функции $\Phi(\mathbf{x})$. Таким образом, при перемещении вдоль траектории системы (1) значение $\Phi(\mathbf{x})$ не возрастает. Формально справедливость этого утверждения следует из неравенства

$$\frac{d\Phi(\mathbf{x})}{dt} = \left(\operatorname{grad}\Phi(\mathbf{x}), \frac{d\mathbf{x}}{dt}\right) = -(\operatorname{grad}\Phi(\mathbf{x}), \operatorname{grad}\Phi(\mathbf{x})), \tag{2}$$

означающего, что $d\Phi(\mathbf{x})/dt < 0$ всюду, за исключением стационарных точек функции $\Phi(\mathbf{x})$.

Другой нестационарный процесс, решение которого при весьма общих предположениях устанавливается к точке минимума функции $\Phi(\mathbf{x})$, описывается системой дифференциальных уравнений

$$\frac{d^2\mathbf{x}}{dt^2} + \gamma\frac{d\mathbf{x}}{dt} + \operatorname{grad}\Phi(\mathbf{x}) = \mathbf{0}, \quad \gamma > 0. \tag{3}$$

Для решений этой системы имеем

$$\frac{d}{dt}\left(\frac{1}{2}\left(\frac{d\mathbf{x}}{dt}, \frac{d\mathbf{x}}{dt}\right) + \Phi(\mathbf{x})\right) = \left(\frac{d\mathbf{x}}{dt}, \frac{d^2\mathbf{x}}{dt^2}\right) +$$

$$+ \left(\operatorname{grad}\Phi(\mathbf{x}), \frac{d\mathbf{x}}{dt}\right) = -\gamma\left(\frac{d\mathbf{x}}{dt}, \frac{d\mathbf{x}}{dt}\right) < 0, \tag{4}$$

если только $d\mathbf{x}/dt \neq 0$. Функцию $\Phi(\mathbf{x})$ в первом случае и $\frac{1}{2}\left(\frac{d\mathbf{x}}{dt}, \frac{d\mathbf{x}}{dt}\right) + \Phi(\mathbf{x})$ — во втором можно рассматривать как энергию материальной системы, движение которой описывается системами уравнений (1) и (3). Соотношения (2) и (4) показывают, что рассматриваемые нестационарные процессы характеризуются оттоком или, как говорят, *диссипацией* энергии.

Чтобы прояснить вопрос о разумном выборе γ, рассмотрим простейшую модель: x — скаляр, $\Phi(x) = \dfrac{a^2x^2}{2}$; тогда (3) приобретает вид

$$x'' + \gamma x' + a^2 x = 0.$$

Соответствующее характеристическое уравнение

$$\lambda^2 + \gamma\lambda + a^2 = 0,$$

его корни $\lambda_{1,2} = -\dfrac{\gamma}{2} \pm \sqrt{\dfrac{\gamma^2}{4} - a^2}$, и при $\lambda_1 \neq \lambda_2$, т. е. при $\gamma \neq 2a$ общее решение есть $c_1 \exp\{\lambda_1 t\} + c_2 \exp\{\lambda_2 t\}$.

Скорость убывания решений рассматриваемого уравнения определяется величиной

$$\sigma(\gamma) = \max(\mathrm{Re}\lambda_1, \mathrm{Re}\lambda_2).$$

При $\gamma \leqslant 2a$ имеем $\gamma^2/4 - a^2 \leqslant 0$, и поэтому

$$\mathrm{Re}\lambda_1 = \mathrm{Re}\lambda_2 = \sigma(\gamma) = -\gamma/2 \geqslant -a.$$

При $\gamma > 2a$ величины λ_1 и λ_2 вещественны и

$$\mathrm{Re}\lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} > \mathrm{Re}\lambda_2 = -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - a^2}.$$

Тогда

$$\sigma(\gamma) = \mathrm{Re}\lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} =$$

$$= -\frac{a^2}{\dfrac{\gamma}{2} + \sqrt{\dfrac{\gamma^2}{4} - a^2}} > -\frac{a^2}{\dfrac{\gamma}{2}} > -a.$$

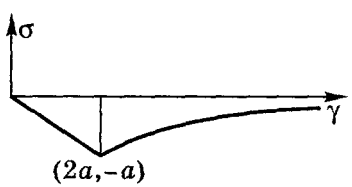


Рис. 7.5.1

Таким образом, график $\sigma(\gamma)$ имеет вид, изображенный на рис. 7.5.1, и $\min\limits_{\gamma} \sigma(\gamma) = \sigma(2a) = -a$. Из результатов рассмотрений этой модельной задачи можно сделать следующие качественные выводы.

1. Если коэффициент трения γ очень мал (в нашем случае $\gamma \ll 2a$), то решение системы (3) медленно устанавливается к положению равновесия; при этом (вследствие условия $\mathrm{Im}\lambda_1$, $\mathrm{Im}\lambda_2 \neq 0$) происходят колебания около положения равновесия.

2. Если γ велико ($\gamma \gg 2a$), то решение также медленно устанавливается; причина состоит в том, что при большом коэффициенте трения γ движение не может приобрести большой скорости.

3. Оптимальное значение γ лежит где-то посередине и зависит от свойств конкретной функции $\Phi(x)$.

Метод установления с помощью решения системы (3) иногда называют *методом тяжелого шарика*. Это название обусловлено следующими соображениями.

Рассмотрим движение материальной точки по поверхности $y = \Phi(x)$ в поле тяжести, направленном в отрицательном направлении оси y. Предположим, что трение пропорционально скорости и точка не может отрываться от поверхности. Тогда ее движение опишется системой уравнений

$$\frac{d^2\mathbf{x}}{dt^2} + \gamma\frac{d\mathbf{x}}{dt} + \frac{\operatorname{grad}\Phi(\mathbf{x})}{1 + \|\operatorname{grad}\Phi(\mathbf{x})\|^2} = 0.$$

Ясно, что решение этой системы с течением времени установится к некоторой стационарной точке функции $\Phi(\mathbf{x})$. Вблизи экстремума $\|\operatorname{grad}\Phi(\mathbf{x})\| \ll 1$ эта система близка к системе (3).

Большинство известных методов установления описывается уравнениями вида

$$A_0\left(\mathbf{x}, \frac{d\mathbf{x}}{dt}\right)\frac{d\mathbf{x}}{dt} + A_1\left(\mathbf{x}, \frac{d\mathbf{x}}{dt}, \operatorname{grad}\Phi(\mathbf{x})\right) = 0 \tag{5}$$

или

$$B_0\left(\mathbf{x}, \frac{d\mathbf{x}}{dt}\right)\frac{d^2\mathbf{x}}{dt^2} + B_1\left(\mathbf{x}, \frac{d\mathbf{x}}{dt}, \operatorname{grad}\Phi(\mathbf{x})\right) = 0, \tag{6}$$

где

$$A_0(\mathbf{X}, \mathbf{0}) \neq 0, \qquad A_1(\mathbf{X}, \mathbf{0}, \mathbf{0}) = 0,$$

$$B_0(\mathbf{X}, \mathbf{0}) \neq 0, \qquad B_1(\mathbf{X}, \mathbf{0}, \mathbf{0}) = 0$$

и выполнены условия диссипативности, обеспечивающие сходимость к точке экстремума $\mathbf{X}$. Вообще говоря, можно обратить операторы A_0 и B_0 и преобразовать эти уравнения к виду, где A_0 и B_0 — тождественные операторы. Однако исходная форма записи часто практически удобнее.

Может показаться, что построение таких нестационарных процессов, устанавливающихся к решению, уже полностью решает проблему отыскания минимума функции. Осталось «лишь» найти решение получившейся системы дифференциальных уравнений, используя какой-либо из численных методов решения задачи Коши.

В действительности сведение решения стационарной задачи к решению нестационарной не всегда дает удовлетворительное решение проблемы минимизации. Остается еще неясным существенный вопрос о выборе величины шагов численного интегрирования. Предположим, что решение нестационарной задачи устанавливается с требуемой точностью к решению стационарной за некоторый промежуток времени T. Если интегрирование производится с малым шагом Δ, то получаемые расчетные точки будут близки к рассматриваемой траектории и можно рассчитывать на попадание в малую окрестность точки минимума. Однако при этом число

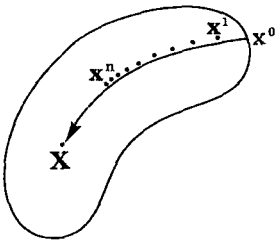


Рис. 7.5.2

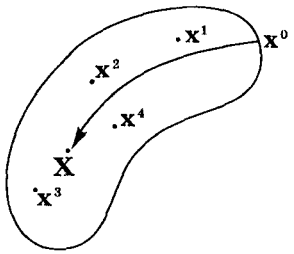


Рис. 7.5.3

шагов T/Δ может оказаться недопустимо большим (рис. 7.5.2). Если шаг интегрирования берется очень большим, то может случиться, что расчетные точки начнут сильно отклоняться от рассматриваемого решения и никогда не попадут в искомую окрестность точки минимума (рис. 7.5.3).

Метод установления применим не только к задачам на экстремум функционала, но и к любым стационарным задачам $F(\mathbf{x}) = 0$. Строится некоторый процесс вида (5) или (6), где вместо $\operatorname{grad}\Phi(\mathbf{x})$ стоит $F(\mathbf{x})$, и такой, что $\mathbf{x}(t) \to \mathbf{X}$ при $t \to \infty$, $\mathbf{X}$ — корень уравнения $F(\mathbf{X}) = 0$.

Рассмотрим подробнее случай системы линейных уравнений $A\mathbf{x} - \mathbf{b} = 0$ в предположении, что жорданова форма матрицы диагональная и все ее собственные значения λ_i лежат в пределах $0 < \mu \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_m \leqslant M$, где $\mathbf{e}_1, \ldots, \mathbf{e}_m$ — соответствующие собственные векторы, образующие полную систему.

Напишем простейшую аппроксимацию метода установления

$$\frac{d\mathbf{x}}{dt} + (A\mathbf{x} - \mathbf{b}) = \mathbf{0} \tag{7}$$

на временной сетке с постоянным шагом

$$\frac{\mathbf{x}^{n+1} - \mathbf{x}^n}{\tau} + (A\mathbf{x}^n - \mathbf{b}) = \mathbf{0}.$$

Соответствующая расчетная формула

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \tau(A\mathbf{x}^n - \mathbf{b})$$

совпадает с расчетной формулой из § 6.3. Погрешности $\mathbf{r}^n = \mathbf{x}^n - \mathbf{X}$ удовлетворяют соотношению $\mathbf{r}^{n+1} = (E - \tau A)\mathbf{r}^n$. При $\mathbf{r}^0 = \sum_1^m c_i\mathbf{e}_i$ имеем $\mathbf{r}^n = \sum_1^m c_i(1 - \tau\lambda_i)^n\mathbf{e}_i$ и скорость убывания погрешности определяется величиной $\max\limits_{\lambda_i}|1 - \tau\lambda_i|$. Там уже было установлено, что наиболее целесообразно брать $\tau = 2/(\mu + M)$, и тогда можно утверждать, что погрешность ведет себя как $O(((M - \mu)/(M + \mu))^n)$. Метод наискорейшего спуска можно также рассматривать как аппроксимацию метода устано-

вления, но уже на сетке с переменным шагом: $\mathbf{x}^{n+1} = \mathbf{x}^n - \tau_n(A\mathbf{x}^n - \mathbf{b})$; шаг τ_n определяется каждый раз из условия $\min\limits_{\tau_n} \Phi(\mathbf{x}^{n+1})$.

Рассмотрим аппроксимацию метода установления

$$\frac{d^2\mathbf{x}}{dt^2} + \gamma\frac{d\mathbf{x}}{dt} + (A\mathbf{x} - \mathbf{b}) = \mathbf{0}$$

на временной сетке с постоянным шагом

$$\frac{\mathbf{x}^{n+1} - 2\mathbf{x}^n + \mathbf{x}^{n-1}}{\tau^2} + \gamma\frac{\mathbf{x}^{n+1} - \mathbf{x}^{n-1}}{2\tau} + (A\mathbf{x}^n - \mathbf{b}) = \mathbf{0};$$

$0 < \gamma$ — скалярный множитель. Погрешности $\mathbf{r}^n$ удовлетворяют соотношениям

$$\frac{\mathbf{r}^{n+1} - 2\mathbf{r}^n + \mathbf{r}^{n-1}}{\tau^2} + \gamma\frac{\mathbf{r}^{n+1} - \mathbf{r}^{n-1}}{2\tau} + A\mathbf{r}^n = \mathbf{0}. \tag{8}$$

Разложим векторы $\mathbf{r}^n$ по собственным векторам матрицы A:

$$\mathbf{r}^n = \sum_{i=1}^{m} c_i^n \mathbf{e}_i.$$

Подставим выражения $\mathbf{r}^{n-1}$, $\mathbf{r}^n$, $\mathbf{r}^{n+1}$ в (8); поскольку векторы $\mathbf{e}_i$ независимы, то коэффициенты при них обращаются в нуль и получается система соотношений

$$\frac{c_i^{n+1} - 2c_i^n + c_i^{n-1}}{\tau^2} + \gamma\frac{c_i^{n+1} - c_i^{n-1}}{2\tau} + \lambda_i c_i^n = 0.$$

Решения этих разностных уравнений записываются в виде

$$c_i^n = C_i^1(z_1^i)^n + C_i^2(z_2^i)^n,$$

если z_1^i, z_2^i — простые корни характеристического уравнения

$$\frac{z^2 - 2z + 1}{\tau^2} + \gamma\frac{z^2 - 1}{2\tau} + \lambda_i z = 0, \tag{9}$$

и в виде

$$c_i^n = C_i^1(z_1^i)^n + C_i^2 n(z_1^i)^n,$$

если корень (9) кратный. Во всех случаях определяющим фактором убывания $\|\mathbf{r}^n\|$ является величина $\max\limits_i \left(\max(|z_1^i|, |z_2^i|)\right)$, которую можно мажорировать сверху величиной

$$\max_{\mu \leqslant \lambda \leqslant M} |z(\lambda)|, \tag{10}$$

где $z(\lambda)$ — максимальный по модулю корень уравнения

$$\frac{z^2 - 2z + 1}{\tau^2} + \gamma\frac{z^2 - 1}{2\tau} + \lambda z = 0.$$

Мы не будем приводить полного решения задачи на экстремум (10), а ограничимся наводящими соображениями и выписыванием ответа.

При $\mathbf{x}^1 = \mathbf{x}^0 + \alpha(A\mathbf{x}^0 - \mathbf{b})$ приближения $\mathbf{x}^n$ записываются в виде (6.6.2). Поэтому рассматриваемый итерационный процесс не может дать лучшего приближения, чем оптимальный линейный итерационный процесс (6.6.19). При $n \to \infty$ оптимальный линейный итерационный процесс переходит в итерационный процесс (6.6.23), который с учетом явного выражения $\lambda_0 = (1 + \sqrt{\mu/M})/(1 - \sqrt{\mu/M})$ может быть записан в виде

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \left(\frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}}\right)^2 (\mathbf{x}^n - \mathbf{x}^{n-1}) - \frac{4}{(\sqrt{M} + \sqrt{\mu})^2}(A\mathbf{x}^n - \mathbf{b}). \qquad (11)$$

Можно подобрать τ и γ так, чтобы итерационный процесс совпадал с рассматриваемым. Для этого следует взять

$$\tau = \frac{2}{\sqrt{M + \gamma}}, \quad \gamma = \frac{2\sqrt{M\mu}}{\sqrt{M} + \mu}.$$

Таким образом, в рамках схемы установления с постоянным γ можно получить итерационный процесс, несущественно отличающийся от оптимального линейного процесса.

Обратим внимание на поведение корней $z(\lambda)$, соответствующих (11); имеем характеристическое уравнение

$$z^2 - \left(\left(1 + \left(\frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}}\right)^2\right) - \frac{4}{(\sqrt{M} + \sqrt{\mu})^2}\lambda\right) z + \left(\frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}}\right)^2 = 0.$$

Запишем это уравнение в виде

$$z^2 - A(\lambda)z + \mu_0^2 = 0, \quad \mu_0 = \left(\sqrt{M} - \sqrt{\mu}\right) \Big/ \left(\sqrt{M} + \sqrt{\mu}\right).$$

Здесь $A(\lambda)$ — линейная функция от λ, причем $A(M) = -2\mu_0$, $A(\mu) = 2\mu_0$. Следовательно, $|A(\lambda)| < 2\mu_0$ при $\mu < \lambda < M$. Поэтому $z_{1,2}(\mu) = \mu_0$, $z_{1,2}(M) = -\mu_0$, $z_{1,2}(\lambda) = -\dfrac{A(\lambda)}{2} \pm \sqrt{\left(\dfrac{A(\lambda)}{2}\right)^2 - \mu_0^2}$ имеют ненулевую мнимую часть и $|z_{1,2}(\mu)| = \mu_0$ при $\mu < \lambda < M$.

Таким образом, указана совокупность коэффициента γ и шага τ, для которой

$$\max_{\mu \leqslant \lambda \leqslant M} |z(\lambda)| = (\sqrt{M} - \sqrt{\mu})/(\sqrt{M} + \sqrt{\mu}).$$

Неулучшаемость этой оценки усматривается из оценки скорости сходимости оптимального линейного итерационного процесса.

Как аналог метода сопряженных градиентов в случае минимизации функционала $\Phi(\mathbf{x})$ общего вида можно рассматривать следующий метод: последующие приближения ищутся в виде

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \alpha_n(\mathbf{x}^n - \mathbf{x}^{n-1}) + \beta_n \operatorname{grad}\Phi(\mathbf{x}^n),$$

α_n и β_n определяются из условия $\min\limits_{\alpha_n,\beta_n} \Phi(\mathbf{x}^n)$.

Соответствие между методами решения стационарных задач путем установления и обычными итерационными методами позволяет строить новые итерационные методы или новые процессы установления.

Рассмотрим, например, расчетные формулы Ньютона

$$\mathbf{x}^{n+1} - \mathbf{x}^n = -\left(F'(\mathbf{x}^n)\right)^{-1}\mathbf{F}(\mathbf{x}^n) \tag{12}$$

решения системы нелинейных уравнений $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Этим формулам можно дать следующую интерпретацию: введем непрерывное время t и будем рассматривать величины $\mathbf{x}^n$ как значения некоторой функции в моменты времени $t_n = n$. Тогда соотношение (12), переписанное в виде

$$\frac{x^{n+1} - x^n}{t_{n+1} - t_n} = -\left(F'(\mathbf{x}^n)\right)^{-1}\mathbf{F}(\mathbf{x}^n),$$

можно интерпретировать как получившееся при аппроксимации системы

$$\frac{d\mathbf{x}}{dt} = -\left(F'(\mathbf{x})\right)^{-1}\mathbf{F}(\mathbf{x}). \tag{13}$$

Решение исходной задачи является стационарной точкой этой системы. При $\mathbf{x} = \mathbf{X} + \boldsymbol{\eta}$ имеем

$$\left(F'(\mathbf{x})\right)^{-1} = \left(F'(\mathbf{X})\right)^{-1} + O(\|\boldsymbol{\eta}\|),$$

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{X}) + F'(\mathbf{x})\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2) = F'(\mathbf{X})\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2).$$

Таким образом,

$$\frac{d\boldsymbol{\eta}}{dt} = \frac{d\mathbf{x}}{dt} = -((F'(\mathbf{X}))^{-1} + O(\|\boldsymbol{\eta}\|)(F'(\mathbf{X})\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|)^2)) =$$
$$= -\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|)^2).$$

Отсюда следует, что $\mathbf{x} = \mathbf{X}$ является асимптотически устойчивым решением (13).

Заменяя производную $d\mathbf{x}/dt$ разностным отношением по значениям функции в некоторых точках $t_0 = 0$, $t_1, \ldots$, получаем соотношения

$$\frac{\mathbf{x}^{n+1} - \mathbf{x}^n}{\Delta_N} = -\left(F'(\mathbf{x}^n)\right)^{-1}\mathbf{F}(\mathbf{x}^n), \qquad \Delta = t_{n+1} - t_n,$$

иначе,

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n(F'(\mathbf{x}^n))^{-1}\mathbf{F}(\mathbf{x}^n). \tag{14}$$

Таким образом, построение нестационарного процесса, соответствующего методу Ньютона, привело нас к получению итерационного процесса (14) более общего вида.

Рассмотренный пример показывает, что переход от некоторого итерационного алгоритма к соответствующему ему нестационарному процессу имеет много общих черт с построением замыкания вычисленного алгоритма (понятие замыкания алгоритма будет введено в гл. 9). Нами была доказана сходимость метода Ньютона лишь при достаточно хорошем начальном приближении к решению. С целью расширения области сходимости иногда прибегают к следующей модификации метода Ньютона.

Задаются функционалом $\Phi(\mathbf{x})$, например $\|\mathbf{F}(\mathbf{x})\|^2$, нижняя грань которого, равная нулю, достигается при решении задачи. Последовательные приближения отыскиваются в виде (14), причем Δ_n определяется из условия

$$\min_{\Delta_n} \Phi\left(\mathbf{x}^n - \Delta_n (F'(\mathbf{x}^n))^{-1}\mathbf{F}(\mathbf{x}^n)\right).$$

Возникает вопрос о практическом нахождении величины Δ_n, при которой достигается этот минимум.

Одна из распространенных процедур определения Δ_n состоит в следующем. Задавая $\lambda \in (0,1)$, $0 \in (0,1]$ и $l > 0$ — целое, при каждом n последовательно вычисляют

$$\mathbf{x}^{n+1,i} = \mathbf{x}^n - \lambda^i (F'(\mathbf{x}^n))^{-1}\mathbf{F}(\mathbf{x}^n), \quad i = 0, \ldots, l,$$

и $q^{n+1,i} = \Phi(x^{n+1,i})$. Если при некотором $k \leqslant l$ оказалось $q^{n+1,k} \leqslant \theta\Phi(\mathbf{x}^n)$, то вычисления прекращают и полагают $\mathbf{x}^{n+1} = \mathbf{x}^{n+1,k}$; в других процедурах находят $\min\limits_{0<i\leqslant l} q^{n+1,i} = q^{n+1,m}$ и полагают $\mathbf{x}^{n+1} = \mathbf{x}^{n+1,m}$.

Если

$$q^{n+1,0}, \ldots, q^{n+1,l} > \theta\Phi(\mathbf{x}^n),$$

то возможны, например, такие варианты:

1) временный переход к другому методу;
2) остановка;
3) изменение значений параметров l, λ, θ.

Отметим в заключение, что методы установления могут применяться и в случае минимизации функции в областях с ограничениями. Тогда уравнения (1), (3) следует дополнить какими-то уравнениями, которым будет подчиняться траектория точки, попавшей на границу.

§ 6. Как оптимизировать?

После выбора модели, целевой функции и параметризации задачи возникает задача минимизации функции обычно большого числа переменных в области, принадлежность к которой задается условием выполнения большого числа ограничений — равенств или неравенств. Наличие ограниче-

ний существенно увеличивает сложность задачи минимизации: как правило, точкой экстремума оказывается некоторая граничная точка области.

Вследствие большой важности решения задач оптимизации для самых различных сторон жизни общества, в настоящее время накопился большой багаж методов и стандартных программ решения задач оптимизации. Поэтому при решении единичной конкретной задачи часто наиболее оправдано обращение к одной из имеющихся стандартных программ.

При решении новых задач оптимизации полезно помнить о следующих моментах. Для рассматриваемых в этой главе методов сходимость приближений всегда доказывалась в предположении наличия достаточно хорошего начального приближения. Это ограничение на метод вызывается существом дела. Так, поверхность уровня многочлена умеренной степени от умеренного числа переменных может содержать весьма много не связанных между собой компонент со сложным взаимным расположением. Поэтому, например, безнадежно рассчитывать на построение алгоритма, позволяющего быстро найти минимум любого многочлена восьмой степени от десяти переменных.

Вышесказанное следует учитывать при использовании стандартных программ, в описании которых указывается на возможность минимизации функций очень большого числа переменных. В лучшем случае оказывается, что:

а) программа эффективно решает задачи минимизации из класса, с которым автор программы обычно имеет дело, или

б) программа, формально говоря, может решить любую задачу минимизации, но требуемое для решения время в подавляющем числе случаев выходит за всякие разумные пределы.

Мы придерживаемся точки зрения, что *всякий «универсальный» метод решения многомерных задач минимизации должен обладать существенными недостатками и в действительности не является универсальным.*

При решении новых задач зачастую приходится разрабатывать специальные, приспособленные именно для этого класса задач, методы отыскания начального приближения к решению. Сложность отыскания приемлемого начального приближения можно проиллюстрировать следующим примером.

Существует ряд стандартных программ наилучшего приближения функций отношением многочленов

$$P_m(x)/Q_n(x) \tag{1}$$

заданных степеней m и n. При разработке стандартной программы вычисления интегралов Френеля проводились следующие эксперименты. Каждая из стандартных программ применялась для приближения рассматриваемой функции на некотором отрезке при всех m, n в пределах $0 < m + n \leqslant 12$. Оказалось, что не менее чем в 3 случаях из 90 возможных каждая программа выдавала ответ, что она не может решить рассматриваемую задачу. У некоторых программ доля таких отказов превосходила половину выданных ответов. В то же время во всех этих про-

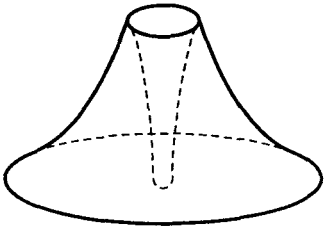


Рис. 7.6.1

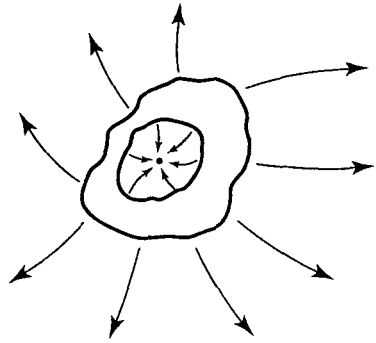


Рис. 7.6.2

граммах используются алгоритмы, для которых доказаны теоремы о сходимости при достаточно хорошем начальном приближении.

Чтобы не создалось впечатления полной безнадежности решения сколько-нибудь сложных задач оптимизации, рассмотрим тот же вопрос с оптимистической точки зрения.

С этих позиций приведенные выше доводы о сложности минимизации многочленов можно рассматривать как малоубедительные — ведь на самом деле нам никогда не потребуется минимизировать произвольный многочлен. Существует мнение, что задачи минимизации функций с очень сложной структурой линий уровня встречаются довольно редко.

Рассмотрим пример задачи, где сам факт необходимости ее решения может быть поставлен под сомнение.

Неудачи при попытках получения хорошего начального приближения могут быть вызваны следующей особенностью поведения рассматриваемой функции. Точка минимума находится в очень узкой «яме» — при удалении от нее во всех направлениях функция резко возрастает, а потом начинает убывать (рис. 7.6.1).

Предположим, что точка минимума находится путем продвижения в направлении, противоположном градиенту функции $\Phi(x)$, иначе — численным интегрированием системы

$$\frac{dx}{dt} + \operatorname{grad} \Phi(x) = 0. \tag{2}$$

Тогда множество начальных условий, исходя из которых мы будем приходить в точку минимума, находится в небольшой окрестности этой точки. Можно говорить, что при рассмотрении окрестности точки минимума в микроскопическом масштабе эта точка оказывается точкой притяжения решения системы (2); при рассмотрении в более крупном масштабе она уже оказывается точкой отталкивания (рис. 7.6.2).

Точно такой же характер поведения последовательных приближений к точке минимума будет и у других итерационных методов.

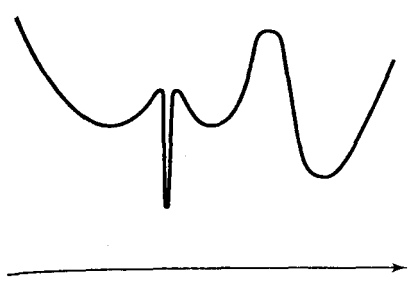


Рис. 7.6.3

Если «яма», где находится точка минимума, очень узкая, то иногда и не стоит искать эту точку минимума. В самом деле, пусть, например, параметры x_i отвечают некоторой реальной работающей системе управления. В работе этой системы неизбежны некоторые сбои, т. е. изменения этих параметров. Если точка минимума функции Ф находится в такой узкой «яме», то малые сбои могут существенно испортить характеристики работы системы. В свете вышесказанного выбор точки экстремума в случае, изображенном на рис. 7.6.3, требует дополнительного изучения.

Некоторые исследователи, обладающие большим опытом решения практических задач оптимизации, утверждают, что подобные целевые функции с узкими «ямами» возникают обычно в случаях, когда математическая модель рассматриваемого явления построена неудачно.

Более существенные причины для оптимизма состоят в следующем. Во многих случаях создание математической модели и ее оптимизация зачастую имеют целью улучшение работы уже существующей системы. В этих случаях параметры реальной системы часто являются хорошим приближением для дальнейшей оптимизации.

При разработке новой системы типична следующая линия поведения. Сначала строится и оптимизируется простейшая модель, учитывающая важнейшие факторы. Затем модель постепенно усложняется за счет учета все новых и новых факторов. Таким образом, последовательно возникают задачи минимизации функций все большего числа параметров. При удачном построении вспомогательных моделей решение каждой из этих задач оптимизации обычно оказывается хорошим начальным приближением для следующей по сложности задачи.

Это обстоятельство часто используется следующим образом. Пусть перед нами стоит задача минимизации функции большого числа параметров $x_1, \ldots, x_n$. Построим функцию меньшего числа параметров $X_1, \ldots, X_m$, $m < n$, приближающую рассматриваемую функцию, иначе — упрощенную модель с определяющими параметрами $X_1, \ldots, X_m$. Произведем оптимизацию этой функции (модели) и на основе ее решения сконструируем начальное приближение. Иногда оказывается полезным произвести несколько шагов такого упрощения функции (модели) и введения новых параметров.

Оптимизация функций (моделей) меньшего числа параметров оказывается более легкой по следующим двум причинам: становится проще структура линий уровня минимизируемой функции; вычисление каждого значения функции обычно требует меньшего объема вычислений.

При построении упрощенных моделей в первую очередь следует учитывать наиболее *важные параметры* задачи.

Что такое важность факторов или параметров? Можно говорить, что важными параметрами являются те, от которых функция сильно зависит. На языке математики это означает, что производные функции по этим параметрам относительно велики. Второстепенными параметрами являются те, от которых рассматриваемая функция зависит слабо, т. е. производные по которым малы. Формально важность параметров можно определить, оценивая производные рассматриваемой функции. Однако для сложных задач такая оценка и особенно математически обоснованный выбор новых параметров $X_1, \ldots, X_m$ весьма трудоемки. Характеристика параметров по более наглядному критерию — их важности — дает возможность руководителю производственной системы подсказать математику первоочередность выбора параметров.

Иногда приемлемый метод оптимизации или способ отыскания хорошего начального приближения можно получить, изучая принципы, которыми руководствуются в своей работе опытный практический работник или руководитель, или приемы их работы.

На одном новом заводе долгое время не удавалось наладить ритмичное производство из-за недостаточного опыта работы операторов. Попытки создания модели производственного процесса, достаточно точной, но в то же время поддающейся анализу средствами математики с помощью ЭВМ, не приводили к успеху. В конце концов пришлось временно отказаться от разработки математической модели и пойти по следующему пути автоматизации и оптимизации производства. В память ЭВМ были записаны режимы работы лучших операторов на родственных предприятиях. Далее, в зависимости от имеющихся на данный момент условий, машина выбирала режим работы, наиболее близкий к режиму работы одного из лучших операторов. Такое мероприятие позволило устранить возникшие трудности.

В другой аналогичной ситуации руководство предприятием не пошло по такому пути. Оно настойчиво требовало от математической группы разработки универсального алгоритма, который по заданным внешним характеристикам конструируемого прибора выдавал бы оптимальный набор внутренних параметров прибора: расположение и размеры деталей конструкции, вес и т. п. Предлагавшиеся математиками алгоритмы оптимизации не оказывались универсальными и в большинстве случаев не приводили к приемлемому решению. Математики предложили подход к решению задачи, имитирующий реальную ситуацию. Конструктор задает компоновку деталей прибора. Компьютер обсчитывает внешние характеристики прибора и выдает их конструктору. На основании полученной информации конструктор вносит изменения в компоновку. Такой диалоговый режим работы позволил бы отказаться от дорогостоящего реального конструирования прибора и его испытаний. Руководство предприятия отказалось от такого подхода к решению проблемы и потратило много времени на бесплодные поиски «более квалифицированных» математиков, способных предложить «универсальный» бездиалоговый алгоритм решения казавшейся ему столь простой задачи. Большие материальные затраты и потеря темпов в разработке новой техники в конце концов привели к понима-

нию того, что предлагавшийся математиками путь решения задачи на данном этапе понимания проблемы был единственно возможным.

Конструкторы приобрели опыт в таком режиме работы. Математикам, анализируя диалог конструкторов с ЭВМ, удалось понять принципы, которыми конструкторы руководствуются при компоновке деталей, и заложив эти принципы в основу алгоритма решения задачи, создать бездиалоговый, «чисто машинный» алгоритм оптимизации конструкции.

Способы нахождения начального приближения и сами итерационные методы часто имеют аналогию с какими-то реальными процессами и в других явлениях. Работа зрительного аппарата и мозга при отыскании какого-либо предмета, по-видимому, организована по следующей схеме. Сначала производится беглый осмотр всего поля зрения в крупном масштабе, на основании полученной информации выбирается участок для дальнейшего просмотра, затем производится просмотр этого участка в крупном масштабе и т. д. Обратим внимание на сходство с методами из § 3.16 и § 4 данной главы.

При необходимости развивать исследования новых трудных задач важное значение имеет правильная организация научных исследований. Если в основном ясно, в каком направлении нужно развивать исследования, концентрация научных усилий обычно производится в этом направлении. Если наиболее рациональный путь к намеченной цели еще не определился, то часто прибегают к дублированию исследований. Несколько независимых организаций ищут решение, каждая на своем пути, иногда без постоянного обмена информацией. Хотя на первый взгляд кажется, что взаимный обмен информацией всегда полезен, постоянный обмен ею может и помешать возникновению и продвижению оригинальных решений проблемы.

Существует гипотеза (необщепризнанная), что сходным образом работает мозг при решении какой-либо проблемы: получаемая информация фиксируется не вполне детерминированным образом в различных его участках; в то же время в каждый момент работы над проблемой эта информация извлекается лишь из локализованного участка мозга.

По аналогии со сказанным выше напрашивается следующий подход к решению задачи оптимизации (а также и любых других задач) в случае, если требуется срочное получение результата. Для решения задачи поочередно или независимо используются несколько известных методов решения подобных задач. В обоих случаях, при поочередном или при независимом использовании, алгоритмы A_p, $p = 1, \ldots, l$, работают циклически; длительность t_{pq} промежутка времени, в течение которого алгоритм A_p работает q-й раз, задается пользователем или определяется в процессе работы. При поочередном использовании каждый алгоритм начинает минимизацию с приближения, полученного предшествующим алгоритмом. При независимом использовании каждый алгоритм начинает минимизацию с приближения, полученного в результате предшествующе-

го применения данного алгоритма. Таким образом, в этом случае алгоритмы работают по принципу «кто быстрее».

Режим независимого использования алгоритмов является аналогом параллельной работы организаций при отсутствии взаимного обмена информацией. Как аналог реальной организации научных исследований с обменом информацией в дискретные моменты времени может рассматриваться следующий режим работы. В начале каждого промежутка времени t_{pq} алгоритм A_p просматривает некоторую совокупность приближений из полученных всеми алгоритмами и выбирает наилучшее приближение исходя из своих позиций. Например, он может просматривать все приближения последнего цикла или приближения, полученные алгоритмами в концах всех промежутков t_{ij}.

Иногда может принести пользу следующая организация работы: после получения алгоритмом A_p приближения, очень хорошего с позиций алгоритма A_s, предоставляется время для работы алгоритма A_s.

Конечно, не следует думать, что непосредственное копирование различных реальных систем всегда позволит наилучшим образом решить рассматриваемую оптимизационную задачу.

Подведем общий итог наших рассуждений. Обычно задачи оптимизации функций большого числа переменных очень трудны; при решении новых задач приходится затрачивать много, иногда бесплодных, усилий, производя пробные просчеты по различным известным и новым алгоритмам. Однако при наличии таких благоприятных факторов, как контакт с практическими работниками и возможность анализа упрощенных моделей, есть все основания сохранять уверенность в благоприятном исходе попыток решения задачи.

Заметим, что часто для успешного решения задач оптимизации необходим диалоговый режим работы исследователя с ЭВМ.

Не имея перед собой конкретной задачи, невозможно дать рекомендацию, каким методом решения системы нелинейных уравнений или минимизации функций следует воспользоваться. Как уже отмечалось выше, велика возможность столкнуться с ситуацией, когда *область сходимости метода* (множество значений нулевого приближения, при которых метод сходится) очень мала.

Опыт решения подобных задач показывает, что в первую очередь стоит попробовать применить методы, имеющие естественную наглядную интерпретацию, например метод установления, или методы, имитирующие действие человека или животного в подобной ситуации. Для выбора начального приближения надо также привлечь естественные наглядные соображения, имитирующие такие действия. На таком пути часто удается довольно быстро построить алгоритм, позволяющий решить задачу.

При однократном решении простых задач иногда проще всего применить простейший итерационный метод, например Ньютона, с выдачей на экран хода итерационного процесса. Задаваясь различными началь-

ными приближениями, часто удается довольно быстро угадать начальное приближение, лежащее в области сходимости метода.

При решении систем уравнений или задач минимизации, возникающих при аппроксимации краевых задач для дифференциальных уравнений, полезно воспользоваться близостью (в соответствующих нормах) решений таких дискретных задач, соответствующих различным шагам сетки. Решение на крупной сетке является хорошим приближением для решения задачи, соответствующей более мелкой сетке. В то же время каждый шаг итерации на крупной сетке менее трудоемок, и с теми же затратами можно провести большее число шагов итерации, начиная с одного или со многих начальных приближений. Таким образом, в этом случае имеет смысл решать задачу на последовательности сеток, т. е. последовательно решать несколько систем алгебраических уравнений (порядок системы m_j возрастает: $m_{j+1} > m_j$). При этом решение j-й системы $\mathbf{X}_j$ используется для получения начального приближения $\mathbf{X}^0_{j+1}$ к решению $(j+1)$-й системы.

Так как векторы $\mathbf{X}_j$ и $\mathbf{X}^0_{j+1}$ имеют, вообще говоря, различную размерность, то для перехода от $\mathbf{X}_j$ к $\mathbf{X}^0_{j+1}$ обычно используют интерполяцию многочленами или сплайнами. Аналогичный подход применим и к другим дискретным задачам, возникающим при аппроксимации задач, связанных с отысканием функций непрерывного аргумента.

В ряде случаев, например при планировании, требуется многократно решать однотипные задачи, причем в режиме реального времени, т. е. решение задачи должно получаться с очень малым отставанием от изменения входных данных задачи. В этом случае для ускорения сходимости следует использовать все многообразие описанных нами методов.

Литература

1. Васильев Ф. П. Численные методы решения экстремальных задач. — М.: Наука, 1980.
2. Васильев Ф. П. Методы решения экстремальных задач. — М.: Наука, 1981.
3. Карманов В. Г. Математическое программирование. — М.: Наука, 1986.
4. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника, 1982.
5. Нестеров Ю.Е. Эффективные методы в нелинейном программировании. — М.: Радио и связь, 1989.
6. Ортега Д., Рейнболдт В. Итерационные методы решения систем уравнений со многими неизвестными. — М.: Мир, 1975.

Глава 8

Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений



Задача решения обыкновенных дифференциальных уравнений сложнее задачи вычисления однократных интегралов, и доля задач, интегрируемых в явном виде, здесь существенно меньше.

Когда говорят об интегрируемости в явном виде, имеют в виду, что решение может быть вычислено при помощи конечного числа «элементарных» операций: сложения, вычитания, умножения, деления, возведения в степень, логарифмирования, потенцирования, вычисления синуса и косинуса и т. п. Уже в период, предшествовавший появлению ЭВМ, понятия «элементарной» операции претерпели изменение. Решения некоторых частных задач настолько часто встречаются в приложениях, что пришлось составить таблицы их значений, в частности таблицы интегралов Френеля, функций Бесселя и ряда других, так называемых *специальных функций*. При наличии таких таблиц исчезает принципиальная разница между вычислением функций $\sin x$, $\ln x$, ... и специальных функций. В том и другом случаях можно вычислять значения этих функций при помощи таблицы, и те и другие функции можно вычислять, приближая их многочленами, рациональными дробями и т. д. Таким образом, в класс задач, интегрируемых в явном виде, включились задачи, решения которых выражаются через специальные функции. Однако и этот, более широкий, класс составляет относительно малую долю задач, предъявляемых к решению. Существенное расширение класса реально решаемых дифференциальных уравнений, а следовательно, и расширение сферы применения математики произошло с разработкой численных методов и активным повсеместным использованием ЭВМ.

В настоящее время затраты человеческого труда при решении на ЭВМ задачи Коши для обыкновенных дифференциальных уравнений сравнимы с затратами на то, чтобы просто переписать заново формулировку этой задачи. При желании можно получить график решения или его изображение на экране. В результате этого для многих категорий научных работников существенно уменьшился интерес к изучению частных способов интегрирования обыкновенных дифференциальных уравнений в явном виде.

Эта глава посвящена описанию основных методов решения задачи Коши для обыкновенных дифференциальных уравнений, исследованию свойств этих методов и оценке их погрешности.

Обратим внимание на то обстоятельство, что, как и в других случаях, первоначальный анализ практической пригодности методов и отбрасывание непригодных методов часто удается произвести, изучая простейшие задачи, где точное и приближенное решения задачи выписываются в явном виде.

§ 1. Решение задачи Коши с помощью формулы Тейлора

Один из простейших по своему описанию методов решения задачи Коши основан на использовании формулы Тейлора.

Пусть требуется найти на отрезке $[x_0, x_0 + X]$ решение дифференциального уравнения

$$y' = f(x, y) \tag{1}$$

при начальном условии $y(x_0) = y_0$; $f(x, y)$ — функция, аналитическая в точке (x_0, y_0). Дифференцируя (1) по x, имеем соотношения

$$y'' = f_x(x, y) + f_y(x, y)y',$$

$$y''' = f_{xx}(x, y) + 2f_{xy}(x, y)y' + f_{yy}(x, y)y'^2 + f_y(x, y)y'', \ldots.$$

Подставляя $x = x_0$ и $y = y_0$ в (1) и в последние соотношения, последовательно получаем значения

$$y'(x_0),\ y''(x_0),\ y'''(x_0), \ldots;$$

Таким образом, можно написать приближенное равенство

$$y(x) \approx \sum_{i=0}^{n} \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i. \tag{2}$$

Если значение $|x - x_0|$ больше радиуса сходимости ряда

$$\sum_i \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i,$$

то погрешность (2) не стремится к нулю при $n \to \infty$ и предлагаемый метод неприменим.

Иногда целесообразно поступить следующим образом. Разобьем отрезок $[x_0, x_0 + X]$ на отрезки $[x_{j-1}, x_j]$, $j = 1, \ldots, N$. Будем последовательно получать приближения y_i к значениям решения $y(x_j)$, $j = 1, \ldots, N$, по следующему правилу. Пусть значение y_i уже найдено, вычисляем значения в точке x_j производных $y_j^{(i)}$ решения исходного дифференциального

уравнения, проходящего через точку (x_j, y_j). На отрезке $[x_j, x_{j+1}]$ полагаем

$$y(x) \approx z_j(x) = \sum_{j=0}^{n} \frac{y_j^{(i)}}{i!}(x - x_j)^i \tag{3}$$

и соответственно берем

$$y_{j+1} = z_j(x_{j+1}). \tag{4}$$

Рассмотрим случай, когда $x_{j+1} - x_j \equiv h$. Если бы значение y_j совпадало со значением точного значения $y(x_j)$, то погрешность от замены y_{j+1} на $z_i(x_{j+1})$ имела бы порядок $O(h^{n+1})$. Поскольку мы вносим погрешность на $O(h^{-1})$ отрезках, то можно ожидать, что при уменьшении шага сетки будет выполняться соотношение

$$\max_{0 \leqslant j \leqslant N} |y_j - y(x_j)| = O(h^n).$$

В ряде случаев такого рода рассуждения приводят к неправильному заключению о наличии факта сходимости приближенного решения к точному, в то время как в действительности этого нет. Поэтому строгое обоснование сходимости методов при уменьшении шага, а также получение оценки погрешности имеет не только теоретическое, но и важнейшее практическое значение.

При использовании этого метода нужно вычислять значения функции f и всех ее производных $f_{x^i y^{m-j}}$ при $m < n$, т. е. вычислять $n(n+1)/2$ значений различных функций. Это требует написания большого числа блоков вычисления производных, что противоречит основной тенденции упрощения отношений между пользователем и ЭВМ.

В настоящее время на некоторых ЭВМ имеются пакеты программ, которые по заданной программе вычисления значений функции строят программу вычисления значений ее производных. Таким образом, при наличии таких пакетов программ, казалось бы, отпадает приведенное выше возражение о сложности использования описанного ранее метода.

Однако этот метод применяется редко. Как правило, программы, создаваемые с помощью таких пакетов, при той же точности результата требуют существенно больших затрат машинного времени, чем программы, основанные на рассматриваемых далее более простых методах типа Рунге-Кутта и Адамса.

В то же время описанный выше алгоритм может быть полезен. Например, при расчетах траекторий движения небесных тел приходится многократно интегрировать системы дифференциальных уравнений вполне определенного вида при различных начальных условиях и различных значениях параметров в правых частях. То обстоятельство, что все время решается одна и та же система дифференциальных уравнений, дает следующее преимущество: конкретные формулы для производных правых частей системы имеют много общего; одновременное вычисление всех этих производных требует относительно малого числа арифметических операций, и рассматриваемый метод иногда оказывается эффективнее других методов численного интегрирования.

§ 2. Методы Рунге—Кутта

В частном случае $n = 1$ формула (1.3) имеет вид

$$y_{j+1} = y_j + hf(x_j, y_j), \quad h = x_{j+1} - x_j. \tag{1}$$

Этот метод называется *методом Эйлера*. Можно построить другой класс расчетных формул, к которому принадлежит метод Эйлера. Укажем сначала простейшие методы этого класса, получаемые из наглядных соображений. Пусть известно значение $y(x)$ и требуется вычислить значение $y(x + h)$. Рассмотрим равенство

$$y(x + h) = y(x) + \int_0^h y'(x + t)\, dt. \tag{2}$$

При замене интеграла в правой части на величину $hy'(x)$ погрешность имеет порядок $O(h^2)$, т. е.

$$y(x + h) = y(x) + hy'(x) + O(h^2).$$

Поскольку $y'(x) = f(x, y(x))$, отсюда имеем

$$y(x + h) = y(x) + hf(x, y(x)) + O(h^2).$$

Отбрасывая член порядка $O(h^2)$ и обозначая $x = x_j$, $x + h = x_{j+1}$, получим расчетную формулу Эйлера (1). Для получения более точной расчетной формулы нужно точнее аппроксимировать интеграл в правой части (2). Воспользовавшись квадратурной формулой трапеции, получим

$$y(x + h) = y(x) + \frac{h}{2}\Big(y'(x) + y'(x + h)\Big) + O(h^3),$$

иначе,

$$y(x + h) = y(x) + \frac{h}{2}\Big(f(x, y(x)) + f(x + h, y(x + h))\Big) + O(h^3), \tag{3}$$

соответствующая расчетная формула

$$y_{j+1} = y_j + \frac{h}{2}\Big(f(x_j, y_j) + f(x_{j+1}, y_{j+1})\Big) + O(h^3) \tag{4}$$

называется *неявной формулой Адамса второго порядка точности*. В некоторых случаях, в частности, когда f линейна по y, это уравнение может быть разрешено относительно y_{j+1}. Обычно же это уравнение неразрешимо явно относительно y_{j+1}, поэтому произведем дальнейшее преобразование алгоритма.

Заменим $y(x + h)$ в правой части (3) на некоторую величину

$$y^* = y(x + h) + O(h^2). \tag{5}$$

Тогда правая часть изменится на величину

$$\frac{h}{2}\Big(f(x + h, y^*) - f(x + h, y(x + h))\Big) = \frac{h}{2} f_y(x + h, \bar{y})\Big(y^* - y(x + h)\Big),$$

где $\bar{y}$ находится между y^* и $y(x+h)$. Вследствие предположения (5) эта величина имеет порядок $O(h^3)$. Таким образом, при условии (5) имеет место соотношение

$$y(x+h) = y(x) + \frac{h}{2}\Big(f(x, y(x)) + f(x+h, y^*)\Big) + O(h^3).$$

Условию (5) удовлетворяет результат вычислений по формуле Эйлера

$$y^* = y(x) + hf(x, y(x)).$$

Последние соотношения определяют пару расчетных формул

$$\begin{aligned} y^*_{j+1} &= y_j + hf(x_j, y_j), \\ y_{j+1} &= y_j + \frac{h}{2}\Big(f(x_j, y_j) + f(x_{j+1}, y^*_{j+1})\Big). \end{aligned} \tag{6}$$

При малых h выражение в правой части (4) удовлетворяет условию сжимаемости (§ 7.1), поэтому уравнение (4) также можно решать методом простой итерации:

$$y^{k+1}_{j+1} = y_j + \frac{h}{2}\Big(f(x_j, y_j) + f(x_{j+1}, y^k_{j+1})\Big).$$

Если y^0_{j+1} вычисляется по методу Эйлера:

$$y^0_{j+1} = y_j + hf(x_j, y_j),$$

то y^1_{j+1}, получаемое на первом шаге итерации, совпадает с y_{j+1}, получаемом по формуле (6). Дальнейшие итерации не приводят к повышению порядка точности по h; в то же время иногда главный член погрешности уменьшается при переходе от y^1_{j+1} к y^2_{j+1}. Если такое уменьшение погрешности компенсирует возрастание вычислительных затрат на шаге, то оно целесообразно.

Можно предложить теоретически обоснованный критерий, позволяющий при малых h выбирать каждый раз наиболее целесообразное число итераций. Однако его использование требует очень большого объема дополнительных вычислений. Поэтому выбор между числом итераций, равным 1 или 2, обычно осуществляется на основе предшествующего опыта, вычислительного эксперимента или просто «волевым» образом.

Построим другую пару формул с погрешностью на шаге такого же порядка. Интеграл в правой части (2) заменим по формуле прямоугольников:

$$y(x+h) = y(x) + hy'\left(x + \frac{h}{2}\right) + O(h^3),$$

или, что все равно,

$$y(x+h) = y(x) + hf\left(x + \frac{h}{2}, y\left(x + \frac{h}{2}\right)\right) + O(h^3).$$

Если

$$y^* = y\left(x + \frac{h}{2}\right) + O(h^2),$$

то, как и в предшествующем случае, имеем

$$y(x+h) = y(x) + hf\left(x + \frac{h}{2}, y^*\right) + O(h^3).$$

В качестве y^* можно взять результат вычислений по формуле Эйлера с шагом $\frac{h}{2}$: $y^* = y(x) + \frac{h}{2} f(x, y(x))$. Этим соотношениям соответствует пара расчетных формул

$$\begin{aligned} y_{j+1/2} &= y_j + \frac{h}{2} f(x_j, y_j), \\ y_{j+1} &= y_j + hf\left(x_j + \frac{h}{2}, y_{j+1/2}\right). \end{aligned} \tag{7}$$

Полученные методы относятся к семейству *методов Рунге—Кутта*, имеющих следующий вид. В процессе вычислений фиксированы некоторые числа

$$\alpha_2, \ldots, \alpha_q, \quad p_1, \ldots, p_q, \quad \beta_{ij}, \quad 0 < j < i \leqslant q;$$

последовательно получаем

$$\begin{aligned} k_1(h) &= hf(x, y), \\ k_2(h) &= hf\left(x + \alpha_2 h, y + \beta_{21} k_1(h)\right), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ k_q(h) &= hf\left(x + \alpha_q h, y + \beta_{q,1} k_1(h) + \cdots + \beta_{q,q-1} k_{q-1}(h)\right) \end{aligned}$$

и полагаем

$$y(x+h) \approx z(h) = y(x) + \sum_{i=1}^{q} p_i k_i(h).$$

Рассмотрим вопрос о выборе параметров α_i, p_i, β_{ij}. Обозначим $\varphi(h) = y(x+h) - z(h)$. Если $f(x, y)$ — достаточно гладкая функция своих аргументов, то $k_1(h)$, $\ldots$, $k_q(h)$ и $\varphi(h)$ — гладкие функции параметра h. Предположим, что $f(x, y)$ настолько гладкая, что существуют производные $\varphi'(h), \ldots, \varphi^{(s+1)}(h)$, а α_i, p_i, β_{ij} выбраны так, что $\varphi'(0) = \cdots = \varphi^{(s)}(0) = 0$. Кроме того, предположим, что существует некоторая гладкая функция $f_0(x, y)$, для которой соответствующее значение $\varphi^{(s+1)}(0) \neq 0$. Согласно формуле Тейлора выполняется равенство

$$\varphi(h) = \sum_{i=0}^{s} \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} = \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1}, \tag{8}$$

где $0 < \theta < 1$. Величина $\varphi(h)$ называется погрешностью метода на шаге, а s — порядком погрешности метода. При $q = 1$ имеем

$$\varphi(h) = y(x+h) - y(x) - p_1 h f(x, y), \quad \varphi(0) = 0,$$

$$\varphi'(0) = (y'(x+h) - p_1 f(x, y))\big|_{h=0} = f(x, y)(1 - p_1),$$

$$\varphi''(h) = y''(x+h);$$

здесь и далее $y = y(x)$. Равенство $\varphi'(0) = 0$ выполняется для всех гладких функций $f(x, y)$ лишь в случае $p_1 = 1$. Этому значению p_1 соответствует метод Эйлера. Для погрешности этого метода на шаге, согласно (8), получаем выражение

$$\varphi(h) = \frac{y''(x + \theta h)h^2}{2}.$$

Рассмотрим случай $q = 2$. Имеем

$$\varphi(h) = y(x+h) - y(x) - p_1 h f(x, y) - p_2 h f(\bar{x}, \bar{y}),$$

где $\bar{x} = x + \alpha_2 h$, $\bar{y} = \beta_{21} h f(x, y)$.

Вычислим производные функции $\varphi(h)$:

$$\varphi'(h) = y'(x+h) - p_1 f(x, y) - p_2 f(\bar{x}, \bar{y}) - p_2 h\big(\alpha_2 f_x(\bar{x}, \bar{y}) + \beta_{21} f_y(\bar{x}, \bar{y}) f(x, y)\big),$$

$$\varphi''(h) = y''(x+h) - 2p_2\big(\alpha_2 f_x(\bar{x}, \bar{y}) + \beta_{21} f_y(\bar{x}, \bar{y}) f(x, y)\big) - p_2 h\big(\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) + 2\alpha_2\beta_{21} f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_{21}^2 f_{yy}(\bar{x}, \bar{y})(f(x, y))^2\big),$$

$$\varphi'''(h) = y'''(x+h) - 3p_2\big(\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) + 2\alpha_2\beta_{21} f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_{21}^2 f_{yy}(\bar{x}, \bar{y})(f(x, y))^2\big) + O(h).$$

Согласно исходному дифференциальному уравнению

$$y' = f, \quad y'' = f_x + f_y f, \quad y''' = f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_y y''.$$

Подставим в выражения $\varphi(h)$, $\varphi'(h)$, $\varphi''(h)$, $\varphi'''(h)$ значение $h = 0$ и воспользуемся этими соотношениями; получим

$$\begin{aligned}
\varphi(0) &= y - y = 0,\\
\varphi'(0) &= (1 - p_1 - p_2) f(x, y),\\
\varphi''(0) &= (1 - 2p_2\alpha_2) f_x(x, y) + (1 - 2p_2\beta_{21}) f_y(x, y) f(x, y),\\
\varphi'''(0) &= (1 - 3p_2\alpha_2^2) f_{xx}(x, y) + (2 - 6p_2\beta_{21}) f_{xy}(x, y) f(x, y) +\\
&\quad + (1 - 3p_2\beta_{21}^2) f_{yy}(x, y)(f(x, y))^2 + f_y(x, y) y''(x).
\end{aligned} \tag{9}$$

Соотношение $\varphi(0) = 0$ выполняется при всех $f(x, y)$, если

$$1 - p_1 - p_2 = 0; \tag{10}$$

соотношение $\varphi''(0) = 0$ выполняется, если

$$1 - 2p_2\alpha_2 = 0 \quad \text{и} \quad 1 - 2p_2\beta_{21} = 0. \tag{11}$$

Таким образом, $\varphi(0) = \varphi'(0) = \varphi''(0) = 0$ при всех $f(x, y)$, если выполнены три указанных выше соотношения (10), (11) относительно четырех параметров. Задавая произвольно один из параметров, получим различные методы Рунге—Кутта с погрешностью второго порядка малости по h. Например, при $p_1 = 1/2$ получаем $p_2 = 1/2$, $\alpha_2 = 1$, $\beta_{21} = 1$, что соответствует паре расчетных формул (6). При $p_1 = 0$ получаем $p_2 = 1$, $\alpha_2 = 1/2$, $\beta_{21} = 1/2$, что соответствует паре расчетных формул (7). В случае уравнения $y' = y$, согласно (9) имеем $\varphi'''(0) = y$ независимо от значений p_1, p_2, α_2, β_{21}. Отсюда следует, что нельзя построить формулы Рунге—Кутта со значениями $q = 2$ и $s = 3$.

В случае $q = 3$ расчетных формул, соответствующих значению $s = 4$, не существует. Наиболее употребительна совокупность расчетных формул при $q = s = 3$:

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right),$$

$$k_3 = hf(x + h, y - k_1 + 2k_2), \quad \Delta y = \frac{1}{6}(k_1 + 4k_2 + k_3).$$

При $q = 4, 5$ нельзя построить расчетных формул рассматриваемого вида со значением $s = 5$; при $q = s = 4$ наиболее употребительна совокупность расчетных формул:

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right), k_3 = hf\left(x + \frac{h}{2}, y + \frac{k_2}{2}\right),$$

$$k_4 = hf(x + h, y + k_3), \quad \Delta y = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

Мы использовали выше формулировку «наиболее употребительный». Эта формулировка отражает исторически сложившуюся тенденцию в использовании численных методов. Казалось бы, в руководстве по численным методам следовало не просто отражать тенденцию, а указать, какая формула из данного семейства расчетных формул является наилучшей. Однако ответ на такой вопрос не прост.

У формул одинакового порядка точности по h главные члены погрешности на шаге часто оказываются непропорциональными. Например, вследствие (8), (9) главный член погрешности формулы (6) равен

$$(B - A)h^3,$$

где

$$B = \frac{1}{6}f_y y'', \quad A = \frac{1}{12}(f_{xx} + 2f_{xy}y' + f_{yy}(y')^2),$$

а у формулы (7) —

$$(B + A/2)h^3.$$

Поэтому можно указать два уравнения таких, что для первого уравнения меньшую погрешность дает метод (6), а для второго уравнения — метод (7).

В подобной ситуации рекомендации в пользу того или другого метода должны основываться на «волевом решении», принятом с учетом традиций и практики использования методов. Понятие практики вычислительной работы является довольно неопределенным. Число различных классов реально встречающихся дифференциальных уравнений существенно превосходит число задач, на которых производится сравнение методов их численного решения, поэтому суждения «с позиций практики» не всегда объективны. Однако несмотря на такую неопределенность, критерий практики часто несет в себе определенную положительную информацию, которая зачастую на данном этапе развития науки не может быть формализована или обоснована.

Если исторически первый из методов рассматриваемого класса оказался приемлемым, то в дальнейшем пользователи привыкают к нему. Замена этого метода на другой, даже более эффективный метод требует определенных затрат времени на «привыкание» пользователей к новому методу (а следовательно, и определенных психологических затрат). Чтобы широкий круг пользователей согласился на подобную перестройку, необходимо существенное преимущество нового метода по какой-либо из характеристик.

При дальнейшем рассмотрении для нас будет существенно, что погрешность метода на шаге $\varphi(h)$ имеет главный член, а именно справедливо представление вида

$$\varphi(h) = \psi(x, y)h^{s+1} + O(h^{s+2}). \tag{12}$$

Наметим основные этапы доказательства этого соотношения. Предположим, что правая часть и все ее производные до порядка $s+1$ включительно ограничены равномерно в области $G: x_0 \leqslant x \leqslant x_0 + X$, $-\infty < y < \infty$. Тогда также будут равномерно ограничены производные всех решений уравнения до порядка $s+2$ включительно. Согласно формуле Тейлора соотношение (8) можно записать в уточненной форме

$$\varphi(h) = \frac{\varphi^{(s+1)}(0)}{(s+1)!}h^{s+1} + \frac{\varphi^{(s+2)}(\theta h)}{(s+2)!}h^{s+2}.$$

Имеем равенство

$$\varphi^{(s+1)}(0) = y^{(s+1)}(0) - z^{(s+1)}(0).$$

Обе величины $y^{(s+1)}(0)$ и $z^{(s+1)}(0)$ явно выражаются через значения в точке (x, y) функции f и ее производных порядка не выше s; примеры таких явных выражений (при $s = 2$) мы уже получали.

Поскольку правая часть дифференцируема $s+1$ раз, то отсюда следует, что функция $\psi(x, y)$ дифференцируема в области G и ее производные ψ_x и ψ_y равномерно ограничены в этой области. Аналогично устанавливается, что величина $\varphi^{(s+2)}(\theta h)$ равномерно ограничена при $x_0 \leqslant x < x+h \leqslant x_0 + X$. Таким образом, соотношение (12) имеет место.

§ 3. Методы с контролем погрешности на шаге

Часто в ходе расчетов бывает целесообразно изменять шаг интегрирования, контролируя величину погрешности метода на шаге. При практической оценке этой величины можно, например, рассуждать следующим образом. Главный член погрешности на шаге интегрирования есть

$$\frac{\varphi^{(s+1)}(0)h^{s+1}}{(s+1)!}.$$

Точка $(x+h,\, z(h))$ находится близко от точки $(x,\, y)$, поэтому погрешность на следующем шаге интегрирования будет иметь такой же главный член. В результате двух шагов будет получено приближение $y^{(1)}$ к значению $y(x+2h)$ такое, что

$$y^{(1)} - y(x+2h) \sim 2\frac{\varphi^{(s+1)}(0)h^{s+1}}{(s+1)!}.$$

Если, исходя из точки $(x,\, y)$, применить метод Рунге—Кутта с шагом $2h$, то получится приближенное значение $y^{(2)}$, для которого

$$y^{(2)} - y(x+2h) \sim \frac{\varphi^{(s+1)}(0)(2h)^{s+1}}{(s+1)!}.$$

Из этих соотношений вытекает представление главного члена погрешности на шаге

$$y^{(1)} - y(x+2h) \sim \frac{y^{(2)} - y^{(1)}}{2^s - 1}.$$

При желании можно уточнить полученное приближенное значение, прибавив к нему величину главного члена погрешности, т. е. положить

$$y(x+2h) \approx y^{(1)} + \frac{y^{(1)} - y^{(2)}}{2^s - 1}. \tag{1}$$

Для более гибкого управления выбором шага интегрирования иногда желательно иметь возможность совершать шаг интегрирования и оценивать погрешность при меньшем количестве вычисляемых значений правых частей.

Примером совокупности формул с теми же характеристиками точности при меньшем числе обращений к правой части может служить совокупность формул

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x + \frac{h}{2}, y + \frac{1}{4}(k_1 + k_2)\right), \quad k_4 = hf(x + h, y - k_2 + 2k_3),$$

$$k_5 = hf\left(x + \frac{2h}{3}, y + \frac{1}{27}(7k_1 + 10k_2 + k_4)\right), \tag{2}$$

$$k_6 = hf\left(x + \frac{h}{5}, y + \frac{1}{625}(28k_1 - 125k_2 + 546k_3 + 54k_4 - 378k_5)\right),$$

$$\Delta y = \frac{1}{6}(k_1 + 4k_3 + k_4)$$

с главным членом погрешности

$$y(x + h) - z(h) = r + O(h^6),$$

$$r = -\frac{1}{336}(42k_1 + 224k_3 + 21k_4 - 162k_5 - 125k_6).$$

Если положить $y(x_0+h) \approx z(h)+r$, то получим метод рассматриваемого типа со значением $s = 5$ и соответственно погрешностью на шаге порядка $O(h^6)$.

В одной распространенной стандартной программе управление шагом интегрирования осуществляется по методу, близкому к горизонтальной процедуре из § 3.17. Задаются мерами погрешности на шаге ε_0 и $\varepsilon_1 < \varepsilon_0$ и некоторым параметром $M > 0$ порядка y. Обычно $\varepsilon_1/\varepsilon_0 \geqslant 2^{-l}$, где l — порядок величины r по h; часто берут $\varepsilon_1/\varepsilon_0 = 2^{-l}$. Если $\psi_n(h) = |r|/\max(M, |y_n|) > \varepsilon_0$, то шаг признается слишком большим и делается попытка интегрирования, начиная с тех же значений (x_n, y_n), с вдвое более мелким шагом $(h/2)$. Если $\psi_n(h) \leqslant \varepsilon_0$, то достигнутая точность признается удовлетворительной. В случае, когда $\varepsilon_1 \leqslant \psi_n(h) \leqslant \varepsilon_0$, следующий шаг берется равным h, а в случае, когда $\psi_n(h) < \varepsilon_1$, — равным $2h$. Такой относительно простой способ выбора переменного шага интегрирования часто позволяет решить задачу с существенно меньшими затратами времени ЭВМ по сравнению со случаем интегрирования с постоянным шагом (при той же точности результата).

§ 4. Оценки погрешности одношаговых методов

Рассмотрим множество всевозможных методов интегрирования, где последовательно получаются приближения y_j к значениям $y(x_j)$, $x_0 < x_1 < \ldots < x_N = x_0 + X$. Пусть в процессе численного интегрирования k фиксировано и при $j \geqslant k$ значения y_j определяются как значения некоторого функционала

$$y_j = \Phi(f;\, x_j,\, \ldots,\, x_{j-k};\, y_{j-1},\, \ldots,\, y_{j-k}). \tag{1}$$

Такой способ численного интегрирования называют *k-шаговым*. Все построенные выше способы интегрирования имеют следующее общее свойство: приближенное значение решения в следующей точке определялось только в зависимости от значения решения в предыдущей точке, следовательно, расчетные формулы, соответствующие этим способам, представимы в виде (1) со значением $k = 1$. Такие методы называются *одношаговыми*.

Рассмотрим специальный способ получения оценки погрешности, применимый лишь к одношаговым методам.

Запишем формулу (1) в виде

$$y_{j+1} = \Phi(f, x_j, x_{j+1} - x_j, y_j). \tag{2}$$

Получаемые в процессе реальных вычислений приближения к значениям $y(x_j)$ связаны не соотношениями (2), а некоторыми соотношениями

$$y_{j+1} = \Phi(f, x_j, x_{j+1} - x_j, y_j) + \delta_{j+1}. \tag{3}$$

Наличие слагаемого δ_{j+1} обусловлено следующими причинами:

1) округлением чисел при вычислениях;

2) погрешностями в значениях правой части $f(x, y)$; эти погрешности вызваны тем, что рассматриваемая нами функция $f(x, y)$ является некоторым приближением к правой части реального дифференциального уравнения; кроме того, зачастую в процессе вычисления значений $f(x, y)$ в ЭВМ эта функция приближается другими функциями, что вносит дополнительные погрешности при вычислениях значений правой части;

3) в некоторых случаях значение y_{j+1} определяется из уравнения, эквивалентного (1), но не разрешенного в явном виде относительно переменной y_{j+1}; тогда величина δ_{j+1} содержит составляющую, являющуюся следствием приближенного решения этого уравнения.

Хотя погрешность δ_{j+1} вызвана не только округлением, ее часто называют *вычислительной погрешностью на шаге*.

Точно так же начальное условие y_0 отличается от значения отыскиваемого решения задачи $y(x_0)$ из-за погрешности в определении исходных данных и округлений. Пусть $y(x)$ — искомое решение дифференциального уравнения, а $y_j(x)$ — решения, удовлетворяющие условиям $y_j(x_j) = y_j$ (рис. 8.4.1).

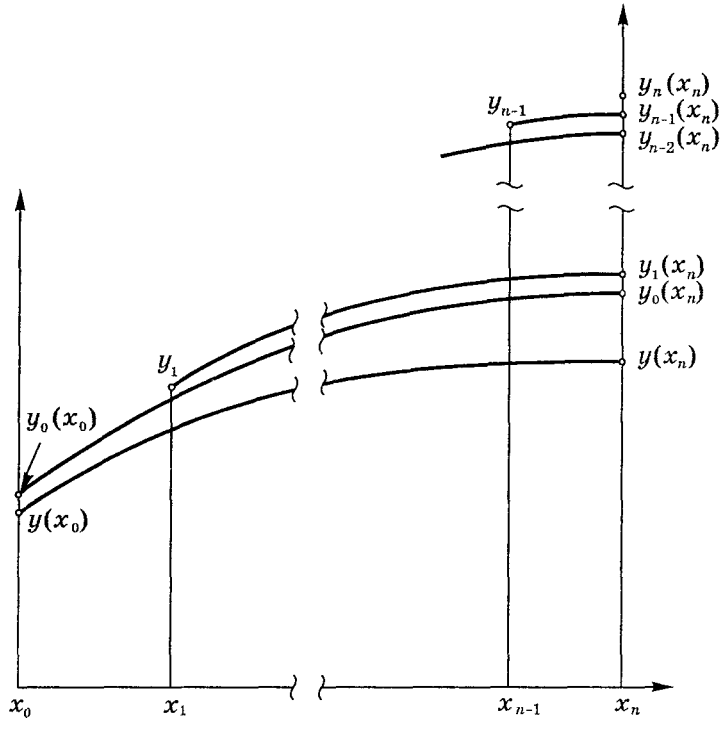


Рис. 8.4.1

Погрешность $R_n = y_n(x_n) - y(x_n)$ можно представить в виде

$$R_n = y_n(x_n) - y_0(x_n) + y_0(x_n) - y(x_n) = \\ = \sum_{j=1}^{n} \Big(y_j(x_n) - y_{j-1}(x_n) \Big) + \Big(y_0(x_n) - y(x_n) \Big). \tag{4}$$

Разность решений дифференциального уравнения в одной точке может быть выражена через их разность в другой точке следующим путем.

Лемма. *Пусть $Y_1(x)$ и $Y_2(x)$ — решения дифференциального уравнения $y' = f(x,y)$, где $f(x,y)$ — непрерывная и непрерывно дифференцируемая по переменной y функция. Тогда*

$$Y_2(\beta) - Y_1(\beta) = \Big(Y_2(\alpha) - Y_1(\alpha) \Big) \exp \left\{ \int_\alpha^\beta f_y(x,\, \tilde{y}(x))\, dx \right\},$$

где $\tilde{y}(x)$ заключено между $Y_1(x)$ и $Y_2(x)$.

Доказательство. Вычтем друг из друга равенства

$$Y_2' = f(x, Y_2), \quad Y_1' = f(x, Y_1).$$

Согласно формуле Лагранжа разность $f(x,Y_2)-f(x,Y_1)$ может быть представлена в виде $f_y(x,\tilde{y})(Y_2-Y_1)$, где $\tilde{y}$ заключено между Y_1 и Y_2. В результате получится линейное дифференциальное уравнение относительно Y_2-Y_1:

$$(Y_2-Y_1)'=f_y(x,\tilde{y})(Y_2-Y_1). \tag{5}$$

Функция

$$f_y(x,\tilde{y}(x))=\frac{f(x,Y_2(x))-f(x,Y_1(x))}{Y_2(x)-Y_1(x)}$$

непрерывна, поскольку числитель и знаменатель — непрерывные функции, а знаменатель отличен от нуля. Из (5) следует утверждение леммы.

Пусть $\alpha=x_j$, $\beta=x_n$, $Y_1(x)=y_{j-1}(x)$, $Y_2(x)=y_j(x)$; тогда вследствие леммы

$$y_j(x_n)-y_{j-1}(x_n)=\Big(y_j(x_j)-y_{j-1}(x_j)\Big)\exp\left\{\int_{x_j}^{x_n}f_y(x,\,\tilde{y}_j(x))\,dx\right\},$$

где $\tilde{y}_j(x)$ заключено между $y_{j-1}(x)$ и $y_j(x)$.

Точно так же

$$y_0(x_n)-y(x_n)=\Big(y_0(x_0)-y(x_0)\Big)\exp\left\{\int_{x_0}^{x_n}f_y(x,\,\tilde{y}_0(x))\,dx\right\}.$$

Теперь равенство (4) можно записать в виде

$$R_n=\sum_{j=1}^{n}\omega_j\exp\left\{\int_{x_j}^{x_n}f_y(x,\,\tilde{y}_j(x))\,dx\right\}+R_0\exp\left\{\int_{x_0}^{x_n}f_y(x,\,\tilde{y}_0(x))\,dx\right\}, \tag{6}$$

где $\omega_j=y_j(x_j)-y_{j-1}(x_j)$, $j=1,\ldots$

Из (3) вытекает соотношение

$$\omega_j=y_j(x_j)-y_{j-1}(x_j)=\rho_j+\delta_j,$$

где

$$\rho_j=\Phi(f,\,x_{j-1},\,x_j-x_{j-1},\,y_{j-1})-y_{j-1}(x_j).$$

Посмотрим, какой смысл имеет величина ρ_j; $\Phi(f,\,x_{j-1},\,x_j-x_{j-1},\,y_{j-1})$ есть число, получаемое в результате вычислений по расчетной формуле (2), $y_{j-1}(x_j)$ — значение в точке x_j точного решения дифференциального уравнения, удовлетворяющего условию $y_{j-1}(x_{j-1})=y_{j-1}$. Таким образом, ρ_j есть погрешность одного шага рассматриваемого метода, если вычисления начинаются с точки $(x_{j-1},\,y_{j-1})$ и производятся без округлений, а шаг равен (x_j-x_{j-1}). Величина ρ_j называется *погрешностью метода на шаге.*

Предположим, что при всех j, соответствующих рассматриваемому отрезку интегрирования $x_0<x_j\leqslant x_0+X$, выполняется неравенство

$$|\rho_j|\leqslant C_1(x_j-x_{j-1})^{s+1}. \tag{7}$$

Пусть

$$L = \sup_{x_0 \leqslant x \leqslant x_0 + X} |f_y| < \infty$$

и

$$H = \max_{0 < j \leqslant N} (x_j - x_{j-1}).$$

Загрубляя (7), имеем

$$|\rho_j| \leqslant C_1 H^s (x_j - x_{j-1}). \tag{8}$$

При $x_0 \leqslant x_j \leqslant x_n \leqslant x_0 + X$ справедливы неравенства

$$\exp\left\{\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x))\, dx\right\} \leqslant \exp\{L(x_n - x_j)\} \leqslant \exp\{LX\}.$$

Воспользовавшись этими неравенствами для оценки правой части (6), получим

$$|R_n| \leqslant \exp\{LX\} \left(\sum_{j=1}^{n} (|\rho_j| + |\delta_j|) + |R_0|\right).$$

Применим теперь к предыдущему неравенству оценку (8), получим

$$\begin{aligned} |R_n| &\leqslant \exp\{LX\} \left(\sum_{j=1}^{n} (C_1 H^s (x_j - x_{j-1}) + |\delta_j|) + |R_0|\right) \leqslant \\ &\leqslant \exp\{LX\}(C_1 H^s (x_n - x_0) + n\delta + |R_0|) \leqslant \\ &\leqslant \exp\{LX\}(C_1 X H^s + N\delta + |R_0|); \end{aligned} \tag{9}$$

здесь $\delta = \max\limits_j |\delta_j|$. Из этого соотношения следует, что $\max\limits_{x_0 < x_n \leqslant x_0 + X} |R_n| \to 0$ при $H \to 0$, если одновременно $N\delta \to 0$, $|R_0| \to 0$. Таким образом, при достаточно мелком шаге интегрирования и малой вычислительной погрешности приближенное решение, получаемое при употреблении метода Рунге—Кутта, близко к точному решению.

Часто решение дифференциального уравнения отыскивается на большом промежутке. Тогда в полученные оценки погрешности входит как множитель очень большое число $\exp\{LX\}$. При LX большом может оказаться, что достижение нужной точности требует столь мелких шагов и столь малой величины вычислительной погрешности на шаге, что использование рассматриваемого метода будет нецелесообразно. Поэтому характеристика методов по признаку — сходится ли приближенное решение к точному при измельчении шага и при достаточно быстром уменьшении вычислительной погрешности или не сходится — является еще недостаточной.

Если $f_y(x, y) \leqslant -b < 0$, то в оценке (9) можно избавиться от множителя, резко растущего с увеличением X. Рассмотрим случай постоянного шага $x_j - x_{j-1} = h$. Тогда

$$\exp\left\{\int_{x_j}^{x_n} f_y(x, \tilde{y}(x))\, dx\right\} \leqslant \exp\{-b(x_n - x_j)\} \leqslant \exp\{-b(n - j)h\}.$$

Пусть $|\omega_j| \leqslant C_1 h^{s+1} + \delta$. Оценивая правую часть (6), получаем

$$|R_n| \leqslant \sum_{j=1}^{n} (C_1 h^{s+1} + \delta) \exp\{-b(n-j)h\} + |R_0| \exp\{-bnh\}. \tag{10}$$

Имеем

$$\sum_{j=1}^{n} \exp\{-b(n-j)h\} \leqslant \sum_{k=0}^{\infty} \exp\{-bkh\} = \frac{1}{1-\exp\{-bh\}}.$$

Таким образом, получаем окончательную оценку погрешности

$$|R_n| \leqslant \frac{C_1 h^{s+1} + \delta}{1-\exp\{-bh\}} + |R_0| \exp\{-bnh\}. \tag{11}$$

Поскольку $1 - \exp\{-bh\} \sim |b|h$, то верна более простая по виду оценка

$$|R_n| \leqslant C_2(h^s + \delta/h) + |R_0| \exp\{-bnh\}. \tag{12}$$

Формально эта оценка не зависит от длины промежутка интегрирования X, однако длина промежутка интегрирования может неявно влиять на значение коэффициента C_2 через оценки производных.

Наличие оценки (11), не ухудшающейся с увеличением промежутка интегрирования, позволяет использовать такие методы для отыскания, например, устойчивых решений дифференциальных уравнений путем установления. Начинаем численное интегрирование с произвольных начальных данных и с течением времени выходим на устойчивое решение. Этот прием часто употребляется при отыскании устойчивых предельных циклов систем обыкновенных дифференциальных уравнений.

В связи с полученной оценкой (12) и возможностью получения аналогичной оценки для случая численного решения задачи Коши для систем дифференциальных уравнений с быстро сближающимися решениями одношаговые методы находят широкое применение в вычислительной практике. В то же время методы, для которых в подобной ситуации погрешность растет неограниченно, практически исчезли из употребления. Заметим, что в случае $f_y \geqslant b > 0$ в соответствии с утверждением леммы решения расходятся с экспоненциальной скоростью, и поэтому погрешность *любого метода должна неограниченно расти при* $x_n \to \infty$.

Другим достоинством одношаговых методов является удобство изменения шага интегрирования и однотипность вычислений во всех расчетных точках (у конкурирующих с ними методов Адамса изменение шага интегрирования и начало вычислений производятся с помощью некоторых специальных формул, которые мы не рассматриваем из-за их громоздкости).

§ 5. Конечно-разностные методы

Среди k-шаговых методов наиболее употребительны методы интегрирования на сетке с постоянным шагом $x_j - x_{j-1} \equiv h = \text{const}$ при помощи соотношений вида

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} F_i(h, x_{n-i}, y_{n-i}) = 0, \tag{1}$$

где a_i — постоянные, F_i — некоторые функции, определяемые функцией $f(x, y)$. В свою очередь, среди таких методов традиционно наиболее употребительны методы

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f_i(x_{n-i}, y_{n-i}) = 0, \tag{2}$$

которые принято называть *конечно-разностными методами* или *конечно-разностными схемами*.

В вычислительной практике применяются формулы (1), (2) со значениями $a_0 \neq 0$, $b_0 = 0$ — *явные*, или *экстраполяционные*, и формулы с $a_0 \neq 0$, $b_0 \neq 0$ — *неявные*, или *интерполяционные*.

Формулы (1), (2) при $a_0 = 0$, $b_0 \neq 0$, называемые *формулами с забеганием вперед*, рассматриваться не будут, поскольку они не нашли распространения в вычислительной практике из-за сложности использования. Тем не менее их следует принимать во внимание при проведении теоретических рассмотрений, так как они расширяют класс используемых конечно-разностных схем. В дальнейшем предполагается, что $a_0 \neq 0$. При $b_0 \neq 0$ уравнение (2) можно записать в виде

$$y_n = \varphi_n(y_n), \quad \varphi_n(y_n) = A_n + h\frac{b_0}{a_0} f(x_n, y_n).$$

Здесь

$$A_n = a_0^{-1} \left(-\sum_{i=1}^{k} a_{-i} y_{n-i} + h \sum_{i=1}^{k} b_{-i} f(x_{n-i}, y_{n-i}) \right)$$

не зависит от y_n. Будем решать это уравнение методом простой итерации:

$$y_n^{k+1} = \varphi_n(y_n^k).$$

Поскольку $\varphi_n'(y) = h\dfrac{b_0}{a_0} f_y(x_n, y)$, то при достаточно малых h выполнено условие сжимаемости отображения $\varphi_n(y)$ и поэтому итерационный процесс сходится. Начальное приближение y_n^0 определяется из какой-либо явной формулы

$$\sum_{i=0}^{l} a_{-i}^1 y_{n-i} - h \sum_{i=0}^{l} b_{-i}^1 f(x_{n-i}, y_{n-i}) = 0, \quad a_0^1 \neq 0.$$

Число итераций на шаге определяется из разумного соотношения между трудоемкостью итераций и точностью получаемых в процессе итераций приближений так же, как в случае формулы (2.4), являющейся частным случаем формулы (2).

Простейшие методы типа (2) получаются на основе квадратурных формул. Всякая квадратурная формула

$$\int_{-ph}^{0} f(x)\,dx = h\sum_{i=0}^{m} b_{-i} f(-ih) + r, \quad m \geqslant 0, \tag{3}$$

где r — остаточный член, порождает соответствующую формулу численного интегрирования обыкновенных дифференциальных уравнений. Действительно, подставляя в (3) соотношение $f(x) = y'(x_n + x)$, имеем

$$\int_{-ph}^{0} f(x)\,dx = y(x_n) - y(x_{n-p}) = h\sum_{i=0}^{m} b_{-i} y'(x_{n-i}) + r.$$

Заменяя $y'(x)$ на $f(x, y(x))$ и отбрасывая r, получим

$$y(x_n) - y(x_{n-p}) \approx h\sum_{i=0}^{m} b_{-i} f\Big(x_{n-i},\, y(x_{n-i})\Big). \tag{4}$$

Соответствующая конечно-разностная схема запишется следующим образом:

$$y_n - y_{n-p} - h\sum_{i=0}^{m} b_{-i} f(x_{n-i},\, y_{n-i}) = 0.$$

Например, формуле трапеций

$$\int_{-h}^{0} f(x)\,dx \approx \frac{h}{2}\Big(f(0) + f(-h)\Big)$$

соответствует интерполяционная формула

$$y_n - y_{n-1} = \frac{h}{2}(f_n + f_{n-1}), \tag{5}$$

формуле Симпсона

$$\int_{-2h}^{0} f(x)\,dx \approx \frac{h}{3}\Big(f(0) + 4f(-h) + f(-2h)\Big)$$

— интерполяционная формула $y_n - y_{n-2} = \dfrac{h}{3}(f_n + 4f_{n-1} + f_{n-2})$; здесь $f_m = f(x_m, y_m)$. Квадратурной формуле прямоугольников, записанной в виде

$$\int_{-2h}^{0} f(x)dx \approx 2hf(x),$$

соответствует экстраполяционная формула

$$y_n - y_{n-2} = 2hf_{n-1}. \tag{6}$$

Если остаточный член квадратуры (3) оценивается через $D(q)\max|f^{(q)}(x)|h^{q+1}$, то погрешность равенства (4) будет оцениваться через $D(q)\max|y^{(q+1)}(x)|h^{q+1}$.

В настоящее время из конечно-разностных методов, как правило, употребляются на практике только методы, соответствующие $p = 1$; их называют *методами Адамса.*

Другие известные методы вида (2) не выдержали испытания на практике по следующим причинам:

1) отсутствует сходимость при уменьшении шага (в предположении отсутствия вычислительной погрешности) даже для бесконечно дифференцируемых правых частей $f(x, y)$;

2) при наличии сходимости происходит экспоненциальный рост (с увеличением X) погрешности в рассмотренном в предыдущем параграфе случае, когда $f_y \leqslant -b < 0$;

3) для некоторых схем при $p > 1$ возникают дополнительные (по сравнению со случаем $p = 1$) неудобства при изменении шага интегрирования.

Явные формулы Адамса обычно записываются в виде

$$y_n - y_{n-1} = h\sum_{i=0}^{m}\gamma_i\nabla^i f_{n-1},$$

а *неявные* — в виде

$$y_n - y_{n-1} = h\sum_{i=0}^{m}\bar{\gamma}_i\nabla^i f_n;$$

здесь, как и в § 2.10,

$$\nabla^i f_n = \sum_{j=0}^{i}(-1)^j C_i^j f_{n-j}.$$

Коэффициенты γ_i, $\bar{\gamma}_i$ вычисляются по формулам

$$\gamma_i = \int_0^1 \prod_{k=1}^{i}\left(1 - \frac{u}{k}\right)du, \quad \bar{\gamma}_0 = 1,$$

$$\bar{\gamma}_i = \gamma_i - \gamma_{i-1} = -\int_0^1 \frac{u}{i}\prod_{k=1}^{i-1}\left(1 - \frac{u}{k}\right)du \quad \text{при} \quad i > 0.$$

Задача 1. Показать, что

$$\gamma_i \sim \frac{\text{const}}{\ln i}, \quad \bar{\gamma}_i \sim \frac{\text{const}}{i\ln i} \quad \text{при} \quad i \to \infty.$$

Обычно методы Адамса используются по следующей схеме. Сначала вычисляется нулевое приближение по явной формуле Адамса, и затем производятся 1–2 итерации на основе неявной формулы (с тем же значением m).

§ 6. Метод неопределенных коэффициентов

Для построения формул численного интегрирования можно также использовать метод неопределенных коэффициентов. Заменим производную $y'(x_n)$ и значение $f(x_n, y(x_n))$ некоторыми выражениями

$$y'(x_n) \approx \sum_{i=0}^{k} \frac{a_{-i} y(x_{n-i})}{h}, \tag{1}$$

$$f\Big(x_n, y(x_n)\Big) \approx \sum_{i=0}^{k} b_{-i} f\Big(x_{n-i}, y(x_{n-i})\Big) \tag{2}$$

(предполагается, что a_{-i} и b_{-i} не зависят от h). Отсюда получаем приближенное равенство

$$\sum_{i=0}^{k} \frac{a_{-i} y(x_{n-i})}{h} \approx \sum_{i=0}^{k} b_{-i} f\Big(x_{n-i}, y(x_{n-i})\Big). \tag{3}$$

Ему соответствует конечно-разностная схема

$$\sum_{i=0}^{k} \frac{a_{-i} y_{n-i}}{h} - \sum_{i=0}^{k} b_{-i} f_{n-i} = 0. \tag{4}$$

Величина

$$r_n = \sum_{i=0}^{k} \frac{a_{-i} y(x_{n-i})}{h} - \sum_{i=0}^{k} b_{-i} f\Big(x_{n-i}, y(x_{n-i})\Big)$$

называется *погрешностью аппроксимации исходного дифференциального уравнения схемой* (4).

Определение. *Разностная схема аппроксимирует дифференциальную на отрезке* $[x_0, x_0 + X]$, *если*

$$\|r\| = \max_{x_0 \leqslant x_n \leqslant x_0 + X} |r_n| \to 0 \quad \text{при} \quad h \to 0.$$

Вспоминая, что $f\Big(x_{n-i}, y(x_{n-i})\Big) = y'(x_{n-i})$ и $x_{n-i} = x_n - ih$, получим

$$r_n = \sum_{i=0}^{k} \frac{a_{-i} y(x_n - ih)}{h} - \sum_{i=0}^{k} b_{-i} y'(x_n - ih).$$

Предположим, что все производные решения до порядка q ограничены:

$$|y^{(p)}(x)| \leqslant M_p < \infty \quad \text{при} \quad x_0 \leqslant x \leqslant x_0 + X, \quad p = 0, \ldots, q.$$

Представим все величины $y(x_n - ih)$ и $y'(x_n - ih)$ с помощью формулы Тейлора следующим образом:

$$y(x_n - ih) = \sum_{p=0}^{q-1} y^{(p)}(x_n) \frac{(-ih)^p}{p!} + \beta_n^i,$$

$$y'(x_n - ih) = \sum_{p=1}^{q-1} y^{(p)}(x_n) \frac{(-ih)^{p-1}}{(p-1)!} + \gamma_n^i,$$

где согласно оценке остаточного члена ряда Тейлора имеем

$$|\beta_n^i| \leqslant M_q (ih)^q / q!, \qquad |\gamma_n^i| \leqslant M_q (ih)^{q-1} / (q-1)!.$$

Подставим выражения $y(x_n - ih)$ и $y'(x_n - ih)$ в правую часть представления r_n и соберем коэффициенты при $y^{(p)}(x_n)$. Получим

$$r_n = E_0 h^{-1} y(x_n) + E_1 y'(x_n) + \cdots + E_{q-1} h^{q-2} y^{(q-1)}(x_n) + \varepsilon_n; \tag{5}$$

здесь

$$E_0 = \sum_{i=0}^{k} a_{-i},$$

$$E_p = \sum_{i=0}^{k} \frac{a_{-i}(-i)^p}{p!} - \sum_{i=0}^{k} \frac{b_{-i}(-i)^{p-1}}{(p-1)!}, \quad p > 0, \tag{6}$$

$$\varepsilon_n = \sum_{i=0}^{k} \frac{a_{-i}\beta_n^i}{h} - \sum_{i=0}^{k} b_{-i}\gamma_n^i = O(h^{q-1}).$$

Имеем

$$|\varepsilon_n| \leqslant D_q M_q h^{q-1},$$

где

$$D_q = \sum_{i=0}^{k} |a_{-i}| \frac{i^q}{q!} + \sum_{i=0}^{k} |b_{-i}| \frac{i^{q-1}}{(q-1)!}.$$

Как правило, производя более аккуратную оценку, можно уменьшить значение D_q в оценке $|\varepsilon_n|$. Если $E_0 = \cdots = E_m = 0$, то $\varepsilon_n = O(h^m)$ и говорят, что схема (4) имеет *m-й порядок аппроксимации*. Всякая схема m-го порядка аппроксимации является схемой q-го порядка аппроксимации при $q < m$. Если $E_0 = \cdots = E_m = 0$, а $E_{m+1} \neq 0$, то говорят, что *порядок аппроксимации строго равен m*.

Согласно (1), (2) для любой гладкой $y(x)$ имеем соотношения

$$\lim_{h\to 0}\sum_{i=0}^{k}\frac{a_{-i}y(x-ih)}{h}=y'(x),$$
$$\lim_{h\to 0}\sum_{i=0}^{k}b_{-i}f\big(x-ih,\,y(x-ih)\big)=f\big(x,\,y(x)\big). \tag{7}$$

Лемма. *Соотношения* (7) *выполнены тогда и только тогда, когда*

$$E_0=E_1=0,\quad b_0+\cdots+b_{-k}=1. \tag{8}$$

Доказательство. Согласно формуле Тейлора имеем

$$y(x-ih)=y(x)-ihy'(x)+O(h^2),$$
$$f\,(x-ih,\,y(x-ih))=f\big(x,\,y(x)\big)+O(h).$$

Подставляя эти соотношения для левых частей в (7), получим

$$\lim_{h\to 0}\left(\left(\sum_{i=0}^{k}\frac{a_{-i}}{h}y(x)\right)+\sum_{i=0}^{k}a_{-i}(-i)y'(x)+O(h)\right)=y'(x),$$
$$\lim_{h\to 0}\left(\left(\sum_{i=0}^{k}b_{-i}\right)f(x,\,y(x))+O(h)\right)=f(x,\,y(x)).$$

Для справедливости этих соотношений необходимо и достаточно выполнения условий

$$\sum_{i=0}^{k}a_{-i}=0,\quad -\sum_{i=0}^{k}ia_{-i}=1,\quad \sum_{i=0}^{k}b_{-i}=1.$$

Левая часть первого из этих равенств равна E_0, разность левых частей второго и третьего равна E_1. Отсюда следует справедливость утверждения леммы.

Уравнения $E_0=\cdots=E_m=0$ образуют однородную систему линейных алгебраических уравнений относительно $2k+2$ неизвестных. Если число неизвестных больше числа уравнений, т. е. $2k+2>m+1$ (или, что то же самое, $2k\geqslant m$), то эта система имеет ненулевое решение. Можно показать, что при $2k=m$ эта система имеет однопараметрическое семейство ненулевых решений

$$a_{-i}=ca^0_{-i},\quad b_{-i}=cb^0_{-i},$$

причем $\omega=\sum_{i=0}^{k}b^0_{-i}\neq 0$. Выбирая $c=\omega^{-1}$, получим разностную схему $2k$-го порядка аппроксимации. Иногда возникает необходимость построить явную схему ($b_0=0$). Решая систему уравнений $b_0=0$, $E_0=\cdots=$

$E_{2k-1} = 0$ и выбирая решение с $\sum_{i=0}^{k} b_{-i} = 1$, получаем схему $(2k-1)$-го порядка аппроксимации.

Примеры. При $k = 2$ общее решение системы уравнений $E_0 = \cdots = E_4 = 0$ имеет вид $a_0 = c$, $a_{-1} = 0$, $a_{-2} = -c$, $b_0 = b_{-2} = \frac{c}{3}$, $b_{-1} = \frac{4}{3}c$. Из условия (8) получаем $c = 1/2$; расчетная схема имеет вид

$$\frac{y_n - y_{n-2}}{2h} - \left(\frac{1}{6}f_n + \frac{2}{3}f_{n-1} + \frac{1}{6}f_{n-2}\right) = 0.$$

Заметим, что выше эта схема была получена из формулы Симпсона.

Если потребовать выполнения равенств $b_0 = E_0 = E_1 = E_2 = E_3 = 0$, то получим схему

$$\frac{y_n + 4y_{n-1} - 5y_{n-2}}{6h} - \left(\frac{2}{3}f_{n-1} + \frac{1}{3}f_{n-2}\right) = 0. \tag{9}$$

Отметим следующее обстоятельство. Если равенства

$$y'(x_n) \approx \sum_{i=0}^{k} \frac{a_{-i}y(x_{n-i})}{h}, \tag{10}$$

$$f\big(x_n, y(x_n)\big) \approx \sum_{i=0}^{k} b_{-i}f\big(x_{n-i}, y(x_{n-i})\big) \tag{11}$$

выполняются с точностью до членов порядка $O(h^m)$, то величина r_n, равная разности погрешностей этих соотношений, имеет порядок $O(h^m)$ и, следовательно, согласно (3) $E_0 = \cdots = E_m = 0$. Однако для выполнения условия $r_n = O(h^m)$ не обязательно требовать, чтобы погрешности приближенных соотношений (10), (11) имели порядок $O(h^m)$. Например, для последней схемы $r_n = O(h^3)$, в то время как погрешности соотношений (10) порядка $O(h)$.

Кроме построенных выше методов типа Рунге—Кутта и конечно-разностных методов следует отметить группу методов, где при нахождении каждого нового значения y_n используется несколько предшествующих значений y_{n-i}, как в конечно-разностных методах, но в то же время на каждом шаге производится несколько вычислений правой части, как в методах Рунге—Кутта.

Пример метода этой группы:

$$y_{n-1/2} = y_{n-2} + \frac{h}{8}(9f_{n-1} + 3f_{n-2}),$$

$$y_n^1 = \frac{1}{5}(28y_{n-1} - 23y_{n-2}) + \frac{h}{15}(32f_{n-1/2} - 60f_{n-1} - 26f_{n-2}),$$

$$y_n = \frac{1}{31}(32y_{n-1} - y_{n-2}) + \frac{h}{93}(64f_{n-1/2} + 15f_n^1 + 12f_{n-1} - f_{n-2}),$$

погрешность на шаге порядка $O(h^6)$; здесь $f_m^k = f(x_m, y_m^k)$.

§ 7. Исследование свойств конечно-разностных методов на модельных задачах

После конструирования нового метода решения задачи, например метода решения дифференциальных уравнений, целесообразно, прежде чем писать программу, посмотреть, как будет работать этот метод на простейших модельных задачах, где точное и приближенное решения вычисляются в явном виде. Если для такой задачи метод дает неудовлетворительный результат, то от применения этого метода, скорее всего, стоит отказаться.

Часто первоначально конструируется не один метод, а некоторое семейство методов, зависящих от одного или нескольких параметров. Изучение модельного примера может позволить сравнить эти методы и выбрать оптимальные значения параметров. На примере решаемой в явном виде задачи можно понять реальную ситуацию, возникающую при реализации метода.

Такой подход часто более предпочтителен, чем подробное теоретическое исследование, поскольку дает большой выигрыш по времени.

Из формулы (6.4) можно находить значение y_n, если известны значения $y_{n-k}, \ldots, y_{n-1}$; поэтому, чтобы начать вычисления, нужно знать y_j в k начальных точках $y_0, y_1, \ldots, y_{k-1}$. Они могут быть найдены каким-либо образом заранее, например с помощью формулы Тейлора или методом Рунге—Кутта.

Рассмотрим вопрос о том, насколько влияют погрешности в начальных данных разностной задачи $y_0, y_1, \ldots, y_{k-1}$ на ее решение.

Пусть y_n^s $(s = 1, 2)$ — решения разностной задачи

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0 \tag{1}$$

при начальных данных $y_0^s, \ldots, y_{k-1}^s$ соответственно. На основании формулы Лагранжа имеем

$$f(x_m, y_m^2) - f(x_m, y_m^1) = l_m \varepsilon_m;$$

здесь $l_m = f_y(x_m, \bar{y}_m)$, $\bar{y}_m$ лежит между y_m^1 и y_m^2, $\varepsilon_m = y_m^2 - y_m^1$.

Вычтем из соотношения (1) при $s = 2$ то же соотношение при $s = 1$. Получим уравнение относительно разности ε_m:

$$\sum_{i=0}^{k} (a_{-i} - h b_{-i} l_{n-i}) \varepsilon_{n-i} = 0. \tag{2}$$

Оказывается, что довольно существенную информацию о поведении погрешности можно получить, рассматривая простейшее дифференциаль-

ное уравнение $y' = 0$. В этом случае (2) превращается в уравнение с постоянными коэффициентами

$$\sum_{i=0}^{k} a_{-i}\varepsilon_{n-i} = 0. \tag{3}$$

Соответствующее характеристическое уравнение имеет вид

$$F_0(\mu) = \sum_{i=0}^{k} a_{-i}\mu^{k-i} = 0. \tag{4}$$

Пусть μ_1 — максимальный по модулю из корней этого уравнения. Поскольку $E_0 = F_0(1)$, то условие $E_0 = 0$ равносильно тому, что $\mu = 1$ является корнем (4). Поэтому $|\mu_1| \geqslant 1$. Сеточная функция $\varepsilon_n = \mathrm{const} \cdot \mu_1^n$ является решением уравнения (4).

Пусть μ_1 вещественно. Рассмотрим ситуацию, когда $\varepsilon_n = y_n^2 - y_n^1 = \delta\mu_1^{n-k+1}$, $|\mu_1| > 1$. Разность между значениями решений y_n^1 и y_n^2 в начальных точках не превосходит δ, в то время как

$$\max_{nh \leqslant X} |\varepsilon_n| = \delta|\mu_1|^{X/h-k+1}$$

экспоненциально растет с ростом числа шагов $N = X/h$.

Таким образом, в случае $\mu_1 = \max|\mu_i| > 1$ малые возмущения начальных данных могут приводить к катастрофическому возмущению решения уже при не очень большом числе шагов.

Пусть μ_1 невещественно; сеточные функции $\varepsilon_{n,1} = \mathrm{Re}\,(\delta\mu_1^{n-k+1})$ и $\varepsilon_{n,2} = \mathrm{Im}\,(\delta\mu_1^{n-k+1})$ будут решениями уравнения (3). Поскольку $\max\{|x|, |y|\} \geqslant |x + \mathrm{i}y|/\sqrt{2}$, то

$$\max\left\{\max_{nh\leqslant X}|\varepsilon_{n,1}|, \max_{nh\leqslant X}|\varepsilon_{n,2}|\right\} \geqslant \frac{\delta}{\sqrt{2}}|\mu_1|^{X/h-k+1}.$$

Таким образом, и в случае μ_1 невещественного, $|\mu_1| > 1$, решение уравнения (1) может сильно исказиться при малом возмущении начальных данных.

Рассмотрим случай, когда все корни уравнения (4) не превосходят по модулю 1, но среди корней, по модулю равных 1, есть p-кратный корень μ_1, $|\mu_1| = 1$, $p > 1$. Для простоты проведем построения для случая μ_1 вещественного. Сеточная функция $\varepsilon_n = \delta\mu_1^{n-k+1}\left(\dfrac{n}{k-1}\right)^{p-1}$ соответствует возмущению начальных данных $y_0, \ldots, y_{k-1}$ не более чем на δ, в то время как в конце отрезка возмущение имеет порядок $\delta(X/h)^{p-1}$.

Такой степенной (по отношению к числу узлов) рост влияния возмущения начальных данных иногда является допустимым. Однако на примере модельного уравнения $y' = My$ можно показать, что в случае p-кратного корня на границе единичного круга возмущение исходных данных сказывается более существенным образом. При $f(x, y) = My$ все

$l_m = M$ и уравнение (3) относительно ε_n является линейным разностным уравнением

$$\sum_{i=0}^{k}(a_{-i} - Mhb_{-i})\varepsilon_{n-i} = 0.$$

Соответствующее характеристическое уравнение имеет вид

$$\sum_{i=0}^{k}(a_{-i} - Mhb_{-i})\mu^{k-i} = 0. \tag{5}$$

Наложим ограничение

$$\sum_{i=0}^{k} b_{-i}\mu_1^{k-i} \neq 0.$$

В противном случае разностное уравнение (1) записывается в виде

$$\left(\sum_{i=0}^{k-1} c_{-i}y_{n-i} - h\sum_{i=0}^{k-1} d_{-i}f(x_{n-i}, y_{n-i})\right) - \\ - \mu_1\left(\sum_{i=0}^{k-1} c_{-i}y_{n-1-i} - h\sum_{i=0}^{k-1} d_{-i}f(x_{n-i-1}, y_{n-i-1})\right) = 0.$$

Рассмотрение таких уравнений не представляет интереса, поскольку решение более простого уравнения

$$\sum_{i=0}^{k-1} c_{-i}y_{n-i} - h\sum_{i=0}^{k-1} d_{-i}f(x_{n-i}, y_{n-i}) = 0$$

также оказывается решением уравнения (5).

Предположим также, что $a_0 \neq 0$. Справедлива

Теорема (без доказательства). 1. *Если $p > 2$, то среди корней уравнения* (5) *есть корень, удовлетворяющий неравенству*

$$|\mu_1(MH)| \geqslant \exp\left\{c|Mh|^{1/p}\right\}, \quad c > 0.$$

2. *Если $p = 2$, то или при $M > 0$, или при $M < 0$ среди корней уравнения* (5) *есть корень, удовлетворяющий неравенству*

$$|\mu_1(Mh)| \geqslant \exp\left\{c|Mh|^{1/p}\right\}.$$

Таким образом, при $p \geqslant 2$ или уравнению $y' = +|M|y$, или уравнению $y' = -|M|y$ соответствует рост возмущения решения в

$$\approx \exp\left\{c|Mh|^{1/p}\frac{X}{h}\right\} = \exp\left\{c|M|^{1/p}Xh^{1/p-1}\right\} \quad \text{раз.}$$

Возмущение решения растет быстрее любой степени числа шагов; такой рост возмущения уже при небольшом числе шагов также является недопустимым.

В связи со сказанным практически пригодными могут оказываться лишь схемы, удовлетворяющие следующему *условию α: все корни характеристического уравнения* (4) *лежат в единичном круге и на границе единичного круга нет кратных корней.*

Можно было бы подумать, что все дело только в округлениях и погрешностях исходных данных: если бы их не было, то, может быть, решение конечно-разностной задачи сходилось бы к решению дифференциальной? На самом деле для любой разностной схемы, не удовлетворяющей условию α, можно построить пример дифференциального уравнения с бесконечно дифференцируемой правой частью, для которого и при отсутствии округлений и погрешностей в исходных данных решение конечно-разностной задачи не стремится к решению дифференциальной при измельчении шага.

На первый взгляд может показаться целесообразным строить схемы с возможно большим порядком аппроксимации m ($E_0 = \cdots = E_m = 0$). Однако оказывается, что все схемы с большим m не удовлетворяют условию α.

Теорема (без доказательства). *В случаях:* а) *схема (6.4) явная, $m > k$;* б) *схема (6.4) неявная, k нечетно, $m > k+1$;* в) *схема (6.4) неявная, k четно, $m > k+2$, среди корней характеристического уравнения* (4) *имеется корень, по модулю больший* 1.

Далее будет показано, что при некоторых дополнительных условиях на погрешности начальных данных разностной задачи и вычислительную погрешность при выполнении условия α решение разностной задачи (1) сходится к решению дифференциальной задачи. Будет приведено выражение главного члена погрешности, из которого видно, что главный член ведет себя примерно одинаково для всех разностных схем одного и того же порядка точности, удовлетворяющих условию α. Однако это не означает, что на практике они являются примерно эквивалентными.

Рассмотрим поведение решений двух разностных схем второго порядка аппроксимации:

$$\frac{y_n - y_{n-1}}{h} = \frac{f(x_n, y_n) + f(x_{n-1}, y_{n-1})}{2}, \tag{6}$$

$$\frac{y_n - y_{n-2}}{2h} = f(x_{n-1}, y_{n-1}) \tag{7}$$

на примере модельного уравнения $y' = My$, $M = \text{const}$. Разностные схемы (6), (7) порождают конечно-разностные уравнения

$$y_n(1 - Mh/2) - y_{n-1}(1 + Mh/2) = 0,$$

$$y_n - 2Mhy_{n-1} - y_{n-2} = 0.$$

В первом случае решение уравнения для погрешности имеет вид

$$\varepsilon_n = \left(\frac{1 + Mh/2}{1 - Mh/2}\right)^n \varepsilon_0;$$

Это решение растет при $M > 0$. Это естественно, поскольку для дифференциальной задачи разность двух решений $\varepsilon(x)$ с различными начальными условиями записывается в виде

$$\varepsilon(x_n) = \exp\{M(x_n - x_0)\}\varepsilon(x_0)$$

и также растет при $M > 0$. При $M < 0$ как ε_n, так и $\varepsilon(x_n)$ убывают. Решение второго уравнения имеет вид $c_1\mu_1^n + c_2\mu_2^n$, где μ_1 и μ_2 — корни характеристического уравнения $\mu_2 - 2Mh\mu - 1 = 0$, т. е. $\mu_{1,2} = Mh \pm \sqrt{1 + (Mh)^2}$. Имеем $\mu_1 = Mh + \sqrt{1 + (Mh)^2} = 1 + Mh + \dfrac{(Mh)^2}{2} + O((Mh)^3) = \exp\{Mh(1 + O(Mh))\}$. Здесь и далее мы пользуемся формулой Тейлора $\sqrt{1+\varepsilon} = 1 + \dfrac{\varepsilon}{2} + O(\varepsilon^2)$. Отсюда следует равенство $\mu_1^n = \exp\{Mnh(1+O(Mh)^2)\}$. Таким образом, слагаемое $c_1\mu_1^n$ соответствует решению разностного уравнения, ведущему себя качественно так же, как решение дифференциального уравнения. Аналогичным образом получаем

$$\mu_2 = Mh - \sqrt{1 + M^2h^2} = -1 + Mh - \frac{M^2h^2}{2} + O((Mh)^3) =$$
$$= -\exp\{-Mh(1 + O(Mh))\}.$$

Имеем $\mu_2^n = (-1)^n \exp\{-Mnh(1 + O(Mh))\}$; слагаемое μ_2^n ведет себя качественно иначе, чем решение дифференциального уравнения, и, что особенно существенно, *оно возрастает по модулю при $M < 0$, в то время как точное решение убывает.* Рассуждая так же, как выше, заключаем, что вычислительная погрешность может исказить решение на величину порядка $\delta\exp\{-Mnh(1 + O(Mh))\}$. При $M < 0$ и большом значении $|Mnh|$ эта величина может оказаться недопустимо большой, особенно на фоне убывающего решения $e^{M(x-x_0)}$.

Поскольку расматриваемый метод дает неудовлетворительный результат для такой простейшей модельной задачи, его вряд ли можно рекомендовать для широкого употребления, тем более в стандартных программах численного интегрирования дифференциальных уравнений.

Мы отбраковали второй метод на примере задачи, где $M < 0$ и величина $\delta e^{M|x|}$ недопустимо большая. В последние сорок лет в приложениях часто стали встречаться задачи с резкими переходными процессами, где решение существенно меняется на малом промежутке времени. Типичной модельной задачей является задача Коши для уравнения $y' = My$, $M < 0$, где величина $|M|X$ настолько велика, что число шагов порядка $|M|X$ является недопустимым. Если при разумном числе шагов по времени $|M|h \gg 1$, то использование первой из рассматрива-

емых разностных схем также может привести к неудовлетворительным результатам. Для этой схемы имеем

$$\mu_1 = \frac{1 + Mh/2}{1 - Mh/2} = \frac{1 + 2/(Mh)}{-1 + 2/(Mh)} \approx -\exp\left\{\frac{4}{Mh} + O\left(\frac{1}{(Mh)^2}\right)\right\}.$$

Таким образом, решение разностного уравнения имеет вид

$$y_n = (-1)^n \exp\left\{\left(\frac{4}{Mh} + O\left(\frac{1}{(Mh)^2}\right)\right) n\right\} y_0 =$$
$$= (-1)^n \exp\left\{\left(\frac{4}{Mh^2} + O\left(\frac{1}{(M^2h^3)}\right)\right)(x_n - x_0)\right\} y_0$$

и существенно отличается от точного решения дифференциальной задачи $y(x_n) = y_0 e^{M(x_n - x_0)}$, например, при $|M|h^2 \gg 1$ (решение разностного уравнения по модулю близко к 1, а решение дифференциального уравнения мало).

В качестве итога проводимого выше анализа свойств первого метода можно заключить, что этот метод применим для решения довольно широкого круга задач. В то же время существуют определенные типы задач, называемые *жесткими* (моделируемые случаем $M < 0$, $|M|X$ очень велико), в которых к применению этого метода нужно отнестись с определенной осторожностью. Для решения таких задач разработаны специальные методы (см. § 9).

§ 8. Оценка погрешности конечно-разностных методов

Произведем оценку погрешности приближенных решений, которые получаются при использовании конечно-разностных методов вида (5.2), удовлетворяющих условию α. Получаемые в процессе реальных вычислений величины y_n, являющиеся приближениями к значениям $y(x_n)$, связаны на самом деле не соотношением (5.2), а соотношением

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = \delta_n, \tag{1}$$

где δ_n может быть отлична от нуля (причины появления δ_n уже были указаны в § 4).

С другой стороны, в § 5, 6 установлено, что значения $y(x_n)$ точного решения дифференциальной задачи удовлетворяют соотношению

$$\sum_{i=0}^{k} a_{-i} y(x_{n-i}) - h \sum_{i=0}^{k} b_{-i} f\Big(x_{n-i}, y(x_{n-i})\Big) = h r_n; \tag{2}$$

при соответствующем подборе коэффициентов a_{-i} и b_{-i} оказывалось, что $r_n = O(h^m)$, где $m > 0$. Вычитая (2) из (1), получим уравнения для погрешности $R_n = y_n - y(x_n)$. На основании формулы Лагранжа имеем равенство

$$f(x_{n-i}, y_{n-i}) - f\Big(x_{n-i}, y(x_{n-i})\Big) = l_{n-i}R_{n-i}, \tag{3}$$

где $l_{n-i} = f_y(x_{n-i}, \tilde{y}_{n-i})$ и $\tilde{y}_{n-i}$ лежит между y_{n-i} и $y(x_{n-i})$. С учетом (3) разность соотношений (1) и (2) запишется в виде

$$\sum_{i=0}^{k} a_{-i}R_{n-i} - h\sum_{i=0}^{k} b_{-i}l_{n-i}R_{n-i} = g_n, \tag{4}$$

где $g_n = \delta_n - hr_n$.

Теорема (об оценке погрешности). *Пусть разностная схема удовлетворяет условию α и $|f_y| \leqslant L$ при $x_0 \leqslant x \leqslant x_0 + X$. Тогда при $x_0 \leqslant x_n \leqslant x_0 + X$ выполняется неравенство*

$$|R_n| \leqslant c(L, X)\left(\max_{0\leqslant i<k}|R_i| + \sum_{j=k}^{n}|g_j|\right), \tag{5}$$

где $c(L, X) < \infty$ — некоторая постоянная, зависящая от коэффициентов a_{-i}, b_{-i} и от L и X.

Доказательство. Для доказательства нам понадобится частный случай леммы из § 3 гл. 6: пусть все собственные значения матрицы A лежат в круге $|\lambda| \leqslant q$ и на границе круга нет кратных корней; тогда можно указать матрицу C такую, что $\|D\|_\infty \leqslant q$, где $D = C^{-1}AC$.

Для удобства оценки преобразуем уравнение (4) в одношаговое векторное уравнение. В дальнейшем для определенности предполагаем h настолько малым, что $|hb_0L| \leqslant |a_0/2|$, тогда коэффициент $a_0 - hb_0l_n$ при R_n в (4) по модулю не менее $|a_0/2|$. Перенося в (4) все слагаемые, не содержащие R_n, в правую часть и деля на коэффициент при R_n, получим равенство

$$R_n = \sum_{i=1}^{k} \frac{-a_{-i} + hb_{-i}l_{n-i}}{a_0 - hb_0l_n}R_{n-i} + \frac{g_n}{a_0 - hb_0l_n}. \tag{6}$$

Положим

$$\frac{-a_{-i} + hb_{-i}l_{n-i}}{a_0 - hb_0l_n} - \left(\frac{-a_{-i}}{a_0}\right) = hv_{in};$$

$$v_{in} = \frac{b_{-i}a_0l_{n-1} - a_{-i}b_0l_n}{(a_0 - hb_0l_n)a_0}, \quad |v_{in}| \leqslant 2\,\frac{|b_{-i}a_0| + |a_{-i}b_0|}{|a_0|^2}\,L.$$

Введем в рассмотрение векторы

$$\mathbf{Z}_n = (R_n, \ldots, R_{n-k+1})^T.$$

Соотношение (6) равносильно равенству

$$\mathbf{Z}_n = A\mathbf{Z}_{n-1} + hV_n\mathbf{Z}_{n-1} + \mathbf{W}_n, \tag{7}$$

где

$$V_n = \begin{pmatrix} v_{1n} & \dots & v_{kn} \\ 0 & \dots & 0 \\ \cdot & \dots & \cdot \\ 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{W}_n = \begin{pmatrix} \dfrac{g_n}{a_0 - hb_0l_n} \\ 0 \\ \cdot \\ 0 \end{pmatrix},$$

$$A = \begin{pmatrix} -\dfrac{a_{-1}}{a_0} & -\dfrac{a_{-2}}{a_0} & \dots & -\dfrac{a_{1-k}}{a_0} & -\dfrac{a_{-k}}{a_0} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

Действительно, приравнивая первые компоненты векторов в правой и левой частях (7), мы получим равенство (6), а приравнивая остальные компоненты, получаем тождества

$$R_{n-i} = R_{n-i}, \quad i = 1, \dots, k-1.$$

Вычислим характеристический многочлен матрицы A:

$$P(\lambda) = \det(A - \lambda E) = \det \begin{pmatrix} -\dfrac{a_{-1}}{a_0} - \lambda & -\dfrac{a_{-2}}{a_0} & \dots & -\dfrac{a_{1-k}}{a_0} & -\dfrac{a_{-k}}{a_0} \\ 1 & -\lambda & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & \dots & 1 & -\lambda \end{pmatrix}.$$

Для этого умножим первый столбец на λ и прибавим ко второму, затем — второй на λ и прибавим к третьему и т. д. В результате получим

$$P(\lambda) = \det \begin{pmatrix} p_1(\lambda) & p_2(\lambda) & \dots & p_{k-1}(\lambda) & p_k(\lambda) \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

где $p_1(\lambda) = -\frac{a_{-1}}{a_0} - \lambda$, $p_2(\lambda) = -\frac{a_{-2}}{a_0} + \lambda\left(-\frac{a_{-1}}{a_0} - \lambda\right)$, $\ldots$, $p_k(\lambda) = -\frac{a_{-k}}{a_0} + \lambda\left(-\frac{a_{1-k}}{a_0} + \lambda\left(-\frac{a_{2-k}}{a_0} + \cdots + \lambda\left(-\frac{a_{-1}}{a_0} - \lambda\right)\cdots\right)\right)$. Раскрывая определитель по последнему столбцу, имеем $P(\lambda) = (-1)^{k+1}p_k(\lambda)$ или, что то же самое,

$$(-1)^k P(\lambda) = \lambda^k + \frac{a_{-1}}{a_0}\lambda^{k-1} + \cdots + \frac{a_{-k}}{a_0} = \\ = \frac{1}{a_0}(a_0\lambda^k + a_{-1}\lambda^{k-1} + \cdots + a_{-k}).$$

Характеристическое уравнение матрицы A оказалось пропорциональным характеристическому уравнению (7.4) разностной схемы. Согласно предположению α все корни характеристического уравнения матрицы A лежат в круге $|z| \leqslant 1$ и на границе круга нет кратных. Поэтому по отношению к матрице A условие леммы выполнено со значением $q = 1$. Следовательно, существует матрица C такая, что $C^{-1}AC = D$ и $\|D\|_\infty \leqslant 1$. Произведем в уравнении (7) замену переменных $\mathbf{Z}_n = C\mathbf{z}_n$. После умножения слева на C^{-1} оно приведется к виду

$$\mathbf{z}_n = D\mathbf{z}_{n-1} + hv_n\mathbf{z}_{n-1} + \mathbf{w}_n, \tag{8}$$

где

$$D = C^{-1}AC, \quad v_n = C^{-1}V_nC, \quad \mathbf{w}_n = C^{-1}\mathbf{W}_n.$$

В матрице V_n ненулевые элементы находятся только в первой строке; поэтому

$$\|V_n\|_\infty \leqslant \left(2\sum_{i=1}^{k}\frac{|b_{-i}a_0| + |a_{-i}b_0|}{a_0^2}\right)L \cdot = vL$$

и

$$\|v_n\|_\infty \leqslant \|C^{-1}\|_\infty \|V_n\|_\infty \|C\|_\infty \leqslant vL\,\|C^{-1}\|_\infty \|C\|_\infty.$$

Имеем

$$\|\mathbf{w}_n\|_\infty \leqslant \|C^{-1}\|_\infty \|\mathbf{W}_n\|_\infty = \|C^{-1}\|_\infty \left|\frac{g_n}{a_0 - hb_0l_n}\right| \leqslant 2\,\|C^{-1}\|_\infty \frac{|g_n|}{|a_0|}.$$

После оценки норм слагаемых в правой части (8) можно записать неравенство

$$\|\mathbf{z}_n\|_\infty \leqslant \beta|g_n| + (1 + \gamma Lh)\,\|\mathbf{z}_{n-1}\|_\infty, \tag{9}$$

где

$$\beta = 2\,\frac{\|C^{-1}\|_\infty}{|a_0|}, \quad \gamma = v\,\|C^{-1}\|_\infty \|C\|_\infty.$$

Для оценки $\|\mathbf{z}_n\|_\infty$ через $\|\mathbf{z}_{n-1}\|_\infty$ и величины $|g_n|$ поступим следующим образом. Выпишем неравенства (9) при $j = n, \ldots, k$. В правой

части оценки $\|\mathbf{z}_n\|_\infty$ через $\|\mathbf{z}_{n-1}\|_\infty$ оценим $\|\mathbf{z}_{n-1}\|_\infty$ через $\|\mathbf{z}_{n-2}\|_\infty$, затем в правой части получившегося неравенства оценим $\|\mathbf{z}_{n-2}\|_\infty$ через $\|\mathbf{z}_{n-3}\|_\infty$ и т. д. В результате получим

$$\|\mathbf{z}_n\|_\infty \leqslant \beta|g_n| + (1+\gamma Lh)\big(\beta|g_{n-1}| + (1+\gamma Lh)(\beta|g_{n-2}| + \cdots + (1+\gamma Lh)(\beta|g_k| + (1+\gamma Lh)\|\mathbf{z}_{k-1}\|_\infty)\dots)\big),$$

или, что то же самое,

$$\|\mathbf{z}_n\|_\infty \leqslant \beta\sum_{j=k}^{n}(1+\gamma Lh)^{n-j}|g_j| + (1+\gamma Lh)^{n-k+1}\|\mathbf{z}_{k-1}\|_\infty. \tag{10}$$

Упростим эту оценку, одновременно несколько загрубив ее. При $x_0 \leqslant x_j \leqslant x_n \leqslant x_0 + X$ имеем $(n-j)h \leqslant X$; поэтому

$$(1+\gamma Lh)^{n-j} \leqslant \exp\{\gamma Lh(n-j)\} \leqslant \exp\{\gamma LX\}.$$

Теперь из (10) получаем

$$\|\mathbf{z}_n\|_\infty \leqslant \exp\{\gamma LX\}\left(\beta\sum_{j=k}^{n}|g_j| + \|\mathbf{z}_{k-1}\|_\infty\right). \tag{11}$$

Справедливы неравенства

$$|R_n| \leqslant \|\mathbf{Z}_n\|_\infty \leqslant \|C\|_\infty\|\mathbf{z}_n\|_\infty,$$
$$\|\mathbf{z}_{k-1}\|_\infty \leqslant \|C^{-1}\|_\infty\|\mathbf{Z}_{k-1}\|_\infty = \|C^{-1}\|_\infty \max_{0\leqslant i<k}|R_i|,$$

и поэтому

$$\|\mathbf{Z}_n\|_\infty \leqslant \|C\|_\infty \exp\{\gamma LX\}\left(\beta\sum_{j=k}^{n}|g_j| + \|C^{-1}\|_\infty\|\mathbf{Z}_{k-1}\|_\infty\right). \tag{12}$$

Далее получаем оценку погрешности

$$|R_n| \leqslant \exp\{\gamma LX\}\left(M_1\sum_{j=k}^{n}|g_j| + M_2\max_{0\leqslant i<k}|R_i|\right),$$

где $M_1 = \beta\|C\|_\infty$, $M_2 = \|C\|_\infty\|C^{-1}\|_\infty$ и γ — некоторые постоянные, зависящие только от коэффициентов a_{-i}, b_{-i} исходной разностной схемы. В частности, M_1 и M_2 зависят только от коэффициентов a_{-i}. Утверждение теоремы доказано.

Подставляя в (12) значение $g_j = \delta_j - hr_j$, получим искомую оценку погрешности

$$|R_n| \leqslant \exp\{\gamma LX\}\left(M_1\sum_{j=k}^{n}\big(|\delta_j| + h|r_j|\big) + M_2\max_{0\leqslant i<k}|R_i|\right). \tag{13}$$

Из оценки (13) видно, что для сходимости решения разностного уравнения к решению дифференциального достаточно выполнения условий

$$\sum_{j=k}^{n} |\delta_j| \to 0, \quad h\sum_{j=k}^{n} |r_j| \to 0, \quad \max_{0\leqslant i<k} |R_i| \to 0.$$

Оценка погрешности (13) во многих случаях является существенно завышенной. Например, для методов Адамса можно получить оценку погрешности, остающуюся ограниченной в случае $f_y \leqslant -b < 0$ при сколь угодно большой длине промежутка интегрирования в предположении, что погрешности округления δ_j и погрешности аппроксимации r_j равномерно ограничены: $|\delta_j| \leqslant \delta$, $|r_j| \leqslant r$. Заметим, что из оценки погрешности для одношаговых методов (§ 4) следует оценка погрешности метода Эйлера, являющегося частным случаем методов Адамса. В то же время из (13) при таких предположениях нельзя получить равномерной ограниченности погрешности интегрирования. Для получения более реального представления о величине погрешности полезно располагать выражением для главного члена погрешности.

Наметим путь получения этого выражения. При достаточно гладкой функции $f(x, y)$ согласно (6.5) справедливо равенство

$$r_n = E_{m+1}h^m y^{(m+1)}(x_n) + o(h^m).$$

Поскольку $\sum\limits_{i=0}^{k} b_{-i} = 1$, то это выражение можно переписать в более удобном для нас виде

$$r_n = h^m \sum_{i=0}^{k} b_{-i}E_{m+1}y^{(m+1)}(x_n) + o(h^m).$$

Предположим, что вычислительная погрешность мала по сравнению с погрешностью аппроксимации, точнее, $\max\limits_j |\delta_j| = \delta = o(h^{m+1})$.

При выполнении условий теоремы об оценке погрешности решение разностной задачи сходится к решению дифференциальной, поэтому справедливо равенство $l_n = f_y(x_n, \tilde{y}_n) = f_y(x_n, y(x_n)) + o(1)$. С учетом выписанных выше соотношений равенство (4) можно переписать в виде

$$L_h^0\left(\frac{R_n}{h^m}\right) = \frac{1}{h}\sum_{i=0}^{k} a_{-i}\frac{R_{n-i}}{h^m} - \sum_{i=0}^{k} b_{-i}\Big(f_y(x_{n-i}, y(x_{n-i}))\frac{R_{n-i}}{h^m} - E_{m+1}y^{(m+1)}(x_n)\Big) = o(1).$$

Таким образом, сеточная функция $z_n = R_n/h^m$ приближенно удовлетворяет сеточному уравнению $L_h^0(z_n) = 0$, которое получается при аппроксимации уравнения

$$z' - \Big(f_y\big(x, y(x)\big)z - E_{m+1}y^{(m+1)}(x)\Big) = 0. \tag{14}$$

В предположении, что $z_0, \dots, z_{k-1} = o(1)$, рассуждая так же, как при доказательстве теоремы об оценке погрешности, получаем, что z_n близко к решению уравнения (14) при начальном условии $z(x_0) = 0$. Это решение можно выписать в явном виде:

$$z_n \approx z(x_n) = -E_{m+1} \int_{x_0}^{x_n} \exp\left\{ \int_x^{x_n} f_y(t,\, y(t))\, dt \right\} y^{(m+1)}(x)\, dx. \tag{15}$$

Таким образом,

$$R_n \approx h^m z(x_n).$$

Сформулируем аналогичный результат, относящийся к случаю интегрирования системы уравнений. Для системы уравнений $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$, когда $\mathbf{y}$ и $\mathbf{f}$ — векторы: $\mathbf{y} = (y_1, \dots, y_l)^T$, $\mathbf{f} = (f_1, \dots, f_l)^T$, выражение главного члена погрешности (15) имеет вид

$$\mathbf{y}_n - \mathbf{y}(x_n) \sim h^m \mathbf{z}(x_n),$$

$$\mathbf{z}(x_n) = \int_{x_0}^{x_n} W(x, x_n) \mathbf{y}^{(m+1)}(x)\, dx.$$

Здесь матрица $W(a, b)$ является решением матричного дифференциального уравнения

$$W'(a, b) = f_y(b, \mathbf{y}(b)) W(a, b)$$

при начальном условии $W(a, a) = E$, где E — единичная матрица, $f_y(x, u)$ — матрица с элементами $\partial f_i / \partial y_j \,|_{x,u}$, $i, j = 1, \dots, l$.

Проведенные выше рассуждения, каждый этап которых может быть строго обоснован, иногда облекают в более грубую форму. Точное решение дифференциальной задачи удовлетворяет соотношению

$$L_h(y(nh)) \approx E_{m+1} h^m y^{(m+1)}(x),$$

где

$$L_h z_n = \frac{1}{h} \sum_{i=0}^{k} a_{n-i} z_{n-i} - \sum_{i=0}^{k} b_{-i} f(x_{n-i},\, z_{n-i}),$$

а приближенное решение — соотношению

$$L_h(y_n) = 0,$$

поэтому их разность удовлетворяет равенству

$$L_h(y_n) - L_h(y(nh)) \approx -E_{m+1} h^m y^{(m+1)}(x).$$

Поскольку y_n и $y(nh)$ близки, это соотношение можно записать в виде

$$L_h'(y(nh))(y_n - y(nh)) \approx -E_{m+1} h^m y^{(m+1)}(x);$$

здесь L_h' — производная оператора L_h. Поскольку операторы L и L_h в определенном смысле близки, то можно написать

$$L'(y(nh))(y_n - y(nh)) \approx -E_{m+1} h^m y^{(m+1)}(x).$$

Напомним, что производная оператора L определяется равенством

$$L'(y(x))\delta = \lim_{t\to 0} \frac{L(y(x)+t\delta(x)) - L(y(x))}{t} =$$
$$= \lim_{t\to 0} \frac{((y+t\delta)' - f(x, y+t\delta)) - (y' - f(x, y))}{t} = \delta' - f_y(x, y(x))\delta.$$

Отсюда получаем приближенное равенство

$$\delta' - f_y(x, y(x))\delta \approx -E_{m+1}h^m y^{(m+1)}(x),$$

где $\delta = y_n - y(x_n)$, и затем — (15).

Второй путь вывода выражения для главного члена погрешности уже не поддается непосредственному обоснованию и в принципе может привести к неверным заключениям. Это видно хотя бы из того, что нигде не проявилось условие α, без которого не имеет места сам факт малости величины $y_n - y(x_n)$. Справедливость получаемых на таком пути результатов требует специального обоснования.

В то же время его следует признать крайне полезным, поскольку не известно ни одного противоречащего примера, когда бы его применение приводило к неправильному выражению для главного члена погрешности *в случае, если решение разностной задачи сходится к решению дифференциальной.*

Для некоторых методов, например Адамса и Рунге—Кутта, выражение главного члена погрешности, как правило, дает реальное представление о величине погрешности. В других случаях, например для метода (7.6), являющегося простейшим случаем так называемого метода Милна, это выражение следует рассматривать как некоторую оценку снизу для реальной величины погрешности.

Рассмотрим случай метода (7.6) и уравнения $y' = My$. Точное решение имеет вид

$$y(x) = y_0 e^{M(x-x_0)},$$

а $E_3 = 1/6$. Согласно (15) имеем

$$z(x_n) = -\frac{1}{6}\int_{x_0}^{x_n} e^{M(x-x_0)} M^3 y_0 e^{M(x-x_0)}\, dx = -\frac{1}{6} M^3 y_0 e^{M(x_n-x_0)}(x_n - x_0). \quad (16)$$

Оценим $|ve^{-v}|$ в области $v > 0$. Поскольку эта функция стремится к нулю при $v \to 0$ и при $v \to \infty$, то ее наибольшее значение принимается в точке, где ее производная равна нулю, т. е. при $v = 1$. Отсюда следует, что $|ve^{-v}| \leqslant e^{-1}$ при $v > 0$. Рассмотрим случай $M < 0$. Подставляя $v = -M(x_n - x_0)$, получим $|Me^{M(x_n-x_0)}(x_n - x_0)| \leqslant e^{-1}$. Отсюда и из (16) следует $|z(x_n)| \leqslant M^2|y_0|/(6e)$, и, таким образом, главный член погрешности равномерно ограничен при $x_0 < x_n < \infty$. В то же время из рассмотрения этого модельного примера, проведенного в § 7, вытекает, что реальная величина погрешности довольно сильно растет вследствие большого влияния погрешности начальных значений y_0, y_1.

Казалось бы, в обрисованной выше ситуации есть какое-то противоречие. Говорится о главном члене погрешности, но в то же время утверждается, что он не является определяющим в реальной величине погрешности. Дело заключается в следующем.

При получении главного члена погрешности имелось в виду, что длина промежутка интегрирования фиксирована, а δ/h и h стремятся к нулю. При рассмотрении модельного примера в § 7 речь шла о поведении погрешности при h фиксированном, $x_n - x_0 \to 0$.

Как показало рассмотрение этого модельного примера, влияние погрешности исходных данных существенно уже при не очень больших значениях $|M|(x_n - x_0)$, поэтому широкое применение метода (7.6) в реальной практике является нецелесообразным, несмотря на малое значение главного члена погрешности при δ/h, $h \to 0$ и $x_n - x_0$ фиксированном.

§ 9. Особенности интегрирования систем уравнений

Проводившиеся выше построения, в частности расчетные формулы, применимы без всяких изменений в случае систем уравнений

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}). \tag{1}$$

Формальное отличие состоит в том, что в соответствующих соотношениях вместо скалярных величин участвуют некоторые матрицы или тензоры.

Для выявления особенностей, которые могут возникнуть при численном интегрировании систем обыкновенных дифференциальных уравнений, рассмотрим модельный пример линейной системы с постоянными коэффициентами

$$\mathbf{y}' = A\mathbf{y}. \tag{2}$$

В случае использования конечно-разностной аппроксимации (5.2) соответствующая система конечно-разностных уравнений имеет вид

$$\sum_{i=0}^{k} \frac{a_{-i}\mathbf{y}_{n-i}}{h} - \sum_{i=0}^{k} b_{-i}A\mathbf{y}_{n-i} = \mathbf{0}. \tag{3}$$

Для простоты предположим, что жорданова форма матрицы простая: $C^{-1}AC = \Lambda$, Λ — диагональная матрица с диагональными элементами $\lambda_1, \ldots, \lambda_l$. Положим $C^{-1}\mathbf{y}_n = \mathbf{z}_n$ и умножим систему (3) слева на C^{-1}. Получим

$$\sum_{i=0}^{k} \frac{a_{-i}\mathbf{z}_{n-i}}{h} - \sum_{i=0}^{k} b_{-i}C^{-1}AC\mathbf{z}_{n-i} = \mathbf{0}$$

или

$$\sum_{i=0}^{k} \frac{a_{-i}\mathbf{z}_{n-i}}{h} - \sum_{i=0}^{k} b_{-i}\Lambda\mathbf{z}_{n-i} = \mathbf{0}.$$

Эта система распадается на систему скалярных конечно-разностных уравнений относительно компонент z_{pn} векторов $\mathbf{z}_n = (z_{1n}, \ldots, z_{in})^T$:

$$\frac{1}{h}\sum_{i=0}^{k} a_{-i}z_{p,n-i} - \sum_{i=0}^{k} b_{-i}\lambda_p z_{p,n-i} = 0, \quad p = 1, \ldots, l. \tag{4}$$

Соотношение (4) совпадает с конечно-разностной аппроксимацией для уравнения

$$z_p' = \lambda_p z_p. \tag{5}$$

Если в (2) перейти к новой неизвестной вектор-функции $\mathbf{z} = C^{-1}\mathbf{y}$ и умножить (2) слева на матрицу C^{-1}, то получится система скалярных уравнений (5), $p = 1, \ldots, l$. В соответствии с определением вектор-функций $\mathbf{z}(x)$ и $\mathbf{z}_n$ имеем равенство

$$\mathbf{z}_n - \mathbf{z}(x_n) = C^{-1}\Big(\mathbf{y}_n - \mathbf{y}(x_n)\Big).$$

Следовательно, для получения решения $\mathbf{y}(x)$ с малой погрешностью необходимо и достаточно, чтобы решения $\mathbf{z}_p(x)$ уравнения (5) получались с малой погрешностью в случае интегрирования с помощью аппроксимации (4). При этом имеется в виду, что есть соответствие между аппроксимациями начальных условий

$$\mathbf{y}_j = C\mathbf{z}_j, \quad j = 0, \ldots, k-1.$$

Проводя аналогичные построения, можно получить тот же вывод и по отношению к методам Рунге-Кутта.

Решение уравнения (5) имеет вид $z_p = z_p^0 \exp\{\lambda_p(x - x_0)\}$ и существенно изменяется при изменении x на расстояние $\Delta x = |1/\lambda_p|$, т. е. характерный размер изменения решения порядка $1/|\lambda_p|$. Если говорить о векторе $\mathbf{z}(x)$ как о едином целом, то характерный размер его изменения — величина порядка $1/\max\limits_p |\lambda_p|$; точно такой же порядок характерного изменения будет и у вектора $\mathbf{y}(x)$.

Шаг интегрирования должен быть существенно меньше характерного размера изменения решения, т. е. $h \ll \dfrac{1}{\max\limits_p |\lambda_p|}$. Отсюда следует оценка снизу для числа шагов интегрирования

$$N = \frac{X}{h} \gg \max_p |\lambda_p| \cdot X.$$

Если число шагов, много большее величины $\max\limits_p |\lambda_p| \cdot X$, неприемлемо по затратам машинного времени, то желательно применить методы, использующие специфику поведения решения.

В случае, когда

$$|\lambda_p| X \gg 1 \quad \text{при всех} \quad p,$$

для описания решения можно было бы применить асимптотические методы. Однако на практике часто встречаются задачи, когда это условие не выполнено, и поэтому применение асимптотических методов невозможно или крайне затруднительно.

Конечно, возникает вопрос, о каких проблемах идет речь, поскольку решение системы $\mathbf{y}' = A\mathbf{y}$ выписывается в явном виде? Дело в том, что мы говорим об этой задаче как о модельной; реально же метод применяется для решения какой-то, как правило, сложной задачи, и мы смотрим, как ведет себя метод в применении к простейшей задаче, где все выписывается в явном виде.

Широкий круг прикладных проблем сводится к решению задачи Коши для так называемых *жестких систем* дифференциальных уравнений. В частности, к таким системам относятся системы уравнений, возникающие при применении методов установления

$$\frac{d\mathbf{x}}{dt} + \nabla f = \mathbf{0}, \qquad \frac{d^2\mathbf{x}}{dt^2} + \gamma \frac{d\mathbf{x}}{dt} + \nabla f = \mathbf{0}$$

при минимизации функций f, у которых линии уровня имеют форму эллипсоидов с большим разбросом полуосей.

В качестве модели таких систем берется система уравнений

$$\mathbf{y}' = A\mathbf{y}, \tag{6}$$

удовлетворяющая определенным условиям на собственные значения матрицы A. Не существует установившегося определения жестких систем. Обычно систему (6) относят к классу жестких, если величина $(\max\limits_p \operatorname{Re}\lambda_p) \cdot X$ не является большим положительным числом, а величина $(\max\limits_p |\lambda_p|) X \gg 1$ и

а) величина $(\max\limits_p |\operatorname{Im}\lambda_p|)\, X$ не является большим положительным числом, или

б) $\dfrac{|\operatorname{Im}\lambda_p|}{b - \operatorname{Re}\lambda_p} \leqslant c$ при умеренных значениях b и c.

Нелинейную систему $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ относят к классу жестких систем, если при всех $\bar{x}$ из некоторого отрезка длины $\bar{X} > 0$, принадлежащего области интегрирования, система уравнений

$$\mathbf{y} = \mathbf{f}_y(\bar{x}, \mathbf{y}(\bar{x}))\mathbf{y}$$

относится к классу жестких систем в смысле приведенного выше определения.

Из проводившихся выше рассуждений для случая системы (6) видно, что численное решение задачи Коши для таких систем требует разработки специальных методов решения. Такие методы в настоящее время разработаны, и на их основе созданы соответствующие комплексы стандартных программ.

Рассмотрим простейшие варианты наиболее распространенных методов решения жестких систем.

1. Пусть уже найдено приближение $\mathbf{y}_n$ к значению $\mathbf{y}(x_n)$ и ищется приближение к значению $\mathbf{y}(x_{n+1})$, $x_{n+1} - x_n = H$. Разложим правую часть $\mathbf{f}(x, \mathbf{y})$ в ряд Тейлора в точке $(x_n, \mathbf{y}_n)$

$$\mathbf{f}(x, \mathbf{y}) \sim \mathbf{f}(x_n, \mathbf{y}_n) + \frac{\partial \mathbf{f}(x_n, \mathbf{y}_n)}{\partial x}(x - x_n) + \mathbf{f}_y\Big(x_n, \mathbf{y}(x_n)\Big)(\mathbf{y} - \mathbf{y}_n) + \cdots . \quad (7)$$

В прикладных задачах, как правило, возникают такие жесткие системы, что $\mathbf{f}(x, \mathbf{y})$ не зависит от x или меняется относительно медленно с изменением x. В этом случае главными членами в правой части являются первый и третий. За $\mathbf{y}_{n+1}$ примем значение в точке x_{n+1} решения системы

$$\mathbf{z}' = \mathbf{f}(x_n, \mathbf{y}_n) + \mathbf{f}_y\Big(x_n, \mathbf{y}(x_n)\Big)(\mathbf{z} - \mathbf{y}_n) \quad (8)$$

при начальном условии $\mathbf{z}(x_n) = \mathbf{y}_n$. Произведем следующую замену переменных $\mathbf{z}(x) - \mathbf{y}_n = \mathbf{u}(x)$, $x - x_n = t$ и введем обозначения

$$\mathbf{f}(x_n, \mathbf{y}_n) = \mathbf{b}, \quad \mathbf{f}_y\Big(x_n, \mathbf{y}(x_n)\Big) = A.$$

Для определения $\mathbf{z}(x_{n+1})$ нужно найти значение $\mathbf{u}(H)$ решения системы $\mathbf{u}' = A\mathbf{u} + \mathbf{b}$ при начальном условии $\mathbf{u}(0) = \mathbf{0}$.

Эта задача решается в явном виде, однако для ее решения требуется знать все собственные векторы и собственные значения матрицы A. Если размерность матрицы A сколько-нибудь большая, то найти их — довольно трудоемкая задача. Поэтому целесообразнее следующий путь нахождения $\mathbf{u}(H)$. Решение системы уравнений $\mathbf{u}' = A(t)\mathbf{u} + \mathbf{b}(t)$ при начальном условии $\mathbf{u}(0) = \mathbf{0}$ записывается в виде

$$\mathbf{u}(t) = \int_0^t W(\tau, t)\mathbf{b}(\tau)d\tau.$$

Матрица $W(\tau, t)$ является решением системы

$$\frac{dW(\tau, t)}{dt} = A(t)W(\tau, t)$$

при начальном условии $W(\tau, \tau) = E$.

В частном случае A, $\mathbf{b} = \text{const}$ имеем $\mathbf{u}(t) = \omega(t)\mathbf{b}$, где матрица $\omega(t)$ имеет вид

$$\omega(t) = \int_0^t \exp\{A(t-\tau)\}\,d\tau = A^{-1}\Big(\exp\{At\} - E\Big).$$

Значение $\omega(H)$ можно было бы попытаться вычислить, пользуясь разложением в ряд Тейлора

$$\omega(H) \sim H\sum_{i=1}^{\infty}\frac{1}{i!}(AH)^{i-1}. \tag{9}$$

Однако в случае жестких систем при реально допустимых значениях H величина $\|A\|H \gg 1$ и

а) для достижения приемлемой точности требуется взять слишком много слагаемых;

б) в тех случаях, когда требуемое для достижения нужной точности число слагаемых допустимо по числу арифметических действий, использование разложения (9) при $\|A\|H \gg 1$ может быть неприемлемо по другой причине: среди учитываемых в правой части слагаемых встречаются очень большие, и относительная погрешность, вызванная округлениями, недопустимо велика.

Матрица $\omega(t)$ удовлетворяет соотношению

$$\omega(t) = \omega\left(\frac{t}{2}\right)\left(2E + A\omega\left(\frac{t}{2}\right)\right), \tag{10}$$

где E — единичная матрица. Поэтому часто бывает целесообразно пойти по следующему пути. Выбираем s такое, что $\|A\|H\cdot 2^{-s} \ll 1$; основываясь на (9), вычисляем

$$\omega\left(\frac{H}{2^s}\right) \approx \frac{H}{2^s}\sum_{i=1}^{s}\frac{1}{i!}\left(A\frac{H}{2^s}\right)^{i-1},$$

а затем $\omega(H/2^{s-1}), \ldots, \omega(H/2), \omega(H)$ с помощью рекуррентной формулы (10).

В случае линейной системы $\mathbf{y}' = A(x)\mathbf{y}$ алгоритм решения задачи может быть несколько упрощен. Здесь в описанном выше алгоритме на каждом шаге возникает необходимость решения системы $\mathbf{y}' = A(x_n)\mathbf{y}$ при начальном условии $\mathbf{y}(x_n) = \mathbf{y}_n$. Тогда

$$\mathbf{y}_{n+1} = \varphi(H)\mathbf{y}_n, \qquad \text{где} \quad \varphi(H) = \exp\{AH\}, \quad A = A(x_n).$$

На первый взгляд может показаться разумным следующий путь. Матрица $\varphi(H)$ удовлетворяет соотношению

$$\varphi(t) = \varphi\left(\frac{t}{2}\right)\varphi\left(\frac{t}{2}\right), \tag{11}$$

поэтому зададимся некоторым s и вычисляем

$$\varphi\left(\frac{H}{2^s}\right) \approx \sum_{i=0}^{k} \frac{1}{i!}\left(A\frac{H}{2^s}\right)^i,$$

а затем

$$\varphi\left(\frac{H}{2^{s-1}}\right), \ldots, \varphi(H),$$

пользуясь рекуррентной формулой (11). Такой путь при большом s приводит к существенному накоплению погрешности. Поэтому положим $\psi(H) = \varphi(H) - E$, вычислим

$$\psi\left(\frac{H}{2^s}\right) \approx \sum_{i=1}^{k} \frac{1}{i!}\left(\frac{AH}{2^i}\right)^s$$

и затем $\psi(H/2^{s-1}), \ldots, \psi(H)$, пользуясь рекуррентной формулой

$$\psi(t) = \psi\left(\frac{t}{2}\right)\left(2E + \psi\left(\frac{t}{2}\right)\right).$$

Далее находим $\varphi(H) = E + \psi(H)$.

При $\mathbf{f}_x \not\equiv \mathbf{0}$ метод точности $O(h^2)$ получается, если в (7) также учесть второе слагаемое. Тогда потребуется аналогичным образом сконструировать точный метод решения вспомогательного уравнения

$$\mathbf{u}' = A\mathbf{u} + \mathbf{b} + \mathbf{c}t, \quad \mathbf{c} = \left.\frac{\partial \mathbf{f}}{\partial x}\right|_{(x_n,\, y_n)}.$$

В случае решения линейной задачи $\mathbf{u}' = A(x)\mathbf{u} + \mathbf{f}(x)$ можно предложить довольно простые методы точности $O(h^4)$.

2. Другая группа методов решения жестких задач строится следующим образом. Зададимся некоторым k и приблизим производную $\mathbf{y}'(x_n)$ односторонней аппроксимацией k-го порядка точности

$$\mathbf{y}'(x_n) \approx \frac{1}{h}\sum_{i=1}^{k}\frac{\nabla^i \mathbf{y}_n}{i} = \sum_{i=0}^{k}\frac{a_{-i}\mathbf{y}_{n-i}}{h}.$$

Выражение $\mathbf{f}(x_n, \mathbf{y}(x_n))$ оставим без изменения. Получим конечно-разностную аппроксимацию

$$\sum_{i=0}^{k}\frac{a_{-i}\mathbf{y}_{n-i}}{h} - \mathbf{f}(x_n, \mathbf{y}_n) = \mathbf{0}. \tag{12}$$

Рассмотрим случай модельного уравнения $\mathbf{y}' = M\mathbf{y}$, когда (12) превращается в конечно-разностное уравнение

$$\sum_{i=0}^{k}\frac{a_{-i}\mathbf{y}_{n-i}}{h} - M\mathbf{y}_n = \mathbf{0}.$$

Решение этого уравнения выписывается через корни характеристического уравнения

$$\sum_{i=0}^{k} a_{-i}\mu^{k-i} - hM\mu^k = 0.$$

При $k = 1, 2$ корни этого уравнения удовлетворяют условию $|\mu| \leqslant 1$ в области значений M : $\operatorname{Re} M < 0$; при $k = 3, 4, 5, 6$ — условию $|\mu| \leqslant 1$ в области значений M : $|\operatorname{Im} M| \leqslant -\alpha_k \operatorname{Re} M$, где $\alpha_k > 0$.

В нелинейном случае значение $\mathbf{y}_n$ требуется находить из нелинейной системы (12).

Алгоритмы решения жестких систем различаются способами нахождения начального приближения к решению (12) и алгоритмами приближенного решения (12). Рассмотрим простейший случай $k = 1$, когда (12) превращается в неявный метод Эйлера:

$$\frac{\mathbf{y}_n - \mathbf{y}_{n-1}}{h} - \mathbf{f}(x_n, \mathbf{y}_n) = \mathbf{0}. \tag{13}$$

Могло бы показаться разумным найти начальное приближение к $\mathbf{y}_n$ с помощью явной формулы Эйлера

$$\mathbf{y}_n^0 = \mathbf{y}_{n-1} + h\mathbf{f}(x_{n-1}, \mathbf{y}_{n-1}),$$

однако это не всегда целесообразно. В случае $\mathbf{f}(x, \mathbf{y}) \equiv M\mathbf{y}$ имеем $\mathbf{y}_n^0 = (1 + Mh)\mathbf{y}_{n-1}$ и при $|Mh| \gg 1$ может случиться, что такое приближение будет слишком далеко от истинного решения. Поэтому более безопасный, но не всегда самый эффективный вариант — это положить $\mathbf{y}_n^0 = \mathbf{y}_{n-1}$. Перепишем (13) в виде системы нелинейных уравнений относительно $\mathbf{y}_n$:

$$\mathbf{y}_n - \mathbf{y}_{n-1} - h\mathbf{f}(x_n, \mathbf{y}_n) = \mathbf{0},$$

и применим итерационный метод Ньютона. В данном конкретном случае интерполяционная формула Ньютона приобретает вид

$$\mathbf{y}_n^{k+1} = \mathbf{y}_n^k - \Big(E - h\mathbf{f}_y(x_n, \mathbf{y}_n^k)\Big)^{-1}\Big(\mathbf{y}_n^k - \mathbf{y}_{n-1} - h\mathbf{f}(x_n, \mathbf{y}_n^k)\Big), \tag{14}$$

где k — номер итерации.

В одном из методов решения жестких систем за $\mathbf{y}_n$ принимается значение $\mathbf{y}_n^1$, получаемое из (14) при $\mathbf{y}_n^0 = \mathbf{y}_{n-1}$. Тогда имеем

$$\mathbf{y}_n = \mathbf{y}_n^1 = \mathbf{y}_{n-1} + h\Big(E - h\mathbf{f}_y(x_n, \mathbf{y}_n)\Big)^{-1}\mathbf{f}(x_n, \mathbf{y}_{n-1}).$$

Рассмотрим случай скалярного уравнения $y' = My$; тогда

$$y_n = y_{n-1} + \frac{Mh}{1 - Mh}y_{n-1} = \frac{1}{1 - Mh}y_{n-1}.$$

Если $\operatorname{Re} M < 0$, в частности, если M действительно и $M < 0$, то при любом h имеем $\left|\dfrac{1}{1 - Mh}\right| < 1$ и погрешности, возникающие в процессе счета, убывают. Поэтому такой метод может быть применен к решению жестких систем.

3. Укажем еще один подобный метод решения задачи Коши для жестких систем довольно распространенного вида

$$\mathbf{z}' = F(\mathbf{z}, x), \quad \mathbf{u}' = G_0(\mathbf{z}, x) + G_1(\mathbf{z}, x)\mathbf{u},$$

с контролем локальной погрешности через некоторую величину δ.

Исходя из приближения $\mathbf{z}_n$, $\mathbf{u}_n$ делаем попытку интегрирования с шагом h_n. Используется аппроксимация первого порядка точности

$$\frac{\mathbf{z}_{n+1} - \mathbf{z}_n}{h_n} = F(\mathbf{z}_n, x_n) + \left[\frac{\partial F}{\partial \mathbf{z}}\Big|_{(\mathbf{z}_n, x_n)}\right](\mathbf{z}_{n+1} - \mathbf{z}_n),$$

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{h_n} = G_0(\mathbf{z}_{n+1}, x_{n+1}) + G_1(\mathbf{z}_{n+1}, x_{n+1})(\mathbf{u}_{n+1} - \mathbf{u}_n).$$

В результате двух шагов получаются приближения $\mathbf{z}^1_{n+2}$ и $\mathbf{u}^1_{n+2}$ к значениям $\mathbf{z}(x_n + 2h_n)$ и $\mathbf{u}(x_n + 2h_n)$; находятся приближения $\mathbf{z}^2_{n+2}$ и $\mathbf{u}^2_{n+2}$ в результате применения той же аппроксимация с двойным шагом $2h_n$; величины $\mathbf{R}_n = \mathbf{z}^1_{n+2} - \mathbf{z}^2_{n+2}$ и $\mathbf{r}_n = \mathbf{u}^1_{n+2} - \mathbf{u}^2_{n+2}$ используются для контроля шага интегрирования.

Если $\Delta_n = \sqrt{\|\mathbf{R}_n\|^2 + \|\mathbf{r}_n\|^2} \leqslant \delta$, то за окончательные приближенные значения $\mathbf{z}(x_n + 2h_n)$ и $\mathbf{u}(x_n + 2h_n)$ принимаются $\mathbf{z}^1_{n+2} + \mathbf{R}_n$ и $\mathbf{u}^1_{n+2} + \mathbf{r}_n$, соответственно. При $\Delta_n \leqslant \delta/4$ следующий шаг берется в два раза больше. При $\Delta_n > \delta$ делается попытка повторного интегрирования, исходя из точки x_n, но с уже вдвое меньшим шагом.

Получаемая в итоге аппроксимация имеет уже второй порядок.

Задача 1. На примере уравнения $u' + Mu = 0$ получить явную формулу, выражающую $u^1_{n+2} + r_n$ через u_n: $u^1_{n+2} + r_n = \lambda(M)u_n$. При каких M выполняется неравенство $|\lambda(M)| \leqslant 1$?

4. Как уже отмечалось ранее, явные методы численного интегрирования для жестких систем обыкновенных дифференциальных уравнений неприемлемы вследствие ограничений на шаг интегрирования. Действительно, рассмотрим модельное уравнение

$$u' + Mu = 0, \quad u(0) = u_0 \tag{15}$$

и применим для его решения метод Эйлера:

$$u^{n+1} = u^n - Mhu^n = (1 - Mh)u^n, \quad u^0 = u_0. \tag{16}$$

Решение задачи (15) при $M > 0$ имеет вид монотонно убывающей экспоненты $u(x) = u_0 e^{-Mx}$, в то время как решение (16) имеет вид $u^n = (1 - Mh)^n u_0$. Таким образом, при $|1 - Mh| > 1$ решение задачи методом Эйлера экспоненциально возрастает и не имеет ничего общего с реальным решением. Более того, если даже $|1 - Mh| < 1$, но $1 - Mh < 0$, то решение (15) экспоненциально убывает, но при переходе от шага к шагу меняет знак, т. е. метод Эйлера в этом случае также неприменим.

Таким образом, из проведенных выше рассуждений можно сделать вывод, что метод Эйлера дает приближенное решение, правильно моделирующее поведение решения дифференциальной задачи при выполнении условия на шаг интегрирования

$$h \leqslant \frac{1}{M}, \tag{17}$$

что при больших по модулю M влечет за собой непомерное увеличение вычислительных затрат.

Может показаться, что явные методы численного интегрирования вообще неприменимы для решения жестких систем обыкновенных дифференциальных уравнений. Однако это не так. Не так давно был предложен метод численного интегрирования жестких систем обыкновенных дифференциальных уравнений с помощью явных методов (в частности, метода Эйлера), которые позволяют существенно сократить вычислительные затраты. Суть метода состоит в переменном шаге интегрирования.

Изложим основные идеи метода на примере модельной задачи. В качестве такой задачи рассмотрим систему обыкновенных дифференциальных уравнений с симметричной неотрицательной матрицей A:

$$\frac{d\mathbf{u}}{dt} + A\mathbf{u} = \mathbf{0}, \quad \mathbf{u}(0) = \mathbf{u}_0. \tag{18}$$

Будем предполагать, что спектр матрицы A может быть разделен на две части: гладкую — $\lambda_1, \ldots, \lambda_k = O(1)$ и жесткую — $\lambda_{k+1}, \ldots, \lambda_m \gg 1$; $\max \lambda_j = M$. Решение задачи (18) имеет вид

$$\mathbf{u}(t) = e^{-At}\mathbf{u}_0. \tag{19}$$

Решение задачи (18) будем искать на отрезке $[0, \tau]$, используя метод Эйлера

$$\mathbf{u}_{n+1} = \mathbf{u}_n - hA\mathbf{u}_n = (E - hA)\mathbf{u}_n,$$

откуда

$$\mathbf{u}_n = (E - hA)^n\mathbf{u}_0. \tag{20}$$

Разлагая $\mathbf{u}_0$ по собственным векторам матрицы A, получим

$$\mathbf{u}_0 = \sum_{j=1}^{m} \alpha_j \mathbf{e}_j.$$

Тогда из (20) имеем

$$\mathbf{u}_n = \sum_{j=1}^{m} \alpha_j (1 - h\lambda_j)^n \mathbf{e}_j. \tag{21}$$

Вследствие (19) норма решения $\|\mathbf{u}(t)\|$ дифференциальной задачи не возрастает. Если потребовать, чтобы норма решения дискретной задачи

также не возрастала, то из (21) можно получить, что шаг интегрирования h должен удовлетворять условию

$$|1 - h\lambda_j| \leqslant 1, \quad j = 1, \ldots, m,$$

откуда получаем оценку для h:

$$h \leqslant 2/M. \tag{22}$$

Условие (22) является необходимым для устойчивости метода Эйлера (20). Даже в случае, когда мы находимся в области, где составляющая решения, соответствующая жесткой части спектра, близка к нулю, мы вынуждены выбирать шаг интегрирования, удовлетворяющий (22); в противном случае происходит экспоненциальное накопление вычислительной погрешности.

Предположим, что мы задали число шагов интегрирования, равное N. Тогда за N шагов, используя метод Эйлера (20) и условие устойчивости (22), мы можем получить приближенное решение до момента времени $\tau = 2N/M$. Поставим следующую задачу. Используя переменный шаг метода Эйлера, получить за N шагов приближенное решение на как можно большем промежутке. Естественно, что при этом используемый метод должен быть устойчивым.

Уточним данную постановку для случая общего уравнения $\dfrac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u})$, $\mathbf{u}(0) = \mathbf{u}_0$. Если $\mathbf{u}_n$ известно, то $\mathbf{u}_{n+1}$ будем искать, используя следующий алгоритм

$$\begin{aligned}
&\mathbf{y}_{n+1/2} = \mathbf{u}_n + h_{n+1}\mathbf{f}(t_n, \mathbf{u}_n), && t_{n+1/2} = t_n + h_{n+1},\\
&\mathbf{y}_{n+1} = \mathbf{y}_{n+1/2} + h_{n+1}\mathbf{f}(t_{n+1/2}, \mathbf{y}_{n+1/2}), && t_{n+1} = t_{n+1/2} + h_{n+1},\\
&\mathbf{u}_{n+1} = \mathbf{y}_{n+1} + \gamma_{n+1}h_{n+1}\left(\mathbf{f}(t_n, \mathbf{u}_n) - \mathbf{f}(t_{n+1/2}, \mathbf{y}_{n+1/2})\right);
\end{aligned} \tag{23}$$

здесь γ_j — некоторые постоянные, которые наряду с h_j надо определить.

Рассмотрение метода (23) будем проводить на примере задачи (18). Для случая задачи (18) метод (23) имеет вид

$$\begin{aligned}
&\mathbf{y}_{n+1/2} = \mathbf{u}_n - h_{n+1}A\mathbf{u}_n, && t_{n+1/2} = t_n + h_{n+1},\\
&\mathbf{y}_{n+1} = \mathbf{y}_{n+1/2} - h_{n+1}A\mathbf{y}_{n+1/2}, && t_{n+1} = t_{n+1/2} + h_{n+1},\\
&\mathbf{u}_{n+1} = \mathbf{y}_{n+1} - \gamma_{n+1}h_{n+1}\left(A\mathbf{u}_n - A\mathbf{y}_{n+1/2}\right).
\end{aligned}$$

С помощью элементарных преобразований отсюда получаем соотношение, связывающее $\mathbf{u}_{n+1}$ и $\mathbf{u}_n$:

$$\begin{aligned}
\mathbf{u}_{n+1} &= (E - h_{n+1}A)^2\mathbf{u}_n - \gamma_{n+1}h_{n+1}\left(A\mathbf{u}_n - A(\mathbf{u}_n - h_{n+1}A\mathbf{u}_n)\right) =\\
&= (E - h_{n+1}A)^2\mathbf{u}_n - \gamma_{n+1}h_{n+1}^2A^2\mathbf{u}_n =\\
&= \left[(E - h_{n+1}A)^2 - \gamma_{n+1}h_{n+1}^2A^2\right]\mathbf{u}_n.
\end{aligned} \tag{24}$$

Таким образом, мы можем записать соотношение, связывающее $\mathbf{u}_N$ и $\mathbf{u}_0$ в виде

$$\mathbf{u}_N = Q_{2N}(A)\mathbf{u}_0, \tag{25}$$

где

$$Q_{2N}(\lambda) = \prod_{i=1}^{N} \left[(1 - h_i\lambda)^2 - \gamma_i h_i^2 \lambda^2\right]. \tag{26}$$

Заметим, что

$$Q_{2N}(\lambda) = 1 - 2\sum_{i=1}^{N} h_i\lambda + \cdots$$

Из постановки задачи видно, что коэффициент при λ в многочлене $Q_{2N}(\lambda)$ как раз и является величиной, которую мы должны минимизировать. Кроме этого должно выполняться условие устойчивости $\|Q_{2N}(A)\| \leqslant 1$, что эквивалентно выполнению неравенства $|Q_{2N}(\lambda)| \leqslant 1$ при $\lambda \in [0, M]$.

Таким образом, мы пришли к следующей постановке задачи. Введем класс P многочленов степени $2N$ со свободным членом 1, не превосходящих по модулю 1 на отрезке $[0, M]$. На этом классе требуется найти многочлен, производная которого в нуле является минимальной, т. е. требуется найти $Q_{2N} \in P$ такой, что

$$Q_{2N}(\lambda) = \arg \min_{P_{2N} \in P} P'_{2N}(0). \tag{27}$$

Справедлива следующая

Лемма.

$$arg \min_{P_{2N} \in P} P'_{2N}(0) = T_{2N}\left(\frac{2\lambda - M}{M}\right) \equiv Q_{2N}(\lambda);$$

здесь T_{2N} — многочлен Чебышева степени $2N$.

Доказательство. Из определения многочленов Чебышева легко видеть, что $T_{2N} \in P$. Предположим, что утверждение леммы неверно. Тогда существует многочлен $P_{2N} \in P$ такой, что $P'_{2N}(0) < Q'_{2N}(0)$. Рассмотрим разность $r(\lambda) = Q_{2N}(\lambda) - P_{2N}(\lambda)$. Эта разность является многочленом степени $2N$ и в точке 0 обращается в нуль, так как P_{2N}, $Q_{2N} \in P$. Пусть $\lambda_0, \ldots, \lambda_{2N}$ — точки экстремума Q_{2N}; при этом $\lambda_0 = 0$, $\lambda_{2N} = M$. Так как Q_{2N} многочлен четной степени, то $\operatorname{sign} Q_{2N}(\lambda_{2j}) > 0$, $\operatorname{sign} Q_{2N}(\lambda_{2j+1}) < 0$. Отсюда следует, что r удовлетворяет условию

$$\operatorname{sign} r(\lambda_{2j}) \geqslant 0, \quad \operatorname{sign} r(\lambda_{2j+1}) \leqslant 0. \tag{28}$$

Найдем количество корней многочлена r на отрезке $[0, M]$. На отрезке $[\lambda_0, \lambda_1]$ всегда имеется два корня. Действительно, $r(\lambda_0) = 0$ и, по предположению леммы, $r'(\lambda_0) > 0$. Поскольку $r(\lambda_1) \leqslant 0$, то это и означает, что

на $[\lambda_0, \lambda_1]$ имеется не менее двух корней. Если до точки λ_{2j+1} во всех узлах λ_k (за исключением λ_0) имело место строгое неравенство (28), то на отрезке $[\lambda_0, \lambda_{2j+1}]$ имеется $2j+2$ корня $r(\lambda)$. Пусть $r(\lambda_{2j+1}) = 0$. Тогда возможны два случая: или $r(\lambda)$ меняет знак в окрестности точки λ_{2j+1}, или знак в окрестности λ_{2j+1} не меняется. В первом случае это означает, что на отрезке $[\lambda_0, \lambda_{2j+2}]$ имеется $2j+3$ корня. Во втором случае точка λ_{2j+1} является кратным корнем $r(\lambda)$. Таким образом, до точки λ_{2j+2} мы рассмотрели все возможные случаи и установили, что $r(\lambda)$ имеет с учетом кратности на $[\lambda_0, \lambda_{2j+2}]$ не менее $2j+3$ корней. Продолжая процесс подсчета корней, окончательно получаем, что на всем отрезке $[\lambda_0, \lambda_{2N}]$ многочлен $r(\lambda)$ степени $2N$ имеет не менее $2N+1$ корней. Полученное противоречие завершает доказательство.

Следует отметить, что исходная задача до сих пор не решена, так как мы лишь нашли функцию из класса P, удовлетворяющую условию экстремума. Однако ниоткуда пока не следует, что эта функция (многочлен) представима в виде (26). Покажем, что это действительно так. Многочлен $Q_{2N}(\lambda) \equiv T_{2N}\left(\frac{2\lambda - M}{M}\right)$ имеет на отрезке $[0, M]$ ровно $2N$ корней: $\lambda_j = \frac{M}{2} + \frac{M}{2}\cos\frac{(2j-1)\pi}{2N}$, $j = 1, \ldots, 2N$, которые расположены симметрично относительно точки $\dfrac{M}{2}$ — середины отрезка $[0, M]$. Вводя величины ξ_j такие, что $\lambda_j = \dfrac{M}{2} + \xi_j$, $j = 1, \ldots, N$, искомый многочлен $T_{2N}\left(\dfrac{2\lambda - M}{M}\right)$ можно записать в виде

$$T_{2N}\left(\frac{2\lambda - M}{M}\right) = \prod_{j=1}^{N}\left(1 - \frac{\lambda}{M/2 + \xi_j}\right)\left(1 - \frac{\lambda}{M/2 - \xi_j}\right) = \prod_{j=1}^{N}\frac{1}{M^2/4 - \xi_j^2}\left[(M/2 - \lambda)^2 - \xi_j^2\right]. \tag{29}$$

Приравнивая при фиксированном j выражение, стоящее под знаком произведения в (29), к выражению под знаком произведения в (26) при том же j, получаем соотношение для определения параметров h_j, γ_j:

$$h_j = \frac{M}{M^2/2 - 2\xi_j^2}, \quad \gamma_j = \frac{4\xi_j^2}{M^2}. \tag{30}$$

Таким образом, многочлен $Q_{2N}(\lambda) = T_{2N}\left(\dfrac{2\lambda - M}{M}\right)$ действительно является решением поставленной задачи. Вычислим значение $Q'_{2N}(0)$. Для многочлена Чебышева $T_{2N}(x)$ справедлива формула

$$T_{2N}(x) = \frac{(x + \sqrt{x^2-1})^{2N} + (x - \sqrt{x^2-1})^{2N}}{2}, \quad x = \frac{2\lambda - M}{M}. \tag{31}$$

Тогда

$$\frac{dT_{2N}}{d\lambda} = N\left[(x+\sqrt{x^2-1})^{2N-1}\left(\frac{2}{M}+\frac{2x}{M\sqrt{x^2-1}}\right)+\right.$$
$$\left.+(x-\sqrt{x^2-1})^{2N-1}\left(\frac{2}{M}-\frac{2x}{M\sqrt{x^2-1}}\right)\right] =$$
$$= \frac{2N}{M}\left[\left(x^{2N-1}+(2N-1)x^{2N-2}\sqrt{x^2-1}+o\left(\sqrt{x^2-1}\right)\right)\left(1+\frac{x}{\sqrt{x^2-1}}\right)+\right.$$
$$\left.+\left(x^{2N-1}-(2N-1)x^{2N-2}\sqrt{x^2-1}+o(\sqrt{x^2-1})\right)\left(1-\frac{x}{\sqrt{x^2-1}}\right)\right] =$$
$$= \frac{2N}{M}\left[2x^{2N-1}+2(2N-1)x^{2N-1}+O\left(\sqrt{x^2-1}\right)\right]. \qquad (32)$$

Так как

$$\left.\frac{dT_{2N}(\frac{2\lambda-M}{M})}{d\lambda}\right|_{\lambda=0} = \left.\frac{dT_{2N}(x)}{d\lambda}\right|_{x=-1},$$

то из (32) имеем

$$\left.\frac{dT_{2N}(\frac{2\lambda-M}{M})}{d\lambda}\right|_{\lambda=0} = -\frac{8N^2}{M}. \qquad (33)$$

Таким образом, за $2N$ шагов численного интегрирования жесткой системы уравнений с помощью метода Эйлера мы можем получить приближенное решение на отрезке $[0, \frac{4N}{M}]$, в то время как при интегрировании той же системы уравнений с помощью *явного метода с переменным шагом* можно получить приближенное решение на отрезке $[0, \frac{8N^2}{M}]$ за то же количество шагов. Отсюда следует, что использование переменного шага интегрирования позволяет увеличить при тех же затратах процессорного времени длину отрезка интегрирования в $\approx N$ раз, т. е. метод будет особо эффективным при больших N. В настоящее время расчеты по указанному методу проводятся до значений N, достигающих порядка 10^6. При этом N выбирается в каждой расчетной точке в зависимости от гладкости решения. Как правило, N выбирают среди чисел $N = 2^l \cdot 3^k$.

Следует отметить тот факт, что при практической реализации рассмотренного выше метода, как и в случае Чебышевского ускорения итерационных методов, очень важной является проблема правильного упорядочивания параметров процесса. В настоящее время эти задачи решены и на их основе создан комплекс программ Лебедева «DUMKA» для решения жестких систем обыкновенных дифференциальных уравнений, который показал свою высокую эффективность на многочисленных задачах этого класса. Особенно эффективен данный метод по сравнению с другими при использовании многопроцессорной вычислительной техники, так как он легко распараллеливается.

§ 10. Методы численного интегрирования уравнений второго порядка

Введением новых неизвестных функций дифференциальные уравнения порядка выше первого и их системы сводятся к системам уравнений первого порядка. Таким образом, при формальном подходе вопрос о численном решении задачи Коши для уравнений высших порядков можно было бы считать исчерпанным.

Однако методы, приспособленные специально для решения уравнений высших порядков, часто более эффективны. При разработке таких методов нужно также иметь в виду широкое распространение систем уравнений высоких порядков специального вида, учет специфики которых может еще более повысить эффективность методов. Например, ряд задач небесной механики сводится к интегрированию систем уравнений

$$\mathbf{y}'' = \mathbf{f}(\mathbf{y}).$$

Рассмотрим несколько более широкий класс уравнений

$$y'' = f(x, y).$$

Как и выше, рассуждения проводятся для одного уравнения, поскольку перенос результатов на случай системы осуществляется автоматически. Традиционно наиболее распространенными методами интегрирования таких уравнений являются *явный*

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=0}^{k} b_{-i} f(x_{n-i},\, y_{n-i}) \tag{1}$$

и *неявный*

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=-1}^{k} b_{-i} f(x_{n-i},\, y_{n-i}) \tag{2}$$

конечно-разностные методы. Эти методы можно было бы получить методом неопределенных коэффициентов, потребовав, чтобы разность

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1})}{h^2} - \sum_{i=-1}^{k} b_{-i} y''(x_{n-i})$$

была величиной как можно более высокого порядка по h. Здесь, как обычно, предполагается $x_{n+1} - x_n \equiv h$.

Одним из наиболее употребительных среди этих методов является *неявный метод четвертого порядка точности вида* (2) (часто называемый *методом Нумерова*) при $k = 1$:

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \Big(\frac{1}{12} f(x_{n-1},\, y_{n-1}) + \frac{10}{12} f(x_n,\, y_n) + \frac{1}{12} f(x_{n+1},\, y_{n+1}) \Big),$$

который особенно удобен в случае линейных задач.

Для целей численного интегрирования уравнения

$$y'' = f(x, y, y')$$

употребительными являются следующие методы: значения y_n и z_n приближений к $y(x_n)$ и $y'(x_n)$ вычисляются из совокупности явных рекуррентных соотношений вида

$$z_{n+1} = z_n + h\sum_{k=0}^{l} a_k \nabla^k f_n, \quad y_{n+1} = y_n + h\sum_{k=0}^{l} b_k \nabla^k z_{n+1}$$

или неявных вида

$$z_{n+1} = z_n + h\sum_{k=0}^{l} c_k \nabla^k f_{n+1}, \quad y_{n+1} = y_n + h\sum_{k=0}^{l} d_k \nabla^k z_{n+1}.$$

Уравнения более высокого порядка, чем второй, на практике часто сводят к системам уравнений второго или первого порядков (см. также § 9.11).

Как и в случае конечно-разностных схем для уравнений первого порядка, отбросим в (1) и (2) слагаемые, содержащие значения f, и рассмотрим получившееся разностное уравнение. В обоих случаях оно имеет вид

$$y_{n+1} - 2y_n + y_{n-1} = 0.$$

Его характеристическое уравнение имеет кратный корень $\mu = 1$. Не нужно пугаться того, что корень оказался кратным: если в разностных схемах, аппроксимирующих уравнения первого порядка, перед значениями f стоял коэффициент порядка h, то здесь стоит коэффициент порядка h^2.

Все упоминавшиеся ранее разностные схемы интегрирования уравнения $y'' = f(x, y)$ могут быть представлены в виде

$$\frac{1}{h^2}\sum_{i=0}^{k} a_{-i} y_{n-i} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0, \quad a_0 \neq 0; \tag{3}$$

при этом для гладкой функции

$$\frac{1}{h^2}\sum_{i=0}^{k} a_{-i} y(x_{n-i}) \to y''(x_n)$$

при фиксированном x_n и $h \to 0$; кроме того, $\sum\limits_{i=0}^{k} b_{-i} = 1$. Пусть r_n — погрешность аппроксимации (результат подстановки в левую часть решения дифференциальной задачи):

$$r_n = \sum_{i=0}^{k} \frac{a_{-i} y(x_{n-i})}{h^2} - \sum_{i=0}^{k} b_{-i} f\Big(x_{n-i}, y(x_{n-i})\Big),$$

и $|r_n| \leqslant c_0 h^m$ равномерно на всем рассматриваемом отрезке интегрирования $[x_0, x_0 + X]$. Как и в случае уравнений первого порядка, из-за округлений при вычислениях и неточности решения уравнения относительно неизвестной y_n в случае неявной схемы ($b_0 \neq 0$) реально получаемое приближенное решение y_n удовлетворяет (3) с некоторой погрешностью. Таким образом, имеем

$$\frac{1}{h^2}\sum_{i=0}^{k} a_{-i} y_{n-i} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = \delta_n.$$

Вычитая из этого соотношения предыдущее, получим уравнение для погрешности $R_n = y_n - y(x_n)$:

$$\frac{1}{h^2}\sum_{i=0}^{k} a_{-i} R_{n-i} - \sum_{i=0}^{k} b_{-i} l_{n-i} R_{n-i} = g_n,$$

где $l_j = f_y(x_j, \tilde{y}_j)$, $g_n = \delta_n - r_n$. Пусть

$$L = \sup_{x_0 \leqslant x \leqslant x_0 + X} |f_y(x, y)| < \infty.$$

Теорема (без доказательства). *Пусть все корни характеристического уравнения разностной схемы*

$$\sum_{i=0}^{k} a_{-i} \mu^{k-i} = 0$$

лежат в единичном круге и на границе круга нет кратных корней, за исключением двукратного корня, равного 1. *Тогда при* $x_0 \leqslant x \leqslant x_0 + X$ *справедлива оценка погрешности*

$$\max_n \left(|R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right) \leqslant \\ \leqslant C(L, X) \left(\max_{x_0 \leqslant x \leqslant x_0 + X} |g_n| + \max_{0 \leqslant j < k} |R_j| + \max_{0 < j < k} \left| \frac{R_j - R_{j-1}}{h} \right| \right). \tag{5}$$

При численной реализации методов решения уравнений второго порядка с целью уменьшить влияние вычислительной погрешности целесообразно преобразовать расчетные формулы к другому виду.

Методы численного решения уравнений второго порядка могут быть использованы при численном решении уравнения, где производная решения не выражается явно через само решение и аргумент; для определенности рассмотрим скалярный случай

$$F(x, y, y') = 0.$$

Если бы это уравнение удалось разрешить относительно y', то получилось бы некоторое уравнение

$$y' = f(x, y).$$

Первый возможный путь решения задачи состоит в формальном применении методов Рунге-Кутта или конечно-разностных методов; для каждых x и y значение f определяется путем численного решения уравнения

$$F(x, y, f) = 0.$$

Хорошее начальное приближение к искомому значению f можно получить интерполяцией ранее найденных значений, поэтому число требуемых итераций (например, в методе секущих) обычно оказывается небольшим.

В случае неявных методов бывает целесообразно сразу решать систему уравнений

$$F(x_n, y_n, f_n) = 0,$$

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f_{n-i} = 0$$

относительно неизвестных y_n, f_n.

Иногда (крайне редко) поступают следующим образом: дифференцируя исходное уравнение по x, получают соотношения

$$\frac{d}{dx}(F(x, y, y')) = F_x(x, y, y') + F_y(x, y, y')y' + F_{y'}(x, y, y')y'' = 0,$$

$$\frac{d^2}{dx^2}(F(x, y, y')) = \cdots = 0$$

и т. д. Если значение $y'(x_0)$ уже найдено, то из этих соотношений можно явным образом определить значения $y''(x_0), \ldots$, а затем получить $y(x_0+h)$ с помощью формулы Тейлора.

Первое из этих соотношений можно переписать в виде

$$y'' = g(x, y, y').$$

Отсюда получается третий путь решения задачи: из уравнения

$$F(x_0, y_0, y'(x_0)) = 0$$

определяется $y'(x_0)$, и далее численно интегрируется уравнение

$$y'' = g(x, y, y').$$

§ 11. Оптимизация распределения узлов интегрирования

Решения дифференциальных уравнений и систем могут иметь различную гладкость на различных участках отрезка интегрирования. На примере оценки погрешности метода Рунге—Кутта было видно, что вклад от погрешности интегрирования на некотором шаге $[x_n, x_{n+1}]$ в суммарную погрешность в точке $x_N = x_0 + X$ равен произведению погрешности

на шаге на множитель $\exp\left\{\int_{x_n}^{x_N} f_y(x,\,\tilde{y}(x))\,dx\right\}$, зависящий от n. Поэтому для некоторых классов дифференциальных уравнений становится особо актуальной задача оптимизации распределения узлов интегрирования. Произведем анализ этой проблемы, не вникая в тонкости обоснования проводимых построений. Для простоты предполагаем, что начальное условие задано точно и округления отсутствуют. Погрешность результата численного интегрирования по методу Рунге—Кутта в точке x_N не превосходит

$$S_N = \sum_{n=1}^{N} |y_n - y_{n-1}(x_n)| \exp\left\{\int_{x_n}^{x_N} f_y(x,\,\tilde{y}_n(x))\,dx\right\}. \tag{1}$$

Предположим, что $x_n = \varphi(n/N)$, $\varphi(0) = x_0$, $\varphi(1) = x_0 + X$, $\varphi(t)$ — гладкая функция. Согласно формуле Лагранжа

$$x_n - x_{n-1} = \varphi\left(\frac{n}{N}\right) - \varphi\left(\frac{n-1}{N}\right) = \frac{1}{N}\varphi'(\bar{t}_n),$$

где $(n-1)/N \leqslant \bar{t}_n \leqslant n/N$, поэтому

$$H = \max_{0<n\leqslant N}(x_n - x_{n-1}) \leqslant \frac{1}{N}\max_{[0,\,1]}\left|\varphi'(t)\right|, \quad H \to 0 \quad \text{при} \quad N \to \infty;$$

следовательно (см. § 4), имеем

$$\max_{x_0\leqslant x_n\leqslant x\leqslant x_0+X} |y_n(x) - y(x)| \to 0$$

при $N \to \infty$. Таким образом, можно написать асимптотические равенства (см. также § 4)

$$\exp\left\{\int_{x_n}^{x_N} f_y(x,\,\tilde{y}_n(x))\,dx\right\} \sim \exp\left\{\int_{x_n}^{x_N} f_y(x,\,y(x))\,dx\right\},$$

$$y_n - y_{n-1}(x_n) \sim \varphi\Big(x_n,\,y(x_n)\Big)(x_n - x_{n-1})^{k+1} \sim$$

$$\sim \varphi\left(\varphi\Big(\frac{n}{N}\Big),\,y\Big(\varphi\Big(\frac{n}{N}\Big)\Big)\right)\left(\frac{1}{N}\varphi'\Big(\frac{n}{N}\Big)\right)^{k+1},$$

используя которые, выражение (1) для S_N представим в виде

$$S_N = \frac{1}{N^k}\sum_{n=1}^{N}\frac{1}{N}\Phi\left(\frac{n}{N}\right) + o\left(\frac{1}{N^k}\right), \tag{2}$$

где

$$\Phi(t) = |\varphi(\varphi(t),\,y(\varphi(t)))|\,(\varphi'(t))^{k+1}\exp\left\{\int_{\varphi(t)}^{x_0+X} f_y(x,\,y(x))\,dx\right\}. \tag{3}$$

При гладкой функции $\Phi(t)$ и $N \to \infty$ величина $\sum_{n=1}^{N} \frac{1}{N} \Phi\left(\frac{n}{N}\right)$ стремится к интегралу

$$I = \int_0^1 \Phi(t)\, dt \tag{4}$$

и, таким образом,

$$S_n \sim N^{-k} \int_0^1 \Phi(t)\, dt.$$

Дальнейшие построения являются некоторым усложнением построений из § 3.12. Примем φ за новую переменную в интеграле (4). Тогда он запишется в виде

$$I = \int_{x_0}^{x_0+X} L(\varphi)\, d\varphi,$$

где

$$L(\varphi) = \Big|\varphi\Big(\varphi,\, y(\varphi)\Big)\Big| \Big(t'(\varphi)\Big)^{-k} \exp\left\{\int_{\varphi}^{x_0+X} f_y(x,\, y(x))\, dx\right\}. \tag{5}$$

Задача минимизации интеграла (4) за счет выбора функции $\varphi(t)$ и задача минимизации этого же интеграла за счет выбора обратной функции $t(\varphi)$ в форме (5) эквивалентны. Вследствие равенства $\dfrac{\partial L}{\partial t} = 0$ уравнение Эйлера

$$\frac{d}{d\varphi}\left(\frac{\partial L}{\partial t'}\right) - \frac{\partial L}{\partial t} = 0$$

в данном случае приобретает вид $\dfrac{d}{d\varphi}\left(\dfrac{\partial L}{\partial t'}\right) = 0$. Отсюда следует, что

$$\frac{\partial L}{\partial t'} = -k\Big|\psi\Big(\varphi,\, y(\varphi)\Big)\Big| \Big(t'(\varphi)\Big)^{-k-1} \exp\left\{\int_{\varphi}^{x_0+X} f_y(x,\, y(x))\, dx\right\} = \text{const}.$$

Возвращаясь к переменной φ, получаем

$$\Big(\varphi'(t)\Big)^{k+1}\Big|\psi\Big(\varphi,\, y(\varphi)\Big)\Big| \exp\left\{\int_{\varphi}^{x_0+X} f_y(x,\, y(x))\, dx\right\} = C_1. \tag{6}$$

Решение этого дифференциального уравнения зависит от C_1 и еще некоторой постоянной C_2. Их значения следует определить из граничных условий $\varphi(0) = x_0$, $\varphi(1) = x_0 + X$.

Отметим одно неочевидное обстоятельство. Уравнение (6) может быть записано в виде

$$\Big(\varphi'(t)\Big)^{k+1}\Big|\psi\Big(\varphi,\, y(\varphi)\Big)\Big| \exp\left\{-\int_0^{\varphi} f_y(x,\, y(x))\, dx\right\} = C, \tag{7}$$

куда не входят ни начальная, ни конечная точки интегрирования. При $C = Ca^{k+1}$ уравнению (7) удовлетворяет также функция $\varphi(at + b)$. Задавшись

некоторым N, осуществим одновременное интегрирование исходного уравнения и уравнения (7), выбирая каждый раз шаги интегрирования из условия $x_n - x_{n-1} \sim N^{-1}\varphi'(t_{n-1})$. Скорее всего, мы придем в конечную точку $x_0 + X$ с числом узлов N_1, отличным от N, и тогда это распределение узлов не будет являться оптимальным распределением, соответствующим данному N. Это естественно, поскольку, начиная вычисления, мы не могли заранее предвидеть хода поведения решения и сказать, какую величину C следует взять в правой части (7). Однако можно показать, что, вследствие указанных свойств решения уравнения (7), функция $\varphi(t)$ будет искомой и распределение узлов $\varphi(n/N_1)$ — оптимальным (с точностью до погрешности численного интегрирования), соответствующим числу узлов N_1.

Непосредственное интегрирование уравнений (1) и (7) встречает затруднение из-за необходимости вычисления значений функции $\psi(x, y(x))$. Вместо непосредственного вычисления значений этой функции целесообразно использовать величину контрольного члена точности на шаге. При численном интегрировании уравнения (7) следует также иметь в виду, что эти соотношения носят асимптотический характер. В окрестности точек, где $\psi(x, y(x)) = 0$, в остаточном члене начинают играть существенную роль слагаемые порядка $(x_n - x_{n-1})^{k+2}$.

Дополнительное численное интегрирование уравнения (7) может сильно усложнить решение задачи. Поэтому к вопросу оптимизации распределения узлов часто подходят следующим образом. Пусть решаются задачи из некоторого определенного класса. Рассмотрим модельную для этого класса задачу, для которой можно в явном виде решить уравнение (7). Постараемся на ее примере установить зависимость шага (или меры погрешности на шаге) от поведения решения, при которой распределение узлов близко к оптимальному. Далее все задачи этого класса интегрируем с шагом, соответствующим этой зависимости.

В других случаях заранее задаются некоторой формой такой зависимости. Пусть система уравнений порядка m интегрируется с контролем точности на шаге. В случае системы уравнений контрольный член r будет некоторым вектором $\mathbf{r}_n = (r_n^1, \ldots, r_n^m)$. В ряде программ шаг интегрирования выбирается из условия

$$\|\mathbf{r}_n\| \Big/ \max\left\{M, \|y_n\|\right\} \approx \varepsilon = \text{const}$$

или из условия

$$\max_{1 \leqslant k \leqslant m} \frac{|r_n^k|}{\max\{|M_k|, |(y_n)_k|\}} \approx \varepsilon = \text{const}.$$

Параметры M, M_k подбирают из соображений оптимальности распределения узлов в условиях конкретной задачи.

Рассмотрим пример оптимизации распределения узлов интегрирования.

Пусть задача Коши $y'(x) = My(x)$ при начальном условии $y(0)$ решается методом Эйлера. Тогда

$$\psi(x, y(x)) = -\frac{1}{2}y''(x) = -\frac{M^2}{2}y(0)\exp\{Mx\}$$

и уравнение (6) имеет вид

$$(\varphi'_t)\frac{M^2}{2}|y(0)|\exp\{M\varphi\}\exp\{M(X-\varphi)\} = \text{const}.$$

Отсюда получаем $\varphi'_t = \text{const}$, т. е. распределение узлов следует взять равномерным.

Наибольшего эффекта решение задачи оптимизации распределения узлов или ее упрощенных вариантов достигает в случае решений с особенностями производных и при решении задач с малыми параметрами при старших производных, например задач типа пограничного слоя.

Литература

1. Бахвалов Н. С. Некоторые замечания к вопросу о численном интегрировании дифференциальных уравнений методом конечных разностей. // ДАН СССР. — 1955. — **104**, N 6, С. 805–808.
2. Винокуров В. А., Ювченко Н. В. Полуявные численные методы решения жестких задач // ДАН. — 1985. — **284**, N 2, С. 272–277.
3. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Дифференциальные уравнения — Минск: Наука и техника, 1982.
4. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т.2 — М.: Наука, 1977.
5. Лебедев В. И. Как решать явными методами жесткие системы дифференциальных уравнений // Вычислительные процессы и системы — М.: Наука, 1991. Вып.8, С. 237–291.
6. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем — М.: Наука, 1979.
7. Современные численные методы решения обыкновенных дифференциальных уравнений // Под ред. Дж. Холла, Дж. Уатта — М.: Мир, 1979.
8. Федоренко Р. П. Жесткие системы обыкновенных дифференциальных уравнений и их численное интегрирование // В кн. Вычислительные процессы и системы. Вып. 8, М.: Наука, 1991. С. 328–380.
9. Федоренко Р. П. Введение в вычислительную физику — М.: Изд-во МФТИ, 1994.
10. Хайрер Э., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи — М.: Мир, 1999.
11. Хайрер Э., Нерсетт С., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи — М.: Мир, 1990.
12. Butcher I. G. A modified multistep method for the numerical integration of ordinary differential equations // J. Assoc. Comput. Math. — 1965, **12**, N 1. P. 124–135.
13. Dahlquist Y. Stability and error bounds in the numerical integration of ordinary differential equations — Uppsala, Almqvist & Wiksells boktr 130 (1959). P. 5–92.

Глава 9

Численные методы решения краевых задач для обыкновенных дифференциальных уравнений



При решении краевых задач возникают дополнительные трудности по сравнению со случаем решения задачи Коши: значительно сложнее исследуется вопрос о существовании решения; после написания сеточной задачи возникает система линейных или нелинейных уравнений, проблема решения которой требует дополнительного изучения,

§ 1. Простейшие методы решения краевой задачи для уравнений второго порядка

Среди краевых задач для обыкновенных дифференциальных уравнений существенную часть составляют задачи для уравнений и систем второго порядка. В частности, такие задачи возникают в баллистике, теории упругости и т. д.

Начнем изучение вопроса с одной частной, но довольно распространенной краевой задачи. Ищется решение уравнения

$$Ly \equiv y'' - p(x)y = f(x) \quad \text{на} \quad (0, X) \tag{1}$$

при граничных условиях

$$y(0) = a, \quad y(X) = b. \tag{2}$$

Зададимся шагом $h = XN^{-1}$, N целое; точки $x_n = nh$ примем за узлы сетки; как обычно, y_n — приближения к значениям $y(x_n)$. После замены производной $y''(x_n)$ на разностное отношение

$$\frac{\delta^2 y_n}{h^2} \equiv \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2}$$

получаем систему уравнений

$$l(y_n) \equiv \frac{\delta^2 y_n}{h^2} - p_n y_n = f_n, \quad n = 1, \ldots, N-1; \tag{3}$$

здесь $p_n = p(x_n)$, $f_n = f(x_n)$; граничные условия заменим соотношениями

$$y_0 = a, \quad y_N = b. \tag{4}$$

Покажем, что при $p(x) \geqslant 0$ система уравнений (3), (4) имеет решение, и дадим оценку погрешности.

Лемма 1. *Пусть* $p(x) \geqslant 0$, $l(z_n) \leqslant 0$, $z_0, z_N \geqslant 0$. *Тогда* $z_n \geqslant 0$ *при всех* n.

Доказательство. Обозначим $\min\limits_{0\leqslant n\leqslant N} z_n$ через d. Предположим, что $d < 0$ и, следовательно, $d \neq z_0, z_N$. Пусть q — наименьшее целое такое, что $z_q = d$; из определения d и q имеем $z_{q-1} > d$, $z_{q+1} \geqslant d$. Тогда

$$l(z_q) = \frac{(z_{q+1} - d) + (z_{q-1} - d)}{h^2} - p_q z_q \geqslant \frac{z_{q-1} - d}{h^2} > 0$$

и мы приходим к противоречию с предположением $d < 0$.

Лемма 2. *Если* $p(x) \geqslant 0$, *то для любой функции* z_n *выполняется неравенство*

$$\max_{0\leqslant n\leqslant N} |z_n| \leqslant \max\left\{|z_0|, |z_N|\right\} + Z\frac{X^2}{8},$$

где

$$Z = \max_{0<n<N} |l(z_n)|.$$

Доказательство. Введем в рассмотрение функцию

$$w_n = |z_0|\left(1 - \frac{nh}{X}\right) + |z_N|\frac{nh}{X} + Z\frac{nh(X - nh)}{2}.$$

Для многочлена второй степени величина $\delta^2 Q/h^2$ совпадает со второй производной, поэтому $\delta^2 w_n/h^2 = -Z$. Из явного вида w_n следует, что $w_n \geqslant 0$, поэтому

$$l(w_n) = \delta^2 w_n/h^2 - p_n w_n \leqslant -Z, \quad l(w_n \pm z_n) \leqslant -Z \pm l(z_n) \leqslant 0.$$

Имеем

$$w_0 \pm z_0 = |z_0| \pm z_0 \geqslant 0, \quad w_N \pm z_N = |z_N| \pm z_N \geqslant 0.$$

Функции $w_n \pm z_n$ удовлетворяют условиям леммы 1, поэтому $w_n \pm z_m \geqslant 0$. Отсюда следует оценка $|z_n| \leqslant |w_n| \leqslant \max\limits_{0\leqslant n\leqslant N} |w_n|$. Имеем неравенства

$$|z_0|\left(1 - \frac{nh}{X}\right) + |z_N|\frac{nh}{X} \leqslant \max\{|z_0|, |z_N|\}\left(\left|1 - \frac{nh}{X}\right| + \left|\frac{nh}{X}\right|\right) = \max\{|z_0, |z_N|\},$$

$nh(X - nh) \leqslant X^2/4$.

Поэтому

$$\max_{0\leqslant n\leqslant N} |w_n| \leqslant \max\left\{|z_0|, |z_N|\right\} + Z\frac{X^2}{8}.$$

Лемма доказана.

Рассмотрим случай, когда функции $p(x)$ и $f(x)$ дважды непрерывно дифференцируемы. В курсе дифференциальных уравнений доказывается, что тогда решение $y(x)$ четырежды непрерывно дифференцируемо.

Пусть r_n — погрешность аппроксимации, соответствующая конечно-разностной схеме (3):

$$r_n = l\big(y(x_n)\big) - f_n = \\ = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1})}{h^2} - p(x_n)y(x_n) - f(x_n). \tag{5}$$

Поскольку $p(x)y(x) + f(x) = y''(x)$, то

$$r_n = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1})}{h^2} - y''(x_n).$$

Из оценок погрешности формул численного дифференцирования (§ 2.15) имеем

$$r_n = \frac{y^{(4)}(\bar{x}_n)h^2}{12}, \qquad \text{где} \quad x_{n-1} \leqslant \bar{x}_n \leqslant x_{n+1}. \tag{6}$$

Из-за округления получаемые в процессе вычислений приближения y_n к значениям $y(x_n)$ удовлетворяют системе (3), (4) с некоторыми погрешностями

$$l(y_n) - f_n = \delta_n. \tag{7}$$

Вычитая (5) из (7), получим уравнение

$$l(R_n) = \delta_n - r_n$$

относительно погрешности приближенного решения $R_n = y_n - y(x_n)$. Воспользовавшись леммой 2, получим

$$|R_n| \leqslant \max\Big\{|R_0|, |R_N|\Big\} + \frac{X^2}{8}\Big(\max_{0<n<N}|r_n| + \max_{0<n<N}|\delta_n|\Big).$$

Согласно оценке (6) имеем

$$|r_n| \leqslant M_4\frac{h^2}{12}, \quad M_4 = \max_{[0,\,X]}\Big|y^{(4)}(x)\Big|.$$

Таким образом, окончательная оценка погрешности имеет вид

$$\max_{0\leqslant n\leqslant N}|R_n| \leqslant \max\Big\{|R_0|, |R_N|\Big\}\frac{M_4X^2h^2}{96} + \frac{X^2}{8}\max_{0<n<N}|\delta_n|.$$

Мы видим, что при повышении точности, с которой удовлетворяются граничные условия и разностное уравнение, при одновременном стремлении шага к нулю решение сеточной задачи сближается с решением дифференциальной задачи.

Описанный метод дает приближенное решение, сходящееся к точному со скоростью $O(h^2)$. Займемся построением более точных схем. Будем предполагать, что функции $p(x)$ и $f(x)$ непрерывно дифференцируемы

четыре раза, тогда решение задачи непрерывно дифференцируемо шесть раз. Еще раз рассмотрим выражение

$$r_n = \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1})}{h^2} - y''(x_n).$$

Подставим сюда представление $y(x_{n\pm1})$ с помощью формулы Тейлора:

$$y(x_{n\pm1}) = y(x_n) \pm y'(x_n)h + y''(x_n)\frac{h^2}{2} \pm y'''(x_n)\frac{h^3}{6} + \\ + y^{(4)}(x_n)\frac{h^4}{24} \pm y^{(5)}(x_n)\frac{h^5}{120} + O(h^6)$$

и получим

$$r_n = \frac{y^{(4)}(x_n)h^2}{12} + O(h^4). \tag{8}$$

Вычтем из $l(y(x_n))$ слагаемое, аппроксимирующее величину $y^{(4)}(x_n)h^2/12$; полученной схеме соответствует погрешность аппроксимации более высокого порядка. Например, можно приблизить $y^{(4)}(x_n)$ выражением

$$\frac{\delta^4 y(x_n)}{h^4} = \frac{y(x_{n+2}) - 4y(x_{n+1}) + 6y(x_n) - 4y(x_{n-1}) + y(x_{n-2})}{h^4};$$

получится конечно-разностная схема

$$\frac{\delta^2 y_n}{h^2} - \frac{\delta^4 y_n}{12h^2} - p_n y_n = f_n. \tag{9}$$

Эту схему можно также построить непосредственно, заменив производную $y''(x_n)$ выражением $h^{-2}\Big(\delta^2 y(x_n) - (1/12)\delta^4 y(x_n)\Big)$, приближающим ее с погрешностью $O(h^4)$.

Уравнение (9) содержит пять неизвестных y_n с ненулевыми коэффициентами. Решение системы, состоящей из уравнений (9) и уравнений, получающихся при аппроксимации граничных условий, более трудоемко, чем решение системы (3). Исходя из других соображений, построим конечно-разностную схему с погрешностью аппроксимации $O(h^4)$ такую, что в каждое уравнение входят только три неизвестных.

Дифференцируя дважды исходное уравнение, имеем $y^{(4)}(x) = (p(x)y + f)''$, поэтому

$$y^{(4)}(x_n) \approx \frac{\delta^2(py+f)\big|_{x_n}}{h^2} = \\ = \frac{(p_{n+1}y(x_{n+1}) + f_{n+1}) - 2(p_n y(x_n) + f_n) + (p_{n-1}y(x_{n-1}) + f_{n-1})}{h^2}.$$

Вычитая из исходной схемы приближение для $y^{(4)}h^2/12$, получим схему

$$\frac{\delta^2 y_n}{h^2} - p_n y_n - \frac{1}{12}\delta^2(p_n y_n + f_n) = f_n$$

или

$$l^{(1)}(y_n) = \frac{\delta^2 y_n}{h^2} - p_n y_n - \frac{1}{12}\delta^2(p_n y_n) = \bar{l}^{(1)}(f_n) = f_n + \frac{1}{12}\delta^2 f_n.$$

Этот метод совпадает с методом Нумерова.

В предположении, что решение непрерывно дифференцируемо восемь раз, рассмотрим погрешность аппроксимации новой схемы $r_n^{(1)} = l^{(1)}(y(x_n)) - \bar{l}^{(1)}(f_n)$. Учитывая, что $p_n y(x_n) + f_n = y''(x_n)$, получим

$$r_n^{(1)} = l^{(1)}(y(x_n)) - \bar{l}^{(1)}(f_n) = \frac{\delta^2 y(x_n)}{h^2} - y''(x_n) - \frac{1}{12}\delta^2 y''(x_n).$$

Воспользовавшись формулой Тейлора, аналогично (8) получаем равенство

$$r_n^{(1)} = -\frac{y^{(6)}(x_n)h^4}{240} + O(h^6).$$

Построим приближение величины $y^{(6)}(x_n)$ посредством значений $y(x_{n-1})$, $y(x_n)$, $y(x_{n+1})$. Таких приближений можно написать очень много, например следующим образом. Согласно уравнению (1)

$$y'' = py + f, \quad y''' = (py+f)', \quad y^{(4)} = (py+f)'',$$
$$y^{(6)} = (py+f)^{(4)} = py^{(4)} + 4p'y^{(3)} + 6p''y'' + 4p^{(3)}y' + p^{(4)}y + f^{(4)},$$

поэтому справедливо приближенное равенство

$$\begin{aligned} y^{(6)}(x_n) \approx p_n &\frac{\delta^2(p_n y(x_n) + f_n)}{h^2} + \\ &+ 4p'(x_n)\frac{(p_{n+1}y(x_{n+1}) + f_{n+1}) - (p_{n-1}y(x_{n-1}) + f_{n-1})}{2h} + \\ &+ 6p''(x_n)(p_n y(x_n) + f_n) + 4p^{(3)}(x_n)\frac{y(x_{n+1}) - y(x_{n-1})}{2h} + \\ &+ p^{(4)}(x_n)y(x_n) + f^{(4)}(x_n). \end{aligned}$$

Прибавляя к $l^{(1)}(y_n)$ выражение, приближающее $y^{(6)}(x_n)h^4/240$, получим конечно-разностную схему с погрешностью аппроксимации $O(h^6)$; при этом в каждое уравнение получившейся алгебраической системы входят только три неизвестных y_n.

Для практической оценки погрешности решения краевой задачи может применяться правило Рунге. Законность его применения основывается на существовании главного члена погрешности.

Задача 1. Пусть функции $p(x)$ и $f(x)$ четырежды дифференцируемы. Доказать, что для решения задачи (3), (4) справедливо соотношение

$$\max_n \left| y_n - y(x_n) - h^2 z(x_n) \right| = O(h^4);$$

здесь $z(x)$ — решение краевой задачи

$$Lz = -y^{(4)}(x)/12, \quad z(0) = 0, \quad z(X) = 0.$$

Аналогичный прием последовательного повышения порядка погрешности аппроксимации может быть применен и по отношению к аппроксимациям граничного условия.

Рассмотрим случай граничного условия $y'(0) - \alpha y(0) = a$. Дискретное приближение высокой точности к такому граничному условию можно получить непосредственно, заменив производную $y'(0)$ по какой-либо формуле численного дифференцирования высокой точности:

$$y'(0) \approx \sum_{i=0}^{l} \frac{c_i y(x_i)}{h}.$$

Однако погрешность аппроксимации будет меньше и решение возникающей алгебраической системы представит меньше трудностей, если идти по описанному выше пути последовательного повышения порядка точности аппроксимации.

Заменим производную $y'(0)$ отношением $\dfrac{y(h) - y(0)}{h}$; тогда получим $l_0^{(1)}(y_n) = \dfrac{y_1 - y_0}{h} - \alpha y_0 - a = 0$. Подставляя в $r_0^{(1)} = \dfrac{y(h) - y(0)}{h} - \alpha y(0) - a$ разложение $y(h) = y(0) + y'(0)h + O(h^2)$, имеем

$$r_0^{(1)} = y'(0) + \frac{y''(0)h}{2} + O(h^2) - \alpha y(0) - a = \frac{y''(0)h}{2} + O(h^2).$$

Таким образом, погрешность аппроксимации граничного условия есть $O(h)$. Поскольку, согласно уравнению (1),

$$y''(0) = p_0 y(0) + f_0,$$

то уравнению

$$l_0^2(y_n) = \frac{y_1 - y_0}{h} - \alpha y_0 - a - (p_0 y_0 + f_0)\frac{h}{2} = 0$$

соответствует второй порядок аппроксимации. Подставляя в $r_0^{(2)} = l_0^{(2)}(y(x_n))$ разложение $y(h) = y(0) + y'(0)h + y''(0)h^2/2 + O(h^3)$, получим $r_0^{(2)} = \dfrac{y^{(3)}(0)h^2}{6} + O(h^3)$. После дифференцирования исходного уравнения (1) имеем

$$y^{(3)}(0) - p_0 y'(0) - p'(0)y(0) - f'(0) = 0;$$

поэтому с учетом граничного условия справедливо равенство

$$y^{(3)}(0) = p_0(\alpha y(0) + a) + p'(0)y(0) + f'(0).$$

Разностному уравнению

$$l_0^{(3)}(y_n) = l_0^{(2)}(y_n) - \Big((p_0\alpha + p'(0))y_0 + p_0 a + f'(0)\Big)\frac{h^2}{6} = 0$$

будет соответствовать уже третий порядок аппроксимации.

Можно было бы сразу написать равенство

$$r_0^{(1)} = y''(0)\frac{h}{2} + y^{(3)}(0)\frac{h^2}{6},$$

затем выразить производные $y''(0)$ и $y^{(3)}(0)$ через $y(0)$ и, вычитая из разностной схемы соответствующее выражение, получить уравнение $l_0^{(3)}(y_n) = 0$. Однако мы обратили основное внимание именно на способ последовательного повышения порядка точности, поскольку его перенесение на случай уравнений в частных производных является наиболее простым и естественным.

§ 2. Функция Грина сеточной краевой задачи

Функция w_n, введенная в § 1 при доказательстве леммы 2, носит название *мажорирующей функции*, и метод получения оценок погрешности с использованием этой функции называется *методом мажорант*, или *методом Гершгорина*.

В ряде случаев, когда метод мажорант неприменим, оценку погрешности приближенного решения можно получить, используя так называемую сеточную функцию Грина. Проводимые ниже построения функции Грина сеточной краевой задачи (1.3), (1.4) кроме всего прочего интересны своей аналогией со случаем дифференциальной краевой задачи.

Функция Грина $G(x, s)$ дифференциальной краевой задачи

$$Ly = y'' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b,$$

определяется как решение уравнения

$$L(G(x, s)) = G_{xx}(x, s) - p(x)G(x, s) = \delta(x - s) \tag{1}$$

при граничном условии $G(0, s) = G(X, s) = 0$; здесь $\delta(x)$ — δ-функция. Функцию Грина можно задать следующими явными формулами. Пусть $W^1(x)$, $W^2(x)$ — решения уравнения $L(W) = 0$ при условиях

$$W^1(0) = 0, \ (W^1)'(0) = 1,$$

$$W^2(X) = 0, (W^2)'(X) = -1;$$

тогда

$$G(x, s) = \begin{cases} \dfrac{W^2(s)W^1(x)}{V^0} & \text{при} \quad 0 \leqslant x \leqslant s, \\ \dfrac{W^1(s)W^2(x)}{V^0} & \text{при} \quad s \leqslant x \leqslant X, \end{cases} \tag{2}$$

где V^0 — значение определителя Вронского:

$$V(x) = \begin{vmatrix} W^1(x) & W^2(x) \\ (W^1)'(x) & (W^2)'(x) \end{vmatrix} = \text{const} = V^0.$$

Решение задачи (1.1), (1.2) записывается с помощью функции Грина в виде

$$y(x) = \int_0^X G(x, s) f(s)\, ds + G'_s(x, X)b - G'_s(x, 0)a. \tag{3}$$

Перейдем к сеточной задаче

$$l(y_n) = \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = f_n, \quad y_0 = a, \quad y_N = b.$$

По аналогии определим функции W_n^1, W_n^2 из соотношений

$$l(W_n^i) = 0, \quad i = 1, 2, \quad n = 1, \ldots, N-1,$$
$$W_0^1 = 0, \quad W_1^1 = h, \quad W_N^2 = 0, \quad W_{N-1}^2 = h.$$

Здесь и далее оператор l применяется при фиксированном верхнем индексе

$$l(W_n^i) = \frac{W_{n+1}^i - 2W_n^i + W_{n-1}^i}{h^2} - p_n W_n^i.$$

Аналогом определителя Вронского является определитель

$$V_n = \begin{vmatrix} W_n^1 & W_n^2 \\ \dfrac{W_n^1 - W_{n-1}^1}{h} & \dfrac{W_n^2 - W_{n-1}^2}{h} \end{vmatrix} = \frac{W_n^2 W_{n-1}^1 - W_n^1 W_{n-1}^2}{h}.$$

Из тождества

$$0 = W_n^1 l(W_n^2) - W_n^2 l(W_n^1) =$$
$$= \frac{W_n^1 W_{n+1}^2 + W_n^1 W_{n-1}^2 - W_n^2 W_{n+1}^1 - W_n^2 W_{n-1}^1}{h^2} = \frac{V_{n+1} - V_n}{h^2}$$

следует, что величина V_n не зависит от n; будем обозначать ее $V(h)$.

Положим

$$G_n^k = \begin{cases} \dfrac{W_k^2 W_n^1}{V(h)} & \text{при} \quad 0 \leqslant n \leqslant k, \\ \dfrac{W_k^1 W_n^2}{V(h)} & \text{при} \quad k \leqslant n \leqslant N. \end{cases} \tag{4}$$

Из определения W_n^i следуют равенства

$$l(G_n^k) = \frac{W_k^2}{V(h)} l(W_n^1) = 0$$

при $n < k$,

$$l(G_n^k) = \frac{W_k^1}{V(h)} l(W_n^2) = 0$$

при $n > k$; при $k = n$ имеем

$$l(G_n^k) = \frac{G_{n+1}^n - 2G_n^n + G_{n-1}^n}{h^2} - p_n G_n^n =$$
$$= \left(\frac{W_{n+1}^2 W_n^1 - 2W_n^1 W_n^2 + W_{n-1}^1 W_n^2}{V(h)h^2} - \frac{p_n W_n^1 W_n^2}{V(h)} \right) =$$
$$= \frac{W_n^1}{V(h)} \left(\frac{W_{n+1}^2 - 2W_n^2 + W_{n-1}^2}{h^2} - p_n W_n^2 \right) + \frac{W_n^2 W_{n-1}^1 - W_n^1 W_{n-1}^2}{V(h)h^2}.$$

Согласно определению W_n^2 первая круглая скобка равна 0; вторая скобка равна $V(h)h$. В итоге при $k = n$ получаем $l(G_n^k) = h^{-1}$; объединяя полученные соотношения, имеем

$$l(G_n^k) = \delta_n^k h^{-1}. \tag{5}$$

Функция $\delta_n^k h^{-1}$ является сеточным аналогом δ-функции, а равенство (5) — аналогом (1). Сеточные функции W_n^i, $i = 1, 2$, являются решениями сеточных задач Коши, соответствующих задачам Коши, определяющим функции $W^i(x)$.

Предположим, что

$$\|p''\|_{C[0,X]} = \sup_{0 \leqslant x \leqslant X} |p''| \leqslant M_0 < \infty; \tag{6}$$

тогда можно показать, что

$$\left\| \left(W^i(x) \right)^{(4)} \right\|_{C[0,X]} < \infty.$$

Оценив близость начальных данных и воспользовавшись далее (недоказанной) теоремой из § 8.10, можно получить оценку близости сеточных и дифференциальных решений задачи Коши вида

$$\max_{0 \leqslant x_n \leqslant X} \left| W_n^i - W^i(nh) \right| \leqslant Mh^2, \quad i = 1, 2. \tag{7}$$

Согласно свойствам V^0 и $V(h)$ имеем

$$V^0 = V(X) = \begin{vmatrix} W^1(X) & 0 \\ (W^1)'(X) & -1 \end{vmatrix} = -W^1(X),$$

$$V(h) = \begin{vmatrix} W_N^1 & 0 \\ \dfrac{W_N^1 - W_{N-1}^1}{h} & -1 \end{vmatrix} = -W_N^1.$$

Поэтому

$$\left| V(h) - V^0 \right| \leqslant Mh^2 \tag{8}$$

на основании оценки (7).

Далее предполагается, что $W^1(X) \neq 0$; в противном случае однородная краевая задача $y'' - p(x)y = 0$, $y(0) = y(X) = 0$ имеет ненулевое решение $W^1(x)$, а следовательно, неоднородная краевая задача (1.1), (1.2) или не имеет решения, или имеет неединственное решение. Будем также предполагать h настолько малым, что $2Mh^2 \leqslant \left|W^1(X)\right| = |V^0|$. В этом случае имеем

$$V(h) \geqslant V^0 - Mh^2 \geqslant |V^0/2| > 0.$$

Сравнивая явные выражения функции Грина дифференциальной (2) и сеточной (4) задач с учетом оценок (7), (8), получаем

$$\left|G_n^k - G(nh,\, kh)\right| \leqslant Qh^2$$

при $0 \leqslant kh,\ nh \leqslant X$.

Лемма. *Если $V(h) \neq 0$, то рассматриваемая сеточная задача* (1.3), (1.4) *однозначно разрешима и ее решение записывается в виде*

$$y_n = h\sum_{k=1}^{N-1} G_n^k f_k + \frac{W_n^2}{W_0^2}\,a + \frac{W_n^1}{W_N^1}\,b. \tag{9}$$

Формула (9) *является сеточным аналогом формулы* (2).

Доказательство. Поскольку согласно определению функции G_n^k здесь выполняется равенство $G_0^k = G_N^k \equiv 0$, то $y_0 = a$, $y_N = b$. Имеем равенство

$$l(y_n) = h\sum_{k=1}^{N-1} l(G_n^k) f_k + \frac{l(W_n^2)}{W_0^2}\,a + \frac{l(W_n^1)}{W_N^1}\,b;$$

коэффициенты при a и b равны нулю по определению функций W_n^i. Воспользовавшись равенством (5), получим $l(y_n) = f_n$. Таким образом, система линейных уравнений (1.3), (1.4) при любых правых частях имеет решение, записываемое в виде (9). Но если система линейных уравнений разрешима при любых правых частях, то ее решение единственно. Лемма доказана.

Соотношение (9) может быть двояким способом использовано для оценки близости решений сеточной и дифференциальной задач.

Первый путь состоит в непосредственном сравнении выражений (2) и (9). Для простоты продемонстрируем его в случае $a = b = 0$, $\delta_n \equiv 0$. Подставляя $x = nh$, перепишем (2) в виде

$$y(nh) = \int_0^{nh} G(nh,\, s) f(s)\, ds + \int_{nh}^{X} G(nh,\, s) f(s)\, ds.$$

При условиях $\|f''\|_{C[0,X]} \leqslant \infty$ и (6) можно показать, что подынтегральные функции в обоих интегралах имеют ограниченную вторую производную. Заменим оба эти интеграла по формуле трапеций с шагом h. Получим

$$y(nh) = h\left(\frac{G(nh,\,0)f(0)}{2} + \sum_{k=1}^{n-1} G(nh,\,kh)f(kh) + \frac{G(nh,\,nh)f(nh)}{2}\right) +$$
$$+ h\left(\frac{G(nh,\,nh)f(nh)}{2} + \sum_{k=n+1}^{N-1} G(nh,\,kh)f(kh) + \frac{G(nh,\,Nh)f(Nh)}{2}\right) + O(h^2).$$

Разбиение интеграла на две части потребовалось в связи с тем, что погрешность формулы трапеций оценивается через вторую производную подынтегральной функции, а функция $G(nh,\,s)$ имеет разрыв первой производной в точке nh.

Поскольку из определения функции Грина $G(x,\,s)$ следует, что $G(nh,\,0) = G(nh,\,N) = 0$, то

$$y(nh) = h\sum_{k=1}^{N-1} G(nh,\,kh)f_k + O(h^2). \tag{10}$$

Из равенств (9), (10) следует

$$y_n - y(nh) = h\sum_{k=1}^{N-1}\Big(G_n^k - G(nh,\,kh)\Big)f_k + O(h^2).$$

Воспользовавшись оценкой (7), получим $y_n - y(nh) = O(h^2)$. Нетрудно проследить, что эта оценка погрешности равномерна по n при $0 \leqslant x_n \leqslant X$, т. е.

$$\max_{0\leqslant x_n\leqslant X}\Big|y_n - y(nh)\Big| = O(h^2).$$

Другой путь состоит в получении уравнения для погрешности и дальнейшей оценки погрешности с помощью формулы Грина. В § 1 было показано, что погрешность R_n удовлетворяет уравнению

$$l(R_n) = \delta_n - r_n,$$

где $|r_n| \leqslant M_4h^2/12$, δ_n — погрешность, обусловленная неточностью решения системы (1.3), (1.4). Пусть $R_0 = R_N = 0$. Из явного вида функции Грина дифференциальной задачи следует ее равномерная ограниченность; с учетом (7) получаем, что функция G_n^k также равномерно ограничена: $|G_n^k| \leqslant D$ при $0 \leqslant kh,\ nh \leqslant X$. Напишем явное представление R_n с помощью формулы Грина (9):

$$R_n = h\sum_{k=1}^{N-1} G_n^k(\delta_k - r_k). \tag{11}$$

Тогда из (11) следует оценка

$$|R_n| \leqslant Dh \sum_{k=1}^{N-1} \Big(|\delta_k| + |r_k| \Big).$$

Загрубляя, получим оценку

$$|R_n| \leqslant DX \left(\frac{M_4}{12} h^2 + \max_{0<k<N} |\delta_k| \right),$$

имеющую тот же порядок по отношению к h и $\max|\delta_k|$, что и полученная в § 1 оценка.

Эта оценка применима и при $\inf p(x) < 0$, когда лемма 1.2 неприменима. Таким образом, использование аппарата функции Грина позволяет расширить множество задач, для которых удается получить оценку погрешности сеточного решения.

§ 3. Решение простейшей краевой сеточной задачи

При численном решении задачи Коши значения решения определяются в последовательных узлах по рекуррентным формулам; в случае краевой сеточной задачи, например (1.3), (1.4), такой возможности нет, поскольку значения решения зависят от граничных условий на обоих концах отрезка интегрирования.

Краевую задачу

$$y'' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b$$

можно было бы решать следующим способом. Возьмем частное решение $y_0(x)$ неоднородного уравнения

$$y_0'' - p(x)y_0 = f(x)$$

и два линейно независимых решения однородного уравнения $y_i'' - p(x)y_i = 0$, $i = 1, 2$. Общее решение неоднородного уравнения запишется в виде

$$y(x) = y_0(x) + C_1 y_1(x) + C_2 y_2(x);$$

постоянные C_1 и C_2 определяем из граничных условий. Приближения к функциям $y_i(x)$, $i = 0, 1, 2$, находим каким-либо численным методом решения задачи Коши, затем определяем C_i и получаем нужное решение.

Экономнее поступить следующим образом. Находим частное решение неоднородного уравнения $y_0'' - p(x)y_0 = f(x)$, удовлетворяющее условию $y_0(0) = a$, и частное решение однородного уравнения $y_1(x)$, удовлетворяющее условию $y_1(0) = 0$. Общее решение неоднородного уравнения, удовлетворяющее условию $y(0) = a$, имеет вид $y_0(x) + Cy_1(x)$; значение C определяется из условия $y_0(X) + Cy_1(X) = b$.

Метод решения краевой задачи, соответствующий этой схеме, принято называть *методом стрельбы* или *методом пристрелки*. Сеточный аналог этого метода заключается в следующем. Задаемся $y_0^0 = a$, $y_0^1 = 0$, произвольными $y_1^1 \neq 0$ и y_1^0 и из уравнений

$$l(y_n^0) = \frac{y_{n+1}^0 - 2y_n^0 + y_{n-1}^0}{h^2} - p_n y_n^0 = f_n, \tag{1}$$

$$l(y_n^1) = \frac{y_{n+1}^1 - 2y_n^1 + y_{n-1}^1}{h^2} - p_n y_n^1 = 0, \tag{2}$$

последовательно определяем $y_2^0, \ldots, y_N^0$, $y_2^1, \ldots, y_N^1$. Затем находим C из уравнения $y_N^0 + Cy_N^1 = b$ и полагаем $y_n = y_n^0 + Cy_n^1$; функция y_n является требуемым решением.

Иногда значения y_n^i, $i = 0, 1$ не хранят в процессе вычислений, а после отыскания C находят $y_1 = y_1^0 + Cy_1^1$ и затем последовательно определяют $y_2, \ldots, y_{N-1}$ из уравнения $l(y_n) = f_n$. Описанный алгоритм формально применим при любых значениях y_1^0 и y_1^1; однако для уменьшения влияния вычислительной погрешности разумнее взять $y_1^0 = a + O(h)$.

Рассмотрим модельное уравнение

$$y'' - p(x)y = 0, \quad p = \text{const} > 0, \quad x \in [0, X]. \tag{3}$$

Решения соответствующего однородного уравнения $y(x) = \exp\{\pm\sqrt{p}x\}$ сильно возрастают (убывают) на рассматриваемом участке, если величина $\sqrt{p}X$ достаточно велика. Таким образом, при $p > 0$ величина $\sqrt{p}\,X$ является параметром, существенно влияющим на характер накопления вычислительной погрешности. Рассмотрим накопление вычислительной погрешности на одном из этапов решения сеточной задачи. Пусть уже найдено значение y_1 и далее вычисления проходят по рекуррентной формуле

$$y_{n+1} = 2y_n - y_{n-1} + ph^2 y_n. \tag{4}$$

Получаемые при реальных вычислениях величины y_n^* связаны равенством

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + ph^2 y_n^* + \delta_n;$$

наличие слагаемого δ_n в правой части обусловлено округлениями. Вычитая отсюда (4), получим уравнение относительно погрешности $\Delta_n = y_n^* - y_n$:

$$\Delta_{n+1} = 2\Delta_n - \Delta_{n-1} + ph^2\Delta_n + \delta_n. \tag{5}$$

Рассмотрим следующую модель накопления погрешности: $\delta_n = \delta = \text{const}$ и погрешности в y_0 и y_1 отсутствуют, т. е. $\Delta_0 = \Delta_1 = 0$. Уравнение (5) можно переписать в виде

$$\frac{\Delta_{n+1} - 2\Delta_n + \Delta_{n-1}}{h^2} = p\Delta_n + \omega, \quad \omega = \frac{\Delta}{h^2}. \tag{6}$$

При $\omega = \text{const}$ совокупность соотношений $\Delta_0 = \Delta_1 = 0$ и (6) образует сеточную задачу, аппроксимирующую задачу Коши

$$\Delta'' = p\Delta + \omega, \quad \Delta(0) = \Delta'(0) = 0.$$

Можно проверить, что условия теоремы из § 8.10 выполнены, поэтому решения этих задач будут близки:

$$\Delta_n \sim \Delta(x_n) = \frac{\operatorname{ch}\{\sqrt{p}\,x_n\} - 1}{p}\,\omega. \tag{7}$$

Асимптотическое равенство понимается здесь в обычном смысле:

$$\Delta_n/\Delta(x_n) \to 1 \quad \text{при} \quad h \to 0, \quad \sqrt{p},\ x_n = \text{const},\ x_n \neq 0. \tag{8}$$

При умножении ω на какой-либо множитель на этот множитель умножится как Δ_n, так и $\Delta(x_n)$, поэтому (7) и соответственно (8) справедливы и при ω, зависящем от h, т. е. в рассматриваемом случае

$$\Delta_n \sim \sigma_n = \frac{\operatorname{ch}\{\sqrt{p}\,x_n\} - 1}{ph^2}\,\delta.$$

Рассмотрим числовой пример: $p = 10$, $x_n = 10$, $h = 10^{-2}$, $\delta = 10^{-2}$; тогда $|\sigma_n| > 10^4$ и нельзя рассчитывать на получение решения с разумной точностью. Отметим, что значение величины $\sqrt{p}x_n$, определившее столь большую величину накопленной вычислительной погрешности, было не очень большим.

При естественной нумерации неизвестных $y_0, \ldots, y_N$ система уравнений (1.3), (1.4) записывается в виде $A\mathbf{y} = \mathbf{c}$, где матрица A трехдиагональная. Напомним, что матрица $A = [a_{ij}]$ называется $(2m+1)$-диагональной, если $a_{ij} = 0$ при $|i - j| > m$.

Для решения систем уравнений такого вида часто наиболее целесообразно применять метод Гаусса при естественном порядке исключения неизвестных. В случае, когда этот метод применяется для решения систем уравнений, возникающих при аппроксимации краевых задач, его называют *методом прогонки*.

Приведем конкретные расчетные формулы метода прогонки в случае системы (1.3), (1.4). Представим граничное условие $y_0 = a$ в виде $y_0 = C_0 y_1 + \varphi_0$, где $C_0 = 0$, $\varphi_0 = a$. Подставляя $y_0 = C_0 y_1 + \varphi_0$ в первое уравнение системы (1.3), (1.4)

$$\frac{y_0 - 2y_1 + y_2}{h^2} - p_1 y_1 = f_1,$$

получаем уравнение, связывающее значения y_1 и y_2. Разрешая это уравнение относительно y_1, имеем

$$y_1 = C_1 y_2 + \varphi_1, \tag{9}$$

где

$$C_1 = \frac{1}{2 + p_1 h^2}, \quad \varphi_1 = C_1(a - f_1 h^2).$$

Подставляя полученное выражение y_1 через y_2 во второе уравнение системы, получим уравнение, связывающее y_2 и y_3, и т. д. Пусть уже получено соотношение

$$y_n = C_n y_{n+1} + \varphi_n; \tag{10}$$

подставим выражение y_n в $(n+1)$-е уравнение системы (1.3), (1.4):

$$\frac{y_n - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1}y_{n+1} = f_{n+1}.$$

Разрешая получившееся уравнение

$$\frac{(C_n y_{n+1} + \varphi_n) - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1}y_{n+1} = f_{n+1}$$

относительно y_{n+1}, имеем

$$y_{n+1} = C_{n+1}y_{n+2} + \varphi_{n+1};$$

здесь

$$C_{n+1} = \frac{1}{2 + p_{n+1}h^2 - C_n}, \quad \varphi_{n+1} = C_{n+1}(\varphi_n - f_{n+1}h^2). \tag{11}$$

Таким образом, коэффициенты уравнений (10), связывающих последовательные значения y_n и y_{n+1}, можно определять из рекуррентных соотношений (11) при начальных условиях $C_0 = 0$, $\varphi_0 = a$. Так как y_N известно, то после нахождения всех коэффициентов C_n, φ_n можно последовательно определять $y_{N-1}, \ldots, y_1$ из соотношений (10). Процесс вычисления коэффициентов C_n, φ_n принято называть *прямым ходом прогонки*, а процесс вычисления неизвестных y_n — *обратным ходом прогонки*. Последовательность производимых вычислений можно изобразить следующей схемой (на этой схеме $a \to b$ означает, что значение a используется при вычислении b):

прямой ход прогонки

$$\begin{array}{ccccccccc} & & p_1 & & p_2 & & & & p_{N-1} \\ & & \downarrow & & \downarrow & & & & \downarrow \\ C_0 = 0 \to & & C_1 & \to & C_2 & \to \cdots \to & & & C_{N-1} \\ & & \downarrow & & \downarrow & & & & \downarrow \\ \varphi_0 = a \to & & \varphi_1 & \to & \varphi_2 & \to \cdots \to & & & \varphi_{N-1} \\ & & \uparrow & & \uparrow & & & & \uparrow \\ & & f_1 & & f_2 & & & & f_{N-1} \end{array};$$

обратный ход прогонки

$$\begin{array}{ccccccccc} C_1 & & C_2 & & & & C_{N-1} & & \\ \downarrow & & \downarrow & & & & \downarrow & & \\ y_1 & \leftarrow & y_2 & \leftarrow \cdots \leftarrow & & & y_{N-1} & \leftarrow & y_N. \\ \uparrow & & \uparrow & & & & \uparrow & & \\ \varphi_1 & & \varphi_2 & & & & \varphi_{N-1} & & \end{array}$$

Названию «метод прогонки» иногда предлагается следующее объяснение. Уравнение

$$y_n = C_n y_{n+1} + \varphi_n$$

получено как следствие граничного условия в точке $x = 0$ и уравнений системы (1.3), соответствующих точкам $x_i \leqslant x_n$. Таким образом, это равенство выполнено для любого решения системы уравнений $l(y_j) = f_j$, $j = 1, \ldots, n$, удовлетворяющего левому граничному условию; граничное условие в точке $x = 0$ «перегоняется» в текущую точку $x = x_n$.

Задача 1. Выписать расчетные формулы метода квадратного корня в случае решения рассматриваемой системы. Сравнить трудоемкость этого метода с трудоемкостью метода прогонки.

Применим для решения системы (1.3), (1.4) метод Гаусса при порядке исключения неизвестных $y_0, y_2, \ldots, y_N, y_1$. Из первого уравнения выражаем y_2 через y_1 и подставляем в остальные уравнения. После этого во второе уравнение входят только неизвестные y_1 и y_3; выражаем y_3 через y_1 и подставляем в остальные уравнения и т. д. Пусть уже выразили через y_1 неизвестные $y_2, \ldots, y_n$ и получили соотношения $y_j = \alpha_j y_1 + \beta_j$ при $2 \leqslant j \leqslant n$; для единообразия добавим сюда равенства

$$y_0 = \alpha_0 y_1 + \beta_0, \quad \text{где} \quad \alpha_0 = 0,\ \beta_0 = a,$$

$$y_1 = \alpha_1 y_1 + \beta_1, \quad \text{где} \quad \alpha_1 = 1,\ \beta_0 = 0.$$

Подставляя выражения y_{n-1} и y_n в уравнение

$$y_{n-1} - (2 + p_n h^2) y_n + y_{n+1} = f_n h^2,$$

получим

$$(\alpha_{n-1} y_1 + \beta_{n-1}) - (2 + p_n h^2)(\alpha_n y_1 + \beta_n) + y_{n+1} = f_n h^2,$$

или

$$y_{n+1} = \alpha_{n+1} y_1 + \beta_{n+1},$$

где

$$\begin{aligned} \alpha_{n+1} &= (2 + p_n h^2)\alpha_n - \alpha_{n-1}, \\ \beta_{n+1} &= (2 + p_n h^2)\beta_n - \beta_{n-1} + f_n h^2. \end{aligned} \tag{12}$$

Таким образом, коэффициенты α_n, β_n можно последовательно вычислять по рекуррентным формулам (12). После получения соотношения $y_N = \alpha_N y_1 + \beta_N$ определяем значение y_1, а затем все y_j по формулам

$$y_j = \alpha_j y_1 + \beta_j. \tag{13}$$

Вследствие (12) значения α_j удовлетворяют однородному конечно-разностному уравнению (2) и начальным условиям $\alpha_0 = 0$, $\alpha_1 = 1$; значения β_j — неоднородному конечно-разностному уравнению (1) и начальным условиям $\beta_0 = a$, $\beta_1 = 0$. Таким образом, α_n совпадает с y_n^1, а β_n — с y_n^0 в методе стрельбы при определенном выборе начальных условий. Функция $z_n = \alpha_n C + \beta_n$ удовлетворяет неоднородному конечно-разностному уравнению и левому граничному условию при любых C, причем $z_1 = \alpha_1 C + \beta_1 = C$. Следовательно, решая уравнение $y_N = \alpha_N C + \beta_N$ относительно C, мы как раз находим значение $C = y_1$, которое отыскивалось в методе стрельбы. Вычисления по формуле (13) соответствуют вычислению значений y_n по формуле $y_n = y_n^1 C + y_n^0$. Полученный метод совпадает с методом стрельбы при

$$y_0^0 = a, \quad y_1^0 = 0, \quad y_0^1 = 1, \quad y_1^1 = 0.$$

Решение системы (1.3), (1.4) описанными выше методами требует $O(N)$ арифметических операций; если при решении этой системы обратиться непосредственно к стандартной программе метода Гаусса, то число операций будет порядка $O(N^3)$; это количество операций состоит из $O(N)$ *содержательных операций* метода прогонки и $O(N^3)$ *несодержательных операций* умножения (деления) нуля на некоторое число, и сложения (вычитания) двух нулей. Таким образом, хотя содержательные операции при решении этой системы методом прогонки и по стандартной программе метода Гаусса одни и те же, использование стандартной программы приводит к существенному увеличению затрат на решение задачи.

Текст программы метода Гаусса может быть преобразован в текст программы метода прогонки, если в преобразованиях над строками и столбцами матрицы изменить начало и конец циклов так, чтобы исключить несодержательные операции.

Рассмотрим теперь случай, когда при $x = 0$ имеем граничное условие $y'(0) - \alpha y(0) = a$. В § 1 при его аппроксимации возникло уравнение, связывающее значения y_0 и y_1; это уравнение может быть записано в виде

$$y_0 = C_0 y_1 + \varphi_0; \tag{14}$$

например, простейшую аппроксимацию $\dfrac{y_1 - y_0}{h} - \alpha y_0 = a$ можно переписать в таком виде при $C_0 = \dfrac{1}{1+\alpha h}$, $\varphi_0 = -\dfrac{ah}{1+\alpha h}$. Далее вычисляем C_n, φ_n по формулам (11) при $n = 1, \ldots, N-1$ и при обратном ходе прогонки $y_{N-1}, \ldots, y_0$ по формулам (10). Если при $x = X$ имеем граничное условие $y'(X) + \beta y(X) = b$, то аналогично (14) имеем уравнение

$$y_N = \bar{C}_N y_{N-1} + \bar{\varphi}_N. \tag{15}$$

Осуществляя прямой ход прогонки, получим равенство

$$y_{N-1} = C_{N-1} y_N + \varphi_N; \tag{16}$$

решая систему (15), (16), находим y_N, y_{N-1} и затем последовательно вычисляем $y_{N-2}, \ldots, y_0$ по формулам (10).

При аппроксимации краевых задач для уравнений высших порядков или для систем дифференциальных уравнений появляются системы уравнений $A\mathbf{y} = \mathbf{c}$ с (l, s)-*диагональными матрицами* A; матрицу $A = [a_{ij}]$ называют (l, s)-*дагональной*, если $a_{ij} = 0$ при $j < i - l$ и при $j > i + s$. Для решения таких уравнений также довольно часто бывает целесообразно применять метод Гаусса, алгоритм которого может быть записан аналогично методу прогонки в виде совокупности рекуррентных формул.

Если $l > s$, то для уменьшения числа арифметических операций целесообразно переобозначить неизвестные и уравнения в обратном порядке, чтобы получить систему с (s, l)-диагональной матрицей.

Задача 2. Подсчитать число операций при решении методом Гаусса системы с (l, s)-диагональной матрицей при $l > s$, $l = s$, $l < s$.

В ряде случаев (l, s)-диагональная матрица системы уравнений записывается естественным образом в виде (p, q)-диагональной матрицы клеточного вида, т. е. $A = [A_{ij}]$, A_{ij} — некоторые матрицы такие, что $A_{ij} = 0$, если $j < i - p$ или $j > i + q$.

Рассмотрим случай системы уравнений

$$\mathbf{y}'' - p(x)\mathbf{y} = \mathbf{f}(x), \tag{17}$$

где $\mathbf{y}$, $\mathbf{f}$ — векторы размерности m, p — матрица размерности $m \times m$, и простейшую аппроксимацию

$$\frac{\mathbf{y}_{n+1} - 2\mathbf{y}_n + \mathbf{y}_{n-1}}{h^2} - p(x_n)\mathbf{y}_n = \mathbf{f}(x_n). \tag{18}$$

Матрица системы естественным способом записывается в виде (17) с $p = q = 1$; A_{ij} — матрицы размерности $m \times m$. В то же самое время эта матрица является $(2m-1, 2m-1)$-диагональной или, что то же самое, $(4m-1)$-диагональной. Для решения этой системы может быть применен метод исключения Гаусса в клеточной форме, который аналогично скалярному случаю может быть записан в виде совокупности рекуррентных матричных соотношений типа формул метода прогонки.

Задача 3. Подсчитать число арифметических операций для метода Гаусса в клеточной форме и для общей процедуры метода Гаусса с исключением несодержательных операций в применении к системе уравнений (18).

При решении ряда задач возникают системы уравнений с матрицей A, отличающейся по структуре от (l, s)-диагональной матрицы наличием ненулевых элементов при $|i-j| \sim n$, т. е. вблизи левого нижнего и правого верхнего углов матрицы. Для решения таких систем также целесообразно применять метод Гаусса с исключением несодержательных операций. В случае отыскания периодического решения сеточного уравнения (3) этот вариант метода Гаусса называют *методом циклической прогонки.*

Рассмотрим пример задачи, сводящейся к системе уравнений такого вида. При сглаживании функций методом регуляризации в § 5.3 возникла следующая задача. Дана периодическая функция f_q целочисленного аргумента q; период равен N. Требуется найти периодическую с тем же периодом функцию u_q, удовлетворяющую системе соотношений

$$\frac{\delta^{2n} u_q}{h^{2n}} + (-\lambda^2)^n u_q = f_q \quad \text{при всех} \quad q.$$

Выпишем эти соотношения при $q = 1, \ldots, N$. Вследствие условия периодичности заменим значения u_q при $q \leqslant 0$ и при $q > N$, входящие в эти соотношения,

соответственно на u_{q+N} и u_{q-N}. В результате этого получится система N уравнений относительно N неизвестных $u_1, \ldots, u_N$: $A\mathbf{u} = \mathbf{f}$. Элементы матрицы $A = [a_{ij}]$ этой системы определяются соотношениями $a_{ij} = a(|i-j|)$, где

$$a(0) = (-1)^n (C_{2n}^n h^{-2n} + \lambda^{2n}),$$

$$a(k) = \begin{cases} (-1)^{n-k} C_{2n}^{n-k} h^{-2n} & \text{при} \quad 0 < k \leqslant n, \\ 0 & \text{при} \quad n < k \leqslant N-n, \\ (-1)^{k+n-N} C_{2n}^{k+n-N} h^{-2n} & \text{при} \quad N-n \leqslant k < N; \end{cases}$$

$$A = \begin{pmatrix} a\ (0) & a\ (1) & \ldots & a\ (n) & 0 & \ldots & 0 & a\ (N-n) & \ldots & a\ (N-1) \\ a(1) & \cdot & \cdot & & & & & & & \\ \vdots & \cdot & \cdot & \cdot & & & & & & \\ a(n) & & \cdot & \cdot & \cdot & & & & & \\ 0 & & & \cdot & \cdot & \cdot & & & & \\ \vdots & & & & \cdot & \cdot & \cdot & & & \\ 0 & & & & & \cdot & \cdot & \cdot & & \\ a(N-n) & & & & & & \cdot & \cdot & \cdot & \\ \vdots & & & & & & & \cdot & \cdot & \cdot \\ a(N-1) & & & & & & & & \cdot & \cdot \end{pmatrix}.$$

Матрица A симметричная и положительно определенная.

Задача 4. Выписать расчетные формулы этого метода при $n = 1$. Показать, что решение этой системы методом Гаусса с исключением несодержательных операций требует $O(n^2 N)$ арифметических операций.

Задача 5. Выписать расчетные формулы метода квадратного корня в конкретном случае решения этой системы при $n = 1$. Подсчитать необходимое число арифметических операций.

Выше, когда при рассмотрении метода прогонки выписывались расчетные формулы и подсчитывалось число арифметических действий, был оставлен в стороне вопрос о возможности переполнения процессора ЭВМ, в частности, вследствие деления на нуль. Кроме того, отсутствие переполнения само по себе не гарантирует избавления от большого влияния вычислительной погрешности.

На модельном примере $p(x) \equiv p = \text{const}$ рассмотрим поведение прогоночных коэффициентов C_n при различном знаке p. Соотношения

$$C_{n+1} = \frac{1}{2 + ph^2 - C_n}, \tag{19}$$

которым удовлетворяют коэффициенты C_n метода прогонки, совпадают с итерационными формулами решения уравнения

$$C = g(C) = \frac{1}{2 + ph^2 - C}$$

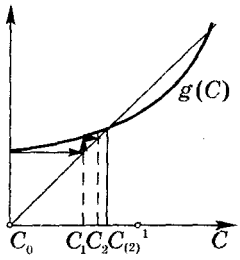


Рис. 9.3.1

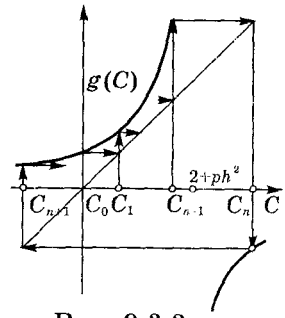


Рис. 9.3.2

при начальном приближении $C_0 = 0$. Уравнение $C = g(C)$ равносильно квадратному уравнению

$$C^2 - (2 + ph^2)C + 1 = 0; \tag{20}$$

его корни

$$C_{(1)},\, C_{(2)} = 1 + \frac{ph^2}{2} \pm \sqrt{ph^2 + \frac{p^2h^4}{4}}.$$

При $ph^2 > 0$ имеем $C_{(1)} > 1$; поэтому $0 < C_{(2)} = C_{(1)}^{-1} < 1$. При $p < 0$ и h малом подкоренное выражение отрицательно и уравнение (20) не имеет вещественных корней. Совместное расположение графиков $y = C$ и $y = g(C)$ и точек (C_n, C_n), (C_n, C_{n+1}) изображено на рис. 9.3.1, 9.3.2. Видно, что при $p \geqslant 0$ значения C_n лежат в пределах $0 < C_n \leqslant 1$ (рис. 9.3.1). При $p < 0$ не исключено (рис. 9.3.2), что при достаточно больших n некоторое значение C_n окажется близким к $2 + ph^2$; тогда следующее значение C_{n+1} будет очень большим и может произойти переполнение.

Рассмотрим вопрос о накоплении погрешности при вычислении коэффициентов C_n. Из (19) получаем, что

$$\frac{\partial C_{n+1}}{\partial C_n} = \frac{1}{(2 + p_n h^2 - C_n)^2} = (C_{n+1})^2.$$

Следовательно, возмущения в коэффициентах C_n связаны соотношением $\delta C_{n+1} \approx C_{n+1}^2 \delta C_n$ и при больших C_n происходит также большой рост погрешности. Отмеченные выше обстоятельства приводят к необходимости более детального изучения влияния вычислительной погрешности, по крайней мере при $p < 0$.

§ 4. Замыкания вычислительных алгоритмов

При предварительном анализе алгоритмов каждый раз удавалось многое понять, рассматривая случай решения модельной задачи. Другим эффективным приемом предварительного анализа является использование понятия *замыкания вычислительного алгоритма*, введенного Соболевым.

Чтобы не загромождать изложения, ограничимся рассмотрением существа проблемы; поэтому многие из проводимых в этом параграфе построений не всегда подробно обосновываются.

Пусть решается некоторая задача

$$Lu = f \tag{1}$$

и пусть

$$L^h u^h = f^h \tag{2}$$

— последовательность задач, зависящих от параметра h (например, шага сетки), решения которых u^h сходятся при $h \to 0$ к решению исходной задачи (1). Предположим, что алгоритм решения задачи (2) состоит в последовательном получении некоторых соотношений

$$L_m^h u_m^h = f_m^h, \quad m = 1, \ldots, M; \tag{3}$$

при этом $L_M^h = E$ — единичный оператор, $u_M^h = f_M^h = (L^h)^{-1} f^h = u^h$, т. е. на M-м шаге получается точное решение u^h. Пусть можно ввести параметр $z(m, h)$, монотонно зависящий от m, такой, что при $h \to 0$ и z фиксированном соотношение (3) переходит в пределе в соотношение

$$L_z y = f_z, \quad 0 \leqslant z \leqslant z_0; \tag{4}$$

здесь

$$z_0 = \lim_{h \to 0} z(M, h).$$

Соотношение (4) называется *замыканием вычислительного алгоритма* (3).

Мы не дали строгого определения замыкания вычислительного алгоритма. Существо дела станет понятнее после рассмотрения замыканий алгоритмов решения краевой задачи.

Определение понятия вычислительного алгоритма не исключает возможности $M = \infty$. В этом случае равенства

$$L_M^h = E, \quad f_M^h = (L^h)^{-1} f^h$$

понимаются в том смысле, что

$$L_m^h \to E, \quad f_m^h \to (L^h)^{-1} f^h \quad \text{при} \quad m \to \infty.$$

Случай $M = \infty$ соответствует итерационным методам решения задачи (1). Если операторы L_z равномерно ограничены по z, то говорят, что алгоритм (3) имеет *регулярное замыкание*. В противном случае говорят, что алгоритм имеет *нерегулярное замыкание*.

Если замыкание алгоритма регулярно, то есть основания предположить, что он устойчив к различным возмущениям, в частности к вычислительной погрешности. Поэтому исследование замыкания алгоритма является удобным способом получения предварительного суждения

о свойствах нового метода на первоначальной стадии изучения вопроса. Такая предварительная оценка свойств метода не всегда оказывается окончательной. Может случиться, что операторы L_z равномерно ограничены, но очень большой постоянной, что равносильно практической неограниченности. С другой стороны, возможно, что неограниченность операторов L_z вызывается неудачной нормировкой уравнений или неудачным выбором норм в пространствах функций u. Отсутствие равномерной ограниченности операторов L_z означает неограниченное возрастание $p(h) = \sup\limits_m \|L_m^h\|$ при $h \to 0$. Не исключена возможность, что величина $p(h)$ растет очень медленно и соответствующее возрастание влияния вычислительной погрешности при стремлении шага к нулю окажется допустимым. Несмотря на высказанные соображения, изучение замыканий вычислительных алгоритмов приносит большую пользу.

При прямом ходе прогонки получаются соотношения

$$y_n - C_n y_{n+1} = \varphi_n. \tag{5}$$

Попытаемся понять, во что переходят эти соотношения при $h \to 0$. Возьмем простейший случай $p \equiv 0$. Тогда $C_0 = 0$, $C_1 = 1/2$, $C_2 = 2/3, \ldots$ и, как нетрудно установить методом индукции, $C_n = 1 - 1/(n+1)$. Пусть фиксирована какая-то точка $x = nh$. Тогда левая часть (5), соответствующая этой точке, равна

$$y_n - \left(1 - \frac{1}{x/h + 1}\right) y_{n+1}.$$

При подстановке вместо величин y_n значений гладкой функции $y(nh)$ предел левой части равен нулю и нет смысла рассматривать подобное замыкание алгоритма. Чтобы получаемое замыкание было осмысленным, нужно подобрать такой множитель $\lambda(h)$, чтобы предел

$$\lim_{h\to 0} \lambda(h) \left(y(x) - \left(1 - \frac{1}{x/h + 1}\right) y(x + h)\right) \tag{6}$$

был конечен и отличен от нуля. Возьмем $\lambda(h) = h^{-1}$; тогда

$$\begin{aligned}
&\frac{1}{h}\left(y(x) - \left(1 - \frac{1}{x/h + 1}\right) y(x + h)\right) = \\
&\quad = \frac{1}{h}\left(y(x) - \left(1 - \frac{h}{x} + O(h^2)\right)\left(y(x) + y'(x)h + O(h^2)\right)\right) = \\
&\quad = \frac{y(x)}{x} - y'(x) + O(h).
\end{aligned}$$

Таким образом, (6) имеет ненулевой предел. Применим такую же нормировку и в общем случае. Поскольку в рассмотренном примере существовал предел

$$\lim_{h\to 0} \frac{C_n - 1}{h} \quad \text{при} \quad nh = x,$$

введем вместо C_n новую переменную $\alpha_n = (1 - C_n)/h$. Напомним, что прогоночные соотношения имеют вид

$$C_{n+1} = \frac{1}{2 + p_{n+1}h^2 - C_n}, \quad \varphi_{n+1} = C_{n+1}(\varphi_n - f_{n+1}h^2). \tag{7}$$

Умножим соотношение (5) на $-h^{-1}$ и положим $C_n = 1 - \alpha_n h$, $\varphi_n = -\beta_n h$. Тогда (5) перепишется в виде

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = \beta_n. \tag{8}$$

Подставив выражения C_n и φ_n через α_n и β_n в (7), получим равенства

$$1 - \alpha_{n+1}h = \frac{1}{1 + p_{n+1}h^2 + \alpha_n h},$$

$$\beta_{n+1}h = (1 - \alpha_{n+1}h)(\beta_n h + f_{n+1}h^2);$$

отсюда

$$\alpha_{n+1} = \frac{\alpha_n + p_{n+1}h}{1 + \alpha_n h + p_{n+1}h^2}, \tag{9}$$

$$\beta_{n+1} = \beta_n + (f_{n+1} - \alpha_{n+1}\beta_n)h - \alpha_{n+1}f_{n+1}h^2. \tag{10}$$

Соотношение (9) можно преобразовать к виду

$$\frac{\alpha_{n+1} - \alpha_n}{h} = p_{n+1} - \alpha_n^2 - h\frac{2p_{n+1}\alpha_n - \alpha_n^3 - (p_{n+1}^2 - p_{n+1}\alpha_n^2)h}{1 + \alpha_n h + p_{n+1}h^2}, \tag{11}$$

а соотношение (10) — к виду

$$\frac{\beta_{n+1} - \beta_n}{h} = f_{n+1} - \alpha_{n+1}\beta_n - f_{n+1}\alpha_{n+1}h. \tag{12}$$

Если предположить, что коэффициенты α_n и β_n при фиксированном $nh = x$ и $h \to 0$ стремятся к некоторым пределам $\alpha(x)$ и $\beta(x)$, то при подстановке в (8) вместо y_n значений $y(nh)$ в пределе получится дифференциальное уравнение

$$y' - \alpha(x)y = \beta(x).$$

Соотношения (11), (12) показывают, что величины α_n, β_n удовлетворяют уравнениям, напоминающим метод Эйлера численного интегрирования дифференциальных уравнений

$$\alpha' = p(x) - \alpha^2, \tag{13}$$

$$\beta' = f(x) - \alpha\beta. \tag{14}$$

Доказательство того факта, что значения α_n, β_n при $nh = x$ фиксированном, $h \to 0$, стремятся к значениям решений этих дифференциальных уравнений, затрудняется следующим обстоятельством. Согласно определению

$$\alpha_0 = (1 - C_0)/h = 1/h, \quad \beta_0 = \varphi_0/h = -a/h. \tag{15}$$

Таким образом, в окрестности точки $x = 0$ начальные условия для численного интегрирования по формулам (11), (12) имеют не совсем обычный характер и развитые нами методы оценки близости решений сеточных и дифференциальных задач в рассматриваемом случае неприменимы. Глядя на соотношения (15), можно было бы предположить, что функция $\alpha(x)$ является решением (13), ведущим себя вблизи нуля, как x^{-1}, а $\beta(x)$ — решением (14), ведущим себя вблизи нуля, как $-ax^{-1}$.

Обоснуем это предположение. Коэффициенты C_n, а следовательно, и α_n получаются из рекуррентных соотношений (7), не зависящих ни от правой части дифференциального уравнения, ни от граничного условия в точке X; начальное условие рекурсии $C_0 = 0$ также от них не зависит. Рассмотрим краевую задачу

$$y''(x) - p(x)y(x) = 0, \quad y(0) = 0, \quad y(X) = 1$$

и соответствующую сеточную задачу

$$l(y_j) = 0, \quad y_0 = 0, \quad y_N = 1.$$

Решения этих задач можно записать в виде

$$y(x) = W^1(x)/W^1(X), \quad y_n = W_n^1/W_N^1$$

(определение $W^1(x)$ и W_n^1 см. в § 2). Поскольку для этой краевой задачи $\varphi_0 = 0$ и все $f_n = 0$, то $\beta_n \equiv 0$ и соотношение (8) имеет вид

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = 0.$$

Подставляя сюда $y_n = W_n^1/W_N^1$, получим

$$\alpha_n = \left(\frac{W_{n+1}^1 - W_n^1}{h}\right)\left(W_{n+1}^1\right)^{-1}. \tag{16}$$

Пользуясь оценками близости

$$\left|W_n^1 - W^1(nh)\right| \leqslant Mh^2 \quad \text{при} \quad 0 \leqslant nh \leqslant X, \tag{17}$$

можно получить оценку

$$\alpha_n = \frac{(W^1)'\big|_{(n+1)h} + O(h)}{W^1((n+1)h) + O(h^2)}. \tag{18}$$

Функция

$$\alpha(x) = (W^1)'|_X/W^1(x)$$

удовлетворяет дифференциальному уравнению (13), поскольку

$$\left(\frac{(W^1)'}{W^1}\right)' = \frac{(W^1)''}{W^1} - \left(\frac{(W^1)'}{W^1}\right)^2 = p(x) - \left(\frac{(W^1)'}{W^1}\right)^2.$$

Согласно определению $W^1(x)$ при малых x имеем $W^1(x) \sim x$, $(W^1)'|_x \sim 1$ и, таким образом, $\alpha(x) = (W^1)'|_x/W(x) \sim x^{-1}$, как и предполагалось.

Функция $\alpha(x)$ обращается в бесконечность в точках, где $W^1(x) = 0$, в частности при $x = 0$. В окрестности каждой такой точки y имеем

$$(W^1)'|_x \sim (W^1)'|_y \neq 0, \quad W^1(x) \sim ((W^1)'|_y)(x - y)$$

и, таким образом,

$$\alpha(x) \sim (x - y)^{-1}. \tag{19}$$

Для точек этой окрестности можно написать цепочки соотношений

$$\frac{W^1_{n+1} - W^1_n}{h} = (W^1)'|_{(n+1)h}(1 + O(h)),$$

$$W^1_{n+1} = W^1\Big((n+1)h\Big) + O(h^2) = W^1\Big((n+1)h\Big)\left(1 + O\left(\frac{h^2}{W^1((n+1)h)}\right)\right) =$$

$$= W^1\Big((n+1)h\Big)\left(1 + O\left(\frac{h^2}{(n+1)h - y}\right)\right)$$

и, таким образом,

$$\alpha_n = \alpha((n+1)h)\Big(1 + O(h)\Big)\left(1 + O\left(\frac{h^2}{(n+1)h - y}\right)\right).$$

Отсюда видно, что относительная погрешность равенства $\alpha_n \approx \alpha((n+1)h)$ является величиной порядка $O(h)$ вплоть до окрестности порядка h точек, где $\alpha(x) = \infty$. Аналогичным образом исследуется поведение функции $\beta(x)$.

Выпишем теперь соотношения (3) и соответствующие замыкания алгоритма. Возьмем $M = 2N - 1$ и при $m = 1, \ldots, N$ за (3) примем совокупность соотношений

$$\frac{y_n - C_n y_{n+1}}{h} = \frac{\varphi_n}{n}, \quad 0 \leqslant n < m,$$

$$\frac{y_{n-1} - 2y_n + y_{n+1}}{h^2} - p_n y_n = f_n, \quad m \leqslant n < N,$$

$$y_N = b,$$

а при $N < m \leqslant M$ — совокупность соотношений

$$\frac{y_n - C_n y_{n+1}}{h} = \frac{\varphi_n}{h}, \quad 0 \leqslant n \leqslant M - m,$$

$$y_n = \psi_n, \quad M - m < n \leqslant N,$$

где ψ_n — точное решение сеточной задачи. Положим $z = mh$. Тогда замыкание алгоритма (4) будет выглядеть следующим образом:

$$L_z y = f_z(x),$$

где для $0 \leqslant z < 1$

$$L_z y = \begin{cases} y' - \alpha(x)y & \text{при} \quad 0 < x < z, \\ y'' - p(x)y & \text{при} \quad z \leqslant x < 1, \\ y & \text{при} \quad x = 1, \end{cases}$$

$$f_z(x) = \begin{cases} \beta(x) & \text{при} \quad 0 < x < z, \\ f(x) & \text{при} \quad z \leqslant x < 1, \\ b & \text{при} \quad x = 1 \end{cases}$$

и для $1 \leqslant z \leqslant 2$

$$L_z y = \begin{cases} y' - \alpha(x)y & \text{при} \quad 0 < x < 2 - z, \\ y & \text{при} \quad 2 - z \leqslant x \leqslant 1, \end{cases}$$

$$f_z(x) = \begin{cases} \beta(x) & \text{при} \quad 0 < x < 2 - z, \\ \psi(x) & \text{при} \quad 2 - z \leqslant x \leqslant 1; \end{cases}$$

здесь $\psi(x)$ — точное решение дифференциальной задачи (равенство $L_z y|_{x=0} = a$ опущено).

Рассмотрим примеры поведения функции $\alpha(x)$ и $\beta(x)$. Пусть $p(x) > 0$; тогда, согласно уравнению $\alpha' = p(x) - \alpha^2$,

$$\alpha' < 0 \quad \text{при} \quad \alpha > p_1 = \sqrt{\max_{[0,\,X]} p(x)}$$

и

$$\alpha' > 0 \quad \text{при} \quad \alpha < p_2 = \sqrt{\min_{[0,\,X]} p(x)}.$$

Соответствующее поле интегральных кривых изображено на рис. 9.4.1. Поэтому решение $\alpha(x)$, удовлетворяющее условию $\alpha(x) \sim x^{-1}$ при $x \to 0$, монотонно убывает по крайней мере до тех пор, пока не попадет в область $p_2 < \alpha < p_1$, где оно и останется. Уравнение (14) линейно относительно $\beta(x)$. Так как коэффициент $\alpha(x)$ конечен при $x \neq 0$, то решение этого уравнения остается ограниченным при всех $x \neq 0$. Таким образом, были бы все основания признать замыкание алгоритма решения рассматриваемой задачи регулярным, если бы не неограниченность коэффициентов $\alpha(x)$ и $\beta(x)$ при $x \to 0$. В рассматриваемом случае значения α_0 и β_0 в точке $n = 0$ являются величинами порядка h^{-1}; далее с ростом n они должны иметь тенденцию к убыванию из-за аналогичной тенденции решений дифференциальных уравнений. Мы уже свыклись с тем, что во многих случаях влияние вычислительной

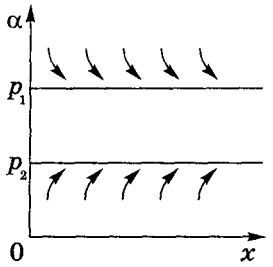


Рис. 9.4.1

погрешности порядка $2^{-t}h^{-p}$ является неизбежным и допустимым и возможно, что такая нерегулярность алгоритма не приведет к недопустимо большому влиянию вычислительной погрешности.

Обратимся к случаю, когда не всюду $p(x) \geqslant 0$. Тогда может оказаться, что в некоторой внутренней точке отрезка $W^1(y) = 0$ и, следовательно, $\alpha(y) = \infty$. В этом случае замыкание алгоритма вычислений следует признать нерегулярным; однако и здесь не следует спешить окончательно отказываться от применения метода прогонки. Если оказалось, что во всех узлах сетки $W^1(nh)h^{-2} \gg 1$, то, согласно (16)–(18), имеем

$$\max_n |\alpha_n| = O(h^{-2}). \tag{20}$$

Такого рода нерегулярность алгоритма также не всегда следует считать катастрофической. Функция $W^1(x)$ — гладкая с производной, отличной от нуля в точках y, где $W^1(y) = 0$. Поэтому часто значение $\min\limits_{n>0} W^1(nh)$ будет оказываться величиной порядка h и соотношение (20) будет выполняться. Таким образом, следует ожидать, что и в случае $p(x) \leqslant 0$ вычисления по методу прогонки в большинстве случаев будут устойчивыми к различного рода возмущениям.

Оказалось, что замыкание алгоритма прогонки, по существу, регулярно, если $W^1(x) \neq 0$ при $x \in (0, X]$. Условие $W^1(x_0) \neq 0$ равносильно условию, что краевая задача

$$y(0) = a, \quad y(x_0) = b, \quad y'' - p(x)y = f(x) \tag{21}$$

разрешима на $[0, x_0]$ при любых a, b и $f(x)$. Таким образом, условие регулярности замыкания алгоритма прогонки можно сформулировать в следующем виде.

Замыкание метода прогонки регулярно, если для любого отрезка $[0, x_0]$ *при* $0 < x_0 \leqslant X$ *краевая задача* (21) *разрешима.*

Получим этот вывод еще одним путем. Обратимся к формулам обратного хода прогонки. Коэффициенты C_j и φ_j при $j < n$ зависят только от a и значений p_j и f_j при $j < n$, т. е. в точках $jh \in [0, nh]$. Следовательно, после отыскания значения y_n вычисления происходят так же, как в случае краевой задачи для уравнения $y'' - p(x)y = f(x)$ на отрезке $[0, nh]$ при заданных $y(0) = a$ и $y(nh) = y_n$. Если эта краевая задача неразрешима при любых правых частях $f(x)$, то возникают сомнения в хороших свойствах соответствующей сеточной задачи. Поэтому в случае, когда при каком-то $x_0 \in (0, X]$ краевая задача (21) разрешима не при любых правых частях, необходим более детальный анализ метода.

Отметим, что регулярность замыкания еще не обеспечивает малости суммарной вычислительной погрешности в силу следующих обстоятельств. Суммарная погрешность определяется погрешностями в ходе вычислений и множителями пропорциональности, с которыми эти погрешности входят в сумарную погрешность. Регулярность замыкания, как правило, обеспечивает лишь малость погрешностей округления в ходе

вычислений. Чтобы множители пропорциональности не были большими, нужна слабая чувствительность решения уравнений замыкания к возмущениям коэффициентов. Однако и этого, вообще говоря, недостаточно. На примере уравнения $y' = My$ при $M < 0$ мы видели, что зачастую решение дифференциальной задачи слабо чувствительно к таким возмущениям, в то время как решение сеточной задачи сильно чувствительно.

§ 5. Обсуждение постановок краевых задач для линейных систем первого порядка

Рассмотрим краевую задачу

$$\mathbf{y}' - A(x)\mathbf{y} = \mathbf{f}(x), \quad B\mathbf{y}(0) = \mathbf{b}, \quad D\mathbf{y}(X) = \mathbf{d}; \tag{1}$$

здесь $\mathbf{y}$, $\mathbf{f}$, $\mathbf{b}$, $\mathbf{d}$ — векторы размерностей соответственно l, l, $l-r$, r, а A, B, D — матрицы размерностей $l\times l$, $(l-r)\times l$, $r\times l$. Всюду в дальнейшем предполагается, что ранг матрицы B равен $l-r$, а ранг матрицы D равен r.

Прежде чем отыскивать практически пригодные методы решения задачи (1), обсудим вопрос о чувствительности решений краевых задач к разного рода возмущениям. В качестве модели возьмем краевую задачу

$$\mathbf{y}' - A\mathbf{y} = 0, \quad A = \text{const}, \quad B\mathbf{y}(0) = \mathbf{b}, \quad D\mathbf{y}(X) = \mathbf{d}. \tag{2}$$

Рассмотрим случай, когда все собственные значения матрицы A различны; в общем случае ход рассуждений изменяется несущественно. Пусть $\lambda_j = \alpha_j + \mathbf{i}\beta_j$ — собственные значения матрицы A, упорядоченные в порядке возрастания α_j, а $\mathbf{e}_j$ — соответствующие собственные векторы, причем $\|\mathbf{e}_j\| = 1$. Общее решение системы $\mathbf{y}' - A\mathbf{y} = 0$ запишется в виде

$$\mathbf{y}(x) = \sum_{j=1}^{l} c_j \mathbf{e}_j \exp\{\lambda_j x\}. \tag{3}$$

Собственные значения λ_j разделим на три группы, присваивая собственным значениям каждой группы соответствующий верхний индекс. Собственные значения λ_j, для которых величина $\exp\{|\alpha_j|X\}$ очень большая, обозначаем как λ_j^+, если $\alpha_j > 0$, и как λ_j^-, если $\alpha_j < 0$. Остальные собственные значения, т. е. те, для которых величина $\exp\{|\alpha_j|X\}$ не очень велика, обозначаем как λ_j^0. Соответственно снабдим верхними индексами собственные векторы $\mathbf{e}_j$; суммирование по индексам j, соответствующим этим группам, будем обозначать $\sum^+$, $\sum^0$, $\sum^-$. Пусть l^+, l^0, l^- — число собственных значений в соответствующих группах. Форма записи решения (3) ставит в неодинаковое положение концы отрезка интегрирования: все функции $\exp\{\lambda_j x\}$ в точке $x = 0$ имеют порядок 1, в то время как

в точке $x = X$ одни из них очень большие, другие очень малые. Удобнее другая форма записи общего решения:

$$\mathbf{y}(x) = \sum^{-} c_j^- \mathbf{e}_j^- \exp\{\lambda_j^- x\} + \sum^{0} c_j^0 \mathbf{e}_j^0 \exp\{\lambda_j^0 x\} + \sum^{+} c_j^+ \mathbf{e}_j^+ \exp\{\lambda_j^+ (x - X)\}.$$

Выпишем систему уравнений $B\mathbf{y}(0) = \mathbf{b}$, $D\mathbf{y}(X) = \mathbf{d}$, из которой следует определить постоянные c_j, соответствующие отыскиваемому решению:

$$B\mathbf{y}(0) = \sum^{-} c_j^- B\mathbf{e}_j^- + \sum^{0} c_j^0 B\mathbf{e}_j^0 + \sum^{+} c_j^+ B\mathbf{e}_j^+ \exp\{-\lambda_j^+ X\} = \mathbf{b},$$

$$D\mathbf{y}(X) = \sum^{-} c_j^- D\mathbf{e}_j^- \exp\{\lambda_j^- X\} + \sum^{0} c_j^0 D\mathbf{e}_j^0 \exp\{\lambda_j^0 X\} + \sum^{+} c_j^+ D\mathbf{e}_j^+ = \mathbf{d}. \tag{4}$$

Систему уравнений (4) можно записать в форме

$$G\mathbf{c} = \mathbf{g};$$

здесь

$$\mathbf{c} = (c_1^-, \ldots, c_l^+)^T, \quad \mathbf{g} = (b_1, \ldots, b_{l-r}, d_1, \ldots, d_r)^T,$$

матрица G имеет клеточный вид:

$$G = \begin{pmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{pmatrix} \begin{matrix} l-r \\ r \end{matrix}, \qquad \begin{matrix} l^- & l^0 & l^+ \end{matrix}$$

в котором клетки записываются следующим образом:

$$G_1^- = [(B\mathbf{e}_j^-)_q], \quad 1 \leqslant j \leqslant l^-, \quad 1 \leqslant q \leqslant l - r,$$

$$G_1^0 = [(B\mathbf{e}_j^0)_q], \quad l^- < j \leqslant l^- + l^0, \quad 1 \leqslant q \leqslant l - r,$$

$$G_1^+ = [(B\mathbf{e}_j^+)_q \exp\{-\lambda_j^+ X\}], \quad l^- + l^0 < j \leqslant l, \quad 1 \leqslant q \leqslant l - r,$$

$$G_2^- = [(D\mathbf{e}_j^-)_q \exp\{\lambda_j^- X\}], \quad 1 \leqslant j \leqslant l^-, \quad 1 \leqslant q \leqslant r,$$

$$G_2^0 = [(D\mathbf{e}_j^0)_q \exp\{\lambda_j^0 X\}], \quad l^- < j \leqslant l^- + l^+, \quad 1 \leqslant q \leqslant r,$$

$$G_2^+ = [(D\mathbf{e}_j^+)_q], \quad l^- + l^0 < j \leqslant l, \quad 1 \leqslant q \leqslant r.$$

Предположим, что $\Delta = \det G \neq 0$, и представим решение системы (4) в виде $\mathbf{c} = G^{-1}\mathbf{g}$. Согласно известным формулам элементы обратной матрицы имеют вид

$$G^{-1} = \left(\frac{(-1)^{i+j} G_{ij}}{\Delta} \right), \tag{5}$$

где G_{ij} — миноры матрицы G. При сформулированных ранее предположениях величины $\exp\{-\lambda_j^+ X\}$, $\exp\{\lambda_j^- X\}$ ничтожно малы, поэтому ничтожно малы элементы матриц G_1^+ и G_2^- и определитель

$$\Delta = \det \begin{pmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{pmatrix}$$

близок к определителю

$$\Delta_0 = \det G^0, \quad G^0 = \begin{pmatrix} G_1^- & G_1^0 & 0 \\ 0 & G_2^0 & G_2^+ \end{pmatrix}.$$

Рассмотрим отдельно случаи:

а) определитель Δ_0 не мал;

б) определитель Δ_0 мал, в частности равен нулю.

Поскольку среди элементов матрицы G нет больших, то, вследствие формулы (5), в случае а) элементы матрицы G^{-1} обычно не очень большие.

Предположим, что правые части граничных условий $\mathbf{b}$ и $\mathbf{d}$ содержат некоторые погрешности $\delta\mathbf{b}$ и $\delta\mathbf{d}$. Пусть $\delta\mathbf{g} = (\delta\mathbf{b}, \delta\mathbf{d})^T$, тогда погрешность вектора $\mathbf{c}$ равна $G^{-1}\delta\mathbf{g}$. Если элементы матрицы G^{-1} не очень велики, то влияние погрешности $\delta\mathbf{g}$ на коэффициенты c_j будет не очень большим. Решение задачи является линейной комбинацией с коэффициентами c_j слагаемых

$$\mathbf{e}_j^- \exp\{\lambda_j^- x\}, \quad \mathbf{e}_j^0 \exp\{\lambda_j^0 x\}, \quad \mathbf{e}_j^+ \exp\{\lambda_j^+ (x - X)\}.$$

Поэтому погрешность приближенного решения, являющаяся следствием погрешностей $\delta\mathbf{b}$ и $\delta\mathbf{d}$, также будет приемлемой.

Заметим, что наши высказывания носят довольно неопределенный характер: «малый», «очень малый», «небольшой», «очень большой»; при анализе конкретной задачи исследователь должен сам решать, насколько приемлем для него тот или иной порядок рассматриваемых величин. В частности, если решается система «большого» порядка, то при «умеренных» значениях коэффициентов системы и «не очень малом» определителе Δ_0 возможно, что миноры, состоящие из сумм произведений большого числа элементов, окажутся «недопустимо большими».

Если определитель Δ_0 очень мал или равен нулю, то, вследствие равенства

$$\det G \det G^{-1} = 1,$$

среди элементов матрицы G^{-1} встретятся большие. Тогда малые возмущения правых частей граничных условий могут приводить к большим возмущениям коэффициентов c_j, а следовательно, и решения задачи.

Зная элементы матрицы G^{-1} и собственные значения матрицы A, из полученных выше соотношений можно получить довольно точную информацию о возмущении решения дифференциальной задачи. Однако получение этой информации само по себе требует большого объема вычислений; перенос этих построений на случай переменной матрицы $A(x)$ потребует еще большего объема вычислений. Попытаемся поэтому получить критерии устойчивости решения к возмущениям $\delta\mathbf{b}$ и $\delta\mathbf{d}$ качественного характера, требующие меньшей информации о задаче, хотя, может быть, и несколько менее надежные. Таким критерием могут служить соотношения между числами l^-, l^0, l^+, $l-r$ и r. Среди элементов первых $l-r$ строк матрицы G^0 ненулевые элементы могут находиться в первых $l^- + l^0$ столбцах, соответствующих матрицам G_1^-, G_1^0. Если $l^+ > r$, то $l^- + l^0 < l-r$, и тогда все миноры порядка $(l-r)\times(l-r)$, лежащие в первых $l-r$ строках, обращаются в нуль. Раскрывая определитель Δ_0 по первым $l-r$ строкам, получаем $\Delta_0 = 0$. Точно так же, если $l^- > l-r$, то все миноры порядка $r\times r$, лежащие в последних r строках, обращаются в нуль, поэтому $\Delta_0 = 0$. Если $l^+ \leqslant r$, а $l^- \leqslant l-r$, то определитель Δ_0 окажется линейной комбинацией произведений элементов матриц B и D и координат собственных векторов $\mathbf{e}_j$, причем коэффициентами при этих произведениях будут произведения чисел $\exp\{\lambda_j^0 X\}$, не очень больших и не очень маленьких по модулю. Можно принять гипотезу, что этот определитель оказывается малым числом довольно редко. Тогда решение задачи (2) будет мало чувствительно к возмущениям $\delta\mathbf{b}$ и $\delta\mathbf{d}$ правых частей граничных условий.

Мы можем сформулировать полученные выводы в качестве следующего предложения. Если $l^+ > r$ или $l^- > l-r$, то *решение дифференциальной задачи сильно чувствительно к возмущениям правых частей граничных условий: если $l^+ \leqslant r$, а $l^- \leqslant l-r$, то, как правило, решение задачи* (2) *будет мало чувствительно к изменениям правых частей граничных условий.*

Первую часть этого утверждения можно переформулировать еще в такой форме: *для малой чувствительности задачи к возмущениям граничных условий необходимо, чтобы число независимых частных решений $\mathbf{e}_j \exp\{\lambda_j x\}$, сильно растущих на $[0, X]$ с ростом x, не превосходило числа граничных условий на правом конце, а число частных решений $\mathbf{e}_j \exp\{\lambda_j x\}$, сильно убывающих на $[0, X]$ с ростом x, не превосходило числа граничных условий на левом конце.*

Эта формулировка при определенных уточнениях может быть перенесена и на случай задачи (1) с переменной матрицей $A(x)$. Строгая переформулировка этого утверждения будет довольно громоздкой; однако если элементы матрицы $A(x)$ относительно плавно меняются на $[0, X]$, то при первоначальном исследовании устойчивости задачи к возмущениям граничных условий зачастую можно ограничиться подсчетом числа собственных значений матрицы $A(x)$ с большим положительным и большим отрицательным значениями величины $X \operatorname{Re}\lambda_j(x)$.

Краевую задачу называют *хорошо обусловленной* (*хорошо поставленной*), если малые возмущения коэффициентов и правых частей уравнения и граничных условий приводят к столь же малым по порядку изменениям решения задачи. Более аккуратное определение хорошей обусловленности можно дать следующим образом. Наряду с краевой задачей (1) рассмотрим краевые задачи

$$\tilde{\mathbf{y}}' - (A(x) + \delta A(x))\tilde{\mathbf{y}} = \mathbf{f}(x) + \delta\mathbf{f}(x),$$
$$(B + \delta B)\tilde{\mathbf{y}}(0) = \mathbf{b} + \delta\mathbf{b}, \quad (D + \delta D)\tilde{\mathbf{y}}(X) = \mathbf{d} + \delta\mathbf{d}$$

с не очень большой мерой возмущения

$$\varepsilon = \max_{0 \leqslant x \leqslant X} (\|\delta A(x)\| + \|\delta\mathbf{f}(x)\|) + \|\delta B\| + \|\delta D\| + \|\delta\mathbf{b}\| + \|\delta\mathbf{d}\|.$$

Если для всех решений таких краевых задач выполняются неравенства

$$\max_{0 \leqslant x \leqslant X} \|\tilde{\mathbf{y}}(x) - \mathbf{y}(x)\| \leqslant M\varepsilon \tag{6}$$

с не очень большим значением постоянной M, то исходную задачу называют *хорошо обусловленной*, в противном случае задачу называют *плохо обусловленной*. Минимальное значение $M(\varepsilon_0)$, при котором неравенство (6) выполняется при всех $\varepsilon \leqslant \varepsilon_0$ ($\varepsilon_0 > 0$ фиксировано), иногда называют *мерой обусловленности* данной задачи (относительно возмущений с нормой, не большей ε_0). Обусловленность задачи характеризует устойчивость решения к возмущениям исходных данных, например к неточности задания коэффициентов уравнения. Поскольку погрешности от округления при вычислениях эквивалентны возмущениям коэффициентов исходного уравнения, то мера обусловленности характеризует и устойчивость численного решения к возможным округлениям при численном решении. Если известна ориентировочная оценка ε возмущения коэффициентов задачи и погрешность порядка $M(\varepsilon)\varepsilon$ допустима, то имеет смысл непосредственное численное решение задачи.

Рассмотрим в качестве примера задачу Коши для системы $y_1' = y_2$, $y_2' = y_1$ при $y_1(0) = 1$, $y_2(0) = 1$ на отрезке $[0, 30]$. Собственные значения матрицы

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

равны ± 1. Величина $\exp(-30)$ очень малая, а $\exp(30)$ — очень большая, $l^- = 1$, $l^0 = 0$, $l^+ = 1$, $r = 0$ и $l^+ = 1 > r = 0$. Поэтому малые возмущения граничного условия должны приводить к очень большим изменениям решения. В данном случае проводившиеся нами в этом параграфе построения не имеют особого смысла, поскольку и без них ясно, что, вследствие сильного роста решений исходного линейного уравнения, погрешность решения растет очень быстро.

Пусть для той же системы рассматривается задача с краевыми условиями $b_1 y_1(0) + b_2 y_2(0) = b$, $d_1 y_1(30) + d_2 y_2(30) = d$; собственные векторы, соответствующие собственным значениям $\lambda_1^- = -1$ и $\lambda_2^+ = 1$, равны соответственно

$(1, -1)^T$ и $(1, 1)^T$, $l^+ = r$, $l^- = l - r$. Вообще говоря, следует ожидать, что задача будет устойчива к возмущениям δb и δd. Решение отыскивается в виде

$$c_1^- \begin{pmatrix} 1 \\ -1 \end{pmatrix} \exp\{-x\} + c_2^+ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \exp\{x - 30\}.$$

Уравнения (4) имеют вид

$$\begin{aligned} (b_1 - b_2)c_1^- + (b_1 + b_2)c_2^+ \exp\{-30\} &= b, \\ (d_2 - d_2)c_1^- \exp\{-30\} + (d_1 + d_2)c_2^+ &= d. \end{aligned} \tag{7}$$

Отсюда

$$c_1^- = \frac{b(d_1 + d_2) - d(b_1 + b_2)\exp\{-30\}}{\Delta},$$

$$c_2^+ = \frac{d(b_1 - b_2) - b(d_1 - d_2)\exp\{-30\}}{\Delta},$$

где

$$\Delta = (b_1 - b_2)(d_1 + d_2) - (d_1 - d_2)(b_1 + b_2)\exp\{-60\}$$

близко к $\Delta_0 = (b_1 - b_2)(d_1 + d_2)$. Если коэффициенты b_i и d_i небольшие, а Δ_0 не мало, то коэффициенты c_1^- и c_2^- мало изменяются при малых изменениях δb и δd. Если $\Delta = 0$, то система (7) имеет ненулевое решение c_1^-, c_2^+ при $b = d = 0$. В этом случае говорят, что «задача лежит на спектре», т. е. имеется в виду, что однородная задача имеет ненулевое решение и бессмысленно говорить об устойчивости решения к возмущениям граничных условий. При $\Delta_0 = 0$ задача плохо обусловлена, поскольку $\Delta \approx \Delta_0$.

Задача 1. Доказать, что решение хорошо обусловленной краевой задачи единственно.

§ 6. Алгоритмы решения краевых задач для систем уравнений первого порядка

В дальнейшем, в § 6–8, предполагается, что рассматриваемая краевая задача хорошо обусловлена.

Простейшим по форме методом решения краевой задачи (5.1) является *метод стрельбы.* Рассмотрим систему уравнений

$$B\mathbf{y}(0) = \mathbf{b}. \tag{1}$$

Поскольку по предположению ранг матрицы B равен $l - r$, то общее решение системы (1) записывается в виде

$$\mathbf{y}_0 + \sum_{j=1}^{r} c_j \mathbf{y}_j;$$

здесь $\mathbf{y}_0$ — произвольное решение неоднородной системы $B\mathbf{y} = \mathbf{b}$, а $\mathbf{y}_1, \ldots, \mathbf{y}_r$ — произвольная система r линейно независимых решений системы $B\mathbf{y} = \mathbf{0}$. Пусть $\mathbf{y}_1, \ldots, \mathbf{y}_r$, $\mathbf{y}_0$ — какой-то набор таких векторов.

Численным интегрированием найдем частное решение неоднородной системы

$$\mathbf{y}_0' = A(x)\mathbf{y}_0 + \mathbf{f}(x) \tag{2}$$

при начальном условии $\mathbf{y}_0(0) = \mathbf{y}_0$ и решения однородной системы

$$\mathbf{y}_j' = A(x)\mathbf{y}_j, \quad j = 1, \ldots, r, \tag{3}$$

при начальных условиях $\mathbf{y}_j(0) = \mathbf{y}_j$. Пусть $\mathbf{y}(x)$ удовлетворяет (2) и левому граничному условию (1). Вектор $\mathbf{y}(0)$ является решением (1), и поэтому его можно записать в виде

$$\mathbf{y}(0) = \mathbf{y}_0 + \sum_{j=1}^{r} c_j \mathbf{y}_j;$$

вектор-функция

$$\mathbf{z}(x) = \mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x)$$

удовлетворяет уравнению (2) и совпадает с $\mathbf{y}(x)$ при $x = 0$. Следовательно,

$$\mathbf{y}(x) = \mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x). \tag{4}$$

Всякая функция вида (4) удовлетворяет соотношениям (1) и (2). Таким образом, многообразие всех решений (2), удовлетворяющих левому граничному условию (1), задается равенством (4). Чтобы найти искомое решение, надо выделить из этого многообразия решение, удовлетворяющее правому граничному условию. Определим коэффициенты c_j из системы r уравнений с r неизвестными

$$D\left(\mathbf{y}_0(X) + \sum_{j=1}^{r} c_j \mathbf{y}_j(X)\right) = \mathbf{d}. \tag{5}$$

В предположении однозначной разрешимости задачи (5.1) определитель этой системы отличен от нуля. Действительно, предположив противное, мы получили бы, что однородная краевая задача ($\mathbf{f} \equiv \mathbf{0}$, $\mathbf{b} = \mathbf{0}$, $\mathbf{d} = \mathbf{0}$) имеет ненулевое решение

$$\mathbf{y}(x) = \sum_{j=1}^{r} c_j \mathbf{y}_j(x).$$

Если $c_1, \ldots, c_r$ — решение системы (5), то вектор-функция

$$\mathbf{y}(x) = \mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x)$$

удовлетворяет уравнению и всем граничным условиям и, следовательно, является решением искомой задачи.

При необходимости экономить память ЭВМ следует найти

$$\mathbf{y}(0) = \mathbf{y}_0(0) + \sum_{j=1}^{r} c_j \mathbf{y}_j(0)$$

или

$$\mathbf{y}(X) = \mathbf{y}_0(X) + \sum_{j=1}^{r} c_j \mathbf{y}_j(X)$$

и затем решить численно задачу Коши вперед или назад.

Если среди решений однородной системы $\mathbf{y}' = A(x)\mathbf{y}$ есть быстро растущие с ростом X, то столь же быстро может возрастать вычислительная погрешность в решениях $\mathbf{y}_0(x), \ldots, \mathbf{y}_r(x)$. В этом случае алгоритм метода стрельбы окажется непригодным для практического использования. По-другому практическую непригодность метода стрельбы в этом случае можно объяснить следующим образом. Пусть метод стрельбы применяется при $A = \text{const}$. Обозначим через $\mathbf{e}_j$ и λ_j собственные векторы и соответствующие собственные значения матрицы A, причем пусть $\operatorname{Re}\lambda_1 \leqslant \cdots \leqslant \operatorname{Re}\lambda_{l-1} < \operatorname{Re}\lambda_l$. Предположим, что

$$\mathbf{y}_j = \sum_{k=1}^{l} \alpha_{jk}\mathbf{e}_k;$$

тогда

$$\mathbf{y}_j(X) = \sum_{k=1}^{l} \alpha_{jk}\mathbf{e}_k \exp\{\lambda_k X\}. \tag{6}$$

Если $\alpha_{jk} \neq 0$ и $\exp\{\operatorname{Re}\lambda_l X\} \gg \exp\{\operatorname{Re}\lambda_{l-1} X\}$, то

$$\mathbf{y}_j(X) \sim \alpha_{jl}\mathbf{e}_l \exp\{\lambda_l X\}.$$

Таким образом, все столбцы матрицы системы (5) оказываются приблизительно пропорциональными вектору $D\mathbf{e}_l$, поэтому решение этой системы $c_1, \ldots, c_r$ будет найдено с большой вычислительной погрешностью.

Рассматриваемую задачу можно трактовать как задачу выделения из многообразия

$$\mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x) \tag{7}$$

вектора, удовлетворяющего правому граничному условию

$$D\mathbf{y}(X) = \mathbf{d}.$$

При каждом фиксированном x множество концов векторов вида (4) образует r-мерную плоскость в l-мерном пространстве; плоскость задается концом вектора $\mathbf{y}_0(x)$, лежащим в этой плоскости, и векторами $\mathbf{y}_1(x), \ldots, \mathbf{y}_r(x)$, лежащими в ней. Если бы эти векторы задавались точно, то было бы несущественно, какими векторами задавать плоскость.

Однако вследствие погрешностей в значениях этих векторов эта плоскость будет несколько смещена и повернута. Предположим, что исходная краевая задача хорошо поставлена и соответственно норма вектора $\mathbf{y}(x)$ в каждой точке $x \in [0, X]$ невелика. В рассматриваемом случае при больших значениях $\operatorname{Re}\lambda_l X$ вектор $\mathbf{y}_0(x)$ имеет большую норму, а векторы $\mathbf{y}_j(x)$ примерно пропорциональны. Небольшие возмущения $\mathbf{y}_0(x)$ и векторов $\mathbf{y}_j(x)$ приведут к существенному изменению положения плоскости в области относительно небольших значений $\|\mathbf{y}\|$, т. е. там, где находятся значения точного решения. Поэтому при таком способе задания многообразия (4) существенно теряется информация о решении.

Чтобы положение плоскости в области небольших значений норм векторов $\|\mathbf{y}\|$, где находится решение, было устойчиво к погрешностям такого способа ее задания, представляется целесообразным задавать плоскость точкой, являющейся проекцией на нее начала координат, или точкой, близкой к ней, и некоторым множеством векторов, лежащих в этой плоскости и образующих ортогональный репер или близкий к таковому. Существо наиболее распространенных методов прогонки решения задачи (5.1) состоит в непрерывном или дискретном (в отдельных точках отрезка) переходе к заданию многообразия (4) при помощи точки проекции начала координат на эту плоскость и ортогонального репера, лежащего в этой плоскости.

В одном из вариантов метода прогонки (метод Годунова) поступают следующим образом. Отрезок интегрирования разбивается на части точками $0 = x_0 < x_1 < \cdots < x_m = X$. Пусть на отрезке $[x_s, x_{s+1}]$ решение задачи отыскивалось в виде линейной комбинации

$$\mathbf{y}(x) = \mathbf{g}_0^s(x) + \sum_{j=1}^{r} c_j \mathbf{g}_j^s(x);$$

здесь $\mathbf{g}_0^s(x)$ — решение неоднородной системы (2); $\mathbf{g}_j^s(x)$, $j = 1, \ldots, r$, — решения однородной системы (3). Путем последовательной ортогонализации и нормировки векторов $\mathbf{g}_j^s(x_{s+1})$, $j = 1, \ldots, r$, получают систему ортонормированных векторов $\mathbf{g}_j^{s+1}(x_{s+1})$, $j = 1, \ldots, r$. Полагают

$$\mathbf{g}_0^{s+1}(x_{s+1}) = \mathbf{g}_0^s(x_{s+1}) - \sum_{j=1}^{r} \Big(\mathbf{g}_0^s(x_{s+1}),\, \mathbf{g}_j^{s+1}(x_{s+1})\Big)\mathbf{g}_j^{s+1}(x_{s+1}),$$

т. е. вычитают из вектора $\mathbf{g}_0^s(x_{s+1})$ его проекцию на плоскость, натянутую на векторы $\mathbf{g}_j^{s+1}(x_{s+1})$, $j = 1, \ldots, r$. Далее опять на отрезке $[x_{s+1}, x_{s+2}]$ ищут $\mathbf{g}_0^{s+1}(x)$ и $\mathbf{g}_j^{s+1}(x)$, $j = 1, \ldots, r$, как решения систем (2) и (3) соответственно. После получения такого представления решения в точке X находят решение задачи на отрезке $[x_{m-1}, x_m]$, воспользовавшись гранич-

ным условием в точке $X = x_m$. Далее, пользуясь формулами перехода между совокупностями векторов

$$\Big(\mathbf{g}_0^s(x_{s+1}), \dots, \mathbf{g}_r^s(x_{s+1})\Big) \quad \text{и} \quad \Big(\mathbf{g}_0^{s+1}(x_{s+1}), \dots, \mathbf{g}_r^{s+1}(x_{s+1})\Big),$$

последовательно находят решение на отрезках $[x_{m-2}, x_{m-1}], \dots, [x_0, x_1]$. Точки $x_1, \dots, x_{m-1}$ выбираются из условия, чтобы система векторов $\mathbf{g}_1^s(x_{s+1}), \dots, \mathbf{g}_r^s(x_{s+1})$ была близка ортонормированной, а вектор $\mathbf{g}_0^s(x_{s+1})$ не образовывал слишком малый угол с подпространством, натянутым на этот базис. Поскольку векторы $\mathbf{g}_1^s(x_s), \dots, \mathbf{g}_r^s(x_s)$ образуют ортонормированную систему, а вектор $\mathbf{g}_0^s(x_s)$ ортогонален к ним, сформулированные условия выполнены при достаточно малой величине $\max\limits_s (x_{s+1} - x_s)$.

Заметим, что в отличие от подробно рассмотренного в § 3,4 метода прогонки для дифференциального уравнения второго порядка методы прогонки, основанные на идее ортогонализации, при достаточно малой величине $\max\limits_s (x_{s+1} - x_s)$ оказываются слабо чувствительными к влиянию вычислительной погрешности для любых хорошо определенных краевых задач.

Непрерывный аналог описанного выше метода ортогональной прогонки Годунова заключается в следующем: находится матрица $Z(0)$ размерности $l \times (l - r)$, столбцы которой образуют ортонормированную систему решений системы уравнений $B\mathbf{z} = \mathbf{0}$, и вектор $\mathbf{u}(0)$ ортогональный к этим векторам, удовлетворяющий системе уравнений $B\mathbf{u}(0) = \mathbf{b}$.

При начальных условиях $Z(0), \mathbf{u}(0)$ решается задача Коши для системы уравнений

$$Z' = AZ - Z(Z^TZ)^{-1}R,$$

$$\mathbf{u}' = \Big(E - Z(Z^TZ)^{-1}Z^T\Big)\Big(A\mathbf{u} + \mathbf{f} - Z(Z^TZ)^{-1}Z^TA^T\mathbf{u}\Big);$$

верхняя треугольная матрица R определяется равенством

$$R + R^T = Z^T(A + A^T)Z.$$

Совокупность этих вычислений называют прямым ходом метода прогонки.

В этом методе вектор $\mathbf{u}(x)$ при каждом x минимизирует $\|\mathbf{y}(x)\|$ среди значений этой нормы у решений системы $\mathbf{y}' = A\mathbf{y} + \mathbf{f}$, удовлетворяющих левому граничному условию.

Так называемый обратный ход метода прогонки заключается в следующем. Определяется вектор $\mathbf{v}(1)$ размерности $l - r$ — решение системы уравнений $D\Big(Z(1)\mathbf{v}(1) + \mathbf{u}(1)\Big) = \mathbf{d}$ и при заданном $\mathbf{v}(1)$ решается в направлении убывания x задача Коши для системы уравнений

$$\mathbf{v}' = R\mathbf{v} + Z^T(A + A^T)\mathbf{u} + Z^T\mathbf{f}.$$

Само решение задачи $\mathbf{y}(x)$ вычисляется по формуле

$$\mathbf{y}(x) = \mathbf{u}(x) + Z(x)\mathbf{v}(x). \tag{8}$$

Возможен и такой вариант действий. Аналогично найденным в процессе прямого хода метода прогонки функциям $Z(x)$ и $\mathbf{u}(x)$, которые мы обозначим как $Z_{\text{лев}}(x)$ и $\mathbf{u}_{\text{лев}}(x)$, находятся функции $Z_{\text{пр}}(x)$ и $\mathbf{u}_{\text{пр}}(x)$, соответствующие граничным условиям на правом конце отрезка интегрирования.

Значения решения в каждой точке x находятся из системы уравнений относительно неизвестных значений $\mathbf{y}(x)$, $\mathbf{v}_{\text{лев}}(x)$ и $\mathbf{v}_{\text{пр}}(x)$:

$$\mathbf{y}(x) = \mathbf{u}_{\text{лев}}(x) + Z_{\text{лев}}(x)\mathbf{v}_{\text{лев}}(x), \quad \mathbf{y}(x) = \mathbf{u}_{\text{пр}}(x) + Z_{\text{пр}}(x)\mathbf{v}_{\text{пр}}(x). \tag{9}$$

Здесь $\mathbf{v}_{\text{лев}}(x)$ и $\mathbf{v}_{\text{пр}}(x)$ — векторы-столбцы размерности $l - r$ и r соответственно.

Часто удобнее применить следующий метод ортогональной прогонки Абрамова. Краевые условия преобразуются к виду

$$B\mathbf{y}(0) = \mathbf{b}, \quad D\mathbf{y}(1) = \mathbf{d},$$

такому, чтобы строки матриц B и D образовывали ортонормированные системы векторов, т. е. выполнялись равенства $BB^T = E$, $DD^T = E$; здесь E — единичные матрицы размерностей $r \times r$ и $(l - r) \times (l - r)$ соответственно.

На отрезке [0,1] решаются задачи Коши для системы уравнений

$$\begin{gathered} Z' + ZA(E - Z^T(ZZ^T)^{-1}Z) = 0, \\ \mathbf{u}' - ZAZ^T(ZZ^T)\mathbf{u} - Zf = 0 \end{gathered} \tag{10}$$

при начальных условиях $Z(0) = B, \mathbf{u}(0) = \mathbf{b}$ в направлении возрастания x и при начальных условиях $Z(1) = D, \mathbf{u}(1) = \mathbf{d}$ в направлении убывания x. Полученные решения обозначаются, соответственно,

$$Z_{\text{лев}}(x), \ \mathbf{u}_{\text{лев}}(x) \ \text{ и } \ Z_{\text{пр}}(x), \ \mathbf{u}_{\text{пр}}(x).$$

Решение задачи в каждой точке находится из системы уравнений

$$Z_{\text{лев}}(x)\,\mathbf{y}(x) = \mathbf{u}_{\text{лев}}(x), \quad Z_{\text{пр}}(x)\,\mathbf{y}(x) = \mathbf{u}_{\text{пр}}(x). \tag{11}$$

В этом методе строки матрицы Z образуют наиболее медленно изменяющийся базис ортогонального дополнения к этому пространству.

Упомянем ряд фактов, свидетельствующих об определенных «хороших» свойствах указанных методов.

Для этих методов выполняются, соответственно, равенства

$$Z^T(x)Z(x) = E \ \text{ или } \ Z(x)Z^T(x) = E, \ \text{где } E \text{ — единичная матрица.} \tag{12}$$

Для первого метода выполнены неравенства $\|\mathbf{u}(x)\| \leqslant \|\mathbf{y}(x)\|$, $\|\mathbf{v}\| \leqslant \|\mathbf{y}(x) - \mathbf{u}(x)\|$; здесь $\|\cdot\|$ — евклидова норма вектора. Для второго метода выполнено неравенство $\|\mathbf{u}(x)\| \leqslant \|\mathbf{y}(x)\|$.

Вследствие равенств (12) возникает соблазн отказаться от обращения матриц $Z(x)Z^T(x)$ или $Z^T(x)Z(x)$. Такое упрощение часто приводит к нежелательному росту погрешности.

§ 7. Нелинейные краевые задачи

Существует большое сходство между методами решения нелинейных краевых задач и нелинейных алгебраических задач. В частности, так же, как и в последнем случае, мы проведем обсуждение различных методов, не приходя в конце концов к конкретной рекомендации по решению произвольных нелинейных краевых задач. По существу, всякий раз решение нелинейной краевой задачи сводится к решению некоторой нелинейной системы уравнений. Различные методы решения нелинейных краевых задач отличаются выбором параметров этих вспомогательных задач и, естественно, методом решения этих задач.

Рассмотрим простейший пример — нелинейную краевую задачу: найти решение уравнения

$$\begin{aligned} &x'' - f(t, x) = 0 \quad \text{при} \quad t \in (0, T), \\ &x(0) - a = 0, \qquad x(T) - b = 0. \end{aligned} \tag{1}$$

Предположим, что известно некоторое приближение к решению $x_n(t)$; в окрестности этого приближения справедливо разложение

$$f(t, x) \approx f\Big(t, x_n(t)\Big) + \frac{\partial f}{\partial x}\Big(t, x_n(t)\Big)(x - x_n(t)),$$

поэтому представляется целесообразным искать следующее приближение к решению $x_{n+1}(t)$ как решение краевой задачи

$$\begin{aligned} &x''_{n+1} - \Big(f(t, x_n(t)) + \frac{\partial f}{\partial x}\Big(t, x_n(t)\Big)(x_{n+1}(t) - x_n(t))\Big) = 0, \\ &x_{n+1}(0) - a = 0, \qquad x_{n+1}(T) - b = 0. \end{aligned} \tag{2}$$

Рассмотрим сеточную аппроксимацию задачи (1):

$$\begin{aligned} &\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} - f(t_k, x_k) = 0, \quad k = 1, \ldots, N-1, \\ &x_0 - a = 0, \qquad x_N - b = 0; \end{aligned} \tag{3}$$

здесь $h = T/N$, x_k — приближения к значениям $x(kh)$. Пусть x_k^n, $k = 0, \ldots, N$, — совокупность величин, образующих n-е приближение к решению системы (3). В окрестности этого приближения справедливо соотношение

$$f(t_k, x_k) \approx f(t_k, x_k^n) + f_x(t_k, x_k^n)(x_k - x_k^n).$$

Поэтому следующее приближение отыскиваем из системы уравнений

$$\frac{x_{k+1}^{n+1} - 2x_k^{n+1} + x_{k-1}^{n+1}}{h^2} - \\ - (f(t_k, x_k^n) + f_x(t_k, x_k^n)(x_k^{n+1} - x_k^n)) = 0, \quad k = 1, \ldots, N-1, \qquad (4)$$
$$x_0^{n+1} - a = 0, \qquad x_N^{n+1} - b = 0,$$

являющейся дискретной аппроксимацией (2). В § 3 мы изучали применение методов стрельбы и прогонки для решения таких линейных уравнений.

Эта же совокупность уравнений (2) относительно следующего приближения к решению получится при формальном применении схемы метода Ньютона. Рассматриваем (1) как операторное уравнение $F(x) = 0$; оператором производной F' будет оператор P, определяемый равенствами

$$P\eta = \begin{cases} \eta'' - f_x\big(t, x(t)\big)\eta & \text{при} \quad t \in (0, T), \\ \eta(0) & \text{при} \quad t = 0, \\ \eta(T) & \text{при} \quad t = T. \end{cases}$$

Уравнение метода Ньютона относительно $x_{n+1}(t)$ записывается в виде

$$x_n''(t) - f(t, x_n(t)) + (x_{n+1}(t) - x_n(t))'' - f_x(t, x_n(t))(x_{n+1}(t) - x_n(t)) = 0,$$
$$x_n(0) - a + (x_{n+1}(0) - x_n(0)) = 0,$$
$$x_n(T) - b + (x_{n+1}(T) - x_n(T)) = 0$$

и совпадает с (2).

Задача 1. Выписать расчетные формулы метода Ньютона для системы (3) и убедиться, что они совпадают с (4).

В обоих вариантах, непрерывном и дискретном, мы построили сначала итерационные методы (2) и (4), линеаризуя правую часть, а потом убедились, что эти методы совпадают с итерационным методом Ньютона.

Заметим, что итерационный метод (2), как правило, не может быть реализован из-за невозможности решения уравнения (2) в явном виде, и на практике имеют дело именно с методом (4).

Метод (4) может представлять практические неудобства из-за необходимости хранить в памяти ЭВМ все значения x_k^n, $k = 0, \ldots, N$. Поэтому часто прибегают к следующему методу, который также будет изложен в непрерывном и дискретном вариантах.

Исходная задача будет решена, если будет найдено значение x_0' производной решения в точке 0. Решая задачу Коши для уравнения $x'' = f(t, x)$ при начальных условиях $x(0) = a$, $x'(0) = x_0'$, получим решение на всем отрезке. Если решить задачу Коши при начальных условиях

$x(0) = a$ и произвольном $x'(0)$, то получим значение $x(T)$, вообще говоря, отличное от b. Значение $x(T)$ решения задачи Коши является некоторой функцией от $x'(0)$: $x(T) = \varphi(x'(0))$. Таким образом, исходная задача сводится к решению нелинейного уравнения

$$F(x'(0)) = \varphi(x'(0)) - b = 0. \tag{5}$$

Нахождение значения $F(\alpha)$ при каждом α требует решения задачи Коши

$$x'' = f(t, x), \quad x(0) = a, \quad x'(0) = \alpha. \tag{6}$$

В связи с большой трудоемкостью нахождения значения $F(\alpha)$ при каждом α для решения уравнения должны применяться методы, требующие вычисления небольшого числа значений функций $F(\alpha)$.

В ряде сложных случаев задача решается в диалоговом режиме; значение $F(\alpha_n)$ для каждого последующего приближения α_n вычисляется ЭВМ, а выбор α_{n+1} осуществляется исследователем, решающим задачу. Подобного рода диалоговые системы применяются на практике много столетий. В случае конкретных работающих промышленных систем руководитель часто подбирает значение α_{n+1}, изучая результат работы системы при ранее взятых значениях $\alpha_1, \ldots, \alpha_n$. В случае артиллерийской стрельбы вычисление значений $F(\alpha_n)$ или $\operatorname{sign} F(\alpha_n)$ осуществляется при помощи посылки снаряда в цель, а выбор α_{n+1} осуществляется лицом, корректирующим стрельбу. Можно найти много аналогов таких алгоритмов в повседневной жизни человека и животных.

Рассмотрим вопрос о решении уравнения (5) методом Ньютона. Согласно формуле дифференцирования решения уравнения второго порядка, справедливо равенство

$$\frac{\partial F(x'(0))}{\partial x'(0)} = \frac{\partial x(T)}{\partial x'(0)} = \eta(T);$$

$\eta(t)$ — решение задачи Коши

$$\begin{aligned} &\eta'' - f_x(t, x(t))\eta = 0, \\ &\eta(0) = 0, \quad \eta'(0) = 1. \end{aligned} \tag{7}$$

Численно или аналитически решая эту задачу, находим значение $\eta(T)$.

Поскольку задачу Коши (6), как правило, требуется решать численно, исследуем сразу дискретный вариант этого метода. Рассмотрим сеточную аппроксимацию (3); зададимся произвольным x_1 и, пользуясь рекуррентной формулой, вытекающей из (3),

$$x_{k+1} = 2x_k - x_{k-1} + h^2 f(t_k, x_k), \quad k = 1, \ldots, N-1, \tag{8}$$

при начальных условиях $x_0 = a$, x_1, находим сеточную функцию, удовлетворяющую сеточному уравнению и левому граничному условию. Значение x_N является некоторой функцией от x_1, $x_N = \varphi(x_1)$, вычисление

каждого значения которой производится с помощью рекуррентного процесса (8). Таким образом, решение задачи (3) сводится к решению скалярного уравнения

$$F(x_1) = \varphi(x_1) - b = 0.$$

Предположим, что для решения этого уравнения используется метод Ньютона. Продифференцировав (8) по x_1, получим

$$\eta_{k+1} = 2\eta_k - \eta_{k-1} + h^2 \frac{\partial f(t_k, x_k)}{\partial x_k} \eta_k, \quad k = 1, \ldots, N-1; \tag{9}$$

здесь $\eta_k = \partial x_k / \partial x_1$. Кроме того, имеем

$$\eta_0 = \frac{\partial x_0}{\partial x_1} = \frac{\partial a}{\partial x_1} = 0; \quad \eta_1 = \frac{\partial x_1}{\partial x_1} = 1.$$

Таким образом, при заданном x_1, производя параллельно вычисления x_k по формуле (8) и η_k по формуле (9), можно найти $x_N(x_1)$ и $\partial x_N / \partial x_1 = \eta_N$ и осуществить следующий шаг метода Ньютона:

$$x_1^{n+1} = x_1^n - \left(\frac{\partial x_N}{\partial x_1^n}\right)^{-1} \Big(\varphi(x_1^n) - b\Big). \tag{10}$$

Если соотношение (9) переписать в виде

$$\frac{\eta_{k+1} - 2\eta_k + \eta_{k-1}}{h^2} - \frac{\partial f(t_k, x_k)}{\partial x_k} \eta_k = 0,$$

то оно превращается в разностную схему, аппроксимирующую уравнение (7).

В случае, когда (1) — скалярное уравнение, вместо (10) часто целесообразно вычислить $\partial x_N / \partial x_1^n$ по приближенной формуле

$$\frac{\partial x_N}{\partial x_1^n} \approx \frac{x_N(x_1^n) - x_N(x_1^n - \Delta_n)}{\Delta_n}, \tag{11}$$

в частности, можно брать $\Delta_n = x_1^n - x_1^{n-1}$.

Заметим, что в случае использования формулы (10) и вычисления производной с помощью (11), нужно обращать внимание на разумный выбор величины Δ_n (см. § 2.16), поскольку величина погрешности значения $x_N(x_1)$, получаемого путем численного интегрирования, часто оказывается довольно большой.

Рассмотрим нелинейную краевую задачу

$$\begin{gathered} \mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{B}(\mathbf{y}(0)) = \mathbf{0}, \quad \mathbf{D}(\mathbf{y}(X)) = \mathbf{0}, \\ \mathbf{y} = (y_1, \ldots, y_l)^T, \quad \mathbf{B} = (b_1, \ldots, b_{l-r})^T, \quad \mathbf{D} = (d_1, \ldots, d_r)^T. \end{gathered} \tag{12}$$

Пусть функции $g_1(\mathbf{y}(0)), \ldots, g_r(\mathbf{y}(0))$ таковы, что система уравнений

$$\begin{gathered} b_1(\mathbf{y}(0)) = 0, \ldots, b_{l-r}(\mathbf{y}(0)) = 0, \\ g_1(\mathbf{y}(0)) = g_1, \ldots, g_r(\mathbf{y}(0)) = g_r \end{gathered} \tag{13}$$

однозначно определяет вектор $\mathbf{y}(0) = \omega_0(\mathbf{g}), \mathbf{g} = (g_1, \ldots, g_r)$. Тогда задача (12) может быть сведена к системе нелинейных уравнений относительно параметров $g_1, \ldots, g_r$.

Довольно часто граничное условие в точке 0 имеет вид

$$y_1(0) = 0, \ldots, y_{l-r}(0) = 0;$$

тогда в качестве $g_1(\mathbf{y}(0)), \ldots, g_r(\mathbf{y}(0))$ целесообразно взять $y_{l-r+1}(0), \ldots, y_l(0)$. Таким образом, здесь в качестве искомых параметров выступают неизвестные компоненты решения в точке $x = 0$.

Решая систему (13) при каждых $g_1, \ldots, g_r$, можно определить вектор $\mathbf{y}(0) = \omega_0(\mathbf{g})$; решая задачу Коши при начальном условии $\mathbf{y}(0)$, определяем $\mathbf{y}(X) = \omega_X(\mathbf{g})$ и затем находим $\psi(\mathbf{g}) = \mathbf{D}(\omega_X(\mathbf{g}))$. Таким образом, решение задачи (12) сводится к решению нелинейной системы уравнений

$$\psi(\mathbf{g}) = \mathbf{0}. \tag{14}$$

Значительную группу методов решения нелинейных систем составляют методы типа метода Ньютона, где наряду со значениями функций ψ_i используются значения их производных $\partial\psi_i/\partial g_k$. Наиболее распространены следующие два способа нахождения этих производных.

1. Для определенности речь пойдет о вычислении производных $\partial\psi_i/\partial g_1$. Пусть мы задались значениями $g_1^0, \ldots, g_r^0$. Из системы уравнений (13) найдем соответствующие $y_1(0), \ldots, y_l(0)$. Дифференцируя уравнения системы (13) по g_1, получаем систему уравнений для отыскания производных $\partial y_j(0)/\partial g_1$:

$$\sum_{j=1}^{l} \frac{\partial b_k}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = 0, \quad k = 1, \ldots, l - r,$$

$$\sum_{j=1}^{l} \frac{\partial g_m}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = \delta_m^1, \quad m = 1, \ldots, r.$$

Решение системы $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ при начальных условиях $y_1(0), \ldots, y_l(0)$ является функцией от параметров g_m; вектор производных

$$\frac{\partial \mathbf{y}(x)}{\partial g_1} = \left(\frac{\partial y_1(x)}{\partial g_1}, \ldots, \frac{\partial y_l(x)}{\partial g_1}\right)^T$$

удовлетворяет системе дифференциальных уравнений

$$\frac{d}{dx}\left(\frac{\partial \mathbf{y}(x)}{\partial g_1}\right) = F_y(x, \mathbf{y}) \frac{\partial \mathbf{y}(x)}{\partial g_1}, \quad F_y = \left[\frac{\partial f_i}{\partial y_j}\right]. \tag{15}$$

После численного решения этой системы получаем вектор $\partial \mathbf{y}(X)/\partial g_1$. Теперь можно найти производные

$$\frac{\partial \psi_i}{\partial g_1} = \sum_{j=1}^{l} \frac{\partial d_i}{\partial y_j} \frac{\partial y_j(X)}{\partial g_1} = \left(\operatorname{grad} d_i, \frac{\partial \mathbf{y}(X)}{\partial g_1} \right). \tag{16}$$

Этот путь может быть целесообразен, если, например, многократно решается задача (12) при одной и той же правой части $f(x, \mathbf{y})$, но при различных граничных условиях.

В качестве стандартного метода определения производных $\partial \psi_i/\partial g_j$ этот метод неудобен тем, что наряду с написанием программы вычисления правых частей уравнения он требует от пользователя также написания программы вычисления производных $\partial f_i/\partial y_j$.

2. Для большинства итерационных методов решения систем уравнений требуется вычислять значения производных правых частей лишь с умеренной точностью. Поэтому для нахождения этих производных можно просто воспользоваться какими-либо формулами численного дифференцирования, например, простейшей:

$$\frac{\partial \psi_i(g_1, \ldots, g_r)}{\partial g_1} \approx \frac{\psi_i(g_1 + \Delta, \ldots, g_r) - \psi_i(g_1, \ldots, g_r)}{\Delta}.$$

Вычисление производных $\partial \psi_i/\partial g_1$ по этой формуле требует дополнительно (по сравнению с нахождением значений $\psi_i(g_1, \ldots, g_r)$) еще одного численного решения задачи Коши, соответствующей параметрам $g_1 + \Delta, g_2, \ldots, g_r$. Поскольку при численном нахождении значений функций ψ_i может иметь место большая погрешность, которая еще более возрастает при делении на Δ, здесь следует обратить внимание на разумный выбор Δ. Для описываемого метода типично одновременное нахождение производных всех функций ψ_i по фиксированной переменной g_j; поэтому здесь может оказаться разумным применение итерационных процессов с поочередным уточнением параметров g_j. Можно представить себе такой итерационный процесс: параметры g_j уточняются в циклическом порядке; уточненное значение параметра $\bar{g}_j$ выбирается из условия минимума некоторой функции

$$\Phi_j \Big(\psi_1(g_1, \ldots, g_r) + \frac{\partial \psi_1(g_1, \ldots, g_r)}{\partial g_j} (\bar{g}_j - g_j), \ldots \\ \ldots, \psi_r(g_1, \ldots, g_r) + \frac{\partial \psi_r g_1, \ldots, g_r)}{\partial g_j} (\bar{g}_j - g_j) \Big)$$

(см. подробнее § 7.3).

Так же как при решении линейных краевых задач, возникает вопрос о применимости метода в условиях реальных округлений. Если решения η системы уравнений в вариациях $\eta' = f_y \eta$ сильно растут с ростом x,

то погрешности в значениях g_j и вычислительные погрешности при численном интегрировании приводят к большим погрешностям в значениях функции ψ_i; это в конечном счете приведет к большой погрешности получаемого решения. Если такая погрешность окажется недопустимой, то следует ввести иную параметризацию задачи.

В настоящее время в вычислительной практике получили распространение методы, занимающие промежуточное положение между методами, соответствующими формулам (3), где в качестве неизвестных параметров выступают значения решения во всех узлах сетки, и описанным выше методом, где в качестве неизвестных параметров выступают значения решения в одной точке. Задаются точками $0 = x_0 < x_1 < \cdots < x_m = X$ такими, что на отрезках $[x_{i-1}, x_i]$, $i = 1, \ldots, m$, решения системы уравнений в вариациях не очень сильно возрастают как при продвижении в направлении положительных x, так и при продвижении в направлении отрицательных x. В качестве отыскиваемых параметров принимаются неизвестные компоненты решения в точках x_0 и x_m и значения решений в точках $x_1, \ldots, x_{m-1}$ (более подробно см. [1]).

При практическом решении конкретных нелинейных задач для обыкновенных дифференциальных уравнений, как и в случае других нелинейных задач, обычно приходится заниматься «доводкой» метода: предлагается какой-то специальный метод получения начального приближения, который затем модернизируется с целью расширения области начальных условий, при которых он сходится для данного конкретного класса задач. В ряде случаев метод решения строится путем имитации на ЭВМ методов, встречающихся в живой природе, или применяемых практиками для решения задач данного класса. Если рассматриваемая краевая задача является задачей на экстремум некоторого функционала, то исходный функционал приближается функционалом, зависящим от конечного числа параметров, и путем линеаризации последнего получают достаточно хорошее приближение. Примеры подобных функционалов для случая линейных задач будут рассмотрены в § 11.

§ 8. Аппроксимации специального типа

Рассмотрим краевую задачу

$$\Big(k(x)y'(x)\Big)' - p(x)y(x) = f(x), \quad y(0) = a, \quad y(X) = b, \tag{1}$$

где $p(x) \geqslant 0$, $k(x) \geqslant k_0 > 0$, $k(x)$ — трижды, $p(x)$, $f(x)$ — дважды непрерывно дифференцируемые функции, за исключением конечного числа точек, где эти функции или их производные k', k'', k''', p', p'', f', f'' могут иметь разрывы первого рода.

Пусть K, P, F — множества точек разрыва соответственно функций $k(x)$, $p(x)$, $f(x)$ или их производных, $\Omega = K \cup P \cup F$.

В случае уравнений с разрывными коэффициентами или решениями иногда не ясно, как понимать уравнение в точках разрыва. Для решения

этого вопроса следует обратиться к интегральным соотношениям, обычно называемым законами сохранения, из которых были получены рассматриваемые дифференциальные уравнения. Уравнение (1) с разрывным $k(x)$ обычно возникает из интегрального соотношения

$$\int_{x_1}^{x_2} \Big(p(x)y(x) + f(x)\Big)\,dx = k(x)y'\Big|_{x_1}^{x_2}, \tag{2}$$

которое выполнено для любых $x_1, x_2 \in [0, X]$. Если $x_2 \to x_0 + 0$, $x_1 \to x_0 - 0$, то левая часть стремится к нулю; переходя к пределу в правой части, получаем, что

$$k(x)y'\Big|_{x_0-0}^{x_0+0} = 0 \quad \text{для любых} \quad x_0 \in [0, X].$$

Исходя из этого соотношения, решением задачи (1) будем называть функцию $y(x)$, удовлетворяющую следующим условиям:

1) $y(x)$ непрерывна на $[0, X]$;

2) $y(x)$ удовлетворяет уравнению всюду на $[0, X]$, за возможным исключением точек множества Ω;

3) функция $w(x) = k(x)y'(x)$, называемая *потоком*, непрерывна на $[0, X]$.

Из условия 3) следует, что функция $y'(x) = w(x)/k(x)$ непрерывна всюду, за исключением точек разрыва функции $k(x)$; в этих точках $y'(x)$ будет иметь разрывы первого рода. Рассмотрим сначала равномерную сетку. При n целых и полуцелых будем употреблять обозначение $x_n = nh$; далее $X = Nh$.

Можно показать, что такое решение существует, а производные $y'(x)$, $y''(x)$, $y'''(x)$ и $y^{(4)}(x)$ непрерывны и равномерно ограничены на множестве $[0, X]\backslash\Omega$.

Если при построении разностной схемы не учитывается факт разрывности $y'(x)$, то может случиться, что решение разностной задачи не сходится к решению задачи (1). Например, такая ситуация возникнет, если раскрыть скобки в первом слагаемом

$$\Big(k(x)y'(x)\Big)' = ky''(x) + k'(x)y'(x)$$

и аппроксимировать его выражением

$$k(x_n)\,\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} + k'(x_n)\,\frac{y_{n+1} - y_{n-1}}{2h}.$$

Дело в том, что хотя в области гладкости коэффициентов погрешности аппроксимации порядка $O(h^2)$, но условие 3) не учтено при построении

схемы. Рассмотрим другую аппроксимацию:

$$(ky')'\big|_{x_n} \approx \frac{(ky')|_{x_{n+1/2}} - (ky')|_{x_{n-1/2}}}{h} \approx$$
$$\approx \frac{k(x_{n+1/2})\,\dfrac{y_{n+1}-y_n}{h} - k(x_{n-1/2})\,\dfrac{y_n - y_{n-1}}{h}}{h}, \tag{3}$$
$$(py)|_{x_n} \approx p(x_n)y_n, \quad f(x_n) \approx f_n.$$

Можно показать, что решение соответствующей разностной задачи

$$\frac{k(x_{n+1/2})\,\dfrac{y_{n+1}-y_n}{h} - k(x_{n-1/2})\,\dfrac{y_n - y_{n-1}}{h}}{h} - p(x_n)y_n = f_n,$$
$$n = 1, \ldots, N-1, \quad y_0 = a, \quad y_N = b,$$

сходится к решению дифференциальной со скоростью $O(h)$. В то же время оказывается, что в случае, когда один из интервалов (x_{n-1}, x_n), (x_n, x_{n+1}) содержит точку разрыва функции $k(x)$, погрешность аппроксимации в точке x_n имеет порядок h^{-1}.

Проверим последнее утверждение и объясним, почему оно не противоречит факту сходимости со скоростью $O(h)$. Погрешность аппроксимации записывается в виде

$$r_n = \frac{v(x_{n+1/2}) - v(x_{n-1/2})}{h} - p(x_n)y(x_n) - f(x_n), \tag{4}$$

где

$$v(x_{n+1/2}) = k(x_{n+1/2})\frac{y(x_{n+1}) - y(x_n)}{h}.$$

Положим $v(x_{n+1/2}) - k(x_{n+1/2})y'(x_{n+1}) = \beta_n$; поскольку $k(x)$ и производная $y'(x)$ равномерно ограничены на $[0, X]$, то величина $\left|\dfrac{y(x_{n+1}) - y(x_n)}{h}\right|$ равномерно ограничена по n и h и $\sup\limits_{n,\,h} |\beta_n| = O(1)$.

Если производная $y'''(x)$ ограничена на $[x_n, x_{n+1}]$, то

$$|\beta_n| \leqslant \frac{1}{24}\sup|k(x)| \cdot \sup_{[x_n,\,x_{n+1}]} |y'''|\,h^2 = O(h^2).$$

Соотношение (4) записывается в виде

$$r_n = \frac{(ky')|_{x_{n+1/2}} - (ky')|_{x_{n-1/2}}}{h} - p(x_n)y(x_n) - f(x_n) + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h}. \tag{5}$$

Положим

$$\frac{(ky')|_{x_{n+1/2}} - (ky')|_{x_{n-1/2}}}{h} - (ky')'\Big|_{x_n} = \alpha_n;$$

если величина $(ky')'''$ ограничена, то $\alpha_n = O(h^2)$. Окончательное выражение погрешности аппроксимации имеет вид

$$r_n = \alpha_n + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h} + \Big((ky')' - py - f\Big)\Big|_{x_n} = \alpha_n + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h}.$$

Рассмотрим случай, когда на (x_n, x_{n+1}) имеется точка разрыва коэффициента $k(x)$; тогда для $\beta_{n+1/2}$ не удается получить оценки лучшей, чем $O(1)$, а для r_n — лучшей чем $O(1/h)$. Тем не менее погрешность решения имеет порядок $O(h)$. Наглядно это можно объяснить следующим образом. Точек, где r_n имеет порядок $O(1/h)$, конечное число; их доля в общем числе узлов порядка $O(h)$, поэтому суммарный вклад от погрешности аппроксимации в таких точках будет $O(h) \cdot O(1/h) = O(1)$.

Значение $\beta_{n+1/2}$ входит в выражение погрешности аппроксимации следующим образом:

$$r_n = \alpha_n + \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h},$$

$$r_{n+1} = \alpha_{n+1} + \frac{\beta_{n+3/2} - \beta_{n+1/2}}{h}.$$

Поэтому $\beta_{n+1/2}$ дает вклад в погрешность аппроксимации в точке x_n, равный $\beta_{n+1/2}/h$, а в точке x_{n+1} — равный $-\beta_{n+1/2}/h$. Погрешность решения записывается в виде (см. § 2)

$$y_q - y(x_q) = h \sum G_q^n r_n.$$

Можно показать, что сеточная функция Грина G_q^n удовлетворяет условию $\left|\dfrac{G_q^{n+1} - G_q^n}{h}\right| \leqslant c$, где постоянная c не зависит от h. Вследствие этого вклад от величины $\beta_{n+1/2}$ есть

$$h(G_q^n - G_q^{n+1})\,\frac{\beta_{n+1/2}}{h} = O(h)\,|\beta_{n+1/2}| = O(h).$$

Из последних соотношений следует, что $\max\limits_q |y_q - y(x_q)| = O(h)$.

Приведенные выше соображения подтверждают широко распространенное эмпирическое правило: *при построении разностных схем не следует зря раскрывать скобок и пользоваться формулой дифференцирования произведения.*

Построим более точную разностную схему исходя из закона сохранения (2). Имеем равенство, которое следует из (2) в результате интегрирования в пределах от $x_{n-1/2}$ до $x_{n+1/2}$:

$$w(x_{n+1/2}) - w(x_{n-1/2}) = \int_{x_{n-1/2}}^{x_{n+1/2}} \Big[p(x)y(x) + f(x)\Big]\,dx. \qquad (6)$$

Так как функция $y(x)$ кусочно-дифференцируема, то

$$y(x) = y(x_n) + O(h) \quad \text{при} \quad x - x_n = O(h).$$

Поэтому (6) можно переписать в виде

$$w(x_{n+1/2}) - w(x_{n-1/2}) = \int_{x_{n-1/2}}^{x_{n+1/2}} \Big(p(x)y(x_n) + f(x) + O(h)\Big)\,dx =$$

$$= h(\bar{p}_n y(x_n) + \bar{f}_n) + O(h^2);$$

здесь

$$\bar{p}_n = \frac{1}{h}\int_{x_{n-1/2}}^{x_{n+1/2}} p(x)\,dx, \quad \bar{f}_n = \int_{x_{n-1/2}}^{x_{n+1/2}} f(x)\,dx.$$

После деления на h получится соотношение

$$\frac{w(x_{n+1/2}) - w(x_{n-1/2})}{h} - \bar{p}_n y(x_n) = \bar{f}_n + O(h).$$

В случае, если интервал (x_{n-1}, x_n) содержит точки разрыва $y'(x)$, погрешность от непосредственной замены выражения $\dfrac{w(x_{n-1/2})}{h}$ на $k(x_{n-1/2})\dfrac{y(x_n) - y(x_{n-1})}{h^2}$ может оказаться величиной порядка h^{-1}.

Для получения лучшей аппроксимации введем в рассмотрение вспомогательную независимую переменную $t = \displaystyle\int_0^x \frac{dx}{k(x)}$. Из ограниченности производной по x и равенства $\dfrac{du}{dt} = k(x)\dfrac{du}{dx}$ следует ограниченность производной по t. Функция $w = k(x)\dfrac{dy}{dx} = \dfrac{dy}{dt}$ имеет ограниченную производную по x, а следовательно, и по t. Таким образом, вторая производная ограничена и можно написать равенство

$$k(x)\frac{dy}{dx}\bigg|_{x_{n-1/2}} = \frac{dy}{dt}\bigg|_{t_{n-1/2}} = \frac{y(t_n) - y(t_{n-1})}{t_n - t_{n-1}} + O(t_n - t_{n-1});$$

здесь

$$t_n = \int_0^{x_n} \frac{dx}{k(x)}; \quad t_n - t_{n-1} = \int_{x_{n-1}}^{x_n} \frac{dx}{k(x)} = O(h).$$

Поэтому

$$w(x_{n-1/2}) = k(x)\frac{dy}{dx}\bigg|_{x_{n-1/2}} = \frac{y(x_n) - y(x_{n-1})}{t_n - t_{n-1}} + O(h).$$

После подстановки $k(x)\left.\dfrac{dy}{dx}\right|_{x_{n\pm1/2}}$ в левую часть получим

$$r_n = \frac{1}{h}\left(\frac{y(x_{n+1}) - y(x_n)}{h\bar{k}_{n+1/2}^{-1}} - \frac{y(x_n) - y(x_{n-1})}{h\bar{k}_{n-1/2}^{-1}}\right) - \bar{p}_n y(x_n) - \bar{f}_n = O(1),$$

где

$$\bar{k}^{-1}_{j+1/2} = \frac{1}{h}\int_{x_j}^{x_{j+1}} \frac{dx}{k(x)}. \tag{7}$$

Соответствующая конечно-разностная схема (предложенная Самарским и Тихоновым) имеет вид

$$L(y_n) = \frac{1}{h}\left(\frac{y_{n+1}-y_n}{h\bar{k}^{-1}_{n+1/2}} - \frac{y_n - y_{n-1}}{h\bar{k}^{-1}_{n-1/2}}\right) - \bar{p}_n y_n = \bar{f}_n. \tag{8}$$

Максимум погрешности аппроксимации у полученной схемы есть $O(1)$, поэтому для получения оценки погрешности привлекается ряд дополнительных соображений.

Если отрезок $[x_{n-1},\, x_{n+1}]$ не содержит точек Ω, то непосредственной проверкой с помощью разложения в ряд Тейлора устанавливается, что погрешность аппроксимации r_n есть $O(h^2)$. В противном случае погрешность аппроксимации представляется в виде

$$\begin{aligned} r_n = L_h(y(x_n)) - \bar{f}_n = \frac{1}{h}\left(\frac{y(x_{n+1})-y(x_n)}{h\bar{k}^{-1}_{n+1/2}} - \frac{y(x_n)-y(x_{n-1})}{h\bar{k}^{-1}_{n-1/2}}\right) - \\ - \bar{p}_n y(x_n) - \bar{f}_n = \frac{\beta_{n+1/2}-\beta_{n-1/2}}{h} + \alpha_n, \end{aligned} \tag{9}$$

где

$$\alpha_n = \frac{w(x_{n+1/2}) - w(x_{n-1/2})}{h} - \bar{p}_n y(x_n) - \bar{f}_n,$$

$$\beta_{n+1/2} = \frac{y(x_{n+1})-y(x_n)}{h\bar{k}^{-1}_{n+1/2}} - w(x_{n+1/2}).$$

Если $(x_n,\, x_{n+1})$ не содержит точек Ω, то разложением в ряд Тейлора устанавливаем, что $\beta_{n+1/2} = O(h^2)$; если $(x_{n-1},\, x_{n+1})$ не содержит точек Ω, то так же устанавливаем, что $\alpha_n = O(h^2)$; в противоположных случаях удается получить лишь оценки

$$\beta_{n+1/2} = O(h), \quad \alpha_n = O(h).$$

Далее, следуя намеченному выше способу оценки с помощью аппарата функции Грина, можно получить оценку

$$\max_n |y_n - y(x_n)| = O(h^2).$$

Ниже будет получена другая оценка погрешности. Из равенств (8) и (9) следует, что

$$r_n = L(y(x_n) - y_n) = L(R_n); \tag{10}$$

здесь R_n — погрешность приближенного решения. Пусть φ_n — сеточная функция, удовлетворяющая, как и R_n, условию

$$\varphi_0 = \varphi_N = 0.$$

Умножим (10) на $h\varphi_n$ и просуммируем в пределах от 1 до $N-1$:

$$h\sum_{n=1}^{N-1} r_n\varphi_n = h\sum_{n=1}^{N-1} L(R_n)\varphi_n. \tag{11}$$

Воспользовавшись выражением (9) для r_n, перепишем это равенство в виде

$$S_1 + S_2 = S_3 + S_4, \tag{12}$$

где

$$S_1(\varphi_n) = h\sum_{n=1}^{N-1} \alpha_n\varphi_n, \quad S_2(\varphi_n) = h\sum_{n=1}^{N-1} \frac{\beta_{n+1/2} - \beta_{n-1/2}}{h}\varphi_n,$$

$$S_3(\varphi_n) = -h\sum_{n=1}^{N-1} \bar{p}_n R_n\varphi_n, \quad S_4(\varphi_n) = h\sum_{n=1}^{N-1} \frac{g_{n+1/2} - g_{n-1/2}}{h}\varphi_n;$$

$$g_{n+1/2} = \bar{k}_{n+1/2}\frac{R_{n+1} - R_n}{h}.$$

Собрав в выражении S_2 подобные члены при одинаковых слагаемых $\beta_{n+1/2}$, получим

$$S_2(\varphi_n) = -h\sum_{n=0}^{N-1} \beta_{n+1/2}\frac{\varphi_{n+1} - \varphi_n}{h}. \tag{13}$$

Заметим, что в правой части дописаны слагаемые $\beta_{N-1/2}\varphi_N$ и $-\beta_{1/2}\varphi_0$, равные нулю в силу условия $\varphi_0 = \varphi_N = 0$.

То же самое выражение для S_2 можно было бы получить, применяя *разностную формулу Абеля суммирования по частям*:

$$\sum_{n=0}^{N-1} (a_{n+1} - a_n)b_n = -\sum_{n=1}^{N} (b_{n+1} - b_n)a_{n+1} + a_N b_N - a_0 b_0. \tag{14}$$

Точно так же получим

$$S_4(\varphi_n) = -h\sum_{n=0}^{N-1} g_{n+1/2}\frac{\varphi_{n+1} - \varphi_n}{h}. \tag{15}$$

Подставим в (12) $\varphi_n = R_n$; имеем

$$S_3(R_n) = -h\sum_{n=1}^{N-1} \bar{p}_n R_n^2 \leqslant 0,$$

$$S_4(R_n) = -h\sum_{n=0}^{N-1} \bar{k}_{n+1/2}\left(\frac{R_{n+1} - R_n}{h}\right)^2 \leqslant 0.$$

Поэтому из (12) получаем

$$\left|S_4(R_n)\right| \leqslant \left|S_3(R_n) + S_4(R_n)\right| \leqslant \left|S_1(R_n)\right| + \left|S_2(R_n)\right|. \tag{16}$$

Из (7) следует, что $\bar{k}_{n+1/2} \geqslant k_0$, поэтому

$$\left|S_4(R_n)\right| \geqslant k_0 S_0(R_n), \quad S_0(R_n) = h\sum_{n=0}^{N-1}\left(\frac{R_{n+1}-R_n}{h}\right)^2.$$

При $R_0 = R_N = 0$ выражение $(S_0(R_n))^{1/2}$ обозначают как $\|R_n\|_{1,h}$.

Очевидно, что оно является сеточным аналогом нормы в пространстве С. Л. Соболева $\overset{\circ}{W}{}_2^1$ функций, удовлетворяющих условиям $\varphi(0) = \varphi(X) = 0$, с нормой

$$\|\varphi\|_1 \equiv \left(\int_0^X \left(\frac{d\varphi}{dx}\right)^2 dx\right)^{1/2}.$$

Нормы

$$\|\varphi_n\|_{0,h} = \left(\sum_{n=0}^{N-1} h|\varphi_n|^2\right)^{1/2} \quad \text{и} \quad \|\varphi_n\|_{C_h} = \max_{0<n<N}|\varphi_n|$$

являются сеточными аналогами норм пространств L_2 и C соответственно.

Теорема (сеточная теорема вложения).

$$\|\varphi_n\|_{0,h} \leqslant \sqrt{X}\|\varphi_n\|_{C_h}, \tag{17}$$

$$\|\varphi_n\|_{C_h} \leqslant \frac{\sqrt{X}}{\sqrt{2}}\|\varphi_n\|_{1,h}. \tag{18}$$

Справедливость первого утверждения непосредственно следует из определения норм и цепочки неравенств

$$\|\varphi_n\|_{0,h} \leqslant \sqrt{h(N-1)\|\varphi_n\|^2_{C_h}} \leqslant \sqrt{X}\|\varphi_n\|_{C_h}.$$

Пусть n_0 — точка, где достигается наибольшее значение $|\varphi_n|$. Рассмотрим случай $n_0 \leqslant N/2$; случай $n_0 > N/2$ сводится к рассматриваемому введением нового индекса $n = N - n$. Имеем равенство

$$\varphi_{n_0} = \sum_{n=0}^{n_0-1}(\varphi_{n+1}-\varphi_n) = \sum_{n=0}^{n_0-1}\sqrt{h}\,\frac{\varphi_{n+1}-\varphi_n}{\sqrt{h}}.$$

Воспользовавшись неравенством для скалярного произведения

$$\left|\sum_{q=1}^{l} a_q b_q\right| \leqslant \sqrt{\sum_{q=1}^{l}|a_q|^2}\cdot\sqrt{\sum_{q=1}^{l}|b_q|^2},$$

получим

$$|\varphi_{n_0}| \leqslant \sqrt{n_0 h}\cdot\sqrt{h\sum_{n=0}^{n_0-1}\left(\frac{\varphi_{n+1}-\varphi_n}{h}\right)^2} \leqslant \sqrt{\frac{X}{2}}\,\|\varphi\|_{1,h}.$$

Теорема доказана.

Из (17), (18) следует, что

$$\|\varphi\|_{0,h} \leqslant \frac{X}{\sqrt{2}}\|\varphi\|_{1,h}. \tag{19}$$

Воспользовавшись (13), получаем оценку

$$|S_2(R_n)| \leqslant \sqrt{h\sum_{n=0}^{N-1}(\beta_{n+1/2})^2}\cdot\sqrt{h\sum_{n=0}^{N-1}\left(\frac{R_{n+1}-R_n}{h}\right)^2} =$$
$$= \sqrt{h\sum_{n=0}^{N-1}(\beta_{n+1/2})^2}\cdot\|R_n\|_{1,h}.$$

Точно так же с учетом (15) имеем

$$|S_1(R_n)| \leqslant \|\alpha_n\|_{L_{0,h}}\|R_n\|_{0,h} \leqslant \frac{\sqrt{X}}{2}\|\alpha_n\|_{0,h}\|R_n\|_{1,h}.$$

Таким образом, из (16) следует

$$k_0\|R_n\|_{1,h}^2 \leqslant \left(\frac{\sqrt{X}}{2}\|\alpha_n\|_{0,h} + \sqrt{h\sum_{n=0}^{N-1}(\beta_{n+1/2})^2}\right)\|R_n\|_{1,h},$$

поэтому

$$\|R_n\|_{1,h} \leqslant \frac{1}{k_0}\sqrt{\frac{X}{2}}\|\alpha_n\|_{0,h} + \frac{1}{k_0}\sqrt{h\sum_{n=0}^{N-1}(\beta_{n+1/2})^2}.$$

Величина α_n порядка $O(h^2)$ за возможным исключением конечного числа n, соответствующих отрезкам $[x_{n-1}, x_{n+1}]$, имеющим общие точки с Ω; для этих точек $\alpha_n = O(h)$. Отсюда следует оценка $\|\alpha_n\|_{0,h} = O(h^{3/2})$.

Точно так же выводится, что

$$\sqrt{h\sum_{n=0}^{N-1}(\beta_{n+1/2})^2} = O(h^{3/2}).$$

Таким образом, $\|R_n\|_{1,h}$, а следовательно, согласно теореме вложения и $\|R_n\|_{C_h}$ есть $O(h^{3/2})$. Напомним, что на самом деле $\|R_n\|_{C_h} = O(h^2)$.

При использовании схемы (8) вычислительный процесс не зависит от положения точек разрыва. Поэтому ее относят к классу *однородных схем*.

Схема (8) на первый взгляд обладает следующим неудобством. Ее коэффициенты $\bar{k}_{q+1/2}$, $\bar{p}_q$, $\bar{f}_q$ записываются как некоторые интегралы. На самом деле можно показать, что если погрешность в значениях этих коэффициентов есть $O(h^2)$, то погрешность приближенного решения оказывается также $O(h^2)$. Поэтому если интервал (x_{q-1}, x_{q+1}) не содержит точек разрывов коэффициентов $k(x)$, $p(x)$, $f(x)$, то без потери порядка точности можно заменить $\bar{p}_q$ на $p(x_q)$, $\bar{f}_q$ на $f(x_q)$ и $\bar{k}_{q\pm1/2}$ на $k(x_{q\pm1/2})$.

В ряде случаев построение разностных схем путем непосредственной аппроксимации производной разностным отношением приводит к недостаточно эффективным разностным схемам. Иногда бывает удобно в окрестности каждого расчетного узла приблизить рассматриваемое уравнение дифференциальным уравнением, интегрируемым в явном виде, и построить разностную схему, точную для его решений.

Рассмотрим дифференциальное уравнение

$$\mu^2 y''(x) + p(x)y(x) = f(x), \tag{20}$$

где μ — малое число; для определенности сначала предполагаем $p(x) > 0$. В случае $p = \text{const}$, $f \equiv 0$ решения этого уравнения $\exp\left\{\pm \mathbf{i}\dfrac{\sqrt{p}x}{\mu}\right\}$ колеблются с периодом $2\pi\mu/\sqrt{p}$, т. е. очень сильно. Характерный размер изменения решения имеет порядок $\mu/\sqrt{p}$, поэтому если не использовать специфику данного уравнения, то для получения высокой точности необходимо выполнение довольно обременительного условия $h \ll \mu/\sqrt{p}$.

В окрестности каждого узла x_n рассматриваемое уравнение близко к уравнению $\mu^2 y'' + p_n y = f_n$, $p_n = p(x_n)$, $f_n = f(x_n)$. Общее решение этого уравнения записывается в виде

$$y(x) = D_1 \exp\left\{\mathbf{i}\frac{\sqrt{p_n}(x - x_n)}{\mu}\right\} + D_2 \exp\left\{-\mathbf{i}\frac{\sqrt{p_n}(x - x_n)}{\mu}\right\} + \frac{f_n}{p_n}; \tag{21}$$

D_1, D_2 — произвольные константы. Найдем схему вида

$$a_n y_{n+1} + b_n y_n + c_n y_{n-1} - d_n = 0, \tag{22}$$

точную на всех решениях вида (21). Для этого подставим (21) в соотношение (22). Получим

$$a_n\left(D_1 \exp\left\{\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\} + D_2 \exp\left\{-\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\}\right) + b_n(D_1 + D_2) +$$
$$+ c_n\left(D_1 \exp\left\{-\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\} + D_2 \exp\left\{\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\}\right) + (a_n + b_n + c_n)\frac{f_n}{p_n} - d_n = 0.$$

Чтобы это равенство выполнялось при всех D_1 и D_2, необходимо и достаточно равенства нулю коэффициентов при D_1 и D_2 и свободного члена. Приравняем их к нулю:

$$\begin{aligned}
&a_n \exp\left\{\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\} + b_n + c_n \exp\left\{-\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\} = 0,\\
&a_n \exp\left\{-\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\} + b_n + c_n \exp\left\{\mathbf{i}\frac{\sqrt{p_n}h}{\mu}\right\} = 0,\\
&(a_n + b_n + c_n)\frac{f_n}{p_n} - d_n = 0.
\end{aligned} \tag{23}$$

Полагая $a_n = 1$, получим

$$c_n = 1, \quad b_n = -2\cos\frac{\sqrt{p_n}h}{\mu}, \quad d_n = \left(2 - 2\cos\frac{\sqrt{p_n}h}{\mu}\right)\frac{f_n}{p_n}.$$

Общее решение системы (23) пропорционально полученному частному решению. Умножим все коэффициенты a_n, b_n, c_n, d_n на $\mu^2 h^{-2}$. Тогда получим схему

$$\mu^2\frac{y_{n+1} - 2\cos\frac{\sqrt{p_n}h}{\mu}y_n + y_{n-1}}{h^2} = \left(2 - 2\cos\frac{\sqrt{p_n}h}{\mu}\right)\frac{\mu^2 f_n}{h^2 p_n} = 0. \qquad (24)$$

Это соотношение мы примем за разностное уравнение, соответствующее узлу x_n. Такая нормировка схемы (24) является наиболее естественной: если вместо y_n подставить в (24) значение $y(x_n)$ и при фиксированном x_n устремить h к 0, то в пределе получится исходное дифференциальное уравнение

$$\mu^2 y'' - p(x_n)y - f(x_n) = 0.$$

Все приведенные выше построения имеют смысл независимо от знака p_n; при $p_n < 0$ в окончательной расчетной формуле имеем

$$\cos\frac{\sqrt{p_n}h}{\mu} = \cos \mathrm{i}\frac{\sqrt{-p_n}h}{\mu} = \operatorname{ch}\frac{\sqrt{-p_n}h}{\mu}.$$

В случае $p_n \equiv \mathrm{const} > 0$ и $f \equiv 0$ расчетная формула (24) переписывается в виде

$$y_{n+1} - 2\cos\frac{\sqrt{p_n}h}{\mu}y_n + y_{n-1} = 0. \qquad (25)$$

То, что эта формула является точной на решениях уравнения (20) при $p_n \equiv \mathrm{const} > 0$ и $f \equiv 0$, можно было бы усмотреть из известной формулы тригонометрии

$$\cos\varphi_1 + \cos\varphi_2 - 2\cos\frac{\varphi_1 + \varphi_2}{2}\cos\frac{\varphi_1 - \varphi_2}{2} = 0,$$

подставив в нее

$$\varphi_1 = \frac{\sqrt{p}(n-1)h}{\mu} + \alpha, \qquad \varphi_2 = \frac{\sqrt{p}(n+1)h}{\mu} + \alpha;$$

α — произвольное число.

Расчетная формула (25) иногда используется для быстрого вычисления таблицы значений $\cos t$ или $\sin t$ на равномерной сетке с невысокой точностью. Необходимость в этом может возникнуть, если, например, решается дифференциальное уравнение $x' = f(x, t)$ с правой частью, содержащей значения $\cos t$ или $\sin t$, и вычисление их значений составляет существенную долю от затрат на вычисление правой части. Заметим, что суммарная вычислительная погрешность при вычислении значений $\cos t$ и $\sin t$ по этим формулам имеет порядок $O(n^2\delta)$ ($\delta = 2^{-t}$ — погрешность округлений).

Рассмотрим уравнение

$$y'(x) = y^2(x) + a^2(x), \tag{26}$$

решения которого могут обращаться в бесконечность. При аналитической функции $a(z)$ решения аналитичны в комплексной плоскости в окрестности вещественной оси и иногда представляет интерес найти значения решения уравнения (26) в некоторой точке вещественной оси, отделенной от исходной несколькими полюсами. Один из возможных путей — это численное интегрирование (26) вдоль некоторой кривой, обходящей особые точки. Другой путь — это построение разностных схем, имеющих высокую точность на особенностях решения.

На отрезке $[x_n, x_{n+1}]$ уравнение (26) приблизим уравнением

$$y'(x) = y^2(x) + a^2_{n+1/2}, \quad a_{n+1/2} = a((x_n + x_{n+1})/2).$$

При $a = \text{const}$ общее решение уравнения $y'(x) = y^2(x) + a^2$ имеет вид $y(x) = a\,\mathrm{tg}(a(x + c))$. Воспользуемся формулой

$$\mathrm{tg}(\varphi + \psi) = \frac{\mathrm{tg}\,\varphi + \mathrm{tg}\,\psi}{1 - \mathrm{tg}\,\varphi\,\mathrm{tg}\,\psi}$$

при $\varphi = a(x_n + c)$, $\psi = a(x_{n+1} - x_n)$ и тем, что $\dfrac{y(x_n)}{a} = \mathrm{tg}\,\varphi$, $\dfrac{y(x_{n+1})}{a} = \mathrm{tg}(\varphi + \psi)$. Получим равенство

$$\frac{1}{a}\,y(x_{n+1}) = \frac{\dfrac{y(x_n)}{a} + \mathrm{tg}(a(x_{n+1} - x_n))}{1 - \dfrac{y(x_n)}{a} + \mathrm{tg}(a(x_{n+1} - x_n))}.$$

Отсюда получаем расчетную формулу для исходного уравнения

$$y_{n+1} = \frac{y_n + a_{n+1/2}\,\mathrm{tg}(a_{n+1/2}(x_{n+1} - x_n))}{1 - \dfrac{y_n}{a_{n+1/2}}\,\mathrm{tg}(a_{n+1/2}(x_{n+1} - x_n))}. \tag{27}$$

За счет некоторого понижения точности ее можно упростить, воспользовавшись приближенным равенством $\mathrm{tg}\,\psi \approx \psi$; получим

$$y_{n+1} = \frac{y_n + a^2_{n+1/2}(x_{n+1} - x_n)}{1 - y_n(x_{n+1} - x_n)}. \tag{28}$$

Обе расчетные формулы (27) и (28) позволяют получить приближение к решению и после прохождения полюсов, если только случайно не оказалось, что расстояние от одного из узлов до ближайшего полюса много меньше, чем $\min\limits_n (x_{n+1} - x_n)^2$. Этого всегда можно избежать, распорядившись выбором шагов вблизи полюса.

Как примеры расчетных формул подобного рода можно рассматривать рекуррентные формулы метода прогонки (4.9), (4.10).

§ 9. Конечно-разностные методы отыскания собственных значений

Рассмотрим простейшую краевую задачу на собственные значения:

$$y''(x) - p(x)y(x) = \lambda\rho(x)y(x), \quad y(0) = 0, \quad y(X) = 0. \tag{1}$$

Зададимся шагом $h = XN^{-1}$ и выпишем сеточную задачу

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = \lambda\rho_n y_n,$$
$$n = 1, \ldots, N-1, \quad y_0 = y_N = 0; \qquad p_n = p(x_n), \quad \rho_n = \rho(x_n). \tag{2}$$

Значения λ, при которых система уравнений (2) имеет ненулевое решение $y_0, \ldots, y_N$, естественно назвать собственными значениями сеточной задачи. Пусть собственные значения задач (1), (2) упорядочены в порядке убывания, т. е. $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$; $\lambda_1^h \geqslant \lambda_2^h \geqslant \ldots$.

Рассмотрим модельный пример: $p(x) \equiv 0$, $\rho(x) \equiv 1$; тогда (1) приобретает вид

$$y''(x) = \lambda y(x), \quad y(0) = y(X) = 0.$$

Можно проверить, что собственные функции этой задачи есть $w^m(x) = \sin(\pi m x/X)$ и соответствующие собственные значения $\lambda_m = -(\pi m/X)^2$. В случае сеточной задачи (2), приобретающей вид

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - \lambda_n y_n = 0, \qquad y_0 = y_N = 0,$$

рассуждаем следующим образом: общее решение разностного уравнения $y_{n+1} - (2 + \lambda h^2)y_n + y_{n-1} = 0$ записывается в виде

$$y_n = C_1\mu_1^n + C_2\mu_2^n,$$

где μ_1, μ_2 — корни характеристического уравнения

$$\mu^2 - (2 + \lambda h^2)\mu + 1 = 0. \tag{3}$$

По формуле Виета $\mu_1\mu_2 = 1$, поэтому $\mu_2 = \mu_1^{-1}$ и $y_n = C_1\mu_1^n + C_2\mu_1^{-n}$. Условия $y_0 = y_N = 0$ дают систему уравнений

$$C_1 + C_2 = 0, \quad C_1\mu_1^N + C_2\mu_2^{-N} = 0;$$

она имеет ненулевое решение, если ее определитель равен 0, а именно, если $\mu_1^{-N} - \mu_1^N = 0$. Отсюда

$$\mu_1 = \exp\{\pi i m/N\}, \quad m = \ldots, -1, 0, 1, \ldots$$

Из (3) можно выразить значения λ^h через значения μ_1:

$$\lambda_m^h = \frac{\mu_1 + \mu_1^{-1} - 2}{h^2} = \frac{\exp\left\{\dfrac{\pi i m}{N}\right\} + \exp\left\{-\dfrac{\pi i m}{N}\right\} - 2}{h^2} =$$
$$= \frac{N^2}{X^2}\left(2\cos\frac{\pi m}{N} - 2\right) = -4\frac{N^2}{X^2}\sin^2\frac{\pi m}{2N}.$$

Для определенности возьмем $C_1 = (2\mathrm{i})^{-1}$; тогда соответствующие собственные функции имеют вид

$$W_n^m = \frac{1}{2\mathrm{i}}\left(\exp\left\{\frac{\pi \mathrm{i} mn}{N}\right\} - \exp\left\{\frac{-\pi \mathrm{i} mn}{N}\right\}\right) = \sin\frac{\pi mn}{N}.$$

Собственные значения $\lambda_1^h, \ldots, \lambda_{N-1}^h$ различны между собой, поэтому соответствующие им собственные функции $W_n^1 = \sin\frac{\pi n}{N}, \ldots, W_n^{N-1} = \sin\frac{\pi(N-1)n}{N}$ также различны. Так как задача (2) является задачей на собственные значения для матрицы размерности $N-1$, то мы получили полную систему собственных функций; каждая из функций W_n^m при $m \leqslant 0$ или при $m \geqslant N$ равна тождественно нулю или пропорциональна одной из перечисленных выше функций $W_n^1, \ldots, W_n^{N-1}$. Мы случайно выбрали такой пример, где в узлах сетки $x_n = nh$ выполняется равенство

$$W_n^m = w^m(nh).$$

В общем случае это равенство не имеет места; однако характер близости собственных значений этих задач типичен и для общего случая. Поскольку, согласно формуле Тейлора, $\cos x = 1 - \frac{x^2}{2} + \cos(\theta x)\frac{x^2}{24}$, где $|\theta| \leqslant 1$, то из выражения для λ_m^n получаем

$$\lambda_m^h = -\frac{N^2}{X^2}\left(\frac{\pi^2 m^2}{N^2} - \frac{\cos\left(\theta_m \dfrac{\pi m}{N}\right)}{12}\left(\frac{\pi m}{N}\right)^4\right), \quad |\theta_m| \leqslant 1,$$

или

$$\lambda_m^h = -\left(\frac{\pi m}{X}\right)^2 + \frac{\cos\theta_m'}{12}\left(\frac{\pi m}{X}\right)^4 h^2, \quad \theta_m' = \theta_m\frac{\pi m}{N}. \tag{4}$$

Из этой формулы видно, что $\lambda_m^h - \lambda_m = O(h^2)$ при фиксированном m; в то же время с ростом m как абсолютная, так и относительная погрешности монотонно возрастают и, например,

$$\frac{\lambda_{N-1}}{\lambda_{N-1}^h} = \frac{-\pi^2(N-1)^2}{-4N^2\sin^2\dfrac{\pi(N-1)}{2N}} \sim \frac{\pi^2}{4}.$$

Равенство (4) можно записать в виде оценки $|\lambda_m^h - \lambda_m| \leqslant C\lambda_m^2 h^2$, где C не зависит от λ_m и h. В случае дважды дифференцируемых функций $p(x)$, $\rho(x)$ также можно получить такую оценку.

Для решения задачи (1) с более высокой точностью можно воспользоваться любыми разностными аппроксимациями уравнения $y''(x) - q(x)y(x) = 0$ более высокой точности. Рассмотрим пример.

В § 1 была построена разностная схема, аппроксимирующая последнее уравнение с погрешностью $O(h^4)$:

$$\frac{\delta^2 y_n}{h^2} - q_n y_n - \frac{1}{12}\delta^2(q_n y_n) = 0 \tag{5}$$

(с несколько иными обозначениями). Уравнение (1) записывается в рассматриваемом виде при $q(x) = p(x) + \lambda\rho(x)$. Отсюда получаем сеточную задачу на собственные значения:

$$\begin{aligned} &\frac{\delta^2 y_n}{h^2} - (p_n + \lambda\rho_n)y_n - \frac{1}{12}\delta^2((p_n + \lambda\rho_n)y_n) = 0, \\ &n = 1, \ldots, N-1, \quad y_0 = y_N = 0. \end{aligned} \tag{6}$$

Можно показать, что для собственных значений этой задачи выполняется оценка

$$|\lambda_m^h - \lambda_m| \leqslant c\lambda_m^3 h^4.$$

Система уравнений (6) имеет вид $A\mathbf{y} - \lambda B\mathbf{y} = \mathbf{0}$, формально несколько более сложный, чем (2), поскольку матрица B не диагональная.

При повышении порядка точности могут возникать сеточные задачи, имеющие на первый взгляд еще более сложный вид, например такая: требуется найти λ, при котором система соотношений

$$\begin{aligned} &\alpha_n(\lambda, h)y_{n-1} - \beta_n(\lambda, h)y_n + \gamma_n(\lambda, h)y_{n+1} = 0, \\ &n = 1, \ldots, N-1, \quad y_0 = y_N = 0, \end{aligned} \tag{7}$$

имеет ненулевое решение y_n.

Рассмотрим случай $\gamma_n \neq 0$. Фиксируем некоторое $w_1 \neq 0$, например $w_1 = h$; зададимся произвольным λ и из сотношения

$$\alpha_n(\lambda, h)w_{n-1}^\lambda - \beta_n(\lambda, h)w_n^\lambda + \gamma_n(\lambda, h)w_{n+1}^\lambda = 0 \tag{8}$$

последовательно определим $w_2^\lambda, \ldots, w_N^\lambda$. Если $w_N^\lambda = 0$, то это λ окажется собственным значением и w_n^λ — собственной функцией; если $w_N^\lambda \neq 0$, то это λ не является собственным значением задачи (7).

Для отыскания собственных значений задачи (7), совпадающих с нулями w_N^λ, можно применить какой-либо итерационный метод отыскания нулей функции и по ее значениям. Этот процесс облегчается следующим обстоятельством, имеющим место во многих случаях и позволяющим получить «вилку» для искомого корня: если функция w_n^λ имеет j перемен знака на $(0, X)$, то $\lambda_j < \lambda < \lambda_{j+1}$. Для вычисления значений w_N^λ при различных λ, как правило, наиболее рационально воспользоваться непосредственно рекуррентными формулами (8).

Предлагаемый алгоритм вычисления значений w_n^λ совпадает с алгоритмом решения сеточной задачи Коши для уравнения $y''(x) = (p(x) + \lambda\rho(x))y(x)$.

Для отыскания собственных значений может применяться также и метод прогонки. Не повторяя идеи метода, ограничимся написанием расчетных формул. Положим $w_n^\lambda / w_{n+1}^\lambda = C_n$, тогда (8) перепишется в виде

$$\alpha_n C_{n-1} C_n - \beta_n C_n + \gamma_n = 0, \tag{9}$$

откуда

$$C_n = \gamma_n / (\beta_n - \alpha_n C_{n-1}). \tag{10}$$

Если w_n^λ и w_{n+1}^λ одного знака, то $C_n > 0$, если разного, то $C_n < 0$. Поэтому, наблюдая за переменами знака у C_n, можно определить число перемен знака у функции w_n^λ. Как мы видели в § 3, коэффициент C_n может оказываться очень большим, поэтому этот метод чаще применяется лишь для отыскания первого собственного значения.

§ 10. Построение численных методов с помощью вариационных принципов

Часто бывает естественно и целесообразно строить численные методы, исходя из естественной постановки задачи как вариационной или пользуясь определением решения как некоторой функции, удовлетворяющей интегральному тождеству.

1. Метод Ритца. Рассмотрим краевую задачу из § 8:

$$\begin{aligned} &Ly \equiv -(k(x)y'(x))' + p(x)y(x) = f(x), \\ &y(0) = a, \quad y(X) = b, \quad k \geqslant k_0 > 0. \end{aligned} \tag{1}$$

Ее решение является точкой экстремума функционала

$$I(y) = \int_0^X (k(y'(x))^2 + py^2(x) - 2f(x)y(x))\, dx \tag{2}$$

на классе функций $W_2^1[0, X]$, удовлетворяющих условию $y(0) = a$, $y(X) = b$. Напомним, что $W_2^1[0, X]$ — это класс функций с ограниченным интегралом

$$I^0(y) = \int_0^X [(y'(x))^2 + y^2(x)]\, dx = \|y\|^2_{W_2^1[0, X]}.$$

Задаются некоторым N и выбирают совокупность функций $\varphi_0^N(x), \ldots, \varphi_N^N(x)$ с ограниченным интегралом $I^0(\varphi_k^N)$, удовлетворяющих условиям

$$\begin{aligned} &\varphi_0^N(0) = a, \quad \varphi_0^N(X) = b, \\ &\varphi_q^N(0) = \varphi_q^N(X) = 0, \quad q = 1, \ldots, N. \end{aligned}$$

Приближенное решение ищется в виде

$$y^N = \varphi_0^N + \sum_{q=1}^{N} c_q \varphi_q^N.$$

Имеем

$$I(y^N) = \sum_{p,\,q=1}^{N} \Lambda(\varphi_p^N,\, \varphi_q^N) c_p c_q - 2\sum_{q=1}^{N} b_q c_q + d_0;$$

здесь

$$d_0 = I(\varphi_0^N),$$

$$\Lambda(\varphi_p^N,\, \varphi_q^N) = \int_0^X \left(k(\varphi_p^N)'(\varphi_q^N)' + p\varphi_p^N \varphi_q^N\right) dx,$$

$$b_q = \int_0^X \left(f\varphi_q^N - p\varphi_0^N \varphi_q^N - k(\varphi_0^N)'(\varphi_q^N)'\right) dx.$$

Находим экстремум функционала $I(y^N)$ по переменным $c_1, \ldots, c_N$ и соответствующую функцию $y^N = \varphi_0^N + \sum_{q=1}^{N} c_q \varphi_q^N$ принимаем за приближенное решение задачи. При этом нахождение коэффициентов c_q сведется к решению системы линейных алгебраических уравнений

$$A\mathbf{c} = \mathbf{b}, \tag{3}$$

где A — матрица с элементами $a_{pq} = \Lambda(\varphi_p^N, \varphi_q^N)$, $\mathbf{b}$ — вектор с компонентами b_q.

Часто бывает удобнее сразу вычесть из решения функцию φ_0^N, удовлетворяющую граничным условиям, т. е. свести исходную задачу к задаче с однородными граничными условиями. В линейном случае (как (1)) обычно φ_0^N берут не зависящим от N. Часто бывает выполнено следующее условие. Краевая задача

$$Ly + \lambda y = 0, \quad y(0) = y(X) = 0$$

имеет только нулевое решение, если $\lambda \geqslant 0$. Тогда функционал $I(y)$ ограничен снизу и искомое решение является не просто точкой экстремума, а точкой минимума функционала $I(y)$. В этом случае описанный выше метод построения приближенного решения называют *методом Ритца.* Существует ряд моментов, существенно влияющих на сходимость метода Ритца.

Чтобы приближенные решения y^N сходились к точному в норме $W_2^1[0, X]$, т. е. чтобы $\|y^N - y\|_{W_2^1} \to 0$ при $N \to \infty$, необходимо и достаточно выполнения следующего условия: для любой функции $g \in W_2^1$ и любого $\varepsilon > 0$ существует линейная комбинация

$$g_N = \varphi_0^N + \sum_{q=1}^{N} c_q \varphi_q^N \quad \text{с} \quad \|g_N - g\|_{W_2^1} \leqslant \varepsilon.$$

Указанное условие обеспечивает сходимость метода Ритца в предположении, что все вычисления производятся точно. Пусть λ_N и λ^N — наименьшее и наибольшее по модулю собственные значения матрицы системы уравнений (3). Чтобы округления не повлияли на приближения y^N, существенно выполнение условия

$$|\lambda^N/\lambda_N| \leqslant M, \tag{4}$$

где M не зависит от N.

Довольно часто не удается построить системы функций, удовлетворяющие условию (4). Тогда ограничиваются использованием систем функций, для которых

$$|\lambda^N/\lambda_N| = O(N^\alpha), \tag{5}$$

где α — не очень большое число. В случаях (4), (5), как правило, удается так организовать процесс решения системы (3), что суммарная вычислительная погрешность будет порядка $O(N^\beta\delta)$.

В ряде случаев нетрудно построить системы функций, удовлетворяющие условию (4), но, как правило, для них матрица A является полностью заполненной. Для задачи (1) такой системой является

$$\varphi_q^N(x) = \sin(\pi qx/X), \quad q = 1, \ldots, N.$$

В то же время для системы функций, соответствующих вариационно-разностному методу (см. далее), $\alpha = \beta = 2$, но зато матрица A трехдиагональная. Для системы функций $\varphi_q^N(x) = x^q(1-x)$ величина $|\lambda^N/\lambda_N|$ растет быстрее любой степени N и матрица заполненная. Если вместо системы функций $\varphi_q^N(x) = x^q(1-x)$ взять систему

$$\varphi_q^N(x) = x(1-x)T_q(2x/X - 1), \tag{6}$$

где $T_q(x)$ — многочлены Чебышева, то при отсутствии округлений получится одно и то же приближение. В то же время система (6) удовлетворяет условию (5) и при практическом использовании накопление погрешности будет не очень большим.

Замечание. Может случиться, что для некоторого набора функций величина $|\lambda^N/\lambda_N|$ растет с ростом N очень быстро, но для достижения нужной точности достаточно небольшого значения N; тогда такой набор функций приемлем при решении данной задачи.

2. Метод Бубнова-Галеркина. Описываемый ниже метод является обобщением метода Ритца и применим в случаях, когда исходная задача не является вариационной. Формально этот метод можно представить следующим образом. Запишем исходную задачу в виде задачи нахождения решения из некоторого интегрального соотношения, справедливого для любой функции ψ из соответствующего класса:

$$(Ly, \psi) = (f, \psi). \tag{7}$$

Под выражением в круглых скобках понимаем скалярное произведение в $L_2[0, X]$. Соотношение (7) в дальнейшем будем называть интегральным тождеством. Приближенное решение ищется в виде линейной комбинации

$$y^N = \varphi_0^N + \sum_{q=1}^{N} c_q \varphi_q^N.$$

Задаются некоторой линейно независимой системой функций $\psi_1^N, \ldots, \psi_N^N$ и требуют выполнения интегральных соотношений

$$(Ly^N, \psi_q^N) = (f, \psi_q^N), \qquad q = 1, \ldots, N. \tag{8}$$

Так же как и в случае метода Ритца, решение исходной задачи сводится к решению системы линейных уравнений (8) относительно неизвестных $c_1^N, \ldots, c_N^N$; в матричной форме система уравнений (8) записывается в виде $A\mathbf{c} = \mathbf{d}$, где $A = [a_{ij}]$ — матрица размерности $N \times N$, $\mathbf{c} = (c_1^N, \ldots, c_N^N)^T$, $\mathbf{d}$ — вектор правой части.

Оба описанных метода применимы в нелинейном случае. Если исходная задача является задачей на экстремум функционала, не являющегося, как (2), квадратичным, то система уравнений (3) относительно $c_1^N, \ldots, c_N^N$, соответствующих точке экстремума $l(y^N)$, будет нелинейной.

Точно так же в случае нелинейного уравнения $L(y) = 0$ метод Бубнова-Галеркина сводится к решению нелинейной системы

$$(L(y^N), \psi_q^N) = 0, \quad q = 1, \ldots, N.$$

3. Вариационно-разностный вариант метода Ритца. *Носителем* $N(f)$ *функции* f назовем замыкание множества точек, где $f \neq 0$. Если носители функций φ_i^N и φ_j^N пересекаются по мере нуль, то $a_{ij} = \Lambda(\varphi_i^N, \varphi_j^N) = 0$. Наличие большого числа нулевых элементов в матрице может привести к существенному уменьшению объема вычислений при решении системы (3). Это обстоятельство явилось стимулом к разработке *вариационно-разностных методов*, соединяющих в себе преимущества метода Ритца и конечно-разностных методов.

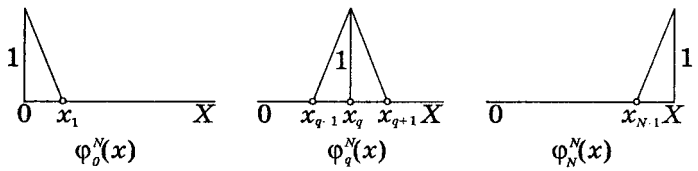


Рис. 9.10.1

Зададимся точками $0 = x_0 < x_1 < \cdots < x_N = X$ и будем отыскивать приближенное решение (1) в виде функции, линейной на каждом из от-

резков $[x_{q-1}, x_q]$ и принимающей заданные значения на концах отрезка $[0, X]$. Это равносильно тому, что решение ищется в виде

$$y_N(x) = \bar{\varphi}_0^N(x) + \sum_{q=1}^{N-1} y_q \varphi_q^N(x), \tag{9}$$

$$\bar{\varphi}_0^N(x) = a\varphi_0^N(x) + b\varphi_N^N(x),$$

$$\varphi_0^N(x) = \begin{cases} \dfrac{x_1 - x}{x_1} & \text{при} \quad 0 \leqslant x \leqslant x_1, \\ 0 & \text{при} \quad x_1 \leqslant x \leqslant x_N, \end{cases}$$

$$\varphi_N^N(x) = \begin{cases} 0 & \text{при} \quad 0 \leqslant x \leqslant x_{N-1}, \\ \dfrac{x - x_{N-1}}{x_N - x_{N-1}} & \text{при} \quad x_{N-1} \leqslant x \leqslant x_N, \end{cases} \tag{10}$$

$$\varphi_q^N(x) = \begin{cases} \dfrac{x - x_{q-1}}{x_q - x_{q-1}} & \text{при} \quad x_{q-1} \leqslant x \leqslant x_q, \\ \dfrac{x_{q+1} - x}{x_{q+1} - x_q} & \text{при} \quad x_q \leqslant x \leqslant x_{q+1}, \\ 0 & \text{в остальных точках} \end{cases}$$

при $q = 1, \ldots, N-1$ (см. рис. 9.10.1).

Система уравнений (3)

$$\partial I(y_N)/\partial y_q = 0, \qquad q = 1, \ldots, N-1, \tag{11}$$

из которой определяются значения $y_1, \ldots, y_{N-1}$, в данном случае оказывается системой с трехдиагональной матрицей. Укажем конкретный вид коэффициентов системы, получающейся в случае функций φ_q^N вида (10):

$$a_{qq} = \int_{x_{q-1}}^{x_q} \left(\frac{k(x)}{(x_q - x_{q-1})^2} + p(x) \frac{(x - x_{q-1})^2}{(x_q - x_{q-1})^2} \right) dx + \\ + \int_{x_q}^{x_{q+1}} \left(\frac{k(x)}{(x_{q+1} - x_q)^2} + p(x) \left(\frac{x_{q+1} - x}{x_{q+1} - x_q} \right)^2 \right) dx, \quad q = 1, \ldots, N-1;$$

$$a_{q,q+1} = \int_{x_q}^{x_{q+1}} \left(\frac{-k(x)}{(x_{q+1} - x_q)^2} + p(x) \frac{(x_{q+1} - x)(x - x_q)}{(x_{q+1} - x_q)^2} \right) dx, \quad q = 1, \ldots, N-2;$$

$$a_{q+1,q} = a_{q,q+1}, \quad q = 2, \ldots, N-2,$$

$$b_q = \int_{x_{q-1}}^{x_q} f(x) \frac{x_q - x}{x_q - x_{q-1}} dx + \int_{x_q}^{x_{q+1}} f(x) \frac{x_{q+1} - x}{x_{q+1} - x_q} dx - \beta_q,$$

где

$$\beta_q = \begin{cases} a\displaystyle\int_0^{x_1} \left(p(x)\frac{x_1 - x}{x_1}\cdot\frac{x}{x_1} - \frac{k(x)}{x_1^2}\right) dx, & q = 1, \\ 0, \quad 1 < q < N-1, \\ b\displaystyle\int_{x_{N-1}}^{x_N} \left(p(x)\frac{x_N - x}{x_N - x_{N-1}}\cdot\frac{x - x_{N-1}}{x_N - x_{N-1}} - \frac{k(x)}{(x_N - x_{N-1})^2}\right) dx, & q = N-1. \end{cases}$$

4. Вариационно-разностный вариант метода Бубнова-Галеркина. Если умножим (1) на произвольную функцию $\psi(x) \in \overset{\circ}{W}{}_2^1([0, X])$ и проинтегрируем первое слагаемое в выражении

$$\int_0^X (L(y) - f(x))\psi(x)\,dx \tag{12}$$

по частям, то получим интегральное тождество

$$\Lambda(y, \psi) = \int_0^X (k(x)y'\psi' + py\psi - f\psi)\,dx = 0. \tag{13}$$

Будем искать решение y_N в виде (9): потребуем, чтобы (13) обращалось в нуль для всех функций $\psi(x)$ вида

$$\psi(x) = \sum_{q=1}^{N-1} \psi_q \varphi_q^N(x),$$

ψ_q — произвольные числа. Поскольку выражение $\Lambda(y, \psi)$ линейно по ψ, то получаем систему уравнений

$$\Lambda(y_N, \psi_q) = 0, \qquad q = 1, \ldots, N-1. \tag{14}$$

В данном случае эта система уравнений совпадает с (3).

Можно было не производить в (12) интегрирования по частям, а непосредственно потребовать удовлетворения (12) для любой функции вида (9). Однако для функций вида (9) выражение (12) содержит слагаемые типа δ-функции, поэтому непосредственное выписывание уравнений системы (14) представляло бы дополнительные технические трудности.

Описанный выше метод решения задачи, называемый *проекционно-разностным*, применим и к задачам более общего вида.

Рассмотрим краевую задачу

$$(k(x)y')' + ay' + (by)' + cy = f + F', \tag{15}$$

уже не являющуюся задачей на экстремум некоторого функционала. Умножим (15) скалярно на $\psi \in \overset{\circ}{W}{}_2^1[0, X]$ и проинтегрируем некоторые из слагаемых по частям; получим

$$\Lambda_1(y, \psi) = \int_0^X (ky'\psi' - ay'\psi + by\psi' - cy\psi + f\psi - F\psi')\,dx = 0.$$

Приближение y_N вида (9) находим из системы уравнений

$$\Lambda_1(y_N, \psi_q) = 0, \qquad q = 1, \ldots, N-1. \tag{16}$$

Заметим, что интегрирование по частям интеграла от $(by)'\psi$ является необходимым лишь в случае разрывного коэффициента b.

Описанные выше методы формально представляют некоторые неудобства, поскольку для построения системы уравнений (16) требуется вычисление некоторых интегралов. Если коэффициенты гладкие, то эти интегралы можно вычислить с помощью квадратурных формул:

$$\begin{aligned}
&\int_{x_{q-1}}^{x_q} k(x)\,dx \approx k\left(\frac{x_{q-1}+x_q}{2}\right)(x_q - x_{q-1}),\\
&\int_{x_{q-1}}^{x_q} p(x)(x_q - x)(x - x_{q-1})\,dx \approx p\left(\frac{x_{q-1}+x_q}{2}\right)\int_{x_{q-1}}^{x_q}(x_q - x)(x - x_{q-1})\,dx =\\
&\qquad = p\left(\frac{x_{q-1}+x_q}{2}\right)\frac{(x_q - x_{q-1})^3}{6},\\
&\int_{x_{q-1}}^{x_q} p(x)(x - x_{q-1})^2\,dx \approx p\left(\frac{x_{q-1}+x_q}{2}\right)\int_{x_{q-1}}^{x_q}(x - x_{q-1})^2\,dx =\\
&\qquad = p\left(\frac{x_{q-1}+x_q}{2}\right)\frac{(x_q - x_{q-1})^3}{3},\\
&\int_{x_{q-1}}^{x_q} f(x)(x - x_{q-1})\,dx \approx f\left(\frac{x_{q-1}+x_q}{2}\right)\frac{(x_q - x_{q-1})^2}{2}.
\end{aligned} \tag{17}$$

Во всех случаях был применен способ вычисления интегралов, при котором подынтегральная функция разлагалась на множители и интеграл от наиболее резко меняющейся функции брался в явном виде. Можно показать, что полученная схема обеспечивает второй порядок точности.

Можно предложить способ построения вариационно-разностных схем более высокого порядка точности; рассмотренный выше способ построения схемы является частным случаем (при $m = 1$) этого способа. Приближение $y_N(x)$ ищется в виде многочлена $P_q^m(x)$ степени m на каждом из отрезков $[x_{q-1}, x_q]$, причем значения многочленов, соответствующих отрезкам $[x_{q-1}, x_q]$ и $[x_q, x_{q+1}]$, совпадают в точке x_q. Минимизируя функционал $I(y_N)$ на множестве таких многочленов, удовлетворяющих граничным условиям в (1), получим систему с клеточно-трехдиагональной матрицей относительно коэффициентов этих многочленов. Иногда удобнее рассматривать как неизвестные не эти коэффициенты, а некоторые другие параметры. Например, можно аналогично случаю приближения сплайнами (§ 4.8) при $m = 3$ принять за неизвестные значения

$$y_N(x_q), \quad y_N''(x_{q+0}), \quad y_N''(x_{q-0}).$$

Можно на каждом из отрезков $[x_{q-1}, x_q]$ взять $m-1$ дополнительную точку и принять за неизвестные значения $y_N(x)$ в этих точках.

Вариационно-разностные методы в определенном смысле «технологичнее» разностных. В случае построения вариационно-разностных методов путем минимизации квадратичного функционала возникает система уравнений с положительно определенной матрицей, что обеспечивает определенную «физичность» получаемых приближений и одновременно облегчает решение системы. Переменность шага интегрирования также не оказывает существенного влияния на сложность программ вариационно-разностных методов. Эти преимущества наиболее эффективно проявляются при решении многомерных задач в областях со сложной геометрией. При одном и том же порядке точности использование вариационно-разностных схем часто требует меньшего объема программирования. Все это послужило причиной того, что они взяты за основу, например, при создании пакетов численных методов решения задач теории упругости. В то же время при решении очень сложных задач, в которых для получения нужной точности требуется число узлов сетки, находящееся на пределе возможностей ЭВМ, часто бывает целесообразно обратиться к сеточным методам.

Заметим, что вариационно-разностные и проекционно-разностные методы называют также *методами конечных элементов.*

5. Построение разностных схем путем аппроксимации функционала. При непосредственном построении разностных аппроксимаций в областях сложного вида иногда оказывается, что получается система уравнений со знаконеопределенной матрицей, в то время как исходная задача была знакоопределена. Чтобы преодолеть этот дефект, не прибегая к использованию вариационно-разностных методов, строят конечно-разностные схемы, используя дискретную аппроксимацию минимизируемого функционала, соответствующего задаче. Приблизим исходный функционал дискретной аппроксимацией:

$$I(y) \approx I_h(y_N) = \sum_{q=1}^{N} k(x_{q-1/2}) \left(\frac{y_q - y_{q-1}}{x_q - x_{q-1}} \right)^2 (x_q - x_{q-1}) + \\ + \sum_{q=1}^{N} \left(\frac{p(x_q) y_q^2 + p(x_{q-1}) y_{q-1}^2}{2} \right) (x_q - x_{q-1}) - \\ - \sum_{q=1}^{N} (f(x_q) y_q + f(x_{q-1}) y_{q-1})(x_q - x_{q-1}); \tag{18}$$

здесь $x_{q-1/2} = (x_q + x_{q-1})/2$.

Первая сумма получилась на основе квадратурной формулы прямоугольников, вторая и третья — формулы трапеций. Приравнивая нулю

производные $\partial f_h/\partial y_q$, получаем систему уравнений

$$2\left[k(x_{q-1/2})\frac{y_q-y_{q-1}}{\Delta x_{q-1}}-k(x_{q+1/2})\frac{y_{q+1}-y_q}{\Delta x_q}\right]+p(x_q)y_q\Delta x_{q-1}+p(x_q)y_q\Delta x_q-$$
$$-f(x_q)(\Delta x_{q-1}+\Delta x_q)=0,\qquad q=1,\ldots,N-1;\quad \Delta x_q=x_{q+1}-x_q.$$

Поделив предыдущее соотношение на $-2\delta_q$, где $\delta_q=(\Delta x_q+\Delta x_{q-1})/2$, получим конечно-разностную схему

$$L_q y_q=\frac{k(x_{q+1/2})\frac{y_{q+1}-y_q}{\Delta x_q}-k(x_{q-1/2})\frac{y_q-y_{q-1}}{\Delta x_{q-1}}}{\frac{\Delta x_q+\Delta x_{q-1}}{2}}-p(x_q)y_q-f(x_q)=0, \qquad (19)$$

совпадающую со схемой (8.3) в случае равномерной сетки.

Выражение $I_h(y_h)$ в (18) является многочленом второй степени от переменных y_i; оно записывается в виде

$$I_h(y_h)=\sum_{i,j=1}^{N-1}a_{ij}y_iy_j+\sum_{i=1}^{N-1}b_iy_i+c;$$

выше учтено то, что $y_0=a$, $y_N=b$. Из (18) видно, что первые две суммы в $I_h(y_h)$ неотрицательны. Поэтому

$$I_h(y_h)\geqslant-\min|f|\left(\sum_{q=0}^{N}|y_q|\delta_q\right).$$

При таком поведении многочлена второй степени в окрестности бесконечности его главная часть $\sum\limits_{i,j=1}^{N-1}a_{ij}y_iy_j$ неотрицательна, т. е. $A=[a_{ij}]\geqslant 0$. Поскольку $a_{ij}=\frac{1}{2}\partial^2 I_h(y_h)/\partial y_i\partial y_j$, то матрица A автоматически оказывается симметричной.

Задача 1. Доказать, что при $k>0$, $p\geqslant 0$ матрица A положительно определена.

Чтобы проведенные рассуждения о неотрицательности матрицы A были справедливы, целесообразно строить аппроксимации функционала, приближая выражения, стоящие под знаком квадрата, не раскрывая скобок.

Пусть, например, исходная задача является задачей на экстремум функционала:

$$I(y)=\int_0^X\left(k(x)(y'+\lambda(x)y)^2+p(x)y^2\right)dx;\quad p\geqslant 0.$$

Интеграл от первого слагаемого приближаем выражением

$$\sum_{q=1}^{N} k(x_{q-1/2}) \left(\frac{y_q - y_{q-1}}{\Delta x_q} + \lambda(x_{q-1/2}) \frac{y_q + y_{q-1}}{2} \right) \Delta x_q,$$

от второго — так же, как в (18). Все проведенные выше рассуждения остаются в силе, и поэтому соответствующая матрица A неотрицательна. Если раскрыть скобки в $(y' + \lambda(x)y)^2$, а потом аппроксимировать интеграл, то может случиться, что условие $A \geqslant 0$ при крупных шагах будет нарушено.

Из последних рассуждений следует заключение, которое особо существенно в многомерном случае.

При построении конечно-разностных методов путем аппроксимации минимизируемого функционала целесообразно записать функционал в виде суммы интегралов от квадратов некоторых выражений и линейной части и аппроксимировать эти выражения, не раскрывая скобок.

6. Случай невариационной задачи. Для невариационных задач разностные схемы можно получать, аппроксимируя интегральное тождество, из которого определяется решение. Применим этот способ к рассматриваемой нами вариационной задаче. Будем аппроксимировать интегральное тождество

$$\Lambda(y, \psi) = \int_0^X (ky'\psi' + py\psi - f\psi)\, dx = 0$$

для любой функции $\psi \in \overset{\circ}{W}{}_2^1 [0, X]$. Имеем

$$\Lambda(y, \psi) \approx \Lambda_h(y_h, \psi_h) = \sum_{q=1}^{N} k(x_{q-1}) \frac{y_q - y_{q-1}}{\Delta x_{q-1}} \frac{\psi_q - \psi_{q-1}}{\Delta x_{q-1}} \Delta x_{q-1} +$$

$$+ \sum_{q=1}^{N} \frac{1}{2} \Big(\big(p(x_q)y_q - f(x_q)\big)\psi_q + \big(p(x_{q-1})y_{q-1} - f(x_{q-1})\big)\psi_{q-1} \Big) \Delta x_{q-1} = 0.$$

Полагая $\psi_0 = \psi_N = 0$ и собирая коэффициенты при одних и тех же ψ_q, получим

$$\Lambda_h(y_h, \psi_h) = \sum_{q=1}^{N-1} \delta_q L_q y_q \cdot \psi_q = 0,$$

где $L_q y_q$ определено формулой (19). Выражение $\Lambda_h(y_h, \psi_h)$ равно нулю для любой сеточной функции ψ, если все $L_q y_q = 0$. Полученная система уравнений совпадает с системой уравнений (19).

§ 11. Улучшение сходимости вариационных методов в нерегулярном случае

Погрешность итерационных методов существенно зависит от точности, с которой можно приблизить решение функциями из пространства, в котором ищется решение. Рассмотрим некоторые задачи, в которых при использовании вариационных методов имеет смысл несколько усложнить построение системы базисных функций.

1. В случае разрывного коэффициента $k(x)$ производная u' также разрывна, поэтому решение плохо приближается кусочно-линейными функциями. В то же время выражение ku' является дифференцируемой функцией. Поэтому для достижения более высокой точности целесообразно искать приближение в виде функции, которая на каждом из отрезков $[x_{q-1}, x_q]$ является решением уравнения $ky' = \text{const}$ или, что то же самое, $(ky')' = 0$. В этом случае получается вариационно-разностная схема, имеющая в сеточной норме $\overset{\circ}{W}{}^{1}_{2,h}$ порядок сходимости $O(h^2)$ даже при измеримых функциях $k(x)$, $p(x)$, $f(x)$. Под нормой $\overset{\circ}{W}{}^{1}_{2,h}$ понимаем выражение

$$\|u_h\|_{\overset{\circ}{W}{}^{1}_{2,h}} = \sqrt{\sum_{n=0}^{N-1} h \left(\frac{u_{n+1} - u_n}{h}\right)^2}, \quad u_0 = u_N = 0,$$

являющееся сеточным аналогом нормы пространства $\overset{\circ}{W}{}^{1}_{2}$.

2. Иногда оказывается, что решение имеет особенность в конечном числе точек, а вдали от них является гладким. Например, относительно решения иногда можно установить, что оно имеет вид

$$y(x) = \sum_{j=1}^{l} C_j \psi_j(x) + u(x),$$

где $\psi_j(x)$ — известные функции, а $u(x)$ — неизвестная гладкая функция. Если некоторые из коэффициентов C_j, например $C_{k+1}, \ldots, C_l$, известны, то следует перейти к новой неизвестной функции

$$y^* = y - \sum_{j=k+1}^{l} C_j \psi_j(x).$$

Далее рассматриваем случай — все C_j неизвестны. Приближение для гладкой части ищем в виде

$$u(x) = \sum_{q=0}^{N} c_q \varphi_q^N(x),$$

где $\varphi_q^N(x)$ те же, что и в (10.9), т. е. решение ищется в виде

$$y(x) = \sum_{j=1}^{l} C_j \psi_j(x) + \sum_{q=0}^{N} c_q \varphi_q^N(x)$$

с неизвестными коэффициентами C_j, c_q и при дополнительном условии на эти коэффициенты, имеющем вид $y(0) = a$, $y(X) = b$. (Здесь для определенности мы приняли, что $\psi_1, \ldots, \psi_l$, $\varphi_0^N, \ldots, \varphi_N^N$ линейно независимы.) Функции $\psi_j(x)$ в отличие от функций φ_q^N обычно имеют носитель, размеры которого не стремятся к нулю при уменьшении шага сетки. Поэтому строки матрицы A, соответствующие функциям ψ_j, как правило, будут полностью заполнены. Матрица системы уже не оказывается трехдиагональной. Перестановкой строк и столбцов ее можно преобразовать к виду, где $a_{ij} = 0$, если одновременно $|i-j| > 1$, $|i|$, $|j| > l+1$. Если решать эту систему методом Гаусса при обратном порядке исключения неизвестных, то общее число арифметических операций оказывается, как и в случае трехдиагональных матриц, порядка $O(N)$. В описываемом случае надо особенно внимательно следить за погрешностью метода решения задачи (включающей в себя погрешность приближенного вычисления интегралов) и вычислительной погрешностью.

Рассмотрим в качестве примера краевую задачу

$$\varepsilon^2 y''(x) - p(x)y(x) = f(x), \quad p(x) > 0,$$
$$y(0) = a, \quad y(X) = b, \quad \varepsilon \text{ — малое.}$$

Формально говоря, решение не имеет особенности. Однако при малом ε имеется пограничный слой, где производные от решения велики и решение плохо приближается функциями вида (10.9). Из теории асимптотических методов известно, что в окрестности точки $x = 0$ решение хорошо приближается линейными комбинациями функций вида $x^k \exp\left\{-\dfrac{\sqrt{p(0)}}{\varepsilon}\, x\right\}$, а в окрестности точки X — функций вида $(X-x)^k \exp\left\{-\dfrac{\sqrt{p(X)}}{\varepsilon}(X-x)\right\}$. Поэтому приближенное решение имеет смысл отыскивать в виде

$$y(x) = \sum_{k=0}^{l} C_k x^k \exp\left\{-\frac{\sqrt{p(0)}}{\varepsilon}\, x\right\} +$$
$$+ \sum_{k=0}^{l} D_k (X-x)^k \exp\left\{-\sqrt{p(X)}\frac{X-x}{\varepsilon}\right\} + \sum_{q=1}^{N} c_q \varphi_n^q(x);$$

здесь C_k, D_k, c_q — неизвестные коэффициенты.

§ 12. Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения

Как было установлено ранее, вычислительная погрешность имеет различный характер роста для различных способов решения дифференциальных уравнений. Рассмотрим теперь такой частный, но важный вопрос: как зависит вычислительная погрешность от формы записи конечно-разностных уравнений? Хотя все изложение ведется на примере задачи Коши, проводимые соображения относятся в равной мере и к случаю решения краевых задач. Для примера обратимся к методу Эйлера:

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{1}$$

При реальных вычислениях будут получаться величины y_n^*, связанные соотношением

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n; \tag{2}$$

наличие слагаемого δ_n является следствием ряда причин — погрешностей при вычислении значений функции $f(x_n, y_n^*)$, погрешностей при округлении произведения $hf(x_n, y_n^*)$ и погрешностей при сложении чисел y_n^* и округленного значения $hf(x_n, y_n^*)$.

Введем обозначение $y_n^* - y_n = \Delta_n$; на основании формулы Лагранжа имеем

$$f(x_n, y_n^*) - f(x_n, y_n) = l_n \Delta_n,$$

где $l_n = f_y(x_n, \bar{y}_n)$. Предположим, что всегда $|f_y(x, y)| \leqslant L$. Вычитая (1) из (2), получаем

$$\Delta_{n+1} = (1 + l_n h)\Delta_n + \delta_n,$$

откуда

$$|\Delta_{n+1}| \leqslant (1 + Lh)|\Delta_n| + \delta, \quad \delta = \max_n \delta_n. \tag{3}$$

Рассмотрим разностное уравнение

$$z_{n+1} = (1 + Lh)z_n + \delta,$$

являющееся *мажорирующим* для (3). Его решение при начальном условии $z_0 = |\Delta_0|$ есть

$$z_n^0 = |\Delta_0|(1 + Lh)^n + \delta \frac{(1 + Lh)^n - 1}{Lh}.$$

Лемма. *При всех* $n \geqslant 0$ *справедливо*

$$|\Delta_n| \leqslant z_n^0. \tag{4}$$

Доказательство. При $n = 0$ утверждение (4) очевидно. Пусть оно верно для некоторого n; тогда имеем

$$|\Delta_{n+1}| \leqslant (1 + Lh)z_n^0 + \delta = z_{n+1}^0.$$

Лемма доказана.

Если интегрирование производится на отрезке $[x_0, x_0 + X]$, то

$$nh \leqslant X, \quad (1 + Lh)^n \leqslant (\exp\{Lh\})^n \leqslant \exp\{LX\}.$$

Согласно формуле Лагранжа при $y \geqslant 0$ имеем $e^y - 1 = ye^{\theta y} \leqslant ye^y$, где $0 \leqslant \theta \leqslant 1$. Отсюда получаем

$$(1 + Lh)^n - 1 \leqslant e^{LX} - 1 \leqslant LX \exp\{LX\}.$$

В итоге имеем оценку

$$|\Delta_n| \leqslant |z_n^0| \leqslant |\Delta_0| \exp\{LX\} + \delta h^{-1} X \exp\{LX\}.$$

Рассмотрим случай $\Delta_0 = 0$. Тогда погрешность $\Delta_n = y_n^* - y_n$ при фиксированном X оценивается сверху через $O(\delta h^{-1})$.

Эта оценка неулучшаема по порядку. Например, при $\Delta_0 = 0$, $f_y \equiv 0$, $\delta_n = \delta$ имеем $\Delta_{n+1} = \Delta_n + \delta$ и, таким образом,

$$\Delta_N = N\delta = \delta X h^{-1}.$$

Приведем некоторые рассуждения, из которых следует, что погрешность округления может оказаться величиной порядка δh^{-1}. Соотношение

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n$$

можно переписать в виде

$$y_{n+1}^* = y_n^* + hf^*(x_n, y_n^*),$$

где

$$f^*(x_n, y_n^*) = f(x_n, y_n^*) + \delta_n h^{-1}.$$

Таким образом, результат численного интегрирования уравнения $y' = f(x, y)$ при наличии округлений будет такой же, как если бы без округлений интегрировалось уравнение со значениями правой части в узлах сетки $f^*(x_n, y_n^*)$. Рассмотрим случай $\delta_n \equiv \delta$. Разность между решениями дифференциальных уравнений

$$y' = f(x, y) \qquad \text{и} \qquad y' = f(x, y) + \delta h^{-1}$$

имеет порядок разности между правыми частями этих уравнений, т. е. δh^{-1}. Нет оснований ожидать, что решения разностных уравнений $y_{n+1} = y_n + hf(x_n, y_n)$ и $y_{n+1}^* = y_n^* + h(f(x_n, y_n^*) + \delta h^{-1})$ будут отличаться на величину, существенно меньшую δh^{-1}. Если δ порядка 2^{-t}, t — разрядность чисел в ЭВМ, то решение разностного уравнения изменится на величину порядка $2^{-t}h^{-1}$.

При получении этого вывода было сделано допущение $\delta_n \equiv \delta$. Рассмотрим задачу вычисления интеграла $\int_0^1 f(x)\,dx$, являющуюся частным случаем рассматриваемой задачи при $y(0) = 0$. Пусть $f(x) = 2/3$, $h = 2^{-k}$, $t-k$ — нечетно. При $x_n > 3/4$ значение y_n лежит в пределах $(1/2, 1)$. Тогда после округления при сложении

$$y_n = 0,\ 1\alpha_2 \ldots \alpha_t,$$

$$hf_n = 0, \underbrace{0\ldots0}_{k}\ \underbrace{10\ldots101}_{t-k}\ \underline{0101}\ldots$$

каждый раз происходит отбрасывание подчеркнутой величины, равной $(1/3)2^{-t}$. Таким образом, на этом участке действительно $\delta_n \equiv \delta$ порядка 2^{-t}.

Перейдем к случаю интегрирования уравнения $y'' = f(x, y)$ при помощи простейшего метода

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} = f(x_n, y_n)$$

по расчетной формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 f(x_n, y_n).$$

Реально получаемые значения y_n^* связаны соотношением

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 f(x_n, y_n^*) + \delta_n. \tag{5}$$

Значения y_n^* можно рассматривать как получаемые без округлений при вычислениях по формуле

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 f^*(x_n, y_n^*),$$

где $f^*(x_n, y_n^*) = f(x_n, y_n^*) + \delta_n h^{-2}$.

Рассмотрим случай $\delta_n \equiv \delta$. Тогда решения дифференциальных уравнений $y'' = f(x, y)$ и $y'' = f(x, y) + \delta h^{-2}$ различаются между собой на величину порядка δh^{-2}.

Задача 1. Показать, что в случае уравнения $y'' = \alpha$, $\alpha = \text{const}$, возможен случай $\delta_n \equiv \delta$, δ порядка 2^{-t}.

Как и для уравнения первого порядка, делаем вывод, что для рассматриваемого алгоритма суммарная вычислительная погрешность может оказаться величиной порядка $2^{-t}h^{-2}$. Требование малости этой величины накладывает ограничение снизу на допустимый шаг интегрирования.

Рассмотрим случай, когда такая величина вычислительной погрешности оказывается недопустимо большой. Можно было бы записать рассматриваемое уравнение в виде системы уравнений первого порядка

$$y' = v, \quad v' = f(x, y) \tag{6}$$

и применить какой-либо метод численного интегрирования этой системы. Как уже отмечалось, при этом произошла бы потеря эффективности,

поскольку методы, применимые для систем общего вида, не учитывают специфики этой системы.

Попытаемся записать рассматриваемую расчетную формулу как некоторую расчетную формулу интегрирования системы (6). Введем новую дискретную переменную

$$\frac{y_n - y_{n-1}}{h} = z_n,$$

тогда уравнение (5) запишется в виде

$$\frac{z_{n+1} - z_n}{h} = f(x_n, y_n).$$

Вычисления последовательных значений y_n, z_n будем производить при помощи пары расчетных формул

$$z_{n+1} = z_n + hf(x_n, y_n), \quad y_{n+1} = y_n + hz_{n+1}. \tag{7}$$

При наличии округлений соответственно имеем

$$z^*_{n+1} = z^*_n + hf(x_n, y^*_n) + \alpha_n, \quad y^*_{n+1} = y^*_n + hz^*_{n+1} + \beta_n,$$

где α_n, $\beta_n = O(2^{-t})$. Эти соотношения можно представить в виде

$$\begin{aligned} &z^*_{n+1} = z^*_n + hf^*(x_n, y^*_n), \quad f^*(x_n, y^*_n) = f(x_n, y^*_n) + \alpha_n h^{-1}, \\ &y^*_{n+1} = y^*_n + h(z^*_{n+1} + g(x_{n+1})), \quad g(x_{n+1}) = \beta_n h^{-1}. \end{aligned} \tag{8}$$

Если формулы (7) можно трактовать как формулы численного интегрирования системы $z' = f(x, y)$, $y' = z$, то формулы (8) соответствуют системе

$$z' = f^*(x, y), \qquad y' = z + g(x).$$

Правые части этих систем различаются на величины порядка $O(2^{-t}h^{-1})$; поэтому есть какие-то основания ожидать, что и решения разностных задач, т. е. решения разностной задачи с округлениями и разностной задачи без округлений, будут различаться на величину того же порядка.

Задача 2. Доказать справедливость утверждения, сформулированного выше.

Конечно, к этому заявлению следует отнестись с осторожностью; мы уже видели, что для некоторых конечно-разностных схем малые погрешности могут приводить к катастрофическому изменению результата.

Приведенные рассуждения о влиянии вычислительной погрешности в конкретных методах интегрирования уравнений первого и второго порядков опираются лишь на учет свойств конечно-разностной схемы, связанных с порядком дифференциального уравнения. Поэтому они переносятся на другие конечно-разностные методы. Например, при интегрировании уравнения $y^{(k)} = f(x, y)$ и прямом использовании схемы

$$\nabla^k y_n - h^k \sum_{i=0}^{m} a_{-i} f(x_{n-i}, y_{n-i}) = 0$$

следует ожидать влияния вычислительной погрешности порядка $2^{-t}h^{-k}$. Таким образом, в случае уравнений высокого порядка еще более актуальна задача преобразования схемы к форме, где влияние вычислительной погрешности будет меньше. Например, аналогично случаю $k=2$ целесообразно ввести вспомогательные переменные

$$z_n^i = \nabla^i y_n h^{-i}, \qquad i = 1, \ldots, k-1.$$

Попытаемся объяснить улучшение свойств разностной схемы (5) при переходе к расчетным формулам (7), оценивая количество хранимой информации. При использовании расчетной формулы (5) при каждом n в памяти ЭВМ хранятся величины y_{n-1} и y_n, и все дальнейшие значения y_j определяются по этим значениям. Пусть для определенности $1/2 \leqslant y_{n-1},\ y_n \leqslant 1$ и $|y_n - y_{n-1}| \leqslant Mh$. В ячейке, содержащей значение $y_{n-1} = 0{,}1\alpha_2 \ldots \alpha_t$, имеется $t-1$ независимых двоичных знаков $\alpha_2, \ldots, \alpha_t$; разряды числа $y_n = 0{,}1\beta_2 \ldots \beta_t$ уже не все несут новую информацию.

Дело заключается в следующем. Пусть l наибольшее целое, такое, что $Mh < 2^{-l+1}$. Тогда имеем

$$y_n - y_{n-1} = \pm 0{,}0 \ldots 0\gamma_{l+1} \ldots \gamma_t,$$

и для задания $y_n - y_{n-1}$, а следовательно, и y_n достаточен $t-l+1$ двоичный знак (знак разности и $\gamma_{l+1}, \ldots, \gamma_t$). Поскольку $l \sim \log_2((Mh)^{-1})$, то общее количество независимой информации, которое имеется в нашем распоряжении при каждом n, составляет $2t - \log_2((Mh)^{-1})$ двоичных разрядов (с точностью до слагаемого, не зависящего от t и h).

В случае вычислений по формуле (7) все разряды чисел y_n и z_n независимы и поэтому информация о решении задается независимыми двоичными разрядами.

Тот факт, что количество независимой информации для второго способа больше, конечно, не означает, что этот способ лучше. Не исключено, что эта дополнительная информация не является содержательной и поэтому не позволяет точнее определить решение. Поясним, почему дополнительная информация при втором способе является содержательной.

Для определения решения дифференциального уравнения второго порядка с точностью $O(\varepsilon)$ требуется задание с такой же по порядку точностью значения решения и его производной в некоторой точке. Для разностного уравнения роль этих величин играют y_n и $(y_n - y_{n-1})/h$. При первом способе задание значений y_n и y_{n-1} с t двоичными знаками позволяет определить $(y_n - y_{n-1})/h$ с погрешностью порядка $O(2^{-t}h^{-1})$. Таким образом, здесь по известной нам информации мы располагаем возможностью найти дальнейшие значения решения сеточной задачи с погрешностью порядка $O(2^{-t}h^{-1})$ (если все последующие вычисления будут производиться абсолютно точно). В случае второго способа мы имеем значения y_n и $(y_n - y_{n-1})/h$ с t двоичными знаками, поэтому обладаем возможностью найти решение сеточной задачи с погрешностью порядка $O(2^{-t})$. Таким образом, эта дополнительная информация действитель-

но оказывается содержательной. На каждом шаге реального численного интегрирования погрешности округления вносят дополнительную неопределенность в компоненты вектора $(y_n, (y_n - y_{n-1})/h)$: во втором случае порядка $O(2^{-t})$, в первом — порядка $O(2^{-t}/h)$. Это и приводит к тому, что суммарная вычислительная погрешность решения во втором случае может оказаться величиной порядка $O(2^{-t}h^{-1})$, а в первом — порядка $O(2^{-t}h^{-2})$.

Рассмотрим метод прогонки решения сеточной краевой задачи (1.3), соответствующей уравнению второго порядка $y''(x) - p(x)y(x) = f(x)$ при $p(x) \equiv p = \text{const}$.

Коэффициенты C_n вычисляются по следующим рекуррентным формулам $C_{n+1} = (2+ph^2-C_n)^{-1}$; в случае $p(x) \equiv p$ целесообразно заранее вычислить $Q = 2 + ph^2$ и вести вычисления по формуле $C_{n+1} = (Q - C_n)^{-1}$. При вычислении суммы $2 + ph^2$ произойдет округление, т. е. получится величина $Q^* = 2 + ph^2 + \delta$, где δ может оказаться величиной порядка 2^{-t}. Это равносильно тому, что без округлений решается уравнение $y''(x) - p^*y(x) = f(x)$, где $p^* = p + \delta h^{-2}$. Рассуждая, как и выше, получим, что такое возмущение коэффициента p может привести к возмущению решения на величину порядка $2^{-t}h^{-2}$.

Если коэффициент C_n существенно больше 1, то при вычислении выражения $ph^2 + 2 - C_n$ погрешность округления может оказаться величиной порядка $2^{-t}h^{-2}|C_n|$. Поскольку вклад от погрешности в одной точке в суммарную погрешность умножается на коэффициент порядка h, то влияние этого округления на окончательный результат порядка $2^{-t}h^{-1}|C_n|$. Если $|C_n| \gg h^{-1}$, то это выражение будет существенно больше чем $2^{-t}h^{-2}$.

Замечание. Возмущения, вносимые другими округлениями при вычислениях C_n и φ_n, также равносильны некоторым возмущениям коэффициентов p и f.

Чтобы погрешность решения системы (1.1) была существенно меньше, необходимо по крайней мере задавать ее в форме, где округления коэффициентов системы равносильны существенно меньшим возмущениям коэффициентов исходной дифференциальной задачи. С этой целью можно, например, перейти к системе

$$\frac{y_n - y_{n-1}}{h} = z_n, \qquad \frac{z_{n+1} - z_n}{h} - p_n y_n = f_n. \tag{9}$$

Соответственно при решении этой системы вместо рекуррентных соотношений (3.11) относительно коэффициентов C_n, φ_n следует перейти к рекуррентным соотношениям (4.9), (4.10) относительно α_n, β_n.

Выше рассматривался случай, когда при всех n погрешность округления оказывалась одной и той же. Если коэффициент $p(x) \not\equiv \text{const}$, то при вычислении величины $2 + p_n h^2$ при различных n округления могут оказаться различных знаков, и поэтому суммарная погрешность может оказаться по порядку меньшей чем $2^{-t}h^{-2}$. В случае задачи Коши для

уравнения $y' = f(x, y)$ при условиях $h = 2^{-t}$, $2^{-t}h^{-2} \ll 1$ вычислительная погрешность часто накапливается медленнее — как $\delta h^{-1/2}$.

Обратим внимание на прием практической оценки вычислительной погрешности путем изменения масштабов, применяемый иногда для экспериментального исследования чувствительности метода к вычислительной погрешности. Пусть некоторым методом решается задача Коши

$$\frac{dy}{dx} = f(x, y), \qquad y(0) = a.$$

Замена переменных $x = \mu t$, $y = \lambda z$ сводит эту задачу к задаче

$$\frac{dz}{dt} = \frac{\mu}{\lambda} f(\mu t, \lambda z), \qquad z(0) = \frac{a}{\lambda}.$$

Предположим, что первая задача интегрировалась с шагами h_i; осуществим численное интегрирование второй задачи тем же методом, но с шагами $h'_i = h_i/\mu$.

При отсутствии округлений будет иметь место равенство $y_i \equiv \lambda z_i$; если λ и μ оба не являются целыми степенями двойки, то разность между реально получаемыми значениями величин y_i и λz_i обычно дает определенное представление о величине вычислительной погрешности. Например, можно взять $\mu = \sqrt{3}$, $\lambda = \sqrt{2}$.

Литература

1. Бахвалов Н. С. Численные методы — М.: Наука, 1975.
2. Бахвалов Н. С. К оптимизации методов решения краевых задач при наличии пограничного слоя // ЖВМиМФ. 1969, **9**, N 4. С. 841–859.
3. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Дифференциальные уравнения — Минск: Наука и техника, 1982.
4. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 2 — М.: Наука, 1977.
5. Самарский А. А., Тихонов А. Н. Об однородных разностных схемах // ЖВМиМФ. — 1961. — **1**, N 1. С. 5–63.
6. Соболев С. Л. Некоторые замечания о численном решении интегральных уравнений // Изв. АН СССР, сер. матем. — 1956. **20**, N 4. С. 413–436.
7. Федорова О. А. Вариационно-разностная схема для одномерного уравнения диффузии // Матем. заметки — 1975. — **17**, N 6. С. 893–898.

Глава 10

Методы решения уравнений в частных производных



В случае решения обыкновенных дифференциальных уравнений мы видели следующую картину. Имеется ряд уравнений, интегрируемых в квадратурах. Решение большинства уравнений можно получить, используя лишь численные методы. Принципиально различных постановок задач довольно немного: в основном это задача Коши, краевая задача для линейных уравнений, краевая задача для нелинейных уравнений. Имеется небольшое количество теоретически исследованных и практически отработанных алгоритмов, позволяющих эффективно решать существенную часть задач, связанных с численным решением обыкновенных дифференциальных уравнений. В частности, ряд численных методов решения задачи Коши был разработан еще в прошлом веке.

В настоящее время разработка методов и алгоритмов решения задачи Коши для обыкновенных дифференциальных уравнений продвинута настолько, что зачастую исследователь, имеющий дело с этой задачей, не занимается выбором метода ее решения, а просто обращается к стандартной программе.

В случае уравнений с частными производными число принципиально различных постановок задач существенно больше. В курсе уравнений с частными производными обычно рассматривается незначительная часть таких постановок, главным образом связанных с линейными уравнениями с постоянными коэффициентами. При этом существует очень малое количество задач, решаемых в явном виде. Многообразие постановок в теории уравнений с частными производными связано с многообразием явлений окружающего нас мира.

Большое количество различных постановок задач, связанных с решением уравнений в частных производных, привело к тому, что теория численных методов в этом направлении дробится на большое число направлений. Использование численных методов с применением ЭВМ сильно расширило возможности в исследовании подобных задач. Например, разработанные за последние пятьдесят лет алгоритмы дают возможность решать с приемлемой затратой машинного времени подавляющее большинство краевых задач, связанных с решением одно- и многомерных уравне-

ний параболического типа с переменными коэффициентами, в частности с коэффициентами, нелинейно зависящими от решения.

Конечно, здесь иначе, чем в случае обыкновенных дифференциальных уравнений, обстоит дело с обоснованием сходимости численных методов и оценкой погрешности. Для широких классов типовых задач такие исследования проведены. Однако для многих важных классов прикладных задач, предъявляемых математикам для решения, не только не доказан, но часто остается неясным сам факт существования решения.

Например, уже для простой на вид системы

$$v_t + u_x = 0, \qquad \left((\operatorname{sign}\gamma)v^{\gamma}\right)_x = 0, \tag{1}$$

описывающей одномерное адиабатическое течение газа в лагранжевых координатах, не доказан при $|\gamma| > 1$ факт существования решения в целом (т. е. на неограниченном промежутке времени).

При таком состоянии вопроса о существовании решения в настоящее время трудно ожидать получения строгих оценок погрешности и теорем сходимости сеточных методов при достаточно общих предположениях. Тем не менее, часто пользуясь полуэмпирическими соображениями, аналогиями по сравнению со случаем линейных уравнений и численными экспериментами на задачах с известным точным решением, математики создают численные методы решения и для таких задач. При этом результаты и анализ численных расчетов наравне с экспериментом оказывают существенное влияние на развитие соответствующих разделов теории уравнений с частными производными. Так, например, обстоит дело в случае решения уравнений газовой динамики типа (1).

Несмотря на отсутствие строгих обоснований чисто математической (в частности, алгоритмической) стороны вопроса, математикам, занимающимся решением подобных прикладных задач, часто приходится брать на себя ответственность за достоверность получаемых численных результатов, включая правильность математической постановки задачи.

Конечно, все сказанное не умаляет роли чисто теоретических исследований. Их результаты, в частности теоремы существования, дают уверенность в правильности постановки, подсказывают информацию о качественных свойствах решения, что крайне важно при выборе алгоритма. Наличие известных частных решений, например в задачах газовой динамики, позволяет производить проверку точности предлагаемых методов. Использование известных частных решений простейших задач часто позволяет облегчить численное решение более сложных задач.

§ 1. Основные понятия теории метода сеток

На первых этапах практического решения задач для уравнений с частными производными применялись в основном вариационные и другие методы, где приближенное решение получается в виде некоторой аналитической формулы. При решении некоторых задач такие методы применяются и в настоящее время.

В последующий период наиболее актуальными для решения стали задачи ядерной проблематики и задачи динамики газа и жидкости, где подобные методы практически неприменимы. На решение этих задач были направлены усилия крупнейших математиков, что имело, в частности, своим следствием создание и широкое продвижение сеточных методов решения уравнений с частными производными. В настоящее время эти методы наряду с вариационно- и проекционно-разностными (метод конечных элементов) являются наиболее распространенными. При решении задач сеточными методами мы получаем совокупность приближенных значений решения в некоторой конечной системе точек. В случае необходимости можно построить формулу (например, интерполяционную) для приближенного представления решения во всей области.

Рассмотрим простейший пример решения задачи сеточным методом. Пусть в полуполосе $0 < x < 1$, $0 < t < \infty$ решается уравнение

$$u_t - u_{xx} = f(x, t) \tag{1}$$

при начальном и граничных условиях

$$u(x, 0) = \varphi(x), \quad u_x(0, t) + a(t)u(0, t) = b(t), \quad u(1, t) = 0. \tag{2}$$

Зададимся некоторыми шагами сетки $h, \tau > 0$ ($1/h = M$ — целое). Точки $(mh, n\tau)$ назовем *узлами сетки* (m, n); пусть u_m^n — приближения к значениям $u(mh, n\tau)$,

$$f_m^n = f(mh, n\tau), \quad a^n = a(n\tau), \quad b^n = b(n\tau),$$

u_h и f_h — функции, определенные на сетке, со значениями u_m^n и f_m^n в узлах сетки. Если искомое решение исходной дифференциальной задачи есть гладкая функция, то выполняются соотношения

$$\begin{aligned} L_h u(mh, n\tau)|_{(m,n)} &\equiv \frac{u(mh, (n+1)\tau) - u(mh, n\tau)}{\tau} - \\ &- \frac{u((m+1)h, n\tau) - 2u(mh, n\tau) + u((m-1)h, n\tau)}{h^2} = \\ &= u_t(mh, n\tau) - u_{xx}(mh, n\tau) + O(h^2 + \tau) = f(mh, n\tau) + O(h^2 + \tau), \end{aligned} \tag{3}$$

$$\begin{aligned} l_h u(mh, n\tau)|_{(0,n)} &= \frac{u(h, n\tau) - u(0, n\tau)}{h} + a(n\tau)u(0, n\tau) = \\ &= u_x(0, n\tau) + a(n\tau)u(0, n\tau) + O(h) = b(n\tau) + O(h). \end{aligned} \tag{4}$$

Исходя из этого можно сделать предположение, что решение системы

$$L_h u_h|_{(m,n)} = f_m^n, \quad 0 < m < M, \quad 0 \leqslant n, \tag{5}$$

$$l_h u_h|_{(0,n)} = b^n, \quad u_M^n = 0, \quad n > 0, \tag{6}$$

$$u_m^0 = \varphi(nh), \quad 0 \leqslant m \leqslant M, \tag{7}$$

является приближением к точному решению исходной задачи. Значения решения системы (5)–(7) можно находить последовательно при каждом n следующим образом: u_m^0 нам заданы, при каждом n величина $u_M^n = 0$, значения u_m^n при $0 < m < M$ находим из (5), а затем u_0^n — из (6).

Докажем теорему, показывающую, что сходимость приближенного решения сеточной задачи к решению дифференциальной при $\tau, h \to 0$ не обязательно имеет место. Пусть в области D с границей $\Gamma = \bigcup_{i=1}^{s} \Gamma_i$ решается краевая задача

$$L(u) = f \tag{8}$$

при граничных условиях

$$l_i(u) = \varphi_i \quad \text{на} \quad \Gamma_i, \qquad i = 1, \ldots, s. \tag{9}$$

Фиксируем какую-либо точку P в пространстве независимых переменных. Значение решения $u(P)$ зависит от коэффициентов уравнения, правой части f и функций φ_i.

Рассмотрим случай, когда уравнение (8) и граничные условия (9) линейные, и изучим вопрос о зависимости $u(P)$ от значений функции φ_1.

> В случае нелинейного уравнения или исследования зависимости от коэффициентов уравнения, которая является нелинейной, проводимые далее рассуждения требуют для своего обоснования некоторых дополнительных формальных построений.

Пусть Φ_1 — некоторый класс функций φ_1. *Областью зависимости* $\Omega_1(P, \Phi_1)$ значения $u(P)$ по граничному условию $l_1 u = \varphi_1$ в классе функций Φ_1 назовем множество точек $Q \in \Gamma_1$, удовлетворяющих следующему соотношению: для любой окрестности точки Q найдутся две функции $\varphi_1^1, \varphi_1^2 \in \Phi_1$ такие, что

1) $\varphi_1^1 = \varphi_1^2$ вне этой окрестности;

2) если u_1 и u_2 — решения задачи (8), (9) при одних и тех же $f, \varphi_2, \ldots, \varphi_s = 0$ и при $\varphi_1 = \varphi_1^1$ и $\varphi_1 = \varphi_1^2$ соответственно, то $u_1(P) \neq u_2(P)$.

При решении задачи (8), (9) методом сеток решение ищется в узлах некоторой сетки в пространстве независимых переменных. Будем считать, что сетка задается некоторым параметром h, причем $h \to 0$ при

измельчении сетки. Если точка P является узлом сетки, то при данном h решение сеточной задачи, аппроксимирующей исходную дифференциальную задачу, зависит от значений φ_1 в некоторых точках границы Γ_1. Множество этих точек обозначим через $\Omega_2(P, h)$.

Теорема Куранта (об областях зависимости). *Для того, чтобы для некоторых $f, \varphi_2, \ldots, \varphi_s$ и всех $\varphi_1 \in \Phi_1$ при $h \to 0$ значения в точке P решений сеточной задачи стремились к значению в точке P решения дифференциальной задачи, необходимо выполнение следующего условия: все точки множества $\Omega_1(P, \Phi_1)$ являются предельными для точек множеств $\Omega_2(P, h)$, иначе говоря,*

$$\Omega_1(P, \Phi_1) \subset \overline{\lim_{h\to 0}}\, \Omega_2(P, h).$$

Доказательство. Предположим противное, т. е. что существует точка $Q \in \Omega_1(P, \Phi_1)$, $Q \notin \overline{\lim_{h\to 0}}\, \Omega_2(P, h)$. Тогда имеется окрестность точки Q, не содержащая точек множеств $\Omega_2(P, h)$ при $h < h_0$. Возьмем функции $\varphi_1^1, \varphi_1^2 \in \Phi_1$, соответствующие этой окрестности согласно приведенному выше определению области зависимости. Пусть u^1 и u^2 — решения задачи (8), (9) при $\varphi_1 = \varphi_1^1$, $\varphi_1 = \varphi_1^2$ соответственно и $f, \varphi_2, \ldots, \varphi_s = 0$. Пусть u^0 — решение задачи (8), (9) при $\varphi_1 = 0$ и заданных $f, \varphi_2, \ldots, \varphi_s$. Тогда $u_1 = u^1 + u^0$, $u_2 = u^2 + u^0$ будут решениями задачи (8), (9) при $\varphi_1 = \varphi_1^1$ и $\varphi_1 = \varphi_1^2$ соответственно и тех же $f, \varphi_2, \ldots, \varphi_s$. Значения решений сеточных задач при соответствующих u_1 и u_2 начальных и граничных условиях будут совпадать при $h < h_0$. Поскольку в то же время $u_1(P) \neq u_2(P)$, то для одного из этих наборов граничных условий решение сеточной задачи не сходится к решению дифференциальной. Теорема доказана.

Рассмотрим пример применения этой теоремы. Пусть в полуплоскости $t \geqslant 0$ решается задача Коши для уравнения $u_t + au_x = 0$ при начальном условии $u(0, x) = u_0(x)$. Зададимся сеткой с узлами в точках $(mh, n\tau)$ и заменим исходную дифференциальную задачу разностной:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^n - u_m^n}{h} = 0, \quad u_m^0 = u_0(mh).$$

Тогда значения u_m^n при $n > 0$ определяем последовательно из соотношения

$$u_m^{n+1} = (1 + a\tau/h)u_m^n - (a\tau/h)u_{m+1}^n.$$

Пусть при измельчении сетки $\tau/h = \kappa = \text{const}$. Тогда значения решения сеточной задачи в точке (x_0, t_0) не зависят от начальных значений условий вне отрезка $[x_0, x_0 + \kappa^{-1}t_0]$. Точное решение дифференциальной задачи есть $u(x, t) = u_0(x - at)$. Поэтому в классе начальных условий, обладающих некоторым ограниченным числом производных, областью зависимости для дифференциальной задачи является точка $x_0 - at_0$. Таким образом, на основании теоремы Куранта необходимым условием сходимости в

этом классе начальных условий является условие $x_0 - at_0 \in [x_0, x_0 + \kappa^{-1}t_0]$ или, иначе, $a \leqslant 0$, $|a\tau/h| = |a|\kappa \leqslant 1$.

Заметим, что в случае рассматриваемой схемы это условие случайно оказывается и достаточным условием сходимости.

Кроме вопроса о сходимости при анализе разностных аппроксимаций, возникает проблема анализа устойчивости получаемого результата относительно погрешностей в исходных данных задачи и при округлениях.

Проиллюстрируем сказанное на этом же примере. Пусть $a > 0$, $a\tau/h = 1$. Тогда значения u_m^n определяются из рекуррентного соотношения

$$u_m^{n+1} = 2u_m^n - u_{m+1}^n.$$

При $u_m^0 \equiv 0$ решением сеточной задачи будет $u_m^n \equiv 0$; пусть теперь $u_0^0 = \varepsilon$ и $u_m^0 = 0$ при $m \neq 0$. Пользуясь этим рекуррентным соотношением, получаем, что тогда в узлах сетки u_m^n принимает следующие значения:

$$u_m^n = \begin{cases} 0 & \text{при} \quad m > 0 \quad \text{или} \quad m < -n, \\ C_n^{-m} 2^{n+m} (-1)^m \varepsilon & \text{при} \quad -n \leqslant m \leqslant 0. \end{cases}$$

Отсюда $\sum\limits_{m=-\infty}^{\infty} |u_m^n| = 3^n\varepsilon$, т. е. с ростом n разность между этим решением и решением $u_m^n \equiv 0$ катастрофически возрастает; поэтому при $a\tau/h = 1$ рассматривая схема не может быть признана пригодной для решения задачи в случае большого числа шагов также и вследствие большого влияния вычислительной погрешности.

Таким образом, при замене решения дифференциальной задачи решением ее разностной аппроксимации возникают следующие проблемы (аналогичные проблемам, возникавшим ранее при рассмотрении методов решения других задач):

1) сходится ли точное решение разностной задачи к решению дифференциальной;

2) насколько сильно изменяется решение разностной задачи, если при вычислениях допускаются некоторые погрешности.

Построим формальный математический аппарат, помогающий при решении этих проблем.

Будем рассматривать задачу (8), (9). Относительно Γ_i будем считать, что Γ_i — заданные части границы Γ, причем различные Γ_i могут иметь непустое пересечение; l_i — некоторые операторы; $f, \varphi_1, \ldots, \varphi_s$ — заданные функции.

Определим некоторое множество точек в пространстве независимых переменных D_h, которое назовем *сеткой* (как правило, D_h выбирают так, чтобы оно принадлежало замкнутой области $\bar{D}$). Точки множества D_h называют *узлами сетки*. Обычно сетка, на которой отыскивается решение, зависит от нескольких параметров (в предыдущем примере от τ

и h); однако во многих типичных случаях при дроблении сетки ее шаги связывают между собой каким-то законом вида $\tau = Ah^\alpha$. Поэтому в дальнейших общих построениях и определениях для простоты мы указываем зависимость только от одного параметра $h > 0$.

Пусть U_h — пространство функций u_h, определенных в узлах сетки D_h, L_h — оператор, преобразующий функции из U_h в функции, определенные на некотором множестве $D_h^0 \subset D_h$; будем предполагать, что $D_h^0 \subset \bar{D}$. Множество функций, определенных в точках D_h^0, будем обозначать через F_h. Для аппроксимации граничных условий (9) выбираются некоторые множества $\Gamma_{ih} \subset \Gamma_i$ и в точках этих множеств определяются значения некоторых операторов над пространством функций U_h. Пусть Φ_{ih} — пространство функций, определенных в точках множеств Γ_{ih}. Если $X \subset Y$ и функция v определена на множестве Y, то ее *следом на множестве* X называют функцию, определенную на множестве X и совпадающую там с v. Если функция v определена на некотором множестве, содержащем D_h, то ее след на этом множестве будем обозначать $[v]_h$.

Пусть U — пространство, к которому мы относим решение задачи (8), (9); F — пространство правых частей f; Φ_i — пространства функций, определенных на Γ_i. Пусть в пространствах функций U, U_h, F, F_h, Φ_i, Φ_{ih} определены нормы

$$\|\cdot\|_U, \quad \|\cdot\|_{U_h}, \quad \|\cdot\|_F, \quad \|\cdot\|_{F_h}, \quad \|\cdot\|_{\Phi_i}, \quad \|\cdot\|_{\Phi_{ih}}.$$

Эти нормы называют *согласованными*, если при $h \to 0$ для любых достаточно гладких функций $u \in U$, $f \in F$, $\varphi_i \in \Phi_i$ выполняются соотношения

$$\|[u]_h\|_{U_h} \to \|u\|_U; \quad \|[f]_h\|_{F_h} \to \|f\|_F; \quad \|[\varphi_i]_h\|_{\Phi_{ih}} \to \|\varphi_i\|_{\Phi_i}.$$

Говорят, что сеточная функция u_h *сходится* к решению задачи (8), (9), если

$$\|u_h - [u]_h\|_{U_h} \to 0 \quad \text{при} \quad h \to 0.$$

Исследование сходимости разностных аппроксимаций имеет смысл производить лишь в нормах, согласованных с некоторыми нормами в пространствах гладких функций. Если отказаться от требования согласованности норм, то условие сходимости иногда может перестать быть содержательным: в случае любой последовательности сеточных функций u_h путем введения некоторого множителя, достаточно быстро убывающего при $h \to 0$, в определение нормы можно добиться, чтобы эта последовательность сходилась к решению задачи u.

Рассмотрим некоторую сеточную задачу

$$L_h(u_h) = f_h, \tag{10}$$

$$l_{ih}(u_h) = \varphi_{ih}, \qquad i = 1, \ldots, s. \tag{11}$$

Говорят, что сеточная задача (10), (11) *аппроксимирует* дифференциальную задачу (8), (9), если выполняется следующее условие: при любых гладких u, f и φ_i

$$z(h) = \| L_h([u]_h) - [L(u)]_h \|_{F_h} + \| f_h - [f]_h \|_{F_h} + \\ + \sum_{i=1}^{s} \left(\| l_{ih}([u]_h) - [l_i(u)]_h \|_{\Phi_{ih}} + \| \varphi_{ih} - [\varphi_i]_h \|_{\Phi_{ih}} \right) \to 0, \quad \text{если} \quad h \to 0. \tag{12}$$

Проиллюстрируем приведенные определения на примере рассмотренной аппроксимации уравнения теплопроводности. Через D будем обозначать множество точек $0 < x < 1$, $0 < t \leqslant T$; пусть Γ_1 — отрезок $[0, 1]$ оси x, Γ_2 — полуинтервал $(0, T]$ оси t, Γ_3 — полуинтервал $(0, T]$ прямой $x = 1$; точки $(0, 0)$, $(1, 0)$ можно было бы отнести и к множествам Γ_2, Γ_3 соответственно. Если $1/h = M$, $N = [T/\tau]$ — целые, то через D_h обозначим множество точек $(mh, n\tau)$ — узлов (m, n), удовлетворяющих условиям $0 \leqslant m \leqslant M$, $0 \leqslant n \leqslant N$. Определим сеточный оператор L_h соотношением

$$L_h u_h|_{(m,\,n)} = \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}. \tag{13}$$

Тогда множество D_h^0 будет состоять из узлов (m, n) таких, что $0 < m < M$, $0 \leqslant n < N$; в остальных узлах (m, n) при $u_h \in U_h$ значения $L_h u_h$ не будут определены. Если правую часть (13) обозначить как $L_h u_h|_{(m,\,n+1)}$, то множество D_h^0 будет состоять из узлов (m, n) таких, что $0 < m < M$, $0 < n \leqslant N$. Правую часть сеточной задачи выберем в виде

$$f_h|_{(m,\,n)} = f(mh, n\tau).$$

Тогда величина $\|f_h - [f]_h\|_{F_h}$, входящая в выражение $z(h)$, есть нуль. Это соотношение выполняется не для всех схем; например, из соображения повышения точности иногда разумнее было бы полагать правую часть разностной задачи в точке (mi, n) равной $f(mh, (n + 0{,}5)\tau)$. В качестве согласованных норм $\|\cdot\|_{U_h}$ и $\|\cdot\|_U$ при исследовании этой задачи обычно выбирают нормы

$$\|u_h\|_{U_h} = \sup_{0 \leqslant n \leqslant N,\, 0 \leqslant m \leqslant M} |u_m^n|, \\ \|u\|_U = \sup_{0 \leqslant t \leqslant T,\, 0 \leqslant x \leqslant 1} |u(x, t)| \tag{14}$$

или же нормы

$$\|u_h\|_{U_h} = \sup_{0\leqslant n\leqslant N} \sqrt{h\sum_{m=0}^{M} |u_m^n|^2},$$
$$\|u\|_U = \sup_{0\leqslant t\leqslant T} \sqrt{\int_0^1 |u(x,\,t)|^2\,dx}. \tag{15}$$

В дальнейшем для простоты изложения будем предполагать, что операторы L, l_i, L_h, l_{ih} линейные.

В этом случае вводится следующее определение *устойчивости (корректности) сеточной задачи (10), (11). Эту задачу называют устойчивой, если при $h \leqslant h_0$ существуют постоянные M_0 и M_i, не зависящие от h, такие, что*

$$\|u_h\|_{U_h} \leqslant M_0\|L_h u_h\|_{F_h} + \sum_{i=1}^{s} M_i \|l_{ih}u_h\|_{\Phi_{ih}}. \tag{16}$$

Как видно из определения, в случае линейной задачи в определение устойчивости не входят функции f_h и φ_{ih}.

Посмотрим, какой смысл в этом определении. Для случая линейных задач *разностная схема* (10), (11) представляет собой систему линейных алгебраических уравнений. Поэтому из (16) следует, что при $f_h \equiv 0$, $\varphi_{ih} \equiv 0$ система уравнений (10), (11) имеет лишь нулевое решение; поэтому на основании теоремы Кронекера-Капелли задача (10), (11) разрешима при любых правых частях f_h, φ_{ih}. Таким образом, в случае линейной задачи из условия устойчивости следует однозначная разрешимость системы сеточных уравнений при любых правых частях.

Если u_h^1 и u_h^2 — решения сеточных задач

$$L_h u_h^1 = f_h^1, \quad l_{ih}u_h^1 = \varphi_{ih}^1, \quad i = 1, \ldots, s,$$
$$L_h u_h^2 = f_h^2, \quad l_{ih}u_h^2 = \varphi_{ih}^2, \quad i = 1, \ldots, s,$$

то при линейных L_h и l_{ih} согласно (16) можно написать

$$\|u_h^1 - u_h^2\|_{U_h} \leqslant M_0\|L_h u_h^1 - L_h u_h^2\|_{F_h} + \sum_{i=1}^{s} M_i \|\, l_{ih}u_h^1 - l_{ih}u_h^2 \,\|_{\Phi_{ih}} =$$
$$= M_0\|\, f_h^1 - f_h^2 \,\|_{F_h} + \sum_{i=1}^{s} M_i \|\, \varphi_{ih}^1 - \varphi_{ih}^2 \,\|_{\Phi_{ih}}. \tag{17}$$

Таким образом, в случае выполнения условия устойчивости решения сеточной задачи мало различаются друг от друга при малом изменении правых частей уравнения и граничных условий.

Пусть $u \in U$. Величину $r_h^0 = L_h[u]_h - f_h$ называют *погрешностью аппроксимации уравнения на решении задачи*, а величины $r_h^i = l_{ih}[u]_h - \varphi_{ih}$ — *погрешностями аппроксимации граничных условий на решении задачи*. Положим

$$\rho_0(h) = \| L_h[u]_h - f_h \|_{F_h}, \quad \rho_i(h) = \| l_{ih}[u]_h - \varphi_{ih} \|_{\Phi_{ih}}.$$

Если u — решение задачи (8), (9), то величину $\rho(h) = \sum_{i=0}^{s} \rho_i(h)$ называют *мерой погрешности аппроксимации разностной схемы (10), (11) дифференциальной задачи (8), (9) на решении*. Если $\rho(h) \to 0$ при $h \to 0$ и u — решение (8), (9), то говорят, что (11), (12) *аппроксимирует* (8), (9) на решении задачи. Порядок величины $\rho(h)$ при $h \to 0$ называют *порядком аппроксимации* на решении.

Выше обсуждалась проблема чувствительности реально получаемого приближенного решения сеточной задачи к округлениям в процессе вычисления этого решения, или, иначе, проблема устойчивости приближенного решения сеточной задачи к погрешностям округления. Решение этой проблемы тесно связано с решением вопроса об устойчивости разностной задачи. Дело в том, что округления, допускаемые при вычислениях, можно рассматривать как возмущения коэффициентов сеточной задачи.

Найдем связь между аппроксимацией, устойчивостью и сходимостью. Предположим, что сеточная аппроксимация (10), (11) удовлетворяет следующим условиям:

1) решение дифференциальной задачи удовлетворяет точно $(s-k)$-сеточным граничным условиям

$$l_{ih}[u]_h = \varphi_{ih}, \quad i = k+1, \ldots, s, \qquad \text{т. е. } \rho_i(h) = 0, \quad i = k+1, \ldots, s;$$

2) на классе функций из U_h, удовлетворяющих однородным граничным условиям

$$l_{ih} u_h = 0, \qquad i = k+1, \ldots, s,$$

выполняется условие устойчивости

$$\|u_h\|_{U_h} \leqslant M_0 \| L_h u_h \|_{F_h} + \sum_{i=1}^{k} M_i \| l_{ih} u_h \|_{\Phi_{ih}}.$$

Теорема Филиппова (о связи устойчивости, аппроксимации и сходимости). *При сформулированных выше условиях выполняется неравенство*

$$\| u_h - [u]_h \|_{U_h} \leqslant \sum_{i=0}^{k} M_i \rho_i(h). \tag{18}$$

Если разностная задача аппроксимирует дифференциальную, то

$$\| u_h - [u]_h \|_{U_h} \to 0 \quad \textit{при} \quad h \to 0.$$

Доказательство. Поскольку $l_{ih}(u_h - [u]_h) = 0$ при $i = k+1, \ldots, s$, то воспользуемся условием 2), подставив в него $u_h - [u]_h$ вместо u_h. Имеем

$$\| u_h - [u]_h \|_{U_h} \leqslant M_0 \| L_h u_h - L_h [u]_h \|_{F_h} + \sum_{i=1}^{k} M_i \| l_{ih} u_h - l_{ih} [u]_h \|_{\Phi_{ih}};$$

подставляя сюда $L_h u_h = f_h$, $l_{ih} u_h = \varphi_h$ и воспользовавшись определением $\rho_i(h)$, получаем (18). Если имеет место аппроксимация, т. е. $\rho_i(h) \to 0$, $i = 0, \ldots, k$, $\rho(h) \to 0$ при $h \to 0$, то из (18) следует справедливость второго утверждения теоремы: $\|u_h - [u]_h\|_{U_h} \to 0$.

В случае гладких решений исследование аппроксимации схемы на решении является относительно несложной задачей и теорема Филиппова переносит центр тяжести на исследование устойчивости сеточной задачи.

Часто случается, что сеточная задача устойчива в одной норме, согласованной с некоторой дифференциальной нормой, но неустойчива в другой. Так может, например, обстоять дело в случае норм, определяемых равенствами (14), (15). В случае гладких решений для практической приемлемости схемы обычно достаточно устойчивости в какой-либо согласованной норме. В случае разрывных решений к разностным аппроксимациям часто предъявляются некоторые дополнительные требования относительно поведения их решений вблизи мест разрыва решений; в этих случаях часто недостаточно устойчивости в произвольной согласованной норме. Например, требование устойчивости в определенных нормах предъявляется в отношении аппроксимаций задач газовой динамики.

Если выполняется условие согласования, то при гладких u, переходя к пределу в (16) при $h \to 0$, получаем неравенство

$$\|u\|_U \leqslant M_0 \|Lu\|_F + \sum_{i=1}^{s} M_i \| l_i u \|_{\Phi_i}. \tag{19}$$

Из этого соотношения следует корректность постановки дифференциальной задачи (8), (9). Такой путь — получение оценок (16), а из них оценок (19) — используется для исследования корректности дифференциальных задач вида (8), (9), для доказательства существования и единственности их решений.

§ 2. Аппроксимация простейших гиперболических задач

Изучение многих важных прикладных задач требует численного решения краевых задач для систем уравнений с частными производными гиперболического типа. Такими системами являются, например, системы уравнений газовой динамики, которые являются квазилинейными системами.

Для решений таких систем типично наличие разрывов. В таких задачах в большинстве случаев отсутствует строгое исследование вопросов устойчивости разностных схем и сходимости, и реальный отбор разностных схем производится на примере простейших модельных задач, где его проще осуществить и теоретически, и путем численного эксперимента.

Простейшими примерами, на которых производился отбор конечно-разностных методов решения задач газовой динамики, являются уравнения

$$u_t + au_x = 0 \tag{1}$$

и

$$u_t + \big(\varphi(u)\big)_x = 0. \tag{2}$$

Далее будут рассмотрены некоторые явные аппроксимации уравнения (1), а затем и (2).

При практическом анализе разностных аппроксимаций задачи Коши для гиперболических и параболических уравнений часто руководствуются следующим критерием, называемым *спектральным признаком устойчивости.*

Пусть на сетке с узлами (m, n) — точками $(x_m, t_n) = (mh, n\tau)$ — построена некоторая разностная схема, например

$$L_h u_h|_{m,n} = \sum_{l,k} a_{lk} u_{m+l}^{n+k} = 0; \tag{3}$$

выпишем все частные решения уравнения $L_h u_h = 0$, имеющие вид

$$u_m^n = (\lambda(\varphi))^n e^{\mathrm{i}m\varphi}.$$

Спектральный признак устойчивости (СПУ). *Если при заданном законе стремления шагов τ и h к нулю существует $C < \infty$, не зависящее от τ и h, такое, что*

$$|\lambda(\varphi)| \leqslant e^{C\tau} \quad \textit{для любых } \varphi, \tag{4}$$

то разностная схема может быть применена для численного решения соответствующей задачи Коши. В противном случае от применения разностной схемы следует воздержаться.

В разумности СПУ можно убедиться, решив следующие задачи.

Задача 1. Пусть определена какая-то норма $\|\cdot\|_n$ на сеточном слое по времени такая, что номер слоя не входит в определение нормы. Этому условию удовлетворяют, например, нормы

$$\|u_m^n\|_n = \sup_m |u_m^n|, \tag{5}$$

$$\|u_m^n\|_n = \sqrt{\sum_m |u_m^n|^2}, \tag{6}$$

но не удовлетворяет норма

$$\|u_m^n\|_n = n \sup_m |u_m^n|.$$

Пусть схема двухслойная и явная, т. е. имеет вид

$$L_h u_h|_{(m,n)} \equiv \alpha u_m^{n+1} + \sum_{|j|\leqslant k} a_j u_{m+j}^n = 0, \quad \alpha \neq 0, \tag{7}$$

и пусть СПУ не выполнен. Доказать, что ни для какого $T > 0$ нельзя указать $Q < \infty$ такое, что при всех $n\tau \leqslant T$ выполнено соотношение $\|u_m^n\|_n \leqslant Q\|u_m^0\|_0$.

Задача 2. Пусть $\tau/h = \text{const}$, разностная схема (3) двухслойная и явная, т. е. имеет вид (7), и начальные условия финитные, т. е. $u_m^n = 0$ при $|m| \geqslant M$. Доказать справедливость оценки

$$\|u_m^n\|_n \leqslant e^{Cn\tau}\|u_m^0\|_0 \leqslant e^{CT}\|u_m^0\|$$

при $n\tau \leqslant T$ в случае нормы (6).

Если краевая задача корректна и условие (4) выполнено, то, как правило, удается построить аппроксимацию граничного условия так, чтобы сеточная задача была устойчива (корректна). В то же время можно привести примеры задач Коши (для систем уравнений), где спектральный признак устойчивости выполнен, а сеточная задача не удовлетворяет условию устойчивости (1.16).

Исследуем при помощи спектрального признака устойчивость ряда сеточных аппроксимаций задачи Коши для уравнения $u_t + au_x = 0$ в полуплоскости $t \geqslant 0$.

Пример 1. Разностная аппроксимация

$$L_h u_h|_{(m,n)} = \frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0. \tag{8}$$

Подставляем сюда $u_m^n = \lambda^n(\varphi)\exp\{\mathbf{i}m\varphi\}$:

$$\frac{\lambda^{n+1}(\varphi)\exp\{\mathbf{i}m\varphi\} - \lambda^n(\varphi)\exp\{\mathbf{i}m\varphi\}}{\tau} +$$
$$+ a\,\frac{\lambda^n(\varphi)\exp\{\mathbf{i}(m+1)\varphi\} - \lambda^n(\varphi)\exp\{\mathbf{i}(m-1)\varphi\}}{2h} = 0.$$

После сокращения на $\lambda^n(\varphi)\exp\{\mathbf{i}m\varphi\}$ получим

$$\lambda(\varphi) = 1 - \frac{a\tau}{2h}(e^{\mathbf{i}\varphi} - e^{-\mathbf{i}\varphi}) = 1 - \mathbf{i}a\frac{\tau}{h}\sin\varphi \tag{9}$$

(см. рис. 10.2.1). На рис. 10.2.1–10.2.4 изображаются наборы узлов (*шаблоны*), по которым выписываются аппроксимации, и кривые, которые описывает точка $\lambda(\varphi)$ на комплексной плоскости при

изменении φ в пределах $0 \leqslant \varphi \leqslant 2\pi$; стрелка означает направление изменения $\lambda(\varphi)$ при изменении φ от 0 до 2π (при $a > 0$); на рис. 10.2.1–10.2.4 цифрой 1 обозначен единичный круг.

Мы имеем $|\lambda(\varphi)| = \sqrt{1 + (a^2\tau^2/h^2)\sin^2\varphi}$,

$$\max_{0\leqslant\varphi\leqslant 2\pi} |\lambda(\varphi)| = \left|\lambda\left(\frac{\pi}{2}\right)\right| = \sqrt{1 + (a^2\tau^2/h^2)}.$$

Если $\tau = Ah^2$ при $\tau, h \to 0$, то

$$\max_{0\leqslant\varphi\leqslant 2\pi} |\lambda(\varphi)| = \sqrt{1 + a^2A^2\tau} = 1 + a^2A^2\tau/2 + O(\tau^2)$$

и условие (4) выполняется; в этом случае следует ожидать устойчивости. Если $\lim\limits_{\tau,h\to 0} \tau/h^2 = \infty$, то

$$\lim_{\tau,h\to 0} \left(\left|\lambda\left(\frac{\pi}{2}\right)\right| - 1\right) \Big/ \tau = \infty$$

и условие (4) не выполняется ни при каком C; тогда сеточная аппроксимация неустойчива.

Эта аппроксимация практически не употребляется вследствие более жесткого по сравнению с другими схемами ограничения на шаг $\tau = O(h^2)$, необходимого для устойчивости, и сильного роста (как $\exp\{a^2A^2n\tau/2\}$) возмущения решения.

Пример 2. Разностная аппроксимация:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_m^n - u_{m-1}^n}{h} = 0 \quad \text{при} \quad a > 0. \tag{10}$$

Аналогично (9) получаем

$$\lambda(\varphi) = 1 - a\tau/h + a\tau/h\exp\{-\mathbf{i}\varphi\}.$$

Если $0 \leqslant a\tau/h \leqslant 1$, то (см. рис. 10.2.2) $|\lambda(\varphi)| \leqslant 1 - a\tau/h + a\tau/h = 1$ и следует ожидать устойчивости. Если $a\tau/h = \kappa > 1$ при $\tau, h \to 0$, то $\lim\limits_{\tau,h\to 0} \lambda(\pi) = 1 - 2\kappa < -1$ и аппроксимация неустойчива вследствие СПУ.

Аналогично показывается, что следует ожидать устойчивости схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^n - u_m^n}{h} = 0 \tag{11}$$

при $a < 0$ и $|a|\tau/h \leqslant 1$.

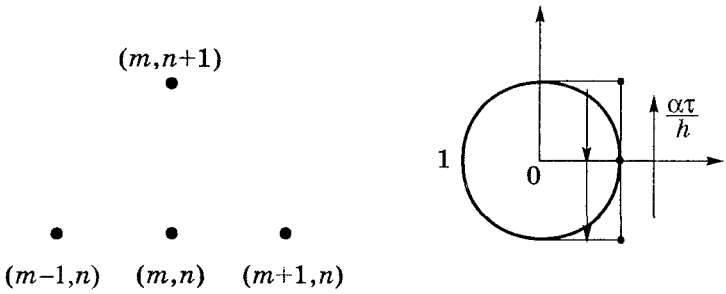
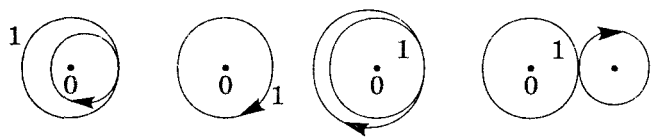
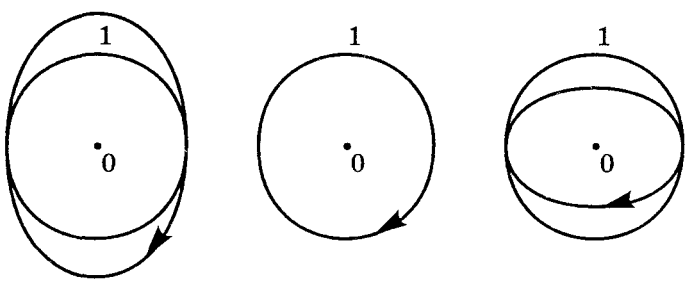


Рис. 10.2.1



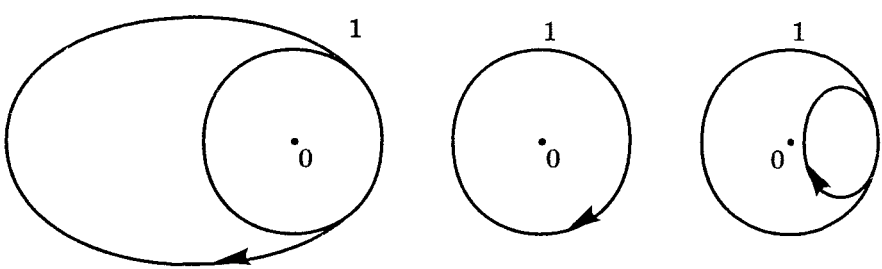
$0 < \frac{\alpha\tau}{h} < 1$ $\frac{\alpha\tau}{h} = 1$ $\frac{\alpha\tau}{h} > 1$ $\frac{\alpha\tau}{h} < 0$

Рис. 10.2.2



$\frac{\alpha\tau}{h} > 1$ $\frac{\alpha\tau}{h} = 1$ $0 < \frac{\alpha\tau}{h} < 1$

Рис. 10.2.3



$\frac{\alpha\tau}{h} > 1$ $\frac{\alpha\tau}{h} = 1$ $0 < \frac{\alpha\tau}{h} < 1$

Рис. 10.2.4

Пример 3. Разностная аппроксимация:

$$\frac{u_m^{n+1}-u_m^n}{\tau}+a\frac{u_{m+1}^n-u_{m-1}^n}{2h}-\frac{h^2}{2\tau}\frac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2}=0, \qquad (12)$$

$$\lambda(\varphi)=\left(\frac{1}{2}-\frac{a\tau}{2h}\right)e^{\mathbf{i}\varphi}+\left(\frac{1}{2}+\frac{a\tau}{2h}\right)e^{-\mathbf{i}\varphi}=\cos\varphi-\frac{a\tau\mathbf{i}}{h}\sin\varphi.$$

При $|a\tau/h|\leqslant 1$ имеем (рис. 10.2.3)

$$|\lambda(\varphi)|\leqslant\left(\frac{1}{2}-\frac{a\tau}{2h}\right)+\left(\frac{1}{2}+\frac{a\tau}{2h}\right)=1$$

и следует ожидать устойчивости. Если же $|a\tau/h|=\kappa>1$ при $\tau,\ h\to 0$, то $\lim\limits_{\tau,h\to 0}\left|\lambda\left(\frac{\pi}{2}\right)\right|=|a\tau/h|=\kappa>1$ и аппроксимация неустойчива. Заметим, что эту аппроксимацию можно рассматривать как (8) с добавленной в нее для устойчивости «вязкостью»:

$$\frac{h^2}{2\tau}\frac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2}\sim\frac{h^2}{2\tau}u_{xx}.$$

Пример 4. Аппроксимация «тренога». Приведенные выше аппроксимации имеют первый порядок по совокупности τ и h. Построим аппроксимацию второго порядка; для простоты будем отправляться от аппроксимации (8). Подставляя разложение Тейлора $u(x,t)$ в точке $(mh,\ n\tau)$, имеем

$$(L_h[u]_h)\Big|_{(m,n)}=\tau u_{tt}(mh,\ n\tau)/2+O(h^2)+O(\tau^2).$$

Из дифференциального уравнения (1) получаем $u_{tt}=a^2u_{xx}$. Приблизим a^2u_{xx} выражением $a^2(\delta_x^2u_h)_{m,n}$. Тогда соответствующая аппроксимация $L_h^{(1)}u_h=0$ примет вид

$$\frac{u_m^{n+1}-u_m^n}{\tau}+a\frac{u_{m+1}^n-u_{m-1}^n}{2h}-\frac{a^2\tau}{2}\frac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2}=0. \qquad (13)$$

Так же как и ранее, получаем

$$\lambda(\varphi)=1-\mathbf{i}(a\tau/h)\sin\varphi+(a^2\tau/h^2)(\cos\varphi-1).$$

Если мы положим $\lambda(\varphi)=x+\mathbf{i}y$, то (см. рис. 10.2.4) можно написать

$$\frac{(x-(1-(a\tau/h)^2))^2}{(a\tau/h)^4}+\frac{y^2}{(a\tau/h)^2}=1,$$

т. е. точки $\lambda(\varphi)$ лежат на некотором эллипсе в комплексной плоскости, расположенном симметрично относительно оси $y=0$.

Поскольку $\lambda(\pi)=1-2a^2\tau^2/h^2<-1$ при $|a\tau/h|>1$, то в случае $|a\tau/h|=\mathrm{const}>1$ спектральное условие устойчивости не выполнено.

При $|a\tau/h| \leqslant 1$ имеем $(a^2\tau^2/h^2)^2 \leqslant a^2\tau^2/h^2$; поэтому можно написать цепочку соотношений

$$|\lambda(\varphi)|^2 = \left(1 + \frac{a^2\tau^2}{h^2}(\cos\varphi - 1)\right)^2 + \frac{a^2\tau^2}{h^2}\sin^2\varphi =$$

$$= 1 + \frac{a^2\tau^2}{h^2}(2\cos\varphi - 2 + \sin^2\varphi) + \left(\frac{a^2\tau^2}{h^2}\right)^2 (\cos\varphi - 1)^2 \leqslant$$

$$\leqslant 1 + \frac{a^2\tau^2}{h^2}(2\cos\varphi - 2 + \sin^2\varphi) + \frac{a^2\tau^2}{h^2}(\cos\varphi - 1)^2 =$$

$$= 1 + \frac{a^2\tau^2}{h^2}(2\cos\varphi - 2 + \sin^2\varphi + \cos^2\varphi - 2\cos\varphi + 1) \equiv 1.$$

Таким образом, при $|a\tau/h| \leqslant 1$ выполнено спектральное условие устойчивости.

Сделаем ряд общих замечаний.

1. Если $a\tau/h = \kappa$ при $\tau, h \to 0$, то для всех выписанных аппроксимаций выполнялось соотношение

$$\lambda(\varphi) = \exp\{-\mathbf{i}\kappa\varphi\} + O(\varphi^{r+1}),$$

где r — порядок аппроксимации по τ и h. Можно показать, что это условие является необходимым для того, чтобы имела место аппроксимация r-го порядка точности.

2. Если $|a\tau/h| = 1$ и $a > 0$, то все рассмотренные аппроксимации имеют вид

$$u_m^{n+1} - u_{m-1}^n = 0.$$

Поскольку решение задачи есть $u = u_0(x - at)$, то в этих случаях они абсолютно точные; при этом $\lambda(\varphi) = \exp\{-\mathbf{i}(a\tau/h)\varphi\} = e^{-\mathbf{i}\varphi}$.

3. В тех случаях, когда аппроксимации (10)–(13) некорректны вследствие СПУ, их можно было бы также забраковать при помощи теоремы Куранта.

4. В остальных случаях, т. е. для (10) при $0 \leqslant a\tau/h \leqslant 1$, для (11) при $-1 \leqslant a\tau/h \leqslant 0$ и для (12), (13) при $|a\tau/h| \leqslant 1$, их устойчивость следует из решения задачи 2.

5. Как и в случае обыкновенных дифференциальных уравнений (гл. 8, § 9) заслуживают внимания методы интегрирования с переменными шагами по времени, в ряде случаев являющиеся весьма эффективными.

Задача 3. Доказать, что аппроксимации (10) при $0 \leqslant a\tau/h \leqslant 1$, (11) при $-1 \leqslant a\tau/h \leqslant 0$ и (12) при $|a\tau/h| \leqslant 1$ обладают следующим свойством: для их решений выполняется неравенство

$$\sum_m |u_m^{n+1}| \leqslant \sum_m |u_m^n|,$$

если

$$\sum_m |u_m^0| < \infty,$$

и неравенство

$$\sup_m \sum_m |u_m^{n+1}| \leqslant \sup_m \sum_m |u_m^n|, \tag{14}$$

если $\sup\limits_m |u_m^0|$ существует.

Задача 4. Доказать, что при использовании этих аппроксимаций при любом n решение сеточной задачи монотонно по m, если монотонно u_m^0.

Это свойство монотонности делает схемы, удовлетворяющие условию (14), весьма удобными при интегрировании разрывных решений. Если при интегрировании разрывных решений употреблять аппроксимации, не обладающие таким свойством, то в разностном решении появляются паразитические волны, имитирующие разрывы и иногда мешающие пониманию истинной картины явления.

Сделаем ряд замечаний по поводу практического употребления этих аппроксимаций. Исторически первой была аппроксимация (12); она обладает следующим недостатком.

Если мы подставим в $L_h[u]_h$ разложение Тейлора $u(x, t)$ в точке $(mh, n\tau)$, то получим

$$\begin{aligned}(L_h[u]_h)|_{(m,n)} &= \tau u_{tt}(mh, n\tau)/2 - (h^2/2\tau)u_{xx}(mh, n\tau) + \cdots = \\ &= (\tau a^2/2 - h^2/2\tau)u_{xx}(mh, n\tau).\end{aligned}$$

Условия $\tau, h \to 0$ еще недостаточно, чтобы эта схема аппроксимировала исходное уравнение. Необходимо еще добавить условие $h^2/\tau \to 0$. Требование корректности $|a\tau/h| \leqslant 1$ и ряд других условий при решении сложных задач приводят к тому, что отношение h^2/τ часто остается большим при малых τ и h. Качественно это ухудшение аппроксимации проявляется в появлении «вязкости» аппроксимации: все неровности решения, включая разрывы, сильно выглаживаются.

В случае, когда коэффициент a меняет знак, возможно совместное использование схем (10), (11), когда используется та или иная схема в зависимости от знака a. Одна из наиболее распространенных схем решения задач газовой динамики (схема Годунова) использует именно эту идею.

Аппроксимация (13) эффективна в случае гладких решений, но при наличии разрывов дает большое число паразитических волн. Поэтому она подвергается модификации в областях больших градиентов решения.

После построения аппроксимаций (10)–(12) в применении к гиперболическим задачам долгое время казалось, что применение схем высокого порядка точности является неоправданным; такой вывод объяснялся тем, что все известные к тому времени аппроксимации второго порядка превращали разрывные решения в решения с большим числом паразитических волн.

Впоследствии теоретический анализ вопроса о качественных свойствах решений аппроксимаций при наличии разрывов показал, что этим свойством обладают все линейные аппроксимации второго порядка точности. Однако теоретические исследования предсказали, а практика подтвердила наличие аппроксимаций третьего порядка точности с удовлетворительным поведением решений разностных уравнений при наличии разрывов.

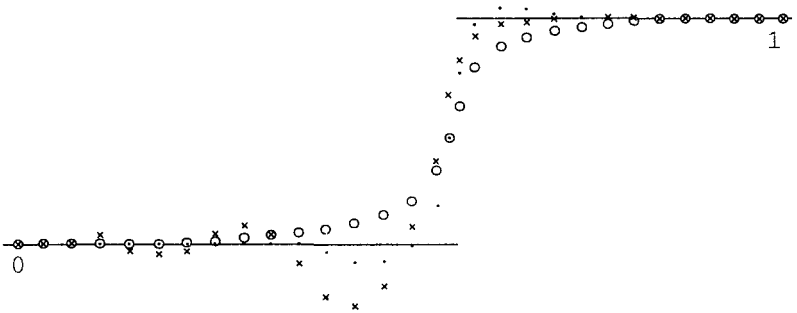


Рис. 10.2.5

На рис. 10.2.5 изображено поведение решений различных разностных аппроксимаций уравнения

$$u_t + u_x = 0 \quad \text{при} \quad u_0(x) = \begin{cases} 0, & \text{если } x < 0, \\ 1, & \text{если } x \geqslant 1; \end{cases}$$

сплошной линией обозначено точное решение, $\circ$, $\times$, $\cdot$ — полученные расчетные значения по схемам (10), (13) и третьего порядка соответственно.

Рассмотрим пример применения СПУ в случае решения системы уравнений.

Пусть решается задача Коши для системы уравнений

$$u_t + av_x = 0, \quad v_t + bu_x = 0, \quad ab > 0;$$

условие $ab > 0$ обеспечивает гиперболичность системы.

Рассмотрим трехслойную разностную схему, аппроксимирующую исходную задачу с погрешностью $O(h^2 + \tau^2)$:

$$\begin{aligned} &\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} + a\,\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0, \\ &\frac{v_m^{n+1} - v_m^{n-1}}{2\tau} + b\,\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0. \end{aligned} \tag{15}$$

Ищем частное решение системы в виде

$$u_m^n = c_1\lambda^n e^{\mathbf{i}m\varphi}, \qquad v_m^n = c_2\lambda^n e^{\mathbf{i}m\varphi}.$$

После подстановки этих выражений в (15) получим

$$\lambda^{n-1}e^{\mathbf{i}m\varphi}\left(c_1\frac{\lambda^2-1}{\tau} + 2c_2a\lambda\frac{\mathbf{i}\sin\varphi}{h}\right) = 0,$$

$$\lambda^{n-1}e^{\mathbf{i}m\varphi}\left(c_2\frac{\lambda^2-1}{\tau} + 2bc_1\lambda\frac{\mathbf{i}\sin\varphi}{h}\right) = 0$$

и, следовательно,

$$c_1\frac{\lambda^2-1}{\tau} + 2ac_2\lambda\frac{\mathbf{i}\sin\varphi}{h} = 0,$$

$$2c_1b\,\lambda\frac{\mathbf{i}\sin\varphi}{h} + c_2\frac{\lambda^2-1}{\tau} = 0.$$

Эта система линейных уравнений относительно коэффициентов c_1, c_2 имеет ненулевое решение, если определитель системы равен нулю. Получаем уравнение, связывающее λ и φ:

$$\det\begin{pmatrix} \dfrac{\lambda^2-1}{\tau} & \dfrac{\lambda a\cdot 2\mathbf{i}\sin\varphi}{h} \\ \dfrac{\lambda b\cdot 2\mathbf{i}\sin\varphi}{h} & \dfrac{\lambda^2-1}{\tau} \end{pmatrix} = 0.$$

Отсюда

$$\left(\frac{\lambda^2-1}{\tau}\right)^2 + \lambda^2\frac{4ab\sin^2\varphi}{h^2} = 0$$

или

$$\lambda^2 \pm 2\mathbf{i}\sqrt{ab}\,\frac{\tau}{h}\sin\varphi\,\lambda - 1 = 0.$$

Таким образом, окончательно

$$\lambda = \mp\mathbf{i}\sqrt{ab}\,\frac{\tau}{h}\sin\varphi \pm \sqrt{-ab\frac{\tau^2}{h^2}\sin^2\varphi + 1}.$$

Если $ab\tau^2/h^2 \leqslant 1$, то подкоренное выражение неотрицательно и

$$|\lambda| = ab\frac{\tau^2}{h^2}\sin^2\varphi + \left(-ab\frac{\tau^2}{h^2}\sin^2\varphi + 1\right) = 1.$$

Таким образом, при $\sqrt{ab}\,\tau/h \leqslant 1$ СПУ выполнен.

Задача 5. Показать, что при $\sqrt{ab}\,\dfrac{\tau}{h} = \kappa = \text{const} > 1$ СПУ не выполнен.

Задача 6. С помощью теоремы об областях зависимости показать, что при $\sqrt{ab}\,\dfrac{\tau}{h} = \kappa = \text{const} > 1$ решение сеточной задачи не обязательно сходится к решению дифференциальной задачи.

Пусть в области $0 \leqslant m \leqslant M$ ищется функция u_m^n, удовлетворяющая (3) и некоторым граничным условиям

$$L_h^1 u_h = 0 \tag{16}$$

относительно значений u_m^n при m, близких к нулю, и

$$L_h^2 u_h = 0 \tag{17}$$

относительно значений u_m^n при m, близких к M.

Примечание. Число уравнений относительно значений u_m^n на каждом слое берется равным числу неизвестных $M+1$, поэтому некоторые из уравнений (3) отбрасываются.

Для краевых сеточных задач с постоянными коэффициентами также имеется СПУ, часто позволяющий довольно просто отбраковать непригодные для счета разностные схемы. Можно показать, что разностная схема, не удовлетворяющая этому СПУ, неустойчива.

Этот СПУ заключается в следующем.

1. Должен быть выполнен спектральный признак устойчивости задачи Коши (отличный от сформулированного ранее). Ищутся всевозможные частные решения (3) вида

$$u_m^n = \lambda^n \varphi_m, \tag{18}$$

$$\|\varphi\|_0 \equiv \sup_{-\infty<m<\infty} |\varphi_m| < \infty. \tag{19}$$

СПУ задачи Коши состоит в том, что при заданном законе стремления τ, h к нулю

$$\overline{\lim} \sup_{\|\varphi\|_0<\infty} |\lambda| \leqslant 1; \tag{20}$$

обозначение $\sup\limits_{\|\varphi\|_0<\infty}$ введено с целью подчеркнуть еще раз, что верхняя грань берется по множеству всех решений (18), удовлетворяющих условию (19).

2. Должно быть выполнено *условие спектральной устойчивости «левой» краевой задачи*, состоящее в следующем.

Рассмотрим «левую» краевую задачу

$$L_h u_h|_{(m,n)} = 0 \quad \text{при} \quad 0 \leqslant m < \infty, \quad L_h^1 u_h = 0$$

(с учетом примечания) и находим ее частные решения вида

$$u_m^n = \lambda^n \varphi_m, \quad \|\varphi\|_+ = \sup_{0\leqslant m<\infty} |\varphi_m| < \infty.$$

СПУ левой краевой задачи имеет вид, аналогичный (20):

$$\overline{\lim} \sup_{\|\varphi\|_+<\infty} |\lambda| \leqslant 1. \tag{21'}$$

3. Точно так же рассматривается «правая» краевая задача

$$L_h u_h = 0 \quad \text{при} \quad -\infty < m \leqslant M, \quad L_h^2 u_h = 0.$$

Ищутся ее частные решения вида $u_m^n = \lambda^n \varphi_m$ такие, что $\|\varphi\|_- = \sup\limits_{-\infty<m\leqslant M} |\varphi_m| < \infty$. СПУ «правой» краевой задачи имеет вид

$$\overline{\lim} \sup_{\|\varphi\|_-<\infty} |\lambda| \leqslant 1. \tag{21''}$$

Заметим, что замена неизвестной переменной $m' = M - m$ переводит «левую» краевую задачу в «правую» и наоборот и соответственно преобразуются друг в друга СПУ «левой» и «правой» задач.

Пример 5. Рассмотрим сеточную краевую задачу

$$\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0, \quad 0 < m < M, \quad Mh = 1, \\ &\frac{u_0^n - 4u_1^n + 3u_2^n}{2h} = 0, \quad u_M^n = 0. \end{aligned} \tag{22}$$

Эта задача аппроксимирует дифференциальное уравнение

$$u_t - u_{xx} = 0 \quad \text{при} \quad 0 < x < 1$$

с краевыми условиями $u_x(0, t) = 0$, $u(1, t) = 0$.

Исследуем спектральную устойчивость сеточной задачи в предположении, что $\tau/h^2 = \kappa = \text{const}$ при стремлении τ, h к нулю.

1. СПУ задачи Коши. Ищем частные решения вида $u_m^n = \lambda^n \varphi_m$. После подстановки в (22) и сокращения на λ^n получаем уравнение

$$\frac{\lambda - 1}{\tau}\varphi_m - \frac{\varphi_{m+1} - 2\varphi_m + \varphi_{m-1}}{h^2} = 0$$

или, что то же самое,

$$\varphi_{m+1} - \left(2 + (\lambda - 1)\frac{h^2}{\tau}\right)\varphi_m + \varphi_{m-1} = 0. \tag{23}$$

Функция φ_m является решением одномерного конечно-разностного уравнения с постоянными коэффициентами (23), поэтому $\varphi_m = c_1\mu_1^m + c_2\mu_2^m$, где $\mu_{1,2}$ — корни характеристического уравнения

$$\mu^2 - \left(2 + (\lambda - 1)\frac{h^2}{\tau}\right)\mu + 1 = 0. \tag{24}$$

Согласно теореме Виета $\mu_1\mu_2 = 1$, поэтому

$$\varphi_m = c_1\mu_1^m + c_2\mu_1^{-m}.$$

Для ограниченности $\|\varphi\|_0$ при всех c_1, c_2 необходимо, чтобы $|\mu_1| = 1$. Полагая $\mu_1 = e^{i\varphi}$, из (24) получаем

$$\lambda = 1 - 4\frac{\tau}{h^2}\sin^2\frac{\varphi}{2}.$$

Если $\kappa = \tau/h^2 \leqslant 1/2$, то $\sup|\lambda| = 1$ и СПУ задачи Коши выполнен. В противном случае он не выполнен.

2. Спектральную неустойчивость «левой» краевой задачи мы докажем, «угадав» последовательность частных решений

$$u_m^n = \lambda^n\varphi_m \quad \text{с} \quad \|\varphi_m\|_+ < \infty$$

таких, что $\lim\limits_{\tau,\,h\to 0}|\lambda| > 1$. Ищем решение «левой» краевой задачи в виде $u_m^n = \lambda^n\mu^m$. Подставляя u_m^n в (22) и сокращая на $\lambda^n\mu^m$, получим

$$\frac{\lambda - 1}{\tau} - \frac{\mu - 2 + u^{-1}}{h^2} = 0. \tag{25}$$

Подставляя u_m^n в левое граничное условие (22), получим уравнение

$$1 - 4\mu + 3\mu^2 = 0.$$

Его корни $\mu_1 = 1$, $\mu_2 = \dfrac{1}{3}$; корню $\mu_2 = \dfrac{1}{3}$, согласно (25), соответствует $\lambda = 1 + \dfrac{4\tau}{3h^2} = 1 + \dfrac{4}{3}\kappa$. Для этого частного решения $u_m^n = \lambda^n\varphi_m = \left(1 + \dfrac{4}{3}\kappa\right)^n\left(\dfrac{1}{3}\right)^m$ имеем соотношения

$$\|\varphi_m\|_+ = 1 \quad \text{и} \quad |\lambda| = 1 + \frac{4}{3}\kappa > 1.$$

Следовательно, СПУ этой задачи не выполнен и следует взять другую аппроксимацию граничного условия $u_x = 0$ в точке 0.

Задача 7. Доказать, что аппроксимация граничного условия

$$\frac{-u_2^n + 4u_1^n - 3u_0^n}{2h} = 0$$

соответствует «левой» краевой задаче, удовлетворяющей СПУ.

Пример 6. Рассмотрим «правую» краевую задачу. Как и в случае задачи Коши, получаем совокупность частных решений вида

$$u_m^n = \lambda^n\varphi_m, \quad \varphi_m = C_1\mu_1^m + C_2\mu_2^m,$$

где λ и μ связаны соотношением (25), причем, как и там, $\mu_1\mu_2 = 1$. Удобно представлять функцию φ_m в виде

$$\varphi_m = C_1\mu_1^{m-M} + C_2\mu_2^{m-M}.$$

Из правого граничного условия (22) получаем $\varphi_M = C_1 + C_2 = 0$, поэтому $\varphi_m = C_1(\mu_1^{m-M} - \mu_2^{m-M})$. Если $|\mu_1| > 1$, то $|\mu_2| < 1$, и наоборот; в обоих этих случаях $\varphi_m \to \infty$ при $m \to -\infty$. Поэтому нас интересуют лишь решения с $|\mu_1| = |\mu_2| = 1$. Тогда $\mu_1 = e^{\mathrm{i}\varphi}$, $\mu_2 = e^{-\mathrm{i}\varphi}$ и

$$\lambda = 1 - \frac{4\tau}{h^2}\sin^2\frac{\varphi}{2}.$$

Как и в случае задачи Коши, получаем, что СПУ выполнен при $\kappa = \tau/h^2 \leqslant 1/2$.

Рассмотренные выше разностные схемы относятся к классу *явных*; значения решения на верхнем слое вычисляются по значениям решения на нижних слоях по формулам вида (ср. с (7)):

$$u_m^{n+1} = \sum_{|j|\leqslant k} a_{m+j}^n u_{m+j}^n;$$

k ограничено при $h \to 0$.

Если в разностное уравнение входят не менее чем два значения решения на верхнем слое, то такую схему относят к классу *неявных схем*. В этом случае значения сеточного решения на верхнем слое находят, решая некоторую систему уравнений

$$C\mathbf{u}^{n+1} = F(\mathbf{u}^n, f). \tag{26}$$

В случае, когда матрица C треугольная (а при решении систем уравнений в частных производных блочно-треугольная), разностную схему называют *полуявной*. Таким образом, полуявные схемы составляют подкласс неявных (C — треугольная матрица), а явные — подкласс полуявных (C — единичная матрица) разностных схем.

В случае неявных (в частности полуявных) схем возникает вопрос о построении алгоритма решения системы уравнений (26), устойчивых к влиянию вычислительной погрешности.

Рассмотрим простейшую полуявную схему для уравнения $u_t + au_x = 0$:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_m^{n+1} - u_{m-1}^{n+1}}{h} = 0. \tag{27}$$

Ищем частные решения вида $u_m^n = \lambda^n e^{\mathrm{i}m\varphi}$; после подстановки такого представления в (27) получим

$$\left(\frac{\lambda - 1}{\tau} + a\frac{\lambda(1 - e^{-\mathrm{i}\varphi})}{h}\right)\lambda^n e^{\mathrm{i}m\varphi} = 0.$$

Таким образом, $\lambda(\varphi) = \left(1 + a\frac{\tau}{h} - a\frac{\tau}{h}e^{-\mathrm{i}\varphi}\right)^{-1}$.

Задача 8. Пусть $\sigma = \sup\limits_{0\leqslant\varphi\leqslant 2\pi} |\lambda(\varphi)|$. Показать, что

1) $\sigma \leqslant 1$ при $a \geqslant 0$;

2) $\sigma \leqslant 1$ при $a\dfrac{\tau}{h} \leqslant -1$;

3) $\sigma = \dfrac{1}{|2\kappa - 1|} > 1$ при $a\dfrac{\tau}{h} = \kappa = \text{const}$, $-1 < \kappa < 0$.

Таким образом, СПУ не выполнен при $-1 < \kappa < 0$.

В реальных вычислениях всегда участвует конечное число значений u_m^{n+1}, соответствующих $(n+1)$-му слою. Если мы выпишем уравнение (27) при $m = 1, \ldots, M$, то получим систему из M уравнений относительно $(M+1)$-го неизвестных $u_0^{n+1}, \ldots, u_M^{n+1}$.

Рассмотрим случай, когда решение ищется в прямоугольнике

$$0 \leqslant x \leqslant X = Mh, \qquad 0 \leqslant t \leqslant T.$$

Функция $u(x, t) = \psi(x - at)$ является решением уравнения $u_t + au_x = 0$. Если $a = 0$, то решение однозначно определяется заданием начальных условий $u(x, 0) = u_0(x)$; при $a > 0$ для нахождения решения достаточно задать $u(x, 0)$ и $u(0, t)$; при $a < 0$ — задать $u(x, 0)$ и $u(X, t)$.

В случае $a > 0$ из граничных условий нам известно также, что $u_0^{n+1} = u(0, (n+1)\tau)$ и, таким образом, число уравнений равно числу неизвестных.

Уравнение (27) можно представить в виде рекуррентной формулы для определения u_m^{n+1}:

$$u_m^{n+1} = \frac{\kappa}{1+\kappa} u_{m-1}^{n+1} + \frac{u_m^n}{1+\kappa}, \quad \kappa = \frac{a\tau}{h}, \quad m = 1, \ldots, M. \tag{28}$$

Если u_{m-1}^{n+1} содержит некоторую погрешность δ_{m-1}^{n+1}, то вследствие (28) она порождает погрешность δ_m^{n+1} в u_m^{n+1}, равную $\dfrac{\kappa}{1+\kappa}\delta_{m-1}^{n+1}$. При $a \geqslant 0$ имеем $\kappa \geqslant 0$ и $0 \leqslant \dfrac{\kappa}{1+\kappa} < 1$. Таким образом, $|\delta_m^{n+1}| < |\delta_{m-1}^{n+1}|$, т. е. погрешности в значениях u_m^{n+1} убывают и можно ожидать, что при вычислениях по формулам (28) накопление погрешности не приведет к нежелательным последствиям.

В случае $a < 0$ известно значение $u_M^{n+1} = u(X, (n+1)\tau)$, и вычисления будем вести по рекуррентной формуле, вытекающей из (27):

$$u_{m-1}^{n+1} = \left(1 + \frac{1}{\kappa}\right) u_m^{n+1} - \frac{1}{\kappa} u_m^n, \quad m = M, \ldots, 1. \tag{29}$$

Если $\kappa \leqslant -1$, то $0 \leqslant 1 + \dfrac{1}{\kappa} < 1$ и погрешность δ_m^{n+1} в значении u_{m-1}^{n+1} породит погрешность $\delta_{m-1}^{n+1} = \left(1 + \dfrac{1}{\kappa}\right)\delta_m^{n+1}$ такую, что $|\delta_{m-1}^{n+1}| < |\delta_m^{n+1}|$. Таким образом, следует ожидать, что накопление вычислительной погрешности не будет существенным.

Замечание. Для расчетной формулы (28) соотношение $|\delta_m^{n+1}| \leqslant |\delta_{m-1}^{n+1}|$ выполнено и при $-1/2 \leqslant \kappa < 0$, а для формулы (29) соотношение $|\delta_{m-1}^{n+1}| \leqslant |\delta_m^{n+1}|$

выполнено и при $-1 < \kappa \leqslant -1/2$. В этих случаях, однако, не выполнен спектральный признак устойчивости $\sigma = \sup\limits_{0 \leqslant \varphi \leqslant 2\pi} |\lambda(\varphi)| < 1$.

Предположим, что в области $t \leqslant T$ нас интересует решение задачи Коши при начальном условии $u(x, 0) = u_0(x)$, определенном при $0 \leqslant x \leqslant X_0$ и $a > 0$. Решение дифференциальной задачи определено в полосе $at \leqslant x \leqslant X_0 + at$.

Рассмотрим краевую задачу в прямоугольнике $0 \leqslant x \leqslant X = X_0 + at$, $0 \leqslant t \leqslant T$, задав произвольные гладкие функции $u(0, t)$ при $0 \leqslant t \leqslant T$ и $u(x, 0)$ при $X_0 \leqslant x \leqslant X$, удовлетворяющие условиям согласования $u(0, 0) = u_0(0)$ и $u_0(X_0) = u(X_0, 0)$.

Применим для решения этой задачи полуявную схему (27), проводя вычисления на верхнем слое по рекуррентным формулам (28). Можно показать, что при гладкой функции $u_0(x)$ в области $at \leqslant x \leqslant X_0 + at$, $0 \leqslant t \leqslant T$ полученное сеточное решение будет близко к решению задачи Коши.

При решении задачи Коши и краевых задач, особенно в случае систем уравнений, часто используется неявная схема

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2h} = 0 \tag{30}$$

с порядком аппроксимации $O(\tau + h^2)$. Иногда эта схема используется как вспомогательная для получения решения на полуцелом слое, т. е. находятся $u_m^{n+1/2}$ из соотношений

$$\frac{u_m^{n+1/2} - u_m^n}{\tau/2} + a\frac{u_{m+1}^{n+1/2} - u_{m-1}^{n+1/2}}{2h} = 0;$$

далее используется явная формула

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^{n+1/2} - u_{m-1}^{n+1/2}}{2h} = 0.$$

Рассмотрим случай той же самой задачи Коши. Если мы выпишем уравнение (30) при $m = 1, \ldots, M-1$, то получим систему $(M-1)$-го уравнения с $(M+1)$-м неизвестным. На каждом слое нам не хватает двух уравнений. Значения u_0^n зададим как и ранее, используя граничные значения, а u_M^n — по произволу. Систему уравнений (30) относительно значений $u_1^{n+1}, \ldots, u_{M-1}^{n+1}$ будем решать методом прогонки.

Можно показать, что при гладкой функции $u_0(x)$ и любом $\varepsilon > 0$ решение сеточной задачи сходится к решению дифференциальной в области, определяемой неравенствами

$$at + \varepsilon \leqslant x \leqslant X_0 + at - \varepsilon, \quad 0 \leqslant t \leqslant T.$$

Для улучшения сходимости вместо задания u_M^{n+1} часто целесообразнее взять так называемое «мягкое» граничное условие $u_M^{n+1} - u_{M-1}^{n+1} = 0$.

§ 3. Принцип замороженных коэффициентов

Часто не удается произвести теоретическое исследование корректности разностной задачи и доказать сходимость ее решения к решению дифференциальной задачи. В некоторых случаях на данном этапе развития математической теории такое исследование в принципе возможно, но требует от исследователя достаточно высокой квалификации и больших затрат времени.

В такой ситуации иногда ограничиваются исследованием устойчивости схем на основе описываемого ниже *принципа замороженных коэффициентов* и последующей экспериментальной проверкой полученных выводов путем расчета тестовых задач, по возможности с известным решением.

Принцип замороженных коэффициентов (ПЗК) заключается в следующем.

1. Для разностной схемы пишется уравнение в вариациях (т. е. уравнение, которому удовлетворяет разность двух бесконечно близких решений). Это уравнение является линейным и в случае линейных задач совпадает с исходным уравнением.

2. Фиксируется некоторая точка P области G и замораживаются коэффициенты этого уравнения, т. е. все значения коэффициентов уравнения в вариациях берутся равными их значениям в этой точке. Если задача нелинейная, то коэффициенты уравнения в вариациях зависят от неизвестной функции и все значения сеточного решения, входящие в это уравнение, берутся равными их значениям в той же точке P.

3. Получившаяся сеточная задача $L_h^P(\delta u_h) = 0$ исследуется на устойчивость методами, которые применяются для исследования устойчивости сеточных задач с постоянными коэффициентами.

Предположим, что сеточная задача устойчива при выполнении условия

$$\varphi(h, P) \geqslant 0 \tag{1}$$

на шаги сетки; это условие, естественно, может зависеть от выбора точки P.

4. За условие устойчивости принимают некоторое условие $\varphi(h) \geqslant 0$, из выполнения которого следует выполнение условия $\varphi(h, P) \geqslant 0$ для всех точек $P \in G$. Часто, особенно в случае нелинейных задач, условие устойчивости $\varphi(h) \geqslant 0$ выбирается с некоторым *«запасом устойчивости»*.

Рассмотрим пример применения принципа замороженных коэффициентов. Пусть решается задача Коши для уравнения

$$u_t + (\varphi(x, t, u))_x - \psi(x, t, u) = 0 \tag{2}$$

при начальном условии $u(x, 0) = u_0(x)$. Через u_m^n будем обозначать приближение к значению решения в точке $(x, t) = (mh, n\tau)$. Аппроксимируем уравнение (2) разностной схемой

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{\varphi(mh, n\tau, u_m^n) - \varphi((m-1)h, n\tau, u_{m-1}^n)}{h} - \psi(mh, n\tau, u_m^n) = 0. \tag{3}$$

Пусть $v_m^n = u_m^n + \delta_m^n$ — другое решение сеточной задачи (3), т. е. δ_m^n — разность между двумя решениями задачи (3). Имеем равенство (получаемое при подстановке v_m^n в (3)):

$$\frac{(u_m^{n+1} + \delta_m^{n+1}) - (u_m^n + \delta_m^n)}{\tau} - \psi(mh, n\tau, u_m^n + \delta_m^n) + \\ + \frac{\varphi(mh, n\tau, u_m^n + \delta_m^n) - \varphi((m-1)h, n\tau, u_{m-1}^n + \delta_{m-1}^n)}{h} = 0.$$

Вычитая из этого равенства соотношение (3), получим

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + \frac{\varphi(mh, n\tau, u_m^n + \delta_m^n) - \varphi(mh, n\tau, u_m^n)}{h} - \\ - \frac{\varphi((m-1)h, n\tau, u_{m-1}^n - \delta_{m-1}^n) - \varphi((m-1)h, n\tau, u_{m-1}^n)}{h} - \\ - (\psi(mh, n\tau, u_m^n + \delta_m^n) - \psi(mh, n\tau, u_m^n)) = 0. \tag{4}$$

С точностью до членов второго порядка малости по величинам δ_j^n выполнены приближенные равенства

$$\varphi(mh, n\tau, u_m^n + \delta_m^n) - \varphi(mh, n\tau, u_m^n) \approx \varphi_u(mh, n\tau, u_m^n)\delta_m^n,$$
$$\varphi((m-1)h, n\tau, u_{m-1}^n + \delta_{m-1}^n) - \varphi((m-1)h, n\tau, u_{m-1}^n) \approx \\ \approx \varphi_u((m-1)h, n\tau, u_{m-1}^n)\delta_{m-1}^n,$$
$$\psi(mh, n\tau, u_m^n + \delta_m^n) - \psi(mh, n\tau, u_m^n) \approx \psi_u(mh, n\tau, u_m^n)\delta_m^n.$$

Таким образом, бесконечно малое приращение решения δ_m^n, его так называемая *вариация*, удовлетворяет уравнению

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + \frac{\varphi_u(mh, n\tau, u_m^n)\delta_m^n - \varphi_u((m-1)h, n\tau, u_{m-1}^n)\delta_{m-1}^n}{h} - \\ - \psi_u(mh, n\tau, u_m^n)\delta_m^n = 0;$$

это уравнение называется *уравнением в вариациях* для (4).

Заморозим коэффициенты, взяв все значения φ_u и ψ_u равными их значениям в некоторой точке

$$\frac{\delta_m^{n+1} - \delta_m^n}{\tau} + a\frac{\delta_m^n - \delta_{m-1}^n}{h} - b\delta_m^n = 0. \tag{5}$$

Отсюда имеем равенство

$$\delta_m^{n+1} = \delta_m^n \left(1 - \frac{a\tau}{h} + b\tau\right) + \delta_{m-1}^n a \frac{\tau}{h}$$

и затем оценку

$$|\delta_m^{n+1}| \leqslant \max_m |\delta_m^n| \left(\left|1 - a\frac{\tau}{h}\right| + |b\tau| + \left|\frac{a\tau}{h}\right|\right). \tag{6}$$

Поскольку (6) выполнено при всех m, то

$$\max_m |\delta_m^{n+1}| \leqslant \max_m |\delta_m^n| \left(\left|1 - \frac{a\tau}{h}\right| + |b\tau| + \left|\frac{a\tau}{h}\right|\right).$$

Положим $\max\limits_m |\delta_m^n| = \|\delta^n\|_C$. Соотношение (6) переписывается в виде

$$\|\delta^{n+1}\|_C \leqslant \left(\left|1 - \frac{a\tau}{h}\right| + |b\tau| + \left|\frac{a\tau}{h}\right|\right) \|\delta^n\|_C.$$

Если $0 \leqslant \dfrac{a\tau}{h} \leqslant 1$, то

$$\left|1 - \frac{a\tau}{h}\right| + |b\tau| + \left|\frac{a\tau}{h}\right| = \left(\left|1 - \frac{a\tau}{h}\right| + \left|\frac{a\tau}{h}\right|\right) + |b\tau| = 1 + |b|\tau \leqslant e^{|b|\tau}.$$

Таким образом,

$$\|\delta^{n+1}\|_C \leqslant e^{|b|\tau} \|\delta^n\|_C. \tag{7}$$

Пользуясь (7), получаем соотношения

$$\begin{aligned}
&\|\delta^1\|_C \leqslant e^{|b|\tau} \|\delta^0\|_C,\\
&\|\delta^2\|_C \leqslant e^{|b|\tau} \|\delta^1\|_C \leqslant e^{|b|2\tau} \|\delta^0\|_C,\\
&\|\delta^3\|_C \leqslant e^{|b|\tau} \|\delta^2\|_C \leqslant e^{|b|3\tau} \|\delta^0\|_C,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\|\delta^N\|_C \leqslant e^{|b|\tau} \|\delta^{N-1}\|_C \leqslant e^{|b|N\tau} \|\delta^0\|_C.
\end{aligned}$$

На ограниченном промежутке времени при $n\tau \leqslant T$ имеем

$$\|\delta^n\|_C \leqslant e^{|b|T} \|\delta^0\|_C,$$

т. е. разностная задача *устойчива по начальным данным.*

Устойчивость разностной схемы (5) доказана при условии $0 \leqslant a\dfrac{\tau}{h} \leqslant 1$.

В соответствии с ПЗК постулируется, что исходная разностная схема (3) должна быть устойчива при условии

$$0 \leqslant a(mh,\, n\tau,\, u_m^n)\, \frac{\tau}{h} \leqslant 1. \tag{8}$$

Для данной разностной схемы (3) можно показать, что это условие является достаточным для ее практической пригодности при малых τ.

В тех случаях, когда не удается строго обосновать устойчивость разностной задачи, рекомендуется создавать «запас устойчивости», сужая

область изменения параметров схемы по сравнению с той, которая получается из принципа замороженных коэффициентов. Например, в данном случае вместо (8) рекомендовалось бы взять условие

$$0 < \kappa \leqslant a\frac{\tau}{h} \leqslant 1 - \kappa.$$

Величина требуемого сужения подбирается из численного эксперимента.

Примеры сужения области устойчивости для различных схем: в 1 раз (сужение не производится); в 1,15 раза; в 1,3 раза; в 1,5 раза; в 2 раза.

Заранее неизвестно, в какой области находятся значения решения сеточной задачи u_m^n, поэтому до реального решения сеточной задачи нельзя выбрать шаг τ таким, чтобы при всех m, n выполнялось условие устойчивости (8). В частности, в связи с этим при решении нестационарных задач шаг по времени часто берется переменным: ищутся приближения u_m^nк значениям решения в точках $(mh,\, t_n)$; шаг $\tau_n = t_{n+1} - t_n$ зависит от n. При каждом n вычисляются

$$A_n^1 = \inf_m a(mh,\, t_n,\, u_m^n), \quad A_n^2 = \sup_m a(mh,\, t_n,\, u_m^n).$$

Если оказалось, что $A_n^1 < 0$, то счет по этой схеме прекращается, поскольку ни при каком $\tau_n > 0$ не удается добиться удовлетворения условия $0 \leqslant a(mh,\, t_n,\, u_m^n)\dfrac{\tau_n}{h}$ при всех m. Если $A_n^1 \geqslant 0$, то шаг $\tau_n = t_{n+1} - t_n$ выбирают таким, чтобы выполнялось условие $A_n^2\tau_n/h \leqslant 1$.

Для линейных и слаболинейных задач во всех известных случаях, когда было проведено строгое исследование устойчивости и коэффициенты уравнения удовлетворяли условию Липшица по всем переменным, имел место следующий факт: если выполнялся критерий устойчивости по ПЗК, то схема действительно была устойчива.

§ 4. Численное решение нелинейных задач с разрывными решениями

Рассмотрим задачу Коши для уравнения

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \tag{1}$$

при начальном условии

$$u_0(x) = \begin{cases} 1 & \text{при} \quad x < 0, \\ 0 & \text{при} \quad x \geqslant 0. \end{cases}$$

Эта задача не имеет непрерывного решения, поэтому разностная задача в обычном смысле не аппроксимирует дифференциальную. Однако тем не менее рассмотрим разностные аппроксимации этой задачи

$$\frac{u_m^{n+1}-u_m^n}{\tau}+u_m^n\frac{u_m^n-u_{m-1}^n}{h}=0, \tag{2}$$

$$\frac{u_m^{n+1}-u_m^n}{\tau}+u_{m-1}^n\frac{u_m^n-u_{m-1}^n}{h}=0, \tag{3}$$

$$\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{u_m^n+u_{m-1}^n}{2}\frac{u_m^n-u_{m-1}^n}{h}=0 \tag{4}$$

при начальных условиях

$$u_m^0=\begin{cases}1 & \text{при} \quad m<0,\\ 0 & \text{при} \quad m\geqslant 0.\end{cases} \tag{5}$$

Нетрудно убедиться, что при $\tau=h$ решением задачи (2), (5) является

$$u_m^n=\begin{cases}1 & \text{при} \quad m<0,\\ 0 & \text{при} \quad m\geqslant 0;\end{cases}$$

решением задачи (3), (5) является

$$u_m^n=\begin{cases}1 & \text{при} \quad m-n<0,\\ 0 & \text{при} \quad m-n\geqslant 0.\end{cases}$$

Решение задачи (4), (5) не выписывается в явном виде. При $\tau=h$ из (4) имеем

$$u_m^{n+1}=u_m^n-\frac{1}{2}\Big((u_m^n)^2-(u_{m-1}^n)^2\Big). \tag{6}$$

Непосредственно вычисляя значения u_m^n, можно убедиться, что

$$u_m^n\approx\begin{cases}1 & \text{при} \quad m-n/2<0, \quad |m-n/2|\gg 1,\\ 0 & \text{при} \quad m-n/2>0, \quad |m-n/2|\gg 1.\end{cases}$$

Более точные теоретические оценки показывают, что

$$u_m^n=\begin{cases}1+O(q^{|m-n/2|}) & \text{при} \quad m-n/2\to-\infty,\\ 0+O(q^{|m-n/2|}) & \text{при} \quad m-n/2\to\infty,\end{cases}$$

где $q<1$; следовательно, решение сеточной задачи (4), (5) близко к разрывной функции

$$u(x,\,t)=\begin{cases}1, & x-t/2<0,\\ 0, & x-t/2>0.\end{cases}$$

Таким образом, решения различных разностных задач, аппроксимирующих на гладком решении одну и ту же дифференциальную задачу, в

случае разрывного решения могут сходиться к различным пределам при стремлении шагов сетки к нулю. Заметим, что само решение такой дифференциальной задачи также не определено однозначно, пока ничего не сказано о том, как проходит линия разрыва решения.

В наиболее типичных случаях условия на линии разрыва являются следствием интегральных законов сохранения, из которых возникла данная дифференциальная задача.

Пусть u — гладкое решение уравнения (1); интегрируя (1) по переменным (x, t) по некоторой области G, получим

$$\int_G \left(\frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial u^2}{\partial x} \right) dx dt = 0$$

или

$$\int_\Gamma \left(\frac{1}{2} u^2 \, dt - u \, dx \right) = 0; \tag{7}$$

здесь Γ — граница области G. Если умножить уравнение (1) на u и проинтегрировать по области G, то получим

$$\int_G \left(\frac{1}{2} \frac{\partial u^2}{\partial t} + \frac{1}{3} \frac{\partial u^3}{\partial x} \right) dx dt = 0$$

или

$$\int_\Gamma \left(\frac{u^3}{3} \, dt - \frac{u^2}{2} \, dx \right) = 0. \tag{8}$$

Если для гладкой функции $u(x, t)$ выполнены условия (7) или (8) для любого контура Γ, то эти условия равносильны. Дело обстоит иначе в случае разрывной $u(x, t)$.

Если $u(x, t)$ — кусочно-гладкая функция, то из (7) можно получить, что в области гладкости функция u является решением уравнения (1), а вдоль линии разрыва $X(t)$ выполнено соотношение

$$\frac{dX}{dt} = \omega = \frac{[u^2/2]}{[u]} = \frac{u_+ + u_-}{2};$$

здесь

$$f_+(x, t) = \lim_{\varepsilon > 0, \varepsilon \to 0} f(x + \varepsilon, t),$$

$$f_-(x, t) = \lim_{\varepsilon < 0, \varepsilon \to 0} f(x - \varepsilon, t), \quad [f] = f_+ + f_-.$$

Точно так же из (8) следует, что в области гладкости функция u является решением уравнения (1), а вдоль линии разрыва выполнено соотношение

$$\frac{dX}{dt} = \omega = \frac{[u^3/3]}{[u^2/2]} = \frac{2(u_+^2 + u_+ u_- + u_-^2)}{3(u_+ + u_-)}.$$

Из вышесказанного видно, что для сходимости решения разностной задачи к разрывному решению уравнения (1) существенно определенное

соответствие между разностной задачей и законом сохранения, соответствующим дифференциальной задаче.

Соображения здравого смысла, численный эксперимент и теоретические оценки погрешности для случая одной неизвестной функции показали, что разностная схема должна обладать *свойством дивергентности*. Первоначально это свойство формулировалось в следующем виде: если ищется решение уравнения

$$\frac{\partial \varphi(x,\, t,\, u)}{\partial t} + \frac{\partial \psi(x,\, t,\, u)}{\partial x} = 0,$$

соответствующее закону сохранения $\int (\psi\, dt - \varphi\, dx) = 0$, то левая часть разностной схемы должна являться линейной комбинацией выражений

$$\varphi(x_m,\, t_n,\, u_m^n), \quad \psi(x_m,\, t_n,\, u_m^n) \tag{9}$$

или близких к ним. Например, разностная схема (4) удовлетворяет условию дивергентности по отношению к закону сохранения (7). По отношению к закону сохранения (8) условию дивергентности удовлетворяет разностная схема

$$\frac{(u_m^{n+1})^2 - (u_m^n)^2}{2\tau} + \frac{(u_m^n)^3 - (u_{m-1}^n)^3}{3h} = 0.$$

Впоследствии оказалось, что условие дивергентности допускает существенное расширение. Например, такому расширенному условию дивергентности по отношению к уравнению $u_t + (\varphi(u))_x = 0$ удовлетворяет следующая разностная схема.

Сначала делается полушаг

$$\frac{u_{m+1/2}^{n+1/2} - \dfrac{u_{m+1}^n + u_m^n}{2}}{\tau} + \varphi'\left(\frac{u_{m+1}^n + u_m^n}{2}\right)\frac{u_{m+1}^n - u_m^n}{h} = 0,$$

а затем полный шаг по формуле

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{\varphi\left(u_{m+1/2}^{n+1/2}\right) - \varphi\left(u_{m-1/2}^{n+1/2}\right)}{h} = 0.$$

Здесь разностная схема записывается в виде линейной комбинации выражений (9) лишь на заключительном шаге.

§ 5. Разностные схемы для одномерного параболического уравнения

После того как мы познакомились с вопросами устойчивости и сходимости для гиперболических задач на нестрогом уровне, перейдем к исследованию разностных схем для параболического уравнения в случае одной пространственной переменной на математическом уровне строгости.

Пусть требуется найти функцию $u(x, t)$, являющуюся решением уравнения

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t) \tag{1}$$

в области $\bar{Q}_T = [0, X] \times [0, T]$ с начальными и краевыми условиями

$$u(x, 0) = u_0(x), \quad u(0, t) = \mu_1(t), \quad u(X, t) = \mu_2(t). \tag{2}$$

Всюду далее будем считать, что функции f, μ_i и u_0 таковы, что существует достаточно гладкое решение задачи (1), (2).

При построении разностной схемы поступим так же, как и ранее. Разобьем исходную область $\bar{Q}_T$ прямоугольной сеткой с шагами $h = X/M$, $\tau = T/N$ соответственно по координатам x и t. Будем искать функцию u^h, определенную в узлах (m, n) сетки $\bar{Q}_{h,\tau} = \{(mh, n\tau) : 0 \leqslant m \leqslant M, 0 \leqslant n \leqslant N\}$, которая является приближением функции u в $\bar{Q}_{h,\tau}$. Обозначим, как и ранее, $u^h(mh, n\tau) = u_m^n$.

Заменим производные в (1) разностными отношениями. Производная $\partial u/\partial t$ в точке $(mh, n\tau)$ может быть заменена разностным отношением многими способами, например

$$\left.\frac{\partial u}{\partial t}\right|_{(mh,\,n\tau)} \approx \frac{u(mh, (n+1)\tau) - u(mh, n\tau)}{\tau} \quad \text{или же}$$

$$\left.\frac{\partial u}{\partial t}\right|_{(mh,\,n\tau)} \approx \frac{u(mh, n\tau) - u(mh, (n-1)\tau)}{\tau}.$$

В зависимости от способа аппроксимации будут получаться различные разностные схемы. Вторую производную по переменной x заменим обычным способом:

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(mh,\,n\tau)} \approx \frac{u((m-1)h, n\tau) - 2u(mh, n\tau) + u((m+1)h, n\tau)}{h^2}.$$

Подставляя эти соотношения вместо соответствующих производных в (1), получим

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} + \varphi_m^n; \tag{3}$$
$$m = 1, \ldots, M-1, \qquad n = 0, \ldots, N-1;$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^{n+1} - 2u_t m^{n+1} + u_{m+1}^{n+1}}{h^2} + \varphi_m^{n+1}; \tag{4}$$
$$m = 1, \ldots, M-1, \qquad n = 0, \ldots, N-1;$$

(вторая схема получена после переобозначения $n \to n+1$). Функция φ_m^n является аппроксимацией $f(x, t)$.

Кроме уравнения (1) необходимо аппроксимировать начальные и граничные условия. Положим

$$u_m^0 = u_0(mh), \quad u_0^n = \mu_1(n\tau), \quad u_M^n = \mu_2(n\tau). \tag{5}$$

Таким образом, уравнения (3), (5) и (4), (5) соответствуют некоторым разностным аппроксимациям краевой задачи для параболического уравнения (1), (2).

Найдем порядок погрешности аппроксимации разностной схемы (3), (5). Для этого подставим в (3) точное решение дифференциальной задачи. Так как

$$\frac{u(x, n\tau+\tau) - u(x, n\tau)}{\tau} = \left.\frac{\partial u}{\partial t}\right|_{(x, n\tau)} + \frac{\tau}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x, n\tau+\xi)}, \quad 0 \leqslant \xi \leqslant \tau,$$

$$\frac{u((m-1)h, t) - 2u(mh, t) + u((m+1)h, t)}{h^2} = \left.\frac{\partial^2 u}{\partial x^2}\right|_{(mh, t)} +$$

$$+ \frac{h^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{(mh+\eta, t)}, \quad 0 \leqslant \eta \leqslant h,$$

то

$$\frac{u(x, t+\tau) - u(x, t)}{\tau} - \frac{u(x-h, t) - 2u(x, t) + u(x+h, t)}{h^2} - \varphi(x, t) =$$

$$= \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + \frac{\tau}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x, t+\xi)} - \frac{h^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{(x+\eta, t)} - \varphi(x, t) =$$

$$= f(x, t) - \varphi(x, t) + O(h^2 + \tau).$$

Таким образом, если положить $\varphi_m^n = f(mh, n\tau)$, то порядок погрешности аппроксимации разностной схемы (3), (5) будет $O(h^2 + \tau)$ (граничные и начальные условия выполнены точно). Аналогично устанавливается, что порядок погрешности аппроксимации схемой (4), (5) задачи (1), (2) также равен $O(h^2 + \tau)$.

Между схемами (3), (5) и (4), (5), однако, имеется принципиальная разница. Выясним ее суть. Из (3) следует соотношение

$$u_m^{n+1} = u_m^n + \frac{\tau}{h^2}(u_{m-1}^n - 2u_m^n + u_{m+1}^n) + \tau\varphi_m^n. \tag{6}$$

В силу того что значения u_m^0 известны, из (6) можно найти значения u_m^1 ($m = 1, \ldots, M-1$) и т. д. Поэтому по известным значениям u_m^n решение u_m^{n+1} на следующем временном слое находится с помощью явных формул (6). Поэтому схема (3), (5) называется *явной*.

Преобразуя (4), имеем

$$-\frac{\tau}{h^2}u_{m-1}^{n+1}+\left(1+\frac{2\tau}{h^2}\right)u_m^{n+1}-\frac{\tau}{h^2}u_{m+1}^{n+1}=u_m^n+\tau\varphi_m^{n+1},\ m=1,\dots,M-1;$$

$$u_0^{n+1}=\mu_1^{n+1}\equiv\mu_1((n+1)\tau),\quad u_M^{n+1}=\mu_2^{n+1}\equiv\mu_2((n+1)\tau). \tag{7}$$

При известных u_m^n, $m=1,\dots,M-1$, соотношения (7) представляют собой систему линейных алгебраических уравнений относительно неизвестных u_m^{n+1}, $m=1,\dots,M-1$. Поэтому схема (4), (5) называется *неявной*.

Система линейных уравнений (7) относительно вектора неизвестных $\mathbf{v}=(u_1^{n+1},\dots,u_{M-1}^{n+1})^T$ может быть записана в виде $A\mathbf{v}=\mathbf{b}$, где матрица A и вектор правой части $\mathbf{b}$ имеют вид

$$A=\begin{pmatrix} 1+\frac{2\tau}{h^2} & -\frac{\tau}{h^2} & 0 & \dots & 0 & 0\\ -\frac{\tau}{h^2} & 1+\frac{2\tau}{h^2} & -\frac{\tau}{h^2} & \dots & 0 & 0\\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot\\ 0 & 0 & 0 & \dots & -\frac{\tau}{h^2} & 1+\frac{2\tau}{h^2} \end{pmatrix},$$

$$b_k=\begin{cases} u_k^n+\tau\varphi_k^{n+1}, & 2\leqslant k\leqslant M-2,\\ u_1^n+\tau\varphi_1^{n+1}+\frac{\tau}{h^2}\mu_1((n+1)\tau), & k=1,\\ u_{M-1}^n+\tau\varphi_{M-1}^{n+1}+\frac{\tau}{h^2}\mu_2((n+1)\tau), & k=M-1. \end{cases}$$

Для решения этой системы можно воспользоваться, например, методом прогонки, описанным в предыдущей главе.

Проведем исследование устойчивости этих разностных схем. Множество узлов вида (m,n), $m=0,\dots,M$, будем называть *n-м слоем*. Пусть u^n — сужение функции u^h на n-й слой, а φ^n — сужение правой части φ^h на внутренние узлы n-го слоя. Введем нормы на слое

$$\|u^n\|=\max_{0\leqslant m\leqslant M}|u_m^n|,\quad \|\varphi^n\|=\max_{0<m<M}|\varphi_m^n|.$$

Разностную схему будем называть *устойчивой в сеточной норме пространства* C, если существует постоянная c_1, не зависящая от шагов сетки h и τ, такая, что имеет место оценка

$$\max_{0\leqslant n\leqslant N}\|u^n\|\leqslant \\ \leqslant c_1\left(\max_{0\leqslant n\leqslant N}\|\varphi^n\|+\max\left\{\max_{0\leqslant n\leqslant N}|\mu_1^n|,\ \max_{0\leqslant n\leqslant N}|\mu_2^n|,\ \|u^0\|\right\}\right). \tag{8}$$

Исследуем вначале устойчивость явной схемы (3), (5). Имеет место

Теорема 1. *Пусть $\tau \leqslant h^2/2$. Тогда разностная схема* (3), (5) *устойчива в сеточной норме пространства C.*

Доказательство. Перепишем (3) в виде

$$u_m^{n+1} = (1-2\rho)u_m^n + \rho u_{m-1}^n + \rho u_{m+1}^n + \tau\varphi_m^n,$$

где $\rho = \tau/h^2$. Если $\max\limits_m |u_m^{n+1}|$ достигается во внутренней точке $(m_0,\, n+1)$, то

$$\max_m |u_m^{n+1}| = \max_m \left|(1-2\rho)u_m^n + \rho u_{m-1}^n + \rho u_{m+1}^n + \tau\varphi_m^n\right| \leqslant$$
$$\leqslant (1-2\rho)\,\|u^n\| + 2\rho\|u^n\| + \tau\|\varphi^n\| = \|u^n\| + \tau\|\varphi^n\|.$$

В противном случае

$$\max_m |u_m^{n+1}| \leqslant \max\left(|\mu_1^{n+1}|,\ |\mu_2^{n+1}|\right).$$

Отсюда следует оценка

$$\|u^{n+1}\| \leqslant \max\left\{|\mu_1^{n+1}|,\ |\mu_2^{n+1}|,\ \|u^n\| + \tau\|\varphi^n\|\right\}, \tag{9}$$

связывающая нормы функции на соседних слоях.

Представим теперь решение u^h задачи (3), (5) в виде $u^h = y^h + v^h$, где y^h — решение задачи (3), (5) с правой частью $\varphi^h \equiv 0$, а v^h — решение задачи (3), (5) с однородными граничными и начальными условиями. В силу оценки (9) для y^h имеем

$$\|y^{n+1}\| \leqslant \max\left\{\max_{0\leqslant k\leqslant N}|\mu_1^k|,\ \max_{0\leqslant k\leqslant N}|\mu_2^k|,\ \|y^n\|\right\} \leqslant \cdots$$
$$\cdots \leqslant \max\left\{\max_{0\leqslant k\leqslant N}|\mu_1^k|,\ \max_{0\leqslant k\leqslant N}|\mu_2^k|,\ \|u^0\|\right\}.$$

С другой стороны, для v^h в силу той же оценки (9) получаем

$$\|v^{n+1}\| \leqslant \|v^n\| + \tau\|\varphi^n\| \leqslant \|v^{n-1}\| + \tau\Big(\|\varphi^n\| + \|\varphi^{n-1}\|\Big) \leqslant \cdots$$
$$\cdots \leqslant \sum_{k=0}^{n}\tau\|\varphi^k\| \leqslant T\max_{0\leqslant k\leqslant N}\|\varphi^k\|,$$

если $(n+1)\tau \leqslant T$. Таким образом, окончательно имеем

$$\|u^n\| \leqslant \|y^n\| + \|v^n\| \leqslant$$
$$\leqslant \max\left\{\max_{0\leqslant k\leqslant N}|\mu_1^k|,\ \max_{0\leqslant k\leqslant N}|\mu_2^k|,\ \|u^0\|\right\} + T\max_{0\leqslant k\leqslant N}\|\varphi^k\|.$$

Так как это неравенство справедливо при любом n, $0 \leqslant n \leqslant N$, то это и означает устойчивость разностной схемы в сеточной норме пространства C. Теорема доказана.

(Этот вывод можно было бы сделать, непосредственно используя соотношение (9) и не вводя в рассмотрение функции y^h и v^h.)

Отметим, что постоянная в (8) получилась в данном случае зависящей от T.

Если $\tau/h^2 = \kappa = \text{const}$, то условие $\kappa \leqslant 1/2$ является необходимым и достаточным условием устойчивости.

Задача 1. Пусть $\lim\limits_{h,\tau\to 0} \dfrac{\tau/h^2 - 1/2}{\tau} = \infty$. Доказать, что схема (3), (5) неустойчива.

Указание. Рассмотреть частные решения

$$u_m^n = \lambda_q^n \sin\frac{\pi m h}{X} q.$$

Доказательство устойчивости разностной схемы (3), (5) было получено при соотношении на шаги сетки $\tau \leqslant h^2/2$. Разностные схемы, которые обладают устойчивостью лишь при определенных соотношениях на шаги сетки, называются *условно устойчивыми*. Соответственно, если схема устойчива при любых сотношениях между шагами сетки, то такая схема называется *безусловно устойчивой*.

Покажем, что схема (4), (5) относится к классу безусловно устойчивых схем. Справедлива

Теорема 2. *При любых h и τ для решения задачи* (4), (5) *имеет место оценка* (8).

Доказательство. По аналогии с доказательством предыдущей теоремы преобразуем (4) к виду

$$u_m^{n+1} + \rho(-u_{m-1}^{n+1} + 2u_m^{n+1} - u_{m+1}^{n+1}) = u_m^n + \tau\varphi_m^{n+1}, \quad 1 \leqslant m \leqslant M-1. \qquad (10)$$

Из всех значений u_m^{n+1}, по модулю равных $\|u^{n+1}\|$, возьмем то, у которого индекс m принимает наименьшее значение. Если $m = 0$ или же $m = M$, то (9) выполнено. Пусть теперь m отлично от 0 и M. Поскольку $|u_m^{n+1}| > |u_{m-1}^{n+1}|$ (по определению m) и $|u_m^{n+1}| \geqslant |u_{m+1}^{n+1}|$, то $|2u_m^{n+1}| > |u_{m-1}^{n+1}| + |u_{m+1}^{n+1}|$; поэтому $\operatorname{sign}(2u_m^{n+1} - u_{m-1}^{n+1} - u_{m+1}^{n+1}) = \operatorname{sign} u_m^{n+1}$. Тогда

$$\|u^{n+1}\| = |u_m^{n+1}| \leqslant \left|u_m^{n+1} + \rho(-u_{m+1}^{n+1} + 2u_m^{n+1} - u_{m-1}^{n+1})\right| =$$
$$= |u_m^n + \tau\varphi_m^{n+1}| \leqslant \|u^n\| + \tau\|\varphi^{n+1}\|;$$

при любых h и τ для схемы (4), (5) получена оценка (9). Завершение доказательства теоремы совпадает с доказательством предыдущей теоремы. Таким образом, схема (4), (5) является безусловно устойчивой.

Между явной (3), (5) и неявной (4), (5) схемами имеется, таким образом, принципиальное отличие. Явной схеме соответствуют явные формулы для вычисления функции на слое по известным значениям на предыдущих слоях. Однако эта схема является условно устойчивой. Это приводит к тому, что при малом шаге h мы вынуждены выбирать слишком мелкий шаг по времени ($\tau \leqslant h^2/2$), чтобы обеспечить устойчивость. Это, в свою очередь, приводит к значительному увеличению затрат времени счета на ЭВМ и не может быть оправдано требованиями точности, если по временной переменной t решение достаточно гладкое. С другой стороны, при использовании неявной схемы можно значительно увеличить шаг по времени τ, однако при переходе от слоя к слою требуется каждый раз решать систему уравнений. Впрочем, в одномерном случае это не представляет проблемы. В частности, используя метод прогонки, можно получить u^{n+1} при известном u^n за $O(M)$ операций, т. е. количество арифметических операций при переходе от слоя к слою по порядку будет тем же, что и в случае явной схемы. Это позволяет сделать вывод о том, что использование неявных схем в одномерном случае часто является более предпочтительным, так как ведет к уменьшению затрат времени счета на ЭВМ.

Перейдем теперь к исследованию устойчивости разностных схем (3), (5) и (4), (5) в других нормах, в частности в сеточных аналогах норм L_2 и W_2^1 по слою. Так как исследование свойств схем (3), (5) и (4), (5) проводится по одной и той же методике, то имеет смысл объединить эти две схемы. Как и ранее, для наглядности сопоставим разностной схеме ее шаблон. В нашем случае для схем (3), (5) и (4), (5) шаблоны имеют вид, изображенный соответственно на рис. 10.5.1 и 10.5.2.

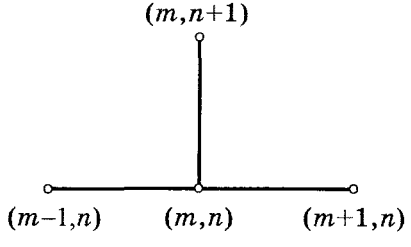


Рис. 10.5.1

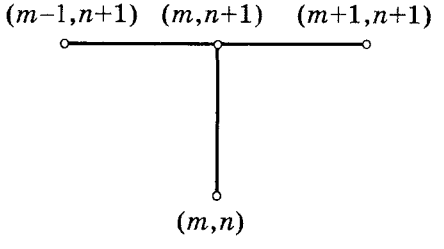


Рис. 10.5.2

Пусть v — функция, определенная на слое и принимающая в m-м узле значение v_m. Обозначим $\Lambda v_m = (v_{m+1} - 2v_m + v_{m-1})/h^2$, $\Lambda u|_{(x,t)} = u(x + h, t) - 2u(x, t) + u(x - h, t)$. Введем более общий, чем (3), (5) и (4), (5), вид схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \sigma \Lambda u_m^{n+1} + (1 - \sigma)\Lambda u_m^n + \varphi_m^n, \quad m = 1, \ldots, M - 1. \tag{11}$$

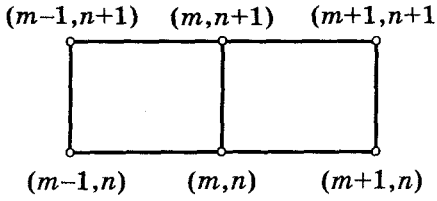


Рис. 10.5.3

Постоянная σ в (11) называется *весом* и обычно берется в пределах $0 \leqslant \sigma \leqslant 1$. В частности, при $\sigma = 0$ выражение (11) переходит в (3) , а при $\sigma = 1$ получаем (4). Разностную схему (11), (5) называют *схемой с весами*. Она имеет шеститочечный шаблон при $\sigma \in (0, 1)$ (рис. 10.5.3). Схема (11), (5) является явной лишь при $\sigma = 0$. Разностную схему (4), (5), чтобы отличить ее от других неявных схем вида (11) с $0 < \sigma < 1$, называют *чисто неявной схемой*.

Пусть

$$L^{\sigma}_{h,\tau}\, u^h\Big|_{(m,\,n)} = \frac{u_m^{n+1} - u_m^n}{\tau} - \sigma\Lambda u_m^{n+1} - (1-\sigma)\Lambda u_m^n.$$

Как и выше, назовем величину

$$r = L^{\sigma}_{h,\tau}[u]_h - \varphi^h, \tag{12}$$

где $[u]_h$ — значения решения в узлах сетки $\bar{Q}_{h,\tau}$, *погрешностью аппроксимации* разностной схемы (11) уравнения (1). Заметим, что граничные и начальные условия для функции u^h выполняются точно. Используя разложение u в ряд Тейлора в точке $(x,\, t+\tau/2) = (mh,\, n\tau+\tau/2)$, имеем

$$\frac{u(x,\,t+\tau) - u(x,\,t)}{\tau} = \frac{\partial u}{\partial t}\Big|_{(x,\,t+\tau/2)} + \frac{\tau^2}{24}\frac{\partial^3 u}{\partial t^3}\Big|_{(x,\,t+\xi)},$$

$$\Lambda u\Big|_{(x,\,t+\frac{\tau}{2}\pm\frac{\tau}{2})} = \Lambda u\Big|_{(x,\,t+\frac{\tau}{2})} \pm \frac{\tau}{2}\Lambda\frac{\partial u}{\partial t}\Big|_{(x,\,t+\frac{\tau}{2})} + O(\tau^2) =$$

$$= \frac{\partial^2 u}{\partial x^2}\Big|_{(x,\,t+\frac{\tau}{2})} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}\Big|_{(x,\,t+\frac{\tau}{2})} \pm \frac{\tau}{2}\Lambda\frac{\partial u}{\partial t}\Big|_{(x,\,t+\frac{\tau}{2})} + O(\tau^2 + h^4).$$

Отсюда

$$r_m^n = L^{\sigma}_{h,\tau}[u] - \varphi^h\Big|_{(m,\,n)} = f\left(x,\, t+\frac{\tau}{2}\right) -$$

$$- \varphi_m^n + \tau(\sigma - 0{,}5)\Lambda\frac{\partial u}{\partial t}\Big|_{(x,\,t+\frac{\tau}{2})} + O(\tau^2 + h^2).$$

Таким образом, если $\varphi_m^n = f(x,\, t+\tau/2)$, то $r = O(h^2+\tau)$ при $\sigma \neq 0{,}5$ и $r = O(h^2+\tau^2)$ при $\sigma = 0{,}5$.

Будем исследовать устойчивость по начальным данным, т. е. будем оценивать чувствительнось решения к возмущению начальных данных. Положим $\mu_k(t) \equiv 0$, $\varphi^h \equiv 0$. Обозначим

$$\|u^n\|_{L_{2,h}} = \left(\sum_{m=1}^{M-1} h(u_m^n)^2 \right)^{1/2}.$$

Назовем разностную схему *устойчивой по начальным данным в норме* $L_{2,h}$, если существует постоянная c_1, не зависящая от шагов сетки h и τ, такая, что для решения u^h задачи (11), (5) с $\mu_1 = \mu_2 = \varphi_h \equiv 0$ справедлива оценка

$$\max_{0 \leqslant n \leqslant N} \|u^n\|_{L_{2,h}} \leqslant c_1 \|u^0\|_{L_{2,h}}. \tag{13}$$

В теории разностных схем установилась традиция, когда не делают различия между матрицей и порождаемым ею линейным оператором.

Обозначим через Λ оператор (матрицу), который функции v со значениями $v_0 = 0, v_1, \ldots, v_{M-1}$, $v_M = 0$ в узлах $0, \ldots, M$ ставит в соответствие функцию со значениями в тех же узлах равными $0, \Lambda v_1, \ldots, \Lambda v_{M-1}, 0$. В случае $\mu_1 = \mu_2 = \varphi_h \equiv 0$ уравнение (11) можно записать в виде

$$u^{n+1} = Su^n.$$

Матрица S называется *матрицей* или *оператором перехода* от слоя к слою. В общем случае S может зависеть от n. Пусть $\{\lambda_i\}$, $i = 1, \ldots, M-1$, — собственные числа матрицы S. Матрица S симметрична и поэтому $\|S\|_2 = \max\limits_i |\lambda_i|$. Функция u_m^0 может быть представлена в виде дискретной суммы Фурье

$$u_m^0 = \sum_{k=1}^{M-1} c_k \sin \frac{\pi k (mh)}{X}.$$

Из равенств

$$\Lambda \sin \frac{\pi k m h}{X} = -\nu_k \sin \frac{\pi k m h}{X}, \qquad \nu_k = \frac{4 \sin^2 \dfrac{\pi k h}{2}}{h^2},$$

следует, что числа $-\nu_k$ являются собственными значениями оператора Λ. Из (11) получаем $(E - \sigma\tau\Lambda)u^1 = (E + \tau(1-\sigma)\Lambda)u^0$, т. е. $S = (E - \sigma\tau\Lambda)^{-1}(E + \tau(1-\sigma)\Lambda)$. Поэтому

$$u_m^1 = S\left(\sum_{k=1}^{M-1} c_k \sin \frac{\pi k (mh)}{X} \right) = \sum_{k=1}^{M-1} \frac{1 - \tau(1-\sigma)v_k}{1 + \tau\sigma v_k} c_k \sin \frac{\pi k (mh)}{X}.$$

Таким образом, собственные числа λ_k матрицы S имеют вид

$$\lambda_k = \frac{1 - \tau(1-\sigma)\nu_k}{1 + \tau\sigma\nu_k}.$$

Выясним, когда будет выполнено условие $|\lambda_k| \leqslant 1$. Заменяя λ_k его выражением через ν_k, получаем условие

$$-1 \leqslant \frac{1-\tau(1-\sigma)\nu_k}{1+\tau\sigma\nu_k} \leqslant 1.$$

Так как ν_k, $\tau > 0$, то эти неравенства эквивалентны соотношениям

$$-1-\tau\sigma\nu_k \leqslant 1+\tau\sigma\nu_k-\tau\nu_k \leqslant 1+\tau\sigma\nu_k.$$

Правое неравенство выполнено всегда, а левое — при $\tau(1-2\sigma)\nu_k \leqslant 2$. При $\sigma \geqslant 1/2$ последнее неравенство будет выполняться при любом $\tau > 0$, а при $\sigma < 1/2$ — если

$$\tau \leqslant \frac{X^2h^2}{2(1-2\sigma)} \leqslant \frac{2}{(1-2\sigma)\max\limits_k \nu_k} \leqslant \frac{h^2}{2(1-2\sigma)}. \tag{14}$$

Таким образом, нами получены достаточные условия устойчивости схемы (11), (5) по начальным данным. А именно, если $\varphi \equiv 0$ и μ_1, $\mu_2 \equiv 0$, то при $\sigma \geqslant 1/2$ разностная схема (11), (5) безусловно устойчива; при $\sigma < 1/2$ схема устойчива, если шаги h и τ связаны соотношением (14), т. е. схема (11), (5) в этом случае условно устойчива.

Предположим, что условие $|\lambda_k| \leqslant 1$ нарушается, т. е. что $|\lambda_{M-1}| \geqslant 1+\delta$, где $\delta > 0$ и не зависит от h и τ. Положим $u_m^0 = \sin\dfrac{\pi(M-1)mh}{X}$. Тогда

$$u_m^n = (1+\bar{\delta})^n \sin\frac{\pi(M-1)mh}{X} \text{ и } \|u^n\| = (1+\bar{\delta})^n\|u^0\|, \text{ где } \bar{\delta} \geqslant \delta > 0.$$

В этом случае $\max\limits_{n,\,n\tau\leqslant T}\|u^n\| \to \infty$ при $\tau \to 0$, т. е. схема неустойчива.

Иногда используют несколько отличное определение устойчивости по начальным данным. Говорят, что разностная схема устойчива по начальным данным, если собственные числа оператора перехода лежат в круге радиуса $1+c\tau$. Покажем, что в рассматриваемом примере это определение согласуется с (13). Действительно, пусть λ_k — собственные числа матрицы S. Тогда $u^n = S^n u^0 = \sum\limits_{k=1}^{M-1} \lambda_k^n c_k \sin\dfrac{\pi kn}{X}$ и $\|u^n\| \leqslant (1+c\tau)^n\|u^0\|$. В этом случае при $n\tau = \text{const}$ и $\tau \to 0$ имеем $\|u^n\| \leqslant e^{cT}\|u^0\|$, $n\tau \leqslant T$.

При исследовании разностных схем для более сложных задач, например при других типах граничных условий, более общем операторе в правой части уравнения (1) и т. п., доказательство устойчивости с использованием принципа максимума или метода Фурье вызывает определенные затруднения, а иногда исследование устойчивости этими приемами является просто невозможным. В этом случае исследование устойчивости разностных схем обычно проводится методом энергетических оценок.

Опишем кратко его суть на дифференциальном уровне. Пусть $u(x, t)$ — решение задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f, \quad u(0, t) = u(X, t) = 0, \quad u(x, 0) = u_0(x). \tag{15}$$

Будем предполагать, что правая часть $f(x, t)$ и функция $u_0(x)$ таковы, что при любом $t \in [0, T]$ существует интеграл $\int_0^X \left(\frac{\partial u}{\partial x}\right)^2 dx < \infty$ и $\frac{\partial u}{\partial t}$ непрерывна по t. Умножим обе части уравнения (15) на u и проинтегрируем по x. Используя формулу интегрирования по частям, получаем *энергетическое тождество*

$$\frac{1}{2}\int_0^X \frac{\partial}{\partial t} u^2\, dx + \int_0^X \left(\frac{\partial u}{\partial x}\right)^2 dx = \int_0^X f u\, dx. \tag{16}$$

Для функции $\varphi(x, t)$, которая при любом $t \in [0, T]$ принадлежит пространству $\overset{\circ}{W}{}_2^1[0, X]$, обозначим $\|\varphi(t)\|_1^2 = \int_0^X \left(\frac{\partial \varphi}{\partial x}\right)^2 dx$.

Если функция $f(x, t)$ такова, что для любой $g \in \overset{\circ}{W}{}_2^1[0, X]$ существует интеграл $\int_0^X f(x, t) g(x)\, dx$, то через $\|f(t)\|_{-1}$ обозначим норму

$$\|f(t)\|_{-1} = \sup_{g \in \overset{\circ}{W}{}_2^1} \frac{1}{\|g\|_1} \int_0^X |fg|\, dx.$$

Согласно определению $\|\cdot\|_{-1}$ имеем

$$\left|\int_0^X f u\, dx\right| \leqslant \|f(t)\|_{-1} \|u(t)\|_1 \leqslant \varepsilon \|u(t)\|_1^2 + \frac{1}{4\varepsilon} \|f(t)\|_{-1}^2.$$

При получении последней оценки было использовано *ε-неравенство*

$$|ab| \leqslant \varepsilon a^2 + \frac{1}{4\varepsilon} b^2, \tag{17}$$

которое следует из соотношений

$$0 \leqslant \left(\sqrt{\varepsilon} a \pm \frac{1}{2\sqrt{\varepsilon}} b\right)^2 = \varepsilon a^2 + \frac{b^2}{4\varepsilon} \pm ab.$$

В данном случае мы обозначили $\|u(t)\|_1$ через a, а $\|f(t)\|_{-1}$ — через b. Тогда из (16) следует неравенство

$$\frac{1}{2}\frac{\partial}{\partial t}\int_0^X u^2\, dx + (1 - \varepsilon)\|u(t)\|_1^2 \leqslant \frac{1}{4\varepsilon}\|f(t)\|_{-1}^2.$$

Для определенности можно было бы положить $\varepsilon = 0{,}5$. Проинтегрируем последнее неравенство по t в пределах от нуля до T. В результате получим

$$\frac{1}{2}\|u(T)\|^2 + (1-\varepsilon)\int_0^T \|u(t)\|_1^2\,dt \leqslant \frac{1}{2}\|u_0\|^2 + \frac{1}{4\varepsilon}\int_0^T \|f(t)\|_{-1}^2\,dt.$$

Здесь $\|u(t)\| = \left(\int_0^X u^2(x,\,t)\,dx\right)^{1/2}$. Последнее неравенство называется *энергетическим неравенством*. Из него, в частности, следует, что решение $u(x,\,t)$ непрерывно зависит от правой части и начальных условий.

Применим несколько похожую схему к исследованию устойчивости сеточной задачи (11), (5). Напомним, что u^n — значение u^h на n-м слое, т. е. $u^n(mh) = u^h(mh,\,n\tau)$. В этом случае уравнение (11) может быть переписано в виде

$$u_t^n = \sigma\Lambda u^{n+1} + (1-\sigma)\Lambda u^n + \varphi^n, \tag{18}$$

где $u_t^n = (u^{n+1} - u^n)/\tau$. В пространстве функций на слое (с нулевыми граничными условиями) введем скалярное произведение и нормы:

$$(v,\,w) = \sum_{m=1}^{M-1} h v_m w_m, \quad \|v\|^2 = (v,\,v),$$

$$\|v\|_1^2 = \sum_{m=0}^{M-1} h\left(\frac{v_{m+1} - v_m}{h}\right)^2.$$

Заметим, что

$$u^{n+1} = \frac{1}{2}(u^{n+1} + u^n) + \frac{\tau}{2}u_t^n, \quad u^n = \frac{1}{2}(u^{n+1} + u^n) - \frac{\tau}{2}u_t^n,$$

поэтому (18) можно преобразовать к виду

$$u_t^n = \frac{1}{2}\Lambda(u^{n+1} + u^n) + \tau(\sigma - 0{,}5)\Lambda u_t^n + \varphi^n. \tag{19}$$

Умножим обе части (19) скалярно на $2\tau u_t^n$. Получим

$$\begin{aligned} 2\tau\|u_t^n\|^2 = {} & \big(\Lambda(u^{n+1} + u^n),\,u^{n+1} - u^n\big) + \\ & + 2\tau^2(\sigma - 0{,}5)(\Lambda u_t^n,\,u_t^n) + 2\tau(u_t^n,\,\varphi^n). \end{aligned} \tag{20}$$

Из формулы суммирования по частям (9.8.14)

$$\sum_{m=1}^{M-1} h\frac{a_{m+1} - a_m}{h}b_m = -\sum_{m=1}^{M} h\frac{b_m - b_{m-1}}{h}a_m + a_M b_M - a_1 b_0,$$

положив $a_m = \dfrac{v_m - v_{m-1}}{h}$, $b_m = v_m$, имеем

$$(\Lambda v,\, v) = \sum_{m=1}^{M-1} h \frac{v_{m+1} - 2v_m + v_{m-1}}{h^2} v_m = \\ = \sum_{m=0}^{M-1} h \left(\frac{v_{m+1} - v_m}{h} \right)^2 = -\|v\|_1^2.$$

Оператор Λ является симметричным (см. гл. 9), поэтому

$$\left(\Lambda u^{n+1} + u^n),\, u^{n+1} - u^n\right) = \left(\Lambda u^{n+1},\, u^{n+1}\right) - (\Lambda u^n,\, u^n) = \\ = \|u^n\|_1^2 - \|u^{n+1}\|_1^2.$$

Используя полученные соотношения, преобразуем (20) к виду

$$2\tau \|u_t^n\|^2 + \|u^{n+1}\|_1^2 + 2\tau^2(\sigma - 0{,}5)\|u_t^n\|_1^2 = \|u^n\|_1^2 + 2\tau(u_t^n,\, \varphi^n). \qquad (21)$$

Полученное равенство по аналогии с непрерывным случаем называют *энергетическим тождеством.*

Оценим скалярное произведение в правой части (21) при помощи ε-неравенства $|(u_t^n,\, \varphi^n)| \leqslant \varepsilon \|u_t^n\|^2 + \dfrac{1}{4\varepsilon}\|\varphi^n\|^2$. Тогда из (21) имеем оценку

$$2\tau \left[(1 - \varepsilon)\|u_t^n\|^2 + \tau(\sigma - 0{,}5)\|u_t^n\|_1^2\right] + \|u^{n+1}\|_1^2 \leqslant \|u^n\|_1^2 + \frac{\tau}{2\varepsilon}\|\varphi^n\|^2. \qquad (22)$$

Выясним, при каких σ выражение в квадратных скобках будет неотрицательным. Заметим, что здесь ε — произвольное положительное число, которое до сих пор не было фиксировано. При $0 < \varepsilon \leqslant 1$ условие $\sigma \geqslant 0{,}5$ является достаточным, чтобы выражение в квадратных скобках было неотрицательным. В этом случае, фиксируя $\varepsilon \leqslant 1$ (например, можно положить $\varepsilon = 1$), из (22) получим

$$\|u^{n+1}\|_1^2 \leqslant \|u^n\|_1^2 + \frac{\tau}{2\varepsilon}\|\varphi^n\|^2. \qquad (23)$$

Проведем более детальное исследование устойчивости при $\sigma < 0{,}5$. С учетом неравенства $\|v_1^2\| \leqslant \dfrac{4}{h^2}\|v^2\|$, которое было установлено ранее, из (22) получаем

$$2\tau \left[1 - \varepsilon + \frac{4(\sigma - 0{,}5)\tau}{h^2}\right] \|u_t^n\|^2 + \|u^{n+1}\|_1^2 \leqslant \|u^n\|_1^2 + \frac{\tau}{2\varepsilon}\|\varphi^n\|^2.$$

Поэтому из выполнения соотношения $1 - \varepsilon + 4(\sigma - 0{,}5)\tau/h^2 \geqslant 0$ следует справедливость (23). Таким образом, для справедливости (23) при $\sigma < 0{,}5$ достаточно выполнения соотношения

$$\tau \leqslant \frac{(1 - \varepsilon)h^2}{4(0{,}5 - \sigma)}. \qquad (24)$$

Используя оценку (23) рекуррентным образом, получаем

$$\|u^n\|_1^2 \leqslant \|u^0\|_1^2 + \sum_{j=0}^{n-1} \frac{\tau}{2\varepsilon}\|\varphi^j\|^2, \quad n\tau \leqslant T. \tag{25}$$

Последнее неравенство как раз и означает устойчивость разностной схемы (11), (5) по начальным данным и правой части. При этом сумма в правой части (25) является квадратурной формулой для интеграла $\int_0^{n\tau} \frac{\|f(t)\|^2}{2\varepsilon}\,dt$.

Отметим, что если соответствующий интеграл $\int_0^\infty \|f(t)\|^2\,dt$ сходится, то из (25) следует ограниченность сеточного решения на бесконечном промежутке времени.

Таким образом, схема (11), (5) по доказанному выше является безусловно устойчивой при $\sigma \geqslant 0{,}5$ и условно устойчивой (шаги h и τ удовлетворяют соотношению (24), $\varepsilon > 0$ не зависит от шагов сетки) при $\sigma < 0{,}5$.

Нами практически не рассматривался вопрос об устойчивости по граничным условиям. Дело заключается в следующем. Возьмем функцию $s(x, t) = \mu_1(t)(X - x)/X + \mu_2(t)x/X$. В этом случае функция $v(x, t) = u(x, t) - s(x, t)$ является решением задачи (1) с однородными граничными условиями и правой частью $f + \partial s/\partial t$. Таким образом, если функции μ_1 и μ_2 имеют производные по t, то граничные условия в задаче (1) могут быть сняты описанным способом. В сеточном случае можно иначе свести задачу (11) с неоднородными граничными условиями (5) к задаче с однородными условиями. Пусть u_m^n — решение сеточной задачи. Положим

$$v_m^n = \begin{cases} u_m^n, & 1 \leqslant m \leqslant M-1, \\ 0, & m = 0, \quad m = M. \end{cases}$$

В этом случае v_m^n удовлетворяет однородным граничным условиям (5), начальным условиям (5) и системе уравнений (11) при φ_m^n, замененной на ψ_m^n, где

$$\psi_m^n = \begin{cases} \varphi_m^n, & 2 \leqslant m \leqslant M-2, \\ \varphi_1^n - \dfrac{1-\sigma}{h^2}\mu_1^n - \dfrac{\sigma}{h^2}\mu_1^{n+1}, & m = 1, \\ \varphi_{M-1}^n - \dfrac{1-\sigma}{h^2}\mu_2^n - \dfrac{\sigma}{h^2}\mu_2^{n+1}, & m = M-1. \end{cases}$$

Таким образом, задача (11), (5) с неоднородными граничными условиями может быть записана как задача с однородными граничными условиями и некоторой измененной правой частью.

Получим оценку скорости сходимости. Будем рассматривать схему (11), (5), так как остальные схемы являются ее частным случаем. Пусть u — решение дифференциальной задачи (1), а u^h — решение разностной задачи (11), (5). По определению $L^{\sigma}_{h,\tau}[u] - \varphi^h = r$, где r — погрешность аппроксимации. Рассмотрим разность $z = [u] - u^h$. Она удовлетворяет уравнению

$$\frac{z^{n+1} - z^n}{\tau} = \sigma \Lambda z^{n+1} + (1-\sigma)\Lambda z^n + \varphi^n - f^n + r^n. \tag{26}$$

Таким образом, z^n_m является решением задачи (11) с правой частью r, равной погрешности аппроксимации, и однородными граничными и начальными условиями (5). Если рассматриваемая схема устойчива, то справедлива оценка (25), откуда

$$\|z^n\|_1^2 \leqslant \sum_{j=0}^{n-1} \frac{\tau}{2\varepsilon}\|r^j\|^2 \leqslant \sum_{j=0}^{N-1} \frac{\tau}{2\varepsilon}\|r^j\|^2 \leqslant \frac{T}{2\varepsilon} \max_{0\leqslant j\leqslant N-1} \|r^j\|^2.$$

В силу того что $r = O(h^2+\tau)$ при $\sigma \neq 0{,}5$ и $r = O(h^2+\tau^2)$ при $\sigma = 0{,}5$, из последней оценки получаем, что при выполнении условия устойчивости решения сеточной задачи (11), (5) u^h сходятся к решению дифференциальной задачи u в сеточной норме $\max\limits_n \|z^n\|_1$. При этом порядок скорости сходимости равен порядку аппроксимации схемы.

Таким образом, решение сеточной задачи сходится к решению дифференциальной задачи со скоростью, по порядку совпадающей с порядком аппроксимации разностной схемы.

Сходимость разностных схем была установлена в сеточной норме пространства $\overset{\circ}{W}{}^1_2$ на слое. Для получения оценки скорости сходимости в других нормах следует воспользоваться соответствующими оценками устойчивости либо использовать сеточные теоремы вложения (см. § 9.8). В частности, из теоремы вложения

$$\max_{1\leqslant m\leqslant M-1} |v_m| \leqslant \frac{\sqrt{X}}{2}\|v\|_1$$

следует сходимость разностной схемы с порядком, равным порядку аппроксимации, в сеточной норме пространства C.

Рассмотрим аппроксимацию граничных условий другого типа. Пусть, например,

$$lu \equiv \frac{\partial u}{\partial x} - \alpha u\Big|_{(0,\,t)} = 0. \tag{27}$$

Заменяя в (27) производную $\partial u/\partial x$ разностным отношением, получаем аппроксимацию граничного условия

$$l_h u^n \equiv \frac{u^n_1 - u^n_0}{h} - \alpha u^n_0 = 0. \tag{28}$$

Оценим погрешность такой аппроксимации. Имеем

$$l_h[u] = \frac{u(h,\,t) - u(0,\,t)}{h} - \alpha u(0,\,t) = \\ = lu + \frac{h}{2}\frac{\partial^2 u}{\partial x^2}\Big|_{(0,\,t)} + O(h^2) = \frac{h}{2}\frac{\partial^2 u}{\partial x^2}\Big|_{(0,\,t)} + O(h^2). \tag{29}$$

Таким образом, построенная аппроксимация (28) граничного условия (27) имеет первый по h порядок аппроксимации. Для повышения порядка аппроксимации воспользуемся приемом, который применялся при аппроксимации условия (27) в краевых задачах для обыкновенных дифференциальных уравнений. Из уравнения (1) выразим $\dfrac{\partial^2 u}{\partial x^2}$. Получим $\dfrac{\partial^2 u}{\partial x^2} = \dfrac{\partial u}{\partial t} - f$. Тогда

$$\frac{\partial^2 u}{\partial x^2}\Big|_{(0,\,t)} = \frac{u(0,\,t+\tau) - u(0,\,t)}{\tau} - f(0,\,t) + O(\tau)$$

либо

$$\frac{\partial^2 u}{\partial x^2}\Big|_{(0,\,t)} = \frac{u(0,\,t) - u(0,\,t-\tau)}{\tau} - f(0,\,t) - O(\tau).$$

Подставив это выражение в (29), получим

$$l_h[u] = \frac{h}{2}\left(\frac{u(0,\,t+\tau) - u(0,\,t)}{\tau} - f(0,\,t)\right) + O(h^2 + \tau).$$

Таким образом, аппроксимация граничного условия (27) с порядком $O(h^2 + \tau)$ будет иметь вид

$$\frac{u_1^n - u_0^n}{h} - \alpha u_0^n - \frac{h}{2}\left(\frac{u_0^{n+1} - u_0^n}{\tau} - f_0^n\right) = 0 \tag{30}$$

либо

$$\frac{u_1^{n+1} - u_0^{n+1}}{h} - \alpha u_0^{n+1} - \frac{h}{2}\left(\frac{u_0^{n+1} - u_0^n}{\tau} - f_0^{n+1}\right) = 0. \tag{31}$$

Можно также использовать линейную комбинацию этих условий

$$\sigma\left(\frac{u_1^{n+1} - u_0^{n+1}}{h} - \alpha u_0^{n+1}\right) + (1-\sigma)\left(\frac{u_1^n - u_0^n}{h} - \alpha u_0^n\right) - \\ - \frac{h}{2}\frac{u_0^{n+1} - u_0^n}{\tau} = -\frac{h}{2}\Big(\sigma f_0^{n+1} + (1-\sigma) f_0^n\Big). \tag{32}$$

Из построения следует, что аппроксимация граничного условия (32) имеет порядок $O(h^2 + \tau)$. Непосредственно можно убедиться, что при $\sigma = 0{,}5$ условие (32) аппроксимирует (27) с порядком $O(h^2 + \tau^2)$.

Заметим, что шаблон, на котором осуществляется аппроксимация, содержит в общем случае только четыре узла $(0, n)$, $(1, n)$, $(0, n+1)$, $(1, n+1)$ (см. рис. 10.5.4). Поэтому структура матрицы линейных уравнений для нахождения u^{n+1} в неявных схемах фактически не изменится: матрица будет иметь трехдиагональный вид. Исследование устойчивости разностных схем, аппроксимирующих краевые условия третьего рода, проводится методом энергетических оценок по схеме, описанной выше. По аналогии с построениями, проводившимися для краевых задач для обыкновенных дифференциальных уравнений, можно строить схемы повышенного порядка точности, например $O(h^4+\tau)$ или же $O(h^4+\tau^2)$.

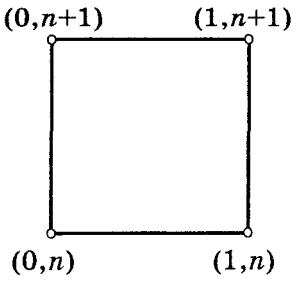


Рис. 10.5.4

§ 6. Разностная аппроксимация эллиптических уравнений

По сравнению с краевыми задачами для обыкновенных дифференциальных уравнений при построении разностных схем в многомерном случае возникают дополнительные трудности, связанные в основном с аппроксимацией граничных условий.

Рассмотрим простейшую краевую задачу — задачу Дирихле для уравнения Пуассона. Пусть область Ω представляет единичный квадрат: $\Omega = \{(x, y),\ 0 < x, y < 1\}$; $\partial\Omega$ — граница Ω. Требуется найти функцию u, дважды непрерывно дифференцируемую внутри Ω и непрерывную в замкнутой области $\bar{\Omega}$, которая внутри области удовлетворяет уравнению

$$-\Delta u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f \tag{1}$$

и принимает на границе заданное значение

$$u|_{\partial\Omega} = \alpha(x, y), \quad (x, y) \in \partial\Omega. \tag{2}$$

В дальнейшем независимые переменные будем обозначать как буквами (x, y) так и (x_1, x_2).

Опишем построение сетки. Разобьем плоскость R^2 прямоугольной сеткой с шагами h_1 и h_2, $h_k = 1/N_k$. Для определенности будем считать, что $N_1 \leqslant N_2$. Точки вида $(x_m, y_n) = (mh_1, nh_2)$ будем называть *узлами сетки* и обозначать их (m, n). Узлы, лежащие внутри Ω, будем называть *внутренними узлами* и множество таких узлов будем обозначать

Ω_h; узлы, которые лежат на $\partial\Omega$, будем называть *граничными узлами* и множество таких узлов будем обозначать $\partial\Omega_h$ (см. рис. 10.6.1).

Пусть $u(x_m, y_n)$ — значение решения в узле (m, n). Сеточную функцию, принимающую в узлах (m, n) значения u_{mn}, будем обозначать u^h.

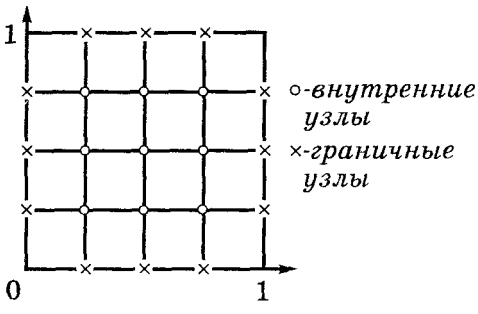


Рис. 10.6.1

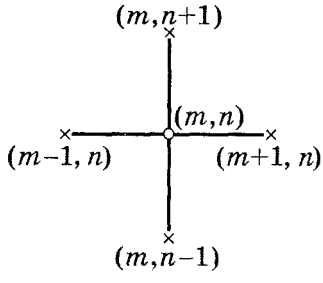


Рис. 10.6.2

Пятиточечный шаблон «крест»; × — точки окрестности узла (m, n)

Заменяя в (1) производные разностными отношениями, получим систему уравнений

$$-\Delta^h u_{mn} = f_{mn}, \quad m = 1, \ldots, N_1 - 1, \quad n = 1, \ldots, N_2 - 1; \tag{3}$$

здесь $\Delta^h = \delta_1^2/h_1^2 + \delta_2^2/h_2^2$, где операторы δ_1^2 и δ_2^2 определены соотношениями

$$\delta_1^2 u_{mn} = u_{m+1,n} - 2u_{mn} + u_{m-1,n}, \quad \delta_2^2 u_{mn} = u_{m,n+1} - 2u_{mn} + u_{m,n-1}.$$

Граничные условия заменим на следующие:

$$u_{mn}|_{\partial\Omega_h} = \alpha(mh_1, nh_2); \quad (mh_1, nh_2) \in \partial\Omega_h. \tag{4}$$

Соотношения (3), (4) будем называть разностной схемой, аппроксимирующей задачу (1), (2). Функция u^h — решение (3), (4) — определена на сетке $\bar{\Omega}_h = \Omega_h \cup \partial\Omega_h$.

Совокупность узлов (m, n), $(m+1, n)$, $(m-1, n)$, $(m, n+1)$, $(m, n-1)$, соответствующих значениям u^h, входящим в уравнение (3), образует шаблон разностной схемы. Уравнение Пуассона (1), таким образом, аппроксимируется на *пятиточечном шаблоне «крест»*. Если (m, n) — внутренний узел, то *окрестностью* этого узла будем называть остальные точки шаблона (рис. 10.6.2). Если в окрестности узла есть точки границы, то этот узел будем называть *приграничным*. Заметим, что значения u^h на границе $\partial\Omega_h$ известны и поэтому могут быть исключены из системы уравнений (3), (4). А именно, подставляя в (3) значения из (4) и

перенося известные члены в правую часть, получим систему линейных алгебраических уравнений

$$L_h u_{mn} = \varphi_{mn}, \quad m = 1, \ldots, N_1 - 1, \quad n = 1, \ldots, N_2 - 1. \tag{5}$$

Нетрудно видеть, что уравнения (5) отличаются от (3) лишь в приграничных узлах. Так, например, в узлах вида $(1, n)$ уравнения (5) будут выглядеть следующим образом:

$$\frac{2u_{1n} - u_{2n}}{h_1^2} + \frac{2u_{1n} - u_{1,n+1} - u_{1,n-1}}{h_2^2} = f_{1n} + \frac{\alpha(0, nh_2)}{h_1^2} \equiv \varphi_{1n}.$$

Число уравнений в системе (5) совпадает с числом неизвестных. Поэтому матрицу системы уравнений (5) можно трактовать как некоторый линейный оператор, отображающий пространство сеточных функций, определенных на Ω_h, в себя.

Опишем подробно структуру матрицы системы уравнений (5) в случае $h_1 = h_2 = h$. Упорядочим компоненты вектора неизвестных $\mathbf{v}$ «естественным» образом:

$$\mathbf{v} = (u_{11}, u_{21}, \ldots, u_{N_1-1,1}, u_{12}, \ldots, u_{N_1-1,N_2-1})^T$$

и умножим обе части (5) на h^2. Тогда матрица системы линейных уравнений будет иметь блочно-трехдиагональную форму.

$$\begin{pmatrix} A_{11} & A_{12} & 0 & \ldots & \cdot & \cdot & 0 \\ A_{21} & A_{22} & A_{23} & \ldots & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \ldots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \ldots & A_{N_1-2,N_2-3} & A_{N_1-2,N_2-2} & A_{N_1-2,N_2-1} \\ 0 & \cdot & \cdot & \ldots & 0 & A_{N_1-1,N_2-2} & A_{N_1-1,N_2-1} \end{pmatrix},$$

где матрицы A_{ij} размера $(N_1 - 1)(N_1 - 1)$ имеют вид

$$A_{ii} = \begin{pmatrix} 4 & -1 & 0 & \ldots & \cdot & 0 \\ -1 & 4 & -1 & \ldots & \cdot & 0 \\ \cdot & \cdot & \cdot & \ldots & \cdot & \cdot \\ 0 & \cdot & \cdot & \ldots & 4 & -1 \\ 0 & \cdot & \cdot & \ldots & -1 & 4 \end{pmatrix}, \quad A_{i,i\pm1} = \begin{pmatrix} -1 & 0 & \ldots & \cdot & 0 \\ 0 & -1 & \ldots & \cdot & 0 \\ \cdot & \cdot & \ldots & \cdot & \cdot \\ 0 & \cdot & \ldots & -1 & 0 \\ 0 & \cdot & \ldots & 0 & -1 \end{pmatrix}.$$

Оценим погрешность аппроксимации схемы (3), (4). При $u \in C^4(\bar{\Omega})$ имеют место соотношения

$$\left.\frac{\delta_k^2 u}{h_k^2}\right|_{(m,n)} = \left.\frac{\partial^2 u}{\partial x_k^2}\right|_{(m,n)} + \frac{h_k^2}{12}\left.\frac{\partial^4 u}{\partial x_k^4}\right|_P, \quad k = 1, 2; \tag{6}$$

поэтому

$$r_{mn} \equiv (-\Delta^h[u]_h)\Big|_{(m,n)} - f_{mn} = -\left(\frac{\delta_1^2 u}{h_1^2} + \frac{\delta_2^2 u}{h_2^2}\right)\Big|_{(m,n)} - f_{mn} =$$

$$= -\left(\frac{h_1^2}{12}\frac{\partial^4 u}{\partial x_1^4}\Big|_{(m,n)} + \frac{h_2^2}{12}\frac{\partial^4 u}{\partial x_2^4}\Big|_{(m,n)}\right) + o(h_1^2 + h_2^2) = O(h_1^2 + h_2^2).$$

При подстановке точного решения u в (4) обнаруживаем, что краевые условия (2) выполняются точно; r^h является погрешностью аппроксимации разностной схемы. Из проведенных рассмотрений следует, что разностная схема (3), (4) имеет второй порядок аппроксимации.

Исследуем разрешимость системы уравнений (3), (4).

Лемма 1 (Сеточный принцип максимума). *Пусть функция v^h определена на $\bar{\Omega}_h$ и в узлах Ω_h выполняется условие $\Delta^h v^h \geqslant 0$. Тогда хотя бы в одной точке границы $\partial\Omega_h$ функция v^h достигает наибольшего значения.*

Доказательство. Допустим противное, т. е. что максимальное значение достигается во внутреннем узле (вообще говоря, таких узлов может быть несколько). Среди всех таких узлов выберем тот, у которого наибольшая абсцисса, т. е. узел (m, n), в котором $v_{mn} > v_{m+1,n}$ и $v_{mn} = \max\limits_{P\in\bar{\Omega}_h} v^h(P)$. Тогда, рассматривая $\Delta^h v^h$ в точке (x_m, y_n), получаем

$$\Delta^h v_{mn} = \frac{v_{m-1,n} - 2v_{mn} + v_{m+1,n}}{h_1^2} + \frac{v_{m,n+1} - 2v_{mn} + v_{m,n-1}}{h_2^2} =$$

$$= \frac{(v_{m-1,n} - v_{mn}) + (v_{m+1,n} - v_{mn})}{h_1^2} +$$

$$+ \frac{(v_{m,n+1} - v_{mn}) + (v_{m,n-1} - v_{mn})}{h_2^2} < 0,$$

что противоречит условию леммы; дело заключается в том, что $v_{m+1,n} - v_{nm} < 0$, а остальные выражения, стоящие в круглых скобках, неотрицательны, поскольку (x_m, y_n) — точка максимума. Таким образом, исходная предпосылка является неверной. Утверждение леммы доказано.

Доказанный принцип максимума справедлив и в случае областей более общего вида.

Аналогично доказывается

Лемма 2. *Пусть v^h определена на $\bar{\Omega}_h$ и в узлах Ω_h выполнено условие $\Delta^h v^h \leqslant 0$. Тогда наименьшее значение достигается функцией v^h хотя бы в одной точке границы $\partial\Omega_h$.*

Из лемм 1 и 2 непосредственно следует

Теорема 1. *Пусть v^h определена на $\bar{\Omega}_h$ и в узлах Ω_h удовлетворяет уравнению*

$$\Delta^h v_{mn} = 0; \quad m = 1, \ldots, N_1 - 1; \quad n = 1, \ldots, N_2 - 1.$$

Тогда v^h достигает своего наибольшего по модулю значения на границе $\partial\Omega_h$.

Теорема 1 является разностным аналогом принципа максимума для гармонических функций. Из нее следует, что система уравнений (3), (4) с $f_{mn} \equiv 0$ и $\alpha_{mn} \equiv 0$ имеет только нулевое решение, поскольку наибольшее по модулю значение u_{mn} равно нулю. Следовательно, определитель системы линейных уравнений (3), (4) (число уравнений равно числу неизвестных) отличен от нуля и при любых f^h и α^h система (3), (4) имеет единственное решение. Заметим также, что отсюда следует однозначная разрешимость системы уравнений (5) при любой правой части φ.

Конкретизируем общие построения § 1 этой главы. Пусть U^h, F^h и G^h — некоторые пространства функций, определенных на $\bar{\Omega}_h$, Ω_h и $\partial\Omega_h$. Введем в них нормы, согласованные с нормами соответствующих пространств в непрерывном случае. Согласно определению (см. § 1) разностная задача (3), (4) будет устойчивой, если существует постоянная c_1, не зависящая от h_1, h_2, такая, что для решения u^h системы уравнений (3), (4) справедлива оценка

$$\|u^h\|_{U^h} \leqslant c_1\Big(\|f^h\|_{F^h} + \|\alpha^h\|_{G^h}\Big). \tag{7}$$

Исследуем устойчивость схемы (3), (4) и оценим близость u к u^h. Пусть нормы в U^h, F^h и G^h заданы следующим образом:

$$\|u^h\|_{U^h} = \max_{\bar{\Omega}_h} |u_{mn}|, \quad \|f^h\|_{F^h} = \max_{\Omega_h} |f_{mn}|, \quad \|\alpha^h\|_{G^h} = \max_{\partial\Omega_h} |\alpha_{mn}|.$$

Вследствие (6) для любого многочлена $Q(x_1, x_2)$ второй степени выполняется равенство

$$\Delta^h Q_{mn} = \Delta\, Q|_{(x_m, y_n)},$$

так как четвертые производные, входящие в (6), обращаются в нуль. Возьмем $R = \sqrt{2}/2 = (\operatorname{diam}\Omega)/2$ и построим вспомогательную функцию

$$Q(x_1, x_2) = \frac{1}{4}\left[R^2 - \left(x_1 - \frac{1}{2}\right)^2 - \left(x_2 - \frac{1}{2}\right)^2\right] \|f^h\|_{F^h} + \|\alpha^h\|_{G^h},$$

которую будем рассматривать в узлах сетки $\bar{\Omega}_h$. Из вышесказанного следует, что в любом внутреннем узле

$$\Delta^h Q_{mn} = \Delta\, Q|_{(m,n)} = -\|f^h\|_{F^h}, \; m = 1, \ldots, N_1 - 1, \; n = 1, \ldots, N_2 - 1.$$

Тогда разность $v^h = u^h - Q$ в узлах Ω_h удовлетворяет неравенству

$$\Delta^h v = f^h + \|f^h\|_{F^h} \geqslant 0.$$

По лемме 1 функция v^h принимает наибольшее значение на границе $\partial\Omega_h$. Но на $\partial\Omega_h$ справедливо отношение

$$v^h = \alpha^h - Q = \alpha^h - \|\alpha^h\|_{G^h} - \frac{1}{4}\left[R^2 - \left(x_1 - \frac{1}{2}\right)^2 - \left(x_2 - \frac{1}{2}\right)^2\right]\|f^h\|_{F^h} \leqslant 0.$$

Таким образом, $v^h \leqslant 0$, т.е. $u^h \leqslant Q$ в $\bar{\Omega}_h$. Аналогично, рассматривая функцию $v^h = u^h + Q$ в Ω_h, устанавливаем, что

$$\Delta^h v^h \leqslant 0 \quad \text{и} \quad v^h\Big|_{\partial\Omega_h} \geqslant 0.$$

Тогда из леммы 2 следует оценка $v^h \geqslant 0$ или же $u^h \geqslant -Q$. Таким образом, всюду в $\bar{\Omega}_h$ установлено неравенство $|u^h| \leqslant Q$ и поэтому

$$\|u^h\|_{U^h} \leqslant \|Q\|_{U^h} \leqslant \frac{1}{4}R^2\|f^h\|_{F^h} + \|\alpha^h\|_{G^h}. \tag{8}$$

Заменяя это неравенство более сильным

$$\|u^h\|_{U_h} \leqslant \|f^h\|_{F^h} + \|\alpha^h\|_{G^h},$$

получаем оценку (7). Таким образом, разностная схема (3), (4) устойчива в сеточном аналоге нормы пространства C.

Оценим сходимость разностной схемы (3), (4). Для этого запишем уравнение для погрешности $R^h(mh, nh) = R_{mn} = u(x_m, y_n) - u_{mn}$:

$$-\Delta^h R_{mn} = -\Delta^h[u]_h\Big|_{(m,n)} + \Delta^h u_{mn} = r_{mn};$$

здесь r_{mn} — погрешность аппроксимации. В силу того, что граничные условия выполняются точно, имеем

$$R^h\Big|_{\partial\Omega_h} = 0.$$

Так как радиус R области Ω равен $\sqrt{2}/2$, то используя оценку (8), получаем

$$\|R^h\|_{U^h} \leqslant \frac{1}{8}\|r^h\|_{F^h} = O(h_1^2 + h_2^2).$$

Таким образом, решение сеточной задачи (3), (4) сходится к точному решению дифференциальной задачи в сеточной норме пространства C. Напомним, что все рассуждения проводились в предположении, что решение задачи (1), (2) обладает достаточной гладкостью, а именно что $u(x, y)$ имеет непрерывные четвертые производные в $\bar{\Omega}$.

Из доказательства сходимости видно, что основным моментом являлось получение оценки (8), характеризующей устойчивость разностной схемы. Сходимость же схемы является следствием аппроксимации и устойчивости, причем порядок скорости сходимости совпадает с порядком аппроксимации. Проведенное выше доказательство сходимости схемы является частным случаем теоремы Филиппова.

Описанный метод дает приближенное решение, сходящееся к точному со скоростью $O(h^2)$. По аналогии с одномерным случаем можно построить разностные схемы, обладающие более высоким порядком сходимости. Наметим пути получения более точных схем. Предположим, что решение u непрерывно дифференцируемо шесть раз в замкнутой области $\bar{\Omega}$. Тогда вместо (6) можно написать равенство

$$\left.\frac{\delta_k^2 u}{h_k^2}\right|_{(m,n)} = \left.\frac{\partial^2 u}{\partial x_k^2}\right|_{(m,n)} + \frac{h_k^2}{12}\left.\frac{\partial^4 u}{\partial x_k^4}\right|_{(m,n)} + O(h_k^4).$$

Дифференцируя (1) два раза по x, получаем

$$\frac{\partial^4 u}{\partial x^4} = -\frac{\partial^4 u}{\partial x^2 \partial y^2} - \frac{\partial^2 f}{\partial x^2}.$$

Таким образом,

$$\left.\frac{\delta_k^2 u}{h_1^2}\right|_{(m,n)} = \left.\frac{\partial^2 u}{\partial x^2}\right|_{(m,n)} - \frac{h_1^2}{12}\left.\left(\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^2 f}{\partial x^2}\right)\right|_{(m,n)} + O(h_1^4).$$

Заменяя производные в правой части разностными отношениями, имеем

$$\left.\frac{\delta_1^2 u}{h_1^2}\right|_{(m,n)} = \left.\frac{\partial^2 u}{\partial x^2}\right|_{(m,n)} - \frac{h_1^2}{12}\left.\left(\frac{\delta_1^2\delta_2^2 u}{h_1^2 h_2^2} + \frac{\delta_1^2 f}{h_1^2}\right)\right|_{(m,n)} + O(h_1^4 + h_2^4).$$

Аналогично устанавливаем, что

$$\left.\frac{\delta_2^2 u}{h_2^2}\right|_{(m,n)} = \left.\frac{\partial^2 u}{\partial y^2}\right|_{(m,n)} - \frac{h_2^2}{12}\left.\left(\frac{\delta_1^2\delta_2^2 u}{h_1^2 h_2^2} + \frac{\delta_2^2 f}{h_2^2}\right)\right|_{(m,n)} + O(h_1^4 + h_2^4).$$

Складывая полученные равенства, получаем

$$\left.\left(\frac{\delta_1^2 u}{h_1^2} + \frac{\delta_2^2 u}{h_2^2}\right)\right|_{(m,n)} = \Delta u|_{(m,n)} - \frac{h_1^2 + h_2^2}{12}\cdot\left.\frac{\delta_1^2\delta_2^2 u}{h_1^2 h_2^2}\right|_{(m,n)} -$$
$$- \frac{1}{12}\left.\left(\delta_1^2 f + \delta_2^2 f\right)\right|_{(m,n)} + O\left(h_1^4 + h_2^4\right).$$

Искомая разностная схема четвертого порядка аппроксимации будет иметь вид

$$-\Delta^h u_{mn} - \frac{h_1^2+h_2^2}{12}\cdot\frac{\delta_1^2\delta_2^2 u}{h_1^2h_2^2}\Big|_{(m,n)} = f_{mn} + \frac{1}{12}\left(\delta_1^2 f_{mn} + \delta_2^2 f_{mn}\right); \tag{9}$$

$$m = 1,\ldots,N_1-1, \quad n = 1,\ldots,N_2-1.$$

Нетрудно видеть, что шаблон схемы (9) состоит из девяти точек (рис. 10.6.3).

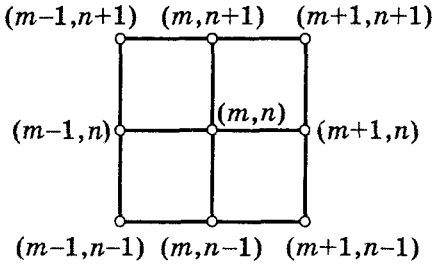


Рис. 10.6.3
Шаблон схемы 4-го порядка

Схемы более высокого порядка, в отличие от одномерного случая, будут содержать тем большее количество узлов, чем выше порядок аппроксимации разностной схемы. В случае, если область Ω является объединением прямоугольников со сторонами, параллельными осям координат, и $h_1 = h_2$, при гладком решении рассмотренная схема будет иметь четвертый порядок сходимости. Более точно, если решение u исходной задачи имеет шестые ограниченные производные в замкнутой области $\bar{\Omega}$, то справедлива оценка

$$\max_{\bar{\Omega}_h}\left|u_{mn} - u(x_m, y_n)\right| \leqslant ch^4,$$

где u_{mn} — решение системы сеточных уравнений (9), (4).

Пусть Ω является объединением конечного числа прямоугольников со сторонами, параллельными осям координат, причем стороны, параллельные оси x_1, лежат на прямых $x_2 = n_2h_2$, n_2 — целое, а параллельные оси x_2 — на прямых $x_1 = n_1h_1$, n_1 — целое. Тогда построение и исследование разностной схемы проводится аналогично.

Расмотрим наиболее простые методы аппроксимации граничных условий в случае области с криволинейной границей. Ограничимся рассмотрением равномерного шага, т. е. $h_1 = h_2 = h$. Рассмотрим множество прямых $x = mh$, $y = nh$. Точки их пересечения между собой, а также точки их пересечения с $\partial\Omega$ будем называть узлами. Через Γ_h обозначим узлы, лежащие на $\partial\Omega$, а через $\partial\Omega_h$ — узлы (mh, nh) с целыми координатами (m, n), лежащие в $\bar{\Omega}$, расстояние от которых до границы $\partial\Omega$, измеряемое вдоль направления какой-либо из осей координат, меньше h. Остальные узлы сетки, лежащие в Ω, обозначим Ω_h. Тогда в каждом

узле (m, n) (т. е. в точке с координатами (mh, nh)) сеточной области Ω_h можно записать уравнение

$$-\Delta^h u_{mn} = f_{mn}, \quad (m, n) \in \Omega_h. \tag{10}$$

Простейший способ аппроксимации граничных условий заключается в сносе граничных условий в узлы $\partial\Omega_h$, т. е. полагаем

$$u_{mn} = \alpha(x, y), \quad (m, n) \in \partial\Omega_h, \tag{11}$$

где (x, y) — ближайшая к узлу (m, n) точка границы. В этом случае Ω_h — внутренние узлы сеточной области, а $\partial\Omega_h$ — граничные. Нетрудно видеть, что при таком способе задания граничных условий порядок аппроксимации будет $O(h)$. Исследование устойчивости и сходимости разностной схемы (10), (11) проводится точно так же, как для схемы (3), (4), однако в данном случае получается порядок сходимости $O(h)$. Это является следствием довольно грубой аппроксимации граничных условий.

Рассмотрим еще один способ аппроксимации граничных условий. Назовем узлы $\partial\Omega_h$ приграничными, а узлы Γ_h — границей сеточной области. В узлах Ω_h уравнение (1) заменим уравнением (10); в Γ_h положим $u^h|_{\Gamma_h} = \alpha^h$, т. е. в этом случае граничные условия выполняются точно. Пусть (m, n) — узел $\partial\Omega_h$. Для определенности будем считать, что узлы $(m+1, n)$, $(m, n+1)$, $(m, n-1)$ не лежат на Γ_h (отрезки, соединяющие их с узлом (m, n), принадлежат Ω), а узел $(m-1, n)$ лежит на Γ_h. (В этом случае узел $(m-1, n)$ соответствует точке $((m-1)h\theta, nh) \in \partial\Omega$, $0 < \theta < 1$ (см. рис. 10.6.4).) Тогда аппроксимация (1) в приграничном узле (m, n) берется в виде

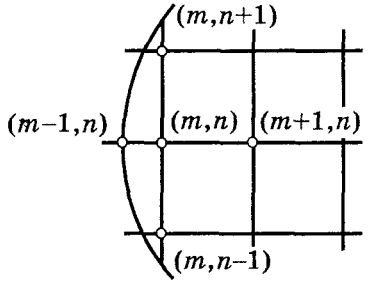


Рис. 10.6.4
Аппроксимация в приграничных узлах.

$$\Delta^h u_{mn} = \frac{1}{h}\left(\frac{u_{m+1,n} - u_{mn}}{h} - \frac{u_{mn} - u_{m-1,n}}{h^*}\right) + \\ + \frac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2}. \tag{12}$$

Здесь h^* — расстояние между узлами $(m-1, n)$ и (m, n). Аналогично осуществляется аппроксимация уравнения (1) в других узлах $\partial\Omega_h$.

Итак, в узлах Ω_h уравнение (1) аппроксимируется обычным образом, а аппроксимация на неравномерной сетке используется только в узлах $\partial\Omega_h$. Поэтому узлы Ω_h называют *регулярными*, а узлы $\partial\Omega_h$ *нерегулярными*. Аппроксимация уравнения (1) в нерегулярных узлах в данном случае имеет порядок $O(1)$.

Для рассмотренной схемы имеет место

Теорема 2 (без доказательства). *Если решение задачи* (1) $u \in C^4(\bar{\Omega})$, *то разностная схема* (10) *с аппроксимацией* (12) *в нерегулярных узлах имеет второй порядок сходимости в сеточной норме пространства* C, *т. е.*

$$\|u^h - u\|_{u^h} \leqslant ch^2.$$

Замечание. Довольно распространенным является другой способ аппроксимации уравнения (1) в нерегулярных узлах. А именно, полагают (см. рис. 10.6.4)

$$\Delta^h u_{mn} = \frac{2}{h + h^*}\left(\frac{u_{m+1,n} - u_{mn}}{h} - \frac{u_{mn} - u_{m-1,n}}{h^*}\right) + \frac{\delta_2^2 u_{mn}}{h^2}. \qquad (13)$$

В этом случае погрешность аппроксимации в таких узлах имеет порядок $O(h)$ и теорема 2 будет справедлива. Однако если систему линейных уравнений путем исключения граничных значений привести к виду (5), то получится система уравнений с несимметричной матрицей. Таким образом, в данном случае сеточная задача теряет важное свойство, присущее исходной задаче — симметричность оператора.

При исследовании краевых задач с другими граничными условиями и эллиптическими дифференциальными операторами более общего вида, а также краевых задач для систем уравнений в частных производных принцип максимума, разностный аналог которого использовался при исследовании устойчивости и сходимости разностной схемы, вообще говоря, не имеет места. Кроме этого, часто бывает необходимо оценивать не только близость получаемого приближения к точному решению, но и близость их производных. Все это приводит к необходимости создания методов исследования разностных схем, не использующих принцип максимума. Как и ранее, проиллюстрируем методику исследования на модельной задаче Дирихле для уравнения Пуассона в прямоугольнике (1), (2) и на соответствующей разностной схеме (3), (4).

Пусть $h_1 = h_2 = h$. Исключая граничные условия (4) из (3), получаем систему уравнений (5). Обозначим через H линейное пространство сеточных функций, определенных на Ω_h. Таким образом, элементы H можно рассматривать как векторы размерности $(N_1 - 1) \times (N_2 - 1)$; компоненты этих векторов являются значениями функций в узлах Ω_h. Матрица L_h системы уравнений (5) порождает линейный невырожденный оператор в H. Тогда если $u^h = \{u_{mn}\}$, $\varphi^h = \{\varphi_{mn}\}$, то система уравнений (5) может быть записана в операторной форме:

$$L_h u^h = \varphi^h. \qquad (14)$$

Здесь L^h — оператор, соответствующий матрице системы уравнений (5). Введем в H скалярное произведение

$$(u, v) = \sum_{P \in \Omega_h} h^2 u(P) v(P), \quad u, v \in H. \tag{15}$$

Скалярное произведение (15) согласовано со скалярным произведением функкций в $L_2(\Omega)$, т. е.

$$\lim_{h \to 0} \Big([g_1]_h, [g_2]_h \Big) = \int_\Omega g_1 g_2 \, dx$$

для любых непрерывных функций из $L_2(\Omega)$.

Пусть $\|v\|^2 = (v, v)$. Справедлива

Лемма 3. *Оператор L_h является симметричным и положительно определенным на H, и для него выполнена оценка*

$$\gamma_1 \|v\|^2 \leqslant (L_h v, v) \leqslant \gamma_2 \|v\|^2, \tag{16}$$

где $0 < a_1 \leqslant \gamma_1$, а $\gamma_2 \leqslant a_2 h^{-2}$; a_i не зависят от h.

Доказательство. Заметим прежде всего, что симметричность оператора L_h следует из симметричности соответствующей ему матрицы. Проведем, однако, доказательство этого факта другим методом. Пусть $v \in H$; обозначим через $\tilde{v}$ сеточную функцию, совпадающую с v на Ω_h и равную нулю на $\partial\Omega_h$. Тогда из определения L_h можно записать

$$(L_h v)_{mn} = (-\Delta^h \tilde{v})_{mn}, \qquad 1 \leqslant m, n \leqslant N - 1.$$

Представим L_h в виде $L_h = L_1 + L_2$, где

$$L_1 v\Big|_{mn} = \frac{-\tilde{v}_{m+1,n} + 2v_{mn} - \tilde{v}_{m-1,n}}{h^2},$$

$$L_2 v\Big|_{mn} = \frac{-\tilde{v}_{m,n+1} + 2v_{mn} - \tilde{v}_{m,n-1}}{h^2},$$

т. е. L_1 и L_2 являются одномерными сеточными операторами, соответствующими дифференциальным операторам $-\dfrac{\partial^2}{\partial x^2}$ и $-\dfrac{\partial^2}{\partial y^2}$. Покажем, что L_k симметричен и положительно определен. Пусть для определенности $k = 1$. Используя формулу суммирования по частям (9.8.14), получим

$$\begin{aligned}
(L_1 v, w) &= \sum_{m,n=1}^{N-1} h^2 \frac{-\tilde{v}_{m+1,n} + 2v_{mn} - \tilde{v}_{m-1,n}}{h^2} \tilde{w}_{mn} = \\
&= \sum_{n=1}^{N-1} h \left(\sum_{m=1}^{N-1} h \frac{-\tilde{v}_{m+1,n} + 2v_{mn} - \tilde{v}_{m-1,n}}{h^2} \tilde{w}_{mn} \right) = \\
&= \sum_{n=1}^{N-1} \sum_{m=1}^{N} h^2 \left(\frac{\tilde{v}_{m,n} - \tilde{v}_{m-1,n}}{h} \right) \left(\frac{\tilde{w}_{m,n} - \tilde{w}_{m-1,n}}{h} \right).
\end{aligned}$$

Функции $\tilde{v}$ и $\tilde{w}$ входят в правую часть равенства симметричным образом, поэтому

$$(L_1 v,\, w) = \sum_{n=1}^{N-1}\sum_{m=1}^{N} h^2 \left(\frac{\tilde{v}_{m,n} - \tilde{v}_{m-1,n}}{h}\right)\left(\frac{\tilde{w}_{m,n} - \tilde{w}_{m-1,n}}{h}\right) = (L_1 w,\, v),$$

т. е. оператор L_1 симметричен. С другой стороны, из разностного аналога теоремы вложения (§ 9.8) получаем

$$(L_1 v,\, v) = \sum_{n=1}^{N-1}\sum_{m=1}^{N} h^2 \left(\frac{\tilde{v}_{m,n} - \tilde{v}_{m-1,n}}{h}\right)^2 \geqslant$$

$$\geqslant \frac{1}{4}\sum_{n=1}^{N-1} h \max_{0<m<N} |v_{mn}|^2 \geqslant \frac{1}{4}\|v\|^2.$$

Используя аналогичную оценку для оператора L_2, получаем

$$(L_h v,\, v) \geqslant \frac{1}{2}\|v\|^2,$$

т. е. левая часть (16) доказана.

В другую сторону оценка получается намного проще. Поскольку $(\tilde{v}_{m,n} - \tilde{v}_{m-1,n})^2 \leqslant 2(\tilde{v}^2_{m,n} + \tilde{v}^2_{m-1,n})$, то

$$(L_1 v,\, v) \leqslant \sum_{n=1}^{N-1}\sum_{m=1}^{N} h^2 \frac{2(\tilde{v}^2_{m,n} + \tilde{v}^2_{m-1,n})}{h^2} = \frac{4}{h^2}\|v\|^2,$$

откуда

$$(L_h v,\, v) \leqslant \frac{4}{h^2}\|v\|^2.$$

Лемма доказана.

Задача 1. В случае прямоугольника со сторонами l_1 и l_2 показать, что минимальным и максимальным собственными значениями оператора L_h являются соответственно

$$\lambda_{\min} = 4\left(\frac{\sin^2 \dfrac{\pi h_1}{2l_1}}{h_1^2} + \frac{\sin^2 \dfrac{\pi h_1^2}{2l_2}}{h_2^2}\right), \qquad \lambda_{\max} = 4\left(\frac{\cos^2 \dfrac{\pi h_1}{2l_1}}{h_1^2} + \frac{\cos^2 \dfrac{\pi h_1^2}{2l_2}}{h_2^2}\right).$$

Таким образом, (16) выполняется при $\gamma_1 = \lambda_{\min}$ и $\gamma_2 = \lambda_{\max}$, что по порядку при $h_1 = h_2$ совпадает с оценками, полученными выше.

Из проведенных выше рассуждений вытекает возможность введения в пространстве H нормы

$$\|v\|_1^2 = (L_h v,\, v),$$

которую называют энергетической. Название связано с тем, что в непрерывном случае при u, имеющем физический смысл отклонения мембраны и $u\,|_{\partial\Omega} = 0$, выражение $\frac{1}{2}(Lu, u)$ пропорционально потенциальной энергии мембраны.

Будем исследовать устойчивость разностной схемы (14) в H. Умножим обе части (14) скалярно в H на u^h; применяя (16), получим

$$\|u^h\|_1^2 = (\varphi^h, u^h) \leqslant \|\varphi^h\|\,\|u^h\| \leqslant \gamma_1^{-1}\|\varphi^h\|\,\|u^h\|_1.$$

Поэтому справедлива оценка

$$\|u^h\|_1 \leqslant \gamma_1^{-1}\|\varphi^h\|, \tag{17}$$

что означает устойчивость разностной схемы. Отсюда, в частности, следует, что система уравнений (14) при $\varphi^h \equiv 0$ имеет лишь тривиальное решение $u^h \equiv 0$, т. е. мы еще раз доказали, что (14) разрешима при любой правой части единственным образом.

Оценим скорость сходимости разностной схемы (3), (4) в энергетической норме. Как и ранее, будем предполагать, что решение u дифференциальной краевой задачи (1), (2) имеет непрерывные четвертые производные в замкнутой области. Тогда погрешность $R^h = [u]_h - u^h$ будет удовлетворять уравнению

$$L_h R^h = r \equiv L_h[u]_h - L_h u^h, \tag{18}$$

где $r = O(h^2)$ — погрешность аппроксимации. Применяя оценку (17), имеем

$$\|R^h\|_1 \leqslant \gamma_1^{-1}\|r\| = O(h^2).$$

Таким образом, рассматриваемая разностная схема имеет второй порядок сходимости по h в энергетической норме.

При исследовании скорости сходимости в энергетической норме мы предполагали, что решение имеет непрерывные четвертые производные. Оказывается, что это требование является завышенным и тот же результат можно получить в предположении, что решение обладает только третьими непрерывными производными в $\bar{\Omega}$. Это связано с тем, что погрешность аппроксимации имеет дивергентный (дипольный) характер. Используя формулу Тейлора, получаем

$$\frac{u(x_{m+1}, y_n) - 2u(x_m, y_n) + u(x_{m-1}, y_n)}{h^2} = \frac{h}{6}\left[\left.\frac{\partial^3 u}{\partial x^3}\right|_{(\xi_{m+1}, y_n)} - \right.$$

$$\left. - \left.\frac{\partial^3 u}{\partial x^3}\right|_{(\xi_m, y_n)}\right] = \frac{h^2}{6}\left(\frac{\psi^{(1)}_{m+1,\,n} - \psi^{(1)}_{mn}}{h}\right), \quad m = 1, \ldots, N-1;$$

здесь $x_{i-1} \leqslant \xi_i \leqslant x_i$. Аналогичная формула получается и для разности по второй переменной. Следовательно погрешность будет удовлетворять уравнению (18) с правой частью, равной

$$r_{mn} = \frac{h^2}{6}\left(\frac{\psi^{(1)}_{m+1,\,n} - \psi^{(1)}_{mn}}{h} + \frac{\psi^{(2)}_{m,\,n+1} - \psi^{(2)}_{mn}}{h}\right).$$

Умножим обе части (18) скалярно в H на R^h. Используя формулы суммирования по частям, правую часть преобразуем к виду

$$(r,\,R^h) = \frac{h^2}{6}\sum_{m,\,n=1}^{N-1} h^2\left(\frac{\psi^{(1)}_{m+1,\,n} - \psi^{(1)}_{mn}}{h}R_{mn} + \frac{\psi^{(2)}_{m,\,n+1} - \psi^{(2)}_{mn}}{h}R_{mn}\right) =$$

$$= \frac{h^2}{6}\sum_{m,\,n=1}^{N-1} h^2\left(\frac{\tilde{R}_{mn} - \tilde{R}_{m-1,\,n}}{h}\psi^{(1)}_{mn} + \frac{\tilde{R}_{mn} - \tilde{R}_{m,\,n-1}}{h}\psi^{(2)}_{mn}\right).$$

Таким образом,

$$\left|(r,\,R^h)\right| \leqslant \frac{h^2}{6}\,\|R^h\|_1\left(\|\psi^{(1)}\| + \|\psi^{(2)}\|\right) \leqslant ch^2\,\|R^h\|_1;$$

поэтому $\gamma_1\,\|R^h\|_1^2 \leqslant ch^2\,\|R^h\|_1$, откуда следует $\|R^h\|_1 = O(h^2)$.

При построении разностной схемы, аппроксимирующей уравнение (1) с краевыми условиями второго или третьего рода, можно воспользоваться методами, которые применялись в одномерном случае. Пусть для определенности рассматривается задача для уравнения (1) с краевым условием третьего рода

$$\left(\frac{\partial u}{\partial n} + \theta u\right)\Big|_{\partial\Omega} = \alpha, \quad \theta(s) \geqslant 0, \quad s \in \partial\Omega. \tag{19}$$

Напомним, что в качестве области Ω мы рассматриваем единичный квадрат. Для определенности рассмотрим аппроксимацию условия (19) на участке границы $x = 1$. Тогда, заменяя производную в (19) разностным соотношением, получим

$$\frac{u_{Nn} - u_{N-1,\,n}}{h} + \theta_{Nn}u_{Nn} = \alpha_{Nn}. \tag{20}$$

Найдем погрешность аппроксимации

$$\frac{u(1,\,nh) - u(1-h,\,nh)}{h} + \theta(1,\,nh)u(1,\,nh) - \alpha(1,\,nh) =$$

$$= \left(\frac{\partial u}{\partial x} - \frac{h}{2}\,\frac{\partial^2 u}{\partial x^2} + \theta u - \alpha\right)\Big|_{(1,\,nh)} + O(h^2) =$$

$$= -\frac{h}{2}\,\frac{\partial^2 u}{\partial x^2}\Big|_{(1,\,nh)} + O(h^2).$$

Таким образом, погрешность аппроксимации граничных условий (20) имеет первый порядок по h. Выражая $\dfrac{\partial^2 u}{\partial x^2}$ из уравнения (1), получаем $\dfrac{\partial^2 u}{\partial x^2} = -f - \dfrac{\partial^2 u}{\partial y^2}$. Тогда

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(1,\,nh)} = -\left.\left(f + \frac{\partial^2 u}{\partial y^2}\right)\right|_{(1,\,nh)} = -f_{Nn} - \frac{u_{N,n+1} - 2u_{Nn} + u_{N,n-1}}{h^2} + O(h^2).$$

Поэтому если рассмотреть аппроксимацию граничного условия (19) вида

$$\frac{u_{Nn} - u_{N-1,n}}{h} + \theta_{Nn} u_{Nn} - \alpha_{Nn} - \\ - \frac{h}{2}\left(f_{Nn} + \frac{u_{N,n+1} - 2u_{Nn} + u_{N,n-1}}{h^2}\right) = 0, \tag{21}$$

то в силу проведенных выше построений получаем, что соотношение (21) аппроксимирует краевое условие (19) с порядком $O(h^2)$. Приводя подобные члены, преобразуем (21) к виду

$$\frac{2u_{Nn} - u_{N-1,n} - 0{,}5(u_{N,n+1} + u_{N,n-1})}{h} + \theta_{Nn}\, u_{Nn} = \alpha_{Nn} + \frac{h}{2} f_{Nn}. \tag{22}$$

Отсюда ясно, как будет записываться аппроксимация граничного условия в других узлах $\partial\Omega_h$. В частности, с помощью подобных рассуждений получаем аппроксимацию в угловом узле $(N,\,0)$:

$$\frac{2u_{N0} - u_{N-1,0} - u_{N1}}{h} + 2\theta_{N0}\, u_{N0} = 2\alpha_{N0} + h f_{N0}. \tag{23}$$

Заметим, что в данном случае граница сеточной области включает угловые точки.

Другой способ аппроксимации граничного условия (19) опирается на то, что берется другая сетка. Рассмотрим множество узлов

$$\bar{\Omega}_h = \{x = (x_1,\, x_2);\ x_k^j = jh - h/2,\ j = 0, \ldots,\, N+1\}, \quad h = N^{-1},$$

и пусть Ω_h — множество узлов сетки, лежащих в Ω. Тогда, как и ранее, уравнение (1) можно аппроксимировать в узлах Ω_h обычной пятиточечной схемой. Краевые условия (19) на такой сетке будем аппроксимировать следующим образом:

$$\frac{u_{N+1,\,n} - u_{Nn}}{h}\,\theta\,(Nh,\, nh) + \frac{u_{N+1,\,n} + u_{Nn}}{2} = \alpha\,(Nh,\, nh)\,. \tag{24}$$

Предположим, что решение дифференциальной задачи может быть продолжено за пределы области Ω с сохранением свойств гладкости. Тогда выражение (24) аппроксимирует краевое условие (19) на участке границы $x = 1$ с порядком $O(h^2)$. Это можно непосредственно проверить, подставляя решение u в (24) и воспользовавшись формулой Тейлора в точке $(1,\, nh)$.

Метод конечных элементов. До сих пор рассматривалась разностная задача Дирихле для уравнения Пуассона, которая строилась непосредственно путем замены производных в дифференциальном уравнении разностными отношениями. Аналогично случаю краевых задач для обыкновенных дифференциальных уравнений рассмотрим способы построения дискретных аппроксимаций, основанные на вариационных и проекционных принципах. Будем рассматривать краевую задачу (1), (2) с однородными граничными условиями.

На множестве непрерывно дифференцируемых функций, обращающихся в нуль на границе $\partial\Omega$, введем норму

$$\|v\|_{H^1} = \left\{ \int_\Omega \left[\left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right] dxdy \right\}^{1/2} = \left(\int_\Omega (\nabla v)^2 \, dxdy \right)^{1/2}. \qquad (25)$$

Замыкание множества таких функций в этой норме является гильбертовым пространством; обозначим его через H^1 (ранее мы его обозначали $\overset{\circ}{W}{}_2^1$).

Рассмотрим задачу о нахождении минимума функционала

$$\min_{v \in H^1} \Phi(v) = \min_{v \in H^1} \left\{ \int_\Omega (\nabla v)^2 \, dxdy - 2 \int_\Omega f v \, dxdy \right\} \qquad (26)$$

на пространстве функций H^1. Если классическое решение u задачи (1), (2) при $\alpha \equiv 0$ существует и принадлежит H^1, то оно дает минимум функционалу (26). Обратное, вообще говоря, неверно; функция, доставляющая минимум функционалу (26) на H^1, не обязательно должна обладать вторыми производными.

Таким образом, задачу нахождения решения (1), (2) можно заменить задачей о нахождении минимума квадратичного функционала (26) на H^1. Решение, получаемое в результате минимизации функционала (26) на H^1, является обобщенным решением краевой задачи (1), (2).

Для построения *вариационно-разностной схемы* воспользуемся методом Ритца. Аппроксимируем H^1 некоторым конечномерным подпространством V^h; в методе Ритца за приближенное решение задачи (26) принимают функцию $v \in V^h$, минимизирующую функционал (26) на подпространстве V^h.

Подпространство V^h построим следующим способом. Пусть

$$\bar{\Omega}_h = \{(x, y);\ x = mh,\ y = nh;\ 0 \leqslant m, n \leqslant N\}.$$

Разобьем $\bar{\Omega}$ на квадратные ячейки со стороной h и вершинами в узлах $\bar{\Omega}_h$. Каждую ячейку $\Omega_{mn} = \{(x, y),\ mh \leqslant x \leqslant (m+1)h,\ nh \leqslant y \leqslant (n+1)h\}$ разобьем диагональю, проходящей через вершины (m, n), $(m+1, n+1)$. Таким образом, вся область $\bar{\Omega}$ будет разбита на прямоугольные треугольники с катетами, равными h. Эти треугольники будем называть *элементарными*, а разбиение области $\bar{\Omega}$ на треугольники — *триангуляцией обла-*

сти $\bar{\Omega}$. В качестве подпространства V^h пространства H^1 возьмем пространство непрерывных в $\bar{\Omega}$ функций, линейных на каждом элементарном треугольнике и обращающихся в нуль на $\partial\Omega$. Каждая функция из V^h *однозначно определяется своими значениями в узлах* $\bar{\Omega}_h$ и, обратно, каждая сеточная функция, которая принимает в узлах сетки заданные значения, однозначно определяет функцию из V^h. При этом функция из V^h называется *кусочно-линейным восполнением сеточной функции.* Таким образом, существует взаимно однозначное соответствие между H и V^h, где H — пространство сеточных функций, определенных на $\bar{\Omega}_h$ и принимающих нулевые значения на $\partial\Omega_h$. Функции $\varphi^h_{mn} \in H$:

$$\varphi^h_{mn} = \begin{cases} 1, & (x, y) = (mh, nh), \\ 0, & (x, y) \neq (mh, nh), \end{cases}$$

образуют базис в H. Соответствующие им кусочно-линейные функции φ_{mn} из V^h, принимающие значения, равные единице в узле (m, n) и нулю в других узлах, будут образовывать базис в V^h.

В качестве приближенного решения задачи (26) возьмем функцию $v^h \in V^h$, которая минимизирует функционал (26) на подпространстве V^h, т. е.

$$\min_{v \in V^h} \Phi(v) = \Phi(v^h). \tag{27}$$

Предположим, что v^h существует; представим ее в виде

$$v^h = \sum_{i,j=1}^{N-1} v_{ij}\varphi_{ij},$$

где v_{ij} — неизвестные коэффициенты разложения. Отметим, что v_{ij}, в силу выбора функций φ_{ij}, является значением v^h в точке (i, j). Таким образом, отыскание приближенного решения состоит в определении коэффициентов v_{ij}. Выпишем уравнения для определения этих коэффициентов. В точке минимума функции $\Phi(\sum v_{ij}\varphi_{ij})$ должны выполняться равенства

$$\frac{\partial \Phi(\sum v_{ij}\varphi_{ij})}{\partial v_{mn}} = 0; \quad m, n = 1, \ldots, N-1.$$

Вычислим левую часть этого соотношения:

$$\frac{\partial \Phi}{\partial v_{mn}} = \frac{\partial}{\partial v_{mn}} \int_\Omega \left[\left(\sum v_{ij} \frac{\partial \varphi_{ij}}{\partial x} \right)^2 + \left(\sum v_{ij} \frac{\partial \varphi_{ij}}{\partial y} \right)^2 - \right.$$

$$\left. - 2f \sum v_{ij}\varphi_{ij} \right] dxdy = 2 \int_\Omega \left[\sum v_{ij} \left(\frac{\partial \varphi_{ij}}{\partial x} \frac{\partial \varphi_{mn}}{\partial x} + \right. \right.$$

$$\left. \left. + \frac{\partial \varphi_{ij}}{\partial y} \frac{\partial \varphi_{mn}}{\partial y} \right) - f\varphi_{mn} \right] dxdy.$$

Следовательно, система уравнений относительно v_{ij} будет иметь вид

$$\sum_{i,j=1}^{N-1} v_{ij} \int_\Omega \left(\frac{\partial \varphi_{ij}}{\partial x} \frac{\partial \varphi_{mn}}{\partial x} + \frac{\partial \varphi_{ij}}{\partial y} \frac{\partial \varphi_{mn}}{\partial y} \right) dxdy = \tag{28}$$

$$= \int_\Omega f \varphi_{mn}\, dxdy, \quad m, n = 1, \ldots, N-1.$$

Количество уравнений в (28) совпадает с количеством неизвестных.

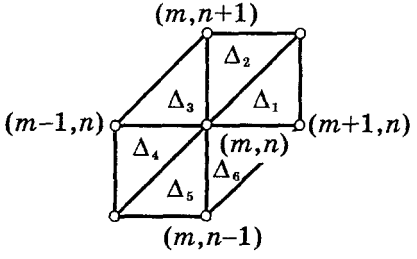


Рис. 10.6.5
Носитель функции φ_{mn}

Функция φ_{mn} отлична от нуля лишь в тех элементарных треугольниках, которые имеют узел (m, n) своей вершиной. Поэтому в каждом из уравнений (28) интегрирование ведется не по всей области Ω, а лишь по пересечению таких треугольников с Ω. Множество точек, где $\varphi_{mn} \neq 0$, образует шестиугольник (см. рис. 10.6.5). Обозначим этот шестиугольник через S_{mn}, а входящие в него треугольники — через $\Delta_1, \ldots, \Delta_6$. Положим

$$J^1_{mn}(i, j) = \int_\Omega \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x}\, dxdy;$$

поскольку $\dfrac{\partial \varphi_{mn}}{\partial x} \equiv 0$ в Δ_2 и Δ_5, то

$$J^1_{mn}(i, j) = \int_\Omega \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x}\, dxdy =$$

$$= \int_{\Delta_1 \cup \Delta_6} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x}\, dxdy + \int_{\Delta_3 \cup \Delta_4} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{ij}}{\partial x}\, dxdy.$$

Отсюда видно, что $J^1_{mn} \neq 0$ лишь при $j = n$ и $i = m-1$, $i = m$, $i = m+1$. Проводя соответствующие вычисления, получаем

$$J^1_{mn}(m, n) = \int_{\Delta_1 \cup \Delta_6} \left(\frac{\partial \varphi_{mn}}{\partial x} \right)^2 dxdy + \int_{\Delta_3 \cup \Delta_4} \left(\frac{\partial \varphi_{mn}}{\partial x} \right)^2 dxdy = 2,$$

$$J^1_{mn}(m+1, n) = J^1_{mn}(m-1, n) = \int_{\Delta_3 \cup \Delta_4} \frac{\partial \varphi_{mn}}{\partial x} \frac{\partial \varphi_{m+1,n}}{\partial x}\, dxdy = -1.$$

Аналогично, для $J^2_{mn}(i, j) = \dfrac{\partial \varphi_{mn}}{\partial y} \dfrac{\partial \varphi_{ij}}{\partial y}\, dxdy$ получаем

$$J^2_{mn}(m, n) = 2, \quad J^2_{mn}(m, n+1) = J^2_{mn}(m, n-1) = -1;$$

в остальных случаях $J^2_{mn}(i, j) = 0$.

Таким образом, уравнение, соответствующее узлу (m, n), записывается в виде

$$4v_{mn} - v_{m+1,n} - v_{m-1,n} - v_{m,n+1} - v_{m,n-1} = h^2 g_{mn},$$

где $g_{mn} = \dfrac{1}{h^2}\displaystyle\int_{\Omega_{mn}} f\varphi_{mn}\,dxdy$. Разделив обе части на h^2, получим систему сеточных уравнений

$$L_h v_{mn} = g_{mn}, \quad 1 \leqslant m, n \leqslant N-1, \tag{29}$$

структура которой полностью совпадает с (5). Единственное отличие заключается в другом способе вычисления правой части. Однако если g_{mn} вычислять приближенно, полагая

$$\frac{1}{h^2}\int_{\Omega_{mn}} f\varphi_{mn}\,dxdy \approx f_{mn}\frac{1}{h^2}\int_{\Omega_{mn}} \varphi_{mn}\,dxdy = f_{mn},$$

то получится разностная схема, полностью идентичная (5).

Так как левая часть системы (29) совпадает с левой частью системы (5), то система уравнений (29) имеет единственное решение при любой f. Справедливы неравенства

$$\|u - u^h\|_{H^1} \equiv \|u - u^h\|_1 \leqslant c_1 h \|f\|_{L_2}, \qquad \|u - u^h\|_{L_2} \leqslant c_1 h^2 \|f\|_{L_2}.$$

Таким образом, в случае описываемого метода требования к гладкости решения существенно меньше, чем в случае применения метода конечных разностей.

Построение разностных схем таким способом особенно целесообразно в случае уравнений и систем с естественными граничными условиями, когда непосредственная аппроксимация граничных условий вызывает затруднения.

В последнее время получили широкое распространение *проекционно-разностные методы* решения краевых задач (*метод конечных элементов*). Описанный выше метод построения разностных схем с помощью метода Ритца является одной из разновидностей метода конечных элементов.

Опишем в общих чертах суть проекционно-разностного подхода на другой модельной задаче (ср. с § 9.11). За основу метода обычно берется интегральное тождество для определения обобщенного решения. Итак, предположим, что в квадрате $\bar{\Omega} = \{x = (x_1, x_2),\ 0 \leqslant x_1, x_2 \leqslant 1\}$ требуется найти решение краевой задачи

$$-\Delta u = f, \quad \left(\frac{\partial u}{\partial n} + \alpha u\right)\Big|_{\Gamma} = 0, \quad u|_{\partial\Omega\setminus\Gamma} = 0, \quad \alpha > 0, \tag{30}$$

где Γ — участок границы, лежащий на прямой $x = 1$. Предположим, что классическое решение этой задачи существует. Умножим обе части уравнения (30) на функцию φ, частные производные которой являются

кусочно-непрерывными и $\varphi\big|_{\partial\Omega\setminus\Gamma} = 0$. Интегрируя по частям и используя краевые условия, получим

$$\int_\Omega \nabla u \nabla\varphi\, dxdy + \int_\Gamma \alpha(s)u(s)\varphi(s)\, ds = \int_\Omega f\varphi\, dxdy. \tag{31}$$

Соотношение (31) называется *интегральным тождеством*; оно имеет место для любой функции $\varphi \in H^1$, где H^1 — пространство, являющееся замыканием множества гладких функций, равных нулю на $\partial\Omega\setminus\Gamma$ в норме (25). Если u является классическим решением задачи (30) и имеет суммируемые в квадрате производные, то оно удовлетворяет (31) и $u \in H^1$. Обратное, вообще говоря, неверно. Можно указать функции u и f, для которых выполнено (31), однако классическое решение задачи (30) существовать не будет. Функцию $u \in H^1$, удовлетворяющую интегральному тождеству (31), называют *обобщенным решением* задачи (30). Обобщенное решение, определяемое из интегрального тождества (31), совпадает с обобщенным решением, определяемым минимизацией функционала

$$\int_\Omega |\nabla v|^2\, dxdy - 2\int_\Omega fv\, dxdy + \int_\Gamma \alpha(s)v^2(s)\, ds.$$

Подобное обстоятельство всегда имеет место в случаях, когда исходный дифференциальный оператор является симметричным и положительно определенным. Если эти условия не выполнены, то задача определения обобщенного решения не может быть сформулирована в терминах минимизации некоторого квадратичного функционала, но может быть сформулирована при помощи интегрального тождества типа (31).

Поступим аналогично предыдущему случаю. Триангулируем область $\bar\Omega$ и введем пространство V^h функций, кусочно-линейных на элементарных треугольниках и обращающихся в нуль на $\partial\Omega\setminus\Gamma$. Приближенным решением задачи (31) назовем функцию $u^h \in V^h$ такую, что для любой $\varphi \in V^h$ выполняется равенство

$$\int_\Omega \nabla u^h \nabla\varphi\, dxdy + \int_0^1 \alpha(y)u^h(1,\, y)\varphi(1,\, y)\, dy = \int_\Omega f\varphi\, dxdy. \tag{32}$$

Таким образом, интегральное тождество (32) совпадает с (31) с той лишь разницей, что в (32) решение и *пробные* функции берутся из подпространства $V^h \subset H^1$.

Функция u^h полностью определяется своими значениями в узлах сетки. Для того чтобы u^h удовлетворяла (32), необходимо и достаточно, чтобы (32) было справедливо для любой функции $\varphi \in V^h$, входящей в базис V^h. В качестве базисных функций возьмем функции из V^h, которые равны единице в одном из узлов $\Omega^h \cup \Gamma_h$ и нулю во всех остальных узлах. Это дает нам систему сеточных уравнений. Обратим внимание на тот факт, что теперь в базис входят функции, которые, вообще говоря, отличны от нуля на Γ.

Представим u^h в виде $u^h = \sum u_{mn}\varphi_{mn}$, где u_{mn} — неизвестные коэффициенты. Подставляя это выражение в (32), получаем

$$\sum_{m,n} u_{mn}\int_\Omega \nabla\varphi_{mn}\nabla\varphi_{ij}\,dxdy + \sum_{m,n} u_{mn}\int_0^1 \alpha(y)\varphi_{mn}(1,\,y)\varphi_{ij}(1,\,y)\,dy = \\ = \int_\Omega f\varphi_{ij}\,dxdy; \quad 1\leqslant i\leqslant N, \quad 1\leqslant j\leqslant N-1. \tag{33}$$

Совокупность соотношений (33) образует систему линейных алгебраических уравнений относительно неизвестных коэффициентов u_{mn} и порождает некоторую *проекционно-разностную схему* (эта схема могла бы быть получена и как вариационно-разностная). Так как любая функция $\varphi\in V^h$ может быть разложена по функциям φ_{mn}, то из выполнения (33) будет следовать справедливость (32). Таким образом, выражения (32) и (33) эквивалентны.

Для доказательства устойчивости (33) положим в (32) $\varphi = u^h$. Тогда

$$\|u^h\|_1^2 \leqslant \int_\Omega (\nabla u^h)^2\,dxdy + \int_0^1 \alpha(y)[u^h(1,\,y)]^2\,dy \leqslant \\ \leqslant \left|\int_\Omega f u^h\,dxdy\right| = \left|(f,\,u^h)\right| \leqslant \|f\|_{-1}\cdot\|u^h\|_1,$$

где

$$\|f\|_{-1} = \sup_{\varphi\in H^1}\frac{|(f,\,\varphi)|}{\|\varphi\|_1}.$$

Таким образом, $\|u^h\|_1 \leqslant \|f\|_{-1}$. Это означает устойчивость проекционно-разностной схемы, т. е. малым изменениям f в норме $\|\cdot\|_{-1}$ соответствуют малые изменения u^h в норме $\|\cdot\|_1$.

Рассмотрим структуру матрицы системы уравнений (33). Если $1\leqslant i\leqslant N-1$, то $\varphi_{ij}|_\Gamma = 0$, поэтому интеграл по Γ в (33) равен нулю. В этом случае выражение (33) совпадает с (28), и, следовательно, уравнение, соответствующее функции φ_{ij}, имеет вид

$$4u_{ij} - u_{i+1,\,j} - u_{i-1,\,j} - u_{i,\,j+1} - u_{i,\,j-1} = h^2 g_{ij},$$

где

$$g_{ij} = \frac{1}{h^2}\int_\Omega f\varphi_{ij}\,dxdy.$$

В точках Γ, т. е. в узлах вида $(N,\,j)$, имеем

$$\frac{2u_{Nj} - u_{N,\,j-1} - u_{N,\,j+1}}{2} + u_{Nj} - u_{N-1,\,j} + \\ + (\alpha_{j-1}^j u_{N,\,j-1} + \alpha_j^j u_{Nj} + \alpha_{j+1}^j u_{N,\,j+1}) = h^2 g_{Nj}.$$

Здесь $\alpha_n^j = \int_\Gamma \varphi_{Nj}\varphi_{Nn}\alpha(y)\,dy$, $h^2 g_{Nj} = \int_\Omega f\varphi_{Nj}\,dxdy$. В случае, когда α_n^j и g_{ij} не вычисляются в явном виде, их можно вычислять приближенно, используя квадратурные формулы. Например,

$$\alpha_n^j = \int_\Gamma \varphi_{Nj}\varphi_{Nn}\alpha(s)\,ds \approx \frac{\alpha(jh)+\alpha(nh)}{2}\int_\Gamma \varphi_{Nj}\varphi_{Nn}\,ds,$$

$$h^2 g_{ij} = \int_\Omega f\varphi_{ij}\,dxdy \approx h^2 f_{ij}.$$

В заключение кратко опишем построение проекционно-разностной схемы в случае криволинейной границы и возможные обобщения этого метода. Пусть для простоты рассматривается задача Дирихле для уравнения Пуассона и Ω — плоская односвязная область с кусочно-гладкой границей $\partial\Omega$, т. е. $\partial\Omega$ состоит из конечного числа гладких дуг, пересекающихся между собой под ненулевыми углами. Зададимся параметром h — шагом сетки. Построим ломаную Γ_h, обладающую следующими свойствами:

а) область Ω_h, ограниченная Γ_h, содержится в $\overline{\Omega}$;

б) между точками $\partial\Omega$ и Γ_h можно установить взаимно однозначное соответствие, т. е. существует взаимно однозначное отображение $\varphi : \partial\Omega \to \Gamma_h$, которое имеет кусочно-непрерывную производную φ', $|\varphi'|$, $|(\varphi^{-1})'| \leqslant C$, где C не зависит от h;

в) расстояние от точек Γ_h до $\partial\Omega$ не превосходит величины $c_1 h^2$, где $c_1 > 0$ — некоторая постоянная, не зависящая от h;

г) длина звеньев ломаной ограничена снизу величиной $c_2 h$, $c_2 > 0$. При принятых допущениях относительно области такое построение всегда возможно.

Разобьем область Ω_h на треугольники (которые будем называть *элементарными*) так, что:

а) длины сторон треугольников находятся в пределах $[c_3 h, c_4 h]$, $c_i > 0$ — постоянные, не зависящие от h;

б) площади треугольников находятся в пределах $[c_5 h^2, c_6 h^2]$;

в) любые два треугольника либо не пересекаются, либо имеют только одну общую сторону, либо общую вершину.

Описанное выше построение называется *квазиравномерной триангуляцией области* Ω; вершины треугольников называются узлами сетки. Можно доказать, что такая триангуляция осуществима. Пусть H — пространство непрерывных функций, кусочно-линейных над элементарными треугольниками Ω_h и обращающихся в нуль на Γ_h. Задаче (1) с однородным граничным условием (2) поставим в соответствие проекционно-разностную задачу — найти функцию $u^h \in H$, удовлетворяющую при любой $\varphi \in H$ соотношению

$$\int_{\Omega_h} \nabla u^h \nabla\varphi\,dxdy = \int_{\Omega_h} f\varphi\,dxdy. \tag{34}$$

Функция u^h полностью определяется своими значениями в узлах сетки. Поэтому если в качестве φ брать базисные функции пространства H (равные единице

в одном узле и нулю в остальных узлах и кусочно-линейные над треугольниками Ω_h), то (34) является системой линейных алгебраических уравнений относительно значений u^h в узлах. Доказательство устойчивости проводится точно так же, как и выше. При исследовании же сходимости следует дополнительно оценить норму решения в W_2^1 в приграничной полосе $\Omega \backslash \Omega_h$. С учетом этих оценок можно получить следующее соотношение:

$$\|u - u^h\|_{W_2^1(\Omega_h)} \leqslant ch,$$

где постоянная c зависит от нормы решения в $W_2^2(\Omega)$.

При построении проекционно-разностной схемы можно использовать более сложные конечные элементы, за счет чего может быть достигнута большая точность. Например, кроме узлов сетки, можно также в качестве узлов рассматривать середины сторон треугольников.

Пусть H — подпространство непрерывных в Ω_h функций, равных нулю на $\partial\Omega_h$, и являющихся полиномом второй степени в каждом элементарном треугольнике Ω_h. Пусть $\partial\Omega_h = \partial\Omega$. В качестве базисных функций в H можно рассмотреть функции, которые равны единице в одном узле, а в других узлах равны нулю и принадлежат H (под узлами здесь понимаются как вершины, так и середины сторон треугольников). Тогда можно получить оценку

$$\|u - u^h\|_{W_2^1(\Omega_h)} \leqslant ch^2.$$

Аналогично могут быть построены проекционно-разностные схемы с более высоким порядком скорости сходимости. При этом используются способы интерполяции, рассмотренные в § 5.5. Заметим, что структура матрицы системы линейных алгебраических уравнений ухудшается; а именно, возрастает количество ненулевых элементов в строке матрицы и ширина ленты.

Заметим, что в случае бигармонического уравнения и области с криволинейной границей такой подход нуждается в уточнении, так как получаемые приближения могут, вообще говоря, не сходиться к точному решению задачи при $h \to 0$.

Кратко осветим историю вопроса. Вариационные и проекционные методы при небольшом числе базисных функций применялись издавна, еще до появления ЭВМ. Применение ЭВМ позволило увеличить число базисных функций; при этом часто возрастало суммарное влияние вычислительной погрешности и погрешности, возникающей при аппроксимации интегралов квадратурными суммами.

Это обстоятельство ставило ограничение на точность, с которой могли быть получены решения с помощью вариационных методов. Были проведены теоретические исследования, показавшие, что для устойчивости вариационных методов существенно выполнение некоторого условия на систему базисных функций, называемого условием *сильной минимальности*. Построение системы базисных функций, удовлетворяющей этому условию, в случае областей сложной формы иногда бывает непросто.

Параллельно шло интенсивное развитие теории и практики применения конечно-разностных методов. Если при использовании классических

вариационных методов для решения линейных задач возникают линейные системы уравнений с полностью заполненной матрицей, то при использовании конечно-разностных уравнений возникают системы уравнений с матрицей, содержащей относительно малое число ненулевых элементов. Это обстоятельство позволяет решать с теми же затратами процессорного времени системы уравнений с существенно большим числом неизвестных. Однако в случае областей сложной формы применение конечно-разностных методов представляет определенные неудобства вследствие неоднородности построения разностных уравнений в приграничных точках.

Получивший в последнее время интенсивное развитие метод конечных элементов свободен от ряда недостатков описанных методов: он не требует специальных усилий по построению системы базисных функций, являющейся сильно минимальной, при его использовании упрощается написание уравнений вблизи границы. Матрица линейной системы уравнений содержит относительно малое число ненулевых элементов. Большая «технологичность» метода позволила создать на его основе ряд *промышленных систем стандартных программ* решения краевых задач, в частности задач теории упругости. При использовании таких систем не требуется знание теории численных методов и тонкостей программирования. Исследователь должен лишь задать триангуляцию области, а часто система и сама осуществляет такую триангуляцию. Эти методы сходятся при меньших требованиях гладкости, чем конечно-разностные методы. В случае квазиравномерных триангуляций базисные функции метода автоматически удовлетворяют условию сильной минимальности.

В то же время увеличивается объем работы при вычислении матрицы системы уравнений. Поэтому при решении крупных задач зачастую все-таки применяют конечно-разностные методы или приходят к составлению систем уравнений с помощью аппроксимации минимизирующего функционала (или интегрального тождества) (см. § 9.12).

Традиционно для решения эллиптических задач применялись методы теории потенциала. С появлением ЭВМ они были практически вытеснены конечно-разностными методами. Однако в последнее время в вычислительную практику стал интенсивно проникать *метод граничных элементов*, имеющий некоторые общие черты с методом потенциала.

§ 7. Решение параболических уравнений с несколькими пространственными переменными

При решении параболических уравнений, как и в случае эллиптических уравнений, переход от одномерного случая к многомерному вызывает существенные затруднения. Поскольку все принципиальные трудности возникают лишь при переходе от одной пространственной переменной к

двум, то в дальнейшем будем рассматривать случай двух пространственных переменных.

Перейдем к построению и исследованию разностных схем. Пусть требуется найти функцию u, являющуюся решением уравнения

$$\frac{\partial u}{\partial t} = \Delta u + f(x, t) \tag{1}$$

в области $Q_T = \Omega \times [0, T]$, $\bar{\Omega} = \{x;\ 0 \leqslant x_i \leqslant 1,\ i = 1, 2\}$ с начальными и граничными условиями

$$u(x, 0) = u_0(x), \quad u(x, t)|_{x \in \Gamma} = \alpha(x, t); \tag{2}$$

здесь $x = (x_1, x_2)$, $\Gamma = \partial\Omega \times [0, T]$.

Попытаемся применить к решению задачи (1), (2) методы, разработанные ранее. Введем в рассмотрение квадратную сетку с шагом $h = 1/M$:

$$\bar{\Omega}_h = \{x;\ x = (ih, jh),\ 0 \leqslant i, j \leqslant M\},$$

а на отрезке $[0, T]$ — сетку с шагом $\tau = T/N$. Будем искать приближенное решение задачи (1), (2) в дискретной системе точек (узлов)

$$\bar{Q}_h = \{(x, t);\ x \in \bar{\Omega}_h,\ t = n\tau,\ n = 0, \ldots, N\}.$$

Множество $\bar{Q}_h$ будем называть *сеточной областью*, а множество точек $\bar{\Omega}_h^n = \{(x, t) \in \bar{Q}_h,\ x \in \Omega_h,\ t = n\tau\}$ при фиксированном $t = n\tau$ будем называть *n-м слоем*.

По аналогии с одномерным случаем построим явную и неявную разностные схемы для задачи (1), (2) и попытаемся выяснить, в чем заключается принципиальное отличие от случая одной пространственной переменной. Заменяя $\dfrac{\partial u}{\partial t}$ в узле (i, j, n) разделенной разностью $\dfrac{u_{ij}^{n+1} - u_{ij}^n}{\tau}$ или $\dfrac{u_{ij}^n - u_{ij}^{n-1}}{\tau}$, а Δu — выражением (см. § 6)

$$\Delta^h u_{ij}^n = \frac{u_{i+1,j}^n + u_{i-1,j}^n + u_{i,j+1}^n + u_{i,j-1}^n - 4u_{ij}^n}{h^2},$$

получаем разностные схемы: явную

$$\begin{gathered} \frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} = \Delta^h u_{ij}^n + f_{ij}^n, \quad 1 \leqslant i, j \leqslant M - 1, \\ u_{ij}^k = \alpha(ih, jh, k\tau),\ (ih, jh) \in \Gamma_h,\ u_{ij}^0 = u_0(ih, jh) \end{gathered} \tag{3}$$

и неявную

$$\begin{gathered} \frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} = \Delta^h u_{ij}^{n+1} + f_{ij}^{n+1}, \quad 1 \leqslant i, j \leqslant M - 1, \\ u_{ij}^k = \alpha(ih, jh, k\tau),\ (ih, jh) \in \Gamma_h,\ u_{ij}^0 = u_0(ih, jh). \end{gathered} \tag{4}$$

При использовании схемы (3) счет ведется по явным формулам — по известным значениям u_{ij}^n из (3) находятся значения u_{ij}^{n+1}. Поэтому проблем с реализацией алгоритма на ЭВМ не возникает. Остается лишь исследовать устойчивость этой схемы.

По-другому обстоит дело в случае схемы (4). Относительно u_{ij}^{n+1}, $1 \leqslant i, j \leqslant M-1$, имеем систему линейных алгебраических уравнений, т. е. схема (4) неявна. Структура матрицы уравнений (4) совпадает со структурой матрицы оператора $-\Delta^h$ (см. § 6). Поэтому при решении этой системы возникают те же трудности, что и в случае эллиптических уравнений. Напомним, что в одномерном случае проблема численного решения уравнений на верхнем слое не возникала, так как можно было воспользоваться методом прогонки.

Введем понятие экономичной разностной схемы. Разностную схему, аппроксимирующую задачу со временем, называют *экономичной*, если она безусловно устойчива и при переходе от слоя к слою требуется количество арифметических операций, пропорциональное числу узлов на слое. (Иногда условие безусловной устойчивости в определении экономичной схемы отсутствует.) Из определения следует, что чисто неявная схема для одномерного уравнения теплопроводности является экономичной.

Перед тем как заниматься построением экономичных разностных схем, исследуем устойчивость разностной схемы в общей постановке.

Введем пространство H функций, определенных на Ω_h; v_{ij} — значение функции $v \in H$ в узле (i, j). Скалярное произведение и норму в H определим как

$$(v, w) = \sum_{i,j=1}^{M-1} h^2 v_{ij} w_{ij}, \quad \|v\|^2 = (v, v).$$

Разностные схемы (3), (4), рассматривавшиеся выше, связывали значения приближенного решения задачи на двух соседних слоях n-м и $(n+1)$-м, поэтому их естественно называть *двухслойными*. Далее будем рассматривать двухслойные разностные схемы вида

$$B \frac{u^{n+1} - u^n}{\tau} + Au^n = \varphi^n, \tag{5}$$

где B и A — симметричные положительно определенные операторы, отображающие H в себя. Как и в одномерном случае, иногда будем различать устойчивость по начальным данным и по правой части.

Обозначим $u_t^n = \dfrac{u^{n+1} - u^n}{\tau}$. Учитывая равенство $u^n = u^{n+1} - \tau u_t^n$, преобразуем (5) к виду

$$(B - \tau A) u_t^n + Au^{n+1} = \varphi^n. \tag{6}$$

Положим $D = B - \tau A$. Так как $u^{n+1} = \dfrac{u^{n+1} + u^n}{2} + \dfrac{\tau}{2} u_t^n$ и оператор D симметричен, то

$$2\tau(Du_t^{n+1}, u^{n+1}) = \tau(Du_t^n, u^{n+1} + u^n) + \tau^2(Du_t^n, u_t^n) =$$

$$= (Du^{n+1}, u^{n+1}) - (Du^n, u^n) + (Du^{n+1}, u^n) - (Du^n, u^{n+1}) +$$

$$+ \tau^2(Du_t^n, u_t^n) = (Du^{n+1}, u^{n+1}) - (Du^n, u^n) + \tau^2(Du_t^n, u_t^n).$$

Поэтому, умножая обе части (6) скалярно в H на $2\tau u^{n+1}$, получим

$$(u^{n+1}, u^{n+1})_D - (u^n, u^n)_D + \tau^2(u_t^n, u_t^n)_D + \\ + 2\tau\|u^{n+1}\|_A^2 = 2\tau(\varphi^n, u^{n+1}). \tag{7}$$

Напомним, что по определению $(v, w)_D = (Dv, w)$, $\|v\|_A^2 = (Av, v)$.

Будем считать, что

$$D = B - \tau A > 0. \tag{8}$$

Тогда (7) можно переписать в виде

$$\|u^{n+1}\|_D^2 - \|u^n\|_D^2 + \tau^2\|u_t^n\|_D^2 + 2\tau\|u^{n+1}\|_A^2 = 2\tau(\varphi^n, u^{n+1}). \tag{9}$$

Соотношение (9) является энергетическим тождеством.

Так как пространство H конечномерно, то существует постоянная κ такая, что $(Dv, v) \leqslant \kappa(Av, v)$ для всех $v \in H$ или, что то же самое,

$$D \leqslant \kappa A. \tag{10}$$

Так как $D = D^* > 0$, то существует оператор (матрица) $D^{1/2}$ симметричный и положительно определенный такой, что $D^{1/2}D^{1/2} = D$. Через $D^{-1/2}$ обозначим оператор $(D^{1/2})^{-1}$. В рассматриваемом случае

$$\|u^{n+1}\|_A^2 \geqslant \kappa^{-1}\|u^{n+1}\|_D^2,$$

$$2\tau\left|(\varphi^n, u^{n+1})\right| = 2\tau\left|(D^{-1/2}\varphi^n, D^{1/2}u^{n+1})\right| \leqslant$$

$$\leqslant 2\tau\|D^{-1/2}\varphi^n\|\,\|D^{1/2}u^{n+1}\| = 2\tau\|\varphi^n\|_{D^{-1}}\|u^{n+1}\|_D \leqslant$$

$$\leqslant \varepsilon\tau\|u^{n+1}\|_D^2 + \frac{\tau}{\varepsilon}\|\varphi^n\|_{D^{-1}}^2$$

и из (9) следует неравенство

$$\left(1 + \frac{2\tau}{\kappa} - \varepsilon\tau\right)\|u^{n+1}\|_D^2 \leqslant \|u^n\|_D^2 + \frac{\tau}{\varepsilon}\|\varphi^n\|_{D^{-1}}^2.$$

Фиксируя ε, например, полагая $\varepsilon = \kappa^{-1}$, отсюда получаем соотношение

$$\left(1 + \frac{\tau}{\kappa}\right)\|u^{n+1}\|_D^2 \leqslant \|u^n\|_D^2 + \tau\kappa\|\varphi^n\|_{D^{-1}}^2,$$

связывающее нормы функции u^n на соседних слоях. Таким образом,

$$\|u^{n+1}\|_D^2 \leqslant \|u^n\|_D^2 + \tau\kappa\|\varphi^n\|_{D^{-1}}^2. \tag{11}$$

Имеем последовательность неравенств

$$\|u^{n+1}\|_D^2 \leqslant \|u^n\|_D^2 + \kappa\tau\|\varphi^n\|_{D^{-1}}^2 \leqslant \|u^{n-1}\|_D^2 + \\ + \kappa\sum_{k=n-1}^{n}\tau\|\varphi^k\|_{D^{-1}}^2 \leqslant \cdots \leqslant \|u^0\|_D^2 + \kappa\sum_{k=0}^{n}\tau\|\varphi^k\|_{D^{-1}}^2.$$

Предполагая, что κ не зависит от шагов сетки h и τ, при $n\tau \leqslant T$ получаем окончательное соотношение

$$\|u^n\|_D^2 \leqslant \|u^0\|_D^2 + c\sum_{k=0}^{n-1}\tau\|\varphi^k\|_{D^{-1}}^2 \leqslant \|u^0\|_D^2 + c\sum_{k=0}^{N-1}\tau\|\varphi^k\|_{D^{-1}}^2 \leqslant \\ \leqslant \|u^0\|_D^2 + cT\max_{0\leqslant k\leqslant N-1}\|\varphi^k\|_{D^{-1}}^2, \tag{12}$$

которое означает устойчивость разностной схемы (5) по начальным данным и по правой части. Таким образом, условие $D = B - \tau A > 0$ является достаточным для устойчивости разностной схемы по правой части и начальным условиям.

Заметим, что выражение $\sum_{k=0}^{N-1}\tau\|\varphi^k\|_{D^{-1}}^2$, стоящее в правой части (12), является квадратурной формулой для интеграла $\int_0^T \|\varphi(t)\|_{D^{-1}}^2\,dt$, а выражение $\max_{0\leqslant k\leqslant N-1}\|\varphi^k\|_{D^{-1}}^2$ — сеточным аналогом нормы $\max_{0\leqslant t\leqslant T}\|\varphi(t)\|_{D^{-1}}^2$.

Найдем необходимое и достаточное условие того, чтобы собственные числа оператора перехода от слоя к слою не превосходили единицы. При выполнении этого условия схема устойчива по начальным данным и норма погрешности не возрастает при переходе от слоя к слою. Для этого положим $\varphi^k \equiv 0$.

Применим к обеим частям (5) оператор $B^{-1/2}$; получим

$$B^{1/2}u^{n+1} = B^{1/2}u^n - \tau B^{-1/2}Au^n.$$

Обозначим $B^{1/2}u^n = y^n$. Тогда $u^n = B^{-1/2}y^n$ и последнее равенство примет вид

$$y^{n+1} = y^n - \tau B^{-1/2}AB^{-1/2}y^n = (E - \tau B^{-1/2}AB^{-1/2})y^n.$$

Оператор $S = E - \tau B^{-1/2}AB^{-1/2}$ симметричен, поэтому собственные числа S лежат на отрезке $[\gamma_1, \gamma_2]$, где

$$\gamma_1 = \min_{v\in H}\frac{(Sv, v)}{(v, v)}, \quad \gamma_2 = \max_{v\in H}\frac{(Sv, v)}{(v, v)}.$$

Таким образом, собственные числа S не превосходят 1, если выполняется соотношение

$$-1 \leqslant \frac{(Sv, v)}{(v, v)} \leqslant 1 \quad \forall v \in H. \tag{13}$$

Преобразуя выражение $(Sv, v)/(v, v)$ и обозначая $v = B^{1/2}z$, получаем

$$\frac{(Sv, v)}{(v, v)} = \frac{(v, v) - \tau(B^{-1/2}AB^{-1/2}v, v)}{(v, v)} = \frac{(Bz, z) - \tau(Az, z)}{(Bz, z)}.$$

Поэтому (13) имеет вид

$$-1 \leqslant \frac{(Bz, z) - \tau(Az, z)}{(Bz, z)} \leqslant 1.$$

Правая часть неравенства выполнена всегда, поскольку $A, B > 0$; левая часть эквивалентна выполнению для любого $z \in H$, $z \neq 0$, неравенства

$$2(Bz, z) - \tau(Az, z) \geqslant 0.$$

Последнее означает, что

$$B \geqslant \frac{\tau}{2}A. \tag{14}$$

Как видим, условие (14), обеспечивающее устойчивость схемы (5), не совпадает с (8). Однако у схем, обладающих свойством (8), имеется существенное достоинство. В ряде практических задач интервал интегрирования T уравнения (1) достаточно велик или же требуется вести счет задачи до выхода на стационарный режим. В этих случаях целесообразно использовать разностные схемы, удовлетворяющие более сильной оценке устойчивости (15) при $q < 1$.

Обозначим $(1 + \tau/\kappa)^{-1} = q$ и $\tau\kappa = \gamma$. Тогда из (11) имеем

$$\begin{aligned}\|u^{n+1}\|_D^2 &\leqslant q(\|u^n\|_D^2 + \gamma\|\varphi^n\|_{D^{-1}}^2) \leqslant q^2\|u^{n-1}\|_D^2 + q\gamma\|\varphi^n\|_{D^{-1}}^2 + \\ &+ \gamma q^2\|\varphi^{n-1}\|_{D^{-1}}^2 \leqslant \cdots \leqslant q^{n+1}\|u^0\|_D^2 + \\ &+ q\gamma(\|\varphi^n\|_{D^{-1}}^2 + q\|\varphi^{n-1}\|_{D^{-1}}^2 + \cdots + q^n\|\varphi^0\|_{D^{-1}}^2).\end{aligned}$$

Пусть $\|\varphi\|_\infty = \max\limits_n \|\varphi^n\|_{D^{-1}} < \infty$. Тогда из последнего неравенства получаем

$$\|u^n\|_D^2 \leqslant \gamma\frac{q}{1-q}\|\varphi\|_\infty^2 + q^n\|u^0\|_D^2. \tag{15}$$

Отсюда следует, что при $\|\varphi\|_\infty < \infty$ решение задачи (5) будет ограничено на бесконечном промежутке времени при выполнении условия (8). Отметим, что из (14) ограниченность решения на бесконечном промежутке времени при наличии правой части, вообще говоря, не следует.

Разностную схему (5) можно рассматривать как итерационный процесс решения уравнения

$$Au = \varphi,$$

где $\varphi^n = \varphi$. В этом случае выполнение условия $B - \tau A > 0$ обеспечивает сходимость итерационного процесса. Действительно, записывая уравнение для погрешности $z^n = u^n - u$, имеем

$$Bz_t^n + Az^n = 0, \quad z^0 = u^0 - u.$$

Из условия $D = B - \tau A > 0$ следует, что D определяет в H норму, которую будем обозначать $\|\cdot\|_D$. Тогда из (9) получаем

$$\|u^{n+1}\|_D^2 + 2\tau\|u^{n+1}\|_A^2 \leqslant \|u^n\|_D^2.$$

Отсюда

$$\left(1 + \frac{\tau}{\kappa}\right)\|u^{n+1}\|_D^2 + \tau\|u^{n+1}\|_A^2 \leqslant \|u^n\|_D^2. \tag{16}$$

Введем норму $\|u^n\|_1^2 = \|u^n\|_D^2 + \tau\left(1 + \frac{\tau}{\kappa}\right)^{-1}\|u^n\|_A^2$; тогда из (16) следует окончательная оценка

$$\|u^{n+1}\|_1^2 \leqslant \left(1 + \frac{\tau}{\kappa}\right)^{-1}\|u^n\|_1^2 \leqslant \cdots \leqslant \left(1 + \frac{\tau}{\kappa}\right)^{-n-1}\|u^0\|_1^2.$$

Таким образом, итерационный метод (5) сходится со скоростью геометрической прогрессии. При этом скорость сходимости определяется величинами τ и κ и условием $D > 0$. Процесс сходится как в норме, определяемой оператором D, так и в норме, определяемой оператором A.

Выясним, при каких условиях будут устойчивы схемы (3), (4). Схема (3) уже имеет вид (5), при этом $B = E$, а $A = -\Delta^h$. Необходимо, таким образом, проверить выполнение условия $B \geqslant \frac{\tau}{2}A$. Имеем (см. § 6)

$$(Av, v) = -\sum_{i,j=1}^{M-1} h^2\Delta^h v_{ij}v_{ij} = \sum_{i,j=0}^{M-1} h^2\left[\left(\frac{\tilde{v}_{i+1,j} - \tilde{v}_{ij}}{h}\right)^2 + \right.$$

$$\left. + \left(\frac{\tilde{v}_{i,j+1} - \tilde{v}_{ij}}{h}\right)^2\right] \leqslant \frac{8}{h^2}\sum_{i,j=1}^{M-1} h^2 v_{ij}^2 = \frac{8}{h^2}(v, v).$$

Таким образом, чтобы (14) было справедливо, достаточно выполнения неравенства $2/\tau \geqslant 8/h^2$, т. е. явная схема (3) условно устойчива при $\tau \leqslant h^2/4$.

Представим теперь неявную схему (4) в виде (5). В этом случае

$$B = E - \tau\Delta^h, \quad A = -\Delta^h, \qquad \text{причем} \qquad A > 0.$$

Таким образом, условие (14), равно как и условие (8), выполнено при любых τ и h, т. е. неявная схема (4) безусловно устойчива.

Если для решения системы уравнений относительно значений решения на верхнем слое применяется так называемый *марш-алгоритм* в его устойчивой форме, то число арифметических операций при переходе от слоя к слою пропорционально числу неизвестных. Тогда неявная схема является экономичной.

Перейдем к изучению других экономичных разностных схем для уравнения (1). Будем рассматривать задачу (1) с однородными граничными условиями, т. е. при $\alpha \equiv 0$. Пусть Λ_1 и Λ_2 — операторы второй разделенной разности по направлениям x_1 и x_2 соответственно, т. е.

$$\Lambda_1 v_{ij} = \frac{\tilde{v}_{i+1,j} - 2\tilde{v}_{ij} + \tilde{v}_{i-1,j}}{h^2}, \quad \Lambda_2 v_{ij} = \frac{\tilde{v}_{i,j+1} - 2\tilde{v}_{ij} + \tilde{v}_{i,j-1}}{h^2}.$$

Здесь, как и ранее, $\tilde{v}$ — функция, совпадающая с v на Ω_h и равная нулю на $\partial\Omega_h$.

Задача 1. Проверить, что функция $\varphi_j \sin(\pi n i h)$, где φ_j — произвольная функция аргумента j, является собственной для оператора Λ_1, а любая функция $\psi_i \sin(\pi m j h)$, где ψ_i — произвольная функция аргумента i, — собственной для оператора Λ_2.

Задача 2. Проверить, что функции $\varphi_{mn} = \sin(\pi m i h)\sin(\pi n j h)$ образуют полную систему собственных функций операторов Λ_1 и Λ_2.

Положим

$$B = (E - \mu\Lambda_1)(E - \mu\Lambda_2), \quad A = -\Delta^h.$$

Оператор B является симметричным и положительно определенным как произведение симметричных положительно определенных и коммутирующих между собой операторов. Операторы такого вида называют *расщепляющимися*. Коммутируемость операторов Λ_1 и Λ_2 можно проверить непосредственно; кроме того, она следует из того факта, что эти операторы имеют общую полную систему собственных функций (см. задачу 2) и, следовательно, записываются в виде

$$(E - \mu\Lambda_1) = C^{-1} M_1 C, \quad (E - \mu\Lambda_2) = C^{-1} M_2 C,$$

где M_1, M_2 — диагональные матрицы, матрица C одна и та же.

Проверим, при каких μ выполняется условие (14). Имеем

$$B = E - \mu(\Lambda_1 + \Lambda_2) + \mu^2 \Lambda_1 \Lambda_2 = E - \mu\Delta^h + \mu^2 \Lambda_1 \Lambda_2,$$

поэтому условие (14) приобретает вид

$$E - \mu\Delta^h + \mu^2 \Lambda_1 \Lambda_2 \geqslant -\frac{\tau}{2}\Delta^h.$$

Так как оператор $E + \mu^2 \Lambda_1 \Lambda_2$ положительно определен, то условие $\mu \geqslant \tau/2$ обеспечивает выполнение (14), т. е. при $\mu \geqslant \tau/2$ разностная схема (5) безусловно устойчива по начальным данным.

Рассмотрим алгоритм реализации схемы (5) в данном случае. Обозначим $z = \dfrac{u^{n+1} - u^n}{\tau}$ и представим (5) в виде

$$(E - \mu\Lambda_1)(E - \mu\Lambda_2) z = \Delta^h u^n + \varphi^n.$$

Последнее уравнение разобьем на два:

$$(E-\mu\Lambda_1)y=\Delta^h u^n+\varphi^n,\quad (E-\mu\Lambda_2)z=y. \tag{17}$$

Функция $g=\Delta^h u^n+\varphi^n$ может быть вычислена во всех точках Ω_h, т. е. можно считать ее известной. Первое уравнение (17) запишем в виде

$$y_{ij}-\mu\frac{y_{i+1,j}-2y_{ij}+y_{i-1,j}}{h^2}=g_{ij},\ i=1,\ldots,M-1;\ j=1,\ldots,M-1. \tag{18}$$

При фиксированном j система (18) относительно неизвестных $y_{1,j}$, $y_{2,j},\ldots,y_{M-1,j}$ представляет собой систему уравнений с трехдиагональной матрицей, которая может быть решена, например, методом прогонки за $O(M)$ арифметических операций. Решая (18) при каждом $j=1,\ldots,M-1$, найдем функцию y во всех узлах Ω_h. Группы неизвестных, связанные уравнениями (18), объединены на рис. 10.7.1 символом $\leftrightarrow$.

Примечание. Если решается неоднородная краевая задача, т. е. $u|_\Gamma\neq 0$, то, вообще говоря, $y=(E-\mu\Lambda_2)\dfrac{u^{n+1}-u^n}{\tau}$ не удовлетворяют граничному условию $y|_\Gamma=0$ и значения $y|_\Gamma$ требуется определять специальным образом.

Аналогично второе уравнение (17) расписывается в виде

$$z_{ij}-\mu\frac{z_{i,j+1}-2z_{ij}+z_{i,j-1}}{h^2}=y_{ij},\ j=1,\ldots,M-1;\ i=1,\ldots,M-1. \tag{19}$$

При фиксированном i (19) является системой уравнений с трехдиагональной матрицей относительно неизвестных $(z_{i,1},\ z_{i,2},\ldots,\ z_{i,M-1})$.

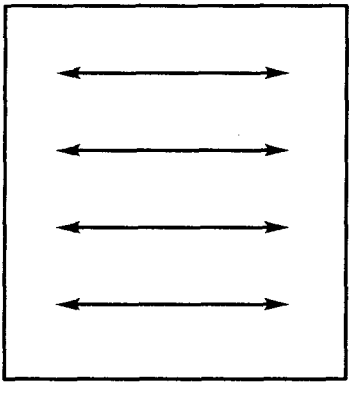


Рис. 10.7.1

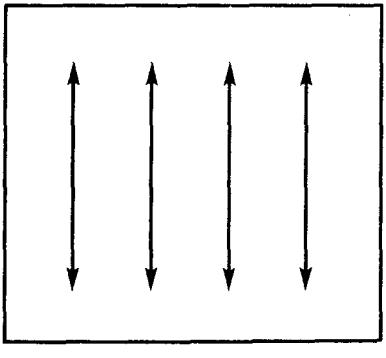


Рис. 10.7.2

Таким образом, функция z может быть найдена из (19) за $O(M^2)$ арифметических операций. Группы неизвестных, связанные уравнениями (19), объединены на рис. 10.7.2 символом $\updownarrow$. Значения u^{n+1} находятся по явным формулам:

$$u^{n+1}=u^n+\tau z. \tag{20}$$

Таким образом, суть представленного алгоритма заключается в следующем: при каждом j решаем систему уравнений (18) с трехдиагональной матрицей. При этом изменение i соответствует изменению абсциссы; поэтому i иногда называют «горизонтальной» переменной функции y. Далее, используя найденное значение y, при каждом i решаем систему уравнений (19). Аргумент j в этом случае называют, соответственно, «вертикальной» переменной. После этого значение u^{n+1} находится по явной формуле (20).

Описанный алгоритм часто называют *методом расщепления.*

Как мы видели, основная его идея заключается в построении оператора при разностной производной по времени таким образом, чтобы этот оператор являлся произведением операторов $B = B_1 B_2$, каждый из которых действует только в одном направлении, и полученная схема аппроксимировала исходную задачу. Так, в данном случае $B_i = E - \mu\Lambda_i$.

Задача 3. Доказать, что при достаточно гладком решении для рассматриваемой схемы имеет место аппроксимация порядка $O(\mu + \tau + h^2)$. Таким образом, при $\mu = \tau/2$ погрешность аппроксимации имеет порядок $O(\tau + h^2)$ и схема абсолютно устойчива.

Близким по своей структуре является метод переменных направлений. Суть его заключается в переходе от u^n к u^{n+1} по формулам

$$\begin{aligned} \frac{u^{n+1/2} - u^n}{0{,}5\tau} &= \Lambda_1 u^{n+1/2} + \Lambda_2 u^n + \varphi^n, \\ \frac{u^{n+1} - u^{n+1/2}}{0{,}5\tau} &= \Lambda_1 u^{n+1/2} + \Lambda_2 u^{n+1} + \varphi^n. \end{aligned} \tag{21}$$

Здесь введен промежуточный вектор неизвестных $u^{n+1/2}$; первое уравнение (21) решается применением прогонки по оси x_1, а второе — применением прогонки по оси x_2.

Построение методов с расщепляющимся оператором на верхнем слое в k-мерном случае можно провести по точно такой же схеме. Пусть $A = A_1 + A_2 + \cdots + A_k$, где A_i — одномерный оператор в i-м направлении. Положим $B = (E + \mu A_1) \ldots (E + \mu A_k)$. Нужная разностная схема будет иметь вид

$$(E + \mu A_1) \ldots (E + \mu A_k) \frac{u^{n+1} - u^n}{\tau} + Au^n = \varphi^n. \tag{22}$$

Реализация этого метода проводится по такому же алгоритму, что и выше. Параметры μ и τ выбираются из условий устойчивости и аппроксимации разностной схемы.

Перейдем к рассмотрению методов решения параболических уравнений в случае, когда область Ω имеет достаточно произвольную форму. В этом случае реализация описанной выше схемы, основанной на представлении оператора B на верхнем слое в виде произведения одномерных операторов, встречает существенные затруднения. Однако применима схема (21).

Другая разностная схема может быть получена из следующих соображений. Представим A в виде $A = R_1 + R_2$, где $R_1^T = R_2$, R_1 и R_2 — правая и левая треугольные матрицы. В частности, если $A = -\Delta^h$, то R_1 и R_2 можно записать следующим образом:

$$R_1 v|_{(i,j)} = \frac{2\tilde{v}_{ij} - \tilde{v}_{i+1,j} - \tilde{v}_{i,j+1}}{h^2}, \quad R_2 v|_{(i,j)} = \frac{2\tilde{v}_{ij} - \tilde{v}_{i-1,j} - \tilde{v}_{i,j-1}}{h^2}. \tag{23}$$

Рассмотрим разностную схему

$$(E + \sigma\tau R_1)(E + \sigma\tau R_2)\frac{u^{n+1} - u^n}{\tau} + Au^n = \varphi^n. \tag{24}$$

Оператор $B = (E+\sigma\tau R_1)(E+\sigma\tau R_2)$ будет симметричным и положительно определенным. Поэтому для устойчивости разностной схемы по начальным данным достаточно проверить выполнение условия (14). Параметр σ (весовой множитель) выбирается таким, чтобы схема была устойчива и аппроксимировала исходное уравнение.

Пусть в качестве примера рассматривается уравнение (1) в прямоугольнике с нулевыми граничными условиями. В этом случае $A = -\Delta^h$, а R_1 и R_2 определяются по формулам (23). Тогда $B = (E + \sigma\tau R_1)(E + \sigma\tau R_2) = E - \sigma\tau\Delta^h + \sigma^2\tau^2 R_1 R_2$. Оператор $R_1 R_2$ является положительно определенным, поэтому $B = B^* > 0$. Условие $B \geqslant \dfrac{\tau}{2}A$ в данном случае заведомо будет выполнено при $\sigma \geqslant 0{,}5$. Таким образом, условие $\sigma \geqslant 0{,}5$ обеспечивает безусловную устойчивость схемы (24).

Найдем порядок аппроксимации разностной схемы (24). Так как

$$R_1 v = \frac{2v_{ij} - v_{i+1,j} - v_{i,j+1}}{h^2} = \frac{1}{h}\left(\frac{v_{ij} - v_{i+1,j}}{h}\right) + \frac{1}{h}\left(\frac{v_{ij} - v_{i,j+1}}{h}\right),$$

то, используя формулу Тейлора, получим

$$R_1[u] = -\frac{1}{h}\left(\frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} + \frac{h}{2}\frac{\partial^2 u}{\partial x_1^2} + \frac{h}{2}\frac{\partial^2 u}{\partial x_2^2}\right) + O(h).$$

Используя аналогичную оценку погрешности аппроксимации оператора R_2, приходим к заключению, что выражение $B = E - \sigma\tau\Delta^h + \sigma^2\tau^2 R_1 R_2$ аппроксимирует единичный оператор E с порядком $O(\sigma\tau + \sigma\tau^2/h^2)$. Если считать, что $\sigma \geqslant 0{,}5$ порядка 1, то схема (24) аппроксимирует исходное уравнение (1) с порядком $O(h^2 + \tau + \tau^2/h^2)$. Таким образом, величина τ/h должна быть достаточно мала. Тем не менее схема (24) все же лучше, чем явная. Для устойчивости явной схемы требуется выполнение условия $\tau \leqslant h^2/4$, в то время как при $\sigma \geqslant 0{,}5$ схема (24) будет безусловно устойчивой и шаг по времени τ может быть выбран существенно большим. Например, можно взять $\tau = ah^2$ или же $\tau = ah^{3/2}$. Погрешность аппроксимации по времени в этих случаях будет иметь порядок $O(h^2)$ и $O(h)$ соответственно. Следует, однако, отметить, что часто решение параболического уравнения по t обладает достаточным количеством производных

и эти производные стремятся к нулю при $t \to \infty$. Это оправдывает применение схемы (24) при расчете нестационарных задач, поскольку шаг по времени τ можно брать достаточно большим.

Выпишем алгоритм, соответствующий разностной схеме (24). Обозначим $(E + \sigma\tau R_2)z$ через y. Тогда, решая уравнение

$$(E + \sigma\tau R_1)y = -Au^n + \varphi^n, \tag{25}$$

можно найти значение y по явным формулам. Действительно, в случае, когда Ω — квадрат (что несущественно), уравнение (25) в узле (i, j) имеет вид

$$y_{ij} + \frac{\sigma\tau}{h^2}(2y_{ij} - y_{i+1,j} - y_{i,j+1}) = (-Au^n + \varphi^n)_{i,j}. \tag{26}$$

По известным значениям y_{iM}, y_{Mj}, $i, j = 1, \ldots, M$, из (26) по явным формулам можно найти $y_{i,M-1}$ и $y_{M-1,j}$, $i, j = 1, \ldots, M-1$. После этого по тем же формулам находим значения $y_{i,M-2}$, $y_{M-2,j}$ и т. д.

Для вычисления значения y_{ij} требуется число арифметических операций, не зависящее от шагов сетки. Поэтому вычисление значений функции y во всех узлах потребует $O(M^2)$ арифметических операций, что по порядку совпадает с количеством узлов на слое. Аналогичным образом решая уравнение

$$(E + \sigma\tau R_2)z = y,$$

за $O(M^2)$ арифметических операций найдем значение z. Решение u^{n+1} находится после этого простым пересчетом по формуле

$$u^{n+1} = u^n + \tau z.$$

Таким образом, переход от u^n к u^{n+1} в схеме (24) требует числа арифметических операций, пропорционального количеству узлов сетки на слое, т. е. схема (24) экономична.

Наиболее эффективна схема (24), если ее рассматривать как итерационный метод для решения стационарной задачи $Au = \varphi$. В этом случае требование аппроксимации по t не играет никакой роли и параметр τ можно выбирать только из соображений наиболее быстрой сходимости итерационного метода. Обычно выбирают $\tau = O(h)$ и $\sigma = 1$, чтобы выполнялось (8). Тогда операторы B и A связаны соотношением

$$\gamma_1 hA \leqslant B \leqslant \gamma_2 A, \tag{27}$$

где γ_1, $\gamma_2 > 0$ не зависят от h, и, решая стационарную задачу с помощью итерационного процесса (24), получим решение с точностью ε за $O\left(h^{-3}\ln(\varepsilon^{-1})\right)$ арифметических операций.

Задача 4. Доказать оценку (27) при $\sigma = 1$, $\tau = O(h)$.

Задача 5. Показать, что при любых σ, $\tau > 0$ отношение γ_2/γ_1 ограничено снизу постоянной, отличной от нуля.

Итерационный процесс можно ускорить, если зафиксировать B и после этого выбирать параметр τ переменным. В частности, если выбирать τ_k как указывалось в § 6.6 (чебышевское ускорение), то решение стационарной задачи с точностью ε может быть получено за $O(h^{-5/2}\ln(\varepsilon^{-1}))$ арифметических операций.

Для решения параболического уравнения в области достаточно произвольной формы существуют также и другие методы. Рассмотрим метод, который сводит исходную задачу к решению последовательности одномерных задач. Изложение метода проведем на примере уравнения (1).

Представим оператор Δ в виде суммы одномерных операторов

$$\Delta = L_1 + L_2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2},$$

а правую часть f — в виде суммы правых частей: $f = f_1 + f_2$. Левая и правая части уравнения (1) равны сумме левых и правых частей уравнения

$$\frac{1}{2}\frac{\partial u}{\partial t} - L_1 u = f_1, \quad \frac{1}{2}\frac{\partial u}{\partial t} - L_2 u = f_2. \tag{28}$$

Опишем переход от n-го к $(n+1)$-му слою. Аппроксимируем первое из уравнений (28) следующим образом:

$$P_h^1 u_{ij}^n \equiv \frac{u_{ij}^{n+1/2} - u_{ij}^n}{\tau} - \frac{u_{i+1,j}^{n+1/2} - 2u_{ij}^{n+1/2} + u_{i-1,j}^{n+1/2}}{h^2} - f_{1,ij}^{n+1/2} = 0. \tag{29}$$

Второе уравнение (28) заменим соотношением

$$P_h^2 u_{ij}^{n+1/2} \equiv \frac{u_{ij}^{n+1} - u_{ij}^{n+1/2}}{\tau} - \frac{u_{i,j+1}^{n+1} - 2u_{i,j}^{n+1} + u_{i,j-1}^{n+1}}{h^2} - f_{2,ij}^{n+1} = 0. \tag{30}$$

Таким образом, алгоритм заключается в последовательном решении уравнений (29), (30). При этом вычисленное значение функции является начальным условием для следующего уравнения.

Ясно, что каждое из уравнений (29), (30) не аппроксимирует исходную задачу. Найдем погрешность аппроксимации. Имеем

$$P_h^1[u] = \frac{1}{2}\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - f_1 + O(h^2 + \tau),$$

$$P_h^1[u] - \left[\frac{\partial u}{\partial t} - \Delta u - f\right] = -\frac{1}{2}\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial y^2} + f_2 + O(h^2 + \tau) \equiv \psi_1.$$

Аналогично

$$P_h^2[u] = \frac{1}{2}\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial y^2} - f_2 + O(h^2 + \tau),$$

$$P_h^2[u] - \left[\frac{\partial u}{\partial t} - \Delta u - f\right] = -\frac{1}{2}\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + f_1 + O(h^2 + \tau) \equiv \psi_2.$$

В общем случае $\psi_i = O(1)$, поэтому и уравнения (29), (30) аппроксимируют (1) с порядком $O(1)$. Однако

$$\psi_1 + \psi_2 = -\frac{\partial u}{\partial t} + \Delta u + f + O(h^2 + \tau) = O(h^2 + \tau).$$

В этом случае говорят, что *схема* (29), (30) *аппроксимирует задачу* (1) *в суммарном (или слабом) смысле*, т. е. хотя каждое из уравнений (29), (30) не аппроксимирует исходную задачу, сумма погрешностей аппроксимаций этих уравнений равна $O(h^2 + \tau)$.

Реализация (29), (30) требует на каждом шаге решения уравнений с трехдиагональной матрицей. Таким образом, этот метод применим для решения уравнения (1), когда область Ω имеет достаточно произвольную форму. Остается лишь выяснить ее устойчивость и сходимость. Оказывается, имеет место

Теорема (без доказательства). *Схема* (29), (30) *устойчива в сеточной норме пространства* C *и при достаточно гладком решении*

$$\|u^n - u(n\tau)\|_{C(\Omega_h)} \leqslant C_1(h^2 + \tau),$$

где C_1 *не зависит от* h *и* τ, *а* $u(n\tau)$ *— значение решения* $u(x, \tau)$ *на* n*-м слое.*

Рассмотренный метод получения разностных схем носит название *метода дробных шагов* или же *метода суммарной аппроксимации*. Его можно применять не только в линейных задачах, но и в нелинейных.

В общем случае для уравнения

$$\frac{\partial u}{\partial t} = P^1(u) + \cdots + P^k(u), \tag{31}$$

где операторы $P^i(u)$, вообще говоря, нелинейные и не обязательно одномерные, схема метода дробных шагов заключается в следующем. Решение на шаге уравнения (31) заменяется последовательным решением на шаге уравнений

$$\frac{1}{k}\frac{\partial u_i}{\partial t} = P^i(u_i), \quad i = 1, \ldots, k.$$

При этом в качестве начального условия на шаге для каждого из уравнений берется значение, вычисленное из предыдущего уравнения.

В создание описанных выше экономичных методов решений многомерных нестационарных задач внесли большой вклад Е. Г. Дьяконов, Г. И. Марчук, А. А. Самарский и Н. Н. Яненко.

§ 8. Методы решения сеточных эллиптических уравнений

В этом параграфе будут рассмотрены методы решения систем сеточных эллиптических уравнений. Рассмотрим простейший пример (см. § 6). Пусть

$$-\frac{u_{m+1,n}-2u_{mn}+u_{m-1,n}}{h_1^2}-\frac{u_{m,n+1}-2u_{mn}+u_{m,n-1}}{h_2^2}=\varphi_{mn}, \tag{1}$$

$$u_{m0}=u_{mN}=0,\ u_{0,n}=u_{M,n}=0;\ m=1,\ldots,M-1;\ n=1,\ldots,N-1,$$

система сеточных уравнений относительно неизвестных u_{mn}, $0\leqslant m\leqslant M$, $0\leqslant n\leqslant N$, получающаяся в результате аппроксимации краевой задачи Дирихле для уравнения Пуассона в квадрате $\bar{\Omega}=\{x=(x_1,x_2),\ 0\leqslant x_i\leqslant 1\}$. Будем для простоты считать, что $h_1=h_2=h=1/N$. Для решения системы уравнений (1) можно предложить большое число методов. Поэтому для того, чтобы их сравнивать, необходимо выбрать один или несколько критериев, по которым будет проводиться сравнение. Условимся критерием качества метода считать количество арифметических операций, которые необходимы либо для получения точного решения, либо для получения решения с некоторой заданной точностью. Так как часто не удается вычислить точное число арифметических операций либо этот подсчет достаточно громоздок, оценивают лишь порядок числа арифметических операций по отношению к числу узлов сетки.

Прежде всего отметим особенности рассматриваемых систем уравнений. Во-первых, это большая размерность (большое число неизвестных в системе). Это связано с желанием получить решение исходной дифференциальной задачи с нужной точностью, что требует достаточно малого шага сетки. Во-вторых, в каждой строке матрицы лишь конечное число элементов отлично от нуля. В частности, в (1) количество ненулевых элементов в строке не превосходит пяти. Все это заставляет разрабатывать специальные эффективные методы, учитывающие особенности систем уравнений такого типа.

Рассмотрим прежде всего, что же дает применение классического метода Гаусса к решению системы (1). Предположим, что граничные узлы исключены из системы уравнений. Тогда вектор неизвестных, которыми являются значения функции во внутренних узлах, имеет размерность $(N-1)^2$. Поэтому прямое применение метода Гаусса к решению системы (1) требует $2/3(N-1)^6+O(N^4)$ арифметических операций. Кроме этого, потребуется хранить в памяти матрицу системы, т. е. потребуется $O(N^4)$ слов ЭВМ для хранения элементов матрицы.

Заметим, однако, что бóльшая часть арифметических операций является несодержательной — это операции над нулевыми элементами ма-

трицы. Выясним, какое потребуется количество арифметических операций, если вычисления проводить только над ненулевыми элементами. Напомним, что матрица A системы уравнений (1) при «естественной» нумерации компонент вектора неизвестных $(u_{11}, u_{12}, \ldots, u_{1,N-1}, u_{21}, \ldots, u_{N-1,N-1})$ имеет вид (с точностью до множителя h^{-2})

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdot & \cdots & \cdot & 0 \\ A_{21} & A_{22} & A_{23} & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & A_{N-2,N-1} \\ 0 & \cdot & \cdot & \cdots & A_{N-1,N-2} & A_{N-1,N-1} \end{pmatrix},$$

где

$$A_{ii} = \begin{pmatrix} 4 & -1 & 0 & \cdot & \cdots & \cdot & 0 \\ -1 & 4 & -1 & 0 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdots & 4 & -1 \\ 0 & \cdot & \cdot & \cdot & \cdots & -1 & 4 \end{pmatrix},$$

а $A_{i,i\pm1}$ — диагональные матрицы с элементами на диагонали, равными -1. Матрицы A_{ii} имеют размерность $(N-1)\times(N-1)$.

Таким образом, матрица A оказывается блочно-трехдиагональной и в каждой строке матрицы не более пяти элементов отличны от нуля; кроме этого, матрица A является ленточной с шириной ленты равной $(2N-1)$: все элементы $a_{ij}=0$ при $|i-j|\geqslant N$. Задача решения системы (1) является частным случаем решения системы m уравнений с m неизвестными

$$A\mathbf{x} = \mathbf{b} \tag{2}$$

с $(2s+1)$-диагональной матрицей.

Для решения такой системы могут быть применены, например, методы Гаусса, квадратного корня, отражений и вращений с исключением несодержательных операций во всех случаях. Под исключением несодержательных операций имеется в виду следующее. Поскольку $a_{ij}=0$ при $|i-j|>s$, то не надо проводить операций по обнулению этих элементов. Поэтому производятся операции лишь по обнулению $O(ms)$ элементов. Кроме того, при реализации каждого шага обнуления все элементы $a_{ij}=0$ при $|i-j|>s$ остаются равными нулю и поэтому каждый шаг требует $O(s)$ арифметических операций. Таким образом, общее число операций при решении системы уравнений (2) оказывается величиной порядка $O(ms^2)$. В данном случае $m=(N-1)^2$, $s=N-1$ и поэтому общее число арифметических операций $O(N^4)$. Матрица A симметричная и положительно определенная. Как отмечалось ранее, в случае $A>0$ при реализации этих методов не возникает операции деления на нуль.

Еще один возможный метод решения этой системы — это метод прогонки в блочной форме. Исходную систему уравнений записываем в виде

$$A_{11}\mathbf{u}_1 + A_{12}\mathbf{u}_2 = \mathbf{f}_1,$$
$$A_{21}\mathbf{u}_1 + A_{22}\mathbf{u}_2 + A_{23}\mathbf{u}_3 = \mathbf{f}_2,$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$A_{N-2,N-3}\mathbf{u}_{N-3} + A_{N-2,N-2}\mathbf{u}_{N-2} + A_{N-2,N-1}\mathbf{u}_{N-1} = \mathbf{f}_{N-2},$$
$$A_{N-1,N-2}\mathbf{u}_{N-2} + A_{N-1,N-1}\mathbf{u}_{N-1} = \mathbf{f}_{N-1},$$

где $\mathbf{u}_k$ — вектор с компонентами $u_{k,1}, u_{k,2}, \dots, u_{k,N-1}$.

Далее последовательно исключаем векторы неизвестных $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{N-2}$ и получаем рекуррентные формулы, аналогичные формулам метода прогонки. При реализации этих формул придется произвести порядка cN операций умножения и обращения матриц размерности $N-1$, что потребует всего порядка $2cN^4$ арифметических операций. Дополнительно требуется произвести порядка $O(N)$ операций умножения матрицы на вектор и сложения векторов, для этого потребуется $O(N^3)$ арифметических операций. Общие вычислительные затраты, таким образом, составят порядка $2cN^4$ арифметических операций.

Описанные выше методы непосредственно применимы в случае эллиптического уравнения или системы эллиптических уравнений с произвольными, в частности, переменными коэффициентами и во всех этих случаях требуют $O(N^4)$ арифметических операций.

В конкретном случае системы (2) при $N = 2^l$ можно предложить метод, требующий $O(N^3 \ln N)$ арифметических операций. Для этого следует применять метод Гаусса в векторной форме при следующем порядке исключения неизвестных: сначала исключаются векторы $\mathbf{u}_1$, $\mathbf{u}_3$, $\mathbf{u}_5, \dots$, т. е. все векторы $\mathbf{u}_{2k+1}$, затем векторы $\mathbf{u}_2$, $\mathbf{u}_6$, $\mathbf{u}_{10}, \dots$, т. е. все векторы $\mathbf{u}_{2(2k+1)}$, затем векторы $\mathbf{u}_4$, $\mathbf{u}_{12}$, $\mathbf{u}_{20}, \dots$, т. е. все векторы $\mathbf{u}_{4(2k+1)}$ и т. д. Оказывается, что при таком способе исключения многократно умножаются и обращаются одни и те же матрицы. Если вместо этого запомнить произведения соответствующих матриц, то общее число операций окажется $O(N^3 \log N)$.

Описанные методы требуют в общем случае для запоминания ленты матрицы $O(N^3)$ слов памяти ЭВМ.

Задача 1. Показать, что при нумерации неизвестных

$$u_{11}, u_{12}, u_{21}, \dots, u_{1,N-1}, u_{2,N-2}, \dots, u_{N-1,1}, \dots, u_{N-1,N-1}$$

количество арифметических операций в методе Гаусса по порядку также равно $O(N^4)$.

Задача 2. Для случая разностной аппроксимации уравнения Лапласа в произвольной двумерной области указать порядок исключения неизвестных, при котором решение системы получается за $O(M^{3/2})$ арифметических операций; здесь M — общее число узлов.

Рассмотрим прямой метод решения системы уравнений (1), основанный на использовании дискретного преобразования Фурье в случае прямоугольника. Здесь мы не будем предполагать, что $M = N$. Преобразуем вначале систему уравнений (1) к операторной форме. Пусть $u(mh_1, nh_2)$ — сеточная функция, определенная при $0 \leqslant m \leqslant M$, $0 \leqslant n \leqslant N$, которая соответствует решению (1), т. е. $u(mh_1, nh_2) = u_{mn}$. Аналогично, $\varphi(mh_1, nh_2) = \varphi_{mn}$; $\varphi_{0n} = \varphi_{Mn} = \varphi_{m0} = \varphi_{mN} = 0$ — сеточная функция, соответствующая правой части (1). Через Λ_k обозначим операторы

$$(\Lambda_1 v)_{mn} = \begin{cases} \dfrac{v_{m+1,n} - 2v_{mn} + v_{m-1,n}}{h_1^2}, & 1 \leqslant m \leqslant M-1, \quad 1 \leqslant n \leqslant N-1, \\ 0 & \text{в остальных случаях;} \end{cases}$$

$$(\Lambda_2 v)_{mn} = \begin{cases} \dfrac{v_{m,n+1} - 2v_{mn} + v_{m,n-1}}{h_2^2}, & 1 \leqslant m \leqslant M-1, \quad 1 \leqslant n \leqslant N-1, \\ 0 & \text{в остальных случаях.} \end{cases}$$

Тогда система уравнений (1) может быть представлена в операторной форме

$$Au \equiv -(\Lambda_1 + \Lambda_2)u = \varphi. \tag{3}$$

Пусть для определенности $Mh_1 = 1$. Заметим, что этого всегда можно достичь, умножая обе части (1) на соответствующий коэффициент. Функция u как функция переменного m представима в виде дискретной суммы Фурье по переменной m (см. § 4.3):

$$\begin{aligned} u(mh_1, nh_2) &= \sum_{j=1}^{M-1} u_j(nh_2) \sin(\pi j m h_1), \\ u_j(nh_2) &= 2 \sum_{k=1}^{M-1} h_1 u(kh_1, nh_2) \sin(\pi j k h_1). \end{aligned} \tag{4'}$$

Аналогично представим φ в виде

$$\begin{aligned} \varphi(mh_1, nh_2) &= \sum_{j=1}^{M-1} \varphi_j(nh_2) \sin(\pi j m h_1), \\ \varphi_j(nh_2) &= 2 \sum_{k=1}^{M-1} h_1 \varphi(kh_1, nh_2) \sin(\pi j k h_1). \end{aligned} \tag{4''}$$

Подставляя полученные разложения в (3), получим

$$-(\Lambda_1+\Lambda_2)u_{mn}=-\sum_{j=1}^{M-1}(u_j(nh_2)\Lambda_1\sin(\pi jmh_1)+\sin(\pi jmh_1)\Lambda_2u_j(nh_2))= \\ =\sum_{j=1}^{M-1}\varphi_j(nh_2)\sin(\pi jmh_1). \tag{5}$$

Напомним, что функции $\sin(\pi jx)$ являются собственными для оператора Λ_1:

$$\Lambda_1\sin(\pi jmh_1)=\frac{\sin(\pi j(m+1)h_1)-2\sin(\pi jmh_1)+\sin(\pi j(m-1)h_1)}{h_1^2}= \\ =-\frac{4\sin^2\dfrac{\pi jh_1}{2}}{h_1^2}\sin(\pi jmh_1)\equiv-\lambda_j\sin(\pi jmh_1).$$

Поэтому выражение (5) может быть заменено эквивалентным

$$\sum_{j=1}^{M-1}\left(\lambda_ju_j(nh_2)+\frac{2u_j(nh_2)-u_j((n-1)h_2)-u_j((n+1)h_2)}{h_2^2}\right)\sin(\pi jmh_1)= \\ =\sum_{j=1}^{M-1}\varphi_j(nh_2)\sin(\pi jmh_1). \tag{6}$$

Функции $\sin(\pi jmh_1)$, $j=1,\dots,M-1$, являются ортогональными в скалярном произведении

$$(v,\,w)=\sum_{i=1}^{M-1}h_1v_iw_i.$$

Поэтому (6) представляет собой систему независимых уравнений

$$\lambda_ju_j(nh_2)+\frac{2u_j(nh_2)-u_j((n-1)h_2)-u_j((n+1)h_2)}{h_2^2}=\varphi_j(nh_2), \\ 1\leqslant n\leqslant N-1;\quad u_j(0)-u_j(Nh_2)=0,\quad j=1,\dots,M-1. \tag{7}$$

Система уравнений (7) может быть получена из (6) также умножением обеих частей (6) скалярно на $\sin(\pi jmh_1)$, $j=1,\dots,M-1$. Выражение (7) при фиксированном j представляет собой систему линейных алгебраических уравнений с трехдиагональной матрицей относительно неизвестных $u_j(h_2)$, $u_j(2h_2)$, $\dots$, $u_j((N-1)h_2)$, которая может быть решена, например, методом прогонки.

Таким образом, алгоритм решения задачи (1) заключается в следующем:

а) находим из (4″) при каждом n, $0<n<N$, коэффициенты $\varphi_j(nh_2)$, $j=1,\dots,M-1$;

б) при $j=1,\dots,M-1$ решаем методом прогонки систему уравнений (7). В результате получаем функции $u_j(nh_2)$, $j=1,\dots,M-1$;

в) из формулы (4′) при $m = 1, \ldots, M-1$, $n = 1, \ldots, N-1$ вычисляем значения функции $u(mh_1, nh_2)$.

Оценим затраченное количество арифметических операций. Пусть $M = 2^q$. В этом случае, используя алгоритм быстрого дискретного преобразования Фурье, найдем все φ_j за $O(MN \log_2 M)$ арифметических операций. При нахождении коэффициентов u_j потребуется $O(MN)$ операций. Наконец, при вычислении u из (4′) с использованием быстрого дискретного преобразования Фурье понадобится $O(MN \log_2 M)$ операций. Поэтому суммарное количество арифметических операций, необходимых для нахождения решения, в данном случае по порядку равно $O(MN \log_2 M)$. В частности, при $M = N$ получаем $O(N^2 \log_2 N)$.

Рассмотренный метод даст решение намного быстрее, чем метод Гаусса. Однако этот метод применим лишь в случае, когда исходная область является прямоугольником, в то время как метод Гаусса применим и в случае областей общей формы.

В случае прямоугольника существует ряд других методов решения с такой же асимптотикой числа действий. Как уже отмечалось, один из вариантов марш-алгоритма с приемлемой величиной вычислительной погрешности решает задачу за $O(N^2)$ арифметических операций при $N = M = 2^k$.

Рассмотрим другие приближенные методы решения системы уравнений (1), допускающие обобщение на случай более общих, чем прямоугольник, областей. В основу этих методов положено то свойство системы уравнений (1), что результат применения матрицы системы к вектору вычисляется по простым формулам и не требует запоминания матрицы. Количество арифметических операций, затрачиваемое на вычисление результата применения матрицы к вектору, по порядку равно $O(MN)$, т. е. пропорционально длине вектора. Ранее мы упоминали, что матрица A системы уравнений (1) симметрична и положительно определена и в рассматриваемом случае ее собственные значения λ_{mn}, $0 < m < M$, $0 < n < N$, лежат в промежутке

$$\lambda_{min} = \frac{4\sin^2\dfrac{\pi h_1}{2}}{h_1^2} + \frac{4\sin^2\dfrac{\pi h_2}{2}}{h_2^2} \leqslant \lambda_{mn} \leqslant \frac{4\cos^2\dfrac{\pi h_1}{2}}{h_1^2} + \frac{4\cos^2\dfrac{\pi h_2}{2}}{h_2^2} = \lambda_{max}.$$

Поэтому для решения системы уравнений

$$Au = \varphi$$

можно применить, например, метод простой итерации

$$\frac{u^{l+1} - u^l}{\tau} + Au^l = \varphi, \quad u^0 = v, \tag{8}$$

где v — вектор начального приближения. Согласно § 6.6 параметр τ целесообразно выбрать из соотношения

$$\tau = \tau_{opt} = \frac{2}{\lambda_{max} + \lambda_{min}}.$$

При этом метод (8) сходится со скоростью геометрической прогрессии и показатель скорости сходимости метода равен

$$q = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = 1 - \frac{2\lambda_{min}}{\lambda_{max} + \lambda_{min}}.$$

При $h = h_1 = h_2$ имеем $\lambda_{min}/\lambda_{max} = \pi^2 h^2/4 + O(h^4)$, поэтому

$$q = 1 - \frac{\pi^2 h^2}{2} + O(h^4).$$

Пусть ε — точность, с которой необходимо найти решение системы уравнений (1). Нормы погрешностей $z^n = u - u^n$ и $z^{n-1} = u - u^{n-1}$ на соседних слоях связаны соотношением

$$\|z^n\| \leqslant q\|z^{n-1}\| \leqslant \cdots \leqslant q^n\|z^0\|.$$

Задача 3. Показать, что при $\tau = \tau_{opt}$ расчетные формулы (8) приобретают вид:

$$u_{mn}^{l+1} = \frac{1}{4}\left(u_{m+1,\,n}^{l} + u_{m-1,\,n}^{l} + u_{m,\,n+1}^{l} + u_{m,\,n-1}^{l}\right) + \frac{h^2}{4}\varphi_{mn}$$

как в случае прямоугольника, так и в случае произвольной области.

Поэтому для выполнения неравенства $\|z^n\| \leqslant \varepsilon\|z^0\|$ достаточно выбрать n так, чтобы выполнялось неравенство $q^n \leqslant \varepsilon$. Отсюда $n|\ln q| \geqslant \ln(\varepsilon^{-1})$. Так как $q = 1 - \pi^2 h^2/2 + O(h^4)$, то $|\ln q| = \left|\ln(1 - \pi^2 h^2/2 + O(h^4))\right| = \pi^2 h^2/2 + O(h^4)$. При малых h имеем требование

$$n \geqslant \frac{2}{\pi^2 h^2}\ln(\varepsilon^{-1}). \tag{9}$$

Так как нас интересует случай малых ε, т. е. когда точность решения системы достаточно высока, то второй член в правой части (9) может быть отброшен. Поэтому в данном случае количество итераций n по порядку должно быть равно $O(h^{-2}\ln(\varepsilon^{-1}))$.

На каждом шаге итерации матрица умножается на вектор. Отсюда может возникнуть впечатление о необходимости запоминания матрицы A. На самом деле это не так. Нам необходим лишь результат применения матрицы A к вектору v, который вычисляется по формулам

$$(Av)_{mn} = \frac{4v_{mn} - v_{m+1,\,n} - v_{m-1,\,n} - v_{m,\,n+1} - v_{m,\,n-1}}{h^2}.$$

Поэтому *матрицу A запоминать не надо.*

Для вычисления значения функции (компоненты вектора) Av в одной точке требуется конечное число арифметических операций, поэтому на каждом шаге итерационного метода (8) затрачивается $O(h^{-2})$ арифметических операций. Общее количество арифметических операций, необходимых для получения решения с точностью ε, таким образом, равно $O\left(h^{-4}\ln(\varepsilon^{-1})\right) = O\left(N^4|\ln\varepsilon|\right)$.

В данном случае была описана схема применения метода простой итерации к решению системы сеточных эллиптических уравнений (1), аппроксимирующих исходную задачу в прямоугольной области; однако все рассуждения справедливы также и для случая произвольной области, если сетка выбирается равномерной, а граничные условия аппроксимируются их «сносом» в ближайший узел сеточной границы. Следует отметить, что при этом, вообще говоря, неизвестны точные границы λ_{min}, λ_{max} спектра матрицы системы. Эти величины можно оценить, например, следующим образом. Пусть сеточный прямоугольник K' со сторонами l_1' и l_2' содержится в $\bar{\Omega}_h$, а сеточный прямоугольник K'' со сторонами l_1'' и l_2'' содержит $\bar{\Omega}_h$. (Узлы K' являются узлами $\bar{\Omega}_h$ и узлы $\bar{\Omega}_h$ являются узлами K''.) Тогда имеет место соотношение

$$\lambda'_{min} \leqslant \lambda_{kl} \leqslant \lambda''_{max},$$

где λ'_{min} — минимальное собственное значение сеточной задачи Дирихле в прямоугольнике K', а λ''_{max} — максимальное собственное значение сеточной задачи Дирихле в прямоугольнике K'', Выбирая $\tau = 2/(\lambda'_{min} + \lambda''_{max})$, можно приближенно находить решение системы уравнений (1) методом простой итерации с той же асимптотикой числа арифметических операций, что и в предыдущем случае.

Задача 4. Показать, что для итерационного процесса с чебышевским набором параметров требуемое число операций $O(N^3|\ln\varepsilon|)$.

Задача 5. Получить такую же оценку числа действий для оптимального линейного итерационного процесса.

Задача 6. Получить такую же оценку числа действий для трехслойного итерационного процесса (см. § 6.6, задача 3) с фиксированным итерационным параметром ω.

Задача 7. Показать, что средняя трудоемкость трехслойного итерационного процесса (см. § 6.6) с фиксированным параметром ω может быть снижена вдвое за счет следующего обстоятельства. При нахождении u^l при l четном вычисляются и запоминаются только значения в точках с четной суммой $m+n$, а при l нечетном — в точках с нечетной суммой $m+n$.

Заметим, что в случае прямоугольника в методе переменных направлений (21) (если его рассматривать как итерационный метод) можно указать последовательность переменных шагов по времени, при которой общее число операций будет равно $O\left(N^2 \ln N|\ln\varepsilon|\right)$.

Заметим, что все описываемые выше методы укладываются в общую схему решения стационарных уравнений путем установления. В частности, двухслойные итерационные методы можно рассматривать как аппроксимацию уравнения

$$\frac{\partial u}{\partial t} = \Delta u + f,$$

а трехслойные — как аппроксимацию уравнения

$$\frac{\partial^2 u}{\partial t^2} + \gamma \frac{\partial u}{\partial t} = \Delta u + f,$$

В случае необходимости решения более сложных стационарных задач для уравнений с частными производными часто идут по такому пути. Строят нестационарный процесс, сходящийся к решению задачи, а затем в качестве итерационного процесса берут дискретную аппроксимацию этого нестационарного процесса. Например, в случае уравнения $\Delta u + f = 0$ один из таких процессов установления описывается уравнением

$$\frac{\partial^2 u}{\partial t \partial x} + \frac{\partial^2 u}{\partial t \partial y} + \gamma \frac{\partial u}{\partial t} = \Delta u + f.$$

Задача 8. Доказать, что при определенном соотношении между шагами по t, x, y сеточная аппроксимация этого уравнения превращается в метод сверхрелаксации для решения сеточного уравнения $\Delta^h u + f = 0$. В случае, когда область G есть квадрат и $h_1 = h_2 = h$, указать итерационный параметр γ, при котором число итераций для получения решения с точностью ε будет порядка $\dfrac{2}{\pi} h^{-1} \ln(\varepsilon^{-1})$.

Эффективными методами решения сеточных эллиптических уравнений являются интенсивно развиваемые в последнее время метод фиктивных компонент и многосеточный метод. По сути дела они также укладываются в общую схему построения итерационных методов, и проблема заключается в выборе соответствующего переобуславливателя.

В итерационном методе фиктивных компонент, предназначенном для решения сеточного уравнения Пуассона в области произвольного вида, на каждом шаге итерационного процесса необходимо решать первую краевую задачу для уравнения Пуассона в некотором прямоугольнике, содержащем внутри себя эту область. Если для решения последней задачи применяется какой-либо из эффективных методов (например, марш-алгоритм), то при любом p для получения решения с точностью $O(h^p)$ потребуется $O(h^{-2}|\ln h|)$ арифметических операций.

Метод Федоренко (называемый также многосеточным методом). К числу наиболее эффективных и употребляемых методов решения сеточных эллиптических задач (включая краевые задачи для систем уравнений Навье—Стокса) относится многосеточный метод, предложенный в шестидесятые годы. Сначала практическое использование этого метода носило эпизодический характер из-за неприспособленности существовавшего тогда программного обеспечения к использованию методов такого типа.

Основная идея этого метода заключается в следующем. Пусть решается сеточная краевая задача $L_h u_h = f_h$. Подбирается некоторый итераци-

онный процесс такой, что уже при небольшом числе итераций обеспечивается определенное сглаживание погрешности. Таким образом, решение исходной задачи сводится к решению задачи с относительно более гладким решением. Решение задачи с гладким решением на сетке с шагом h близко к решению задачи на сетке с более крупным шагом, например с шагом $2h$. Предлагается решить уравнение для погрешности на более грубой сетке и затем, проинтерполировав на исходную сетку, получить существенно лучшее приближение к решению. Для приближенного решения задачи на сетке с шагом $2h$ применяется аналогичная процедура с переходом к решению задачи на сетке с шагом $4h$ и так далее.

Часто одна итерация на сетке с шагом h состоит в дву- или трехкратном применении описанной процедуры перехода к решению задачи на сетке с шагом $2h$.

Рассмотрим этот метод на примере краевой задачи

$$-u'' = f \quad \text{при} \quad 0 < x < 1, \quad u(0) = u(1) = 0.$$

Соответствующая сеточная краевая задача $L_h u_h = f^h$ имеет вид:

$$-\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} = f_n = f(nh), \quad n = 1, \dots, N-1; \quad u_0 = u_N = 0.$$

Пусть $N = h^{-1}$ четное. Далее, как правило не оговариваясь, предполагаем что все рассматриваемые сеточные функции принадлежат подпространству функций U_h, удовлетворяющих условию $u_0 = u_N = 0$.

Опишем подробнее процедуру сведения решения на сетке с шагом h к решению задачи на сетке с шагом $2h$. Рассмотрим итерационный процесс

$$u_h^{k+1} = u_h^k - \tau(L_h u_h^k - f^h), \quad k = 0, \dots, m-1. \tag{10}$$

Вычитая это соотношение из равенства

$$u_h = u_h - \tau(L_h u_h - f^h),$$

получим уравнение относительно погрешности $r_h^k = u_h - u_h^k$:

$$r_h^{k+1} = (E - \tau L_h) r_h^k.$$

В качестве начального приближения возьмем $u_h^0 = 0$; тогда

$$r_h^0 = u_h, \quad r_h^m = (E - \tau L_h)^m u_h.$$

Далее в рассуждениях об уменьшении величины погрешности мы имеем в виду уменьшение погрешности решения u_h относительно погрешности приближения $u_h = 0$.

Функции $\varphi_q(n) = \sqrt{2} \sin \dfrac{\pi q n}{N}$, $q = 1, \dots, N-1$, образуют полную ортонормированную систему в пространстве функций U_h относительно скалярного произведения

$$(f, g)_h = h \sum_{n=1}^{N-1} f_n g_n$$

и являются собственными функциями для оператора L_h:

$$L_h\varphi_q = \frac{2-2\cos\dfrac{\pi q}{N}}{h^2}\,\varphi_q.$$

Поэтому

$$(E-\tau L_h)\varphi_q = \Bigl(1-4\tau h^{-2}\sin^2\frac{\pi q}{2N}\Bigr)\varphi_q. \tag{11}$$

Отсюда видно, что итерационный процесс (10) сходится при $0<\tau\leqslant h^2/2$. Если $\tau=\theta h^2$, где $0<\theta<1/2$ не зависит от h, то составляющие погрешности, соответствующие функциям φ_q со значениями q порядка N, то есть сильно колеблющиеся, умножаются на множители существенно меньшие единицы. Вследствие этого и происходит определенное сглаживание погрешности.

Положим далее $\tau=h^2/4$; тогда проводимые выкладки имеют наиболее простой вид. В частности,

$$(E-\tau L_h)u_h\Big|_{nh} = Q_h u_h\Big|_{nh}, \quad \text{где} \quad Q_h u_h\Big|_{nh} = \frac{u_{n+1}+2u_n+u_{n-1}}{4},$$

и соотношение (11) приобретает вид

$$Q_h\varphi_q = \frac{1+\cos\frac{\pi qn}{N}}{2}\,\varphi_q.$$

Обозначим $c_q=\frac{1+\cos\frac{\pi qn}{N}}{2}=\cos^2\frac{\pi q}{2N}$ и $s_q=\frac{1-\cos\frac{\pi qn}{N}}{2}=\sin^2\frac{\pi q}{2N}$. Таким образом,

$$Q_h\varphi_q = c_q\varphi_q, \quad {Q_h}^m\varphi_q = {c_q}^m\varphi_q,$$

Определим оператор Π_h^{2h} перехода с сетки с шагом h на сетку с шагом $2h$:

$$\Pi_h^{2h}g_h\Big|_{nh} = \frac{g_{n+1}+2g_n+g_{n-1}}{4}, \quad n \text{ — четное},$$

и оператор Π_{2h}^h перехода с сетки с шагом $2h$ на сетку с шагом h:

$$\Pi_{2h}^h g_{2h}\Big|_{nh} = \begin{cases} g_n & \text{при } n \text{ — четном,} \\ \dfrac{g_{n-1}+g_{n+1}}{2} & \text{при } n \text{ — нечетном.} \end{cases}$$

Лемма. *Справедливы неравенства*

$$\|\Pi_h^{2h}g_h\|_{2h}^2 \leqslant \|g_h\|_h^2, \quad \|\Pi_{2h}^h g_{2h}\|_h^2 \leqslant \|g_{2h}\|_{2h}^2. \tag{12}$$

Доказательство. Запишем неравенство Коши—Буняковского для скалярного произведения векторов $(1/s,\dots,1/s)$ и $(a_1,\dots,a_s)$:

$$\left(\frac{a_1+\dots+a_s}{s}\right)^2 \leqslant \frac{{a_1}^2+\dots+{a_s}^2}{s}.$$

Поэтому

$$\left(\frac{g_{n-1}+2g_n+g_{n+1}}{4}\right)^2=\left(\frac{g_{n-1}+g_n+g_n+g_{n+1}}{4}\right)^2\leqslant\frac{g_{n-1}^2+g_n^2+g_n^2+g_{n+1}^2}{4}.$$

Если $g_h\in U_h$, то после суммирования по четным n и домножения на $2h$, получаем первое из утверждений леммы.

Аналогично имеем

$$\left(\frac{g_{n-1}+g_{n+1}}{2}\right)^2\leqslant\frac{g_{n-1}^2+g_{n+1}^2}{2}.$$

Просуммируем по нечетным n и добавим сумму $\sum\limits_{j=1}^{N/2-1}g_{2j}^2$. После умножения на h получим второе утверждение леммы.

Если бы удалось решить точно уравнение

$$L_h r_h^m=Q_h{}^m f_h, \tag{13}$$

то, прибавив $r_h{}^m$ к $u_h{}^m$, мы получили бы точное решение задачи u_h.

Предположим, что известен алгоритм приближенного решения уравнения

$$L_{2h}r_{2h}=g_{2h},$$

такой, что приближенное решение может быть записано в виде

$$\tilde r_{2h}=A_{2h}^{\varepsilon_1}g_{2h}=r_{2h}-S_{2h}^{\varepsilon_1}r_{2h},$$

где $A_{2h}^{\varepsilon_1}$ и $S_{2h}^{\varepsilon_1}$ — некоторые линейные операторы и

$$\|S_{2h}^{\varepsilon_1}r_{2h}\|_{2h}\leqslant\varepsilon_1\|r_{2h}\|_{2h}. \tag{14}$$

Перенесем правую часть (13) на сетку с шагом $2h$ и применим этот алгоритм к уравнению

$$L_{2h}r_{2h}=\Pi_h^{2h}Q_h{}^m f_h.$$

Точное решение этого уравнения записывается в виде

$$r_{2h}=L_{2h}^{-1}\Pi_h^{2h}Q_h{}^m f_h=L_{2h}^{-1}\Pi_h^{2h}Q_h{}^m L_h u_h.$$

Проинтерполировав получившееся приближенное решение $\tilde r_{2h}=r_{2h}-S_{2h}^{\varepsilon_1}r_{2h}$ на сетку с шагом h, получим приближение

$$\tilde r_h=\Pi_{2h}^h\left(r_{2h}-S_{2h}^{\varepsilon_1}r_{2h}\right);$$

положим

$$u_h^1=u_h^m+\tilde r_h.$$

Погрешность получившегося приближения равна $R_h^1=r_h^m-\tilde r_h$. Выразив эту погрешность через решение исходной задачи, получим

$$R_h^1=z_h^1+z_h^2,$$

где

$$z_h^1 = {Q_h}^m u_h - \Pi_{2h}^h r_{2h} = \left({Q_h}^m - \Pi_{2h}^h L_{2h}^{-1} \Pi_h^{2h} {Q_h}^m L_h\right) u_h,$$
$$z_h^2 = \Pi_{2h}^h S_{2h}^{\varepsilon_1} r_{2h} = \Pi_{2h}^h S_{2h}^{\varepsilon_1} L_{2h}^{-1} \Pi_h^{2h} {Q_h}^m L_h u_h. \qquad (15)$$

Функции φ_q являются собственными для операторов L_h и L_{2h}:

$$L_h \varphi_q = \frac{4\sin^2 \frac{\pi q}{2N}}{h^2} \varphi_q = \frac{4s_q}{h^2} \varphi_q, \quad L_{2h} \varphi_q = \frac{4\sin^2 \frac{\pi q}{N}}{(2h)^2} \varphi_q = \frac{4 s_q c_q}{h^2} \varphi_q.$$

Справедливо равенство

$$\Pi_h^{2h} \varphi_{q\,2h} = c_q \varphi_{q_h}.$$

Поэтому

$$r_{2h} = L_{2h}^{-1} \Pi_h^{2h} L_h {Q_h}^m u_h = \sum_{k=1}^{N-1} \left(\frac{h^2}{4 s_q c_q}\right) c_q \left(\frac{4 s_q}{h^2}\right) {c_q}^m a_q \varphi_q = \sum_{q=1}^{N-1} {c_q}^m a_q \varphi_q.$$

Совпадение функции $L_{2h}^{-1} \Pi_h^{2h} L_h {Q_h}^m u_h$ и ${Q_h}^m u_h$ в узлах сетки с шагом $2h$ носит случайный характер и не имеет места в других случаях.

Справедливы равенства

$$\varphi_{N-q}(n) = \sin \frac{\pi (N-q) n}{N} = \begin{cases} \sin \frac{\pi q n}{N} = \varphi_q(n) & \text{при нечетном } n, \\ -\sin \frac{\pi q n}{N} = -\varphi_q(n) & \text{при четном } n, \end{cases}$$
$$\varphi_{N/2}(n) = 0 \quad \text{при четном} \quad n, \qquad (16)$$
$$s_{N/2} = c_{N/2} = \frac{1}{2},$$
$$s_q + c_q = 1, \quad c_q = s_{N-q}, \quad s_q = c_{N-q}.$$

Поэтому

$$r_{2h} = \sum_{q=1}^{N/2-1} ({c_q}^m a_q - {s_q}^m a_{N-q}) \varphi_q.$$

Согласно равенству Парсеваля

$$\|z_{2h}\|_{2h} = \sum_{q=1}^{N/2-1} ({c_q}^m a_q - {s_q}^m a_{N-q})^2, \quad \|u_h\|_h = \sum_{q=1}^{N-1} a_q^2.$$

Из неравенства Коши—Буняковского следует

$$({c_q}^m a_q - {s_q}^m a_{N-q})^2 \leqslant ({c_q}^{2m} + {s_q}^{2m})({a_q}^2 + {a_{N-q}}^2).$$

Вследствие (16) имеем

$${c_q}^{2m} + {s_q}^{2m} \leqslant c_q + s_q = 1.$$

Поэтому

$$({c_q}^m a_q - {s_q}^m a_{N-q})^2 \leqslant {a_q}^2 + {a_{N-q}}^2.$$

После суммирования по q получаем

$$\|z_{2h}\|_{2h}^2 \leqslant \|u_h\|_h^2.$$

Отсюда и из (12), (14) получаем оценку

$$\|z_h^2\|_h = \|\Pi_{2h}^h S_{2h}^{\varepsilon_1} L_{2h}^{-1} \Pi_h^{2h} L_h {Q_h}^m\|_h \leqslant \varepsilon_1 \|u_h\|_h.$$

Исходя из равенств

$$\Pi_{2h}^h \varphi_q\Big|_{nh} = \begin{cases} \varphi_q(n) & \text{при четном } n, \\ \cos\frac{\pi q}{N}\varphi_q & \text{при нечетном } n \end{cases}$$

и равенств (16), подберем a_q и b_q так, чтобы при всех n выполнялось равенство

$$\Pi_{2h}^h \varphi_q|_{nh} = a_q\varphi_q(n) + b_q\varphi_{N-q}(n).$$

Получим уравнения

$$a_q - b_q = 1 = c_q + s_q, \quad a_q + b_q = \cos\frac{\pi q}{N} = c_q - s_q,$$

и следовательно, $a_q = c_q$, $b_q = -s_q$. Поэтому

$$\Pi_{2h}^h r_{2h} = \sum_{q=1}^{N-1} {c_q}^m a_q (c_q\varphi_q - s_q\varphi_{N-q}) = \sum_{q=1}^{N-1} ({c_q}^{m+1} a_q - {s_q}^m c_q a_{N-q})\varphi_q.$$

Подставляя это разложение в (15), получим равенство

$$z_h^1 = {Q_h}^m - \Pi_{2h}^h r_{2h} = \sum_{q=1}^{N-1} \left({c_q}^m a_q - ({c_q}^{m+1} a_q - {s_q}^m c_q a_{N-q})\right)\varphi_q = \\ = \sum_{q=1}^{N-1} ({c_q}^m s_q a_q + {s_q}^m c_q a_{N-q})\,\varphi_q.$$

Согласно неравенству Коши—Буняковского

$$({c_q}^m s_q a_q + {s_q}^m c_q a_{N-q})^2 \leqslant ({c_q}^{2m} {s_q}^2 + {s_q}^{2m} {c_q}^2)({a_q}^2 + {a_{N-q}}^2).$$

Первый множитель записывается в виде

$$g_m(y) = (1-y)^{2m} y^2 + y^{2m}(1-y)^2, \text{ где } y = s_q.$$

Если ввести обозначение $\bar{g}_m = \max\limits_{[0;1]} g_m(y)$, то при всех q будет справедливо неравенство

$$(c_q{}^m s_q a_q + s_q{}^m c_q a_{N-q})^2 \leqslant \bar{g}_m(a_q{}^2 + a_{N-q}{}^2),$$

и поэтому

$$\|z_h^1\|_h^2 = \sum_{q=1}^{N-1} (c_q{}^m s_q a_q + s_q{}^m c_q a_{N-q})^2 \leqslant$$

$$\leqslant \bar{g}_m \sum_{q=1}^{N-1} (a_q{}^2 + a_{N-q}{}^2) \leqslant 2\bar{g}_m \sum_{q=1}^{N-1} a_q{}^2 = 2\bar{g}_m \|u_h\|_h^2.$$

Отсюда получаем оценку $\|z_h^1\|_h \leqslant G_m \|u_h\|_h$. где $G_m = \sqrt{2\bar{g}_m}$.

Из этой оценки и оценки для $\|z_h^1\|$ следует, что

$$\|R_h^1\|_h \leqslant (G_m + \varepsilon_1)\|R_h^0\|_h. \tag{17}$$

Здесь $R_h^0 = u_h$ — ошибка начального приближения: $u_h^0 = 0$.

Оценим величину G_2. Справедливо равенство $g_2(y) = (1-2u)u^2$, где $u = y(1-y)$. Согласно формуле дифференцирования сложной функции

$$\frac{dg_2(y)}{dy} = u(2-6u)(1-2y).$$

Отсюда получаем, что производная функции $g_2(y)$ обращается в нуль в точках $0, 1/2, 1, 1/2 \pm \mathbf{i}\sqrt{1/12}$; поэтому, исследуя график функции $g_2(y)$, получаем $\bar{g}_2 = g_2(1/2) = 1/32, G_2 = 0{,}25$.

Задача 10. Показать, что

$$G_m < \sqrt{\frac{2}{e^2 m^2} + \frac{1}{2 \cdot 4^m}} < 0{,}59/m$$

при $m > 2$.

Подведем итог проведенных построений. Начав с приближения $u_h^0 = 0 = u_h - u_h$, мы получили новое приближение $u_h^1 = u_h - S_h^{\varepsilon_0} u_h$, где $\varepsilon_0 = G_m + \varepsilon_1$,

$$S_h^{\varepsilon_0} = Q_h{}^m - \Pi_{2h}^h L_{2h}^{-1} \Pi_h^{2h} Q_h{}^m L_h +$$

$$\Pi_{2h}^h S_{2h}^{\varepsilon_1} L_{2h}^{-1} \Pi_h^{2h} Q_h{}^m L_h.$$

Таким образом, воспользовавшись алгоритмом уменьшения погрешности в $1/\varepsilon_1$ раз на сетке с шагом $2h$ и произведя дополнительно $O(mN)$ арифметических операций, мы добились уменьшения погрешности на сетке с шагом h в $1/\varepsilon_0$ раз, где $\varepsilon_0 = G_m + \varepsilon_1$.

Далее фиксируем $m = 2$ и возьмем $\varepsilon_1 = 0{,}25$. Из оценки (17) следует, что при использовании описанной процедуры норма погрешности решения на сетке с шагом h, умножится на множитель не больший 0,5. После

повторного применения этой процедуры норма погрешности решения на сетке с шагом h умножится на множитель не больший 0,25.

В итоге после двукратного использования алгоритма уменьшения погрешности в 4 раза на сетке с шагом $2h$, осуществляя дополнительно не более $C(2)N$ арифметических операций, мы получаем алгоритм уменьшения погрешности в 4 раза на сетке с шагом h.

Обозначим число арифметических операций, достаточное для уменьшения погрешности в 4 раза на сетке с шагом $h_l = 2^{-l}$ через $Z(l)$.

Полученный выше результат можно записать в виде неравенства

$$Z(l) \leqslant 2Z(l-1) + C(2)\,2^l.$$

Функция $W(l) = C(2)\,l\,2^l$ удовлетворяет уравнению

$$W(l) = 2W(l-1) + C(2)\,2^l.$$

При $l = 1$ рассматриваемую сеточную задачу можно решить, совершив не более чем 3 арифметические операции. Поскольку $C(2) > 2$, то $Z(1) \leqslant W(1)$.

Индукцией по l можно получить оценку $Z(l) \leqslant W(l) = O(N \log_2 N)$ при $N = 2^l$.

Если требуется добиться уменьшения погрешности в M раз, то будет достаточно $[1+\log_4 M]$ итераций, и, таким образом, общее число действий, требуемое для решения задачи, будет $O(N \log_2 N \log_2 M)$.

Эта оценка хуже, чем оценка $O(N)$ числа действий методов прогонки или стрельбы.

Однако в рассматриваемом методе можно уменьшить число арифметических операций. Во-первых, можно показать, что при некотором видоизменении итерационного процесса (10) при решении задачи на сетке с шагом h достаточно лишь один раз обращаться к решению задачи на сетке с шагом $2h$.

В результате этого число операций, требуемое для уменьшения нормы погрешности на сетке с шагом h в 4 раза снижается до $O(N)$.

Можно предложить другой вариант уменьшения нормы погрешности, например, в 4 раза на сетке с шагом $h = 2^{-l}$ с затратой $O(N)$ арифметических действий. Положим $\eta_j = 1/(4(l-j+1)), m_j = [4(l-j+1) \times (l-j+2)0{,}59]+1$ при $j = 1, \ldots, m_l$. Применим описанный выше алгоритм последовательного сведения решения задачи на сетке с шагом $h_j = 2^{-j}$ к решению задачи на сетке с шагом $h_j = 2^{-(j-1)}$; при этом для каждого $j = l, \ldots, 2$ производим m_j итераций по формуле (10) при $\tau = \tau_j = h_j^2/4$. Задачу на сетке с шагом $h_1 = 1/2$ решаем точно.

Задача 11. Доказать, что при таком выборе η_j и m_j погрешность решения на сетке с шагом $h = 2^{-l}$ уменьшится в 4 раза, а общее число арифметических действий, затрачиваемое при этом, есть $O(N)$.

При решении задачи существенно используется то, что при всех j выполняются наравенства $\eta_j \geqslant \eta_{j-1} + 0{,}59/m_j$.

Во вторых, можно принять во внимание следующее обстоятельство. При дважды дифференцируемой $f(x)$ решения задач на сетках с шагами 2^{-l} и $2^{-(l-1)}$ отличаются на $O(2^{-2l})$. Поэтому можно поступить следующим образом.

Задаемся некоторой последовательностью убывающих величин $\delta_l = \text{const} \cdot 2^{-2l}$. Последовательно решаем задачи на сетках с шагами $h_l = 2^{-(l-1)}$ с погрешностью итерационного процесса порядка δ_{l-1}, и получившиеся приближения берем в качестве исходных для итераций на сетке с шагом $h_l = 2^{-l}$. При этом оказывается, что при каждом l требуется лишь конечное число итераций описанного выше вида. В результате решение исходной сеточной задачи будет получено с погрешностью $O(1/N^2)$ при общем числе арифметических операций $O(N)$.

В рассматриваемом случае при некотором видоизменении итерационного процесса (10) в результате одного полного цикла итерации — спуска от шага $h = 2^{-l}$ до шага $h = 2^{-1}$ и возвращения к шагу $h = 2^{-l}$ с общей затратой $O(N)$ арифметических операций получается точное решение сеточной задачи. Однако этот факт носит случайный характер и не относится к более сложным задачам.

Рассмотрим случай размерности задачи $s > 1$. Пусть используется рассмотренный алгоритм сведения решения задачи на сетке с шагом h к решению задачи на сетке с шагом $2h$ и число таких сведений при каждом шаге равно $q > 1$.

Для весьма широкого класса задач можно доказать существование $\varepsilon(q) > 0$ и $m(q) < \infty$, удовлетворяющих следующим соотношениям. После q-кратного применения описанной выше процедуры, состоящей из $m(q)$ итераций (10), перехода к задаче на сетке с шагом $2h$, решения задачи на этой сетке с помощью алгоритма $S_{2h}^{\varepsilon(q)}$ и перехода на сетку с шагом h, некоторая норма погрешности решения задачи на сетке с шагом h уменьшается в $1/\varepsilon(q)$ раз.

Естественно, что операторы перехода с одной сетки на другую Π_h^{2h} и Π_{2h}^h будут иными, чем выше.

Обозначим число арифметических операций, достаточное для уменьшения погрешности в $1/\varepsilon(q)$ раз на сетке с шагом $h = 2^{-l}$, через $Z(l)$. Справедливо неравенство $Z(l) \leqslant qZ(l-1) + C_0(s,q)2^{sl}$.

Отсюда при $q < 2^s$ можно получить неравенство $Z(l) \leqslant C_1(s,q)2^{sl}$.

Таким образом, общее число действий, достаточное для уменьшения погрешности на сетке с шагом h в $1/\varepsilon(q)$ раз, будет порядка $\mathrm{O}(h^{-s})$, то есть порядка $\mathrm{O}(N)$, где N — общее число узлов сеточной задачи.

Высказанные выше утверждения относятся и к случаю аппроксимаций метода конечных элементов.

Литература

1. Бахвалов Н. С. О сходимости одного релаксационного метода при естественных ограничениях на эллиптический оператор. // ЖВМиМФ — 1966, т. 6, N 5, с. 861–883.
2. Березин И. С., Жидков Н. П. Методы вычислений. Т. 2. — М.: Физматгиз, 1962.
3. Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления. — М.: Наука, 1984.
4. Годунов С. К., Рябенький В. С. Введение в теорию разностных схем. — М.: Наука, 1962.
5. Годунов С. К., Рябенький В. С. Разностные схемы. — М.: Наука, 1977.
6. Годунов С. К., Забродин А. В. и др. Численное решение многомерных задач газовой динамики. — М.: Наука, 1976.
7. Денисов А. М. Введение в теорию обратных задач — М.: Изд-во МГУ, 1994.
8. Джордж А., Лю Д. Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.
9. Дьяконов Е. Г. Минимизация вычислительной работы. Асимптотически оптимальные алгоритмы для эллиптических задач. — М.: Наука, 1989.
10. Зенкевич О., Морган К. Конечные элементы и аппроксимация. — М.: Мир, 1980.
11. Кобельков Г. М. Решение задачи о стационарной свободной конвекции. // ДАН СССР. — 1980, **225**, N 2, с. 277-282.
12. Кобельков Г. М. О методах решения уравнений Навье-Стокса. // Вычислительные процессы и системы. — М.: Наука, 1991, вып.8, с. 204–236.
13. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Уравнения в частных производных. — Минск: Наука и техника, 1982.
14. Локуциевский О. В., Гавриков М. Б. Начала численного анализа. — М.: ТОО «Янус», 1995.
15. Марчук Г. И. Методы вычислительной математики. — М.: Наука, 1980.
16. Марчук Г. И., Агошков В. И. Введение в проекционно-разностные методы. — М.: Наука, 1981.
17. Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1981.
18. Марчук Г. И., Шайдуров В. В. Повышение точности решений разностных схем. — М.: Наука, 1979.
19. Марчук Г. И., Яненко Н. Н. Применение метода расщепления (дробных шагов) для решения задач математической физики. — В кн.: Некоторые вопросы вычислительной и прикладной математики.— Новосибирск: Наука, 1966.
20. Рябенький В. С., Филиппов А. Ф. Об устойчивости разностных уравнений. — М.: Гостехиздат, 1956.
21. Самарский А. А. Теория разностных схем. — М.: Наука, 1982.
22. Самарский А. А., Андреев В.Б. Разностные методы для решения эллиптических уравнений. — М.: Наука, 1976.
23. Самарский А. А., Гулин А. В. Устойчивость разностных схем. — М.: Наука, 1973.
24. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.

25. Самарский А. А., Попов Ю. П. Разностные методы решения задач газовой динамики. — М.: Наука, 1980.
26. Самарский А. А., Капорин И. Е., Кучеров А. Б., Николаев Е. С. Некоторые современные методы решения сеточных уравнений. // Изв. вузов. Сер. мат., 1983, N 7 (254). С. 3–12.
27. Саульев В. К. Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз, 1960.
28. Стрэнг Г., Фикс Дж. Теория метода конечных элементов. — М.: Мир, 1977.
29. Федоренко Р. П. Релаксационный метод решения разностных эллиптических уравнений.// ЖВМиМФ — 1961, т.1, N 5. С. 922–927.
30. Федоренко Р. П. Введение в вычислительную физику. — М.: Изд-во МФТИ, 1994.
31. Шайдуров В. В. Многосеточные методы конечных элементов. — М.: Наука, 1989.
32. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука, 1967.
33. Hackbusch W. Multi-Grid Methods and Applications. — Springer-Verlag, Berlin—Heidelberg, 1985.

Глава 11

Численные методы решения интегральных уравнений

В этой главе мы дадим краткое описание алгоритмов решения интегральных уравнений, не вникая подробно в вопросы оценки погрешности.

Задача решения интегральных уравнений возникает как вспомогательная при решении краевых задач для дифференциальных уравнений с частными производными и как самостоятельная при исследовании работы ядерных реакторов, при решении так называемых обратных задач геофизики, при обработке результатов наблюдений и т. п. Мы ограничимся рассмотрением случая интегральных уравнений с одной неизвестной функцией и одной независимой переменной.

§ 1. Решение интегральных уравнений методом замены интеграла квадратурной суммой

В теории численных методов решения интегральных уравнений рассматриваются следующие типичные задачи. Найти решение интегрального уравнения Фредгольма первого рода

$$Gy = \int_a^b K(x, s)y(s)\,ds = f(x), \tag{1}$$

интегрального уравнения Фредгольма второго рода

$$y - \lambda Gy = y - \lambda \int_a^b K(x, s)y(s)\,ds = f(x), \tag{2}$$

интегрального уравнения Вольтерра первого рода

$$Gy = \int_a^x K(x, s)y(s)\,ds = f(x), \tag{3}$$

интегрального уравнения Вольтерра второго рода

$$y - \lambda Gy = y - \lambda \int_a^x K(x, s)y(s)\,ds = f(x) \tag{4}$$

и задачи на собственные значения

$$Gu = \lambda u. \tag{5}$$

В последнем случае ищутся числа λ, при которых задача (5) имеет ненулевое решение.

Воспользуемся какой-либо формулой численного интегрирования

$$J(\psi) = \int_a^b \psi(x)\,dx \approx S_m(\psi) = \sum_{j=1}^{m} c_j \psi\left(x_j^{(m)}\right), \tag{6}$$

где c_j, вообще говоря, зависят от m. Имеем равенство

$$J(\psi) = S_m(\psi) + R_m(\psi), \tag{7}$$

где $R_m(\psi)$ — остаточный член квадратурной формулы (6).

Рассмотрим для примера уравнение Фредгольма второго рода (2). С помощью соотношения (7) его можно переписать в виде

$$y(x) - \lambda \sum_{j=1}^{m} c_j K\left(x, x_j^{(m)}\right) y\left(x_j^{(m)}\right) - R_m(\lambda K y) = f(x); \tag{8}$$

остаточный член $R_m(\lambda K y)$ при вычислении интеграла $\lambda \int_a^b K(x, s) y(s)\,ds$ с помощью квадратуры (6) является функцией переменной x. Полагая в (8) $x = x_i^{(m)}$, $i = 1, \ldots, m$, получим систему уравнений

$$y\left(x_i^{(m)}\right) - \lambda \sum_{j=1}^{m} c_k K\left(x_i^{(m)}, x_j^{(m)}\right) y\left(x_j^{(m)}\right) = f\left(x_i^{(m)}\right) + R_m(\lambda K y)\Big|_{x_i^{(m)}}.$$

Отбрасывая остаточный член, приходим к системе линейных алгебраических уравнений

$$y_i - \lambda \sum_{j=1}^{m} c_j K\left(x_i^{(m)}, x_j^{(m)}\right) y_j = f_i, \quad f_i = f\left(x_i^{(m)}\right), \quad i = 1, \ldots, m. \tag{9}$$

Для решения этой задачи могут быть применены стандартные методы решения систем линейных алгебраических уравнений.

Будем рассматривать случай вещественных $K(x, s)$ и $f(x)$. Обратим внимание на следующее обстоятельство. Если ядро $K(x, s)$ интегрального оператора G симметрично ($K(s, x) \equiv K(x, s)$), то оператор $E - G$, входящий в левую часть исходного уравнения (2), также симметричен.

Однако матрица системы уравнений (9) не обязательно будет симметричной. Мы видели ранее, что решение систем уравнений с симметричной матрицей в определенном смысле предпочтительнее решения системы уравнений с несимметричной матрицей: шире класс точных и итерационных методов, которые могут быть применены для решения таких систем.

Систему уравнений (9) можно преобразовать к виду, в котором матрица системы будет симметричной. Для этого умножим i-е уравнение системы (9) на c_i; получим систему уравнений

$$c_i y_i - \lambda \sum_{j=1}^{m} c_i c_j K\left(x_i^{(m)}, x_j^{(m)}\right) y_j = c_i f_i, \quad i = 1, \ldots, m, \tag{10}$$

уже с симметричной матрицей.

Другой возможный способ симметризации состоит в следующем. Умножим i-е уравнение в (9) на $\sqrt{c_i}$ и положим $\sqrt{c_i} y_i = z_i$. Получим систему уравнений

$$z_i - \lambda \sum_{j=1}^{m} \sqrt{c_i}\sqrt{c_j} K\left(x_i^{(m)}, x_j^{(m)}\right) z_j = \sqrt{c_i} f_i. \tag{11}$$

В случае $c_i > 0$ второй способ симметризации является более предпочтительным, поскольку разброс собственных значений у матрицы системы (11), как правило, меньше, чем у матрицы системы (10).

Заметим, что в случае, когда в квадратуре (6) все веса одинаковы:

$$c_1^{(m)} = \cdots = c_m^{(m)} = (b-a)/m, \tag{12}$$

необходимость в таких симметризациях отпадает.

Задача 1. Рассмотреть случай комплексного самосопряженного ядра $K(x, s) = \overline{K(x, s)}$. Проверить, что в этом случае при использовании описанных способов получается система уравнений с самосопряженным ядром.

Конечно, как и в случае решения произвольных систем линейных алгебраических уравнений, при использовании методов Гаусса или квадратного корня в процессе вычислений может возникнуть операция деления на нуль или переполнение.

Задача 2. Рассмотреть случай формулы прямоугольников, когда выполнено соотношение (12), и постоянного ядра

$$K(x, s) \equiv K = \text{const}.$$

Применить для решения системы (9) метод исключения Гаусса при естественном порядке исключения неизвестных $y_1, \ldots, y_m$. Показать, что при $1 - \lambda K(b-a) \geqslant \varepsilon > 0$ в ходе исключения по методу Гаусса (§ 6.1) абсолютные величины всех встречающихся элементов a_{ij}^l равномерно ограничены сверху некоторой постоянной $\kappa(\varepsilon)$, зависящей только от ε и не зависящей от m:

$$\left|a_{ij}^l\right| \leqslant \kappa(\varepsilon) < \infty.$$

Показать, что

$$\overline{\lim}_{m\to\infty} \sup_{i,j,l} \left|a_{ij}^{l}\right| = \infty$$

при $1 - \lambda K(b-a) < 0$.

Задача 3. Показать, что при условии

$$|\lambda| \sum_{j=1}^{m} \left| c_j K\left(x_i^{(m)}, x_j^{(m)}\right)\right| \leqslant 1 - \varepsilon, \quad \varepsilon > 0 \tag{13}$$

при решении системы (9) методом Гаусса всегда

$$\left|a_{ij}^{l}\right| \leqslant \kappa_1(\varepsilon) < \infty \quad \text{при любых} \quad m, i, j, l.$$

При расмотрении задачи на собственные значения (5) при аппроксимации интеграла квадратурной суммой (6) возникает алгебраическая задача на собственные значения:

$$\sum_{j=1}^{m} c_j K\left(x_i^{(m)}, x_j^{(m)}\right) y_j = \lambda y_j. \tag{14}$$

Для ее решения могут быть применены стандартные методы решения задач на собственные значения.

В случае $c_j > 0$ и симметричного ядра $K(x, s)$, как и выше, задачу на собственные значения можно преобразовать в задачу на собственные значения для симметричной матрицы с помощью введения новых переменных $\sqrt{c_i} y_i = z_i$.

Задача 4. Пусть все $c_j K\left(x_i^{(m)}, x_j^{(m)}\right) \geqslant 0$ и среди них есть ненулевые. Доказать, что максимальное по модулю собственное значение λ_1 задачи (14) положительно. Доказать, что среди собственных векторов, соответствующих этому собственному значению, имеется вектор, у которого все компоненты неотрицательны.

Указание. Один из возможных путей решения задачи заключается в рассмотрении итерационного процесса (гл. 6)

$$y_i^{(m+1)} = \sum_{j=1}^{m} c_j K\left(x_i^{(m)}, x_j^{(m)}\right) y_j^{(m)}$$

нахождения собственного вектора при начальном условии $\left(y_1^{(0)}, \ldots, y_m^{(0)}\right)^T$ с положительными компонентами $y_i^{(0)}$.

Интегральное уравнение Вольтерра второго рода (4) можно записать в виде

$$y(x) - \lambda \int_a^b \tilde{K}(x, s) y(s)\, ds = f(x)$$

с ядром $\tilde{K}(x, s) = \begin{cases} K(x, s) & \text{при } s \leqslant x, \\ 0 & \text{при } s > x \end{cases}$ и применить описанный выше алгоритм решения интегрального уравнения Фредгольма второго рода.

Однако на таком пути мы можем получить методы с плохой сходимостью при $m \to \infty$: погрешность квадратурной формулы (6) может оказаться большой, поскольку при $s > x$ подынтегральная функция равна нулю и на всем отрезке $[a, b]$ разрывна (если $K(x, s) \neq 0$) или не обладает высокой гладкостью.

Поэтому целесообразнее поступить следующим образом. Зададимся каким-либо набором точек $a \leqslant x_1^{(1)} < \cdots < x_m^{(m)} \leqslant b$. Выпишем соотношения

$$y\left(x_i^{(m)}\right) - \lambda \int_a^{x_i^{(m)}} K(x_i^m, s)y(s)\,ds = f\left(x_i^{(m)}\right)$$

и для вычисления интегралов

$$\int_a^{x_i^{(m)}} K(x_i^m, s)y(s)\,ds \tag{15}$$

применим какие-либо квадратурные формулы достаточно высокой точности, использующие значения подынтегральной функции в точках $x_1^{(m)}, \ldots, x_m^{(m)}$. Если при вычислении интеграла используются значения подынтегральной функции лишь в точках $x_1^{(m)}, \ldots, x_i^{(m)}$, то матрица системы уравнений (9) будет левой треугольной и отыскание решения системы (9) существенно упростится.

Рассмотрим следующую схему решения задачи (4). Пусть $K(x, s)$ и $f(s)$ дифференцируемы l раз. Можно показать, что тогда решение $y(x)$ также l раз дифференцируемо. Положим $x_i^{(m)} = a + (i-1)H$, $H = (b-a)/(m-1)$. Для вычисления интегралов $\int_a^{x_i^{(m)}} K(x, s)f(s)\,ds$ применим формулу Грегори с порядком точности $O(H^l(x_i^{(m)} - a))$. Такие формулы определены при $i \geqslant l$. При $i < l$ для вычисления этих интегралов применим какие-либо квадратурные формулы по узлам $x_0^{(m)}, \ldots, x_{l-1}^{(m)}$ с точностью порядка $O(H^l)$ или $O(H^{l+1})$.

В итоге получим систему уравнений (9), у которой выше главной диагонали ненулевые элементы могут находиться только в первых l строках и столбцах. Решаем систему из первых l уравнений системы (9) относительно l неизвестных $y_1, \ldots, y_l$; затем находим остальные y_i, решая систему уравнений с левой треугольной матрицей.

Если ядро $K(x, s)$ и правая часть $f(s)$ — аналитические функции, то иногда целесообразнее использовать идею широко известного метода Батчера решения дифференциальных уравнений. В этом случае получится система уравнений с полностью заполненной матрицей, но при этом погрешность решения убывает как q^m, $q < 1$.

Если $K(x, x) \neq 0$, то дифференцируя по x уравнение Вольтерра первого рода (3), мы сведем его решение к решению интегрального уравнения Вольтерра второго рода

$$K(x, x)y(x) + \int_a^x K_x(x, s)y(s)\,ds = f_x(x). \tag{16}$$

§ 2. Решение интегральных уравнений с помощью замены ядра на вырожденное

Другой классический способ решения интегральных уравнений, применяемый в случае задач (1.2), (1.4), состоит в замене $K(x, s)$ — ядра интегрального оператора на вырожденное.

Вырожденным называется ядро, представимое в виде конечной суммы

$$K(x, s) = \sum_{j=1}^{q} c_j(x)d_j(s), \quad q < \infty.$$

Пусть

$$K(x, s) \approx K^0(x, s) = \sum_{j=1}^{q} c_j(x)d_j(s). \tag{1}$$

Для определенности будем предполагать, что $c_1(x), \ldots, c_q(x)$ линейно независимы и $d_1(s), \ldots, d_q(s)$ также линейно независимы. В противном случае ядро $K^0(x, s)$ можно записать в виде (1) с меньшим значением q.

В случае (1) есть основания ожидать, что решение уравнения (1.2) близко к решению уравнения

$$y(x) - \lambda \int_a^b K^0(x, s)y(s)\,ds = f(x). \tag{2}$$

Подставляя выражение $K^0(x, s)$ в (2), получим равенство

$$y(x) = f(x) + \lambda \int_a^b \sum_{j=1}^{q} c_j(x)d_j(s)y(s)\,ds. \tag{3}$$

Следовательно,

$$y(x) = f(x) + \lambda \sum_{j=1}^{q} A_j c_j(x), \tag{4}$$

где

$$A_j = \int_a^b d_j(s)y(s)\,ds.$$

Таким образом, решение уравнения (2) сводится к определению коэффициентов A_j.

Подставляя выражение (4) для $y(x)$ в (3), получим уравнение

$$\lambda\sum_{i=1}^{q}A_ic_i(x)-\lambda\int_a^b\sum_{i=1}^{q}c_i(x)d_i(s)\left(f(s)+\lambda\sum_{j=1}^{q}A_jc_j(s)\right)ds=0;$$

при получении этого уравнения в двух случаях индекс суммирования j переобозначен через i. Последнее уравнение можно переписать в виде

$$\sum_{i=1}^{q}B_ic_i(x)=0,$$

где

$$B_i=A_i-\int_a^b d_i(s)f(s)\,ds-\lambda\sum_{j=1}^{q}A_j\int_a^b d_i(s)c_j(s)\,ds.$$

Вследствие линейной независимости функций $c_i(x)$ имеем $B_i=0$. Таким образом, мы получили систему уравнений относительно неизвестных A_i:

$$A_i-\lambda\sum_{j=1}^{q}(d_i,\,c_j)A_j=(d_i,\,f);\tag{5}$$

здесь $(g,\,f)=\int_a^b g(x)f(x)\,dx$ — скалярное произведение. После решения этой системы получаем приближение к решению задачи, которое имеет вид

$$y(x)\approx f(x)+\lambda\sum_{j=1}^{q}A_jc_j(x).$$

Если применяется метод из § 1, то при больших m решение системы (1.9) методом Гаусса может оказаться невозможным из-за недопустимо большого (порядка m^3) числа операций.

Если

$$\sum_{j=1}^{m}\left|\lambda c_jK\left(x_i^{(m)},\,x_j^{(m)}\right)\right|\leqslant q<1,\tag{6}$$

то для решения системы уравнений (1.9) можно применить метод простой итерации

$$y_i^{n+1}=\lambda\sum_{j=1}^{m}c_jK\left(x_i^{(m)},\,x_j^{(m)}\right)y_j^n+f_i.\tag{7}$$

Если переписать этот итерационный процесс в векторном виде

$$\mathbf{y}^{n+1}=B\mathbf{y}^n+\mathbf{f},$$

то вследствие (6) имеем $\|B\|_\infty\leqslant q<1$ и погрешность итерационного процесса убывает как $O(q^n)$. Если q не очень близко к 1, то применение итерационного процесса (7) целесообразнее применения метода Гаусса.

Если итерационный процесс сходится медленно или не сходится, то применяется, например, следующий прием. В уравнение (1.2) вводится новая неизвестная функция

$$z(x) = y(x) - \lambda \int_a^b K^0(x,\, s) y(s)\, ds$$

и оно преобразуется к виду

$$z(x) - \int_a^b H(x,\, s) z(s)\, ds = g(s). \tag{8}$$

Оказывается, что $\|H(x,\, s)\|_\infty \to 0$, если $\|K^0(x,\, s) - K(x,\, s)\|_\infty \to 0$. Поэтому при достаточно малой величине $\|K^0(x,\, s) - K(x,\, s)\|_\infty$ и разумном выборе квадратурной формулы окажется, что система уравнений (1.9), соответствующая уравнению (8), удовлетворяет условию (6), и итерационный процесс (7) быстро сходится.

Рассмотрим другой способ получения уравнения вида (8) с малой $\|H\|_\infty$. Определим интегральный оператор

$$Q_0 g = \int_a^b K_0(x,\, s) g(s)\, ds.$$

Обозначим через G_0 оператор, обратный к оператору $E - \lambda Q_0$, т. е. $G_0 f$ равно решению уравнения $y - \lambda Q_0 y = f$. Применяя к (1.2) оператор G_0, получим уравнение

$$G_0(E - \lambda Q)\, y = G_0 f,$$

которое может быть записано в виде

$$y(x) - \int_a^b R(x,\, s) y(s)\, ds = h(x), \tag{9}$$

где опять-таки $\|R\|_\infty \to 0$, если $\|K - K_0\|_\infty \to 0$.

В вычислительной практике наиболее употребителен дискретный вариант описанных алгоритмов. По аналогии со случаем вырожденных ядер можно определить класс вырожденных матриц. Пусть M_q^m — множество квадратных матриц размерности m, представимых в виде

$$S = \sum_{j=1}^{q} \mathbf{c}_j \mathbf{d}_j^T;$$

матрица $\mathbf{c}_j \mathbf{d}_j^T$ является квадратной матрицей размерности m и ранга 1.

Аналогично способу решения интегральных уравнений Фредгольма второго рода можно построить способ решения алгебраических систем с «вырожденной» матрицей

$$\mathbf{y} - \lambda S \mathbf{y} = \mathbf{f}, \quad S \in M_q^m;$$

вектор $\mathbf{y}$ отыскивается в виде $\mathbf{y} = \mathbf{f} + \sum_{j=1}^{q} A_j \mathbf{c}_j$. Относительно неизвестных A_j получаем систему уравнений (5), где $(\cdot,\, \cdot)$ — обычное скалярное произведение векторов.

Рассмотрим простейший случай. Пусть в (1.6) применяется формула прямоугольников

$$\int_a^b \psi(x)\,dx \approx \sum_{i=1}^{m} \frac{1}{m}\,\psi\left(x_i^{(m)}\right), \quad x_i^{(m)} = a + \frac{(2i-1)}{2m}(b-a);$$

система уравнений (1.9) имеет вид

$$y_i - \frac{\lambda}{m}\sum_{j=1}^{m} K\left(x_i^{(m)},\, x_j^{(m)}\right) y_j = f_i \tag{10}$$

или в векторной форме

$$\mathbf{y} - \lambda K^{(m)}\mathbf{y} = \mathbf{f}.$$

Пусть $m = sl$, l целое, s нечетное. Определим матрицу S размерности $m \times m$ по правилу: ее элементы s_{ij} равны

$$\frac{1}{m} K\left(x^{(l)}_{\alpha(m,i,j)},\, x^{(l)}_{\beta(m,i,j)}\right),$$

где $\left(x^{(l)}_{\alpha(m,i,j)},\, x^{(l)}_{\beta(m,i,j)}\right)$ — ближайшая к $(x_i^{(m)},\, x_j^{(m)})$ из точек $(x_\alpha^{(l)},\, x_\beta^{(l)})$. Близость измеряется как максимум модулей разностей первых и вторых компонент. Таким образом,

$$\left|x^{(l)}_{\alpha(m,i,j)} - x_i^{(m)}\right| < \frac{b-a}{2l}, \qquad \left|x^{(l)}_{\beta(m,i,j)} - x_j^{(m)}\right| < \frac{b-a}{2l}$$

(напомним, что s нечетно).

Задача 1. Показать, что $S \in M_l^m$.

Далее вводим новый вектор неизвестных $\mathbf{z} = \mathbf{y} - \lambda S\mathbf{y}$ (аналог перехода к уравнению (8)) или умножаем обе части системы (10) слева на матрицу $(E - \lambda S)^{-1}$ (аналог перехода к уравнению (9)). В обоих случаях получаем новую систему уравнений вида

$$\mathbf{u} - P\mathbf{u} = \mathbf{g}. \tag{11}$$

Справедливо следующее утверждение.

Если ядро $K(x,\, s)$ непрерывно, то

$$\|P\|_\infty \leqslant \omega(l), \qquad \omega(l) \to 0 \quad \textit{при } l \to \infty.$$

Таким образом, получаем систему уравнений, для решения которой может быть эффективно применен метод простой итерации.

При реальном решении задач часто в явном виде системы уравнений (8), (9), (11) не выписываются; на каждом шаге необходимые вспомогательные величины, например в случае (11) значения векторов $(E - \lambda S)^{-1}\mathbf{h}$ при различных $\mathbf{h}$, вычисляются заново. В результате этого трудоемкость метода оказывается довольно малой.

Рассмотрим, например, дискретный вариант перехода к системе (9). Имеем систему

$$(E-\lambda S)^{-1}\left(E-\lambda K^{(m)}\right)\mathbf{y}=(E-\lambda S)^{-1}\mathbf{f}.$$

Итерационный процесс записывается в виде

$$\mathbf{y}^{n+1}=\mathbf{y}^n-(E-\lambda S)^{-1}\left((E-\lambda K^{(m)})\mathbf{y}^n-\mathbf{f}\right).$$

Вычисление вектора $\mathbf{w}^n=\left(E-\lambda K^{(m)}\right)\mathbf{y}^n-\mathbf{f}$ требует $O(m^2)$ арифметических операций.

Вектор $\mathbf{z}^n=(E-\lambda S)^{-1}\mathbf{w}^n$ находим на каждой итерации, решая систему уравнений

$$\mathbf{z}-\lambda S\mathbf{z}=\mathbf{w}^n$$

с матрицей $S\in M_l^m$.

Вычисление коэффициентов системы (5) требует $O(l^2m)$ операций и при известных A_j вычисление $\mathbf{z}$ требует $O(lm)$ операций. Таким образом, при $l=O(\sqrt{m})$ на каждом шаге производится $O(m^2)$ операций, по порядку столько же, сколько и в методе простой итерации.

§ 3. Интегральные уравнения Фредгольма первого рода

Задача решения интегрального уравнения Фредгольма первого рода

$$Qy=\int_a^b K(x,\,s)y(s)\,ds=f(x) \tag{1}$$

относится к классу *некорректных задач*.

Поясним, что это означает. Пусть ядро $K(x,\,s)$ вещественно и симметрично, т. е. $K(s,\,x)=K(x,\,s)$. Предположим также, что $K(x,\,s)$ и $f(x)$ непрерывны. Тогда существует полная ортонормированная система собственных функций φ^n оператора Q:

$$Q\varphi_n=\int_a^b K(x,\,s)\varphi_n(s)\,ds=\lambda_n\varphi_n(x),$$

$$(\varphi_i,\,\varphi_j)=\int_a^b \varphi_i(s)\varphi_j(s)\,ds=\delta_i^j,$$

где δ_i^j — символ Кронекера. При этом

$$K(x,\,s)=\sum_{n=1}^{\infty}\lambda_n\varphi_n(x)\varphi_n(s),$$

сходимость ряда в правой части понимается в норме:

$$\|F(x,\,s)\| = \sqrt{\int_a^b \int_a^b |F(x,\,s)|^2\,dxds}.$$

Из предыдущего соотношения следует, что

$$\|K\|^2 = \sum_{n=1}^{\infty} |\lambda_n|^2$$

и, следовательно, $\lambda_n \to 0$ при $n \to \infty$.

Рассмотрим случай, когда $\lambda_n \neq 0$ при $1 \leqslant n \leqslant n_0$ и все $\lambda_n = 0$ при $n > n_0$. Тогда ядро имеет вид

$$K(x,\,s) = \sum_{n=1}^{n_0} \lambda_i \varphi_n(x)\varphi_n(s),$$

т. е. является вырожденным. В случае вырожденного ядра

$$\int_a^b K(x,\,s)y(s)\,ds = \sum_{n=1}^{n_0} \lambda_n \int_a^b \varphi_n(x)\varphi_n(s)y(s)\,ds = \sum_{n=1}^{n_0} \lambda_n(\varphi_n,\,y)\varphi_n(x) = f(x).$$

Следовательно, задача (1) может иметь решение только в том случае, когда $f(x)$ является линейной комбинацией $\varphi_1(x), \ldots, \varphi_{n_0}(x)$, т. е. записывается в виде

$$f(x) = \sum_{n=1}^{n_0} f_n \varphi_n(x).$$

Задача 1. Проверить, что этим решением является

$$y(x) = y_0(x) = \sum_{n=1}^{n_0} \frac{f_n}{\lambda_n} \varphi_n(x).$$

Задача 2. Проверить, что любая функция $y(x)$, представимая в виде

$$y(x) = y_0(x) + \sum_{n=n_0+1}^{\infty} c_n \varphi_n(x),$$

где $\sum\limits_{n=n_0+1}^{\infty} |c_n|^2 < \infty$, также будет решением уравнения (1).

Таким образом, в рассматриваемом случае *задача* (1) *может не иметь решения; в случае, когда она имеет решение, это решение неединственно.*

Рассмотрим случай, когда все $\lambda_n \neq 0$. Если $\|f\| = \sqrt{\int_a^b |f(x)|^2\,dx} < \infty$, то $f(x)$ представима сходящимся в норме пространства L_2 рядом Фурье

$$f(x) = \sum_{n=1}^{\infty} c_n \varphi_n(x), \qquad c_n = (f,\, \varphi_n), \qquad \sum_{n=1}^{\infty} c_n^2 = \|f\|^2.$$

Здесь и далее сходимость рядов понимается в смысле нормы пространства L_2. Положим $S = \sum\limits_{n=1}^{\infty} \left|\dfrac{c_n}{\lambda_n}\right|^2$.

Задача 3. Доказать, что при $S < \infty$ функция $y(x) = \sum\limits_{n=1}^{\infty} \dfrac{c_n}{\lambda_n} \varphi_n(x)$ является решением уравнения (1).

Задача 4. Доказать, что при $S = \infty$ задача 1 не имеет решения.

Задача 5. Пусть все $\lambda_n \neq 0$. Показать, что задача (1) не может иметь двух различных решений (решения, отличающиеся на множестве меры нуль, считаются совпадающими).

Таким образом, возможна следующая ситуация. *Задача* (1) *может иметь не более одного решения, однако при этом решение существует лишь для множества правых частей, удовлетворяющих условию* $S < \infty$.

Задача 6. Рассмотреть случай решения уравнения (1) описанным выше методом, когда имеется бесконечное число λ_n, отличных от нуля, и бесконечное число λ_n, равных нулю.

При изучении многих задач, в частности в задачах *интерпретации результатов наблюдений*, или, как говорят, в задачах их *обработки*, часто возникает следующая ситуация. Имеется некоторая функция $y(x)$; мы наблюдаем не ее, а функцию $f(x) = \int_a^b K(x,\, s)y(s)\,ds$, причем в значения этой функции вносятся возмущения $\delta f(x)$.

Таким образом, задача

$$\int_a^b K(x,\, s)y(s)\,ds = f(x)$$

имеет решение, но нам реально требуется решать задачу

$$\int_a^b K(x,\, s)\tilde{y}(s)\,ds = \tilde{f}(x), \tag{2}$$

где $\tilde{f}(x) = f(x) + \delta f(x)$; норма δf погрешности измерения $f(x)$ мала:

$$\|\delta f(x)\| \leqslant \varepsilon. \tag{3}$$

Разность между решениями задач (2) и (1), которую можно записать в виде $\delta y(x) = \tilde{y}(x) - y(x)$, является решением интегрального уравнения

$$\int_a^b K(x,\, s)\delta y(s)\, ds = \delta f(x). \tag{4}$$

Пусть $\delta f(x) = \sum\limits_{n=1}^{\infty} \alpha_n \varphi_n(x)$. Условие (3) означает, что $\sum\limits_{n=1}^{\infty} |\alpha_n|^2 \leqslant \varepsilon^2$.

Рассмотрим сначала случай, когда все $\lambda_n \neq 0$. Если ряд $\sum\limits_{n=1}^{\infty} \left|\dfrac{\alpha_n}{\lambda_n}\right|^2$ расходится, то уравнение (4) не имеет решения. Если даже этот ряд сходится, то нельзя гарантировать, что погрешность $\delta y(x)$ будет стремиться к нулю при $\varepsilon \to 0$.

В самом деле, среди всех правых частей δf с $\|\delta f\| \leqslant \varepsilon$ имеется правая часть $\delta f_n = \varepsilon \varphi_n(x)$, соответствующая такому n, что $|\lambda_n| \leqslant \varepsilon$. Тогда $\delta y = \dfrac{\varepsilon}{\lambda_n}\varphi_n(x)$, т. е. $\|\delta y\| = \varepsilon/|\lambda_n| > 1$.

Для решения рассматриваемой задачи можно применить *метод регуляризации по Тихонову*. Никто не обязывает нас непосредственно решать задачу (2) с возмущенной правой частью. Можно попытаться заменить эту задачу некоторой «близкой» задачей, решение которой будет «близко» к $y(x)$.

Мы уже изучали некоторые способы регуляризации на примере решения систем линейных уравнений.

При решении интегральных уравнений Фредгольма первого рода в качестве такой близкой к (1) задаче рассмотрим уравнение

$$\mu y_\mu(x) + \int_a^b K(x,\, s) y_\mu(s)\, ds = \tilde{f}(x), \quad \mu > 0; \tag{5}$$

параметр μ иногда называют *параметром регуляризации*.

Теорема. *Пусть все* $\lambda_n > 0$,

$$y(x) = \sum_{n=1}^{\infty} y_n \varphi_n(x), \quad \|y\|^2 = \sum_{n=1}^{\infty} |y_n|^2 < \infty.$$

Тогда справедливо неравенство

$$\|y_\mu - y\| \leqslant \omega(\varepsilon,\, \mu),$$

где $\omega(\varepsilon,\, \mu) \to 0$ *при* $\varepsilon/\mu,\ \mu \to 0$ ($\omega(\varepsilon,\, \mu)$, *вообще говоря, зависит от* $y(x)$).

Доказательство. Сравним решение уравнения (5) с точным решением задачи (1). Имеем

$$f(x) = \sum_{n=1}^{\infty} \lambda_n y_n \varphi_n(x) = \sum_{n=1}^{\infty} f_n \varphi_n(x),$$

где $f_n = \lambda_n y_n$, $y_n = (y(x), \varphi_n(x))$ и

$$\tilde{f}(x) = \sum_{n=1}^{\infty} (f_n + \alpha_n)\varphi_n(x).$$

Подставляя $y_\mu(x) = \sum\limits_{n=1}^{\infty} y_n^\mu \varphi_n(x)$ в (5), получим

$$\sum_{n=1}^{\infty} (\mu + \lambda_n) y_n^\mu \varphi_n(x) = \sum_{n=1}^{\infty} (f_n + \alpha_n)\varphi_n(x).$$

Таким образом,

$$y_n^\mu = \frac{f_n + \alpha_n}{\mu + \lambda_n}, \quad y_\mu(x) = \sum_{n=1}^{\infty} \frac{f_n + \alpha_n}{\mu + \lambda_n}\varphi_n(x) = \sum_{n=1}^{\infty} \frac{\lambda_n y_n + \alpha_n}{\mu + \lambda_n}\varphi_n(x).$$

Рассмотрим разность

$$R_\mu(x) = y_\mu(x) - y(x) = \sum_{n=1}^{\infty} \left(\frac{\lambda_n y_n + \alpha_n}{\mu + \lambda_n} - y_n \right) \varphi_n(x).$$

Имеем равенство

$$\frac{\lambda_n y_n + \alpha_n}{\mu + \lambda_n} - y_n = \frac{\alpha_n - \mu y_n}{\mu + \lambda_n}.$$

Таким образом, погрешность $R_\mu(x)$ можно представить в виде суммы двух слагаемых $R_\mu^1(x)$ и $R_\mu^2(x)$:

$$R_\mu(x) = R_\mu^1(x) + R_\mu^2(x), \tag{6}$$

где

$$R_\mu^1(x) = \sum_{n=1}^{\infty} \frac{\alpha_n}{\mu + \lambda_n}\varphi_n(x), \qquad R_\mu^2(x) = \sum_{n=1}^{\infty} \frac{-\mu y_n}{\mu + \lambda_n}\varphi_n(x).$$

Вследствие ортонормированности системы функций $\varphi_n(x)$ имеем

$$\|R_\mu^1\| = \sqrt{\sum_{n=1}^{\infty} \left| \frac{\alpha_n}{\mu + \lambda_n} \right|^2}, \qquad \|R_\mu^2\| = \sqrt{\sum_{n=1}^{\infty} \left| \frac{-\mu y_n}{\mu + \lambda_n} \right|^2}.$$

Поскольку $\mu, \lambda_n \geqslant 0$, то $\mu + \lambda_n \geqslant \mu$ и поэтому

$$\|R_\mu^1\| \leqslant \sqrt{\sum_{n} \left| \frac{\alpha_n}{\mu} \right|^2} = \frac{1}{\mu}\|\delta f\| \leqslant \frac{\varepsilon}{\mu}. \tag{7}$$

Перейдем к оценке $\|R_\mu^2\|$. Рассмотрим сначала более простой и относительно часто встречающийся случай, когда дополнительно выполнено условие

$$\sqrt{\sum_{n=1}^{\infty}\left|\frac{y_n}{\lambda_n}\right|^2} = F_0 < \infty. \tag{8}$$

Тогда

$$\|R_\mu^2\| \leqslant \sqrt{\sum_{i=1}^{\infty}\left|\frac{\mu y_n}{\lambda_n}\right|^2} = \mu F_0. \tag{9}$$

Из соотношений (6), (7), (9) следует, что

$$\|R_\mu\| \leqslant \omega(\varepsilon, \mu) = \frac{\varepsilon}{\mu} + \mu F_0. \tag{10}$$

Следовательно, при дополнительном предположении (8) теорема доказана, поскольку

$$\|R_\mu\| \leqslant \omega(\varepsilon, \mu) = \frac{\varepsilon}{\mu} + \mu F_0$$

и $\omega(\varepsilon, \mu) \to 0$ при ε/μ, $\mu \to 0$.

Таким образом, при достаточно малых μ, ε/μ мы получаем решение задачи с малой погрешностью.

Для получения наилучшей оценки стремления погрешности к нулю найдем $\min\limits_{\mu}\omega(\varepsilon, \mu)$. В точке минимума $\mu = \mu_0$ имеем $\omega'_\mu = -\dfrac{\varepsilon}{\mu_0^2} + F_0 = 0$, т. е. $\mu_0 = \sqrt{\varepsilon/F_0}$. Из (10) получаем, что при $\mu = \sqrt{\varepsilon/F_0}$

$$\|R_\mu\| \leqslant 2\sqrt{\varepsilon F_0} \to 0.$$

Проведем теперь доказательство теоремы без предположения (8).

Представим выражение $\|R_\mu^2\|$ в виде

$$\|R_\mu^2\| = \sqrt{R^1_{\mu,N} + R^2_{\mu,N}},$$

где

$$R^1_{\mu,N} = \sum_{n=1}^{N}\left|\frac{-\mu y_n}{\mu+\lambda_n}\right|^2, \qquad R^2_{\mu,N} = \sum_{n=N+1}^{\infty}\left|\frac{-\mu y_n}{\mu+\lambda_n}\right|^2.$$

Справедливы оценки

$$R^1_{\mu,N} \leqslant \sum_{n=1}^{N}\left|\frac{\mu y_n}{\lambda_n}\right|^2 = \mu^2 F_1^2(N),$$

где $F_1^2(N) = \sum_{n=1}^{N} \left|\frac{y_n}{\lambda_n}\right|^2$, и

$$R_{\mu,N}^2 \leqslant \sum_{n=N+1}^{\infty} |y_n|^2.$$

Покажем, что $\|R_\mu^2\| \to 0$ при $\mu \to 0$. Для этого достаточно показать, что для любого $\delta > 0$ существует $\mu(\delta)$ такое, что

$$\|R_\mu^2\| \leqslant \delta \quad \text{при} \quad \mu \leqslant \mu(\delta).$$

Возьмем произвольное $\delta > 0$. Поскольку ряд $\sum_{n=1}^{\infty} |y_n|^2$ сходится, то существует $N(\delta)$ такое, что

$$\sum_{n=N(\delta)+1}^{\infty} |y_n|^2 \leqslant \frac{\delta^2}{2}.$$

Если $\mu^2 \leqslant \left(\mu(\delta)\right)^2 = \delta^2 \Big/ \left(2(F_1(N(\delta)))^2\right)$, то $R_{\mu,N}^1 \leqslant \delta^2/2$ и

$$\|R_\mu^2\| = \sqrt{R_{\mu,N}^1 + R_{\mu,N}^2} \leqslant \sqrt{\delta^2/2 + \delta^2/2} = \delta.$$

Таким образом, имеем

$$\|R_\mu\| \leqslant \|R_\mu^1\| + \|R_\mu^2\|, \quad \|R_\mu^1\| \leqslant \varepsilon/\mu, \quad \|R_\mu^2\| \to 0 \quad \text{при} \quad \mu \to 0.$$

Таким образом, утверждение теоремы справедливо и без предположения (8).

Описанный выше метод регуляризации применим и в случае, когда некоторые из λ_n могут обращаться в нуль.

Пусть N_1 — множество n таких, что $\lambda_n > 0$, N_0 — множество n таких, что $\lambda_n = 0$. (Каждое из этих множеств может быть как конечным, так и бесконечным; одновременно оба конечными быть не могут.)

Если $y(x) = \sum_n y_n \varphi_n(x)$, $\sum_n |y_n|^2 < \infty$ — решение уравнения (1), то при $\sum_{n \in N_0} |a_n|^2 < \infty$ функция $\sum_n (y_n + a_n)\varphi_n(x)$ также будет решением уравнения (1).

Положим $y^0(x) = \sum_{n \in N_1} y_n \varphi_n(x)$; согласно вышесказанному $y^0(x)$ также является решением уравнения (1).

Задача 7. Пусть $y(x) \neq y^0(x)$ — решение уравнения (1). Показать, что

$$\|y\| > \|y^0\|. \tag{11}$$

Решение уравнения (1) с минимальной нормой (в случае неединственного решения) называется *нормальным*. Из (11) следует, что $y_0(x)$ — *нормальное решение задачи*.

Теорема. *Пусть* $y^0(x) = \sum\limits_{n\in N_1} y_n\varphi_n(x)$, $\|y^0\|^2 = \sum\limits_{n\in N_1} |y_n|^2 < \infty$ — *решение уравнения* (1). *Тогда справедливо неравенство* $\|y_\mu - y^0\| \leqslant \omega(\varepsilon, \mu)$, *где* $\omega(\varepsilon, \mu) \to 0$ *при* ε/μ, $\mu \to 0$.

Доказательство несущественным образом отличается от проведенного выше в случае, когда все $\lambda_m > 0$. Решение $y^0(x)$ задачи (1) записывается в виде

$$y^0(x) = \sum_{n\in N_1} y_n^0\varphi_n(x) = \sum_{n=1}^{\infty} Y_n\varphi_n(x),$$

где

$$Y_n = \begin{cases} y_n^0 & \text{при} \quad n \in N_1, \\ 0 & \text{при} \quad n \in N_0. \end{cases}$$

Погрешность $y_\mu(x) - y^0(x)$ запишется в виде

$$y_\mu(x) - y^0(x) = \sum_{n=1}^{\infty} \left(\frac{Y_n + \alpha_n}{\mu + \lambda_n} - Y_n \right) \varphi_n = S_1 + S_2,$$

где

$$S_1 = \sum_{n\in N_1} \left(\frac{y_n + \alpha_n}{\mu + \lambda_n} - y_n \right) \varphi_n, \quad S_2 = \sum_{n\in N_0} \frac{\alpha_n}{\mu}\varphi_n.$$

Оценка для слагаемого S_2 имеет вид

$$\|S_2\| = \left\| \sum_{n\in N_0} \frac{\alpha_n}{\mu}\varphi_n \right\| = \sqrt{\sum_{n\in N_0} \left| \frac{\alpha_n}{\mu} \right|^2} \leqslant \frac{1}{\mu} \sqrt{\sum_{n=1}^{\infty} |\alpha_n|^2} = \frac{\varepsilon}{\mu}.$$

Оценка для S_1 производится так же, как и в случае доказанной ранее теоремы.

Метод регуляризации применяется для решения самых разнообразных задач, в частности нелинейных.

Рассмотрим случай, когда ядро $K(x, s)$ несимметрично. Определим оператор Q^* соотношением

$$Q^*y = \int_a^b K(s, x)y(s)\,ds.$$

Обозначим через $y_\mu(x)$ решение уравнения

$$\mu y_\mu(x) + Q^*Qy_\mu = Q^*\tilde{f}(x).$$

Выберем параметр $\mu = \mu(\varepsilon)$ из условия

$$\|Qy_{\mu(\varepsilon)} - \tilde{f}\| = \varepsilon.$$

Справедлива

Теорема (без доказательства). *Пусть уравнение* (1) *разрешимо и* $y^0(x)$ — *нормальное решение уравнения* (1), *т. е. решение с минимальной нормой. Тогда* $\|y_{\mu(\varepsilon)} - y^0\| \to 0$ *при* $\varepsilon \to 0$.

Литература

1. Березин И. С. Жидков Н. П. Методы вычислений. Т. 2. — М.: Физматгиз, 1962.
2. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интегральные уравнения, некорректные задачи и улучшение сходимости. — Минск: Наука и техника, 1984.
3. Морозов В. А. Регулярные методы решения некорректно поставленных задач. — М.: Наука, 1987.
4. Романов В. Г. Обратные задачи математической физики. — М.: Наука, 1984.
5. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1986.
6. Денисов А. М. Введение в теорию обратных задач — М.: Изд-во МГУ, 1994.

Заключение

Мы закончили обсуждение традиционных вопросов теории численных методов. При этом мы, возможно слишком часто, обращали внимание на «подводные камни», встречающиеся при применении того или иного численного метода. У читателя могло возникнуть превратное впечатление, что применение численных методов для решения реальных задач настолько сложная и безнадежная задача, что от него следует отказаться. Чтобы исправить такое впечатление, посмотрим на этот же вопрос с оптимистических позиций.

Существует большое число задач, где есть хорошо отработанные численные методы и созданные на их основе стандартные программы решения задач.

Стандартные программы решения многих типов прикладных задач входят в математическое обеспечение, поставляемое вместе с ЭВМ.

Если вам впервые встретилась единичная задача, то, как правило, целесообразнее воспользоваться стандартной программой или самому составить программу, основанную на простейшем методе решения. При решении единичных задач, требующих умеренного объема вычислений, часто идут на более чем 100-кратное увеличение объема вычислений по сравнению с наиболее эффективными методами, лишь бы побыстрее получить результат, воспользовавшись при этом стандартной программой или алгоритмами, для реализации которых можно быстро составить и отладить программу.

На практике регулярно встречаются задачи минимизации функций большого числа переменных, в которых применение стандартных программ не приводит к положительному результату. Тем не менее, опыт авторов показывает, что при решении новой задачи надо все равно начинать с попытки использовать стандартную программу. Если, например, в 30% случаев применение стандартных программ оказывается эффективным, то их применение заведомо можно считать оправданным даже в том случае, когда использование стандартной программы не приводит к положительному результату; при обращении к стандартной программе будет составлено большое число блоков окончательной программы и зачастую будет накоплена полезная информация о свойствах минимизируемой функции.

Среди «пионеров» использования вычислительной техники встречается убеждение, что чужими стандартными программами пользоваться нельзя и программу всегда следует писать самому, заново. Такое суждение было

оправдано на первоначальном этапе использования ЭВМ, когда теория численных методов была развита недостаточно и созданные на ее основе алгоритмы часто были ненадежными. При современном уровне теории численных методов и жестких требованиях к *тестированию* программ с такой точкой зрения согласиться нельзя.

Рассмотрим ситуацию, когда требуется путем численного эксперимента исследовать какой-либо физический процесс. Часто, набив много шишек и потратив иногда годы на расчет сложных моделей, начинающий исследователь приходит к пониманию того, что целесообразнее было начать с расчета простейшей модели и изучать ее с помощью простейших проверенных методов, иногда требующих повышенных вычислительных затрат. Лишь в случае полного доверия к постановке задачи имеет смысл заниматься расчетом сложной модели с применением громоздких по своей структуре методов.

Таким образом, на первоначальном этапе исследования задачи обычно мы имеем дело с простейшей моделью и простейшими методами решения, часто основанными на использовании стандартных программ. Если же мы начинаем исследование задачи с изучения сложных моделей, то скорее всего мы поступаем в чем-то неправильно. Целесообразнее переходить к рассмотрению более сложных задач и применению более сложных методов, имея за плечами опыт использования ЭВМ при решении простейших задач простейшими методами, поскольку при такой последовательности действий использование более сложных численных методов уже не будет казаться чрезмерно трудной проблемой.

Рассмотрение самими математиками постановки прикладной задачи с ее истоков, с простейших моделей, еще важно и в связи со следующим обстоятельством. Часто, стремясь «приспособить» задачу для численного решения, специалист, не знакомый с численными методами и возможностями ЭВМ, исходит лишь из сложности внешнего вида математической постановки. В результате этого иногда

а) происходит замена исходной задачи задачей, не имеющей к ней отношения,

б) задача, передающаяся численному решению при возможностях современных ЭВМ, становится задачей, не поддающейся такому решению.

Еще один довод в пользу более всестороннего изучения постановки задачи состоит в следующем. Часто в процессе построения простейшей модели мы получаем представление о том, в каком направлении будет идти усложнение метода и программы решения задачи. В результате этого мы сможем заложить в исходный вариант программы, предназначенной для исследования простейшей модели, возможности для дальнейшего усложнения программы.

Во всех случаях целесообразно составлять программу поблочно. Дело в том, что при решении сколько-нибудь сложных задач мы заранее часто не можем сказать, будет ли выбранный нами метод решения подходящим для решения этой задачи.

Например, при решении краевой задачи для дифференциальных уравнений у нас может быть неуверенность в разумности выбора разностной схемы во внутренних точках или вблизи границы. Вычисления во внутренних точках и вблизи границы ведутся по различным формулам и поэтому лучше, чтобы они были выделены в отдельные блоки с тем, чтобы можно было производить независимую *отладку методов*. Также важно предусмотреть возможность быстрого и удобного изменения параметров задачи, например шага сетки при решении дифференциальных уравнений или числа узлов квадратуры. Отладка программы и апробация метода решения задачи при малом числе узлов проходят быстрее и позволяют лучше использовать возможности ЭВМ. Например, при малом числе узлов можно быстро проверить сходимость итерационного процесса.

Иногда профессиональные программисты рекомендуют проводить отладку программы при помощи *тестов*. Берется минимально допустимое число узлов, при котором включаются в работу все блоки программы (и все циклы). При таком числе узлов задача решается без помощи написанной программы, а затем результаты расчета сравниваются с результатами расчета с помощью ЭВМ. Для типичных конечно-разностных методов решения дифференциальных уравнений минимальное число узлов по каждой оси, при котором включаются в работу все блоки и циклы, находится в пределах от 2 до 5.

В процессе работы над отладкой программы в программу вносятся дополнительные команды выдачи на дисплей или на печать промежуточных данных с целью сравнения их с заранее просчитанным тестом. Такой режим отладки особенно удобен для начинающего программиста.

Специалисты, имеющие большой опыт работы с ЭВМ, предпочитают отлаживать программу крупными модулями, сочетая отладку программы с проверкой качества метода (отладкой метода). Для этого стараются строить программы решения так, чтобы при определенных значениях параметров она превращалась в программу решения задачи с известным решением.

Например, уравнения движения снежной лавины при определенных значениях параметров превращаются в уравнения мелкой воды, для которых в случае кусочно-постоянных начальных данных известно точное решение (так называемое решение задачи о распаде разрыва). Сравнивая решение, полученное в результате расчетов, с точным решением, можно судить о правильности программы и качестве метода.

Построение решения задачи о распаде разрыва само по себе требует относительно большой вспомогательной работы. Вместо непосредственного построения такого решения для заданной системы часто поступают следующим образом.

Вместо конкретных значений коэффициентов системы ставятся произвольные параметры, производится одновременная подборка этих параметров и системы функций, которая может быть решением задачи о

распаде разрыва. В результате получится некоторая задача с известным решением, отличающаяся от интересующей нас задачи лишь значениями числовых параметров. На такой задаче мы можем отладить все наиболее существенные моменты метода и программы.

Проиллюстрируем способы построения задач с известным частным решением на примере дифференциальных уравнений.

Пусть решается краевая задача

$$L(u) = \frac{\partial}{\partial x}\left((1+\gamma u^2)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left((1+\gamma)\frac{\partial u}{\partial y}\right) - e^{x+y} = 0$$

в квадрате $G: 0 < x, y < 1$; на границе квадрата Γ ставится краевое условие

$$\left(\frac{\partial u}{\partial n} + \sigma u - g\right)\Big|_\Gamma = 0.$$

Возьмем некоторую функцию $u^0(x, y)$ и вычислим функции

$$f^0(x, y) = L(u^0(x, y)), \quad g^0|_\Gamma = \left(\frac{\partial u^0}{\partial n} + \sigma u^0 - g\right)\Big|_\Gamma.$$

Краевая задача

$$L(u) - f^0(x, y) = 0, \quad \left(\frac{\partial u}{\partial n} + \sigma u - g - g^0\right)\Big|_\Gamma = 0$$

будет задачей с известным точным решением $u = u^0(x, y)$.

Часто желательно иметь в распоряжении сеточную задачу с известным точным решением. Если, например, по простейшим явным формулам решается задача Коши, то такое решение легко вычислить непосредственно.

В случае краевых задач или при решении задачи Коши по неявным формулам часто целесообразно поступить как и выше. Строим сеточную задачу $L_h(u_h) = 0$, $l_h(u_h) = 0$ (с небольшим числом узлов), берем некоторую сеточную функцию u_h^0 и вычисляем

$$f_h^0 = L_h(u_h^0), \quad g_h^0 = l_h(u_h).$$

Сеточной задачей с известным точным решением $u_h = u_h^0$ является задача

$$L_h(u_h) = f_h^0, \quad l_h(u_h) = g_h^0.$$

Иногда возникает необходимость иметь в распоряжении сеточную задачу с большим числом узлов и с точным решением. В этом случае может оказаться целесообразным поручить ЭВМ вычисление f_h^0 и g_h^0. Часто такую задачу получают, взяв в качестве u_h^0 множество значений в узлах сетки некоторого многочлена невысокой (обычно 1 – 4-й) степени. Степень многочлена выбирается из условия, чтобы погрешность аппроксимации дифференциальной задачи сеточной равнялась нулю.

Еще раз выделим основную идею описанного выше приема построения тестов. *Вместо непосредственного построения теста для исходной задачи* (построение решения по заданной правой части) *строится тест для обратной задачи* (по заданному решению строится правая часть) и он оказывается тестом для прямой задачи с той же структурой, но с другими числовыми данными.

Применение этого приема обычно существенно сокращает затраты на построение теста.

Список литературы

1. Абрамов А. А. О численном решении некоторых алгебраических задач, возникающих в теории устойчивости // ЖВМ и МФ. — 1984. — **24**, N 3. — С. 339-347.
2. Бахвалов Н. С. Об оптимальных оценках скорости сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций // Численные методы решения дифференциальных и интегральных уравнений и квадратурные формулы. — М.: Наука, 1964. — С. 5–63.
3. Бахвалов Н. С. К оптимизации методов решения краевых задач при наличии пограничного слоя // ЖВМиМФ. — 1969. — Т. 9, N 4. — С. 841–859.
4. Бахвалов Н. С. Численные методы. — М.: Наука, 1975.
5. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. — М.: Наука, 1987.
6. Бахвалов Н. С., Кобельков Г. М., Поспелов В. В. Сборник задач по методам вычислений. — М.: Изд-во МГУ, 1989.
7. Бейкер Дж., Грейвс—Моррис П. Аппроксимации Паде. — М.: Мир, 1986.
8. Березин И. С. Жидков Н. П. Методы вычислений. Т. 1. — М.: Наука, 1966.
9. Березин И. С. Жидков Н. П. Методы вычислений. Т. 2. — М.: Физматгиз, 1962.
10. Васильев Ф. П. Численные методы решения экстремальных задач. — М.: Наука, 1980.
11. Васильев Ф. П. Методы решения экстремальных задач. — М.: Наука, 1981.
12. Винокуров В. А., Ювченко Н. В. Полуявные численные методы решения жестких задач // ДАН. — 1985. — **284**, N 2, С. 272–277.
13. Воеводин В. В. Численные методы алгебры. Теория и алгоритмы. — М.: Наука, 1966.
14. Воеводин В. В. Вычислительные основы линейной алгебры. — М.: Наука, 1977.

15. Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления. — М.: Наука, 1984.

16. Воеводин В. В., Арушанян О. Б. Структура и организация бмблиотеки численного анализа НИВЦ МГУ // Численный анализ на ФОРТРАНе. Вычислительные методы и инструментальные системы. — М.: Изд-во МГУ, 1979.

17. Волков Е. А. Численные методы. — М.: Наука, 1982.

18. Годунов С. К., Забродин А. В. О разностных схемах второго порядка точности для многомерных задач // ЖВМ и МФ. — 1962. — **2**, N 4. — С. 706–708.

19. Годунов С. К., Рябенький В. С. Введение в теорию разностных схем. — М.: Наука, 1962.

20. Годунов С. К., Рябенький В. С. Разностные схемы. — М.: Наука, 1977.

21. Годунов С. К., Забродин А. В. и др. Численное решение многомерных задач газовой динамики. — М.: Наука, 1976.

22. Годунов С. К. Решение систем линейных уравнений. — Новосибирск: Наука, 1980.

23. Джордж А., Лю Д. Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.

24. Дробышевич В. И., Дымников В. П., Ривин Г. С. Задачи по вычислительной математике. — М.: Наука, 1980.

25. Дьяконов Е. Г. Минимизация вычислительной работы. — М.: Наука, 1989.

26. Дьяконов Е. Г. О некоторых модификациях проекционно-разностных методов // Вестник МГУ. Сер. вычислит. мат. и киберн. — 1977. — **1**, N 2. — С. 13–19.

27. Завьялов Ю. С., Квасов Б. И., Мирошниченко В.Л. Методы сплайн-функций. — М.: Наука, 1980.

28. Зенкевич О., Морган К. Конечные элементы и аппроксимация. — М.: Мир, 1980.

29. Икрамов Х. Д. Численное решение матричных уравнений. — М.: Наука, 1984.

30. Карманов В. Г. Математическое программирование. — М.: Наука, 1986.

31. Кобельков Г. М. Решение задачи о стационарной свободной конвекции // ДАН СССР. — 1980. — **225**, N 2. С. 277–282.

32. Кобельков Г. М. О методах решения уравнений Навье-Стокса // Вычислительные процессы и системы. — М.: Наука, 1991, вып. 8. С. 204–236.

33. Колмогоров А. Н., Тихомиров В. М. ε-энтропия и ε-емкость множеств в функциональных пространствах // УМН. — 1959. — **14**, вып. 2. С. 3–86.

34. Копченова Н. В., Марон И. А. Вычислительная математика в примерах и задачах. — М.: Наука, 1972.

35. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Дифференциальные уравнения. — Минск: Наука и техника, 1982.

36. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Уравнения в частных производных. — Минск: Наука и техника, 1982.

37. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника, 1982.

38. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интегральные уравнения, некорректные задачи и улучшение сходимости. — Минск: Наука и техника, 1984.

39. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интерполирование и интегрирование. — Минск: Наука и техника, 1983.

40. Крылов В. И., Шульгина А. Т. Справочная книга по численному интегрированию. — М.: Наука, 1966.

41. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т.1. — М.: Наука, 1976.

42. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т.2. — М.: Наука, 1977.

43. Ланцош К. Практические методы прикладного анализа. — М.: ГИФМЛ, 1961.

44. Лебедев В. И. Функциональный анализ и вычислительная математика. — М.: Изд-во ВИНИТИ, 1987. — (ОВМ АН СССР).

45. Лебедев В. И. Как решать явными методами жесткие системы дифференциальных уравнений // Вычислительные процессы и системы. — М.: Наука, 1991, вып. 8. С. 237–291.

46. Лоусон Ч., Хентон Р. Численное решение задач метода наименьших квадратов. — М.: Наука, 1986.

47. Мак-Кракен Д., Дорн У. Численные методы и программирование на ФОРТРАНе. — М.: Мир, 1977.

48. Марчук Г. И. Методы вычислительной математики. — М.: Наука, 1980.

49. Марчук Г. И., Агошков В. И. Введение в проекционно-разностные методы. — М.: Наука, 1981.

50. Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1981.

51. Марчук Г. И., Шайдуров В. В. Повышение точности решений разностных схем. — М.: Наука, 1979.

52. Марчук Г. И., Яненко Н. Н. Применение метода расщепления (дробных шагов) для решения задач математической физики. — В кн.: Некоторые вопросы вычислительной и прикладной математики.— Новосибирск: Наука, 1966.

53. Мысовских И. П. Интерполяционные кубатурные формулы. — М.: Наука, 1981.

54. Никифоров А. Ф., Суслов С. К., Уваров В. Б. Классические ортогональные полиномы дискретной переменной. — М.: Наука, 1985.

55. Никифоров А. Ф., Уваров В. Б. Специальные функции. — М.: Наука, 1979.

56. Никольский С. М. Квадратурные формулы. — М.: Наука, 1979.

57. Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983.

58. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем. — М.: Наука, 1979.

59. Романов В.Г. Обратные задачи математической физики. — М.: Наука, 1984.

60. Рябенький В. С., Филиппов А. Ф. Об устойчивости разностных уравнений. — М.: Гостехиздат, 1956.

61. Самарский А. А. Теория разностных схем. — М.: Наука, 1982.

62. Самарский А. А., Андреев В. Б. Разностные методы для решения эллиптических уравнений. — М.: Наука, 1976.

63. Самарский А. А., Гулин А. В. Устойчивость разностных схем. — М.: Наука, 1973.

64. Самарский А. А., Капорин И. Е., Кучеров А. Б., Николаев Е. С. Некоторые современные методы решения сеточных уравнений // Изв. вузов. Сер. мат., 1983, N 7(254). с. 3–12.

65. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.

66. Самарский А. А., Попов Ю. П. Разностные методы решения задач газовой динамики. — М.: Наука, 1980.

67. Саульев В. К. Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз, 1960.

68. Соболев С.Л. Введение в теорию кубатурных формул. — М.: Наука, 1974.

69. Современные численные методы решения обыкновенных дифференциальных уравнений // Под ред. Дж. Холла, Дж. Уатта. — М.: Мир, 1979.

70. Стечкин С. Б., Субботин Ю. Н. Сплайны в вычислительной математике. — М.: Наука, 1976.

71. Стрэнг Г., Фикс Дж. Теория метода конечных элементов. — М.: Мир, 1977.

72. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1986.

73. Уилкинсон Дж., Райни К. Справочник алгоритмов на языке Алгол. Линейная алгебра. — М.: Машиностроение, 1976.

74. Фаддеев Л. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М.: Физматгиз, 1963.

75. Форсайт Дж. и др. Машинные методы математических вычислений. — М.: Мир, 1980.

76. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука, 1967.

77. Axelsson O. Numerical linear algebra. — Cambrige, 1996.

78. Dahlquist Y. Stability and error bounds in the numerical integration of ordinary differential equations. — Uppsala, Almqyist & Wiksells boktr 130 (1959). P. 5–92.

79. Butcher I. G. A modified multistep method for the numerical integration of ordinary differential equations // J. Assoc. Comput. Math. — 1965. — **12**, No. 1. — P. 124–135.

80. Stroud A. H. and Secrest D. Gaussian Quadrature Formulas. — Englewood Cliffs, N. Y.: Prentice-Hall, 1966.

Предметный указатель

Алгоритм ненасыщаемый, 65
Аппроксимация, 379
  дифференциального уравнения разностной схемой, 379

Большое число, 23

Валле-Пуссена теорема, 179
Вариация, 522
Верная цифра, 23
Весовая функция, 89
Восполнение, 559
Выделение весовой функции, 151

Гарантированные оценки погрешности на классе функций, 232
Главный член погрешности, 140

Дивергентность разностной схемы, 527
Дискретные коэффициенты Фурье, 173
Дифференцирование численное, 76

Замыкание алгоритма, 437
  нерегулярное, 437
  регулярное, 437
Значащие цифры, 23

Интегрирование
  осциллирующих функций, 116
  системы уравнений, 396
Интерполирование, 36
  с кратными узлами, 49
Интерполяционная формула
  Лагранжа, 41
  Ньютона, 46
  для интерполирования назад, 69
  для интерполирования вперед, 69
Интерполяция, 36
  квадратичная, 72
  линейная, 71
  тригонометрическая, 173
Итерационные методы построения многочлена наилучшего равномерного приближения с использованием спектрально эквивалентных операторов, 300

Квадратуры
  Гаусса, 106
  Грегори, 143
  Лобатто, 111
  Ньютона—Котеса, 94
  Ромберга, 149
  Симпсона, 88
  Филона, 117
  Эйлера, 142
  обобщенные, 122
  прямоугольников, 86
  составные, 122
  трапеций, 87
Количество арифметических операций, 41
Конечно-разностное уравнение, 51
Конечно-разностные
  методы, 376
  схемы, 376

Ленточная структура, 257
Линейная оценка погрешности, 28
Линейное разностное уравнение, 52

Мажорирующее разностное уравнение, 488
Матрица
  Грама, 167
  ортогонализации, 100
  отражений, 262
Мера
  обусловленности

матрицы, 305
системы, 305
погрешности аппроксимации, 504
Метод
Адамса, 378
Гаусса, 253
Зейделя, 290
Монте-Карло, 232
Ньютона решения нелинейных уравнений, 330
Ритца, 477
Рунге–Кутта, 363
Эйлера, 363
вариационно-разностный, 479
верхней релаксации, 290
вилки, 336
квадратного корня, 259
конечных элементов, 558
наименьших квадратов, 203
наискорейшего спуска, 290
неопределенных коэффициентов, 39, 202
оврагов, 342
оптимальный, 63
парабол, 336
покоординатного спуска, 288
пристрелки, 429
прогонки, 430
проекционно-разностный, 561
простой итерации, 265, 326
регуляризации, 205
релаксации, 289
сверхрелаксации, 290
секущих, 335
сопряженных градиентов, 294
спуска, 336
стрельбы, 449
суммарной аппроксимации, 579
циклической прогонки, 434
штрафа, 340
Методы
интерполяционные, 376
экстраполяционные, 376
Многочлен наилучшего равномерного приближения, 178
Многочлены
Лагерра, 104
Лежандра, 104
Чебышева, 58
второго рода, 104
первого рода, 104
Эрмита, 104
Якоби, 103
наименее уклоняющиеся от нуля, 60

Наилучшее равномерное приближение, 178
Недетерминированный метод, 242
Некорректные задачи, 608
Неравенство
Бесселя, 169
Чебышева, 233
ε-неравенство, 537
Неустойчивость, 35
Неявная схема, метод, 376, 530
Норма энергетическая, 555
Нормы
векторов и матриц, 250
эквивалентные, 266

Область
зависимости, 498
сходимости метода, 358
Обобщенное решение, 562
Обратная интерполяция, 76
Обратный ход метода
Гаусса, 254
прогонки, 431
Обусловленность
матрицы системы, 305
системы, 305
Однородные схемы, 469
Односторонние формулы численного дифференцирования, 80
Одношаговые методы численного интегрирования, 371
Оператор расщепляющийся, 573
Оптимальные квадратуры, 129
Оптимальный
линейный итерационный процесс, 279, 283
по порядку итерационный процесс, 42, 63
Оптимизация
методов, 63

оценки погрешности интерполирования, 63
распределения узлов интегрирования, 131
скорости сходимости итерационного процесса, 275
Ортогональная система, 99
Ортогональные многочлены, 101
Ортонормированная система элементов, 101
Остаточный член формулы Лагранжа, 43
Очень большое число, 23

Параметр регуляризации, 611
Переобуславливатель, 301
Планирование эксперимента, 37
Плохо обусловленные системы, 307
Повышение порядка точности разностной схемы, 419
Погрешность
абсолютная, 22
аппроксимации дифференциального уравнения разностной схемой, 504, 534
вычислительная, 17
квадратуры на классе функций, 129
математической модели, 17
метода, 17
на классе задач, 63
на шаге, 371
неустранимая, 17
относительная, 22
предельная, 27
Порядок
метода, 329
погрешности аппроксимации, 380
Почти собственные значения, 270
Преобразование Фурье
быстрое, 175
дискретное, 173
Принцип замороженных коэффициентов, 521
Проблема собственных значений
полная, 315
частичная, 315
δ^2-процесс ускорения сходимости, 273
Прямое произведение формул интегрирования, интерполирования, дифференцирования, 219
Прямой ход метода
Гаусса, 254
прогонки, 431

Разности
вперед, 65
высшего порядка, 65
конечные, 65
назад, 65
разделенные, 43
центральные, 65
Разностная схема, 377
экономичная, 568
Ряд Фурье дискретный, конечный, 173

Сеточная функция Грина, 423
Сжатое отображение, 326
Симметризация системы уравнений, 276
Система жесткая, 398
Слой, 530
Согласованные нормы, 501
Спектральная эквивалентность, 300
Спектральный признак устойчивости, 506, 515
Сплайн, 191
интерполяционный, 197
локальный, 198
Стандартные программы, 47
Строго нормированное пространство, 165
Схема Эйткена, 47
Схемы, точные на решениях специального вида, 461
Сходимость, 501

Таблица разделенных разностей, 45
Теорема
Валле-Пуссена, 179
Чебышева, 179
Точки чебышевского альтернанса, 179
Триангуляции, 558
Тригонометрическая интерполяция, 173

Узел
внутренний, 544

граничный, 544
нерегулярный, 551
приграничный, 544
регулярный, 551
сетки, 543
Узлы интерполяции, 36
Уравнения в конечных разностях, 51
Условие
α, 386
сильной минимальности, 565
Устойчивость, 503
безусловная, 532
по начальным данным, 535
условная, 532

Формула
Абеля, 467
Ромберга, 149
Симпсона, 88
Филона, 117
прямоугольников, 86
с кратными узлами, 96
трапеций, 87
Формулы
Адамса, 363, 378
Грегори, 143
численного дифференцирования, 76
Функция сильно растущая, 23

Характеристическое уравнение, 55
разностной схемы, 384
Хорошо
обусловленная (поставленная) краевая задача, 448
обусловленные системы, 307

Шаблон, 507
Шаг таблицы, 65

Экстраполяция, 36
Элемент наилучшего приближения, 165
Элементарный треугольник, 558
Энергетическое
неравенство, 538
тождество, 537

Явная схема, 376, 529
Явный метод, 376, 529

Оглавление

Предисловие 7

Введение 8

1 Погрешность результата численного решения задачи 17
§ 1. Источники и классификация погрешности ... 17
§ 2. Запись чисел в ЭВМ ... 21
§ 3. Абсолютная и относительная погрешности. Формы записи данных ... 22
§ 4. О вычислительной погрешности ... 25
§ 5. Погрешность функции ... 27
§ 6. Обратная задача ... 32

2 Интерполяция и численное дифференцирование 35
§ 1. Постановка задачи приближения функций ... 36
§ 2. Интерполяционный многочлен Лагранжа ... 39
§ 3. Оценка остаточного члена интерполяционного многочлена Лагранжа ... 43
§ 4. Разделенные разности и их свойства ... 43
§ 5. Интерполяционная формула Ньютона с разделенными разностями ... 45
§ 6. Разделенные разности и интерполирование с кратными узлами ... 48
§ 7. Уравнения в конечных разностях ... 51
§ 8. Многочлены Чебышева ... 58
§ 9. Минимизация оценки остаточного члена интерполяционной формулы ... 62
§ 10. Конечные разности ... 65
§ 11. Интерполяционные формулы для таблиц с постоянным шагом ... 68
§ 12. Составление таблиц ... 71
§ 13. О погрешности округления при интерполяции ... 74
§ 14. Применения аппарата интерполирования. Обратная интерполяция ... 75
§ 15. Численное дифференцирование ... 76
§ 16. О вычислительной погрешности формул численного дифференцирования ... 83
§ 17. Рациональная интерполяция ... 84

3 Численное интегрирование 86
§ 1. Простейшие квадратурные формулы. Метод неопределенных коэффициентов ... 86
§ 2. Оценки погрешности квадратуры ... 89
§ 3. Квадратурные формулы Ньютона—Котеса ... 94
§ 4. Ортогональные многочлены ... 99
§ 5. Квадратурные формулы Гаусса ... 106
§ 6. Практическая оценка погрешности элементарных квадратурных формул ... 113
§ 7. Интегрирование быстро осциллирующих функций ... 116
§ 8. Повышение точности интегрирования за счет разбиения отрезка на равные части ... 119
§ 9. О постановках задач оптимизации ... 124
§ 10. Постановка задачи оптимизации квадратур ... 129
§ 11. Оптимизация распределения узлов квадратурной формулы 130
§ 12. Примеры оптимизации распределения узлов ... 137
§ 13. Главный член погрешности ... 140
§ 14. Правило Рунге практической оценки погрешности ... 144
§ 15. Уточнение результата интерполяцией более высокого порядка точности ... 148
§ 16. Вычисление интегралов в нерегулярном случае ... 150
§ 17. Принципы построения стандартных программ с автоматическим выбором шага ... 157

4 Приближение функций и смежные вопросы 164
§ 1. Наилучшие приближения в линейном нормированном пространстве ... 164
§ 2. Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении ... 166
§ 3. Тригонометрическая интерполяция. Дискретное преобразование Фурье ... 171
§ 4. Быстрое преобразование Фурье ... 175
§ 5. Наилучшее равномерное приближение ... 178
§ 6. Примеры наилучшего равномерного приближения ... 181
§ 7. О форме записи многочлена ... 187
§ 8. Интерполяция и приближение сплайнами ... 191

5 Многомерные задачи 201
§ 1. Метод неопределенных коэффициентов ... 202
§ 2. Метод наименьших квадратов и регуляризация ... 203
§ 3. Примеры регуляризации ... 206
§ 4. Сведение многомерных задач к одномерным ... 212
§ 5. Интерполяция функций в треугольнике ... 220
§ 6. Оценка погрешности численного интегрирования на равномерной сетке ... 222
§ 7. Оценка снизу погрешности численного интегрирования ... 225
§ 8. Метод Монте-Карло ... 232

§ 9. Обсуждение правомерности использования недетерминированных методов решения задач ... 236
§ 10. Ускорение сходимости метода Монте-Карло ... 239
§ 11. О выборе метода решения задачи ... 243

6 Численные методы алгебры 250
§ 1. Методы последовательного исключения неизвестных ... 253
§ 2. Метод отражений ... 262
§ 3. Метод простой итерации ... 265
§ 4. Особенности реализации метода простой итерации на ЭВМ 268
§ 5. δ^2-процесс практической оценки погрешности и ускорения сходимости ... 271
§ 6. Оптимизация скорости сходимости итерационных процессов 275
§ 7. Метод Зейделя ... 285
§ 8. Метод наискорейшего градиентного спуска ... 290
§ 9. Метод сопряженных градиентов ... 294
§ 10. Итерационные методы с использованием спектрально-эквивалентных операторов ... 300
§ 11. Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация ... 304
§ 12. Проблема собственных значений ... 315
§ 13. Решение полной проблемы собственных значений при помощи QR-алгоритма ... 320

7 Решение систем нелинейных уравнений и задач оптимизации 324
§ 1. Метод простой итерации и смежные вопросы ... 326
§ 2. Метод Ньютона решения нелинейных уравнений ... 330
§ 3. Методы спуска ... 336
§ 4. Другие методы сведения многомерных задач к задачам меньшей размерности ... 341
§ 5. Решение стационарных задач путем установления ... 345
§ 6. Как оптимизировать? ... 352

8 Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений 360
§ 1. Решение задачи Коши с помощью формулы Тейлора ... 361
§ 2. Методы Рунге—Кутта ... 363
§ 3. Методы с контролем погрешности на шаге ... 369
§ 4. Оценки погрешности одношаговых методов ... 371
§ 5. Конечно-разностные методы ... 376
§ 6. Метод неопределенных коэффициентов ... 379
§ 7. Исследование свойств конечно-разностных методов на модельных задачах ... 383
§ 8. Оценка погрешности конечно-разностных методов ... 388
§ 9. Особенности интегрирования систем уравнений ... 396
§ 10. Методы численного интегрирования уравнений второго порядка ... 409

§ 11. Оптимизация распределения узлов интегрирования 412

9 Численные методы решения краевых задач для обыкновенных дифференциальных уравнений 417
§ 1. Простейшие методы решения краевой задачи для уравнений второго порядка 417
§ 2. Функция Грина сеточной краевой задачи 423
§ 3. Решение простейшей краевой сеточной задачи 428
§ 4. Замыкания вычислительных алгоритмов 436
§ 5. Обсуждение постановок краевых задач для линейных систем первого порядка 444
§ 6. Алгоритмы решения краевых задач для систем уравнений первого порядка 449
§ 7. Нелинейные краевые задачи 455
§ 8. Аппроксимации специального типа 461
§ 9. Конечно-разностные методы отыскания собственных значений 473
§ 10. Построение численных методов с помощью вариационных принципов 476
§ 11. Улучшение сходимости вариационных методов в нерегулярном случае 485
§ 12. Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения 488

10 Методы решения уравнений в частных производных 495
§ 1. Основные понятия теории метода сеток 497
§ 2. Аппроксимация простейших гиперболических задач 505
§ 3. Принцип замороженных коэффициентов 521
§ 4. Численное решение нелинейных задач с разрывными решениями 524
§ 5. Разностные схемы для одномерного параболического уравнения 528
§ 6. Разностная аппроксимация эллиптических уравнений 543
§ 7. Решение параболических уравнений с несколькими пространственными переменными 566
§ 8. Методы решения сеточных эллиптических уравнений 580

11 Численные методы решения интегральных уравнений 599
§ 1. Решение интегральных уравнений методом замены интеграла квадратурной суммой 599
§ 2. Решение интегральных уравнений с помощью замены ядра на вырожденное 604
§ 3. Интегральные уравнения Фредгольма первого рода 608

Заключение 617

Список литературы 622

Предметный указатель 627