ЧИСЛЕННЫЕ МЕТОДЫ

Н. Н. К а л и т к и н

В книге излагаются основные численные методы решения широкого круга математических задач, возникающих при исследовании физических и технических проблем. Изложенные методы пригодны как для расчетов на ЭВМ, так и для «ручных» расчетов. Для каждого метода даны практические рекомендации по применению. Для лучшего понимания алгоритмов приведены примеры численных расчетов.

Книга предназначена для студентов, аспирантов В преподавателей университетов и технических институтов, научных работников и инженеров-исследователей, а также для всех, имеющих дело с численными расчетами.

ОГЛАВЛЕНИЕ

Предисловие редактора
Предисловие
Глава I
Что такое численные методы?
1. Решение задачи (13). 2. Численные методы (15). 3. История прикладной математики (16). 13
§ 2. Приближенный анализ
1. Понятие близости (17). 2. Структура погрешности (22). 3. Корректность (24). 17

Г л а в а II
Аппроксимация функций
§ 1. Интерполирование
1. Приближенные формулы (27). 2. Линейная интерполяция (27). 3. Интерполяционный многочлен Ньютона (29). 4. Погрешность многочлена Ньютона (31). 5. Применения интерполяции (34). 6. Интерполяционный многочлен Эрмита (36). 7. Сходимость интерполяции (39). 8. Нелинейная интерполяция (41). Интерполяция сплайнами (44). Монотонная интерполяция (46). 11. Многомерная интерполяция (47). 26 27
§ 2. Среднеквадратичное приближение
1. Наилучшее приближение (51). 2. Линейная аппроксимация (53). 3. Суммирование рядов Фурье (56). Метод наименьших, квадратов (59. Нелинейная аппроксимация (62). 51
§ 3. Равномерное приближение
1. Наилучшие приближения (66). 2. Нахождение равномерного приближения (68). 66

Г л а в а III
Численное дифференцирование
1. Полиномиальные формулы (70). 2. Простейшие формулы (72). 3. Метод Рунге — Ромберга (74). 4. Квазиравномерные сетки (78). 5. Быстропеременные функции (80). 6. Регуляризация дифференцирования (81). 69

Г л а в а IV
Численное интегрирование
§ 1. Полиномиальная аппроксимация
1. Постановка задачи (85). Формула трапеций (86). 3. Формула Симпсона (88). 4. Формула средних (89). 5. Формула Эйлера (91). 6. Процесс Эйткена (92). 7. Формулы Гаусса— Кристоффеля (94). 8. Формулы Маркова (97). 9. Сходимость квадратурных формул (98). 84 85
§ 2. Нестандартные формулы
1. Разрывные функции (100). 2. Нелинейные формулы (100). 3. Метод Филона (103). 4. Переменный предел интегрирования (105). 5. Несобственные интегралы (105). 100
§ 3. Кратные интегралы
1. Метод ячеек (108). 2. Последовательное интегрирование (111). 108
§ 4. Метод статистических испытаний
1. Случайные величины (113). 2. Разыгрывание случайной величины (114). 3. Вычисление интеграла (117). 4. Уменьшение дисперсии (119). 5. Кратные интегралы (121). 6. Другие задачи (123). 113
Задачи 124

Глава V
Системы уравнений
§ 1. Линейные системы
1. Задачи линейной алгебры (126). 2. Метод исключения Гаусса (128). 3. Определитель и обратная матрица (130). 4. 0 других прямых методах (132). 5. Прогонка (132). Метод квадратного корня (135). 7. Плохо обусловленные системы (137). 126
§ 2. Уравнение с одним неизвестным
1. Исследование уравнения (138). 2. Дихотомия (139). 3. Удаление корней (140). 4. Метод простых итераций (141). 5. Метод Ньютона (143). 6. Процессы высоких порядков (145). Метод секущих (145). 8. Метод парабол (146). 9. Метод квадрирования (148). 138
§ 3. Системы нелинейных уравнений
1. Метод простых итераций (150). 2. Метод 150

Ньютона (152). 3. Метод спуска (153). 4. Итерационные методы решения линейных систем (153). -
Задачи 155
Глава VI
Алгебраическая проблема собственных значений
§ 1. Проблема и простейшие методы 156
1. Элементы теории (156). 2. Устойчивость (159). 3. Метод интерполяции (162). 4. Трехдиагональные матрицы (164). 5. Почти треугольные матрицы (165). 6. Обратные итерации (166).
§ 2. Эрмитовы матрицы 170
1. Метод отражения (170). 2. Прямой метод вращении (175). 3. Итерационный метод вращении (177).
§ 3. Неэрмитовы матрицы 181
1. Метод элементарных преобразований (181). 2. Итерационные методы (186). 3. Некоторые частные случаи (187).
§ 4. Частичная проблема собственных значений 189
1. Особенности проблемы (189). 2. Метод линеаризации (189). 3. Степенной метод (190). 4. Обратные итерации со сдвигом (191).
Задачи 193
Глава VII
Поиск минимума
1. Постановка задачи (194). 2. Золотое сечение (196). 3. Метод парабол (198). 4. Стохастические задачи (200). 194
§ 2. Минимум функции многих переменных 201
1. Рельеф функции (201). 2. Спуск по координатам (203). 3. Наискорейший спуск (207). 4. Метод оврагов (209). 5. Сопряженные направления (210). 6. Случайный поиск (214).
§ 3. Минимум в ограниченной области 215
1. Формулировка задачи (215). 2. Метод штрафных функций (216). 3. Линейное программирование (217). 4. Симплекс-метод (220). 5. Регуляризация линейного программирования (221).
§ 4. Минимизация функционала 223
1. Задачи на минимум функционала (223). 2. Метод пробных функций (226). 3. Метод Ритца (230). 4. Сеточный метод (240).
Задачи 236
Глава VIII
Обыкновенные дифференциальные уравнения
§ 1. Задача Коши 237
1. Постановка задачи (237). 2. Методы решения (238). 3. Метод Пикара (240). 4. Метод малого параметра (242). 5. Метод ломаных (243). 6. Метод Рунге—Кутта (246). 7. Метод Адамса (250). 8. Неявные схемы (252). 9. Специальные методы (353). 10. Особые точки (257). 11. Сгущение сетки (258).
§ 2. Краевые задачи 261
1. Постановки задач (261). 2. Метод стрельбы (262). 3. Уравнения высокого порядка (266). 4. Разностный метод; линейные задачи (268). 5. Разностный метод; нелинейные задачи (271). 6. Метод Галеркина (276). 7. Разрывные коэффициенты (279).
§ 3. Задачи на собственные значения 280
1. Постановка задач (280). 2. Метод стрельбы (281). 3. Фазовый метод (282). 4. Разностный метод (284). 5. Метод дополненного вектора (286). 6. Метод Галеркина (288).
Задачи 289
Г л а в а IX
Уравнения в частных производных
1. О постановках задач (290). 2. Точные методы решения (292). 3. Автомодельность и подобие (294); 4. Численные методы (296). 290
§ 2. Аппроксимация 299
1. Сетка и шаблон (299). 2. Явные и неявные схемы (301). 3. Невязка (302). 4; Методы составления схем (303). 5. Аппроксимация и ее порядок (307).
§ 3. Устойчивость 311
1. Неустойчивость (311). 2. Основные понятия (312). 3. Принцип максимума (315). 4. Метод разделения переменных (318). 5. Метод энергетических неравенств (322). 6. Операторные неравенства (323).
§ 4. Сходимость 324
1. Основная теорема (324). 2. Оценки точности (327). 3. Сравнение схем на тестах (331).
333
Глава X
Уравнение переноса
1. Задачи и решения (334). 2. Схемы бегущего счета (336). 3. Геометрическая интерпретация устойчивости (341). 4. Многомерное уравнение (344). 5. Перенос с поглощением (346). 6. Монотонность схем (348). 7. Диссипативные схемы (351). 334
§ 2. Квазилинейное уравнение 354
1. Сильные и слабые разрывы (354). 2. Однородные схемы (357). 3. Псевдовязкость (359). 4. Ложная сходимость (362). 5.

Консервативные схемы (363).
Г л а в а XI 366
Параболические уравнения
§ 1. Одномерные уравнения 368
1. Постановки задач (368). 2. Семейство неявных схем (369). 3. Асимптотическая устойчивость неявной схемы (374). 4. Монотонность (376). 5. Явные схемы (378). 6. Наилучшая схема (380). 7. Криволинейные координаты (384). 8. Квазилинейное уравнение (386).
§ 2. Многомерное уравнение 389
1. Экономичные схемы (389). 2. Продольно-поперечная схема (391). 3. Локально-одномерный метод (394). 4. Метод Монте-Карло (399).
Задачи 399
Глава XII
Эллиптические уравнения
§ 1. Счет на установление 401
1. Стационарные решения эволюционных задач (401). 2. Оптимальный шаг (404). 3. Чебышевский набор шагов (409).
§ 2. Вариационные и вариационно-разностные методы 413
1. Метод Ритца (413). 2. Стационарные разностные схемы (414). 3. Прямые методы решения (415). 4. Итерационные методы (420).
Задачи 423
Глава XIII
Гиперболические уравнения
§ 1. Волновое уравнение 424
1. Схема «крест» (424). 2. Неявная схема (427). 3. Двуслойная акустическая схема. (429). 4. Инварианты (434). 5. Явная многомерная схема (435). 6. Факторизованные схемы (436).
§ 2. Одномерные уравнения газодинамики 439
1. Лагранжева форма записи (439). 2. Псевдовязкость (442). 3. Схема «крест» (444). 4. Неявная консервативная схема (447). 5. 0 других схемах (450).
Задачи 451
Глава XIV
Интегральные уравнения
§ 1. Корректно поставленные задачи 452
1. Постановки задач (452). 2. Разностный метод (455). 3. Метод последовательных приближений (458). 4. Замена ядра вырожденным (460). 5. Метод Галеркина (461).
§ 2. Некорректные задачи 462
1. Регуляризация (462). 2. Вариационный метод регуляризации (465). 3. Уравнение Эйлера (469). 4. Некоторые приложения (473). 5. Разностные схемы (476).
Задачи 478
Г л а в а XV
Статистическая обработка эксперимента
1. Ошибки эксперимента (480). 2. Величина и доверительный интервал (482). 3. Сравнение величин (490). 4. Нахождение стохастической зависимости (494).
Задачи 500
Приложение Ортогональные многочлены 501
Литература 505
Предметный указатель 509

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Автомодельные решения 294
Адамса метод 250
Анализ регрессии 495, 496
Анизотропная теплопроводность 394, 395
Аппроксимационная вязкость 351
Аппроксимация 308
— абсолютная 310
— безусловная 310
— дробно-линейная 63
— краевых условий 385, 393, 427
— локальная 309
— условная 310
Асимметрия 487
Бегущая температурная волна 295
Бегущий счет 337, 344, 379
Бесселя формулы 62
Большие задачи 388
Включение точки 388
Вольтерра уравнение второго рода 454
— первого рода 462
Выбор веса 60, 486, 497
Выравнивающая замена переменных 42
Вырожденное ядро 460
Вычисление корней многочлена 147, 148
— кратных интегралов методом Монте-Карло 121
— — — — последовательного интегрирования 111
— — — — ячеек 108
— несобственных интегралов 105
— обратной матрицы 131
— определителя 130
Галеркина метод 276, 288, 461
Гарвика прием 146
Геометрическая интерпретация устойчивости 341, 379

Гивенса метод вращении 175
Гильбертово пространство 20
Двухкруговые итерации 449
Дервюдье метод 189
Дирихле задача 401
Дисбаланс 365
Дисперсионный анализ 495
Диссипативные схемы 353
Дифференцирование быстропеременных функций 80
— интерполяционного многочлена Ньютона 70
— — — —, погрешность 71
— на квазиравномерных сетках 80
— на равномерной сетке 73
Дихотомия 139, 263
Доверительная вероятность 483
Доверительный интервал 483
Допустимое решение 356
Жорданов набор шагов 411
Жорданова подматрица 157
— форма матрицы 157
Замораживание коэффициентов 320
Зейделя метод 155
Инварианты акустические 434
Интегрирование осциллирующих функций 103
— разрывных функций 100
Интегро-интерполяционный метод 304
Интерполяционный многочлен Ньютона 30
— — —, погрешность 32
— — —, —, апостериорная оценка 33
— — Эрмита 36
— — —, погрешность 37
Интерполяция квазилинейная 43
— лагранжева 28
— линейная 28
— многомерная 47
— — на произвольной сетке 50
— — последовательная 49
— — треугольная 49
Интерполяция монотонная 47
— нелинейная 41
— обратная 35
— сплайнами 44
—, сходимость 39
— эрмитова 36
Квадратурные формулы, априорные оценки точности 99
— —, веса 86
— — Гаусса — Кристоффеля 94
— — Маркова 97
— — нелинейные 100
— —, погрешность 86
— — Симпсона 88
— — средних 89
— —, сходимость 98
— — трапеций 86
— — —, погрешность 87
— —, узлы 86
— — Эйлера — Маклорена 91
Комплексная организация расчета 274, 287, 409
Конечные разности 31
Консервативные схемы 365. 447
Корректность 24
Корреляционный анализ 497
Коши задача 238, 291
— — плохо обусловленная 240
Коэффициент парной корреляции 497
— перекоса матрицы 161
Коэффициентная устойчивость 384
Краевые задачи 261, 291
— — нестационарные 291
Критерии установления 408
Куранта условие 338, 436
Лагерра многочлены 503
Лежандра многочлены 501
Линеаризация разностной схемы 321
Линейное программирование 217
Локально-одномерные схемы 396
Матриц виды 132, 158
— нормы 21
Матрица вращения 175
— отражения 170
— сдвинутая 191
Метод баланса 304, 363, 380
— баллистический 262
— вращений итерационный 177
— — —, выбор оптимального элемента 179
— — прямой 175
— выбранных точек 63
— выравнивания 42
— декомпозиции 419
Метод дополненного вектора 286
— золотого сечения 196
— исключения Гаусса, выбор главного элемента 130
— — —, обратный ход 129
— — —, прямой ход 129
— итерированного веса 64, 68
— касательных 143
— квадратного корня 135
— квадрирования 148
— линеаризации 143, 152, 263, 274
— ломаных 243
— малого параметра 242
— моментов 461
— наименьших квадратов 59, 224
— — —, выбор весов 60
— — —, оптимальное число коэффициентов 60
— неопределенных коэффициентов 305

— оврагов 209
— отражений 170
— парабол 146, 198
— последовательных приближений 141, 150, 272, 458 — — —, стохастические задачи 142
— простых итераций 141, 150
— прямых 298
— разностной аппроксимации 303
— секущих 145, 264
— сопряженных направлений 210
— стрельбы 262, 266, 281
— —, линейные задачи 264, 267
— уменьшения невязки 307
— фиктивных точек 306
— штрафных функций 216
Минимизация функционала по аргументу 223
Многочлены обобщенные 28
— ортогональные 501
— — на системе точек 503
Модуль непрерывности 19
Монотонность схем 376, 384
Наилучшая схема 381
Наилучшее приближение 51
— — равномерное 66
— — среднеквадратичное 53
Наискорейший спуск 207
— —, сходимость 208
Направление 299
Невязка 302
Независимые измерения 491
Непрерывный аналог метода Ньютона 288
— функционал 227
Неявные схемы 252, 301
Нормальное распределение 483, 487
Нормальное решение 222, 476
Нормы 19
— векторов 21
— матриц 21
— — подчиненные 22
— — согласованные 22
— негативные 322
— энергетические 308
Ньютона интерполяционный многочлен 30
— метод 143, 152, 263, 274
Обратные итерации 166
— — с переменным сдвигом 192
— — со сдвигом 191
Овраг 203
— разрешимый 203
Однородные схемы 358
Операторов виды 323
—— свойства 323
Оптимальное управление 226
Особые точки дифференциальных уравнений 257
Оценки погрешности апостериорные 33, 330
— — априорные 33, 328
Ошибки грубые 481, 489
— систематические 481
— случайные 481
Первое дифференциальное приближение 352
Пикара метод 240
Плохая обусловленность 25, 240
— — линейных алгебраических систем 127, 130, 137, 476
Подобие 296
Погрешность метода 23
— неустранимая 22
— округления 23
Показатель симметрии 384, 440
Полностью консервативные схемы 366, 450
Попеременно-треугольная схема 421
Порядок точности 325, 327
— — не целый 93, 340
Последовательность точек ЛПt 121
— функций минимизирующая 227
Потенциал скоростей 429
Предиктор-корректор 247
Преобладание диагонального элемента 134, 154
Преобразование подобия матриц 158
Признак равномерной устойчивости 314, 316, 319
Принцип максимума 315
Прогонка 132
Прогонка дифференциальная 266
Продольно-поперечная схема 391
Пространство *C* 19
Псевдовязкость 359
— квадратичная 361, 443
— линейная 362, 442
Псевдослучайные числа 115
Разделенные разности 29
— — с кратными узлами 37
Разрывные коэффициенты 279, 380
Разыгрывание случайной величины 117
— — — многомерной 122
— — — равномерно распределенной 115
Регуляризация дифференцирования по Тихонову 474
— — по шагу 83
— — сглаживанием 83
— линейного программирования 221
— суммирования ряда по Тихонову 58, 475
— — — по числу членов 57
Регуляризирующий оператор 464
Рельеф функции 201
Решение уравнения обратной интерполяцией 35
Ритца метод 230, 413

Рунге — Кутта метод 246
— — —, оценка точности 249
Рунге метод 75, 259, 332
— — рекуррентный 77, 331
Рунге — Ромберга метод 76
Сглаживание функции 60, 62, 474
Сетки квазиравномерные 78
— специальные 279, 383
Сильный разрыв 357
Симплекс-метод 220
Слабый разрыв 355
Слой 299
Случайная величина 114
— —, плотность распределения 114
— —, равномерно распределенная 114
— —, ——, разыгрывание 115
— —, разыгрывание 117
Собственные значения 156, 280
Согласованные измерения 492
Сплайн 46
— многомерный 235
Способ параллельных касательных 211
Спуск по координатам 203
Стандарт 484
— выборки 485
— —, несмещенная оценка 484
Степенной метод 190
Стохастическая зависимость 495
Стохастическая задача нахождения минимума 194
Стьюдента коэффициенты 485
— критерий 485
Субтабулирование 34
Схема двуслойная 313
— —, каноническая форма 318
— «крест» 425, 435, 444
— ломаных 243
— с весами 370
—с выделением особенностей 358, 430
— с полусуммой 371
Сходимость 325
— векторов по направлению 21
— квадратичная 145
— кубическая 145
— линейная 145
— ложная 362
— равномерная 19
— среднеквадратичная 20
Счет на установление 190, 403
— — —, критерий установления 408
— — —, оптимальный шаг 404
Тихоновский стабилизатор 405
Точки повышенной точности численного дифференцирования 72
Треугольный оператор 421
Удаление найденных корней 140
Узлы сетки нерегулярные 300
— — регулярные 300
Уменьшение дисперсии метода Монте-Карло 119
Устойчивость 24, 312
— асимптотическая 314, 374
— безусловная 313
— по начальным данным 313
— — — — равномерная 313
— слабая 25, 314
— собственных значений и векторов матриц 159
— условная 313
Фазовый метод 282
Факторизованные схемы 437
Филона формулы 103
Фишера коэффициенты 494
— критерий 493
Фредгольма уравнение второго рода 453
— — первого рода 462
Фурье преобразование быстрое 416
— — дискретное 62
Характеристический многочлен 156
Хаусхолдера метод отражений 170
Центральные моменты распределения 487
Циклическая прогонка 434
Чебышева критерий 486
— многочлены 503
Чебышевская система функций 28
Чебышевский набор шагов 409
— — — упорядоченный 412
Чисто неявная схема 371
Шаблон 297, 300
Эйлера метод 243
— уравнение 469
Эйткена экстраполяционный процесс 92
Экономичные схемы 391
Экстраполяция 33
— многомерная 48
Эксцесс 487
Эрмита многочлены интерполяционные 36
— — ортогональные 503
Явно-неявная схема 342
Явные схемы 301
Якоби метод вращении 177
— многочлены ортогональные 501

ПРЕДИСЛОВИЕ РЕДАКТОРА

Современное развитие физики и техники тесно связано с использованием электронных вычислительных машин (ЭВМ). В настоящее время ЭВМ стали обычным оборудованием многих

институтов и конструкторских бюро. Это позволило от простейших расчетов и оценок различных конструкций или процессов перейти к новой стадии работы—детальному математическому моделированию (вычислительному эксперименту), которое существенно сокращает потребность в натурных экспериментах, а в ряде случаев может их заменить.

В основе вычислительного эксперимента лежит решение уравнений математической модели численными методами. Изложению численных методов посвящено немало книг. Однако большинство этих книг ориентировано на студентов и научных работников математического профиля. Поэтому в настоящее время ощущается потребность в книге, рассчитанной на широкий круг читателей различных специальностей и сочетающей достаточную полноту изложения с разумной степенью строгости при умеренном объеме.

Предлагаемая книга отвечает этим требованиям. Она достаточно полно освещает тот круг вопросов, знание которого наиболее часто требуется в практике вычислений, и содержит ряд разделов, которые редко включают в учебные пособия. Умеренный объем достигнут за счет тщательного отбора материала и включения в книгу только наиболее эффективных и часто используемых на практике методов. Материал изложен четко и сжато, при этом большое внимание уделено рекомендациям по практическому применению алгоритмов; изложение пояснено рядом примеров. Для обоснования алгоритмов использован несложный математический аппарат, знакомый студентам физических и инженерных специальностей.

Книга рассчитана на читателя, который занимается не столько разработкой численных методов, сколько их применением к прикладным проблемам. Однако в процессе работы над книгой читатель знакомится с основными идеями построения вычислительных алгоритмов и с их обоснованием и приобретает знания, достаточные для разработки новых алгоритмов. Эта книга является по существу введением в численные методы. Овладев ею, читатель затем может углубить свои знания, обратившись к руководствам по теории разностных схем и по методам численного решения отдельных классов задач.

Книга написана специалистом по теоретической и математической физике. Она возникла в результате работы автора над рядом актуальных проблем физики в Институте прикладной математики АН СССР и преподавания на физическом факультете МГУ.

Несомненно, книга окажется полезной широкому кругу читателей — студентам, аспирантам, научным сотрудникам и инженерам математических, физических и технических специальностей.

А. А. Самарский

ПРЕДИСЛОВИЕ

Сложные вычислительные задачи, возникающие при исследовании физических и технических проблем, можно разбить на ряд элементарных—таких как вычисление интеграла, решение дифференциального уравнения и т. п. Многие элементарные задачи являются несложными и хорошо изучены. Для этих задач уже разработаны методы численного решения, и нередко имеются стандартные программы решения их на ЭВМ. Есть и достаточно сложные элементарные задачи; методы решения таких задач сейчас интенсивно разрабатываются (например, решение уравнений бесстолкновительной плазмы).

Поэтому полная программа обучения численным методам должна состоять из ряда этапов. Во-первых, это освоение логарифмической линейки, клавишных вычислительных машин и программирования на ЭВМ. Во-вторых, основы численных методов, содержащие изложение классических элементарных задач (включая основные сведения о разностных схемах). В-третьих, курс теории разностных схем. И в-четвертых — ряд специальных курсов, которые сейчас нередко называют методами вычислительной физики: численное решение задач газодинамики, аэродинамики, переноса излучения, квантовой физики, квантовой химии и т. д.

Эта книга является введением в численные методы. Она начинается с простейших задач интерполирования функций и кончается недавно возникшим разделом вычислительной математики — методами решения некорректно поставленных задач. Книга написана на основе годового курса лекций, читавшихся автором сначала инженерам-конструкторам, а после переработки—студентам физического факультета МГУ. Для каждой задачи существует множество методов решения. Например, хорошо обусловленную систему линейных уравнений можно решать методами Гаусса, Жордана, оптимального исключения, окаймления, отражений, ортогонализации и рядом других. Интерполяционный многочлен записывают в формах Лагранжа, Ньютона, Грегори—Ньютона, Бесселя, Стирлинга, Гаусса и Лапласа—Эверетта. Подобные методы обычно являются вариациями одного-двух основных методов, и если даже

в каких-то частных случаях имеют преимущества, то незначительные. Кроме того, многие методы создавались до появления ЭВМ, и ряд из них в качестве существенного элемента включает интуицию вычислителя. Появление ЭВМ потребовало переоценки старых методов, что до конца еще не сделано, и до сих пор по традиции большое количество неэффективных методов кочует из учебника в учебник. Отчасти это объясняется тем, что эффективность многих методов сильно зависит от мелких деталей алгоритма, почти не поддающихся теоретическому анализу; поэтому окончательный отбор лучших методов можно сделать только на основании большого опыта практических расчетов.

В этой книге сделана попытка такого отбора, опирающаяся на многолетний опыт решения большого числа разнообразных задач математической физики. Для большинства рассмотренных в книге задач изложены только наиболее эффективные методы с широкой областью применимости. Несколько методов для одной и той же задачи даны в том случае, если они имеют существенно разные области применимости, или если для данной задачи еще не разработано достаточно удовлетворительных методов.

Часто приходится слышать, что наступила эпоха ЭВМ, а «ручные» расчеты являются архаизмом. На самом деле это далеко не так. Прежде чем поручать ЭВМ большую задачу, надо сделать много оценочных расчетов и на их основе понять, какие методы окажутся эффективными для данной задачи. Конечно, даже в мелких расчетах ЭВМ с хорошим математическим обеспечением и набором периферийных устройств (телетайп, дисплей, графико-построитель) оказывает большую пользу. Однако логарифмическая линейка и клавишные машины еще долго будут необходимы. Поэтому большинство методов, изложенных здесь, в равной мере пригодны для ЭВМ и «ручных» расчетов.

Основное внимание в книге уделено выработке практических навыков у читателя. Поэтому в первую очередь изложены алгоритмы, даны рекомендации по их применению и отмечены «маленькие хитрости»—те незначительные на первый взгляд практические приемы, которые сильно повышают эффективность алгоритма. Теоретическое обоснование методов приведено лишь в той мере, в какой оно необходимо для лучшего усвоения и практического применения.

В книгу включен ряд сведений, не относящихся к необходимому минимуму, но полезных читателю для лучшего понимания тонких деталей вычислительных процессов. Чтобы не увеличивать объем книги и избежать сложных выкладок, эти сведения приведены, как правило, без доказательств, но со ссылками на дополнительную литературу. Некоторые сведения даны в форме задач в конце каждой главы. Предполагается, что читатели знакомы с основами высшей математики, включая краткие сведения об уравнениях в частных производных. Необходимые дополнительные сведения, которые не содержатся в обязательных курсах университетов и втузов, сообщаются здесь в соответствующих разделах.

Книга разделена на главы; параграфы и пункты. В начале каждой главы кратко изложено ее содержание. Нумерация таблиц и рисунков—единая по всей книге, а нумерация формул—самостоятельная в каждой главе. Если ссылка не выходит за пределы данной главы, то указывается только номер формулы; если выходит—то номер главы и номер формулы. В конце книги дан список литературы. Приведенные в нем учебники и монографии рекомендуются для углубленного изучения отдельных разделов. Журнальные статьи даны для указания на оригинальные работы, их список не претендует на полноту; более полная библиография имеется в рекомендованных учебниках.

Общий подход к теории и практике вычислений, определивший стиль этой книги, сложился у меня под влиянием А. А. Самарского и В. Я. Гольдина за много лет совместной работы. Ряд актуальных тем был включен по инициативе, А. Г. Свешникова и В. Б. Гласко. Много ценных замечаний сделали А. В. Гулин, Б. Л. Рождественский, И. М. Соболь, И. В. Фрязинов, Е. В. Шикин и сотрудники кафедры прикладной математической физики МИФИ. В оформлении рукописи мне помогли Л. В. Кузьмина и В. А. Кра-сноярова. Я пользуюсь случаем искренне поблагодарить всех названных лиц, и в особенности Александра Андреевича Самарского.

Н. Н. Калиткин

ГЛАВА I

ЧТО ТАКОЕ ЧИСЛЕННЫЕ МЕТОДЫ?

Глава I является вводной. В § 1 рассмотрены роль математики при решении физико-технических задач и место численных методов среди других математических методов и кратко изложена история численных методов. В § 2 разобраны основные понятия приближенного анализа: корректность постановки задач, определение близости точного и приближенного решений, структура погрешности.

§ 1. Математические модели и численные методы

1. Решение задачи. Физиков математика интересует не сама по себе, а как средство решения физических задач. Рассмотрим поэтому, как решается любая реальная задача — например, нахождение светового потока конструируемой лампы, производительности проектируемой химической установки или себестоимости продукции строящегося завода.

Одним из способов решения является эксперимент. Построим эту лампу, установку или завод и измерим интересующую нас характеристку. Если характеристика оказалась неудачной, то изменим проект и построим новый завод и т. д. Ясно, что мы получим достоверный ответ на вопрос, но слишком медленным и дорогим способом.

Другой способ — математический анализ конструкции или явления. Но такой анализ применяется не к реальным явлениям, а к некоторым математическим моделям этих явлений. Поэтому первая стадия работы — это *формулировка математической модели* (постановка задачи). Для физического процесса модель обычно состоит из уравнений, описывающих процесс; в эти уравнения в виде коэффициентов входят характеристики тел или веществ, участвующих в процессе. Например, скорость ракеты при вертикальном полете в вакууме определяется уравнением

$$\left(M-\int_0^t m(\tau)\,d\tau\right)\left(\frac{dv}{dt}+g\right)=cm(t), \qquad (1)$$

где M — начальная масса ракеты, $m(t)$ — заданный расход горю-

чего, g — ускорение поля тяготения, а c — скорость истечения газов, зависящая от калорийности топлива и среднего молекулярного веса продуктов сгорания.

Любое изучаемое явление бесконечно сложно. Оно связано с другими явлениями природы, возможно, не представляющими интереса для рассматриваемой задачи. Математическая модель должна охватывать важнейшие для данной задачи стороны явления. Наиболее сложная и ответственная работа при постановке задачи заключается в выборе связей и характеристик явления, существенных для данной задачи и подлежащих формализации и включению в математическую модель.

Если математическая модель выбрана недостаточно тщательно, то, какие бы методы мы ни применяли для расчета, все выводы будут недостаточно надежны, а в некоторых случаях могут оказаться совершенно неправильными. Так, уравнение (1) непригодно для запуска ракеты с поверхности Земли, ибо в нем не учтено сопротивление воздуха.

Вторая стадия работы — это *математическое исследование*. В зависимости от сложности модели применяются различные математические подходы. Для наиболее грубых и несложных моделей зачастую удается получить аналитические решения; это излюбленный путь многих физиков-теоретиков. Например, уравнение (1) легко интегрируется при $g = \text{const}$ и $m(t) = \text{const}$:

$$v = c \ln [M/(M - mt)] - gt.$$

Из-за грубости модели физическая точность этого подхода невелика; нередко такой подход позволяет оценить лишь порядки величин.

Для более точных и сложных моделей аналитические решения удается получить сравнительно редко. Обычно теоретики пользуются приближенными математическими методами (например, разложением по малому параметру), позволяющими получить удовлетворительные качественные и количественные результаты. Наконец, для наиболее сложных и точных моделей основными методами решения являются численные; как правило, они требуют проведения расчетов на ЭВМ. Эти методы зачастую позволяют добиться хорошего количественного описания явления, не говоря уже о качественном.

Во всех случаях математическая точность решения должна быть несколько (в 2 — 4 раза) выше, чем ожидаемая физическая точность модели. Более высокой математической точности добиваться бессмысленно, ибо общую точность ответа это все равно не повысит. Но более низкая математическая точность недопустима (для облегчения решения задачи нередко в ходе работы делают дополнительные математические упрощения; это снижает ценность результатов).

Наконец, третья стадия работы — это *осмысливание математического решения* и сопоставление его с экспериментальными данными. Если расчеты хорошо согласуются с контрольными экспериментами, то это свидетельствует о правильном выборе модели; такую модель можно использовать для расчета процессов данного типа. Если же расчет и эксперимент не согласуются, то модель необходимо пересмотреть и уточнить.

2. Численные методы являются одним из мощных математических средств решения задачи. Простейшие численные методы мы используем всюду, например, извлекая квадратный корень на листке бумаги. Есть задачи, где без достаточно сложных численных методов не удалось бы получить ответа; классический пример — открытие Нептуна по аномалиям движения Урана.

В современной физике таких задач много. Более того, часто требуется выполнить огромное число действий за короткое время, иначе ответ будет не нужен. Например, суточный прогноз погоды должен быть вычислен за несколько часов; коррекцию траектории ракеты надо рассчитать за несколько минут (напомним, что для расчета орбиты Нептуна Леверье потребовалось полгода); режим работы прокатного стана должен исправляться за секунды. Это немыслимо без мощных ЭВМ, выполняющих тысячи или даже миллионы операций в секунду.

Современные численные методы и мощные ЭВМ дали возможность решать такие задачи, о которых полвека назад могли только мечтать. Но применять численные методы далеко не просто. Цифровые ЭВМ умеют выполнять только арифметические действия и логические операции. Поэтому помимо разработки математической модели, требуется еще разработка алгоритма, сводящего все вычисления к последовательности арифметических и логических действий. Выбирать модель и алгоритм надо с учетом скорости и объема памяти ЭВМ: чересчур сложная модель может оказаться машине не под силу, а слишком простая — не даст физической точности.

Сам алгоритм и программа для ЭВМ должны быть тщательно проверены. Даже проверка программы нелегка, о чем свидетельствует популярное утверждение: «В любой сколь угодно малой программе есть по меньшей мере одна ошибка». Проверка алгоритма еще более трудна, ибо для сложных алгоритмов не часто удается доказать сходимость классическими методами. Приходится использовать более или менее надежные «экспериментальные» проверки, проводя пробные расчеты на ЭВМ и анализируя их (смотри, например, главу IX, § 4, п. 3).

Строгое математическое обоснование алгоритма редко бывает исчерпывающим исследованием. Например, большинство доказательств сходимости итерационных процессов справедливо только при точном выполнении всех вычислений; практически же число

сохраняемых десятичных знаков редко происходит 5—6 при «ручных» вычислениях и 10—12 при вычислениях на ЭВМ. Плохо поддаются теоретическому исследованию «маленькие хитрости» — незначительные на первый взгляд детали алгоритма, сильно влияющие на его эффективность. Поэтому окончательную оценку метода можно дать только после опробования его в практических расчетах.

К чему приводит пренебрежение этими правилами — видно из принципа некомпетентности Питера: «ЭВМ многократно увеличивает некомпетентность вычислителя».

Для сложных задач разработка численных методов и составление программ для ЭВМ очень трудоемки и занимают от нескольких недель до нескольких лет. Стоимость комплекса отлаженных программ нередко сравнима со стоимостью экспериментальной физической установки. Зато проведение отдельного расчета по такому комплексу много быстрей и дешевле, чем проведение отдельного эксперимента. Такие комплексы позволяют подбирать оптимальные параметры исследуемых конструкций, что не под силу эксперименту.

Однако численные методы не всесильны. Они не отменяют все остальные математические методы. Начиная исследовать проблему, целесообразно использовать простейшие модели, аналитические методы и прикидки. И только разобравшись в основных чертах явления, надо переходить к полной модели и сложным численным методам; даже в этом случае численные методы выгодно применять в комбинации с точными и приближенными аналитическими методами.

Современный физик или инженер-конструктор для успешной работы должен одинаково хорошо владеть и «классическими» методами, и численными методами математики.

3. История прикладной математики. Раздел математики, имеющий дело с созданием и обоснованием численных алгоритмов для решения сложных задач различных областей науки, часто называют прикладной математикой; американцы применение численных методов к физическим задачам называют вычислительной физикой. Главная задача прикладной математики — фактическое нахождение решения с требуемой точностью; этим она отличается от классической математики, которая основное внимание уделяет исследованию условий существования и свойств решения.

В истории прикладной математики можно выделить три основных периода.

Первый начался 3—4 тысячи лет назад. Он был связан с ведением конторских книг, вычислением площадей и объемов, расчетами простейших механизмов; иными словами — с несложными задачами арифметики, алгебры и геометрии. Вычислительными средствами служили сначала собственные пальцы, а затем

— счеты. Исходные данные содержали мало цифр, и большинство выкладок выполнялось точно, без округлений.

Второй период начался с Ньютона. В этот период решались задачи астрономии, геодезии и расчета механических конструкций, сводящиеся либо к обыкновенным дифференциальным уравнениям, либо к алгебраическим системам с большим числом неизвестных. Вычисления выполнялись с округлением; нередко от результата требовалась высокая точность, так что приходилось сохранять до 8 значащих цифр.

Вычислительные средства стали разнообразнее: таблицы элементарных функций, затем — арифмометр и логарифмическая линейка; к концу этого периода появились неплохие клавишные машины с электромотором. Но скорость всех этих средств была невелика, и вычисления занимали дни, недели и даже месяцы.

Третий период начался примерно с 1940 г. Военные задачи — например, наводка зенитных орудий на быстро движущийся самолет — требовали недоступных человеку скоростей и привели к разработке электронных систем. Появились электронные вычислительные машины (ЭВМ).

Скорость даже простейших ЭВМ настолько превосходила скорость механических средств, что стало возможным проводить вычисления огромного объема. Это позволило численно решать новые классы задач; например, процессы в сплошных средах, описывающиеся уравнениями в частных производных.

Сначала для решения эти задач использовались численные методы, разработанные в «доэлектронный» период. Но применение ЭВМ быстро привело к переоценке методов. Многие старые методы оказались неподходящими для автоматизированных расчетов. Стали быстро разрабатываться новые методы, ориентированные прямо на ЭВМ (например, метод Монте-Карло).

Мощности ЭВМ быстро растут. Если в 50-е гг. в СССР вступила в строй первая «Стрела» со скоростью 2000 операций в секунду и памятью 1024 ячейки, то сейчас во многих вычислительных центрах страны работают БЭСМ-6 со скоростью в 300 раз больше и памятью в 30 раз больше. А наилучшие современные ЭВМ имеют скорость до 30 миллионов операций в секунду при практически неограниченной оперативной памяти с прямой адресацией. Становятся возможными расчеты все более сложных задач. Это служит стимулом для разработки новых численных методов.

§ 2. Приближенный анализ

1. Понятие близости. Если требуется определить некоторую величину y по известной величине x, то символически задачу можно записать в виде $y = A(x)$. Здесь и y, и x могут быть числами, совокупностью чисел, функцией одного или нескольких

переменных, набором функций и т.д. Если оператор A настолько сложен, что решение не удается явно выписать или точно вычислить, то задачу решают приближенно.

Например, пусть надо вычислить $y=\int_a^b x(t)\,dt$. Можно приближенно заменить $x(t)$ многочленом $\bar{x}(t)$ или другой функцией, интеграл от которой легко вычислить. А можно заменить интеграл суммой $\sum_i x(t_i)\,\Delta t_i$, вычислить которую тоже несложно. Таким образом, приближенный метод заключается в замене исходных данных на близкие данные $\bar{x}$ и (или) замене оператора на близкий оператор $\bar{A}$, так чтобы значение $\bar{y}=\bar{A}(\bar{x})$ легко вычислялось. При этом мы ожидаем, что значение $\bar{y}$ будет близко к искомому решению.

Но что такое «близко»? Очевидно, для двух чисел x_1 и x_2 надо требовать малости $|x_1-x_2|$; а близость двух функций можно определить разными способами. Эти вопросы рассматриваются в функциональном анализе, некоторые понятия которого будут сейчас изложены.

Множество элементов x любой природы называется *метрическим пространством,* если в нем введено расстояние $\rho(x_1, x_2)$ между любой парой элементов (*метрика*), удовлетворяющее следующим аксиомам:

$$\begin{aligned}
&\text{а) } \rho(x_1, x_2) \text{ — вещественное неотрицательное число,}\\
&\text{б) } \rho(x_1, x_2)=0, \text{ только если } x_1=x_2,\\
&\text{в) } \rho(x_1, x_2)=\rho(x_2, x_1),\\
&\text{г) } \rho(x_1, x_3)\leqslant\rho(x_1, x_2)+\rho(x_2, x_3).
\end{aligned}\tag{2}$$

Последовательность элементов x_n метрического пространства называется *сходящейся* (по метрике) к элементу x, если $\rho(x_n, x)\to 0$ при $n\to\infty$. Последовательность x_n называется *фундаментальной*, если для любого $\varepsilon>0$ найдется такое $k(\varepsilon)$, что $\rho(x_n, x_m)<\varepsilon$ при всех n и $m>k$.

Метрическое пространство называют *полным*, если любая фундаментальная последовательность его элементов сходится к элементу того же пространства. Примером неполного пространства является множество рациональных чисел $x=(n/m)$ с метрикой $\rho(x_1, x_2)=|x_1-x_2|$. Последовательность $x_k=(1+1/k)^k$ ему принадлежит, является фундаментальной, а сходится к иррациональному числу e, т.е. не к элементу данного пространства. Если переменные y, x принадлежат неполным пространствам, то обосновать сходимость численных методов очень трудно: даже если удается доказать, что при $x_n\to x$ последовательность y_n фундаментальная, то отсюда еще не следует, что она сходится к элементу данного пространства, т. е. к решению допустимого класса.

Элементами наших множеств будут числа, векторы, матрицы, функции и т. п. Сами множества обычно являются линейными нормированными пространствами, ибо в них определены операции сложения элементов и умножения их на число и введена *норма* каждого элемента $\|x\|$, причем выполнены следующие аксиомы:

$$x_1+x_2=x_2+x_1,\quad (x_1+x_2)+x_3=x_1+(x_2+x_3);$$

существует единственный элемент θ такой, что $x+\theta=x$ для любого x (будем использовать для θ обозначение 0); для всякого x существует единственный элемент $-x$ такой, что $x+(-x)=\theta$; (3)

$$a(x_1+x_2)=ax_1+ax_2;\quad (a+b)x=ax+bx;$$

$a(bx)=(ab)x$; $1\cdot x=x$; $0\cdot x=\theta$ единствен; $\|x\|\geqslant 0$ — вещественное число; $\|ax\|=|a|\cdot\|x\|$; $\|x\|=0$ только при $x=0$; $\|x_1+x_2\|\leqslant\|x_1\|+\|x_2\|$.

Линейное нормированное пространство есть частный случай метрического пространства, а норма определяется метрикой. Полное линейное нормированное пространство называется *банаховым*. Практически всегда величины, с которыми мы будем оперировать, являются элементами банаховых пространств; это важно при доказательстве сходимости численных методов.

Рассмотрим некоторые примеры банаховых пространств, с которыми нам часто придется встречаться. Выполнимость аксиом (3) и полноту читатели легко проверят сами.

а) Множество всех действительных чисел с нормой $\|x\|=|x|$.

б) Пространство C — множество функций $x(t)$, определенных и непрерывных при $0\leqslant t\leqslant 1$, с чебышевской нормой $\|x\|_c=\max|x(t)|$. Сходимость в этом пространстве называется *равномерной*. Условие $0\leqslant t\leqslant 1$ здесь и в следующем примере принято для удобства; оно не является существенным, и можно определять функции на любом конечном отрезке.

Класс непрерывных функций часто еще сужают, накладывая на функции дополнительные требования: липшиц-непрерывности, однократной или многократной дифференцируемости и т. д. Напомним некоторые определения.

Функция $x(t)$ называется *равномерно-непрерывной* на отрезке, если для сколь угодно малого $\omega>0$ найдется такое δ, что $|x(t_1)-x(t_2)|\leqslant\omega$ для любой пары точек отрезка, удовлетворяющих условию $|t_1-t_2|\leqslant\delta$. Таким образом, устанавливается функциональная связь между ω и δ. Величина $\omega(\delta)$ называется *модулем непрерывности* функции. Функция, непрерывная во всех

точках замкнутого отрезка $a \leqslant t \leqslant b$, является на этом отрезке ограниченной и равномерно-непрерывной (теорема Кантора); следовательно, пространство C — множество ограниченных и равномерно-непрерывных функций. Если $\omega(\delta) \leqslant K\delta$, где K — некоторая константа, то функцию называют *липшиц-непрерывной*. Нетрудно видеть, что если функция имеет ограниченную производную, то она липшиц-непрерывна, причем $K = \sup |x'(t)|$.

в) Пространство L_p — множество функций $x(t)$, определенных при $0 \leqslant t \leqslant 1$ и интегрируемых по модулю с p-й степенью, если норма определена

$$\| x \|_{L_p} = \left[\int_0^1 |x(t)|^p \, dt \right]^{1/p}.$$

Сходимость в такой норме называют сходимостью *в среднем*. Пространство L_2 называют *гильбертовым*, а сходимость в нем — *среднеквадратичной*.

Разницу между равномерной близостью и близостью в среднем иллюстрирует рис. 1. Функция x_2 равномерно близка к функции x_1, а функция x_3 близка в среднем, т. е. мало отличается от x_1 на большей части отрезка, но может сильно отличаться от нее на небольших участках.

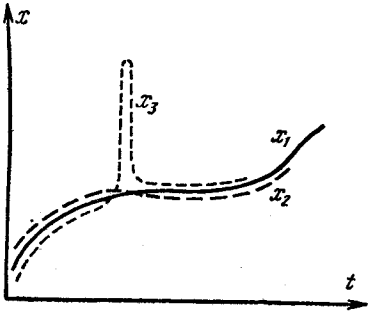


Рис. 1.

Выбирая метрические пространства, т. е. выбирая множества X, Y и определяя в них метрики, мы тем самым уславливаемся, в каких классах функций можно брать начальные данные и искать решение. Поэтому в конкретной задаче выбор пространств должен в первую очередь определяться физическим смыслом задачи, и лишь во вторую — чисто математическими соображениями (такими, например, как возможность доказать сходимость). Например, при расчете прочности самолета нужна равномерная близость приближенного решения к точному, а близости в среднем недостаточно: перенапряжение в маленьком участке может разрушить конструкцию. А в задаче о нагреве тела потоком тепла даже норма L_1 удовлетворительна, ибо температура тела определяется интегралом от потока по времени.

Нетрудно показать, что между разными нормами (если они существуют) выполняются определенные соотношения. Если функции $x(t)$ определены при $0 \leqslant t \leqslant 1$, тогда

$$\| x(t) \|_{L_1} \leqslant \| x(t) \|_{L_2} \leqslant \ldots \leqslant \| x(t) \|_C. \qquad (4)$$

В самом деле, например:

$$\| x(t) \|_{L_p}^p = \int_0^1 | x(t) |^p \, dt \leqslant \int_0^1 \max | x(t) |^p \, dt = \max | x(t) |^p = \| x(t) \|_C^p.$$

Следовательно, из равномерной сходимости вытекает сходимость в среднем, в частности — среднеквадратичная. Поэтому чебышевскую норму называют *более сильной*, чем гильбертова.

г) Координатные бесконечномерные пространства, элементами которых являются счетные множества чисел $x = \{x_1, x_2, \ldots\}$. По аналогии с пространствами функций, в них обычно вводят норму $\| x \|_c = \sup | x_i |$ или

$$\| x \|_{l_p} = \lim_{n \to \infty} \left(\frac{1}{n} \sum_{i=1}^{n} | x_i |^p \right)^{1/p},$$

а само пространство называют соответственно c или l_p.

д) Конечномерные пространства $c^{(n)}$, $l_p^{(n)}$, элементами которых являются группы из n чисел $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$; их можно считать координатами векторов в n-мерном пространстве, $l_2^{(n)}$ называют евклидовым. Нормы векторов вводят по аналогии со случаем (г), например,

$$\| \boldsymbol{x} \|_p = \left(\frac{1}{n} \sum_{i=1}^{n} | x_i |^p \right)^{1/p}.$$

Для конечномерных векторов между разными нормами существуют соотношения

$$\| \boldsymbol{x} \|_1 \leqslant \| \boldsymbol{x} \|_2 \leqslant \| \boldsymbol{x} \|_c \leqslant \sqrt{n} \, \| \boldsymbol{x} \|_2 \leqslant n \, \| \boldsymbol{x} \|_1, \tag{5}$$

которые легко проверить. Поэтому из сходимости в одной из этих норм следует сходимость во всех остальных нормах. Нормы, обладающие таким свойством, называют *эквивалентными*.

Отметим, что если последовательность векторов $\boldsymbol{x}_m$ не сходится, но $\boldsymbol{x}_m / \| \boldsymbol{x}_m \|$ сходится, то говорят о сходимости векторов *по направлению*.

е) В пространстве квадратных матриц порядка n наиболее употребительны следующие нормы:

$$\| A \|_c = \max_i \left(\sum_{j=1}^{n} | a_{ij} | \right), \quad \| A \|_1 = \max_j \left(\sum_{i=1}^{n} | a_{ij} | \right),$$

$$\| A \|_M = n \cdot \max_{i,j} | a_{ij} |, \qquad \| A \|_E = \left(\sum_{i,j=1}^{n} | a_{ij} |^2 \right)^{1/2}, \tag{6}$$

$$\| A \|_2 = \sqrt{\max \mu_i},$$

где μ_i — собственные значения эрмитовой матрицы $A^H A$ (здесь A^H — матрица, эрмитово сопряженная по отношению к A). Первые две нормы не имеют специальных названий, третья называется максимальной, четвертая — сферической или евклидовой и пятая — спектральной. Между ними выполняются некоторые соотношения, аналогичные (5).

Интересна связь между нормами матриц и векторов, на которые матрицы действуют. Норма матрицы называется *согласованной* с нормой вектора, если $\|Ax\| \leqslant \|A\| \cdot \|x\|$. Наименьшая из норм матрицы, согласованных с данной нормой вектора: $\|A\| = \sup(\|Ax\| / \|x\|)$, называется нормой матрицы, *подчиненной* данной норме вектора.

Приведем пример подчиненной нормы. Из цепочки неравенств

$$\|Ax\|_c = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \max_i \left[\left(\max_j |x_j| \right) \sum_{k=1}^{n} |a_{ik}| \right] =$$

$$= \|x\|_c \cdot \max_i \left(\sum_{k=1}^{n} |a_{ik}| \right) = \|A\|_c \cdot \|x\|_c \qquad (7)$$

следует, что $\|A\|_c$ согласована с $\|x\|_c$. Кроме того, для любой матрицы A существует такой вектор x, что неравенство (7) обращается в равенство. Для его нахождения положим $x_j = \pm 1$; знаки выберем так, чтобы они совпадали со знаками элементов a_{ij} той строки матрицы i, в которой $\sum_{j=1}^{n} |a_{ij}|$ максимальна. Тогда именно сумма по этой строке будет максимальна в левой части (7), и неравенство превратится в равенство. Это означает, что $\|A\|_c$ есть наименьшая из норм, согласованных с $\|x\|_c$: если мы возьмем еще меньшую $\|A\|$, то при этом векторе x для нее знак неравенства (7) будет обратным, т. е. она не будет согласованной. Следовательно, $\|A\|_c$ подчинена $\|x\|_c$.

Без доказательства укажем, что $\|A\|_1$ подчинена $\|x\|_1$, и спектральная норма подчинена $\|x\|_2$. Сферическая норма согласована с $\|x\|_2$, а максимальная норма согласована со всеми рассмотренными выше векторными нормами.

2. Структура погрешности. Есть четыре источника погрешности результата: математическая модель, исходные данные, приближенный метод и округления при вычислениях. Погрешность математической модели связана с физическими допущениями и здесь рассматриваться не будет.

Исходные данные зачастую неточны; например, это могут быть экспериментально измеренные величины. В прецизионных физических измерениях точность доходит до 10^{-12}, но уже характерная астрономическая и геодезическая точность равна 10^{-6}, а во многих физических и технических задачах погрешность измерения бывает 1 — 10%. Погрешность исходных данных δx приводит к так называемой *неустранимой* (она не зависит от математика) погрешности решения $\delta y = A(x + \delta x) - A(x)$. В следующем пункте будут рассмотрены случаи, когда неустранимая погрешность может становиться недопустимо большой.

Погрешность метода связана с тем, что точные оператор и исходные данные заменяются приближенными. Например, заменяют интеграл суммой, производную — разностью, функцию — многочленом или строят бесконечный итерационный процесс и обрывают его после конечного числа итераций. Методы строятся обычно так, что в них входит некоторый параметр; при стремлении параметра к определенному пределу погрешность метода стремится к нулю, так что эту погрешность можно регулировать. Погрешность метода мы будем исследовать при рассмотрении конкретных методов.

Погрешность метода целесообразно выбирать так, чтобы она была в $2-5$ раз меньше неустранимой погрешности. Бóльшая погрешность метода снижает точность ответа, а заметно меньшая — невыгодна, ибо это обычно требует значительного увеличения объема вычислений.

Вычисления как на бумаге, так и на ЭВМ выполняют с определенным числом значащих цифр. Это вносит в ответ *погрешность округления*, которая накапливается в ходе вычислений.

Рассмотрим накопление погрешности при простейших вычислениях. Пусть исходные данные x_i известны с относительной погрешностью $\Delta_i > 0$, т. е. заключены между $x_i(1-\Delta_i)$ и $x_i(1+\Delta_i)$; их абсолютные погрешности равны $\Delta_i|x_i|$. Тогда при сложении или вычитании двух чисел результат равен $x_1 \pm x_2$ с абсолютной погрешностью не более $\Delta_1|x_1|+\Delta_2|x_2|$, т. е. при этих операциях абсолютные погрешности складываются. При умножении (делении) результат равен x_1x_2 (x_1/x_2) с относительной погрешностью не более $\Delta_1+\Delta_2$, т. е. складываются относительные погрешности. На современных ЭВМ числа записываются с 10—12 десятичными знаками, поэтому в расчете на них погрешность единичного округления $\Delta = 10^{-10} \div 10^{-12}$ обычно пренебрежимо мала по сравнению с погрешностью метода и неустранимой погрешностью.

При решении больших задач выполняются миллиарды действий. Казалось бы, начальные ошибки возрастут в 10^9 раз и погрешность ответа будет огромной. Однако при отдельных действиях фактические погрешности чисел могут иметь разные знаки и компенсировать друг друга. Согласно статистике при N одинаковых действиях среднее значение суммарной ошибки превышает единичную примерно в $\sqrt{N}$ раз, а вероятность заметного уклонения суммарной ошибки от среднего значения очень мала. Значит, если нет систематических причин, то случайное накопление ошибок не слишком существенно.

Систематические причины возникают, например, если алгоритм таков, что в нем есть вычитание близких по величине чисел: хотя абсолютная ошибка при этом невелика, относительная ошибка $\Delta = (\Delta_1|x_1|+\Delta_2|x_2|)/(x_1-x_2)$ может стать большой. Например, при нахождении корней квадратного уравнения по

обычной формуле

$$x^2+px-q=0, \quad x_{1,2}=-0{,}5p\pm\sqrt{0{,}25p^2+q}$$

в случае, когда $0<q\ll p$, относительная ошибка округления для положительного корня x_1 велика. Это надо заранее предусмотреть и преобразовать формулу так, чтобы избавиться от подобных вычитаний:

$$x_1=q/(0{,}5p+\sqrt{0{,}25p^2+q}).$$

Этот пример очень прост. Существуют гораздо более сложные алгоритмы, где ошибки округления очень опасны: например, нахождение корней многочлена очень высокой степени (глава V, § 2, п.8) или итерационное решение разностных схем для эллиптических уравнений при помощи чебышевского набора параметров (глава XII, § 1). В этих случаях только после серьезного исследования удалось так видоизменить алгоритм, чтобы довести ошибки округления до безопасного уровня.

Отметим, что в большинстве подобных задач неприятностей можно избежать, проводя расчет с двойной или тройной точностью. Такая возможность реализована в хороших математических обеспечениях ЭВМ; это в несколько раз увеличивает время расчета, зато позволяет пользоваться уже известными алгоритмами, а не разрабатывать новые.

При любых расчетах справедливо правило: надо удерживать столько значащих цифр, чтобы погрешность округления была существенно меньше всех остальных погрешностей.

3. Корректность. Задача $y=A(x)$ называется *корректно поставленной*, если для любых входных данных x из некоторого класса решение y существует, единственно и устойчиво по входным данным. Рассмотрим это определение подробнее.

Чтобы численно решать задачу $y=A(x)$, надо быть уверенным в том, что искомое решение существует. Естественно также требовать единственности решения точной задачи: численный алгоритм — однозначная последовательность действий, и она может привести к одному решению. Но этого мало.

Нас интересует решение y, соответствующее входным данным x. Но реально мы имеем входные данные с погрешностью $x+\delta x$ и находим $y+\delta y=A(x+\delta x)$. Следовательно, неустранимая погрешность решения равна $\delta y=A(x+\delta x)-A(x)$. Если решение непрерывно зависит от входных данных, т. е. всегда $\|\delta y\|\to 0$ при $\|\delta x\|\to 0$, то задача называется *устойчивой* по входным данным; в противном случае задача неустойчива по входным данным.

Рассмотрим классический пример неустойчивости — задачу Коши для эллиптического уравнения в полуплоскости $y\geqslant 0$:

$$u_{xx}+u_{yy}=0, \quad u(x,\ 0)=0, \quad u_y(x,\ 0)=\varphi(x). \tag{8}$$

Входными данными является $\varphi(x)$. Если $\bar{\varphi}(x) = 0$, то задача имеет только тривиальное решение $\bar{u}(x, y) = 0$. Если же $\varphi_n(x) = \frac{1}{n}\cos nx$, то решением будет

$$u_n(x, y) = \frac{1}{n^2}\cos nx \cdot \operatorname{sh} ny.$$

Очевидно, $\varphi_n(x)$ равномерно сходятся к $\bar{\varphi}(x)$ при $n \to \infty$; но при этом если $y \neq 0$, то $u_n(x, y)$ неограничено и никак не может сходиться к $\bar{u}(x, y)$. Этот пример связан с физической задачей о тяжелой жидкости, налитой поверх легкой; при этом действительно возникает так называемая релей-тейлоровская неустойчивость.

Отсутствие устойчивости обычно означает, что даже сравнительно небольшой погрешности δx соответствует весьма большое δy, т. е. получаемое в расчете решение будет далеко от искомого. Непосредственно к такой задаче численные методы применять бессмысленно, ибо погрешности, неизбежно появляющиеся при численном расчете, будут катастрофически нарастать в ходе вычислений.

Правда, сейчас развиты методы решения многих некорректных задач. Но они основаны на решении не исходной задачи, а близкой к ней вспомогательной корректно поставленной задачи, содержащей параметр α; при $\alpha \to 0$ решение вспомогательной задачи должно стремиться к решению исходной задачи. Примеры таких методов (называемых регуляризацией) даны в следующих двух главах, а их строгое обоснование приведено в главе XIV, § 2.

На практике даже не всякую устойчивую задачу легко решить. Пусть $\|\delta y\| \leqslant C\|\delta x\|$, причем константа C очень велика. Задача формально устойчива, но фактическая неустранимая ошибка может быть большой. Этот случай называют *слабой* устойчивостью (или плохой обусловленностью). Примером является такая задача:

$$y''(x) = y(x), \tag{9а}$$

$$y(0) = 1, \quad y'(0) = -1. \tag{9б}$$

Общее решение дифференциального уравнения (9а) есть:

$$y(x) = 0{,}5\,[y(0) + y'(0)]\,e^x + 0{,}5\,[y(0) - y'(0)]\,e^{-x}.$$

Начальным условиям (9б) соответствует точное решение $y(x) = e^{-x}$; но небольшая погрешность начальных данных может привести к тому, что в решении добавится член вида εe^x, который при больших аргументах много больше искомого решения.

Очевидно, для хорошей практической устойчивости расчета константа C должна быть не слишком велика. Так, если начальные данные известны точно, т. е. могут быть заданы с точностью до ошибок округления $\Delta \sim 10^{-12}$, то необходимо, чтобы $C \ll 10^{12}$. Если же начальные данные найдены из эксперимента с точностью

$\delta x \sim 0{,}001$, а требуемая точность решения $\delta y \sim 0{,}1$, то допустимо $C \leqslant 100$.

Даже если задача устойчива, то численный алгоритм может быть неустойчивым. Например, если производные заменяются разностями, то приходится вычитать близкие числа и сильно теряется точность. Эти неточные промежуточные результаты используются в дальнейших вычислениях, и ошибки могут сильно нарастать.

По аналогии можно говорить о корректности алгоритма $\bar{y} = \bar{A}(\bar{x})$, подразумевая существование и единственность приближенного решения для любых входных данных $\bar{x}$ некоторого класса, и устойчивость относительно всех ошибок в исходных данных и промежуточных выкладках. Однако в общем случае этим определением трудно пользоваться; только в теории разностных схем (глава IX) оно применяется успешно.

ЗАДАЧИ

1. Доказать выполнимость всех соотношений (4). Рассмотреть, как меняется форма записи этих соотношений при задании функции на произвольном конечном отрезке $a \leqslant t \leqslant b$.

2. Доказать утверждения о согласованности и подчиненности норм матриц, приведенные в конце п. 1 § 2.

ГЛАВА II

АППРОКСИМАЦИЯ ФУНКЦИЙ

В главе II рассмотрены способы построения приближенных формул для заданной функции. В § 1 изложен способ интерполяции; он несложен и обеспечивает хорошую точность на небольших отрезках. В § 2 рассмотрена среднеквадратичная аппроксимация, частным случаем которой является метод наименьших квадратов; она позволяет строить приближенные формулы, пригодные на больших отрезках. В § 3 кратко изложены основные сведения о равномерной аппроксимации.

§ 1. Интерполирование

1. Приближенные формулы. Если задана функция $y(x)$, то это означает, что любому допустимому значению x сопоставлено значение y. Но нередко оказывается, что нахождение этого значения очень трудоемко. Например, $y(x)$ может быть определено как решение сложной задачи, в которой x играет роль параметра, или $y(x)$ измеряется в дорогостоящем эксперименте. При этом можно вычислить небольшую таблицу значений функции, но прямое нахождение функции при большом числе значений аргумента будет практически невозможно.

Функция $y(x)$ может участвовать в каких-либо физико-технических или чисто математических расчетах, где ее приходится многократно вычислять. В этом случае выгодно заменить функцию $y(x)$ приближенной формулой, т. е. подобрать некоторую функцию $\varphi(x)$, которая близка в некотором смысле к $y(x)$ и просто вычисляется. Затем при всех значениях аргумента полагают $y(x) \approx \approx \varphi(x)$. Близость получают введением в аппроксимирующую функцию свободных параметров $\boldsymbol{a} = \{a_1, a_2, \ldots, a_n\}$ и соответствующим их выбором.

Подбор удачного вида функциональной зависимости $\varphi(x; \boldsymbol{a})$ — искусство; некоторые советы по этому поводу будут даны в § 1, п. 8. А определение наилучших (в требуемом смысле) параметров формулы делается стандартными методами, которые и будут рассмотрены в этой главе.

2. Линейная интерполяция. Пусть функция $y(x)$ известна только в узлах некоторой сетки x_i, т. е. задана таблицей. Если

потребовать, чтобы $\varphi(x;\ \boldsymbol{a})$ совпадала с табличными значениями в n выбранных узлах сетки, то получим систему

$$\varphi(x_i;\ a_1,\ a_2,\ \ldots,\ a_n) = y(x_i) \equiv y_i, \qquad 1 \leqslant i \leqslant n, \tag{1}$$

из которой можно определить параметры a_k. Этот способ подбора параметров называется *интерполяцией* (точнее, *лагранжевой* интерполяцией). По числу используемых узлов сетки будем называть интерполяцию одноточечной, двухточечной и т. д.

Если $\varphi(x;\ \boldsymbol{a})$ нелинейно зависит от параметров, то интерполяцию назовем *нелинейной*; в этом случае нахождение параметров из системы (1) может быть трудной задачей. Сейчас мы рассмотрим *линейную* интерполяцию, когда $\varphi(x;\ \boldsymbol{a})$ линейно зависит от параметров, т. е. представима в виде так называемого *обобщенного многочлена*

$$\varphi(x;\ a_1,\ a_2,\ \ldots,\ a_n) = \sum_{k=1}^{n} a_k\ \varphi_k(x). \tag{2}$$

Очевидно, функции $\varphi_k(x)$ можно считать линейно-независимыми, иначе число членов в сумме и параметров можно было бы уменьшить. На систему функций $\varphi_k(x)$ надо наложить еще одно ограничение. Подставляя (2) в (1), получим для определения параметров a_k следующую систему линейных уравнений:

$$\sum_{k=1}^{n} a_k \varphi_k(x_i) = y_i, \qquad 1 \leqslant i \leqslant n. \tag{3}$$

Чтобы задача интерполяции всегда имела единственное решение, надо, чтобы при любом расположении узлов (лишь бы среди них не было совпадающих) определитель системы (3) был бы отличен от нуля:

$$\Delta \equiv \operatorname{Det}\{\varphi_k(x_i)\} = \begin{vmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \ldots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \ldots & \varphi_n(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_1(x_n) & \varphi_2(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} \neq 0 \quad \text{при} \quad x_i \neq x_j. \tag{4}$$

Система функций, удовлетворяющих требованию (4), называется *чебышевской*. Таким образом, при линейной интерполяции надо строить обобщенный многочлен по какой-нибудь чебышевской системе функций.

Для линейной интерполяции наиболее удобны обычные многочлены, ибо они легко вычисляются и на клавишной машине и на ЭВМ. Другие системы функций сейчас почти не употребляются, хотя в теории подробно рассматривают интерполяцию тригонометрическими многочленами и экспонентами. Поэтому мы не приводим выражения обобщенного многочлена (2) через табулированные значения функции y_i; вывести это выражение несложно.

3. Интерполяционный многочлен Ньютона. Рассмотрим систему $\varphi_k(x) = x^k$, $0 \leqslant k \leqslant n$; для удобства узлы интерполяции также перенумеруем с нулевого по n-й. Легко заметить, что определитель (4) в этом случае есть определитель Вандермонда

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = \prod_{n \geqslant k > m \geqslant 0} (x_k - x_m). \tag{5}$$

Следовательно, алгебраический интерполяционный многочлен $\mathcal{P}_n(x)$ всегда существует и единствен (с точностью до формы записи). Применим для его вывода следующий прием.

Определим *разделенные разности* табулированной функции $y(x)$ при помощи соотношений

$$\begin{aligned} y(x_i, x_j) &= [y(x_i) - y(x_j)] / (x_i - x_j), \\ y(x_i, x_j, x_k) &= [y(x_i, x_j) - y(x_j, x_k)] / (x_i - x_k), \end{aligned} \tag{6}$$

и т. д. Разделенные разности первого, второго и более высоких порядков имеют размерности производных соответствующих порядков; в главе III показано, что они дают приближенные значения производных. Разделенные разности любого порядка можно выразить непосредственно через узловые значения функции, но вычислять их удобнее по рекуррентному соотношению (6).

Пусть $\mathcal{P}(x)$ есть многочлен степени n. Рассмотрим, что представляют собой его разделенные разности. Вычитая из него константу $\mathcal{P}(x_0)$, получим многочлен $\mathcal{P}(x) - \mathcal{P}(x_0)$, который обращается в нуль при $x = x_0$ и поэтому делится нацело на $x - x_0$. Следовательно, первая разделенная разность многочлена n-й степени $\mathcal{P}(x, x_0) = [\mathcal{P}(x) - \mathcal{P}(x_0)] / (x - x_0)$ есть многочлен степени $n - 1$ относительно x и в силу симметрии выражения — относительно x_0. Аналогично, вторая разность $\mathcal{P}(x, x_0, x_1)$ есть многочлен степени $n - 2$; в самом деле, из (6) видно, что числитель этой разности обращается в нуль при $x = x_1$, и значит, нацело делится на $x - x_1$, а степень при этом понижается на единицу. Продолжая эти рассуждения, можно показать, что разность $\mathcal{P}(x, x_0, x_1, \dots, x_{n-1})$ есть многочлен нулевой степени, т. е. константа, а более высокие разделенные разности тождественно равны нулю.

Перепишем соотношения (6) для случая, когда функция есть многочлен и первый аргумент равен x:

$$\begin{aligned} \mathcal{P}(x) &= \mathcal{P}(x_0) + (x - x_0)\, \mathcal{P}(x, x_0), \\ \mathcal{P}(x, x_0) &= \mathcal{P}(x_0, x_1) + (x - x_1)\, \mathcal{P}(x, x_0, x_1) \end{aligned}$$

и т. д. Эта цепочка соотношений конечна, ибо $(n+1)$-я разделенная

разность многочлена тождественно равна нулю. Последовательно подставляя эти соотношения друг в друга, получим формулу

$$\mathcal{P}(x) = \mathcal{P}(x_0) + (x-x_0)\,\mathcal{P}(x_0, x_1) + \\ + (x-x_0)(x-x_1)\,\mathcal{P}(x_0, x_1, x_2) + \dots \\ \dots + (x-x_0)(x-x_1)\dots(x-x_{n-1})\,\mathcal{P}(x_0, x_1, \dots, x_n), \quad (7)$$

по которой многочлен n-й степени выражается при помощи разделенных разностей через свои значения в узлах $x_0, \dots, x_n$. Но значения интерполяционного многочлена в этих узлах по определению совпадают со значениями искомой функции, и поэтому разделенные разности $y(x)$ и $\mathcal{P}(x)$ тоже совпадают. Подставляя в (7) разделенные разности искомой функции и заменяя точное равенство на приближенное, получим интерполяционную формулу Ньютона

$$y(x) \approx y(x_0) + \sum_{k=1}^{n} (x-x_0)(x-x_1)\dots(x-x_{k-1})\,y(x_0, x_1, \dots, x_k). \quad (8)$$

Формула Ньютона удобна для вычислений и на ЭВМ, и на клавишной машине. Легко составить следующую таблицу 1 разделенных разностей для табулированной функции $y(x)$ и произвести вычисления по формуле (8).

Таблица 1

x_0	$y(x_0)$			
		$y(x_0, x_1)$		
x_1	$y(x_1)$		$y(x_0, x_1, x_2)$	
		$y(x_1, x_2)$		$y(x_0, x_1, x_2, x_3)$
x_2	$y(x_2)$		$y(x_1, x_2, x_3)$	
		$y(x_2, x_3)$		
x_3	$y(x_3)$			

Замечание 1. За точностью расчета удобно следить, визуально оценивая скорость убывания членов суммы (8). Если они убывают медленно, то на хорошую точность рассчитывать, вообще говоря, нельзя (подробнее см. пп. 6, 7). Если убывание быстрое, то оставляют только те члены, которые больше допустимой погрешности; тем самым определяют, сколько узлов требуется подключить в расчет.

Пример. Покажем, как вычислять синус в первом квадранте, используя четыре известных значения. Составим таблицу 2 с четырьмя узлами, причем для удобства вычисления положим $y(x) = \sin(30^\circ \cdot x)$. Для проверки точности, используя разности верхней косой строки, вычислим

$$y(1{,}5) \approx 0 + 0{,}750 - 0{,}050 + 0{,}006 = 0{,}706.$$

Таблица 2

x_i	$y(x_i)$	$y(x_i, x_{i+1})$	$y(x_i, x_{i+1}, x_{i+2})$	$y(x_i, \ldots, x_{i+3})$
0	0,000	0,500		
1	0,500		—0,067	
		0,366		—0,016
2	0,866		—0,116	
		0,134		
3	1,000			

Это приближенное значение мало отличается от точного значения $y(1,5) = \sin 45° \approx 0{,}707$. Таким образом, достаточно помнить только верхнюю косую строку таблицы 2, чтобы вычислять синус с точностью около 0,001.

Замечание 2. При заданном числе узлов многочлен Ньютона удобнее вычислять по схеме Горнера, записывая его в виде

$$y(x) = y(x_0) + (x - x_0)[y(x_0, x_1) + (x - x_1)[y(x_0, x_1, x_2) + \ldots]].$$

Но если надо контролировать точность расчета и определять нужное число узлов, то удобнее форма (8).

Замечание 3. Для расчетов по формуле Ньютона безразличен порядок, в котором перенумерованы узлы интерполяции; это полезно помнить при подключении новых узлов в расчет.

Мы ограничились здесь общими формулами, пригодными для таблиц с переменным шагом. Во многих учебниках для таблиц с постоянным шагом вводят *конечные разности* $\Delta^n y$, связанные с разделенными разностями соотношением $\Delta^n y = n! y(x_0, x_1, \ldots, x_n)$. Но это дань историческим традициям, ибо разделенные разности не менее удобны в расчетах, чем конечные.

Есть много разных форм записи интерполяционного многочлена общего вида: Ньютона, Лагранжа, Гаусса, Грегори — Ньютона, Лапласа — Эверетта и другие. Наиболее удобной для вычислений с контролем точности и на ЭВМ и вручную является форма Ньютона (8). Большинство остальных форм рассчитано на определенные частные случаи расположения узлов интерполяции, но те выгоды, которые при этом получаются, обычно несущественны при расчетах на ЭВМ.

4. Погрешность многочлена Ньютона. Выше мы рассмотрели эмпирическое правило определения погрешности интерполяции по убыванию членов суммы (8). Проведем теперь строгое исследование погрешности метода, проистекающей от замены искомой функции интерполяционным многочленом Ньютона.

Погрешность удобно представить в следующем виде:

$$y(x) - \mathscr{P}_n(x) = \omega_n(x)\, r(x), \qquad \omega_n(x) = \prod_{i=0}^{n} (x - x_i), \tag{9}$$

ибо эта погрешность заведомо равна нулю во всех узлах интерполяции. Введем вспомогательную функцию $q(\xi) = y(\xi) - \mathscr{P}_n(\xi) - \omega_n(\xi)\, r(x)$, где x играет роль параметра и принимает любое фиксированное значение. Очевидно, $q(\xi) = 0$ при $\xi = x_0, x_1, \ldots, x_n$ и при $\xi = x$, т. е. обращается в нуль в $n+2$ точках.

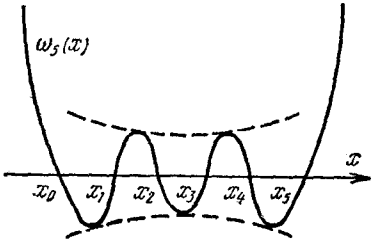


Рис. 2.

Предположим, что $y(x)$ имеет $n+1$ непрерывную производную; тогда то же справедливо для $q(\xi)$. Между двумя нулями гладкой функции лежит нуль ее производной. Последовательно применяя это правило, получим, что между крайними из $n+2$ нулей функции лежит нуль $n+1$-й производной. Но $q^{(n+1)}(\xi) = y^{(n+1)}(\xi) - (n+1)!\, r(x)$, и если в какой-то точке ξ^*, лежащей между указанными выше нулями, она обращается в нуль, то $r(x) = y^{(n+1)}(\xi^*)/(n+1)!$. Заменяя погрешность (9) максимально возможной, получаем оценку погрешности:

$$|y(x) - \mathscr{P}_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |\omega_n(x)|, \qquad M_{n+1} = \max |y^{(n+1)}(\xi)|, \tag{10}$$

где максимум производной берется по отрезку между наименьшим и наибольшим из значений $x, x_0, x_1, \ldots, x_n$.

Оценить $\omega_n(x)$ при произвольном расположении узлов интерполяции сложно. Однако таблицы чаще всего имеют постоянный шаг $h = x_{i+1} - x_i$, а узлы интерполяции берутся из таблицы подряд. Тогда $\omega_n(x)$ имеет примерно такой вид, как показано на рис. 2 для $n = 5$: вблизи центрального узла интерполяции экстремумы невелики, вблизи крайних узлов — несколько больше, а если x выходит за крайние узлы интерполяции, то $\omega_n(x)$ быстро возрастает.

Можно подобрать узлы интерполяции так, чтобы на заданном отрезке $\max |\omega_n(x)|$ был меньше, чем у любого другого многочлена той же степени. Для этого $\omega_n(x)$ должен быть многочленом Чебышева первого рода (см. Приложение). Узлы этого многочлена расположены сравнительно редко в середине рассматриваемого отрезка и сгущаются у его концов. Но вне выбранного отрезка многочлен $\omega_n(x)$ по-прежнему будет быстро возрастать. Этот способ интерполяции довольно громоздок, а выигрыш в точности невелик; поэтому его используют только для специальных целей — например, при построении аппроксимирующих формул.

Термин *интерполяция* в узком смысле употребляют, если x заключено между крайними узлами интерполяции; если он выходит из этих пределов, то говорят об *экстраполяции*. Очевидно, что при экстраполяции далеко за крайний узел ошибка может быть велика, поэтому экстраполяция мало надежна. На практике рекомендуется пользоваться преимущественно интерполяцией.

При интерполяции на равномерной сетке выгодно выбирать из таблицы узлы так, чтобы искомая точка x попадала ближе к центру этой конфигурации узлов — это обеспечит более высокую точность. Для упрощения вычислений рассмотрим случай нечетного $n=2k+1$. Из симметрии полинома $\omega_n(x)$ очевидно, что в центральном интервале экстремум достигается точно в середине (см. рис. 2). Этот экстремум равен

$$\left[\frac{h}{2}\cdot\frac{3h}{2}\cdot\frac{5h}{2}\cdot\ldots\cdot\frac{(2k+1)\,h}{2}\right]^2=\left[\frac{(2k+1)!\,h^{k+1}}{k!2^{2k+1}}\right]^2.$$

Подставим эту величину в оценку (10). После несложных преобразований с использованием формулы Стирлинга $p!\approx\sqrt{2\pi p}\,(p/e)^p$ получим оценку ошибки в центральном интервале

$$|y(x)-\mathscr{P}_n(x)|<\sqrt{2/\pi n}\,M_{n+1}(h/2)^{n+1}. \tag{11}$$

Если величины производных $y(x)$ можно оценить, то отсюда легко определить число узлов, достаточное для получения заданной точности.

Из оценки (11) видно, что если перейти от таблиц с крупным шагом к таблицам с более мелким шагом, то погрешность метода будет убывать, как h^{n+1}. Поэтому говорят, что многочлен Ньютона $\mathscr{P}_n(x)$ имеет погрешность $O(h^{n+1})$ и обеспечивает $n+1$-й порядок точности интерполяции.

В главе III мы увидим, что между разделенными разностями и производными соответствующих порядков существует соотношение $y^{(n)}(x)\approx n!\,y(x_0, x_1, \ldots, x_n)$. Если учесть это при определении величины членов суммы (8), то нетрудно заметить, что эмпирическая оценка погрешности по первому отброшенному члену близка к оценке (10), хотя является менее строгой. Оценки (10) и (11) можно провести до вычисления интерполяционного многочлена, т. е. это *априорные* оценки точности. Оценка же по первому отброшенному члену делается после выполнения вычислений, т. е. является *апостериорной*. Поскольку обычно величины производных искомой функции заранее неизвестны, а в ходе вычисления многочлена Ньютона они фактически определяются, то на практике удобнее пользоваться апостериорной оценкой.

Далее мы не раз сможем убедиться, что строгие априорные оценки используются в основном при теоретическом исследовании методов. При практическом контроле точности расчетов обычно

употребляют менее строгие (хотя тоже имеющие теоретическое обоснование), но более удобные апостериорные оценки.

5. Применения интерполяции. Интерполяция применяется во многих задачах, а не только для вычисления табулированной функции при любых значениях аргумента.

При помощи разделенных разностей контролируется точность таблиц. Для этого составляют таблицы разделенных разностей различных порядков для соседних узлов и анализируют их поведение.

Например, в таблице 3 приведена зависимость коэффициента теплопроводности высокотемпературной фазы циркония от температуры. Там же вычислены первая и вторая конечные разности. Видно, что вторая разность меняется беспорядочно, так что интерполировать более чем по двум точкам бессмысленно. По величине второй разности можно сказать, что случайная погрешность λ составляет около единицы третьего знака в большинстве точек, но в двух первых она может доходить до единицы второго знака (для систематической погрешности измерений эти соображения неприменимы).

Таблица 3

Теплопроводность циркония

$T°$, К	$\lambda \cdot 10^4$, кал/см·г·сек	$\Delta_1\lambda \cdot 10^6$	$\Delta_2\lambda \cdot 10^6$
1200	561		
		79	
1300	640		−24
		55	
1400	695		−21
		21	
1500	716		−2
		19	
1600	735		−2
		17	
1700	752		+2
		19	
1800	771		−2
		17	
1900	788		−3
		14	
2000	802		+5
		19	
2100	821		

Подобный контроль полезен при анализе результатов измерений в физике и технике.

Интерполяцию применяют для *субтабулирования* — сгущения таблиц. Алгоритмы непосредственного вычисления многих функ-

ций очень сложны. Поэтому при математическом табулировании обычно функцию непосредственно вычисляют в небольшом числе узлов, т. е. таблицы имеют крупный шаг. Затем при помощи интерполяции высокого порядка точности сетку узлов сгущают и составляют подробную таблицу. Шаг этой таблицы выбирают таким, чтобы простейшая интерполяция (двухточечная) обеспечивала требуемую точность.

В связи с этим отметим, что при ручных расчетах выгодны подробные таблицы, ибо они допускают применение простейших способов интерполяции, легко выполняемых на бумаге или клавишных машинах, а время поиска нужных узлов интерполяции невелико по сравнению со временем выполнения алгебраических действий. Наоборот, при расчетах на ЭВМ задание подробных таблиц невыгодно, поскольку они занимают много места в оперативной памяти, а время поиска становится много больше времени выполнения алгебраических действий; выгоднее таблицы с большим шагом, хотя при этом требуются более сложные и точные способы интерполяции.

Задачей *обратного интерполирования* называют нахождение x для произвольного y, если задана таблица $y_i = y(x_i)$. Для монотонных функций между прямым и обратным интерполированием нет разницы: можно читать таблицу наоборот, как задание $x_i = x(y_i)$. Единственное отличие будет в том, что «обратная» таблица $x(y_i)$ будет иметь переменный шаг, даже если «прямая» таблица имела постоянный. Но все наши формулы рассчитаны на переменный шаг. Отметим, что для достижения заданной точности прямая и обратная интерполяции требуют, вообще говоря, разного числа узлов.

Важный пример обратного интерполирования — решение уравнения $y(x) = 0$. Вычислим несколько значений функции $y(x_i)$, т. е. составим небольшую таблицу. Запишем ее в виде $x_i = x(y_i)$ и при помощи интерполяции найдем приближенное значение $x(0)$. Этот способ дает хорошие результаты, если функция достаточно гладкая, а корень лежит между рассчитанными узлами. Если корень расположен далеко от узлов, то способ ненадежен, ибо применяется экстраполяция.

Пример. Решим уравнение

$$y(x) \equiv (1+x)\, e^{0{,}5x} - 2{,}5 = 0. \qquad (12)$$

Составим таблицу 4 значений функции; первым запишем столбец значений y, ибо в дальнейших вычислениях эта величина будет аргументом. Найдем разделенные разности и произведем вычисления по верхней косой строке:

$$x(0) \approx x_0 + (0 - y_0)\, x(y_0,\ y_1) + (0 - y_0)(0 - y_1)\, x(y_0,\ y_1,\ y_2) = 0{,}744.$$

Точное решение есть $x(0) = 0{,}732$, так что ошибка получилась

небольшой. Для повышения точности в этом способе целесообразно взять новые узлы, близко расположенные к грубо найденному корню, а не увеличивать число узлов.

Таблица 4

y_i	x_i	$x(y_i, y_{i+1})$	$x(y_0, y_1, y_2)$
−1,500	0	0,540	−0,076
−0,574	0,5	0,365	
0,797	1,0		

В этом курсе будут рассмотрены и другие примеры применения интерполирования.

6. Интерполяционный многочлен Эрмита. Пусть табулирована не только функция, но и ее производные вплоть до некоторого порядка. Тогда можно потребовать, чтобы в узлах интерполяции совпадали не только значения искомой функции $y(x)$ и интерполяционной функции $\varphi(x)$, но и значения их производных вплоть до некоторого порядка. Такую интерполяцию будем называть *эрмитовой*; если $\varphi(x)$ — алгебраический многочлен n-й степени, то он называется интерполяционным многочленом Эрмита и обозначается $\mathscr{H}_n(x)$.

Покажем, как построить этот многочлен. По $n+1$ узлу построим интерполяционный многочлен Ньютона $\mathscr{P}_n(x; x_0, x_1, \ldots, x_n)$. Поскольку значения функции $y(x)$ и многочлена в узлах совпадают, то их средние наклоны на участках между узлами равны. Мысленно будем приближать узел x_n к узлу x_{n-1}; при этом средний наклон будет стремиться к производной. Значит, после совпадения узлов получим многочлен, который в узле x_{n-1} правильно передает не только значение функции, но и значение первой производной. Символически обозначим его как $\mathscr{P}_n(x; x_0, x_1, \ldots, x_{n-1}, x_{n-1})$ *).

Слияние трех узлов в один обеспечивает передачу не только наклона, но и кривизны, т. е. первой и второй производных и т. д. Таким образом, многочлен

$$\mathscr{H}_n(x) = \mathscr{P}_n(x; \underbrace{x_0, x_0, \ldots, x_0}_{m_0}, \underbrace{x_1, x_1, \ldots, x_1}_{m_1}, \underbrace{x_p, x_p, \ldots, x_p}_{m_p}), \qquad (13)$$

$$\sum_{k=0}^{p} m_k = n+1,$$

*) Чтобы отличать его обозначение от разделенной разности, мы отделяем аргумент от узлов, по которым составлен многочлен, точкой с запятой.

в узле x_k правильно передает значение функции и ее производных вплоть до порядка $m_k - 1$ и имеет минимально необходимую для этого степень. Оценка погрешности метода (10) в этом случае принимает следующий вид:

$$|y(x) - \mathscr{H}_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |\Omega_n(x)|, \qquad \Omega_n(x) = \prod_{k=0}^{p} (x - x_k)^{m_k}. \tag{14}$$

Очевидно, если сетка имеет шаг h, а точка x лежит между крайними узлами интерполяции, то $\Omega_n(x) = O(h^{n+1})$; следовательно, порядок точности эрмитовой интерполяции равен $n+1$, т. е. числу коэффициентов интерполяционного многочлена.

Заметим, что обычный многочлен Ньютона с таким же числом коэффициентов (т. е. той же степени) также имеет погрешность $O(h^{n+1})$. Однако на одной и той же сетке численная величина погрешности многочлена Ньютона будет больше, чем у многочлена Эрмита: его вспомогательный многочлен $\omega_n(x)$ содержит больше узлов, чем $\Omega_n(x)$, и поэтому в него входят бóльшие сомножители. Очевидно также, что чем более высокие производные используются при построении интерполяционного многочлена Эрмита заданной степени, тем меньше требуемое число узлов, и тем меньше будет численная величина его погрешности (хотя порядок точности остается одним и тем же).

Выражением (13) нельзя пользоваться буквально. Если формально подставить в формулу Ньютона (8) совпадающие узлы, то потребуется вычислить разделенные разности, у которых некоторые узлы являются кратными. Выражения (6) для таких разностей содержат неопределенность типа 0/0. Если кратность каждого узла не больше чем двойная, то эту неопределенность можно раскрыть с помощью предельного перехода, например,

$$\begin{aligned} y(x_0, x_0) &= \lim_{x_1 \to x_0} \frac{y(x_0) - y(x_1)}{x_0 - x_1} = y'(x_0), \\ y(x_0, x_0, x_1) &= \frac{1}{x_0 - x_1} [y'(x_0) - y(x_0, x_1)], \\ y(x_0, x_0, x_1, x_1) &= \frac{1}{(x_0 - x_1)^2} [y'(x_0) - 2y(x_0, x_1) + y'(x_1)]. \end{aligned} \tag{15}$$

Если узлы имеют более высокую кратность, то удобнее дифференцировать формулу Ньютона (8). Например, если ее продифференцировать $m-1$ раз, то обратятся в нуль все члены, содержащие разделенные разности порядка меньше $m-1$. Затем положим $x = x_0 = x_1 = \ldots$; тогда обратятся в нуль множители перед разделенными разностями порядка больше $m-1$, и мы получим

$$y(\underbrace{x_0, x_0, \ldots, x_0}_{m}) = \frac{1}{(m-1)!} y^{(m-1)}(x_0). \tag{16}$$

Но узлы более чем двойной кратности почти не встречаются в практике вычислений, ибо вторые и более высокие производные искомой функции редко табулируются.

Рассмотрим наиболее употребительные частные случаи интерполяционного многочлена Эрмита.

Первый случай — многочлен, который в одном узле x_0 совпадает с функцией и всеми ее заданными производными:

$$\mathscr{T}_n(x;\ x_0,\ x_0,\ \ldots) = \\ = y(x_0) + (x-x_0)\,y'(x_0) + \frac{1}{2}(x-x_0)^2\,y''(x_0) + \ldots \qquad (17)$$

Очевидно, это отрезок ряда Тейлора; в этом случае $\Omega_n(x) = (x-x_0)^{n+1}$, и оценка (11) переходит в известную оценку точности ряда Тейлора.

Второй случай — многочлен, передающий в двух узлах значения функции и ее первой производной:

$$\mathscr{T}_n(x;\ x_0,\ x_0,\ x_1,\ x_1) = y(x_0) + (x-x_0)\{y'(x_0) + \\ + (x-x_0)[y(x_0,\ x_0,\ x_1) + (x-x_1)\,y(x_0,\ x_0,\ x_1,\ x_1)]\}; \qquad (18)$$

разделенные разности сюда надо подставить из соотношения (15). Функция $\Omega_n(x) = (x-x_0)^2(x-x_1)^2$ внутри интервала интерполирования не превышает $(h/2)^4$, так что погрешность формулы (18) не более $0{,}026M_4h^4$; эта формула имеет четвертый порядок точности.

Для сравнения приведем без вывода общее выражение интерполяционного многочлена Эрмита

$$\mathscr{H}_n(x) = \sum_{k=0}^{p}\sum_{m=0}^{\alpha_k-1}\sum_{q=0}^{\alpha_k-1-m}\frac{y^{(m)}(x_k)}{m!q!}\left\{(x-x_k)^{m+q}\times\right. \\ \left.\times\prod_{i\neq k}(x-x_i)^{\alpha_i}\right\}\left\{\frac{d^q}{dx^q}\prod_{j\neq k}(x-x_j)^{-\alpha_j}\right\}_{x=x_k}.$$

Оно настолько громоздко, что пользоваться им для вычислений практически невозможно. Если все $\alpha_i = 1$, то обе внутренние суммы превращаются в одно слагаемое с $m = q = 0$, и многочлен Эрмита переходит в многочлен Ньютона в форме Лагранжа. Если все $\alpha_i = 2$, то получим

$$\mathscr{H}_n(x) = \sum_{k=0}^{p}\left\{(x-x_k)\,y_k' + \left(1 - 2\sum_{\substack{i=0\\ i\neq k}}^{p}\frac{x-x_k}{x_k-x_i}\right)y_k\right\}\prod_{\substack{j=0\\ j\neq k}}^{p}\left(\frac{x-x_j}{x_k-x_j}\right)^2;$$

можно проверить, что в случае двух узлов последнее выражение совпадает с (18) с точностью до формы записи. Но даже и это выражение оказывается очень громоздким.

Такая ситуация довольно часто встречается в прикладной математике. Общие формулы, рассчитанные на все случаи жизни, нередко оказываются настолько сложными, что их не применяют

ни в одном конкретном случае. К тому же, в практических расчетах, как мы увидим далее, нецелесообразно использовать многочлены высоких степеней, поэтому в общих формулах нет серьезной необходимости. Трудоемкость же вычислений часто оказывается существенно меньшей при применении рекуррентных процедур типа формулы разделенных разностей (6).

7. Сходимость интерполяции. При каких условиях погрешность метода стремится к нулю, т. е. когда и как интерполяционный многочлен сходится к $y(x)$? На практике мы имеем два способа перехода к пределу. Первый состоит в том, чтобы, сохраняя степень интерполяционного многочлена, уменьшить шаг сетки, т. е. воспользоваться более подробными таблицами. Второй — сохраняя шаг сетки, увеличивать число используемых узлов, т. е. увеличивать степень многочлена.

Уменьшение шага. Если $y(x)$ имеет непрерывные производные вплоть до $n+1$-й, то при интерполяции многочленом $\mathcal{P}_m(x)$ степени $m \leqslant n$ погрешность метода есть $O(h^{m+1})$. В этом случае при фиксированной степени многочлена и уменьшении шага сетки погрешность $|y(x) - \mathcal{P}_m(x)|$ неограниченно убывает. Если ограничена производная, входящая в оценку ошибки, то интерполяционный многочлен равномерно сходится к $y(x)$ на ограниченном отрезке $a \leqslant x \leqslant b$.

Строго говоря, для каждого значения x выбирают свои узлы интерполяции, ближайшие (на данной сетке) к точке x, т. е. составляют свой многочлен $\mathcal{P}_m(x)$. При этом точка x заведомо лежит между крайними узлами интерполяции, используемыми в данном многочлене. Поэтому входящий в оценку погрешности (10) полином $\omega_m(x)$ ограничен равномерно по x: $|\omega_m(x)| < \max_i |x - x_i|^{m+1} \leqslant (mh)^{m+1}$, где h — шаг сетки (для неравномерных сеток — максимальный шаг). Для заданной точности ε определим шаг сетки из условия $M_{m+1}(mh)^{m+1} \leqslant \varepsilon \cdot (m+1)!$, где $M_{m+1} = \max_{[a,\,b]} |y^{(m+1)}(x)|$. Тогда для всех сеток с данным и более мелким шагом и любой точки отрезка $a \leqslant x \leqslant b$ погрешность интерполяционного многочлена $\mathcal{P}_m(x)$, узлы которого выбраны указанным выше образом, будет не более ε.

Аналогичные утверждения справедливы для интерполяционного многочлена Эрмита.

Увеличение числа узлов. Увеличивать степень интерполяционного многочлена далеко не всегда целесообразно. Во-первых, неизвестно, как быстро растет максимум производной M_m с увеличением ее порядка. Во-вторых, у функции может быть лишь конечное число производных. Рассмотрим интерполяцию на отрезке $a \leqslant x \leqslant b$, когда число узлов, используемых для построения интерполяционного многочлена, неограниченно возрастает.

Известно, что если $y(x)$ — целая функция, то при произвольном расположении узлов на $[a, b]$ многочлен $\mathscr{P}_n(x)$ равномерно сходится к $y(x)$ при $n \to \infty$. Но целая функция — это функция, разложимая в степенной ряд с бесконечным радиусом сходимости. Гораздо чаще приходится импонировать не целые функции, так что практическая ценность этого утверждения невелика.

Если же на $[a, b]$ функция имеет непрерывные производные сколь угодно высоких порядков, то это не гарантирует сходимости при произвольном расположении узлов. Например, возьмем функцию

$$y(x) = 0 \text{ при } -1 \leqslant x \leqslant 0, \quad y(x) = e^{-1/x} \text{ при } 0 < x \leqslant 1.$$

Ее график приведен на рис. 3. Все производные этой функции на $[-1, +1]$ непрерывны. Но если разместить все узлы интерполяции левее точки $x = 0$, то, очевидно, $\mathscr{P}_n(x) \equiv 0$, и никакой сходимости быть не может.

Правда, в этом примере расположение узлов было грубо неравномерным. Но равномерное

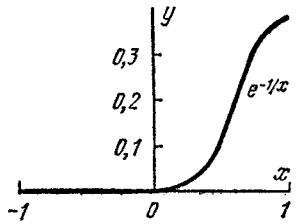


Рис. 3.

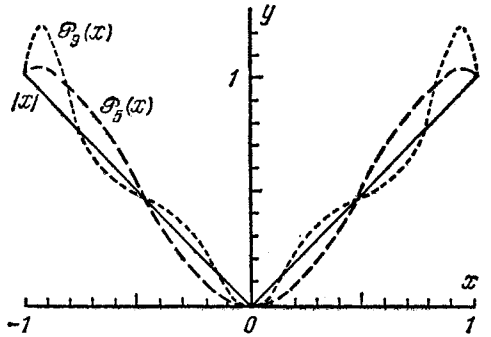


Рис. 4.

расположение не всегда спасает. С. Н. Бернштейн в 1916 г. доказал, что для функции $y(x) = |x|$ на отрезке $[-1, +1]$, покрытом равномерной сеткой узлов, значения $\mathscr{P}_n(x)$ между узлами интерполяции неограниченно возрастают при $n \to \infty$. Это иллюстрируется рис. 4, где даны графики функции и двух многочленов разных степеней.

Более того, для любой наперед заданной системы узлов можно найти такую непрерывную функцию, что построенные по этим узлам и функции многочлены Ньютона не будут равномерно сходиться.

Но сходимости в среднем для многочленов Ньютона всегда можно добиться следующим несложным выбором узлов. Пусть $\Phi_n(x)$ — система многочленов, ортогональных с весом $\rho(x)$ на отрезке $[a, b]$, и $x_m^{(n)}$ — нули этих многочленов. Используем эти точки в качестве узлов интерполяций; тогда

$$\int_a^b [\mathscr{P}_n(x) - y(x)]^2 \rho(x)\, dx \to 0$$

при $n \to \infty$ для любой непрерывной функции.

Для многочленов Эрмита получены более сильные результаты. Пусть функция $y(x)$ непрерывна на $[-1, +1]$; возьмем в качестве узлов нули многочленов Чебышева первого рода $T_n(x)$ (см. Приложение); фиксируем в этих узлах значения функции, а вместо ее производной возьмем любые числа c_{in}, удовлетворяющие условию $\lim\limits_{n \to \infty} \max\limits_i |c_{in} \ln n/n| = 0$. Построенный по всем этим зна-

чениям многочлен $\mathscr{H}_{2n-1}(x)$ равномерно сходится к $y(x)$ при $n \to \infty$. Очевидно, если $y(x)$ имеет ограниченную производную, то в качестве c_{in} можно брать значение производной в узлах.

Но и для многочленов Эрмита неудачный выбор узлов может испортить сходимость. Например, ряд Тейлора (17) расходится, если $|x - x_0|$ больше расстояния от x_0 до ближайшей особой точки в комплексной плоскости.

Выводы. На практике интерполировать многочленом высокой степени нежелательно. Если 3—5 узлов (точнее, свободных параметров) не обеспечивают требуемой точности, то обычно надо не увеличивать число узлов, а уменьшать шаг таблицы.

8. Нелинейная интерполяция. Полиномиальная интерполяция по оценке (11) имеет погрешность $\sim M_{n+1}(h/2)^{n+1}$, и при повышении порядка точности формулы на единицу погрешность меняется примерно в $hM_{n+2}/2M_{n+1}$ раз. Если шаг достаточно мал, то погрешность при этом уменьшается. Но если шаг велик, или производные быстро растут с увеличением порядка, то погрешность может увеличиваться при увеличении порядка точности формулы. С этим часто приходится сталкиваться при работе с быстро меняющимися функциями.

Таблица 5

x_i	$y(x_i)$	$y(x_i, x_{i+1})$		
0	1			
		10		
1	11		50	
		110		170
2	121		560	
		1230		
3	1351			

Пример 1. Пусть требуется найти значение $y(0{,}5)$, если функция задана таблицей 5 (в ней выписаны не только значения функции, но и разделенные разности). Используя интерполяционный многочлен Ньютона и ведя вычисления по верхней строке таблицы 5, запишем последовательно члены все более высоких порядков:

$$\varphi(0{,}5) = 1 + 5 - 12{,}5 + 63{,}75 - \ldots$$

Этот ряд содержит быстро возрастающие члены и совсем не похож на сходящийся; поэтому вычислить функцию с его помощью не удается. Функция слишком быстро меняется или, что то же самое, шаг сетки слишком велик для данной функции (рис. 5, *а*).

Как интерполировать такие функции, если более подробных таблиц нет? Универсального рецепта, пригодного для любой функции, не существует. Однако для конкретной функции нередко удается найти свой способ интерполяции, дающей разумную точность. Такая интерполяция обычно нелинейна.

Для этого нужно располагать дополнительной информацией о качественном поведении функции. Часто ее можно получить, зная физический смысл $y(x)$. Например, проходящий через погло-

щающую среду свет ослабляется примерно по экспоненциальному закону; сопротивление движению в газе зависит от скорости примерно как v^m, где $m \approx 1$ для ламинарного движения, $m \approx 2$ для турбулентного и $m > 2$ вблизи звукового барьера. Нередко помогают формальные математические соображения — изучение графика функции и сравнение его с графиками хорошо изученных функций (в первую очередь элементарных).

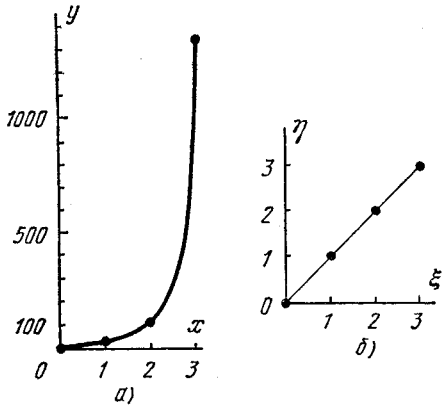


Рис. 5.

Выяснив качественное поведение функции, стараются подобрать такое преобразование переменных $\eta = \eta(y)$, $\xi = \xi(x)$, чтобы в новых переменных график $\eta(\xi)$ мало отличался от прямой на протяжении нескольких шагов таблицы. Тогда в переменных $\eta(\xi)$ интерполяция многочленом невысокой степени будет давать хорошую точность. Вычисления заключаются в составлении таблицы для новых переменных $\eta_i = \eta(\xi_i)$, интерполяции по ней и нахождении $y = y(\eta)$ обратным преобразованием. Этот способ называют *методом выравнивания*.

Пример 2. Проиллюстрируем метод выравнивания на примере функции, заданной таблицей 5. Нетрудно заметить, что зависимость близка к показательной, $y(x) \approx 10^x$; значит, в переменных $\xi = x$, $\eta = \lg y$ график будет почти прямым (рис. 5, *б*). Составим новую таблицу 6 и проведем интерполяцию по формуле Ньютона

$$\eta^* = \eta(0{,}5) = 0 + 0{,}5207 + 0 - 0{,}0004 \approx 0{,}5203.$$

Теперь члены ряда быстро убывают, обеспечивая хорошую точность; считая, что точность η^* примерно равна последнему члену ряда, обратным преобразованием получим, что $y(\eta^*) \approx 3{,}314 \pm 0{,}1\%$. Очевидно, что удачно подобранное выравнивание позволило получить высокую точность интерполяции.

Замечание 1. Для каждой конкретной функции подбирают свой вид нелинейной интерполяции. Для других функций этот вид, как правило, будет давать плохую точность.

Замечание 2. Оценка погрешности такой интерполяции содержит старшие производные $\eta(\xi)$. Их трудно найти, поэтому на практике удобнее оценивать точность по скорости убывания членов в формуле Ньютона, как было сделано выше. Употреби-

телен также следующий прием: для одного из узлов x_i вычисляют $y(x_i)$ интерполяцией по соседним узлам и сравнивают с табличным значением y_i.

Таблица 6

ξ_i	η_i	$\eta(\xi_i, \xi_{i+1})$		
0	0,0000			
		1,0414		
1	1,0414		0,0000	
		1,0414		0,0011
2	2,0828		0,0032	
		1,0478		
3	3,1306			

Пример 3. Отбросим в таблице 6 узел $\xi = 1$ и связанные с ним разделенные разности. По оставшимся трем узлам приближенно вычислим отброшенное значение $\eta(1) \approx 1{,}0382$ или $y(1) \approx \approx 10{,}92$. Последняя величина отличается от табличного значения на 0,8%. Это вычисление велось фактически с шагом $h_0 = 2$ многочленом второй степени, имеющим погрешность $O(h^3)$. Значит, при вычислениях с шагом $h = 1$ погрешность должна уменьшиться в $(h_0/h)^3 = 8$ раз и составить 0,1%. Это хорошо согласуется с оценкой по последнему члену ряда, сделанной выше.

Замечание 3. Оба прямых преобразования $\eta(y)$, $\xi(x)$ и обратное преобразование $y(\eta)$ должны выражаться несложными формулами, иначе метод выравнивания будет малопригодным на практике. Удобны преобразования типа логарифмирования, вычисления экспонент, тригонометрических функций и другие, имеющиеся в библиотеках стандартных программ современных ЭВМ (или легко выполнимые на логарифмической линейке).

Замечание 4. В исходных переменных интерполяция нелинейна относительно параметров; в данном примере она имела вид $\varphi(x) = \exp\left(\sum_{k=0}^{n} a_k x^k\right)$. Однако в переменных η, ξ она линейна по параметрам. Такая нелинейность мало осложняет работу, поэтому интерполяцию подобного вида будем называть *квазилинейной*.

Встречаются случаи, когда метод выравнивания неприменим. Например, если $y(x) \approx a(x+b)^c$, то не удается найти такие координаты, которые превращали бы график в прямую и не содержали бы явно параметров a, b, c. Тогда зависимость от параметров не сводится к линейной и отыскать параметры и выполнить интерполяцию нелегко. Такую интерполяцию будем называть *существенно* нелинейной; на практике она используется крайне редко.

Замечание 5. Если выравнивающие преобразования переменных просты, то иногда удается явно выразить $\varphi(x)$ через

табличные значения функции в исходных переменных. Например, двухточечная интерполяция многочленом Ньютона в выравнивающих переменных имеет следующий вид:

$$\eta \approx \eta_0 + (\eta_1 - \eta_0)\,(\xi - \xi_0)/(\xi_1 - \xi_0), \qquad \xi_0 \leqslant \xi \leqslant \xi_1.$$

Если при выравнивании используется преобразование $\xi = x$, $\eta = \ln y$, то, возвращаясь к исходным переменным, получим

$$y(x) \approx y_0\,(y_1/y_0)^{(x-x_0)/(x_1-x_0)}, \qquad x_0 \leqslant x \leqslant x_1. \tag{19}$$

Но при большем числе узлов интерполяции подобные формулы становятся настолько громоздкими, что более выгодно не пользоваться ими, а проводить вычисления в выравнивающих переменных.

9. Интерполяция сплайнами. Когда надо провести график функции по известным точкам $y(x_i)$, $0 \leqslant i \leqslant N$, то обычно пользуются лекалом. Однако если точки расположены редко, то нелегко бывает подобрать участок лекала, проходящий сразу через много точек. Тогда опытные инженеры берут гибкое лекало — металлическую линейку, ставят ее на ребро и изгибают, придерживая в нескольких местах пальцами так, чтобы ее ребро проходило сразу через все точки (рис. 6).

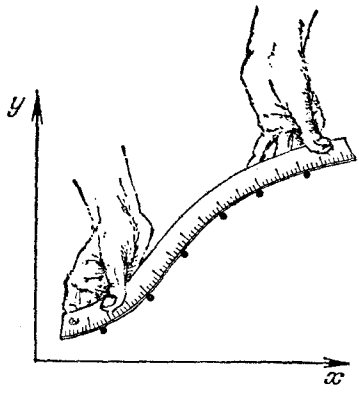


Рис. 6.

Этот способ интерполяции можно описать математически. Гибкая линейка — это упругий брусок; из курса сопротивления материалов известно, что уравнение его свободного равновесия есть $\varphi^{IV}(x) = 0$. Значит, в промежутке между каждой парой соседних узлов интерполяционная функция является многочленом третьей степени, который удобно записать в таком виде:

$$\varphi(x) = a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3,$$
$$x_{i-1} \leqslant x \leqslant x_i. \tag{20}$$

Коэффициенты многочлена на каждом интервале определяют из условий в узлах. Очевидно, в узлах многочлен должен принимать табличные значения функции:

$$y_{i-1} = \varphi(x_{i-1}) = a_i, \qquad 1 \leqslant i \leqslant N, \tag{21}$$

$$y_i = \varphi(x_i) = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \qquad h_i = x_i - x_{i-1}. \tag{22}$$

Число этих уравнений вдвое меньше числа неизвестных коэффи-

циентов, поэтому для определенности задачи нужны дополнительные условия. Для их получения вычислим первую и вторую производные многочлена (20):

$$\varphi'(x) = b_i + 2c_i(x - x_{i-1}) + 3d_i(x - x_{i-1})^2,$$
$$\varphi''(x) = 2c_i + 6d_i(x - x_{i-1}) \text{ при } x_{i-1} \leqslant x \leqslant x_i,$$

и потребуем непрерывности этих производных (т. е. гладкости линейки) во всех точках, включая узлы. Приравнивая во внутреннем узле x_i правые и левые пределы производных, получим

$$b_{i+1} = b_i + 2c_i h_i + 3d_i h_i^2, \qquad 1 \leqslant i \leqslant N-1, \tag{23}$$
$$c_{i+1} = c_i + 3d_i h_i, \qquad 1 \leqslant i \leqslant N-1. \tag{24}$$

Недостающие два условия обычно получают из естественного предположения о нулевой кривизне графика на концах:

$$^1/_2\varphi''(x_0) = c_1 = 0, \quad ^1/_2\varphi''(x_N) = c_N + 3d_N h_N = 0, \tag{25}$$

что соответствует свободно отпущенным концам линейки. Но если есть дополнительные сведения об асимптотике функции, то можно записать другие краевые условия.

Уравнения (21) — (25) образуют систему линейных уравнений для определения $4N$ неизвестных коэффициентов. Эту систему можно решить методом исключений Гаусса, описанным в главе V. Но гораздо выгоднее сначала привести ее к специальному виду. Уравнение (21) сразу дает нам все коэффициенты a_i. Из уравнений (24) и (25) следует

$$d_i = (c_{i+1} - c_i)/3h_i \text{ при } 1 \leqslant i \leqslant N-1,$$
$$d_N = -c_N/3h_N. \tag{26}$$

Подставим соотношение (26) в (22), одновременно исключая оттуда $a_i = y_{i-1}$; тогда получим

$$b_i = [(y_i - y_{i-1})/h_i] - {}^1/_3 h_i(c_{i+1} + 2c_i), \quad 1 \leqslant i \leqslant N-1,$$
$$b_N = [(y_N - y_{N-1})/h_N] - {}^2/_3 h_N c_N. \tag{27}$$

Исключим теперь из (23) величины b_i и b_{i+1} при помощи (27), соответственно увеличивая во втором случае индекс на единицу, а величину d_i — на основании (26). Останется система линейных уравнений для коэффициентов c_i, легко приводящаяся к следующему виду:

$$c_1 = 0,$$
$$h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_i c_{i+1} =$$
$$= 3[(y_i - y_{i-1})/h_i - (y_{i-1} - y_{i-2})/h_{i-1}] \text{ при } 2 \leqslant i \leqslant N,$$
$$c_{N+1} = 0. \tag{28}$$

Матрица этой системы трехдиагональна, т. е. ненулевыми в ней являются только элементы главной диагонали и двух соседних. Такие системы экономно решаются методом прогонки, изложенным в главе V. После нахождения коэффициентов c_i остальные коэффициенты нетрудно вычислить по формулам (21), (26) и (27).

Можно рассмотреть более общую задачу интерполяции функции *сплайном* — многочленом n-й степени:

$$S(x) = \sum_{k=0}^{n} a_{ik} x^k, \qquad x_{i-1} \leqslant x \leqslant x_i,$$

коэффициенты которого кусочно-постоянны и который в узлах интерполяции принимает заданные значения и непрерывен вместе со своими $n-1$ производными. При нечетной степени многочлена $n = 2p-1$ можно рассматривать сплайновую интерполяцию как решение задачи лангранжевой интерполяции при дополнительном условии

$$\int_a^b [S^{(p)}(x)]^2 \, dx = \min.$$

Из этого условия следует уравнение $S^{(2p)}(x) = 0$ для интерполирующей функции, условия непрерывности $2p-2$ производных во внутренних узлах и естественные ограничения на производные в крайних узлах.

На практике употребительны два случая. Один — подробно рассмотренный здесь случай $n = 3$. Второй — $n = 1$, когда сплайн совпадает с многочленом Ньютона первой степени и соответствует аппроксимации графика ломаной, построенной по узлам; определение коэффициентов при этом очевидно.

Сплайновая интерполяция напоминает лагранжеву тем, что она требует знания в узлах только значений функции, но не ее производных. По области применения она занимает промежуточное положение между линейной и нелинейной лагранжевой интерполяцией. Ее целесообразно применять тогда, когда сетка недостаточно подробна для интерполяции многочленом Ньютона, но еще не настолько редка, чтобы необходимо было прибегать к нелинейной интерполяции. Если функция так же резко меняется за один шаг сетки, как в таблице 5, то сплайновая интерполяция не гарантирует хорошей точности.

Наиболее успешно применяют сплайновые интерполяции при разностном решении краевых задач для эллиптических уравнений в частных производных с гладкими коэффициентами.

10. Монотонная интерполяция. Монотонность — важное свойство функций. Например, возьмем таблицы синусов с шагом аргумента 1°; тогда на каждые 89 интервалов, в которых функция будет монотонна, придется всего 1 интервал, содержащий экстремум. Поэтому при интерполяции нередко желательно сохранять монотонность функции.

Если интерполяционная функция $\varphi(x; a)$ монотонна по x, то интерполяция будет монотонной. Классический пример — двухточечная интерполяция многочленом Ньютона $\mathcal{P}_1(x) = a_0 + a_1 x$. Другим примером может служить двухточечная квазилинейная интерполяция (19). Очевидно, если интерполяция квазилинейная двухточечная, а преобразования $\eta(y)$, $\xi(x)$ монотонны, то интерполяция будет монотонной.

При трехточечной интерполяции монотонность может нарушиться. В таблице 5 (п. 8) функция, по-видимому, монотонна. Но если использовать многочлен Ньютона с тремя узлами, т. е. оставить в формуле Ньютона только три члена, то получим $y(1/2) \approx -6{,}5$, что нарушает монотонность. Очевидно, это результат использования немонотонной интерполяционной функции — параболы $\mathcal{P}_2(x) = a_0 + a_1 x + a_2 x^2$.

Двухточечная интерполяция имеет погрешность $O(h^2)$, и ее точность не всегда достаточна; а увеличение числа узлов может внести немонотонность. Конечно, если интерполяционный ряд Ньютона (8) хорошо сходится, а монотонность все-таки нарушена, то это означает, что функция на самом деле немонотонна. Но нередко мы вынуждены ограничиваться заданным в формуле (или программе для ЭВМ) числом узлов. Если при этом надо сохранить монотонность, то можно поступать следующим образом.

Найдем такие соседние точки сетки, чтобы выполнялось $x_l \leqslant x \leqslant x_{l+1}$. Проведем вычисления по заданной многоточечной интерполяционной формуле и получим $y^* = \varphi(x)$. Если это значение лежит между значениями y_i, y_{i+1}, то считаем ответ правильным. Если оно выходит за пределы интервала, определяемого значениями y_i, y_{i+1}, то вместо y^* в качестве ответа берем ближайшее из этих двух значений, т. е. полагаем

$$\begin{aligned} y(x) &= y^* = \varphi(x) && \text{при} && \min(y_i, y_{i+1}) \leqslant y^* \leqslant \max(y_i, y_{i+1}), \\ y(x) &= \min(y_i, y_{i+1}) && \text{при} && y^* < \min(y_i, y_{i+1}), \\ y(x) &= \max(y_i, y_{i+1}) && \text{при} && y^* > \max(y_i, y_{i+1}). \end{aligned} \qquad (29)$$

Эта *монотонная* интерполяция бывает полезна, например, при составлении разностных схем для уравнений в частных производных (глава X, § 1, п. 6).

11. Многомерная интерполяция. Двумерные таблицы широко распространены в физике и технике; например, таковыми являются таблицы термодинамических функций газов, где независимыми переменными обычно являются температура и плотность. Трехмерные таблицы составляют и используют значительно реже, но не потому, что таких зависимостей нет, а потому, что таблицы слишком громоздки. Четырехмерных таблиц практически нет, хотя в физике немало задач с большим числом параметров; так, проводимость плазмы $\sigma(T, \rho, E, H)$ зависит от ее температуры и плотности, и напряженностей электрического (если сказываются нелинейные эффекты) и магнитного полей.

Отметим некоторые существенные стороны многомерной интерполяции. Для простоты ограничимся двумерными таблицами $z(x, y)$; обобщить все результаты на большее число измерений нетрудно.

1) Чтобы объем таблиц был приемлем, приходится шаги по аргументам брать довольно большими. Это предъявляет жесткие требования к способу интерполяции. Часто приходится пользоваться методом выравнивания, т. е. подбирать замену перемен-

ных $\zeta(z)$, $\xi(x)$, $\eta(y)$, преобразующую описываемую функцией поверхность в плоскость.

Например, законы зависимости давления горячих газов от температуры и плотности $P(T, \rho)$ близки к степенным. Поэтому при составлении таблиц свойств газов выгодно табулировать $\zeta = \lg P$ при аргументах $\xi = \lg T$, $\eta = \lg \rho$ и сетки по новым аргументам брать равномерными (к сожалению, физики редко это делают). Сходные закономерности справедливы для других термодинамических функций, коэффициентов теплопроводности и электропроводности и еще многих свойств веществ.

В дальнейшем мы будем предполагать, что выравнивающие переменные уже подобраны, и таблицы составлены в новых переменных. Тогда в качестве интерполирующей функции можно использовать многочлен невысокой степени.

2) Не любое число узлов интерполяции выгодно. Если для одной переменной степень многочлена была взаимно однозначно связана с числом узлов, то для двух переменных многочлен n-ой степени $\mathcal{P}_n(x, y) = \sum_{k+m=0}^{n} a_{km} x^k y^m$ имеет $(n+1)(n+2)/2$ узлов. Если число узлов не соответствует этой формуле, то часть коэффициентов при высших степенях должна задаваться принудительно (в частности, нулями); для выбора этих коэффициентов редко есть разумные основания.

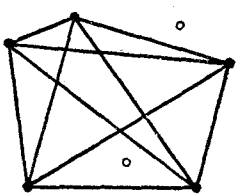


Рис. 7.

3) В многомерном случае иначе определяется понятие экстраполяции. Возьмем узлы интерполяции и соединим их попарно прямыми (в случае большего числа измерений — гиперплоскостями). Крайние отрезки ограничивают выпуклую область (рис. 7). Если искомая точка попадает в эту область, то имеет место интерполяция; если не попадает, то экстраполяция.

4) Не всякое расположение узлов допустимо. В одномерном случае узлы не должны были совпадать. Теперь же для интерполяции многочленом $\mathcal{P}_1(x, y)$ необходимо, чтобы узлы не лежали на одной прямой в плоскости (x, y). В самом деле, система трех уравнений $a + bx_i + cy_i = z_i$ имеет определитель

$$\Delta(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3) = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2), \quad (30)$$

который обращается в нуль, если узлы лежат на одной прямой. При интерполяции многочленом $\mathcal{P}_2(x, y)$ требуется, чтобы узлы не лежали на кривой второго порядка и т. д.

Такие условия, а также условие отсутствия экстраполяции проверять в общем случае сложно. Поэтому для хорошей интерполяции сетка должна быть регулярно построенной, а не представлять собой совокупность беспорядочно расположенных точек; узлы из нее следует выбирать определенным образом. В дальнейшем ограничимся наиболее удобной прямоугольной сеткой (рис. 8, 9); желательно, чтобы она была равномерной.

На прямоугольной сетке удобна *последовательная* интерполяция. Пусть заданы $z_{ij} = z(x_i, y_j)$ и требуется найти $z(x, y)$. Выберем на сетке прямоугольник из km узлов, в который попадает искомая точка (рис. 8). Сначала проведем лагранжеву интерполяцию по строкам, т. е. при каждом фиксированном j_0 найдем значение $z(x, y_{j_0})$ по значениям z_{i_0}. Затем проведем лагранжеву интерполяцию по столбцу, т. е. по значениям $z(x, y_j)$ найдем искомое значение $z(x, y)$.

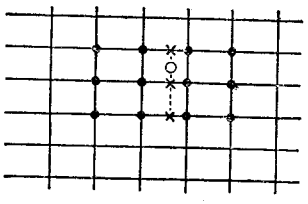


Рис. 8.

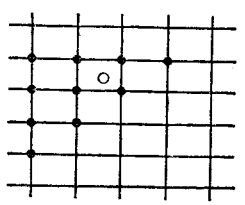


Рис. 9.

Последовательная интерполяция имеет ряд преимуществ. Она позволяет брать по каждой переменной свое число узлов. Легко написать ее общую формулу, аналогичную одномерной формуле Лагранжа:

$$\mathcal{P}_{km}(x, y) = \sum_{i=0}^{k} \sum_{j=0}^{m} z_{ij} \prod_{\substack{p=0 \\ p \neq i}}^{k} \prod_{\substack{q=0 \\ q \neq j}}^{m} \frac{(x - x_i)(y - y_j)}{(x_p - x_i)(y_q - y_j)}, \tag{31}$$

хотя вычисления удобнее производить, последовательно применяя одномерные формулы Ньютона. Формулу (31) можно обобщить, используя для каждого аргумента свою квазилинейную интерполяцию, т. е. по строкам делая замену $\xi(x)$, $\zeta_1(z)$, а по столбцам — $\eta(y)$, $\zeta_2(z)$; такие выравнивающие замены подобрать проще, чем единую замену. Однако последовательная интерполяция завышает степень интерполирующего многочлена; например, если по обоим направлениям берется двухточечная интерполяция, т. е. многочлен первой степени, то результирующий многочлен будет квадратичным многочленом вида $\mathcal{P}_{2,2}(x, y) = \alpha + \beta x + \gamma y + \delta xy$.

Многочлен минимальной степени получается при *треугольной* интерполяции. Если взять треугольную конфигурацию узлов

интерполяции, изображенную на рис. 9 или повернутую на угол, кратный 90°, то число узлов будет равно $(n+1)(n+2)/2$. Это число однозначно определяет многочлен n-й степени, который удобно записать в форме Ньютона, вводя разделенные разности функции двух переменных:

$$z(x_0, x_1; y) = [z(x_0, y) - z(x_1, y)]/(x_0 - x_1),$$
$$z(x; y_0, y_1) = [z(x, y_0) - z(x, y_1)]/(y_0 - y_1) \tag{32}$$

и т. д. Такими же рассуждениями, как в одномерном случае, можно показать, что интерполяционный многочлен лагранжева типа имеет следующий вид:

$$\mathcal{P}_n(x, y) = \sum_{i=0}^{n} \sum_{j=0}^{n-i} z(x_0, \ldots, x_i; y_0, \ldots, y_j) \prod_{p=0}^{i-1} (x - x_p) \prod_{q=0}^{j-1} (y - y_q). \tag{33}$$

В одномерном случае переменная y и индексы j, q исчезают, так что формула (33) переходит в обычную формулу Ньютона.

Многомерная интерполяция настолько громоздка, что обычно используется только многочлен первой или второй степени; читателям предлагается записать формулы (31) — (33) для этих случаев. Многочлены более высоких степеней используются много реже. По той же причине интерполяция эрмитова типа для многих переменных практически не употребляется. Сплайновая интерполяция используется в основном при разностном решении уравнений в частных производных.

Иногда мы вынуждены работать с функцией, заданной на нерегулярной сетке (например, с функцией, измеренной экспериментально). Тогда обычно ограничиваются интерполяционным многочленом первой степени; его коэффициенты находят по трем выбранным узлам, приравнивая в них многочлен табличным значениям функции:

$$z \approx a + bx + cy,$$
$$z_i = a + bx_i + cy_i, \quad i = 1, 2, 3. \tag{34}$$

Вычислять коэффициенты a, b, c на самом деле не нужно. Заметим, что равенства (34) означают, что столбец $\{z, z_1, z_2, z_3\}$ есть линейная комбинация трех столбцов, стоящих в правой части при коэффициентах. Следовательно, составленный из всех четырех столбцов определитель равен нулю:

$$\begin{vmatrix} z & 1 & x & y \\ z_1 & 1 & x_1 & y_1 \\ z_2 & 1 & x_2 & y_2 \\ z_3 & 1 & x_3 & y_3 \end{vmatrix} = 0.$$

Раскрывая этот определитель по первому столбцу и вспоминая формулу (30), получим следующее выражение для интерполяционного многочлена:

$$z=[z_1\Delta(r, r_2, r_3)+z_2\Delta(r_1, r, r_3)+ \\ +z_3\Delta(r_1, r_2, r)]/\Delta(r_1, r_2, r_3). \quad (35)$$

Эту процедуру вывода формулы нетрудно обобщить на многочлен любой степени при произвольном расположении узлов, но сами формулы для многочленов высокой степени получаются громоздкими и неудобными для вычислений.

§ 2. Среднеквадратичное приближение

1. Наилучшее приближение. Интерполяция позволяет легко аппроксимировать функцию $y(x)$. Однако точность такой аппроксимации гарантирована лишь в небольшом интервале порядка нескольких шагов сетки. Для другого интервала приходится заново вычислять коэффициенты интерполяционной формулы. Нам же всегда желательно иметь единую приближенную формулу $y\approx\varphi(x)$, пригодную для большого отрезка $a\leqslant x\leqslant b$. Поэтому далее будем сравнивать заданную и аппроксимирующую функции на большом отрезке.

При интерполяции мы приравниваем значения $y(x)$ и $\varphi(x)$ в узлах. Если $y(x_i)$ определены неточно — например, из эксперимента, — то точное приравнивание неразумно. Поэтому нередко целесообразней приближать функцию не по точкам, а в среднем, т. е. в норме L_p.

Пусть заданы функция $y(x)$ и множество функций $\varphi(x)$, принадлежащие линейному нормированному пространству функций. Нас интересуют две задачи. Первая — аппроксимация с заданной точностью: по заданному ε найти такую $\varphi(x)$, чтобы выполнялось неравенство $\|y(x)-\varphi(x)\|\leqslant\varepsilon$. Второе — нахождение *наилучшего приближения*, т. е. функции $\bar{\varphi}(x)$, удовлетворяющей соотношению

$$\|y(x)-\bar{\varphi}(x)\|=\inf\|y(x)-\varphi(x)\|=\nu. \quad (36)$$

Существует ли наилучшее приближение и единственно ли оно (для данных функции и множества)? Это имеет место не при любом выборе пространства и множества. Например, в пространстве L_1, $-1\leqslant x\leqslant+1$, выберем функцию $y(x)=1$ и множество $\varphi(x)=cx$; тогда

$$\|y-\varphi\|_{L_1}=\int_{-1}^{+1}|1-cx|\,dx=\begin{cases}2 & \text{при } |c|\leqslant 1,\\ \dfrac{c^2+1}{|c|}>2 & \text{при } |c|>1.\end{cases}$$

В самом деле, при $|c|\leqslant 1$ эта норма равна площади заштрихованной трапеции на рис. 10, *а*, т. е. двум. При $|c|>1$ эта

норма, согласно рис. 10, *б*, равна площади заштрихованной трапеции (которая опять равна двум) плюс площади заштрихованных треугольников. Значит, для любого c, по модулю меньшего единицы, $\varphi = cx$ минимизирует норму отклонения, т. е. наилучшее приближение здесь существует, но оно не единственно.

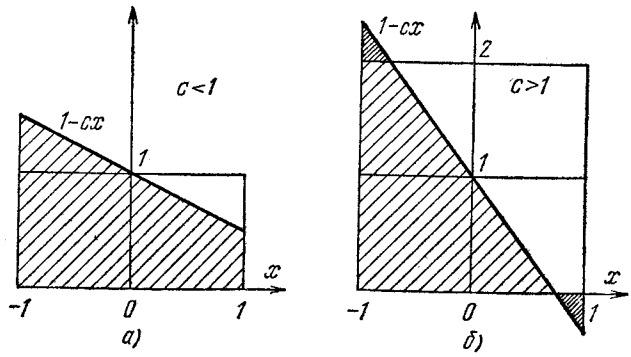


Рис. 10.

Выведем достаточное условие существования наилучшего приближения. Пусть в линейном пространстве функций выбрано множество, образованное функциями вида

$$\varphi(x) = \sum_{k=1}^{n} a_k \varphi_k(x), \tag{37}$$

где функции $\varphi_k(x)$ можно считать линейно-независимыми. Это множество есть линейное подпространство нашего пространства. Изменим один из коэффициентов суммы (37) на величину δa_k; из неравенства треугольника (1.3) следует

$$\|\|y - (\varphi + \delta\varphi)\| - \|y - \varphi\|\| \leqslant \|\delta\varphi\| = |\delta a_k| \cdot \|\varphi_k\|,$$

т. е. норма $\|y - \varphi\|$ непрерывно зависит от a_k. Очевидно, $\|\varphi\|$ также есть непрерывная функция коэффициентов a_k.

Рассмотрим нормы как функции координат a_k. Сфера

$$\sum_{k=1}^{n} a_k^2 = 1$$

есть замкнутое ограниченное множество, поэтому $\|\varphi\|$ на этой сфере имеет точную нижнюю грань μ и в силу непрерывности достигает ее при некотором $\tilde{\varphi}(x)$. Очевидно, $\mu > 0$; в противном случае $\tilde{\varphi}(x) \equiv 0$, что противоречит линейной независимости $\varphi_k(x)$.

Возьмем шар $\sum_{k=1}^{n} a_k^2 \leqslant R^2 = |\nu + \|y\| + \varepsilon|^2 / \mu^2$, где ε — какое-то положительное число. В силу однородности нормы функции вне этого шара $\|\varphi\| \geqslant \mu R = \nu + \|y\| + \varepsilon$ и, следовательно, $\|y - \varphi\| \geqslant \geqslant \|\varphi\| - \|y\| \geqslant \nu + \varepsilon$. Значит, вне этого шара норма погрешности заведомо далека от нижней грани. Только внутри шара $y(x)$ и $\varphi(x)$ достаточно близки по норме. Но шар — ограниченное и замкнутое множество значений координат a_k, поэтому непрерывная функция координат $\|y - \varphi\|$ достигает на нем точной нижней грани.

Следовательно, в любом линейном нормированном пространстве при линейной аппроксимации (37) наилучшее приближение существует, хотя не во всяком линейном пространстве оно единственно.

На практике используются пространства L_2 и C. В этом параграфе рассмотрим приближения в пространстве L_2, т. е. среднеквадратичную аппроксимацию.

2. Линейная аппроксимация. Рассмотрим гильбертово пространство $L_2(\rho)$ действительных функций, интегрируемых с квадратом с весом $\rho(x) > 0$ на $[a, b]$. Норма в нем равна $\|f\|_{L_2} = \sqrt{(f, f)}$, где скалярное произведение определено следующим образом:

$$(f, \varphi) = \int_a^b \rho(x) f(x) \varphi(x)\, dx.$$

Физический смысл весовой функции будет пояснен в п. 4. Выберем в качестве аппроксимирующей функции линейную комбинацию (37). Подставляя ее в условие наилучшего приближения (36), получим

$$\|y - \varphi\|_{L_2}^2 = (y, y) - 2\sum_{k=1}^{n} a_k (y, \varphi_k) + \sum_{k, m=1}^{n} a_k a_m (\varphi_k, \varphi_m) = \min.$$

Приравнивая нулю производные по коэффициентам, получим систему линейных уравнений

$$\sum_{m=1}^{n} (\varphi_k, \varphi_m) a_m = (y, \varphi_k), \qquad 1 \leqslant k \leqslant n. \tag{38}$$

Ее определитель есть определитель Грама функций $\varphi_k(x)$; поскольку функции линейно-независимы, он отличен от нуля. Следовательно, *наилучшее среднеквадратичное приближение существует и единственно*. Для его вычисления необходимо решить систему линейных уравнений (38).

Линейно-независимую систему функций можно ортогонализировать. Пусть $\varphi_k(x)$ уже образуют ортонормированную систему,

т. е. $(\varphi_k, \varphi_m) = \delta_{km}$; тогда формулы (38) резко упрощаются и становятся удобными для вычислений

$$a_k = (\varphi_k, y) = \int_a^b \rho(x)\, y(x)\, \varphi_k(x)\, dx \quad \text{при} \quad (\varphi_k, \varphi_m) = \delta_{km}. \qquad (39)$$

Это коэффициенты Фурье, так что *наилучшее приближение есть отрезок обобщенного ряда Фурье.*

Если функции $\varphi_k(x)$ образуют полную ортонормированную систему, то в силу равенства Парсеваля

$$\| y - \varphi \|_{L_2}^2 = \sum_{k=n+1}^{\infty} a_k^2.$$

Значит, при $n \to \infty$ норма погрешности неограниченно убывает, т. е. наилучшее приближение среднеквадратично сходится к $y(x)$, и возможна аппроксимация с любой точностью.

Отметим, что если $\varphi_k(x)$ не ортогональны, то при $n \to \infty$ определитель Грама обычно быстро стремится к нулю, система (38) становится плохо обусловленной, т. е. ее решение связано с большой потерей точности (см. главу V), и больше $5-6$ членов суммы (37) брать нецелесообразно. Численная ортогонализация базиса при этом тоже приводит к большой потере точности. Поэтому если нужно большое число членов, то надо или проводить ортогонализацию точно (аналитически), или пользоваться готовыми системами ортогональных функций.

При интерполяции мы обычно полагали $\varphi_k(x) = x^k$. Для среднеквадратичной аппроксимации удобнее в качестве $\varphi_k(x)$ брать многочлены, ортогональные с заданным весом. Наиболее употребительны из них многочлены Якоби (частным случаем которых являются многочлены Лежандра и Чебышева), Лагерра и Эрмита. Для аппроксимации периодических функций используют тригонометрический ряд; он соответствует $\rho(x) = 1$. Сводка формул для ортогональных полиномов приведена в Приложении.

Все перечисленные выше системы функций полные, так что наилучшие приближения по ним среднеквадратично сходятся при $n \to \infty$, если $y(x)$ интегрируема с квадратом с заданным весом. При более сильных ограничениях имеет место сходимость во всех точках и даже равномерная сходимость. Приведем без доказательства некоторые результаты.

а) Ряд по многочленам Якоби $P_n^{\alpha, \beta}(x)$ сходится к непрерывной функции $y(x)$ равномерно на $[-1, +1]$, если существует непрерывная $y^{(p)}(x)$ при некотором $p \geqslant 2 + 2 \max(\alpha, \beta)$ и если $\max(\alpha, \beta) \geqslant -1/2$. В частности, для многочленов Чебышева первого рода достаточно $p = 1$, а для многочленов Чебышева второго рода $p = 3$. Для многочленов Лежандра доказан более сильный результат: ряд сходится равномерно, если существует ограниченная $y'(x)$.

б) Если функция $y(x)$ кусочно-непрерывная и кусочно-гладкая на $[0, \infty)$ и существует

$$\int_0^\infty e^{-x/2} x^{(2\alpha-1)/4} \mid y(x) \mid dx,$$

то ряд по многочленам Лагерра $L_n^{(\alpha)}(x)$ сходится к функции в точках ее непрерывности и к полусумме односторонних пределов $\frac{1}{2}(y_+ + y_-)$ в точках разрыва. Эта сходимость, вообще говоря, не равномерная.

в) Если функция $y(x)$ кусочно-непрерывная и кусочно-гладкая на $(-\infty, +\infty)$ и существует

$$\int_{-\infty}^{+\infty} e^{-x^2} y^2(x) \mid x \mid dx,$$

то ряд по многочленам Эрмита $H_n(x)$ сходится так же, как в предыдущем абзаце.

г) Если $y(x)$ периодическая и непрерывная, причем ее модуль непрерывности удовлетворяет условию $\omega(\delta) \leqslant C\delta^\alpha$, $0 < \alpha \leqslant 1$, то ее тригонометрический ряд Фурье равномерно сходится к ней на всем периоде (признак Липшица); в частности, это условие выполняется для функции с ограниченной производной. Если функция имеет ограниченную p-ю производную $\mid y^{(p)}(x) \mid \leqslant M_p$, а все младшие производные непрерывны, то для погрешности тригонометрического ряда Фурье и величин отдельных коэффициентов справедливы оценки

$$\mid R_n(x) \mid < AM_n \frac{\ln n}{n^p}, \quad a_k = O(k^{-(p+1)}),$$

где A — константа. Видно, что при больших p ряд сходится быстро. Но если $y(x)$ кусочно-непрерывна, то сколько бы ни было у нее кусочно-непрерывных и ограниченных производных, ее коэффициенты Фурье убывают не быстрей $a_k = O(1/k)$, и ряд сходится медленно (или даже расходится).

Замечание 1. Сходимость не во всех рассмотренных случаях была равномерной. Более того, не существует такого веса $p(x)$, чтобы любая непрерывная функция $y(x)$ разлагалась в равномерно сходящийся ряд по полиномам, ортогональным с этим весом. Дю Буа-Реймондом и Л. Фейером были построены примеры периодических непрерывных функций, у которых тригонометрический ряд Фурье в отдельных точках расходится.

Замечание 2. Сходимость среднеквадратичного приближения тем лучше, чем меньше у функции $y(x)$ особенностей — разрывов ее самой или ее производных. Если можно выделить основные особенности в виде несложной функции $y_0(x)$ и аппроксимировать разность $y(x) - y_0(x)$, точность аппроксимации существенно улучшается.

Например, периодически продолжим функцию, изображенную сплошной линией на рис. 11, и аппроксимируем ее тригонометрическим рядом Фурье. Этот ряд сходится в каждой точке, но неравномерно, ибо периодическое продолжение $y(x)$ разрывно. Если же мы положим $y_0(x) = x$, то функция $y(x) - y_0(x)$, изображенная пунктиром на рис. 11, имеет непрерывное периодическое продолжение, и ее ряд Фурье сходится к ней равномерно. Скорость сходимости ряда при этом также возрастает.

Замечание 3. Алгебраический многочлен $P_n(x)=\sum_{k=0}^{n} a_k x^k$ наилучшего среднеквадратичного приближения обладает свойством, напоминающим лагранжеву интерполяцию: разность $y(x)-P_n(x)$ на интервале (a, b) имеет не менее $n+1$ нуля. В самом

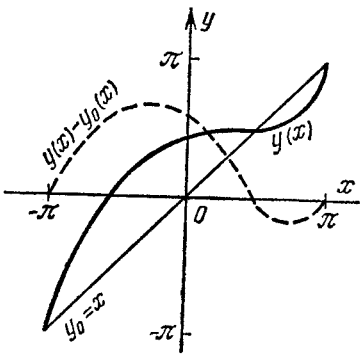


Рис. 11.

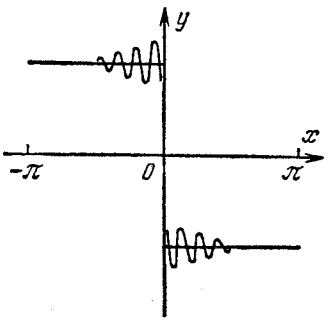


Рис. 12.

деле, предположим обратное: нули этой разности суть x_j, где $j=1, 2, \ldots, m\leqslant n$. Составим многочлен

$$Q_m(x)=\prod_{j=1}^{m}(x-x_j)\equiv\sum_{k=0}^{m} b_k x^k;$$

тогда произведение $[y(x)-P_n(x)]\,Q_m(x)$ не меняет знак, следовательно,

$$\int_a^b \rho(x)\,[y(x)-P_n(x)]\,Q_m(x)\,dx=\sum_{l=0}^{m} b_l\,[(x^l,\ y)-\sum_{k=0}^{n}(x^l,\ x^k)\,a_k]\neq 0.$$

Но если в (38) положить $\varphi_k(x)=x^k$, то квадратные скобки в сумме должны обратиться в нуль. Полученное противоречие доказывает наше утверждение.

3. Суммирование рядов Фурье. Нахождение наилучшего приближения приводит к суммированию рядов. Казалось бы, просуммировать ряд нетрудно. Но, во-первых, он далеко не всегда сходится равномерно, даже при наличии сходимости в каждой точке. Так, если $y(x)=1$ на первой половине периода и $y(x)=0$ на второй, то максимум частной суммы тригонометрического ряда Фурье стремится к 1,09 при $n\to\infty$ (явление Гиббса, рис. 12), хотя в любой точке, кроме точки разрыва, этот ряд сходится к функции.

Во-вторых, если надо суммировать много членов ряда, то происходит большое накопление погрешности входных данных и даже погрешности округления. Например, ряд Тейлора для $y(x)=\sin x$ сходится при любых значениях аргумента. Вычислим $\sin 2550°$, используя ЭВМ с 16 значащими цифрами и прекращая вычисления, когда очередной член ряда будет менее 10^{-8}. Получим бессмысленный ответ: $\sin 2550° = 29{,}5$!

Причина состоит в том, что вычисления с заданным количеством цифр эквивалентны внесению погрешности в коэффициенты ряда. Погрешности вносятся и в том случае, если находить коэффициенты по формулам (39) не аналитически, а численно. А бесконечные ряды, вообще говоря, неустойчивы по отношению к погрешности коэффициентов. В самом деле, изменим все коэффициенты a_k ряда Фурье на малые величины $\varepsilon\varphi_k(\xi)$; тогда сумма ряда изменится на

$$\sum_{k=1}^{\infty} \varepsilon\varphi_k(x)\,\varphi_k(\xi) = \varepsilon\delta(x-\xi),$$

т. е. при $x=\xi$ изменение суммы бесконечно велико. Таким образом, *суммирование бесконечного ряда Фурье является некорректной задачей*, и требуется какая-то регуляризация суммирования.

Регуляризация по числу членов. Простейшей регуляризацией является использование небольшого отрезка ряда

$$\varphi(x;\ N) = \sum_{k=1}^{N(\varepsilon)} a_k\varphi_k(x),$$

где верхний предел суммирования есть функция ошибок ε отдельных коэффициентов. Чем меньше ε, тем больше допустимое $N(\varepsilon)$.

Оценим оптимальное число членов для тригонометрического ряда Фурье. Ошибка из-за отбрасывания далеких членов ряда равна

$$\delta_1 = \sum_{k=N+1}^{\infty} a_k\varphi_k(x),$$

а ошибка из-за погрешности коэффициентов составляет

$$\delta_2 = \sum_{k=1}^{N} \delta a_k\varphi_k(x).$$

При увеличении N на единицу первая ошибка убывает на величину $a_{N+1}\varphi_{N+1}(x)$, а вторая возрастает на $\delta a_{N+1}\varphi_{N+1}(x)$. Очевидно, при малых N коэффициенты a_N велики, и преобладает убывание первой ошибки, а при достаточно больших N преобладает возрастание второй. Оптимальной является ситуация, когда

скорости изменения этих ошибок равны, т. е. при $a_{N+1} \approx \delta a_{N+1}$. Получается естественный вывод: *надо суммировать только те члены ряда, коэффициенты a_k которых превышают уровень ошибки δa_k*. Суммирование следующих членов ряда только ухудшает точность и может привести к бессмысленному результату, как видно из примера с вычислением $\sin 2550°$ (в котором роль ошибок коэффициентов играют погрешности округления при вычислении максимальных членов суммы).

Ранее отмечалось, что если $y(x)$ имеет ограниченную p-ю производную, то $a_N = O(N^{-(p+1)})$. Отсюда следует, что по порядку величины оптимальное число членов $N = O(\delta a^{-1/(p+1)})$, а достигаемая при этом погрешность $\delta_1 + \delta_2 = O(\delta a^{p/(p+1)})$. Для достаточно гладких функций оптимальное число членов оказывается небольшим и при уменьшении δa растет, но довольно медленно. Достигаемая точность тем выше, чем более высокие производные имеет функция.

Регуляризация форм-фактором. Описанный способ напоминает обрезание шумов в радиотехнике. Но подавлять шумы можно и с помощью форм-фактора, лишь ослабляющего высокие частоты. Для этого каждый член ряда (37) делят на соответственно подобранную величину $1 + b_k$ и суммируют достаточно большое число членов ряда

$$\sum_{k=1}^{\infty} a_k \varphi_k(x)/(1+b_k), \qquad b_k \geqslant 0,$$

где при малых номерах $b_k \approx 0$, а при больших номерах они достаточно быстро возрастают, причем $b_k \to \infty$. Регуляризация по числу членов означает, что выбрано $b_k = 0$ при $k \leqslant N$ и $b_k = \infty$ при $k > N$. Естественный способ выбора регуляризирующих множителей предложил А. Н. Тихонов [44], показавший, что если ортогональная система $\varphi_k(x)$ есть система собственных функций задачи Штурма — Лиувилля:

$$\frac{d}{dx}\left[p(x)\frac{d\varphi}{dx}\right] - [\lambda + q(x)]\varphi(x) = 0, \qquad \varphi'(a) = \varphi'(b) = 0,$$

то сумму обобщенного ряда Фурье следует заменить на

$$\varphi(x;\ \alpha) = \sum_{k=1}^{\infty} \frac{a_k}{1+\alpha\lambda_k}\varphi_k(x), \qquad \alpha > 0. \tag{40}$$

Поскольку собственные значения λ_k положительны и быстро растут при $k \to \infty$, то ошибки на высоких частотах хорошо подавляются. В главе XIV, § 2 будет показано, что суммирование ряда (40) устойчиво, а сумма $\varphi(x;\ \alpha)$ равномерно сходится

к $y(x)$ при $\varepsilon = \max |\delta a_k| \to 0$, если параметр $\alpha \to 0$ по определенному закону. Там же будет рассмотрен выбор параметра регуляризации α; сейчас отметим, что оптимальное $\bar{\alpha} = \alpha(\varepsilon)$ монотонно стремится к нулю при $\varepsilon \to 0$.

Попытки улучшить сходимость тригонометрических рядов Фурье предпринимались давно. В методе Фейера рассматриваются частные суммы ряда Фурье:

$$\varphi(x;\ n) = a_0/2 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx),$$

и составляется функция

$$\psi(x;\ N) = \frac{1}{N} \sum_{n=0}^{N-1} \varphi(x;\ n).$$

Эта функция при $N \to \infty$ равномерно сходится к $y(x)$, если последняя непрерывна. Скорость сходимости невелика; если ограничиться небольшим числом членов, то все резкие колебания функции будут сильно сглажены. Реально для хорошей передачи одного резкого скачка надо взять около 20 гармоник, а 10 гармоник дают невысокую точность.

Более быструю сходимость и меньшее сглаживание функции дает метод σ-множителей Ланцоша. В нем частная сумма $\varphi(x;\ n)$ осредняется по отрезку $x \pm \pi/(2n)$, т. е. по одному полупериоду наивысшей гармоники. Это приводит к умножению каждого члена частной суммы на $\sigma_k = (2\pi/(\pi k)) \sin(\pi k/(2n))$. Метод Ланцоша позволяет даже почленно дифференцировать ряд Фурье, причем выполнение всех выкладок приводит к несложной формуле

$$\bar{\varphi}'(x;\ n) = \frac{n}{2\pi}\left[\bar{\varphi}\left(x + \frac{\pi}{n};\ n\right) - \bar{\varphi}\left(x - \frac{\pi}{n};\ n\right)\right].$$

На метод Ланцоша похож метод С. Н. Бернштейна, в котором полагают

$$\psi(x;\ n) = \frac{1}{2}\left[\varphi(x;\ n) + \varphi\left(x + \frac{2\pi}{2n+1};\ n\right)\right].$$

Это обеспечивает равномерную сходимость для любой непрерывной функции $y(x)$.

Однако последние три метода не слишком точны, и область их применимости узка; поэтому с появлением регуляризации по А. Н. Тихонову их почти перестали употреблять.

4. Метод наименьших квадратов. Если вещественные функции заданы таблично, т. е. на конечном множестве точек, то их скалярное произведение определяется формулой

$$(f,\ \varphi) = \sum_{i=1}^{N} \rho_i f(x_i)\, \varphi(x_i), \qquad \rho_i > 0, \tag{41}$$

где N — полное число узлов таблицы. Тогда условие наилучшего среднеквадратичного приближения примет вид

$$\delta_\varphi^2 \sum_{i=1}^{N} \rho_i \equiv \sum_{i=1}^{N} \rho_i [y(x_i) - \varphi(x_i)]^2 = \min. \tag{42}$$

Выберем линейную аппроксимацию

$$\varphi(x) = \sum_{k=1}^{n} a_k \varphi_k(x)$$

с числом членов $n \leqslant N$. Тогда коэффициенты аппроксимации находятся из уравнений (38), где скалярные произведения надо брать согласно (41); эти уравнения можно получить и непосредственно, подставляя обобщенный многочлен в (42) и приравнивая нулю производные по коэффициентам. Описанный способ нахождения аппроксимации называется *методом наименьших квадратов*.

Метод наименьших квадратов широко используют для обработки экспериментальных кривых, точки которых измерены с заметной погрешностью ε. В этом случае весу ρ_i придают смысл точности измерения данной точки: чем выше точность, тем большее значение веса приписывают точке *). Аппроксимирующая кривая будет проходить ближе к точкам с большим весом. Сходные соображения используют в математической постановке задачи: выбирают весовую функцию $\rho(x)$ большой при тех значениях аргумента, где нужно получить более высокую локальную точность аппроксимации.

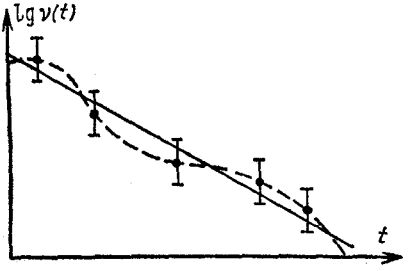


Рис. 13.

Если число коэффициентов аппроксимации n взять равным числу узлов N, то среднеквадратичная аппроксимация совпадет с лагранжевой интерполяцией. Очевидно, при наличии значительных ошибок эксперимента интерполяция неразумна. Это хорошо видно из рис. 13, показывающего описание измерений радиоактивного распада в выравнивающих переменных интерполяционным многочленом (пунктир) и прямой, найденной методом наименьших квадратов. Поскольку при $n \approx N$ среднеквадратичная аппроксимация близка к интерполяции, то хорошее сглаживание ошибок эксперимента будет при $n \ll N$; но если n слишком мало, то для описания сложной кривой коэффициентов может не хватить. Должно существовать какое-то оптимальное число коэффициентов; оно зависит от функции $y(x)$, числа узлов N, их расположения, весов и от выбранной системы $\varphi_k(x)$.

Оптимальное число коэффициентов определяют следующим образом. Выбирают некоторое n, находят из условия (42) соот-

*) Обычно полагают $\rho_i = \varepsilon_i^{-2}$.

ветствующие коэффициенты $a_k^{(n)}$, $1 \leqslant k \leqslant n$, вычисляют полученное при этом среднеквадратичное уклонение δ_n и сравнивают его с известной погрешностью эксперимента. Если $\delta_n \gg \varepsilon$, т. е. математическая погрешность аппроксимации много больше физической погрешности исходных данных, то число коэффициентов недостаточно для описания $y(x)$, и надо увеличить n. Если $\delta_n \ll \varepsilon$, то старшие коэффициенты аппроксимации физически недостоверны, и надо уменьшить n. Если $\delta_n \approx \varepsilon$, то число коэффициентов оптимально.

Обычно начинают расчет с $n=1$, когда наверняка $\delta_1 \ll \varepsilon$, и увеличивают число коэффициентов до тех пор, пока не выполнится условие $\delta_n \approx \varepsilon$. Если при этом $n \ll N$, то вид аппроксимирующей функции выбран удачно. Если же $n_{\text{опт}} \sim N$, то следует поискать более подходящий вид аппроксимирующей функции.

Описанная процедура напоминает регуляризацию суммирования ряда Фурье по числу членов. Сглаживать экспериментальные кривые можно и регуляризацией по А. Н. Тихонову (см. главу XIV, § 2); при таком сглаживании не требуется предположений о виде аппроксимирующей функции, но она успешно выполняется только при довольно большом числе узлов N. При очень малом N нахождение оптимального числа коэффициентов становится трудной задачей; требуется очень удачно подобрать вид $\varphi(x)$, а для определения достоверности результатов необходимо привлечь аппарат статистики (см. главу XV).

Отметим некоторые употребительные частные случаи метода наименьших квадратов.

Первый — полиномиальная аппроксимация, когда $\varphi_k(x) = x^k$ при $0 \leqslant k \leqslant n$. Система (38) принимает при этом вид

$$\sum_{k=0}^{n} (x^m,\ x^k)\, a_k = (y,\ x^m), \qquad 0 \leqslant m \leqslant n, \tag{43}$$

$$(x^m,\ x^k) = \sum_{i=1}^{N} \rho_i x_i^{m+k}, \qquad (y,\ x^m) = \sum_{i=1}^{N} \rho_i y_i x_i^m.$$

Поскольку степени на любом отрезке образуют чебышевскую систему, то определитель Грама отличен от нуля и задача (43) имеет единственное решение. Но система степеней не ортогональна, и при больших значениях n задача (43) плохо обусловлена. Можно обойти эту трудность, строя и используя многочлены, ортогональные с заданным весом на заданной системе точек; но к этому прибегают только в задачах, связанных с особенно тщательной статистической обработкой эксперимента. Обычно же ограничиваются небольшими степенями $n \approx 2 \div 5$, когда обусловленность задачи (43) удовлетворительна.

Второй случай — типичная радиотехническая задача о тригонометрической аппроксимации периодического сигнала, измеренного через равные доли периода, т. е. на равномерной сетке

$x_p = 2\pi p/N$, где $0 \leqslant p \leqslant N-1$. Вес в этом случае можно считать постоянным $\rho_p = 1$. Система комплексных функций $\varphi_k(x) = \exp(ikx)$ ортогональна с неединичной нормой на этой сетке; в самом деле, их скалярное произведение равно

$$(\varphi_n, \varphi_m) = \sum_{p=0}^{N-1} \varphi_n^*(x_p)\,\varphi_m(x_p) = \sum_{p=0}^{N-1} \exp\left[\frac{2\pi i}{N}(m-n)\,p\right] = N\delta_{nm}.$$

Поэтому коэффициенты аппроксимации можно находить по формулам (39) при условии введения нормирующего множителя, что приводит к так называемым формулам Бесселя

$$\begin{gathered} y(x) \approx \sum_{k=0}^{n} a_k \exp(ikx), \\ a_k = \frac{1}{N}\sum_{p=0}^{N-1} y(x_p)\exp(-ikx_p), \qquad x_p = \frac{2\pi}{N}p. \end{gathered} \tag{44}$$

Благодаря ортогональности системы функций эти формулы без потери точности можно использовать при больших n и N (разумеется, $n \leqslant N-1$). Особенно часто выбирают $N = 12$, ибо тогда все коэффициенты очень просто вычисляются.

Третий случай — это несложное сглаживание экспериментальных таблиц, точки которых измерены со значительными ошибками. Возьмем несколько соседних точек, и в этом узком интервале построим среднеквадратичную аппроксимацию с одним-двумя параметрами. Центральной точке припишем то значение, которое дает аппроксимация. Для равноотстоящих точек и единичного веса это приводит к несложным формулам. Например, для трех точек при аппроксимации многочленом первой степени из (43) нетрудно получить

$$\bar{y}_i = {}^1/_3\,(y_{i-1} + y_i + y_{i+1}). \tag{45}$$

В радиотехнике этот способ сглаживания называют *фильтром*, ибо он ослабляет высокочастотные колебания, мало влияя на низкочастотные.

Все способы сглаживания надо применять осторожно, поскольку при этом можно исказить поведение функции.

5. Нелинейная аппроксимация. Линейная, особенно линейная полиномиальная, аппроксимация часто не соответствует характеру функции. Например, многочлен высокой степени быстро растет при $|x| \to \infty$; поэтому даже несложную функцию $y(x) = 1/(1+x^2)$ многочлен плохо аппроксимирует на большом отрезке. Поскольку аппроксимация проводится в широком интервале изменения аргумента, использование нелинейной зависимости от коэффициентов здесь еще выгодней, чем при интерполяции.

На практике используют два вида зависимости. Один — *квазилинейная* зависимость, сводящаяся выравнивающей заменой переменных $\eta(y)$, $\xi(x)$ к линейной, которая подробно изучена в предыдущих пунктах. Этот способ очень эффективен и часто используется при обработке эксперимента, ибо априорные сведения о физике процесса помогают найти хорошую замену переменных. Надо только иметь в виду, что приближение, наилучшее в новых переменных, не будет наилучшим в смысле скалярного произведения в старых переменных. Поэтому на выбор веса в новых переменных надо обращать особое внимание.

Классический пример — задача о радиоактивном распаде облученного образца, в которой удобны переменные $\eta = \lg y$ и t, где $y(t)$ — скорость распада. В этих переменных кривая обычно аппроксимируется ломаной, звенья которой соответствуют распаду все более долгоживущих членов радиоактивного ряда.

Другой употребительный вид зависимости от коэффициентов — *дробно-линейная*, когда аппроксимирующая функция рациональна:

$$\varphi(x) = P_n(x)/Q_m(x) = \left(\sum_{k=0}^{n} a_k x^k\right) \Big/ \left(\sum_{q=0}^{m} b_q x^q\right). \qquad (46)$$

Нередко используется и отношение обобщенных многочленов. Такая аппроксимация позволяет передать полюсы функции $y(x)$ — им соответствуют нули знаменателя требуемой кратности. Зачастую можно воспроизвести асимптотическое поведение $y(x)$ при $x \to \infty$ за счет соответствующего выбора величины $n - m$; например, если $y(\infty) = \text{const} \neq 0$, то надо положить $n = m$. При этом сами n, m можно брать достаточно большими, чтобы располагать многими коэффициентами аппроксимации.

Однако квадрат погрешности $\| y - (P_n/Q_m) \|^2_{L_2}$ уже не будет квадратичной функцией коэффициентов, так что найти коэффициенты рациональной функции нелегко. Можно по аналогии со среднеквадратичной аппроксимацией многочленами выдвинуть гипотезу, что погрешность $y(x) - [P_n(x)/Q_m(x)]$ имеет на $[a, b]$ число нулей, не меньшее числа свободных коэффициентов (сравните с замечанием 3 в п. 2). Тогда задача сводится к лагранжевой интерполяции по этим нулям x_p и коэффициенты a_k, b_q находятся из системы линейных уравнений:

$$y(x_p) \sum_{q=0}^{m} b_q x_p^q = \sum_{k=0}^{n} a_k x_p^k, \quad 0 \leqslant p \leqslant n+m; \quad b_0 = 1. \qquad (47)$$

Разумеется, точное положение нулей неизвестно; их выбирают произвольно, обычно равномерно распределяя на отрезке $[a, b]$. Этот способ называют методом *выбранных точек*. Полученное этим методом приближение $\varphi(x)$ вовсе не будет наилучшим. Кроме

того, метод выбранных точек неразумен, как и всякая интерполяция, если $y(x_p)$ имеют заметную погрешность.

Наилучшее приближение можно найти методом *итерированного веса*. Заметим, что задача

$$\| Q_m(x)\, y(x) - P_n(x) \|^2_{L_2} = \min$$

легко решается: стоящее слева выражение есть квадратичная функция коэффициентов a_k, b_q, и дифференцирование по ним приводит к линейной системе для определения коэффициентов, сходной с (38). Новая задача отличается от исходной по существу тем, что вместо веса $\rho(x)$ используется другой вес $\rho(x)\,Q_m^2(x)$, поэтому ее решение не является наилучшим приближением. Запишем исходную задачу в новой форме:

$$\begin{gathered} \| y - (P_n/Q_m) \|^2_{L_2} = \int_a^b \bar{\rho}(x)\,[Q_m(x)\,y(x) - P_n(x)]^2\,dx = \min, \\ \bar{\rho}(x) = \rho(x)/Q_m^2(x), \end{gathered} \tag{48}$$

и будем решать ее простым итерационным процессом

$$\begin{gathered} \bar{\rho}^{(s)}(x) = \rho(x)\,[Q_m^{(s-1)}(x)]^{-2}, \\ \int_a^b \bar{\rho}^{(s)}(x)\,[Q_m^{(s)}(x)\,y(x) - P_n^{(s)}(x)]^2\,dx = \min; \end{gathered} \tag{49}$$

за нулевое приближение можно взять $Q_m^{(0)}(x) \equiv 1$. На каждой итерации вес известен по предыдущей итерации, поэтому коэффициенты $a_k^{(s)}$, $b_q^{(s)}$ легко находятся из условия минимума квадратичной формы. Практика показывает, что коэффициенты наилучшего приближения слабо зависят от выбора веса, поэтому обычно итерации сходятся быстро.

а) Рассмотрим некоторые примеры аппроксимации рациональной функцией. Положим

$$y(x) = \ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + \dots;$$

заменяя два первых члена ряда дробью, получим $\ln(1+x) \approx 2x/(2+x)$. Эта несложная формула обеспечивает точность $\sim 1\%$ при $-1/2 \leqslant x \leqslant 1$ и очень удобна для оценок.

б) В теории вероятностей важную роль играет интеграл ошибок $\Phi(x)$, для которого известны разложения в ряды:

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2}\,d\xi = \frac{2}{\sqrt{\pi}}\left(x - \frac{1}{3}x^3 + \frac{1}{10}x^5 - \frac{1}{42}x^7 + \dots\right),$$

$$\Phi(x) \approx 1 - \frac{1}{x\sqrt{\pi}}\,e^{-x^2}\left(1 - \frac{1}{2}x^{-2} + \frac{3}{8}x^{-4} - \frac{15}{8}x^{-6} + \dots\right).$$

Первый ряд абсолютно сходится, но при $x > 1$ сходимость очень медленная; второй ряд сходится асимптотически при больших значениях x. Заменяя первые члены каждого ряда дробями, получим

$$\Phi(x) \approx \frac{6x}{\sqrt{\pi}\,(3+x^2)} \quad \text{при} \quad x \lesssim 1,$$

$$\Phi(x) \approx 1 - \frac{2x}{\sqrt{\pi}\,(1+2x^2)}\, e^{-x^2} \quad \text{при} \quad x \gtrsim 1.$$

В указанных диапазонах изменения аргумента погрешность первой формулы не превышает 0,4%, а погрешность второй формулы — 2,4%. Таким образом, точность этих аппроксимаций вполне достаточна для многих практических приложений.

в) Положим $y(x) = \operatorname{arctg} x$ при $0 \leqslant x < \infty$. Эта функция монотонна, причем $y(x) \approx x$ при $x \to 0$ и $y(+\infty) = \pi/2$. Легко построить дробь

$$\varphi(x) = x \Big/ \left(1 + \frac{2}{\pi} x\right),$$

удовлетворяющую тем же условиям. Она дает грубую аппроксимацию арктангенса; локальная погрешность в точке $x = 1$ составляет 30%. Несложное видоизменение этой формулы

$$\operatorname{arctg} x \approx x \Big/ \sqrt{1 + \left(\frac{2}{\pi} x\right)^2}$$

дает вчетверо лучшую точность.

г) Тангенс в первой четверти можно грубо аппроксимировать формулой

$$\operatorname{tg} x \approx x \Big/ \left(\frac{\pi}{2} - x\right),$$

передающей поведение вблизи нуля и наличие полюса при $x = \pi/2$.

д) В задачах рассеяния часто встречается одна из специальных функций — интегральная экспонента:

$$\operatorname{Ei}(x) = \int_x^\infty \frac{e^{-t}}{t}\, dt = \ln \frac{1}{x} - C + \sum_{k=1}^{\infty} (-1)^{k-1} \frac{x^k}{k \cdot k!}. \tag{50}$$

Ряд, в который она разлагается, сходится при любых положительных значениях аргумента. Но только при $x \leqslant 1$ сходимость достаточно быстрая, и ряд пригоден для вычисления функции. Если учесть асимптотику $\operatorname{Ei}(x) \approx e^{-x}/x$ при $x \to \infty$, то рациональную аппроксимацию при $x \geqslant 1$ целесообразно искать в следующем виде:

$$\operatorname{Ei}(x) \approx \frac{e^{-x}}{x} \left(\sum_{k=0}^{n} a_k x^k\right) \Big/ \left(\sum_{q=0}^{n} b_q x^q\right), \quad a_n = b_n = 1, \tag{51}$$

где не полиномиальная часть асимптотики выделена отдельным множителем. Оказывается, уже $n = 3$, т. е. шесть свободных коэффициентов обеспечивают точность $10^{-4}\%$.

Отметим, что рациональными функциями при небольшом числе коэффициентов можно удовлетворительно аппроксимировать функции с разрывами производной вроде $y(x) = |x|$, которые плохо поддаются аппроксимации другими способами.

§ 3. Равномерное приближение

1. Наилучшие приближения. Поскольку чебышевская норма сильнее нормы L_p, то принято считать, что равномерная аппроксимация лучше аппроксимации в среднем. Поэтому поиску равномерных и особенно *наилучших равномерных* приближений, определяемых условием

$$\Delta(y, \varphi) = \min, \text{ где } \Delta(y, \varphi) = \max_{a \leqslant x \leqslant b} |y(x) - \varphi(x)|, \qquad (52)$$

где минимум ищется на множестве функций $\varphi(x)$, посвящено много работ. В частности, получены следующие результаты (доказательства большинства из них приведены в учебнике И. С. Березина и Н. П. Жидкова [4]).

а) Если выбрана линейная аппроксимация (37) с чебышевской системой функций $\varphi_k(x)$, то равномерное наилучшее приближение единственно *). Доказательство существования наилучшего приближения для этого случая было приведено в § 2, п. 1.

б) Чтобы обобщенный многочлен $\varphi(x)$ по чебышевской системе функций $\varphi_k(x)$, $1 \leqslant k \leqslant n$, был наилучшим равномерным приближением к $y(x)$ на $[a, b]$, необходимо и достаточно, чтобы на этом отрезке нашлось не менее $n+1$ таких точек, в которых погрешность $\delta(x) = y(x) - \varphi(x)$ попеременно принимает значения $+\Delta$ и $-\Delta(y, \varphi)$. Следовательно, погрешность имеет на (a, b) не менее n нулей, как и у многочленов наилучшего среднеквадратичного приближения. Впервые этот результат был получен П. Л. Чебышевым в 1859 г. для алгебраических многочленов.

в) Для функции $y(x)$, имеющей p непрерывных производных, причем $y^{(p)}(x)$ удовлетворяет условию Липшица с константой l_p, Д. Джексоном в 1911 г. получены некоторые оценки скорости сходимости наилучших равномерных приближений. При аппроксимации алгебраическим многочленом n-й степени на отрезке $-1 \leqslant x \leqslant 1$:

$$\Delta_n \leqslant (C_0 e)^{p+1} l_p / \left[\sqrt{2\pi(p+1)}\, n^{p+1}\right] = O(1/n^{p+1}), \qquad (53)$$

а при аппроксимации периодической функции с периодом 2π тригонометрическим многочленом такой же степени:

$$\Delta_n \leqslant l_p (C_0/n)^{p+1} = O(1/n^{p+1}), \qquad (54)$$

где C_0 — универсальная константа ($C_0 < 137$). С. Н. Бернштейн доказал, что из сходимости приближений со скоростью $O(1/n^{p+1+\varepsilon})$, $\varepsilon > 0$, следует наличие у функции ограниченной $p+1$-й производной, поэтому оценки Джексона почти неулучшаемы.

Таким образом, эти приближения для достаточно гладких функций быстро сходятся при $n \to \infty$, а для липшиц-непрерывных, но не гладких функций следует полагать $p = 0$, т. е. для них приближения сходятся медленно. Для произвольной функции, непрерывной на конечном отрезке $a \leqslant x \leqslant b$, равномерные приближения алгебраическими и тригонометрическими многочленами также сходятся (теорема, доказанная К. Вейерштрассом в 1885 г.; но скорость сходимости, как показал С. Н. Бернштейн в 1938 г., может быть сколь угодно малой. Именно, как бы медленно ни убывали члены монотонной последовательности $\delta_n \to 0$, $\delta_n \geqslant \delta_{n+1} > 0$, всегда найдется такая непрерывная функция $y(x)$, для которой $\Delta(y, P_n(x)) = \delta_n$. Соответствующая оценка Джексона для алгебраической аппроксимации произвольной функции, непрерывной

при $-1 \leqslant x \leqslant 1$ (и тем самым равномерно-непрерывной), есть

$$\Delta_n \leqslant (1/2 C_0 + 2)\,\omega\,(2/n), \tag{55}$$

а для тригонометрической аппроксимации непрерывной функции с периодом 2π:

$$\Delta_n \leqslant \left(\frac{1}{2\pi} C_0 + 2\right)\omega\,(2\pi/n), \tag{56}$$

где ω — модуль непрерывности функции.

Наилучшее равномерное приближение рациональной функцией (отношением многочленов) имеет такой же порядок точности, как в оценках (53)—(56), где под n надо подразумевать полное число свободных коэффициентов, которое на единицу меньше суммарной степени числителя и знаменателя.

г) Многочлены наилучшего равномерного приближения не обеспечивают хорошей сходимости (а иногда и просто сходимости) производных $\varphi'(x)$ к $y'(x)$. Если нужна сходимость производных, то приходится строить другие многочлены, которые имеют меньшую скорость сходимости. Например, многочлены С. Н. Бернштейна

$$B_n(x) = \sum_{k=0}^{n} C_n^k (1-x)^{n-k} x^k y\left(\frac{k}{n}\right), \quad 0 \leqslant x \leqslant 1, \tag{57}$$

равномерно сходятся к любой непрерывной функции $y(x)$, но не быстрее чем $O(1/n)$, сколь бы гладкой функция ни была; зато если существует непрерывная производная $y^{(p)}(x)$, то производные многочленов С. Н. Бернштейна $B_n^{(p)}(x)$ равномерно сходятся к ней на указанном отрезке при $n \to \infty$.

д) Наибольший практический интерес представляет соотношение между точностями, достигаемыми при наилучшей равномерной и наилучшей среднеквадратичной аппроксимациях. Пусть для произвольной функции $y(x)$ с периодом 2π тригонометрический многочлен наилучшего равномерного приближения есть $R_n(x)$. Доказано (см. монографию В. Л. Гончарова [9], стр. 186), что тригонометрический многочлен наилучшего среднеквадратичного приближения той же степени $Q_n(x)$ имеет погрешность не более:

$$\|y(x) - Q_n(x)\|_C \leqslant (4{,}5 + \ln n)\,\|y(x) - R_n(x)\|_C. \tag{58}$$

Сходные оценки существуют и для наилучших аппроксимаций алгебраическими многочленами на отрезке $-1 \leqslant x \leqslant 1$. Из неравенства (58) следует, что при небольших n погрешность многочленов наилучшего среднеквадратичного приближения даже в $\|\cdot\|_C$ не сильно превосходит погрешность многочленов наилучшего равномерного приближения (например, при $n \leqslant 12$ не более чем в 7 раз).

Из оценок (53)—(56) следует, что для функций с непрерывными старшими производными, не слишком большими по абсолютной величине, наилучшие равномерные приближения обеспечивают высокую точность уже при небольших $n \approx 5 \div 10$. Значит, для таких функций наилучшие среднеквадратичные приближения будут обеспечивать в $\|\cdot\|_C$ почти ту же точность, что и наилучшие равномерные приближения. Только для недостаточно гладких функций

среднеквадратичные приближения не сходятся или плохо сходятся в $\|\cdot\|_C$, но в этом случае и наилучшие равномерные приближения сходятся настолько медленно, что практически их трудно использовать.

Описанные в § 2 алгоритмы нахождения наилучших среднеквадратичных приближений намного проще, чем известные алгоритмы нахождения наилучших равномерных приближений. По всем указанным причинам на практике много удобнее искать наилучшие среднеквадратичные, а не равномерные приближения; как отмечалось в § 2, для улучшения их сходимости следует явно выделять в простой форме основные особенности функции и ее младших производных и аппроксимировать оставшуюся достаточно гладкую часть. К нахождению равномерных приближений прибегают в основном при разработке алгоритмов для стандартных программ вычисления функций, когда добиваются очень высокой точности при минимальном числе членов суммы.

2. Нахождение равномерного приближения. Для функции, заданной на отрезке $[a, b]$, не найдено способа определения коэффициентов наилучшего равномерного приближения за конечное число действий. Рассмотрим простой итерационный процесс нахождения коэффициентов.

Чебышевскую норму можно рассматривать как предел $\|\cdot\|_{L_p}$ при $p\to\infty$ и единичном весе. В пространстве L_p задачу нахождения наилучшего приближения $\|y-\varphi\|_{L_p}^p=\min$ удобно решать итерированием веса:

$$\int_a^b \rho^{(s)}(x)\,[y(x)-\varphi^{(s+1)}(x)]^2\,dx=\min,$$
$$\rho^{(s)}(x)=|y(x)-\varphi^{(s)}(x)|^{p-2}; \tag{59}$$

для начала итерационного процесса можно положить $\rho^{(0)}(x)\equiv 1$. Если $\varphi(x)$ является обобщенным многочленом, то на каждой итерации задача на минимум опять сводится к решению системы линейных (относительно коэффициентов a_k) уравнений. Для решения полной задачи $\|y-\varphi\|_C=\min$ надо выбрать последовательность $p\to\infty$, для каждого фиксированного p провести итерации (59) до сходимости, а затем в коэффициентах $a_k^{(p)}$ произвести предельный переход при $p\to\infty$ (т. е. оценить, начиная с какого p_0 коэффициенты перестают меняться в пределах заданной точности при дальнейшем увеличении p).

Двойной предельный переход требует больших численных расчетов. Поэтому целесообразно объединить предельные переходы $s\to\infty$ и $p\to\infty$. Для этого на первой итерации по s положим $p=2$, на второй возьмем $p=4$, на третьей — $p=6$ и т. д.

Вместо (59) получим следующую задачу:

$$\int_a^b \rho^{(s)}(x)\,[y(x)-\varphi^{(s+1)}(x)]^2\,dx=\min,$$
$$\rho^{(s)}(x)=[y(x)-\varphi^{(s)}(x)]^{2s},\quad s=0,\ 1,\ 2,\ \ldots \tag{60}$$

Здесь начальное условие для итераций $\rho^{(0)}(x)\equiv 1$ получается естественно при $s=0$. Этот итерационный процесс не исследован теоретически и мало опробован в практических расчетах, но поскольку обычно коэффициенты аппроксимации слабо зависят от выбора веса, то следует ожидать быстрой сходимости процесса.

ЗАДАЧИ

1. Доказать, что разделенная разность n-го порядка выражается через узловые значения функции следующим образом:

$$y(x_0,\ x_1,\ \ldots,\ x_n)=\sum_{k=0}^{n} y(x_k)\prod_{\substack{i=0\\ i\neq k}}^{n}(x_k-x_i)^{-1}.$$

2. Вывести оценку (11).

3. Написать оценки погрешности типа (11) для трех случаев интерполяционного многочлена Эрмита 7-й степени: $\mathcal{P}(x;\ x_0,\ x_1,\ \ldots,\ x_7)$, $\mathcal{P}(x;\ x_0,\ x_0,\ x_1,\ x_1,\ x_2,\ x_2,\ x_3,\ x_3)$ и $\mathcal{P}(x;\ x_0,\ x_0,\ x_0,\ x_0,\ x_1,\ x_1,\ x_1,\ x_1)$; сравнить их порядки точности и численные коэффициенты.

4. Применить формулу (19) к вычислению $y(0,\ 5)$ в таблице 5; оценить точность.

5. Вывести формулы типа (19) для случаев, когда функция на малых отрезках приближенно представима в виде $y(x)\approx ax^b$ или $y(x)\approx a(x+b)^m$, где m — заданное число.

6. Разобрать интерполяцию сплайном второй степени; по аналогии со случаем $n=3$ найти экономный способ вычисления коэффициентов.

7. Оценить погрешность округления при вычислении $\sin 2550^\circ$ по формуле Тейлора на ЭВМ с 16 десятичными знаками.

8. Доказать, что прямая, проведенная методом наименьших квадратов, проходит через точку с координатами

$$\bar{x}=\left(\sum_i \rho_i x_i\right)\Big/\left(\sum_i \rho_i\right),\quad \bar{y}=\left(\sum_i \rho_i y_i\right)\Big/\left(\sum_i \rho_i\right),$$

которая является «центром тяжести».

9. Вывести формулы Бесселя (44) для случая, когда тригонометрические функции заданы в действительной форме: $\varphi_0=1$, $\varphi_1=\sin x$, $\varphi_2=\cos x$, $\varphi_3=\sin 2x$ и т. д.

10. Вывести формулы сглаживания типа (45) для центральной точки по пяти точкам при среднеквадратичной аппроксимации многочленом первой и второй степени.

11. Написать систему уравнений для определения коэффициентов $a_k^{(s)}$, $b_q^{(s)}$, минимизирующих (49).

12. Доказать, что коэффициенты a_k формул Бесселя (44) связаны с коэффициентами обычного ряда Фурье α_k соотношениями $a_k=\sum_{p=-\infty}^{+\infty}\alpha_{k+pN}$.

ГЛАВА III

ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

В главе III рассмотрено численное дифференцирование функции, заданной на некоторой сетке. Введены квазиравномерные сетки, полезные во многих приложениях. Обсуждена некорректность задачи дифференцирования, проявляющаяся при сильном уменьшении шага, и изложены некоторые способы регуляции. Показано, как можно повышать точность и оценивать погрешность при сгущении сетки.

1. Полиномиальные формулы. Численное дифференцирование применяется, если функцию $y(x)$ трудно или невозможно продифференцировать аналитически — например, если она задана таблицей. Оно нужно также при решении дифференциальных уравнений при помощи разностных методов.

При численном дифференцировании функцию $y(x)$ аппроксимируют легко вычисляемой функцией $\varphi(x; \boldsymbol{a})$ и приближенно полагают $y'(x) = \varphi'(x; \boldsymbol{a})$. При этом можно использовать различные способы аппроксимации, изложенные в главе II. Сейчас мы рассмотрим простейший случай — аппроксимацию интерполяционным многочленом Ньютона (2.8). Вводя обозначение $\xi_i = x - x_i$, запишем этот многочлен и продифференцируем его почленно:

$$\varphi(x) = y(x_0) + \xi_0 y(x_0, x_1) + \xi_0\xi_1 y(x_0, x_1, x_2) + \\ + \xi_0\xi_1\xi_2 y(x_0, x_1, x_2, x_3) + \dots$$

$$\varphi'(x) = y(x_0, x_1) + (\xi_0 + \xi_1)\, y(x_0, x_1, x_2) + \\ + (\xi_0\xi_1 + \xi_0\xi_2 + \xi_1\xi_2)\, y(x_0, x_1, x_2, x_3) + \dots,$$

$$\varphi''(x) = 2y(x_0, x_1, x_2) + 2(\xi_0 + \xi_1 + \xi_2)\, y(x_0, x_1, x_2, x_3) + \dots$$

Общая формула имеет следующий вид:

$$\varphi^{(k)}(x) = k!\left[y(x_0, x_1, \dots, x_k) + \left(\sum_{i=0}^{k} \xi_i\right) y(x_0, x_1, \dots, x_{k+1}) + \right. \\ + \left(\sum_{i>j\geqslant 0}^{i=k+1} \xi_i\xi_j\right) y(x_0, x_1, \dots, x_{k+2}) + \\ \left. + \left(\sum_{i>j>l\geqslant 0}^{i=k+2} \xi_i\xi_j\xi_l\right) y(x_0, x_1, \dots, x_{k+3}) + \dots\right]. \quad (1)$$

Обрывая ряд на некотором числе членов, получим приближенное выражение для соответствующей производной. Наиболее простые выражения получим, оставляя в формуле (1) только первый член:

$$y'(x) \approx y(x_0, x_1) = [y(x_0) - y(x_1)] / (x_0 - x_1),$$
$$\frac{1}{2} y''(x) \approx y(x_0, x_1, x_2) = \frac{1}{x_0 - x_2} \left(\frac{y_0 - y_1}{x_0 - x_1} - \frac{y_1 - y_2}{x_1 - x_2} \right), \qquad (2)$$
$$\frac{1}{k!} y^{(k)}(x) \approx y(x_0, x_1, \ldots, x_k) = \sum_{p=0}^{k} y_p \prod_{\substack{i=0 \\ i \neq p}}^{k} (x_p - x_i)^{-1}.$$

При написании последней формулы использованы результаты задачи 1 к главе II. Все формулы (1) — (2) рассчитаны на произвольную неравномерную сетку.

Исследование точности полученных выражений при численных расчетах удобно делать при помощи апостериорной оценки, по скорости убывания членов ряда (1). Если шаг сетки достаточно мал, то погрешность близка к первому отброшенному члену. Пусть мы используем узлы x_i, $0 \leqslant i \leqslant n$. Тогда первый отброшенный член содержит разделенную разность $y(x_0, x_1, \ldots, x_{n+1})$, которая согласно (2) примерно равна $y^{(n+1)}(x) / (n+1)!$. Перед ней стоит сумма произведений различных множителей ξ_i; каждое произведение содержит $n+1-k$ множителей, а вся сумма состоит из C_{n+1}^k слагаемых. Отсюда следует оценка погрешности формулы (1) с $n+1$ узлами:

$$R_n^{(k)} \lesssim \frac{M_{n+1}}{(n+1-k)!} \max_i |\xi_i|^{n+1-k}, \qquad M_{n+1} = \max |y^{(n+1)}|. \qquad (3)$$

В частности, если сетка равномерная, то $\max |\xi_i| < nh$, откуда

$$R_n^{(k)} < M_{n+1} \left(\frac{en}{n+1-k} h \right)^{n+1-k} = O(h^{n+1-k}). \qquad (4)$$

Эти оценки можно несколько улучшить за счет более детального рассмотрения множителей ξ_i. Заметим, что строгое априорное исследование погрешности формулы (1), аналогичное выводу остаточного члена многочлена Ньютона в форме Коши (2.10), для произвольного расположения узлов приводит к той же оценке (3).

Таким образом, *порядок точности формулы* (1) *по отношению к шагу сетки равен числу оставленных в ней членов*, или, что то же самое, он равен числу узлов интерполяции минус порядок производной. Поэтому минимальное число узлов, необходимое для вычисления k-й производной, равно $k+1$; оно приводит к формулам (2) и обеспечивает первый порядок точности. Эти выводы соответствуют общему принципу: при почленном дифференцировании ряда скорость его сходимости уменьшается.

В главе II рекомендовалось использовать в формулах интерполяции не более 4—6 узлов. Если еще учесть ухудшение сходимости ряда при дифференцировании, то можно сделать вывод: даже если функция задана хорошо составленной таблицей на довольно подробной сетке, то практически численным дифференцированием можно хорошо определить первую и вторую производные, а третью и четвертую — лишь удовлетворительно. Более высокие производные редко удается вычислить с приемлемой точностью.

Замечание 1. Кубическая сплайновая интерполяция (2.20) обладает тем свойством, что первая и вторая производные интерполяционного многочлена всюду непрерывны. Обычно дифференцирование кубического сплайна позволяет определить эти производные с хорошей точностью. Если надо вычислить более высокие производные, то целесообразно строить сплайны высоких порядков. Из-за большой трудоемкости этот способ редко используется; теоретически он мало исследован.

Замечание 2. Если табулирована не только функция, но и ее производные, то следует составлять и дифференцировать интерполяционный многочлен Эрмита. Производные при этом вычисляются намного точнее, чем при дифференцировании интерполяционного многочлена Ньютона с тем же числом свободных параметров по формулам (1).

2. Простейшие формулы. Чаще всего используются равномерные сетки, на которых вид формул (1) заметно упрощается, а точность нередко повышается.

Рассмотрим сначала причину повышения точности. Остаточный член общей формулы (1) есть многочлен $\sum\prod(x-x_i)$ степени $n+1-k$ относительно x. Если x равен корню этого многочлена, то главный остаточный член обращается в нуль, т. е. в этой точке формула имеет порядок точности на единицу больше, чем согласно оценке (4). Эти *точки повышенной точности* будем обозначать $x_k^{(p)}$, где k — порядок производной, а $p=n+1-k$ — число оставленных в формуле (1) членов. Очевидно, p-членная формула имеет p точек повышенной точности.

У одночленной формулы (2) для k-й производной точка повышенной точности на произвольной сетке определяется условием $\sum\xi_i=\sum(x-x_i)=0$, что дает

$$x_k^{(1)}=(x_0+x_1+\ldots+x_k)/(k+1); \tag{5}$$

в этой точке одночленная формула имеет погрешность $O(h^2)$ вместо обычной $O(h)$. Для двучленной формулы задача нахождения точек повышенной точности приводит к квадратному уравнению, корни которого действительны, но формула для их нахождения громоздка (см. задачу 2). Если $p>2$, то найти точки

повышенной точности очень сложно, за исключением одного частного случая, который мы сейчас рассмотрим.

Пусть p нечетно, а узлы в формуле (1) *выбраны так, что они расположены симметрично относительно точки x; тогда x является одной из точек повышенной точности $x_k^{(p)}$.*

Доказательство. В самом деле, при этом величины $\xi_i = x - x_i$ имеют попарно равные абсолютные величины, но противоположные знаки. В остаточном члене множитель $w = \sum \prod \xi_i$ имеет нечетную степень, и при одновременном изменении знаков всех ξ_i он должен изменить знак. Но поскольку одновременное изменение знаков ξ_i сводится при таком расположении узлов лишь к перемене их нумерации, то величина w должна сохраниться, что возможно только при $w = 0$. Утверждение доказано.

Замечание 1. Доказательство справедливо для неравномерной сетки.

Замечание 2. Число узлов предполагалось произвольным; очевидно, симметричное расположение узлов относительно точки $x_k^{(p)}$ означает, что при нечетном числе узлов точка $x_k^{(p)}$ совпадает с центральным узлом, а при четном — лежит между средними узлами.

Замечание 3. Повышение точности достигается не только в самих точках повышенной точности, но и в достаточно малой их окрестности, где изменение производной не превышает погрешности формулы; для точки $x_k^{(1)}$ это окрестность размером $O(h^2)$, для $x_k^{(2)} - O(h^3)$ и т. д.

На произвольной сетке условие симметрии реализуется только в исключительных случаях. Но если сетка равномерна, то каждый ее узел симметрично окружен соседними узлами. Это позволяет составить несложные формулы хорошей точности для вычисления производных в узлах сетки.

Например, возьмем три соседних узла x_0, x_1, x_2 и вычислим первую и вторую производные в среднем узле. Выражая в одночленных формулах (2) разделенные разности через узловые значения функции, легко получим

$$y'(x_1) = (y_2 - y_0)/2h + O(h^2), \qquad h = x_{i+1} - x_i = \text{const}, \tag{6}$$

$$y''(x_1) = (y_2 - 2y_1 + y_0)/h^2 + O(h^2). \tag{7}$$

Формулу (6) часто записывают в несколько ином виде, удобном для определения производной в средней точке интервала сетки:

$$\begin{aligned} y'_{i+1/2} &\equiv y'(x_{i+1/2}) = (y_{i+1} - y_i)/h + O(h^2), \\ x_{i+1/2} &= x_i + {}^1/_2 h. \end{aligned} \tag{8}$$

Аналогично можно вывести формулы более высокого порядка точности или для более высоких производных. Например, трех-

членная формула (1) для первой производной в середине интервала по четырем соседним узлам дает

$$y'_{5/2} = (-y_3 + 27y_2 - 27y_1 + y_0)/(24h) + O(h^4), \qquad (9)$$

а для второй производной в центральном узле по пяти узлам

$$y''_2 = (-y_4 + 16y_3 - 30y_2 + 16y_1 - y_0)/(12h^2) + O(h^4). \qquad (10)$$

Все формулы (6) — (10) имеют четный порядок точности. Заметим, что все эти формулы написаны для случая равномерной сетки; применение их на произвольной неравномерной сетке для первой производной приводит к низкой точности $O(h)$, а для второй производной — к грубой ошибке.

На равномерной сетке для априорной оценки точности формул часто применяют способ разложения по формуле Тейлора — Маклорена. Предположим, например, что функция $y(x)$ имеет непрерывную четвертую производную, и выразим значения функции в узлах $x_{i\pm1}$ через значения функции и ее производных в центре симметрии узлов (в данном случае этим центром является узел x_i):

$$y(x_{i\pm1}) = y(x_i \pm h) = y_i \pm hy'_i + \frac{1}{2}h^2y''_i \pm \frac{1}{6}h^3y'''_i + \\ + \frac{1}{24}h^4y^{IV}(\eta_\pm), \qquad y(x_i) = y_i, \quad (11)$$

где η_+ есть некоторая точка интервала (x_i, x_{i+1}), а η_- есть некоторая точка интервала (x_{i-1}, x_i). Подставляя эти разложения во вторую разность, стоящую в правой части формулы (7) для второй производной, получим

$$\frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}) = y''_i + \frac{h^2}{24}[y^{IV}(\eta_+) + y^{IV}(\eta_-)] = y''_i + O(h^2). \quad (12)$$

Это подтверждает ранее сделанную оценку и уточняет величину остаточного члена, который оказался равным $h^2y^{IV}(\eta)/12$. Такой способ получения остаточного члена проще, чем непосредственное вычисление по формуле (1). Особенно часто он применяется при исследовании аппроксимации разностных схем (см. главу IX).

3. Метод Рунге — Ромберга. При вычислении одной и той же величины формулы с бо́льшим числом узлов дают более высокий порядок точности, но они более громоздки. Для оценки их точности надо привлекать дополнительный узел, что требует еще более сложных вычислений. Рассмотрим более простой способ получения высокого порядка точности.

Из формулы (12) видно, что погрешность простейшей формулы (7) для четырежды дифференцируемой функции имеет вид $R = h^2\psi(\eta)$, где η — некоторая точка вблизи узла x_i. Если $y^{IV}(x)$ липшиц-непрерывна, то оценку нетрудно уточнить: $R = h^2\psi(x_i) +$

$+O(h^3)$. Пусть в общем случае имеется некоторая приближенная формула $\zeta(x, h)$ для вычисления величины $z(x)$ по значениям на равномерной сетке с шагом h, а остаточный член этой формулы имеет следующую структуру:

$$z(x) - \zeta(x, h) = \psi(x) h^p + O(h^{p+1}). \quad (13)$$

Произведем теперь расчет по той же приближенной формуле для той же точки x, но используя равномерную сетку с другим шагом rh. Тогда получим значение $\zeta(x, rh)$, связанное с точным значением соотношением

$$z(x) - \zeta(x, rh) = \psi(x)(rh)^p + O((rh)^{p+1}). \quad (14)$$

Заметим, что $O((rh)^{p+1}) \approx O(h^{p+1})$. Имея два расчета на разных сетках, нетрудно оценить величину погрешности. Для этого вычтем (13) из (14) и получим *первую формулу Рунге*:

$$R \approx \psi(x) h^p = \frac{\zeta(x, h) - \zeta(x, rh)}{r^p - 1} + O(h^{p+1}). \quad (15)$$

Первое слагаемое справа есть главный член погрешности. Таким образом, расчет по второй сетке позволяет оценить погрешность расчета на первой сетке (с точностью до членов более высокого порядка).

Можно исключить найденную погрешность (15) из формулы (13) и получить результат с более высокой точностью по *второй формуле Рунге*:

$$z(x) = \zeta(x, h) + \frac{\zeta(x, h) - \zeta(x, rh)}{r^p - 1} + O(h^{p+1}). \quad (16)$$

Этот метод оценки погрешности и повышения точности результата очень прост, применим в большом числе случаев и исключительно эффективен. Рассмотрим два примера его применения к численному дифференцированию.

Таблица 7

x	$y = \lg x$
1	0,000
2	0,301
3	0,478
4	0,602
5	0,699

Пример 1. Пусть функция $y(x) = \lg x$ задана таблицей 7 и требуется вычислить $y'(3)$. Выберем для вычислений простейшую формулу (6). Полагая $h = 1$, т. е. производя вычисления по точкам $x = 2$ и $x = 4$, получим $y'(3) \approx 0{,}151$. Увеличивая шаг вдвое ($r = 2$), т. е. вычисляя производную по точкам $x = 1$ и $x = 5$, получим $y'(3) \approx 0{,}175$. Проводя вычисления по формуле Рунге (16), где согласно оценке (6) берется $p = 2$, получим уточненное значение $y'(3) \approx 0{,}143$; это всего 2% отличается от искомого значения $y'(3) \approx 0{,}145$.

Пример 2. Выведем формулу высокой точности из формулы низкой точности. Возьмем простейшую формулу для вычисления

первой производной в середине интервала (8) и запишем ее, выбирая сначала соседние узлы, а затем более удаленные:

$$y'_{3/2}(h) \approx (y_2 - y_1)/h, \quad y'_{3/2}(3h) \approx (y_3 - y_0)/3h.$$

Порядок точности формулы $p=2$, а коэффициент увеличения шага $r=3$, поэтому уточнение методом Рунге дает формулу (9):

$$y'_{3/2} \approx y'_{3/2}(h) + \frac{1}{8}[y'_{3/2}(h) - y'_{3/2}(3h)] = \frac{1}{24h}(y_0 - 27y_1 + 27y_2 - y_3).$$

Отсюда видно, что для получения высокого порядка точности не обязательно производить вычисления непосредственно по формулам высокого порядка точности; можно произвести вычисления по простым формулам низкой точности на разных сетках и затем уточнить результат методом Рунге. Последний способ предпочтительней еще потому, что величина поправки (15) дает апостериорную оценку точности.

Метод Рунге обобщается на случай произвольного числа сеток. Пусть функция $y(x)$ имеет достаточно высокие непрерывные производные. Тогда в разложениях Тейлора типа (11) можно удерживать большое число членов и подстановка их в формулы типа (6)—(10) приводит к представлению остаточного члена в виде ряда

$$z(x) - \zeta(x, h) = \sum_{m \geqslant p} \psi_m(x) h^m. \tag{17}$$

Пусть расчет проведен на q различных сетках с шагами h_j, $1 \leqslant j \leqslant q$. Тогда из остаточного члена можно исключить первые $q-1$ слагаемых. Для этого перепишем соотношение (17), оставляя первые $q-1$ члены погрешности:

$$z(x) - \sum_{m=p}^{p+q-2} \psi_m(x) h_j^m = \zeta(x, h_j) + O(h^{p+q-1}), \quad 1 \leqslant j \leqslant q.$$

Это система линейных уравнений относительно величин $z(x)$, $\psi_m(x)$. Решая ее по правилу Крамера, получим уточненное значение по *формуле Ромберга*

$$z(x) = \begin{vmatrix} \zeta(x, h_1) & h_1^p & h_1^{p+1} & \dots & h_1^{p+q-2} \\ \zeta(x, h_2) & h_2^p & h_2^{p+1} & \dots & h_2^{p+q-2} \\ \dots & \dots & \dots & \dots & \dots \\ \zeta(x, h_q) & h_q^p & h_q^{p+1} & \dots & h_q^{p+q-2} \end{vmatrix} \times \begin{vmatrix} 1 & h_1^p & h_1^{p+1} & \dots & h_1^{p+q-2} \\ 1 & h_2^p & h_2^{p+1} & \dots & h_2^{p+q-2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & h_q^p & h_q^{p+1} & \dots & h_q^{p+q-2} \end{vmatrix}^{-1} + O(h^{p+q-1}). \tag{18}$$

Эта формула приводит к повышению порядка точности результата

на $q-1$ по сравнению с исходной формулой $\zeta(x, h)$, т. е. каждая лишняя сетка позволяет повысить порядок точности на единицу.

Формула Ромберга удобна тем, что ее можно применять при любом числе равномерных сеток и любом соотношении их шагов. Ее недостатками являются сравнительная громоздкость и отсутствие в промежуточных выкладках апостериорных оценок точности. Если сетки выбраны так, что сгущение сеток происходит всегда в одно и то же число раз (т. е. $h_j = rh_{j-1} = \ldots = r^{j-1}h_1$), то вместо формулы Ромберга удобнее рекуррентно применять метод Рунге.

Для этого берут последовательные пары сеток (h_1, h_2), (h_2, h_3), (h_3, h_4) и т. д. По каждой паре производят уточнение методом Рунге, исключая тем самым главный член погрешности $\psi_p(x)\,h^p$. Поэтому в уточненных величинах главный член погрешности будет иметь вид $\tilde{\psi}_{p+1}(x)\,h^{p+1}$, где шаг можно условно принять для первой пары сеток за h_1, для второй — за h_2 и т. д. (это верно, только если h_j/h_{j-1} одинаково для всех пар сеток). Уточненные значения таким же образом группируют в пары и исключают ошибку следующего порядка $O(h^{p+1})$. Всего можно произвести $q-1$ уточнение, на единицу меньше числа сеток. При каждом уточнении вычисляется погрешность (15), дающая апостериорную оценку точности на данном этапе вычислений. Пример такого вычисления будет дан в главе IV.

Замечание 1. Если исходная формула для вычисления $\zeta(x, h)$ имеет симметричный вид, то на равномерной сетке обычно все нечетные члены ряда (17) обращаются в нуль. При этом пользоваться общей формулой (18) можно, но невыгодно, ибо она не учитывает дополнительной информации о нулевых коэффициентах. Следует оставить в сумме (17) только степени h^p, h^{p+2}, h^{p+4}, ... и соответственно изменить формулу Ромберга. Аналогично изменяется рекуррентная процедура Рунге: при очередном исключении ошибки порядок точности повышается не на 1, а на 2. Примером может служить данный выше вывод формулы (9) из формулы (8), когда после первого уточнения погрешность уменьшилась с $O(h^2)$ сразу до $O(h^4)$.

Замечание 2. Допустимое число членов суммы (17) связано с количеством существующих у функции непрерывных производных. Поэтому для недостаточно гладких функций бессмысленно брать большое число сеток. Практически даже для «хороших» функций используют не более 3 — 5 сеток; обычно отношение r их шагов стараются выбрать равным 2.

Замечание 3. Метод Рунге — Ромберга можно применять только в том случае, если ошибка представима в виде (17), где коэффициенты $\psi_m(x)$ одинаковы для *всех* сеток. Строго говоря, при численном дифференцировании эти коэффициенты зависят от положения узлов сетки. Но если выбранные конфигурации узлов

на всех сетках подобны относительно точки x (рис. 14, *а*), то зависимость от узлов одинакова для всех сеток и сводится к величине шага. Тогда метод Рунге — Ромберга применим. Если же правило подобия нарушено (рис. 14, *б*, *в*), то метод применять нельзя.

Поэтому при численном дифференцировании метод Рунге—Ромберга удается применять только для нахождения производных в узлах или серединах интервалов равномерных (или описанных далее квазиравномерных) сеток. Но эти случаи являются достаточно важными в практических приложениях. Особенно широко применяется описанный метод при численном интегрировании и разностных методах решения задач для дифференциальных и интегральных уравнений.

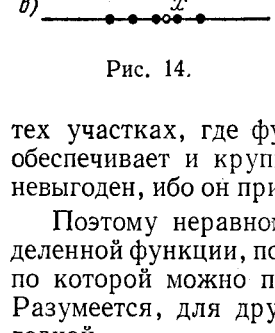


Рис. 14.

4. Квазиравномерные сетки. При тех значениях аргумента, где функция резко меняется, шаг таблицы должен быть малым, иначе точность вычисления по этой таблице будет плохой. А на тех участках, где функция меняется медленно, хорошую точность обеспечивает и крупный шаг таблицы; мелкий шаг при этом даже невыгоден, ибо он приводит к сильному увеличению объема таблицы.

Поэтому неравномерная сетка, удачно подобранная для определенной функции, позволяет построить таблицу небольшого объема, по которой можно производить вычисления с хорошей точностью. Разумеется, для других функций эта сетка может быть малопригодной.

Каждая конкретная сетка либо равномерна (т. е. ее шаг $h_i = x_{i+1} - x_i$ постоянен), либо неравномерна. Но нам нередко приходится сгущать сетку, т. е. рассматривать на $[a, b]$ последовательность сеток $x_i^{(N)}$, $0 \leqslant i \leqslant N$, с возрастающим числом интервалов N. Разумеется, если таблица уже задана, то сетку сгущать невозможно, но можно ее разреживать, выбрасывая из таблицы половину, две трети и т. д. точек. Это также является некоторым способом построения последовательности сеток. Сгущение сеток широко применяется при численном решении дифференциальных и интегральных уравнений.

Среди последовательностей сеток важное место занимают *квазиравномерные сетки*. Будем называть сетки квазиравномерными, если существует дважды непрерывно дифференцируемая функция $x = \xi(t)$, преобразующая отрезок $0 \leqslant t \leqslant 1$ в отрезок $a \leqslant x \leqslant b$ так, что каждой сетке $x_i^{(N)}$ соответствует равномерная сетка $t_i^{(N)} = i/N$, причем на этом отрезке $\xi'(t) \geqslant \varepsilon > 0$, а $\xi''(t)$ ограничена.

Если эти условия выполнены, то шаг сетки $h_i \approx \xi'(t_i)/N$, а разность двух соседних шагов есть $h_i - h_{i-1} \approx \xi''(t_i)/N^2$. Значит,

при большом числе узлов разность соседних шагов $\Delta h \sim h^2$, т. е. много меньше длины шага, и соседние интервалы почти равны. Поэтому такие сетки и называют квазиравномерными или почти равномерными. Однако отношение длин далеких друг от друга интервалов $h_i/h_j \approx \xi'(t_i)/\xi'(t_j)$ может быть большим.

Для того чтобы сгустить квазиравномерную сетку x_i, надо сгустить равномерную сетку t_i (увеличить N) и по ней вычислить новую сетку. Середину интервала $x_{i+1/2}$ квазиравномерной сетки надо вычислять при помощи того же преобразования, полагая

$$x_{i+1/2} = \xi\left(\frac{i+1/2}{N}\right);$$

брать $x_{i+1/2}$ равной полусумме соседних узлов x_i, x_{i+1} нельзя.

Рассмотрим некоторые примеры.

а) Если надо детально передать поведение функции вблизи одного из концов отрезка $[a, b]$, то удобно преобразование

$$x = a + (b-a)(e^{\alpha t} - 1)/(e^{\alpha} - 1). \tag{19}$$

Значение $\alpha > 0$ соответствует малому шагу сетки у левого конца отрезка, значение $\alpha < 0$ — у правого. Шаги этой сетки составляют геометрическую прогрессию со знаменателем $q = h_{i+1}/h_i = e^{\alpha/N}$. Отношение первого и последнего шага сетки примерно равно e^{α}; при большом α оно может быть очень большим. Такая сетка полезна, например, в задачах атомной физики, где волновые функции наиболее быстро меняются вблизи ядра.

б) На полупрямой $[a, \infty)$ тоже можно построить квазиравномерную сетку; например, таким преобразованием:

$$x = a + \alpha \operatorname{tg}(\pi t/2), \quad 0 \leqslant t < 1. \tag{20}$$

Параметр α управляет сеткой; чем он меньше, тем гуще узлы сетки при $x \to a$ и реже при $x \to \infty$. Последний интервал этой сетки (x_{N-1}, x_N) бесконечно велик, ибо точка x_N — бесконечно удаленная (отсюда ясно, что середину интервала квазиравномерной сетки надо находить при помощи основного преобразования!). Эта сетка полезна при вычислении интегралов с бесконечным верхним пределом.

в) Преобразование $x = \alpha \operatorname{tg}(\pi t/2)$ при $-1 < t < 1$ позволяет построить квазиравномерную сетку на бесконечной прямой. Первый и последний интервалы этой сетки бесконечны.

г) Преобразование $x = t^2$, $0 \leqslant t \leqslant 1$, определяет не квазиравномерную сетку. Здесь не выполнено условие строгой положительности $\xi'(t)$. По этому преобразованию строится такая сетка:

$$x_0 = 0, \quad x_1 = 1/N^2, \quad x_2 = 4/N^2, \ \ldots; \quad h_0 = 1/N^2, \quad h_1 = 3/N^2, \ \ldots$$

В результате разность двух соседних шагов — первого и второго — вдвое больше одного из них при любом N. Значит, вблизи точки $x = 0$ сетка не стремится к равномерной при $N \to \infty$.

Если сетка квазиравномерна, то производные на ней вычисляются либо проще, либо точнее, чем на произвольной неравномерной сетке. Например, если на такой сетке взять подряд три узла x_0, x_1, x_2, то $x_1 = (x_0 + x_2)/2 + (h_0 - h_1)/2 = (x_0 + x_2)/2 + O(h^2)$

и аналогично $x_1 = (x_0 + x_1 + x_2)/3 + O(h^2)$. Это означает, что узел x_1 расположен вблизи точки повышенной точности для этих узлов, в окрестности размером $O(h^2)$. Из сделанного в п. 2 замечания следует, что одночленные формулы (2), рассчитанные на произвольную сетку:

$$y_1' \approx y(x_0, x_2), \quad y_1'' \approx 2y(x_0, x_1, x_2),$$

в узлах квазиравномерной сетки обеспечивают точность $O(h^2)$. Пользоваться в этом случае формулами типа (7), рассчитанными на равномерную сетку, не следует — на квазиравномерной сетке их точность будет хуже.

На квазиравномерных сетках справедливо разложение остаточного члена в ряд (17), если порождающее сетки преобразование $x = \xi(t)$ достаточное число раз непрерывно дифференцируемо. В этом случае для повышения точности расчетов можно употреблять метод Рунге — Ромберга, подставляя в формулы (16) — (18) вместо h величину $1/N$. Для квазиравномерных сеток этот метод особенно выгоден, ибо для них прямые формулы высокого порядка точности очень громоздки.

Только в одном пункте квазиравномерные сетки уступают равномерным. На них ряд для остаточного члена (17) даже в случае симметричной формулы содержит обычно все степени $1/N$, поэтому каждая лишняя сетка позволяет повысить порядок точности лишь на единицу, а не на двойку.

Квазиравномерные сетки часто используют при решении сложных задач математической физики, когда необходимо при малом числе узлов детально передать особенности решения.

5. Быстропеременные функции. Если функция (точнее, ее разделенные разности) значительно меняется на протяжении нескольких интервалов сетки, то интерполяция обобщенным многочленом обычно недостаточно точна для дифференцирования этой функции. Для таких функций особенно полезна квазилинейная интерполяция, производимая при помощи выравнивающих переменных.

Если $\xi(x)$, $\eta(y)$ — выравнивающие переменные, то для искомой производной справедливо соотношение

$$y_x' = \xi_x' \eta_\xi' / \eta_y'. \tag{21}$$

Выравнивающие преобразования подбирают несложными, чтобы их производные ξ_x', η_y' находились точно. Остается только численно найти η_ξ' способами, изложенными в предыдущих пунктах.

Например, пусть имеются таблицы энергии $E(T, \rho)$ многократно ионизованной плазмы тяжелых атомов. Рассмотрим нахождение теплоемкости $c_v = (\partial E/\partial T)_v$; она отличается от теплоемкости идеального газа, поскольку в нее входит энергия, идущая на отрыв от ионного остова новых электронов при повышении температуры. Ранее упоминалось, что зависимость $E(T)$ напоминает степенную со слабо переменным показателем и выравнивающим является пре-

образование $\eta = \ln E$, $\xi = \ln T$. Легко видеть, что формула (21) принимает вид

$$c_v = (E/T)\,\eta'_\xi = (E/T)\,[\partial(\ln E)/\partial(\ln T)];$$

последнюю производную находят численным дифференцированием (см. задачу 7).

Если в исходных переменных сетка была равномерной или квазиравномерной, то обычно она квазиравномерна и в выравнивающих переменных, ибо выравнивающее преобразование на ограниченном отрезке почти всегда обладает требуемыми свойствами производных. В этом случае результат можно уточнять методом Рунге — Ромберга.

Формула двукратного дифференцирования при помощи выравнивающих переменных достаточно сложна:

$$y''_{xx} = [\xi''_{xx}\eta'_x + (\xi'_x)^2\,\eta''_{\xi\xi} - \eta''_{yy}\,(\xi'_x\eta'_\xi/\eta'_y)^2]/\eta'_y, \qquad (22)$$

и ее применение не всегда обеспечивает хорошую точность. Но для быстропеременной функции двукратное дифференцирование интерполяционного многочлена Лагранжа еще более ненадежно. Поэтому вторую и более высокие производные быстропеременной функции трудно найти численно.

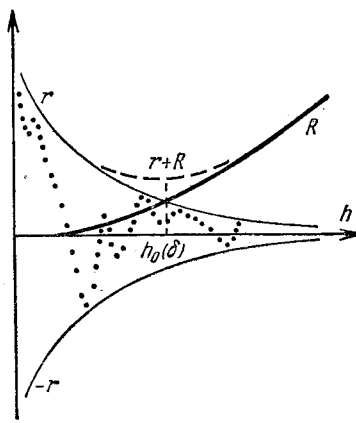


Рис. 15.

6. Регуляризация дифференцирования. При численном дифференцировании приходится вычитать друг из друга близкие значения функции. Это приводит к уничтожению первых значащих цифр, т. е. к потере части достоверных знаков числа. Если значения функции известны с малой точностью, то встает естественный вопрос — останется ли в ответе хоть один достоверный знак?

Для ответа на этот вопрос исследуем ошибки при численном дифференцировании. При интерполировании обобщенным многочленом производная k-го порядка определяется согласно (2)—(3) формулой типа

$$y^{(k)}(x) = h^{-k}\sum_q C_q(x)\,y(x_q) + R_k(x). \qquad C_q(x) = O(1). \qquad (23)$$

Если формула имеет порядок точности p, то, значит, ее остаточный член равен $R_k(x) \approx C(x)\,h^p$. Этот остаточный член определяет погрешность метода, и он неограниченно убывает при $h \to 0$. Его зависимость от шага изображена на рис. 15 жирной линией.

Но есть еще неустранимая погрешность, связанная с погрешностью функции $\delta y(x)$. Поскольку точный вид этой погрешности неизвестен, можно оценить только мажоранту неустранимой погрешности $r_k = \delta \cdot h^{-k} \sum_q |C_q|$; она неограниченно возрастает при $h \to 0$ (тонкая линия на рис. 15). Фактически же неустранимая погрешность будет нерегулярно зависеть от величины шага, беспорядочно осциллируя в границах, определяемых мажорантой (точки на рис. 15).

Пока шаг достаточно велик, при его убывании неустранимая погрешность мала по сравнению с погрешностью метода; поэтому полная погрешность убывает. При дальнейшем уменьшении шага неустранимая погрешность становится заметной, что проявляется в не вполне регулярной зависимости результатов вычислений от величины шага. Наконец, при достаточно малом шаге неустранимая погрешность становится преобладающей, и при дальнейшем уменьшении шага результат вычислений становится все менее достоверным.

Полная погрешность мажорируется суммой $R_k + r_k$ (штриховая кривая на рисунке). Оптимальным будет шаг, соответствующий минимуму этой кривой. Нетрудно подсчитать, что

$$h_0(\delta) = \left(k\delta \sum_q |C_q|/pC\right)^{1/(p+k)} = O\left(\delta^{1/(p+k)}\right),$$
$$\min(R_k + r_k) \doteq C h_0^p \left(1 + \frac{p}{k}\right) = O\left(\delta^{p/(p+k)}\right). \tag{24}$$

Меньший шаг невыгоден, а меньшая погрешность, вообще говоря, недостижима (хотя отдельные вычисления случайно окажутся более точными, но мы этого не сможем узнать). Эта минимальная ошибка тем меньше, чем меньше погрешность входных данных и порядок вычисляемой производной и чем выше порядок точности формулы.

Очевидно, *при* $\delta y(x) \to 0$ *можно получить сколь угодно высокую точность результата, если шаг стремится к нулю, будучи всегда не менее* $h_0(\delta)$. Но если допустить $h < h_0(\delta)$, то результат предельного перехода может быть неправильным.

Эта тонкость связана с некорректностью задачи дифференцирования. Рассмотрим погрешность входных данных вида $\delta y(x) = m^{-1} \sin m^2 x$. Она приводит к погрешности первой производной $\delta y'(x) = m \cos m^2 x$. При $m \to \infty$ погрешность функции в $\|\cdot\|_C$ неограниченно убывает, а погрешность производной в той же норме неограниченно растет. Значит, нет непрерывной зависимости производной от функции, т. е. *дифференцирование некорректно.* Особенно сильно это сказывается при нахождении производных высокого порядка.

Изложенный выше способ определения оптимального шага и запрещение вести расчет шагом меньше оптимального есть некоторый способ регуляризации дифференцирования, так называемая *регуляризация по шагу*. Этот способ в простейшей форме давно применялся физиками, которые при однократном численном дифференцировании всегда выбирали такой шаг, чтобы $|y(x+h) - y(x)| \geqslant \delta$.

К этой задаче применим и метод регуляризации А. Н. Тихонова; он будет изложен в главе XIV, § 2.

Физики издавна употребляют (без строгого обоснования) еще один способ регуляризации — дифференцирование предварительно сглаженной кривой, причем сглаживание обычно выполняют методом наименьших квадратов. Роль параметра регуляризации здесь играет отношение числа свободных параметров n аппроксимирующей кривой к числу узлов сетки N; для хорошего сглаживания должно выполняться условие $n \ll N$.

Рассмотрим, как это делается в простейшем случае. Выберем около искомой точки не очень большой интервал изменения аргумента, чтобы двучленная аппроксимация $y(x) \approx a + bx$ обеспечивала удовлетворительную точность. Но этот интервал должен содержать довольно много узлов сетки, т. е. быть не слишком малым.

Система уравнений (2.43) для определения коэффициентов среднеквадратичной аппроксимации принимает следующий вид:

$$\begin{aligned} a\sum \rho_i + b\sum \rho_i x_i &= \sum \rho_i y_i, \\ a\sum \rho_i x_i + b\sum \rho_i x_i^2 &= \sum \rho_i x_i y_i, \end{aligned} \tag{25}$$

где сумма берется по узлам сетки x_i, лежащим в этом интервале. Введем на этом интервале средние значения

$$\bar{x} = \Big(\sum \rho_i x_i\Big)\Big/\Big(\sum \rho_i\Big), \quad \bar{y} = \Big(\sum \rho_i y_i\Big)\Big/\Big(\sum \rho_i\Big). \tag{26}$$

Тогда первое уравнение (25) можно записать в виде $a + b\bar{x} = \bar{y}$ (см. задачу 8 к главе II). Умножая его на $\bar{x}\sum \rho_i$ и вычитая из второго уравнения (25), получим

$$y'(x) \approx b = \Big[\sum \rho_i (x_i y_i - \bar{x}\bar{y})\Big]\Big/\Big[\sum \rho_i (x_i^2 - \bar{x}^2)\Big]. \tag{27}$$

Пользуясь определением средних (26), произведем несложное преобразование знаменателя в (27):

$$\begin{aligned} \sum \rho_i (x_i^2 - \bar{x}^2) &= \sum \rho_i (x_i^2 + \bar{x}^2) - 2\bar{x}^2 \sum \rho_i = \\ &= \sum \rho_i (x_i^2 + \bar{x}^2) - 2\bar{x} \sum \rho_i x_i = \sum \rho_i (x_i - \bar{x})^2 \end{aligned}$$

и аналогично преобразуем числитель. Тогда выражение (27)

приводится к виду, напоминающему коэффициент парной корреляции величин x и y:

$$y'(x) \approx b = \left[\sum \rho_i (x_i - \bar{x})(y_i - \bar{y})\right] / \left[\sum \rho_i (x_i - \bar{x})^2\right]. \qquad (28)$$

Последняя формула несколько выгодней для численных расчетов, чем предыдущая, ибо ошибки округления в ней меньше.

Двучленная среднеквадратичная аппроксимация дает удовлетворительные результаты, только если $\delta/\sqrt{N} \ll \Delta$, где δ — погрешности отдельных значений функции, N — число точек в выбранном участке, а Δ — нелинейная часть приращения функции на данном участке. Если это соотношение нарушено, то надо строить сглаживающие аппроксимации с 3—4 членами и дифференцировать их.

В заключение отметим, что выравнивающие переменные позволяют вести расчет крупным шагом или с малым числом свободных параметров. Поэтому предварительное приведение к выравнивающим переменным существенно ослабляет влияние погрешности начальных данных и позволяет теми же способами регуляризации добиться большей точности.

ЗАДАЧИ

1. Составить формулу вычисления $y'(x)$ на основании интерполяционного многочлена Эрмита (2.18) и сравнить ее с простейшей формулой $y'(x) \approx y'(x_0) + (x - x_0)\, y'(x_0, x_1)$. Найти погрешности обеих формул. Какая из формул точнее и почему?

2. Показать, что у двучленной формулы (1) есть две точки повышенной точности, определяемые соотношением

$$x_k^{(2)} = \left[\sqrt{k+1} \sum_{i=0}^{k+1} x_i \pm \sqrt{\sum_{i>j\geqslant 0}^{i=k+1} (x_i - x_j)^2}\right] \Big/ \left[(k+2)\sqrt{k+1}\right],$$

в которых достигается третий порядок точности.

3. Аналогично (6)—(10) получить формулы для вычисления y^{III} и y^{IV} в среднем узле по пяти узлам равномерной сетки.

Таблица 8

T, эв	E, $\frac{кдж}{г}$
2,04	2250
1,15	720
0,646	303
0,363	176
0,204	64,8
0,115	24,8

4. Способом разложения по формуле Тейлора найти остаточные члены формул (8)—(10).

5. Получить для второй производной формулу высокой точности (10) из простейшей формулы (7) методом Рунге.

6. Строго обосновать рекуррентное применение метода Рунге.

7. В таблице 8 приведены данные по энергии плазмы алюминия при плотности 10^{19} атом/см^3. Составить таблицу теплоемкости c_v и оценить ее точность, полагая, что: а) значения энергии вычислены точно, б) значения энергии имеют погрешность $\pm 10\%$.

8. Используя среднеквадратичную аппроксимацию функции параболой $y(x) \approx a + bx + cx^2$, найти выражение для ее первой и второй производной через значения функции в узлах.

ГЛАВА IV

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

В главе IV изложены основные методы численного интегрирования. В § 1 выведены формулы вычисления однократных интегралов, основанные на полиномиальной аппроксимации подынтегральной функции: простейшие формулы трапеций и средних и некоторые формулы более высокой точности, в том числе формулы наивысшей алгебраической точности (Гаусса — Кристоффеля и Маркова). Исследованы погрешности этих формул и характер их сходимости. В § 2 рассмотрены способы интегрирования функций, для которых полиномиальная аппроксимация не обеспечивает приемлемой точности. В § 3 описанные методы перенесены на случай кратных интегралов. В § 4 изложены основы метода Монте-Карло применительно к вычислению интегралов.

§ 1. Полиномиальная аппроксимация

1. Постановка задачи. Пусть требуется найти определенный интеграл

$$F = \int_a^b f(x)\,\rho(x)\,dx, \qquad \rho(x) > 0, \tag{1}$$

где функция $f(x)$ непрерывна на отрезке $[a, b]$, а весовая функция $\rho(x)$ непрерывна на интервале (a, b). Выразить интеграл через элементарные функции удается редко, а компактный и удобный для доведения до числа ответ получается еще реже. Поэтому обычно заменяют $f(x)$ на такую аппроксимирующую функцию $\varphi(x, a) \approx f(x)$, чтобы интеграл от нее легко вычислялся в элементарных функциях.

Чаще всего $f(x)$ заменяют некоторым обобщенным интерполяционным многочленом. Поскольку такая аппроксимация линейна относительно параметров, то функция при этом заменяется некоторым линейным выражением, коэффициентами которого служат значения функции в узлах:

$$f(x) = \sum_{i=0}^{n} f(x_i)\,\varphi_i(x) + r(x), \tag{2}$$

где $r(x)$ — остаточный член аппроксимации. Подставляя (2) в (1),

получим формулу численного интегрирования (*квадратурную формулу*)

$$F = \sum_{i=0}^{n} c_i f(x_i) + R,$$

$$c_i = \int_a^b \varphi_i(x)\,\rho(x)\,dx, \qquad R = \int_a^b r(x)\,\rho(x)\,dx, \tag{3}$$

где величины x_i называют *узлами*, c_i — *весами*, а R — *погрешностью* или *остаточным членом* формулы. Интеграл приближенно заменяется суммой, похожей на интегральную сумму, причем узлы и коэффициенты этой суммы не зависят от функции $f(x)$. Интерполяционный многочлен (2) может быть не только лагранжева, но и эрмитова типа; в последнем случае в сумму (3) войдут производные функции в узлах.

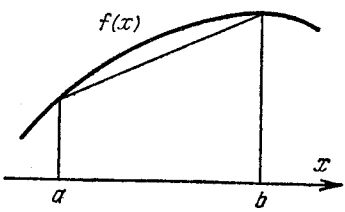


Рис. 16.

Лучше всего изучена замена $f(x)$ алгебраическим многочленом; она рассматривается в этом параграфе. Обычно будем полагать $\rho(x) \equiv 1$. Случаи не единичного веса будут особо оговариваться.

2. Формула трапеций. Заменим функцию на отрезке $[a, b]$ многочленом Лагранжа первой степени с узлами $x_0 = a$, $x_1 = b$. Это соответствует замене кривой на секущую. Искомый интеграл, равный площади криволинейной фигуры, заменяется на площадь трапеции (рис. 16); из геометрических соображений нетрудно написать для него *формулу трапеций*

$$F = \int_a^b f(x)\,dx \approx \frac{1}{2}(b-a)[f(a) + f(b)]. \tag{4}$$

Это одна из простейших квадратурных формул. Найдем ее погрешность. Для этого разложим $f(x)$ по формуле Тейлора, выбирая середину отрезка за центр разложения и предполагая наличие у функции требуемых по ходу рассуждений непрерывных производных:

$$f(x) = f(\bar{x}) + (x-\bar{x})f'(\bar{x}) + \frac{1}{2}(x-\bar{x})^2 f''(\bar{x}) + \ldots,$$
$$\bar{x} = \frac{1}{2}(a+b). \tag{5}$$

Погрешность есть разность точного и приближенного значений интеграла. Подставляя в (4) разложение (5), получим главный

член погрешности

$$R=\int_a^b f(x)\,dx-\frac{b-a}{2}[f(a)+f(b)]\approx-\frac{1}{12}(b-a)^3 f''(\bar{x}), \tag{6}$$

где члены, отброшенные при замене точного равенства приближенным, содержат старшие производные и более высокие степени длины отрезка интегрирования. Заметим, что содержащие $f(\bar{x})$ и $f'(\bar{x})$ члены разложения (5) уничтожились и не дали вклада в погрешность; это нетрудно было предвидеть, ибо формула трапеций по самому выводу точна для многочлена первой степени.

Вообще говоря, длина отрезка $b-a$ не мала, поэтому остаточный член (6) может быть велик. Для повышения точности на отрезке $[a, b]$ вводят достаточно густую сетку $a=x_0<x_1<x_2<\ldots$ $\ldots<x_N=b$. Интеграл разбивают на сумму интегралов по шагам сетки и к каждому шагу применяют формулу (4). Получают *обобщенную формулу трапеций*

$$\int_a^b f(x)\,dx\approx\frac{1}{2}\sum_{i=1}^{N}(x_i-x_{i-1})(f_{i-1}+f_i),$$
$$R\approx-\frac{1}{12}\sum_{i=1}^{N}(x_i-x_{i-1})^3 f''(\bar{x}_i). \tag{7}$$

На равномерной сетке она упрощается:

$$\int_a^b f(x)\,dx\approx h\left(\frac{1}{2}f_0+f_1+f_2+\ldots+f_{N-1}+\frac{1}{2}f_N\right),$$
$$R\approx-\frac{1}{12}\sum_{i=1}^{N}h^3 f''(\bar{x}_i)\approx-\frac{1}{12}h^2\int_a^b f''(x)\,dx, \tag{8}$$
$$h=x_i-x_{i-1}=\text{const}.$$

Поскольку в оценке (6) были отброшены члены, содержащие более высокие степени длины интервала, то выражение остаточного члена (8) является асимптотическим, т. е. выполняющимся при $h\to 0$ с точностью до членов более высокого порядка малости. Но для справедливости этой оценки необходимо существование непрерывной $f''(x)$; если $f''(x)$ кусочно-непрерывна, то удается сделать лишь мажорантную оценку

$$|R|\leqslant\frac{b-a}{12}h^2M_2,\qquad M_2=\max_{[a,\,b]}|f''(x)|. \tag{9}$$

Таким образом, обобщенная формула трапеций имеет второй порядок точности относительно шага сетки. На равномерной сетке это видно непосредственно, а на квазиравномерной сетке, порожденной преобразованием $x=\xi(t)$, остаточный член (7) можно привести к виду

$$R \approx -\frac{1}{12N^2}\int_a^b (\xi_t')^2 f''(x)\,dx, \tag{10}$$

если используемые в этой формуле производные непрерывны. Для произвольной неравномерной сетки асимптотическая оценка в виде суммы (7) справедлива, но неудобна для использования; можно пользоваться мажорантной оценкой (9), подразумевая под шагом $h=\max(x_i-x_{i-1})$.

3. Формула Симпсона. Вычислим интеграл по обобщенной формуле трапеций сначала на равномерной сетке с шагом h, а затем на сетке с вдвое более крупным шагом; вторая сетка получается из первой выбрасыванием узлов через один. Из вида остаточного члена (8) следует, что результат, полученный по формуле трапеций, можно уточнять методом Рунге. Проводя такое уточнение для отрезка, содержащего узлы x_0, x_1, x_2, получим формулу Симпсона

$$\begin{aligned} F &\approx \frac{1}{3}[4F_{\text{трап}}(h)-F_{\text{трап}}(2h)] = \\ &= \frac{1}{3}\left[4h\left(\frac{1}{2}f_0+f_1+\frac{1}{2}f_2\right)-2h\left(\frac{1}{2}f_0+\frac{1}{2}f_2\right)\right] = \\ &= \frac{1}{3}h(f_0+4f_1+f_2), \qquad h=x_i-x_{i-1}. \end{aligned} \tag{11}$$

Обобщенная формула Симпсона для равномерной сетки и четного числа шагов N имеет вид

$$F \approx \frac{h}{3}(f_0+4f_1+2f_2+4f_3+2f_4+\ldots+2f_{N-2}+4f_{N-1}+f_N). \tag{12}$$

Для квазиравномерных или произвольных неравномерных сеток формул такого типа не составляют.

Исключение главного члена погрешности формулы трапеций означает, что мы перешли к аппроксимации параболой, и формула Симпсона точна для любого многочлена второй степени. Однако нетрудно проверить, что для $f(x)=x^3$ эта формула также дает точный результат, т. е. она точна для многочлена третьей степени! Это объясняется тем, что на равномерной сетке остаточный член формулы трапеций разлагается только по четным степеням шага и однократное применение метода Рунге увеличивает порядок точности на два.

Как и для формулы трапеций, погрешность формулы Симпсона вычисляется подстановкой разложения (5), в котором теперь надо удержать большее число членов и для каждой пары интервалов (x_{i-1}, x_i) и (x_i, x_{i+1}) за центр разложения взять узел x_i. Из предыдущего рассуждения видно, что главный вклад в погрешность дает только пятый член разложения $(1/24)(x-x_i)^4 f^{\mathrm{IV}}(x_i)$. Подставляя этот член в выражение суммарной погрешности двух соседних

интервалов, легко найдем

$$R_i = \int_{x_{i-1}}^{x_{i+1}} f(x)\,dx - \frac{h}{3}(f_{i-1}+4f_i+f_{i+1}) \approx -\frac{h^5}{90} f^{IV}(x_i).$$

После суммирования по парам соседних интервалов получим

$$R \approx -\frac{h^4}{180}\int_a^b f^{IV}(x)\,dx = O(h^4), \tag{13}$$

т. е. формула Симпсона имеет четвертый порядок точности, а численный коэффициент в остаточном члене очень мал. Благодаря этим обстоятельствам формула Симпсона обычно дает хорошую точность при сравнительно небольшом числе узлов (если четвертая производная функции не слишком велика).

Асимптотическая оценка (13) выведена в предположении существования непрерывной четвертой производной. Если $f^{IV}(x)$ кусочно-непрерывна, то справедлива только мажорантная оценка, аналогичная (9).

Пример. Вычислим интеграл $F = \int_0^1 e^x\,dx \approx 1{,}7183$. В таблице 9 приведены результаты вычислений по формулам трапеций и Симпсона при разных шагах. Вторая формула обеспечивает гораздо более высокую точность при том же шаге.

Таблица 9

h	Трапеций	Симпсона
1	1,8591	—
0,5	1,7539	1,7189
0,25	1,7272	1,7183

Заметим, однако, что формулу Симпсона можно было вообще не вводить. Проведем расчеты по формуле трапеций на последовательности сгущающихся вдвое сеток и применим однократный процесс Рунге не к формуле, а непосредственно к найденному на каждой сетке значению интеграла. Результат будет тот же, что и при расчете по формуле Симпсона; попутно будет оценена фактическая погрешность формулы трапеций.

К самой формуле Симпсона, как следует из вида ее остаточного члена, тоже можно применять метод Рунге. Это эквивалентно применению рекуррентного процесса Рунге к формуле трапеций.

4. Формула средних. Если на отрезке $[a, b]$ взять единственный узел квадратурной формулы x_0, то функция аппроксимируется многочленом нулевой степени — константой $f(x_0)$. Поскольку симметрия формулы численного интегрирования приводит к повышению ее точности, то выберем в качестве единственного узла середину отрезка интегрирования $\bar{x} = {}^1/_2(a+b)$. Приближенно заменяя площадь криволинейной трапеции площадью прямоугольника (рис. 17), получим *формулу средних*

$$F = \int_a^b f(x)\,dx \approx (b-a) f(\bar{x}), \qquad \bar{x} = \frac{1}{2}(a+b). \tag{14}$$

Погрешность этой формулы вычислим стандартным приемом — подстановкой разложения (5); в данном случае за центр разложения надо брать середину отрезка, т. е. узел квадратурной формулы. Несложные выкладки показывают, что

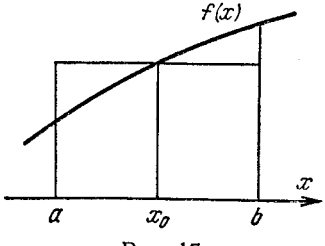


Рис. 17.

$$R = \int_a^b f(x)\,dx - (b-a)\,f(\bar{x}) \approx$$

$$\approx \frac{1}{24}(b-a)^3 f''(\bar{x}). \qquad (15)$$

При вычислении уничтожился не только первый, но и второй член разложения Тейлора. Это связано с симметричным построением формулы средних и означает, что формула точна для любой линейной функции.

Так же как и для формулы трапеций, для повышения точности вводится достаточно подробная сетка x_i и составляется обобщенная формула средних

$$F \approx \sum_{i=1}^{N} (x_i - x_{i-1})\, f\left(\frac{x_i + x_{i-1}}{2}\right), \qquad R \approx \frac{1}{24} \sum_{i=1}^{N} (x_i - x_{i-1})^3 f''(\bar{x}_i). \qquad (16)$$

На равномерной сетке она имеет вид

$$F \approx h \sum_{i=1}^{N} f_{i-1/2}, \qquad R \approx \frac{h^2}{24} \int_a^b f''(x)\,dx. \qquad (17)$$

Замечание 1. Остаточный член формулы средних примерно вдвое меньше, чем у формулы трапеций. Поэтому если значения функции одинаково легко определяются в любых точках, то лучше вести расчет по более точной формуле средних. Формулу трапеций употребляют в тех случаях, когда функция задана только в узлах сетки, а в серединах интервалов неизвестна.

Замечание 2. Знаки главного члена погрешности у формул трапеций и средних разные. Поэтому, если есть расчеты по обеим формулам, то точное значение интеграла лежит, как правило, в вилке между ними. Деление этой вилки в отношении 2:1 дает уточненный результат, соответствующий формуле Симпсона.

Замечание 3. К формуле средних тоже можно применять метод Рунге и либо непосредственно уточнять значение интеграла, либо строить формулы повышенной точности. Те формулы, которые при этом будут получаться, и те, которые были рассмотрены в предыдущих пунктах, — частные случаи так называемых формул Котеса.

5. Формула Эйлера. Аппроксимация подынтегральной функции интерполяционным многочленом Эрмита приводит к квадратурным формулам, содержащим производные в узлах. Мы не будем рассматривать общий случай и выведем только формулу с первой производной. Для этого приближенно выразим остаточный член формулы трапеций (6) через значения производных в узлах

$$R_{\text{трап}} \approx -\frac{1}{12}(b-a)^3 f''(\bar{x}) \approx \frac{1}{12}(b-a)^2 [f'(a) - f'(b)]. \quad (18)$$

Прибавляя эту величину к правой части формулы трапеций, получим *формулу Эйлера* (или Эйлера — Маклорена)

$$F \approx \frac{1}{2}(b-a)[f(a)+f(b)] + \frac{1}{12}(b-a)^2 [f'(a) - f'(b)]. \quad (19)$$

Остаточный член этой формулы вычислим стандартной подстановкой разложения Тейлора для $f(x)$. Предполагая существование непрерывной четвертой производной, выпишем в формуле (5) пять членов разложения и подставим их в выражение погрешности. Выполнив выкладки, получим асимптотическую оценку

$$R = \int_a^b f(x)\,dx - \frac{1}{2}(b-a)[f(a)+f(b)] - \frac{1}{12}(b-a)^2[f'(a) - f'(b)] \approx$$

$$\approx \frac{1}{720}(b-a)^5 f^{\text{IV}}(\bar{x}). \quad (20)$$

Видно, что из-за симметрии формулы уничтожился член, содержащий $f'''(x)$; значит, формула Эйлера точна для многочлена третьей степени. Ее остаточный член имеет тот же вид, что и остаточный член формулы Симпсона; но его численный коэффициент в пересчете на один интервал оказывается вчетверо меньше.

Отметим, что остаточный член (20) можно выразить, аналогично (18), через разности третьих производных в узлах и т. д. Так строят формулы Эйлера — Маклорена высших порядков, но в практических вычислениях они применяются редко, и мы не будем их рассматривать.

Обобщенную формулу Эйлера можно написать на произвольной сетке. Особенно простой вид приобретает формула на равномерной сетке, ибо производные во внутренних узлах сетки при этом взаимно уничтожаются:

$$F \approx h\left(\frac{1}{2}f_0 + f_1 + f_2 + \ldots + f_{N-1} + \frac{1}{2}f_N\right) + \frac{1}{12}h^2(f'_0 - f'_N),$$

$$R \approx \frac{1}{720}h^4 \int_a^b f^{\text{IV}}(x)\,dx. \quad (21)$$

Это показывает, что небольшая добавка к формуле трапеций сильно увеличивает точность, повышая ее с $O(h^2)$ до $O(h^4)$.

Если в таблице заданы значения только самой функции, а не ее производных, то обобщенную формулу Эйлера можно применять, подставляя разностные выражения для f'_0, f'_N. Но эти выражения должны иметь второй порядок точности, чтобы соответствовать общей точности формулы (получающиеся при этом формулы называются *формулами Грегори*). Если заменить производные односторонними разностями $f'_0 \approx (f_1 - f_0)/h$, $f'_N \approx (f_N - f_{N-1})/h$, то общий порядок точности понижается до третьего.

6. Процесс Эйткена. У всех рассмотренных выше обобщенных формул на равномерных и квазиравномерных сетках ошибку можно разложить в ряд по степеням шага типа (3.17). Значит, к ним ко всем применим метод Рунге. Но для его применения надо знать, каков порядок точности исходной формулы.

Предположим, что порядок точности p существует, но неизвестен нам. Оказывается, и в этом случае можно уточнить результат, если расчеты проведены на трех (или более) сетках.

Чтобы упростить алгоритм расчета, выберем три сетки с постоянным отношением шагов, т. е. с шагами $h_1 = h$, $h_2 = qh$, $h_3 = q^2h$. Обозначим приближенное значение интеграла на k-й сетке через F_k и ограничимся главным членом погрешности; тогда можно написать

$$F = F_k + \alpha h_k^p, \qquad k = 1, 2, 3. \tag{22}$$

Это система трех уравнений для определения неизвестных F, α, p. Вводя вспомогательные обозначения $\beta = \alpha h^p$, $\gamma = q^p$, преобразуем эту систему к следующему виду:

$$F - F_1 = \beta, \quad F - F_2 = \beta\gamma, \quad F - F_3 = \beta\gamma^2. \tag{23}$$

Перемножая крайние уравнения (23) и сравнивая с квадратом среднего уравнения, получим $\beta^2\gamma^2 = (F - F_1)(F - F_3) = (F - F_2)^2$; отсюда легко получить уточненное значение интеграла

$$F = F_1 + (F_1 - F_2)^2/(2F_2 - F_1 - F_3). \tag{24}$$

Попарно вычитая уравнения (23) друг из друга, получим

$$F_2 - F_1 = \beta(1 - \gamma), \quad F_3 - F_2 = \beta\gamma(1 - \gamma),$$

или

$$q^p = \gamma = (F_3 - F_2)/(F_2 - F_1).$$

Следовательно, эффективный порядок точности исходной формулы (22) равен

$$p = (\ln q)^{-1} \ln[(F_3 - F_2)/(F_2 - F_1)]. \tag{25}$$

Описанный алгоритм был предложен Эйткеном в 1937 г. для ускорения сходимости итерационных процессов последовательного приближения, в которых ошибка убывает примерно по геометрической прогрессии (см. главу V, § 2).

Погрешность численного интегрирования при изменении шага в q раз меняется приблизительно в q^p раз; поэтому если сетки последовательно сгущаются в одно и то же число раз, то ошибка убывает именно по требуемому закону.

Замечание. Вычислять уточненное значение следует именно по формуле (24), не преобразовывая ее. В данной записи из F_1 вычитается поправка, в которой числитель и знаменатель имеют одинаковый порядок малости, поэтому заметной потери точности не происходит. Если же привести все члены в формуле к общему знаменателю, то в вычислениях придется удерживать много знаков, чтобы избежать потери точности при округлениях.

Пример. Рассмотрим вычисление интеграла $F=\int\limits_0^1 \sqrt{x}\,dx=2/3$. У подынтегральной функции даже первая производная не ограничена, поэтому все приведенные ранее априорные оценки погрешности неприменимы. Мы не знаем, каков здесь эффективный порядок точности каждой из рассмотренных ранее формул численного интегрирования. Составим таблицу 10 значений функции и вычислим интеграл по формулам трапеций и Симпсона при разных шагах (таблица 11).

Таблица 10

x	$f(x)=\sqrt{x}$
0,00	0,0000
0,25	0,5000
0,50	0,7071
0,75	0,8660
1,00	1,0000

Таблица 11

h	Трапеций	Симпсон	Эйткен
1,00	0,5000	—	—
0,50	0,6036	0,6381	—
0,25	0,6433	0,6565	0,6680

Видно, что обе формулы дают результаты невысокой точности. Плохая точность формулы Симпсона означает, что формула трапеций фактически имеет не второй порядок точности и уточнение методом Рунге здесь бессмысленно. А уточнение первого столбца таблицы процессом Эйткена существенно улучшает результат; попутно выясняется, что в данном примере эффективный порядок точности формулы трапеций $p\approx 1{,}38$.

Эффективный порядок точности оказался не целым числом! С этим приходится встречаться, если функция имеет особенность, а формула интегрирования явно этого не учитывает, или если особенность имеет сама формула (это возможно в нелинейных формулах интегрирования, рассмотренных в § 2).

Если никаких особенностей нет, то эффективный порядок точности может только слегка отличаться от теоретического благодаря наличию в погрешности не только главного члена, но и членов более высокого порядка малости. В этом случае при $h\to 0$ эффективный порядок стремится к теоретическому.

На этом основан быстрый метод контроля программ для ЭВМ. Зададим функцию, не имеющую особенностей, проведем расчеты на сгущающихся сетках и проверим, согласуется ли эффективный порядок точности с теоретическим. Сильное расхождение свидетельствует об ошибке в программе.

7. Формулы Гаусса — Кристоффеля. Параметрами формулы интегрирования (3) являются узлы и веса. Однако, строя формулы трапеций, Симпсона, Эйлера, мы заранее задавали узлы и по ним находили веса. Поэтому мы не полностью использовали возможности общей формулы. Только в формуле средних мы подобрали положение узла из соображений симметрии, что привело к существенному улучшению формулы.

Формула (3) с n узлами содержит всего $2n$ параметров; столько же коэффициентов у многочлена степени $2n-1$. Значит, параметры можно подобрать так, чтобы квадратурная формула (3)

$$F=\int_a^b f(x)\,\rho(x)\,dx \approx \sum_{k=1}^{n} c_k f(x_k)$$

была точна для любого многочлена степени не выше $2n-1$ *). Покажем, как находятся узлы и веса этих формул.

Будем считать, что вес положителен $\rho(x)>0$ и непрерывен на (a, b); он может обращаться в нуль или в бесконечность на концах отрезка так, чтобы существовал $\int_a^b \rho(x)\,dx$. Известно **), что при выполнении этих условий существует полная система алгебраических многочленов $P_m(x)$, ортогональных на $[a, b]$ с заданным весом:

$$\int_a^b P_k(x)\,P_m(x)\,\rho(x)\,dx=\delta_{km}\,\|P_k(x)\|^2_{L_2}. \tag{26}$$

Все нули этих многочленов действительны и расположены на интервале (a, b).

Составим по узлам интегрирования многочлен n-й степени $\omega_n(x)=\prod_{k=1}^{n}(x-x_k)$. Функция $f(x)=\omega_n(x)\,P_m(x)$ при $m\leqslant n-1$ есть многочлен степени не выше $2n-1$; значит, для нее формула Гаусса — Кристоффеля точна. Тогда получим

$$\int_a^b \omega_n(x)\,P_m(x)\,\rho(x)\,dx=\sum_{k=1}^{n} c_k\omega_n(x_k)\,P_m(x_k)=0, \tag{27}$$

*) Первую такую формулу для $\rho(x)\equiv 1$ построил Гаусс. Случай произвольного веса рассмотрел Кристоффель.

**) Смотри, например, [24].

так как $\omega_n(x_k)=0$. Значит, многочлен $\omega_n(x)$ ортогонален всем многочленам $P_m(x)$ степени $m\leqslant n-1$.

Если разложить $\omega_n(x)$ в ряд по нашим ортогональным многочленам и этот ряд подставить в условие ортогональности (27), то получим

$$\omega_n(x)=\sum_{k=0}^{n} b_k P_k(x),$$

$$0=\int_a^b \omega_n(x)\,P_m(x)\,\rho(x)\,dx=b_m\|P_m\|^2, \qquad m\leqslant n-1,$$

т. е. все коэффициенты разложения $b_m=0$ при $m\leqslant n-1$. Это значит, что $\omega_n(x)$ с точностью до численного множителя совпадает с $P_n(x)$. Значит, *узлами формулы Гаусса — Кристоффеля являются нули многочленов соответствующей степени* $P_n(x)$, *ортогональных на* $[a, b]$ *с весом* $\rho(x)$.

Веса интегрирования нетрудно определить, если узлы уже найдены. Функция

$$\psi_m(x)=\prod_{k=1,\ k\neq m}^{n}(x-x_k)/(x_m-x_k)$$

есть многочлен степени $n-1$, т. е. для нее формула Гаусса — Кристоффеля точна. Подставляя ее в формулу (3) и учитывая, что эта функция равна нулю во всех узлах, кроме m-го, получим веса формулы Гаусса — Кристоффеля

$$c_m=\int_a^b \rho(x)\left\{\prod_{k=1,\ k\neq m}^{n}(x-x_k)/(x_m-x_k)\right\}dx. \tag{28}$$

Из этого выражения ничего нельзя сказать о знаке веса. Но если подставить в формулу интегрирования многочлен $\psi_m^2(x)$ степени $2n-2$, для которого формула также точна, то получим соотношение

$$c_m=\int_a^b \psi_m^2(x)\,\rho(x)\,dx>0,$$

из которого видно, что *все веса положительны*. Подставляя в формулу Гаусса $f(x)=1$, получим соотношение

$$\sum_{k=1}^{n} c_k=\int_a^b \rho(x)\,dx, \tag{29}$$

из которого следует *равномерная ограниченность весов*.

Для наиболее употребительных весовых функций $\rho(x)$ узлы и веса формул Гаусса — Кристоффеля приведены в Приложении вместе с соответствующими ортогональными многочленами.

Формулы Гаусса — Кристоффеля называют также формулами наивысшей алгебраической точности, поскольку для произвольного многочлена степени выше $2n-1$ формула (3) с n узлами уже не может быть точной.

Заметим, что в принципе можно не обращаться к ортогональным многочленам, а просто подставить в (3) функции вида $f(x)=x^m$ и получить систему уравнений для определения узлов и весов интегрирования

$$\sum_{k=1}^{n} c_k x_k^m = M_m, \quad 0 \leqslant m \leqslant 2n-1, \qquad M_m = \int_a^b x^m \rho(x)\, dx. \tag{30}$$

(Величины M_m называются моментами весовой функции.) Однако это нелинейная система; найти и исследовать ее решение очень трудно даже при небольших n.

Рассмотрим некоторые частные случаи.

а) Собственно формула Гаусса соответствует $\rho(x)=1$. Линейным преобразованием аргумента можно перейти к отрезку $a=-1$, $b=1$. На нем ортогональны с единичным весом многочлены Лежандра. Если обозначить их узлы и соответствующие веса через ξ_k, γ_k, то обратным линейным преобразованием можно получить узлы и веса для произвольного отрезка

$$x_k = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\xi_k, \qquad c_k = \frac{1}{2}(b-a)\gamma_k, \quad 1 \leqslant k \leqslant n. \tag{31}$$

В частности, при $n=1$ получаем формулу средних. Погрешность формулы Гаусса (выражение для которой мы приводим без вывода) пропорциональна той производной, которая соответствует низшей неучтенной степени аргумента; верхняя граница погрешности равна

$$\max|R| = \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} M_{2n} \approx \frac{b-a}{2{,}5\sqrt{n}} \left(\frac{b-a}{3n}\right)^{2n} M_{2n},$$

$$M_{2n} = \max_{[a,\ b]} |f^{(2n)}(x)|.$$

Формула Гаусса рассчитана на функции, имеющие достаточно высокие производные, причем не слишком большие по абсолютной величине. Для таких функций формула обеспечивает очень высокую точность при небольшом числе узлов, ибо численный коэффициент в остаточном члене быстро убывает с ростом n.

б) Формула Эрмита позволяет интегрировать на отрезке $[-1, +1]$ с весом $\rho(x)=1/\sqrt{1-x^2}$. При этих условиях ортогональны многочлены Чебышева первого рода $T_n(x)$. Соответствующие

узлы и веса интегрирования равны

$$\xi_k = \cos[\pi(k - 1/2)n], \quad \gamma_k = \pi/n, \quad 1 \leqslant k \leqslant n. \tag{32}$$

Отметим, что веса во всех узлах одинаковы. На произвольный отрезок эти узлы и веса преобразуются так же, как в формуле Гаусса. Погрешность формулы Эрмита не превышает

$$\max|R| = \pi M_{2n}/[2^{n-1}(2n)!].$$

в) По формулам Гаусса — Кристоффеля возможно вычисление несобственных интегралов на полупрямой $0 \leqslant x < \infty$; если весовая функция равна $\rho(x) = x^\alpha e^{-x}$, то ортогональными будут многочлены Лагерра $L_n^\alpha(x)$. То же относится к интегралам на всей прямой $-\infty < x < +\infty$ при весе $\rho(x) = e^{-x^2}$, только ортогональными будут многочлены Эрмита. Соответствующие примеры имеются в § 2.

8. Формулы Маркова. Потребуем, чтобы квадратурная формула (3) была точна для многочлена как можно более высокой степени при дополнительном условии, что одна или обе границы также являются узлами интегрирования. Очевидно, если число узлов равно n и одна из границ — узел, то формула может быть точна для многочлена степени $2n-2$; если обе границы являются узлами — то для многочлена степени $2n-3$.

Соответствующие формулы называют формулами Маркова. Они рассчитаны на гладкие функции и по точности мало уступают формулам Гаусса — Кристоффеля (для той же точности в формулах Маркова надо брать на один узел больше).

Общие формулы разбирать не будем. Приведем только таблицу 12 узлов и весов для случая, когда оба конца отрезка являются узлами, а весовая функция $\rho(x) = 1$; для простоты положим $a = -1$, $b = +1$, ибо преобразовать узлы и веса для произвольного отрезка можно по формулам (31). Отметим, что в этом случае формула Маркова при $n = 2$ совпадает с формулой трапеций, а при $n = 3$ — с формулой Симпсона.

Таблица 12

n	ξ_k	γ_k
2	-1; 1	1; 1
3	-1; 0; 1	1/3; 4/3; 1/3
4	-1; $-1/5$; 1/5; 1	1/6; 5/6; 5/6; 1/6
5	-1; $-3/7$; 0; 3/7; 1	1/10; 49/90; 64/90; 49/90; 1/10

Замечание 1. Формулы Маркова и формулы Гаусса — Кристоффеля рассчитаны на получение очень высокой точности уже при небольшом числе узлов $n \approx 4 - 10$. Поэтому для них не строят обобщенных формул типа (7); исключениями являются только перечисленные выше формулы средних, трапеций и Симпсона.

Замечание 2. В математической литературе подробно разбираются так называемые квадратурные формулы Чебышева. Это формулы типа (3), в которых все веса одинаковы и равны

$c_k = (b-a)/n$, а положение узлов подбирается так, чтобы формула была точна для многочлена как можно более высокой степени. Однако серьезного практического значения эти формулы не имеют. Для функций высокой гладкости удобнее формулы Гаусса, а для недостаточно гладких функций — обобщенные формулы трапеций и средних.

9. Сходимость квадратурных формул. Стремится ли сумма (3) при $n \to \infty$ к точному значению интеграла, и если стремится, то с какой скоростью?

Обобщенная формула средних (16) является интегральной суммой. Следовательно, для любой непрерывной функции она сходится к точному значению интеграла при стремлении к нулю $\max(x_i - x_{i-1})$. Это справедливо и для обобщенной формулы трапеций (7). Она тоже является интегральной суммой, соответствующей несколько иному выбору интервалов: нулевой интервал — от x_0 до $x_{1/2}$, первый — от $x_{1/2}$ до $x_{3/2}$, второй — от $x_{3/2}$ до $x_{5/2}$ и т. д.

Обобщенная формула Симпсона получается линейной комбинацией двух обобщенных формул трапеций на равномерной сетке. При сгущении сетки каждая из последних формул сходится к общему пределу — точному значению интеграла. Значит, и формула Симпсона сходится для любой непрерывной функции.

Более тонкими рассуждениями можно доказать сходимость формул Гаусса — Кристоффеля при $n \to \infty$ для любой непрерывной функции.

Значительно сложнее вопрос о скорости сходимости; он связан с оценкой остаточного члена формул. Напомним, что если $R = O(h^p)$, то мы называем формулу сходящейся с p-м порядком точности.

В большинстве квадратурных формул мы находили вид главного члена погрешности; он выражался через интеграл от некоторой производной функции. Попутно мы отмечали, что если заменить под интегралом производную на максимум ее модуля, т. е. заменить $\int_a^b f^{(p)}(x)\,dx \sim (b-a)\,M_p$, то мы получим мажорантную оценку погрешности. Такие оценки определяют скорость сходимости. Согласно этим оценкам погрешность формул средних и трапеций есть $O(h^2)$, а формулы Симпсона — $O(h^4)$.

Эти оценки пригодны, если функция имеет ту производную, которая входит в оценку остаточного члена, причем эта производная соответственно непрерывна или кусочно-непрерывна. Наличие у функции более высоких производных не улучшает оценку. Зато если у функции нет требуемой ограниченной производной, то сходимость может быть хуже, как мы видели в примере из п. 6.

Скорость сходимости наиболее распространенных квадратурных формул для недостаточно гладких функций сейчас хорошо изучена. В таблице 13 приведены полученные в [25] мажорантные оценки погрешности некоторых формул на классе функций, имеющих на $[a, b]$ кусочно-непрерывную p-ю производную, ограниченную по модулю константой M_p (примерно такие же оценки получаются, если $f^{(p)}(x)$ не ограничена, но интегрируема с квадратом). Стрелка в таблице 13 означает перенос оценки из предыдущего столбца.

Таблица 13

Формула \ p	1	2	3	4	Много
Трапеций	$\frac{b-a}{4}hM_1$	$\frac{b-a}{12}h^2M_2$	→	→	→
Средних	$\frac{b-a}{4}hM_1$	$\frac{b-a}{24}h^2M_2$	→	→	→
Симпсона	$\frac{5}{18}(b-a)hM_1$	$\frac{4}{81}(b-a)h^2M_2$	$\frac{b-a}{72}h^3M_3$	$\frac{b-a}{180}h^4M_4$	→
Гаусса при $n=4$, $b-a=2$	$0{,}276M_1$	$0{,}022M_2$	$0{,}0024M_3$	$0{,}0003M_4$	$10^{-8}M_8$

Из таблицы 13 видно, что для функций малой гладкости, имеющих лишь первую или вторую производную, лучшие результаты дает обобщенная формула средних. Для функций высокой гладкости выгодны формулы Гаусса (отметим, что для функций малой гладкости формулы Гаусса дают примерно ту же точность, что и простейшие формулы, но формулы Гаусса с большим числом узлов довольно сложны и поэтому невыгодны для таких функций). Простой и рекуррентный метод Рунге для обобщенных формул также целесообразно применять только при достаточно высокой гладкости функций: если существует кусочно-непрерывная ограниченная $f^{(p)}(x)$, то можно рассчитывать лишь на точность $O(h^p)$.

Рассмотрим корректность численного интегрирования. Существование и единственность суммы (3) очевидны, и надо исследовать только устойчивость. Во всех рассмотренных выше формулах веса положительны, поэтому при варьировании подынтегральной функции вариация суммы не превышает

$$|\delta F| = \left|\sum_{k=0}^{n} c_k \delta f_k\right| \leqslant \left(\sum_{k=0}^{n} |c_k|\right) \cdot \max_k |\delta f_k| \leqslant \|\delta f\|_C \int_a^b \rho(x)\,dx,$$

так что устойчивость по входным данным есть.

Строго говоря, квадратурные формулы (3) неустойчивы относительно ошибок округления. Эти ошибки носят случайный

характер, но в среднем растут, как $\sqrt{n}$, при увеличении числа узлов, так что график полной ошибки похож на пунктирную линию на рис. 15. Но эта неустойчивость слабая, и она проявляется только при расчете с небольшим (3 — 5) числом цифр.

§ 2. Нестандартные формулы

1. Разрывные функции. Пусть функция и ее производные кусочно-непрерывны; в точках разрыва подразумевается существование односторонних производных всех требуемых порядков.

Разобъем отрезок $[a, b]$ на отрезки так, чтобы на этих отрезках функция и некоторое число p ее низших производных были непрерывны; на концах этих отрезков в качестве значений функции и производных возьмем соответствующие односторонние пределы.

Представим интеграл в виде суммы интегралов по отрезкам непрерывности. Применим к каждому отрезку квадратурную формулу порядка точности q, $q \leqslant p$. Если одновременно и одинаково сгущать сетки на всех отрезках непрерывности, то порядок точности ответа будет q, как и для непрерывных достаточно гладких функций. В этом случае методом Рунге — Ромберга можно повысить порядок точности до p.

Если же применять квадратурные формулы к разрывным или не гладким функциям, не выделяя особые точки указанным образом, то при сгущении сетки сходимость хотя и будет, но с невысокой скоростью и без четко выраженного порядка точности. Мажорантную оценку ошибки вида $O(h^\nu)$ при этом обычно можно найти, но асимптотической оценки вида $R \approx \alpha h^\nu$, как правило, не существует. При этом применять метод Рунге или процесс Эйткена будет нельзя.

Пример. Рассмотрим

$$F = \int_{-1}^{2} (x|x|)\,dx = (7/3).$$

Здесь подынтегральная функция непрерывная и гладкая, но вторая производная имеет разрыв при $x = 0$. Если для этой функции выделить отрезки непрерывности, то формула Симпсона дает точный ответ. Если же сгущать равномерную сетку делением пополам, то точка $x = 0$ никогда не будет узловой и следует ожидать плохой сходимости. Это подтверждается расчетами, приведенными в таблице 14.

2. Нелинейные формулы. Ранее мы видели, что нелинейная аппроксимация может существенно повысить точность расчетов, особенно для быстропеременных функций. В случае интегрирования подбор подходящего приближения становится очень слож-

ным, ибо интеграл от аппроксимирующей функции должен точно вычисляться, иначе метод будет практически бесполезен.

Таблица 14

Формула \ h	3	3/2	3/4	3/8	0
Трапеций	4,5000	2,6250	2,4375	2,3555	2,3333
Симпсона	—	2,0000	2,3750	2,3282	2,3333

Обычно стараются найти выравнивающие переменные, в которых уже два свободных параметра обеспечивали бы удовлетворительную аппроксимацию. На отрезке $[a, b]$ вводят сетку и на каждом интервале сетки функцию заменяют нелинейной интерполяционной функцией, в которой параметры выражены через табличные значения функции. Например, если функция близка к экспоненте, то согласно (2.19)

$$f(x) \approx f_{i-1} \exp\left[(x - x_{i-1}) \ln (f_i/f_{i-1})/(x_i - x_{i-1})\right], \qquad x_{i-1} \leqslant x \leqslant x_i.$$

Если на каждом интервале проинтегрировать это выражение вместо исходной функции, то получим обобщенную квадратурную формулу

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} (x_i - x_{i-1})(f_i - f_{i-1})/\ln (f_i/f_{i-1}). \tag{33}$$

Разумеется, для неэкспоненциальных функций эта формула не обеспечит хорошей точности.

Такие формулы напоминают обобщенную формулу трапеций, ибо они построены при помощи двухпараметрической интерполяции лагранжева типа, которая для каждого интервала сетки выполняется отдельно. Если воспользоваться интерполяцией эрмитова типа, то получим формулы, сходные с обобщенной формулой средних. Например, если по-прежнему считать $f(x) \approx \approx \alpha e^{\beta x}$ и потребовать правильной передачи функции и производной в точке $x_{i-1/2}$, то получим $\beta_i = f_{i-1/2}/f'_{i-1/2}$, $\alpha_i = f_{i-1/2} \times \times \exp(-\beta_i x_{i-1/2})$. Использование этой аппроксимации на каждом шаге дает такую квадратурную формулу:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f'_{i-1/2} \{\exp[\beta_i (x_i - x_{i-1/2})] - \exp[-\beta_i (x_{i-1/2} - x_{i-1})]\}, \qquad \beta_i = f_{i-1/2}/f'_{i-1/2}. \tag{34}$$

Эта и предыдущая формулы написаны для произвольной сетки.

Можно показать, что если исходная и аппроксимирующая функции имеют непрерывные вторые производные, то формулы

такого типа имеют второй порядок точности. Оценим, например, погрешность формулы (34). Для этого разложим на интервале (x_{i-1}, x_i) функцию $\varphi(x) = \ln f(x)$ в ряд Тейлора

$$\varphi(x) = \varphi_{i-1/2} + (x - x_{i-1/2})\,\varphi'_{i-1/2} + \frac{1}{2}(x - x_{i-1/2})^2\,\varphi''_{i-1/2} + \ldots$$

Используемые здесь производные можно найти, поскольку $\varphi'(x) = d(\ln f)/dx = f'(x)/f(x)$ и т. д. Возвращаясь к функции $f(x)$ и учитывая, что $|\varphi''(x)| \ll 1$ в силу исходного предположения о почти экспоненциальном виде функции, получим

$$f(x) = e^{\varphi(x)} =$$

$$= f_{i-1/2} \exp\left\{(x - x_{i-1/2})\,\varphi'_{i-1/2} + \frac{1}{2}(x - x_{i-1/2})^2\,\varphi''_{i-1/2} + \ldots\right\} =$$

$$= f_{i-1/2} e^{(x - x_{i-1/2})\varphi'_{i-1/2}} \left[1 + \frac{1}{2}(x - x_{i-1/2})^2\,\varphi''_{i-1/2} + \ldots\right].$$

Если проинтегрировать последнее выражение по отрезку $[x_{i-1}, x_i]$ и просуммировать по всем отрезкам, то единица в квадратных скобках приведет к квадратурной формуле (34), а второй член даст главную часть погрешности

$$R \approx \frac{1}{2} \sum_{i=1}^{N} f_{i-1/2}\varphi''_{i-1/2} \int_{x_{i-1}}^{x_i} (x - x_{i-1/2})\, e^{(x - x_{i-1/2})\varphi'_{i-1/2}}\, dx \approx$$

$$\approx \sum_{i=1}^{N} A_i(f, f', f'', x) \cdot (x_i - x_{i-1})^3,$$

где коэффициенты A_i выражаются некоторым образом через значения интегрируемой функции и ее производных. Заменяя эти коэффициенты их максимальными значениями, получим оценку погрешности

$$|R| \leqslant \max_i |A_i| \cdot \max_k (x_k - x_{k-1})^2 \cdot \sum_{i=1}^{N} (x_i - x_{i-1}) = O(\max_i h_i^2),$$

что и требовалось доказать.

Нелинейные формулы повышенного порядка точности, аналогичные формулам Симпсона или Гаусса, не употребляют, ибо их слишком сложно строить. Повышенный порядок точности получают (разумеется, если функции имеют требуемые непрерывные производные) таким приемом: строят подходящую несложную нелинейную формулу невысокого порядка точности и проводят по ней расчеты на последовательности сгущающихся равномерных или квазиравномерных сеток; полученные результаты уточняют методом Рунге — Ромберга или процессом Эйткена. Однако этот прием применим только при не очень крупном шаге (см. п. 3).

Отметим, что линейные однородные квадратурные формулы (3) имеют те же свойства, что и сам интеграл: при умножении функции на число сумма умножается на то же число, а при сложении функций соответствующие квадратурные суммы складываются. Для нелинейных квадратурных формул эти свойства могут не выполняться (то же относится к задачам интерполяции и дифференцирования). Например, формулы (33) и (34) не аддитивны.

3. Метод Филона. В радиотехнических задачах часто встречаются функции $f(x)$, описывающие несущее высокочастотное колебание $e^{i\omega x}$ с модулированной амплитудой. Это быстропеременные функции, и их производные $f^{(p)}(x) \sim \omega^p$ велики. Поэтому при интегрировании их по формулам § 1 приходится брать настолько мелкий шаг, чтобы выполнялось условие $\omega h \ll 1$, т. е. чтобы одна осцилляция содержала бы много узлов интегрирования. Это приводит к большому объему вычислений.

Для уменьшения объема вычислений надо использовать априорные сведения о подынтегральной функции. Такие функции можно представить в виде $f(x) = y(x)\exp(i\omega x)$, где частота ω известна, а амплитуда $y(x)$ мало меняется за период основного колебания. Выбирая для $y(x)$ несложные полиномиальные аппроксимации, можно получить квадратурные формулы, называемые формулами Филона [45].

Построим, например, аналог формулы средних. Для этого при вычислении интеграла по отдельному интервалу сетки заменим амплитуду ее значением в середине интервала. Погрешность этой замены определим, разлагая амплитуду по формуле Тейлора

$$y(x) \approx y_{k-1/2}, \quad x_{k-1} \leqslant x \leqslant x_k,$$

$$r(x) = y(x) - y_{k-1/2} \approx (x - x_{k-1/2}) \left(\frac{dy}{dx}\right)_{k-1/2}.$$

Умножим амплитуду на несущую частоту, проинтегрируем по интервалу (x_{k-1}, x_k) и сложим интегралы по всем интервалам. После несложных выкладок получим квадратурную формулу

$$F = \int_{x_0}^{x_N} y(x)\, e^{i\omega x}\, dx \approx \frac{2}{\omega} \sum_{k=1}^{N} f_{k-1/2} \sin\left(\frac{\omega}{2} h_k\right), \qquad h_k = x_k - x_{k-1}, \tag{35}$$

и ее погрешность

$$R = \int_{x_0}^{x_N} r(x)\, e^{i\omega x}\, dx \approx$$

$$\approx \frac{2i}{\omega^2} \sum_{k=1}^{N} y'_{k-1/2} \left(\sin\frac{\omega h_k}{2} - \frac{\omega h_k}{2} \cos\frac{\omega h_k}{2}\right) \exp(i\omega x_{k-1/2}). \tag{36}$$

Если шаг интегрирования настолько мал, что $\omega h \ll 1$, то тригонометрические функции в этих формулах можно разложить в быстро сходящиеся ряды. При этом нетрудно видеть, что формула (35) действительно переходит в обобщенную формулу средних (16), и то же имеет место для погрешности.

Для формулы средних погрешность есть малая величина $O(h^2)$. Однако для квадратурной формулы (35) малость шага гарантирует малость погрешности, только если $\omega h \lesssim 1$, что при высокой несущей частоте требует очень малого шага $h < \omega^{-1}$. Если же шаг не настолько мал и удовлетворяет условию $\omega^{-1} < h \ll 1$, то погрешность по порядку величины есть $R = O(N\omega^{-2}y_x')$. Следовательно, для малости погрешности (36) необходимо, чтобы амплитуда была почти постоянна, т. е. ее производная должна быть малой. Кроме того, можно сделать важный вывод: *если* $\omega h > 1$, *то метод Рунге — Ромберга для уточнения результата применять нельзя*, ибо при этом зависимость погрешности от шага носит сложный (не степенной) характер.

Поэтому для построения формул Филона высокой точности приходится использовать более сложные аппроксимации амплитуды. Например, воспользуемся линейным приближением

$$y(x) \approx y_{k-1} + \frac{y_k - y_{k-1}}{x_k - x_{k-1}}(x - x_{k-1}), \qquad x_{k-1} \leqslant x \leqslant x_k.$$

Умножая на несущую частоту и интегрируя по одному интервалу, легко получим

$$\int_{x_{k-1}}^{x_k} y(x)\, e^{i\omega x}\, dx \approx \frac{f_k - f_{k-1}}{i\omega} + \frac{2i}{\omega^2 h_k}(y_k - y_{k-1})\, e^{i\omega x_k - 1/2} \sin\frac{\omega h_k}{2};$$

суммирование по всем интервалам сетки дает

$$F \approx \frac{f_N - f_0}{i\omega} + \frac{2i}{\omega^2} \sum_{k=1}^{N} \frac{\sin\frac{\omega h_k}{2}}{h_k}(y_k - y_{k-1})\, e^{i\omega x_k - 1/2}. \tag{37}$$

Для равномерной сетки эту формулу удобнее представить в виде

$$F \approx \frac{4}{\omega^2 h} \sin^2\frac{\omega h}{2} \sum_{k=1}^{N-1} f_k + \left(\frac{1}{i\omega} + \frac{1 - e^{-i\omega h}}{\omega^2 h}\right) f_N - \left(\frac{1}{i\omega} + \frac{e^{i\omega h} - 1}{\omega^2 h}\right) f_0. \tag{38}$$

Легко проверить, что при $\omega h \lesssim 1$ полученные квадратурные формулы переходят в обобщенную формулу трапеций. Если же $\omega h > 1$, то погрешность этих формул по порядку величины равна $R = O(N\omega^{-3}y_{xx}'')$; она мала, если закон изменения амплитуды близок к линейному.

Аналогично строятся формулы Филона для квадратичного или более сложных законов изменения амплитуды.

Если амплитуда почти постоянна или почти линейна и т. д., то применение соответствующей формулы Филона нередко позволяет интегрировать довольно крупным шагом $h > \omega^{-1}$ (например, шагом, равным длине несущей волны или еще более крупным). Описанные же в § 1 полиномиальные формулы требовали бы гораздо более мелкого шага $h \ll \omega^{-1}$, т. е. большего объема вычислений.

Однако формулы (35) — (38) годятся только в том случае, если несущая частота постоянна. Если частота «плывет» (например, при фазовой модуляции колебаний), то надо составлять другие формулы.

4. Переменный предел интегрирования. Пусть надо вычислить

$$F(x) = \int_a^x f(\xi)\,\rho(\xi)\,d\xi.$$

В принципе при каждом значении x его можно рассматривать как интеграл с постоянными пределами и вычислять одним из приведенных выше способов. Однако если надо определять интеграл для очень многих значений x, то это невыгодно. Целесообразнее выбрать сетку и численным интегрированием высокой точности составить таблицу значений интеграла на этой сетке $F_n = F(x_n)$. Тогда

$$F(x) = F_n + \int_{x_n}^{x} f(\xi)\,\rho(\xi)\,d\xi, \qquad x_n \leqslant x < x_{n+1},$$

и последний интеграл можно вычислять по простым формулам, ибо промежуток интегрирования мал.

Возможен другой способ. Имея таблицу $F(x_n)$, можно находить $F(x)$ интерполяцией по этой таблице. Так как одновременно всюду известна производная интеграла $F'(x) = f(x)\,\rho(x)$, то можно воспользоваться эрмитовой интерполяцией, обеспечивающей высокую точность.

5. Несобственные интегралы. Для интегралов с бесконечными пределами есть несколько приемов вычисления.

Прием 1 — введение такой замены переменных, чтобы превратить пределы интегрирования в конечные. Например, для интеграла

$$\int_a^\infty f(x)\,dx, \quad a > 0,$$

замена $x = a/(1-t)$ превращает полупрямую $[a, \infty)$ в отрезок $[0, 1]$. Если после преобразования подынтегральная функция

вместе с некоторым числом производных остается ограниченной, то можно находить интеграл стандартными численными методами.

Прием 2 — обрезание верхнего предела. Выберем настолько большое b, чтобы

$$\int_b^\infty f(x)\,dx$$

был меньше допустимой ошибки вычислений. Тогда его можно отбросить, а

$$\int_a^b f(x)\,dx$$

вычислить по квадратурной формуле. Вблизи верхнего предела подынтегральная функция мала, поэтому вычисление выгодно вести на квазиравномерных сетках, шаг которых велик при $x \approx b$. Для уменьшения объема вычислений целесообразно приближенно вычислить отброшенную часть интеграла и учесть как поправку; это позволяет выбирать меньшее значение b.

Прием 3 — использование формул Гаусса — Кристоффеля. Из подынтегральной функции надо выделить положительный множитель, который можно рассматривать как вес для данных пределов интегрирования. Например, дадим способ вычисления интегральной экспоненты (2.50). Сдвигая нижний предел, приведем интеграл к форме

$$\mathrm{Ei}(x) = \int_0^\infty \frac{e^{-(x+t)}}{x+t}\,dt = e^{-x}\int_0^\infty \frac{e^{-t}}{x+t}\,dt.$$

Рассматривая e^{-t} как весовую функцию и обозначая через ξ_i, γ_i нули многочленов Лагерра $L_n^{(0)}(t)$ и соответствующие веса квадратурной формулы, получим

$$\mathrm{Ei}(x) \approx e^{-x}\sum_{i=1}^{n} \frac{\gamma_i}{x+\xi_i}. \tag{39}$$

Это выражение можно использовать как аппроксимирующую формулу. Например, одному и двум узлам интегрирования соответствуют

$$\mathrm{Ei}(x) \approx \frac{e^{-x}}{x+1}, \quad \mathrm{Ei}(x) \approx \frac{(x+3)\,e^{-x}}{x^2+4x+2}.$$

Если первая из этих формул пригодна лишь при больших аргументах, то вторая дает удовлетворительную точность $\sim 5\%$ уже при $x=1$, а при больших аргументах точность еще лучше.

Прием 4 — построение нелинейных квадратурных формул, применимых на бесконечном интервале. Например, формула (34) при $\beta_i < 0$ допускает стремление x_i к бесконечности, если $x_{i-1/2}$

остается конечным. Для практического применения таких формул удобно ввести квазиравномерную сетку на $[a, \infty)$, ибо ее последний интервал обладает требуемым свойством: его правая граница удалена в бесконечность, а середина остается конечной. Кроме того, на квазиравномерных сетках можно уточнять результат методом Рунге — Ромберга.

Если пределы интегрирования конечны, значит, $f(x)$ обращается в бесконечность в каких-то точках отрезка $[a, b]$. Будем считать, что вблизи особой точки $|f(x)| \leqslant M \cdot |x - \bar{x}|^{\alpha}$, где $-1 < \alpha < 0$; случай $\alpha \leqslant -1$, когда интеграл существует в смысле главного значения, надо разбирать отдельно. Особые точки разбивают отрезок на части. Рассмотрим приемы вычисления интеграла по отдельному отрезку, у которого особыми точками являются только одна или обе границы.

Прием 1 — аддитивное выделение особенности. Постараемся разбить подынтегральную функцию на сумму $f(x) = \varphi(x) + \psi(x)$, где $\varphi(x)$ — ограниченная функция, а $\psi(x)$ интегрируется аналитическими методами. Тогда $\int_a^b \psi(x)\,dx$ вычисляем точно, а $\int_a^b \varphi(x)\,dx$ находим обычными численными методами. Заметим, что обычно разбиение на сумму делается выделением особенности в наиболее простом виде. Например, если $f(x) = 1/\sqrt{x(1+x^2)}$, а интеграл вычисляется от точки $x = 0$, то основная особенность имеет вид $\psi(x) = 1/\sqrt{x}$; если положить $\varphi(x) = f(x) - \psi(x) = 1/\sqrt{x(1+x^2)} - 1/\sqrt{x}$, то полученная функция будет ограничена, что и требуется.

Прием 2 — мультипликативное выделение особенности. Представим подынтегральную функцию в виде $f(x) = \varphi(x)\rho(x)$, где $\varphi(x)$ ограничена, а $\rho(x)$ положительна и интегрируема на отрезке. Тогда можно рассматривать $\rho(x)$ как весовую функцию и применять квадратурные формулы Гаусса — Кристоффеля. Если на обоих концах отрезка функция имеет особенности степенного вида, то узлами интегрирования будут нули многочленов Якоби. Например,

$$\int_{-1}^{1} \frac{e^x}{\sqrt{1-x^2}}\,dx \approx \frac{\pi}{n} \sum_{i=1}^{n} e^{x_i}, \qquad x_i = \cos\frac{\pi}{n}\left(i - \frac{1}{2}\right). \tag{40}$$

Здесь использовались многочлены Чебышева первого рода (см. Приложение).

Прием 3 — построение нестандартных квадратурных формул, явно учитывающих характер особенности. Так, для приведенного выше интеграла (40) на отдельном интервале сетки (x_{i-1}, x_i) можно аппроксимировать подынтегральную функцию выражением $\exp(x_{i-1/2})/\sqrt{1-x^2}$, поскольку числитель — медленно меняющаяся

гладкая функция и основная особенность связана со знаменателем. Эта аппроксимация легко интегрируется и приводит к квадратурной формуле

$$\int_{-1}^{1} \frac{e^x}{\sqrt{1-x^2}}\,dx \approx \sum_{i=1}^{N} (\arcsin x_i - \arcsin x_{i-1})\, e^{x_i - 1/2},$$

$$x_0 = -1, \quad x_N = 1. \tag{41}$$

По погрешности аппроксимации подынтегральной функции можно заключить, что остаточный член этой формулы на произвольной сетке не превышает $O(h_{\max}^2)$. На специальной сетке $x_i = \cos\frac{\pi i}{n}$ эта формула еще более точна, ибо при этом она переходит в квадратурную формулу Гаусса — Кристоффеля (40), но это уже случайное обстоятельство. Обычно хорошо составленная нестандартная формула имеет один и тот же порядок точности на равномерных и неравномерных сетках.

§ 3. Кратные интегралы

1. Метод ячеек. Рассмотрим двукратный интеграл по прямоугольнику G $(a \leqslant x \leqslant b,\ \alpha \leqslant y \leqslant \beta)$. По аналогии с формулой средних можно приближенно заменить функцию на ее значение в центральной точке прямоугольника. Тогда интеграл легко вычисляется:

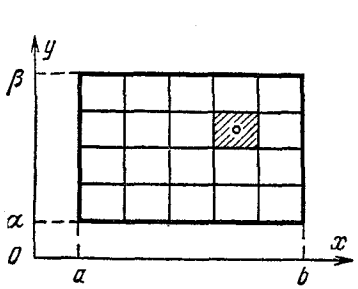


Рис. 18.

$$\int_{\alpha}^{\beta}\int_{a}^{b} f(x,\ y)\,dx\,dy \approx Sf(\bar{x},\ \bar{y}),$$

$$S = (b-a)(\beta-\alpha), \tag{42}$$

$$\bar{x} = \frac{1}{2}(a+b), \qquad \bar{y} = \frac{1}{2}(\alpha+\beta).$$

Для повышения точности можно разбить область на прямоугольные ячейки (рис. 18). Приближенно вычисляя интеграл в каждой ячейке по формуле средних и обозначая через S_i, $\bar{x}_i$, $\bar{y}_i$ соответственно площадь ячейки и координаты ее центра, получим

$$I = \iint_G f(x,\ y)\,dx\,dy \approx \sum_i S_i f(\bar{x}_i,\ \bar{y}_i). \tag{43}$$

Справа стоит интегральная сумма; следовательно, для любой непрерывной $f(x,\ y)$ она сходится к значению интеграла, когда периметры всех ячеек стремятся к нулю.

Оценим погрешность интегрирования. Формула (42) по самому ее выводу точна для $f(x, y) = \text{const}$. Но непосредственной подстановкой легко убедиться, что формула точна и для любой линейной функции, т. е. она соответствует аппроксимации поверхности $z = f(x, y)$ плоскостью. В самом деле, разложим функцию по формуле Тейлора

$$f(x, y) = f(\bar{x}, \bar{y}) + \xi f'_x + \eta f'_y + \frac{1}{2}\xi^2 f''_{xx} + \xi\eta f''_{xy} + \frac{1}{2}\eta^2 f''_{yy} + \ldots, \quad (44)$$

где $\xi = x - \bar{x}$, $\eta = y - \bar{y}$, а все производные берутся в центре ячейки. Подставляя это разложение в правую и левую части квадратурной формулы (42) и сравнивая их, аналогично одномерному случаю легко получим выражение погрешности этой формулы

$$R \equiv \int_{\alpha}^{\beta}\int_{a}^{b} f(x, y)\, dx\, dy - Sf(\bar{x}, \bar{y}) \approx$$

$$\approx \frac{1}{24} S\left[(b-a)^2 f''_{xx} + (\beta-\alpha)^2 f''_{yy}\right], \quad (45)$$

ибо все члены разложения, нечетные относительно центра симметрии ячейки, взаимно уничтожаются.

Пусть в обобщенной квадратурной формуле (43) стороны прямоугольника разбиты соответственно на N и M равных частей. Тогда погрешность интегрирования (45) для единичной ячейки равна

$$R_i \approx \frac{1}{24} S_i\left[\left(\frac{b-a}{N}\right)^2 f''_{xx} + \left(\frac{\beta-\alpha}{M}\right)^2 f''_{yy}\right]_i.$$

Суммируя это выражение по всем ячейкам, получим погрешность обобщенной формулы

$$R \approx \frac{1}{24}\left[\left(\frac{b-a}{N}\right)^2 \iint_G f''_{xx}\, dx\, dy + \left(\frac{\beta-\alpha}{M}\right)^2 \iint_G f''_{yy}\, dx\, dy\right] =$$

$$= O(N^{-2} + M^{-2}), \quad (46)$$

т. е. формула имеет второй порядок точности. При этом, как и для одного измерения, *можно применять метод Рунге — Ромберга, но при одном дополнительном ограничении: сетки по каждой переменной сгущаются в одинаковое число раз, т. е. отношение N/M остается постоянным.*

Обобщим формулу ячеек на более сложные области. Легко сообразить, что для линейной функции $f(x, y)$ формула типа (42) будет точна в области произвольной формы, если под S подразумевать площадь области, а под $\bar{x}$, $\bar{y}$ — координаты центра

тяжести, вычисляемые по обычным формулам

$$S=\iint_G dx\,dy,\quad \bar{x}=\frac{1}{S}\iint_G x\,dx\,dy,\quad \bar{y}=\frac{1}{S}\iint_G y\,dx\,dy. \qquad (47)$$

Разумеется, практическую ценность это имеет только для областей простой формы, где площадь и центр тяжести легко определяются; например, для треугольника, правильного многоугольника, трапеции. Но это значит, что обобщенную формулу (43) можно применять к областям, ограниченным ломаной линией, ибо такую область всегда можно разбить на прямоугольники и треугольники.

Для области с криволинейной границей формулу (43) применяют иным способом. Наложим на область G прямоугольную сетку (рис. 19). Те ячейки сетки, все точки которых принадлежат области, назовем *внутренними*; если часть точек ячейки принадлежит области, а часть — нет, то назовем ячейку *граничной*. Площадь внутренней ячейки равна произведению ее сторон. Площадью граничной ячейки будем считать площадь той ее части, которая попадает внутрь G; эту площадь вычислим приближенно, заменяя в пределах данной ячейки истинную границу области на хорду. Эти площади подставим в (43) и вычислим интеграл.

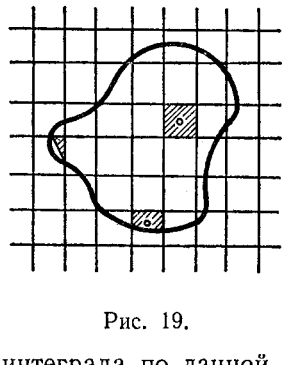


Рис. 19.

Оценим погрешность формулы (43). В каждой внутренней ячейке ошибка составляет $O(N^{-2})$ по отношению к значению интеграла по данной ячейке. В каждой граничной ячейке относительная ошибка есть $O(N^{-1})$, ибо центр прямоугольной ячейки не совпадает с центром тяжести входящей в интеграл части. Но самих граничных ячеек примерно в N раз меньше, чем внутренних. Поэтому при суммировании по ячейкам общая погрешность будет $O(N^{-2})$, если функция дважды непрерывно дифференцируема, а граница области есть кусочно-гладкая кривая; это означает второй порядок точности.

Вычисление площади граничной ячейки довольно трудоемко, ибо требует определения положения границы внутри ячейки. Можно вычислять интегралы по граничным ячейкам более грубо или вообще не включать их в сумму (43). Погрешность при этом будет $O(N^{-1})$, и для хорошей точности потребуется более подробная сетка.

Метод ячеек переносится на большее число измерений. Мы видели, что к области произвольной формы его трудно применять; поэтому всегда желательно заменой переменных преобразо-

вать область интегрирования в прямоугольный параллелепипед (это относится практически ко всем методам вычисления кратных интегралов).

2. Последовательное интегрирование. Снова рассмотрим интеграл по прямоугольнику, разбитому сеткой на ячейки (рис. 18). Его можно вычислить последовательным интегрированием

$$I=\int_{\alpha}^{\beta}\int_{a}^{b} f(x,\ y)\,dx\,dy=\int_{\alpha}^{\beta}F(y)\,dy,$$

$$F(y)=\int_{a}^{b} f(x,\ y)\,dx.$$

Каждый однократный интеграл легко вычисляется на данной сетке по квадратурным формулам типа (3). Последовательное интегрирование по обоим направлениям приводит к кубатурным формулам, которые являются *прямым произведением* одномерных квадратурных формул

$$F(y_j)\approx\sum_i c_i f(x_i,\ y_j),\quad I=\sum_j \bar{c}_j F(y_j),$$

или

$$I\approx\sum_{i,\ j} c_{ij} f(x_i,\ y_j),\qquad c_{ij}=c_i\bar{c}_j. \tag{48}$$

Например, если по каждому направлению выбрана обобщенная формула трапеций, а сетка равномерная, то веса кубатурной формулы равны $c_{ij}/(h_x h_y)=1$, $^1/_2$ и $^1/_4$ соответственно для внутренних, граничных и угловых узлов сетки. Легко показать, что для дважды непрерывно дифференцируемых функций эта формула имеет второй порядок точности и к ней применим метод Рунге — Ромберга.

Вообще говоря, для разных направлений можно использовать квадратурные формулы разных порядков точности p и q. Тогда главный член погрешности имеет вид $R=O\left(h_x^p+h_y^q\right)$. Это надо учитывать в методе Рунге: при сгущении сеток надо сохранять отношение h_x^p/h_y^q постоянным, чтобы закон убывания погрешности был известным. Многократно сгущать сетку при этом условии нелегко, если $p\neq q$; поэтому желательно для всех направлений использовать квадратурные формулы одинакового порядка точности.

Можно подобрать веса и положение линий сетки так, чтобы каждая одномерная квадратурная формула была точна для многочлена максимальной степени, т. е. была бы формулой Гаусса;

тогда

$$c_{ij}=\frac{1}{4}(b-a)(\beta-\alpha)\gamma_i\gamma_j, \quad x_i=\frac{1}{2}(a+b)+\frac{1}{2}(b-a)\xi_i,$$
$$y_j=\frac{1}{2}(\alpha+\beta)+\frac{1}{2}(\beta-\alpha)\xi_j, \quad 1\leqslant i,\ j\leqslant n, \tag{49}$$

где ξ, γ — нули многочленов Лежандра и соответствующие веса. Эти формулы рассчитаны на функции высокой гладкости и дают для них большую экономию в числе узлов по сравнению с более простыми формулами. Например, для m измерений кубатурная формула Симпсона с 3^m узлами и формула (48)—(49) с 2^m узлами дают примерно одинаковую точность, хотя формула Гаусса при $m=2$ имеет вдвое меньше узлов, а при $m=3$ — втрое меньше, чем кубатурная формула Симпсона.

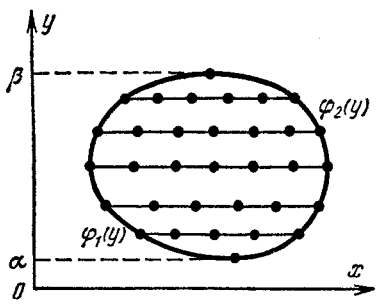


Рис. 20.

Произвольная область. Метод последовательного интегрирования можно применять к области произвольной формы, например, с криволинейной границей. Для этого проведем через область хорды, параллельные оси x, и на них введем узлы, расположенные на каждой хорде так, как нам требуется (рис. 20). Представим интеграл в виде

$$I=\iint\limits_G f(x,\ y)\,dx\,dy=\int\limits_\alpha^\beta F(y)\,dy,$$
$$F(y)=\int\limits_{\varphi_1(y)}^{\varphi_2(y)} f(x,\ y)\,dx.$$

Сначала вычислим интеграл по x вдоль каждой хорды по какой-нибудь одномерной квадратурной формуле, используя введенные узлы. Затем вычислим интеграл по y; здесь узлами будут служить проекции хорд на ось ординат.

При вычислении интеграла по y имеется одна тонкость. Если область ограничена гладкой кривой, то при $y\to\alpha$ длина хорды стремится к нулю не линейно, а как $\sqrt{y-\alpha}$; значит, вблизи этой точки $F(y)\sim\sqrt{y-\alpha}$. То же будет при $y\to\beta$. Поэтому интегрировать непосредственно $F(y)$ по формулам высокого порядка точности бессмысленно. Целесообразно выделить из $F(y)$ основную особенность в виде веса $\rho(y)=\sqrt{(\beta-y)(y-\alpha)}$, которому соответствуют ортогональные многочлены Чебышева второго рода

(см. Приложение). Тогда второе интегрирование выполняется по формулам Гаусса — Кристоффеля

$$I=\int_{\alpha}^{\beta} F(y)\,dy \approx \sum_{j=1}^{n} \frac{\beta-\alpha}{2}\gamma_j \psi\left(\frac{\alpha+\beta}{2}+\frac{\beta-\alpha}{2}\xi_j\right), \qquad (50)$$

где $\psi(y)=F(y)/\rho(y)$, а $\xi_j=\cos[\pi j/(n+1)]$ и γ_j — нули и веса многочленов Чебышева второго рода.

Чтобы можно было применять эту формулу, надо ординаты хорд на рис. 20 заранее выбрать в соответствии с узлами (50). Если это не было сделано, то придется ограничиться интегрированием $F(y)$ по обобщенной формуле трапеций, причем ее эффективный порядок точности в этом случае будет ниже второго (см. пример в § 1, п. 6).

Кроме методов ячеек и последовательного интегрирования есть другие методы, в которых используется кубатурная формула вида $I \approx \sum_i c_i f(r_i)$.

Можно поставить задачу — найти оптимальные узлы и веса, т. е. дающие минимальную погрешность на заданном классе функций. Частный случай этой задачи — нахождение весов и узлов, при которых формула точна для многомерного многочлена максимальной степени.

Оптимальные узлы и веса удается найти только для областей наиболее простой формы, таких как квадрат, круг, сфера. Зато их использование заметно уменьшает объем расчетов. Это хорошо видно из таблицы 15, в клетках которой приведены минимальные числа узлов, при которых m-мерная кубатурная формула может быть точна для многочлена степени n при последовательном интегрировании по формулам Гаусса и при использовании оптимальных m-мерных коэффициентов.

Таблица 15

m \ n		0	1	2	3	4	5	6	7
1	Последовательные Гаусса и оптимальные	1	1	2	2	3	3	4	4
2	Последовательные Гаусса	1	1	4	4	9	9	16	16
	Оптимальные	1	1	2	4	5	7	10	12
3	Последовательные Гаусса	1	1	8	8	27	27	64	64
	Оптимальные	1	1	3	5	9	14	21	30

§ 4. Метод статистических испытаний

1. Случайные величины. Пусть мы измеряем значение некоторой величины ξ (например, отклонение при стрельбе), на которую влияет большое число различных факторов. Мы не можем

учесть их действие, поэтому заранее не известно, какое значение примет эта величина.

Величину ξ называют *случайной с плотностью распределения* $\rho(x)$, если вероятность того, что величина примет значения между x_1 и x_2, равна $\int_{x_1}^{x_2} \rho(x)\,dx$. По смыслу вероятности, $\rho(x)$ неотрицательна и нормирована

$$\rho(x) \geqslant 0, \qquad \int_{-\infty}^{+\infty} \rho(x)\,dx = 1. \tag{51}$$

Очевидно, если значения ξ всегда заключены между a, b, то $\rho(x) = 0$ вне указанных пределов и интеграл (51) надо брать только по отрезку $[a, b]$. Величина ξ может быть дискретной, т. е. принимать только определенные значения x_i с вероятностями ρ_i (например, уровни энергии квантовой системы). Дискретную величину можно формально объединить с непрерывной, если положить

$$\rho(x) = \sum_i \rho_i \delta(x - x_i), \qquad \rho_i > 0, \qquad \sum_i \rho_i = 1,$$

где $\delta(x - x_i)$ есть δ-функция.

Если по значениям случайной величины вычисляется какая-либо функция $f(\xi)$, то значения этой функции также являются случайными величинами. Такую функцию иногда называют случайной.

Равномерно распределенная величина. Рассмотрим следующую случайную функцию:

$$\gamma(\xi) = \int_{-\infty}^{\xi} \rho(x)\,dx. \tag{52}$$

Она принимает значения $0 \leqslant \gamma \leqslant 1$ и монотонно зависит от ξ. Вероятность того, что γ лежит между $\gamma_1 = \gamma(\xi_1)$ и $\gamma_2 = \gamma(\xi_2)$, равна вероятности того, что ξ лежит между ξ_1 и ξ_2. А последняя вероятность есть $\int_{\xi_1}^{\xi_2} \rho(x)\,dx = \gamma_2 - \gamma_1$, т. е. она равна длине интервала по γ и не зависит от положения этого интервала. Это значит, что $\gamma(\xi)$ с равной вероятностью принимает любое значение на отрезке $[0, 1]$. Поэтому ее называют *случайной величиной, равномерно распределенной на отрезке* $[0, 1]$. Плотность распределения γ равна $\bar{\rho}(\gamma) = 1$ при $0 \leqslant \gamma \leqslant 1$ и $\bar{\rho}(\gamma) = 0$ вне этого отрезка.

2. Разыгрывание случайной величины. Из всех случайных величин проще всего разыгрывать (моделировать) равномерно распределенную величину γ. Рассмотрим, как это делается.

Возьмем какое-то устройство, на выходе которого с вероятностью $^1/_2$ могут появляться цифры 0 или 1; появление той или другой цифры должно быть случайным. Таким устройством может быть бросаемая монета, игральная кость (четно — 0, нечетно — 1) или специальный генератор, основанный на подсчете числа радиоактивных распадов или всплесков радиошума за определенное время (четно или нечетно).

Запишем γ как двоичную дробь $\gamma_{II} = 0$, $\alpha_1\alpha_2\alpha_3 \ldots$ и на место последовательных разрядов будем ставить цифры, выдаваемые генератором: например, $\gamma_{II} = 0{,}010110\ldots$ Поскольку в первом разряде с равной вероятностью могут стоять 0 или 1, это число с равной вероятностью лежит в левой или правой половине отрезка $0 \leqslant \gamma \leqslant 1$. Поскольку во втором разряде тоже 0 и 1 равновероятны, число с равной вероятностью лежит в каждой половине этих половин и т. д. Значит, двоичная дробь со случайными цифрами действительно с равной вероятностью принимает любое значение на отрезке $0 \leqslant \gamma < 1$.

Строго говоря, разыграть можно только конечное количество разрядов k. Поэтому распределение будет не вполне требуемым; математическое ожидание $M\gamma$ окажется меньше $^1/_2$ на величину $\sim 2^{-k-1}$ (ибо значение $\gamma = 0$ возможно, а значение $\gamma = 1$ невозможно). Чтобы этот фактор не сказывался, следует брать многоразрядные числа; правда, в методе статистических испытаний точность ответа обычно не бывает выше $0{,}1\% = 10^{-3}$, а условие $\varepsilon < 2^{-k}$ дает $k > 10$, что на современных ЭВМ перевыполнено с большим запасом.

Псевдослучайные числа. Реальные генераторы случайных чисел не свободны от систематических ошибок: несимметричность монеты, дрейф нуля и т. д. Поэтому качество выдаваемых ими чисел проверяют специальными тестами. Простейший тест — вычисление для каждого разряда частоты появления нуля; если частота заметно отлична от $^1/_2$, то имеется систематическая ошибка, а если она слишком близка к $^1/_2$, то числа не случайные — есть какая-то закономерность. Более сложные тесты — это вычисление коэффициентов корреляции последовательных чисел

$$\varkappa = \sum_i (\gamma_i - {}^1/_2)(\gamma_{i+1} - {}^1/_2)$$

или групп разрядов внутри числа; эти коэффициенты должны быть близкими к нулю.

Если какая-то последовательность чисел удовлетворяет этим тестам, то ее можно использовать в расчетах по методу статистических испытаний, не интересуясь ее происхождением. Разработаны алгоритмы построения таких последовательностей; символи-

чески их записывают рекуррентными формулами

$$\gamma_i = f(\gamma_{i-1}) \text{ или}$$
$$\gamma_i = f(\gamma_{i-1}, \gamma_{i-2}, \ldots, \gamma_{i-k}). \tag{53}$$

Такие числа называют *псевдослучайными* и вычисляют на ЭВМ. Это обычно удобнее, чем использование специальных генераторов. Но для каждого алгоритма есть свое предельное число членов последовательности, которое можно использовать в расчетах; при большем числе членов теряется случайный характер чисел, например — обнаруживается периодичность.

Первый алгоритм получения псевдослучайных чисел был предложен Нейманом. Возьмем число из $2r$ цифр (для определенности десятичных) и возведем его в квадрат. У квадрата оставим $2r$ средних цифр, откинув r последних и r (или $r-1$) первых. Полученное число снова возведем в квадрат и т. д. Значения γ_i получаются умножением этих чисел на 10^{-2r}. Например, положим $r=1$ и выберем начальное число 46; тогда получим

$$\begin{cases} 46 \to 2116 \to 121 \to 144 \to 196 \to 361 \to 1296 \ldots \\ \gamma = 0{,}46 \quad 0{,}11 \quad 0{,}12 \quad 0{,}14 \quad 0{,}19 \quad 0{,}36 \quad 0{,}29 \ldots \end{cases}$$

Но распределение чисел Неймана недостаточно равномерно (преобладают значения $\gamma < 1/2$, что хорошо видно на приведенном примере), и сейчас их редко употребляют.

Наиболее употребителен сейчас несложный и неплохой алгоритм, связанный с выделением дробной части произведения

$$\gamma_i = \{A\gamma_{i-1}\}, \tag{54}$$

где A — очень большая константа (фигурная скобка обозначает дробную часть числа). Качество псевдослучайных чисел сильно зависит от выбора величины A: это число в двоичной записи должно иметь достаточно «случайный» вид, хотя его последний разряд следует брать единицей. Величина γ_0 слабо влияет на качество последовательности, но было отмечено, что некоторые значения оказываются неудачными.

При помощи экспериментов и теоретического анализа исследованы и рекомендуются такие значения: $A = 5^{13}$ и $\gamma_0 = 2^{-36}$ для БЭСМ-4; $A = 5^{17}$ и $\gamma_0 = 2^{-40}$ для БЭСМ-6. Для некоторых американских ЭВМ рекомендуются $A = 5^{17}$ и $\gamma_0 = 2^{-42}$; эти цифры связаны с количеством разрядов в мантиссе и порядке числа, поэтому для каждого типа ЭВМ они свои.

Замечание 1. В принципе формулы типа (54) могут давать очень длинные хорошие последовательности, если записывать их в нерекуррентном виде $\gamma_n = \{A^n\gamma_0\}$ и все умножения выполнять без округления. Обычное округление на ЭВМ ухудшает качество псевдослучайных чисел, но тем не менее до 10^5 членов последовательности обычно годятся.

Замечание 2. Качество последовательности улучшается, если ввести в алгоритм (54) небольшие случайные возмущения; например, после нормализации числа γ_i полезно засылать в последние двоичные разряды его мантиссы двоичный порядок числа $A\gamma_{i-1}$.

Строго говоря, закономерность псевдослучайных чисел должна быть незаметна по отношению к требуемому частному применению. Поэтому в несложных или удачно сформулированных задачах можно использовать последовательности не очень хорошего качества, но при этом необходимы специальные проверки.

Произвольное распределение. Для разыгрывания случайной величины с неравномерным распределением $\rho(x)$ можно воспользоваться формулой (52). Разыграем γ и определим ξ из равенства

$$\gamma_i = \int_{-\infty}^{\xi_i} \rho(x)\, dx.$$

Если интеграл берется в конечном виде и формула несложна, то это наиболее удобный способ. Для некоторых важных распределений — Гаусса, Пуассона — соответствующие интегралы не берутся и разработаны специальные способы разыгрывания.

3. Вычисление интеграла. Значение случайной функции $f(\xi)$ заключено между $f(x)$ и $f(x+dx)$, если ξ заключено между x и $x+dx$; вероятность этого события равна $\rho(x)\,dx$. Нетрудно понять, что математическое ожидание случайной функции и ее дисперсия соответственно равны

$$Mf(\xi) = \int_{-\infty}^{+\infty} f(x)\,\rho(x)\,dx, \tag{55}$$

$$Df(\xi) = \int_{-\infty}^{+\infty} [f(x) - Mf(\xi)]^2\,\rho(x)\,dx = Mf^2(\xi) - [Mf(\xi)]^2. \tag{56}$$

Таким образом, *одномерный интеграл можно рассматривать как математическое ожидание случайной функции* $f(\xi)$, аргумент которой есть случайная величина с плотностью распределения $\rho(x)$. На этом основан первый способ статистического вычисления интегралов.

Математическое ожидание можно приближенно вычислить на основании центральной предельной теоремы теории вероятностей; *если* η *есть случайная величина, то среднее арифметическое многих испытаний*

$$\zeta_N = \frac{1}{N}\sum_{i=1}^{N} \eta_i$$

тоже есть случайная величина с тем же математическим ожиданием $M\zeta_N = M\eta$, *причем при* $N \to \infty$ *распределение* ζ_N *стремится к гауссову (нормальному) распределению с дисперсией* $D\zeta_N = \frac{1}{N} D\eta$.

При большом числе испытаний дисперсия ζ_N мала, т. е. значение среднеарифметического с хорошей вероятностью будет близко к математическому ожиданию. Поэтому можно положить

$$\int_{-\infty}^{+\infty} f(x)\,\rho(x)\,dx \approx \frac{1}{N}\sum_{i=1}^{N} f(\xi_i), \tag{57}$$

где ξ — случайная величина с плотностью распределения $\rho(x)$. Оценим дисперсию отдельного испытания по формуле (56), заменяя в ней математические ожидания на суммы типа (57); тогда дисперсия среднеарифметического приближенно равна

$$\Delta_N \approx \frac{1}{N} Df(\xi) \approx \frac{1}{N-1}\left\{\frac{1}{N}\sum_{i=1}^{N} f^2(\xi_i) - \left[\frac{1}{N}\sum_{i=1}^{N} f(\xi_i)\right]^2\right\}. \tag{58}$$

Появление делителя $N-1$ вместо N перед фигурной скобкой обосновывается в теории вероятностей; правда, это существенно только при очень малых числах испытаний.

Ответ в методе статистических испытаний носит вероятностный характер и в принципе может сколь угодно сильно отличаться от точного значения интеграла. Однако, согласно свойствам нормального распределения, с вероятностью 99,7% ошибка не превосходит $3\sqrt{\Delta_N}$. Вероятной называют ошибку $0{,}675\sqrt{\Delta_N}$, соответствующую 50%-ной вероятности; реальная ошибка обычно близка к этой величине — примерно вдвое больше или меньше. Таким образом, выполняя расчеты по формулам (57) — (58), мы одновременно с интегралом получаем неплохую апостериорную оценку ошибки.

При увеличении числа испытаний N погрешность ответа будет убывать примерно, как $1/\sqrt{N}$. Скорости современных ЭВМ позволяют использовать в расчетах $N \sim 10^6$; поэтому на точность выше 0,1% в методе статистических испытаний трудно рассчитывать. В сложных задачах погрешность возрастает до 1—10%.

Поскольку погрешность имеет вероятностный характер, то зависимость $1/\sqrt{N}$ относится не к самой погрешности, а лишь к ширине доверительного интервала. Поэтому нельзя приписывать методу статистических испытаний строгий порядок точности (вроде $p = 1/2$) и *нельзя применять метод Рунге — Ромберга к расчетам, сделанным с различными* N.

Второй способ статистического вычисления применяется к интегралам вида $\int_0^1 f(x)\,dx$, причем на отрезке интегрирования $0 \leqslant f(x) \leqslant 1$. Произвольный интеграл можно привести к такому виду линейной заменой масштабов.

Возьмем случайные числа γ_i, равномерно распределенные на единичном отрезке. Будем рассматривать последовательные пары чисел $(\gamma_{2i}, \gamma_{2i+1})$ как координаты (x_i, y_i) точек в единичном квадрате на плоскости x, y (рис. 21). Эти точки будут случайными и равномерно распределенными в этом квадрате. Поэтому вероятность попадания точки под кривую $y = f(x)$ равна площади, заключенной под кривой, т. е. искомому интегралу. Условие попадания точки под кривую есть $\gamma_{2i+1} < f(\gamma_{2i})$; та доля общего числа испытаний, которая удовлетворяет этому условию, дает приближенное значение интеграла.

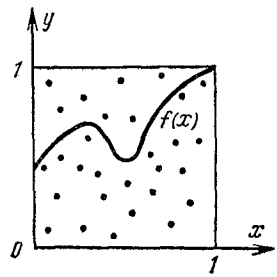


Рис. 21.

4. Уменьшение дисперсии. Точность метода статистических испытаний можно увеличить, выбирая специальную случайную величину. Обозначим $g(x) = f(x)\rho(x)$, тогда исходный интеграл примет вид $F = \int_{-\infty}^{+\infty} g(x)\,dx$. Положим $g(x) = \bar{f}(x)\,\bar{\rho}(x)$, где функция $\bar{\rho}(x) \geqslant 0$ и нормирована на единицу, так что ее можно считать плотностью распределения некоторой случайной величины. Как надо выбрать $\bar{\rho}(x)$, чтобы сделать вычисления наиболее точными, т. е. дисперсию результата — минимальной?

В дисперсии отдельного испытания (56) последнее слагаемое $[Mf(\xi)]^2$ равно квадрату искомого интеграла и тем самым не зависит от выбора $\bar{\rho}(x)$. Значит, надо требовать

$$M\bar{f}^2(\xi) = \int_{-\infty}^{+\infty} \bar{f}^2(x)\,\bar{\rho}(x)\,dx = \min.$$

Добавляя условие нормировки плотности, перепишем эту задачу следующим образом:

$$\int_{-\infty}^{+\infty} [g^2(x)/\bar{\rho}(x)]\,dx = \min, \qquad \int_{-\infty}^{+\infty} \bar{\rho}(x)\,dx = 1.$$

Приравняем нулю вариационные производные по плотности

$$\int_{-\infty}^{+\infty} [g(x)/\bar{\rho}(x)]^2\,\delta\bar{\rho}(x)\,dx = 0, \qquad \int_{-\infty}^{+\infty} \delta\bar{\rho}(x)\,dx = 0.$$

Очевидно, для равенства вариационных производных нулю необходимо и достаточно, чтобы $[g(x)/\bar{p}(x)]^2 = \text{const}$, или $\bar{p}(x) = c|g(x)|$. При этом дисперсия не только минимальна, но даже равна нулю, если $g(x)$ знакопостоянно. В самом деле, тогда $\bar{f}(x) \equiv 1$, и даже одно испытание сразу даст точный результат.

Конечно, на практике взять $\bar{p}(x) = c|g(x)|$ не удается. Для разыгрывания случайной величины с такой плотностью необходимо решить уравнение

$$\gamma_i = \int_{-\infty}^{\xi_i} |g(x)|\,dx,$$

т. е. вычислить искомый интеграл, да еще с переменным верхним пределом! Поэтому обычно подбирают $\bar{p}(x)$ так, чтобы

$$\int_{-\infty}^{\xi} \bar{p}(x)\,dx$$

просто вычислялся, а само $\bar{p}(x)$ было по возможности ближе к $g(x)$.

Смысл увеличения точности нетрудно понять. Если $\bar{p}(x) \sim |g(x)|$, то $\bar{f}(x)$ почти постоянна и все отдельные испытания дают близкие результаты.

Пример. Вычислим

$$F = \int_1^{\infty} e^{-x^2}\,dx.$$

Положим $\rho(x) = cxe^{-x^2}$, где константу $c = 2e$ определим из условия нормировки. Случайную величину с такой плотностью разыграем по формуле

$$\gamma_i = \int_{\xi_i}^{+\infty} \rho(x)\,dx = e^{1-\xi_i^2}, \quad \xi_i = \sqrt{1 - \ln \gamma_i}.$$

Здесь удобнее считать переменным нижний предел интегрирования, что также допустимо. Теперь легко получаем

$$F = \frac{1}{2e}\int_1^{\infty} \frac{1}{x}\,\rho(x)\,dx \approx \frac{1}{2eN}\sum_{i=1}^{N} (1 - \ln \gamma_i)^{-1/2},$$

$$\Delta_N \approx \frac{1}{N-1}\left[\frac{1}{2eN}\sum_{i=1}^{N} (1 - \ln \gamma_i)^{-1} - F^2\right].$$

Приемы уменьшения дисперсии позволяют уменьшать объем вычислений; они широко применяются не только при вычислении интегралов. Например, Бюффон заметил, что можно определить число π, бросая иглу на сетку параллельных линий и регистри-

руя процент случаев, когда игла пересекается с линией (рис. 22). Но для получения трех верных знаков требуется примерно 10^4 испытаний. Оказывается, если брать скрепленные крестом иголки, то для той же точности надо в 25 раз меньше испытаний, а три скрепленные снежинкой иголки дают экономию в 135 раз.

Замечание. Нередко подынтегральная функция имеет на разных участках существенно разное поведение, и ввести хороший единый вес на всем отрезке интегрирования не удается. Тогда выгодно представить интеграл в виде суммы интегралов по отдельным участкам и вычислять каждый из них со своим весом. Это уменьшает дисперсию результата.

5. Кратные интегралы. Второй способ легко обобщается на многомерные интегралы $I=\int\limits_G f(x, y, z)\,dx\,dy\,dz$ по единичному кубу G (для определенности мы выбираем трехмерное пространство), если $0\leqslant f(x, y, z)\leqslant 1$ внутри этого куба. Рассмотрим куб G в четырехмерном пространстве x, y, z, u и случайные равномерно распределенные в нем точки; координатами этих точек будут последовательные четверки случайных чисел γ_{4i+k}, $k=0, 1, 2, 3$. Доля случайных точек, удовлетворяющая неравенству $\gamma_{4i+3}<f(\gamma_{4i}, \gamma_{4i+1}, \gamma_{4i+2})$, даст приближенное значение искомого интеграла.

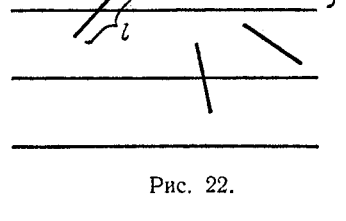


Рис. 22.

Напомним, что чем больше число измерений, тем более жесткими тестами надо проверять качество случайных или псевдослучайных чисел, используемых в расчете.

Замечание 1. Для функций произвольного вида можно получить при том же числе узлов точность в несколько раз более высокую, если использовать не случайные точки, а отрезки так называемых $ЛП_\tau$-последовательностей. Это последовательности многомерных точек, которые обеспечивают более равномерное распределение и самих точек в пространстве, и всех их проекций на грани и ребра многомерного куба. Особенно выгодно в расчетах с такими последовательностями выбирать числа точек $N=2^r$, ибо фактическая ошибка при этом оказывается обычно много меньше, чем по оценке дисперсии.

Замечание 2. Для гладких функций можно получить более хорошие результаты при несложном построении сетки, если выбрать число узлов $N=k^m$, где m — число измерений. Разобьем единичный куб на N кубиков со стороной $1/k$, в каждом кубике выберем одну случайную точку и вычислим по этим точкам интеграл. Дисперсия этого метода есть $\Delta_N=O(N^{-1/2-1/m})$, т. е. она меньше

оценки $O(N^{-1/2})$, получающейся при обычном применении метода Монте-Карло.

Первый способ. Дисперсия второго способа велика, и обычно первый способ статистического вычисления интегралов точнее. Пусть $\rho(x, y, z) \geqslant 0$ есть многомерная плотность распределения некоторой случайной величины. Тогда, аналогично одномерному случаю,

$$\int_G f(x, y, z)\,\rho(x, y, z)\,dx\,dy\,dz = Mf(\xi, \eta, \zeta) \approx$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} f(\xi_i, \eta_i, \zeta_i). \qquad (59)$$

Как найти случайную трехмерную точку с заданным распределением плотности по тройке равномерно распределенных случайных чисел γ_{3i}, γ_{3i+1}, γ_{3i+2}? Для этого надо свести разыгрывание многомерной плотности к последовательным разыгрываниям одномерных случайных величин с плотностями $R(x)$, $R(y)$, $R(z)$.

Для разыгрывания координаты x построим одномерную плотность распределения по этой координате при произвольных остальных координатах

$$R(x) = \int_{-\infty}^{+\infty}\int \rho(x, y, z)\,dy\,dz.$$

Очевидно, функция $R(x)$ неотрицательна и нормирована на единицу, т. е. удовлетворяет предъявляемым к плотности требованиям (51). Поэтому формула разыгрывания есть

$$\gamma_{3i} = \int_{-\infty}^{\xi_i} R(x)\,dx.$$

Теперь одна координата разыграна. Надо найти плотность распределения по второй координате при фиксированной первой координате и произвольной третьей. Если первую координату фиксировать, а по третьей проинтегрировать, то полученное выражение не удовлетворяет условию нормировки (интеграл по y не равен 1). Нормируя его, получим искомую плотность

$$R(y;\ \xi_i) = R^{-1}(\xi_i)\int_{-\infty}^{+\infty} \rho(\xi_i, y, z)\,dz.$$

Вторая координата разыгрывается по формуле

$$\gamma_{3i+1} = \int_{-\infty}^{\eta_i} R(y;\ \xi_i)\,dy.$$

Плотность распределения по третьей координате при фиксированных первых двух координатах пропорциональна $\rho(\xi_i, \eta_i, z)$. Для нормировки надо положить

$$R(z;\ \xi_i, \eta_i) = R^{-1}(\xi_i)\,R^{-1}(\eta_i;\ \xi_i)\,\rho(\xi_i, \eta_i, z);$$

тогда интеграл по z равен единице. Соответственно формула разыгрывания

имеет вид

$$\gamma_{3l+2} = \int_{-\infty}^{\zeta_l} R(z; \xi_l, \eta_l)\, dz.$$

Подставляя полученные координаты в (59), вычислим искомый интеграл. Все, что говорилось в п. 3 о точности расчета, полностью относится к многомерному случаю.

Нелегко подобрать такой вид плотности $\rho(x, y, z)$, чтобы она содержала основные особенности подынтегральной функции и при этом явно бы вычислялись все интегралы, возникающие при разыгрывании координат. Обычно пытаются выделить плотность вида $\rho(x, y, z) = \rho_1(x)\,\rho_2(y)\,\rho_3(z)$, ибо тогда каждая координата разыгрывается независимо от остальных по формуле вида (52), и легче подобрать интегрируемые выражения для одномерных плотностей; к общему виду прибегают, только если точность такого представления недостаточна.

Какими методами удобнее вычислять интегралы — сеточными или статистическими? Точность метода статистических испытаний невелика, и для однократных интегралов он явно невыгоден. Для многих измерений положение резко меняется.

Пусть функция m переменных интегрируется по сеточным формулам p-го порядка точности, причем сетка имеет n шагов по каждой переменной. Тогда полное число узлов есть $N = n^m$, а погрешность расчета $\varepsilon \sim n^{-p}$ (разумеется, предполагается существование p-х кусочно-непрерывных производных функции). Поэтому число узлов, требуемое для достижения данной точности ε, есть $N \sim (1/\varepsilon)^{m/p}$; оно экспоненциально растет при увеличении числа измерений.

При интегрировании методом статистических испытаний погрешность $\varepsilon \sim N^{-1/2}$. Поэтому полное число узлов есть $N \sim (1/\varepsilon)^2$ независимо от числа измерений.

Очевидно, если число измерений $m < 2p$, то сеточные методы требуют меньшего числа узлов и более выгодны. Если $m > 2p$, то статистические методы выгодней. И чем больше число измерений, тем больший выигрыш дают статистические методы.

В многомерном случае редко можно рассчитывать на лучший порядок точности, чем $p = 2$; тогда трехмерные интегралы выгодней вычислять сеточными методами, а пятимерные — уже статистическими. Если же функция имеет только первые производные, то $p = 1$, и статистические методы становятся выгодными даже для трехкратных интегралов.

6. Другие задачи. Методы статистических испытаний применяют не только к численному интегрированию, а и во многих других случаях: задачи массового обслуживания, нахождение критических параметров ядерного реактора, расчет защиты от излучения и т. д.

Например, рассмотрим расчет надежности сложной конструкции, состоящей из многих элементов. Каждый элемент обычно испытывают на изготовляющем

его заводе и снимают так называемую *кривую отказов* (рис. 23, *а*); это вероятность выхода элемента из строя после t часов работы. Чтобы снять такую кривую, надо заставить большую партию элементов работать до поломки. Ясно, что испытывать так готовую конструкцию слишком дорого.

Рассмотрим конструкцию, состоящую из четырех элементов, причем поломка любого элемента выводит конструкцию из строя. Самый ненадежный элемент мы дублируем так, что после поломки элемента включается дублер (рис. 23, *б*). Тогда конструкция сломается, если сломаются оба третьих элемента или любой другой. Если время жизни отдельного элемента есть t_k, то время жизни конструкции равно

$$T = \min(t_1,\ t_2,\ t_3 + t_3',\ t_4). \tag{60}$$

Проведем математическое испытание конструкции. Разыграем выход каждого элемента из строя при помощи равномерно распределенных случайных чисел γ_i. Откладывая γ_1 на оси ординат кривой отказов первого элемента,

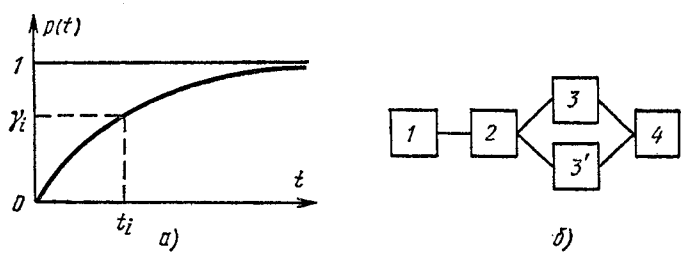


Рис. 23.

получим на оси абсцисс его время жизни (рис. 23, *а*). Время жизни второго элемента определим по числу γ_2 и т. д.; разумеется, отказ каждого дублирующего элемента надо разыграть отдельно. Затем по формуле (60) найдем время жизни конструкции в данном испытании.

Повторяя такое испытание много раз можно найти среднее время работы конструкции

$$T = \frac{1}{N}\sum_i T_i$$

и построить ее кривую отказов. Если надо испытать слегка измененную конструкцию, это можно сделать по той же программе, изменив в ней только формулу (60).

ЗАДАЧИ

1. Составить из обобщенных формул трапеций (8) и средних (17) такую линейную комбинацию, чтобы сократились главные части их погрешностей. Показать, что при этом получается обобщенная формула Симпсона (12).

2. Доказать для формулы трапеций на квазиравномерной сетке асимптотическую оценку погрешности (10).

3. Построить трехточечные разностные выражения для $f'(x)$ на концах отрезка интегрирования. Подставляя их в формулу Эйлера (21), вывести квадратурную формулу Грегори. Найти погрешность этой формулы.

4. Для примера интегрирования функции $f(x) = x|x|$, приведенного в таблице 14, найти по таблице 13 мажорантную оценку погрешности примененных квадратурных формул. Проверить, насколько фактическая ошибка

на каждой сетке отличается от мажорантной. Убедиться, что фактическая погрешность не имеет вида $R \approx \alpha h^{\nu}$.

5. Найти погрешность нелинейной квадратурной формулы (33) на равномерной сетке.

6. Найти погрешность квадратурной формулы Филона (38), аналогичной формуле трапеций.

7. Для слабо меняющихся функций формулы средних и трапеций близки по точности. Почему их аналоги для быстро осциллирующих функций (35) и (37) имеют существенно разную точность?

8. Вывести формулу Филона, соответствующую квадратичной аппроксимации амплитуды по трем соседним узлам. Сравнить ее с формулой Симпсона.

9. Найти асимптотическое выражение погрешности квадратурной формулы (41).

10. Найти формулу для определения числа π способом Бюффона (§ 4, п. 4) и дисперсию этого способа; для этого удобно свести бросания к вычислению интеграла статистическими методами. Сделать то же для иголок, скрепленных крестом и снежинкой.

ГЛАВА V

СИСТЕМЫ УРАВНЕНИЙ

В главе V рассмотрены методы решения систем алгебраических уравнений. В § 1 изложено решение линейных систем методом исключения Гаусса, а также вычисление определителя и обращение матрицы; дан обзор других методов решения этих задач. В § 2 приведены различные методы нахождения корня одного трансцендентного уравнения. В § 3 некоторые из этих методов обобщены на системы нелинейных уравнений.

§ 1. Линейные системы

1. Задачи линейной алгебры. Выделяют четыре основные задачи линейной алгебры: решение системы линейных уравнений $A\boldsymbol{x}=\boldsymbol{b}$, где A — квадратная матрица и $\boldsymbol{x}$, $\boldsymbol{b}$ — векторы; вычисление определителя; нахождение обратной матрицы; определение собственных значений и собственных векторов матрицы. В этом параграфе мы подробно рассмотрим первую задачу и попутно решим вторую и третью. Четвертая задача существенно сложней, и ей посвящена следующая глава.

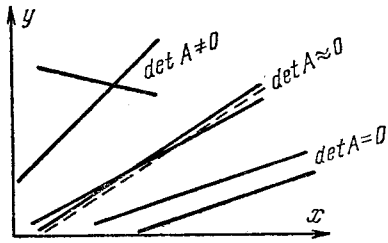


Рис. 24.

Известно, что если $\det A=0$, то система линейных уравнений или не имеет решения, или имеет бесчисленное множество решений. Если же $\det A\neq 0$, то система имеет решение, притом единственное. Дальше мы будем рассматривать только последний случай.

Все эти случаи хорошо иллюстрируются геометрически на системе двух уравнений (рис. 24). Каждому уравнению соответствует прямая в плоскости x, y, а точка пересечения этих прямых есть решение системы (для n уравнений решение есть точка пересечения всех n гиперплоскостей в n-мерном пространстве). Если $\det A=0$, то наклоны прямых равны, и они либо параллельны, либо совпадают. В противном случае прямые имеют единственную точку пересечения.

На практике кроме существования и единственности решения важна еще устойчивость относительно погрешностей правой части и элементов матрицы. Формально перепишем линейную систему в виде $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Варьируя это равенство и определяя вариацию обратной матрицы из соотношения $\delta E = \delta(AA^{-1}) = A\delta A^{-1} + \delta A\, A^{-1} = 0$, получим

$$\delta \boldsymbol{x} = A^{-1}(\delta \boldsymbol{b} - \delta A \cdot \boldsymbol{x}).$$

Формально устойчивость есть, ибо при $\det A \neq 0$ обратная матрица существует. Но если матрица A^{-1} имеет большие элементы, то можно указать такой вид погрешности исходных данных, который сильно изменит решение. В этом случае систему называют плохо обусловленной (по-видимому, плохая обусловленность была известна еще Гауссу). Очевидно, у плохо обусловленных систем $\det A \approx \approx 0$; однако заметим, что этот признак плохой обусловленности является необходимым, но недостаточным.

Плохо обусловленная система геометрически соответствует почти параллельным прямым. При этом небольшое изменение наклона или сдвиг одной прямой сильно меняют положение точки пересечения (рис. 24, пунктир). В многомерном случае геометрическая картина может быть более сложной. Так, для трех переменных возможен случай плохой обусловленности, когда соответствующие трем уравнениям плоскости пересекаются под большими углами (т. е. далеки от параллельности), но линии их попарного пересечения почти параллельны.

В теоретических исследованиях обусловленность часто характеризуют числом $\varkappa = \|A\| \cdot \|A^{-1}\|$. Это число зависит от того, какая норма матриц выбрана, но при любой норме $\varkappa \geqslant 1$. Чем больше это число, тем хуже обусловленность системы; обычно $\varkappa \sim 10^3 - 10^4$ уже означает плохую обусловленность.

В практических расчетах этим определением плохой обусловленности пользуются редко, ибо для его проверки надо находить обратную матрицу, что при плохо обусловленной матрице A нелегко сделать. Чаще ограничиваются проверкой условия $\det A \approx 0$, хотя оно является необходимым, но недостаточным, что видно из простого примера. Положим $A = \varepsilon E$, где E — единичная матрица; тогда $\det A = \varepsilon^n$, и даже при не очень малых ε детерминант высокого порядка n очень мал. Но система с диагональной матрицей хорошо обусловлена, и для нее критерий $\varkappa = \|A\| \cdot \|A^{-1}\| = 1$ наиболее благоприятен.

Методы решения линейных систем делятся на прямые и итерационные. Прямые методы дают решение за конечное число действий, просты и наиболее универсальны; они рассматриваются в этом параграфе. Для систем небольшого порядка $n \lesssim 200$ применяются практически только прямые методы. Итерационные методы приведены в § 3; они выгодны для систем специального вида, со слабо заполненной матрицей очень большого порядка $n \approx 10^3 \div 10^5$. Сравнительно недавно для решения плохо обусловленных систем стали применять методы регуляризации.

2. Метод исключения Гаусса. Как известно из курса линейной алгебры, решение системы линейных уравнений можно выразить по правилу Крамера через отношение определителей. Но этот способ неудобен для вычислений, ибо определитель найти не проще, чем непосредственно решить исходную систему

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned} \tag{1}$$

или короче

$$\sum_{k=1}^{n} a_{ik}x_k = b_i, \qquad 1 \leqslant i \leqslant n.$$

Далее мы увидим, что решить эту систему можно примерно за $^2/_3n^3$ арифметических действий. Но даже если использовать для вычисления определителей наиболее быстрый метод, описанный в п.3, то для нахождения всех требуемых по правилу Крамера определителей надо $^2/_3n^4$ действий! Таким образом, формула Крамера удобна для теоретического исследования свойств решения, но очень невыгодна для его численного нахождения.

Начнем исследование системы (1) с частного случая, когда численное решение находится особенно просто. Пусть матрица системы треугольная, т. е. все элементы ниже главной диагонали равны нулю. Тогда из последнего уравнения сразу определяем x_n. Подставляя его в предпоследнее уравнение, находим x_{n-1} и т. д. Общие формулы имеют вид

$$x_k = \frac{1}{a_{kk}}\left(b_k - \sum_{l=k+1}^{n} a_{kl}x_l\right), \quad k = n,\ n-1,\ \ldots,\ 1, \tag{2}$$

если $a_{kl} = 0$ при $k > l$.

Метод Гаусса для произвольной системы основан на приведении матрицы системы к треугольной. Вычтем из второго уравнения системы (1) первое, умноженное на такое число, чтобы уничтожился коэффициент при x_1. Затем таким же образом вычтем первое уравнение из третьего, четвертого и т. д. Тогда исключатся все коэффициенты первого столбца, лежащие ниже главной диагонали.

Затем при помощи второго уравнения исключим из третьего, четвертого и т. д. уравнений коэффициенты второго столбца. Последовательно продолжая этот процесс, исключим из матрицы все коэффициенты, лежащие ниже главной диагонали.

Запишем общие формулы процесса. Пусть проведено исключение коэффициентов из $k-1$ столбца. Тогда остались такие

уравнения с ненулевыми элементами ниже главной диагонали:

$$\sum_{j=k}^{n} a_{ij}^{(k)} x_j = b_i^{(k)}, \qquad k \leqslant i \leqslant n. \tag{3}$$

Умножим k-ю строку на число

$$c_{mk} = a_{mk}^{(k)} / a_{kk}^{(k)}, \qquad m > k, \tag{4}$$

и вычтем из m-й строки. Первый ненулевой элемент этой строки обратится в нуль, а остальные изменятся по формулам

$$\begin{aligned} a_{ml}^{(k+1)} &= a_{ml}^{(k)} - c_{mk} a_{kl}^{(k)}, \\ b_m^{(k+1)} &= b_m^{(k)} - c_{mk} b_k^{(k)}, \qquad k < m, \ l \leqslant n. \end{aligned} \tag{5}$$

Производя вычисления по этим формулам при всех указанных индексах, исключим элементы k-го столбца. Будем называть такое исключение *циклом* процесса. Выполнение всех циклов называется *прямым ходом* исключения.

Запишем треугольную систему, получающуюся после выполнения всех циклов. При приведении системы к треугольному виду освободятся клетки в нижней половине матрицы системы (1). На освободившиеся места матрицы поставим множители c_{mk}; их следует запоминать, ибо они потребуются при обращении матрицы или уточнении решения. Получим

$$\sum_{k=i}^{n} a_{ik}^{(i)} x_k = b_i^{(i)}, \qquad 1 \leqslant i \leqslant n,$$

$$\begin{array}{ccccccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ c_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\ c_{31} & c_{32} & a_{33}^{(3)} & \dots & a_{3n}^{(3)} & b_3^{(3)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & c_{n3} & \dots & a_{nn}^{(n)} & b_n^{(n)} \end{array} \tag{6}$$

Треугольная система (6) легко решается *обратным ходом* по формулам (2), в которых всем коэффициентам надо приписать вверху (в скобках) индекс строки.

Сделаем несколько замечаний. Исключение по формулам (4)—(5) нельзя проводить, если в ходе расчета на главной диагонали оказался нулевой элемент $a_{kk}^{(k)} = 0$. Но в первом столбце промежуточной системы (3) все элементы не могут быть нулями: это означало бы, что $\det A = 0$. Перестановкой строк можно переместить ненулевой элемент на главную диагональ и продолжить расчет.

Если элемент на главной диагонали $a_{kk}^{(k)}$ мал, то эта строка умножается на большие числа c_{mk}, что приводит к значительным ошибкам округления при вычитаниях. Чтобы избежать этого, каждый цикл всегда начинают с перестановки строк. Среди элементов столбца $a_{mk}^{(k)}$, $m \geqslant k$, находят *главный*, т. е. наибольший по модулю в k-м столбце, и перестановкой строк переводят его на главную диагональ, после чего делают исключения. *В методе Гаусса с выбором главного элемента погрешность округления обычно невелика.* Только для плохо обусловленных систем устойчивость этого метода оказывается недостаточной.

Погрешность округления можно еще уменьшить, если выбирать в каждом цикле элемент $a_{ml}^{(k)}$, m, $l \geqslant k$, максимальный по модулю во всей матрице. Однако точность при этом возрастает не сильно по сравнению со случаем выбора главного элемента, а расчет заметно усложняется, ибо требуется перестановка не только строк, но и столбцов. Этот способ невыгоден для ЭВМ и применяется только при расчетах с небольшим количеством знаков на клавишных машинах.

Для контроля расчета полезно найти *невязки*:

$$r_k = b_k - \sum_{i=1}^{n} a_{ki} x_i, \quad 1 \leqslant k \leqslant n. \tag{7}$$

Если они велики, то это означает грубую ошибку в расчете (ошибка в программе, сбой ЭВМ). Если они малы, а система хорошо обусловлена, то решение найдено достаточно аккуратно. Правда, для плохо обусловленных систем малость невязок не гарантирует хорошей точности решения.

Метод Гаусса с выбором главного элемента надежен, прост и наиболее выгоден для линейных систем общего вида с плотно заполненной матрицей. Он требует примерно n^2 ячеек в оперативной памяти ЭВМ, так что на БЭСМ-4 можно решать системы до 60 порядка. При вычислениях производится $\sim {}^2/_3 n^3$ арифметических действий; из них половина сложений, половина умножений и n делений.

3. Определитель и обратная матрица легко вычисляются методом исключения. В самом деле, вычитание строки из строки не меняет значение определителя. Значит, в процессе исключения элементов (4)—(5) абсолютная величина определителя не меняется, а знак может измениться благодаря перестановке строк. Определитель же треугольной матрицы (6) равен произведению диагональных элементов. Поэтому он вычисляется по формуле:

$$\det A = \pm \prod_{k=1}^{n} a_{kk}^{(k)}, \tag{8}$$

где знак зависит от того, четной или нечетной была суммарная перестановка строк. Для вычисления определителя требуется примерно n^2 ячеек памяти и $^2/_3 n^3$ арифметических действий.

На примере вычисления определителя можно убедиться в экономичности хороших численных методов. Вспомним формальное определение определителя как суммы всевозможных произведений элементов, взятых из разных строк и столбцов. Таких произведений имеется $n! \approx \sqrt{2\pi n}\,(n/e)^n$, и прямое их вычисление уже при небольших $n \approx 30$ требует астрономического числа действий — более 10^{30}, что вряд ли когда-нибудь станет под силу ЭВМ. А метод исключения легко позволяет вычислять определители сотого и более порядка.

Перейдем к вычислению обратной матрицы. Обозначим ее элементы через α_{lm}. Тогда соотношение $AA^{-1} = E$ можно записать так:

$$\sum_{k=1}^{n} a_{ik}\alpha_{kl} = \delta_{il}, \qquad 1 \leqslant i,\ l \leqslant n. \tag{9}$$

Видно, что если рассматривать l-й столбец обратной матрицы как вектор, то он является решением линейной системы (9) с матрицей A и специальной правой частью (в которой на l-м месте стоит единица, а на остальных — нули).

Таким образом, для обращения матрицы надо решить n систем линейных уравнений с одинаковой матрицей A и разными правыми частями. Приведение матрицы A к треугольной по формулам (4)—(5) делается при этом только один раз. В дальнейшем при помощи чисел c_{mk} по формуле (5) преобразуются все правые части, и для каждой правой части делается обратный ход.

При хорошей организации вычислений для обращения матрицы этим методом требуется примерно $2n^2$ ячеек оперативной памяти ЭВМ и $2n^3$ арифметических действий (можно уложиться в $^3/_2 n^2$ ячеек, но это сильно усложняет программу и увеличивает время счета). Заметим, что при обращении матриц контролировать расчет вычислением невязки $R = E - AA^{-1}$ невыгодно: перемножение матриц требует столько же действий ($2n^3$), как и обращение матрицы!

Любопытно отметить, что обращение матрицы сводится к решению n систем линейных уравнений, а требует лишь втрое больше действий, чем решение одной системы уравнений. Это объясняется тем, что при решении линейной системы бóльшая часть вычислений связана с приведением матрицы к треугольному виду, что при обращении матрицы делается только один раз. Обратный ход и преобразования правых частей выполняются много быстрее.

Поэтому, если требуется несколько раз решить линейную систему с одной и той же матрицей, то выгодно привести матрицу к треугольной форме (6) только однажды, используя величины c_{mk} во всех последующих вычислениях.

4. О других прямых методах. Есть очень много других прямых методов решения задач линейной алгебры. Рассмотрим формальные характеристики наиболее известных методов.

Метод оптимального исключения имеет ту же скорость и требует той же памяти, что и метод Гаусса. Но если матрица вводится в оперативную память ЭВМ не вся сразу, а построчно, то для метода Гаусса требуется $^1/_2n^2$ ячеек, а для метода оптимального исключения достаточно $^1/_4n^2$ ячеек. Практическая ценность этого преимущества невелика, ибо построчный ввод означает много обращений к внешней памяти, т. е. сильное увеличение времени расчета. Кроме того, при построчном вводе невозможно выбрать главный элемент, что сказывается на ошибках округления.

Метод окаймления мало отличается от метода оптимального исключения и имеет те же характеристики.

Метод отражений требует вдвое большего числа действий, чем метод Гаусса (оперативная память та же).

Метод ортогонализации втрое медленнее метода Гаусса. Им интересовались в надежде на то, что он позволит решать плохо обусловленные системы. Но выяснилось, что при больших n сама ортогонализация приводит к большой потере точности (сравните с разложением функции по неортогональным функциям), и лучше использовать методы регуляризации.

Метод Жордана имеет ту же скорость, что и метод Гаусса; при решении линейных систем он не дает никаких преимуществ. Но при обращении матрицы он требует меньшей оперативной памяти — всего n^2 ячеек.

Для решения хорошо обусловленных линейных систем общего вида метод Гаусса является одним из лучших; при обращении матрицы немного выгоднее метод Жордана. Но для систем специального вида (например, содержащих много нулевых элементов) существуют более быстрые методы. Некоторые из них будут изложены далее.

5. Прогонка. Пусть матрица A содержит много нулевых элементов, расположенных в матрице не беспорядочно, а плотными массивами на заранее известных местах. Тогда расчет по методу Гаусса можно организовать так, чтобы не включать эти элементы. Тем самым объем вычислений и требуемая память уменьшаются, зачастую очень сильно.

На рис. 25 приведены структуры матриц, которые нередко встречаются в задачах физики и техники и допускают такое ускорение расчета; горизонтальными линиями изображены положения ненулевых элементов, окаймлены границы массивов нулевых и ненулевых элементов. К таким матрицам относятся ленточные (*а*), ящичные (*в*), квазитреугольные (*д*), почти треугольные (*е*) и многие другие *). Можно показать, что при обходе нулевых элементов решение системы с почти треугольной матрицей требует

*) Напомним принятую терминологию. Матрица называется верхней треугольной, если все элементы ниже главной диагонали равны нулю ($a_{ik}=0$ при $i>k$); аналогично определяется нижняя треугольная матрица. Почти треугольной называется матрица, элементы которой удовлетворяют соотношению $a_{ik}=0$ при $i>k+1$, т. е. ненулевые элементы имеются не только в верхнем треугольнике, но и в примыкающей к нему «боковой диагонали». Трехдиагональной называется матрица, у которой ненулевые элементы имеются только на главной диагонали и примыкающих к ней, т. е. $a_{ik}=0$ при $|i-k|>1$. Нетрудно записать определения других типов матриц, изображенных на рис. 25.

всего $2n^2$ действий, а с ленточной — даже $^1/_2k^2n$, где k — ширина ленты, т. е. выигрыш во времени счета очень велик.

Выбор наибольшего элемента в таких расчетах делать нельзя, ибо перестановка столбцов разрушает специальную структуру матрицы. В матрицах с симметричной структурой недопустим

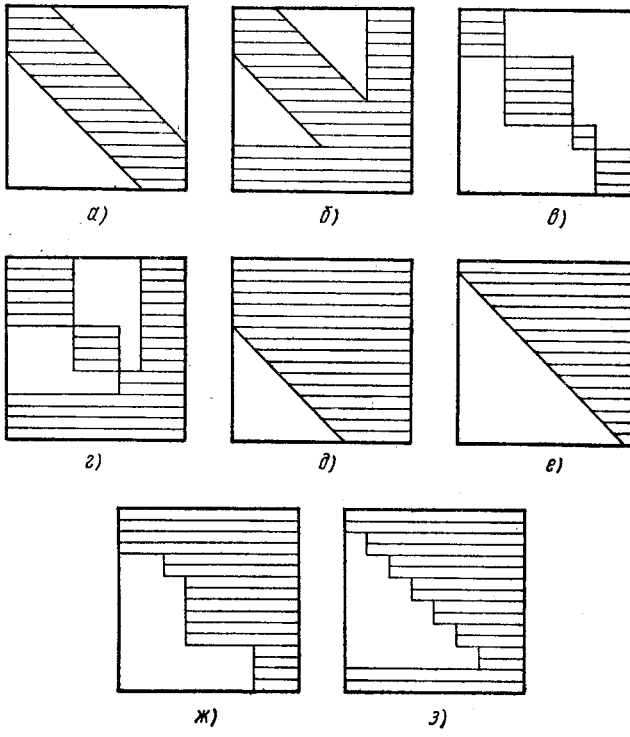


Рис. 25.

даже выбор главного элемента. Но обычно в этом нет необходимости, поскольку подобные физические задачи приводят, как правило, к хорошо обусловленным матрицам с большими элементами на главной диагонали, для которых ошибки округления в методе Гаусса невелики.

Наиболее важным частным случаем метода Гаусса является *метод прогонки*, применяемый к системам с трехдиагональной матрицей (они часто встречаются при решениях краевых задач для дифференциальных уравнений второго порядка). Такие системы обычно записывают в каноническом виде

$$\begin{gathered} a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i, \quad 1 \leqslant i \leqslant n, \\ a_1 = c_n = 0. \end{gathered} \tag{10}$$

Формула (10) называется разностным уравнением второго порядка, или трехточечным уравнением. В этом случае прямой ход (без выбора главного элемента) сводится к исключению элементов a_i. Получается треугольная система, содержащая в каждом уравнении только два неизвестных, x_i и x_{i+1}. Поэтому формулы обратного хода имеют следующий вид:

$$x_i = \xi_{i+1} x_{i+1} + \eta_{i+1}, \quad i = n,\ n-1,\ \ldots,\ 1. \tag{11}$$

Уменьшим в формуле (11) индекс на единицу и подставим в уравнение (10):

$$a_i(\xi_i x_i + \eta_i) - b_i x_i + c_i x_{i+1} = d_i.$$

Выражая отсюда x_i через x_{i+1}, получим

$$x_i = \frac{c_i}{b_i - a_i\xi_i} x_{i+1} + \frac{a_i\eta_i - d_i}{b_i - a_i\xi_i}.$$

Чтобы это выражение совпало с (11), надо, чтобы стоящие в его правой части дроби были равны соответственно ξ_{i+1} и η_{i+1}. Отсюда получим удобную запись формул прямого хода

$$\begin{aligned} &\xi_{i+1} = c_i/(b_i - a_i\xi_i), \\ &\eta_{i+1} = (a_i\eta_i - d_i)/(b_i - a_i\xi_i), \quad i = 1,\ 2,\ \ldots,\ n. \end{aligned} \tag{12}$$

Попутно можно найти определитель трехдиагональной матрицы

$$\det A = \prod_{i=1}^{n} (a_i\xi_i - b_i). \tag{13}$$

Вычисления по формулам прогонки (12) — (11) требуют всего $3n$ ячеек памяти и $9n$ арифметических действий, т. е. они гораздо экономнее общих формул метода исключения.

В формулах прямого и обратного хода начало счета «замаскировано»: для начала (*развязки*) расчета формально требуется задать величины ξ_1, η_1 и x_{n+1}, которые неизвестны. Однако перед этими величинами в формулах стоят множители a_1 или $\xi_{n+1} \sim c_n$, равные нулю. Это позволяет начать вычисления, полагая, например, $\xi_1 = \eta_1 = x_{n+1} = 0$.

Покажем, что если выполнено условие *преобладания диагональных элементов*

$$|b_i| \geqslant |a_i| + |c_i| \tag{14}$$

(причем хотя бы для одного i имеет место неравенство), то в формулах прямого хода (12) не возникает деления на нуль, и тем самым исходная система (10) имеет единственное решение. Для этого предположим, что $|\xi_i| < 1$ при некотором значении индекса. Тогда легко проверяется цепочка неравенств

$$|\xi_{i+1}| = |c_i|/|b_i - a_i\xi_i| \leqslant |c_i|/(|b_i| - |a_i||\xi_i|) \leqslant$$
$$\leqslant |c_i|/(|c_i| + |a_i| - |a_i| \times |\xi_i|) < 1.$$

Поскольку можно положить $\xi_1 = 0$, отсюда по индукции следует $|\xi_i| < 1$; значит, $|b_i - a_i\xi_i| > |c_i| \geqslant 0$, что и требовалось доказать. При выполнении условия (14) формулы прогонки не только безавостны, но и устойчивы относительно ошибок округления и позволяют успешно решать системы уравнений с несколькими сотнями неизвестных.

Условие (14) является достаточным, но не необходимым условием устойчивости прогонки. Конечно, можно построить примеры неустойчивости при несоблюдении этого условия. Но в практических расчетах для хорошо обусловленных систем типа (10) прогонка часто оказывается достаточно устойчивой даже при нарушении условия преобладания диагональных элементов.

Заметим, что к линейным системам с трехдиагональной матрицей обычно приводят трехточечные разностные схемы для дифференциальных уравнений второго порядка (глава VIII, § 2).

6. Метод квадратного корня. Этот метод пригоден только для линейных систем с эрмитовой *) матрицей $A = A^H$, и формулы расчета при этом несколько сложней, чем в методе Гаусса. Зато метод квадратного корня вдвое быстрей метода Гаусса.

Метод основан на представлении эрмитовой матрицы системы в виде произведения трех матриц

$$A = S^H D S. \tag{15}$$

Здесь D — диагональная матрица с элементами $d_{ii} = \pm 1$, S — верхняя треугольная матрица ($s_{ik} = 0$ при $i > k$), а S^H — эрмитово сопряженная к ней нижняя треугольная матрица. Для полной определенности разложения потребуем вещественности и положительности диагональных элементов $s_{ii} > 0$.

Перепишем соотношение (15) в следующем виде:

$$a_{kl} = \sum_{i=1}^{n} s_{ik}^* d_{ii} s_{il} = \sum_{i=1}^{\min(k,\, l)} d_{ii} s_{ik}^* s_{il};$$

ограничение верхнего предела в сумме связано с обращением в нуль элементов S ниже главной диагонали. Последнее равенство можно записать в такой форме:

$$a_{kk} = \sum_{i=1}^{k} d_{ii} |s_{ik}|^2,$$

$$a_{kl} = \sum_{i=1}^{k} d_{ii} s_{ik}^* s_{il}, \quad k < l,$$

*) Напомним, что матрица называется эрмитовой, если $a_{lk} = a_{ki}^*$; эрмитова матрица с вещественными элементами является симметричной.

или окончательно

$$d_{kk} = \operatorname{sign}\left(a_{kk} - \sum_{i=1}^{k-1} d_{ii} |\, s_{ik} \,|^2\right),$$

$$s_{kk} = \sqrt{\left|a_{kk} - \sum_{i=1}^{k-1} d_{ii} |\, s_{ik} \,|^2\right|}, \tag{16}$$

$$s_{kl} = (a_{kl} - \sum_{i=1}^{k-1} d_{ii} s_{ik}^* s_{il})/(s_{kk} d_{kk}) \quad \text{при} \quad k+1 \leqslant l \leqslant n.$$

В этих формулах сначала полагаем $k=1$ и последовательно вычисляем все элементы первой строки матрицы S; при $k=1$ все суммы в формулах (16) отсутствуют. Затем полагаем $k=2$ и вычисляем вторую строку и т. д.

Когда все элементы матриц найдены, то решение линейной системы $A\boldsymbol{x}=\boldsymbol{b}$ сводится к последовательному решению трех систем, двух треугольных и одной диагональной:

$$S^H \boldsymbol{z} = \boldsymbol{b}, \quad D\boldsymbol{y} = \boldsymbol{z}, \quad S\boldsymbol{x} = \boldsymbol{y}, \tag{17}$$

что делается обычным обратным ходом по формулам

$$\begin{aligned}
&y_1 = b_1/s_{11}d_{11},\\
&y_i = (b_i - \sum_{l=1}^{i-1} d_{ll} y_l s_{li}^*)/s_{ii}d_{ii}, \quad i = 2,\ 3,\ \ldots,\ n;\\
&x_n = y_n/s_{nn},\\
&x_i = (y_i - \sum_{l=i+1}^{n} s_{il} x_l)/s_{ii}, \quad i = n-1,\ n-2,\ \ldots,\ 1.
\end{aligned} \tag{18}$$

Определитель матрицы вычисляется по формуле

$$\det A = \prod_{i=1}^{n} d_{ii} s_{ii}^2. \tag{19}$$

Метод квадратного корня требует примерно $^1/_3 n^3$ арифметических действий, т. е. при больших n он вдвое быстрей метода Гаусса, и занимает вдвое меньше ячеек памяти. Это понятно, ибо метод использует информацию о симметрии матрицы.

Кроме того, для ленточной, ящичной и некоторых других структур матрицы A (рис. 25, *а—г*) матрица S будет иметь аналогичную структуру, т. е. содержать массивы нулевых элементов на заранее известных местах. Учет этого позволяет сильно сократить объем вычислений, как и в методе Гаусса. Однако заметим, что для ленточных матриц с узкой лентой, особенно для трехдиагональных, метод квадратного корня по скорости мало отличается от метода Гаусса и может быть даже медлен-

ней, ибо среди производящихся в нем действий есть извлечение корня, медленно выполняемое на ЭВМ.

Наиболее частый на практике случай эрмитовой матрицы — это вещественная симметричная матрица A. Тогда никаких комплексных чисел при вычислениях не возникает, так что матрица S тоже вещественная. Если вдобавок матрица A положительно определенная (для этого необходима и достаточна положительность всех ее главных миноров), то все $d_{ii}=1$, и формулы (16) — (19) можно немного упростить.

Расчет по формулам (16) невозможен, если при некотором значении индекса элемент $s_{kk}=0$ (в частности, $s_{11}=0$ при $a_{11}=0$). От этого можно избавиться, переставляя на место a_{kk} другой диагональный элемент $a_{ll}\neq 0$, т. е. надо переставить и строки и столбцы, на пересечении которых лежат эти два элемента.

Метод квадратного корня применяют в основном при численном решении интегральных уравнений Фредгольма с симметричным ядром, ибо эта задача сводится к линейной системе с симметричной матрицей, обычно не содержащей нулевых элементов (при регуляризации таких задач симметрия матрицы сохраняется).

7. Плохо обусловленные системы. Если система $A\boldsymbol{x}=\boldsymbol{b}$ плохо обусловлена, то это значит, что погрешности коэффициентов матрицы и правых частей или погрешности округления при расчетах могут сильно исказить решение. Для уменьшения погрешностей округления можно было бы провести на ЭВМ расчет с двойным или тройным числом знаков. Но при наличии погрешности коэффициентов это бесполезно, и нужно регуляризовать исходную задачу.

Исходную систему (1) можно переписать в эквивалентной форме $(A\boldsymbol{x}-\boldsymbol{b},\ A\boldsymbol{x}-\boldsymbol{b})=0$. Если коэффициенты матрицы или правые части известны не точно, то решение также является приближенным. Поэтому на самом деле мы можем требовать только приближенного равенства $(A\boldsymbol{x}-\boldsymbol{b},\ A\boldsymbol{x}-\boldsymbol{b})\approx 0$. Задача становится неопределенной, и для определенности надо добавить какие-то дополнительные условия.

Таким условием может быть требование, чтобы решение как можно меньше отклонялось от заданного вектора $\boldsymbol{x}_0$, т. е. чтобы скалярное произведение $(\boldsymbol{x}-\boldsymbol{x}_0,\ \boldsymbol{x}-\boldsymbol{x}_0)$ было минимально. Тогда регуляризованная задача формулируется следующим образом:

$$(A\boldsymbol{x}-\boldsymbol{b},\ A\boldsymbol{x}-\boldsymbol{b})+\alpha\,(\boldsymbol{x}-\boldsymbol{x}_0,\ \boldsymbol{x}-\boldsymbol{x}_0)=\min,\quad \alpha>0. \tag{20а}$$

Это можно переписать в эквивалентной форме

$$(\boldsymbol{x},\ A^HA\boldsymbol{x})-2\,(\boldsymbol{x},\ A^H\boldsymbol{b})+(\boldsymbol{b},\ \boldsymbol{b})+ \\ +\alpha[(\boldsymbol{x},\ \boldsymbol{x})-2\,(\boldsymbol{x},\ \boldsymbol{x}_0)+(\boldsymbol{x}_0,\ \boldsymbol{x}_0)]=\min, \tag{20б}$$

где α — малый положительный управляющий параметр и A^H —

эрмитово сопряженная матрица. Варьируя $\boldsymbol{x}$ в (20), получим следующее уравнение:

$$(A^H A + \alpha E)\,\boldsymbol{x} = A^H \boldsymbol{b} + \alpha \boldsymbol{x}_0, \tag{21}$$

где E — единичная матрица. Решая его (например, методом исключения Гаусса), найдем регуляризованное значение $\boldsymbol{x}_\alpha$, зависящее от параметра α.

Остановимся на выборе параметра. Если $\alpha = 0$, то система (21) переходит в плохо обусловленную систему вида (1). Если же α велико, то регуляризованная система (21) будет хорошо обусловленной благодаря присутствию в левой части хорошо обусловленной матрицы αE; но сама система (21) при большом α сильно отличается от исходной системы, и регуляризованное решение $\boldsymbol{x}_\alpha$ не будет близким к искомому решению. Поэтому слишком малое или слишком большое α непригодны. Очевидно, оптимальным будет наименьшее значение α, при котором обусловленность системы (21) еще удовлетворительна.

Для фактического нахождения оптимума вычисляют невязку $\boldsymbol{r}_\alpha = A\boldsymbol{x}_\alpha - \boldsymbol{b}$ и сравнивают ее по норме с известной погрешностью правых частей $\delta \boldsymbol{b}$ и с влиянием погрешности коэффициентов матрицы $\delta A \cdot \boldsymbol{x}$. Если α слишком велико, то невязка заметно больше этих погрешностей, если слишком мало — то заметно меньше. Проводят серию расчетов с различными α; оптимальным считают тот, в котором $\| \boldsymbol{r}_\alpha \| \approx \| \delta \boldsymbol{b} \| + \| \delta A \cdot \boldsymbol{x} \|$.

Для выбора $\boldsymbol{x}_0$ нужны дополнительные соображения; если их нет, то полагают $\boldsymbol{x}_0 = 0$.

Обоснование изложенного метода дано в главе XIV. Заметим, что матрица системы (21) эрмитова, так что для ее решения можно применять метод квадратного корня.

§ 2. Уравнение с одним неизвестным

1. Исследование уравнения. Пусть задана непрерывная функция $f(x)$ и требуется найти все или некоторые корни уравнения

$$f(x) = 0. \tag{22}$$

Эта задача распадается на несколько задач. Во-первых, надо исследовать количество, характер и расположение корней. Во-вторых, найти приближенные значения корней. В-третьих, выбрать из них интересующие нас корни и вычислить их с требуемой точностью.

Первая и вторая задачи решаются аналитическими и графическими методами. Например, многочлен

$$P_n(x) = \sum_{k=0}^{n} a_k x^k$$

имеет n комплексных корней, не обязательно различных, и все корни лежат внутри круга

$$|x_p| \leqslant 1 + \frac{1}{|a_n|} \max(|a_0|, |a_1|, \ldots, |a_{n-1}|).$$

Правда, эта оценка не оптимальная; модули всех корней могут быть много меньше правой части неравенства, в чем легко убедиться на примере многочлена $P(x) = \sum_{k=1}^{n} (x-k)$.

Когда ищутся только действительные корни уравнения, то полезно составить таблицу значений $f(x)$. Если в двух соседних узлах таблицы функция имеет разные знаки, то между этими узлами лежит нечетное число корней уравнения (по меньшей мере один). Если эти узлы близки, то, скорее всего, корень между ними только один. Но выявить по таблице корни четной кратности сложно.

По таблице можно построить график функции $y = f(x)$ и графически найти точки его пересечения с осью абсцисс. Этот способ более нагляден и дает неплохие приближенные значения корней. Во многих задачах техники такая точность уже достаточна. В технике еще популярны графические методы решения уравнений (номография). Построение графика зачастую позволяет выявить даже корни четной кратности.

Иногда удается заменить уравнение (22) эквивалентным ему уравнением $\varphi(x) = \psi(x)$, в котором функции $y_1 = \varphi(x)$ и $y_2 = \psi(x)$ имеют несложные графики. Например, уравнение $x \sin x - 1 = 0$ удобно преобразовать к виду $\sin x = 1/x$. Абсциссы точек пересечения этих графиков будут корнями исходного уравнения.

Приближенные значения корней уточняют различными итерационными методами. Рассмотрим наиболее эффективные из них.

2. Дихотомия (деление пополам). Пусть мы нашли такие точки x_0, x_1, что $f(x_0) f(x_1) \leqslant 0$, т. е. на отрезке $[x_0, x_1]$ лежит не менее одного корня уравнения. Найдем середину отрезка $x_2 = (x_0 + x_1)/2$ и вычислим $f(x_2)$. Из двух половин отрезка выберем ту, для которой $f(x_2) f(x_{\text{гран}}) \leqslant 0$, ибо один из корней лежит на этой половине. Затем новый отрезок опять делим пополам и выберем ту половину, на концах которой функция имеет разные знаки, и т. д. (рис. 26).

Если требуется найти корень с точностью ε, то продолжаем деление пополам до тех пор, пока длина отрезка не станет меньше 2ε. Тогда середина последнего отрезка даст значение корня с требуемой точностью.

Дихотомия проста и очень надежна: к простому корню она сходится для любых непрерывных функций $f(x)$, в том числе недифференцируемых; при этом она устойчива к ошибкам округления. Скорость сходимости невелика; за одну итерацию точность

увеличивается примерно вдвое, т. е. уточнение трех цифр требует 10 итераций. Зато точность ответа гарантируется.

Перечислим недостатки метода. Для начала расчета надо найти отрезок, на котором функция меняет знак. Если в этом отрезке несколько корней, то заранее неизвестно, к какому из них сойдется процесс (хотя к одному из них сойдется). Метод неприменим к корням четной кратности. Для корней нечетной высокой кратности он сходится, но менее точен и хуже устойчив к ошибкам округления, возникающим при вычислении $f(x)$. Наконец, на системы уравнений дихотомия не обобщается.

Дихотомия применяется тогда, когда требуется высокая надежность счета, а скорость сходимости малосущественна.

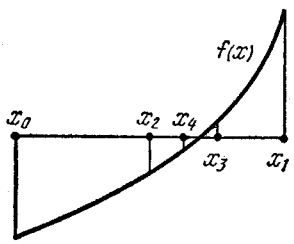


Рис. 26.

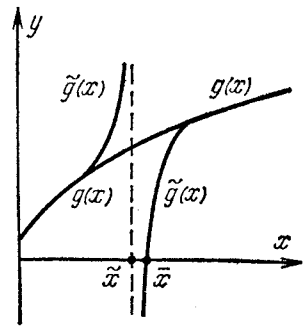


Рис. 27.

3. Удаление корней. Один из недостатков дихотомии — сходимость неизвестно к какому корню — имеется почти у всех итерационных методов. Его можно устранить удалением уже найденного корня.

Если $\bar{x}_1$ есть простой корень уравнения (22) и $f(x)$ липшицнепрерывна, то вспомогательная функция $g(x)=f(x)/(x-\bar{x}_1)$ непрерывна, причем все нули функций $f(x)$ и $g(x)$ совпадают, за исключением $\bar{x}_1$, ибо $g(\bar{x}_1)\neq 0$. Если $\bar{x}_1$ — кратный корень уравнения (22), то он будет нулем $g(x)$ кратности на единицу меньше; остальные нули обеих функций по-прежнему будут одинаковы.

Поэтому найденный корень можно удалить, т. е. перейти к функции $g(x)$. Тогда нахождение остальных нулей $f(x)$ сведется к нахождению нулей $g(x)$. Когда мы найдем какой-нибудь корень $\bar{x}_2$ функции $g(x)$, то этот корень тоже можно удалить, вводя новую вспомогательную функцию $\varphi(x)=g(x)/(x-\bar{x}_2)=f(x)/(x-\bar{x}_1)\times\times(x-\bar{x}_2)$. Так можно последовательно найти все корни $f(x)$.

Строго говоря, мы находим лишь приближенное значение корня $\tilde{x}\approx\bar{x}$. А функция $\tilde{g}(x)=f(x)/(x-\tilde{x}_1)$ имеет нуль в точке $\bar{x}_1$ и полюс в близкой к ней точке $\tilde{x}_1$ (рис. 27); только на некото-

ром расстоянии от этого корня она близка к $g(x)$. Чтобы это не сказывалось при нахождении следующих корней, надо вычислять каждый корень с высокой точностью, особенно если он кратный или вблизи него расположен другой корень уравнения.

Кроме того, в любом методе окончательные итерации вблизи определяемого корня рекомендуется делать не по функциям типа $g(x)$, а по исходной функции $f(x)$. Последние итерации, вычисленные по функции $g(x)$, используются при этом в качестве нулевого приближения. Особенно важно это при нахождении многих корней, ибо чем больше корней удалено, тем меньше нули вспомогательной функции $G(x)=f(x)/\prod_i (x-\tilde{x}_i)$ соответствуют остальным нулям функции $f(x)$.

Учитывая эти предосторожности и вычисляя корни с 8—10 верными десятичными цифрами, зачастую можно определить десятка два корней, о расположении которых заранее ничего не известно (в том числе корней высокой кратности $p \sim 5$).

4. Метод простых итераций. Заменим уравнение (22) эквивалентным ему уравнением $x=\varphi(x)$. Это можно сделать многими способами, например, положив $\varphi(x) \equiv x+\psi(x)f(x)$, где $\psi(x)$ — произвольная непрерывная знакопостоянная функция. Выберем некоторое нулевое приближение x_0 и вычислим дальнейшие приближения по формулам

$$x_{n+1}=\varphi(x_n), \quad n=0,\ 1,\ 2,\ \ldots \tag{23}$$

Очевидно, если x_n стремится к некоторому пределу $\bar{x}$, то этот предел есть корень исходного уравнения.

Исследуем условия сходимости. Если $\varphi(x)$ имеет непрерывную производную, тогда

$$x_{n+1}-\bar{x}=\varphi(x_n)-\varphi(\bar{x})=(x_n-\bar{x})\,\varphi'(\xi), \tag{24}$$

где точка ξ лежит между точками x_n и $\bar{x}$. Поэтому если всюду $|\varphi'(x)| \leqslant q < 1$, то отрезки $|x_n-\bar{x}|$ убывают не медленней членов геометрической прогрессии со знаменателем $q<1$ и последовательность x_n сходится при любом нулевом приближении. Если $|\varphi'(\bar{x})|>1$, то в силу непрерывности $|\varphi'(x)|$ больше единицы и в некоторой окрестности корня; в этом случае итерации не могут сходиться. Если $|\varphi'(\bar{x})|<1$, но вдали от корня $|\varphi'(x)|>1$, то итерации сходятся, если нулевое приближение выбрано достаточно близко к корню; при произвольном нулевом приближении сходимости может не быть.

Эти рассуждения переносятся на липшиц-непрерывные функции практически без изменений.

Очевидно, что чем меньше q, тем быстрей сходимость. Вблизи корня асимптотическая сходимость определяется величиной $|\varphi'(\bar{x})|$ и будет особенно быстрой при $\varphi'(\bar{x})=0$.

Значит, успех метода зависит от того, насколько удачно выбрано $\varphi(x)$. Например, для извлечения квадратного корня, т. е. решения уравнения $x^2=a$, можно положить $\varphi(x)=a/x$ или $\varphi(x)={}^1/_2[x+(a/x)]$ и соответственно написать такие итерационные процессы:

$$x_{n+1}=\frac{a}{x_n} \quad \text{или} \quad x_{n+1}=\frac{1}{2}\left(x_n+\frac{a}{x_n}\right). \tag{25}$$

Первый процесс вообще не сходится, а второй сходится при любом $x_0>0$; сходится он очень быстро, ибо $\varphi'(\bar{x})=0$. Второй процесс используют при извлечении корня на клавишных машинах.

Каков практический критерий сходимости, т. е. когда надо прекращать итерации (23)? Из (24) видно, что если $\varphi'(x)<0$, то итерации попеременно оказываются то с одной, то с другой стороны корня, так что корень заключен в интервале (x_n, x_{n-1}). Это надежная, хотя несколько грубая оценка. Но она неприменима при $\varphi'^{s}(x)>0$, когда итерации сходятся к корню монотонно, т. е. с одной стороны.

Вблизи корня итерации сходятся примерно как геометрическая прогрессия со знаменателем $q=(x_n-x_{n-1})/(x_{n-1}-x_{n-2})$. Чтобы сумма дальнейших ее членов не превосходила ε, должен выполняться критерий сходимости

$$\left|q\frac{x_n-x_{n-1}}{1-q}\right|=\frac{(x_n-x_{n-1})^2}{|2x_{n-1}-x_n-x_{n-2}|}<\varepsilon. \tag{26}$$

При выполнении этого условия итерации можно прекращать.

Легко заметить, что выражение в левой части есть поправка Эйткена (4.24). Если последние три простые итерации уточнить процессом Эйткена, то это обычно заметно повышает точность расчета и позволяет ограничиться меньшим числом итераций.

Метод простых итераций и почти все другие итерационные методы имеют важное достоинство: в них не накапливаются ошибки вычислений. Ошибка вычислений эквивалентна некоторому ухудшению очередного приближения. Но это отразится только на числе итераций, а не на точности окончательного результата. Подобные методы устойчивы даже по отношению к грубым ошибкам (сбоям ЭВМ), если только ошибка не выбрасывает очередное приближение за пределы области сходимости.

При обработке эксперимента возникают *стохастические задачи*, где ошибки определения функции велики и носят случайный характер. Погрешность функции δf приводит к погрешности корня $\delta\bar{x}=\delta f(\bar{x})/f'(\bar{x})$. Однако поскольку ошибки носят случайный характер, то методами статистики можно определить корень гораздо более точно, чем по указанной оценке. Рассмотрим простые итерации

$$x_{n+1}=x_n-a_n f(x_n)\operatorname{sign} f'(x) \tag{27a}$$

при дополнительных условиях

$$a_n > 0, \quad \sum_{n=1}^{\infty} a_n^2 < \infty, \quad \sum_{n=1}^{\infty} a_n = \infty, \tag{27б}$$

которым удовлетворяет, например, последовательность $a_n = (1/n)$. Доказано [47], что $x_n \to \bar{x}$ при $n \to \infty$ с вероятностью единица. Использование в формуле (27а) знака производной не означает, что надо вычислять эту производную: достаточно лишь определить ее знак по разности двух значений функции.

Напомним, что стремлением к пределу с вероятностью единица называется сходимость к пределу $\bar{x}$ в подавляющем большинстве случаев (т. е. при разных нулевых приближениях и разных выборах последовательностей a_n), хотя в отдельных случаях процесс может не сходиться или сходиться к другому пределу. Стохастические процессы сходятся медленно, поэтому к детерминированным задачам их нецелесообразно применять.

5. Метод Ньютона. Он называется также методом *касательных* или методом *линеаризации*. Если x_n есть некоторое приближение к корню $\bar{x}$, а $f(x)$ имеет непрерывную производную, то уравнение (22) можно преобразовать следующим образом:

$$0 = f(\bar{x}) = f(x_n + (\bar{x} - x_n)) = f(x_n) + (\bar{x} - x_n) f'(\xi).$$

Приближенно заменяя $f'(\xi)$ на значение в известной точке x_n, получим такой итерационный процесс:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{28}$$

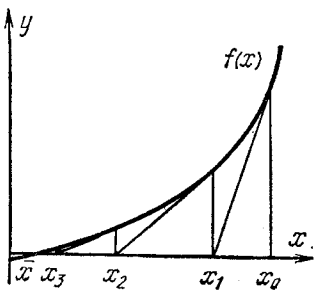


Рис. 28.

Геометрически этот процесс означает замену на каждой итерации графика $y = f(x)$ касательной к нему (рис. 28).

Метод Ньютона можно рассматривать как частный случай метода простых итераций, если положить $\varphi(x) = x - [f(x)/f'(x)]$. Тогда $\varphi'(x) = (ff''/f'^2)$. Если $\bar{x}$ есть p-кратный корень уравнения (22), то вблизи него $f(x) \approx a(x - \bar{x})^p$; отсюда нетрудно получить $\varphi'(\bar{x}) = (p-1)/p$, т. е. $0 \leqslant \varphi'(\bar{x}) < 1$. Для простого корня $p = 1$ и $\varphi'(\bar{x}) = 0$. Используя результаты п. 4, можно сформулировать следующие условия сходимости итераций (28). *Если нулевое приближение выбрано достаточно близко к корню, ньютоновские итерации сходятся*; скорость сходимости велика для простого корня и соответствует скорости геометрической прогрессии для кратного корня. При произвольном нулевом приближении итерации сходятся, если всюду $|ff''| < (f')^2$; в противном случае сходимость будет не при любом нулевом приближении, а только в некоторой окрестности корня.

Сходимость вблизи любого корня монотонна, что легко видеть из рис. 28; но вдали от корня возможна немонотонность итераций. Отметим, что рис. 28 указывает еще на одно достаточное

условие сходимости итераций. Пусть $f''(x) \geqslant 0$ справа от корня на отрезке $[\bar{x}, a]$; если x_0 выбрано также справа от корня на этом отрезке, то итерации (28) сходятся, причем монотонно. То же будет, если $f''(x) \leqslant 0$ слева от корня на отрезке $[b, \bar{x}]$, и на этом же отрезке выбрано нулевое приближение. Таким образом, итерации сходятся к корню с той стороны, с которой $f(x) f''(x) \geqslant 0$.

Оценим скорость сходимости вблизи простого корня. По определению простых итераций, $\bar{x} - x_n = \varphi(\bar{x}) - \varphi(x_{n-1})$. Разлагая правую часть по формуле Тейлора и учитывая равенство $\varphi'(\bar{x}) = 0$, получим

$$x_n - \bar{x} = {}^1/_2 (x_{n-1} - \bar{x})^2 \varphi''(\xi), \qquad \xi \in (x_{n-1}, \bar{x}), \tag{29}$$

т. е. погрешность очередного приближения примерно равна квадрату погрешности предыдущего приближения. Например, если $(n-1)$-я итерация давала 3 верных знака, то n-я даст примерно 6 верных знаков, а $(n+1)$-я — примерно 12 знаков. Это иллюстрирует быструю сходимость вблизи корня. Разумеется, вдали от корня подобные соображения неприменимы.

Таблица 16

$f(x) \equiv x^2 - 4 = 0$

n	x_n, метод Ньютона	x_n, метод секущих
0	1,0000	1,0000
1	2,5000	2,5000
3	2,0500	1,8571
4	2,0001	1,9836

Если вычисляется корень высокой кратности, то $f'(x)$ в знаменателе формулы (28) становится малой вблизи корня. Чтобы не было потери точности, отношение $f(x)/f'(x)$ надо вычислять достаточно аккуратно. К остальным погрешностям расчета метод Ньютона хорошо устойчив.

Для нахождения корней произвольной дифференцируемой функции чаще всего применяют метод Ньютона, особенно если известны разумные начальные приближения для корней.

Пример. Рассмотрим решение уравнения $f(x) \equiv x^2 - a = 0$. Тогда общая формула метода Ньютона (28) принимает вид

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

Мы получили вторую формулу (25), которая, как отмечалось раньше, позволяет очень быстро находить квадратный корень с помощью только сложения и деления. Для иллюстрации в таблице 16 приведен ход итераций при извлечении квадратного корня из $a = 4$. Видно, что сходимость очень быстрая; несмотря на неважное нулевое приближение, уже третья итерация дает точность 0,005%. Попутно можно заметить, что вблизи корня итерации сходятся с одной стороны, т. е. монотонно, хотя первая итерация дает переброс на другую сторону корня.

6. Процессы высоких порядков. В методе простых итераций выберем функцию $\varphi(x)$ так, чтобы выполнялось

$$\varphi'(\bar{x})=\varphi''(\bar{x})=\ldots=\varphi^{(p-1)}(\bar{x})=0, \quad \varphi^{(p)}(\bar{x})\neq 0. \tag{30}$$

Итерационный процесс (23) с такой функцией называют *стационарным процессом p-го порядка*. Скорость сходимости этого процесса вблизи корня можно получить из следующих равенств:

$$x_{n+1}-\bar{x}=\varphi(x_n)-\varphi(\bar{x})=\frac{1}{p!}(x_n-\bar{x})^p\varphi^{(p)}(\xi), \quad \xi\in(x_n,\bar{x}). \tag{31а}$$

Если $|\varphi^{(p)}(x)|\leqslant M_p$, то отсюда следует

$$|x_n-\bar{x}|\leqslant(M_p/p!)^{(p^n-1)/(p-1)}\,|x_0-\bar{x}|^{p^n}. \tag{31б}$$

Сходимость при $p=1$ называют линейной (это собственно метод простых итераций), при $p=2$ — квадратичной (например, метод Ньютона), а при $p=3$ — кубической. Очевидно, чем больше p, тем быстрей сходятся итерации вблизи корня; к сожалению, тем меньше область гарантированной сходимости этих приближений.

Фактически у процессов высокого порядка выход на их асимптотическую скорость сходимости (31) обычно наступает только тогда, когда итерации уже почти сошлись, т. е. для получения всех верных разрядов числа осталось сделать одну — три итерации. Поэтому такие процессы (за исключением метода парабол) редко употребляются.

7. Метод секущих *) [48]. В методе Ньютона требуется вычислять производную функции, что не всегда удобно. Можно заменить производную первой разделенной разностью, найденной по двум последним итерациям, т. е. заменить касательную секущей. Тогда вместо процесса (28) получим

$$x_{n+1}=x_n-\frac{(x_n-x_{n-1})f(x_n)}{f(x_n)-f(x_{n-1})}. \tag{32}$$

Для начала процесса надо задать x_0 и x_1 (рис. 29). Такие процессы, где для вычисления очередного приближения надо знать два предыдущих, называют *двухшаговыми*.

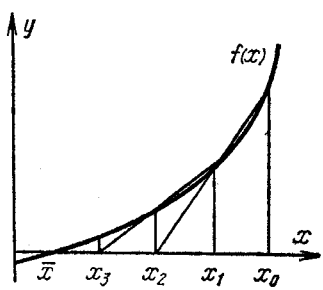


Рис. 29.

Эти, казалось бы, небольшие изменения сильно влияют на характер итераций. Например, сходимость итераций может быть немонотонной не только вдали от корня, но и в малой окрестности корня. Скорость сходимости также изменяется. Иллюстрацией служит приведенный в таблице 16 расчет по методу секущих; для удобства сравнения с методом Ньютона первые два

*) В математической литературе это название нередко употребляют для другого метода, который по его геометрической интерпретации следовало бы называть методом хорд.

приближения взяты одинаковыми. Видно, что метод секущих сходится медленнее.

Скорость сходимости можно оценить, разлагая все функции в (32) по формуле Тейлора с центром $\bar{x}$. Получим с точностью до бесконечно малых более высокого порядка

$$x_{n+1}-\bar{x}=a\,(x_n-\bar{x})\,(x_{n-1}-\bar{x}),\qquad a=\frac{f''(\bar{x})}{2f'(\bar{x})}. \tag{33}$$

Решение этого рекуррентного соотношения естественно искать в виде $x_{n+1}-\bar{x}=a^{\alpha}(x_n-\bar{x})^{\beta}$, аналогичном (29) или (31а). Подставляя эту форму в соотношение (33), получим

$$\alpha\beta=1,\quad \beta^2-\beta-1=0. \tag{34}$$

Только положительный корень β квадратного уравнения (34) соответствует убыванию ошибки, т. е. сходящемуся процессу. Следовательно, в методе секущих

$$\begin{gathered}x_{n+1}-\bar{x}=a^{1/\beta}(x_n-\bar{x})^{\beta},\\ \beta=\tfrac{1}{2}(\sqrt{5}+1)\approx 1{,}62,\quad (1/\beta)\approx 0{,}62,\end{gathered} \tag{35}$$

в то время как в методе Ньютона ошибка убывает быстрей (соответствуя $\beta=2$). Но в методе Ньютона на каждой итерации надо вычислять и функцию, и производную, а в методе секущих — только функцию. Поэтому при одинаковом объеме вычислений в методе секущих можно сделать вдвое больше итераций и получить более высокую точность.

В знаменателе формулы (32) стоит разность значений функции. Вдали от корня это несущественно; но вблизи корня, особенно корня высокой кратности, значения функции малы и очень близки. Возникает потеря значащих цифр, приводящая к «разболтке» счета. Это ограничивает точность, с которой можно найти корень; для простых корней это ограничение невелико, а для кратных может быть существенным.

Заметим, что приводить формулу (32) к общему знаменателю не следует: увеличится потеря точности в расчетах.

От «разболтки» страхуются так называемым *приемом Гарвика*. Выбирают не очень малое ε, ведут итерации до выполнения условия $|x_{n+1}-x_n|<\varepsilon$ и затем продолжают расчет до тех пор, пока $|x_{n+1}-x_n|$ убывают. Первое же возрастание обычно означает начало «разболтки»; тогда расчет прекращают и последнюю итерацию не используют.

8. Метод парабол. Метод секущих можно рассматривать как замену функции $f(x)$ интерполяционным многочленом первой степени, проведенным по узлам x_n, x_{n-1}. По трем последним

итерациям можно построить интерполяционный многочлен второй степени, т. е. заменить график функции параболой. Запишем интерполяционный многочлен в форме Ньютона

$$\mathcal{P}_2(x) = f(x_n) + (x - x_n) f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1}) f(x_n, x_{n-1}, x_{n-2}).$$

Приравнивая его нулю, получим квадратное уравнение

$$az^2 + bz + c = 0, \tag{36а}$$

где

$$\begin{gathered} z = x - x_n, \quad a = f(x_n, x_{n-1}, x_{n-2}), \\ b = a(x_n - x_{n-1}) + f(x_n, x_{n-1}), \quad c = f(x_n). \end{gathered} \tag{36б}$$

Тот из двух корней квадратного уравнения (36), который меньше по модулю, определяет новое приближение $x_{n+1} = x_n + z$.

Очевидно, для начала расчета надо задать три первых приближения x_0, x_1, x_2 (обычно наугад выбирают три числа), т. е. процесс является *трехшаговым*.

Метод парабол построен по образцу методов третьего порядка. Однако замена производных разделенными разностями приводит к существенному уменьшению скорости сходимости. Рассуждениями, аналогичными рассуждениям в п. 7, можно показать, что вблизи простого корня выполняется соотношение

$$|x_{n+1} - \bar{x}| \approx \left| \frac{f'''(\bar{x})}{6f'(\bar{x})} \right|^{0{,}42} |x_n - \bar{x}|^{1{,}84}, \tag{37}$$

т. е. сходимость даже медленнее квадратичной. Вблизи кратного корня сходимость еще медленнее (хотя и более быстрая, чем линейная). Заметим, что строить аналогичные методы с использованием интерполяционного многочлена еще более высокой степени невыгодно: сходимость все равно будет медленней квадратичной, а расчет сильно усложняется.

В методе парабол «разболтка» счета вблизи корня сказывается еще сильней, чем в методе секущих, ибо в расчете участвуют вторые разности. Тем не менее корни можно найти с хорошей точностью; для определения оптимального числа итераций удобно пользоваться приемом Гарвика, описанном в п. 7.

Метод парабол имеет важное достоинство. Даже если все предыдущие приближения действительны, уравнение (36) может привести к комплексным числам. Поэтому процесс может естественно сойтись к комплексному корню исходного уравнения. В методах простых итераций, касательных или секущих для сходимости к комплексному корню может потребоваться задание комплексного начального приближения (если $f(x)$ вещественна при вещественном аргументе).

Корни многочлена. Метод парабол оказался исключительно эффективным для нахождения всех корней многочлена высокой степени. Если $f(x)$ — алгебраический многочлен, то, хотя

сходимость метода при произвольном начальном приближении и не доказана, на практике итерации всегда сходятся к какому-нибудь корню, причем быстро. Для многочлена частное $f(x)/(x-\bar{x})$ есть тоже многочлен; поэтому последовательно удаляя найденные корни, можно найти все корни исходного многочлена.

Замечание 1. Если $f(x)$ — многочлен высокой степени, то возникают дополнительные трудности. Многочлен быстро возрастает при увеличении аргумента, поэтому в программе для ЭВМ должна быть страховка от переполнения. Обычно вводят масштабные множители, величина которых связана с диапазоном изменения аргумента.

Замечание 2. Наибольшие по модулю корни многочлена высокой степени могут быть очень чувствительны к погрешности коэффициентов при старших степенях. Например, корнями многочлена

$$P_{20}(x)=\prod_{k=1}^{20}(x-k)$$

являются последовательные целые числа $\bar{x}_k=1, 2, \ldots, 20$. А слегка измененный многочлен $\tilde{P}_{20}(x)=P_{20}(x)-2^{-23}x^{19}$ имеет такие корни:

$$1{,}0;\ 2{,}0;\ \ldots;\ 8{,}0;\ 8{,}9;\ 10{,}1\pm 0{,}6i;\ \ldots;\ 19{,}5\pm 1{,}9i;\ 20{,}8$$

(здесь приведен только один знак после запятой). Кратные или близкие корни могут быть слабо устойчивыми даже при меньших степенях многочлена.

Замечание 3. Для удаления вычисленных корней надо делить многочлен. Это вносит погрешность округления в коэффициенты и влияет на точность нахождения следующих корней. На практике отмечено, что если сначала удалять меньшие по модулю корни, точность падает мало, но если начать удаление с больших корней, точность может упасть катастрофически. Поэтому за начальное приближение берут $x_0=-1$, $x_1=+1$. $x_2=0$; тогда итерации обычно сходятся к наименьшему по модулю корню. Его удаляют и по такому же начальному приближению ищут следующий корень и т. д. При такой организации вычислений потеря точности будет небольшой.

9. Метод квадрирования. Этот метод позволяет найти все корни многочлена. Запишем многочлен n-й степени двумя способами:

$$P_n(x)=\sum_{k=0}^{n}a_kx^k=a_n\prod_{l=1}^{n}(x-\bar{x}_l), \tag{38}$$

где $\bar{x}_l$ — корни многочлена. Сравнивая обе формы записи, получим *равенства*

Виета

$$\sum_{k=1}^{n} \bar{x}_k = -\frac{a_{n-1}}{a_n}, \quad \sum_{k>l} \bar{x}_k \bar{x}_l = +\frac{a_{n-2}}{a_n}, \quad \sum_{k>l>m} \bar{x}_k \bar{x}_l \bar{x}_m = -\frac{a_{n-3}}{a_n}, \ \ldots \tag{39}$$

Предположим, что корни многочлена сильно различаются по абсолютной величине: $|\bar{x}_1| \gg |\bar{x}_2| \gg \ldots \gg |\bar{x}_n|$. Тогда из равенств Виета получаются приближенные значения корней

$$\bar{x}_1 \approx -\frac{a_{n-1}}{a_n}, \ \bar{x}_2 \approx -\frac{a_{n-2}}{a_{n-1}}, \ \ldots, \ \bar{x}_n \approx -\frac{a_0}{a_1}. \tag{40}$$

Если модули корней мало различаются, то эти формулы слишком неточны. Но квадраты модулей будут различаться сильней, чем сами модули. Поменяем в многочлене (38) знаки всех корней, что эквивалентно смене знака у всех нечетных коэффициентов. Умножая измененный многочлен на исходный, получим

$$\left[\sum_{k=1}^{n} a_k x^k\right] \cdot \left[\sum_{l=1}^{n} (-1)^{n-l} a_l x^l\right] = a_n^2 \prod_{m=1}^{n} (x^2 - \bar{x}_m^2) \equiv Q_n(z), \quad z = x^2. \tag{41}$$

Многочлен $Q_n(z)$ имеет n-ю степень и называется *квадрированным многочленом*. Его корни равны квадратам корней исходного многочлена. Нахождение квадрированного многочлена сравнительно трудоемко; его коэффициенты можно вычислять по формулам

$$b_l = \sum_{m=-\alpha}^{\alpha} (-1)^{l-m} a_{l+m} a_{l-m}, \quad \alpha = \min(l, \ n-l). \tag{42}$$

Для фактического выполнения этих вычислений удобно записать произведения коэффициентов с нужными знаками в форме таблицы 17; знаки ставятся в шахматном порядке. Произведения суммируются по косым строкам, как указано стрелками.

Таблица 17

	a_0	a_1	a_2	...	a_n
a_0	a_0a_0	$-a_0a_1$	a_0a_2	...	$\pm a_0a_n$
a_1	$-a_1a_0$	a_1a_1	$-a_1a_2$	...	$\mp a_1a_n$
a_2	a_2a_0	$-a_2a_1$	a_2a_2	...	$\pm a_2a_n$
...	. . . . .				
a_n	$\pm a_na_0$	$\mp a_na_1$	$\pm a_na_2$	...	a_na_n

Для квадрированного многочлена корни по формулам (40), где вместо a_k надо вставить b_k, определятся точнее. Если результат будет мало отличаться от предыдущего, то на нем можно остановиться. Если отличие заметное, то квадрирование надо повторить. Повторяя квадрирование много раз, получим

быстро сходящийся итерационный процесс (можно показать, что его сходимость будет квадратичной).

При различающихся по модулю корнях после многократного квадрирования выполняются соотношения

$$\left|\frac{b_{n-1}}{b_n}\right| \geqslant \left|\frac{b_{n-2}}{b_{n-1}}\right| \geqslant \ldots \geqslant \left|\frac{b_0}{b_1}\right|.$$

Если среди корней есть равные по модулю (в частности, кратные), то это соотношение нарушается. Например, если второй корень будет двукратным, то получим

$$\left|\frac{b_{n-1}}{b_n}\right| \geqslant \left|\frac{b_{n-2}}{b_{n-1}}\right| \sim \left|\frac{b_{n-3}}{b_{n-2}}\right| \geqslant \left|\frac{b_{n-4}}{b_{n-3}}\right| \geqslant \ldots$$

Если корни равны только по модулю, но $\bar{x}_k \neq \bar{x}_l$ (например, комплексно-сопряженные корни), то это случайное совпадение. Такие корни удобнее всего разделять сдвигом. Рассмотрим многочлен $R(x) = P_n(x - \xi)$, где ξ — случайно выбранное комплексное число. Корни многочлена $R(x)$ будут уже не равны между собой по модулю, ибо они отличаются от корней исходного многочлена на комплексную величину ξ.

Кратные корни разделить сдвигом нельзя. Для них надо составлять специальные формулы, явно учитывающие, какой корень какую кратность имеет. Определить кратность корней можно по поведению отношений соседних коэффициентов.

Обратный переход от корней квадрированных уравнений к корням исходного уравнения осуществляется последовательным извлечением квадратного корня. При этом появляются ложные корни; их надо обнаружить подстановкой в исходное уравнение и отбросить.

Метод квадрирования позволяет легко выполнить все расчеты на клавишных машинах. Он не требует задания какого-либо нулевого приближения. Но для программирования на ЭВМ этот метод не особенно удобен. Во-первых, после нескольких квадрирований в расчете возникают обычно большие числа, и возможно переполнение, от которого приходится страховаться введением масштабных множителей. Во-вторых, при наличии кратных корней требуется произвести довольно громоздкий логический анализ и применить нестандартные формулы вычисления.

§ 3. Системы нелинейных уравнений

1. Метод простых итераций. Систему нелинейных уравнений можно кратко записать в векторном виде

$$\boldsymbol{f}(\boldsymbol{x}) = 0 \tag{43}$$

или более подробно в координатном виде

$$f_k(x_1, x_2, \ldots, x_n) = 0, \quad 1 \leqslant k \leqslant n.$$

Такие системы решают практически только итерационными методами. Нулевое приближение в случае двух переменных можно найти графически: построить на плоскости (x_1, x_2) кривые $f_1(x_1, x_2) = 0$ и $f_2(x_1, x_2) = 0$ и найти точки их пересечения. Для трех и более переменных (а также для комплексных корней) удовлетворительных способов подбора нулевых приближений нет.

Рассмотрим метод *простых итераций*, называемый также методом *последовательных приближений*. Аналогично одномерному

случаю, заменим нелинейную систему (43) эквивалентной системой специального вида $\boldsymbol{x}=\boldsymbol{\varphi}(\boldsymbol{x})$. Выберем некоторое нулевое приближение и дальнейшие приближения найдем по формулам

$$\boldsymbol{x}^{(s+1)}=\boldsymbol{\varphi}(\boldsymbol{x}^{(s)})$$

или

$$x_k^{(s+1)}=\varphi_k\left(x_1^{(s)},\ x_2^{(s)},\ \ldots,\ x_n^{(s)}\right),\qquad 1\leqslant k\leqslant n. \tag{44}$$

Если итерации сходятся, то они сходятся к решению уравнения (предполагается, что решение существует).

Сходимость итераций исследуем так же, как для одной переменной. Обозначим компоненты решения через $\bar{x}_k$ и преобразуем погрешность очередной итерации

$$x_k^{(s+1)}-\bar{x}_k=\varphi_k\left(x_1^{(s)},\ \ldots,\ x_n^{(s)}\right)-\varphi_k\left(\bar{x}_1,\ \ldots,\ \bar{x}_n\right)=$$

$$=\varphi_k(\boldsymbol{x}^{(s)})-\varphi_k(\bar{\boldsymbol{x}})=\left[\partial\varphi_k(\xi_k)/\partial l\right]\rho(\boldsymbol{x}^{(s)},\ \bar{\boldsymbol{x}})=$$

$$=\sum_{i=1}^{n}\left(x_i^{(s)}-\bar{x}_i\right)\left[\partial\varphi_k(\xi_k)/\partial x_i\right],$$

где l — направление, соединяющее многомерные точки $\boldsymbol{x}^{(s)}$ и $\bar{\boldsymbol{x}}$, а ξ_k — некоторая точка, лежащая между ними на этом направлении. Это равенство означает, что вектор погрешности нового приближения равен матрице производных, умноженной на вектор погрешности предыдущего приближения. Если какая-нибудь норма матрицы производных $\{\partial\varphi_k(\xi_k)/\partial x_i\}$, согласованная с некоторой нормой вектора, меньше единицы, то норма погрешности убывает от итерации к итерации по геометрической прогрессии. Это означает линейную сходимость метода.

На практике удобней рассматривать матрицу с элементами $M_{ki}=\max|\partial\varphi_k/\partial x_i|$. Нормы этой матрицы мажорируют соответствующие нормы матрицы производных, поэтому *достаточным условием сходимости является* $\|M_{ki}\|<1$. При использовании разных норм матриц это условие принимает такие формы:

$$\sum_{i=1}^{n}M_{ki}<1,\quad\text{или}\quad\sum_{k=1}^{n}M_{ki}<1,\quad\text{или}\quad\sum_{k,i=1}^{n}M_{ki}^2<1. \tag{45}$$

Каждая норма матрицы согласована с определенной нормой вектора, но в конечномерном пространстве все нормы эквивалентны. Значит, если итерации сходятся в одной норме, то они сходятся и во всех остальных нормах.

Поскольку сходимость линейная, то оканчивать итерации можно по критерию сходимости (26), выполнение которого надо проверять для каждой компоненты. Линейная сходимость довольно медленна; поэтому полезно уточнять результат процессом Эйткена по трем последним итерациям.

Сами вычисления в методе последовательных приближений просты. Но зато сложно найти такую систему $\boldsymbol{x}=\boldsymbol{\varphi}(\boldsymbol{x})$, которая была бы эквивалентна исходной системе $\boldsymbol{f}(\boldsymbol{x})=0$ и одновременно обеспечивала бы сходимость.

Сходимость метода нередко удается улучшить. В методе простых итераций найденное приближение $x_k^{(s+1)}$ используется только для вычисления следующей итерации. Можно использовать его уже на данной итерации для вычисления следующих компонент:

$$x_k^{(s+1)}=\varphi_k\left(x_1^{(s+1)},\ x_2^{(s+1)},\ \ldots,\ x_{k-1}^{(s+1)},\ x_k^{(s)},\ x_{k+1}^{(s)},\ \ldots,\ x_n^{(s)}\right). \qquad (46)$$

Сходимость этого варианта метода тоже линейная; детально ее исследовать не будем. При ручных расчетах можно еще ускорить сходимость за счет перестановок отдельных уравнений на основе анализа их невязок $r_k^{(s)}=x_k^{(s)}-\varphi_k(x^{(s)})=x_k^{(s)}-x_k^{(s+1)}$; но для расчетов на ЭВМ это неудобно, ибо обычно такой анализ полуинтуитивен и плохо алгоритмизируется.

2. Метод Ньютона. Пусть известно некоторое приближение $\boldsymbol{x}^{(s)}$ к корню $\bar{\boldsymbol{x}}$. Как и для одной переменной, запишем исходную систему (43) в виде $\boldsymbol{f}(\boldsymbol{x}^{(s)}+\Delta\boldsymbol{x})=0$, где $\Delta\boldsymbol{x}=\bar{\boldsymbol{x}}-\boldsymbol{x}^{(s)}$. Разлагая эти уравнения в ряды и ограничиваясь первыми дифференциалами, т. е. линеаризуя функцию, получим

$$\sum_{i=1}^{n}\frac{\partial f_k(x^{(s)})}{\partial x_i}\Delta x_i^{(s)}=-f_k(\boldsymbol{x}^{(s)}),\quad 1\leqslant k\leqslant n. \qquad (47)$$

Это система уравнений, линейных относительно приращений $\Delta x_i^{(s)}$; все коэффициенты этой системы выражаются через последнее приближение $\boldsymbol{x}^{(s)}$. Решив эту систему (например, методом исключения), найдем новое приближение $\boldsymbol{x}^{(s+1)}=\boldsymbol{x}^{(s)}+\Delta\boldsymbol{x}^{(s)}$.

Как и для одной переменной, метод Ньютона можно формально свести к методу последовательных приближений, положив $\boldsymbol{\varphi}(\boldsymbol{x})=\boldsymbol{x}-[\partial\boldsymbol{f}/\partial\boldsymbol{x}]^{-1}\boldsymbol{f}(\boldsymbol{x})$, где $[\partial\boldsymbol{f}/\partial\boldsymbol{x}]^{-1}$ есть матрица, обратная матрице производных. Аналогично проводится теоретический анализ условий сходимости. Однако достаточное условие сходимости, записанное в координатной форме, здесь имеет настолько сложный вид, что проверить его выполнимость почти никогда не удается. Отметим только очевидный результат: *в достаточно малой окрестности корня итерации сходятся, если* $\det[\partial\boldsymbol{f}/\partial\boldsymbol{x}]\neq 0$, *причем сходимость квадратичная.*

Следовательно, если нулевое приближение выбрано удачно, то метод Ньютона сходится, причем очень быстро (обычно за 3—5 итераций). Поэтому на практике этот метод используют чаще всего.

В отличие от метода простых итераций, для метода Ньютона хорошим критерием окончания итераций является условие

$\| x^{(s)} - x^{(s+1)} \| \leqslant \varepsilon$. В самом деле, вблизи корня ньютоновские итерации сходятся квадратично, поэтому если этот критерий выполнен, то $\| x^{(s+1)} - \bar{x} \| \approx \varepsilon^2 \ll \varepsilon$. Выбирая $\varepsilon \approx 10^{-5} - 10^{-6}$, можно получить решение с десятком верных знаков.

Вычисления в методе Ньютона несколько сложнее, чем при простых итерациях, ибо на каждой итерации требуется находить матрицу производных и решать систему линейных уравнений. Поэтому в некоторых учебниках рекомендуют такой прием: вычислить матрицу $[\partial f/\partial x]^{-1}$ только на начальной итерации и использовать ее на всех остальных итерациях.

Однако сходимость при этом видоизменении становится линейной, причем обычно не с малой константой, ибо матрица производных на начальной итерации может заметно отличаться от окончательной. Поэтому скорость сходимости заметно уменьшается и требуемое число итераций возрастает.

3. Методы спуска. Рассмотрим функцию $\Phi(x) = \sum_{k=1}^{n} |f_k(x)|^2$. Она неотрицательна и обращается в нуль в том и только в том случае, если $f(x) = 0$. Таким образом, решение исходной системы уравнений (43) будет одновременно нулевым минимумом скалярной функции многих переменных $\Phi(x)$.

Иногда бывает проще искать такой минимум, чем решать систему уравнений. Методы поиска минимума будут рассмотрены в главе VII. В основном это итерационные методы спуска, т. е. движения в направлении убывания функции. Все методы спуска для гладких функций сходятся, но зачастую — довольно плохо. Поэтому на хорошую точность полученного решения трудно рассчитывать; однако этим способом обычно можно найти разумное приближение, которое потом можно уточнять методом Ньютона.

Надо помнить, что метод спуска в зависимости от выбора нулевого приближения может сойтись к любому минимуму функции. А функция $\Phi(x)$ может иметь ненулевые локальные минимумы, которые не являются решениями исходной системы уравнений.

4. Итерационные методы решения линейных систем иногда дополняют, а иногда заменяют прямые методы.

Решая линейную систему (1) общего вида методом исключения, попутно можно проверить, насколько хорошо она обусловлена. Решение, найденное прямым методом, из-за ошибок округления будет приближенным. Нетрудно проверить, что поправки к нему удовлетворяют уравнениям

$$\sum_{i=1}^{n} a_{ki} \Delta x_i = r_k, \quad 1 \leqslant k \leqslant n, \tag{48}$$

где r_k — невязки (7). Это линейная система с той же матрицей, что исходная система (1). Решим ее, используя ранее найденные коэффициенты c_{mk} (т. е. почти не увеличивая общего объема вычислений). Ранее отмечалось, что в методе Гаусса невязки малы. Если величины Δx_i тоже окажутся малыми, то система хорошо обусловлена; если большими — то плохо. В последнем случае зачастую удается уточнить решение, рассматривая (48) как итерационный процесс Ньютона и делая 2—3 итерации.

Для уточнения обратной матрицы тоже есть итерационный процесс с квадратичной сходимостью

$$A_{S+1}^{-1} = A_S^{-1} + A_S^{-1} R_S, \quad R_S = E - A A_S^{-1}. \tag{49}$$

Однако каждая итерация этого процесса требует $4n^3$ арифметических действий, т. е. вдвое больше, чем прямое обращение матрицы по методам Гаусса или Жордана. Поэтому в практических вычислениях этот процесс теперь не применяют.

При очень плохой обусловленности матрицы оба описанных метода уточнения могут потребовать вычислений с двойным и более числом знаков, но тогда лучше применять регуляризирующие алгоритмы.

Есть важная группа задач, приводящая к линейным системам с сотнями и тысячами неизвестных. Это решение двумерных и трехмерных уравнений в частных производных эллиптического типа при помощи разностных схем. Матрицы таких систем слабо заполнены, но расположение нулевых элементов таково, что метод исключения не может полностью использовать особенности структуры матрицы и приводит к большому объему вычислений. Кроме того, в методе исключения матрицы таких систем не помещаются в оперативной памяти ЭВМ, а обращение к внешней памяти еще более увеличивает время расчета.

Подобные линейные системы зачастую выгодно решать итерационными методами. Современные итерационные методы мы рассмотрим в главе XII, посвященной эллиптическим уравнениям. Здесь упомянем только два старых метода, уже не используемых в практических вычислениях, но удобных для некоторых теоретических оценок.

Один из них — стационарный метод простых итераций, основанный на приведении системы $Ax = b$ к эквивалентной системе специального вида:

$$x = Cx + d. \tag{50}$$

Запись итерационного процесса очевидна. Для сходимости итераций достаточно, если какая-нибудь норма $\| C \| < 1$. В этой задаче известно даже необходимое и достаточное условие сходимости — модули всех собственных значений матрицы C должны быть меньше единицы; но на практике этим условием трудно воспользоваться, ибо найти собственные значения обычно сложнее, чем решить линейную систему.

К виду (50) систему можно привести, например, выделением диагональных элементов

$$x_k = \frac{1}{a_{kk}} \left(b_k - \sum_{i \neq k} a_{ki} x_i \right), \qquad 1 \leqslant k \leqslant n. \tag{51}$$

В этой записи легко учесть наличие нулей в матрице A и при умножении матрицы на вектор суммировать только по ненулевым элементам. При использовании различных норм матрицы достаточные условия сходимости итераций принимают вид

$$\sum_{i \neq k} \left| \frac{a_{ki}}{a_{kk}} \right| < 1, \quad \text{или} \quad \sum_{k \neq i} \left| \frac{a_{ki}}{a_{kk}} \right| < 1, \quad \text{или} \quad \sum_k \sum_{i \neq k} \left| \frac{a_{ki}}{a_{kk}} \right|^2 < 1, \tag{52}$$

что означает *преобладание диагональных элементов* матрицы (сравните условия устойчивости метода прогонки (14)).

Если метод сходится, то корень уравнения существует. Таким образом, преобладание диагональных элементов матрицы в смысле одного из неравенств (52) является признаком того, что система линейных уравнений (1) имеет решение, т. е. $\det A \neq 0$. Этим признаком мы будем часто пользоваться при исследовании разрешимости неявных разностных схем.

Заметим, что чем меньше $\| C \|$, тем быстрей сходятся итерации. Наилучшие современные методы основаны на специальных способах выбора матрицы C, при которых ее норма становится минимальной в некотором смысле.

Второй — метод Зейделя. Для уравнения (51) он имеет такой вид:

$$x_k^{(s+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{i=1}^{k-1} a_{ki} x_i^{(s+1)} - \sum_{i=k+1}^{n} a_{ki} x_i^{(s)} \right), \quad 1 \leqslant k \leqslant n. \qquad (53)$$

Необходимое и достаточное условие его сходимости не совпадает с условием сходимости простых итераций, и в разных условиях он может быть выгоден или невыгоден. Отметим один признак сходимости: если матрица A — эрмитова и положительно определенная, то метод Зейделя в форме (53) совпадает с методом покоординатного спуска для решения задачи на минимум квадратичной формы $(x, Ax) - 2(b, x) = \min$; как будет показано в главе VII, метод покоординатного спуска сходится, что обеспечивает сходимость метода Зейделя в данном случае.

ЗАДАЧИ

1. Записать для почти треугольной матрицы (рис. 25, *д*) формулы метода исключения Гаусса с обходом нулевых элементов.

2. Показать, что при преобразовании эрмитовых матриц, изображенных на рис. 25, *а*—*г*, в произведение (15) структура треугольных матриц подобна структуре исходных матриц.

3. Доказать, что первый итерационный процесс (25) не сходится, а второй сходится при любом (положительном) нулевом приближении.

4. Найти асимптотическую скорость сходимости метода секущих (32) вблизи корня кратности p.

5. Доказать, что скорость сходимости метода парабол вблизи простого корня определяется формулой (37); исследовать сходимость вблизи кратного корня.

6. Доказать, что метод квадрирования сходится квадратично.

7. Вывести формулы метода квадрирования для случая, когда наибольший по модулю корень — двукратный.

8. Доказать, что итерационный процесс (49) для нахождения обратной матрицы сходится квадратично.

ГЛАВА VI

АЛГЕБРАИЧЕСКАЯ ПРОБЛЕМА СОБСТВЕННЫХ ЗНАЧЕНИЙ

В главе VI рассмотрены методы нахождения собственных значений и собственных векторов квадратных матриц. В § 1 изложены необходимые сведения по линейной алгебре, рассмотрена устойчивость проблемы собственных значений и даны простые (но сравнительно медленные) численные методы решения. Наиболее быстрые численные методы нахождения всех собственных значений и собственных векторов эрмитовых матриц разобраны в § 2, а неэрмитовых матриц — в § 3. В § 4 изложены методы, которые оказываются более выгодными при определении не всех, а некоторых собственных значений и собственных векторов.

§ 1. Проблема и простейшие методы

1. Элементы теории. Напомним некоторые сведения из курса линейной алгебры. Если A — квадратная матрица n-го порядка и $Ax = \lambda x$ при $x \neq 0$, то число λ называется *собственным значением* матрицы, а ненулевой вектор x — соответствующим ему *собственным вектором*. Перепишем задачу в виде

$$(A - \lambda E)\, x = 0, \qquad x \neq 0. \tag{1}$$

Для существования нетривиального решения задачи (1) должно выполняться условие

$$\det (A - \lambda E) = 0. \tag{2}$$

Этот определитель является многочленом n-й степени от λ; его называют *характеристическим многочленом*. Значит, существует n собственных значений — корней этого многочлена, среди которых могут быть одинаковые (кратные). В принципе можно вычислить характеристический многочлен и найти все его корни.

Если найдено некоторое собственное значение, то, подставляя его в однородную систему (1), можно определить соответствующий собственный вектор. Будем нормировать собственные векторы *). Тогда каждому простому (не кратному) собственному

*) Нормировкой (на единицу) вектора x называют умножение его на $\| x \|^{-1}$; нормированный вектор имеет единичную длину.

значению соответствует один (с точностью до направления) собственный вектор, а совокупность всех собственных векторов, соответствующих совокупности простых собственных значений, линейно-независима. Таким образом, если все собственные значения матрицы простые, то она имеет n линейно-независимых собственных векторов, которые образуют базис пространства.

Кратному собственному значению кратности p может соответствовать от 1 до p линейно-независимых собственных векторов. Например, рассмотрим такие матрицы четвертого порядка:

$$A=\begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}, \quad C_4=\begin{bmatrix} a & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{bmatrix}, \quad B=\begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{bmatrix}. \tag{3}$$

У каждой из них характеристическое уравнение принимает вид $\det(A-\lambda E)=(a-\lambda)^4=0$, а следовательно, собственное значение $\lambda=a$ и имеет кратность $p=4$. Однако у первой матрицы есть четыре линейно-независимых собственных вектора

$$\boldsymbol{e}_1=\begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad \boldsymbol{e}_2=\begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad \boldsymbol{e}_3=\begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \quad \boldsymbol{e}_4=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}; \tag{4}$$

это легко проверить, поочередно подставляя векторы (4) в равенство (1). У второй же матрицы имеется только один собственный вектор $\boldsymbol{e}_1$. В самом деле, пусть ее собственный вектор $\boldsymbol{x}$ имеет компоненты x_i; тогда уравнение (1) примет для нее вид

$$(G_4-\lambda E)\,\boldsymbol{x}=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}\cdot\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}x_2\\x_3\\x_4\\0\end{pmatrix}=0, \qquad \lambda=a,$$

откуда $x_2=x_3=x_4=0$, а $x_1=1$ в силу условия нормировки. Вторую матрицу называют простой *жордановой* (или классической) *подматрицей*. Третья матрица имеет так называемую каноническую жорданову форму (по диагонали стоят либо числа, либо жордановы подматрицы, а остальные элементы равны нулю). Ее собственными векторами являются $\boldsymbol{e}_1$ и $\boldsymbol{e}_2$; в этом легко убедиться при помощи выкладки, аналогичной только что проделанной.

Таким образом, если среди собственных значений матрицы есть кратные, то ее собственные векторы не всегда образуют базис. Однако и в этом случае собственные векторы, соответствующие различным собственным значениям, являются линейно-независимыми.

Задача на собственные значения легко решается для некоторых простых форм матрицы: диагональной, трехдиагональной, треугольной или почти треугольной. Например, определитель треугольной (в частности, диагональной) матрицы равен произведению диагональных элементов. В этом случае $A-\lambda E$ тоже треугольная или диагональная матрица. Поэтому *собственные значения треугольной (диагональной) матрицы равны диагональным элементам*. Легко проверить, что диагональная матрица имеет n собственных ортонормированных векторов $e_i=\{0, \ldots, 0, 1, 0, \ldots, 0\}^T$, соответствующих собственным значениям $\lambda_i=a_{ii}$; наоборот, матрица с такими собственными векторами диагональна.

Многие численные методы решения задач на собственные значения основаны на приведении матрицы к одной из перечисленных выше простых форм при помощи преобразования подобия. Матрица $G=F^{-1}AF$ называется *подобной* матрице A. Пусть η, y суть собственное значение и собственный вектор матрицы G; тогда $\eta y=Gy=F^{-1}AFy$, что после умножения слева на матрицу F дает $\eta(Fy)=A(Fy)$. Отсюда видно, что η и Fy суть собственное значение и собственный вектор матрицы A. Следовательно, *преобразование подобия не меняет собственных значений матрицы* и по определенному закону преобразует ее собственные векторы.

Особенно удобны преобразования подобия при помощи унитарных матриц *). Если ортонормированный базис преобразовать унитарной матрицей, то он останется ортонормированным. Если подобно преобразовать эрмитову матрицу при помощи унитарной, то она остается эрмитовой; в самом деле,

$$B=U^HAU, \quad B^H=(U^HAU)^H=U^HA^H(U^H)^H=U^HAU=B.$$

Если A — нормальная матрица, то при подобном унитарном преобразовании она остается нормальной; читателям предлагается проверить это.

Известно, что для любой матрицы A есть такое унитарное преобразование, что U^HAU является верхней треугольной матри-

*) Напомним принятую терминологию. Матрица B называется эрмитово сопряженной к матрице A, если она получена из нее транспонированием (зеркальным отражением от главной диагонали) с последующей заменой элементов комплексно-сопряженными, т. е. $B=A^H$, если $b_{ik}=\overline{a_{ki}}$. Матрица эрмитова, если она эрмитово сопряжена самой себе: $A=A^H$, и косоэрмитова, если она удовлетворяет соотношению $A=-A^H$. Вещественная эрмитова матрица называется симметричной, а косоэрмитова — кососимметричной. Унитарной называется матрица, обратная своей эрмитово сопряженной: $U^H=U^{-1}$; вещественные унитарные матрицы называют ортогональными. Матрица называется нормальной, если она перестановочна со своей эрмитово сопряженной, т. е. $AA^H=A^HA$. Легко видеть, что эрмитовые, косоэрмитовые и унитарные матрицы являются частными случаями нормальных.

цей; если A — нормальная матрица, то это унитарное преобразование приводит ее к диагональной форме.

Непосредственно для практических вычислений теорема Шура ничего не дает, ибо неизвестен способ нахождения такого унитарного преобразования. Но одно косвенное следствие является важным. После указанного преобразования нормальная матрица A становится диагональной; тогда ее новые собственные векторы образуют ортонормированный базис e_i. Следовательно, собственные векторы исходной нормальной матрицы получаются из ортонормированного базиса e_i унитарным преобразованием и сами образуют ортонормированный базис.

Это существенно, ибо в практике вычислений часто встречаются нормальные матрицы, особенно их такие частные случаи, как эрмитовы, косоэрмитовы и унитарные матрицы. Ортогональные же преобразования обеспечивают наибольшую устойчивость алгоритма по отношению к ошибкам округления. Действия с неортогональными базисами и преобразованиями при больших порядках матрицы нередко приводят к «разболтке» счета (это уже отмечалось в главе II в связи с вопросами аппроксимации).

Не всякую матрицу с кратными собственными значениями можно подобно преобразовать к диагональной форме, но ее заведомо можно преобразовать к канонической жордановой форме. Если же матрица имеет только простые собственные значения, то существует преобразование подобия (не обязательно унитарное), приводящее ее к диагональной. В самом деле, такая матрица имеет n линейно-независимых собственных векторов. Матрица F, столбцами которой являются координаты этих векторов, преобразует базис e_i в базис из собственных векторов. Значит, преобразование подобия с матрицей F приводит A к диагональной форме.

2. Устойчивость. Для исследования устойчивости проблемы собственных значений надо наряду с матрицей A рассмотреть эрмитово сопряженную к ней матрицу A^H. Поскольку при транспонировании матрицы ее определитель не меняется, а замена всех матричных элементов комплексно сопряженными величинами приводит к замене определителя тоже комплексно сопряженным числом, то $\det(A^H-\lambda^* E)=[\det(A-\lambda E)]^*$. Отсюда видно, что если λ_i есть собственное значение матрицы A, то $\det(A^H-\lambda_i^* E)=0$, то есть λ_i^* есть собственное значение матрицы A^H. Следовательно, *собственные значения эрмитово сопряженных матриц комплексно-сопряжены друг другу.*

Обозначим собственные векторы матриц A и A^H соответственно через x_i и y_i. Докажем, что *собственные векторы сопряженных матриц, соответствующие различным* (точнее, не комплексно-сопряженным друг другу) *собственным значениям, взаимно*

ортогональны. Для этого напишем тождества

$$A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i, \quad A^H \boldsymbol{y}_j = \lambda_j^* \boldsymbol{y}_j.$$

Скалярно умножим *) первое равенство слева на $\boldsymbol{y}_j$, а второе — справа на $\boldsymbol{x}_i$ и вычтем одно из другого; получим

$$(\boldsymbol{y}_j,\ A\boldsymbol{x}_i) - (A^H \boldsymbol{y}_j,\ \boldsymbol{x}_i) = (\boldsymbol{y}_j,\ \lambda_i \boldsymbol{x}_i) - (\lambda_j^* \boldsymbol{y}_j,\ \boldsymbol{x}_i).$$

Выражение в левой части этого равенства равно нулю. Вынося λ из скалярных произведений правой части этого равенства, получим $(\lambda_i - \lambda_j)(\boldsymbol{y}_j,\ \boldsymbol{x}_i) = 0$ или

$$(\boldsymbol{y}_j,\ \boldsymbol{x}_i) = 0 \quad \text{при} \quad \lambda_i \neq \lambda_j, \tag{5}$$

что и требовалось доказать.

Отсюда следует, что у эрмитовых матриц собственные значения вещественны, а собственные векторы образуют ортогональную систему (поскольку $\boldsymbol{y}_j = \boldsymbol{x}_j$).

Рассмотрим устойчивость проблемы собственных значений. Для простоты ограничимся случаем, когда собственные векторы матрицы образуют базис, а данное собственное значение — простое.

Если немного изменить матричные элементы, то поправки к собственному значению и соответствующему вектору с точностью до величин второго порядка малости удовлетворяют линеаризованному уравнению

$$A\,\delta\boldsymbol{x}_i + \delta A \cdot \boldsymbol{x}_i \approx \delta\lambda_i \cdot \boldsymbol{x}_i + \lambda_i\,\delta\boldsymbol{x}_i. \tag{6}$$

Разложим поправку $\delta\boldsymbol{x}_i$ по невозмущенным собственным векторам. Вектор $\boldsymbol{x}_i + \delta\boldsymbol{x}_i$ определен с точностью до множителя; подберем этот множитель так, чтобы диагональный коэффициент разложения обратился в нуль:

$$\delta\boldsymbol{x}_i = \sum_{j=1}^{n} \xi_{ij}\boldsymbol{x}_j, \qquad \xi_{ii} = 0.$$

Подставляя это разложение в (6) и умножая слева на различные собственные векторы сопряженной матрицы, получим

$$(\boldsymbol{y}_i,\ \boldsymbol{x}_i)\,\delta\lambda_i \approx (\boldsymbol{y}_i,\ \delta A \cdot \boldsymbol{x}_i), \quad \xi_{ij}(\lambda_j - \lambda_i)(\boldsymbol{y}_j,\ \boldsymbol{x}_j) \approx (\boldsymbol{y}_j,\ \delta A \cdot \boldsymbol{x}_i).$$

Поскольку вариация матрицы может быть любой, то максимумы правых частей обоих последних равенств равны $\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| \times \times \max|\delta a_{kl}|$. Тогда максимально возможные ошибки собственного значения и компонент собственного вектора не превышают (с точ-

*) Напомним, что для комплексных векторов скалярное произведение равно $(\boldsymbol{a},\ \boldsymbol{b}) = \sum_{k=1}^{n} a_k^* b_k$.

ностью до отброшенных в ходе выкладок бесконечно малых более высокого порядка) следующих значений:

$$|\delta\lambda_i| \lesssim \varkappa_i \max_{k,\,l} |\delta a_{kl}|, \quad |\xi_{ij}| \lesssim \frac{\varkappa_j}{|\lambda_i - \lambda_j|} \max_{k,\,l} |\delta a_{kl}|. \tag{7}$$

Здесь через $\varkappa_i$ обозначен так называемый i-й *коэффициент перекоса* матрицы

$$\varkappa_i = \frac{\sqrt{(x_i,\ x_i)\,(y_i,\ y_i)}}{(x_i,\ y_i)} = \frac{1}{\cos\varphi_i}, \tag{8}$$

где φ_i есть угол между соответствующими векторами данной матрицы и эрмитово сопряженной к ней.

Заметим, что для эрмитовой матрицы все коэффициенты перекоса равны единице, поскольку соответствующие векторы ортогональны. А для типичной жордановой клетки

$$A = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}, \quad x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad A^{\text{н}} = \begin{bmatrix} a^* & 0 \\ 1 & a^* \end{bmatrix}, \quad y_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

выполняется условие $(x_1,\ y_1) = 0$, т. е. коэффициент перекоса обращается в бесконечность. Очевидно, что для любых матриц $|\varkappa_i| \geqslant 1$.

Выводы из оценки (7) можно сформулировать следующим образом. Собственное значение устойчиво относительно вариации матричных элементов, если соответствующий ему коэффициент перекоса мал; если этот коэффициент перекоса очень велик, то устойчивость может быть плохой. Собственный вектор устойчив по матричным элементам, если *все* коэффициенты перекоса матрицы невелики, а данное собственное значение — простое.

Значит, все собственные значения эрмитовых матриц мало чувствительны к погрешностям матричных элементов. А собственные значения жордановых подматриц могут быть очень чувствительны к погрешностям. Проиллюстрируем последнее на примере неэрмитовой матрицы 20-го порядка:

$$A = \begin{bmatrix} 20 & 20 & 0 & 0 & \dots & 0 & 0 \\ 0 & 19 & 20 & 0 & \dots & 0 & 0 \\ 0 & 0 & 18 & 20 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 2 & 20 \\ \varepsilon & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \tag{9}$$

где через ε обозначено малое возмущение нулевого углового элемента. Характеристическое уравнение этой матрицы имеет вид

$$\det(A - \lambda E) = \prod_{k=1}^{20} (k - \lambda) - 20^{19}\varepsilon = 0. \tag{10}$$

При $\varepsilon = 0$ младший коэффициент характеристического многочлена есть $a_0 = 20! \approx 2{,}5 \times 10^{18}$, а наименьшее по модулю собственное значение $\lambda_{20} = 1$. Если же положить $\varepsilon = 20^{-19} \times 20! \approx 5 \times 10^{-7}$, то коэффициент a_0 обращается в нуль, а тогда $\lambda_{20} = 0$. Таким образом, и коэффициенты и сами корни характеристического многочлена могут быть очень чувствительны к малым погрешностям матричных элементов, что означает слабую устойчивость задачи. Это согласуется со сделанным в § 2 главы V замечанием о том, что корни многочлена высокой степени нередко чувствительны к погрешностям коэффициентов.

Но для эрмитовых матриц собственные значения хорошо устойчивы по матричным элементам. Даже для неэрмитовых матриц опасна вариация не любого коэффициента; например, к возмущениям элементов главной диагонали собственные значения мало чувствительны.

3. Метод интерполяции. Если мы найдем характеристический многочлен, то все его корни нетрудно вычислить, например, методом парабол. В методе парабол для нахождения одного корня обычно требуется около 10 раз вычислить многочлен. Поэтому важно найти способ быстрого вычисления характеристического многочлена.

Те методы решения проблемы собственных значений, которые позволяют определить характеристический многочлен за конечное число действий, называются прямыми. Методы, в которых характеристический многочлен определяется как предел некоторого итерационного процесса, называются итерационными. Это разделение носит несколько условный характер, ибо даже если характеристический многочлен найден за конечное число действий, то его корни приходится определять итерационным процессом. Однако оно имеет практический смысл, поскольку нахождение характеристического многочлена высокой степени гораздо более трудоемко, чем отыскание его корней.

Простейшим прямым методом является *метод интерполяции* (предложенный, по-видимому, Ш. Е. Микеладзе в 1948 г.) Известно, что многочлен n-й степени однозначно определяется своими значениями в $n+1$ узле. Произвольно выберем $n+1$ значение $\lambda^{(k)}$ в качестве таких узлов. Вычислим в них значение $f(\lambda^{(k)}) = \det(A - \lambda^{(k)} E)$ и построим по этим значениям интерполяционный многочлен Ньютона при помощи формул (2.6) и (2.8). В силу единственности этот многочлен будет характеристическим. Он при этом получается в форме многочлена с заданными коэффициентами, так что дальнейшие вычисления для нахождения его корней потребуют малого числа действий.

Описанный алгоритм несложен и легко программируется на ЭВМ. В нем следует использовать стандартную программу вычисления определителя методом исключения (глава V, § 1, п. 3). При этом характеристический многочлен определяется примерно за $^2/_3 n^4$ арифметических действий, из которых половину составляют сложения и половину — умножения. Видно, что для мат-

риц невысокого порядка $n \leqslant 10$ нахождение характеристического многочлена методом интерполяции требует не более 0,5 сек на ЭВМ БЭСМ-4, что вполне приемлемо.

Если известны границы, в которых расположены собственные значения, то целесообразно размещать узлы интерполяции $\lambda^{(k)}$ в этих границах, причем приблизительно равномерно. Это уменьшает ошибки округления, возникающие при нахождении разделенных разностей в формуле Ньютона, т. е. улучшает устойчивость алгоритма. Для определения границ спектра можно воспользоваться оценкой $|\lambda_i| \leqslant \|A\|$, справедливой для любой нормы матрицы (это следует из того, что спектральная норма $\|A\|_\lambda = \max|\lambda_i|$ есть наименьшая из норм матрицы). Хотя эта оценка в среднем завышена, но она достаточно разумна для тех матриц, с которыми приходится встречаться на практике, и тех норм (см. § 2 главы I), которые просто вычисляются.

Метод интерполяции прост и применим для матриц произвольной структуры, а также для более сложных проблем. Например, общую задачу $\det[p_{ik}(\lambda)] = 0$, где каждый элемент матрицы есть некоторый многочлен от λ, решают практически только этим методом; разумеется, число узлов по λ выбирают в соответствии со степенью результирующего многочлена. Лишь в частном случае этой задачи — так называемой обобщенной проблемы собственных значений $\det(A - \lambda B) = 0$, разработаны более экономичные методы.

Однако, чем выше порядок матрицы, тем менее выгоден метод интерполяции. Во-первых, число выполняемых арифметических действий возрастает с ростом порядка очень быстро — как n^4. Во-вторых, при составлении интерполяционного многочлена Ньютона вычисляются разделенные разности, что при высоких порядках приводит к большой потере точности. Поэтому при $n \gtrsim 10$ (а в случае кратных или близких собственных значений и при меньших n) метод интерполяции дает плохие результаты. Для матриц высокого порядка применяют более сложные, но зато более устойчивые и экономичные методы, изложенные в следующих параграфах.

Существуют прямые методы Леверье, А. Н. Крылова, А. М. Данилевского, Самуэльсона и Ланцоша, позволяющие вычислить все коэффициенты характеристического многочлена произвольной матрицы примерно за n^3 арифметических действий. Они экономичней метода интерполяции.

Однако в § 2 главы V отмечалось, что корни многочлена высокой степени могут быть очень чувствительны к погрешностям коэффициентов. Кроме того, и коэффициенты и сами корни характеристического многочлена нередко слабо устойчивы по матричным элементам, как было показано в п. 2. Поэтому указанные выше экономичные методы оказались достаточно устойчивыми только для матриц невысокого порядка $n \lesssim 10$, а при наличии кратных или близких собственных значений — для еще меньшего порядка. Но при таких порядках матрицы экономия по сравнению с методом интерполяции невелика и не оправдывает применения этих довольно сложных и капризных методов.

4. Трехдиагольные матрицы. В интерполяционном методе мы находили явное выражение для характеристического многочлена только затем, чтобы иметь способ экономного вычисления этого многочлена при заданных λ. Однако для трехдиагональных матриц (даже очень высокого порядка) есть способ быстрого вычисления $\det(A-\lambda E)$ без нахождения явного выражения характеристического многочлена. Это существенно, ибо матрицы общего вида можно привести к трехдиагольной форме преобразованием подобия.

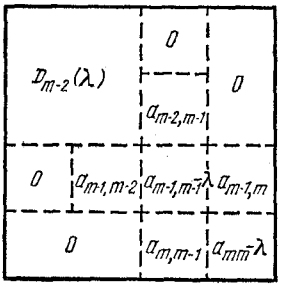


Рис. 30.

Рассмотрим этот способ. Обозначим главный минор m-го порядка матрицы $A-\lambda E$ через $D_m(\lambda)$. Разложим такой минор по элементам последней строки; в ней всего два ненулевых элемента (рис. 30), так что получим

$$D_m(\lambda) = (a_{mm}-\lambda)D_{m-1}(\lambda) - a_{m,m-1}B_{m,m-1}(\lambda), \quad (11)$$

где через $B_{m,m-1}(\lambda)$ обозначен минор, дополняющий элемент $a_{m,m-1}$. Этот минор содержит в последнем столбце только один ненулевой элемент $a_{m-1,m}$, поэтому его целесообразно разложить по элементам последнего столбца:

$$B_{m,m-1}(\lambda) = a_{m-1,m}D_{m-2}(\lambda). \quad (12)$$

Подставляя (12) в (11), найдем рекуррентное соотношение, выражающее минор высшего порядка через низшие:

$$D_m(\lambda) = (a_{mm}-\lambda)D_{m-1}(\lambda) - a_{m,m-1}a_{m-1,m}D_{m-2}(\lambda). \quad (13)$$

Для начала расчета по рекуррентной формуле надо задать два первых минора. Удобно формально положить

$$D_{-1}(\lambda) = 0, \quad D_0(\lambda) = 1. \quad (14)$$

Подставляя эти значения в (13) и вычисляя $D_1(\lambda)$ и $D_2(\lambda)$, легко убедиться, что результат получается правильным. Следовательно, такой способ начала счета приемлем.

Однократный расчет величины определителя по формулам (11)—(14) требует всего $5n$ арифметических действий, причем среди них нет делений, и вычисления очень быстрые и устойчивые. Таким образом, имеется быстрый способ нахождения характеристического многочлена при заданном значении λ.

Имеется способ нахождения характеристического многочлена, более экономичный при многократных вычислениях. Преобразуя рекуррентное соотношение (13), можно получить коэффициенты характеристического многочлена в форме Горнера. Однократное же вычисление многочлена по схеме Горнера требует всего $2n$ действий. Однако устойчивость этого процесса для высоких порядков матрицы, по-видимому, хуже, а формулы расчета более сложны.

Еще один несложный способ вычисления характеристического многочлена заключается в следующем. Вычтем заданное значение λ из диагональных элементов a_{ii}, а затем найдем определитель получившейся трехдиагональной матрицы по формулам прямого хода прогонки (5.12)—(5.13). Однако этот способ менее экономичен и устойчив, чем расчет по формуле (13).

Корни многочлена $D_n(\lambda)$ удобнее всего находить методом парабол (см. § 2 главы V). Этот метод для многочленов не слишком высокой степени ($n \lesssim 50$) достаточно устойчив и позволяет найти все корни с 5—7 верными знаками, даже если среди корней есть кратные. В библиотеках многих ЭВМ имеются стандартные программы вычислений всех корней многочлена методом парабол.

Иногда для нахождения всех корней характеристического многочлена употребляют метод Ньютона, но детали такого алгоритма хуже отработаны. В основном метод Ньютона применяют к частичной проблеме собственных значений.

Заметим, что при любом способе вычислений для нахождения всех корней требуется удалять уже полученные корни, т. е. переходить к вспомогательной функции $G(\lambda) = D_n(\lambda) / \prod_{i=1}^{k} (\lambda - \lambda_i)$. Поскольку явного выражения характеристического многочлена мы не выписываем, то для вычисления $G(\lambda)$ надо находить отдельно числитель и знаменатель при требуемых значениях λ. Это немного увеличивает объем расчетов.

Описанным способом все собственные значения трехдиагональной матрицы находятся довольно легко, причем для этого требуется всего около $50n^2$ арифметических действий *), т. е. способ экономичен. Поэтому для трехдиагональных матриц этот способ является основным.

5. Почти треугольные матрицы. Для такой матрицы также можно написать формулы, позволяющие легко вычислить определитель при заданном значении λ. Это удобно делать методом исключения Гаусса, учитывая большое количество нулей в матрице определителя.

Используем формулы метода исключения (5.3)—(5.5). Для определенности будем считать нашу матрицу верхней почти треугольной. Тогда видно, что $c_{mk} = 0$ при $m > k+1$, а каждый цикл исключения сводится всего лишь к вычитанию двух строк. Достаточно при этом помечать изменяющиеся величины штрихом, опуская верхний индекс цикла. После этого формулы k-го цикла примут вид

$$c_k = \frac{a_{k+1,\,k}}{a'_{kk}}, \quad a'_{k+1,\,i} = a_{k+1,\,i} - c_k a'_{ki}, \quad k+1 \leqslant i \leqslant n, \tag{15}$$

причем $a'_{k+1,\,k} = 0$. Последовательно полагая $k = 1, 2, \ldots, n-1$, аннулируем все поддиагональные элементы. После этого определитель легко вычисляется

*) Метод парабол обычно сходится менее чем за 10 итераций, а одна итерация требует $5n$ действий. Эти цифры мы будем использовать при описании других методов.

по формуле Чио (5.8):

$$\det A = \prod_{k=1}^{n} a'_{kk}, \quad a'_{11} = a_{11}. \tag{16}$$

Поскольку нас интересует $\det(A - \lambda E)$, то для его вычисления надо в формулах (15)—(16) вместо нетрихованных величин a_{kk} всюду подставить $a_{kk} - \lambda$. Этот способ позволяет вычислить определитель за n^2 арифметических действий.

Как и для трехдиагональной матрицы, корни характеристического многочлена можно находить методом парабол. Тогда нахождение всех корней потребует около $10n^3$ действий. Видно, что метод оказывается не быстрым, но довольно простым и устойчивым. Дальше мы увидим, что есть заметно более быстрые способы нахождения собственных значений почти треугольной матрицы, основанные на преобразовании матрицы к трехдиагональной форме. Но они более сложны и менее устойчивы.

6. Обратные итерации. Если собственное значение известно, то собственный вектор удовлетворяет системе (1). Но любой численный метод дает вместо точного собственного значения λ_i приближенное значение $\tilde{\lambda}_i$, так что $\det(A - \tilde{\lambda}_i E) \neq 0$, хотя отличается от нуля очень мало. В таком случае задача $(A - \tilde{\lambda}_i E)\boldsymbol{x} = \boldsymbol{0}$ при использовании приближенного собственного значения имеет только тривиальное решение $\boldsymbol{x} = \boldsymbol{0}$. Поэтому в численных расчетах находить собственные векторы непосредственно из системы (1) нельзя.

Для нахождения собственных векторов удобен метод *обратной итерации*, заключающийся в следующем. Выберем наудачу вектор $\boldsymbol{b}$ и рассмотрим линейную неоднородную систему

$$(A - \tilde{\lambda}_i E)\boldsymbol{x} = \boldsymbol{b}. \tag{17}$$

Определитель этой системы отличен от нуля, так что она имеет единственное решение. Покажем, что найденный из нее вектор $\boldsymbol{x}$ окажется почти равным собственному вектору $\boldsymbol{x}_i$, соответствующему данному собственному значению λ_i.

Для простоты ограничимся случаем, когда матрица n-го порядка имеет n линейно-независимых собственных векторов $\boldsymbol{x}_j$ — например, матрица нормальная (для случая произвольных матриц ниже приведен численный пример). Тогда собственные векторы образуют базис, по которому можно разложить векторы $\boldsymbol{x}$ и $\boldsymbol{b}$:

$$\boldsymbol{x} = \sum_{j=1}^{n} \xi_j \boldsymbol{x}_j, \quad \boldsymbol{b} = \sum_{j=1}^{n} \beta_j \boldsymbol{x}_j. \tag{18}$$

Подставляя это разложение в систему (17), перенося все члены влево и учитывая, что $A\boldsymbol{x}_j = \lambda_j \boldsymbol{x}_j$, получим

$$\sum_{j=1}^{n} [\xi_j (\lambda_j - \tilde{\lambda}_i) - \beta_j] \boldsymbol{x}_j = \boldsymbol{0}. \tag{19}$$

Поскольку собственные векторы линейно-независимы, то их линейная комбинация обращается в нуль только в том случае, когда

все ее коэффициенты равны нулю. Поэтому из (19) следует

$$\xi_j = \frac{\beta_j}{\lambda_j - \tilde{\lambda}_i}. \tag{20}$$

Видно, что если $\lambda_j \approx \tilde{\lambda}_i$, то коэффициент ξ_j будет очень большим; в противном случае он невелик. Рассмотрим следствия из этого в трех основных случаях.

Первый случай — собственное значение λ_i простое. Тогда из всех коэффициентов ξ_j, $1 \leqslant j \leqslant n$, только один коэффициент ξ_i оказывается очень большим. Это означает, что найденный вектор $\boldsymbol{x}$ почти совпадает с собственным вектором $\boldsymbol{x}_i$ (с точностью до нормировочного множителя), что и требовалось доказать. Заметим, что поскольку найденный вектор $\boldsymbol{x}$ оказывается очень большим, то его обычно нормируют.

Из (20) видно, что при обратной итерации (т. е. при переходе от $\boldsymbol{b}$ к $\boldsymbol{x}$) компонента β_i усиливается по сравнению с другими компонентами β_j примерно во столько раз, во сколько погрешность данного собственного значения меньше разности соседних собственных значений. Поэтому чем точнее найдено $\tilde{\lambda}_i$ (очевидно, хорошая точность особенно важна при наличии близких собственных значений), тем ближе $\boldsymbol{x}$ будет к $\boldsymbol{x}_i$. Если собственные значения найдены слишком грубо, или случайно вектор $\boldsymbol{b}$ выбран неудачно, так что β_i очень мало, то разница между $\boldsymbol{x}$ и $\boldsymbol{x}_i$ может оказаться заметной. Тогда подставляют найденный вектор $\boldsymbol{x}$ в правую часть уравнения (17) вместо $\boldsymbol{b}$ и организуют итерационный процесс

$$(A - \tilde{\lambda}_i E)\,\boldsymbol{x}^{(s)} = \boldsymbol{x}^{(s-1)}, \quad \boldsymbol{x}^{(0)} = \boldsymbol{b}. \tag{21}$$

Обычно он сходится настолько быстро, что двух итераций вполне достаточно. Напомним, что на каждой итерации обязательно надо нормировать найденные $\boldsymbol{x}^{(s)}$, чтобы не получать в расчетах слишком больших чисел, вызывающих переполнение на ЭВМ.

Замечание 1. Очень эффективен один простой способ выбора $\boldsymbol{b}$. В качестве его компонент в декартовых координатах возьмем последовательные многоразрядные псевдослучайные числа γ_k (см. § 4 главы IV). Тогда вероятность того, что β_i окажется очень малым, будет ничтожна.

Второй случай — собственное значение λ_i кратно; например, $\lambda_1 = \lambda_2 = \ldots = \lambda_p$, $1 < p \leqslant n$. Напомним, что в этом случае собственные векторы $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ определены неоднозначно; любая их линейная комбинация удовлетворяет уравнению

$$A\left(\sum_{j=1}^{p} \alpha_j \boldsymbol{x}_j\right) = \sum_{j=1}^{p} \alpha_j A\boldsymbol{x}_j = \lambda_1 \sum_{j=1}^{p} \alpha_j \boldsymbol{x}_j$$

и является собственным вектором. Т. е. они порождают p-мерное

подпространство, любой базис которого можно взять в качестве системы искомых собственных векторов.

Теперь из (20) следует, что большими оказываются коэффициенты $\xi_1, \xi_2, \ldots, \xi_p$, причем степень их усиления одинакова; остальные коэффициенты остаются малыми. Значит, найденный из (17) вектор $\boldsymbol{x}$ будет приближенно линейной комбинацией $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, а тем самым — искомым собственным вектором. Если точность полученного приближения недостаточна, то обратную итерацию повторяют снова по формуле (21).

Чтобы найти все собственные векторы для кратного собственного значения, возьмем столько линейно-независимых векторов $\boldsymbol{b}^{(k)}$, какова кратность корня. Обратными итерациями получим столько же векторов $\boldsymbol{x}^{(k)}$, которые и будут искомыми; они будут линейно-независимыми, поскольку преобразование (17) невырожденное. Остается только ортогонализовать найденные векторы, если это требуется по условиям задачи.

Напомним, что в качестве декартовых координат векторов $\boldsymbol{b}^{(k)}$ целесообразно брать псевдослучайные числа; здесь это имеет то дополнительное преимущество, что векторы автоматически получаются линейно-независимыми.

Третий случай — когда матрица имеет кратные корни, но число ее собственных векторов меньше n — выходит за рамки нашего доказательства. Однако метод обратных итераций здесь также применим в той форме, которая описана для кратных корней. Разница лишь в том, что если p-кратному собственному значению соответствуют всего q собственных векторов ($q < p$), то из полученных обратной итерацией векторов $\boldsymbol{x}^{(k)}$ только q будут линейно-независимыми. Это выясняется при их ортогонализации: первые q векторов ортогонализуются «без приключений», а при ортогонализации следующих векторов их компоненты обращаются почти в нуль (в пределах погрешности расчета).

Каков объем расчетов в методе обратной итерации? Нахождение собственного вектора требует (при одной итерации) не более $^2/_3 n^3$ действий, так что для нахождения всех их надо около n^4 арифметических действий. Таким образом, при больших порядках матрицы метод неэкономичен, но при $n \leqslant 10$ вполне удовлетворителен. Особенно употребителен этот метод из-за своей простоты, универсальности и хорошей устойчивости алгоритма.

В некоторых частных случаях расчеты существенно упрощаются и ускоряются. Наиболее важен случай трехдиагональной матрицы. При этом линейная система уравнений (17) для определения компонент собственных векторов также будет трехдиагональной, и ее решают экономичным методом прогонки по несложным формулам (5.10) — (5.12). Для вычисления одного собственного вектора в этом случае требуется $10n$, а для всех — $10n^2$ арифметических действий.

Для почти треугольной матрицы в методе обратных итераций требуется решать линейную систему с почти треугольной матрицей, что делается специальным вариантом метода исключения. Если учесть, что случайный вектор в правой части (17) можно задавать уже после приведения матрицы в методе исключения Гаусса к треугольной форме, то нахождение каждого собственного вектора требует $^3/_2n^2$ действий (тот же прием для трехдиагональной матрицы позволяет сократить число действий до $7n$). А для нахождения всех собственных векторов требуется соответственно $^3/_2n^3$ арифметических действий.

Отметим одну существенную деталь. Поскольку $\det(A-\tilde{\lambda}_iE)\approx\approx 0$, то при нахождении собственных векторов в формулах прямого хода метода исключений (прогонки) на главной диагонали появится хотя бы один очень малый элемент. Чтобы формально можно было вести расчет, диагональные элементы не должны обращаться в нуль; для этого надо, чтобы погрешность собственного значения была не слишком мала, т. е. составляла бы 10—15 последних двоичных разрядов числа на ЭВМ. Если корни характеристического многочлена находят методом парабол (или секущих), то такая погрешность получается естественно, ибо из-за ошибок округления эти методы перестают сходиться в очень малой окрестности корня. Но если корни определялись методом Ньютона, то при этом могли быть найдены верно все знаки собственного значения; тогда, чтобы избежать деления на нуль, приходится специально вносить в λ_i небольшие погрешности.

Пример. Возьмем жорданову подматрицу C_4 четвертого порядка (3) и приближенное собственное значение $\tilde{\lambda}=a-\varepsilon$. В качестве $\boldsymbol{b}$ выберем вектор с единичными декартовыми координатами. Тогда уравнение (17) примет вид

$$\begin{bmatrix}\varepsilon & 1 & 0 & 0\\ 0 & \varepsilon & 1 & 0\\ 0 & 0 & \varepsilon & 1\\ 0 & 0 & 0 & \varepsilon\end{bmatrix}\cdot\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}\equiv\begin{pmatrix}\varepsilon x_1+x_2\\ \varepsilon x_2+x_3\\ \varepsilon x_3+x_4\\ \varepsilon x_4\end{pmatrix}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}. \tag{22а}$$

Последовательно находим компоненты вектора $\boldsymbol{x}$:

$$\begin{aligned}&x_1=\varepsilon^{-1},\quad x_3=-\varepsilon^{-2}+\varepsilon^{-1},\\ &x_2=\varepsilon^{-3}-\varepsilon^{-2}+\varepsilon^{-1},\\ &x_1=-\varepsilon^{-4}+\varepsilon^{-3}-\varepsilon^{-2}+\varepsilon^{-1}.\end{aligned} \tag{22б}$$

Затем нормируем вектор, умножив все компоненты на $-\varepsilon^{-4}$:

$$x_1=1+O(\varepsilon),\quad x_2=O(\varepsilon),\quad x_3=O(\varepsilon^2),\quad x_4=O(\varepsilon^3). \tag{22в}$$

Полученный вектор $\boldsymbol{x}=\boldsymbol{e}_1+O(\varepsilon)$ приближенно равен собственному вектору жордановой матрицы (см. п. 1), что нам и требовалось. Попутно заметим, что в промежуточных выкладках (22б) возникали высокие обратные степени погрешности ε, чего не бывает у матриц с n собственными векторами. Это показывает, что случай матриц, содержащих жордановы подматрицы высокого порядка, труден для численных расчетов на ЭВМ, ибо в них легко возникают переполнения.

§ 2. Эрмитовы матрицы

1. Метод отражения. Существуют экономичные и устойчивые методы нахождения всех собственных значений матриц высокого порядка*). Они основаны на приведении матрицы преобразованием подобия к трехдиагональной или другим простым формам, для которых проблема собственных значений решается легко.

Сейчас мы рассмотрим *метод отражений*, который позволяет подобно преобразовать произвольную матрицу к почти треугольной форме за $(10/3)n^3$ арифметических действий, а эрмитову матрицу к трехдиагональной форме — всего за $^4/_3 n^3$ действий. Поскольку для трехдиагональной матрицы все собственные значения находятся очень экономично, то для эрмитовых матриц метод отражения является самым быстрым из известных методов решения полной проблемы собственных значений. Рассмотрим его.

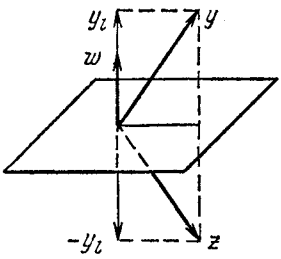


Рис. 31.

Произведем в n-мерном векторном пространстве отражение относительно некоторой гиперплоскости, проходящей через начало координат. Преобразование полностью определяется заданием нормали $\boldsymbol{w}$ к гиперплоскости. Эта нормаль есть нормированный вектор-столбец

$$(\boldsymbol{w},\ \boldsymbol{w}) \equiv \boldsymbol{w}^H \boldsymbol{w} = \sum_{i=1}^{n} w_i^* w_i = 1, \tag{23}$$

где $\boldsymbol{w}^H$ есть вектор-строка, эрмитово сопряженный к столбцу. Возьмем произвольный вектор $\boldsymbol{y}$ и разложим его на две составляющие: параллельно нормали $\boldsymbol{y}_l = \boldsymbol{w}(\boldsymbol{w},\ \boldsymbol{y})$ и перпендикулярно ей. При отражении вектора его перпендикулярная составляющая остается неизменной, а параллельная — меняет знак (рис. 31), поэтому отраженный вектор $\boldsymbol{z}$ отличается от исходного на удвоенную величину параллельной компоненты

$$\boldsymbol{z} = \boldsymbol{y} - 2\boldsymbol{w}(\boldsymbol{w},\ \boldsymbol{y}). \tag{24}$$

Это преобразование вектора можно записать в канонической форме умножения на *матрицу отражения* R:

$$\boldsymbol{z} = R\boldsymbol{y}, \quad R = E - 2\boldsymbol{w}\boldsymbol{w}^H, \tag{25}$$

*) Практически все описанные в §§ 2—3 методы хорошо применимы к матрицам, порядок которых не превышает сотни. Для матриц произвольного типа с $n > 100$ удовлетворительных методов решения общей проблемы собственных значений пока нет.

где умножение столбца w справа на строку той же длины w^H дает, по правилам умножения прямоугольных матриц, квадратную матрицу. Заметим, что равенства (24)—(25) в координатной форме записываются следующим образом:

$$z_i = y_i - 2w_i \sum_{j=1}^{n} w_j^* y_j, \qquad R_{ij} = \delta_{ij} - 2w_i w_j^*. \tag{26}$$

Исследуем свойства матрицы отражения. Эта матрица эрмитова, что непосредственно вытекает из следующей цепочки преобразований:

$$R^H = (E - 2ww^H)^H = E - 2(w^H)^H w^H = E - 2ww^H = R. \tag{27}$$

Возведем матрицу отражения в квадрат:

$$R^2 = (E - 2ww^H)(E - 2ww^H) = E - 4ww^H + 4ww^H\, ww^H.$$

Преобразуем последний член правой части, используя ассоциативность умножения матриц и условие нормировки (23):

$$ww^H\, ww^H = w(w^H w)w^H = ww^H.$$

Тогда последний член сократится с предпоследним, и мы получим

$$RR = E, \text{ или } R = R^{-1}, \tag{28}$$

т. е. *матрица отражения равна своей обратной*. А сравнивая (27) и (28), убедимся, что $R^H = R^{-1}$, так что *матрица отражений унитарна*.

Последнее свойство для нас наиболее важно, поскольку для эрмитовых матриц наиболее выгодны унитарные преобразования подобия. В § 1 было показано, что они сохраняют эрмитовость матрицы. Поэтому если мы унитарным преобразованием подобия приведем матрицу к верхней почти треугольной форме, то в силу эрмитовости она будет трехдиагональной.

Заметим, что произведение любого числа унитарных матриц есть также унитарная матрица. В самом деле, если матрицы $U, V, \ldots, W$ унитарны, то

$$(UV\ldots W)^{-1} = W^{-1}\ldots V^{-1}U^{-1} = W^H\ldots V^H U^H = (UV\ldots W)^H.$$

Поэтому если мы применяем к эрмитовой матрице последовательность унитарных преобразований подобия, то она эквивалентна одному результирующему унитарному преобразованию подобия, и эрмитовость матрицы сохраняется.

Покажем, что для произвольной матрицы A можно подобрать такую конечную последовательность отражений, которая приводит матрицу к верхней почти треугольной форме. Для этого очередное отражение должно уничтожать самый длинный ненулевой столбец в нижней части матрицы A. Действие первых двух отражений показано на рис. 32, где жирными точками обозначены ненулевые элементы матрицы, а кружками — нулевые; третье отражение обращает в нуль обведенные элементы третьего столбца.

Будем считать, что уже уничтожен $q-1$ столбец, и разобьем матрицу A на клетки, как показано на рис. 32. Квадратная клетка A_1 есть верхняя почти треугольная матрица, а в прямоугольной клетке A_3 только последний столбец отличен от нуля.

Сделаем отражение при помощи вектора $\boldsymbol{w}^q = \{w_i^q\}$, у которого первые q компонент нулевые:

$$w_i^q = 0 \text{ при } i \leqslant q$$

(дальше верхний индекс вектора будем обычно опускать). Тогда видно, что если матрицу отражения разбить на клетки того же размера, что у матрицы A, то недиагональные клетки будут нулевыми:

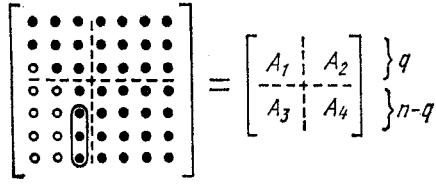


Рис. 32.

$$R = R^{-1} = \left[\begin{array}{c|c} E & 0 \\ \hline 0 & W \end{array}\right], \qquad W_{ij} = \delta_{ij} - 2w_i w_j^*, \quad q+1 \leqslant i;\ j \leqslant n. \tag{29}$$

Из курса линейной алгебры известно, что если матрицы одинаковым образом разбиты на клетки, то они перемножаются по таким же формулам, как если бы эти клетки были обыкновенными элементами. Тогда искомое преобразование подобия принимает следующий вид:

$$B = R^{-1}AR = \left[\begin{array}{c|c} E & 0 \\ \hline 0 & W \end{array}\right] \cdot \left[\begin{array}{c|c} A_1 & A_2 \\ \hline A_3 & A_4 \end{array}\right] \cdot \left[\begin{array}{c|c} E & 0 \\ \hline 0 & W \end{array}\right] = \left[\begin{array}{c|c} A_1 & A_2 W \\ \hline WA_3 & WA_4 W \end{array}\right]. \tag{30}$$

Левая верхняя клетка результирующей матрицы B имеет нужную нам форму — верхнюю почти треугольную. Поскольку у клетки A_3 только последний столбец отличен от нуля, то у клетки $B_3 = WA_3$ элементы последнего столбца равны

$$b_{iq} = a_{iq} - \alpha w_i, \quad q+1 \leqslant i \leqslant n, \tag{31}$$

где

$$\alpha = 2 \sum_{j=q+1}^{n} w_j^* a_{jq}; \tag{32}$$

остальные же элементы этой клетки равны нулю.

Нам надо так подобрать вектор $\boldsymbol{w}$, чтобы обратились в нуль все элементы столбца (31), кроме верхнего. Очевидно, для этого надо положить

$$w_i = a_{iq}/\alpha \text{ при } q+2 \leqslant i \leqslant n \tag{33}$$

и найти, чему равна постоянная α. Заметим, что умножение вектора $\boldsymbol{w}$ на множитель $e^{i\varphi}$ не меняет матрицы отражения; тогда из (32) следует, что α также определена только с точностью до этого множителя. Его всегда можно выбрать так, чтобы α была вещественной положительной, что будем предполагать выполненным.

Из (31) следует, что

$$w_{q+1} = (a_{q+1,\,q} - b_{q+1,\,q})/\alpha. \tag{34}$$

Подставляя (33)—(34) в условие нормировки (23), получим

$$\alpha^2 = \sum_{i=q+1}^{n} |a_{iq}|^2 + |b_{q+1,\,q}|^2 - (b^*_{q+1,\,q} a_{q+1,\,q} + b_{q+1,\,q} a^*_{q+1,\,q}). \tag{35}$$

Подстановка тех же выражений в формулу (32) дает

$$\alpha^2 = 2 \sum_{i=q+1}^{n} |a_{iq}|^2 - 2 b^*_{q+1,\,q} a_{q+1,\,q}. \tag{36}$$

Последнее слагаемое в правой части этого равенства должно быть вещественным, поскольку остальные члены вещественны. Поэтому должно выполняться соотношение $\arg b^*_{q+1,\,q} = \pi k - \arg a_{q+1,\,q}$, где k — любое целое число. Для улучшения счетной устойчивости алгоритма выгодно полагать $k = \pm 1$: тогда последний член в правой части (36) будет положительным, и величина α никогда не станет близкой к нулю. Таким образом,

$$\arg b_{q+1,\,q} = \pi + \arg a_{q+1,\,q}. \tag{37}$$

Учитывая этот выбор аргумента и приравнивая друг другу правые части равенств (35) и (36), получим

$$|b_{q+1,\,q}| = \left(\sum_{i=q+1}^{n} |a_{iq}|^2 \right)^{1/2}. \tag{38}$$

Заменяя в (36) сумму при помощи равенства (38) и учитывая (37), упростим выражение для α:

$$\alpha = [2\,|b_{q+1,\,q}|\,(|b_{q+1,\,q}| + |a_{q+1,\,q}|)]^{1/2}. \tag{39}$$

Формулы (37)—(39) и (33)—(34) полностью определяют матрицу очередного отражения. Эти формулы составлены так, что для вещественной матрицы A при вычислениях не возникает комплексных величин, а формула (37) для вычисления аргумента принимает при этом вид

$$\operatorname{sign} b_{q+1,\,q} = -\operatorname{sign} a_{q+1,\,q}.$$

Последовательно полагая $q = 1, 2, \ldots, n-2$, определяя соответствующие векторы $\boldsymbol{w}^q$ и производя отражения, мы приведем произвольную матрицу A к верхней почти треугольной форме. Если исходная матрица A была эрмитова, то результирующая матрица будет трехдиагональной.

Рассмотрим, как экономно организовать вычисления. Формулы для определения матрицы отражения не требуют большого объема

расчетов. Основное число действий уходит на перемножение матричных клеток в формуле (30). Заметим, что клетка A_1 не меняется, а в клетке $B_3 = WA_3$ имеется только один ненулевой элемент $b_{q+1,\,q}$, уже вычисленный при нахождении матрицы отражения; следовательно, эти клетки не нужно специально вычислять. При нахождении остальных двух клеток умножение на матрицу отражения W надо выполнять специальным образом; например, умножим A_4 на W справа, тогда

$$A_4W = A_4(E - 2\boldsymbol{w}\boldsymbol{w}^H) = A_4 - 2(A_4\boldsymbol{w})\,\boldsymbol{w}^H. \qquad (40)$$

Вместо того, чтобы перемножать две матрицы, мы пользуемся ассоциативностью умножения и сводим вычисление к двукратному умножению матрицы на вектор, что примерно в n раз быстрее. Умножение на W слева выполняется аналогично. Заметим, что если матрица A эрмитова, то возможна дополнительная экономия: тогда $B_2 = B_3^H$ и $B_4 = B_4^H$, так что клетку B_2 можно вообще не вычислять, а в клетке B_4 достаточно найти только нижнюю половину элементов.

Устойчивость численного алгоритма теоретически исследована недостаточно. Однако практика вычислений показала, что преобразования унитарными матрицами достаточно устойчивы. Поэтому основное, на что надо обращать внимание, — это чтобы ошибки округления не сказались бы на унитарности матриц отражения. Для контроля следует проверять выполнение условия нормировки (23); если оно соблюдается с очень высокой точностью (верны почти все двоичные разряды), то устойчивость обычно хорошая.

Когда матрица A приведена к трехдиагональной (или верхней почти треугольной) форме, то для этой формы собственные значения λ_i и собственные векторы $\boldsymbol{y}_i$ находятся легко. Найденные собственные значения одновременно являются собственными значениями исходной матрицы A. Для нахождения собственных векторов $\boldsymbol{x}_i$ исходной матрицы надо применить преобразование отражения

$$\boldsymbol{x}_i = (R_1R_2 \ldots R_{n-2})\,\boldsymbol{y}_i. \qquad (41)$$

Вычисления по этой формуле также надо делать экономично, выполняя каждое умножение на очередную матрицу R_q как два умножения на вектор по формуле (24).

Подсчет числа операций показывает, что для эрмитовых матриц метод отражения позволяет найти все собственные значения примерно за $((4/3)\,n^3 + 50n^2)$ арифметических действий, а все собственные векторы — еще за $(2n^3 + 10n^2)$ действий. Это самый быстрый из известных методов. Его скорость настолько велика, что позволяет на ЭВМ класса БЭСМ-6 вести расчет для матриц порядка $n \sim 100$; фактическую границу его применимости определяет устойчивость, которая при расчете с обычной точностью

теряется при меньших значениях n (расчет с двойным числом знаков более устойчив, но и более трудоемок).

К сожалению, метод отражений не дает дополнительной экономии в случае ленточных и других аналогичных специальных структур матриц, потому что такие структуры не сохраняются в процессе расчета.

Для неэрмитовых матриц метод отражения менее выгоден, ибо результирующая матрица является верхней почти треугольной, а для нее описанный в § 1 метод нахождения собственных значений довольно трудоемок. В итоге такой подход требует около $13n^3$ арифметических действий для вычисления всех собственных значений и еще $(9/2)\,n^3$ действий для нахождения всех собственных векторов. Зато численная устойчивость алгоритма хорошая.

2. Прямой метод вращений. Этот метод немного уступает по скорости методу отражений, однако формулы расчета в нем несколько проще. Он также устойчив и позволяет привести преобразованием подобия неэрмитову матрицу к почти треугольной форме, а эрмитову — к трехдиагональной.

Метод основан на специально подобранном вращении координатной системы. Поэтому исследуем свойства вращений. Любое вращение можно свести к последовательности элементарных (плоских) вращений — поворотов в двумерной плоскости, проходящей через k-ю и l-ю оси координат; остальные оси координат при этом неподвижны. Для комплексных векторов матрица элементарного вращения имеет следующий вид:

$$U_{kl}=\begin{bmatrix} 1 & & & & & \\ & 1 & & 0 & & \\ & & 1 & & & \\ & & \alpha & \cdots & -\beta^* & \\ & & \vdots & 1 & \vdots & \\ & & \beta & \cdots & \alpha & \\ & 0 & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \dots k\text{-я} \\ \\ \dots l\text{-я} \\ \\ \end{matrix}, \qquad \begin{matrix} \alpha=\cos\varphi, \\ \beta=e^{-i\psi}\sin\varphi, \\ \alpha^2+|\beta|^2=1. \end{matrix} \tag{42}$$

Остальные недиагональные элементы этой матрицы — нули. Для вещественных векторов надо полагать $\psi=0$. Непосредственным перемножением легко проверить, что $U_{kl}U_{kl}^H=E$, т. е. матрица вращения унитарна (напомним, что преобразования с унитарными матрицами обычно устойчивы).

Очень важно помнить, что если из-за погрешностей расчета окажется, что $\alpha^2+|\beta|^2\neq 1$, или β и β^* в (42) не точно сопряжены друг другу, то унитарность матрицы нарушается. Тогда при преобразовании подобия нарушится эрмитовость матрицы A и сильно ухудшится устойчивость всех методов вращения, которые будут описаны далее. Поэтому в численных расчетах следует определять α и β по таким формулам, чтобы указанные соотношения выполнялись с очень высокой точностью.

Построим формулы для преобразования матрицы A при элементарном вращении. Матрица $B=AU_{kl}$ отличается от матрицы A элементами k-го и l-го столбцов; остальные элементы у них совпадают:

$$\begin{aligned} &b_{ik}=a_{ik}\alpha+a_{il}\beta, \quad b_{il}=-a_{ik}\beta^*+a_{il}\alpha, \quad 1\leqslant i\leqslant n, \\ &b_{ij}=a_{ij} \quad \text{при} \quad j\neq k,\ l \quad \text{и} \quad 1\leqslant i\leqslant n. \end{aligned} \tag{43}$$

Аналогично, матрица $C=U_{kl}^{H}B$ отличается от матрицы B только элементами k-й и l-й строк:

$$c_{kl}=b_{ki}\alpha+b_{li}\beta^{*},\quad c_{li}=-b_{ki}\beta+b_{li}\alpha,\quad 1\leqslant i\leqslant n,$$
$$c_{ji}=b_{ji}\quad\text{при}\quad j\neq k,\ l\quad\text{и}\quad 1\leqslant i\leqslant n. \tag{44}$$

Следовательно, матрица $C=U^{H}AU$ отличается от матрицы A лишь двумя строками и двумя столбцами. Формулы для вычисления элементов этих строк и столбцов написать нетрудно, но в этом нет необходимости; удобнее программировать на ЭВМ непосредственно формулы (43)—(44). Заметим, что если матрица A эрмитова, то матрица C также будет эрмитова; тогда в изменившихся столбцах и строках достаточно вычислить только половину элементов и тем самым вдвое уменьшить объем расчетов.

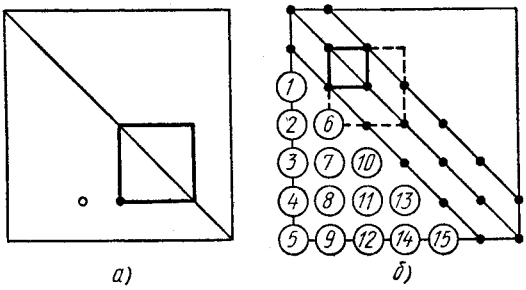


Рис. 33.

Найдем такую последовательность элементарных вращений, которая приводит произвольную (неэрмитову) матрицу A к верхней почти треугольной форме.

Можно так подобрать угол поворота в матрице U_{kl}, чтобы уничтожить элемент $c_{l,\,k-1}$, расположенный непосредственно перед левым нижним углом подматрицы плоского поворота (рис. 33, *а*). Из формул (43) и (44) видно, что для этого надо положить

$$\alpha=\frac{|a_{k,\,k-1}|}{\sqrt{|a_{k,\,k-1}|^2+|a_{l,\,k-1}|^2}},\quad \beta=\frac{\alpha a_{l,\,k-1}}{a_{k,\,k-1}}. \tag{45а}$$

Сами углы вычислять нет необходимости, ибо в формулы для преобразования матричных элементов они не входят. Отметим, что для вещественных матриц величина β тоже будет вещественной; тогда формулы (45) удобнее записать следующим образом:

$$\alpha=\frac{a_{k,\,k-1}}{\sqrt{a_{k,\,k-1}^2+a_{l,\,k-1}^2}},\quad \beta=\frac{a_{l,\,k-1}}{\sqrt{a_{k,\,k-1}^2+a_{l,\,k-1}^2}}. \tag{45б}$$

Теперь будем аннулировать те элементы матрицы и в том порядке, как это указано цифрами на рис. 33, б. Первый элемент уничтожается при помощи матрицы U_{23}, обозначенной на рисунке сплошным квадратом. Второй уничтожается вращением U_{24}, обозначенным пунктирным квадратом. При втором вращении в матрице A меняются элементы вторых и четвертых строк и столбцов. Значит, аннулированный элемент «1», лежащий в третьей строке, так и останется равным нулю.

Продолжая эти рассуждения, можно убедиться, что однажды уничтоженный элемент при такой последовательности исключения будет оставаться равным нулю. Поэтому после окончания всех исключений матрица станет верхней почти треугольной матрицей ($a_{ij}=0$ при $i>j+1$). Это справедливо для произвольной (неэрмитовой) матрицы.

Если исходная матрица A эрмитова, то благодаря сохранению эрмитовости при унитарном преобразовании подобия она приводится к трехдиагональной форме. В этом случае для экономии времени при каждом вращении достаточно

вычислять только изменившиеся элементы нижней половины матрицы (уже обратившиеся в нуль элементы в дальнейшие расчеты не включают).

Для полученной трехдиагональной (или почти треугольной) матрицы можно вычислять собственные значения и собственные векторы способами, изложенными в § 1. Найденные собственные значения будут одновременно собственными значениями исходной матрицы. А собственные векторы x_i исходной матрицы связаны с собственными векторами трехдиагональной матрицы соотношением

$$x_i = U_{23}U_{24} \dots U_{n-1,\,n}y_i. \tag{46}$$

Проще всего вычислять их, последовательно умножая требуемый вектор y слева на матрицы вращения. Структура матриц такова, что при умножении на U_{kl} меняются только k-я и l-я компоненты вектора

$$\begin{aligned} &x_k = \alpha y_k - \beta^* y_l, \quad x_l = \beta y_k + \alpha y_l, \\ &x_j = y_j \quad \text{при} \quad j \neq k,\ l. \end{aligned} \tag{47}$$

Предварительное перемножение самих матриц вращения потребовало бы большего числа действий (это особенно невыгодно, если нужна только часть собственных векторов).

На приведение эрмитовой матрицы к трехдиагональной форме и нахождение всех собственных значений в методе вращений требуется около $2n^3 + 50n^2$ арифметических действий и n^2 ячеек оперативной памяти. Для нахождения каждого собственного вектора надо затратить еще $3n^2$ действий. Собственные значения и собственные векторы в этом методе определяются устойчиво (если унитарность U_{kl} не нарушена ошибками округления).

3. Итерационный метод вращений. Несмотря на свою быстроту, описанные выше прямые методы не вполне удовлетворительны. Так, их алгоритм состоит из разнородных частей: преобразования исходной матрицы, вычисления корней многочлена, нахождения собственных векторов обратными итерациями. Кроме того, их формулы не упрощаются для некоторых употребительных специальных форм матриц (например, ленточных); тем самым они невыгодны для таких матриц. Поэтому разработан и используется ряд итерационных методов, в общем случае более медленных, но обладающих какими-то частными преимуществами.

Для эрмитовых матриц наиболее известен итерационный метод вращений, предложенный Якоби в 1846 г.; но в численных расчетах он начал использоваться только после появления работы [52]. Метод основан на подборе такой бесконечной последовательности элементарных вращений, которая в пределе преобразует эрмитову матрицу A в диагональную. При этом используются преобразования вращения с матрицами (42) такого же типа, как и для прямого метода вращений, но последовательность поворотов и их углы подбираются совершенно иным способом.

Рассмотрим, как действует элементарное вращение на сферическую норму матрицы (точнее, квадрат этой нормы):

$$S = \|A\|_E^2 = \sum_{i,\,j=1}^{n} |a_{ij}|^2. \tag{48}$$

Для определенности рассмотрим сначала умножение справа, $B = AU_{kl}$. Из формулы (43) видно, что при этом элементы k-го и l-го столбцов меняются так, что попарные суммы квадратов модулей сохраняются:

$$|b_{ik}|^2+|b_{il}|^2=|a_{ik}|^2+|a_{il}|^2, \quad 1\leqslant i\leqslant n;$$

элементы остальных столбцов остаются неизменными. Отсюда следует, что $\|AU\|_E=\|A\|_E$, т. е. сферическая норма матрицы A не меняется при умножении справа на матрицу вращения. Аналогичное утверждение легко доказать для умножения на матрицу U или U^H слева *). Описываемый метод основан на сохранении сферической нормы при вращениях.

Разобьем сумму, входящую в сферическую норму (48), на диагональную и недиагональную части:

$$S_1=\sum_{i=1}^{n}|a_{ii}|^2, \quad S_2=\sum_{\substack{i,\,j=1\\ i\neq j}}^{n}|a_{ij}|^2. \tag{49}$$

При элементарном преобразовании вращения $U_{kl}^H A U_{kl}$ недиагональные элементы a_{ik}, a_{il} и a_{ki}, a_{li} при $i\neq k$, l меняются так, что попарные суммы квадратов их модулей сохраняются; это легко видеть из формул (43)—(44). Кроме этих элементов вне диагонали есть еще один меняющийся элемент — это a_{kl}. Поэтому величина S_2 меняется при элементарном вращении настолько, насколько изменится $|a_{kl}|^2$. Будем подбирать вращения так, чтобы S_2 уменьшалась.

Чтобы максимально уменьшить S_2 за одно вращение, подберем угол поворота так, чтобы аннулировать элемент a_{kl}. Для простоты ограничимся вещественными эрмитовыми (т. е. симметричными) матрицами. Тогда $a_{kl}=a_{lk}$ — вещественные числа, и матрицы вращений U_{kl} тоже вещественны. Из формул (44) и (43) с учетом вещественности всех величин следует

$$c_{kl}=b_{kl}\alpha+b_{ll}\beta=a_{kl}(\alpha^2-\beta^2)+(a_{ll}-a_{kk})\alpha\beta.$$

Полагая $c_{kl}=0$ и вспоминая условие нормировки (42), получим систему уравнений для определения параметров поворота

$$\begin{aligned}\alpha^2+\beta^2&=1,\\ a_{kl}(\alpha^2-\beta^2)&=(a_{kk}-a_{ll})\alpha\beta.\end{aligned} \tag{50}$$

Возводя второе уравнение (50) в квадрат и исключая из него β^2 при помощи первого уравнения, получим биквадратное уравнение

*) Это является частным случаем общего утверждения, которое мы не доказываем: сферическая норма любой матрицы не меняется при умножении с любой стороны на унитарную матрицу.

для определения α:

$$\alpha^4 - \alpha^2 + a_{kl}^2 [4a_{kl}^2 + (a_{kk} - a_{ll})^2]^{-1} = 0.$$

Можно выбрать любой из четырех корней этого уравнения, тогда β определится однозначно. Для определенности положим

$$\alpha = \sqrt{\frac{1}{2}\left(1 + \frac{1}{\sqrt{1+\mu^2}}\right)}, \quad \text{где} \quad \mu = \frac{2a_{kl}}{a_{kk} - a_{ll}},$$
$$\beta = (\operatorname{sign} \mu) \sqrt{\frac{1}{2}\left(1 - \frac{1}{\sqrt{1+\mu^2}}\right)}. \tag{51}$$

Сами углы поворота находить не требуется.

Итак, при каждом вращении S_2 уменьшается, а S_1 соответственно увеличивается, поскольку $S = S_1 + S_2$ сохраняется. Если подобрать такую последовательность вращений, чтобы $S_2 \to 0$, то все недиагональные элементы после достаточного числа поворотов станут пренебрежимо малыми и матрица A преобразуется в диагональную. *Диагональные элементы полученной диагональной матрицы и будут искомыми собственными значениями.* Но уничтожить все недиагональные элементы за конечное число поворотов нельзя, ибо, в отличие от прямого метода вращений, здесь при очередном повороте ранее уничтоженный элемент снова может стать ненулевым.

Какой именно недиагональный элемент целесообразно аннулировать при очередном повороте? Конечно, если уничтожать максимальный по модулю внедиагональный элемент, то скорость убывания S_2 будет наибольшей. При ручных расчетах это наилучший способ. Но на ЭВМ перебор элементов матрицы для определения максимального элемента требует неприемлемо большого числа действий. А если аннулировать элементы в заранее определенном порядке — циклом, то сходимость будет очень медленной. Наиболее выгодным оказалось уничтожение так называемого *оптимального элемента.*

Составим суммы квадратов модулей внедиагональных элементов строк:

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|^2. \tag{52}$$

Выберем из этих сумм наибольшую, а в ней выберем наибольший по модулю элемент; его называют *оптимальным.* Поиск оптимального элемента сводится к перебору двух строк (строки сумм r_i, а в выбранной сумме — строки $|a_{ij}|$), т. е. требует малого числа действий. Суммы (52) вычисляются тоже экономично, ибо при каждом вращении из них меняются только две — r_k

и r_l, причем их можно вычислять по таким формулам:

$$\begin{aligned} r'_k &= r_k + a'^2_{kk} - a^2_{kk} - a^2_{kl}, \\ r'_l &= r_l + r_k - r'_k; \end{aligned} \tag{53}$$

штрихи относятся к значениям после вращения.

Доказательство сходимости. Оптимальный элемент составляет не менее $1/(n-1)$ части суммы (52) своей строки, а эта сумма — не менее $1/n$ части S_2. Следовательно, за одно вращение недиагональная часть сферической нормы уменьшается не менее чем на $\frac{2}{n(n-1)}$ долю своей величины (ибо уничтожаются два симметричных элемента). Значит, за N вращений S_2 убывает не медленнее, чем

$$\left[1 - \frac{2}{n(n-1)}\right]^N \approx \exp(-2N/n^2),$$

и тем самым стремится к нулю при $N \to \infty$. Следовательно, процесс Якоби с выбором оптимального (и тем более максимального) элемента всегда сходится.

Исследуем сходимость вблизи решения, считая собственные значения простыми. Пусть все внедиагональные элементы уже малы, $a_{ij} \sim \varepsilon$. Тогда из формул (51) следует, что угол поворота имеет тот же порядок малости: $\beta \sim \varepsilon$, $\alpha \approx 1 - O(\varepsilon^2)$. Подстановка в формулы (43)—(44) показывает, что при этом неуничтожаемые внедиагональные элементы меняются на $O(\varepsilon^2)$. Значит, за один *цикл* вращений *) все внедиагональные элементы станут $\sim \varepsilon^2$, что означает квадратичную сходимость.

Итак, вдали от решения сходимость не хуже линейной, а вблизи решения - квадратичная, т. е. быстрая. Это позволяет получать все собственные значения с высокой точностью. Обычно процесс сходится за $6-8$ циклов вращений, или за $(3 \div 4)\, n^2$ элементарных вращений. Интересно, что чем больше кратных собственных значений, тем быстрей сходится метод.

Поскольку собственные векторы диагональной матрицы суть $\boldsymbol{e}_i$, то собственными векторами матрицы A будут столбцы матрицы $U = \prod U_{kl}$. Заметим, что если внедиагональные элементы $a_{ij} = O(\varepsilon)$, то $S_2 = O(\varepsilon^2)$, так что диагональные элементы отличаются от собственных значений на $O(\varepsilon^2)$. Поэтому для нахождения собственных значений достаточно положить $\varepsilon \approx 10^{-6}$, чтобы получить правильно 10 знаков. Но чтобы получить с той же точностью

*) Циклом будем называть процесс с последовательным перебором всех поддиагональных элементов, или условно — $n(n-1)/2$ последовательных поворотов при другом способе выбора аннулируемых элементов.

собственные векторы, надо или вычислять их по формулам

$$x_i = \left(\prod U_{kl}\right) y_i, \quad y_i = \{y_{ij}\},$$
$$y_{ii} = 1 \text{ и } y_{ij} = \frac{a_{ji}}{a_{ii} - a_{jj}} \text{ при } j \neq i \tag{54}$$

или делать еще один цикл вращений.

Хотя теоретически в методе Якоби могут накапливаться ошибки, фактически устойчивость и точность очень высоки. Кратные собственные значения получаются столь же точно, как и простые, а собственные векторы практически ортогональны друг другу.

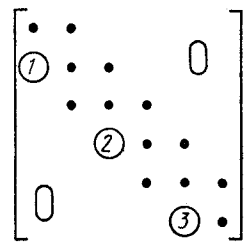


Рис. 34.

Метод Якоби с выбором оптимального элемента требует обычно около $30n^3$ арифметических действий и n^2 ячеек памяти для нахождения всех собственных значений. Для нахождения всех собственных векторов требуется еще около $20n^3$ действий. Таким образом, этот метод раз в 10 медленнее метода отражений. Основное его достоинство — надежность и единообразность вычислений, что позволяет легко запрограммировать метод. Итерационный метод вращений применяется там, где важна точность, надежность и простота расчета и менее существен объем вычислений.

Замечание. Этот метод можно ускорить в 1,5—2 раза, не теряя его достоинств. У произвольных матриц недиагональная часть сферической нормы в среднем много больше диагональной, $S_2/S_1 \sim n$, а у трехдиагональных в среднем $S_2/S_1 \sim 2$. Значит, трехдиагональная матрица является выгодным начальным приближением для итерационного метода Якоби. Поэтому целесообразно предварительно привести исходную эрмитову матрицу к трехдиагональной форме при помощи прямого метода вращений и затем первым ходом метода Якоби аннулировать все нечетные или все четные поддиагональные элементы (рис. 34). После этого можно переходить на обычный вариант итерационного метода вращений с выбором оптимального элемента.

§ 3. Неэрмитовы матрицы

1. Метод элементарных преобразований. В принципе можно привести преобразованием подобия неэрмитову матрицу к почти треугольной форме при помощи отражений или вращений. Для матриц такой формы задача на собственные значения решается сравнительно быстро и устойчиво способом, описанным в § 1. Однако существует втрое более быстрый (хотя несколько менее устойчивый) метод элементарных преобразований; он позволяет привести произвольную матрицу к трехдиагональной форме всего за $2n^3$ арифметических действий. Для неэрмитовых матриц это самый быстрый из известных методов.

Метод является двухходовым. Первым ходом матрица приводится к верхней почти треугольной форме, а вторым — к трехдиагональной форме. Каждый ход состоит из последовательности элементарных преобразований подобия, напоминающих отражения; преобразования первого хода поочередно обращают в нуль столбцы в нижней части матрицы, а преобразования второго хода — строки в верхней части матрицы.

Первый ход. На его q-м шаге для преобразования подобия используется матрица следующего вида:

$$N=\begin{bmatrix} E_q & 0 \\ 0 & N_q \end{bmatrix}\begin{matrix}\}q\\ \}n-q\end{matrix}, \qquad N_q\equiv N_q(v)=\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ v_{q+2} & 1 & 0 & \dots & 0 \\ v_{q+3} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ v_n & 0 & 0 & \dots & 1 \end{bmatrix}. \tag{55}$$

Исследуем свойства этой матрицы. Нетрудно проверить, что $N^HN\neq E$, так что эта матрица *не унитарна* (именно поэтому элементарные преобразования менее устойчивы по отношению к ошибкам округления). Изменим знаки всех компонент v_i, т. е. возьмем матрицу $N(-v)$; непосредственным перемножением легко убеждаемся, что $N(v)N(-v)=E$. Следовательно, обратная к (55) матрица определяется просто:

$$N^{-1}(v)=N(-v). \tag{56}$$

Аналогично методу отражений, матрицы (55) применяются для обращения в нуль нижней части q-го столбца, если уже аннулированы предыдущие столбцы (рис. 32). Разобьем матрицу A на клетки тех же размеров, что и в (55), и запишем преобразование подобия на данном шаге

$$B=N^{-1}AN=\begin{bmatrix} E & 0 \\ 0 & N_q(-v) \end{bmatrix}\cdot\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}\cdot\begin{bmatrix} E & 0 \\ 0 & N_q(v) \end{bmatrix}=$$
$$=\begin{bmatrix} A_1 & A_2N_q(v) \\ N_q(-v)A_3 & N_q(-v)A_4N_q(v) \end{bmatrix}=\begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}. \tag{57}$$

У клетки A_3, а следовательно, и у клетки $B_3=N_q(-v)A_3$ только последний столбец является ненулевым; элементы этого столбца результирующей матрицы получаются поэлементным перемножением клеток:

$$b_{q+1,\,q}=a_{q+1,\,q}, \quad b_{iq}=a_{iq}-v_ia_{q+1,\,q} \text{ при } q+2\leqslant i\leqslant n. \tag{58}$$

Поэтому, чтобы обратить в нуль все элементы клетки B_3, кроме углового элемента $b_{q+1,\,q}$, надо положить

$$v_i=\frac{a_{iq}}{a_{q+1,\,q}} \text{ при } q+2\leqslant i\leqslant n. \tag{59}$$

Последняя формула определяет матрицу искомого элементарного

преобразования. Она существенно проще, чем формулы для нахождения нужной матрицы отражения в п. 2.

Само преобразование (57) очень несложно. Благодаря специальной структуре матрицы N умножение на нее выполняется так же быстро, как умножение на вектор. Например, поэлементно перемножая матрицы, найдем клетку $B_2 = A_2 N_q(v)$:

$$b_{ij} = a_{ij} \text{ при } q+2 \leqslant j \leqslant n,\ 1 \leqslant i \leqslant q,$$
$$b_{i,\,q+1} = a_{i,\,q+1} + \sum_{j=q+2}^{n} a_{ij} v_j \text{ при } 1 \leqslant i \leqslant q; \tag{60}$$

при умножении справа на N_q меняется только первый столбец клетки A_2, а остальные столбцы клетки B_2 равны соответствующим столбцам клетки A_2. Произведение $C_4 = A_4 N_q(v)$ вычисляется также по формулам (60), только с одним отличием: первый индекс элементов принимает значения $q+1 \leqslant i \leqslant n$. Умножение слева на N_q приводит к другим выражениям; так, для четвертой клетки $B_4 = N_q(-v) C_4$ поэлементное перемножение дает

$$b_{q+1,\,j} = c_{q+1,\,j} \text{ при } q+1 \leqslant j \leqslant n,$$
$$b_{ij} = c_{ij} - v_i c_{q+1,\,j} \text{ при } q+1 \leqslant j \leqslant n,\ q+2 \leqslant i \leqslant n, \tag{61}$$

т. е. меняются почти все элементы клетки. Формулы (58)—(61) полностью определяют очередной шаг первого хода. Они экономичны, так что метод элементарных преобразований позволяет привести произвольную матрицу к почти треугольной форме всего за $(5/3)\,n^3$ арифметических действий, т. е. вдвое быстрей, чем в методе отражений.

Но для эрмитовых матриц метод элементарных преобразований невыгоден, ибо при неунитарных преобразованиях эрмитовость не сохраняется. Тем самым результирующая матрица будет почти треугольной, а не трехдиагональной, как в методе отражений; вдобавок выигрыша в скорости по сравнению с методом отражений в этом случае нет.

Однако расчет по полученным формулам еще недостаточно устойчив. Если в ходе расчета на очередном шаге возникает малый ведущий элемент $a_{q+1,\,q}$, то согласно формулам (59) компоненты v_i будут велики. При вычислении остальных клеток матрицы элементы умножаются на эти компоненты, и погрешность сильно возрастает. Чтобы сделать метод устойчивым, выбирают *главный* (т. е. наибольший по модулю) элемент аннулируемого столбца

$$|a_{rq}| = \max_i |a_{iq}| \text{ при } q+2 \leqslant r,\ i \leqslant n \tag{62}$$

и перестановкой $(q+1)$-й и r-й строк делают его ведущим. Тогда будет выполняться неравенство $|v_i| \leqslant 1$, и погрешность практически не будет нарастать. Формально перестановку двух строк

можно записать как умножение слева на *матрицу перестановки* P:

$$A' = P_{q+1,r}A, \quad P_{q+1,r} = \begin{matrix} & \begin{matrix} q+1 & r \end{matrix} \\ \begin{matrix} \\ \\ q+1 \\ \\ r \\ \\ \end{matrix} & \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 0 & 1 & \\ & & 1 & & \\ & & & 1 & \\ & & 1 & 0 & \\ & & & & 1 \end{bmatrix} \end{matrix} \tag{63}$$

(все элементы, стоящие вне главной диагонали, кроме отмеченных здесь, равны нулю). Непосредственным перемножением легко убедиться, что $P = P^{-1} = P^H$, так что свойства матрицы перестановки похожи на свойства матрицы отражения.

Но перестановка строк меняет собственные значения матрицы. Чтобы собственные значения не менялись, надо сделать преобразование подобия $A'' = P^{-1}AP$. Используя свойства матриц перестановок и ассоциативность умножения, получим

$$A'' = P^{-1}AP = PAP = (PA)\,P = A'P.$$

Отсюда видно, что после перестановки строк надо умножить полученную матрицу справа на P. Поэлементным перемножением легко убедиться, что умножение справа на P есть просто перестановка соответствующих столбцов матрицы A' (перестановка выполняется на ЭВМ быстрее, чем арифметические действия).

Таким образом, устойчивое элементарное преобразование подобия имеет следующий вид:

$$B = N(-v)\,PAPN(v). \tag{64а}$$

Важно помнить, что если перемножение этих матриц проводить в строго определенном порядке

$$A \to PA \to N(-v)\,(PA) \to (N(-v)\,PA)\,P \to$$
$$\to (N(-v)\,PAP)\,N(v), \tag{64б}$$

то устойчивость будет хорошей (хотя, по-видимому, метод отражений более устойчив). Если же после перестановки строк сразу переставить столбцы, то это может привести к уменьшению ведущего элемента и потере устойчивости.

В принципе после приведения матрицы к почти треугольной форме можно определить собственные значения способом, описанным в § 1. Но решение этим способом полной проблемы собственных значений для почти треугольной матрицы в несколько раз более трудоемко, чем приведение произвольной матрицы к почти треугольной форме. Поэтому, чтобы использовать высокую скорость метода элементарных преобразований, надо привести матрицу к еще более удобной форме. Ниже приведен способ быстрого преобразования почти треугольной матрицы в трехдиагональную.

Второй ход. Сначала заметим, что если матрица A была *нижней* почти треугольной (т. е. $a_{ij} = 0$ при $i+1 < j$), то при

преобразовании подобия (57) с матрицей N она останется нижней почти треугольной. В самом деле, клетка A_1 при этом преобразовании не меняется. В клетке A_2 ненулевым был только левый нижний элемент $a_{q,\,q+1}$; из формул (60) видно, что в клетке B_2 тоже только он будет отличен от нуля, причем $b_{q,\,q+1}=a_{q,\,q+1}$. Клетка B_4 также имеет нужную форму; в этом нетрудно убедиться, произведя поэлементное умножение.

Мысленно транспонируем все преобразования первого хода; согласно правилам матричной алгебры для этого надо транспонировать все матрицы, а во всех произведениях изменить порядок перемножения матриц на обратный. При транспонировании $N^{-1}(v)$ получим матрицу

$$M=\left[\begin{array}{c|c} E_q & 0 \\ \hline 0 & M_q \end{array}\right], \qquad M_q=\begin{bmatrix} 1 & -v_{q+2} & -v_{q+3} & \dots & -v_n \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \tag{65}$$

причем $M^{-1}(v)=M(-v)$. Тогда предыдущий вывод принимает такую форму: если для верхней почти треугольной матрицы производится преобразование подобия при помощи матрицы (65), то результирующая матрица остается верхней почти треугольной.

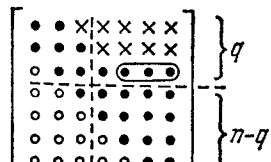


Рис. 35.

Применим цепочку преобразований подобия $M^{-1}AM$ к матрице, полученной в результате первого хода. На каждом шаге элементарную матрицу $M(v)$ будем подбирать так, чтобы аннулировать элементы правой половины очередной строки (см. рис. 35, где точками обозначены ненулевые элементы, кружками — элементы, обращаемые в нуль на первом ходе, крестиками — обращаемые в нуль на начальных шагах второго хода; элементы, аннулируемые на очередном шаге, обведены). Благодаря такому выбору результирующая матрица должна стать нижней почти треугольной. Но в силу сказанного выше она одновременно остается верхней почти треугольной. Тем самым она будет трехдиагональной, что и требовалось.

Все формулы второго хода легко получить, применяя описанное выше транспонирование к формулам первого хода. Например, пусть аннулированы первые $q-1$ строк, и надо аннулировать q-ю строку. Тогда подбор элементов искомого преобразования производится аналогично формуле (59):

$$v_j=\frac{a_{q,\,j}}{a_{q,\,q+1}} \quad \text{при} \quad q+2\leqslant j\leqslant n, \tag{66}$$

и т. д. Поскольку на втором ходе в клетке A_3 имеется только

один ненулевой элемент, а в клетке A_4 около половины элементов — нулевые, то преобразование подобия почти треугольной матрицы к трехдиагональной форме требует всего $n^3/3$ арифметических действий, т. е. является очень быстрым.

Однако, если на некотором шаге второго хода ведущий элемент $a_{q,\,q+1}$ окажется очень малым, то расчет становится неустойчивым. А улучшать устойчивость перестановкой строк и столбцов здесь уже нельзя, ибо такая перестановка нарушает структуру матрицы (она перестает быть верхней почти треугольной). Поэтому, чтобы ослабить влияние ошибок округления, нередко производят расчет второго хода и вычисление характеристического многочлена трехдиагональной матрицы с двойной точностью. В отдельных случаях и это не помогает; тогда производят какую-либо перестановку столбцов исходной матрицы и такую же перестановку строк и повторяют расчет с самого начала. Правда, на практике срывы процесса довольно редки.

Всего двухходовой метод элементарных преобразований требует $2n^3$ действий для приведения матрицы к трехдиагональной форме, около $60n^2$ действий для нахождения всех собственных значений и собственных векторов трехдиагональной матрицы, и еще $2n^3$ действий для преобразования этих векторов в собственные векторы исходной матрицы. Это всего лишь в 7—8 раз больше, чем нужно для решения очень простой задачи — линейной системы того же порядка!

Таким образом, метод элементарных преобразований является очень быстрым и в большинстве случаев устойчивым.

2. Итерационные методы. Существует много методов, основанных на бесконечной последовательности преобразований подобия, приводящей матрицу к некоторым специальным формам, для которых полная проблема собственных значений легко решается. Итерационные методы сложнее прямых, а для матриц произвольного вида заметно уступают прямым методам по скорости (и зачастую — по устойчивости). Но поскольку известные прямые методы не совсем удовлетворительны, то пренебрегать итерационными методами не следует. Ниже даны краткие сведения о наиболее известных итерационных методах; подробное изложение их алгоритмов имеется, например, в монографиях [5, 41].

Метод обобщенных вращений (развитый Эберлейн и В. В. Воеводиным в 1962—1965 гг.) основан на преобразовании матрицы к квазидиагональной форме, когда по главной диагонали расположены клетки, порядки которых равны кратности соответствующих собственных значений, а все остальные элементы матрицы равны нулю (разумеется, приближенно, ибо процесс итерационный). Если все собственные значения простые, то процесс сходится к диагональной матрице.

Для клеток, соответствующих кратным собственным значениям, надо находить собственные значения и собственные векторы специальным алгоритмом, т. е. в метод включаются дополнительные процедуры. Это неудобство имеется во всех итерационных методах. Но поскольку порядок таких клеток обычно невелик, то это не вызывает серьезных затруднений.

Шаг процесса состоит из двух частей. На первом полушаге делается элементарное преобразование матрицей типа N или M, в которой только одна из компонент v отлична от нуля; ее величина подбирается так, чтобы как

можно сильнее уменьшить $\| A \|_E$. Поскольку для любой матрицы $\| A \|_E^2 \geqslant \sum_{i=1}^{n} |\lambda_i|^2$, причем только для нормальных матриц имеет место равенство, то такое преобразование приближает матрицу к нормальной. Второй полушаг — это вращение типа Якоби; для вещественной матрицы угол поворота φ определяется из условия

$$\operatorname{tg} 2\varphi = (a_{kl} + a_{lk})/(a_{kk} - a_{ll}).$$

Процесс организован так, что полный шаг для эрмитовых матриц точно совпадает с циклическим вариантом метода Якоби. Значит, вычисления в общем случае требуют более $50n^3$ арифметических действий, т. е. метод довольно медленный. Зато он является одним из наиболее устойчивых.

Ортогональный степенной метод (предложенный В. В. Воеводиным в 1962 г.) основан на преобразовании матрицы к квазитреугольной форме, когда на главной диагонали стоят клетки, а ниже их все элементы равны нулю. У таких матриц собственные значения равны собственным значениям диагональных клеток, но собственные векторы определяются сложней и значительно менее точно.

Ортогональный степенной метод устойчив и всегда сходится. Скорость сходимости линейная, со знаменателем типа $\mu = \max |\lambda_i/\lambda_{i+1}|$, где λ_i — собственные значения, расположенные в порядке возрастания модулей (причем кратные значения считаются за одно). Следовательно, требуемое число итераций довольно велико, особенно если среди собственных значений есть близкие. Одна итерация требует $(10/3)\,n^3$ арифметических действий, так что метод оказывается весьма медленным.

Треугольный степенной метод (предложен Бауэром в 1957 г.) также основан на преобразовании матрицы к квазитреугольной форме. Сходимость его тоже линейная, но одна итерация требует всего $(5/3)\,n^3$ действий, а при небольшом усложнении алгоритма — даже $(2/3)\,n^3$ действий. Зато этот метод менее устойчив, чем ортогональный степенной метод, особенно если собственные значения комплексные, или в расчетах появляются матрицы с близкими к нулю главными минорами. Зачастую для сохранения устойчивости приходится видоизменять алгоритм.

LR-алгоритм (предложен Рутисхаузером и Бауэром в 1955 г.) тоже содержит преобразование матрицы к квазитреугольной форме. Он разработан только для вещественных матриц с вещественными собственными значениями. Метод всегда сходится, причем вблизи решения квадратично; одна итерация требует $(7/3)\,n^3$ действий. Таким образом, по скорости этот метод превосходит ортогональный степенной; зато он уступает ему по устойчивости.

QR-алгоритм (предложен В. Н. Кублановской и Френсисом в 1961 г.) основан на преобразовании матрицы к квазитреугольной форме. По устойчивости и характеру сходимости он аналогичен ортогональному степенному методу. Этот метод очень выгоден для верхних почти треугольных матриц: в ходе преобразований их структура не разрушается, и благодаря этому одна итерация требует всего $6n^2$ арифметических действий (т. е. время расчета уменьшается в $n/2$ раз по сравнению с общим случаем). Детали этого алгоритма хорошо отработаны, и существуют основанные на нем стандартные программы.

LP-алгоритм (предложен Рутисхаузером в 1955 г.) рассчитан только на вещественные матрицы с вещественными собственными значениями. Он близок к треугольному степенному методу, не очень устойчив и сходится медленно (построены даже примеры зацикливания процесса). Зато для почти треугольных матриц он требует всего $2n^2$ действий на одну итерацию, а для ленточных матриц дает еще бо́льшую экономию.

3. Некоторые частные случаи. *Косоэрмитова* матрица A умножением на i превращается в эрмитову матрицу $B = iA$; для эрмитовой же матрицы проблема собственных значений решается

гораздо легче, чем для неэрмитовых. Для комплексных матриц этот способ наиболее удобен. Для вещественных косоэрмитовых (кососимметричных) матриц он несколько менее выгоден, ибо после умножения на i приходится все остальные действия выполнять с комплексными числами.

Обобщенная проблема собственных значений

$$Ax = \lambda Bx \tag{67}$$

особенно легко решается, если матрицы A и B эрмитовы и одна из них — положительно определенная. Будем считать, что положительно определена вторая матрица (если положительно определена матрица A, то задачу (67) надо переписать в виде $Bx = \lambda^{-1}Ax$).

Разложим матрицу B методом квадратного корня (см. главу V, § 1) в произведение двух треугольных $B = S^H S$; благодаря положительной определенности матрицы B в этом разложении отсутствует диагональная матрица. Тогда можно переписать исходную задачу (67) в следующем виде:

$$Cy = \lambda y, \text{ где } y = Sx, \quad C = (S^H)^{-1} A S^{-1}. \tag{68}$$

Вычисление треугольной матрицы S, ее обращение и нахождение матрицы C выполняются за $4n^3$ действий. Легко проверить, что C — эрмитова матрица; таким образом, задача свелась к хорошо изученной.

Замечание. Запись задачи (67) в виде $(B^{-1}A)x = \lambda x$ невыгодна, ибо матрица $B^{-1}A$ не будет, вообще говоря, эрмитовой.

Нормальную матрицу A по теореме Шура можно привести к диагональной форме унитарным преобразованием подобия. Например, если уничтожены все элементы нижней половины, то в силу нормальности матрицы полученная верхняя треугольная матрица будет диагональной.

Чтобы реализовать эту идею, были предложены разные варианты итерационного метода вращений: циклическое аннулирование поддиагональных элементов, или уменьшение поддиагональной части сферической нормы. Однако они оказались неудачными: были построены примеры, в которых эти процессы сходились к недиагональным матрицам.

Удовлетворительным оказался довольно искусственный вариант. Рассмотрим преобразование $V_{kl}^H A U_{kl}$, производимое унитарными матрицами вращения (оно не является преобразованием подобия). Если углы поворота справа и слева разные, то их можно подобрать так, что на каждом повороте недиагональная часть сферической нормы уменьшается, а бесконечная цепочка преобразований приводит матрицу к диагональной форме: $V^H A U \to D$, где $V = \prod V_{kl}$, $U = \prod U_{kl}$.

Теперь рассмотрим преобразование подобия $B = U^H A U$, выполненное найденной матрицей поворота. Можно доказать, что в матрице B равны нулю (разумеется, приближенно) все недиагональные элементы, кроме тех, которые лежат на пересечении строк и столбцов, для которых диагональные элементы вспомогательной матрицы D равны между собой по модулю: $b_{ik} = 0$, если $|d_{ii}| \neq |d_{kk}|$.

Сделаем такую перестановку столбцов и строк с одинаковыми номерами (а это — преобразование подобия), чтобы в матрице D равные по модулю диагональные элементы заняли соседние места вдоль диагонали. Такая перестановка приводит матрицу B к квазидиагональной форме, после чего задача на собственные значения легко решается (если размеры клеток невелики).

§ 4. Частичная проблема собственных значений

1. Особенности проблемы. Во многих задачах интересны не все собственные значения, а только небольшая их часть.

Матрицы очень высоких порядков обычно получаются при конечно-разностном решении задач на собственные значения для дифференциальных уравнений. В этом случае достаточно вычислить только несколько низших собственных значений, соответствующих малому числу нулей собственной функции (а высокие собственные значения матрицы все равно плохо аппроксимируют соответствующие собственные значения дифференциального оператора в силу свойств конечно-разностных методов). В задачах диффузии нейтронов имеют физический смысл только одно или два собственных значения, и т. д.

В этих случаях решать полную проблему собственных значений невыгодно. Обычно применяют итерационные процессы, сходящиеся к одному собственному значению и собственному вектору. Большинство этих процессов для особенно важных на практике ленточных матриц записывается очень экономно.

Для описанных ниже процессов нужно задавать нулевое приближение. Если оно удачно выбрано, то число итераций заметно уменьшается. Зачастую хорошее нулевое приближение можно получить из физических соображений.

2. Метод линеаризации. Запишем задачу на собственные значения (1) через компоненты собственного вектора:

$$F_i(\boldsymbol{x}, \lambda) \equiv \sum_{k=1}^{n} a_{ik}x_k - \lambda x_i = 0, \quad 1 \leqslant i \leqslant n. \tag{69}$$

Задачу (69) можно рассматривать как систему уравнений с $n+1$ неизвестным $x_1, x_2, \ldots, x_n, \lambda$; эта система нелинейна благодаря наличию членов λx_i. Для решения нелинейной системы целесообразно применить метод Ньютона. Давая всем переменным малые приращения δx_i, $\delta\lambda$ и линеаризуя уравнения (69) относительно приращений, получим

$$\sum_{k=1}^{n} a_{ik}\delta x_k - \lambda^{(s)}\delta x_i - x_i^{(s)}\delta\lambda = -F_i(\boldsymbol{x}^{(s)}, \lambda^{(s)}), \quad 1 \leqslant i \leqslant n; \tag{70}$$

здесь индекс s обозначает номер итерации. Система (70) содержит n

уравнений, линейных относительно неизвестных приращений. Этих неизвестных $n+1$; но поскольку собственный вектор определен с точностью до множителя, то, не нарушая общности, можно положить или $\delta x_1 = 0$, или $\delta x_n = 0$. После этого число уравнений будет равно числу неизвестных приращений.

Напомним, что ньютоновский процесс вблизи решения сходится квадратично. Число итераций зависит от выбора нулевого приближения. При удачном приближении достаточно 3—5 итераций. А если за 10 итераций процесс не сошелся, то скорее всего он не сойдется, и надо изменить нулевое приближение. Заметим, что при неудачном нулевом приближении процесс иногда сходится не к искомому собственному значению, а к другому.

Матрица линейной системы (70) отличается от матрицы A по структуре мало — только добавлением одного ненулевого столбца. Поэтому для матрицы A общего вида на одну итерацию требуется $^2/_3 n^3$ арифметических действий, для почти треугольной матрицы — всего $2n^2$ действий, а для ленточной матрицы даже $m^2 n/2$ действий (где m — ширина ленты).

Метод линеаризации успешно применяется к матрицам порядка $n \approx 100 - 1000$. Он особенно выгоден для трехдиагональных матриц; для них получение одного собственного значения и собственного вектора требует обычно $\sim 50\ n$ арифметических действий.

3. Степенной метод (*счет на установление*) применяется для получения наибольшего по модулю собственного значения. Пусть $|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots$ Построим такой итерационный процесс:

$$x^{(s+1)} = Ax^{(s)}. \tag{71}$$

Он не сходится в обычном смысле. Разложим нулевое приближение по собственным векторам матрицы: $x^{(0)} = \sum_i \xi_i x_i$. Тогда легко убедиться, что $x^{(s)} = \sum_i \lambda_i^s \xi_i x_i$ и при достаточно большом числе итераций $x^{(s)} \approx \lambda_1^s \xi_1 x_1$, т. е. вектор $x^{(s)}$ сходится к собственному вектору по направлению. Очевидно, при этом $x^{(s+1)} \approx \lambda_1 x^{(s)}$.

Процесс сходится линейно со знаменателем $q \approx |\lambda_2/\lambda_1|$. Считается, что процесс практически сошелся, если отношения соответствующих координат векторов $x^{(s+1)}$ и $x^{(s)}$ с требуемой точностью одинаковы и не меняются на последних итерациях. При этом для более точного получения собственного значения целесообразно положить

$$|\lambda_1| \approx \frac{|x^{(s+1)}|}{|x^{(s)}|} = \sqrt{\frac{(x^{(s+1)},\ x^{(s+1)})}{(x^{(s)},\ x^{(s)})}}. \tag{72}$$

Отметим, что при расчетах на ЭВМ на каждой итерации после вычисления λ_1 вектор $x^{(s+1)}$ надо нормировать, чтобы не получать переполнений или исчезновений чисел.

Формально при $\xi_1 = 0$ итерации сходятся к следующему собственному значению. Однако из-за ошибок округления ξ_1 не может быть точно нулем, а при малом ξ_1 процесс по-прежнему сходится к первому собственному значению, только за большее число итераций.

Если наибольшее собственное значение кратное, но соответствующий элементарный делитель матрицы линеен, то итерации сходятся обычным образом. Но если $\lambda_1 \neq \lambda_2$, а их модули равны или если элементарный делитель матрицы нелинеен (жорданова клетка), то процесс не сходится.

Если $|\lambda_1| \approx |\lambda_2|$, то сходимость очень медленная; этот случай нередко встречается в простейших итерационных методах решения разностных схем для эллиптических уравнений (глава XII). Тогда сходимость можно ускорить процессом Эйткена (см. главу IV, § 1).

Одна итерация для матрицы общего вида требует $2n^2$ арифметических действий, а для ленточной матрицы—$2mn$ действий. Из-за медленной сходимости степенной метод применяют только к матрицам, содержащим очень много нулевых элементов (и даже к ним—довольно редко).

В математической литературе описана вариация степенного метода, имеющая квадратичную сходимость: $x^{(s)} = A_s x^{(0)}$, где $A_s = A_{s-1} A_{s-1}$ и $A_0 = A$. Однако если матрица A имеет много нулевых элементов, то ее степени уже такими не будут. Поэтому этот вариант обычно не экономичен.

4. Обратные итерации со сдвигом. Напишем итерационный процесс, обратный по отношению к степенному процессу:

$$x^{(s+1)} = A^{-1} x^{(s)}. \tag{73}$$

Очевидно, он сходится в указанном в п. 3 смысле к наибольшему по модулю собственному значению матрицы A^{-1}, т. е. к наименьшему по модулю собственному значению матрицы A (ибо собственные значения матриц A и A^{-1} обратны друг другу). Все, что говорилось в предыдущем пункте о характере сходимости, разумеется, справедливо и в этом случае; сходимость будет довольно медленной.

Однако здесь положение можно существенно улучшить *методом сдвига*, который заключается в следующем. Пусть нам приближенно известно некоторое, не обязательно наименьшее, собственное значение $\tilde{\lambda}_i$. Тогда так называемая *сдвинутая матрица* $(A - \tilde{\lambda}_i E)$ будет иметь собственные значения $\lambda - \tilde{\lambda}_i$. У этой матрицы интересующее нас собственное значение $\lambda_i - \tilde{\lambda}_i$ будет намного меньше по модулю, чем остальные. Поэтому обратные итерации со сдвинутой матрицей (которые мы запишем в несколько иной форме)

$$(A - \tilde{\lambda}_i E) x^{(s+1)} = x^{(s)}, \tag{74a}$$

будут быстро сходиться и определят требуемое нам собственное значение $\lambda_i - \tilde{\lambda}_i$. Напомним, что после каждой итерации надо нормировать вектор, чтобы избежать переполнений. С учетом этого вместо (74а) получим последовательность формул

$$(A - \tilde{\lambda}_i E)\boldsymbol{y}^{(s)} = \boldsymbol{x}^{(s)},$$
$$\lambda_i^{(s)} - \tilde{\lambda}_i = \left\langle \frac{x_k^{(s)}}{y_k^{(s)}} \right\rangle, \quad \boldsymbol{x}^{(s+1)} = \frac{\boldsymbol{y}^{(s)}}{\|\boldsymbol{y}^{(s)}\|}. \tag{74б}$$

Здесь индекс k относится к компонентам векторов, а скобки $\langle \ldots \rangle$ означают некоторое усреднение по всем компонентам: например, среднеарифметическое.

Если исходное приближение было хорошим, то иногда процесс сходится за несколько итераций; тогда выгодно непосредственно решать линейную систему (73). Если же требуемое число итераций велико, то лучше обратить матрицу $(A - \tilde{\lambda}_i E)$. Выгодней всего при решении линейной системы (74) методом исключения Гаусса использовать полученные на первой же итерации вспомогательные коэффициенты c_{mk} (см. главу V, § 1, п. 1) на каждой последующей итерации; но это не предусмотрено в обычных стандартных программах.

Если сдвиг постоянный, то итерации сходятся линейно. Можно получить квадратичную сходимость, если уточнять сдвиг в ходе расчета следующим образом:

$$(A - \lambda_i^{(s)} E)\boldsymbol{y}^{(s)} = \boldsymbol{x}^{(s)},$$
$$\lambda_i^{(s+1)} = \lambda_i^{(s)} + \left\langle \frac{x_k^{(s)}}{y_k^{(s)}} \right\rangle, \quad \boldsymbol{x}^{(s+1)} = \frac{\boldsymbol{y}^{(s)}}{\|\boldsymbol{y}^{(s)}\|}. \tag{75}$$

Для матриц, имеющих ортогональную систему собственных векторов (например, эрмитовых матриц), сходимость вблизи корня будет даже кубической. Заметим, что допускать слишком точное совпадение $\tilde{\lambda}_i$ с собственным значением нельзя, ибо матрица системы (75) становится плохо обусловленной; об этом уже говорилось в § 1, п. 6 в связи с нахождением собственных векторов. Поэтому, когда в ходе итераций у величины $\lambda_i^{(s)}$ устанавливаются (т. е. перестают меняться) 5—7 знаков, то итерации следует прекращать.

Замечание 1. Переменный сдвиг собственного значения (75) нельзя включать с первой итерации; сначала надо получить грубую сходимость итераций с постоянным сдвигом.

Замечание 2. Обратные итерации особенно удобны, если матрица заранее приведена преобразованием подобия к почти треугольной форме. Тогда одна обратная итерация выполняется

методом исключения с выбором главного элемента всего за $2n^2$ действий. Теоретически для ленточных матриц возможна еще бо́льшая экономия, но преобразование подобия почти треугольной матрицы к трехдиагональной форме не всегда устойчиво.

Выводы. Обратные итерации с постоянным и особенно с переменным сдвигом — очень эффективный метод расчета. Для нахождения собственных векторов этот метод считается наиболее точным. Сходимость при хорошем подборе $\tilde{\lambda}$ настолько быстрая, что метод пригоден и для близких или случайно равных по модулю собственных значений (ибо после сдвига они хорошо различаются), и даже при наличии у матрицы нелинейного элементарного делителя.

ЗАДАЧИ

1. Доказать, что если матрица n-го порядка имеет n собственных векторов $e_i = \{\delta_{in}\}$, $1 \leqslant k \leqslant n$, то она диагональна.

2. Найти собственные векторы треугольной матрицы, считая все собственные значения простыми.

3. Доказать, что нормальная матрица при унитарном преобразовании подобия остается нормальной.

4. Показать, что если матрица A ленточная, то преобразование подобия матрицами отражения (30) не сохраняет ее структуры.

5. Какие элементы необходимо вычислять в формулах (43) — (44) при преобразовании подобия матрицами вращения для эрмитовой матрицы A?

6. Доказать, что сферическая норма произвольной матрицы не меняется при умножении с любой стороны на унитарную матрицу.

7. В итерационном методе вращений вывести для определения параметров поворота комплексных матриц формулу, аналогичную (51).

8. Показать, что в итерационном методе вращений формулы (54) определяют собственные векторы с точностью $O(\varepsilon^2)$, где ε — максимум модулей внедиагональных элементов; если же в этих формулах положить $y_i = e_i$, то точность ухудшается до $O(\varepsilon)$.

9. Получить все формулы расчета матричных элементов для второго хода метода элементарных преобразований.

10. Написать формулы восстановления собственных векторов исходной матрицы по собственным векторам трехдиагональной матрицы в методе элементарных преобразований.

11. Показать, что если матрица A ленточная, то элементарное преобразование подобия (57) разрушает ее структуру.

12. а) Какой вид примут формулы метода линеаризации (70), если недостающее уравнение получать из условия нормировки собственного вектора $\sum_{i=1}^{n} |x_i|^2 = 1$? б) Как построить экономичный алгоритм решения полученной при этом линейной системы, если матрица A является трехдиагональной?

13. Доказать, что метод обратных итераций с переменным сдвигом (75) сходится квадратично вблизи простого собственного значения.

ГЛАВА VII

ПОИСК МИНИМУМА

В главе VII рассмотрены способы нахождения такого значения аргумента, которое минимизирует некоторую зависящую от него скалярную величину. В § 1 изложена задача о минимуме функции одного переменного, лежащая в основе всех более сложных задач. В § 2 рассмотрена задача о минимуме функции многих переменных в неограниченной области. В § 3 область изменения переменных ограничена; наряду с общим случаем рассмотрена частная задача линейного программирования, важная в приложениях к экономике. В § 4 разобрана задача о минимизации функционала, когда аргумент сам является функцией одного или нескольких переменных.

§ 1. Минимум функции одного переменного

1. Постановка задачи. Пусть имеется некоторое множество X, состоящее из элементов x, принадлежащих какому-нибудь метрическому пространству, и на нем определена скалярная функция $\Phi(x)$. Говорят, что $\Phi(x)$ *имеет локальный минимум на элементе* $\bar{x}$, если существует некоторая конечная ε-окрестность этого элемента, в которой выполняется

$$\Phi(\bar{x}) < \Phi(x), \quad \|x - \bar{x}\| \leqslant \varepsilon. \tag{1}$$

У функции может быть много локальных минимумов. Если же выполняется

$$\Phi(\bar{x}) = \inf_X \Phi(x), \tag{2}$$

то говорят о достижении функцией *абсолютного минимума на данном множестве* X.

Естественно требовать, чтобы функция $\Phi(x)$ была непрерывной или, по крайней мере, кусочно-непрерывной, а множество X было компактно*) и замкнуто**) (в частности, если X само является

*) Множество компактно, если из каждого бесконечного и ограниченного его подмножества можно выделить сходящуюся последовательность.

**) Множество замкнуто, если предел любой сходящейся последовательности его элементов принадлежит этому множеству.

пространством, то это пространство должно быть банаховым). Если эти требования не соблюдены, то вряд ли возможно построить разумный алгоритм нахождения решения. Например, если $\Phi(x)$ не является кусочно-непрерывной, то единственным способом решения задачи является перебор всех элементов x, на которых задана функция; этот способ нельзя считать приемлемым. Чем более жестким требованиям удовлетворяет $\Phi(x)$ (таким, как существование непрерывных производных различного порядка), тем легче построить хорошие численные алгоритмы.

Перечислим наиболее важные примеры множеств, на которых приходится решать задачу нахождения минимума. Если множество X является числовой осью, то (1) или (2) есть задача на минимум функции одного вещественного переменного. Если X есть n-мерное векторное пространство, то мы имеем дело с задачей на минимум функции n переменных. Если X есть пространство функций $x(t)$, то (1) называют задачей на минимум функционала.

Для нахождения абсолютного минимума есть только один способ: найти все локальные минимумы, сравнить их и выбрать наименьшее значение. Поэтому задача (2) сводится к задаче (1), и мы будем в основном заниматься задачей поиска локальных минимумов.

Известно, что решение задачи (1) удовлетворяет уравнению

$$\frac{\delta\Phi}{\delta x} = 0. \tag{3}$$

Если множество X есть числовая ось, то написанная здесь производная является обычной производной, и тогда уравнение (3) есть просто одно (нелинейное) уравнение с одним неизвестным. Для n-мерного векторного пространства соотношение (3) оказывается системой нелинейных уравнений $\partial\Phi/\partial x_i = 0$, $1 \leqslant i \leqslant n$. Для пространства функций уравнение (3) является дифференциальным или интегро-дифференциальным. В принципе такие уравнения можно решать численными методами, описанными в главах V и XIV. Однако эти уравнения нередко имеют сложный вид, так что итерационные методы их решения могут очень плохо сходиться или вообще не сходиться. Поэтому в данной главе мы рассмотрим численные методы, применимые непосредственно к задаче (1), без приведения ее к форме (3).

Пусть X является некоторым множеством, принадлежащим какому-то пространству. Тогда (1) называют задачей на минимум в ограниченной области. В частности, если множество X выделено из пространства с помощью ограничивающих условий типа равенств, то задачу (1) называют задачей на условный экстремум; такие задачи методом неопределенных множителей Лагранжа часто можно свести к задачам на безусловный экстремум. Однако при

численном решении обычно удобнее иметь дело непосредственно с исходной задачей (1), хотя при ее решении в ограниченной области возникают свои трудности.

Функция $\Phi(x)$ может иметь на множестве X более одного локального минимума. В конкретных прикладных задачах далеко не всегда удается заранее исследовать свойства функции. Поэтому желательно, чтобы численный алгоритм позволял определить число минимумов и их расположение и аккуратно найти абсолютный минимум.

Задачу называют *детерминированной*, если погрешностью вычисления (или экспериментального определения) функции $\Phi(x)$ можно пренебречь. В противном случае задачу называют *стохастической*. Мы будем рассматривать в основном детерминированные задачи. Для решения стохастических задач есть специальные методы, но они очень медленные, и применять их к детерминированным задачам невыгодно.

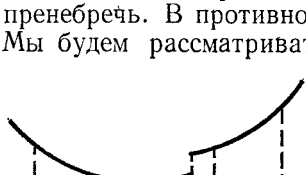


Рис. 36.

2. Золотое сечение. В этом параграфе мы рассмотрим задачу нахождения минимума функции одной действительной переменной. Эта одномерная задача нередко возникает в практических приложениях. Кроме того, большинство методов решения многомерных задач сводится к поиску одномерного минимума.

Сейчас мы рассмотрим метод золотого сечения, применимый к недифференцируемым функциям. Будем считать, что $\Phi(x)$ задана и кусочно-непрерывна на отрезке $a \leqslant x \leqslant b$, и имеет на этом отрезке (включая его концы) только один локальный минимум. Построим итерационный процесс, сходящийся к этому минимуму.

Вычислим функцию на концах отрезка, а также в двух внутренних точках x_1, x_2, сравним все четыре значения функции между собой и выберем среди них наименьшее. Пусть наименьшим оказалось $\Phi(x_1)$. Очевидно, минимум расположен в одном из прилегающих к нему отрезков (рис. 36). Поэтому отрезок $[x_2, b]$ можно отбросить и оставить отрезок $[a, x_2]$. Первый шаг процесса сделан.

На отрезке $[a, x_2]$ снова надо выбрать две внутренние точки, вычислить в них и на концах отрезка значения функции, и сделать следующий шаг процесса. Но на предыдущем шаге вычислений мы уже нашли $\Phi(x)$ на концах нового отрезка a, x_2 и в одной его внутренней точке x_1. Поэтому достаточно выбрать внутри $[a, x_2]$ еще одну точку x_3, определить в ней значение функции и провести необходимые сравнения. Это вчетверо уменьшает объем вычислений на одном шаге процесса.

Как выгодно размещать точки? Всякий раз мы делим оставшийся отрезок на три части (причем одна из точек деления уже определена предыдущими вычислениями) и затем отбрасываем один из крайних отрезков. Очевидно, надо, чтобы следующий отрезок был поделен подобно предыдущему. Для этого должны выполняться соотношения

$$b-x_2=x_1-a, \qquad \frac{x_1-a}{b-a}=\frac{x_2-x_1}{x_2-a}.$$

Решение этих уравнений дает

$$\frac{b-x_2}{b-a}=\frac{x_1-a}{b-a}=\xi, \quad \xi=\frac{2}{3+\sqrt{5}}\approx 0{,}38. \tag{4}$$

После проведения очередного вычисления отрезок сокращается в $1-\xi\approx 0{,}62$ раза; после n вычислений функции он составляет $(1-\xi)^{n-3}$ долю первоначальной величины (три первых вычисления в точках a, b, x_1 еще не сокращают отрезок). Следовательно, при $n\to\infty$ длина оставшегося отрезка стремится к нулю как геометрическая прогрессия со знаменателем $1-\xi\approx 0{,}62$, т. е. метод золотого сечения всегда сходится, причем линейно.

Запишем алгоритм вычисления. Для единообразия записи обозначим

$$a=x_0, \quad b=x_1,$$

а поочередно вводимые внутренние точки будут x_2, x_3, ... На первом шаге полагаем согласно (4)

$$x_2=x_0+\xi(x_1-x_0), \quad x_3=x_1-\xi(x_1-x_0). \tag{5}$$

После сравнения может быть отброшена точка с любым номером, так что на следующих шагах оставшиеся точки будут перенумерованы беспорядочно. Пусть на данном отрезке есть четыре точки x_i, x_j, x_k, x_l, из которых какие-то две являются концами отрезка. Выберем ту точку, в которой функция принимает наименьшее значение; пусть это оказалось x_i:

$$\Phi(x_i)<\Phi(x_j), \quad \Phi(x_k), \quad \Phi(x_l). \tag{6}$$

Затем отбрасываем ту точку, которая более всего удалена*) от x_i; пусть этой точкой оказалась x_l:

$$|x_l-x_i|>|x_j-x_l|, \ |x_k-x_l|. \tag{7}$$

Определим порядок расположения оставшихся трех точек на числовой оси; пусть, для определенности,

$$x_k<x_i<x_j. \tag{8}$$

*) Это верно не при всяких делениях отрезка, но для деления в соответствии (4) это справедливо.

Тогда новую внутреннюю точку введем таким соотношением *):

$$x = x_j + x_k - x_i, \tag{9}$$

и присвоим ей очередной номер. Минимум находится где-то внутри последнего отрезка, $x_k \leqslant \bar{x} \leqslant x_j$. Поэтому итерации прекращаем, когда длина этого отрезка станет меньше заданной погрешности δ:

$$x_j - x_k < \delta. \tag{10}$$

Метод золотого сечения является наиболее экономичным аналогом метода дихотомий применительно к задачам на минимум. Он применим даже к недифференцируемым функциям и всегда сходится; сходимость его линейна. Если на отрезке $[a, b]$ функция имеет несколько локальных минимумов, то процесс сойдется к одному из них (но не обязательно к наименьшему).

Этот метод нередко применяют в технических или экономических задачах оптимизации, когда минимизируемая функция недифференцируема, а каждое вычисление функции — это дорогой эксперимент.

Метод золотого сечения рассчитан на детерминированные задачи. В стохастических задачах из-за ошибок эксперимента можно неправильно определить соотношения между значениями функций в точках; тогда дальнейшие итерации пойдут по ложному пути. Поэтому если различия функций в выбранных точках стали того же порядка, что и ошибки эксперимента, то итерации надо прекращать. Поскольку вблизи минимума чаще всего $\delta\Phi \sim (\delta x)^2$, то небольшая погрешность функции приводит к появлению довольно большой области неопределенности $\delta x \sim \sqrt{\delta\Phi}$.

3. Метод парабол. Метод золотого сечения надежный, но медленный. Если $\Phi(x)$ дифференцируема, то можно построить гораздо более быстрые методы, основанные на решении уравнения $\Phi'(x) = 0$. Напомним, что корень $\bar{x}$ этого уравнения является точкой минимума, если $\Phi''(\bar{x}) > 0$, и точкой максимума при $\Phi''(\bar{x}) < 0$.

На практике часто $\Phi(x)$ имеет и первую производную и вторую. Поэтому для нахождения нулей первой производной применяют метод линеаризации, что приводит к такому итерационному процессу:

$$x_{s+1} = x_s - \frac{\Phi'(x_s)}{\Phi''(x_s)}; \tag{11}$$

в простейших задачах нулевое приближение можно выбрать графически. Формулу (11) можно получить несколько иным способом. Разложим $\Phi(x)$ в точке x_s по формуле Тейлора, ограничившись

*) См. предыдущую сноску.

тремя членами, т. е. аппроксимируем кривую параболой

$$\Phi(x) \approx \Phi(x_s) + (x - x_s)\Phi'(x_s) + \frac{1}{2}(x - x_s)^2 \Phi''(x_s);$$

минимум этой параболы достигается в точке, определяемой формулой (11). Итерационный процесс (11) является ньютоновским; вблизи простого корня уравнения $\Phi'(x) = 0$, т. е. вблизи экстремума с ненулевой второй производной, он сходится квадратично. Если же $\Phi''(\bar{x}) = 0$, то сходимость в достаточно малой окрестности экстремума есть, но она более медленная — линейная.

Обычно для первой и тем более второй производной получаются очень громоздкие выражения. Поэтому выгоднее заменить их конечно-разностными аппроксимациями. Наиболее часто берут симметричные разности (3.6)—(3.7) с постоянным шагом, что приводит к формуле

$$x_{s+1} = x_s - \frac{h}{2} \frac{\Phi(x_s + h) - \Phi(x_s - h)}{\Phi(x_s + h) - 2\Phi(x_s) + \Phi(x_s - h)}. \qquad (12)$$

Это эквивалентно замене кривой на интерполяционную параболу, построенную по трем точкам $x_s - h$, x_s, $x_s + h$. Обычно выбирают вспомогательный шаг $h \approx 0{,}1 - 0{,}01$ при ручных расчетах с небольшим числом знаков и $h \approx 0{,}01 - 0{,}001$ при расчетах на ЭВМ; тогда характер сходимости вблизи экстремума вплоть до расстояний $\sim h^2$ практически не отличается от квадратичного. Формула (12) наиболее часто употребляется в практических расчетах.

Этот способ кажется неэкономным, ибо на каждой итерации надо вычислять три значения функции. Построение параболы по трем последовательным итерациям, как это делалось в методе парабол при нахождении корней многочлена, дает

$$2x_{s+1} = x_s + x_{s-1} - \frac{\Phi(x_s,\ x_{s-1})}{\Phi(x_s,\ x_{s-1},\ x_{s-2})} \qquad (13)$$

и требует только одного вычисления функции за итерацию. Однако ранее уже отмечалось, что такая замена производных разделенными разностями уменьшает скорость сходимости. Можно показать, используя описанную в главе V, § 2, п. 7 технику, что вблизи невырожденного минимума

$$|x_{s+1} - \bar{x}| \approx \left| \frac{\Phi'''(\bar{x})}{6\Phi''(\bar{x})} \right|^{0{,}325} |x_s - \bar{x}|^{1{,}325}. \qquad (14)$$

Во-первых, отсюда видно, что $|x_{s+1} - \bar{x}| < |x_s - \bar{x}|$, только если выполнено условие $|x_s - \bar{x}| < |6\Phi''/\Phi'''|$; это приблизительно показывает размеры окрестности корня, в которой итерации сходятся. Эта окрестность может быть небольшой, если $\Phi'''(\bar{x})$ велика.

Во-вторых, асимптотическая скорость сходимости определяется показателем степени при $|x_s - \bar{x}|$ в правой части соотношения (14). Этот показатель невелик; поэтому сходимость настолько медленна, что три итерации по этой формуле только немного сильней уменьшают погрешность, чем одна итерация по формуле (12). А поскольку формула (13) недостаточно испытана на практике, то нет уверенности, что она окажется лучше.

Заметим, что во всех вариантах метода парабол для успешной работы необходимы «кухонные» поправки к алгоритму. В ходе вычислений надо проверять, движемся ли мы к минимуму: вторая разность, стоящая в знаменателе формулы (12), или вторая производная в знаменателе формулы (11) должна быть положительной. Если она отрицательна, то итерации сходятся к максимуму, и надо сделать какой-то шаг в обратном направлении, причем достаточно большой.

Вычислив новое приближение, надо обязательно проверить, уменьшилась ли функция. Если оказалось, что

$$\Phi(x_{s+1}) > \Phi(x_s),$$

то значение x_{s+1} нельзя использовать и надо просто сделать от точки x_s какой-то шаг в сторону убывания функции. Обычно делают шаг величиной $\tau(x_{s+1} - x_s)$ с $\tau = 1/2$ и проверяют условие убывания функции; если оно снова не выполнено, то уменьшают τ вдвое и делают шаг опять из точки x_s, и так до тех пор, пока не добьются убывания функции.

Фактическая скорость работы программы очень сильно зависит от того, насколько тщательно обдуманы эти поправки к алгоритму.

Если функция имеет несколько локальных минимумов, то итерационный метод может сойтись к любому из них. Удалять найденные минимумы можно только в том случае, когда мы располагаем явным выражением для $\Phi'(x)$ и решаем не исходную задачу (1), а уравнение $\Phi'(x) = 0$; тогда удаляют уже найденные корни этого уравнения при помощи техники, описанной в главе V.

Если так сделать не удается, то выбирают несколько начальных приближений в разных участках отрезка $[a, b]$, и из каждого начального приближения проводят какой-нибудь итерационный процесс поиска минимума. Некоторые из этих итерационных процессов могут сходиться к одному и тому же локальному минимуму, а некоторые — к другим. Остается сравнить найденные локальные минимумы между собой и выбрать наименьший (если это требуется по условиям задачи).

Описанный способ не дает гарантии того, что будут найдены все минимумы (и тем самым, что будет найден абсолютный минимум). Но для недостаточно изученной функции такой гарантии не дают никакие способы.

4. Стохастические задачи. Опишем один алгоритм, рассчитанный на стохастические задачи. Он основан на предположении, что ошибки определения функции $\Phi(x)$ имеют статистическую природу, т. е. они целиком случайны, а систематической погрешности нет. Тогда можно определить минимум со сколь угодно высокой точностью (фактически игнорируя область неопределенности $\delta x \sim \sqrt{\delta\Phi}$), если воспользоваться таким итерационным

процессом:

$$x_{n+1} = x_n - \frac{a_n}{b_n}[\Phi(x_n + b_n) - \Phi(x_n - b_n)], \tag{15}$$

где a_n, b_n — последовательности положительных чисел, удовлетворяющие следующим условиям:

$$a_n,\ b_n \underset{n\to\infty}{\to} 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} \left(\frac{a_n}{b_n}\right)^2 < \infty. \tag{16}$$

При выполнении этих условий $x_n \to \bar{x}$ с вероятностью единица при $n \to \infty$ (напомним, что стремление с вероятностью единица означает «почти всегда стремится», а не «обязательно стремится»). Условиям (16) удовлетворяют, например, $a_n = 1/n$ и $b_n = n^{-1/3}$.

Этот алгоритм является обобщением алгоритма Роббинса — Монро, описанного в главе V, на задачи поиска минимума. Он сходится весьма медленно, ибо изменение аргумента за шаг равно $|x_{n+1} - x_n| \approx 2a_n |\Phi'(x_n)|$, а величины a_n убывают очень медленно, как видно из второго условия (16). Поэтому применять этот алгоритм к детерминированным задачам невыгодно.

§ 2. Минимум функции многих переменных

1. Рельеф функции. Основные трудности многомерного случая удобно рассмотреть на примере функции двух переменных $\Phi(x, y)$. Она описывает некоторую поверхность в трехмерном пространстве с координатами x, y, Φ. Задача $\Phi(x, y) = \min$ означает поиск низшей точки этой поверхности.

Как в топографии, изобразим рельеф этой поверхности линиями уровня. Проведем равноотстоящие плоскости $\Phi = \text{const}$ и найдем линии их пересечения с поверхностью $\Phi(x, y)$; проекции этих линий на плоскость x, y называют линиями уровня. Направление убывания функции будем указывать штрихами, рисуемыми около линий уровня. Полученная картина напоминает топографическое изображение рельефа горизонталями. По виду линий уровня условно выделим три типа рельефа: котловинный, овражный и неупорядоченный.

При *котловинном* рельефе линии уровня похожи на эллипсы (рис. 37, *а*). В малой окрестности невырожденного минимума рельеф функции котловинный. В самом деле, точка минимума гладкой функции определяется необходимыми условиями

$$\frac{\partial \Phi}{\partial x} = \frac{\partial \Phi}{\partial y} = 0, \tag{17}$$

и разложение функции по формуле Тейлора вблизи минимума

имеет вид

$$\Phi(x, y) = \Phi(\bar{x}, \bar{y}) + \frac{1}{2}(\Delta x)^2 \Phi_{xx} + \\ + \Delta x \Delta y \Phi_{xy} + \frac{1}{2}(\Delta y)^2 \Phi_{yy} + \dots, \qquad (18)$$

причем квадратичная форма (18) — положительно определенная *), иначе эта точка не была бы невырожденным минимумом. А линии уровня знакоопределенной квадратичной формы — это эллипсы.

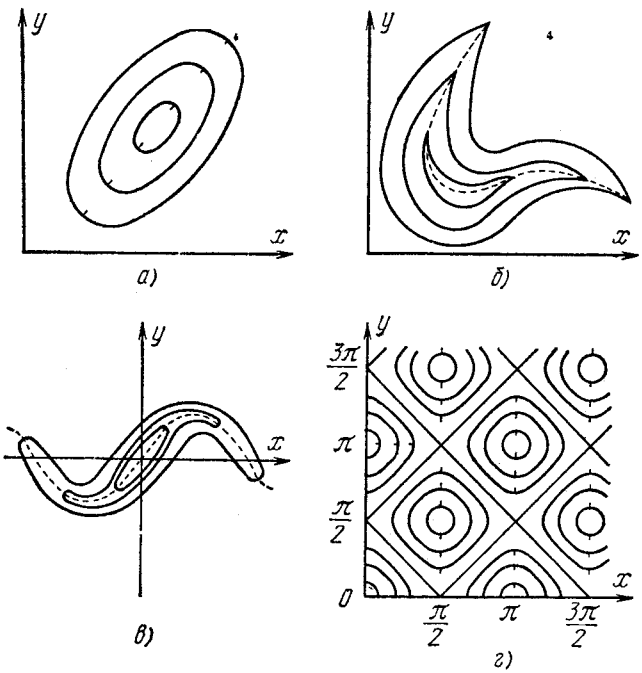


Рис. 37.

Случай, когда все вторые производные равны в этой точке нулю и минимум определяется более высокими производными, по существу ничего нового не дает, и мы не будем его специально рассматривать (линии уровня вместо эллипсов будут похожими на них кривыми четвертого порядка).

Отметим, что условию (17) удовлетворяют также точки максимумов и седловые точки. Но в точках максимумов квадратичная

*) Квадратичная форма $\sum_{i,\ k} a_{ik} z_i z_k$ называется положительно определенной, если при любых z_i (за исключением обращающихся одновременно в нуль) она положительна.

форма (18) отрицательно определенная, а в седловинах она знакопеременна.

Вблизи минимума функция мало меняется при заметных изменениях переменных. Поэтому даже если мы не очень точно определим те значения переменных, которые должны минимизировать функцию, то само значение функции при этом обычно будет мало отличаться от минимального.

Рассмотрим *овражный* тип рельефа. Если линии уровня кусочно-гладкие, то выделим на каждой из них точку излома. Геометрическое место точек излома назовем *истинным оврагом*, если угол направлен в сторону возрастания функции, и *гребнем* — если в сторону убывания (рис. 37, *б*). Чаще линии уровня всюду гладкие, но на них имеются участки с большой кривизной; геометрические места точек с наибольшей кривизной назовем *разрешимыми* оврагами или гребнями (рис. 37, *в*). Например, рельеф функции

$$\Phi(x,\ y) = 10\,(y - \sin x)^2 + 0{,}1\,x^2, \tag{19}$$

изображенный на этом рисунке, имеет ярко выраженный извилистый разрешимый овраг, «дно» которого — синусоида, а низшая точка — начало координат.

В физических задачах овражный рельеф указывает на то, что вычислитель не учел какую-то закономерность, имеющую вид связи между переменными. Обнаружение и явный учет этой закономерности облегчают решение математической задачи. Так, если в примере (19) ввести новые переменные $\xi = x$, $\eta = y - \sin x$, то рельеф становится котловинным.

Неупорядоченный тип рельефа (рис. 37, *г*) характеризуется наличием многих максимумов, минимумов и седловин. Примером может служить функция

$$\Phi(x,\ y) = (1 + \sin^2 x)\,(1 + \sin^2 y), \tag{20}$$

рельеф которой изображен на этом рисунке; она имеет минимумы в точках с координатами $\bar{x}_k = \pi k$, $\bar{y}_l = \pi l$ и максимумы в точках, сдвинутых относительно минимумов на $\pi/2$ по каждой координате.

Все эффективные методы поиска минимума сводятся к построению траекторий, вдоль которых функция убывает; разные методы отличаются способами построения таких траекторий. Метод, приспособленный к одному типу рельефа, может оказаться плохим на рельефе другого типа.

2. Спуск по координатам. Казалось бы, для нахождения минимума достаточно решить систему уравнений типа (17) методом линеаризации или простых итераций и отбросить те решения, которые являются седловинами или максимумами. Однако в реальных задачах минимизации эти методы обычно сходятся в настолько малой окрестности минимума, что выбрать подходящее

нулевое приближение далеко не всегда удается. Проще и эффективнее провести спуск по координатам. Изложим этот метод на примере функции трех переменных $\Phi(x, y, z)$.

Выберем нулевое приближение x_0, y_0, z_0. Фиксируем значения двух координат $y=y_0$, $z=z_0$. Тогда функция будет зависеть только от одной переменной x; обозначим ее через $f_1(x) = \Phi(x, y_0, z_0)$. Используя описанные в § 1 методы, найдем минимум функции одной переменной $f_1(x)$ и обозначим его через x_1. Мы сделали шаг из точки (x_0, y_0, z_0) в точку (x_1, y_0, z_0) по направлению, параллельному оси x; на этом шаге значение функции уменьшилось.

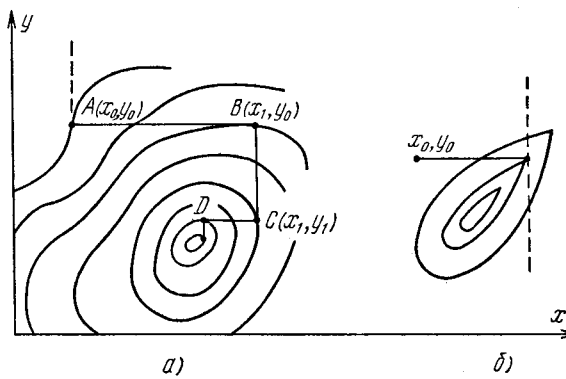


Рис. 38.

Затем из новой точки сделаем спуск по направлению, параллельному оси y, т. е. рассмотрим $f_2(y) = \Phi(x_1, y, z_0)$, найдем ее минимум и обозначим его через y_1. Второй шаг приводит нас в точку (x_1, y_1, z_0). Из этой точки делаем третий шаг — спуск параллельно оси z и находим минимум функции $f_3(z) = \Phi(x_1, y_1, z)$. Приход в точку (x_1, y_1, z_1) завершает *цикл спусков*.

Будем повторять циклы. На каждом спуске функция не возрастает, и при этом значения функции ограничены снизу ее значением в минимуме $\overline{\Phi} = \Phi(\bar{x}, \bar{y}, \bar{z})$. Следовательно, итерации сходятся к некоторому пределу $\tilde{\Phi} \geqslant \overline{\Phi}$. Будет ли здесь иметь место равенство, т. е. сойдутся ли спуски к минимуму и как быстро?

Это зависит от функции и выбора нулевого приближения. На примере функции двух переменных легко убедиться, что существуют случаи сходимости спуска по координатам к искомому минимуму и случаи, когда этот спуск к минимуму не сходится.

В самом деле, рассмотрим геометрическую трактовку спуска по координатам (рис. 38). Будем двигаться по выбранному направлению, т. е. по некоторой прямой в плоскости x, y.

В тех участках, где прямая пересекает линии уровня, мы при движении переходим от одной линии уровня к другой, так что при этом движении функция меняется (возрастает или убывает, в зависимости от направления движения). Только в той точке, где данная прямая касается линии уровня (рис. 38, *а*), функция имеет экстремум вдоль этого направления. Найдя такую точку, мы завершаем в ней спуск по первому направлению, и должны начать спуск по второму направлению (поскольку направления мы сейчас выбираем параллельно координатным осям, то второе направление перпендикулярно первому).

Пусть линии уровня образуют истинный овраг. Тогда возможен случай (рис. 38, *б*), когда спуск по одной координате приводит нас на «дно» оврага, а любое движение по следующей координате (пунктирная линия) ведет нас на подъем. Никакой дальнейший спуск по координатам невозможен, хотя минимум еще не достигнут; процесс спуска по координатам в данном случае не сходится к минимуму.

Наоборот, если функция достаточно гладкая, то в некоторой окрестности минимума процесс спуска по координатам сходится к этому минимуму. Пусть функция имеет непрерывные вторые производные, а ее минимум не вырожден. Для простоты опять рассмотрим функцию двух переменных $\Phi(x, y)$. Выберем некоторое нулевое приближение x_0, y_0 и проведем линию уровня через эту точку. Пусть в области G, ограниченной этой линией уровня, выполняются неравенства, означающие положительную определенность квадратичной формы (18):

$$\Phi_{xx} \geqslant a > 0,\ \Phi_{yy} \geqslant b > 0,\ |\Phi_{xy}| \leqslant c,\ ab > c^2. \qquad (21)$$

Докажем, что тогда спуск по координатам из данного нулевого приближения сходится к минимуму, причем линейно.

Значения функции вдоль траектории спуска не возрастают; поэтому траектория не может выйти из области G, и неравенства (21) будут выполняться на всех шагах. Рассмотрим один из циклов, начинающийся в точке A (рис. 38, *а*). Предыдущий цикл окончился поиском минимума по направлению y, следовательно, $(\Phi_y)_A = 0$ и $|\Phi_x|_A = \xi_1 \neq 0$. Первый шаг нового цикла спускает нас по направлению x в точку B, в которой $\Phi_x = 0$ и $|\Phi_y| = \eta \neq 0$. Поскольку вторые производные непрерывны, можно применить теорему о среднем; получим

$$\xi_1 = |(\Phi_x)_A - (\Phi_x)_B| = |\Phi_{xx}|\,\rho_{AB} \geqslant a\,\rho_{AB},$$
$$\eta = |(\Phi_y)_A - (\Phi_y)_B| = |\Phi_{xy}|\,\rho_{AB} \leqslant c\,\rho_{AB},$$

где через ρ обозначены расстояния между точками. Отсюда получаем $c\,\xi_1 \geqslant a\,\eta$. Выполним второй шаг цикла — спуск по направлению y в точку C, после которого $(\Phi_y)_C = 0$ и $|\Phi_x|_C = \xi_2$.

Аналогичные рассуждения дают соотношение $c\eta \geqslant b\xi_2$. Объединяя эти неравенства, найдем

$$\xi_2 \leqslant q\xi_1, \quad q = \frac{c^2}{ab}, \quad 0 < q < 1.$$

Следовательно, за один цикл Φ_x уменьшается в q раз: то же справедливо для Φ_y, если рассмотреть цикл, сдвинутый на один шаг, т. е. начинающийся в точке B и кончающийся в точке D.

Значит, когда число циклов $n \to \infty$, то все первые производные линейно стремятся к нулю:

$$|\Phi_x|_n \leqslant q^n |\Phi_x|_0 \to 0 \text{ и } |\Phi_y|_n \sim q^n \to 0.$$

Первые производные одновременно обращаются в нуль в точке минимума и вблизи него являются линейными однородными функциями приращений координат. Поэтому координаты точек спуска линейно стремятся к координатам точки минимума, т. е. в данном случае спуск по координатам сходится, причем линейно.

Случай (21) заведомо реализуется в достаточно малой окрестности невырожденного минимума, ибо эти условия эквивалентны требованию положительной определенности квадратичной формы (18). Таким образом, вблизи невырожденного минимума достаточно гладкой функции спуск по координатам линейно сходится к минимуму. В частности, для квадратичной функции этот метод сходится при любом нулевом приближении.

Фактическая скорость сходимости будет неплохой при малых q, когда линии уровня близки к эллипсам, оси которых параллельны осям координат. Для эллипсов, сильно вытянутых под значительным углом к осям координат, величина $q \approx 1$ и сходимость очень медленная.

Если сходимость медленная, но траектория уже попала в близкую окрестность минимума, то итерации можно уточнять процессом Эйткена; разумеется, при этом надо брать в качестве исходных значения не на трех последних спусках, а на трех *циклах* спусков (т. е. не точки A, B, C, а точки B, D и третья точка, которой нет на рис. 38, *а*).

Разрешимый овраг напоминает сильно вытянутую котловину (см. рис. 38, *б*). При попадании траектории спуска в такой овраг сходимость становится настолько медленной, что расчет практически невозможно вести. Отметим, что в стохастических задачах наличие ошибок эквивалентно превращению истинных оврагов и гребней в разрешимые; расчет при этом можно продолжать, хотя практическая ценность такого расчета невелика: сходимость очень медленная.

Метод спуска по координатам несложен и легко программируется на ЭВМ. Но сходится он медленно, а при наличии оврагов — очень плохо. Поэтому его используют в качестве первой попытки при нахождении минимума.

Пример. Рассмотрим квадратичную функцию $\Phi(x, y) = x^2 + y^2 + xy$ и выберем нулевое приближение $x_0 = 1$, $y_0 = 2$. Выполняя вычисления, получим

$$x_1 = -1,\ y_1 = {}^1/_2;\ x_2 = -{}^1/_4,\ y_2 = {}^1/_8;\ x_3 = -{}^1/_{16},\ y_3 = {}^1/_{32}.$$

Уточнение по Эйткену дает $\tilde{x} = \tilde{y} = 0$, т. е. точное положение минимума (заметим, что делать уточнение с использованием нулевого приближения нельзя; читателям предлагается объяснить, почему).

3. **Наискорейший спуск.** Спускаться можно не только параллельно осям координат. Вдоль любой прямой $\boldsymbol{r} = \boldsymbol{r}_0 + \boldsymbol{a}t$ функция зависит только от одной переменной, $\Phi(\boldsymbol{r}_0 + \boldsymbol{a}t) = \varphi(t)$, и минимум на этой прямой находится описанными в § 1 методами.

Наиболее известным является метод наискорейшего спуска, когда выбирается $\boldsymbol{a} = -(\operatorname{grad}\Phi)_{\boldsymbol{r}=\boldsymbol{r}_0}$, т. е. направление, в котором функция быстрее всего убывает при бесконечно малом движении из данной точки. Спуск по этому направлению до минимума определяет новое приближение $\boldsymbol{r}_1$. В этой точке снова определяется градиент и делается следующий спуск.

Однако этот метод значительно сложнее спуска по координатам, ибо требуется вычислять производные и градиент (это нередко делают конечно-разностными методами) и переходить к другим переменным. К тому же, по сходимости наискорейший спуск не лучше спуска по координатам. При попадании траектории в истинный овраг спуск прекращается, а в разрешимом овраге сильно замедляется.

Если функция является положительно определенной квадратичной функцией

$$\Phi(\boldsymbol{r}) = (\boldsymbol{r},\ A\boldsymbol{r}) + (\boldsymbol{b},\ \boldsymbol{r}) + c, \tag{22}$$

то формулы наискорейшего спуска приобретают несложный вид. Вдоль прямой $\boldsymbol{r} = \boldsymbol{r}_n + \boldsymbol{a}t$ функция (22) квадратично зависит от параметра t:

$$\varphi(t) \equiv \Phi(\boldsymbol{r}_n + \boldsymbol{a}t) = \Phi(\boldsymbol{r}_n) + (2A\boldsymbol{r}_n + \boldsymbol{b},\ \boldsymbol{a})\,t + (\boldsymbol{a},\ A\boldsymbol{a})\,t^2. \tag{23}$$

Из уравнения $(d\varphi/dt) = 0$ легко находим ее минимум

$$\bar{t} = -(2A\boldsymbol{r}_n + \boldsymbol{b},\ \boldsymbol{a})/2\,(\boldsymbol{a},\ A\boldsymbol{a}), \tag{24}$$

дающий нам следующую точку спуска:

$$\begin{gathered} \boldsymbol{r}_{n+1} = \boldsymbol{r}_n + \boldsymbol{a}\bar{t}, \\ \Phi(\boldsymbol{r}_{n+1}) = \Phi(\boldsymbol{r}_n) - \frac{(2A\boldsymbol{r}_n + \boldsymbol{b},\ \boldsymbol{a})^2}{4\,(\boldsymbol{a},\ A\boldsymbol{a})}. \end{gathered} \tag{25}$$

Направление наискорейшего спуска определяется градиентом квадратичной функции (22):

$$\boldsymbol{a} = -(\operatorname{grad}\Phi)_{\boldsymbol{r}_n} = -(2A\boldsymbol{r}_n + \boldsymbol{b}). \tag{26}$$

Подставляя это значение в формулы (24) — (25), получим окончательные выражения для вычисления последовательных спусков.

Если воспользоваться разложением всех движений по базису, состоящему из собственных векторов матрицы A, то можно доказать, что для квадратичной функции метод наискорейшего спуска линейно сходится, причем

$$|r_{n+1}-\bar{r}|\leqslant q|r_n-\bar{r}|, \text{ где } q=\frac{\lambda_{\max}-\lambda_{\min}}{\sqrt{\lambda_{\max}^2+\lambda_{\min}^2}}<1; \qquad (27)$$

здесь λ — собственные значения положительно определенной матрицы A (они вещественны и положительны). Если $\lambda_{\min}\ll\lambda_{\max}$, что соответствует сильно вытянутым эллипсам — линиям уровня, то $q\approx 1$ и сходимость может быть очень медленной. Есть такие начальные приближения (рис. 39), когда точно реализуется наихудшая возможная оценка, т. е. в (27) имеет место равенство.

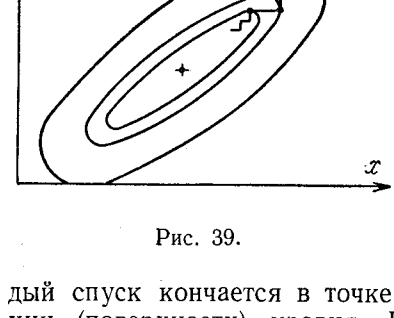


Рис. 39.

Причины нетрудно понять. Во-первых, в данной точке любую прямую, в том числе невыгодную для спуска, можно сделать направлением градиента, если специально подобрать изменение масштабов по осям. Во-вторых, каждый спуск кончается в точке, где его направление касается линии (поверхности) уровня. Градиент перпендикулярен поверхности уровня. Следовательно, в методе наискорейшего спуска каждый спуск перпендикулярен предыдущему. В двумерном случае это означает, что мы совершаем спуск по координатам, повернутым так, что одна ось параллельна градиенту в начальной точке.

Для улучшения метода наискорейшего спуска предлагают «кухонные» поправки к алгоритму — например, совершают по каждому направлению спуск не точно до минимума. Наиболее любопытным представляется такое видоизменение алгоритма. Будем делать по направлению, противоположному градиенту, только бесконечно малый шаг и после него вновь уточнять направление спуска. Это приводит к движению по кривой $r(t)$, являющейся решением системы обыкновенных дифференциальных уравнений:

$$\frac{dr}{dt}=-\operatorname{grad}\Phi(r(t)). \qquad (28)$$

Вдоль этой кривой $d\Phi/dt=(d\Phi/dr)(dr/dt)=-(\operatorname{grad}\Phi)^2<0$, т. е. функция убывает, и мы движемся к минимуму при $t\to+\infty$.

Уравнение (28) моделирует безынерционное движение материальной точки вниз по линии градиента. Можно построить и другие уравнения — например, дифференциальное уравнение второго порядка, моделирующее движение точки при наличии вязкого трения.

Однако от идеи метода еще далеко до надежного алгоритма. Фактически систему дифференциальных уравнений (28) надо численно интегрировать (см. главу VIII). Если интегрировать с большим шагом, то численное решение будет заметно отклоняться от линии градиента. А при интегрировании малым шагом сильно возрастает объем расчетов. Кроме того, если рельеф имеет извилистые овраги, то трудно ожидать хорошей сходимости этого метода.

Алгоритмы наискорейшего спуска и всех его видоизменений сейчас недостаточно отработаны. Поэтому метод наискорейшего спуска для сложных нелинейных задач с большим числом переменных ($m \gtrsim 5$) редко применяется, но в частных случаях он может оказаться полезным.

4. Метод оврагов. Рассмотрим задачу $\Phi(r) = \min$. Выберем произвольно точку ρ_0 и спустимся из нее (например, по координатам), делая не очень много шагов, т. е. не требуя высокой точности сходимости. Конечную точку спуска обозначим r_0. Если рельеф овражный, эта точка окажется вблизи дна оврага (рис. 40).

Теперь выберем другую точку ρ_1 не слишком далеко от первой. Из нее также сделаем спуск и попадем в некоторую точку r_1. Эта точка тоже лежит вблизи дна оврага. Проведем через точки r_0 и r_1 на дне оврага прямую — приблизительную линию дна оврага, передвинемся по этой линии в сторону убывания функции и выберем новую точку

$$\rho_2 = r_1 \pm (r_1 - r_0)\, h, \quad h = \text{const} > 0. \tag{29}$$

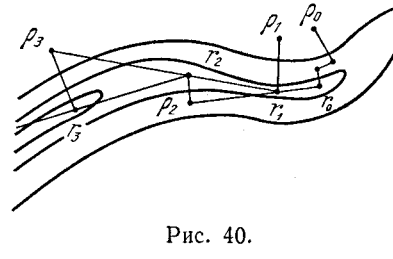


Рис. 40.

В формуле (29) выбирается плюс, если $\Phi(r_1) < \Phi(r_0)$, и минус в обратном случае, так что движение направлено в сторону понижения дна оврага. Величина h называется *овражным шагом* и для каждой функции подбирается в ходе расчета.

Дно оврага не является отрезком прямой, поэтому точка ρ_2 на самом деле лежит не на дне оврага, а на склоне. Из этой точки снова спустимся на дно и попадем в некоторую точку r_2. Затем соединим точки r_1 и r_2 прямой, наметим новую линию дна оврага и сделаем новый шаг по оврагу. Продолжим процесс до тех пор, пока значения функции на дне оврага, т. е. в точках r_n, убывают. В случае, когда

$$\Phi(r_{n+1}) > \Phi(r_n),$$

процесс надо прекратить и значение r_{n+1} не использовать.

Метод оврагов рассчитан на то, чтобы пройти вдоль оврага и выйти в котловину около минимума. В этой котловине значения минимума лучше уточнять другими методами.

Методом оврагов удается находить минимумы достаточно сложных функций от 5—10 переменных. Но этот метод довольно капризен. Для каждой функции приходится подбирать свой овражный шаг, визуально наблюдать за ходом расчета и вносить коррективы. Программирование этого метода на ЭВМ несложно.

5. Сопряженные направления. Методы наискорейшего спуска или спуска по координатам даже для квадратичной функции требуют бесконечного числа итераций. Однако можно построить такие направления спуска, что для квадратичной функции

$$\Phi(\boldsymbol{r}) = (\boldsymbol{r},\ A\boldsymbol{r}) + (\boldsymbol{b},\ \boldsymbol{r}) + c \tag{30}$$

(где $\boldsymbol{r}$ есть n-мерный вектор) с симметричной положительно определенной матрицей A процесс спуска сойдется точно к минимуму за конечное число шагов.

Положительно определенная матрица позволяет ввести норму вектора следующим образом:

$$\|\boldsymbol{x}\|^2 = (\boldsymbol{x},\ A\boldsymbol{x}) > 0 \quad \text{при } \boldsymbol{x} \neq 0. \tag{31}$$

Нетрудно проверить, что все аксиомы нормы при этом выполнены. Определение (31) означает, что под скалярным произведением двух векторов $\boldsymbol{x}$ и $\boldsymbol{y}$ теперь подразумевается величина $(\boldsymbol{x},\ A\boldsymbol{y})$. Векторы, ортогональные в смысле этого скалярного произведения

$$(\boldsymbol{x},\ A\boldsymbol{y}) = 0, \tag{32}$$

называют *сопряженными* (по отношению к данной матрице A). Ниже мы увидим, что поочередный спуск по сопряженным направлениям особенно выгоден при поиске минимума.

На этом основана большая группа методов: сопряженных градиентов, сопряженных направлений, параллельных касательных и другие. Для квадратичной функции они применяются с одинаковым успехом. На произвольные функции наиболее хорошо обобщается метод *сопряженных направлений*, у которого детали алгоритма тщательно отработаны; этот метод излагается в данном пункте.

а) Сначала рассмотрим, как применяется этот метод к квадратичной форме (30). Для этого нам потребуются некоторые свойства сопряженных векторов. Пусть имеется некоторая система попарно сопряженных векторов $\boldsymbol{x}_i$. Нормируем каждый из этих векторов в смысле нормы (31); тогда соотношения между ними примут вид

$$(\boldsymbol{x}_i,\ A\boldsymbol{x}_j) = \delta_{ij}. \tag{33}$$

Докажем, что взаимно сопряженные векторы линейно-независимы. Из равенства $\boldsymbol{x}_1 = \sum_{i=2} \alpha_i \boldsymbol{x}_i$ следует $(\boldsymbol{x}_1,\ A\boldsymbol{x}_1) = \sum_{i=2} \alpha_i (\boldsymbol{x}_1,\ A\boldsymbol{x}_i) = 0,$

что противоречит положительной определенности матрицы. Это противоречие доказывает наше утверждение. Значит, система n сопряженных векторов является базисом в n-мерном пространстве. Для данной матрицы имеется бесчисленное множество базисов, состоящих из взаимно сопряженных векторов.

Пусть мы нашли некоторый сопряженный базис $\boldsymbol{x}_i$, $1 \leqslant i \leqslant n$. Выберем произвольную точку $\boldsymbol{r}_0$. Любое движение из этой точки можно разложить по сопряженному базису

$$\boldsymbol{r} = \boldsymbol{r}_0 + \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i. \tag{34}$$

Подставляя это выражение в правую часть формулы (30), преобразуем ее с учетом сопряженности базиса (33) к следующему виду:

$$\Phi(\boldsymbol{r}) = \Phi(\boldsymbol{r}_0) + \sum_{i=1}^{n} [\alpha_i^2 + 2\alpha_i (\boldsymbol{x}_i,\ A\boldsymbol{r}_0) + \alpha_i (\boldsymbol{x}_i,\ \boldsymbol{b})]. \tag{35}$$

Последняя сумма состоит из членов, каждый из которых соответствует только одной компоненте суммы (34). Это означает, что движение по одному из сопряженных направлений $\boldsymbol{x}_i$ меняет только один член суммы (35), не затрагивая остальных.

Совершим из точки $\boldsymbol{r}_0$ поочередные спуски до минимума по каждому из сопряженных направлений $\boldsymbol{x}_i$. Каждый спуск минимизирует свой член суммы (35), так что *минимум квадратичной функции точно достигается после выполнения одного цикла спусков*, то есть за конечное число действий.

Поясним геометрический смысл сопряженного базиса. Если осями координат сделать главные оси эллипсоидов уровня квадратичной функции, то один цикл спусков по этим координатам приводит точно в минимум. Если перейти к некоторым аффинным координатам, то функция останется квадратичной, но коэффициенты квадратичной формы изменятся. Можно формально рассмотреть нашу квадратичную функцию с измененными коэффициентами как некоторую новую квадратичную форму в декартовых координатах и найти главные оси ее эллипсоидов. Положение этих главных осей в исходных аффинных координатах будет некоторой системой сопряженных направлений. Разный выбор аффинных координат естественно приводит к разным сопряженным базисам.

б) Сопряженный базис можно построить способом *параллельных касательных плоскостей*.

Пусть некоторая прямая параллельна вектору $\boldsymbol{x}$, а квадратичная функция достигает на этой прямой минимального значения в точке $\boldsymbol{r}_0$. Подставим уравнение этой прямой $\boldsymbol{r} = \boldsymbol{r}_0 + \alpha \boldsymbol{x}$ в выражение (30) и потребуем выполнения условия минимума

функции $\varphi(\alpha) \equiv \Phi(\boldsymbol{r}_0 + \alpha \boldsymbol{x})$ в точке $\boldsymbol{r} = \boldsymbol{r}_0$, т. е. при $\alpha = 0$. Для этого воспользуемся выражением (35), где в сумме оставим только один член:

$$\varphi(\alpha) = \Phi(\boldsymbol{r}_0) + \alpha^2 + \alpha(\boldsymbol{x},\ 2A\boldsymbol{r}_0 + \boldsymbol{b}),$$

и положим $(d\varphi/d\alpha)_{\alpha=0} = 0$. Отсюда следует уравнение, которому удовлетворяет точка минимума:

$$(\boldsymbol{x},\ 2A\boldsymbol{r}_0 + \boldsymbol{b}) = 0. \tag{36}$$

Пусть на какой-нибудь другой прямой, параллельной первой, функция принимает минимальное значение в точке $\boldsymbol{r}_1$; тогда аналогично найдем $(\boldsymbol{x},\ 2A\boldsymbol{r}_1 + \boldsymbol{b}) = 0$. Вычитая это равенство из (36), получим

$$(\boldsymbol{x},\ A(\boldsymbol{r}_1 - \boldsymbol{r}_0)) = 0. \tag{37}$$

Следовательно, *направление, соединяющее точки минимума на двух параллельных прямых, сопряжено направлению этих прямых.*

Таким образом, всегда можно построить вектор, сопряженный произвольному заданному вектору $\boldsymbol{x}$. Для этого достаточно провести две прямые, параллельные $\boldsymbol{x}$, и найти на каждой прямой минимум квадратичной формы (30). Вектор $\boldsymbol{r}_1 - \boldsymbol{r}_0$, соединяющий эти минимумы, сопряжен $\boldsymbol{x}$. Заметим, что прямая касается линии уровня в той точке, где функция на данной прямой принимает минимальное значение; с этим связано название способа.

Пусть имеются две параллельные m-мерные плоскости, порожденные системой сопряженных векторов $\boldsymbol{x}_i$, $1 \leqslant i \leqslant m < n$. Пусть квадратичная функция достигает своего минимального значения на этих плоскостях соответственно в точках $\boldsymbol{r}_0$ и $\boldsymbol{r}_1$. Аналогичными рассуждениями можно доказать, что вектор $\boldsymbol{r}_1 - \boldsymbol{r}_0$, соединяющий точки минимума, сопряжен всем векторам $\boldsymbol{x}_i$. Следовательно, если задана неполная система сопряженных векторов $\boldsymbol{x}_i$, то этим способом всегда можно построить вектор $\boldsymbol{r}_1 - \boldsymbol{r}_0$, сопряженный всем векторам этой системы.

Рассмотрим один цикл процесса построения сопряженного базиса. Пусть уже построен базис, в котором последние m векторов взаимно сопряжены, а первые $n - m$ векторов не сопряжены последним. Найдем минимум квадратичной функции (30) в какой-нибудь m-мерной плоскости, порожденной последними m векторами базиса. Поскольку эти векторы взаимно сопряжены, то для этого достаточно произвольно выбрать точку $\boldsymbol{r}_0$ и сделать из нее спуск поочередно по каждому из этих направлений (до минимума!). Точку минимума в этой плоскости обозначим через $\boldsymbol{r}_1$.

Теперь из точки $\boldsymbol{r}_1$ сделаем поочередный спуск по первым $n - m$ векторам базиса. Этот спуск выведет траекторию из первой плоскости и приведет ее в некоторую точку $\boldsymbol{r}_2$. Из точки $\boldsymbol{r}_2$

снова совершим по последним m направлениям спуск, который приведет в точку r_3. Этот спуск означает точное нахождение минимума во второй плоскости, параллельной первой плоскости. Следовательно, направление $r_3 - r_1$ сопряжено последним m векторам базиса.

Если одно из несопряженных направлений в базисе заменить направлением $r_3 - r_1$, то в новом базисе уже $m+1$ направление будет взаимно сопряжено.

Начнем расчет циклов с произвольного базиса; для него можно считать, что $m=1$. Описанный процесс за один цикл увеличивает на единицу число сопряженных векторов в базисе. Значит, за $n-1$ цикл все векторы базиса станут сопряженными, и следующий цикл приведет траекторию в точку минимума квадратичной функции (30).

в) Хотя понятие сопряженного базиса определено только для квадратичной функции, описанный выше процесс построен так, что его можно формально применять для произвольной функции. Разумеется, что при этом находить минимум вдоль направления надо методом парабол, не используя нигде формул, связанных с конкретным видом квадратичной функции (30).

В малой окрестности минимума приращение достаточно гладкой функции обычно представимо в виде симметричной положительно определенной квадратичной формы типа (18). Если бы это представление было точным, то метод сопряженных направлений сходился бы за конечное число шагов. Но представление приближенно, поэтому число шагов будет бесконечным; зато сходимость этого метода вблизи минимума будет квадратичной.

Благодаря квадратичной сходимости метод сопряженных направлений позволяет находить минимум с высокой точностью. Методы с линейной сходимостью обычно определяют экстремальные значения координат менее точно.

Замечание 1. Реально даже для квадратичной функции процесс не всегда укладывается в n циклов. Построение сопряженного базиса означает ортогонализацию в метрике, порожденной матрицей A. Ранее отмечалось, что в процессе ортогонализации теряется точность; при большом числе переменных погрешность настолько возрастает, что процесс приходится повторять.

Замечание 2. Теоретически безразлично, какое из несопряженных направлений выкинуть из базиса в конце цикла. Обычно выкидывают то направление, при спуске по которому на данном цикле функция изменилась менее всего. Поскольку для произвольной функции понятие сопряженности ввести нельзя, то направление наиболее слабого убывания выкидывают независимо от того, под каким номером оно стоит в базисе. Любопытно, что это оказывается выгодным даже для квадратичной функции, хотя

на основании этого критерия иногда можно выкинуть сопряженное направление, оставив несопряженные; зато уменьшается потеря точности при ортогонализации.

Замечание 3. Описанный выше цикл метода включает два спуска по сопряженным направлениям и один — по несопряженным. Более выгоден цикл, при котором сразу после нахождения нового сопряженного направления по нему делают спуск из точки r_3, приходя в некоторую точку r_4. Тогда спуск из r_2 в r_4 будет спуском в плоскости всех новых сопряженных направлений, т. е. его можно считать первой группой нового цикла спусков. Поэтому из точки r_4 сразу можно спускаться по несопряженным направлениям.

При этом новое направление ставят в базис на последнее место и выкидывают то направление, на котором функция слабее всего уменьшилась при спусках от точки r_1 до точки r_4. Наименее выгодным может оказаться и новое направление; тогда следующий цикл спусков будет сделан со старым базисом.

Метод сопряженных направлений является, по-видимому, наиболее эффективным методом спуска. Он неплохо работает и при вырожденном минимуме, и при разрешимых оврагах, и при наличии слабо наклонных участков рельефа — «плато», и при большом числе переменных — до двух десятков.

6. Случайный поиск. Методы спуска неполноценны на неупорядоченном рельефе. Если локальных экстремумов много, то спуск из одного нулевого приближения может сойтись только к одному из локальных минимумов, не обязательно абсолютному. Тогда для исследования задачи применяют случайный поиск.

Предполагают, что интересующий нас минимум (или все минимумы) лежит в некоторой замкнутой области; линейным преобразованием координат помещают ее внутрь единичного n-мерного куба. Выбирают в этом кубе N случайных точек способами, описанными в § 4 главы IV; если о расположении экстремумов заранее ничего не известно, то наилучшие результаты дают $ЛП_\tau$-последовательности точек.

Даже при миллионе пробных точек вероятность того, что хотя бы одна точка попадет в небольшую окрестность локального минимума, ничтожно мала. В самом деле, пусть диаметр котловины около минимума составляет 10% от пределов изменения каждой координаты. Тогда объем этой котловины составляет $0{,}1^n$ часть объема n-мерного куба. Уже при $n > 6$ ни одна точка в котловину не попадет.

Поэтому берут небольшое число точек $N \approx (5-20)\,n$ и каждую точку рассматривают как нулевое приближение. Из каждой точки совершают спуск, быстро попадая в ближайший овраг или котловину; когда шаги спуска сильно укорачиваются, его прекращают, не добиваясь высокой точности. Этого уже достаточно,

чтобы судить о величине функции в ближайшем локальном минимуме с удовлетворительной точностью.

Сравнивая (визуально или при помощи программы) окончательные значения функции на всех спусках между собой, можно изучить расположение локальных минимумов функции и сопоставить их величины. После этого можно отобрать нужные по смыслу задачи минимумы и провести в них дополнительные спуски для получения координат точек минимума с высокой точностью.

Обычно в прикладных задачах нужно в первую очередь добиться того, чтобы исследуемая функция приняла минимальное или почти минимальное значение. Но вблизи минимума значение функции слабо зависит от изменения координат. Зачем тогда нужно находить координаты точки минимума с высокой точностью? Оказывается, что это имеет не только теоретический, но и практический смысл.

Пусть, например, координаты — это размеры деталей механической конструкции, а минимизируемая функция есть мера качества конструкции. Если мы нашли минимум точно, то мы находимся в самом центре котловины около минимума. В этом случае вариации координат влияют на функцию слабее, чем в точках, расположенных ближе к краям котловины. А безопасные вариации координат имеют в данном примере смысл допусков на точность обработки деталей. Значит, при аккуратном вычислении координат минимума мы можем разрешить бо́льшие допуски, т. е. удешевить обработку деталей.

Метод случайного поиска зачастую позволяет найти все локальные минимумы функции от $10-20$ переменных со сложным рельефом. Он полезен и при исследовании функции с единственным минимумом; в этом случае можно обойтись заметно меньшим числом случайных точек. Недостаток метода в том, что надо заранее задать область, в которой выбираются случайные точки. Если мы зададим слишком широкую область, то ее труднее детально исследовать, а если выберем слишком узкую область, то многие локальные минимумы могут оказаться вне ее. Правда, положение несколько облегчается тем, что при спусках траектории могут выйти за пределы заданной области и сойтись к лежащим вне этой области минимумам.

§ 3. Минимум в ограниченной области

1. Формулировка задачи. Пусть в n-мерном векторном пространстве задана скалярная функция $\Phi(\boldsymbol{x})$. Рассмотрим задачу на минимум с дополнительными условиями двух типов:

$$\Phi(\boldsymbol{x}) = \min, \quad \varphi_i(\boldsymbol{x}) = 0, \quad 1 \leqslant i \leqslant m,$$
$$\psi_j(\boldsymbol{x}) \geqslant 0, \quad 1 \leqslant j \leqslant p. \tag{38}$$

Условия типа равенств выделяют в пространстве некоторую $(n-m)$-мерную поверхность; поэтому должно выполняться неравенство $m < n$. Условия типа неравенств выделяют n-мерную область, ограниченную гиперповерхностями $\psi_j(\boldsymbol{x}) = 0$; число таких условий

может быть произвольным. Следовательно, задача (38) есть поиск минимума функции n переменных в $(n-m)$-мерной области G.

Функция может достигать минимального значения как внутри области, так и на ее границе. Эта задача и особенно последний случай трудны для расчета. Вид дополнительных условий в любой реальной задаче не слишком прост, так что явно ввести в области G собственную $(n-m)$-мерную систему координат практически никогда не удается. Значит, при численном расчете мы вынуждены вести спуск не на $(n-m)$-мерной поверхности, а во всем n-мерном пространстве. Тогда даже если нулевое приближение лежит в области G, естественная траектория спуска сразу выходит из этой области; особенно сложно «заставить» траекторию идти вдоль границы области.

В математических задачах экономики поиск минимума при дополнительных условиях называют (в зависимости от типа функций) линейным, нелинейным и т. д. программированием.

2. Метод штрафных функций. Рассмотрим задачу на абсолютный минимум во всем n-мерном пространстве для такой вспомогательной функции:

$$F(\boldsymbol{x}) \equiv \Phi(\boldsymbol{x}) + \\ + \mu\left\{\sum_{i=1}^{m} \varphi_i^2(\boldsymbol{x}) + \sum_{j=1}^{p} \psi_j^2(\boldsymbol{x})\,[1 - \operatorname{sign} \psi_j(\boldsymbol{x})]\right\} = \min, \quad \mu > 0. \qquad (39)$$

Прибавляемые к $\Phi(\boldsymbol{x})$ члены взяты таким образом, что они обращаются в нуль, если дополнительные условия в (38) выполнены. Если же условия нарушены, то эти члены положительны, т. е. они увеличивают $F(\boldsymbol{x})$, причем тем больше, чем сильнее нарушены дополнительные условия. Это своеобразный штраф за нарушение условий.

Если коэффициент штрафа μ достаточно велик, то за границами области G функция $F(\boldsymbol{x})$ быстро возрастает. Значит, минимум $F(\boldsymbol{x})$ расположен или внутри области G, или снаружи вблизи ее границы. Если он лежит в области G, то он совпадает с минимумом $\Phi(\boldsymbol{x})$, ибо там дополнительные члены в условии (39) обращаются в нуль. Если же минимум $F(\boldsymbol{x})$ лежит снаружи, то минимум $\bar{\boldsymbol{x}}$ исходной функции лежит на границе; при разумных предположениях о свойствах функций $\Phi(\boldsymbol{x})$, $\varphi_i(\boldsymbol{x})$ и $\psi_j(\boldsymbol{x})$ доказано, что его отличие от минимума $\bar{\boldsymbol{x}}_\mu$ вспомогательной функции не превышает

$$|\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}_\mu| \leqslant \frac{\text{const}}{\mu}, \qquad (40)$$

где величина константы зависит от конкретных свойств функций (38). Поэтому если взять последовательность $\mu_k \to \infty$ и найти для нее минимумы $\bar{\boldsymbol{x}}_k$ вспомогательной функции $F(\boldsymbol{x};\ \mu_k)$, то $\bar{\boldsymbol{x}}_k \to \bar{\boldsymbol{x}}$.

Задачу (39) на безусловный экстремум удобнее всего решать методом случайного поиска со спуском по сопряженным направлениям: здесь естественно задана область, где надо выбирать случайные точки.

При малых значениях μ согласно оценке (40) точность может быть плохой. Но при большом μ благодаря дополнительным членам в (39) вблизи границы области появляются глубокие овраги и крутые откосы, так что методы спуска сходятся медленно. Полезен следующий прием, заметно ускоряющий сходимость.

Сначала берут небольшое μ_1 и легко находят соответствующий минимум $\bar{x}_1$. Затем берут большее значение μ_2, а значение $\bar{x}_1$ используют в качестве начального приближения для спуска; поэтому спуск будет не длинный, и новый минимум $\bar{x}_2$ определится быстро. Эту процедуру повторяют до тех пор, пока «штраф» — фигурная скобка в (39) — не станет достаточно малым. Тогда можно считать, что точка $\bar{x}_k$ близка к границе области G и хорошо аппроксимирует минимум $\bar{x}$.

Метод штрафных функций медленный и не слишком надежный. Он применим только при небольшом числе переменных $n \lesssim 10$. Но существенно более хороших методов для общей нелинейной задачи (38) пока нет. Перспективным кажется метод штрафных оценок, являющийся комбинацией описанного метода и метода неопределенных множителей Лагранжа; однако он еще мало изучен.

3. Линейное программирование. При оптимизации экономических планов возникают задачи на минимум линейной функции n переменных при наличии линейных дополнительных условий трех типов:

$$L(x) \equiv \sum_{i=1}^{n} c_i x_i = \min, \tag{41а}$$

$$x_i \geqslant 0, \quad 1 \leqslant i \leqslant n, \tag{41б}$$

$$\sum_{i=1}^{n} a_{ji} x_i = b_j, \quad 1 \leqslant j \leqslant m, \tag{41в}$$

$$\sum_{i=1}^{n} a_{ji} x_i \leqslant b_j, \quad m < j \leqslant M. \tag{41г}$$

Каждое из условий типа неравенств (41б) или (41г) определяет полупространство, ограниченное гиперплоскостью; все эти условия вместе определяют выпуклый n-мерный многогранник J, являющийся пересечением соответствующих полупространств. С математической точки зрения условия (41б) и (41г) однотипны; но по традиции их записывают указанным образом. Условия типа равенств (41в) выделяют из n-мерного пространства $(n-m)$-мерную

плоскость. Ее пересечение с областью J дает *выпуклый* $(n-m)$-*мерный многогранник* G; наша задача состоит в том, чтобы найти минимум линейной функции (41а) в этом многограннике G.

Примером такой задачи является распределение производства однотипной продукции по разным заводам. Пусть x_i — выпускаемое i-м заводом количество продукции (оно должно быть неотрицательным), c_i — себестоимостью одного изделия на этом заводе, a_{ji} при $j>m$ — расход сырья j-го вида и a_{ji} при $2\leqslant j\leqslant m$ — расход заработной платы и других аналогичных показателей j-го вида при выпуске единицы продукции на данном заводе. Положим $a_{1i}=1$; тогда b_1 будет суммарным выпуском продукции по всем заводам, b_j, $2\leqslant j\leqslant m$, — полной заработной платой и аналогичными данными по всей отрасли, суммы (41г) — расходом сырья по всем заводам, а L — себестоимостью общей продукции. Требуется, чтобы себестоимость продукции была минимальной, выпуск продукции, расход заработной платы и т. д. — заданными, а фонды сырья b_j, $m<j$, не перерасходовались. Нас интересует, как распределить неотрицательные плановые задания x_i по заводам так, чтобы удовлетворить всем этим требованиям.

Отметим терминологию, установившуюся в экономике. Вектор $\boldsymbol{x}$, удовлетворяющий всем дополнительным условиям, называют *планом*; если он, к тому же, соответствует вершине многогранника G, то *опорным планом*. Решение экстремальной задачи (41) называют *оптимальным планом*, столбцы прямоугольной матрицы A — *векторами условий*, а столбец $\boldsymbol{b}$ — *вектором ограничений*. В задачах экономики обычно все коэффициенты a, b, $c\geqslant 0$, хотя для последующего изложения это несущественно.

Многогранник условий G — выпуклый (он может быть и неограниченным). Поэтому внутри него линейная функция $L(\boldsymbol{x})$ не может достигать минимума. Ее минимум (если он существует) достигается обязательно в какой-нибудь вершине многогранника. При вырождении он может достигаться во всех точках ребра или даже p-мерной ограничивающей плоскости ($p<n-m$). Поэтому теоретически задача линейного программирования проста. Достаточно вычислить значения функции в конечном числе точек — в вершинах многогранника и найти среди этих значений наименьшее.

Сложность заключается в другом. Типичное в экономике число переменных — это сотни и даже тысячи. При этом число вершин многогранника G становится астрономическим. Для того чтобы оценить это число, рассмотрим способ нахождения вершин.

Находить вершины самого многогранника G неудобно. Лучше преобразовать задачу к канонической форме, не содержащей условий третьего типа. Для этого введем в качестве новых переменных невязки условий третьего типа:

$$x_i=b_{i+m-n}-\sum_{q=1}^{n}a_{i+m-n,\,q}\,x_q\geqslant 0,\quad n<i\leqslant N,\quad N=n+M-m. \tag{42}$$

Доопределим коэффициенты экстремальной задачи (41) следующим образом:

$$c_i=0,\quad a_{ji}=\delta_{j,\,i+M-N}\quad\text{при } n<i\leqslant N,\ 1\leqslant j\leqslant M. \tag{43}$$

Тогда задача линейного программирования примет *каноническую форму*:

$$L(x) \equiv \sum_{i=1}^{n} c_i x_i = \min, \tag{44а}$$

$$x_i \geqslant 0, \quad 1 \leqslant i \leqslant N, \tag{44б}$$

$$\sum_{i=1}^{N} a_{ji} x_i = b_j, \quad 1 \leqslant j \leqslant M \quad (M < N). \tag{44в}$$

Многогранник новых канонических условий образован пересечением новой $(N-M)$-мерной плоскости условий с первым координатным углом. Значит, все его вершины лежат на координатных гиперплоскостях, т. е. у каждой вершины часть координат — нули, а остальные координаты положительны.

Будем считать, что строки новой матрицы A линейно-независимы: в противном случае или одно условие лишнее, или система условий несовместна. Тогда ранг этой прямоугольной матрицы равен M, и среди ее столбцов найдется по крайней мере один набор из M линейно-независимых столбцов. Все линейно-независимые наборы столбцов матрицы A соответствуют точкам пересечения плоскости условий с координатными гиперплоскостями.

Чтобы найти вершину, возьмем один такой набор столбцов. Для удобства записи перенумеруем переменные так, чтобы первыми стояли столбцы, соответствующие этому набору (базису). Перепишем условия второго типа (44в) в следующем виде:

$$\sum_{i=1}^{M} a_{ji} x_i = b_j - \sum_{i=M+1}^{N} a_{ji} x_i, \quad 1 \leqslant j \leqslant M. \tag{45}$$

Обозначим через α_{ji}, $1 \leqslant j$, $i \leqslant M$, элементы матрицы, обратной к базисной квадратной матрице, стоящей в левой части системы (45). Приравнивая внебазисные координаты нулю и решая эту систему, получим координаты точки пересечения плоскости условий с координатной гиперплоскостью

$$\begin{aligned} x_i &= \sum_{j=1}^{M} \alpha_{ij} b_j, \quad 1 \leqslant i \leqslant M, \\ x_i &= 0, \qquad\qquad M < i \leqslant N. \end{aligned} \tag{46}$$

Если найденные координаты неотрицательны, точка пересечения принадлежит первому координатному углу, т. е. является вершиной многогранника канонических условий. Если хотя бы одно $x_i < 0$, эту точку надо отбросить и исследовать другой набор столбцов матрицы A. Если мы забракуем все точки, это

означает, что условия первого и второго рода образуют несовместную систему.

Различные столбцы матрицы A могут образовать C_N^M наборов. Поэтому в самом неблагоприятном случае ($M \approx {}^1/_2 N$) многогранник условий может иметь до $C_N^{N/2} \approx 2^N$ вершин. Если $N \sim 100$, то это число настолько велико, что простой перебор вершин невозможен. Нетрудно подсчитать, что для ЭВМ типа БЭСМ-6 простой перебор посилен только при $N \leqslant 15$.

4. Симплекс-метод позволяет найти решение задачи линейного программирования за гораздо меньшее число действий. Изложим идею метода.

Найдем какую-нибудь вершину многогранника и все ребра, выходящие из этой вершины. Пойдем вдоль того из ребер, по которому функция убывает. Придем в следующую вершину, найдем выходящие из нее ребра и повторим процесс. Когда мы придем в такую вершину, что вдоль всех выходящих из нее ребер функция возрастает, то минимум достигнут. Поскольку $L(\boldsymbol{x})$ — линейная функция, а многогранник условий выпуклый, то этот процесс всегда сходится к решению задачи, причем за конечное число шагов.

При канонической форме записи многогранника условий из каждой его вершины исходит $N - M$ ребер. Выбирая одно ребро, мы выбрасываем из рассмотрения вершины, лежащие на остальных траекториях. Следовательно, за k шагов мы рассматриваем $(N-M)^{-k}$-ю часть вершин, проходя мимо остальных. Нам надо найти искомую вершину среди C_N^M вершин многогранника. Приравнивая число вершин C_N^M величине $(N-M)^k$, получим, что минимум достигается примерно за $k \sim N$ шагов, т. е. достаточно быстро.

Выведем формулы шага. Первую вершину находим по формуле (46). Чтобы найти ребро, надо одну из внебазисных переменных x_l сделать положительной; тогда координаты точек ребра можно выразить через нее из (45) при помощи обратной матрицы

$$\tilde{x}_i = x_i - x_l \sum_{j=1}^{M} \alpha_{ij} a_{jl}, \qquad 1 \leqslant i \leqslant M, \qquad \tilde{x}_l = x_l; \tag{47}$$

остальные внебазисные координаты остаются равными нулю. Будем увеличивать x_l до тех пор, пока одна из базисных координат не обратится в нуль. Это будет при

$$\bar{x}_l = \min_{1 \leqslant i \leqslant M} \left(\frac{x_i}{S_{il}} \right), \qquad S_{il} = \sum_{j=1}^{M} \alpha_{ij} a_{jl} > 0; \tag{48}$$

минимум ищется только среди тех индексов i, для которых $S_{il} > 0$, ибо только эти координаты вдоль данного ребра уменьшаются и,

следовательно, могут обратиться в нуль. Если все суммы S_{il} при данном l отрицательны, то это ребро неограниченное и весь многогранник условий — тоже.

Подставляя найденное $\bar{x}_l$ в формулы (47), получим координаты новой вершины и вычислим в ней значение функции $L(\boldsymbol{x}) = L_l$. Поочередно меняя каждую внебазисную переменную, найдем все $N - M$ ребер, выходящих из исходной вершины и проводящих в смежные вершины. Сравним все значения функции в смежных вершинах L_l и выберем из них наименьшее. Если оно меньше, чем значение функции в исходной вершине L_0, то переместимся в наинизшую из новых вершин и повторим процесс. Если же $\min L_l \geqslant L_0$, то минимум уже достигнут в исходной вершине.

Для всех неограниченных ребер, исходящих из вершины, надо проверять знак производной функции

$$\Delta_l = \frac{dL}{dx_l} = c_l - \sum_{i=1}^{M} c_i \sum_{j=1}^{M} \alpha_{ij} a_{jl} = c_l - \sum_{i=1}^{M} c_i S_{il}. \qquad (49)$$

Если эта величина отрицательна, то задача линейного программирования вообще не имеет решения ($\min L(\boldsymbol{x}) = -\infty$). Если же она неотрицательна, то это ребро не ведет к минимуму, и оно нас не интересует.

Нетрудно оценить, что для выполнения всех шагов и получения минимума требуется примерно до $10N^2M^2$ арифметических действий. Это уже приемлемо для крупных современных ЭВМ.

Симплекс-метод является примером высоко специализированного метода. Он пригоден только для нахождения минимума линейной функции в многомерном выпуклом многограннике определенного вида — симплексе. Зато он позволяет решать задачи с огромным числом переменных.

5. Регуляризация линейного программирования. Задача линейного программирования часто оказывается плохо обусловленной. Так, себестоимость единицы продукции или норм расхода сырья на разных заводах не должна сильно отличаться. Поэтому даже заметное перераспределение заказов между заводами слабо влияет на суммарную стоимость продукции. Соответственно малая вариация суммарной стоимости приводит к большой вариации распределения заказов.

По тем же причинам небольшое изменение себестоимости или других показателей на отдельных заводах сильно меняет оптимальный план, так что решение очень чувствительно к вариациям коэффициентов. А сами эти коэффициенты не вполне точно известны. Поэтому на практике задача (41) нередко оказывается настолько плохо обусловленной, что не удается даже проверить, совместна ли система дополнительных условий, т. е. может ли существовать решение поставленной задачи.

Для регуляризации задачи линейного программирования воспользуемся тем же способом, что и для решения плохо обусловленных линейных систем (см. главу V, § 1). Будем искать *нормальное* решение $\boldsymbol{x}$, т. е. наименее уклоняющееся от некоторого заданного вектора $\boldsymbol{x}_0$. Обычно в качестве $\boldsymbol{x}_0$ берут ранее составленный план. Тогда регуляризованное решение будет почти не уступать оптимальному по величине $L(\boldsymbol{x})$ и в то же время мало отличаться от старого плана, так что перестройка планов будет небольшой.

Возьмем исходную задачу в канонической форме (44) и рассмотрим формулы регуляризации. Надо минимизировать положительную функцию $L(\boldsymbol{x})$ или, что то же самое, функцию $L^2(\boldsymbol{x})$. Дополнительным условием служит система уравнений $A\boldsymbol{x}=\boldsymbol{b}$ с прямоугольной матрицей. Поскольку коэффициенты системы известны не точно, то достаточно найти приближенное решение. Тогда требование приближенного соблюдения этих условий эквивалентно введению штрафной функции $\mu\|A\boldsymbol{x}-\boldsymbol{b}\|^2$, т. е. постановке следующей задачи:

$$L^2(\boldsymbol{x})+\mu\|A\boldsymbol{x}-\boldsymbol{b}\|^2=\min, \quad \mu>0. \tag{50}$$

Здесь норму будем определять, как $\|\boldsymbol{y}\|^2=(\boldsymbol{y},\boldsymbol{y})$; тогда все минимизируемые выражения будут квадратичными функциями $\boldsymbol{x}$, что облегчит вычисления. Заметим, что пока мы не учитывали требования неотрицательности компонент решения.

Условием близости решения к заданному вектору можно считать малость величины $\|\boldsymbol{x}-\boldsymbol{x}_0\|$ или, в более общем виде, малость величины

$$\Omega[\boldsymbol{x}]=\sum_{i=1}^{N} p_i(x_i-x_{i0})^2, \quad p_i>0. \tag{51}$$

Эту величину также можно считать штрафом и прибавлять в качестве дополнительного слагаемого в левую часть (50); тогда получаем регуляризованную задачу

$$M[\boldsymbol{x}]=L^2(\boldsymbol{x})+\mu\|A\boldsymbol{x}-\boldsymbol{b}\|^2+\lambda\Omega[\boldsymbol{x}]=\min, \quad \mu,\lambda>0. \tag{52}$$

Отклонение регуляризованного решения от $\boldsymbol{x}_0$ не должно быть большим. Но $\boldsymbol{x}_0$ есть некоторый план; следовательно, его компоненты неотрицательны. Значит, если у решения, найденного из условия (52), и будут отрицательные компоненты, то небольшие по абсолютной величине, что в итоге несущественно. Поэтому при решении регуляризованной задачи (52) условия неотрицательности (44б) обычно можно не принимать во внимание.

Величина $M[\boldsymbol{x}]$ является квадратичной формой, так что нахождение ее минимума (путем обычного дифференцирования по координатам) сводится к решению системы линейных уравнений. Поскольку задача регуляризована, то полученная линейная система

будет хорошо обусловлена; тогда ее решение даже при большом числе неизвестных $N \sim 200$ легко вычислить методом исключения Гаусса.

Более сложен вопрос о выборе параметров регуляризации μ и λ. Величину μ подбирают так, чтобы для найденного регуляризованного решения выполнялось условие $\|A\boldsymbol{x} - \boldsymbol{b}\| \approx \|\delta \boldsymbol{b}\|$, где $\delta \boldsymbol{b}$ — допустимая погрешность вектора $\boldsymbol{b}$, связанная с тем, что его компоненты и коэффициенты матрицы A известны неточно. Аналогичным образом величину λ связывают с погрешностями коэффициентов c_i и с допустимыми отклонениями функции $L(\boldsymbol{x})$ от своего минимального значения.

При численном решении задачи (52) приходится находить серию регуляризованных решений, соответствующих разным значениям параметров μ и λ, и выбирать оптимальные параметры. Несмотря на это, общий объем вычислений в описанном методе, по-видимому, не больше, чем в симплекс-методе для нерегуляризованной задачи (44).

§ 4. Минимизация функционала

1. Задачи на минимум функционала. Если каждой функции $y(x)$ из некоторого множества функций Y сопоставлено число $\Phi[y(x)]$, то говорят, что на множестве Y задан функционал. Задача минимизации функционала формулируется так: *найти функцию $\bar{y}(x) \in Y$, на которой функционал достигает своей точной нижней грани на этом множестве:*

$$\Phi[\bar{y}(x)] = \inf \Phi[y(x)], \qquad y(x),\ \bar{y}(x) \in Y. \tag{53}$$

Иногда эту задачу называют минимизацией функционала *по аргументу*, а просто минимизацией называют нахождение числа $\bar{\Phi} = \inf \Phi[y(x)]$, когда не требуется определять функцию, минимизирующую этот функционал.

Не всякий функционал и не на всяком множестве имеет минимум. Например, решения задачи (53) не существует, если функционал не ограничен снизу на заданном множестве: решения может также не существовать, если множество не компактно в себе, или функционал разрывен и т. д. (хотя условия непрерывности или компактности не являются, вообще говоря, необходимыми). Но мы не исследуем постановки задач и дальше будем предполагать, что конкретные решаемые нами задачи типа (53) корректно поставлены.

Дадим несколько примеров задач на минимум функционала. Пусть требуется решить операторное уравнение

$$Ay(x) = f(x), \qquad a \leqslant x \leqslant b. \tag{54}$$

Составим функционал

$$\Phi[y(x)]=\int_a^b\{Ay(x)-f(x)\}^2\rho(x)\,dx,\quad \rho(x)>0. \tag{55}$$

Очевидно, он равен нулю при $Ay(x)=f(x)$ и положителен, если $Ay(x)\neq f(x)$ на сколь угодно малом, но конечном интервале Δx. Таким образом, найдя функцию $\bar{y}(x)$, на которой функционал (55) достигает своего абсолютного минимума, мы получим решение уравнения (54). Заметим, что этот функционал ограничен снизу на любом множестве функций и непрерывно зависит от $Ay(x)$. Описанный способ решения операторных уравнений называется *методом наименьших квадратов*.

Если задача (54) некорректно поставлена (например, неустойчива по правой части), то наиболее употребительным общим методом регуляризации является замена исходной задачи на задачу минимизации функционала А. Н. Тихонова:

$$M[y(x),\alpha]=\int_a^b\{Ay(x)-f(x)\}^2\rho(x)\,dx+\alpha\Omega[y(x)]=\min,\quad \alpha>0, \tag{56}$$

где так называемый *стабилизатор* $\Omega[y(x)]$ — специально подобранный положительный функционал, обладающий свойствами нормы; он несколько напоминает штрафную функцию. В главе XIV будет показано, что для стабилизаторов типа

$$\Omega[y(x)]=\int_a^b\{p(x)\,y^2(x)+q(x)\,y'^2(x)\}\,dx,\qquad p(x),q(x)>0, \tag{57}$$

решение задачи (56) непрерывно зависит от $f(x)$, причем при правильном подборе α оно одновременно достаточно близко в чебышевской норме к решению $\bar{y}(x)$ уравнения (54).

Уравнение (54) может привести и к другим функционалам. Пусть оператор A аддитивен, положителен и симметричен, так что $(y,\ Ay)>0$ при $y\neq 0$ и $(z,\ Ay)=(Az,\ y)$, где под скалярным произведением подразумевается интеграл от произведения функций. Рассмотрим функционал

$$\Phi[y(x)]=(y,\ Ay)-2(y,f), \tag{58}$$

где

$$(y,\ z)=\int_a^b y(x)\,z(x)\,dx.$$

Покажем, что задача на минимум этого функционала эквивалентна задаче решения операторного уравнения (54).

В самом деле, запишем произвольную функцию $y(x)$ в следующем виде:

$$y(x) = \bar{y}(x) + \lambda z(x). \tag{59}$$

Подставляя это выражение в правую часть формулы (58), получим

$$\Phi[y(x)] = \Phi[\bar{y}(x)] + 2\lambda(z,\ A\bar{y} - f) + \lambda^2(z,\ Az). \tag{60}$$

Если $\bar{y}(x)$ есть решение уравнения (54), то второе слагаемое в правой части (60) обращается в нуль; последний же член в правой части неотрицателен благодаря положительности оператора A. Значит, $\Phi[\bar{y}] = \inf \Phi[y]$, т. е. функционал (58) достигает минимума на решении операторного уравнения (54).

Наоборот, если $\bar{y}(x)$ в представлении (59) есть функция, на которой функционал (58) достигает минимума, то первая вариация функционала на этой функции равна нулю. Следовательно, $(d\Phi/d\lambda)_{\lambda=0} = 0$, каково бы ни было $z(x)$. Применяя это условие к (60) и одновременно полагая $z(x) = A\bar{y}(x) - f(x)$, получим

$$(A\bar{y} - f,\ A\bar{y} - f) = 0,$$

что выполняется только при $A\bar{y}(x) = f(x)$. Это означает, что функция, на которой функционал (58) достигает минимума, является решением операторного уравнения (54). Утверждение доказано.

Классическим примером применения описанного приема является краевая задача

$$\begin{gathered} -\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + q(x)\,y(x) = f(x), \\ p(x) > 0, \quad q(x) > 0, \quad y(-\infty) = y(+\infty) = 0. \end{gathered} \tag{61}$$

Интегрированием по частям легко убедиться в симметричности и положительности дифференциального оператора и получить следующее выражение для функционала (58):

$$\Phi[y(x)] = \int_{-\infty}^{+\infty} \left\{p(x)\left(\frac{dy}{dx}\right)^2 + q(x)\,y^2(x) - 2f(x)\,y(x)\right\} dx. \tag{62}$$

Отметим, что оператор A включает в себя не только дифференциальное (или интегральное) уравнение, но также краевые условия, если последние имеются. Краевые условия должны некоторым образом войти в функционал, соответственно изменив его вид. Например, для задачи на ограниченном отрезке с краевыми условиями третьего рода

$$-\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + q(x)\,y(x) = f(x), \quad p(x),\ q(x) > 0, \tag{63а}$$

$$\alpha_0 y(a) + \alpha_1 y'(a) = \alpha, \quad \beta_0 y(b) + \beta_1 y'(b) = \beta, \tag{63б}$$

надо минимизировать в классе достаточно гладких функций функционал

$$\Phi[y(x)]=\int_a^b\left\{p(x)\left(\frac{dy}{dx}\right)^2+q(x)\,y^2(x)-2f(x)\,y(x)\right\}dx+$$

$$+\frac{p(a)}{\alpha_1}[2\alpha y(a)-\alpha_0 y^2(a)]+\frac{p(b)}{\beta_1}[\beta_0 y^2(b)-2\beta y(b)]. \qquad (64)$$

От функций, минимизирующих этот функционал, уже не надо требовать удовлетворения краевым условиям — они автоматически будут им удовлетворять.

В теоретической физике встречаются функционалы более сложные, чем квадратичные. Например, в статистической модели атома Томаса — Ферми при температуре абсолютного нуля энергия выражается через электронную плотность следующим образом:

$$E[\rho(\boldsymbol{r})]=\int_V dv\left\{\frac{3(3\pi^2)^{2/3}\hbar^2}{10m}\rho^{5/3}(\boldsymbol{r})-\frac{Ze^2}{r}\rho(\boldsymbol{r})+\frac{e^2}{2}\rho(\boldsymbol{r})\int_V\frac{\rho(\boldsymbol{r}')\,dv'}{|\boldsymbol{r}-\boldsymbol{r}'|}\right\}. \qquad (65)$$

Поскольку при нулевой температуре и заданном объеме энергия минимальна, то нахождение электронной плотности сводится к задаче на условный экстремум для этого функционала (дополнительное условие заключается в том, что полное число электронов равно заряду ядра).

К еще более сложным функционалам приводят задачи *оптимального управления*, в которых ищется минимум функционала $\Phi[y(x)]$, причем функция $y(x)$ является решением задачи Коши для дифференциального уравнения $\frac{dy}{dx}=F(x,\ y(x),\ u(x)),\ y(0)=y_0$. Требуется найти такую *управляющую* функцию $u(x)$, при которой заданный функционал минимален. К задачам оптимального управления относится, например, определение оптимального режима расхода горючего $u(t)$ при запуске ракеты, приводящего к максимальной высоте подъема Φ при заданном начальном количестве горючего.

2. Метод пробных функций. Общая схема численного решения заключается в сведении задачи (53) к поиску минимума функции многих переменных. Рассмотрим класс V_n *пробных функций* заданного вида $v_n(x;\ \boldsymbol{a})=v_n(x;\ a_1,\ a_2,\ \ldots,\ a_n)$, содержащих n свободных параметров и принадлежащих множеству V_n. На этом классе функций рассматриваемый функционал будет функцией n переменных — свободных параметров:

$$\Phi[v_n(x;\ \boldsymbol{a})]=F_n(\boldsymbol{a})\equiv F_n(a_1,\ a_2,\ \ldots,\ a_n); \qquad (66)$$

численное нахождение минимума функции многих переменных было подробно рассмотрено в предыдущих параграфах. Найдя

минимум функции $F_n(\boldsymbol{a})$ и соответствующие ему значения параметров $\bar{\boldsymbol{a}}$, мы определим функцию $v_n(x; \bar{\boldsymbol{a}})$, на которой функционал достигает своего минимума в классе V_n.

Можно ли считать найденную функцию $v_n(x; \bar{\boldsymbol{a}})$ приближенным значением искомого решения $\bar{y}(x)$? Чтобы выяснить это, рассмотрим предельный переход $n \to \infty$.

Построим бесконечную последовательность классов функций V_n (принадлежащих заданному множеству Y) с увеличивающимся числом параметров так, чтобы каждая функция предыдущего класса получалась из функции последующего класса фиксированием некоторого значения последнего параметра:

$$v_{n-1}(x; a_1, a_2, \ldots, a_{n-1}) = v_n(x; a_1, a_2, \ldots, a_{n-1}, \bar{a}_n). \quad (67)$$

Тогда каждый класс V_n вложен в классы с бóльшим индексом. Если обозначить через Φ_n минимум функционала на этом классе

$$\Phi_n = \Phi[v_n(x; \bar{\boldsymbol{a}})] = \inf_{V_n} \Phi[v_n(x; \boldsymbol{a})], \quad (68)$$

то

$$\Phi_1 \geqslant \Phi_2 \geqslant \Phi_3 \geqslant \ldots \geqslant \bar{\Phi} = \inf_{Y} \Phi[y(x)].$$

Последовательность Φ_n не возрастает и ограничена снизу; значит, она сходится к пределу, который больше или равен $\bar{\Phi}$. Если $\lim_{n\to\infty} \Phi_n = \bar{\Phi}$, то последовательность функций $v_n(x; \bar{\boldsymbol{a}})$, на которых достигается минимум функционала в классах V_n, называют *минимизирующей* (или минимизирующей функционал).

Рассмотрим два понятия, нужных для дальнейшего изложения.

Будем называть функционал $\Phi[y(x)]$ *непрерывным*, если он непрерывно зависит от $y(x)$, т. е. если фиксировать $y(x)$, то для любого $\varepsilon > 0$ найдется такое $\delta(\varepsilon)$, что при $\|y(x) - \tilde{y}(x)\| < \delta(\varepsilon)$ будет выполняться неравенство $|\Phi[y] - \Phi[\tilde{y}]| < \varepsilon$. Очевидно, наличие или отсутствие этого свойства зависит как от вида функционала, так и от выбора нормы функции. Например, наиболее распространенные функционалы имеют вид

$$\Phi[y(x)] = \int_a^b f(x, y(x), y'(x), \ldots, y^{(p)}(x))\, dx, \quad (69)$$

где f — непрерывная функция всех своих аргументов. Их можно рассматривать в пространстве $C^{(p)}$ с нормой $\|y\| = \max\{|y(x)|, |y'(x)|, \ldots, |y^{(p)}(x)|\}$; тогда непрерывность функционала очевидна. А в чебышевском пространстве $C^{(0)}$ такой функционал уже не будет, вообще говоря, непрерывно зависеть от $y(x)$.

Бесконечная система функций заданного вида $\{v_n\}$ называется *полной*, если при $n \to \infty$ она может аппроксимировать в данной норме со сколь угодно высокой точностью любую функцию множества Y. Это значит, что для любой заданной функции $y(x) \in Y$ и любого $\delta > 0$ существует такое N, что при $n > N$ в классах V_n найдутся функции $\tilde{v}_n(x)$, удовлетворяющие условию $\| y(x) - \tilde{v}_n(x) \| < \delta$. Понятие полноты также существенно связано не только с выбором системы $v_n(x; \boldsymbol{a})$, но также с выбором нормы и множества Y.

Достаточные условия сходимости $v_n(x; \bar{\boldsymbol{a}})$ искомому решению дает следующая

Теорема. а) *Если система функций* $v_n(x; \boldsymbol{a})$ *полная, а функционал* $\Phi[y(x)]$ *непрерывен, то последовательность* $v_n(x; \bar{\boldsymbol{a}})$ *является минимизирующей,*

б) *если требования пункта* (а) *выполнены и функционал удовлетворяет дополнительному условию*

$$\Phi[y(x)] - \Phi[\bar{y}(x)] \geqslant \alpha \| y(x) - \bar{y}(x) \|^{\beta}, \quad \alpha, \beta > 0, \tag{70}$$

то последовательность $v_n(x, \bar{\boldsymbol{a}})$ *сходится к решению* $\bar{y}(x)$ *задачи* (53).

Доказательство. Поскольку функционал непрерывен, то для искомого решения $\bar{y}(x)$ задачи (53) и для заданного ε найдется такое δ, что если $\|y - \bar{y}\| < \delta$, то $\varepsilon > \Phi[y] - \Phi[\bar{y}] \geqslant 0$ (в последнем неравенстве не надо ставить знак модуля, ибо $\Phi[\bar{y}]$ есть минимальное значение функционала). Но система $\{v_n\}$ полная; следовательно, для функции $\bar{y}(x)$ и данного δ существует такое N, что во всех классах V_n при $n > N$ найдутся функции $v_n(x; \tilde{\boldsymbol{a}})$, удовлетворяющие условию $\| v_n(x; \tilde{\boldsymbol{a}}) - \bar{y}(x) \| < \delta$. Тогда выполняется неравенство $\varepsilon > \Phi[v_n(x; \tilde{\boldsymbol{a}})] - \bar{\Phi} \geqslant 0$. Поскольку $\Phi_n = \inf \Phi[v_n(x; \boldsymbol{a})]$, то отсюда следует неравенство $\varepsilon > \Phi_n - \bar{\Phi} \geqslant 0$. Оно означает, что

$$\lim_{n \to \infty} \Phi_n = \bar{\Phi},$$

так что первое утверждение теоремы доказано.

Применяя к последнему неравенству условие (70), получим $\| v_n(x, \bar{\boldsymbol{a}}) - \bar{y}(x) \| \leqslant (\varepsilon/\alpha)^{1/\beta}$, так что второе утверждение теоремы также доказано.

Замечание 1. Сходимость $v_n(x; \bar{\boldsymbol{a}}) \to \bar{y}(x)$ доказана в смысле той нормы, которая входила в определения полноты системы функций, непрерывности функционала и условие (70). Пусть в исходных определениях подразумевались разные нормы; в условиях полноты — аппроксимация в $\| \cdot \|_1$, в условии непрерывности функционала — малость $\| \delta y \|_2$ и в условии (70) — неравенство при $\| \cdot \|_3$. Если существует такая норма $\| \cdot \|_4$, которая не сильнее $\| \cdot \|_1$ и $\| \cdot \|_3$, но не слабее $\| \cdot \|_2$, то при переходе к этой норме все не-

равенства сохранятся *). Тогда из теоремы следует сходимость минимизирующей последовательности в $\|\cdot\|_1$.

Замечание 2. Пусть функционал $\Phi[y]$ определен на множестве Y, но при этом известно, что искомое решение принадлежит некоторому подмножеству Y_0. Например, функционал (64) определен на множестве кусочно-гладких функций, а решение является кусочно-гладкой функцией, удовлетворяющей краевым условиям (63б). В этом случае достаточно искать решение только среди функций подмножества Y_0 и проверять полноту системы пробных функций $\{v_n\}$ и непрерывность функционала лишь по отношению к этому подмножеству. Это может существенно облегчить решение поставленной задачи.

Замечание 3. Нетрудно доказать, что если функционал непрерывен, то для сходимости последовательности $v_n(x; \bar{a})$ к $\bar{y}(x)$ необходимо, чтобы эта последовательность была минимизирующей.

Замечание 4. Существуют функционалы, для которых последовательности $v_n(x; \bar{a})$ являются минимизирующими, но при этом ни к какой предельной функции не сходятся. Это нередко встречается в задачах оптимального управления. Такие задачи относятся к некорректно поставленным и требуют регуляризации.

В задачах для конкретных функционалов исследование сходимости сводится к выбору подходящей полной системы функций $\{v_n\}$ и нормы и проверке условий теоремы. Норму обычно выбирают из соображений простоты доказательства, но эта норма не должна быть слишком слабой, иначе результат не будет представлять практической ценности.

Метод пробных функций в своей наиболее общей постановке применяется не часто. Если функционал имеет достаточно сложный вид, как в примере (65), или если выбрана система функций $v_n(x; \boldsymbol{a})$, нелинейно зависящих от свободных параметров, то получающаяся при этом функция $F(\boldsymbol{a})$ имеет достаточно общий вид. Обычно ее минимум удается найти численными методами, только если число переменных (свободных параметров) не превышает $n \sim 10-20$. Такого числа параметров не всегда достаточно, чтобы уверенно констатировать сходимость.

Поэтому для конкретных функционалов сложного вида обычно стараются исследовать качественный характер решения и выбирают пробные функции с небольшим ($n \sim 3-10$) числом параметров так, чтобы по своему качественному поведению — асимптотике, полюсам и т. д. — они были бы близки к искомому решению. Проводят исследование непрерывности функционала и полноты системы. Затем выполняют расчеты с различным числом

*) Напомним, что норма $\|\cdot\|_1$ называется более сильной, чем $\|\cdot\|_2$, если для любой допустимой функции $y(x)$ выполняется неравенство $\|y\|_1 \geqslant C\|y\|_2$, где $C = \text{const}$.

параметров и смотрят, сходятся ли полученные значения Φ_n и функции $v_n(x; \bar{a})$ к какому-то пределу.

Если последовательность $\{v_n\}$ выбрана удачно, то величина Φ_n будет близка к своему пределу Φ уже при небольшом n. Например, для функционала энергии атома (65) пробная функция всего с четырьмя параметрами $\rho(x) \approx \left(\sum_{k=1}^{4} a_k x^k\right)^{-3/2}$ обеспечивает точность расчета полной энергии существенно лучше 1%. Само искомое решение (в данном примере — распределение электронов в атоме) находится при этом с меньшей, но удовлетворительной точностью.

Однако оценить фактическую точность найденного приближения на основании таких расчетов не удается. Далее мы рассмотрим два частных случая метода пробных функций, когда можно получить и более высокую точность, и неплохую оценку погрешности.

3. Метод Ритца. Ряд важных математических задач сводится к минимизации квадратичного функционала. Примером является решение корректно или некорректно поставленных задач для линейного операторного уравнения (54), приводящее к одному из функционалов (55), (56) или (58). Если в качестве пробных функций взять обобщенные многочлены

$$v_n(x; \boldsymbol{a}) = \varphi_0(x) + \sum_{k=1}^{n} a_k \varphi_k(x), \tag{71}$$

то на них квадратичный функционал будет квадратичной функцией параметров a_k. Задача на нахождение минимума квадратичной функции $F(\boldsymbol{a})$ посредством дифференцирования по переменным a_k сводится к системе алгебраических линейных уравнений; ее нетрудно численно решить даже при числе параметров $n \sim 100 - 200$ *). Этот частный случай метода пробных функций называют методом Ритца.

Обсудим выбор функций $\varphi_k(x)$. Его целесообразно связать с краевыми условиями для задач типа (54), которые обычно линейны. Пусть, для определенности, это условия первого рода

$$y(a) = \alpha, \quad y(b) = \beta. \tag{72}$$

Выберем какую-нибудь гладкую функцию $\varphi_0(x)$ так, чтобы она

*) В отдельных случаях число параметров бывает еще больше. Например, в квантовой химии при решении уравнения Шредингера для несферического многоцентрового поля молекулы берут $n \sim 1000$.

удовлетворяла этим краевым условиям, например,

$$\varphi_0(x) = \alpha + \frac{\beta - \alpha}{b - a}(x - a), \tag{73а}$$

или

$$\varphi_0(x) = \alpha + (\beta - \alpha)\sin\frac{\pi(x-a)}{2(b-a)}. \tag{73б}$$

Остальные функции выберем так, чтобы они удовлетворяли однородным краевым условиям типа (72) и при этом образовывали бы полную систему. Например, согласно теореме Вейерштрасса любую непрерывную функцию можно аппроксимировать со сколь угодно высокой точностью алгебраическими или тригонометрическими многочленами. Поэтому можно положить

$$\varphi_k(x) = (x-a)^k(b-x), \quad k = 1, 2, \ldots, \tag{73в}$$

или

$$\varphi_k(x) = \sin\frac{\pi k(x-a)}{b-a}, \quad k = 1, 2, \ldots \tag{73г}$$

В этом случае пробные функции (71) при любых коэффициентах a_k удовлетворяют неоднородным краевым условиям (72) и являются полными на множестве непрерывных функций, удовлетворяющих этим краевым условиям. Согласно замечанию 3 к теореме п. 2 такой выбор пробных функций допустим.

Пример. Рассмотрим задачу на минимум квадратичного функционала (58) с вещественным симметричным положительным оператором A:

$$\Phi[y(x)] = (y, Ay) - 2(f, y) = \min. \tag{74}$$

Подставляя в этот функционал пробные функции Ритца (71), получим квадратичную функцию свободных параметров

$$\Phi[v_n(x; a)] = F(a) = \sum_{k=1}^{n}\sum_{m=1}^{n} a_k a_m(\varphi_k, A\varphi_m) + 2\sum_{k=1}^{n} a_k[(\varphi_k, A\varphi_0) - (\varphi_k, f)] + (\varphi_0, A\varphi_0 - 2f) = \min.$$

Приравнивая нулю производные этой квадратичной функции по параметрам, получим для определения параметров линейную систему уравнений

$$\sum_{m=1}^{n} a_m(\varphi_k, A\varphi_m) = -(\varphi_k, A\varphi_0 - f), \quad 1 \leqslant k \leqslant n. \tag{75}$$

Дадим схему исследования сходимости, не останавливаясь на деталях. В этом примере удобно ввести норму, связанную с данным положительным оператором A:

$$\|y\|_A^2 = (y, Ay). \tag{76}$$

Сделаем естественное предположение, что эта норма не слабее $\|\cdot\|_C$. В самом деле, для операторов A типа (61) такая норма содержит интеграл от квадрата функции и ее производной, а среднеквадратичная близость и функций, и их производных есть более сильное требование, чем равномерная близость функций. Для таких операторов система тригонометрических функций (73г) будет *полной* по норме (76). Действительно, для любой функции $y(x)$, непрерывно дифференцируемой r раз, ее тригонометрический ряд Фурье среднеквадратично сходится к ней вместе со своими r-ми производными. А сходимость по норме (76) отличается от среднеквадратичной только наличием весовых множителей $p(x)$, $q(x)$ под интегралом (62), что несущественно.

Найдем вариацию функционала (74) на произвольной функции

$$\delta\Phi[y]=\Phi[y+\delta y]-\Phi[y]=(\delta y,\ A\delta y)+2(\delta y,\ Ay-f). \tag{77}$$

Первое слагаемое этой вариации равно $\|\delta y\|_A^2$, т. е. является бесконечно малой второго порядка; второе слагаемое, по предположению о силе нормы (76), является бесконечно малой не ниже первого порядка относительно $\|\delta y\|_A$. Отсюда следует *непрерывность* функционала. Наконец, заметим, что решение $\bar{y}$ искомой задачи (74) удовлетворяет уравнению $Ay=f$. Подставляя это решение в (77), получим

$$\delta\Phi[\bar{y}]=\|\delta y\|_A^2.$$

Таким образом, последнее условие (70) теоремы о сходимости выполнено и метод Ритца в данном примере сходится.

Заметим, что для не квадратичных функционалов $\Phi[y]$ линейные по параметрам пробные функции (71) не дают никаких преимуществ, ибо получающиеся функции параметров $F(\boldsymbol{a})=\Phi[v_n(x;\boldsymbol{a})]$ все равно оказываются не квадратичными. Поэтому метод Ритца фактически применяют только для квадратичных функционалов.

4. Сеточный метод. Введем сетку по аргументу x и заменим все производные и интегралы, входящие в функционал, некоторыми разностями и суммами узловых значений функции $y_k=y(x_k)$. Тогда функционал аппроксимируется некоторой вспомогательной функцией многих переменных — значений решения в узлах:

$$\Phi[y(x)]\approx F(y_0,\ y_1,\ y_2,\ \ldots,\ y_n)=\min. \tag{78}$$

Решая задачу $F(y_0,\ \ldots,\ y_n)=\min$ численными методами, мы непосредственно получим приближенные значения решения в узлах сетки. Зная их, решение при остальных значениях аргумента (не совпадающих с узлами сетки) можно найти интерполяцией.

Например, рассмотрим сферически-симметричный сжатый атом в модели Томаса — Ферми; его энергия задается функционалом (65), где интегралы берутся по сферической атомной ячейке радиуса R. Вводя равномерную сетку $0=r_0<r_1<\ldots<r_n=R$ и вычисляя интегралы по формуле прямоугольников, получим

$$E\approx\frac{R}{n}\sum_{i=1}^{n}(\alpha r_i^2\rho_i^{5/3}-\beta r_i\rho_i+\gamma r_i^2\rho_i\varphi_i), \tag{79а}$$

где атомный потенциал $\varphi(r)$ сам зависит от неизвестной электронной плотности $\rho(r)$:

$$\varphi_i \approx \frac{R}{nr_i}\sum_{j=1}^{i} r_j^2\rho_j + \frac{R}{n}\sum_{j=i+1}^{n} r_j\rho_j, \qquad (79б)$$

а коэффициенты α, β, γ выражаются через физические константы. Надо найти минимум энергии при дополнительном условии нормировки

$$\int \rho(r)\, dv = Z,$$

причем это условие также надо приближенно записать в сеточной форме.

Выражения (79а), (79б) достаточно сложные, и при большом числе узлов сетки найти минимум численными методами трудно. Очевидно, что для произвольных функционалов число узлов сетки, которое практически возможно использовать в расчетах, очень невелико: оно не превышает $n \sim 10-20$. Однако даже при таком числе узлов нередко удается получить неплохую точность при умеренном объеме расчетов, используя прием сгущения сеток.

Для этого выполняют серию расчетов на сгущающихся вдвое сетках с числами интервалов $n=1$, 2, 4, 8 и 16. Поскольку порядок точности выбранных разностных формул дифференцирования и интегрирования обычно известен, то проводят уточнение результатов, полученных на разных сетках, рекуррентным методом Рунге. При этом непосредственно наблюдают, сходится ли численный расчет к пределу при увеличении n, и производят апостериорную оценку погрешности.

На каждой сетке минимум функции $F(y_0, \ldots, y_n)$ находят обычно каким-либо итерационным методом спуска. Для уменьшения числа итераций (а тем самым, объема вычислений) организуют расчет следующим образом. Сначала выполняют расчет на самой редкой сетке, где неизвестных мало (при $n=1$ всего два — y_0 и y_1) и объем вычислений заведомо невелик даже при плохом нулевом приближении. Найденный на этой сетке профиль $y(x)$ интерполируют на следующей, более подробной сетке, и используют на ней в качестве нулевого приближения. Вновь найденный профиль снова интерполируют и т. д.

Для квадратичных функционалов при использовании линейных формул численного дифференцирования и интегрирования задача (78), как и в методе Ритца, сводится к нахождению минимума квадратичной функции. Например, возьмем функционал (62), но на ограниченном отрезке $a \leqslant x \leqslant b$; введем на этом отрезке равномерную сетку с шагом h и аппроксимируем интеграл при

помощи обобщенных формул трапеции и средних:

$$\Phi[y] \approx h \sum_{i=1}^{n} \left\{ p_{i-1/2} \left(\frac{y_i - y_{i-1}}{h} \right)^2 + \right.$$
$$\left. + \frac{1}{4} q_{i-1/2} (y_i + y_{i-1})^2 - f_{i-1/2} (y_i + y_{i-1}) \right\}. \qquad (80)$$

Отыскание минимума опять сводится к решению линейной системы уравнений с неизвестными y_i, $0 \leqslant i \leqslant n$, легко выполняемому численными методами. Это позволяет брать очень большое число узлов. Тогда имеет смысл ставить вопрос о теоретическом исследовании сходимости приближенного решения к искомому при $n \to \infty$. Обоснования сходимости мы не будем давать. Укажем только один случай, когда применима сформулированная в п. 2 теорема.

Пусть функционал имеет вид (69), т. е. явно содержит функцию и ее производные вплоть до p-й. Построим последовательность сеток так, чтобы предыдущая сетка содержалась в последующей; это можно делать сгущением сеток вдвое, причем сетки могут быть даже неравномерными.

В качестве пробных функций возьмем сплайны $S(x)$ порядка не ниже p (см. главу II, § 1). Эти сплайны являются многочленами степени p, коэффициенты которых линейно выражаются через узловые значения искомой функции y_i; их производные порядка ниже p непрерывны, а p-я производная всюду существует и кусочно-непрерывна. Нетрудно убедиться в том, что условие вложения классов пробных функций V_n во все последующие классы при этом выполнено, и что такими пробными функциями при $n \to \infty$ можно со сколь угодно высокой точностью аппроксимировать любую p раз непрерывно дифференцируемую функцию вместе с ее производными вплоть до $y^{(p)}(x)$. Следовательно, система сплайн-функций обладает свойствами, нужными для применения теоремы о сходимости.

В качестве примера рассмотрим квадратичный функционал типа (62), содержащий первую производную:

$$\Phi[y] = \int_a^b \left\{ p(x) \left(\frac{dy}{dx} \right)^2 + q(x) y^2(x) - 2f(x) y(x) \right\} dx. \qquad (81)$$

Сплайн должен иметь порядок тоже не ниже первого. Ограничимся простейшим сплайном первого порядка — ломаной линией, проведенной через точки (x_i, y_i):

$$S(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}} (x - x_{i-1}), \qquad x_{i-1} \leqslant x \leqslant x_i. \qquad (82)$$

Надо разбить интеграл (81) на сумму интегралов по отдельным интервалам сетки и каждый из этих интегралов вычислить, используя заданный закон интерполяции (82). Например, поскольку $S'(x)=(y_i-y_{i-1})/(x_i-x_{i-1})$ при $x_{i-1}<x<x_i$, то

$$\int_a^b p(x)\,[y'(x)]^2\,dx \approx \sum_{i=1}^{n}\left(\frac{y_i-y_{i-1}}{x_i-x_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\,dx.$$

Аналогично вычисляются остальные слагаемые в (81).

Получающиеся выражения имеют более сложный вид, чем при не сплайновой аппроксимации (80); использование сплайнов высших порядков привело бы к еще более сложным выражениям (зато получающиеся при этом сеточные схемы имели бы более высокий порядок точности). Тем не менее, поскольку сами сплайны линейно зависят от узловых значений функции, то подстановка их в квадратичный функционал приводит к задаче на минимум квадратичной формы. Поэтому такой подстановкой пользуются даже для многомерных функционалов, к которым сводятся краевые задачи для эллиптических уравнений в частных производных.

Коснемся построения сплайнов в многомерных задачах. Если область G двумерна, то ее можно разбить на треугольные ячейки (у граничных ячеек одна сторона может быть не прямой). В каждой ячейке g_i по узловым значениям функции двух переменных $z(x, y)$ в трех вершинах ячейки однозначно строится простейший линейный сплайн $S(x, y)=a_i+b_ix+c_iy$, где $(x, y)\in g_i$; он соответствует аппроксимации поверхности $z(x, y)$ плоскостью. Сплайновые плоскости соседних ячеек пересекаются по прямым, проходящим через выбранные узлы поверхности $z(x, y)$; эти прямые проектируются точно на границы ячеек. Следовательно, двумерный линейный сплайн, построенный указанным образом, является непрерывным и кусочно-гладким в области G.

Описанный способ построения линейного сплайна естественно обобщается на случай любого числа измерений. При этом область G следует разбить на многомерные симплексы.

Но для построения сплайнов более высокого порядка этот несложный алгоритм не годится: в этом случае он не гарантирует непрерывности и требуемой гладкости сплайновой поверхности на границах ячеек. Требования непрерывности функции и некоторого числа ее производных на границах ячеек надо формулировать в виде дополнительных уравнений, которым должны удовлетворять коэффициенты сплайнов. Надо, чтобы полное число уравнений равнялось полному числу коэффициентов; это будет не при любой форме ячейки.

Например, рассмотрим двумерный кубический сплайн $S(x, y) = \sum_{k+m=0}^{3} a_{km}x^k y^m$ в прямоугольных ячейках со сторонами, параллельными осям координат. Потребуем непрерывности на границах сплайна вместе со вторыми производными. Этот сплайн имеет 10 коэффициентов в расчете на одну ячейку. Совпадение сплайна с функцией $z(x, y)$ в вершинах ячейки дает 4 уравнения. Потребуем на обеих сторонах ячейки, параллельных оси x, непрерывности $S(x, y)$, S_y и S_{yy}; дифференцируя их по x, нетрудно убедиться, что тогда на этих сторонах величины S_x, S_{xx} и S_{xy} тоже будут непрерывны. Аналогично потребуем непрерывности величин $S(x, y)$, S_x и S_{xx} на сторонах, параллельных оси y. Это дает по 3 уравнения на каждой стороне ячейки, но эти уравнения связывают коэффициенты двух ячеек. Следовательно, всего непрерывность дает 6 уравнений в расчете на одну ячейку. Таким образом, полное число уравнений равно полному числу коэффициентов, и сплайн определяется однозначно (с точностью до условий на границе области G).

ЗАДАЧИ

1. Вывести итерационную формулу (12) поиска минимума функции одной переменной $\Phi(x)$, заменяя истинную кривую интерполяционной параболой, проведенной через три точки $x_s - h$, x_s, $x_s + h$.

2. Дать аналогичный вывод формулы (13), строя интерполяционную параболу по точкам x_s, x_{s-1}, x_{s-2}.

3. Доказать оценку (14) для скорости сходимости процесса (13); для этого можно воспользоваться схемой доказательства, данного в главе V, § 2, п. 7.

4. Написать уравнение для линий уровня квадратичной формы (18); найти главные оси полученных эллипсов и определить отношение длин главных осей.

5. Написать линейную систему уравнений, решение которой минимизирует регуляризованную задачу линейного программирования (52).

6. Построить какую-нибудь полную систему функций в методе Ритца, если вместо краевого условия первого рода (72) задано условие второго рода $y'(a) = \alpha$, $y'(b) = \beta$.

7. Провести аккуратное доказательство сходимости метода Ритца для функционала (62), используя схему, данную в § 4, п. 3.

8. Написать систему уравнений для определения сеточных значений функции y_i, к которой приводится задача минимума функционала (62) после разностной замены (80). Убедиться, что эта линейная система имеет трехдиагональную матрицу и решение ее методом прогонки устойчиво.

9. Показать, что двумерный квадратичный сплайн

$$S(x, y) = \sum_{k+m=0}^{2} a_{km}x^m y^k$$

на треугольной сетке и трехмерный квадратичный сплайн на сетке из тетраэдров определяются однозначно (с точностью до условий на границе области G).

ГЛАВА VIII

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

В главе VIII рассмотрены основные методы численного решения различных типов задач для обыкновенных дифференциальных уравнений. В § 1 изложены постановка и методы решения задачи с начальными условиями (задачи Коши); эти методы применяются и при решении других типов задач. В § 2 даны постановки и методы решения краевых задач, а в § 3 — задач на собственные значения.

§ 1. Задача Коши

1. Постановка задачи. Обыкновенными дифференциальными уравнениями можно описать задачи движения системы взаимодействующих материальных точек, химической кинетики, электрических цепей, сопротивления материалов (например, статический прогиб упругого стержня) и многие другие. Ряд важных задач для уравнений в частных производных также сводится к задачам для обыкновенных дифференциальных уравнений. Так бывает, если многомерная задача допускает разделение переменных (например, задачи на нахождение собственных колебаний упругих балок и мембран простейшей формы, или определение спектра собственных значений энергии частицы в сферически-симметричном поле), или если ее решение зависит только от некоторой комбинации переменных (так называемые автомодельные решения). Таким образом, решение обыкновенных дифференциальных уравнений занимает важное место среди прикладных задач физики, химии и техники.

Конкретная прикладная задача может приводить к дифференциальному уравнению любого порядка, или к системе уравнений любого порядка. Но известно, что обыкновенное дифференциальное уравнение p-го порядка

$$u^{(p)}(x) = f(x,\ u,\ u',\ u'',\ \ldots,\ u^{(p-1)})$$

при помощи замены $u^{(k)}(x) \equiv u_k(x)$ можно свести к эквивалентной системе p уравнений первого порядка

$$u_k'(x) = u_{k+1}(x), \quad 0 \leqslant k \leqslant p-2,$$
$$u_{p-1}'(x) = f(x,\ u_0,\ u_1,\ \ldots,\ u_{p-1}),$$

где $u_0(x) \equiv u(x)$. Аналогично, произвольную систему дифференциальных уравнений любого порядка можно заменить некоторой эквивалентной системой уравнений первого порядка. Поэтому в дальнейшем мы будем, как правило, рассматривать системы уравнений первого порядка

$$u_k'(x) = f_k(x, u_1, u_2, \ldots, u_p), \quad 1 \leqslant k \leqslant p, \tag{1а}$$

записывая их для краткости в векторной форме

$$\boldsymbol{u}'(x) = \boldsymbol{f}(x, \boldsymbol{u}(x)), \tag{1б}$$

$$\boldsymbol{u} = \{u_1, u_2, \ldots, u_p\}, \quad \boldsymbol{f} = \{f_1, f_2, \ldots, f_p\}.$$

Известно, что система p-го порядка (1а) имеет множество решений, которое в общем случае зависит от p параметров $\boldsymbol{c} = \{c_1, c_2, \ldots, c_p\}$ и может быть записано в форме $\boldsymbol{u} = \boldsymbol{u}(x; \boldsymbol{c})$. Для определения значений этих параметров, т. е. для выделения единственного (или нужного) решения, надо наложить p дополнительных условий на функции $u_k(x)$.

Различают три основных типа задач для обыкновенных дифференциальных уравнений: задачи Коши, краевые задачи и задачи на собственные значения.

Задача Коши (задача с начальными условиями) имеет дополнительные условия вида

$$u_k(\xi) = \eta_k, \quad 1 \leqslant k \leqslant p, \tag{2}$$

т. е. заданы значения всех функций в одной и той же точке $x = \xi$. Эти условия можно рассматривать как задание координат начальной точки $(\xi, \eta_1, \eta_2, \ldots, \eta_p)$ интегральной кривой в $(p+1)$-мерном пространстве $(x, u_1, u_2, \ldots, u_p)$. Решение при этом обычно требуется найти на некотором отрезке $\xi \leqslant x \leqslant X$ (или $X \leqslant x \leqslant \xi$), так что точку $x = \xi$ можно считать начальной точкой этого отрезка.

Напомним *), что если правые части (1) непрерывны и ограничены в некоторой окрестности начальной точки $(\xi, \eta_1, \eta_2, \ldots, \eta_p)$, то задача Коши (1) — (2) имеет решение, но, вообще говоря, не единственное. Если правые части не только непрерывны, но и удовлетворяют условию Липшица по переменным u_k, то решение задачи Коши единственно и непрерывно зависит от координат начальной точки, т. е. задача корректно поставлена. Если вдобавок правые части имеют непрерывные производные по всем аргументам вплоть до q-го порядка, то решение $\boldsymbol{u}(x)$ имеет $q+1$ непрерывную производную по x.

2. Методы решения. Их можно условно разбить на точные, приближенные и численные. К точным относятся методы, позволяющие выразить решение дифференциального уравнения

*) См., например, [37].

через элементарные функции, либо представить его при помощи квадратур от элементарных функций. Эти методы изучаются в курсах обыкновенных дифференциальных уравнений. Нахождение точного решения задачи (1) — (2), а тем более — общего решения системы (1) облегчает качественное исследование этого решения и дальнейшие действия с ним.

Однако классы уравнений, для которых разработаны методы получения точных решений, сравнительно узки и охватывают только малую часть возникающих на практике задач. Например, доказано, что решение несложного уравнения

$$u'(x) = x^2 + u^2(x) \tag{3}$$

не выражается через элементарные функции. А уравнение

$$u'(x) = \frac{u - x}{u + x} \tag{4}$$

можно точно проинтегрировать и найти общее решение

$$\frac{1}{2}\ln(x^2 + u^2) + \operatorname{arctg}\frac{u}{x} = \text{const}. \tag{5}$$

Однако для того, чтобы составить таблицу значений $u(x)$, надо численно решить трансцендентное уравнение (5), а это нисколько не проще, чем непосредственно численно проинтегрировать уравнение (4)!

Приближенными будем называть методы, в которых решение получается как предел $u(x)$ некоторой последовательности $y_n(x)$, причем $y_n(x)$ выражаются через элементарные функции или при помощи квадратур. Ограничиваясь конечным числом n, получаем приближенное выражение для $u(x)$. Примером может служить метод разложения решения в обобщенный степенной ряд, рассматриваемый в курсах обыкновенных дифференциальных уравнений; некоторые другие приближенные методы будут изложены в этой главе. Однако эти методы удобны лишь в том случае, когда большую часть промежуточных выкладок удается сделать точно (например, найти явное выражение коэффициентов ряда). Это выполнимо лишь для сравнительно простых задач (таких как линейные), что сильно сужает область применения приближенных методов.

Численные методы — это алгоритмы вычисления приближенных (а иногда — точных) значений искомого решения $u(x)$ на некоторой выбранной сетке значений аргумента x_n. Решение при этом получается в виде таблицы. Численные методы не позволяют найти общего решения системы (1); они могут дать только какое-то частное решение, например, решение задачи Коши (1) — (2). Это основной недостаток численных методов. Зато эти методы применимы к очень широким классам уравнений и всем типам

задач для них. Поэтому с появлением быстродействующих ЭВМ численные методы решения стали одним из основных способов решения конкретных практических задач для обыкновенных дифференциальных уравнений.

Численные методы можно применять только к корректно поставленным (или регуляризованным) задачам. Заметим, однако, что для успешного применения численных методов формальное выполнение условий корректности может оказаться недостаточным. Надо, чтобы задача была *хорошо обусловлена*, т. е. малые изменения начальных условий приводили бы к достаточно малому изменению интегральных кривых. Если это условие не выполнено, т. е. задача *плохо обусловлена* (*слабо устойчива*), то небольшие изменения начальных условий или эквивалентные этим изменениям небольшие погрешности численного метода могут сильно исказить решение.

В качестве примера плохой обусловленности рассмотрим задачу

$$u'(x) = u - x, \qquad 0 \leqslant x \leqslant 100, \tag{6а}$$

$$u(0) = 1. \tag{6б}$$

Общее решение уравнения (6а) содержит одну произвольную постоянную

$$u(x;\ c) = 1 + x + ce^x.$$

При начальном условии (6б) она равна $c = 0$, так что $u(100) = 101$. Однако небольшое изменение начального условия $\bar{u}(0) = 1{,}000001$ слегка меняет постоянную: $\bar{c} = 10^{-6}$; тогда $\bar{u}(100) \approx 2{,}7 \times 10^{37}$, т. е. решение изменилось очень сильно.

В этом параграфе рассмотрены методы решения задачи Коши. Для простоты записи мы почти всюду ограничимся случаем одного уравнения первого порядка. Алгоритмы для случая системы p уравнений (1б) легко получаются из алгоритмов, составленных для одного уравнения, формальной заменой $u(x)$ и $f(x, u)$ на $\boldsymbol{u}(x)$ и $\boldsymbol{f}(x, \boldsymbol{u})$.

3. **Метод Пикара.** Это приближенный метод решения, являющийся обобщением метода последовательных приближений (см. главу V, § 2). Рассмотрим задачу Коши для уравнения первого порядка

$$u'(x) = f(x,\ u(x)), \qquad \xi \leqslant x \leqslant X, \qquad u(\xi) = \eta. \tag{7}$$

Интегрируя дифференциальное уравнение, заменим эту задачу эквивалентным ей интегральным уравнением типа Вольтерра

$$u(x) = \eta + \int_{\xi}^{x} f(\tau,\ u(\tau))\, d\tau. \tag{8}$$

Решая это интегральное уравнение методом последовательных приближений, получим итерационный процесс Пикара

$$y_s(x) = \eta + \int_{\xi}^{x} f(\tau,\ y_{s-1}(\tau))\, d\tau, \qquad y_0(x) \equiv \eta \tag{9}$$

(приближенное решение, в отличие от точного, мы будем обозначать через y). На каждой итерации этого процесса интегрирование выполняется либо точно, либо численными методами, описанными в главе IV.

Докажем сходимость метода, предполагая, что в некоторой ограниченной области $G(x, u)$ правая часть $f(x, u)$ непрерывна и удовлетворяет по переменной u условию Липшица $|f(x, u_1) - f(x, u_2)| \leqslant L|u_1 - u_2|$.

Поскольку область $G(x, u)$ ограничена, то выполняются соотношения $|x - \xi| \leqslant a$, $|u - \eta| \leqslant b$. Обозначим погрешность приближенного решения через $z_s(x) = y_s(x) - u(x)$. Вычитая (8) из (9) и используя условие Липшица, получим

$$|z_s(x)| \leqslant L \int_{\xi}^{x} |z_{s-1}(\tau)|\, d\tau.$$

Решая это рекуррентное соотношение и учитывая, что $|z_0(x)| = |\eta - u(x)| \leqslant b$, найдем последовательно

$$|z_1(x)| \leqslant bL(x-\xi), \quad |z_2(x)| \leqslant \frac{1}{2} bL^2 (x-\xi)^2, \ldots,$$

$$|z_s(x)| \leqslant \frac{1}{s!} bL^s (x-\xi)^s.$$

Отсюда следует оценка погрешности

$$|z_s(x)| \leqslant \frac{b}{s!}(aL)^s \approx \frac{b}{\sqrt{2\pi s}} \left(\frac{eaL}{s}\right)^s. \tag{10}$$

Видно, что $\max |z_s(x)| \to 0$ при $s \to \infty$, т. е. *приближенное решение равномерно сходится к точному во всей области* $G(x, u)$.

Пример. Применим метод Пикара к задаче Коши для уравнения (3), решение которого не выражается через элементарные функции

$$u'(x) = x^2 + u^2, \qquad u(0) = 0.$$

В этом случае квадратуры (9) вычисляются точно, и мы легко получаем

$$y_0(x) = 0, \quad y_1(x) = \frac{1}{3} x^3, \quad y_2(x) = \frac{1}{3} x^3 \left(1 + \frac{1}{21} x^4\right),$$

$$y_3(x) = \frac{1}{3} x^3 \left(1 + \frac{1}{21} x^4 + \frac{2}{693} x^8 + \frac{1}{19845} x^{12}\right),$$

и т. д. Видно, что при $x \leqslant 1$ эти приближения быстро сходятся и позволяют вычислить решение с высокой точностью.

Из этого примера видно, что метод Пикара выгодно применять, если интегралы (9) удается вычислить через элементарные функции. Если же правая часть уравнения (7) более сложна, так что эти интегралы приходится находить численными методами, то метод Пикара становится не слишком удобным.

Метод Пикара легко обобщается на системы уравнений способом, описанным в п. 2. Однако на практике чем выше порядок системы, тем реже удается точно вычислять интегралы в (9), что ограничивает применение метода в этом случае.

Имеется много других приближенных методов. Например, С. А. Чаплыгин предложил метод, являющийся обобщением алгебраического метода Ньютона на случай дифференциальных уравнений. Другой способ обобщения метода Ньютона предложил Л. В. Канторович в 1948 г. В обоих этих методах, так же как и в методе Пикара, итерации выполняются при помощи квадратур. Однако квадратуры в них имеют гораздо более сложный вид, чем (9), и редко берутся в элементарных функциях. Поэтому эти методы почти не применяют.

4. Метод малого параметра. Достаточно простыми оказываются вычисления методом малого параметра, предложенным Пуанкаре в 1892 г. Пусть правая часть уравнения $u' = f(x, u; \lambda)$ зависит от параметра и известно частное решение $y_0(x)$ при некотором значении параметра $\lambda = \lambda_0$. Будем искать решение в виде ряда

$$u(x) = \sum_{n=0}^{\infty} (\lambda - \lambda_0)^n y_n(x). \tag{11}$$

Подставляя этот ряд в исходное уравнение и разлагая $f(x, u; \lambda)$ по формуле Тейлора по степеням $(\lambda - \lambda_0)$, получим для определения $y_n(x)$ линейные уравнения

$$y_n'(x) = \alpha_n(x)\, y_n(x) + v_n(x), \qquad n = 1, 2, 3, \ldots \tag{12}$$

Здесь коэффициенты $\alpha_n(x)$ выражаются через производные $f(x, u; \lambda)$ при $u = y_0(x)$, $\lambda = \lambda_0$, а функции $v_n(x)$ выражаются через $y_k(x)$, $0 \leqslant k < n$. Тем самым нахождение $y_n(x)$ сводится к квадратурам. Достаточным условием сходимости ряда (11) является аналитичность $f(x, u; \lambda)$ по всем аргументам.

При практическом применении метода малого параметра специально вводят в правую часть уравнения (7) параметр так, чтобы при некотором его значении легко находилось частное решение; после этого действуют по описанной схеме. Например, для уравнения (3) можно прибавить к правой части член λu^2, положив, таким образом, $f(x, u; \lambda) = x^2 + (1 + \lambda) u^2$; тогда при $\lambda_0 = -1$ сразу видно частное решение $y_0(x) = \frac{1}{3} x^3 + c$, где постоянная c определяется из начального условия.

Метод малого параметра естественно переносится на уравнения высоких порядков или на системы уравнений. При этом

вместо цепочки последовательно решаемых линейных уравнений (12) возникают цепочки систем линейных дифференциальных уравнений. Однако все выкладки становятся существенно более громоздкими.

5. Метод ломаных. Это простейший численный метод. В практике вычислений он употребляется очень редко из-за невысокой точности. Но на его примере удобно пояснить способы построения и исследования численных методов.

Рассмотрим задачу Коши (7) и выберем на отрезке $[\xi, X]$ некоторую сетку $\{x_n,\ 0 \leqslant n \leqslant N\}$ значений аргумента так, чтобы выполнялись соотношения $\xi = x_0 < x_1 < x_2 < \ldots < x_N = X$ (сетка может быть неравномерной). Разлагая решение $u(x)$ по формуле Тейлора на интервале сетки $x_n \leqslant x \leqslant x_{n+1}$ и обозначая $u(x_n) = u_n$, получим

$$u_{n+1} = u_n + h_n u_n' + \frac{1}{2} h_n^2 u_n'' + \ldots, \quad h_n = x_{n+1} - x_n. \tag{13}$$

Стоящие в правой части производные можно найти, дифференцируя уравнение (7) требуемое число раз:

$$u' = f(x,\ u), \quad u'' = \frac{d}{dx} f(x,\ u) = f_x + f f_u \tag{14}$$

и т. д. В принципе, если $f(x, u)$ имеет q-е непрерывные производные по совокупности аргументов, то в разложении (13) можно удержать члены вплоть до $O(h^{q+1})$.

Однако использовать для расчетов формулу (13) с большим числом членов невыгодно. Во-первых, даже при сравнительно простой правой части выражения для производных могут оказаться громоздкими. Во-вторых, если правая часть известна лишь приближенно, то находить ее производные нежелательно. В простейшем случае, подставляя (14) в (13) и ограничиваясь только первым членом разложения, получим *схему ломаных* *):

$$y_{n+1} = y_n + h_n f(x_n,\ y_n), \quad h_n = x_{n+1} - x_n. \tag{15}$$

Поскольку при такой замене можно найти только приближенные значения искомой функции в узлах, то будем обозначать эти значения через y_n в отличие от точных значений $u_n = u(x_n)$. Для численного расчета по схеме ломаных достаточно задать начальное значение $y_0 = \eta$. Затем по формуле (15) последовательно вычисляем величины $y_1, y_2, \ldots, y_N$.

Геометрическая интерпретация этой схемы дана на рис. 41, где изображено поле интегральных кривых. Использование только первого члена формулы Тейлора означает движение не по интегральной кривой, а по касательной к ней. На каждом шаге мы

*) Она была предложена Эйлером и называется также схемой Эйлера.

заново находим касательную; следовательно, траектория движения будет ломаной линией.

Исследуем сходимость метода ломаных, предполагая правую часть $f(x, u)$ непрерывной и ограниченной вместе со своими первыми производными: $|f| \leqslant M_1$, $|f_x| \leqslant M_2$, $|f_u| \leqslant M_3$ (отсюда следует, что $|u''| \leqslant M_4 = M_2 + M_1 M_3$).

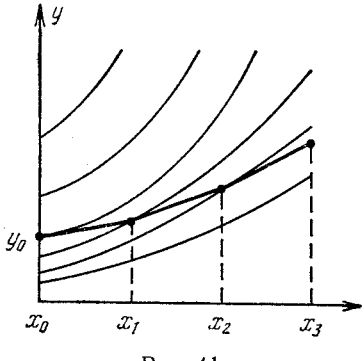


Рис. 41.

Рассмотрим погрешность приближенного решения $z_n = y_n - u_n$. Вычитая (15) из (13), получим соотношение, связывающее погрешности в соседних узлах сетки:

$$z_{n+1} = z_n + h_n [f(x_n, y_n) - f(x_n, u_n)] - \frac{1}{2} h_n^2 u_n'' = z_n (1 + h f_u)_n - \frac{1}{2} h_n^2 u_n'' \quad (16)$$

(члены более высокого порядка малости здесь опущены). Последовательно применяя рекуррентное соотношение (16), выразим погрешность на произвольном шаге через погрешность начальных данных

$$z_m = z_0 \prod_{n=0}^{m-1} (1 + h f_u)_n - \frac{1}{2} \sum_{n=0}^{m-1} h_n^2 u_n'' \prod_{k=n+1}^{m-2} (1 + h f_u)_k. \quad (17)$$

Отсюда нетрудно дать асимптотическую оценку погрешности. Заметим, что при малых шагах сетки

$$\prod_{n=0}^{m-1} (1 + h f_u)_n \approx \prod_{n=0}^{m-1} \exp (h f_x)_n = \exp \left[\sum_{n=0}^{m-1} (h f_u)_n \right] \approx \exp \left[\int_{x_0}^{x_{m-1}} f_u (\tau, u(\tau))\, d\tau \right],$$

причем в качестве верхнего предела интеграла можно взять x_m, ибо ошибка при этом остается в пределах общей точности преобразований. Аналогично преобразуя второй член (17), получим

$$z_m = z_0 \exp \left(\int_{x_0}^{x_m} f_u\, d\tau \right) - \frac{1}{2} \int_{x_0}^{x_m} d\tau\, h(\tau)\, u''(\tau) \exp \left(\int_{\tau}^{x_m} f_u\, d\mu \right). \quad (18)$$

Здесь $h(x)$ — непрерывная функция, дающая в каждом узле x_n величину шага h_n; в качестве такой функции можно выбрать линейный сплайн.

Рассмотрим структуру погрешности (18). Первое слагаемое справа связано с погрешностью начального значения $z_0 = y_0 - u_0$, которая умножается на ограниченную (благодаря ограниченности производных) величину. Начальное значение можно задать точно и считать, что $z_0 = 0$. Остановимся на втором слагаемом. Оно обусловлено тем членом формулы Тейлора (13), который был отброшен при выводе схемы ломаных (15). Оценим это слагаемое сверху; заменяя все функции под интегралами их модулями и вынося $\max h(x)$ за знак интеграла, получим

$$|z_m| \leqslant M(x_m) \max_{0 \leqslant n \leqslant m} h_n = O(\max h_n), \tag{19а}$$

где

$$M(x_m) = \frac{1}{2} \int_{x_0}^{x_m} d\tau \, |u''(\tau)| \exp\left(\int_{\tau}^{x_m} |f_u| d\mu \right) \leqslant \frac{M_4}{2M_3} e^{M_3(x_m - x_0)}. \tag{19б}$$

Таким образом, при $h \to 0$ *приближенное решение сходится к точному равномерно* (на ограниченном отрезке $|x - x_0| \leqslant a$) *с первым порядком точности.*

Замечание 1. Оценка погрешности (19) является мажорантной. Для функций со знакопеременными производными эта оценка может быть сильно завышена по сравнению с асимптотической оценкой (18).

Замечание 2. Экспоненциальный член в оценке (18) характеризует расхождение интегральных кривых (см. рис. 41); если он очень велик, то исходная задача Коши плохо обусловлена.

Пример. Проинтегрируем по схеме Эйлера задачу Коши для уравнения (3):

$$u'(x) = x^2 + u^2, \quad 0 \leqslant x \leqslant 1, \quad u(0) = 0.$$

В таблице 18 даны численные решения $y(x)$, полученные на сетках с шагами $h = 1$, $1/2$ и $1/4$; столбик $\tilde{y}(x)$ будет пояснен в п. 10. Приведено также точное решение $u(x)$, вычисленное методом Пикара (см. пример в п. 3). Видно, что схема Эйлера для получения удовлетворительной точности требует гораздо более малого шага, чем использованный здесь.

Таблица 18

x_n	y_n			$\tilde{y}_n$	$u(x)$
	$h=1$	$h=0{,}5$	$h=0{,}25$	$h=0{,}25$	
0,00	0,000	0,000	0,000	0,000	0,000
0,25	—	—	0,000	0,008	0,005
0,50	—	0,000	0,016	0,031	0,042
0,75	—	—	0,078	0,114	0,143
1,00	0,000	0,125	0,220	0,316	0,350

Ограничимся только написанными членами, так как уже они обеспечивают четвертый порядок точности. Для вычисления решения в следующей точке запишем дифференциальное уравнение в интегральной форме

$$u_{n+1} = u_n + \int_{x_n}^{x_{n+1}} f(x,\ u(x))\,dx = u_n + \int_{x_n}^{x_{n+1}} F(x)\,dx \qquad (29)$$

и подставим в него интерполяционный многочлен (28). Получим формулу Адамса для переменного шага

$$y_{n+1} = y_n + h_n F(x_n) + \frac{1}{2} h_n^2 F(x_n,\ x_{n-1}) +$$
$$+ \frac{1}{6} h_n^2 (2h_n + 3h_{n-1}) F(x_n,\ x_{n-1},\ x_{n-2}) +$$
$$+ \frac{1}{12} h_n^2 (3h_n^2 + 8h_n h_{n-1} + 4h_n h_{n-2} + 6h_{n-1}^2 + 6h_{n-1}h_{n-2}) \times$$
$$\times F(x_n,\ x_{n-1},\ x_{n-2},\ x_{n-3}),\ \text{где}\ h_n = x_{n+1} - x_n. \qquad (30)$$

Эта формула имеет четвертый порядок точности. Если отбросить последнее слагаемое, получим формулу третьего порядка точности. Аналогично получаются формулы низших порядков. Формула первого порядка совпадает со схемой ломаных.

Чаще пользуются менее громоздким вариантом формулы (30), рассчитанным на постоянный шаг интегрирования. Вместо разделенных разностей вводят конечные разности $\Delta^p F_n = p!\,F(x_n,\ x_{n-1},\ \ldots,\ x_{n-p})$, приблизительно равные p-й производной в точке $(x_n + x_{n-p})/2$, и получают

$$y_{n+1} = y_n + hF_n + \frac{1}{2} h^2 \Delta^1 F_n + \frac{5}{12} h^3 \Delta^2 F_n + \frac{3}{8} h^4 \Delta^3 F_n. \qquad (31)$$

Остаточный член этой формулы равен $(251/750)\,h^5 F^{\mathrm{IV}}(x)$.

Метод без изменений переносится на системы уравнений первого порядка типа (25).

Чтобы начать расчет методом Адамса, недостаточно знать $y(x_0)$. Для начала расчета по формуле (30) надо знать величину решения в четырех точках x_0, x_1, x_2, x_3 (а при формуле p-го порядка точности — в p точках). Поэтому надо вычислить недостающие значения y_n каким-либо другим методом — методом Рунге — Кутта, или разложением по формуле Тейлора (13) — (14) с достаточно большим числом членов. При работе на ЭВМ это вдвое увеличивает объем программы. Кроме того, формулы (30) громоздки, а несложные формулы (31) рассчитаны только на постоянный шаг и требуют нестандартных действий при смене шага: надо перейти к формулам (30), сделать по ним четыре шага и снова вернуться

(21), *равномерно сходится к точному решению с погрешностью* $O(\max h_n^3)$, т. е. *двучленная схема Рунге—Кутта имеет второй порядок точности.*

Формула (21) имеет неплохую точность и нередко используется в численных расчетах. При этом обычно полагают либо

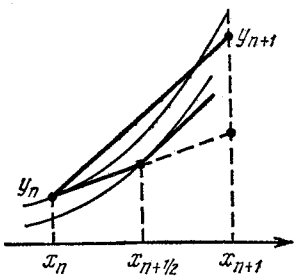


Рис. 42.

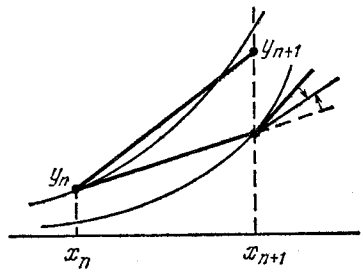


Рис. 43.

$\alpha = 1$, либо $\alpha = 1/2$. В первом случае получается схема особенно простого вида

$$y_{n+1} = y_n + hf\left(x_n + \frac{1}{2}h,\ y_n + \frac{1}{2}hf_n\right). \tag{22}$$

Ее смысл поясняется рис. 42. Сначала делаем половинный шаг по схеме ломаных, находя $y_{n+1/2} = y_n + \frac{1}{2}hf_n$. Затем в найденной точке определяем наклон интегральной кривой $y'_{n+1/2} = f(x_{n+1/2},\ y_{n+1/2})$. По этому наклону определяем приращение функции на целом шаге $y_{n+1} = y_n + hy'_{n+1/2}$.

Геометрическая интерпретация второго случая

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n,\ y_n) + f(x_n + h,\ y_n + hf_n)] \tag{23}$$

изображена на рис. 43. Здесь мы сначала грубо вычисляем по схеме ломаных значение функции $\bar{y}_{n+1} = y_n + hf_n$ и наклон интегральной кривой $\bar{y}'_{n+1} = f(x_{n+1},\ \bar{y}_{n+1})$ в новой точке. Затем находим средний наклон на шаге $y'_{n+1/2} = (y'_n + \bar{y}'_{n+1})/2$ и по нему уточняем значение y_{n+1}. Схемы подобного типа нередко называют «предиктор — корректор».

Таблица 19

x_n	y_n		$\tilde{y}_n$	$u(x)$
	$h = 1$	$h = 0{,}5$	$h = 0{,}5$	
0,00	0,000	0,000	0,000	0,000
0,50	—	0,031	0,042	0,042
1,00	0,250	0,317	0,339	0,350

В таблице 19 приведен численный расчет по схеме (22) того же примера, который рассмотрен в таблице 18 (п.5). Из таблицы видно, что схема второго порядка точности дает существенно лучшие результаты, чем схема ломаных; уже расчет на грубой сетке с $h=0{,}5$ можно считать удовлетворительным.

Методом Рунге—Кутта можно строить схемы различного порядка точности. Например, схема ломаных (15) есть схема Рунге—Кутта первого порядка точности. Наиболее употребительны схемы четвертого порядка точности, образующие семейство четырехчленных схем. Приведем без вывода ту из них, которая записана в большинстве стандартных программ ЭВМ:

$$y_{n+1}=y_n+\frac{h}{6}(k_1+2k_2+2k_3+k_4),$$

$$k_1=f(x_n,\ y_n),\quad k_2=f\left(x_n+\frac{h}{2},\ y_n+\frac{h}{2}k_1\right),\qquad (24)$$

$$k_3=f\left(x_n+\frac{h}{2},\ y_n+\frac{h}{2}k_2\right),\quad k_4=f(x_n+h,\ y_n+hk_3)$$

(при величинах k_m и шаге h следует также ставить индекс сетки n, но для простоты записи мы его опускаем).

Формулы более высокого порядка точности практически не употребляются. Пятичленные формулы имеют всего лишь четвертый порядок точности; шестичленные имеют шестой порядок, но слишком громоздки. Кроме того, высокий порядок реализуется лишь при наличии у правой части непрерывных производных соответствующего порядка.

Схемы Рунге—Кутта имеют ряд важных достоинств. 1) Все они (кроме схемы ломаных) имеют хорошую точность. 2) Они являются явными, т. е. значение y_{n+1} вычисляется по ранее найденным значениям за определенное число действий по определенным формулам. 3) Все схемы допускают расчет переменным шагом; значит, нетрудно уменьшить шаг там, где функция быстро меняется, и увеличить его в обратном случае. 4) Для начала расчета достаточно выбрать сетку x_n и задать значение $y_0=\eta$; далее вычисления идут по одним и тем же формулам. Все эти свойства схем очень ценны при расчетах на ЭВМ.

На случай систем уравнений схемы Рунге—Кутта легко переносятся, как во всех других методах, при помощи формальной замены y, $f(x,\ y)$ на $\boldsymbol{y}$, $\boldsymbol{f}(x,\ \boldsymbol{y})$. Нетрудно произвести покомпонентную запись этих схем. Например, для системы двух уравнений

$$u'(x)=f(x,\ u(x),\ v(x)),$$
$$v'(x)=g(x,\ u(x),\ v(x)),\qquad (25)$$

обозначая через y, z приближенные значения функций $u(x)$,

$v(x)$, запишем аналогичную (24) четырехчленную схему следующим образом:

$$y_{n+1} = y_n + \frac{1}{6} h (k_1 + 2k_2 + 2k_3 + k_4),$$
$$z_{n+1} = z_n + \frac{1}{6} h (q_1 + 2q_2 + 2q_3 + q_4), \tag{26а}$$

где

$$k_1 = f(x_n, y_n, z_n), \quad q_1 = g(x_n, y_n, z_n),$$
$$k_2 = f\left(x_n + \frac{1}{2} h, \; y_n + \frac{1}{2} hk_1, \; z_n + \frac{1}{2} hq_1\right),$$
$$q_2 = g\left(x_n + \frac{1}{2} h, \; y_n + \frac{1}{2} hk_1, \; z_n + \frac{1}{2} hq_1\right),$$
$$k_3 = f\left(x_n + \frac{1}{2} h, \; y_n + \frac{1}{2} hk_2, \; z_n + \frac{1}{2} hq_2\right), \tag{26б}$$
$$q_3 = g\left(x_n + \frac{1}{2} h, \; y_n + \frac{1}{2} hk_2, \; z_n + \frac{1}{2} hq_2\right),$$
$$k_4 = f(x_n + h, \; y_n + hk_3, \; z_n + hq_3),$$
$$q_4 = g(x_n + h, \; y_n + hk_3, \; z_n + hq_3).$$

Напомним, что именно эта схема четвертого порядка точности (разумеется, записанная для системы произвольного числа уравнений) лежит в основе большинства стандартных программ численного решения задачи Коши на ЭВМ.

Замечание. Погрешности различных схем Рунге—Кутта связаны с максимумами модулей соответствующих производных $f(x, u)$ громоздкими выражениями типа (18)—(19). Наглядное представление о величине этих погрешностей можно получить в одном частном случае, когда $f = f(x)$. При этом дифференциальное уравнение сводится к квадратуре, а все схемы численного интегрирования переходят в квадратурные формулы. Легко убедиться, что схема (22) переходит в формулу средних (4.16), схема (23)—в формулу трапеций (4.7) с шагом h, а схема (24) — в формулу Симпсона (4.11) с шагом $h/2$. Напомним, что мажоранты остаточных членов этих формул на равномерной сетке с указанными шагами соответственно равны

$$R_{\text{сред}} = \frac{b-a}{24} h^2 \max |f''|, \quad R_{\text{трап}} = \frac{b-a}{12} h^2 \max |f''|,$$
$$R_{\text{симп}} = \frac{b-a}{2880} h^4 \max |f^{\text{IV}}|. \tag{27}$$

Численные коэффициенты в остаточных членах (27) малы; это является одной из причин хорошей точности схем Рунге—Кутта.

Какими из формул Рунге—Кутта целесообразно пользоваться в каждом конкретном случае и как выбирать шаг сетки?

Если правая часть дифференциального уравнения непрерывна и ограничена вместе со своими четвертыми производными (и эти производные не слишком велики), то хорошие результаты дает схема четвертого порядка (24) благодаря очень малому коэффициенту в остаточном члене и быстрому возрастанию точности при уменьшении шага. Если же правая часть не имеет указанных производных, то предельный порядок точности этой схемы не может реализоваться. Тогда не худшие (хотя, по-видимому, и не лучшие) результаты дают схемы меньшего порядка точности, равного порядку имеющихся производных; например, для двукратно непрерывно дифференцируемых правых частей — несложные схемы (21)—(23).

Шаг сетки следует выбирать настолько малым, чтобы обеспечить требуемую точность расчета; других ограничивающих шаг условий в методе Рунге—Кутта нет. Но выражения остаточных членов типа (18)—(19) слишком громоздки; поэтому априорными оценками точности для выбора шага в практических расчетах не пользуются. Удобнее делать расчеты со сгущением сетки, давая апостериорную оценку точности (подробнее это будет рассмотрено в п. 11).

Встречаются важные задачи, в которых функции являются достаточно гладкими, но настолько быстро меняющимися, что схемы Рунге—Кутта как низкого, так и высокого порядка точности требуют неприемлемо малого шага для получения удовлетворительного результата. Такие задачи требуют использования (а нередко — разработки) специальных методов, ориентированных на данный узкий класс задач.

7. Метод Адамса. Будем рассматривать правую часть уравнения $f(x, u)$ не на всей плоскости ее аргументов x, u, а только на определенной интегральной кривой $u(x)$, соответствующей искомому решению. Тогда она будет функцией только одного аргумента x; обозначим ее через

$$F(x) \equiv f(x, u(x)).$$

Пусть нам уже известно приближенное решение в нескольких точках сетки: y_n, y_{n-1}, ..., y_{n-m}. Тогда в этих точках известны также $F(x_k) = f(x_k, y_k)$. В окрестности этих узлов можно приближенно заменить $F(x)$ интерполяционным многочленом; запишем его для неравномерной сетки в форме Ньютона (2.8):

$$F(x) = F(x_n) + (x - x_n) F(x_n, x_{n-1}) + \\ + (x - x_n)(x - x_{n-1}) F(x_n, x_{n-1}, x_{n-2}) + \\ + (x - x_n)(x - x_{n-1})(x - x_{n-2}) F(x_n, x_{n-1}, x_{n-2}, x_{n-3}) + \ldots \quad (28)$$

Ограничимся только написанными членами, так как уже они обеспечивают четвертый порядок точности. Для вычисления решения в следующей точке запишем дифференциальное уравнение в интегральной форме

$$u_{n+1} = u_n + \int_{x_n}^{x_{n+1}} f(x,\ u(x))\,dx = u_n + \int_{x_n}^{x_{n+1}} F(x)\,dx \qquad (29)$$

и подставим в него интерполяционный многочлен (28). Получим формулу Адамса для переменного шага

$$y_{n+1} = y_n + h_n F(x_n) + \frac{1}{2} h_n^2 F(x_n,\ x_{n-1}) +$$
$$+ \frac{1}{6} h_n^2 (2h_n + 3h_{n-1}) F(x_n,\ x_{n-1},\ x_{n-2}) +$$
$$+ \frac{1}{12} h_n^2 (3h_n^2 + 8h_n h_{n-1} + 4h_n h_{n-2} + 6h_{n-1}^2 + 6h_{n-1}h_{n-2}) \times$$
$$\times F(x_n,\ x_{n-1},\ x_{n-2},\ x_{n-3}),\ \text{где}\ h_n = x_{n+1} - x_n. \qquad (30)$$

Эта формула имеет четвертый порядок точности. Если отбросить последнее слагаемое, получим формулу третьего порядка точности. Аналогично получаются формулы низших порядков. Формула первого порядка совпадает со схемой ломаных.

Чаще пользуются менее громоздким вариантом формулы (30), рассчитанным на постоянный шаг интегрирования. Вместо разделенных разностей вводят конечные разности $\Delta^p F_n = p!\,F(x_n,\ x_{n-1},\ \ldots,\ x_{n-p})$, приблизительно равные p-й производной в точке $(x_n + x_{n-p})/2$, и получают

$$y_{n+1} = y_n + hF_n + \frac{1}{2} h^2 \Delta^1 F_n + \frac{5}{12} h^3 \Delta^2 F_n + \frac{3}{8} h^4 \Delta^3 F_n. \qquad (31)$$

Остаточный член этой формулы равен $(251/750)\,h^5 F^{\mathrm{IV}}(x)$.

Метод без изменений переносится на системы уравнений первого порядка типа (25).

Чтобы начать расчет методом Адамса, недостаточно знать $y(x_0)$. Для начала расчета по формуле (30) надо знать величину решения в четырех точках x_0, x_1, x_2, x_3 (а при формуле p-го порядка точности — в p точках). Поэтому надо вычислить недостающие значения y_n каким-либо другим методом — методом Рунге — Кутта, или разложением по формуле Тейлора (13) — (14) с достаточно большим числом членов. При работе на ЭВМ это вдвое увеличивает объем программы. Кроме того, формулы (30) громоздки, а несложные формулы (31) рассчитаны только на постоянный шаг и требуют нестандартных действий при смене шага: надо перейти к формулам (30), сделать по ним четыре шага и снова вернуться

к формулам (31). Все это делает метод Адамса неудобным для расчетов на ЭВМ.

Внешне этот метод привлекателен тем, что за один шаг приходится только один раз вычислять $f(x, u)$, которая может быть очень сложной. А в четырехчленной схеме Рунге — Кутта того же порядка точности $f(x, u)$ вычисляется за шаг четыре раза. Однако коэффициент в остаточном члене (27) схемы Рунге — Кутта (24) меньше в 960 раз, чем в схеме (31)! Значит, при одинаковой точности схема Рунге — Кутта (24) позволяет брать шаг в $\sqrt[4]{960} = 5{,}7$ раза крупнее, т. е. фактически вычислять $f(x, u)$ даже меньшее число раз, чем в методе Адамса.

Поэтому сейчас метод Адамса и аналогичные методы (например, Милна) употребляются реже метода Рунге — Кутта.

8. Неявные схемы. Предыдущие методы были явными, т. е. значение y_{n+1} определялось за заранее известное число действий. Пример неявной схемы получим, если запишем дифференциальное уравнение в интегральной форме (29), а интеграл по одному интервалу сетки приближенно вычислим по формуле трапеций

$$y_{n+1} = y_n + \frac{1}{2} h [f(x_n, y_n) + f(x_{n+1}, y_{n+1})]. \qquad (32)$$

Решая это алгебраическое уравнение, можно определить y_{n+1}, которое и будет приближенным значением искомого решения $u(x_n)$. Схема (32) имеет второй порядок точности, допускает счет неравномерным шагом, не требует специальных приемов для начала счета.

Но у этой схемы есть серьезные недостатки. Во-первых, неизвестно, имеет ли уравнение (32) вещественный корень, т. е. разрешима ли задача. Можно привести пример, когда при большом шаге корня нет. Пусть $f(x, u) = u^2$ и $u(0) = 1$; тогда на первом шаге $y_1 = 1 + \frac{1}{2} h (1 + y_1^2)$ и при $h > (1 + \sqrt{2})^{-1}$ вещественного корня нет.

Во-вторых, даже если корень есть, то как его найти? Метод Ньютона применять нежелательно, так как для этого надо дифференцировать $f(x, u)$. Метод деления пополам не обобщается на системы уравнений. Остается метод последовательных приближений

$$y_{n+1}^{(s)} = y_n + \frac{1}{2} h [f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(s-1)})]. \qquad (33)$$

Однако он сходится к корню, только если $h|f_u| < 2$, т. е. при достаточно малом шаге. Если в ходе расчета f_u возрастает, то итерации (33) могут перестать сходиться.

От последней трудности можно избавиться, заодно уменьшив объем вычислений. Для этого ограничим заранее число итераций

и рассмотрим (33) как самостоятельную явную схему. Очевидно, вопроса о существовании корня при этом не возникает; y_{n+1} всегда определяется, даже если алгебраическое уравнение (32) вещественного корня не имеет.

Роль числа итераций хорошо видна на примере уравнения $u' = f(u)$. Естественное нулевое приближение есть $y_{n+1}^{(0)} = y_n$, так что первая и вторая итерации

$$y_{n+1}^{(1)} = y_n + hf(y_n).$$

$$y_{n+1}^{(2)} = y_n + \frac{1}{2} h[f(y_n) + f(y_n + hf_n)]$$

являются соответственно схемой ломаных (15) первого порядка точности и схемой Рунге — Кутта второго порядка точности (23) типа «предиктор — корректор». Дальнейшие итерации уже не увеличат порядка точности, так как он не может быть выше, чем в исходной схеме (32); они влияют только на коэффициенты в остаточном члене и увеличивают время счета.

Таким образом, неявные схемы с заданным числом итераций мало отличаются от схем Рунге — Кутта и бывают удобны лишь для некоторых нестандартных задач. Но они приводят к интересной идее ограничения числа итераций.

Есть эмпирическое правило, в общем случае не обоснованное. Пусть для решения дифференциального уравнения написана неявная схема p-го порядка точности. Разрешим ее методом последовательных приближений аналогично (33) и зададим число итераций. Тогда при одной итерации получим схему первого порядка точности, при двух — второго и так далее, при p итерациях — p-го порядка точности. Дальнейшее увеличение числа итераций уже не увеличивает порядок точности.

Это правило оказывается полезным даже для уравнений *в частных производных*. По существу схема с заданным числом итераций есть новая явная схема. Поэтому здесь не возникает вопроса о существовании корня или сходимости итераций, подобных (33).

9. Специальные методы. Из всех численных методов интегрирования обыкновенных дифференциальных уравнений, рассчитанных на произвольные уравнения (точнее, на классы уравнений, у которых правые части имеют определенное число непрерывных и ограниченных производных), наилучшие результаты и при расчетах на ЭВМ, и при ручных расчетах дают методы Рунге — Кутта. Поэтому, приступая к решению какой-либо конкретной задачи Коши, обычно пробуют решить ее одной из описанных в п. 6 схем.

Но выше отмечалось, что встречаются задачи с быстропеременными решениями, когда все схемы Рунге — Кутта для получения удовлетворительной точности требуют неприемлемо малого

шага. Характерным примером такой задачи является система уравнений химической кинетики.

Сначала разберем задачу химического распада одного вещества

$$u'(t) = -\alpha(t, u)\,u, \quad u(0) = u_0 > 0, \quad \alpha(t, u) > 0; \tag{34}$$

здесь u — концентрация вещества, t — время, α — скорость распада, которую считаем зависящей от t и u (ибо она зависит от температуры, а температура определяется выделением тепла при реакции и внешними условиями охлаждения). Запишем для уравнения (34) схему ломаных (15)

$$y_{n+1} = y_n\,[1 - \tau\alpha(t_n, y_n)], \quad \tau = t_{n+1} - t_n. \tag{35}$$

По смыслу задачи, концентрация вещества должна быть положительной. Но если скорость распада настолько велика, что хотя бы в одной точке $\alpha_n > 1/\tau$, то численное решение (35) будет знакопеременным, что физически бессмысленно. Применение вместо (15) схем Рунге — Кутта более высокого порядка точности лишь немного ослабляет указанное ограничение шага, не устраняя его (см. задачу 7).

Для одного уравнения (34) эта трудность несущественна: если скорость распада $\alpha(t, u)$ велика, то вещество распадается за малое время $t \sim \alpha^{-1}$, так что число шагов сетки $N = t/\tau$ будет умеренным. Но в системах уравнений химической кинетики присутствуют вещества с самыми различными константами распада или синтеза (нередко от $\alpha \sim 10^8$ сек$^{-1}$ до $\alpha \sim 10^{-2}$ сек$^{-1}$). Тогда общий промежуток времени будет определяться самой медленной реакцией ($t \sim 100$ сек), а допустимый шаг интегрирования — самой быстрой реакцией ($\tau \sim 10^{-8}$ сек). Ясно, что такой объем расчетов — около 10^{10} шагов — совершенно неприемлем *).

Для подобных задач приходится использовать специальные методы, разработанные именно для данных узких классов уравнений; для других классов уравнений эти методы обычно оказываются непригодными. Способы построения специальных методов основаны на изучении и использовании свойств общих решений исследуемого класса уравнений. Рассмотрим некоторые способы.

Большинство способов основано на том, что для *исходного* уравнения $u'(x) = f(x, u)$ стараются найти такое *вспомогательное* уравнение $v'(x) = g(x, v)$, чтобы решение последнего возможно более просто выражалось через элементарные функции, и при этом на заметном отрезке изменения аргумента выполнялось бы $u(x) \approx v(x)$. Иными словами, ищется приближенное решение, имеющее достаточно простой вид.

*) Кроме того, при таком числе шагов существенно сказывается некорректность задачи, связанная с ошибками округления.

Для нахождения приближенных решений можно применить метод Пикара или другие аналогичные методы. Нередко удается добиться успеха, слегка упрощая правую часть исходного уравнения. Например, если в задаче (34) положить $\alpha(t, u) \approx \alpha_0 = \text{const}$, тогда вспомогательное уравнение будет $v'(t) = -\alpha_0 v$, а его решением при заданном начальном условии является $v(t) = u_0 \times \times \exp[-\alpha_0 (t - t_0)]$.

Первый способ построения специальных схем удобен для знакопеременных решений (например, быстро осциллирующих). В нем рассматривается разность $w(x) = u(x) - v(x)$. Вычитая вспомогательное уравнение из исходного, получим уравнение, которому удовлетворяет эта разность

$$w'(x) = f(x, v(x) + w) - g(x, v(x)); \tag{36}$$

здесь $v(x)$ — известная функция. Если $v(x)$ действительно является хорошим приближением к решению, то функция $w(x)$ невелика, поэтому уравнение (36) должно легко интегрироваться обычными схемами Рунге — Кутта.

Второй способ выгоден для знакопостоянных решений (например, растущих по экспоненциальному или степенному закону). В нем рассматривается отношение $w(x) = u(x)/v(x)$, а для систем уравнений — отношения $w_k(x) = u_k(x)/v_k(x)$. Нетрудно убедиться, что это отношение удовлетворяет уравнению

$$w'(x) = \frac{1}{v(x)}[f(x, wv(x)) - wg(x, v(x))], \tag{37}$$

где $v(x)$ — известное приближенное решение. Аналогично предыдущему случаю, полученное уравнение должно хорошо интегрироваться численно схемами Рунге — Кутта.

Пример. Если для уравнения распада (34) воспользоваться приближенным решением $v(t) = u_0 \exp[-\alpha_0 (t - t_0)]$, то специальная схема (37) примет следующий вид:

$$w'(t) = -[\alpha_0 - \alpha(t, wv(t))]w;$$

при слабо меняющейся $\alpha(t, u)$ малость правой части очевидна.

Третий способ заключается в том, что вспомогательное уравнение рассматривается не на большом промежутке изменения аргумента, а на одном шаге сетки $x_n \leqslant x \leqslant x_{n+1}$. Берется его приближенное решение $v_n(x)$, удовлетворяющее начальному условию $v_n(x_n) = y_n \approx u(x_n)$. Поскольку интервал сетки невелик, то на нем приближенное решение будет близко к точному, поэтому можно положить $u(x_{n+1}) \approx y_{n+1} = v_n(x_{n+1})$. Этот способ означает написание такой разностной схемы, которой решение вспомогательного уравнения удовлетворяет точно, а решение исходного уравнения — приближенно, но с малой погрешностью.

Пример. Рассмотрим уравнение, возникающее в задачах так называемой дифференциальной прогонки:

$$u'(x) = -[u^2 + \rho(x)], \quad \rho(x) > 0. \tag{38}$$

Если положить $\rho(x) \approx \text{const} = \rho_{n+1/2}$ при $x_n \leqslant x \leqslant x_{n+1}$, то вспомогательное уравнение примет вид

$$v'(x) = -(v^2 + \rho), \quad \rho = \text{const}.$$

Оно интегрируется в элементарных функциях

$$\operatorname{arctg} \frac{v_{n+1}}{\sqrt{\rho}} - \operatorname{arctg} \frac{v_n}{\sqrt{\rho}} = -h\sqrt{\rho}.$$

Это соотношение явно разрешается, давая такую специальную схему:

$$y_{n+1} = \sqrt{\rho_{n+1/2}}\, \frac{y_n - \sqrt{\rho_{n+1/2}}\, \operatorname{tg}\left(h\sqrt{\rho_{n+1/2}}\right)}{\sqrt{\rho_{n+1/2}} + y_n \operatorname{tg}\left(h\sqrt{\rho_{n+1/2}}\right)}. \tag{39а}$$

Если можно считать $h\sqrt{\rho} \ll 1$, то схема (39а) упрощается:

$$y_{n+1} = \frac{y_n - h\rho_{n+1/2}}{1 + hy_n}. \tag{39б}$$

Схемы (39а) и (39б) дают неплохие результаты даже в тех случаях, когда условие устойчивости прогонки нарушено, а точное решение задачи (38) имеет полюсы.

При использовании третьего способа обычно удается построить схемы первого или второго порядка точности, но с малым остаточным членом (точнее, мала по величине комбинация производных, входящая множителем в остаточный член); схемы более высокого порядка точности построить этим путем трудно. Первый и второй способы позволяют использовать схемы Рунге — Кутта высокого порядка точности, но остаточный член при этом будет не очень мал, ибо решения $u(x)$ и $v(x)$ на большом отрезке изменения аргумента могут заметно отличаться, и правые части уравнений (36) или (37) становятся большими. Однако первый и второй способы также можно применить к одному интервалу сетки; на этом пути можно построить специальные схемы высокого порядка точности с малым остаточным членом. Заметим, что все эти способы по существу эквивалентны специально подобранным нелинейным интерполяциям искомого решения.

Упомянем четвертый способ, заключающийся в построении так называемых точных разностных схем, которым точно удовлетворяет решение исходной задачи. Коэффициенты таких

схем обычно являются функционалами от коэффициентов исходного уравнения (и могут зависеть также от искомого решения). Но техника построения точных схем более сложна, и мы их не будем рассматривать, отсылая читателя к монографии [30].

10. Особые точки. Решение может иметь в отдельных точках отрезка интегрирования особенности, обусловленные обращением в бесконечность правой части $f(x, u)$ или какой-нибудь ее производной. Сначала рассмотрим случай, когда начальная точка $x = x_0$ является особой. Есть три основных способа численного интегрирования таких решений. Рассмотрим их на примере задачи

$$u'(x) = \frac{1}{2\sqrt{x}} + u^2(x), \quad u(0) = 0, \tag{40}$$

где правая часть в начальной точке обращается в бесконечность; очевидно, начинать интегрирование по схеме Рунге — Кутта любого порядка точности при этом невозможно.

Первый способ — это найти такую замену переменных, которая преобразует уравнение к виду, не имеющему особенностей. Для задачи (40) достаточно сделать замену аргумента $x = t^2$; тогда эта задача принимает вид

$$\frac{du}{dt} = 1 + 2tu^2, \quad u(0) = 0,$$

который допускает применение стандартных численных методов.

Второй способ — построить в небольшой окрестности особой точки приближенное решение, выраженное через элементарные (или другие легко вычисляющиеся) функции. Например, выбирая нулевое приближение $y_0(x) \equiv 0$ и применяя к задаче (40) метод Пикара, получим

$$y_1(x) = \sqrt{x}, \quad y_2(x) = \sqrt{x} + \frac{1}{2}x^2, \ \ldots$$

Отступим от особой точки на конечное расстояние в некоторую точку x_1 и вычислим в ней решение с требуемой точностью на основе найденного приближения. Точка x_1 уже не особая; ее можно считать первым узлом разностной сетки и вести из нее интегрирование стандартными численными методами.

Следует помнить, что если точка x_1 лежит близко к x_0, то правая часть уравнения или ее производные еще велики в этой точке и стандартные численные методы дают заметную погрешность вблизи точки x_1. Поэтому желательно выбирать точку x_1 подальше от x_0. Но тогда, чтобы вычислить $u(x_1)$ с нужной точностью, необходимо строить достаточно хорошее приближенное решение: например, брать высокие приближения метода Пикара.

Третий способ — составить для данной задачи специальную схему, позволяющую вести численное интегрирование непосредственно от особой точки. Например, проинтегрируем уравнение (40) по одному интервалу сетки, и первое слагаемое в подынтегральном выражении проинтегрируем точно, а второе — по формуле прямоугольников с использованием левого конца интервала; тогда получим

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} \left[\frac{1}{2\sqrt{\xi}} + y^2(\xi)\right] d\xi \approx$$

$$\approx y_n + (\sqrt{x_{n+1}} - \sqrt{x_n}) + (x_{n+1} - x_n)\, y_n^2. \qquad (41)$$

Это явная схема, напоминающая схему ломаных (15). Она построена по образцу схем первого порядка точности. Но имеет ли эта схема на самом деле точность $O(h)$ — заранее не очевидно, ибо производные правой части уравнения (40) не ограничены; этот вопрос требует дополнительного исследования.

Если решение имеет особенности во внутренних точках отрезка интегрирования, то при этом обычно нельзя сказать заранее, в каких именно точках: правая часть $f(x, u)$ зависит от решения, которое нам не известно. В этом случае целесообразно применять третий способ — составлять специальные схемы, не теряющие своей применимости вблизи особых точек. Примером является схема (39), позволяющая вести сквозной расчет даже при наличии у решения особенностей типа полюсов.

11. Сгущение сетки. Как получить требуемую точность расчета? Априорные оценки точности для этого мало полезны. Во-первых, остаточные члены выражаются через производные решения, которое до начала расчета не известно. Во-вторых, априорные оценки обычно являются мажорантными и могут во много раз превосходить фактическую ошибку расчета.

Имеются стандартные программы численного интегрирования дифференциальных уравнений с так называемым «автоматическим выбором шага». В них каждый шаг выбирается так, чтобы вносимая на нем погрешность не превышала заданной величины. Но при этом не учитывается, что эта погрешность в ходе дальнейших расчетов умножается на величину типа экспоненты в (18), т. е. может сильно возрасти. Кроме того, общее число шагов заранее не определено. В результате фактическая точность расчета по подобным программам обычно неизвестна.

Поэтому основным практическим приемом является апостериорная оценка точности. Для ее получения расчет проводят на двух или более сгущающихся сетках и применяют правило Рунге или Рунге — Ромберга (см. главу III, п. 3). Напомним, в чем оно заключается.

Вспомним априорную оценку погрешности схемы ломаных (18). Запишем ее, опуская первое слагаемое, связанное с неточным

заданием начальных данных:

$$z(x) = \frac{1}{2}\int_{x_0}^{x} d\tau h(\tau)\, u''(\tau) \exp\left(\int_{\tau}^{x} f_u\, d\mu\right); \qquad (42)$$

здесь $h(x)$ есть некоторая функция, значение которой в каждом узле сетки дает величину шага. Для схем более высокого порядка точности p остаточный член имеет аналогичную структуру, но содержит $h^p(x)$ и соответствующие производные решения или правой части $f(x, u)$.

Если сетка равномерная, $h(x) = h = \text{const}$, то остаточный член типа (43) для схемы p-го порядка точности пропорционален h^p. Поэтому при сгущении равномерной сетки применима оценка точности по Рунге. Если имеются численные решения на двух сетках $y(x; h)$ и $y(x; rh)$, где $r > 1$, то погрешность решения на сетке с меньшим шагом составляет

$$\Delta y(x; h) \approx \frac{y(x; h) - y(x; rh)}{r^p - 1}. \qquad (43)$$

Вместо оценки точности можно погрешность (43) прибавить к численному решению, уточнив его:

$$\tilde{y}(x; h) = y(x; h) + \frac{y(x; h) - y(x; rh)}{r^p - 1}, \qquad (44)$$

но тогда вопрос о погрешности уточненного решения остается открытым.

Приведенное рассуждение справедливо и в том случае, если сетки с разным числом узлов не равномерны, но их можно описать функциями $h(x)$, отношение которых есть $h_{\mathrm{I}}(x) / h_{\mathrm{II}}(x) = r = \text{const}$. Это выполняется, например, для квазиравномерных сеток (описанных в главе III, п. 4).

При выводе оценок типа (18) старшими членами формулы Тейлора (13) пренебрегают. Если их учесть (считая правую часть уравнения непрерывно дифференцируемой достаточное число раз), то погрешность выразится суммой, где последующие слагаемые содержат более высокие степени $h(x)$ и соответствующие производные. В этом случае можно уточнять численное решение по правилу Ромберга или по рекуррентному правилу Рунге, используя расчеты на k различных сетках. Применение этих правил эквивалентно построению некоторой схемы более высокого порядка точности $q = p + k - 1$, где p — порядок точности исходной схемы. Разумеется, фактически получить точность $O(h^q)$ можно только для q раз непрерывно дифференцируемых решений $u(x)$.

Правило Рунге применимо для сеток с любым отношением шагов r. Но используют его преимущественно для целого r, когда все узлы менее подробной сетки являются узлами более

подробной; особенно удобно сгущать сетки вдвое (рис. 44). При этом как для равномерных, так и для квазиравномерных сеток условие совпадения узлов выполняется.

В тех узлах, которые являются общими для нескольких сеток, можно уточнить $y(x)$ непосредственно по правилу Рунге (44). Так, в n-м узле можно увеличить порядок точности на двойку, в $(n+2)$-м — на единицу, а в $(n+1)$-м — нельзя увеличить (рис. 44). Разумеется, если мы не уточняем решение, а лишь оцениваем погрешность, то достаточно найти ее по формуле (43) только в части узлов.

Рис. 44.

Однако можно уточнить функцию во всех узлах самой подробной сетки, если немного усложнить вычисления. Например, для двух нижних сеток на рис. 44 это делается так. Используем совпадающие узлы сеток для определения поправок к значениям функции

$$\Delta_m = [y(x_m;\ h) - y(x_m;\ rh)]/(r^p - 1), \quad m = n,\ n+2. \tag{45а}$$

Значение поправок в остальных узлах найдем простейшей интерполяцией. Для равномерных или квазиравномерных сеток можно положить

$$\Delta_{n+1} = \frac{1}{2}(\Delta_n + \Delta_{n+2}). \tag{45б}$$

Затем вычислим уточненные значения

$$\tilde{y}(x_m;\ h) = y(x_m;\ h) + \Delta_m,\ m = n,\ n+1,\ n+2. \tag{45в}$$

Этот способ легко обобщается на произвольное число сеток. Такое уточнение выгодно для специальных схем третьего типа, имеющих невысокий порядок точности; выполнить уточнение обычно проще, чем составить специальную схему высокого порядка точности.

Примеры применения правила Рунге даны в таблице 18 (п. 5) и таблице 19 (п. 6), содержащих численное решение задачи

$$u' = x^2 + u^2, \quad 0 \leqslant x \leqslant 1, \quad u(0) = 0.$$

В таблице 18 интегрирование выполнено по схеме ломаных (15), и для уточнения использованы сетки с $h = 1$ и $h = 0{,}5$; видно, что, несмотря на плохую точность исходной схемы, уточненное решение не сильно отличается от искомого. В таблице 19 уточнено численное решение, найденное по неплохой схеме Рунге — Кутта второго порядка точности (22); это уточнение уже близко к искомому решению, несмотря на очень грубую сетку.

§ 2. Краевые задачи

1. Постановки задач. Краевая задача — это задача отыскания частного решения системы (1а):

$$\frac{d}{dx}u_k(x) = f_k(x, u_1, u_2, \ldots, u_p), \quad 1 \leqslant k \leqslant p,$$

на отрезке $a \leqslant x \leqslant b$, в которой дополнительные условия налагаются на значения функций $u_k(x)$ более чем в одной точке этого отрезка. Очевидно, что краевые задачи возможны для систем порядка не ниже второго.

Свое первоначальное название этот тип задач получил по простейшим случаям, когда часть дополнительных условий задается на одном конце отрезка, а остальная часть — на другом (т. е. только в точках $x = a$ и $x = b$). Примером является задача нахождения статического прогиба $u(x)$ нагруженной струны с закрепленными концами

$$u''(x) = -f(x), \quad a \leqslant x \leqslant b, \quad u(a) = u(b) = 0; \tag{46}$$

здесь $f(x)$ — внешняя изгибающая нагрузка на единицу длины струны, деленная на упругость струны.

Для уравнений или систем более высокого порядка, где число дополнительных условий больше двух, постановки краевых условий более разнообразны. При этом возможны случаи, когда часть условий задана во внутренних точках отрезка $[a, b]$; их нередко называют внутренними краевыми условиями. Например, статический прогиб нагруженного упругого бруска удовлетворяет уравнению четвертого порядка

$$u^{IV}(x) = f(x), \quad a \leqslant x \leqslant b; \tag{47а}$$

если этот брусок лежит в точках x_i, $1 \leqslant i \leqslant 4$, на опорах, то дополнительные условия имеют вид

$$u(x_i) = 0, \ 1 \leqslant i \leqslant 4, \ a \leqslant x_1 < x_2 < x_3 < x_4 \leqslant b, \tag{47б}$$

т. е. все они заданы в разных точках.

Сами дополнительные условия могут связывать между собой значения нескольких функций в одной точке (или даже в разных точках); тогда для системы p-го порядка (1) они примут вид

$$\begin{gathered}\varphi_k(u_1(\xi_k), u_2(\xi_k), \ldots, u_p(\xi_k)) = \eta_k, \\ 1 \leqslant k \leqslant p, \quad a \leqslant \xi_k \leqslant b.\end{gathered} \tag{48}$$

Существуют задачи с еще более сложными по форме дополнительными условиями, например, условиями нормировки

$$\int_a^b u_k^2(x)\,dx = 1, \tag{49}$$

обычными в квантовой механике, и т. д.

Несмотря на разнообразие форм краевых условий, краевые задачи решаются в основном одними и теми же численными методами, что оправдывает их объединение в один тип. Остановимся на методах решения.

Найти точное решение краевой задачи в элементарных функциях удается редко: для этого надо найти общее решение системы (1) и суметь явно определить из краевых условий значения входящих в него постоянных.

К приближенным методам решения краевых задач относятся разложение в ряды Фурье, методы Ритца и Галеркина. Ряды Фурье применяют к линейным задачам; этот метод излагается в курсах математической физики (см. [2, 40]) и здесь рассматриваться не будет. Остальные два метода применимы и к некоторым нелинейным задачам. Метод Ритца разбирался в главе VII, а метод Галеркина будет рассмотрен в этом параграфе.

Для численного решения краевых задач применяют метод стрельбы и разностный метод. Метод стрельбы основан на сведении краевой задачи к некоторой задаче Коши для той же системы уравнений. В разностном методе задача приближенно заменяется решением алгебраической системы уравнений с очень большим числом неизвестных (неизвестными являются значения решения в узлах сетки). В случае нелинейных задач оба метода являются итерационными; при этом построение хорошо сходящихся итерационных процессов само оказывается достаточно сложным.

2. Метод стрельбы (называемый также *баллистическим*). Это численный метод, заключающийся в сведении краевой задачи к некоторой задаче Коши для той же системы дифференциальных уравнений. Рассмотрим его на примере простейшей задачи для системы двух уравнений первого порядка с краевыми условиями достаточно общего вида

$$u'(x) = f(x, u, v), \quad v'(x) = g(x, u, v), \quad a \leqslant x \leqslant b, \tag{50а}$$

$$\varphi(u(a), v(a)) = 0, \quad \psi(u(b), v(b)) = 0. \tag{50б}$$

Выберем произвольно значение $u(a) = \eta$, рассмотрим левое краевое условие как алгебраическое уравнение $\varphi(\eta, v(a)) = 0$ и определим удовлетворяющее ему значение $v(a) = \zeta(\eta)$. Возьмем значения $u(a) = \eta$, $v(a) = \zeta$ в качестве начальных условий задачи Коши для системы (50а) и проинтегрируем эту задачу Коши

любым численным методом (например, по схемам Рунге — Кутта). При этом получим решение $u(x;\ \eta)$, $v(x;\ \eta)$, зависящее от η, как от параметра.

Значение ζ выбрано так, что найденное решение удовлетворяет левому краевому условию (50б). Однако правому краевому условию это решение, вообще говоря, не удовлетворяет: при его подстановке левая часть правого краевого условия, рассматриваемая как некоторая функция параметра η:

$$\bar{\psi}(\eta) = \psi(u(b;\ \eta)\ v(b;\ \eta)), \tag{51}$$

не обратится в нуль. Надо каким-либо способом менять параметр η, пока не подберем такое значение, для которого $\bar{\psi}(\eta) \approx 0$ с требуемой точностью. Таким образом, решение краевой задачи (50) сводится к нахождению корня одного алгебраического уравнения

$$\bar{\psi}(\eta) = 0. \tag{52}$$

Эта алгебраическая задача изучена в главе V, § 2. Рассмотрим, какие методы ее решения целесообразно применять в данном случае.

Простейшим является *метод дихотомии*. Делают пробные «выстрелы» — расчеты с наудачу выбранными значениями η_i, до тех пор, пока среди величин $\bar{\psi}(\eta_i)$ не окажется разных по знаку. Пара таких значений η_i, η_{i+1} образует «вилку». Деля ее последовательно пополам до получения нужной точности, производим «пристрелку» параметра η. Благодаря этому процессу весь метод получил название стрельбы.

Однако нахождение каждого нового значения функции $\bar{\psi}(\eta)$ требует численного интегрирования системы (50а), т. е. достаточно трудоемко. Поэтому корень уравнения (52) желательно находить более быстрым методом, чем дихотомия.

Если правые части уравнений (50а) и левые части краевых условий (50б) имеют непрерывные и ограниченные первые производные, то $\bar{\psi}(\eta)$ также будет иметь непрерывную производную *). В этом случае можно построить аналог *метода Ньютона*. Нам пока известен только способ вычисления $\bar{\psi}(\eta)$, а нужно научиться определять также производную

$$\frac{d\bar{\psi}(\eta)}{d\eta} = \frac{\partial\psi}{\partial u(b)}\frac{\partial u(b;\ \eta)}{\partial\eta} + \frac{\partial\psi}{\partial v(b)}\frac{\partial v(b;\ \eta)}{\partial\eta}. \tag{53}$$

Входящие сюда производные по параметру от решения задачи Коши можно найти, если продифференцировать по этому параметру систему (50а). Вводя обозначения

$$\mu(x;\ \eta) = \frac{\partial u(x;\ \eta)}{\partial\eta}, \quad \nu(x;\ \eta) = \frac{\partial v(x;\ \eta)}{\partial\eta} \tag{54}$$

*) Это следует из теорем о зависимости решения задач Коши от параметра (см.[37]).

и дифференцируя (50а) по параметру, получим

$$\frac{d\mu}{dx}=f_u(x,\ u,\ v)\,\mu(x)+f_v(x,\ u,\ v)\,\nu(x),$$
$$\frac{d\nu}{dx}=g_u(x,\ u,\ v)\,\mu(x)+g_v(x,\ u,\ v)\,\nu(x),\quad a\leqslant x\leqslant b \tag{55а}$$

Одно из начальных условий для этой системы очевидно: $\mu(a)=\partial u(a)/\partial\eta=1$; второе условие нетрудно найти, дифференцируя левое краевое условие (50б) по η. Отсюда получим

$$\mu(a)=1,\quad \nu(a)=-\varphi_u(\eta,\ \zeta)/\varphi_v(\eta,\ \zeta). \tag{55б}$$

Интегрируя систему (55а) с начальными условиями (55б) совместно с задачей Коши для системы (50а), определим вспомогательные функции $\mu(x)$, $\nu(x)$. Подставляя их значения при $x=b$ в (53), найдем значение производной правого краевого условия по пристрелочному параметру. Новое значение параметра определяется по формуле касательных (5.28):

$$\eta_{s+1}=\eta_s-[\bar{\psi}(\eta_s)/\bar{\psi}'(\eta_s)]. \tag{56}$$

Однако описанный способ требует интегрирования лишней пары дифференциальных уравнений, что приводит к усложнению и двукратному увеличению трудоемкости каждой итерации. Поэтому им пользуются не часто.

Можно избежать этого усложнения, если решать уравнение (52) разностным аналогом метода Ньютона — *методом секущих*. Для этого первые два расчета делают с наудачу выбранными значениями η_0, η_1, а следующие значения параметра вычисляют по формуле (5.32):

$$\eta_{s+1}=\eta_s-\frac{(\eta_s-\eta_{s-1})\,\bar{\psi}(\eta_s)}{\bar{\psi}(\eta_s)-\bar{\psi}(\eta_{s-1})}. \tag{57}$$

Вместо этого процесса можно использовать *метод парабол*, в котором также не требуется располагать явным выражением производных, а достаточно лишь знать об их существовании. Напомним, что последние три метода быстро сходятся вблизи корня; сходимость вдали от корня зависит от того, насколько удачно выбрано нулевое приближение.

Линейные задачи решаются методом стрельбы особенно просто. Пусть система (50а) и краевые условия (50б) линейны;

$$u'(x)=\alpha_1(x)\,u+\beta_1(x)\,v+\gamma_1(x),$$
$$v'(x)=\alpha_2(x)\,u+\beta_2(x)\,v+\gamma_2(x),\quad a\leqslant x\leqslant b, \tag{58а}$$
$$p_1u(a)+q_1v(a)=r_1\qquad p_2u(b)+q_2v(b)=r_2. \tag{58б}$$

Тогда начальные условия соответствующей задачи Коши примут вид

$$u(a)=\eta,\quad v(a)=\zeta=(r_1-p_1\eta)/q_1. \tag{58в}$$

Нетрудно сообразить, что решение задачи Коши (58а), (58в) будет линейно зависеть от параметра η, поэтому $\bar{\psi}(\eta)$ также

будет линейной функцией. Но линейная функция одного аргумента полностью определяется своими значениями в любых двух точках η_0 и η_1, а ее график является прямой, т. е. совпадает со своей секущей. Значит, найденное по формуле секущих (57) значение η_2 является точным корнем уравнения (52), так что расчет с этим значением параметра даст искомое решение. Таким образом, для решения линейной краевой задачи (58а)—(58б) достаточно трижды решить задачу Коши.

Замечание. Для линейных задач можно несколько уменьшить объем расчетов, если воспользоваться тем, что общее решение линейной неоднородной системы равно сумме ее какого-нибудь частного решения и общего решения соответствующей однородной системы. Найдем частное решение неоднородной системы (58а), (58в), соответствующее значению $\eta_0 = 0$, и обозначим его через $u_0(x)$, $v_0(x)$. Затем рассмотрим соответствующую однородную задачу Коши

$$u'(x) = \alpha_1(x)\,u + \beta_1(x)\,v, \quad v'(x) = \alpha_2(x)\,u + \beta_2(x)\,v,$$
$$u(a) = \eta_1 = 1, \qquad v(a) = -p_1/q_1;$$

вычислим ее решение и обозначим его через $u_1(x)$, $v_1(x)$. Тогда общее решение неоднородной задачи Коши, удовлетворяющее (в силу выбора начальных условий) левому краевому условию (58б), является однопараметрическим семейством

$$u(x) = u_0(x) + cu_1(x), \qquad v(x) = v_0(x) + cv_1(x). \tag{59}$$

Значение параметра c выбираем так, чтобы удовлетворить правому краевому условию (58б):

$$c = -\frac{p_2u_0(b) + q_2v_0(b) - r_2}{p_2u_1(b) + q_2v_1(b)}.$$

Затем найдем искомое решение по формуле (59), что позволяет избежать третьего интегрирования задачи Коши.

Метод стрельбы прост, применим как к линейным, так и к нелинейным задачам и позволяет использовать при численном интегрировании схемы Рунге—Кутта (или другие) высокого порядка точности. К большинству задач типа (50) он применяется успешно.

Затруднения возникают в тех случаях, когда краевая задача (50) хорошо обусловлена, а соответствующая ей задача Коши плохо обусловлена. При этом численное интегрирование задачи Коши определяет функцию $\bar{\psi}(\eta)$ с большой погрешностью, что осложняет организацию итераций.

В этом случае пробуют поставить начальные условия на другом конце отрезка $x = b$, т. е. интегрировать задачу Коши справа налево; нередко при этом устойчивость улучшается. Если изме-

нение направления интегрирования не помогает, то такую краевую задачу решают либо специальными, либо разностными методами.

Одним из специальных методов для линейных краевых задач является *дифференциальная прогонка* (ее идея предложена в [1], а подробное описание алгоритма имеется, например, в [3, 4]). Этот метод хорошо устойчив именно в том случае, когда задача Коши для исходной линейной системы плохо обусловлена; этот факт вызывал одно время большой интерес к прогонке. Однако при хорошей устойчивости линейной задачи Коши прогонка становится недостаточно устойчивой. Поэтому в настоящее время дифференциальная прогонка употребляется не часто. Обычно используются ее разностные аналоги, рассматриваемые ниже; они обеспечивают удовлетворительную устойчивость расчета в большинстве интересных случаев.

3. Уравнения высокого порядка или системы большого числа уравнений имеют соответствующее число краевых условий, и способы задания этих условий достаточно разнообразны. Поэтому к таким задачам применять метод стрельбы много труднее, чем к простейшей задаче (50).

Рассмотрим тот (сравнительно несложный) случай, когда для системы p уравнений

$$\frac{du_k(x)}{dx} = f_k(x,\ u_1,\ u_2,\ \ldots,\ u_p), \qquad 1 \leqslant k \leqslant p, \quad a \leqslant x \leqslant b, \tag{60а}$$

дополнительные условия заданы только на концах отрезка и имеют следующий вид:

$$\varphi_k(u_1(a),\ \ldots,\ u_p(a)) = 0 \qquad 1 \leqslant k \leqslant m, \tag{60б}$$

$$\varphi_k(u_1(b),\ \ldots,\ u_p(b)) = 0 \qquad m+1 \leqslant k \leqslant p. \tag{60в}$$

Для определенности, будем полагать $m \geqslant p/2$.

Выберем за исходный тот конец отрезка $[a, b]$, где задана *бóльшая* часть краевых условий; в нашем случае это будет левый конец $x = a$. В качестве пристрелочных параметров возьмем $p - m$ каких-то функций $u_k(x)$ из полного набора, например,

$$u_q(a) = \eta_q, \qquad 1 \leqslant q \leqslant p - m. \tag{61а}$$

Если подставить эти значения в левые краевые условия (60б), то эти условия образуют систему алгебраических уравнений относительно начальных значений остальных функций; решая эту систему, найдем

$$u_q(a) = \psi_q(\eta_1,\ \eta_2,\ \ldots,\ \eta_{p-m}), \qquad p - m + 1 \leqslant q \leqslant p. \tag{61б}$$

Рассмотрим задачу Коши для системы уравнений (60а) с начальными условиями (61а, б). Решение этой задачи, которое можно найти численным интегрированием, удовлетворяет левому краевому условию (60б) и зависит от параметров $\eta = \{\eta_1, \eta_2, \ldots, \eta_{p-m}\}$. Подстановка этого решения в правые краевые условия (60в)

определяет вспомогательные функции параметров

$$\bar{\varphi}_k(\eta) = \varphi_k(u_1(b, \eta), \ldots, u_p(b, \eta)), \quad m+1 \leqslant k \leqslant p; \quad (62а)$$

те значения параметров, которые удовлетворяют системе алгебраических уравнений

$$\bar{\varphi}_k(\eta_1, \eta_2, \ldots, \eta_{p-m}) = 0, \quad m+1 \leqslant k \leqslant p, \quad (62б)$$

определяют искомое решение краевой задачи (60).

Напомним, что решение системы алгебраических уравнений высокого порядка само по себе является нелегкой задачей. Здесь оно осложняется тем, что вычисление функций $\bar{\varphi}_k(\eta)$ очень трудоемко, ибо требует численного интегрирования системы дифференциальных уравнений. Явный вид этих функций неизвестен, так что преобразовать систему (62б) к эквивалентной форме $\eta_q = \bar{\varphi}_q(\eta)$ и применять метод последовательных приближений затруднительно. А если мы захотим, как в п. 2, построить аналог метода Ньютона, то для вычисления матрицы производных $(\partial\bar{\varphi}_k/\partial\eta_q)$ надо будет дополнительно записать и численно интегрировать систему $p(p-m)$ дифференциальных уравнений.

Отсюда видно, что «пристрелка» большого числа параметров очень сложна. Поэтому для нелинейных задач метод стрельбы употребляют в основном тогда, когда $p-m=1$. Такие постановки краевых задач нередко встречаются в системах большого числа уравнений.

Линейные уравнения. В этом случае метод стрельбы сильно упрощается и позволяет легко решать задачи при любом числе параметров $p-m$. В самом деле, функции $\bar{\varphi}_k(\eta)$ будут линейными, т. е. они однозначно определяются по своим значениям в $p-m+1$ точке η^s, $1 \leqslant s \leqslant p-m+1$. Значит, выполнив $p-m+1$ интегрирование задачи Коши (60а), (61) с разными наборами параметров, можно найти искомый набор параметров η. Тогда $(p-m+2)$-е интегрирование даст решение краевой задачи (60).

Вычисления при этом удобно вести следующим образом. Сначала возьмем некоторый набор параметров $\eta_1^0, \eta_2^0, \ldots, \eta_{p-m}^0$ и обозначим полученные значения функций (62а) через $\bar{\varphi}_k^0 = \bar{\varphi}_k(\eta^0)$, $m+1 \leqslant k \leqslant p$. Затем изменим первый параметр на величину $\Delta\eta = 1$, т. е. возьмем набор $\eta_1^0+1, \eta_2^0, \eta_3^0, \ldots, \eta_{p-m}^0$ и обозначим полученные значения функций через $\bar{\varphi}_k^1$. Затем возьмем набор η_1^0, $\eta_2^0+1, \eta_3^0, \ldots, \eta_{p-m}^0$ и т. д. Выполнив полный цикл вычислений, можно записать каждую функцию в виде многомерного интерполяционного многочлена Ньютона первой степени (2.33):

$$\bar{\varphi}_k(\eta) = \bar{\varphi}_k^0 + \sum_{q=1}^{p-m} (\bar{\varphi}_k^q - \bar{\varphi}_k^0)(\eta_q - \eta_q^0), \quad m+1 \leqslant k \leqslant p.$$

Приравнивая эти функции нулю, получим систему линейных алгебраических уравнений для определения искомых параметров η_q:

$$\sum_{q=1}^{p-m} (\overline{\varphi}_k^q - \overline{\varphi}_k^0)\,\eta_q = \sum_{q=1}^{p-m} (\overline{\varphi}_k^q - \overline{\varphi}_k^0)\,\eta_q^0 - \overline{\varphi}_k^0, \qquad m+1 \leqslant k \leqslant n. \quad (63)$$

Заметим, что можно уменьшить на единицу число интегрирований системы линейных дифференциальных уравнений, если воспользоваться приемом, описанным в п. 2; но при большом значении $p-m$ это лишь незначительно сокращает общий объем вычислений, а организацию расчета усложняет.

4. Разностный метод; линейные задачи. Подробно рассмотрим разностный метод на примере простейшей краевой задачи для линейного уравнения второго порядка с краевыми условиями первого рода

$$u''(x) - p(x)\,u(x) = f(x), \qquad a \leqslant x \leqslant b, \quad (64а)$$

$$u(a) = \alpha, \qquad u(b) = \beta. \quad (64б)$$

Введем на $[a, b]$ сетку $a = x_0 < x_1 < x_2 < \ldots < x_N = b$, которую для упрощения выкладок будем считать равномерной. Приближенно выразим вторую производную от решения через значения решения в узлах сетки $u_n = u(x_n)$; например, воспользуемся простейшей аппроксимацией (3.7):

$$u''(x_n) \approx \frac{1}{h^2}(u_{n-1} - 2u_n + u_{n+1}), \qquad h = x_{n+1} - x_n = \text{const}. \quad (65)$$

Такую аппроксимацию можно записать в каждом внутреннем узле сетки x_n, $1 \leqslant n \leqslant N-1$. Если подставить ее в уравнение (64а), то уравнение станет приближенным; точно удовлетворять этому уравнению будет уже не искомое решение $u(x)$, а некоторое приближенное решение $y_n \approx u(x_n)$. Выполняя эту подстановку и обозначая $p_n = p(x_n)$ и $f_n = f(x_n)$, получим

$$y_{n-1} - (2 + h^2 p_n)\,y_n + y_{n+1} = h^2 f_n, \qquad 1 \leqslant n \leqslant N-1. \quad (66а)$$

Эта система состоит из $N-1$ алгебраического уравнения, а неизвестными в ней являются приближенные значения решения в узлах сетки. Число неизвестных y_n, $0 \leqslant n \leqslant N$, равно $N+1$, т. е. оно больше, чем число уравнений (66а). Недостающие два уравнения легко получить из краевых условий (64б):

$$y_0 = \alpha, \quad y_N = \beta. \quad (66б)$$

Решая алгебраическую систему (66а, б), найдем приближенное решение.

При таком подходе возникает три вопроса. 1) Существует ли (вещественное) решение алгебраической системы типа (66)? 2) Как

фактически находить это решение? 3) Сходится ли разностное решение к точному в какой-либо норме при стремлении шага сетки к нулю?

В качестве иллюстрации проведем полное исследование рассмотренного выше примера, дополнительно требуя $p(x) > 0$.

Сначала рассмотрим вопрос о существовании разностного решения. Исходная задача (64) была линейной, разностная аппроксимация (65) — тоже линейна. Благодаря этому система (66а, б) оказалась системой линейных алгебраических уравнений. Поскольку $p_n > 0$, то в матрице этой системы диагональные элементы преобладают: в каждой строке модуль диагонального элемента больше суммы модулей остальных элементов. Как отмечалось в главе V, § 3, п. 4, при этом решение линейной системы существует и единственно.

Вычислить решение линейной системы всегда можно методом исключения Гаусса. В данном случае благодаря использованию трехточечной аппроксимации (65) система (66) имеет трехдиагональную матрицу. Поэтому решение экономично находится частным случаем метода Гаусса — методом алгебраической прогонки (см. главу V, § 1, п. 5).

Докажем утверждение: *если $p(x)$ и $f(x)$ дважды непрерывно дифференцируемы, то разностное решение равномерно сходится к точному с погрешностью $O(h^2)$ при $h \to 0$.*

При сделанном предположении $u(x)$ имеет четвертую непрерывную производную; тогда для погрешности аппроксимации (65) справедливо соотношение (3.12):

$$\frac{1}{h^2}(u_{n-1} - 2u_n + u_{n+1}) - u''(x_n) = \frac{1}{12}h^2 u^{IV}(\xi_n), \qquad x_{n-1} < \xi_n < x_{n+1}.$$

Значит, точное решение удовлетворяет разностному уравнению

$$u_{n-1} - (2 + h^2 p_n) u_n + u_{n+1} = h^2 f_n + \frac{h^4}{12} u^{IV}(\xi_n), \quad 1 \leqslant n \leqslant N-1.$$

Вычитая из него уравнение (66а), получим уравнение, которому удовлетворяет погрешность $z_n = y_n - u(x_n)$; его удобно записать в следующем виде:

$$(2 + h^2 p_n) z_n = z_{n-1} + z_{n+1} + \frac{h^4}{12} u^{IV}(\xi_n), \quad 1 \leqslant n \leqslant N-1, \quad (67а)$$

$$z_0 = 0, \quad z_N = 0. \quad (67б)$$

Последние два уравнения являются очевидным следствием того, что уравнение (66б) точно передает граничное условие первого рода.

Выберем такую точку x_{n_0}, где $|z_n|$ достигает своего максимума; очевидно, это не граничная точка. Учитывая условие $p_n > 0$,

сравним в этой точке модули правой и левой частей уравнения (67а):

$$(2+h^2p_{n_0})\,|\,z_{n_0}\,|\leqslant|\,z_{n_0-1}\,|+|\,z_{n_0+1}\,|+\frac{h^4}{12}\,|\,u^{\mathrm{IV}}(\xi_{n_0})\,|.$$

Заменяя в правой части $|\,z_{n_0\pm1}\,|$ на $|\,z_{n_0}\,|$, мы только усилим неравенство и после сокращений получим оценку погрешности

$$\max|\,z_n\,|\leqslant\frac{h^2}{12}\left|\frac{u^{\mathrm{IV}}(\xi_{n_0})}{p_{n_0}}\right|\leqslant\frac{h^2}{12}\max\left|\frac{u^{\mathrm{IV}}(x)}{p(x)}\right|. \tag{68}$$

Утверждение доказано.

Сейчас была найдена мажорантная оценка погрешности. При некоторых дополнительных ограничениях можно получить асимптотическую оценку типа $z_n=\zeta(x_n)\,h^2+o(h^2)$, где $\zeta(x)$ — некоторая функция (общая теорема о таких оценках будет доказана в главе IX). Из наличия асимптотической оценки следует возможность применения правила Рунге—Ромберга для апостериорной оценки точности или для уточнения решения при помощи расчетов на сгущающихся сетках.

Остановимся на устойчивости расчета. Если $p(x)>0$, то задача Коши для уравнения (64а) плохо обусловлена, причем чем больше $p(x)$, тем хуже ее устойчивость. А из оценки (68) видно, что погрешность нашего разностного решения при большом $p(x)$ мала. Отсюда видно, что хорошо построенные разностные схемы нечувствительны к неустойчивости задачи Коши.

В обратном случае $p(x)<0$ не выполняется достаточное условие устойчивости алгебраической прогонки (5.14). Однако в практике численных расчетов нарушение этого условия обычно не вызывает заметного ухудшения устойчивости. Только в редких случаях, когда определитель алгебраической системы (66) почти равен нулю, точность расчета резко падает из-за возрастания ошибок округления.

Чтобы легко опознать и исключить такую потерю устойчивости, можно провести расчет на трех (или более) сетках с различными шагами. Если при убывании h все разностные решения близки между собой и стремятся к некоторому пределу со скоростью $O(h^2)$ *), то это свидетельствует о хорошей устойчивости.

Пример. Возьмем частный случай задачи (64), соответствующий $p(x)<0$:

$$u''(x)+u(x)=-x,\qquad u(0)=u\left(\frac{\pi}{2}\right)=0. \tag{69}$$

и воспользуемся разностной схемой (66). Если выбрать шаг сетки $h=\pi/4$, то алгебраическая система фактически будет со-

*) В общем случае — со скоростью, соответствующей порядку точности схемы.

стоять из одного уравнения, а при $h=\pi/8$ — из трех уравнений. Разностное решение y_n для этих случаев приведено в таблице 20. К этим двум решениям применено правило Рунге, и уточненное решение $\tilde{y}_n$ тоже представлено в таблице; для сравнения дано точное решение $u(x)=\frac{\pi}{2}\sin x - x$. Из таблицы видно, что рассмотренная разностная схема дает неплохие результаты даже на сетке с большим шагом.

Таблица 20

x_n	y_n		$\tilde{y}_n$	$u(x)$
	$h=\pi/4$	$h=\pi/8$		
$\pi/8$	—	0,2122	0,2090	0,2084
$\pi/4$	0,3503	0,3311	0,3247	0,3253
$3\pi/8$	—	0,2778	0,2746	0,2731

Заметим, что, зная разностное решение в узлах сетки, можно интерполяцией получить приближенное решение при произвольных значениях x. Точность интерполяции целесообразно согласовывать с точностью разностного решения: например, для схемы (66) интерполировать многочленом первой степени, имеющим точность $O(h^2)$, а уточненное решение $\tilde{y}$ интерполировать многочленом второй степени.

5. Разностный метод; нелинейные задачи. Выше была построена несложная разностная схема для простейшей задачи. Перейдем к более общим случаям.

Наибольшие трудности вызывают нелинейные задачи. Рассмотрим краевую задачу для нелинейного уравнения второго порядка

$$u''(x)=f(x,\ u),\qquad u(a)=\alpha,\qquad u(b)=\beta \tag{70}$$

с краевыми условиями первого рода. Будем предполагать, что $f(x,\ u)$ ограничена и непрерывна вместе со своими вторыми производными, так что существует ограниченная и непрерывная $u^{\mathrm{IV}}(x)$. Обозначим через $M_1=\max|f_u|$, $M_2=\max|u^{\mathrm{IV}}|$.

Аналогично п. 4, введем на $[a,\ b]$ равномерную сетку x_n и заменим вторую производную разностным выражением (65). Подставляя его в дифференциальное уравнение (70), получим систему нелинейных алгебраических уравнений

$$\begin{gathered} y_{n-1}-2y_n+y_{n+1}=h^2f(x_n,\ y_n),\qquad 1\leqslant n\leqslant N-1,\\ y_0=\alpha,\quad y_N=\beta; \end{gathered} \tag{71}$$

последние два уравнения аппроксимируют краевые условия.

Докажем сходимость разностного решения к точному, дополнительно предполагая, что $f_u \geqslant m_1 > 0$. Поскольку для погрешности аппроксимации производной (65) справедливо соотношение (3.12):

$$u_{n-1} - 2u_n + u_{n+1} = h^2 u''(x_n) + \frac{h^4}{12} u^{IV}(\xi_n), \quad \xi_n \in (x_{n-1}, x_{n+1}),$$

точное решение удовлетворяет разностным уравнениям

$$u_{n-1} - 2u_n + u_{n+1} = h^2 f(x_n, u_n) + \frac{h^4}{12} u^{IV}(\xi_n), \quad 1 \leqslant n \leqslant N-1,$$

$$u_0 = \alpha, \quad u_N = \beta.$$

Вычитая эти уравнения из (71), обозначая погрешность $z_n = y_n - u_n$ и учитывая, что $f(x_n, y_n) - f(x_n, u_n) = (f_u)_n z_n$, получим для погрешности систему уравнений

$$\begin{gathered} z_{n-1} - (2 + h^2 f_u)_n z_n + z_{n+1} = \frac{h^4}{12} u^{IV}(\xi_n), \quad 1 \leqslant n \leqslant N-1, \\ z_0 = 0, \quad z_N = 0. \end{gathered} \tag{72}$$

Пусть x_{n_0} есть узел, в котором $|z_n|$ максимален. В этом узле перепишем соотношение (72) в форме неравенства

$$(2 + h^2 f_u)_{n_0} |z_{n_0}| \leqslant |z_{n_0-1}| + |z_{n_0+1}| + \frac{h^4}{12} |u^{IV}(\xi_{n_0})|.$$

Усилим это неравенство, заменяя в правой части $|z_{n_0 \pm 1}|$ на $|z_{n_0}|$; тогда получим

$$|z_{n_0}| = \|z_n\|_c \leqslant \frac{h^2}{12} \frac{|u^{IV}(\xi_{n_0})|}{(f_u)_{n_0}} \leqslant \frac{h^2 M_2}{12 m_1}. \tag{73}$$

Это означает, что *при* $h \to 0$ *разностное решение равномерно сходится к точному со вторым порядком точности.*

Займемся фактическим нахождением разностного решения. Алгебраические системы общего вида решают методами последовательных приближений или линеаризации. Однако, если взять метод последовательных приближений в естественной форме (5.44):

$$2y_n^{(s+1)} = y_{n-1}^{(s)} + y_{n+1}^{(s)} - h^2 f(x_n, y_n^{(s)}),$$

то нетрудно убедиться, что критерии сходимости этого метода (5.45) не выполняются. Положение улучшается, если придать методу последовательных приближений специфическую форму

$$\begin{gathered} y_{n-1}^{(s)} - 2y_n^{(s)} + y_{n+1}^{(s)} = h^2 f(x_n, y_n^{(s-1)}), \quad 1 \leqslant n \leqslant N-1, \\ y_0^{(s)} = \alpha, \quad y_N^{(s)} = \beta. \end{gathered} \tag{74}$$

Тогда для определения $y_n^{(s)}$ на каждой итерации получается линейная система, решаемая алгебраической прогонкой. Исследуем сходимость итераций (74).

Рассмотрим *погрешность итерации* $\zeta_n^{(s)} = y_n^{(s)} - y_n$. Она удовлетворяет системе уравнений, получаемой вычитанием (71) из (74):

$$\zeta_{n-1}^{(s)} - 2\zeta_n^{(s)} + \zeta_{n+1}^{(s)} = d_n^{(s)}, \quad 1 \leqslant n \leqslant N-1,$$
$$\zeta_0^{(s)} = 0, \quad \zeta_N^{(s)} = 0, \tag{75}$$
$$d_n^{(s)} = h^2 f\left(x_n,\ y_n^{(s-1)}\right) - h^2 f\left(x_n,\ y_n\right) \approx h^2 f_u\left(x_n,\ y_n\right) \zeta_n^{(s-1)}.$$

Решим эту трехдиагональную систему методом прогонки. Для данной системы рекуррентные соотношения (5.12) для коэффициентов прогонки нетрудно преобразовать к такому виду:

$$\xi_{n+1} = \frac{1}{2-\xi_n} = \frac{n}{n+1}, \quad 0 \leqslant n \leqslant N-2, \quad \xi_N = 0,$$

$$\eta_{n+1} = \xi_{n+1}\left(\eta_n - d_n^{(s)}\right) = -\frac{1}{n+1}\sum_{k=1}^{n} k d_k^{(s)}, \quad 1 \leqslant n \leqslant N-1.$$

Формулы обратного хода прогонки (5.11) также преобразуются

$$\zeta_n^{(s)} = \xi_{n+1}\zeta_{n+1}^{(s)} + \eta_{n+1} = -n \sum_{k=n+1}^{N} \frac{1}{k(k-1)} \sum_{p=1}^{k-1} p d_p^{(s)} \tag{76}$$

и дают искомое решение системы (75).

Для правых частей системы (75) выполняется неравенство

$$\left| d_n^{(s)} \right| \leqslant q^{(s)}, \quad q^{(s)} = h^2 M_1 \| \zeta_n^{(s-1)} \|_c.$$

Подставляя его в (76), получим

$$\left| \zeta_n^{(s)} \right| \leqslant q^{(s)} n \sum_{k=n+1}^{N} \frac{1}{k(k-1)} \sum_{p=1}^{k-1} p = \frac{1}{2} q^{(s)} n (N-n) \leqslant \frac{1}{8} q^{(s)} N^2.$$

Отсюда следует

$$\| \zeta_n^{(s)} \|_c \leqslant \bar{q} \| \zeta_n^{(s-1)} \|_c, \quad \text{где} \quad \bar{q} = \frac{1}{8} N^2 h^2 M_1 = \frac{1}{8} (b-a)^2 M_1. \tag{77}$$

Это означает, что итерации (74) сходятся при выполнении условия

$$\frac{1}{8}(b-a)^2 M_1 < 1, \quad M_1 = \max \left| \frac{\partial f}{\partial u} \right|. \tag{78}$$

Из соотношения (78) следует, что сходимость линейная, т. е. довольно медленная.

Условие (78) является достаточным, но оно близко к необходимому: более сложные оценки показывают, что если $(b-a)^2 M_1 > \pi^2$, то итерации (74) могут расходиться.

Целесообразнее решать уравнения (71) методом Ньютона. Соответствующие формулы нетрудно записать, линеаризуя правые части этих уравнений:

$$y_n^{(s+1)} = y_n^{(s)} + \Delta_n^{(s)}, \quad 0 \leqslant n \leqslant N,$$
$$\Delta_{n-1}^{(s)} - (2 + h^2 f_u)_n \Delta_n^{(s)} + \Delta_{n+1}^{(s)} = h^2 f_n^{(s)} - y_{n-1}^{(s)} + 2y_n^{(s)} - y_{n+1}^{(s)}, \quad (79)$$
$$1 \leqslant n \leqslant N-1,$$
$$\Delta_0^{(s)} = \Delta_N^{(s)} = 0.$$

Линеаризованную систему также решают алгебраической прогонкой. Сходимость итераций исследуют описанными выше приемами. Потребуем, чтобы $f_u \geqslant m_1 > 0$. Тогда сравнивая (79) и (71), можно получить для поправки $\zeta_n^{(s)} = y_n^{(s)} - y_n$ такое неравенство:

$$\| \zeta_n^{(s)} \|_c \leqslant \left\| \frac{f_{uu}(x, u)}{2f_u(x, u)} \right\|_{C(x)} \| \zeta_n^{(s-1)} \|_c^2.$$

Это означает, что если нулевое приближение взято не слишком далеко от корня (например, удовлетворяет условию

$$\| \zeta_n^{(0)} \|_c \leqslant \| 2f_u/f_{uu} \|_{C(x)},$$

то итерации (79) сходятся, причем квадратично. Поэтому метод Ньютона обычно выгодней метода последовательных приближений, несмотря на более громоздкие формулы.

Замечание. Если итерации (79) или (74) сходятся, то в силу непрерывности и гладкости функции $f(x, u)$ они сходятся к решению системы (71). Тем самым устанавливается существование разностного решения в этих случаях.

Для нелинейных задач очень эффективна *комплексная организация расчета*, позволяющая при небольшом объеме вычислений получать высокую точность. Опишем ее.

Возьмем первую сетку с очень малым числом интервалов $N = 2 - 8$; остальные сетки получим из нее последовательным сгущением вдвое. На первой сетке начальное приближение выберем каким-либо приближенным способом: методом Галеркина, или разложением по малому параметру. Поскольку для первой сетки порядок алгебраической системы мал, качество нулевого приближения здесь малосущественно.

Когда итерации сошлись, полученное разностное решение интерполируем (например, линейно) на второй сетке и возьмем на ней в качестве нулевого приближения. Тогда итерации обычно быстро сходятся; в методе Ньютона достаточно 2—4 итераций. Интерполируем это решение на следующей сетке и т. д. Общий объем расчетов при этом невелик и примерно эквивалентен 5—8 итерациям последней сетки.

В заключение разностное решение на всех сетках уточним по рекуррентному правилу Рунге. Это настолько повышает точность, что даже в сложных задачах позволяет ограничиться небольшим числом интервалов последней сетки ($N = 32 - 128$).

Если проводится серия расчетов при варьировании параметров исходной задачи, то целесообразно результат расчета одного варианта брать в качестве нулевого приближения для первой сетки следующего варианта.

Рассмотрим некоторые другие усложнения задачи.

1) Сетка может быть неравномерной. В этом случае надо использовать соответствующую аппроксимацию производных, например,

$$u''(x_n) \approx \frac{2}{x_{n+1} - x_{n-1}} \left(\frac{u_{n+1} - u_n}{x_{n+1} - x_n} - \frac{u_n - u_{n-1}}{x_n - x_{n-1}} \right).$$

Напомним, что эта аппроксимация имеет погрешность $O(h^2)$ на квазиравномерных сетках и $O(h)$ на произвольных сетках. Исследование разностной схемы (71), проведенное выше, легко обобщается на случай неравномерной сетки.

2) Можно использовать аппроксимации, явно учитывающие вид общего решения исходного дифференциального уравнения; при этом получаются специальные схемы (см. § 1, п. 9). Составим, например, для задачи (64) с $p(x) < 0$ такую схему, чтобы она была точна при $p = \text{const}$, $f(x) = \text{const}$. При этом ограничении общее решение уравнения (64а) имеет вид

$$\tilde{u}(x) = -\frac{p}{f} + A \sin\left(\sqrt{|p|}\, x\right) + B \cos\left(\sqrt{|p|}\, x\right),$$

где A, B — произвольные постоянные. Легко проверить, что на равномерной сетке подстановка этого решения в разностную схему

$$y_{n-1} - 2y_n \cos\left(h\sqrt{-p_n}\right) + y_{n+1} = 2f_n \left[1 - \cos\left(h\sqrt{-p_n}\right)\right], \quad 1 \leqslant n \leqslant N-1, \qquad (80)$$

(краевые условия учитываются аналогично (71)) дает тождество. Следовательно, эта схема точна в указанном смысле. Она позволяет получать хорошую точность расчета быстро осциллирующих решений даже на грубой сетке, если $p(x)$ и $f(x)$ являются медленно меняющимися функциями.

Однако заметим, что применять правило Рунге для уточнения разностных решений, полученных по схемам типа (80), можно не всегда. Причина этого была подробно рассмотрена в связи с формулами Филона (глава IV, § 2, п. 3).

3) Дифференциальное уравнение может иметь более высокий порядок. Аппроксимация старших производных требует большего числа узлов, и каждое уравнение типа (71) или (66а) будет содержать соответственно большее число неизвестных. Поэтому для решения алгебраической линейной (или линеаризованной) системы вместо алгебраической прогонки надо использовать несколько более трудоемкие способы. Но принципиальных осложнений это не вызывает.

4) Возможны более сложные краевые условия. Рассмотрим, например, нелинейное условие третьего рода

$$u'(a) = \varphi(u(a)). \tag{81}$$

Если подставить в него аппроксимацию $u_0' \approx (u_1 - u_0)/h$, то ее погрешность $O(h)$ велика, что ухудшает общую точность расчета. Чтобы записать разностное краевое условие повышенной точности, рассмотрим формулу Тейлора

$$u(x_1) = u(x_0) + hu'(x_0) + \frac{1}{2}h^2u''(x_0) + \ldots$$

и на основании уравнения (70) положим $u''(x_0) = f(x_0, u_0)$, а из краевого условия (81) возьмем $u'(x_0) = \varphi(u_0)$. Тогда получим

$$\frac{1}{h}(y_1 - y_0) = \varphi(y_0) + \frac{h}{2}f(x_0, y_0). \tag{82}$$

Другие способы аппроксимации краевых условий будут рассмотрены в главе IX.

Подведем итоги. Разностный метод имеет свои трудности, связанные в основном с решением алгебраической системы уравнений. Однако эти трудности успешно преодолеваются. Метод естественно переносится на уравнения высокого порядка, причем трудоемкость вычислений почти не возрастает. Его численная устойчивость обычно хорошая.

Поэтому для уравнений второго порядка разностный метод успешно конкурирует с методом стрельбы, а для уравнений более высокого порядка, особенно при сложной постановке краевых условий, оказывается выгоднее стрельбы.

6. Метод Галеркина. Краевая задача для уравнения $A(u(x)) = 0$ сводилась в главе VII, § 4 к отысканию минимума функционала типа (Au, Au) или (u, Au). Затем решение $u(x)$ приближенно заменялось отрезком разложения по некоторой полной системе функций, а коэффициенты разложения находились из условия минимума функционала. Этот способ для функционалов первого типа называют методом наименьших квадратов, а для второго — методом Ритца.

Метод наименьших квадратов неудобен тем, что под интегралом возникают квадраты старших производных, входящих в оператор A, и вычисления становятся громоздкими. Метод Ритца имеет тот недостаток, что не для всякого оператора A удается найти эквивалентный функционал (обычно нужна самосопряженность оператора). Более удобен на практике метод Б. Г. Галеркина (или Бубнова — Галеркина), свободный от этих недостатков. Изложим этот метод.

Пусть дано уравнение с некоторыми краевыми условиями (для определенности — первого рода)

$$A(u(x)) = f(x), \quad a \leqslant x \leqslant b, \quad u(a) = \alpha, \quad u(b) = \beta. \tag{83}$$

Как и в методе Ритца (см. главу VII, § 4, п. 3), будем искать приближенное решение в виде суммы

$$u(x) \approx y_n(x) = \varphi_0(x) + \sum_{k=1}^{n} c_k \varphi_k(x), \tag{84}$$

где $\varphi_0(x)$ — некоторая непрерывная функция, удовлетворяющая неоднородным краевым условиям (83), а $\varphi_k(x)$, $1 \leqslant k < \infty$, — какая-то система линейно-независимых функций, полная в классе непрерывных функций, определенных на отрезке $[a, b]$ и обращающихся в нуль на его концах.

Докажем, что если для некоторой функции $F(x)$ и полной системы функций $\varphi_k(x)$ выполняется соотношение

$$\int_a^b F(x)\,\varphi_k(x)\,dx = 0 \quad \text{при} \quad 1 \leqslant k < \infty,$$

то $F(x) \equiv 0$ на $[a, b]$. Для этого из полной системы $\varphi_k(x)$ последовательной ортогонализацией построим полную ортогональную систему $\psi_k(x)$. Очевидно, тогда

$$\varphi_k(x) = \sum_{m=1}^{k} \zeta_{km} \psi_m(x),$$

причем $\zeta_{kk} \neq 0$, иначе $\varphi_k(x)$ были бы линейно-зависимы. Разлагая по новой системе

$$F(x) = \sum_{l=1}^{\infty} \gamma_l \psi_l(x),$$

придем к соотношению

$$0 = \int_a^b F(x)\,\varphi_k(x)\,dx = \sum_{m=1}^{k} \gamma_m \zeta_{km} = 0, \qquad k = 1, 2, \ldots$$

Полагая $k = 1$, получим $\gamma_1 = 0$. Полагая $k = 2$, получим $\gamma_2 = 0$ и т. д. Следовательно, все $\gamma_l = 0$ и $F(x) \equiv 0$. Отметим, что если исходная система $\varphi_k(x)$ уже ортогональна, то доказательство становится тривиальным.

Таким образом, если бы мы нашли такую функцию $u(x)$, чтобы $A(u(x)) - f(x)$ было ортогонально $\varphi_k(x)$ при любых $k \geqslant 1$, то это означало бы, что $A(u(x)) = f(x)$ и задача (83) была бы решена *). Если же ортогональность есть только при $k \leqslant n$, то в разложе-

*) Ортогональности $A(u) - f$ к $\varphi_0(x)$ не требуется, ибо $\varphi_0(x)$ не входит в полную систему функций $\varphi_k(x)$.

ние $A(u)-f$ по $\varphi_k(x)$ входят γ_{n+1} и более старшие коэффициенты, т. е. $A(u)\approx f$.

Возьмем вместо $u(x)$ приближенное решение в форме (84) и потребуем, чтобы

$$\int_a^b [A(y_n(x))-f(x)]\varphi_k(x)\,dx=0, \qquad 1\leqslant k\leqslant n. \tag{85}$$

Это дает нам алгебраическую систему для определения коэффициентов c_k. Найдя из нее коэффициенты, получим приближенное решение (84). В этом и заключается метод Галеркина. Вопрос об условиях сходимости $y_n(x)$ при $n\to\infty$ к точному решению и о скорости сходимости здесь не рассматривается.

Если оператор $A(u)$ нелинейный, то система (85) тоже будет нелинейной. При этом больше чем 3—4 коэффициента трудно найти. Если же оператор линейный, то алгебраическая система (85) линейна и можно решать задачу с большим числом коэффициентов. Отметим, что для линейных уравнений второго порядка метод Галеркина приводит точно к тем же уравнениям, что и метод Ритца.

Пример. Рассмотрим задачу (69). Положим $\varphi_0(x)=0$ и выберем полную систему функций $\varphi_k(x)=x^k((\pi/2)-x)$, $1\leqslant k<\infty$. Тогда, если ограничиться одним членом суммы (84), то легко получить, что

$$c_1=5\pi/(40-\pi^2)\approx 0{,}521, \qquad n=1.$$

Если возьмем два члена суммы, то получим

$$c_1\approx 0{,}815, \qquad c_2\approx 0{,}377, \qquad n=2.$$

Соответствующие приближенные решения, вычисленные в нескольких точках отрезка, приведены в таблице 21; для сравнения там же дано точное решение. Хотя в этой задаче решение является плавно меняющейся функцией, метод Галеркина при небольшом числе членов дает неважные результаты.

Таблица 21

x	$y_1(x)$	$y_2(x)$	$u(x)$
$\pi/8$	0,241	0,445	0,208
$\pi/4$	0,322	0,685	0,325
$3\pi/8$	0,241	0,582	0,273

Обратим внимание на то, что при увеличении n не только добавляются новые коэффициенты, но и меняются старые, что не очень удобно. Легко заметить, что если задача линейная, а система $\varphi_k(x)$ ортогональная, то уже найденные c_k не будут меняться при увеличении n. Поэтому ортогональные системы обычно удобнее неортогональных.

Метод Галеркина для нелинейных задач используют лишь для нахождения грубого приближения; для линейных задач им можно

найти решение с хорошей точностью. Результат очень чувствителен к тому, насколько удачно выбрана система функций $\varphi_k(x)$ для данной задачи.

Отметим также, что при нелинейном краевом условии вида, например, $u'(a) = g(u(a))$ линейная комбинация (84) с произвольными коэффициентами c_k уже не будет удовлетворять этому краевому условию. Поэтому метод Галеркина применим только к задачам с линейными (относительно $u(x)$ и ее производных) краевыми условиями, хотя допустим и нелинейный оператор $A(u)$.

7. Разрывные коэффициенты. Во всех предыдущих пунктах явно или неявно предполагалось, что правые части рассматриваемых дифференциальных уравнений непрерывны вместе с некоторым числом своих производных. Однако в задачах о слоистых средах коэффициенты уравнений (коэффициентами являются различные свойства вещества — плотность, теплопроводность, упругость и т. д.) обычно разрывны на границах раздела двух сред, т. е. во внутренних точках $[a, b]$.

Бегло рассмотрим, как переносятся на этот случай развитые выше методы. Сделаем это на примере уравнения

$$\frac{d}{dx}\left[k(x)\frac{du}{dx}\right] - q(x)u(x) = f(x). \tag{86а}$$

Сначала обсудим характер решения. Если $q(x)$ или $f(x)$ кусочно-непрерывны, то $u''(x)$ также лишь кусочно-непрерывна. Очевидно, в точке разрыва аппроксимировать вторую производную разностным соотношением нельзя.

Еще сложней случай разрыва $k(x)$ в некоторой точке $\bar{x}$. При этом решение краевой задачи становится, вообще говоря, не единственным. Существует множество обобщенных решений, каждое из которых удовлетворяет своему условию согласования в точке $\bar{x}$. Для выделения единственного решения требуется поставить в этой точке внутреннее краевое условие; оно выбирается из физических соображений и должно входить в полную постановку задачи.

Пусть, например, (86а) есть уравнение теплопроводности в стержне, составленном из разных материалов, а $k(x)$ — коэффициент теплопроводности. Тогда дополнительным условием будет непрерывность температуры и теплового потока $W = -ku_x$ в точке соединения

$$[u(x)]_{\bar{x}-0}^{\bar{x}+0} = 0, \quad \left[k(x)\frac{du}{dx}\right]_{\bar{x}-0}^{\bar{x}+0} = 0. \tag{86б}$$

Поэтому и в методе стрельбы, и в разностном методе все точки разрыва коэффициентов выбирают в качестве узлов сетки; такие сетки называют *специальными*.

В методе стрельбы до прихода (для определенности слева) в такую точку пользуются «левыми» значениями коэффициентов.

Придя в эту точку, при помощи внутреннего краевого условия формируют новые начальные условия. Например, в задаче (86) это будут условия

$$u(\bar{x}+0)=u(\bar{x}-0), \quad u_x(\bar{x}+0)=\frac{k(\bar{x}-0)}{k(\bar{x}+0)}u_x(\bar{x}-0).$$

Затем продолжают численное интегрирование, пользуясь уже «правыми» значениями коэффициентов.

В разностном методе для точки разрыва вместо аппроксимации дифференциального уравнения (86а) можно записать аппроксимацию внутреннего краевого условия (86б), или можно составить такую разностную схему, которая применима во всех точках, включая точку $\bar{x}$ *).

В методе Галеркина систему функций $\varphi_k(x)$ следует выбирать так, чтобы линейная комбинация (84) при любых значениях коэффициента c_k удовлетворяла внутреннему краевому условию.

§ 3. Задачи на собственные значения

1. Постановки задач. Задачи на собственные значения — это краевые задачи для системы p уравнений первого порядка

$$\boldsymbol{u}'(x)=\boldsymbol{f}(x,\ \boldsymbol{u};\ \lambda_1, \lambda_2, \ldots, \lambda_q),$$
$$\boldsymbol{u}=\{u_1, u_2, \ldots, u_p\}, \quad \boldsymbol{f}=\{f_1, f_2, \ldots, f_p\},$$

в которых правые части зависят от параметров λ_r, значения которых неизвестны и должны быть определены из самой задачи; число дополнительных (краевых) условий соответственно равно $p+q$. Функции $u_k(x)$, $1\leqslant k\leqslant p$, и значения параметров λ_r, $1\leqslant r\leqslant q$, удовлетворяющие всем уравнениям и краевым условиям, называются *собственными функциями* и *собственными значениями* задачи.

Задачи на собственные значения часто встречаются в физике и технике **). Например, определение собственных колебаний струны приводит к задаче для линейного уравнения второго порядка с одним параметром

$$\frac{d}{dx}\left[k(x)\frac{du}{dx}\right]+\lambda\rho(x)u(x)=0,$$

а собственных колебаний упругого стержня — к линейному уравнению четвертого порядка (краевые условия зависят от способа закрепления струны или стержня). Дифференциальное уравнение

*) Это так называемые *консервативные* схемы, способ построения которых будет изложен в следующих главах.

**) Много примеров таких задач приведено в [17].

второго порядка возникает при нахождении спектра атома водорода. Нахождение уровней энергии многоэлектронного атома в приближении Хартри — Фока приводит к задаче для системы нелинейных уравнений, в которой число функций и число параметров равно числу электронов атома.

Исследование корректности постановки задачи на собственные значения еще более сложно, чем для краевых задач. Исследованы в основном линейные задачи с одним параметром. Однако в курсах теории колебаний и квантовой механики имеется немало примеров, из которых видно, что в зависимости от постановки задачи собственные значения могут существовать или не существовать, быть вещественными или комплексными; спектр собственных значений может быть дискретным, сплошным, состоящим из полос или являющимся комбинацией перечисленных случаев.

Наиболее употребительными численными методами решения задач на собственные значения являются метод стрельбы и разностный метод, подробно рассмотренные ниже. Из приближенных методов упомянем методы Ритца и Галеркина.

2. Метод стрельбы. В задачах на собственные значения имеются естественные пристрелочные параметры — величины λ_r; поэтому такие задачи нередко решают методом стрельбы. Основные черты этого метода те же, что и для краевых задач; рассмотрим детали метода на двух примерах.

Простейший пример — задача для одного уравнения первого порядка с одним параметром и двумя краевыми условиями

$$u'(x) = f(x,\ u;\ \lambda), \quad u(a) = \alpha, \quad u(b) = \beta. \tag{87}$$

Если отбросить правое краевое условие и выбрать некоторое значение λ, то (87) превратится в задачу Коши. Численно интегрируя ее, получим решение $u(x;\ \lambda)$, удовлетворяющее левому краевому условию и зависящее от параметра λ. Вообще говоря, $u(b;\ \lambda) \neq \beta$, т. е. это решение не удовлетворяет правому краевому условию. Тогда будем варьировать λ до тех пор, пока не получим $u(b;\ \lambda) \approx \beta$ с требуемой точностью. Разумеется, при варьировании используют обычные методы нахождения корня алгебраического уравнения, как это было сделано в § 2, п. 2.

Другой пример — это классическая задача на собственные значения уравнения второго порядка при нулевых краевых условиях

$$u''(x) + p(x)\,u'(x) + [\lambda + q(x)]\,u(x) = 0, \quad u(a) = u(b) = 0. \tag{88}$$

Уравнение имеет второй порядок и содержит одно собственное значение; следовательно, задача требует трех дополнительных условий. Но в силу линейности и однородности решение определено с точностью до множителя; это и есть неявное задание третьего условия. Формально третье условие здесь удобно задать в форме $u'(a) = 1$ (что возможно, если $p(a)$ и $q(a)$ конечны).

Тогда можно взять для исходного уравнения задачу Коши с начальными условиями $u(a)=0$, $u'(a)=1$ и вести пристрелку параметра λ до выполнения правого краевого условия.

Заметим, что линейность уравнения и краевых условий не упрощает стрельбу, ибо зависимость $u(x;\ \lambda)$ от параметра все равно остается нелинейной.

Метод стрельбы удобно применять, если стрельба является однопараметрической, как это было в рассмотренных примерах. Если это требование не выполнено, то алгоритмы стрельбы сильно усложняются и становятся менее надежными; тогда выгодней использовать разностный метод.

Метод стрельбы трудно применять также в том случае, если задача Коши плохо обусловлена. Тогда малая вариация λ может резко изменить решение $u(x)$ и даже вывести его за пределы представимых на ЭВМ чисел. При этом невозможно организовать процесс решения алгебраического уравнения типа

$$u(b;\ \lambda)=0.$$

Иногда, как и в краевых задачах, помогает смена направления интегрирования (но ее применяют только, если от этого не увеличивается число параметров пристрелки).

3. Фазовый метод. Классическая задача для уравнения второго порядка (88) имеет много важных физических приложений. В частности, к этому уравнению приводит квантовомеханическая задача об уровнях энергии частицы, движущейся в заданном одномерном (например, сферически-симметричном) поле. В последнем случае задача Коши для уравнения (88) оказывается очень плохо обусловленной: общее решение уравнения обращается в бесконечность на обоих концах отрезка ($x=0$ и $x=\infty$). Поэтому применять метод стрельбы трудно. Но эта задача настолько важна, что для нее разработаны специальные схемы. Рассмотрим одну из них — *фазовый метод*.

Воспользуемся тем, что качественное поведение решения известно. Решение имеет осциллирующий характер, причем амплитуда может сильно зависеть от координаты. Введем амплитуду ρ и фазу φ решения при помощи соотношения

$$u(x)=\rho(x)\sin\varphi(x). \tag{89а}$$

Это соотношение неоднозначно определяет амплитуду и фазу. Для определенности подчиним их дополнительному соотношению

$$u'(x)=\rho(x)\cos\varphi(x). \tag{89б}$$

Наглядный смысл его состоит в том, что если взять вектор с координатами u, u', т. е. перейти в фазовую плоскость, то ρ и φ будут амплитудой и фазой этого вектора.

Дифференцируя (89а) и (89б) и сравнивая их между собой, получим соотношения

$$u'' = \rho' \cos\varphi - \varphi' \rho \sin\varphi,$$
$$\rho' \sin\varphi = (1 - \varphi')\, \rho \cos\varphi.$$

Исключая при помощи этих соотношений и формул (89) функцию $u(x)$ и ее производные из уравнения (88), после несложных преобразований расщепим (88) на уравнения для амплитуды и фазы:

$$\rho'(x) = -\rho(x)\{p(x)\cos\varphi(x) + [q(x) + \lambda - 1]\sin\varphi(x)\}\cos\varphi(x), \quad (90)$$
$$\varphi'(x) = \cos^2\varphi(x) + p(x)\sin\varphi(x)\cos\varphi(x) + [\lambda + q(x)]\sin^2\varphi(x). \quad (91)$$

Граничные условия (88) при этом естественно приписываются фазе. Если надо найти решение, соответствующее квантовому числу n, т. е. имеющее n полуволн на $[a, b]$, то следует положить

$$\varphi(a) = 0, \quad \varphi(b) = \pi n. \quad (92)$$

Таким образом, мы получили задачу на собственные значения (91)—(92) только для уравнения фазы. Она легко решается методом стрельбы, поскольку задача Коши для уравнения (91) хорошо обусловлена. Важной особенностью этой задачи является то, что правому краевому условию (92) удовлетворяет только одно определенное λ_n из всего спектра исходной задачи (88). Поэтому стрельба всегда сходится именно к требующемуся собственному значению.

После нахождения фазы уравнение для амплитуды легко интегрируется в квадратурах

$$\rho(x) = \rho(a) \exp\left\{-\int_a^x [p(\xi)\cos\varphi(\xi) + (q(\xi) + \lambda - 1)\sin\varphi(\xi)]\cos\varphi(\xi)\, d\xi\right\}.$$

Амплитуда определена с точностью до множителя и не меняет знака, как и должно быть по смыслу задачи.

Замечание 1. Задача (88) может иметь и другие типы краевых условий. Если исходное краевое условие имеет вид $u'(b) = 0$, то для фазы надо взять условие $\varphi(b) = \pi(n - 1/2)$. Несколько сложней асимптотическое условие $u(\infty) = 0$, возникающее в задаче на отрезке $a \leqslant x < \infty$; обычно в таких задачах выполняется $p(\infty) = q(\infty) = 0$. Тогда нетрудно построить асимптотику решения $u(x) \sim \exp(-\sqrt{-\lambda}\, x)$ при $x \to \infty$ и получить отсюда асимптотическое краевое условие для фазы

$$\cos\varphi(x) + \sqrt{-\lambda}\sin\varphi(x) \to 0 \quad \text{при} \quad x \to +\infty.$$

Замечание 2. Фаза $\varphi(x)$ может быть немонотонной функцией. Однако, если при некотором $\tilde{x}$ значение фазы $\varphi(\tilde{x}) = \pi k$, то $\varphi'(\tilde{x}) = 1$; поэтому каждую линию $\varphi = \pi k$ интегральная кривая пересекает лишь однажды, а немонотонность может проявляться только между этими линиями. При таком поведении интегральных кривых стрельба с использованием дихотомии надежно сходится к собственному значению, а при использовании метода Ньютона область сходимости нередко оказывается очень узкой.

Замечание 3. Для преодоления последнего недостатка предложена замена функций, несколько более сложная, чем (89), но зато делающая $\varphi(x)$ монотонной функцией. При этом стрельба с использованием метода Ньютона сходится за небольшое число итераций.

4. Разностный метод обычно используется в тех случаях, когда стрельба оказывается многопараметрической, или если задача Коши для исходного дифференциального уравнения плохо обусловлена.

Формулируется он так же, как для краевых задач. Введем на $[a, b]$ сетку $\{x_n,\ 0 \leqslant n \leqslant N\}$ и заменим в исходной задаче все производные некоторыми разностными соотношениями. Тогда вместо дифференциального уравнения и краевых условий получим систему алгебраических уравнений

$$F_k(x_0, x_1, \ldots, x_N, y_0, y_1, \ldots, y_N; \lambda) = 0, \quad 0 \leqslant k \leqslant N+1 \qquad (93)$$

(для простоты записи мы ограничиваемся случаем одного собственного значения). Эта система содержит $N+2$ уравнения, и из нее надо определить такое же число неизвестных: $\lambda, y_0, y_1, \ldots, y_N$.

Возникают те же вопросы, что и в краевых задачах. Имеет ли алгебраическая система (93) решение? Если имеет, то как его фактически вычислить? Если разностное решение найдено, то насколько оно близко к точному решению? Сейчас мы рассмотрим линейные задачи, для которых на эти вопросы ответить легче.

Пусть исходная задача является линейной и однородной относительно $u(x)$, как, например, задача (88). Воспользуемся линейными разностными аппроксимациями производных. Тогда система (93) будет относительно y_n линейной однородной, т. е. это будет алгебраическая задача на собственные значения матрицы. Так, для задачи (88) при простейших аппроксимациях на равномерной сетке получим систему

$$\left(1 - \frac{1}{2} hp_n\right) y_{n-1} - (2 - h^2 q_n - \lambda h^2)\, y_n + \left(1 + \frac{1}{2} hp_n\right) y_{n+1} = 0,$$
$$1 \leqslant n \leqslant N-1, \qquad (94)$$

где $y_0 = y_N = 0$ в силу краевых условий. Эта система содержит $N-1$ уравнение; из нее надо определить $\lambda, y_1, y_2, \ldots, y_{N-1}$.

Задача (94) имеет спектр собственных значений, состоящий из $N-1$ числа (по порядку матрицы). Первые собственные значения являются приближениями к первым собственным значениям λ_m из дискретного спектра исходной задачи (88). Если разностная схема составлена так, что матрица алгебраической системы (93) является эрмитовой, то приближенные собственные значения будут вещественными.

Собственные значения и собственные векторы линейной системы (93) вычисляют методами, описанными в главе VI. Поскольку во многих приложениях матрица системы трехдиагональная (реже — пятидиагональная), а нужны только несколько первых собственных значений, то выгодно применять метод Дервюдье (см. главу VI, § 4, п. 2). При небольшом числе интервалов сетки удобно также находить корни характеристического многочлена методом парабол, вычисляя сам многочлен по рекуррентным соотношениям (см. главу VI, § 1, п. 4).

Сходимость разностного решения к точному при $h \to 0$ хорошо исследована только для задач Штурма — Лиувилля *)

$$\frac{d}{dx}\left[k(x)\frac{du}{dx}\right]+[\lambda r(x)-q(x)]u(x)=0, \quad u(a)=u(b)=0.$$

Оказывается, что простейшая схема (94) дает не очень хорошие, а при разрывных коэффициентах — даже неверные результаты. Следует составлять консервативные разностные схемы (они будут подробно рассмотрены в главах X и XI). Если коэффициенты уравнения непрерывны вместе со своими вторыми производными, то простейшие консервативные схемы обеспечивают равномерную сходимость y_n к $u(x)$ с погрешностью $O(h^2)$. Так называемая *наилучшая* консервативная схема обеспечивает погрешность $O(h^2)$ даже при коэффициентах, кусочно-непрерывных со своими вторыми производными, если выбраны *специальные* разностные сетки (в которых эти точки разрыва являются узлами).

Пример. Рассмотрим частный случай задачи Штурма — Лиувилля

$$u''(x)+\lambda u(x)=0, \quad u(0)=u(1)=0. \tag{95}$$

Точное решение этой задачи есть $\lambda_m=\pi^2m^2$, $u_m(x)=\sin \pi m x$, $m=1, 2, \ldots$; оно нужно для сравнения с численными расчетами. Простейшей разностной схемой для этой задачи является схема (94), в которой надо положить $p_n=q_n=0$. Эта схема имеет второй порядок точности.

Выполняя расчеты для сеток с числом интервалов $N=2, 3, 4$, приближенно определим три первых собственных значения.

*) Это исследование и доказательства приведенных ниже утверждений см. в [30].

Они представлены в таблице 22 вместе с точными значениями λ_m. Из таблицы видно, что с малой погрешностью определяются только те собственные значения, номер которых заметно меньше N. При сгущении сетки приближенные значения быстро стремятся к точным. Очень эффективным оказывается уточнение по правилу Рунге — Ромберга, также приведенное в таблице; уточнение λ_2 по двум сеткам дает неплохую точность, а уточнение λ_1 по трем сеткам — отличную.

Таблица 22

N	2	3	4	Уточненное по Рунге	Точное
λ_1	8,00	9,00	9,37	9,88	9,87
λ_2	—	27,0	32,0	38,4	39,5
λ_3	—	—	54,6	—	88,8

На этом примере хорошо видно, что сочетание схемы невысокого (обычно второго) порядка точности с правилом Рунге выгодно: оно обеспечивает высокую точность расчета при несложном алгоритме. Схемы высокого порядка точности обычно довольно громоздки, и организация расчета по ним сложнее.

5. Метод дополненного вектора. Для разностного метода, особенно в случае сложных нелинейных задач, важным и трудным является вопрос о фактическом вычислении разностного решения, ибо алгебраическая система (93) имеет заведомо высокий порядок. Для многих задач удобно находить это решение методом *дополненного вектора*. Изложим этот метод.

Заметим сначала, что метод стрельбы (и многие конкретные разностные алгоритмы) можно схематически описать следующим образом. Выбирается некоторое приближение $\lambda^{(0)}$; затем вычисляется соответствующее ему приближение $y^{(0)}(x)$. По этой функции находится новое приближение $\lambda^{(1)}$ и т. д. При этом собственное значение и собственная функция считаются элементами разных метрических пространств.

Будем рассуждать иначе. Разностную собственную функцию $\boldsymbol{y} = \{y_0, y_1, \ldots, y_N\}$ можно считать вектором в $(N+1)$-мерном пространстве. Увеличивая размерность пространства на единицу, рассмотрим собственное значение как новую компоненту этого вектора, $y_{N+1} \equiv \lambda$. Новый вектор $\boldsymbol{Y} = \{y_0, y_1, \ldots, y_N, y_{N+1}\}$ назовем *дополненным*. Относительно компонент дополненного вектора алгебраическая система (93) перепишется в каноническом виде

$$F_k(y_0, y_1, \ldots, y_N, y_{N+1}) = 0, \quad 0 \leqslant k \leqslant N+1. \tag{96}$$

Эта система нелинейна, даже если исходная задача была линейной относительно $u(x)$, как в примере (88).

Решать систему (96) будем методом Ньютона. Линеаризуя (96), получим на каждой итерации систему уравнений

$$\sum_{p=0}^{N+1} \frac{\partial F_k(Y^{(s)})}{\partial y_p} \delta y_p^{(s)} = -F_k(Y^{(s)}), \quad 0 \leqslant k \leqslant N+1, \qquad (97)$$

линейную относительно приращений неизвестных $\delta y_p^{(s)} = y_p^{(s+1)} - y_p^{(s)}$. Если искомое решение алгебраической системы (96) не особенное, т. е. в нем $\det(\partial F/\partial Y) \neq 0$, то при не слишком плохом нулевом приближении итерации (97) быстро сходятся к разностному решению. Отметим, что для линейных задач на собственные значения этот итерационный процесс совпадает с методом Дервюдье (см. главу VI, § 4, п. 2).

Удовлетворительное нулевое приближение для итераций (97) можно найти приближенными методами (метод Галеркина, разложение по малому параметру и т. д.), а в прикладных задачах его нередко удается получить из качественных соображений. Исключительно эффективна в таких задачах *комплексная организация расчета*, подробно описанная в § 2, п. 5.

Замечание 1. Метод дополненного вектора особенно полезен для уравнений, у которых задача Коши плохо обусловлена: он подавляет такую неустойчивость.

Замечание 2. Метод легко переносится на более общие задачи вида $A(u(x), \lambda) = 0$, где оператор A может быть интегро-дифференциальным (краевые условия предполагаются включенными в определение оператора). Вводя сетку x_n и аппроксимируя разностными выражениями все производные и интегралы, входящие в оператор, получим алгебраическую систему (96) и решим ее итерационным процессом (97) *).

Замечание 3. Недостатком метода является то, что при неудачном выборе нулевого приближения итерации (97) могут не сойтись, или в задачах со спектром собственных значений итерации могут сойтись не к искомому собственному значению.

Замечание 4. В методе дополненного вектора требуется решать систему линейных уравнений (97). Это легко делать, только если матрица системы целиком помещается в оперативной памяти ЭВМ (например, на БЭСМ-6 это будет при числе неизвестных $N \lesssim 150$). Это приводит к ограничению допустимого числа интервалов сетки.

Если требуется решить задачу для системы большого числа дифференциальных уравнений (например, уравнения Хартри—Фока для многоэлектронного атома), то даже при довольно грубой сетке число узловых значений

*) Здесь обсуждается только вопрос о вычислении разностного решения. Вопрос о его сходимости к точному решению при $h \to 0$ надо рассматривать отдельно; он связан со свойствами оператора A и выбором аппроксимации.

всех функций будет велико, и метод дополненного вектора применять трудно. В подобных задачах успешно применяется так называемый *непрерывный аналог метода Ньютона*, имеющий линейную сходимость итераций, но зато позволяющий оперировать с очень большим числом неизвестных. Этот метод является специальным вариантом метода последовательных приближений, организованным так, что итерации всегда сходятся.

6. Метод Галеркина. Многие приближенные методы пригодны для нахождения собственных значений и собственных функций задач, у которых краевые условия линейны относительно функции и ее производных. Среди этих методов к наиболее простым вычислениям приводит метод Галеркина.

Метод формулируется почти так же, как для краевых задач. Ищем решение задачи $A(u(x), \lambda) = f(x)$ в виде линейной комбинации отрезка полной системы функций $\varphi_k(x)$, $k \geqslant 1$:

$$u(x) \approx y_n(x) = \varphi_0(x) + \sum_{k=1}^{n} c_k \varphi_k(x), \qquad a \leqslant x \leqslant b, \tag{98}$$

выбранной так, чтобы удовлетворялись краевые условия. Потребуем, чтобы выполнялись условия ортогональности

$$\int_a^b [A(y_n(x), \lambda) - f(x)]\, \varphi_k(x)\, dx = 0, \qquad 1 \leqslant k \leqslant N. \tag{99}$$

Эти условия образуют алгебраическую систему n уравнений с $n+1$ неизвестным $c_1, c_2, \ldots, c_n, \lambda$. Недостающее уравнение надо получить из одного из краевых условий.

По тем же соображениям, что и в краевых задачах, удобнее пользоваться ортогональными системами функций $\varphi_k(x)$. В линейных задачах вычисления при этом заметно упрощаются.

Пример. Рассмотрим задачу (95)

$$u''(x) + \lambda u(x) = 0, \quad u(0) = u(1) = 0$$

и воспользуемся полной системой многочленов $\varphi_k(x) = x^k(1-x)$, которые заметно отличаются от точного решения задачи. Одним из дополнительных условий является условие нормировки решения. Благодаря линейности задачи его можно формулировать разными способами; для удобства вычислений зададим его в форме $y_n(0) = 1$, что означает $c_1 = 1$. Тогда, полагая $n = 1$ и 2, легко получим первые приближения

$$\begin{aligned} n = 1, &\quad \lambda^{\mathrm{I}} = 10, && y^{\mathrm{I}}(x) = x(1-x), \\ n = 2, &\quad \lambda^{\mathrm{I}} = 10, && y^{\mathrm{I}} = x(1-x), \\ &\quad \lambda^{\mathrm{II}} = 21, && y^{\mathrm{II}} = x(1-x) - \frac{11}{13} x^2(1-x). \end{aligned}$$

Первое собственное значение определилось с хорошей точностью, второе — с много худшей.

Методом Галеркина можно довольно хорошо находить наименьшие собственные значения. Но точность определения собственных функций обычно заметно хуже.

Обоснование метода Галеркина сложно. В частном случае, если дифференциальный оператор A линеен и однороден относительно $u(x)$, система (99) является задачей на определение собственных значений матрицы. Для задачи Штурма—Лиувилля метод Галеркина приводит к тем же самым алгебраическим уравнениям, что и метод Ритца (сходимость которого в задачах Штурма—Лиувилля доказана).

ЗАДАЧИ

1. Доказать теорему о сходимости метода Пикара, сформулированную в § 1, п. 3.

2. Вывести оценку (10) скорости сходимости метода Пикара.

3. В методе малого параметра вывести формулы для коэффициентов α_n и функций $v_n(x)$ в уравнениях (12).

4. Найти приближенное решение уравнения (3) методом малого параметра.

5. Для системы двух уравнений (25) написать схемы Рунге—Кутта второго порядка точности, аналогичные (22) и (23).

6. Для уравнения химического распада (34) составить схемы Рунге—Кутта второго и четвертого порядка точности и выяснить ограничения на шаг в этих схемах, следующие из положительности решения.

7. Составить для уравнения химического распада (34) специальную схему интегрирования по третьему способу из § 1, п. 8.

8. Вычисляя в (41) интеграл от второго слагаемого по формуле трапеций, получить неявную специальную схему; исследовать ее точность и найти ограничение на шаг сетки.

9. Написать формулы метода стрельбы применительно к краевой задаче (46) для одного дифференциального уравнения второго порядка.

10. Составить формулы метода Ньютона для нахождения корня уравнения (62б), возникающего при решении краевой задачи (60) методом стрельбы.

11. Решить краевую задачу (69) методом Галеркина, выбрав ортогональную систему функций $\varphi_k(x) = \sin 2kx$; сравнить результат с примером, приведенным в § 2, п. 6.

12. Для итерационного процесса при решении задачи на собственные значения (87) баллистическим методом составить а) формулы метода секущих, б) формулы метода Ньютона.

13. Для задачи на собственные значения (95) найти разностным методом при $N=2$ и $N=4$ первую собственную функцию и уточнить ее по правилу Рунге; ответ сравнить с точным решением.

ГЛАВА IX

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

В главе IX рассмотрены методы численного решения задач для уравнений в частных производных. В § 1 обсуждены некоторые постановки задач и дан обзор методов, которыми решаются подобные задачи. Остальные параграфы содержат изложение основ наиболее широко применяемого и хорошо изученного метода — разностного. В § 2 рассмотрены способы построения разностных схем и введено понятие аппроксимации. В § 3 даны методы исследования устойчивости разностных схем. В § 4 доказаны основные теоремы о сходимости разностного решения к точному.

§ 1. Введение

1. О постановках задач. Движение систем малого числа частиц математически описывают, как правило, обыкновенными дифференциальными уравнениями. Если же число частиц очень велико, то следить за движением отдельных частиц практически невозможно. При этом удобнее рассматривать систему частиц как сплошную среду, характеризуя ее состояние средними величинами: плотностью, температурой в данной точке и т. д.

Математические модели сплошной среды приводят к уравнениям в частных производных, которым удовлетворяют упомянутые средние величины. Например, изменение температуры в неподвижном теле описывается уравнением теплопроводности

$$c(u, \boldsymbol{r}, t)\frac{\partial u}{\partial t} = \operatorname{div}[k(u, \boldsymbol{r}, t)\operatorname{grad} u] + q(u, \boldsymbol{r}, t), \qquad (1)$$

где u — температура, c — теплоемкость, k — коэффициент теплопроводности и q — плотность источников тепла.

К уравнениям в частных производных приводят задачи газодинамики, теплопроводности, переноса излучения, распространения нейтронов, теории упругости, электромагнитных полей, процессов переноса в газах, квантовой механики и многие другие.

Независимыми переменными в физических задачах обычно являются время t и координаты $\boldsymbol{r}$; бывают и другие переменные, например, скорости частиц $\boldsymbol{v}$ в задачах переноса. Решение требуется найти в некоторой области изменения независимых пере-

менных $G(t, \boldsymbol{r}, \boldsymbol{v}, \ldots)$. Полная математическая постановка задачи содержит дифференциальное уравнение, а также дополнительные условия, позволяющие выделить единственное решение среди семейства решений дифференциального уравнения. Дополнительные условия обычно задаются на границе области G.

Если одной из переменных является t, то чаще всего рассматривают области вида

$$G(t, \boldsymbol{r}, \ldots) = g(\boldsymbol{r}, \ldots) \times [t_0, T], \tag{2}$$

т. е. решение ищут в некоторой пространственной области $g(\boldsymbol{r}, \ldots)$ на отрезке времени $t_0 \leqslant t \leqslant T$. В этом случае дополнительные условия, заданные при $t = t_0$, называют *начальными*, а дополнительные условия, заданные на границе $\Gamma(\boldsymbol{r})$ области $g(\boldsymbol{r})$, — *граничными* или *краевыми*.

Задачу, у которой имеются только начальные условия, называют *задачей Коши*. Например, для уравнения теплопроводности (1) в неограниченном пространстве можно поставить задачу с начальными условиями

$$u(\boldsymbol{r}, t_0) = \mu(\boldsymbol{r}). \tag{3}$$

Если $\mu(\boldsymbol{r})$ — кусочно-непрерывная ограниченная функция, то решение задачи (1), (3) единственно в классе ограниченных функций (при некоторых ограничениях на коэффициенты уравнения; см. [40]).

Задачу с начальными и граничными условиями называют *смешанной краевой задачей* или *нестационарной краевой задачей*. Для уравнения (1) дополнительные условия такой задачи могут иметь, например, вид

$$u(\boldsymbol{r}, t_0) = \mu(\boldsymbol{r}), \quad \boldsymbol{r} \in g(\boldsymbol{r}), \qquad u(\boldsymbol{r}, t)_\Gamma = \mu_1(\boldsymbol{r}, t), \quad t_0 \leqslant t \leqslant T. \tag{4}$$

Для этого уравнения допустимы и другие граничные условия, например, содержащие нормальную производную решения.

Встречаются задачи, в которых область $G(t, \boldsymbol{r})$ имеет другой вид. Примером является задача с условиями на характеристиках (см. [40]), возникающая при изучении процессов сушки, сорбции газов и многих других.

При исследовании установившихся состояний или стационарных (не зависящих от времени) процессов в сплошной среде формулируются математические задачи, не зависящие от времени. Их решение ищется в области $g(\boldsymbol{r})$, а дополнительные условия являются граничными. Такие задачи называют *краевыми*.

Мы ограничимся рассмотрением корректно поставленных задач, когда для некоторого класса начальных и граничных данных решение (в заданном классе функций) существует, единственно и непрерывно зависит от этих данных. Будем также предполагать, что решение непрерывно зависит от всех коэффициентов уравнения,

Для уравнений в частных производных существуют физически интересные задачи, являющиеся некорректно поставленными: обратные задачи теплопроводности, задачи на развитие неустойчивостей и другие. Так, рассмотренный в главе I пример Адамара связан с возникновением релей-тейлоровской неустойчивости, когда слой тяжелой жидкости налит поверх слоя легкой жидкости. Но здесь мы такие задачи не будем рассматривать (см. [39] и приведенную там библиографию).

В этой главе излагаются методы численного решения уравнений в частных производных и способы обоснования этих методов. Они применимы к широким классам уравнений и различным типам задач для них. Но примеры, иллюстрирующие изложение и конкретные применения этих методов, рассмотренные в главах X — XIII, касаются наиболее распространенных и хорошо изученных задач для уравнений первого и второго порядков, линейных относительно производных.

Напомним классификацию таких уравнений. Они имеют следующий вид (для простоты мы ограничиваемся случаем двух переменных):

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0. \tag{5}$$

Коэффициенты уравнения (5), вообще говоря, зависят от u, x, y. Если коэффициенты не зависят от переменных, то это линейное уравнение с постоянными коэффициентами. Если F линейно зависит от u, а остальные коэффициенты от u не зависят, то это линейное уравнение с переменными коэффициентами. Если коэффициенты зависят от u, то уравнение (5) называется квазилинейным.

Если $A \equiv B \equiv C \equiv 0$, но $D \neq 0$ и $E \neq 0$, то уравнение (5) имеет первый порядок и называется уравнением переноса.

Уравнения второго порядка классифицируются по знаку дискриминанта $B^2 - AC$: у гиперболических уравнений дискриминант положителен, у параболических — равен нулю, у эллиптических — отрицателен.

Те физические процессы, которые описываются разными перечисленными здесь типами уравнений, существенно отличаются друг от друга. Соответственно полные постановки задач для этих типов уравнений имеют свои особенности, подробно рассмотренные в [40]; мы будем кратко напоминать их в соответствующих главах.

Заметим, что уравнение с переменными коэффициентами может иметь разный тип в разных точках области $G(x, y)$. В практике вычислений встречается немало подобных задач, причем нередко — еще неисследованных теоретически. При этом сформулировать полную постановку задачи и обосновать ее корректность зачастую бывает нелегко.

2. Точные методы решения. В курсах уравнений математической физики изложен ряд методов, позволяющих для некоторых классов задач найти точное решение (см. [40]). К таким методам

относятся метод распространяющихся волн, метод разделения переменных, метод функций источника и другие.

Например, для простейшей задачи теплопроводности

$$\begin{gathered} u_t = ku_{xx}, \quad 0 \leqslant x \leqslant a, \quad 0 \leqslant t, \quad k = \text{const} > 0, \\ u(0, t) = u(a, t) = 0, \quad u(x, 0) = \mu(x), \end{gathered} \tag{6}$$

где функция $\mu(x)$ кусочно-непрерывна, методом разделения переменных решение представляется в виде ряда

$$u(x, t) = \sum_{n=1}^{\infty} \alpha_n e^{-\pi^2 n^2 kt/a^2} \sin \frac{\pi n x}{a}, \tag{7а}$$

где величины α_n являются коэффициентами Фурье начальных данных

$$\alpha_n = \frac{2}{a} \int_0^a \mu(x) \sin \frac{\pi n x}{a}\, dx. \tag{7б}$$

Таким образом, получено явное выражение решения через начальные данные.

Подставляя (7б) в (7а) и меняя порядок интегрирования и суммирования, выразим решение через начальные данные и функцию источника

$$u(x, t) = \int_0^a G(x, \xi, t)\, \mu(\xi)\, d\xi, \tag{8а}$$

где функция источника равна

$$G(x, \xi, t) = \frac{2}{a} \sum_{n=1}^{\infty} e^{-\pi^2 n^2 kt/a^2} \sin \frac{\pi n x}{a} \sin \frac{\pi n \xi}{a}. \tag{8б}$$

Для задачи Коши на бесконечной прямой выражение для функции источника имеет следующий вид (см. [40]):

$$G_\infty(x, \xi, t) = \frac{1}{2\sqrt{\pi k t}} e^{-(x-\xi)^2/4kt}. \tag{8в}$$

Точные методы позволяют получить явное выражение решения через начальные данные, что облегчает дальнейшие действия с решением. Например, выражения (7) — (8) позволяют многое сказать о качественном поведении решения.

В самом деле, в формуле (7а) пространственные гармоники $\sin(\pi n x/a)$ множатся на величины $\exp(-\pi^2 n^2 kt/a^2)$, затухающие при возрастании времени; это затухание тем быстрей, чем больше номер гармоники. Но чем меньше амплитуды высоких гармоник,

тем более плавно меняется функция. Следовательно, с течением времени решение задачи (6) должно сглаживаться.

Наоборот, при движении в обратную сторону по времени амплитуды высоких гармоник возрастают тем быстрей, чем больше n; при $n \to \infty$ скорость роста гармоник *неограниченно* увеличивается. Отсюда легко понять, что обратная задача теплопроводности неустойчива.

Заметим, что функция источника на бесконечной прямой положительна: $G_\infty(x, \xi, t) > 0$ при $t > 0$. Следовательно, если в решение (8а) с бесконечными пределами интегрирования подставить начальные данные вида

$$\mu(x) > 0 \text{ на } [a, b], \qquad \mu(x) = 0 \text{ вне } [a, b],$$

то при $t > 0$ решение будет отлично от нуля в любой точке бесконечной прямой. Это означает, что в случае линейной теплопроводности скорость распространения тепла бесконечна.

Таким образом, точные методы очень полезны. Однако они применимы в основном к линейным задачам в областях простой формы (прямоугольник, круг и т. п.), когда дифференциальное уравнение и краевые условия линейны относительно $u(\boldsymbol{r}, t)$ и ее производных. При этом выкладки удается довести до конца обычно лишь для уравнений с постоянными или кусочно-постоянными коэффициентами.

3. Автомодельность и подобие. Для уравнений в частных производных существуют такие частные решения, когда $u(x, t)$ является функцией одной переменной ξ, роль которой играет некоторая комбинация независимых переменных x, t. Такие решения называются *автомодельными*.

Построим пример автомодельного решения. Рассмотрим одномерное уравнение теплопроводности, в котором коэффициент теплопроводности зависит от температуры по степенному закону

$$\frac{\partial u(x, t)}{\partial t} = \frac{\partial}{\partial x}\left[k(u)\frac{\partial u(x, t)}{\partial x}\right], \qquad k(u) = k_0 u^m, \quad k_0 > 0, \quad m > 0. \tag{9}$$

Такая зависимость часто встречается в физических задачах; например, коэффициент электронной теплопроводности плазмы приблизительно пропорционален $u^{5/2}$.

Будем искать частное решение уравнения (9), имеющее вид бегущей волны

$$u(x, t) = f(\xi), \quad \xi = x - ct.$$

При подстановке такого решения уравнение (9) преобразуется в обыкновенное дифференциальное уравнение

$$c\frac{df}{d\xi} + \frac{d}{d\xi}\left(k_0 f^m \frac{df}{d\xi}\right) = 0. \tag{10}$$

Однократное интегрирование этого уравнения дает соотношение

$$f(\xi)+\frac{k_0}{c}f^m(\xi)\frac{df}{d\xi}=\text{const}. \tag{11}$$

Если функция $f(\xi)$ обращается в нуль хотя бы в одной точке ξ_0, то константа в правой части (11) равна нулю и соотношение (11) нетрудно проинтегрировать еще раз:

$$f(\xi)=\left[\frac{cm}{k_0}(\xi_0-\xi)\right]^{1/m} \quad \text{при } \xi\leqslant\xi_0.$$

Доопределим решение при $\xi>\xi_0$, полагая $f(\xi)=0$, что удовлетворяет уравнению (9). Таким образом, получим искомое решение

$$\begin{aligned} &u(x,\ t)=\left[\frac{cm}{k_0}(x_0-x+ct)\right]^{1/m}, && x\leqslant x_0+ct,\\ &u(x,\ t)=0, && x>x_0+ct. \end{aligned} \tag{12}$$

Чтобы это решение могло существовать, начальные и граничные условия должны быть с ним согласованы. Например, если уравнение (9) рассматривается при $t>0$ на полупрямой $x\geqslant x_0$, то следует задать условия

$$\begin{aligned} &u(x,\ 0)=0, \quad x\geqslant x_0,\\ &u(x_0,\ t)=\left(\frac{c^2m}{k_0}t\right)^{1/m}, \quad t\geqslant 0. \end{aligned} \tag{13}$$

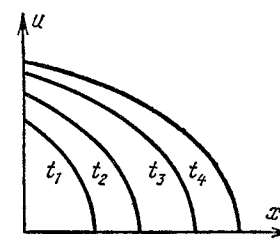


Рис. 45.

Автомодельное решение (12) представляет собой температурную волну, бегущую с постоянной скоростью по нулевому фону температуры (рис. 45). Скорость движения волны c определяется скоростью роста температуры на границе (13). Точка $\bar{x}=x_0+ct$ является фронтом волны. Профиль температуры всюду непрерывен, но на фронте он имеет вертикальную касательную $(\partial u/\partial x)_{\bar{x}}=\infty$, и производная в этой точке терпит разрыв.

Автомодельные решения довольно часто удается найти для линейных и квазилинейных уравнений или систем уравнений, коэффициенты которых зависят от переменных x, t и решения u по степенным законам. Для построения решения надо «угадать» подходящую комбинацию независимых переменных и описанным выше приемом свести уравнение в частных производных к обыкновенному дифференциальному уравнению. Выразить решение этого уравнения через элементарные функции, подобно (12), удается далеко не всегда, но найти это решение численным интегрированием несравненно проще, чем численно решить исходное уравнение в частных производных.

Если уравнение в частных производных описывает сложный физический процесс, то автомодельные решения дают отдельные режимы протекания процесса и позволяют исследовать многие его особенности. Поэтому автомодельные решения широко используются в современной физике (см. [36]).

Автомодельность является частным случаем *подобия*. В теории подобия при помощи анализа физических размерностей коэффициентов уравнения ищутся такие преобразования всех переменных и функций, относительно которых уравнение инвариантно. Например, уравнение (9) не изменится при таком преобразовании:

$$x \to \alpha x, \quad t \to \alpha t, \quad u \to \alpha^{1/m} u. \tag{14}$$

Если для уравнения известно преобразование подобия, то, найдя каким-либо способом одно частное решение, мы при помощи этого преобразования получим целое семейство решений. Это особенно ценно, если задача настолько сложна, что частные решения удается находить только трудоемкими численными методами.

Разумеется, автомодельные решения или преобразования подобия существуют далеко не для всех классов уравнений, а лишь при некоторых видах коэффициентов уравнения и начальных и граничных условиях. Однако многие важные физические задачи точно или приближенно удовлетворяют этим ограничениям.

4. Численные методы. Задачи для нелинейных уравнений с коэффициентами достаточно общего вида или даже линейные задачи, но в областях сложной формы, редко удается решить классическими методами. Основным способом решения таких задач являются численные методы. Среди них чаще всего применяют *разностные методы* благодаря их универсальности и наличию хорошо разработанной теории.

Для применения разностного метода в области изменения переменных $G(r, t)$ вводят некоторую сетку. Все производные, входящие в уравнение и краевые условия, заменяют разностями (или другими алгебраическими комбинациями) значений функции $u(r, t)$ в узлах сетки. Получающиеся при этом алгебраические уравнения называют *разностной схемой*. Решая полученную алгебраическую систему, найдем приближенное (разностное) решение в узлах сетки.

Как и в главе VIII, возникают вопросы: существует ли решение алгебраической системы и единственно ли оно; как это решение фактически вычислить (за возможно меньшее число действий); при каких условиях это разностное решение стремится к точному и какова скорость сходимости? Есть еще два вопроса, которые для обыкновенных дифференциальных уравнений были несложными: как выбрать сетку и как составить разностную схему на этой сетке?

Пример. Составим простейшие разностные схемы для одномерной задачи линейной теплопроводности на ограниченном отрезке

$$u_t = ku_{xx}, \quad 0 < x < a, \quad 0 < t \leqslant T, \tag{15а}$$

$$u(x, 0) = \mu(x), \quad u(0, t) = \mu_1(t), \quad u(a, t) = \mu_2(t). \tag{15б}$$

Решение ищется в области $G = [0 \leqslant x \leqslant a] \times [0 \leqslant t \leqslant T]$.

Введем в G прямоугольную сетку (для простоты равномерную), образованную пересечением линий $x_n = nh$, $0 \leqslant n \leqslant N$, и $t_m = m\tau$, $0 \leqslant m \leqslant M$; величины h, τ являются шагами сетки по переменным x, t (рис. 46). Значения функции в узлах сетки будем обозначать $u_n^m = u(x_n, t_m)$.

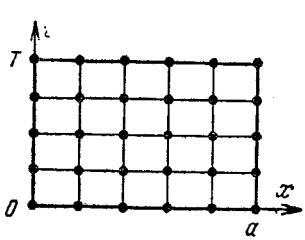
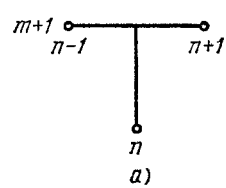


Рис. 46.



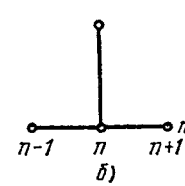


Рис. 47.

Возьмем около узла (x_n, t_m) конфигурацию узлов, изображенную на рис. 47, *а*. Заменим в уравнении (15а) производную u_t разностным отношением $(u_n^{m+1} - u_n^m)/\tau$, а производную u_{xx} — отношением $(u_{n+1}^{m+1} - 2u_n^{m+1} + u_{n-1}^{m+1})/h^2$. Тогда дифференциальное уравнение приближенно заменится (аппроксимируется) разностной схемой *)

$$\frac{1}{\tau}(y_n^{m+1} - y_n^m) = \frac{k}{h^2}(y_{n+1}^{m+1} - 2y_n^{m+1} + y_{n-1}^{m+1}), \quad 1 \leqslant n \leqslant N-1. \tag{16}$$

Число уравнений (16) меньше числа неизвестных y_n^{m+1}, $0 \leqslant n \leqslant N$; недостающие уравнения выводятся из начальных и граничных данных (15б):

$$y_n^0 = \mu(x_n), \quad 0 \leqslant n \leqslant N, \quad y_0^{m+1} = \mu_1(t_{m+1}), \quad y_N^{m+1} = \mu_2(t_{m+1}). \tag{17}$$

Конфигурацию узлов, используемую для составления разностной схемы, называют *шаблоном*.

Для одной и той же задачи можно составить много разностных схем. Например, если для задачи (15) выбрать изображенный

*) Напомним, что разностной схеме удовлетворяет разностное решение, которое мы обозначаем y_n^m.

на рис. 47, *б* шаблон, то вместо (16) получим другую схему:

$$\frac{1}{\tau}(y_n^{m+1}-y_n^m)=\frac{k}{h^2}(y_{n+1}^m-2y_n^m+y_{n-1}^m),\quad 1\leqslant n\leqslant N-1. \quad (18)$$

Начальные и граничные условия для этой схемы можно записать в форме (17).

В этой главе рассмотрены способы составления и исследования разностных схем, применимые для разных типов задач. В следующих главах излагаются те разностные схемы, которые дают хорошие результаты при решении некоторых распространенных типов уравнений математической физики, возникающих в задачах переноса, теплопроводности и диффузии, акустики и газодинамики, стационарных электрических полей.

Есть численные методы, близкие к разностным. Например, в *методе прямых* сетка вводится только для части переменных; эти переменные рассматриваются как дискретные, а одна переменная (обычно время t) остается непрерывной. Производные по дискретным переменным заменяются разностями. При этом уравнение в частных производных аппроксимируется *дифференциально-разностными уравнениями*, которые представляют собой систему большого числа обыкновенных дифференциальных уравнений. Метод прямых оказывается в некоторых случаях удобным.

Для некоторых важных классов задач развиты специальные численные методы, обычно основанные на каких-либо грубых физических моделях процессов. Так, для задач многомерной газодинамики разработан метод частиц в ячейке; для задач разреженной плазмы предложен метод «водяного мешка» и ряд других (см. [6]); в задачах переноса нейтронов комбинируют разностный метод с разложением угловой части функции распределения частиц по сферическим гармоникам и т. д.

Численные методы позволяют решить сложнейшие задачи для систем многомерных уравнений. Однако для сложных задач численные методы очень трудоемки и рассчитаны на использование мощных ЭВМ. В этих случаях даже вывод разностной схемы, составление программы и ее отладка могут занимать несколько месяцев, а разработка математической модели или новых типов разностных схем нередко требует нескольких лет.

Поэтому численные методы целесообразно использовать в сочетании с аналитическими методами. Например, ищут такие упрощенные постановки задачи или частные случаи, когда можно найти точные или автомодельные решения и преобразования подобия. При помощи преобразования подобия по каждому найденному численному решению строят семейство решений. Все это позволяет с меньшими затратами труда провести детальное исследование исходной задачи.

§ 2. Аппроксимация

1. Сетка и шаблон. Для большинства разностных схем узлы сетки лежат на пересечении некоторых прямых линий (в многомерных задачах — гиперплоскостей), проведенных либо в естественной системе координат, либо в специально подобранной по форме области G.

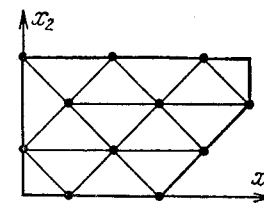


Рис. 48.

Для двумерных задач в прямоугольной области G наиболее часто употребляют прямоугольную сетку (см. рис. 46), которую мы ввели при составлении схем (16) и (18). Заметно реже используют треугольную (рис. 48) или шестиугольную сетку. Для трехмерных задач наиболее употребительна сетка из прямоугольных параллелепипедов (рис. 49); другие виды сеток, например из прямоугольных трехгранных призм (рис. 50), используются редко.

Существуют некоторые разностные схемы, например, для задач двумерной и трехмерной газодинамики, где узлы сетки расположены неупорядоченно. Но такие схемы сколько-нибудь заметного распространения не получили.

Если одна из переменных имеет физический смысл времени t, то сетку обычно строят так, чтобы среди ее линий (или гиперплоскостей) были линии $t = t_m$. Совокупность узлов сетки, лежащих на такой линии или гиперплоскости, называют *слоем*.

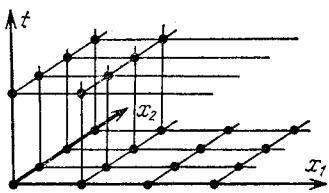


Рис. 49.

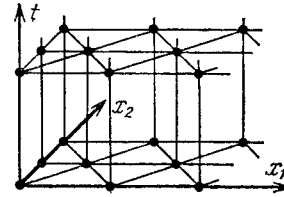


Рис. 50.

На каждом слое выделяют *направления* — линии, вдоль которых меняется только одна пространственная координата. Например, для переменных x, y, t есть направление x ($t = \text{const}$, $y = \text{const}$) и направление y ($t = \text{const}$, $x = \text{const}$).

Если область $G(x, t)$ имеет форму прямоугольника, то часть узлов прямоугольной сетки естественно ложится на границу области (см. рис. 46); эти узлы называются *граничными*, а остальные узлы — *внутренними*. Начальные и краевые условия, наложенные на решение на границе $\Gamma(G)$, можно в этом случае

считать заданными в граничных узлах сетки. Именно так было сделано при выводе соотношений (17) в примере из § 1, п. 4.

В случае двух пространственных переменных x, y граница области $G(x, y)$ нередко бывает ломаной линией. Для таких областей всегда можно ввести такую треугольную сетку, чтобы естественные узлы сетки (пересечения линий сетки) легли на границу (см. рис. 48). Иногда удается добиться этого, используя прямоугольную сетку.

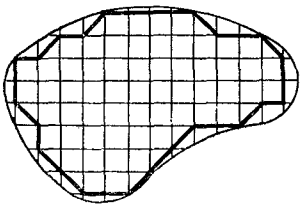


Рис. 51.

Если граница $\Gamma(G)$ криволинейная, то естественные узлы сетки на границу могут не попадать (рис. 51). В этом случае можно взять точки пересечения линий сетки с границей в качестве дополнительных узлов; тогда краевые условия следует задавать в этих узлах. Можно сделать иначе: границу $\Gamma(G)$ приближенно заменить ломаной, проходящей через ближайшие к границе естественные узлы (жирная линия на рис. 51); тогда краевые условия, заданные на $\Gamma(G)$, надо каким-либо образом перенести на эту ломаную.

Если область G является кругом (кольцом), цилиндром или шаром, то часто переходят к системам координат, связанных с видом области: полярным, цилиндрическим или сферическим. Если в таких координатах ввести прямоугольную сетку, то естественные узлы сетки лягут на границу. Иногда для построения хорошей сетки в областях сложной формы прибегают к конформному отображению на квадрат, в котором введена прямоугольная сетка.

Составляя разностные схемы (16) и (18), мы использовали во всех внутренних точках области однотипную разностную аппроксимацию производных. Иными словами, при написании каждого разностного уравнения около некоторого узла сетки бралось одно и то же количество узлов, образующее строго определенную конфигурацию. Эту конфигурацию узлов называют *шаблоном* данной разностной схемы (см. рис. 47).

Узлы, в которых разностная схема записана на шаблоне, называются *регулярными*, а остальные узлы — *нерегулярными*. Нерегулярными являются обычно граничные узлы, а иногда также лежащие вблизи границы узлы (такие, что взятый около этого узла шаблон выходит за границу области). Так, в примере из § 1, п. 4 нерегулярными были граничные узлы, и в них разностная схема имела нестандартную форму (17).

Составление разностной схемы начинается с выбора шаблона. Шаблон не всегда определяет разностную схему однозначно, но существенно влияет на ее свойства; например, далее мы увидим, что на шаблоне рис. 47, *б* нельзя составить хорошей схемы для задачи (15). Для каждого типа уравнений и краевых задач тре-

буется свой шаблон. В следующих главах сформулированы (на основе свойств решаемых уравнений) некоторые общие соображения, которые позволяют подбирать шаблоны, пригодные для построения хороших разностных схем.

Существуют разностные схемы, вообще не имеющие шаблона (пример такой схемы будет приведен в главе X). Но логическая структура таких схем сложна, что вызывает заметное увеличение объема программ и времени счета на ЭВМ. Поэтому такие схемы мало употребительны.

2. Явные и неявные схемы. Обсудим вопрос о фактическом вычислении разностного решения. Бо́льшая часть физических проблем приводит к уравнениям, содержащим время в качестве одной из переменных. Для таких уравнений ставится обычно смешанная краевая задача, типичным случаем которой является (15).

К подобным задачам применяют послойный алгоритм вычислений. Рассмотрим его на примере схем (18) и (16).

В схеме (18) на исходном слое $m=0$ решение известно в силу начального условия. Положим $m=0$ в уравнениях (18). Тогда при каждом значений индекса n уравнение содержит только одно неизвестное y_n^1; отсюда можно определить y_n^1 при $1\leqslant n\leqslant N-1$. Значения y_0^1 и y_N^1 определяются из краевых условий (17). Таким образом, значения решения на первом слое вычислены. По ним аналогичным образом вычисляется решение на втором слое и т. д.

Схема (18) в каждом уравнении содержит только одно значение функции на следующем слое; это значение нетрудно явно выразить через известные значения функции на данном слое. Поэтому такие схемы называются *явными*.

Схема (16) содержит в каждом уравнении несколько неизвестных значений функции на новом слое; подобные схемы называются *неявными*. Для фактического вычисления решения перепишем схему (16) с учетом краевого условия (17) в следующей форме:

$$\begin{gathered} y_{n-1}^{m+1}-\left(2+\frac{h^2}{k\tau}\right)y_n^{m+1}+y_{n+1}^{m+1}=\frac{h^2}{k\tau}\,y_n^m,\quad 1\leqslant n\leqslant N-1,\\ y_0^{m+1}=\mu_1(t_{m+1}),\quad y_N^{m+1}=\mu_2(t_{m+1}). \end{gathered} \tag{19}$$

На каждом слое схема (19) представляет собой систему линейных уравнений для определения величин y_n^{m+1}; правые части этих уравнений известны, поскольку содержат значения решения с предыдущего слоя. Матрица линейной системы трехдиагональна, и решение можно вычислить алгебраической прогонкой.

Рассмотренный сейчас алгоритм достаточно типичен. Он используется во многих неявных разностных схемах для одномерных и многомерных задач. Дальше мы будем вместо индекса времени m часто применять сокращенные обозначения:

$$u(x_n,\ t_m)=u_n,\ u(x_n,\ t_{m+1})=\hat{u}_n,\ u(x_n,\ t_{m-1})=\check{u}_n. \tag{20}$$

В этих обозначениях разностная схема (18) примет вид

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{k}{h^2}(y_{n+1} - 2y_n + y_{n-1}).$$

3. Невязка. Рассмотрим операторное уравнение общего вида (не обязательно линейное):

$$Au = f, \text{ или } Au - f = 0. \tag{21}$$

Заменяя оператор A разностным оператором A_h, правую часть f — некоторой сеточной функцией φ_h, а точное решение u — разностным решением y, запишем разностную схему:

$$A_h y = \varphi_h, \text{ или } A_h y - \varphi_h = 0. \tag{22}$$

Если подставить точное решение u в соотношение (22), то решение, вообще говоря, не будет удовлетворять этому соотношению: $A_h u - \varphi_h \neq 0$. Величину

$$\psi = \varphi_h - A_h u \equiv (Au - f) - (A_h u - \varphi_h) \tag{23}$$

называют *невязкой*.

Невязку обычно оценивают при помощи разложения в ряд Тейлора. Например, найдем невязку явной разностной схемы (18) для уравнения теплопроводности (15а). Запишем это уравнение в каноническом виде:

$$Au \equiv \left(\frac{\partial}{\partial t} - k\frac{\partial^2}{\partial x^2}\right) u = 0.$$

Поскольку в данном случае $f = \varphi_h = 0$, то

$$\psi_n = (Au - A_h u)_n =$$

$$= \left(\frac{\partial u}{\partial t}\right)_n - k\left(\frac{\partial^2 u}{\partial x^2}\right)_n - \frac{1}{\tau}(\hat{u}_n - u_n) + \frac{k}{h^2}(u_{n+1} - 2u_n + u_{n-1}).$$

Разложим решение по формуле Тейлора около узла (x_n, t_m), предполагая существование непрерывных четвертых производных по x и вторых по t:

$$\hat{u}_n = u_n + \tau u_t(x_n, t_m) + \frac{1}{2}\tau^2 u_{tt}(x_n, \theta_m),$$

$$u_{n\pm1} = u_n \pm h u_x(x_n, t_m) + \frac{1}{2}h^2 u_{xx}(x_n, t_m) \pm \tag{24}$$

$$\pm \frac{1}{6}h^3 u_{xxx}(x_n, t_m) + \frac{1}{24}h^4 u_{xxxx}(\xi_{n\pm1}, t_m),$$

где $t_m < \theta_m < t_{m+1}$, $x_{n-1} < \xi_{n-1} < x_n < \xi_{n+1} < x_{n+1}$. Подставляя эти разложения в выражение невязки и пренебрегая, в силу непре-

рывности производных, отличием величин $\xi_{n\pm1}$, θ_m от x_n, t_m, найдем

$$\psi_n = \left(-\frac{\tau}{2} u_{tt} + \frac{kh^2}{12} u_{xxxx}\right)_n = O(\tau + h^2). \quad (25)$$

Таким образом, невязка (25) стремится к нулю при $\tau \to 0$ и $h \to 0$.

Выражение (25) дает невязку только в регулярных узлах схемы (18). Сравнивая (17) и (15б), легко получим невязку в нерегулярных узлах $\psi_0 = \psi_N = 0$.

Замечание 1. Решение задачи теплопроводности с постоянным коэффициентом (15) в области $G = (0 < x < a) \times (0 < t \leqslant T)$ непрерывно дифференцируемо бесконечное число раз. Однако учет пятых и более высоких производных в разложениях (24) прибавляет к невязке (25) только члены более высокого порядка малости по τ и h, т. е., по существу, не меняет вида невязки.

Замечание 2. Пусть по каким-либо причинам решение исходной задачи дифференцируемо небольшое число раз; например, в задачах с переменным коэффициентом теплопроводности, гладким, но не имеющим второй производной, решение имеет лишь третьи непрерывные производные. Тогда в разложении (24) последними будут члены $\pm h^3 u_{xxx}(\xi_{n\pm1}, t_m)/6$, не точно компенсирующие друг друга. Это приведет к появлению в невязке (25) члена типа $hu_{xxx} = O(h)$, т. е. невязка будет иметь меньший порядок малости, чем для четырежды непрерывно дифференцируемых решений.

Замечание 3. Преобразуем выражение невязки с учетом того, что входящая в него функция $u(x, t)$ есть точное решение исходного уравнения и для нее выполняются соотношения

$$u_{tt} = \frac{\partial}{\partial t}(ku_{xx}) = k\frac{\partial^2}{\partial x^2}(u_t) = k^2 u_{xxxx}.$$

Подставляя это выражение в (25), получим

$$\psi_n = \left(\frac{1}{12} kh^2 - \frac{1}{2} k^2\tau\right)(u_{xxxx})_n. \quad (26)$$

Если выбрать шаги по пространству и времени так, чтобы $\tau = h^2/(6k)$, то главный член невязки обратится в нуль и останутся только члены более высокого порядка малости по τ и h (которые мы опускали).

Этот прием применяется при построении разностных схем повышенной точности.

4. Методы составления схем. Есть три основных способа составления разностных схем на заданном шаблоне: метод разностной аппроксимации, интегро-интерполяционный метод и метод неопределенных коэффициентов.

Метод разностной аппроксимации заключается в том, что каждая производная, входящая в дифференциальное уравнение и краевые условия, заменяется каким-либо разностным

выражением (включающим только узлы шаблона). Именно так были составлены схемы (16) и (18). Этот метод очень прост и в дополнительных пояснениях не нуждается.

Метод разностной аппроксимации позволяет легко составить схему первого или второго порядка аппроксимации на прямоугольной сетке для уравнений с непрерывными (и достаточно гладкими) коэффициентами. Однако этот метод трудно или даже невозможно применять в более сложных случаях: для уравнений с разрывными коэффициентами, на не прямоугольных сетках, для уравнений высокого порядка на неравномерных сетках и т. д.

Схемы повышенной точности в этом методе составляют, исследуя выражение невязки аналогично замечанию 3 в п. 3.

Интегро-интерполяционный метод, один из вариантов которого называется *методом баланса*, наиболее надежен и применим во всех случаях. В этом методе после выбора шаблона область $G(r, t)$ разбивают на ячейки, определенным образом связанные с шаблоном. Дифференциальное уравнение интегрируют по ячейке и по формулам векторного анализа приводят к интегральной форме, соответствующей физическому закону сохранения. Приближенно вычисляя полученные интегралы по каким-либо квадратурным формулам, составляют разностную схему.

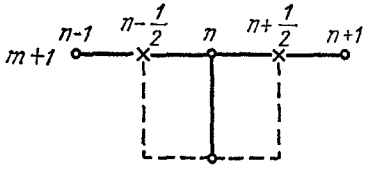


Рис. 52.

Например, для уравнения теплопроводности с переменным коэффициентом $u_t = (ku_x)_x$ выберем шаблон, изображенный на рис. 52 (см. также рис. 47, *а*), и сопоставим ему ячейку, показанную пунктиром. Обозначая средние точки интервалов сетки полуцелыми индексами, выполним интегрирование по ячейке:

$$0 = \int_{t_m}^{t_{m+1}} dt \int_{x_{n-1/2}}^{x_{n+1/2}} dx\,[u_t - (ku_x)_x] =$$

$$= \int_{x_{n-1/2}}^{x_{n+1/2}} (\hat{u} - u)\,dx - \int_{t_m}^{t_{m+1}} [(ku_x)_{n+1/2} - (ku_x)_{n-1/2}]\,dt.$$

Это соотношение является точным. В правой части приближенно вычислим первый интеграл по формуле средних, а второй — по формуле правых прямоугольников. Получим следующее выражение:

$$(\hat{y}_n - y_n)(x_{n+1/2} - x_{n-1/2}) = \tau[(\hat{k}\hat{y}_x)_{n+1/2} - (\hat{k}\hat{y}_x)_{n-1/2}].$$

Заменяя в правой части производные разностями и учитывая,

что на равномерной сетке $x_{n+1/2}-x_{n-1/2}=h$, получим разностную схему

$$\frac{1}{\tau}(\hat{y}_n-y_n)=\frac{1}{h^2}\left[\hat{k}_{n+1/2}(\hat{y}_{n+1}-\hat{y}_n)-\hat{k}_{n-1/2}(\hat{y}_n-\hat{y}_{n-1})\right]. \quad (27)$$

Если $k=\text{const}$, то схема совпадает с неявной схемой (16).

Интегро-интерполяционный метод особенно полезен для уравнений с негладкими или разрывными коэффициентами, поскольку именно интегральная запись законов сохранения выделяет из всех математически допустимых решений таких уравнений физически правильное обобщенное решение.

Метод неопределенных коэффициентов заключается в том, что в качестве разностной схемы берут линейную комбинацию значений разностного решения в узлах шаблона. Коэффициенты этой линейной комбинации определяют из условия, чтобы невязка схемы имела как можно более высокий порядок малости относительно τ и h.

Например, для уравнения $u_t=ku_{xx}$ на шаблоне рис. 52 будем искать разностную схему в следующем виде:

$$\alpha\hat{y}_{n-1}+\beta\hat{y}_n+\gamma\hat{y}_{n+1}+\delta y_n=0. \quad (28)$$

Подставим сюда разложения (24), ограничиваясь для простоты членами $O(\tau)$ и $O(h^2)$, и вычтем схему (28) из исходного уравнения. Получим невязку (индекс n всюду опускаем)

$$\psi=u_t-ku_{xx}-(\alpha+\beta+\gamma+\delta)u+\tau\delta u_t+$$
$$+(\alpha-\gamma)hu_x-\frac{1}{2}(\alpha+\gamma)h^2u_{xx}+\delta\cdot O(\tau^2)+(\alpha-\gamma)O(h^3)+\ldots$$

Чтобы сократились выписанные здесь члены, надо положить

$$\alpha+\beta+\gamma+\delta=0,\quad \tau\delta=-1,\quad \alpha-\gamma=0,$$
$$\frac{1}{2}(\alpha+\gamma)h^2=-k.$$

Отсюда находим коэффициенты:

$$\alpha=\gamma=-\frac{k}{h^2},\quad \beta=\frac{2k}{h^2}+\frac{1}{\tau},\quad \delta=-\frac{1}{\tau}.$$

Подставляя их в (28), получим разностную схему (16).

Метод неопределенных коэффициентов применим на косоугольных сетках. Например, при его помощи нетрудно составить пятиточечную схему для уравнения $u_t=ku_{xx}$ на треугольной сетке с шаблоном рис. 53:

$$\frac{1}{2\tau}(\hat{y}_n+\hat{y}_{n-1})=\left(\frac{1}{8\tau}+\frac{k}{h^2}\right)(y_{n-1}+y_{n+1})+\left(\frac{3}{4\tau}-\frac{2k}{h^2}\right)y_n. \quad (29)$$

Возможны случаи, когда часть коэффициентов схемы типа (28) определяют из условия наивысшего порядка малости невязки, а часть коэффициентов выбирают из других соображений.

Метод неопределенных коэффициентов (как и метод разностной аппроксимации) применим к уравнениям с непрерывными и достаточное число раз дифференцируемыми коэффициентами и решениями. Из-за сравнительной громоздкости он применяется реже двух ранее описанных методов.

Краевые условия. Остановимся на записи разностной схемы в нерегулярных узлах (на границе или вблизи нее). В этих узлах для записи разностных уравнений необходимо привлекать краевые условия.

Например, в разностных схемах (16) и (18) для уравнения теплопроводности $u_t = ku_{xx}$ нерегулярными являются граничные узлы $n=0$, $n=N$. Для первой краевой задачи

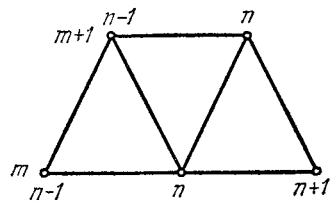


Рис. 53.

$$u(0,\ t) = \mu_1(t),\quad u(a,\ t) = \mu_2(t)$$

в этих узлах нетрудно написать разностные уравнения (17):

$$y_0 = \mu_1(t_m),\quad y_N = \mu_2(t_m),$$

которые являются точными (их невязка равна нулю).

Более сложен случай второй краевой задачи для того же уравнения (дальше будем рассматривать только левое условие):

$$u_x(0,\ t) = \mu_1(t),\quad u_x(a,t) = \mu_2(t). \tag{30}$$

Можно аппроксимировать производную односторонней разностью:

$$\frac{1}{h}(\hat{y}_1 - \hat{y}_0) = \mu_1(t_{m+1}). \tag{31}$$

Однако невязка этого разностного уравнения равна

$$\psi_0 = (\hat{u}_x)_0 - \frac{1}{h}(\hat{u}_1 - \hat{u}_0) = -\frac{h}{2}u_{xx} = O(h), \tag{32}$$

т. е. имеет меньший порядок малости, чем невязка (25) в регулярных узлах. Это приводит к понижению общей точности расчета.

Рассмотрим способы написания разностного краевого условия нормальной точности $O(h^2)$. Сделаем это на примере явной схемы (18).

Способ фиктивных точек очень нагляден. Введем вне отрезка $0 \leqslant x \leqslant a$ фиктивную точку $x_{-1} = x_0 - h$ и будем считать исходное уравнение справедливым при $x_{-1} \leqslant x$. Тогда разностное

уравнение (18) можно написать при $n=0$:

$$\frac{1}{\tau}(\hat{y}_0 - y_0) = \frac{k}{h^2}(y_{-1} - 2y_0 + y_1).$$

Заменим в левом краевом условии (30) производную симметричной разностью:

$$\frac{1}{2h}(y_1 - y_{-1}) = \mu_1(t_m).$$

Исключая из последних двух уравнений фиктивную точку, получим разностный аналог краевого условия:

$$\frac{1}{h}(y_1 - y_0) = \mu_1(t_m) + \frac{h}{2k\tau}(\hat{y}_0 - y_0). \tag{33}$$

Заметим, что это уравнение содержит только одно значение с нового слоя $\hat{y}_0$, т. е. оно явное.

Метод уменьшения невязки менее нагляден, но более универсален. Выразим $u(x_1, t)$ при помощи формулы Тейлора:

$$u(x_1, t) = u(x_0, t) + h\,u_x(x_0, t) + \frac{1}{2}h^2 u_{xx} + \ldots$$

На основании краевого условия (30) положим $u_x(x_0, t) = \mu_1(t)$, а из уравнения теплопроводности (15а) найдем $u_{xx} = u_t/k$. Подставляя эти величины в формулу Тейлора, получим

$$u(x_1, t) = u(x_0, t) + h\mu_1(t) + \frac{h^2}{2k}u_t + \ldots$$

Заменяя здесь $u_t \approx (\hat{y}_0 - y_0)/\tau$, снова придем к краевому условию (33).

В последнем способе можно учесть большее число членов ряда Тейлора и получить краевые условия не только нормальной, но и повышенной точности.

5. Аппроксимация и ее порядок. Пусть имеется область G переменных $x = \{x_1, \ldots, x_p\}$ с границей Γ и поставлена корректная задача для некоторого уравнения с граничными условиями:

$$Au(x) - f(x) = 0, \quad x \in G, \tag{34а}$$

$$Ru(x) - \mu(x) = 0, \quad x \in \Gamma. \tag{34б}$$

Введем в области $G + \Gamma$ сетку с шагом h, состоящую из множества внутренних (регулярных) узлов ω_h и множества граничных (нерегулярных) узлов γ_h. Заменим задачу (34) в регулярных узлах разностным аналогом уравнения (34а):

$$A_h y_h(x) - \varphi_h(x) = 0, \quad x \in \omega_h, \tag{35а}$$

а в нерегулярных узлах — разностным аналогом краевых условий (34б):

$$R_h y_h(x) - \chi_h(x) = 0, \quad x \in \gamma_h \tag{35б}$$

(индексом h отмечены величины, определенные только на сетке; мы будем опускать его там, где это не вызовет недоразумений).

Близость разностной схемы (35) к исходной задаче (34) будем определять по величине невязки:

$$\psi_h(x) = (Au - f) - (A_h u - \varphi_h), \quad x \in \omega_h,$$
$$\nu_h(x) = (Ru - \mu) - (R_h u - \chi_h), \quad x \in \gamma_h.$$

Определение. *Разностная схема* (35) *аппроксимирует задачу* (34), *если*

$$\|\psi\|_{\varphi_h} \to 0, \ \|\nu\|_{\chi_h} \to 0 \ \textit{при}\ h \to 0; \tag{36}$$

аппроксимация имеет p-й порядок, если

$$\|\psi\|_{\varphi_h} = O(h^p), \ \|\nu\|_{\chi_h} = O(h^p) \ \textit{при}\ h \to 0. \tag{37}$$

Обсудим вопрос о выборе норм в этом определении.

Функции $u(x)$, $f(x)$, $\mu(x)$ определены обычно на отрезке $a \leqslant x \leqslant b$ или во всех точках некоторой области пространства большего числа измерений. Для них можно ввести такие нормы, как чебышевская (локальная):

$$\|u(x)\|_C = \max_{a \leqslant x \leqslant b} |u(x)|, \tag{38а}$$

или гильбертова (среднеквадратичная):

$$\|u(x)\|_{L_2} = \left[\int_a^b \rho(x)\, u^2(x)\, dx\right]^{1/2}, \quad \rho(x) > 0 \tag{38б}$$

(выражения написаны для одномерного случая). Часто используют связанные с оператором A энергетические нормы, напоминающие формулы для полной энергии колебательной системы, например:

$$\|u(x)\| = \left\{\int_a^b [\rho_1(x)\, u_x^2(x) + \rho_0(x)\, u^2(x)]\, dx\right\}^{1/2}, \tag{38в}$$
$$\rho_1(x) > 0, \quad \rho_0(x) > 0.$$

Употребляются и другие нормы.

Напомним (см. главу I), что из локальной близости функций следует их среднеквадратичная близость; поэтому $\|\cdot\|_C$ называют более сильной, чем $\|\cdot\|_{L_2}$. Нетрудно проверить, что энергетическая норма (38в) сильней $\|\cdot\|_C$.

Выбор той или иной нормы в конкретной задаче определяется двумя соображениями. Желательно, чтобы разностное решение y

было близко к точному решению в возможно более сильной норме; например, в задачах на разрушение конструкций малость деформаций в $\|\cdot\|_{L_2}$ не гарантирует сохранения конструкции, а малость в $\|\cdot\|_C$ — гарантирует. С другой стороны, чем слабее $\|\cdot\|_u$, тем легче построить сходящуюся в этой норме разностную схему и исследовать ее.

Заметим, что функции $u(x)$, $f(x)$, $\mu(x)$ принадлежат, вообще говоря, разным классам. Например, если $u(x)$ есть четырежды дифференцируемая функция и $A=(d^2/dx^2)$, то $f(x)$ является дважды дифференцируемой функцией. Поэтому каждую из этих функций можно оценивать в своей норме: $\|\cdot\|_u$, $\|\cdot\|_f$, $\|\cdot\|_\mu$.

Функции y_h, φ_h, χ_h определены только на сетке, поэтому для них надо ввести сеточные нормы $\|\cdot\|_{y_h}$, $\|\cdot\|_{\varphi h}$, $\|\cdot\|_{\chi_h}$. Их вводят так, чтобы при $h\to 0$ они переходили в выбранные $\|\cdot\|_u$, $\|\cdot\|_f$, $\|\cdot\|_\mu$. За разностные аналоги чебышевской и гильбертовой норм можно принять соответственно

$$\|y\|_c=\max_{0\leqslant n\leqslant N}|y_n|,\quad \|y\|_{l_2}=\left(\sum_{n=1}^{N}\rho_n y_n^2 h_n\right)^{1/2}. \tag{39}$$

В выборе разностных аналогов норм существует некоторый произвол. Например, сумму в (39) можно брать по $n=0, 1, \ldots, N-1$, что соответствует выбору другой квадратурной формулы для интеграла (38б). Этим пользуются, определяя сеточные нормы так, чтобы облегчить доказательство сходимости.

Как надо понимать $h\to 0$? Для равномерной сетки это не требует пояснений. На неравномерных сетках рассматривают совокупность шагов h_n как некоторую сеточную функцию и вводят какую-либо норму шага, например:

$$h=\|h_n\|_c=\max_n h_n \qquad \text{или} \qquad h=\|h_n\|_{l_2}=\left(\sum_{n=0}^{N-1} h_n^3\right)^{1/2}.$$

Эту норму считают «величиной шага» в определениях аппроксимации, порядка аппроксимации и т. д.

Если невязку оценивают в $\|\cdot\|_c$, то аппроксимацию называют *локальной*. Для уравнений с достаточно гладкими решениями наличие локальной аппроксимации и ее порядок легко проверяются; в таких задачах нередко ограничиваются установлением локальной аппроксимации. Однако наиболее сильные результаты по сходимости разностных схем связаны с использованием более слабых норм для невязки (но сильных норм для решения).

Замечание 1. Факт наличия или отсутствия аппроксимации и порядок аппроксимации зависят не только от операторов A и A_h *), но также от классов, к которым принадлежат $u(x)$, $f(x)$, и от выбора норм. Чем сильнее норма или чем шире классы функций, тем ниже, вообще говоря, порядок аппроксимации

*) Операторы краевых условий для краткости обычно будем опускать.

(последнее видно по замечанию 2 к п. 3, если оценивать невязку в $\|\cdot\|_c$).

Замечание 2. Как правило, решение $u(x)$ исходной задачи (34) неизвестно, так что использовать его для получения невязки затруднительно. В этом случае берут достаточно широкий класс V функций $v(x)$, которому $u(x)$ заведомо принадлежит (обычно это класс функций, непрерывных вместе с достаточным числом своих производных). Если на всех функциях класса V имеется аппроксимация порядка p:

$$\|Av(x) - A_h v(x) + \varphi_h(x) - f(x)\|_{\varphi_h} = O(h^p) \quad \text{при} \quad h \to 0,$$

то, очевидно, аппроксимация на решении $u(x)$ имеет порядок не ниже p.

В подобных случаях аппроксимация на решении $u(x)$ может иметь порядок выше p. В замечании 3 к п. 3 мы видели, что для уравнения $u_t = ku_{xx}$ явная разностная схема (18) при $\tau = h^2/(6k)$ имеет в классе сколь угодно гладких функций $v(x)$ аппроксимацию $O(h^2)$, а на решении — более высокого порядка (четвертого, как нетрудно проверить).

Случай многих переменных имеет некоторые особенности. Определение аппроксимации остается в основном прежним; надо только требовать стремления к нулю шагов по всем переменным. Порядок аппроксимации может быть разный по разным переменным. Например, для двух переменных соотношение

$$\|\psi\|_{\varphi_h} = O(\tau^p + h^q) \quad \text{при} \quad \tau \to 0, \ h \to 0 \tag{40}$$

означает p-й порядок по времени и q-й по пространству. Это хорошо видно на примере схемы (18) с невязкой (25).

Аппроксимация вида (40), погрешность которой стремится к нулю при любом законе стремления шагов к нулю, называется *безусловной* или *абсолютной*. Если же погрешность аппроксимации стремится к нулю при одних законах убывания шагов и не стремится при других, то аппроксимацию называют *условной*. Например, если

$$\|\psi\|_{\varphi_h} = O\left(\tau^p + h^q + \frac{\tau^r}{h^s}\right) \quad \text{при} \quad \tau \to 0, \quad h \to 0, \tag{41}$$

то аппроксимация условная: кроме $\tau \to 0$, $h \to 0$, надо дополнительно требовать, чтобы $(\tau^r/h^s) \to 0$.

Если аппроксимация условная, то разностный оператор A_h при разных законах изменения $\tau(h)$ может аппроксимировать разные дифференциальные операторы. Например, можно проверить, что

$$A_h y = \frac{1}{\tau}(\hat{y}_n - y_n) + \frac{1}{2h}(y_{n+1} - y_{n-1}) - \frac{1}{\tau}(y_{n+1} - 2y_n + y_{n-1}) \tag{42}$$

при $\tau = ch$ аппроксимирует оператор

$$A = \frac{\partial}{\partial t} + \frac{\partial}{\partial x},$$

а при $\tau = ch^2$ — оператор

$$A = \frac{\partial}{\partial t} + \frac{\partial}{\partial x} - \frac{1}{c}\frac{\partial^2}{\partial x^2}.$$

Поэтому, если нет специальных соображений, лучше пользоваться разностными схемами с безусловной аппроксимацией.

§ 3. Устойчивость

1. Неустойчивость. Для некоторых разностных уравнений малые ошибки, допущенные на каком-либо этапе вычисления решения, при дальнейших выкладках сильно возрастают и делают невозможным получение сколько-нибудь пригодного результата.

При численном дифференцировании и суммировании рядов Фурье мы встречались с некорректными задачами, где бесконечно малая ошибка входных данных может привести к большой ошибке решения. Теперь рассмотрим пример неустойчивой разностной схемы решения задачи Коши для уравнения $u'(x) = \alpha u(x)$. Выберем следующую схему:

$$\frac{\sigma}{h}(y_{n+1} - y_n) + \frac{1-\sigma}{h}(y_n - y_{n-1}) = \alpha y_n. \tag{43}$$

При $\sigma = 1$ она переходит в схему ломаных (8.15), устойчивость которой была доказана. Рассмотрим случай $\sigma \neq 1$.

Для простоты отбросим ошибку аппроксимации и исследуем только рост ошибки начальных данных. Тогда ошибка δy_n будет удовлетворять тому же уравнению (43), ибо оно линейное однородное. Удобно исследовать рост ошибки специального вида $\delta y_n = z^n$. Подставляя ее в (43), получим

$$\sigma z^2 + (1 - 2\sigma - \alpha h) z - (1 - \sigma) = 0. \tag{44}$$

Если $h \ll 1$, то корни этого квадратного уравнения равны $z_1 = 1 + O(h)$, $z_2 = 1 - (1/\sigma) + O(h)$. Тогда при $\sigma < 1/2$ будет $|z_2| > 1$, т. е. ошибка такого вида возрастает за шаг в несколько раз. Значит, $\delta y(x) = z_2^n = z_2^{(x-x_0)/h}$ неограниченно возрастает при $h \to 0$, и счет неустойчив.

Если учесть еще ошибку аппроксимации, то получим типичные графики зависимости погрешности решения от шага, приведенные на рис. 54. Сплошная линия соответствует устойчивой схеме, штрихи — неустойчивой. При уменьшении шага ошибка сначала для всех схем убывает, потому что уменьшается погрешность аппроксимации. Для устойчивых схем при $h \to 0$ ошибка стремится к конечной величине, связанной с ошибкой начальных данных. Если сама ошибка начальных данных исчезает при $h \to 0$,

то мы получим график, изображенный пунктиром; т. е. устойчивые схемы позволяют получить сколь угодно высокую точность (если отсутствуют ошибки округления).

Если же схема неустойчива, то при малых шагах погрешность начальных или любых других данных сильно возрастает в ходе расчета и при $h \to 0$ ошибка стремится к бесконечности. Значит, график ошибки имеет ненулевой минимум, и мы в принципе не можем получить сколь угодно высокую точность приближенного решения.

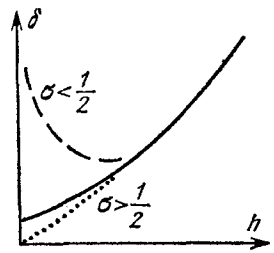


Рис. 54.

Правда, для неустойчивых схем есть некоторый оптимальный шаг, дающий наилучшую точность (это напоминает сходимость асимптотических рядов). Но эта наилучшая точность обычно настолько плоха, что считать по неустойчивым схемам практически невозможно.

Теоретически счет по неустойчивым схемам возможен, если начальные данные таковы, что погрешность их задания при $h \to 0$ убывает быстрее, чем нарастает неустойчивость. Но класс начальных данных, удовлетворяющих этому условию, обычно крайне узок и не охватывает даже малой части интересных случаев. Как правило, погрешности входных данных и аппроксимации убывают при $h \to 0$ по степенному закону, а величина $z_2^{(x-a)/h}$ возрастает много быстрей.

Отметим, что на устойчивость могут сильно влиять способы аппроксимации не только старших производных уравнения, но и младших производных и особенно краевых условий.

2. Основные понятия. *Разностная схема* (35)

$$A_h y = \varphi \quad (x \in \omega_h), \qquad R_h y = \chi \quad (x \in \gamma_h)$$

устойчива, если решение системы разностных уравнений непрерывно зависит от входных данных φ, χ *и эта зависимость равномерна относительно шага сетки.* Иными словами, для каждого $\varepsilon > 0$ найдется такое $\delta(\varepsilon)$, не зависящее от шага h (по крайней мере, для достаточно малых h), что

$$\|y^{\mathrm{I}} - y^{\mathrm{II}}\|_{y_h} \leqslant \varepsilon$$

если

$$\|\varphi^{\mathrm{I}} - \varphi^{\mathrm{II}}\|_{\varphi_h} \leqslant \delta, \qquad \|\chi^{\mathrm{I}} - \chi^{\mathrm{II}}\|_{\chi_h} \leqslant \delta. \tag{45}$$

Если разностная схема (35) линейна, то разностное решение линейно зависит от входных данных. В этом случае $\delta(\varepsilon) = K\varepsilon$, где K — константа, не зависящая от h. Поэтому для линейных схем определение устойчивости (45) принимает следующий вид:

$$\|y^{\mathrm{I}} - y^{\mathrm{II}}\|_{y_h} \leqslant M\|\varphi^{\mathrm{I}} - \varphi^{\mathrm{II}}\|_{\varphi_h} + M_1\|\chi^{\mathrm{I}} - \chi^{\mathrm{II}}\|_{\chi_h}, \tag{46}$$

где M, M_1 — константы, не зависящие от h. Напомним, что в (45) и (46) вариации решения и входных данных рассматриваются каждая в своей норме.

Дальше мы встретимся с примерами разностных схем, устойчивых при одном выборе норм и неустойчивых — при другом.

Если независимых переменных несколько, то вводят понятия условной и безусловной устойчивости. Устойчивость называется *безусловной*, если (45) или (46) выполняется при произвольном соотношении шагов по различным переменным, лишь бы они были достаточно малы. Если для выполнения (45) или (46) шаги по разным переменным должны удовлетворять дополнительным соотношениям, то устойчивость называется *условной*. Например, дальше будет доказано, что явная схема (18) для уравнения теплопроводности устойчива только при $\tau \leqslant h^2/2k$.

Непрерывную зависимость разностного решения от φ называют устойчивостью *по правой части*, а непрерывную зависимость от χ — устойчивостью *по граничным условиям*. Устойчивость по граничному условию на гиперплоскости $t = t_0$ называют устойчивостью *по начальным данным*.

Все простейшие типы уравнений, кроме эллиптического, в качестве одной из переменных содержат время. Для таких уравнений обычно ставится эволюционная задача — смешанная задача Коши. Даже эллиптические уравнения нередко численно решаются посредством счета на установление, т. е. при помощи постановки вспомогательной задачи Коши. Поэтому исследованию устойчивости эволюционных задач мы уделим особое внимание.

Рассмотрим разностные схемы, содержащие только один известный и один новый слой, как (16) или (18). Такие схемы называют *двуслойными*. Их можно составить для любого уравнения. В самом деле, дифференциальное уравнение любого порядка по времени можно свести к системе уравнений первого порядка по времени, а для аппроксимации первой производной по времени достаточно двух слоев.

Для двуслойных схем решение смешанной задачи Коши на некотором слое t^* можно рассматривать как начальные данные для всех последующих слоев.

Двуслойная разностная схема называется *равномерно устойчивой* по начальным данным, если при постановке начальных данных на любом слое t^* $(t_0 \leqslant t^* < T)$ она по ним устойчива, причем устойчивость равномерна по t^*. Запишем условие равномерной устойчивости, ограничиваясь случаем линейных схем:

$$\|y^{\mathrm{I}}(t) - y^{\mathrm{II}}(t)\| \leqslant K \|y^{\mathrm{I}}(t^*) - y^{\mathrm{II}}(t^*)\|, \quad t_0 \leqslant t^* < t < T, \quad (47)$$

где константа K не зависит от t^* и h; здесь y^{I}, y^{II} — решения разностной схемы $A_h y = \varphi$ с разными начальными данными и одной и той же правой частью.

Очевидно, из равномерной устойчивости по начальным данным следует обычная устойчивость по начальным данным (но не наоборот).

Признак равномерной устойчивости. *Если* $A_h y^{\mathrm{I}} = A_h y^{\mathrm{II}}$, *то для равномерной устойчивости по начальным данным достаточно, чтобы при всех* m *выполнялось*

$$\|\hat{y}^{\mathrm{I}} - \hat{y}^{\mathrm{II}}\| \leqslant (1 + C\tau)\,\|y^{\mathrm{I}} - y^{\mathrm{II}}\|, \qquad \tau = t_{m+1} - t_m, \quad C \geqslant 0. \tag{48}$$

Доказательство. Условие (48) означает, что если на некотором слое имеется ошибка δy, то при переходе на следующий слой $\|\delta y\|$ возрастает не более чем в $(1 + C\tau) \leqslant e^{C\tau}$ раз. Для перехода от t^* к t надо сделать $m = (t - t^*)/\tau$ шагов по времени; при этом ошибка возрастет не более чем в $e^{Cm\tau} = e^{C(t - t^*)} < e^{C(T - t_0)}$ раз. Отсюда следует:

$$\|\delta y(t)\| \leqslant K\|\delta y(t^*)\|, \qquad K = e^{C(T - t_0)}, \tag{49}$$

что и требовалось доказать.

Признак (48) мы будем часто использовать при доказательстве устойчивости конкретных схем.

Из (49) видно, что если константа C велика, то, хотя схема формально устойчива, фактическая ошибка может сильно возрастать в ходе расчета; в этом случае схема является *слабо устойчивой*. Очевидно, чем больше промежуток времени $T - t_0$, на котором ищется решение, тем меньшая величина C обеспечивает хорошую устойчивость расчета. При $T \to \infty$ схема будет устойчивой, только если $C = 0$.

Если точное решение задачи сильно возрастает или убывает с течением времени, то более интересна не абсолютная ошибка, а относительная $\|\delta y(t)\| / \|y(t)\|$. Можно классифицировать устойчивость по нарастанию относительной ошибки. Пусть, например, $u(x, t) \sim \exp(C_0 t)$. Тогда разностную схему, удовлетворяющую признаку (48), будем называть слабо устойчивой при $\exp[(C - C_0)(T - t_0)] \geqslant 1$, хорошо устойчивой — в обратном случае и *асимптотически* устойчивой при $T \to \infty$, если $C \leqslant C_0$.

Для многослойных схем определение и признаки равномерной устойчивости по начальным данным имеют более сложный вид; мы не будем их рассматривать.

Теорема. *Пусть двуслойная разностная схема* $A_h y = \varphi$ *равномерно устойчива по начальным данным и такова, что если два разностных решения* $A_h y^k = \varphi^k$ *равны на некотором слое,* $y^{\mathrm{I}} = y^{\mathrm{II}}$, *то на следующем слое выполняется соотношение*

$$\|\hat{y}^{\mathrm{I}} - \hat{y}^{\mathrm{II}}\| \leqslant \alpha\tau\|\varphi^{\mathrm{I}} - \varphi^{\mathrm{II}}\|, \qquad \alpha = \mathrm{const}. \tag{50}$$

Тогда разностная схема устойчива по правой части.

Доказательство. Наряду с решением y рассмотрим решение $\tilde{y}$, соответствующее возмущенной правой части $A_h \tilde{y} = \tilde{\varphi}$; поскольку исследуется устойчивость только по правой части, то можно считать, что $\tilde{y}(t_0) = y(t_0)$.

Введем последовательность сеточных функций $w_m(t)$, определенных при $t \geqslant t_{m-1}$ следующими условиями:

$$w_1(t_0) = y(t_0),$$
$$w_{m+1}(t_m) = w_m(t_m), \quad m = 1, 2, \ldots,$$
$$A_h w_m = \begin{cases} \tilde{\varphi} & \text{при } t_{m-1} \leqslant t < t_m, \\ \varphi & \text{при } t_m \leqslant t. \end{cases} \tag{51}$$

Эти функции определены так, что $w_m(t) = \tilde{y}(t)$ при $t_{m-1} \leqslant t \leqslant t_m$. Заметим, что в тех же обозначениях можно записать $w_0(t) \equiv y(t)$.

Сравним функции $w_m(t)$ и $w_{m+1}(t)$. На слое t_m они совпадают по определению. Тогда из (50) и (51) следует, что

$$\| w_{m+1}(t_{m+1}) - w_m(t_{m+1}) \| \leqslant \alpha\tau \| \varphi - \tilde{\varphi} \|.$$

При $t \geqslant t_{m+1}$ эти функции удовлетворяют разностной схеме с одной и той же правой частью φ, но с разными начальными данными на слое t_{m+1}. Поэтому в силу определения (47) на последнем слое t_M будут выполняться неравенства

$$\| w_{m+1}(t_M) - w_m(t_M) \| \leqslant K \| w_{m+1}(t_{m+1}) - w_m(t_{m+1}) \| \leqslant \alpha\tau K \| \varphi - \tilde{\varphi} \|.$$

Отсюда при помощи неравенства треугольника получим

$$\| y(t_M) - \tilde{y}(t_M) \| \leqslant \sum_{m=0}^{M-1} \| w_{m+1}(t_M) - w_m(t_M) \| \leqslant$$
$$\leqslant \alpha\tau M K \| \varphi - \tilde{\varphi} \| = \alpha (t_M - t_0) K \| \varphi - \tilde{\varphi} \|,$$

т. е. имеет место устойчивость по правой части, что и требовалось доказать.

Следствие. *Если неравенства* (48) *и* (50) *выполнены, то разностная схема устойчива и по начальным данным, и по правой части.*

Замечание. Теорема была доказана для конечного промежутка времени. В бесконечной по t области, если выполнено условие (50), можно доказать следующие достаточные признаки устойчивости по правой части:

а) Если при переходе со слоя на слой ошибка начальных данных не возрастает ($C \leqslant 0$), то схема устойчива по возмущениям правой части с конечным суммарным импульсом

$$\int\int | \delta\varphi(x, t) | \, dx \, dt < \varepsilon.$$

б) Если при переходе со слоя на слой ошибка начальных данных убывает как $(1 - C\tau)$, $C > 0$, то схема устойчива по отношению к постоянно действующим возмущениям $| \delta\varphi(x, t) | < \varepsilon$.

3. Принцип максимума. Есть несколько способов исследования устойчивости разностных схем: принцип максимума, метод

разделения переменных, метод операторных неравенств и некоторые другие. Сейчас мы рассмотрим принцип максимума, который применяют к уравнениям переноса, а также к параболическим и эллиптическим уравнениям. Он позволяет доказывать устойчивость в $\|\cdot\|_c$.

Сформулируем признак устойчивости явных и неявных двухслойных линейных разностных схем. Запишем двухслойную схему в следующем виде:

$$\sum_k \alpha_k \hat{y}_{n+k} = \sum_l \beta_l y_{n+l} + \varphi_n, \tag{52}$$

где суммирование на каждом слое производится по узлам шаблона около n-го узла. Коэффициенты α_k перенумеруем так, чтобы $|\alpha_0| = \max\limits_k |\alpha_k|$. Тогда:

а) *схема равномерно устойчива по начальным данным, если*

$$(1 + C\tau)\,|\alpha_0| \geqslant \sum_{k \neq 0} |\alpha_k| + \sum_l |\beta_l|, \qquad C = \text{const}. \tag{53}$$

б) *схема устойчива по правой части, если выполнено* (53) *и*

$$|\alpha_0| - \sum_{k \neq 0} |\alpha_k| \geqslant \frac{\varkappa}{\tau}, \qquad \varkappa = \text{const} > 0. \tag{54}$$

Доказательство. а) Фиксируем правую часть (52) и внесем ошибку δy на исходном слое. Тогда ошибка $\delta\hat{y}$ на новом слое удовлетворяет уравнению

$$\sum_k \alpha_k \delta\hat{y}_{n+k} = \sum_l \beta_l \delta y_{n+l}.$$

Отсюда для любого узла n следует неравенство

$$|\alpha_0| \cdot |\delta\hat{y}_n| \leqslant \sum_{k \neq 0} |\alpha_k| \cdot |\delta\hat{y}_{n+k}| + \sum_l |\beta_l| \cdot |\delta y_{n+l}|.$$

Применим это неравенство к узлу $\bar{n}$, в котором $|\delta\hat{y}_n|$ достигает своего максимума; при этом в правой части заменим $|\delta\hat{y}_{n+k}|$ и $|\delta y_{n+l}|$ их максимальными значениями, что только усилит неравенство. Тогда получим

$$|\alpha_0| \max_n |\delta\hat{y}_n| \leqslant \max_n |\delta\hat{y}_n| \sum_{k \neq 0} |\alpha_k| + \max_n |\delta y_n| \sum_l |\beta_l|,$$

или

$$\|\delta\hat{y}\|_c \left(|\alpha_0| - \sum_{k \neq 0} |\alpha_k| \right) \leqslant \|\delta y\|_c \sum_l |\beta_l|.$$

Но в силу неравенства (53)

$$\sum_l |\beta_l| \leqslant (1+C\tau)|\alpha_0| - \sum_{k\neq 0}|\alpha_k| \leqslant (1+C\tau)\left(|\alpha_0| - \sum_{k\neq 0}|\alpha_k|\right).$$

Поэтому

$$\|\delta\hat{y}\|_c \leqslant (1+C\tau)\|\delta y\|_c,$$

т. е. выполнен признак (48). Первое утверждение доказано.

б) Зафиксируем в (52) решение на исходном слое и внесем погрешность в правую часть. Тогда погрешность решения на новом слое удовлетворяет уравнению

$$\sum_k \alpha_k \delta\hat{y}_{n+k} = \delta\varphi_n.$$

Отсюда следует неравенство

$$|\alpha_0|\cdot|\delta\hat{y}_n| \leqslant \sum_{k\neq 0}|\alpha_k|\cdot|\delta\hat{y}_{n+k}| + |\delta\varphi_n|.$$

Аналогично предыдущему, выберем узел $\bar{n}$ и заменим справа все величины их максимумами. Легко получим, что

$$\|\delta\hat{y}\|_c\left(|\alpha_0| - \sum_{k\neq 0}|\alpha_k|\right) \leqslant \|\delta\varphi\|_c.$$

Отсюда с учетом (54) следует, что

$$\|\delta\hat{y}\|_c \leqslant \frac{\tau}{\varkappa}\|\delta\varphi\|_c,$$

т. е. выполнено условие (50). Второе утверждение доказано.

Замечание 1. Доказательство непосредственно применимо к схемам с переменными (зависящими от x, t) коэффициентами. Его можно обобщить на некоторые квазилинейные схемы, в которых коэффициенты зависят от y.

Замечание 2. Краевые условия двуслойных линейных схем также имеют форму (52). Поэтому данный признак позволяет устанавливать устойчивость по краевым условиям.

Замечание 3. Принцип максимума дает достаточное условие устойчивости; невыполнение критериев (53) и (54) еще не означает неустойчивости схемы.

Изложенным методом обычно удается доказать устойчивость только схем точности $O(\tau)$, да и то не всех; для обоснования устойчивости схем более высокого порядка точности по τ применяют другие методы.

Пример. Рассмотрим нестационарную краевую задачу для уравнения теплопроводности с постоянным коэффициентом (15):

$$u_t = ku_{xx} + f, \quad u(0, t) = \mu_1(t), \quad u(a, t) = \mu_2(t).$$

Запишем для нее неявную схему (16) — (17) на равномерной сетке:

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{k}{h^2}(\hat{y}_{n+1} - 2\hat{y}_n + \hat{y}_{n-1}) + \varphi_n, \quad 1 \leqslant n \leqslant N-1,$$
$$\hat{y}_0 = \mu_1(\hat{t}), \quad \hat{y}_N = \mu_2(\hat{t}).$$

Переписывая эту схему в форме (52), получим

$$\alpha_0 = \frac{1}{\tau} + \frac{2k}{h^2}, \quad \alpha_{-1} = \alpha_1 = \frac{k}{h^2}, \quad \beta_0 = \frac{1}{\tau} \quad \text{при} \quad 1 \leqslant n \leqslant N-1;$$
$$\alpha_0 = 1, \quad \beta_0 = 0 \quad \text{при} \quad n = 0 \text{ и } n = N;$$

остальные коэффициенты равны нулю. Видно, что при любых соотношениях шагов по t и x условие (54) выполнено в регулярных узлах, а условие (53) — во всех узлах сетки. Следовательно, схема безусловно устойчива по начальным данным, правой части и краевым условиям.

Для эллиптических уравнений обычно дается другая формулировка принципа максимума. Кроме того, для нестационарных задач имеется ряд модификаций принципа максимума: метод роста единичной ошибки, метод индекса разностной схемы и т. д. Мы их рассматривать не будем.

4. Метод разделения переменных. Этот метод применяется для строгого обоснования многих линейных схем и нестрогого, но плодотворного исследования большинства нелинейных задач, возникающих в практике вычислений. При его помощи устанавливается устойчивость в $\|\cdot\|_{l_2}$.

Рассмотрим применение метода к линейным двуслойным схемам, записанным в *канонической форме*:

$$B\frac{\hat{y}-y}{\tau} + Ay = \varphi, \tag{55}$$

где B, A — некоторые разностные операторы, действующие на y (или $\hat{y}$) как функцию пространственной переменной. Например, для явной схемы (18) имеем

$$B = E, \quad Ay_n = -\frac{k}{h^2}(y_{n+1} - 2y_n + y_{n-1}).$$

При фиксированной правой части погрешность решения удовлетворяет однородному уравнению

$$B\,\delta\hat{y} = (B - \tau A)\,\delta y. \tag{56}$$

Будем искать для этого уравнения частное решение с разделяющимися переменными

$$\delta y(x_n, t_m) = \rho_q^m e^{iqx_n}, \quad q = 0, \pm 1, \pm 2, \ldots \tag{57}$$

При этом $\delta\hat{y} = \rho_q\,\delta y$, так что ρ_q есть *множитель роста* q-й гармоники при переходе со слоя на слой. Подставляя (57) в (56),

получим уравнение для определения ρ_q:

$$\rho_q B e^{iqx} = (B - \tau A)\, e^{iqx}. \tag{58}$$

Будем считать, что схема (55) имеет постоянные коэффициенты и задана на равномерной сетке. Тогда уравнение (58) после сокращения множителя $\exp(iqx)$ не будет зависеть от координаты x (или ее индекса n). Следовательно, величина ρ_q не будет зависеть от x или t.

Признак устойчивости. *Схема* (55) *с постоянными коэффициентами устойчива по начальным данным, если для всех* q *выполняется неравенство*

$$|\rho_q| \leqslant 1 + C\tau, \quad C = \text{const}. \tag{59}$$

Доказательство. Система функций e^{iqx} $(0 \leqslant q \leqslant N-1)$ полна и ортогональна на равномерной сетке $\{x_n,\ 0 \leqslant n \leqslant N\}$. Разложим произвольную ошибку начальных данных $\delta y(x, t_0)$ в ряд Фурье по этой системе (см. гл. II, § 2, п. 4):

$$\delta y(x_n, t_0) = \sum_{q=0}^{N-1} a_q e^{iqx_n}.$$

Поскольку для линейного уравнения (55) справедлив принцип суперпозиции, то метод разделения переменных дает для ошибки на слое t_m следующее выражение:

$$\delta y(x_n, t_m) = \sum_{q=0}^{N-1} a_q \rho_q^m e^{iqx_n}.$$

Используя ортогональность гармоник, получаем отсюда

$$\| \delta y(t_m) \|_{l_2}^2 = N \sum_{q=0}^{N-1} |\rho_q|^{2m} |a_q|^2 \leqslant$$

$$\leqslant \max_q |\rho_q|^{2m} N \sum_{q=0}^{N-1} |a_q|^2 = \max_q |\rho_q|^{2m} \| \delta y(t_0) \|_{l_2}^2.$$

При помощи условия (59) преобразуем это неравенство к виду

$$\| \delta y(t_m) \|_{l_2} \leqslant (1 + C\tau)^m \| \delta y(t_0) \|_{l_2},$$

что совпадает с признаком (48). Утверждение доказано.

Замечание 1. Из признака устойчивости (59) и дополнительного условия (54) следует устойчивость схемы по правой части в $\| \cdot \|_{l_2}$.

Замечание 2. Фактически константа C в (59) не должна быть большой, иначе устойчивость будет слабой (см. п. 2). Поэтому при проверке этого признака обычно полагают $C = 0$.

Признак неустойчивости. *Если хотя бы для одного q величину $|\rho_q|$ нельзя мажорировать величиной $1+C\tau$, то схема* (55) *неустойчива.*

Доказательство. Пусть в начальных данных имеется ошибка вида εe^{iqx} с данным q. Тогда к моменту $t=m\tau$ она возрастет в ρ_q^m раз, что по модулю больше величины $(1+C\tau)^m = (1+C\tau)^{t/\tau} > Ct$ при сколь угодно большом C. Неограниченный рост ошибки означает неустойчивость схемы.

Пример. Исследуем устойчивость явной схемы (18) для уравнения теплопроводности. Для этой схемы уравнение (58) принимает вид

$$(\rho_q - 1)\, e^{iqx_n} = \frac{k\tau}{h^2}\left(e^{iq(x_n+h)} - 2e^{iqx_n} + e^{iq(x_n-h)}\right).$$

Отсюда вытекает, что множитель роста

$$\rho_q = 1 - \frac{4k\tau}{h^2}\sin^2\frac{qh}{2}.$$

Тогда условие (59) с учетом замечания 2 приобретает вид $-1 \leqslant \rho_q \leqslant 1$. Это неравенство выполняется для любого q, только если

$$2k\tau/h^2 \leqslant 1, \quad \text{или} \quad \tau \leqslant \frac{h^2}{2k}. \tag{60}$$

Таким образом, явная схема (18) условно устойчива.

Метод разделения переменных применим к многослойным линейным схемам с постоянными коэффициентами, в частности к схемам, аппроксимирующим задачи для дифференциальных уравнений второго порядка по времени (соответствующие примеры будут рассмотрены в главе XIII). Сейчас остановимся на двух нестрогих обобщениях этого метода.

Замораживание коэффициентов. Если линейное дифференциальное уравнение имеет переменные коэффициенты или используется неравномерная сетка, то задача сводится к линейной разностной схеме с переменными коэффициентами. В этом случае уравнение (58) содержит неустранимую зависимость от n. Следовательно, множитель роста ρ_q также зависит от n и его нельзя считать постоянным для данной гармоники.

«Заморозим» коэффициенты схемы, т. е. возьмем в качестве постоянных коэффициентов значения коэффициентов схемы в некотором узле n, и найдем ρ_q из (58). Будем считать разностную схему устойчивой, если при любых q и n выполняется признак (59).

Этот способ оказался очень эффективным приемом исследования устойчивости схем. В настоящее время он обоснован для многих классов параболических и эллиптических уравнений с гладкими коэффициентами (в ряде случаев достаточно непре-

рывности коэффициентов) и для некоторых узких классов гиперболических уравнений. В практике вычислений для любых уравнений с гладкими коэффициентами и решениями критерии устойчивости, полученные этим способом, хорошо согласуются с результатами численных расчетов.

Однако способ «замороженных» коэффициентов применим не всегда. Для ряда задач с разрывными, недифференцируемыми и даже кусочно-гладкими коэффициентами построены примеры *), в которых использование этого способа приводит к ошибочным заключениям.

Линеаризация. Сложные задачи математической физики приводят к нелинейным разностным схемам

$$B_1(\hat{y}) + B_2(y) = \varphi, \tag{61}$$

где B_1, B_2 — нелинейные операторы, действующие на $\hat{y}$ и y как на функции пространственной переменной. Нарастание ошибок (пока эти ошибки малы) описывается линеаризованным уравнением

$$\frac{\delta B_1(\hat{y})}{\delta \hat{y}}\delta\hat{y} + \frac{\delta B_2(y)}{\delta y}\delta y = \delta\varphi. \tag{62}$$

Обычно B и y являются N-мерными векторами; тогда $\delta B/\delta y$ является матрицей производных $(\partial B_l/\partial y_k)$. Устойчивость уравнения (62), линейного относительно ошибок, можно исследовать способом «замороженных» коэффициентов. Уравнение для множителя роста q-й гармоники принимает вид

$$\left(\rho_q \frac{\delta B_1}{\delta \hat{y}} + \frac{\delta B_2}{\delta y}\right) e^{iqx} = 0. \tag{63}$$

Способ линеаризации при исследовании многих сложных разностных схем (например, схем, возникающих в задачах газодинамики) дает критерии устойчивости, хорошо подтверждаемые практикой численных расчетов. Однако он не является строго обоснованным и в некоторых случаях может привести к неверным результатам.

Метод разделения переменных можно строго обобщить на многие классы линейных схем с переменными коэффициентами (на неравномерных сетках), а также применять его для доказательства устойчивости по краевым условиям. Для этого надо вместо гармоник $\exp(iqx)$ использовать систему $\gamma_q(x)$ собственных функций разностной задачи

$$\rho_q B\gamma_q(x) = (B - \tau A)\gamma_q(x)$$

и соответствующие собственные значения ρ_q. Однако точно найти спектр разностной схемы удается лишь в сравнительно простых случаях, так что исследовать этим методом устойчивость схем для сложных задач математической физики удается не часто.

*) См., например, [33], стр. 383.

5. Метод энергетических неравенств. Метод основан на использовании энергетических норм, порождаемых самими разностными операторами. При его помощи доказана устойчивость и даны априорные оценки точности многих разностных схем с переменными коэффициентами, некоторых квазилинейных схем и т. д.

Рассмотрим идею метода *) на примере стационарной (не содержащей времени) разностной схемы

$$Ay = \varphi,$$

где разностный оператор A — линейный, самосопряженный и положительный. В этом случае существует обратный оператор A^{-1}, который тоже является линейным, самосопряженным и положительным.

При помощи положительного оператора можно ввести норму

$$\| y \|_A^2 = (Ay, y) > 0 \quad \text{при} \quad y \neq 0, \tag{64}$$

где $(,)$ — скалярное произведение на сетке; аналогично строится норма $\| \cdot \|_{A^{-1}}$, называемая *негативной*. Проделаем цепочку преобразований:

$$\| y \|_A^2 = (Ay, y) = (\varphi, A^{-1}\varphi) = (A^{-1}\varphi, \varphi) = \| \varphi \|_{A^{-1}}^2.$$

Отсюда вытекает соотношение $\| y \|_A = \| \varphi \|_{A^{-1}}$, которое означает устойчивость по правой части.

Пусть, например, оператор $A = A_h$ является второй разностью, т. е. аналогом $-d^2/dx^2$, а $u(x)$ и ее производные достаточно быстро убывают при $|x| \to \infty$. Тогда непрерывный аналог нормы (64) есть (сеточное выражение не так наглядно, и мы его не приводим)

$$\| u \|_A^2 = \int_{-\infty}^{+\infty} \left(-\frac{d^2u}{dx^2} \right) u(x)\, dx = \int_{-\infty}^{+\infty} \left(\frac{du}{dx} \right)^2 dx;$$

как отмечалось в п. 2, эта норма сильней, чем $\| \cdot \|_C$. Оператор A^{-1} в этом случае является двойной суммой — аналогом двойного интеграла, и порожденная им норма равна

$$\| \varphi \|_{A^{-1}}^2 = \int_{-\infty}^{+\infty} \varphi(x)\, dx \int_{-\infty}^{x} d\xi \int_{-\infty}^{\xi} \varphi(\eta)\, d\eta = \int_{-\infty}^{+\infty} dx \left(\int_{-\infty}^{x} \varphi(\xi)\, d\xi \right)^2.$$

Это слабая норма. Из приведенных рассуждений видно, что метод энергетических неравенств для ряда задач позволяет доказывать устойчивость при использовании сильных норм для решения y и слабых норм для правой части φ **).

*) Подробное изложение метода см. в [30, 33].

**) В подобных случаях нередко удается доказать сходимость схем с более высоким порядком точности, чем при использовании других методов.

Для конкретной реализации этого метода надо проверить, обладает ли оператор A требуемыми свойствами, определить скалярное произведение на сетке, построить сеточный оператор A^{-1} и проверить аппроксимацию в $\|\cdot\|_{A^{-1}}$. Все эти действия связаны обычно с громоздкими вычислениями.

6. Операторные неравенства. Общая теория устойчивости разностных схем, основанная на установлении неравенств между разностными операторами, образующими схему, построена А. А. Самарским (см. [30, 33]). Она позволяет для многих классов линейных схем получить необходимые и достаточные условия устойчивости и априорные оценки точности. Рассмотрим одно из таких условий устойчивости.

Напомним некоторые свойства операторов, отображающих гильбертово пространство H в себя *). Оператор A называется *неотрицательным* ($A \geqslant 0$), если $(Ax, x) \geqslant 0$ для любого ненулевого $x \in H$, называется *положительным* ($A > 0$) при $(Ax, x) > 0$ и *положительно определенным* при $(Ax, x) \geqslant \varepsilon (x, x)$, $\varepsilon > 0$. Неравенство $A \geqslant B$ понимается в том смысле, что $A - B \geqslant 0$.

Оператор A называют самосопряженным, если $(Ax, y) = (x, Ay)$ для любых $x, y \in H$. Квадратным корнем из самосопряженного неотрицательного оператора A называют такой оператор B, что $B \cdot B = A$; его обозначают $A^{1/2}$, он существует и является самосопряженным и неотрицательным.

Исследуем устойчивость двуслойной линейной разностной схемы, записанной в канонической форме:

$$B \frac{\hat{y} - y}{\tau} + Ay = \varphi. \tag{65}$$

Теорема. *Если операторы A и B самосопряженные, не зависят от номера слоя n, и выполняется условие*

$$B \geqslant \frac{\tau}{2} A > 0, \tag{66}$$

то схема (65) *устойчива по начальным данным в энергетической норме* $\|\cdot\|_A$:

$$\|\hat{y}\|_A \leqslant \|y\|_A. \tag{67}$$

Доказательство. Для исследования устойчивости по начальным данным достаточно рассмотреть однородное уравнение (65). Полагая $\varphi = 0$ и умножая (65) слева на $A^{1/2}B^{-1}$, получим

$$A^{1/2} \frac{\hat{y} - y}{\tau} + A^{1/2} B^{-1} A y = 0.$$

*) Более подробно о свойствах операторов см., например, в [20].

Полагая $\eta = A^{1/2}y$ и замечая, что $Ay = A^{1/2}\eta$, преобразуем это уравнение в явную разностную схему:

$$\hat{\eta} = S\eta, \qquad S = E - \tau A^{1/2}B^{-1}A^{1/2},$$

где E — единичный оператор; оператор S является самосопряженным.

Перепишем неравенство (66) в следующем виде:

$$0 < B^{-1} \leqslant \frac{2}{\tau} A^{-1}.$$

Умножая его слева и справа на положительный оператор $A^{1/2}$, получим

$$0 < \tau A^{1/2}B^{-1}A^{1/2} \leqslant 2E.$$

Вычитая это неравенство из E, получим

$$-E \leqslant E - \tau A^{1/2}B^{-1}A^{1/2} \equiv S < E.$$

Это означает, что

$$\| \hat{\eta} \|_{l_2}^2 = (\hat{\eta}, \hat{\eta}) = (S\eta, S\eta) \leqslant (\eta, \eta) = \| \eta \|_{l_2}^2. \tag{68}$$

Норма $\| \cdot \|_{l_2}$ просто связана с энергетической нормой:

$$\| \eta \|_{l_2}^2 = (\eta, \eta) = (A^{1/2}y, A^{1/2}y) = (Ay, y) = \| y \|_A^2. \tag{69}$$

Из (68) и (69) следует (67), что и требовалось доказать.

Замечание. При доказательстве не требовалось постоянства коэффициентов схемы (65). Тем самым, признак устойчивости (66) справедлив для разностных схем с переменными коэффициентами.

В этом параграфе излагалась техника исследования устойчивости уже составленной схемы. А как надо составлять схему, чтобы она была устойчивой? Некоторые математические способы построения устойчивых схем были предложены А. А. Самарским в [30]. Высказывались идеи о рассмотрении разностных схем как некорректных задач с дискретными переменными и регуляризации их по А. Н. Тихонову.

Для ряда конкретных задач на основании физических аналогий (скорости распространения возмущений) можно предсказать, будет ли схема устойчива и как ее надо составить, чтобы она была устойчива. В следующих главах будет приведено много таких примеров.

§ 4. Сходимость

1. Основная теорема. В этом параграфе мы рассмотрим задачу для дифференциального уравнения с граничными условиями

$$Au(x) = f(x), \quad x \in G, \qquad Ru(x) = \mu(x), \quad x \in \Gamma, \tag{70}$$

которая на сетке, состоящей из множества регулярных узлов ω_h и множества нерегулярных узлов γ_h, аппроксимирована разностной схемой

$$A_h y(x) = \varphi(x), \quad x \in \omega_h, \qquad R_h y(x) = \chi(x), \quad x \in \gamma_h. \tag{71}$$

В конечном итоге нас будет интересовать близость разностного решения $y(x)$ к точному решению $u(x)$; поскольку $y(x)$ определено только на сетке $\omega_h + \gamma_h$, то сравнивать эти решения надо в сеточной норме.

Определение. *Разностное решение $y(x)$ сходится к решению $u(x)$ задачи* (70), *если*

$$\| y(x) - u(x) \|_{y_h} \to 0 \text{ при } h \to 0; \tag{72}$$

разностное решение имеет порядок точности p, если

$$\| y(x) - u(x) \|_{y_h} = O(h^p) \text{ при } h \to 0. \tag{73}$$

Анализируя сходимость схемы ломаных (8.15) для обыкновенного дифференциального уравнения, мы видели, что погрешность решения вызвана погрешностью начальных данных и погрешностью аппроксимации, усиливающимися (или ослабляющимися) в ходе расчета. Интуитивно ясно, что для хорошей точности расчета достаточно, чтобы эти погрешности были малы и в ходе расчета не сильно возрастали.

Строго говоря, в любых расчетах присутствуют ошибки округления; поэтому при $h \to 0$ надо одновременно увеличивать количество десятичных знаков, удерживаемое в расчете. Но в современных ЭВМ относительная ошибка округления на одну операцию не превышает 10^{-10}, т. е. пренебрежимо мала по сравнению с ошибками аппроксимации при тех шагах h, которые фактически используются. Поэтому в большинстве случаев ошибками округления можно пренебречь.

Определение. *Разностная схема* (71) *корректна, если ее решение существует и единственно при любых входных данных φ и χ, принадлежащих заданным классам функций, и схема устойчива.*

Строго говоря, для нелинейных схем разностное решение может быть не единственным или существовать не при всяких входных данных. В этом случае схему называют корректной в окрестности решения $y[\varphi, \chi]$, если (по крайней мере при достаточно малом h) для любых $\tilde{\varphi}$, $\tilde{\chi}$, достаточно близких к φ, χ, в некоторой малой окрестности $y[\varphi, \chi]$ имеется единственное решение $\tilde{y}[\tilde{\varphi}, \tilde{\chi}]$, устойчивое по $\tilde{\varphi}$, $\tilde{\chi}$ в смысле определения (45).

Отметим, что если граница области G состоит из нескольких кусков Γ_k, то обычно операторы R_k и правые части $\mu_k(x)$ граничных условий на этих кусках различны. Разностные операторы

R_{hk} и правые части χ_k на соответствующих множествах нерегулярных узлов γ_{hk} также будут различны. Для того чтобы решение разностной схемы (71) существовало, все они должны быть согласованы между собой, т. е. должны удовлетворять определенным соотношениям на линиях или в точках стыка кусков границы. Например, для первой краевой задачи теплопроводности

$$u_t = ku_{xx},\ 0 < x < a,\ t > 0,$$
$$u(x,\ 0) = \mu(x),\ 0 \leqslant x \leqslant a,$$
$$u(0,\ t) = \mu_1(t),\ u(a,\ t) = \mu_2(t),\ t \geqslant 0,$$

условиями согласования будут соотношения $\mu(0) = \mu_1(0)$, $\mu(a) = \mu_2(0)$, или, соответственно, $\chi(0) = \chi_1(0)$, $\chi(a) = \chi_2(0)$.

Теорема*). *Если решение* $u[f,\ \mu]$ *задачи* (70) *существует, разностная схема* (71) *корректна и аппроксимирует задачу* (70) *на данном решении, то разностное решение сходится к точному.*

Доказательство. Напишем цепочку преобразований:

$$A_h u = A_h u - Au + f = A_h u - Au + f - \varphi + \varphi = -\psi(x) + \varphi(x),$$

где $\psi(x)$ есть, по определению, невязка разностной схемы. Делая аналогичное преобразование для краевых условий, получим

$$\begin{aligned} A_h u(x) &= \varphi(x) - \psi(x), \qquad x \in \omega_h, \\ R_h u(x) &= \chi(x) - \nu(x), \qquad x \in \gamma_h. \end{aligned} \tag{74}$$

Равенства (74) представляют собой разностную схему (71) с правыми частями, измененными на величину невязки. Поскольку разностная схема устойчива, то для любого $\varepsilon > 0$ найдется такое $\delta(\varepsilon)$, что $\| y - u \|_{y_h} \leqslant \varepsilon$, если $\| \psi \|_{\varphi_h} \leqslant \delta(\varepsilon)$, $\| \nu \|_{\chi_h} \leqslant \delta(\varepsilon)$.

В силу аппроксимации для любого $\delta > 0$ найдется такое $h_0(\delta)$, что $\| \psi \|_{\varphi_h} \leqslant \delta$, $\| \nu \|_{\chi_h} \leqslant \delta$ при $h \leqslant h_0(\delta)$.

Следовательно, для любого $\varepsilon > 0$ найдется такое $h_0(\delta(\varepsilon))$, что $\| y - u \|_{y_h} \leqslant \varepsilon$ при $h \leqslant h_0$. Сходимость доказана.

Замечание 1. Некоторые начальные или граничные условия аппроксимируются точно; примером являются граничные условия первого рода $u(a,\ t) = \mu(t)$, если узел x_N сетки расположен на границе $x = a$. Устойчивости по таким условиям можно не требовать, ибо никакой ошибки в расчет они не вносят (кроме ошибки округления).

Устойчивость по правой части требуется почти во всех случаях, поскольку погрешность аппроксимации в (74) эквивалентна некоторой погрешности правой части.

---

*) Ее кратко формулируют так: «Из аппроксимации и устойчивости следует сходимость».

Замечание 2. Аппроксимацию часто проверяют не на решениях задачи (70), а на некотором широком классе функций, которому принадлежит решение (обычно на классе функций, непрерывных и ограниченных вместе с некоторым числом своих производных). Из замечания 2 в § 2, п. 5 следует, что такая аппроксимация достаточна для доказательства теоремы о сходимости.

Замечание 3. При исследовании аппроксимации и устойчивости конкретных разностных схем нередко используют разные нормы для одной и той же функции. Например, при установлении локальной аппроксимации для $\varphi(x)$ берется $\|\varphi\|_c$, а при спектральном исследовании устойчивости — $\|\varphi\|_{l_2}$. Доказательство сходимости в этом случае справедливо, только если аппроксимация установлена в нормах $\|\varphi\|$, $\|\chi\|$ более сильных (или тех же самых), чем нормы, использованные для правых частей в определении устойчивости.

Замечание 4. Если аппроксимация или устойчивость условные, то сходимость имеет место при выполнении условий устойчивости и аппроксимации (т. е. при определенных соотношениях между шагами по разным переменным).

Замечание 5. Устойчивость является, как нетрудно убедиться, необходимым условием сходимости. В самом деле, если схема неустойчива, то найдутся такие сколь угодно малые ошибки входных данных, которым соответствует значительная погрешность решения. Сходимости при этом не может быть.

Пример. Рассмотрим явную схему (18) для уравнения теплопроводности (15). В § 2, п. 3 была установлена аппроксимация этой схемы с погрешностью (25), равной $\|\psi\|_c = O(\tau + h^2)$. В § 3, п. 4 было доказано, что она условно устойчива в $\|\cdot\|_{l_2}$ при $\tau \leqslant h^2/(2k)$. С учетом замечания 3 отсюда следует сходимость в норме $\|y-u\|_{l_2}$, если выполнено условие $\tau \leqslant h^2/(2k)$.

Отметим, что на самом деле имеет место сходимость в $\|y-u\|_c$; но для доказательства этого факта надо обосновать устойчивость схемы в нормах $\|\varphi\|_c$, $\|y\|_c$.

2. Оценки точности. Для линейных задач оценки погрешности, как априорные мажорантные, так и апостериорные асимптотические, можно получить на основании приведенных ниже теорем.

Теорема 1 *). *Если условия теоремы из* п. 1 *выполнены, операторы* A_h *и* R_{hk} *линейные, а порядок аппроксимации равен* p, *то сходимость имеет порядок не ниже* p.

Доказательство. Пусть задача (70) и разностная схема (71) линейны, а граница Γ состоит из кусков Γ_k $(k=1, 2, \ldots, K)$.

---

*) Ее кратко формулируют так: «Для линейных схем порядок точности не ниже порядка аппроксимации».

Условие устойчивости (46) для линейной схемы принимает вид

$$\| y \|_y \leqslant M_0 \| \varphi \|_\varphi + \sum_{k=1}^{K} M_k \| \chi_k \|_{\chi_k} \tag{75}$$

(начальные условия, если задача их содержит, входят в сумму по граничным условиям). Рассмотрим погрешность разностного решения $z(x) = y(x) - u(x)$. Вычтем соотношение (74)

$$A_h u = \varphi - \psi, \quad R_{hk} u = \chi_k - \nu_k$$

из разностной схемы (71) и заметим, что благодаря линейности схемы $A_h y - A_h u = A_h (y - u) = A_h z$. Тогда $z(x)$ удовлетворяет схеме с разностными операторами (71):

$$A_h z(x) = \psi(x) \quad x \in \omega_h, \qquad R_{hk} z(x) = \nu_k(x), \quad x \in \gamma_{hk}, \tag{76}$$

где в правых частях стоят невязки. Применяя к (76) условие устойчивости (75), получим

$$\| z \|_y \leqslant M_0 \| \psi \|_\varphi + \sum_{k=1}^{K} M_k \| \nu_k \|_{\chi_k}. \tag{77}$$

Поскольку схема (71) имеет порядок аппроксимации p, то

$$\|\psi\|_\varphi \leqslant \alpha_0 h^p, \qquad \|\nu_k\|_{\chi_k} \leqslant \alpha_k h^p, \quad 1 \leqslant k \leqslant K. \tag{78}$$

Подставляя эти выражения в (77), получим *априорную мажорантную* оценку погрешности:

$$\|y - u\|_y \leqslant M h^p, \qquad M = \sum_{k=0}^{K} M_k \alpha_k, \tag{79}$$

что доказывает теорему.

Замечание 1. Для доказательства требовалась линейность только разностных операторов, но фактически линейными разностными операторами можно аппроксимировать только линейные дифференциальные или интегральные операторы.

Замечание 2. Если условия теоремы 1 выполнены, то порядок точности может быть выше порядка аппроксимации. В таких случаях более полное исследование задачи нередко показывает, что для сходимости в данной норме $\| \cdot \|_y$ достаточно устойчивости по более слабой норме $\| \cdot \|_\varphi$, в которой порядок аппроксимации выше.

Замечание 3. При оценках погрешности конкретных схем константы M_k определяются в ходе доказательства устойчивости; они постоянны для данной схемы. Величины α_k выражаются обычно через нормы некоторых производных $u(x)$ и тем самым зависят от решения (см. выражения невязки (25) или (26)).

Замечание 4. Для нелинейных схем можно сформулировать аналогичную теорему. При этом следует пользоваться опре-

делением устойчивости (45), которое можно записать так:

$$\|y-\tilde{y}\|\leqslant\varepsilon,\ \text{если}\ \|\varphi-\tilde{\varphi}\|\leqslant\delta_0(\varepsilon),\ \|\chi_k-\tilde{\chi}_k\|\leqslant\delta_k(\varepsilon),\quad 1\leqslant k\leqslant K. \tag{80}$$

Тогда, если $\delta_k(\varepsilon)=(\varepsilon/M_k)^{m_k}$, $0\leqslant k\leqslant K$, то порядок точности будет не ниже $q=\min\limits_{0\leqslant k\leqslant K}(p/m_k)$; при $m_k\equiv 1$ снова приходим к теореме 1.

Замечание 5. Для случая многих переменных порядок аппроксимации по разным переменным может быть неодинаковым. Очевидно, порядок точности по разным переменным также может быть различным.

Пример. Явная схема (17)—(18) для первой краевой задачи теплопроводности (15), разобранная в примере к п. 1, имеет погрешность аппроксимации (25):

$$\|\psi\|_c\leqslant\frac{1}{12}kh^2\|u_{xxxx}\|_c+\frac{1}{2}\tau\|u_{tt}\|_c.$$

Начальные данные и краевые условия аппроксимируются точно, и устойчивости по ним можно не требовать; согласно замечанию 1 в § 3, п. 4 условие устойчивости по правой части имеет вид

$$\|\delta\hat{y}\|_{l_2}\leqslant\tau\|\delta\varphi\|_{l_2}\quad\text{или}\quad\|\delta y(t)\|_{l_2}\leqslant(t-t_0)\max_t\|\delta\varphi\|_{l_2}.$$

Отсюда следует априорная оценка

$$\|y-u\|_{l_2}\leqslant(t-t_0)\max_t\left(\frac{1}{2}\tau\|u_{tt}\|_{l_2}+\frac{1}{12}kh^2\|u_{xxxx}\|_{l_2}\right)=O(\tau+h^2), \tag{81}$$

т. е. схема имеет первый порядок точности по времени и второй — по пространству.

Для практических вычислений важное значение имеет следующая

Теорема 2. *Пусть задача* (70) *и разностная схема* (71) *линейны, разностная схема корректна и аппроксимирует задачу так, что существуют*

$$\bar{\psi}(x)=\lim_{h\to 0}h^{-p}(Au-A_hu+\varphi-f),\quad x\in G, \tag{82а}$$

$$\bar{\nu}_k(x)=\lim_{h\to 0}h^{-p}(R_ku-R_{hk}u+\chi_k-\mu_k),\quad x\in\Gamma_k, \tag{82б}$$

Пусть существует решение $\bar{z}(x)$ *задачи*

$$A\bar{z}(x)=\bar{\psi}(x),\quad x\in G,\quad R_k\bar{z}(x)=\bar{\nu}_k(x),\quad x\in\Gamma_k, \tag{83}$$

и на этом решении разностные операторы A_h, R_{hk} *аппроксимируют дифференциальные операторы* A, R_k. *Тогда погрешность*

решения (71) *имеет следующую асимптотику:*

$$y(x)-u(x)=h^p\bar{z}(x)+o(h^p) \text{ при } h\to 0,\ x\in\omega_h+\gamma_h. \qquad (84)$$

Доказательство. Пользуясь линейностью операторов, нетрудно установить следующее равенство:

$$A_h[h^{-p}(y-u)-\bar{z}]=h^{-p}(Au-A_hu+\varphi-f)-\bar{\psi}+(A\bar{z}-A_h\bar{z});$$

аналогичные равенства записываются для граничных условий. При $h\to 0$ правые части всех этих равенств стремятся по норме к нулю: последняя скобка — на основании предположения об аппроксимации на функции $\bar{z}(x)$, а остальные члены — согласно условию (82).

Тогда, благодаря устойчивости разностных операторов A_h, R_{hk}, выражение в квадратных скобках в левой части этих равенств стремится по норме к решению задачи (71) с нулевой правой частью, которое тождественно равно нулю. Теорема доказана.

Замечание 1. Теорему можно обобщить на случай многих переменных, даже если порядок аппроксимации по разным переменным неодинаковый. В случае двух переменных возможна следующая асимптотика погрешности:

$$y(x,\ t)-u(x,\ t)=\tau^q\bar{z}_1(x,\ t)+h^p\bar{z}_2(x,\ t)+o(\tau^q+h^p), \qquad (85)$$

или иная, в зависимости от характера аппроксимации.

Теорема 2 обосновывает использование метода Рунге для апостериорной асимптотической оценки погрешности или для уточнения результата.

Например, явная схема (18) для уравнения $u_t=ku_{xx}$ имеет невязку (26), равную $\psi(x,\ t)=\left(\frac{1}{12}kh^2-\frac{1}{2}k^2\tau\right)u_{xxxx}(x,\ t)$. Поскольку решение этого уравнения дифференцируемо любое число раз, то легко проверить выполнение условий теоремы 2 и определить погрешность:

$$y(x,\ t)-u(x,\ t)=\left(\frac{kh^2}{12}-\frac{k^2\tau}{2}\right)\bar{z}(x,\ t)+o(\tau+h^2), \qquad (86)$$

где $\bar{z}$ удовлетворяет уравнению $\bar{z}_t-k\bar{z}_{xx}=u_{xxxx}$ и соответствующим начальным и граничным условиям.

Возьмем сетку ω^{I} с шагами h и τ и сгущенную сетку ω^{II} с шагами h/r и τ/r^2 (обычно полагают $r=2$). На второй сетке погрешность по каждой переменной, как видно из (86), уменьшается в r^2 раз. Обозначая разностные решения на этих сетках соответственно через y^{I} и y^{II}, определим погрешность (см. гл. III, п. 3 и гл. VIII, § 1, п. 11):

$$z^{\mathrm{II}}=y^{\mathrm{II}}-u\approx\frac{1}{r^2-1}(y^{\mathrm{I}}-y^{\mathrm{II}}).$$

Эту погрешность можно использовать для оценки точности разностного решения, а можно вычесть ее из разностного решения, тем самым уточнив его.

Замечание 2. Если функция $\bar{z}(x)$ такова, что к ней самой применима теорема 2, то можно использовать рекуррентный метод Рунге, несколько раз сгущая сетку.

Изложенная в этой главе теория разностных схем применима к разностным схемам, аппроксимирующим корректно поставленные задачи для обыкновенных дифференциальных уравнений, уравнений в частных производных и интегральных уравнений. Теория переносится на решение уравнений в частных производных методом прямых. Хотя в большинстве формулировок фигурировало только одно уравнение и одна переменная, но теория очевидным образом обобщается на системы уравнений или случай многих переменных.

Теория разностных схем применяется также для доказательства существования решения точной задачи (70) и установления его свойств. В качестве примера приведем без доказательства одно утверждение:

Теорема 3. Если для задачи (70) существует хотя бы одна корректная разностная схема (71), аппроксимирующая задачу на функциях $u(x) \in U$, то решение $u(x)$ задачи (70) в классе U существует и единственно. Если правые части $\varphi(x)$, $\chi(x)$ непрерывны равномерно по h, то $u(x)$ непрерывно зависит от $f(x)$, $\mu(x)$.

3. Сравнение схем на тестах. Для любой задачи даже на фиксированных сетке и шаблоне можно составить много разностных схем. Естественно, возникает вопрос: какую из схем использовать при решении реальной задачи? Как правило, традиционных оценок сходимости и точности для ответа недостаточно. Это связано с несовершенством теоретических методов исследования схем.

1) Для большинства нелинейных задач (например, газодинамических) нет доказательства сходимости или хотя бы устойчивости разностных схем. Соображения об их устойчивости и сходимости основаны на анализе линеаризованных задач.

2) Оценки точности схем являются асимптотическими при стремлении шага к нулю. Но быстродействие и память современных ЭВМ не настолько велики, чтобы можно было относительно сложные реальные задачи считать достаточно малым шагом. Например, для трехмерной задачи сетка из 27 000 узлов, соответствующая оперативной памяти ЭВМ БЭСМ-6, содержит всего 30 интервалов по каждой переменной.

Реально может оказаться, что схема первого порядка точности на грубых сетках даст более точный результат, чем схема второго порядка точности, хотя на подробных сетках соотношение будет обратным.

3) Обычно априорные оценки точности схем далеки от оптимальных. Они бывают завышены в десятки и сотни раз, и только в исключительных случаях удается получить неулучшаемые оценки. Но даже эти неулучшаемые оценки относятся к достаточно широкому классу решений, а для конкретного решения могут быть сильно завышены.

4) Даже наличие доказательства сходимости разностной схемы не гарантирует хорошего качества полученного по схеме решения. Сходимость в гильбертовой норме обеспечивает передачу только некоторых интегральных характеристик решения. Сходимость в чебышевской норме обеспечивает хорошее качество решения лишь при достаточной подробной сетке. В расчетах на грубых сетках при сходящейся схеме нередко возникает «разболтка», делающая результаты расчета фактически неприемлемыми.

Большой опыт численных расчетов показывает, что, помимо аппроксимации и устойчивости, разностные схемы должны удовлетворять добавочным критериям, обеспечивающим передачу некоторых качественных свойств решения. Хорошо известным критерием является консервативность схем или, в более общей форме, инвариантность разностных уравнений относительно определенной группы преобразований. Другие употребительные критерии — это аппроксимационная вязкость схем или диссипативность первого дифференциального приближения и монотонность схем.

Вероятно, в дальнейшем будет создана достаточно строгая качественная теория разностных схем, позволяющая ответить на многие вопросы. Но даже после создания такой теории важным элементом работы останется экспериментальное исследование схем, т. е. проверка их на небольшой системе тестов.

Тестом может служить задача, которая содержит специфические трудности данного класса задач и точное решение которой известно. Это решение может задаваться формулой или находиться численно; в качестве тестов нередко используют автомодельные решения. Для проверки схемы следует провести серию из трех или более расчетов задачи-теста с последовательным сгущением сеток и сравнить разностное решение с точным.

Точность схемы оценивают по норме погрешности разностного решения. Для более полного изучения схемы проверяют сходимость в разных нормах (обычно в C и L_2). При этом обязательно сравнивают фактическую скорость убывания погрешности при $h \to 0$ с теоретическим порядком точности схемы.

Возможны случаи, когда ожидаемый теоретический порядок точности не совпадает с фактическим. О чем это может свидетельствовать? Отметим некоторые типичные ситуации.

а) Метод теоретического исследования был строгим, а фактический порядок точности ниже теоретического. Возможны две причины. 1) Численный расчет был неправильным; например,

программа для ЭВМ содержала ошибки. 2) При теоретическом анализе аппроксимация определялась на функциях более гладких, чем использованное в качестве теста решение $u(x, t)$.

б) Метод теоретического исследования был строгим, а фактический порядок точности выше теоретического. Это означает, что теоретическое исследование было недостаточно полным. Может быть, при доказательстве устойчивости использовались более сильные нормы для правых частей, чем в действительности необходимо; или погрешность аппроксимации определялась не на решении данной задачи, а на заметно более широком классе функций.

в) Метод исследования был нестрогим; например, устойчивость нелинейной схемы изучалась методом разделения переменных. В этом случае теоретическое исследование вообще не дает ответа, а лишь позволяет сделать довольно вероятный прогноз. Сравнение же на тестах позволяет установить здесь точный характер сходимости, правда, только на отдельных примерах.

Исследовать надо не только разностные схемы, но и сетки. Разные классы задач предъявляют разные требования к сеткам. Но лишь в отдельных случаях эти требования удается четко сформулировать; например, если точное решение имеет разрыв или другую особенность, то желательно совместить с ней узел сетки. В остальных же случаях приходится сравнивать сетки тоже на тестах. Зачастую удачный выбор сетки повышает точность расчета не меньше, чем усовершенствование разностной схемы.

ЗАДАЧИ

1. Для уравнения (9) найти автомодельное решение вида $u(x, t)=f(\xi)$, $\xi=x/t$, и соответствующие ему начальные и граничные условия.

2. Найти выражения невязки для случая, рассмотренного в § 1, п. 3, замечание 3.

3. Вывести разностную схему (29) и найти ее невязку.

4. Определить невязку разностного краевого условия (33) и сравнить ее с (32).

5. При доказательстве теоремы в § 3, п. 2 использовано определение равномерной устойчивости (48) для линейных схем; обобщить это определение и доказательство теоремы на случай нелинейных схем.

6. Доказать утверждения, сделанные в замечании к теореме из § 3, п. 2.

7. Доказать замечание 1 об устойчивости по правой части в § 3, п. 4.

8. Доказать утверждение, сделанное в замечании 4 к теореме 1 из § 4, п. 2, и дать для нелинейных схем априорную оценку точности типа (79).

9. Обобщить теорему 2 из § 4, п. 2 на случай разного порядка аппроксимации по различным переменным.

ГЛАВА X

УРАВНЕНИЕ ПЕРЕНОСА

В главе X рассмотрены основные разностные схемы для простейшего уравнения в частных производных — уравнения переноса. В § 1 построены схемы бегущего счета для линейного уравнения переноса, как одномерного, так и многомерного. На их примере дана геометрическая интерпретация устойчивости разностных схем и введены понятия монотонности, аппроксимационной вязкости и первого дифференциального приближения разностных схем, полезные при качественном анализе разностных решений.

В § 2 рассмотрено простейшее квазилинейное уравнение переноса и исследованы качественные особенности его решений. Введено понятие консервативности разностных схем и изложен метод псевдовязкости; на их основе построены схемы для решения данной задачи.

§ 1. Линейное уравнение

1. Задачи и решения. Существует много задач о распространении частиц в веществе: определение нейтронных потоков в реакторе, теплопроводности в газах, обусловленной диффузией атомов и электронов, и т. д. Такие задачи приводят к уравнению переноса, которое может быть интегро-дифференциальным. Например, основное уравнение кинетической теории газов — уравнение Больцмана имеет следующий вид:

$$\frac{\partial u_i}{\partial t} + \boldsymbol{v}_i \frac{\partial u_i}{\partial \boldsymbol{r}} + \boldsymbol{F}_i \frac{\partial u_i}{\partial \boldsymbol{v}_i} = \sum_j \int (u_i' u_j' - u_i u_j)\, d\boldsymbol{v}_i' d\boldsymbol{v}_j, \qquad (1)$$

$$u_i = u_i(\boldsymbol{r},\ \boldsymbol{v}_i;\ t), \quad u_i' = u_i(\boldsymbol{r},\ \boldsymbol{v}_i',\ t).$$

Здесь u_i — функция распределения i-го сорта частиц; она зависит от времени, координаты и скорости частицы. Интегральный член в (1) описывает столкновения частиц.

Решение нелинейных интегро-дифференциальных уравнений типа (1) очень сложно и выходит за пределы нашего курса. Мы ограничимся изучением только линейного дифференциального уравнения переноса:

$$\frac{\partial u}{\partial t} + c(\boldsymbol{x},\ t)\,\mathrm{grad}\, u = f(\boldsymbol{x},\ t), \qquad \boldsymbol{x} = \{x_1,\ x_2,\ \ldots,\ x_p\}, \qquad (2)$$

где c — вектор скорости переноса. Как будет видно в дальнейшем, для этого уравнения многомерность не вносит принципиальных осложнений. Все основные идеи можно пояснить на одномерном уравнении

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = f(x,\ t), \tag{3}$$

где скорость c будем считать постоянной, если специально не оговорено противное.

Если в уравнении (3) правая часть $f = 0$, то общее решение этого уравнения имеет вид бегущей волны:

$$u(x,\ t) = \varphi(x - ct) \tag{4}$$

(отсюда видно, что c есть скорость переноса). Для определенности положим $c > 0$, тогда волна бежит слева направо. Вид решения (4) подсказывает, как можно корректно поставить полную задачу для уравнения (3).

Смешанная задача Коши. Зададим начальные и граничные данные на отрезках, показанных на рис. 55 жирными линиями:

$$u(x,\ 0) = \mu_1(x), \quad 0 \leqslant x \leqslant a,$$
$$u(0,\ t) = \mu_2(t), \quad 0 \leqslant t \leqslant T. \tag{5}$$

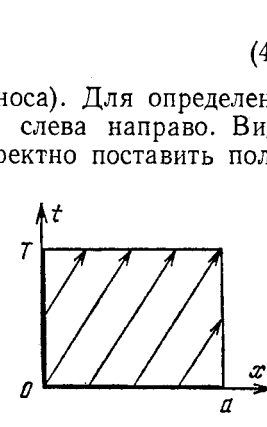


Рис. 55.

Тогда решение задачи (3), (5) однозначно определено в области $G = [0 \leqslant x \leqslant a] \times [0 \leqslant t \leqslant T]$. Если начальные и граничные данные непрерывны вместе со своими p-ми производными, причем выполнены условия согласования в точке стыка кусков границы (для случая $f(x,\ t) \equiv 0$ они имеют следующий вид:

$$\frac{d^q \mu_2(0)}{dt^q} = (-c)^q \frac{d^q \mu_1(0)}{dx^q}, \quad 0 \leqslant q \leqslant p) \tag{6}$$

и $f(x,\ t)$ непрерывна вместе с $(p-1)$-ми производными, то решение $u(x,\ t)$ непрерывно в G вместе с p-ми производными.

Задача Коши. Зададим начальные данные на полубесконечной прямой: $u(x,\ 0) = \mu(x)$ при $-\infty < x \leqslant a$. Тогда решение однозначно определено в области $G = (-\infty < x \leqslant a] \times [0 \leqslant t < +\infty)$. Гладкость решения соответствует гладкости начальных данных $\mu(x)$ и правой части $f(x,\ t)$.

Характеристики уравнения (3) имеют вид $x - ct = \text{const}$ и при постоянной скорости c являются прямыми линиями. Решение (4) однородного уравнения (3) постоянно вдоль такой линии; поэтому говорят, что начальные и граничные условия переносятся по характеристикам.

Решение неоднородного уравнения (3) меняется вдоль характеристик. Это изменение легко найти, если перейти к новым координатам, связанным с характеристиками:

$$\xi = x - ct, \quad \eta = x + ct. \tag{7}$$

При их помощи уравнение (3) преобразуется к виду

$$2c\frac{\partial u}{\partial \eta} = \varphi(\xi, \eta), \quad \varphi(\xi, \eta) = f\left(\frac{\xi+\eta}{2}, \frac{\eta-\xi}{2c}\right). \tag{8}$$

Следовательно, вдоль характеристики $\xi = \text{const}$ решение u можно найти, интегрируя по η обыкновенное дифференциальное уравнение (8), в котором ξ играет роль параметра. Так можно определить решение в любой точке области G, поскольку при $c = \text{const}$ характеристики покрывают всю область.

Этот способ построения точного решения легко обобщается на уравнение с переменным коэффициентом $c(x, t)$. Он показывает, что для корректной постановки задачи необходимо, чтобы через любую точку области G проходила одна и только одна характеристика. Это выполняется, если функция $c(x, t)$ непрерывна во всей области $G + \Gamma$.

Сохранение монотонности является важным свойством однородного уравнения переноса. Если для него поставлена задача Коши с монотонными начальными данными $u(x, 0) = \mu(x)$, $-\infty < x \leqslant a$, то в любой момент t профиль $u(x, t)$ тоже будет монотонным *). Монотонность сохраняется и в смешанной задаче Коши, если граничное значение $u(0, t)$ тоже монотонно зависит от t и согласовано с начальными данными.

В уравнении переноса монотонность является тривиальным следствием из вида общего решения (4). Однако во многих уравнениях начальная монотонность решения сохраняется, хотя общее решение не имеет вида одной бегущей волны. При определенных условиях это имеет место даже в задачах теплопроводности и газодинамики. Поэтому монотонность — достаточно общее и важное свойство многих уравнений.

2. Схемы бегущего счета. Эти схемы предназначены для решения смешанной задачи Коши (3), (5). Они легко обобщаются на случай любого числа измерений. Схемы бегущего счета являются наиболее простыми и позволяют численно решать даже очень сложные задачи переноса с хорошей точностью при умеренном объеме вычислений.

Рассмотрим задачу (3), (5) и построим в области $G = [0 \leqslant x \leqslant a] \times [0 \leqslant t \leqslant T]$ прямоугольную сетку, для простоты равномерную с шагами h и τ. Выберем четыре шаблона, изображен-

*) *Профилем* (по x_α) называют зависимость функции $F(x_1, \ldots, x_p, t)$ от одной из пространственных переменных x_α.

ные на рис. 56—59. Составим на трехточечных шаблонах (рис. 56—58) простейшие схемы с использованием односторонних производных:

$$\left.\begin{aligned} &\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{c}{h}(y_n - y_{n-1}) = \varphi_n, \\ &\varphi_n = f\left(x_n - \frac{h}{2},\ t_m + \frac{\tau}{2}\right), \end{aligned}\right\} \tag{9}$$

$$\frac{1}{\tau}(\hat{y}_{n-1} - y_{n-1}) + \frac{c}{h}(\hat{y}_n - \hat{y}_{n-1}) = \varphi_n, \tag{10}$$

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{c}{h}(\hat{y}_n - \hat{y}_{n-1}) = \varphi_n, \tag{11}$$

а на четырехточечном шаблоне (рис. 59) — схему с симметризованными производными:

$$\frac{1}{2\tau}(\hat{y}_n + \hat{y}_{n-1} - y_n - y_{n-1}) + \frac{c}{2h}(\hat{y}_n + y_n - \hat{y}_{n-1} - y_{n-1}) = \varphi_n. \tag{12}$$

Правую часть мы для определенности выбираем в центре ячейки, соответствующей шаблону, хотя возможен и другой выбор.

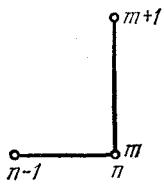


Рис. 56.

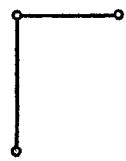


Рис. 57.

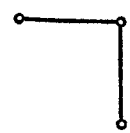


Рис. 58.

Рис. 59.

Организация расчета по этим схемам очень проста. Хотя формально схема (9) является явной, а остальные три — неявными, фактически при расчете смешанной задачи Коши они ведут себя, как явные.

В самом деле, во всех четырех схемах значение $\hat{y}_n$ явно выражается через значения $\hat{y}_{n-1}$, y_n, y_{n-1} (или любые два из них). Значение решения на нулевом слое $y_n^0 = \mu_1(x_n)$ известно из начального условия. На следующем (первом) слое значение $\hat{y}_0 = \mu_2(t_1)$ в силу граничного условия, и можно вычислить $\hat{y}_1$; зная $\hat{y}_1$, можно вычислить $\hat{y}_2$, затем $\hat{y}_3$. Так последовательно вычисляются слева направо все $\hat{y}_n$ первого слоя. Затем, зная решение на первом слое, точно так же вычисляем его на втором слое, на третьем и т. д.

Замечание 1. Явная схема (9) пригодна для решения задачи Коши на полубесконечной (или бесконечной) прямой; неявные схемы бегущего счета к такой задаче неприменимы. Правда, в практике численных расчетов задача Коши для уравнения переноса в неограниченной области почти не встречается.

Из описанного алгоритма видно, что для каждой из схем (9) — (12) разностное решение при любых φ_n существует и единственно. Поэтому для доказательства сходимости остается исследовать аппроксимацию и устойчивость схем. Заметим, что краевое условие $u(0, t) = \mu_2(t)$ для всех схем аппроксимируется точно; поэтому устойчивости по нему не требуется.

Схема (9). Исследуем ее погрешность аппроксимации. Пусть начальные и граничные данные дважды непрерывно дифференцируемы и удовлетворяют условиям согласования типа (6) с $p=2$, а правая часть $f(x, t)$ имеет непрерывные первые производные. Тогда решение $u(x, t)$ дважды непрерывно дифференцируемо; разложим его по формуле Тейлора в узле (x_n, t_m):

$$\hat{u}_n = u_n + \tau u_t + \frac{1}{2}\tau^2 u_{tt},$$

$$u_{n-1} = u_n - h u_x + \frac{1}{2} h^2 u_{xx},$$

$$\varphi_n = f_n + \frac{1}{2}\tau f_t - \frac{1}{2} h f_x.$$

Отсюда легко определим невязку схемы (9):

$$\psi_n = \left(\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} - f\right)_n - \left[\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{c}{h}(u_n - u_{n-1}) - \varphi_n\right] =$$

$$= \frac{\tau}{2}(f_t - u_{tt}) + \frac{h}{2}(c u_{xx} - f_x) = O(\tau + h). \quad (13)$$

При сделанных предположениях схема (9) имеет аппроксимацию в $\|\cdot\|_c$ с первым порядком.

Устойчивость исследуем при помощи принципа максимума. Критерий равномерной устойчивости по начальным данным (9.53) с константой $C=0$ принимает вид

$$\frac{1}{\tau} \geqslant \frac{c}{h} + \left|\frac{1}{\tau} - \frac{c}{h}\right|.$$

Он выполняется только при так называемом условии Куранта:

$$c\tau \leqslant h. \quad (14)$$

Таким образом, схема (9) является условно устойчивой в $\|\cdot\|_c$.

Методом разделения переменных можно доказать необходимость условия (14). Рассматривая отдельную гармонику $\exp(iqx)$ и подставляя в (9) величины

$$\varphi_n = 0, \quad y_n = e^{iqx}, \quad y_{n-1} = e^{iq(x-h)}, \quad \hat{y}_n = \rho_q y_n,$$

легко получим множитель роста этой гармоники:

$$\rho_q = 1 - \frac{c\tau}{h}(1 - e^{-iqh}). \quad (15)$$

Если $c\tau > h$, то для тех гармоник, у которых $\cos qh = -1$, множитель роста равен

$$|\rho_q| = \left|1 - \frac{2c\tau}{h}\right| = \frac{2c\tau}{h} - 1 > 1,$$

т. е. амплитуды этих гармоник неограниченно нарастают при $\tau \to 0$. Устойчивости нет, что и требовалось доказать.

Непосредственно видно, что дополнительное условие устойчивости по правой части (9.54) выполняется, причем $\varkappa = 1$. Поэтому схема устойчива по правой части в $\|\cdot\|_c$ при выполнении условия (14).

Тогда из теорем о сходимости следует, что если решение $u(x, t)$ непрерывно вместе со своими вторыми производными, то схема (9) при выполнении условия Куранта (14) сходится в $\|\cdot\|_c$ со скоростью $O(\tau + h)$, т. е. с первым порядком точности.

Схема (10) исследуется аналогично; при исследовании аппроксимации разложение по формуле Тейлора удобнее вести около узла $(x_{n-1}, t_m + \tau)$. На дважды непрерывно дифференцируемых решениях эта схема при выполнении условия устойчивости

$$c\tau \geqslant h \tag{16}$$

обеспечивает сходимость со скоростью $O(\tau + h)$.

Схема (11) безусловно устойчива и на дважды непрерывно дифференцируемых решениях сходится со скоростью $O(\tau + h)$.

Схема (12) симметричная, и при исследовании ее аппроксимации целесообразно разлагать $u(x, t)$ по формуле Тейлора около центра ячейки $\left(x_n - \frac{h}{2}, t_m + \frac{\tau}{2}\right)$. Тогда после довольно громоздких выкладок определяем невязку:

$$\psi = -\tau^2\left(\frac{1}{24}u_{ttt} + \frac{c}{8}u_{ttx}\right) - h^2\left(\frac{1}{8}u_{txx} + \frac{c}{24}u_{xxx}\right) = O(\tau^2 + h^2). \tag{17}$$

Схема имеет второй порядок аппроксимации, если решение $u(x, t)$ трижды непрерывно дифференцируемо.

Устойчивость схемы (12) при помощи принципа максимума установить не удается. Однако можно провести исследование методом разделения переменных. Для гармоники $\exp(iqx)$ нетрудно получить выражение для множителя роста. Полагая в (12)

$$\varphi_n = 0, \quad y_n = \exp(iqx_n), \quad \hat{y} = \rho_q y,$$

найдем

$$\rho_q = e^{-iqh}\frac{(h + c\tau) + (h - c\tau)e^{iqh}}{(h + c\tau) + (h - c\tau)e^{-iqh}}. \tag{18}$$

Отсюда видно, что $|\rho_q| = 1$ для любой гармоники при любых соотношениях шагов. Следовательно, схема равномерно устойчива по начальным данным в $\|\cdot\|_{l_2}$, причем устойчивость безусловная.

Дополнительный критерий устойчивости по правой части (9.54) после умножения на τ принимает для схемы (12) следующий вид:

$$1+\frac{c\tau}{h}-\left|1-\frac{c\tau}{h}\right|\geqslant\varkappa,\qquad \varkappa=\text{const}>0.$$

Убедимся, что для $\varkappa=2$ это неравенство выполняется при любых τ и h. В самом деле, если $c\tau\leqslant h$, то левая часть неравенства равна 2. Если же $c\tau>h$, то левая часть неравенства равна $(2c\tau/h)>2$. Поскольку критерий выполнен, то схема безусловно устойчива по правой части.

Из сказанного выше следует, что на трижды непрерывно дифференцируемых решениях $u(x, t)$ схема (12) безусловно сходится в норме $\|\cdot\|_{l_2}$ со скоростью $O(\tau^2+h^2)$. Судя по результатам численных расчетов, схема обеспечивает второй порядок точности и в $\|\cdot\|_c$.

Замечание 2. Схемы бегущего счета сходятся на решениях меньшей гладкости и даже на разрывных решениях (разумеется, не равномерно, а в среднем). Например, теоретический анализ и примеры численных расчетов [65, 66] показали, что схема (11) сходится на кусочно-непрерывных решениях в $\|\cdot\|_{l_p}$ с погрешностью $O((\tau+h)^{1/2p})$. Любопытно, что порядок точности оказался не целым!

Замечание 3. Схемы бегущего счета очевидным образом обобщаются на случай неравномерной сетки. Например, схему (9) можно записать следующим образом:

$$\frac{1}{\tau}(\hat{y}_n-y_n)+\frac{c_n}{h_n}(y_n-y_{n-1})=\varphi_n,\qquad h_n=x_n-x_{n-1}. \tag{19}$$

Критерии устойчивости (14) и (16) принимают при этом соответственно вид:

$$\tau\leqslant\min_n(h_n/c_n)\quad\text{и}\quad\tau\geqslant\max_n(h_n/c_n). \tag{20}$$

Интересно сравнить схемы (9) — (12) между собой. Схема (12) имеет второй порядок точности и на достаточно гладких решениях при не слишком больших шагах τ и h дает лучшие результаты на примерах-тестах. Но на разрывных решениях или для быстропеременных решений на грубой сетке она оказывается плохой; в этих случаях удовлетворительные результаты дают схемы (9) — (11).

Схемы (9) — (11) имеют первый порядок точности. Первые две из них условно устойчивы, что неудобно при численных расчетах. Схема (11) безусловно устойчива и очень надежна в расчете; однако по точности она уступает схемам (9) и (10), в чем нетрудно убедиться, сравнив невязки этих схем.

Дальше мы увидим, что схемы (9) и (10) можно объединить в единую явно-неявную схему, безусловно устойчивую и превосходящую схему (11) по точности.

3. Геометрическая интерпретация устойчивости. Ограничимся устойчивостью по начальным данным. Рассмотрим однородное уравнение (3) с $f(x, t) = 0$, общее решение которого имеет вид $u(x, t) = \varphi(x - ct)$, т. е. переносится по характеристикам $x - ct = \text{const}$ без изменения.

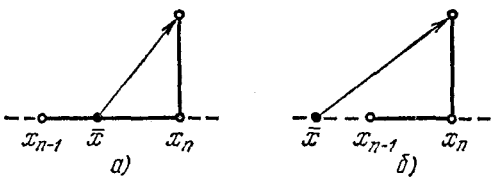


Рис. 60.

Рассмотрим схему (9) с шаблоном, изображенным на рис. 60 (см. также рис. 56). Построим характеристику, проходящую через искомый узел (x_n, t_{m+1}); она обозначена стрелкой на рис. 60. Эта характеристика пересекает исходный слой t_m в точке $\bar{x} = x_n - c\tau$. Схему (9) без правой части можно интерпретировать следующим образом. Линейно интерполируя разностное решение между узлами исходного слоя, найдем

$$y(\bar{x}) = y_{n-1}\frac{x_n - \bar{x}}{h} + y_n\frac{\bar{x} - x_{n-1}}{h} = \frac{c\tau}{h}y_{n-1} + \left(1 - \frac{c\tau}{h}\right)y_n. \qquad (21)$$

Затем найденное значение перенесем без изменения по характеристике в искомый узел, т. е. положим $\hat{y}_n = y(\bar{x})$.

Если выполнено условие устойчивости схемы $c\tau \leqslant h$, то $x_{n-1} \leqslant \bar{x} < x_n$; в противном случае $\bar{x} < x_{n-1}$. Иными словами, схема (9) устойчива, если $\hat{y}_n$ вычисляется по ранее найденным значениям y при помощи интерполяции (рис. 60, *а*); схема неустойчива, если используется экстраполяция (рис. 60, *б*).

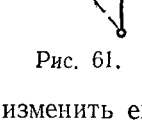


Рис. 61.

Причина этого состоит в том, что при точной постановке задачи в узел (x_n, t_{m+1}) приходят возмущения только из точки $\bar{x}$ исходного слоя t_m. Если точка $\bar{x}$ лежит вне отрезка $[x_{n-1}, x_n]$, то, сохраняя непрерывность и гладкость решения, можно сильно изменить его на этом отрезке (на слое t_m), не меняя значения $u(\bar{x}, t_m)$. Значение $\hat{u}_n = u(\bar{x}, t_m)$ при этом сохраняется, а значение $\hat{y}_n$ сильно изменяется, поскольку оно вычисляется по изменившимся значениям y_n, y_{n-1}. Значит, $\hat{y}_n$ не может сходиться к $\hat{u}_n$.

Схемы (10) и (11) тоже можно интерпретировать как линейную интерполяцию по двум уже вычисленным значениям, с последующим переносом по характеристике. В частности, безусловная устойчивость схемы (11) связана с тем, что приходящая в искомый узел характеристика (стрелка на рис. 61) при любых τ и h пересекает отрезок, соединяющий исходные узлы (пунктир на рисунке).

Схема (12) интерпретируется тоже как интерполяция, но не двухточечная линейная, а трехточечная квадратичная (что, естественно, приводит к более высокому порядку точности). Какую бы сторону ячейки на рис. 59 ни пересекала приходящая в узел (x_n, t_{m+1}) характеристика — горизонтальную или вертикальную, эта сторона связывает узлы с ранее вычисленными значениями y; поэтому экстраполяции здесь нет, что приводит к безусловной устойчивости схемы (12).

Таким образом, прослеживая положение характеристик, нетрудно так выбрать шаблон и составить на нем разностную схему, чтобы схема была устойчива. Приведем несколько примеров.

Явно-неявная схема. Будем считать, что шаги по времени $\tau_m = t_{m+1} - t_m$ и по пространству $h_n = x_{n+1} - x_n$ не постоянны, а коэффициент $c(x, t)$ уравнения (3) переменный. Приступая к вычислению $\hat{y}_n$, проверим критерий Куранта (14) в данной ячейке. Если он выполнен, то проведем вычисления по схеме (9):

$$\frac{1}{\tau_m}(\hat{y}_n - y_n) + \frac{\hat{c}_n}{h_{n-1}}(y_n - y_{n-1}) = \varphi_n \quad \text{при} \quad \hat{c}_n\tau_m \leqslant h_{n-1}. \tag{22а}$$

В противном случае воспользуемся схемой (10):

$$\frac{1}{\tau_m}(\hat{y}_{n-1} - y_{n-1}) + \frac{\hat{c}_n}{h_{n-1}}(\hat{y}_n - \hat{y}_{n-1}) = \varphi_n \quad \text{при} \quad \hat{c}_n\tau_m > h_{n-1}. \tag{22б}$$

Очевидно, явно-неявная схема (22) безусловно устойчива, причем ее невязка меньше, чем у безусловно устойчивой схемы (11). Схему (22) обычно применяют в тех случаях, когда точное решение является недостаточно гладким или быстропеременным.

Схема без шаблона. Проведем через искомый узел (x_n, t_{m+1}) характеристику и определим точку ее пересечения с исходным слоем $\bar{x} = x_n - c\tau$. Найдем на исходном слое такую пару узлов x_p, x_{p+1}, между которыми заключена точка $\bar{x}$. Определим $y(\bar{x})$ линейной интерполяцией по значениям y_p, y_{p+1}:

$$y(\bar{x}) = \frac{x_{p+1} - \bar{x}}{x_{p+1} - x_p} y_p + \frac{\bar{x} - x_p}{x_{p+1} - x_p} y_{p+1}, \qquad x_p \leqslant \bar{x} < x_{p+1}. \tag{23}$$

Перенесем вычисленное значение по характеристике в искомый узел, т. е. положим $\hat{y}_n = y(\bar{x})$. Очевидно, схема (23) абсолютно устойчива; но по точности и удобству вычислений она уступает схеме (22) и поэтому редко применяется.

В схеме (23) положение узлов p, $p+1$ относительно узла n не фиксировано. Если скорость $c(x, t)$ переменна или сетка x_n неравномерна, то $n - p$ будет непостоянной величиной. Таким образом, эта схема не имеет шаблона.

Случай $c<0$. В этом случае наклон характеристик на плоскости (x, t) отрицателен; характеристики зеркально отражены относительно вертикали по сравнению со случаем $c>0$. Соответственно меняется постановка задачи: для отрезка $0\leqslant x\leqslant a$ граничное условие теперь должно задаваться справа, при $x=a$.

Очевидно, шаблоны для устойчивых схем можно получить зеркальным отражением соответствующих шаблонов рис. 56—59. Например, вместо шаблона рис. 56 берут шаблон рис. 62, получая устойчивую при $|c|\tau\leqslant h$ схему. Направление бегущего счета также меняется: расчет на каждом слое ведут справа налево.

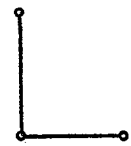


Рис. 62.

Отметим, что шаблоны рис. 57 и 58 зеркальны друг другу; это означает, что при $c<0$ схема (10) становится безусловно устойчивой, а схема (11) — условно устойчивой. Симметричная схема (12) не меняется при отражении, так что она устойчива при любом знаке скорости; но направление счета, разумеется, зависит от знака скорости.

Знакопеременная $c(x, t)$. В этом случае задача в области $G(x, t)$ поставлена корректно, если заданы значения решения на тех и только тех границах, с которых характеристики идут внутрь области.

Пусть, например, скорость $c(x, t)$ непрерывна в области $G=[0\leqslant x\leqslant a]\times[0\leqslant t\leqslant T]$ и меняет знак только при $x=\tilde{x}$, причем $c(0, t)>0$, $c(a, t)<0$. Вид характеристик в этом случае изображен на рис. 63. Корректной будет постановка задачи с двумя граничными условиями:

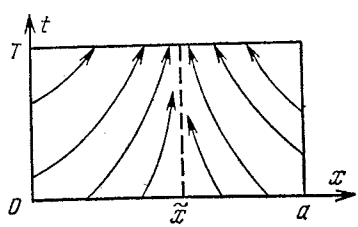


Рис. 63.

$$\begin{aligned} &u_t+c(x, t)\,u_x=f(x, t),\\ &u(x, 0)=\mu_1(x),\\ &u(0, t)=\mu_2(t),\\ &u(a, t)=\mu_3(t). \end{aligned} \qquad (24)$$

Фактически здесь имеется зона влияния каждой границы; эти зоны разделены линией $x=\tilde{x}$ (пунктир на рисунке). В каждой зоне можно построить схему бегущего счета со своим направлением движения.

Можно поступить и иначе. Возьмем шаблон рис. 64 и построим на нем неявную схему

$$\frac{1}{\tau}(\hat{y}_n-y_n)+\frac{\hat{c}_n}{2h}(\hat{y}_{n+1}-\hat{y}_{n-1})=\varphi_n, \qquad 1\leqslant n\leqslant N-1. \qquad (25)$$

По направлению характеристики (стрелки на рисунке) видно,

что при любом знаке c и любых шагах τ и h значение $\hat{y}_n$ вычисляется интерполяцией. Методом разделения переменных нетрудно показать, что схема (25) безусловно устойчива при любом знаке c.

Схема (25) содержит три точки нового слоя. В главе IX отмечалось, что в подобных случаях разностное решение находят прогонкой. Достаточное условие устойчивости прогонки (5.14) в этом случае выполняется только при $|c|\tau \leqslant h$, хотя обычно можно вести расчет и при нарушении этого условия.

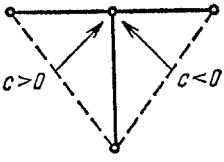


Рис. 64.

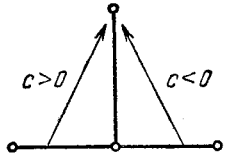


Рис. 65.

Замечание. Геометрическая интерпретация дает необходимое, но не достаточное условие устойчивости. Например, рассмотрим явную схему на шаблоне рис. 65:

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{c}{2h}(y_{n+1} - y_{n-1}) = \varphi_n. \tag{26}$$

При $|c|\tau \leqslant h$ она соответствует интерполяции на исходном слое. Однако она неустойчива при любом соотношении шагов (*абсолютно* неустойчива), что легко доказать методом разделения переменных. В самом деле, подставляя в (26)

$$\varphi_n = 0, \quad y_n = e^{iqx}, \quad y_{n\pm 1} = e^{iq(x\pm h)}, \quad \hat{y}_n = \rho y_n,$$

получим множитель роста отдельной гармоники:

$$\rho_q = 1 - \frac{c\tau}{h}\cos qh.$$

Для гармоники с $\cos qh = -1$ этот множитель $\rho_q = 1 + (c\tau/h)$ неограниченно велик при $h \to 0$. Значит, устойчивости нет.

Поэтому геометрическую интерпретацию используют как способ быстрой оценки качества шаблона и схемы и отбраковки заведомо плохих схем. Устойчивость выбранных при ее помощи схем обязательно проверяют методами, изложенными в главе IX (в большинстве случаев отобранные этим способом схемы оказываются устойчивыми).

4. Многомерное уравнение. Схемы бегущего счета естественно обобщаются на многомерное уравнение переноса. Рассмотрим, для определенности, задачу с двумя пространственными переменными в области $G = [0 \leqslant x_1 \leqslant a] \times [0 \leqslant x_2 \leqslant b] \times [0 \leqslant t \leqslant T]$:

$$u_t + c_1 u_{x_1} + c_2 u_{x_2} = f(x_1, x_2, t); \tag{27а}$$

$$u(x_1, x_2, 0) = \mu_1(x_1, x_2), \quad u(0, x_2, t) = \mu_2(x_2, t), \quad u(x_1, 0, t) = \mu_3(x_1, t). \tag{27б}$$

Скорости переноса по осям c_1, c_2 считаем положительными и, для простоты, постоянными.

Построим, например, многомерный аналог абсолютно устойчивой схемы (11). Введем по переменной x_1 сетку $\{x_{1n},\ 0 \leqslant n \leqslant N\}$, а по переменной x_2 — сетку $\{x_{2m},\ 0 \leqslant m \leqslant M\}$. Значения решения в узлах этой сетки обозначим следующим образом:

$$u(x_{1n},\ x_{2m},\ t) = u_{nm}, \quad u(x_{1n},\ x_{2m},\ t+\tau) = \hat{u}_{nm}. \tag{28}$$

Возьмем шаблон, изображенный жирными линиями на рис. 66, и составим на нем схему

$$\frac{1}{\tau}(\hat{y}_{nm} - y_{nm}) + \frac{c_1}{h_1}(\hat{y}_{nm} - \hat{y}_{n-1,m}) + \frac{c_2}{h_2}(\hat{y}_{nm} - \hat{y}_{n,m-1}) = \varphi_{nm}, \tag{29}$$

где h_1, h_2 — шаги по соответствующим направлениям.

Исследовать схему (29) несложно. Из принципа максимума сразу следует безусловная устойчивость этой схемы. Ее невязка определяется разложением по формуле Тейлора и равна $O(\tau + h_1 + h_2)$. Следовательно, схема (29) сходится в $\|\cdot\|_c$ с первым порядком точности *).

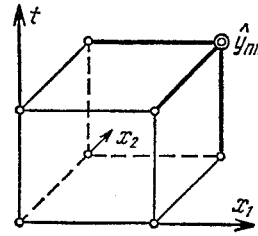


Рис. 66.

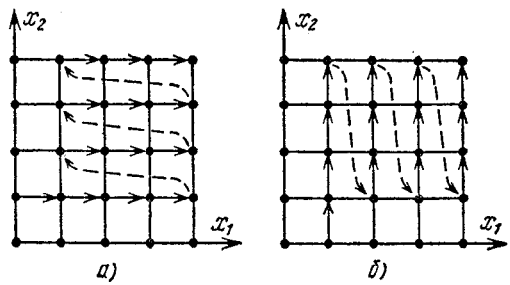
Вычисления проводятся послойно. Значение $\hat{y}_{nm}$ в узле, отмеченном на рис. 66 двойным кружком, выражается по формуле (29) через значения в нескольких других вершинах ячейки. Когда решение на слое t вычислено, то его значения на слое $t+\tau$ можно вычислять по этой формуле вдоль направлений x_1 (см. рис. 67, *а*, где последовательность вычислений указана стрелками).

Заметим, что последовательность вычислений может быть иной. Например, можно вести расчет на слое вдоль направлений x_2 (рис. 67, *б*). В принципе, не обязательна даже послойная организация расчета; достаточно, чтобы последовательность расчета соответствовала какому-то порядку заполнения первого координатного угла в пространстве (x_1, x_2, t) ячейками, при котором новая ячейка прикладывается тремя гранями к ранее уложенным ячейкам или координатным плоскостям.

Двумерный аналог симметричной схемы (12), имеющий второй порядок точности, нетрудно написать методом баланса. Для этого проинтегрируем уравнение (27а) по ячейке, преобразуем трехкратные интегралы в двукратные и вычислим последние по формуле трапеций. Детали настолько очевидны, что мы на них не будем останавливаться.

*) Разумеется, если решение дважды непрерывно дифференцируемо. В дальнейшем обычно будем опускать такие оговорки, подразумевая существование у решения требуемого числа непрерывных производных.

Таким образом, в уравнении переноса многомерность не приводит к принципиальным усложнениям. Вычислительный алгоритм остается простым и экономичным. В декартовых координатах даже формулы расчета имеют обычно простой вид, хотя в криволинейных координатах (цилиндрических, сферических и т. д.) они могут быть громоздкими.



а) *б)*

Рис. 67.

5. Перенос с поглощением. Для неоднородного уравнения переноса (3) от способа аппроксимации правой части $f(x, t)$ зависит только порядок аппроксимации. Для получения схемы второго порядка (12) следует выбирать φ_n так, чтобы выполнялось условие

$$\varphi_n - f\left(x_n - \frac{h}{2},\ t_m + \frac{\tau}{2}\right) = O(\tau^2 + h^2).$$

В схемах первого порядка достаточно было требовать, чтобы

$$\varphi_n - f\left(x_n - \frac{h}{2},\ t_m + \frac{\tau}{2}\right) = O(\tau + h),$$

для этого можно, например, положить φ_n равным $f(x, t)$ в любой точке ячейки. На устойчивость выбор φ_n не влияет.

Положение несколько изменится, если правая часть f зависит от решения u. Рассмотрим это на примере простейшей линейной зависимости $f = -bu$, когда уравнение переноса принимает вид

$$u_t + cu_x = -bu. \tag{30}$$

Будем искать решение этого уравнения в виде $u(x, t) = v(x, t)\exp(-bt)$. Подставляя его в (30), получим для $v(x, t)$ однородное уравнение переноса $v_t + cv_x = 0$, общее решение которого является бегущей волной $v(x, t) = v(x - ct, 0)$. Следовательно, общее решение задачи Коши для уравнения (30) имеет вид

$$u(x, t) = e^{-bt}u(x - ct, 0). \tag{31}$$

Оно описывает перенос частиц по характеристикам при наличии поглощения (если $b > 0$) или размножения (если $b < 0$) частиц.

Дальше мы ограничимся случаем $b>0$, когда точное решение экспоненциально убывает со временем. Рассмотрим два варианта явной схемы (9):

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{c}{h}(y_n - y_{n-1}) = -b\hat{y}_n, \tag{32}$$

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{c}{h}(y_n - y_{n-1}) = -by_n, \tag{33}$$

отличающиеся только аппроксимацией члена с поглощением. Обе схемы имеют первый порядок аппроксимации. Исследуем их устойчивость методом разделения переменных.

Делая стандартную подстановку гармоники $\exp(iqx)$, получим для схемы (32) множитель роста

$$\rho_q = \frac{1}{1+b\tau}\left[1 - \frac{c\tau}{h}(1 - e^{-iqh})\right].$$

Если выполнено условие Куранта $c\tau \leqslant h$, то для любых гармоник справедливо неравенство $|\rho_q| \leqslant (1+b\tau)^{-1} < 1$, так что схема (32) не только устойчива, но и хорошо обусловлена: ошибки не нарастают, а неограниченно убывают при $t \to \infty$.

Для схемы (33) множитель роста

$$\rho_q = 1 - \frac{c\tau}{h}(1 - e^{-iqh}) - b\tau.$$

Если положить $c\tau = h$, то для гармоники с $\exp(-iqh) = -1$ выполняется соотношение $|\rho_q| = 1 + b\tau$, т. е. устойчивость слабая. Таким образом, характер устойчивости схем (32) и (33) является не вполне одинаковым.

Это различие проявляется сильней, если рассмотреть асимптотическую устойчивость схем (соответствующую поведению относительной погрешности $\|\delta y\|/\|u\|$ при $t \to \infty$). Точное решение убывает, как e^{-bt}, так что его гармоники за один шаг затухают, как $e^{-b\tau} \approx (1+b\tau)^{-1}$. Гармоники схемы (32) затухают не медленнее, так что схема (32) асимптотически устойчива при $c\tau \leqslant h$. Наоборот, у схемы (33) при $c\tau = h$ нет асимптотической устойчивости: гармоника с $\exp(-iqh) = -1$ не только не убывает, а даже возрастает.

Этот пример показывает, что на устойчивость может влиять способ аппроксимации не только высших производных данного уравнения, но и всех остальных членов.

Замечание. Общее решение (31) положительно, если начальные данные были положительны. Нетрудно показать, что схема (32) сохраняет это свойство общего решения. Если же схему (33) переписать в форме

$$\hat{y}_n = \left(1 - \frac{c\tau}{h} - b\tau\right)y_n + \frac{c}{h}y_{n-1},$$

то нетрудно видеть, что при достаточно большом коэффициенте $b>0$ (и не слишком малом шаге τ) возможны случаи, когда $\hat{y}_n$ становится отрицательным при y_n, $y_{n-1}>0$. Фактически это приводит к дополнительному ограничению на шаг τ схемы (33). В задачах с сильным поглощением это ограничение, формально не связанное с устойчивостью, может оказаться достаточно жестким.

6. Монотонность схем. В п. 1 отмечалось, что решение однородного уравнения переноса (3), соответствующее монотонным начальным данным, в любой момент времени имеет монотонный профиль. Сохраняется ли это свойство у разностного решения? Иными словами, пусть профиль y_n монотонен; будет ли монотонным профиль $\hat{y}_n$?

Однородные разностные схемы, сохраняющие монотонность профиля разностного решения, называются *монотонными*.

Признак монотонности. *Явная двуслойная линейная однородная схема*

$$\hat{y}_n=\sum_l \beta_l y_{n+l} \tag{34}$$

монотонна тогда и только тогда, когда все $\beta_l\geqslant 0$.

Доказательство. Из (34) следует равенство

$$\hat{y}_{n-1}-\hat{y}_n=\sum_l \beta_l (y_{n-1+l}-y_{n+l}). \tag{35}$$

Если профиль y_n монотонен (для определенности—невозрастающий), то все скобки в правой части (35) неотрицательны. Тогда, если все $\beta_l\geqslant 0$, то $\hat{y}_{n-1}-\hat{y}_n\geqslant 0$ и профиль $\hat{y}_n$ также невозрастающий. Достаточность условия $\beta_l\geqslant 0$ доказана.

Предположим, что хотя бы один коэффициент $\beta_k<0$. Выберем такой невозрастающий профиль:

$$y_{n+l}=1 \quad \text{при} \quad l\leqslant k-1,$$
$$y_{n+l}=0 \quad \text{при} \quad l\geqslant k.$$

Подставляя его в (35), получим

$$\hat{y}_{n-1}-\hat{y}_n=\beta_k (y_{n-1+k}-y_{n+k})=\beta_k<0,$$

т. е. монотонность нарушена: имеется локальное возрастание профиля $\hat{y}_n$. Необходимость условия $\beta_l\geqslant 0$ доказана.

Замечание 1. Признак монотонности относится к разностным схемам, аппроксимирующим как уравнение переноса, так и любые другие типы уравнений.

Замечание 2. Если двуслойная линейная однородная схема неявна, то ее можно преобразовать к явной форме (34), где пре-

делы суммы по l бесконечны, и затем применить признак монотонности.

Теорема. *Двуслойная линейная монотонная схема для уравнения переноса $u_t + cu_x = 0$ не может иметь второй или более высокий порядок точности.*

Доказательство. Предположим, что имеется линейная монотонная схема второго (или более высокого) порядка точности. Запишем ее в форме (34), где все $\beta_l \geqslant 0$. Построим равномерную сетку $x_n = nh$.

Выберем в качестве начальных данных задачи Коши квадратичную функцию

$$u(x,\ 0) = \left(\frac{x}{h} - \frac{1}{2}\right)^2 - \frac{1}{4}, \quad y_n^2 = \left(n - \frac{1}{2}\right)^2 - \frac{1}{4} \geqslant 0. \tag{36}$$

В этом случае решение есть также квадратичная функция и его третьи производные равны нулю. Невязка схем второго порядка точности выражается через третьи производные. Поэтому при квадратичных начальных данных (36) разностное решение для нашей схемы должно совпадать с точным решением.

На первом слое точное и разностное решения равны соответственно

$$u(x,\ \tau) = \left(\frac{x - c\tau}{h} - \frac{1}{2}\right)^2 - \frac{1}{4}, \quad y_n^1 = \left(n - \frac{c\tau}{h} - \frac{1}{2}\right)^2 - \frac{1}{4}. \tag{37}$$

Подставляя разностные решения на исходном (36) и новом (37) слоях в разностную схему (34), получим равенство

$$\left(n - \frac{c\tau}{h} - \frac{1}{2}\right)^2 - \frac{1}{4} = \sum_l \beta_l \left[\left(l - \frac{1}{2}\right)^2 - \frac{1}{4}\right].$$

В правой части этого равенства стоит неотрицательная величина. Но левая часть при не целом $c\tau/h$ в одной из точек x_n отрицательна. Полученное противоречие доказывает теорему.

Следствие. *Линейные монотонные схемы для уравнения переноса могут иметь только первый порядок точности.*

Примеры. Схема (9) явная, и при выполнении условия устойчивости $c\tau \leqslant h$ ее коэффициенты неотрицательны. Следовательно, она монотонна.

Безусловно устойчивая схема (11) неявная. Запишем ее в следующем виде:

$$\hat{y}_n = \frac{1}{h + c\tau}\,(c\tau \hat{y}_{n-1} + h y_n). \tag{38}$$

Уменьшая индексы на единицу, получим выражение $\hat{y}_{n-1}$ через $\hat{y}_{n-2}$. Подставим его в правую часть (38). Продолжая процедуру

уменьшения индекса, приведем схему к явной форме:

$$\hat{y}_n = \frac{h}{h+c\tau} \sum_{l=0}^{\infty} \left(\frac{c\tau}{h+c\tau}\right)^l y_{n-l}. \tag{39}$$

Все коэффициенты здесь положительны; следовательно, схема (11) монотонна при любых τ и h.

Схема (12) линейна и имеет второй порядок точности на трижды непрерывно дифференцируемых решениях уравнения переноса. Из теоремы следует, что эта схема немонотонна.

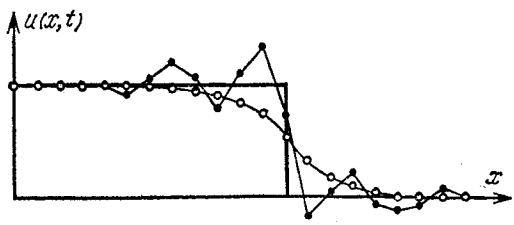


Рис. 68.

Различие монотонных и немонотонных схем особенно четко проявляется при расчетах задач с разрывными точными решениями (см. рис. 68, жирная линия — точное решение). Расчет по монотонной схеме (11) дает сглаженное разностное решение (кружки), а расчет по немонотонной схеме (12) — характерную «разболтку» (точки); эта «разболтка» не является неустойчивостью.

Сходную «разболтку» дают немонотонные схемы на быстропеременных решениях, особенно если шаг сетки не мал. Именно поэтому приходится решать подобные задачи при помощи монотонных схем, несмотря на их невысокую точность $O(\tau+h)$.

Наоборот, если решение достаточно гладкое и шаг сетки мал, то даже расчет по немонотонным схемам не нарушает монотонности решения. Например, для схем второго порядка точности монотонность разностного решения обычно сохраняется, если $|hu_{xx}/u_x| \lesssim 1$. В этих случаях для расчетов используют схемы точности $O(\tau^2+h^2)$ или более высокой.

Таким образом, фактически немонотонность проявляется на сетках со сравнительно большим шагом. Особенно сильно она сказывается при расчетах многомерных задач, ибо для них скорость или объем оперативной памяти даже лучших ЭВМ не позволяют брать малый шаг. В то же время расчет таких задач по монотонным схемам с погрешностью $O(\tau+h)$ дает хорошее качественное поведение разностного решения, но невысокую точность.

Теорема о монотонности доказана только для линейных схем. Были попытки построить нелинейные монотонные схемы второго порядка точности.

В частности, были предложены нелинейные монотонные схемы [70], имеющие на достаточно гладких решениях аппроксимацию $O(\tau^2+h^2)$ почти во всех точках; эффективный порядок точности этих схем, определенный на задачах-тестах, близок ко второму при большом h и стремится к первому при $h\to 0$. Эти схемы дают неплохие результаты при расчетах многомерных задач с быстропеременными решениями.

Другое перспективное направление связано с использованием схем третьего порядка точности. Как показали исследования, их фактическая немонотонность на разрывных решениях существенно слабее, чем у схем второго порядка точности: амплитуда «разболтки» меньше, и «разболтка» быстро затухает при удалении от разрыва.

7. Диссипативные схемы. Когда устойчивость линейных разностных схем исследуется методом разделения переменных, то для каждой гармоники определяют ее множитель роста ρ_q при переходе со слоя на слой. Отметим, что число пробных гармоник не бесконечно. Поскольку рассматривается разностная, т. е. дискретная, задача Фурье на сетке $\{x_n,\ 0\leqslant n\leqslant N\}$, то надо использовать только гармоники $\exp(2\pi iq/N)$, $0\leqslant q\leqslant N-1$, образующие полную систему по отношению к функциям, периодическим на этой сетке.

Схема устойчива, если

$$|\rho_q|\leqslant 1+\alpha_q\tau\approx\exp(\alpha_q\tau),$$

где α_q — не зависящие от h константы. Если хотя бы у одной гармоники $\alpha>0$, то устойчивость слабая. Если для всех гармоник $\alpha<0$, то схема заведомо хорошо обусловлена; но это требование слишком жесткое, и ему удовлетворяет очень мало схем. Рассмотрим более мягкое требование, также обеспечивающее хорошую обусловленность.

Схема обладает аппроксимационной вязкостью, если $\alpha_q<0$ *при* $q\neq 0$ *и* $\alpha_0=0$.

Это требование реализуется у многих схем. Например, схема (9) имеет множитель роста (15), который с учетом замечания об ограниченности числа гармоник принимает вид

$$\rho_q=1-\frac{c\tau}{h}(1-e^{-2\pi qi/N}),\qquad 0\leqslant q\leqslant N-1.$$

Легко проверить, что если $c\tau<h$, то $\rho_0=1$ и $|\rho_q|<1$ при $q=1,\ 2,\ \ldots,\ N-1$. Если $c\tau=h$, то для всех гармоник $|\rho_q|=1$. При $c\tau>h$ большинство гармоник неограниченно растет. Таким образом, схема (9) обладает аппроксимационной вязкостью при $c\tau<h$; это условие почти совпадает с условием устойчивости.

Аналогично доказывается, что схема (10) обладает аппроксимационной вязкостью при $c\tau>h$, а схема (11) — при любом соотношении шагов τ и h.

У схемы второго порядка точности (12) множитель роста (18) таков, что $|\rho_q|=1$ для всех гармоник. Следовательно, схема (12) не обладает аппроксимационной вязкостью.

При наличии аппроксимационной вязкости гармоники с $q \neq 0$ затухают, напоминая тем самым точное решение (9.7а) параболического уравнения. В точном решении параболического уравнения разрывы начальных данных сглаживаются со временем. Из рис. 68 было видно, что расчет по схеме с аппроксимационной вязкостью (11) приводит к аналогичному сглаживанию разрыва начальных данных, а расчет по схеме без аппроксимационной вязкости (12) — нет.

Понятие аппроксимационной вязкости применимо к линейным схемам. Для произвольных разностных схем, как линейных, так и нелинейных, можно ввести понятие первого дифференциального приближения.

Пусть дифференциальное уравнение $Au = f$ имеет решение, у которого непрерывны производные достаточно высокого порядка, и составлена разностная схема $A_h y = \varphi$ порядка аппроксимации p. Невязка этой схемы выражается через некоторые производные от решения $u(x)$, и ее можно представить в следующем виде:

$$\psi(x) \equiv Au - A_h u + \varphi - f = h^p Bu + o(h^p), \tag{40}$$

где B — некоторый дифференциальный оператор (обычно оператор B содержит производные, порядок которых на p превышает порядок старших производных дифференциального оператора A). Запишем равенство (40) двумя способами:

$$A_h u - \varphi = Au - f + O(h^p),$$

$$A_h u - \varphi = Au - h^p Bu - f + o(h^p).$$

Это означает, что разностная схема $A_h y = \varphi$ аппроксимирует дифференциальное уравнение $Au = f$ с порядком p и аппроксимирует уравнение $(A - h^p B)u = f$ с порядком выше p.

Первым дифференциальным приближением разностной схемы $A_h y = \varphi$ называют уравнение

$$(A - h^p B)u = f, \quad Bu = \lim_{h \to 0} [h^{-p} \psi(x)]. \tag{41}$$

Разностная схема аппроксимирует свое первое дифференциальное приближение более точно, чем исходное уравнение. Поэтому следует ожидать, что свойства разностного решения будут во многом напоминать свойства точных решений уравнения (41).

Пусть уравнение (41) является диссипативным, т. е. описывает какой-либо физический процесс с затуханием: теплопроводность (это сильное затухание), колебания в вязкой среде (слабое затухание) и т. д. Такие процессы приводят к более или менее быстрому сглаживанию разрывов начальных данных.

Обычно в этих случаях разностное решение тоже имеет сглаженный вид.

Наоборот, если уравнение (41) является недиссипативным, например чисто колебательным, то разрывы его решений не сглаживаются. В разностном решении при этом легко возникает слабо затухающая (или совсем не затухающая) «разболтка».

Примеры. Рассмотрим однородную разностную схему (9), полагая $f = \varphi = 0$. Ее невязка (13) принимает вид $\psi = (ch u_{xx} - \tau u_{tt})/2$. Учитывая, что для однородного уравнения переноса (3) выполняется условие $u_{tt} = c^2 u_{xx}$, преобразуем невязку к виду $\psi = c(h - c\tau) u_{xx}/2$. Отсюда легко получить первое дифференциальное приближение разностной схемы (9):

$$u_t = \frac{c}{2}(h - c\tau) u_{xx} - c u_x. \tag{42}$$

Если $c\tau < h$, то уравнение (42) относится к параболическому типу. Действительно, выше было показано, что расчет по разностной схеме (9) приводит к сглаживанию разрывов (если $c\tau < h$).

Случай $c\tau > h$ для уравнения (42) соответствует обратной задаче теплопроводности, которая относится к некорректно поставленным задачам. С этим обстоятельством связана неустойчивость схемы (9) при нарушении условия Куранта.

Рассмотрим однородную разностную схему (12). Если учесть, что для однородного уравнения переноса выполняется соотношение

$$u_{ttt} = -c u_{ttx} = c^2 u_{txx} = -c^3 u_{xxx},$$

то главный член невязки (17) этой схемы принимает вид $\psi = c(h^2 - c^2\tau^2) u_{xxx}/12$. Следовательно, ее первым дифференциальным приближением является уравнение

$$u_t + c u_x - \frac{1}{12} c (h^2 - c^2\tau^2) u_{xxx} = 0, \tag{43}$$

которое не содержит диссипативных членов. Действительно, из рис. 68 было видно, что схема (12) не сглаживает разрывы решения.

Если разностная схема обладает аппроксимационной вязкостью или ее первое дифференциальное приближение является уравнением с диссипативными членами, то схему называют *диссипативной*. Обычно в расчетах по таким схемам разболтки не возникает или она невелика; поэтому понятие диссипативности плодотворно используется при качественном анализе разностных схем. Однако это понятие не является строгим, и полученные при его помощи выводы надо проверять другими методами.

§ 2. Квазилинейное уравнение

1. Сильные и слабые разрывы. Решение линейного уравнения переноса может иметь разрывы только в том случае, если они содержатся в начальных или граничных данных. В квазилинейном уравнении даже при непрерывных и достаточно гладких начальных данных могут возникать разрывы решения. Характер этих разрывов удобно исследовать на простейшем квазилинейном уравнении переноса

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0, \quad x,\ t,\ u>0, \tag{44}$$

которым мы и ограничимся в данном параграфе. Оно напоминает линейное уравнение переноса, в котором роль скорости переноса играет величина самого решения $u(x, t)$.

Полная постановка задачи при знакопеременной скорости сложна; мы рассмотрим только наиболее важный случай $u>0$. Тогда начальные и граничные значения решения, заданные на положительных полуосях координат x, t, определяют решение в первом квадранте. Эти значения переносятся по характеристикам $x-ut=\text{const}$. Рассмотрим характер решения при четырех основных типах начальных данных.

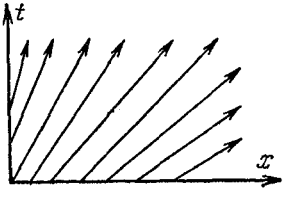


Рис. 69.

Первый случай. Начальные и краевые значения — непрерывные функции, причем $u(x, 0)$ монотонно не убывает, $u(0, t)$ монотонно не возрастает и они непрерывно согласованы в начале координат.

Наклон (тангенс угла наклона) характеристик в каждой точке плоскости (x, t) равен $1/u(x, t)$. В данном случае наклон монотонно и непрерывно убывает слева направо. Поэтому первый квадрант всюду плотно покрыт характеристиками (рис. 69), причем через каждую его точку проходит одна и только одна характеристика. Эта характеристика переносит в данную точку граничное значение. Решение однозначно определено и непрерывно во всем первом квадранте. Если краевые значения гладки (и согласованы в начале координат), то решение также будет гладким.

Второй случай. Краевые значения монотонны указанным выше образом, но имеют разрывы. Для простоты положим $u(0, t)=a$, $u(x, 0)=a$ при $x<x_0$, $u(x, 0)=b>a$ при $x>x_0$, так что разрыв не нарушает предыдущее условие монотонности.

Левее разрыва характеристики на плоскости (x, t) имеют наклон $1/a$, а правее разрыва — меньший наклон $1/b$ (рис. 70, *а*).

Проведем обе характеристики из точки разрыва начальных данных; на рисунке они показаны жирными стрелками. Левее левой и правее правой из них через каждую точку плоскости проходит одна и только одна характеристика, т. е. решение определено и единственно. А между ними нет ни одной характеристики и решение не определено.

Потребуем корректности задачи, т. е. устойчивости решения относительно бесконечно малых возмущений начальных данных. Это позволит нам доопределить решение. Сгладим разрыв начальных данных, заменив его непрерывным монотонным переходом на бесконечно узком интервале. Тогда в пустом угле появится «веер» характеристик и наклон каждой характеристики определит значение решения на ней (рис. 70, *б*).

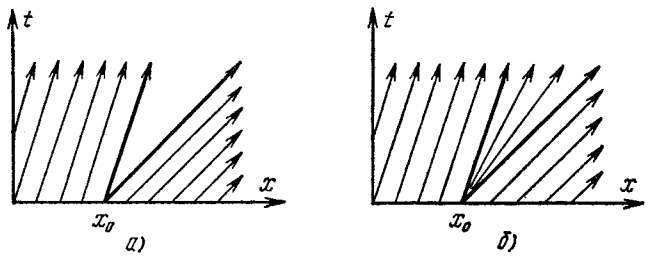


Рис. 70.

Легко видеть, что доопределенное решение будет иметь следующий вид:

$$\begin{aligned} &u(x,\ t)=a && \text{при} \quad x-x_0\leqslant at,\\ &u(x,\ t)=(x-x_0)/t && \text{при} \quad at\leqslant x-x_0\leqslant bt,\\ &u(x,\ t)=b && \text{при} \quad bt\leqslant x-x_0. \end{aligned} \tag{45}$$

Поэтому оно непрерывно на всей плоскости, кроме точки $x=x_0$, $t=0$. Значит, такой разрыв начальных данных сглаживается со временем. Но след разрыва остается: на жирных характеристиках производные решения будут разрывны. Во всех остальных точках решение будет гладким, если начальные данные были гладкими.

Разрыв производных называют *слабым разрывом* решения. Слабые разрывы квазилинейного уравнения переноса распространяются по характеристикам, как и в линейном уравнении переноса.

Третий случай. Пусть нарушено данное выше условие монотонности. Опять положим $u(x,\ 0)=a$ при $x<x_0$, $u(x,\ 0)=b$ при $x>x_0$, но теперь потребуем, чтобы $a>b>0$. Тогда характеристики будут иметь вид, изображенный на рис. 71.

В угле, образованном жирными характеристиками, через каждую точку проходят две характеристики, приносящие в нее разные значения функции! Вне этого угла решение однозначно определено, а внутри угла оно неоднозначно.

В этом случае непрерывное решение построить не удается. Сглаживание разрыва начальных данных не помогает: ход характеристик на некотором расстоянии от точки $x=x_0$, $t=0$ все равно не меняется, так что неоднозначность остается. Значит, однозначное решение должно быть разрывным, т. е. оно будет обобщенным решением дифференциального уравнения.

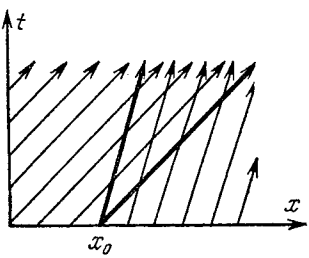


Рис. 71.

Обобщенное решение удовлетворяет некоторому интегральному уравнению, которое получается из определенной дивергентной формы записи данного дифференциального уравнения. Разные дивергентные формы записи одного и того же уравнения приводят к разным разрывным решениям, хотя гладкие решения для всех дивергентных форм одинаковы. Дивергентная форма, соответствующая физическому закону сохранения, определяет правильное решение (его называют также *допустимым*).

Уравнение (44) не имеет физического смысла, и естественного закона сохранения для него нет. Постулируем такую дивергентную форму:

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)=0. \tag{46}$$

Будем искать решение, имеющее единственный разрыв. Пусть наклон линии разрыва соответствует скорости D, т. е. разрыв бежит, как волна. По поведению характеристик видно (рис. 72), что искомое решение имеет вид

$$\begin{aligned} &u(x,\ t)=a \text{ при } x-x_0<Dt,\\ &u(x,\ t)=b \text{ при } x-x_0>Dt. \end{aligned} \tag{47}$$

Проинтегрировав (46) по площади прямоугольника со сторонами τ и $h=D\tau$, получим

$$\int(\hat{u}-u)\,dx+\frac{1}{2}\int(u^2_{\text{прав}}-u^2_{\text{лев}})\,dt=(a-b)\,h+\frac{1}{2}(b^2-a^2)\,\tau=0.$$

Отсюда скорость распространения разрыва равна

$$D=\frac{1}{2}(a+b), \tag{48}$$

Разрыв самого решения называют *сильным разрывом* (а в газодинамике — ударной волной). Сильный разрыв квазилинейного уравнения распространяется не по характеристике. В теории квазилинейных уравнений доказывается, что только такое обобщенное решение устойчиво относительно малых возмущений начальных данных.

Четвертый случай, когда функция $u(x, 0)$ непрерывна, но убывает на каком-то интервале, сводится к третьему. Попрежнему пересечение характеристик приводит к образованию сильного разрыва (рис. 73). Местная скорость разрыва будет определяться по формуле типа (48) приносимыми в данную точку значениями решения и уже не будет постоянной. Существенно,

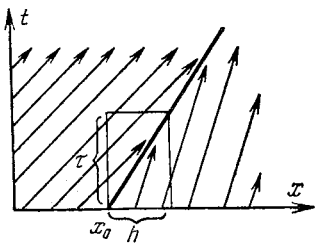


Рис. 72.

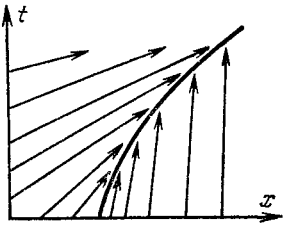


Рис. 73.

что здесь при непрерывных и гладких начальных данных с течением времени возникают сильные разрывы решения. Число разрывов со временем тоже может измениться.

Замечание 1. Если вместо (46) мы постулируем другой закон сохранения, например:

$$\frac{\partial}{\partial t}(\ln u) + \frac{\partial u}{\partial x} = 0, \tag{49}$$

то скорость ударной волны изменится. Но для слабых ударных волн, на которых решение мало меняется, скорость ударной волны будет отличаться от (48) в $1 + O(\varepsilon^2)$ раз, где

$$\varepsilon = (b - a)/(b + a),$$

т. е. изменится очень мало.

Замечание 2. Разрывные решения линейных уравнений можно рассматривать как предел последовательности непрерывных и гладких решений. Для квазилинейных уравнений это сделать не удается.

2. Однородные схемы. Выше встречались случаи, когда на недостаточно гладких решениях (т. е. имеющих малое число непрерывных производных) порядок аппроксимации и порядок

точности разностной схемы был ниже, чем на более гладких решениях. Особенно сильно ухудшается точность расчета, если искомое решение содержит сильные или слабые разрывы; некоторые разностные схемы при этом приводят даже к грубо ошибочным результатам.

Удовлетворительные результаты на таких решениях дают два типа схем — однородные схемы (которые не надо путать с линейными однородными схемами для линейных однородных уравнений) и схемы с явным выделением особенностей решения.

Однородные схемы (называемые также схемами сквозного счета) более просты и широко распространены в практике численных расчетов. В этих схемах шаблон и разностные аналоги производных выбираются так, чтобы нужная аппроксимация была обеспечена всюду, в том числе и на особенностях решения. Поэтому весь расчет ведется по однотипным разностным уравнениям без явного выделения особенностей.

Например, из рис. 68 видно, что схема (11) позволяет рассчитывать перенос разрывных начальных данных без явного выделения этого разрыва.

Однородные схемы не громоздки, требуют умеренного объема вычислений, и каждая хорошая схема пригодна для широкого класса задач. Программы для ЭВМ, составленные на их основе, также позволяют без заметных переделок рассчитывать широкий круг задач. Зато точность расчета по однородным схемам обычно ниже, чем в схемах с выделением особенностей.

По однородным схемам успешно проводят расчеты даже таких сложных задач, как задачи многомерной магнитной газодинамики, в которых возникает большое число ударных, тепловых и других волн, являющихся разрывами.

Схемы с выделением особенностей. В них каждую особенность решения выделяют и детально описывают. В промежутках между особенностями решение непрерывно и достаточно гладко; в этих промежутках дифференциальное уравнение аппроксимируют разностной схемой.

Уравнения, описывающие особенности, служат своеобразными внутренними краевыми условиями, связывающими между собой разностные уравнения в соседних промежутках.

Особенности решения могут быть связаны с разрывами или нарушением гладкости начальных данных и коэффициентов уравнения, с возникновением ударных волн, с образованием слабых разрывов при столкновении ударной волны с какой-либо особенностью решения. Число особенностей с течением времени может меняться. К каждому типу особенностей нужен индивидуальный подход. Очевидно, явно учесть все особенности можно только в наиболее простых задачах.

Схемы с выделением даже небольшого количества особенностей очень громоздки. Они нестандартны, т. е. каждый сравнительно узкий класс задач требует разработки своей схемы расчета. Но зато их точность существенно выше, чем точность прочих схем. Поэтому их используют в задачах, имеющих мало существенных особенностей и требующих особенно высокой точности расчета. Такие схемы мы рассматривать не будем.

3. Псевдовязкость. Основную трудность для вычислений по разностным схемам представляют сильные разрывы решения. Эффективный прием расчета задач с разрывными решениями заключается в следующем. Подберем такую «малую» добавку к исходному уравнению, чтобы его разрывные решения превратились в непрерывные и достаточно гладкие. Тогда составить разностную схему для численного расчета этих гладких решений уже несложно.

Гладкие решения присущи уравнениям с диссипативными членами типа вязкого трения. Поэтому добавляемый в исходное уравнение член должен играть роль вязкости. Его называют *псевдовязкостью*, а также *искусственной* или *математической вязкостью*.

Рассмотрим подробно указанный способ на примере квазилинейного уравнения переноса (44). Заменим его следующим уравнением:

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+\frac{\varepsilon^2}{2}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)^2=0, \tag{50}$$

где последний член является псевдовязкостью, а коэффициент ε^2 мал.

Очевидно, на дважды непрерывно дифференцируемых функциях последний член при малых ε невелик, так что для всех достаточно гладких решений исходного уравнения (44) найдутся близкие к ним гладкие решения уравнения (50).

Выясним, нет ли среди гладких решений уравнения (50) такого, которое напоминало бы ударную волну (47):

$$u(x,\ t)=\begin{cases} a & \text{при } x<Dt,\\ b<a & \text{при } x>Dt,\end{cases}$$

движущуюся со скоростью $D=(a+b)/2$. Будем искать автомодельное решение в виде бегущей волны

$$u_\varepsilon(x,\ t)=f(x-Dt).$$

Подставляя его в (50), получим

$$(\varepsilon^2 f''+f-D)f'=0.$$

Приравнивая каждый из сомножителей нулю, получим два типа решений:

$$f_1=\text{const},$$
$$f_2=D+\text{const}\cdot\sin\frac{x-Dt}{\varepsilon}.$$

Из них можно сконструировать решение, похожее на размытую

волну шириной $\sim\varepsilon$:

$$u_\varepsilon(x,\ t)=\begin{cases} a & \text{при } x-Dt<-\frac{\pi\varepsilon}{2},\\ \frac{a+b}{2}-\frac{a-b}{2}\sin\frac{x-Dt}{\varepsilon} & \text{при } -\frac{\pi\varepsilon}{2}<x-Dt<\frac{\pi\varepsilon}{2},\\ b & \text{при } \frac{\pi\varepsilon}{2}<x-Dt. \end{cases} \quad (51)$$

Это решение не только непрерывно, но даже имеет кусочно-непрерывную вторую производную. При $\varepsilon\to 0$ оно переходит в ударную волну (47).

Таким образом, и гладкие и разрывные решения исходного уравнения (44) можно рассматривать как предел соответствующих гладких решений уравнения (50) при $\varepsilon\to 0$. Значит, вместо численного решения квазилинейного уравнения переноса можно численно решать уравнение (50) при достаточно малом ε. Решения последнего уравнения гладки, и их можно находить при помощи обычных однородных разностных схем.

Замечание 1. Коэффициент псевдовязкости обычно связывают с шагом сетки. Например, если в уравнении (50) положить

$$\varepsilon=\nu h,\ \nu=\text{const}, \quad (52)$$

то любой сильный разрыв «размазывается» на одно и то же число $\pi\nu\approx 3\nu$ интервалов сетки. Тогда при $h\to 0$ уравнение с псевдовязкостью (50) автоматически переходит в исходное уравнение (44), а сглаженная ударная волна (51) — в ударную волну (47).

Пример. Составим простейшую (далеко не лучшую) разностную схему для уравнения (50), а тем самым — и для уравнения (44); сетку для простоты предполагаем равномерной:

$$\frac{1}{\tau}(\hat{y}_n-y_n)+\frac{1}{h}y_n(y_n-y_{n-1})+\frac{\varepsilon^2}{2h^3}[(y_{n+1}-y_n)^2-(y_n-y_{n-1})^2]=0. \quad (53)$$

Это явная схема, так что разностное решение существует и единственно. Не проводя полного исследования схемы, определим только условие ее устойчивости.

Схема (53) нелинейна, поэтому сначала линеаризуем ее и получим уравнение для роста погрешности δy:

$$\frac{1}{\tau}(\delta\hat{y}_n-\delta y_n)+\frac{1}{h}\delta y_n(y_n-y_{n-1})+\frac{1}{h}y_n(\delta y_n-\delta y_{n-1})+$$
$$+\frac{\varepsilon^2}{2h^3}[(\delta y_{n+1}-\delta y_{n-1})(y_{n+1}-2y_n+y_{n-1})+$$
$$+(y_{n+1}-y_{n-1})(\delta y_{n+1}-2\delta y_n+\delta y_{n-1})]=0. \quad (54)$$

Коэффициенты при δy переменные; применяя принцип «замороженных» коэффициентов, будем считать их постоянными. Попутно произведем замены $y_n - y_{n-1} \approx h u_x$ и т. д. Тогда (54) примет вид

$$\frac{1}{\tau}(\delta \hat{y}_n - \delta y_n) + u_x \delta y_n + \frac{u}{h}(\delta y_n - \delta y_{n-1}) + \frac{\varepsilon^2}{2h} u_{xx}(\delta y_{n+1} - \delta y_{n-1}) + \\ + \frac{\varepsilon^2}{h^2} u_x (\delta y_{n+1} - 2\delta y_n + \delta y_{n-1}) = 0. \quad (55)$$

Рассматривая рост ошибки, имеющей вид $\delta y_n \sim \exp(iqx_n)$, и делая в (55) стандартную подстановку:

$$\delta y_n = e^{iqx}, \quad \delta y_{n\pm 1} = e^{iq(x\pm h)}, \quad \delta \hat{y}_n = \rho \delta y_n,$$

определим множитель роста гармоники:

$$\rho_q = 1 - \frac{u\tau}{h}(1 - e^{-iqh}) - \tau\left[u_x + i\frac{\varepsilon^3}{h} u_{xx} \sin qh - \frac{4\varepsilon^2}{h^2} u_x \sin^2 \frac{qh}{2}\right]. \quad (56)$$

Если согласно (52) выбран коэффициент псевдовязкости $\varepsilon \sim h$, то величина в квадратных скобках ограничена равномерно по шагу h. Тогда последний член в (56) есть $O(\tau)$ и не нарушает устойчивости. Первые же два члена аналогичны множителю (15) и приводят к условию устойчивости типа Куранта:

$$u(x, t)\tau \leqslant h, \quad (57)$$

где роль скорости играет величина решения u (напомним, однако, что для нелинейных схем этот способ исследования устойчивости является не строгим, а лишь правдоподобным).

Схема (53) является примером однородной схемы для расчета задач с произвольным числом движущихся разрывов, причем число разрывов может меняться с течением времени. Заметим, что для обеспечения хорошей точности расчета зона сглаживания разрыва должна быть небольшой (3—5 интервалов) и сумма зон сглаживания всех разрывов должна быть мала по сравнению с общим числом узлов сетки N. Тем самым, фактически общее число разрывов не может быть большим.

Замечание 2. Псевдовязкость вида (50) обеспечивает сходимость к тем обобщенным решениям уравнения (44), которые соответствуют дивергентной форме (46). Для другого уравнения или даже для другой дивергентной формы того же уравнения эта псевдовязкость, вообще говоря, непригодна.

Замечание 3. Псевдовязкость (50), называемая квадратичной, имеет один заметный недостаток: не все решения уравнения (50) являются дважды дифференцируемыми. В самом деле, нетрудно проверить, что кусочно-гладкое решение (45) также удовлетворяет этому уравнению. На таких решениях однородные

схемы, рассчитанные обычно на дважды или трижды непрерывно дифференцируемые функции, имеют пониженный порядок аппроксимаций.

Этот недостаток устраняется, если использовать для уравнения (44) другую псевдовязкость, называемую линейной:

$$u_t + uu_x = \varepsilon u_{xx}, \qquad \varepsilon = O(h). \tag{58}$$

Уравнение (58) напоминает уравнение теплопроводности, все решения которого многократно дифференцируемы. Его нетрудно исследовать аналогично уравнению (50). Однако линейная псевдовязкость тоже не лишена недостатков.

4. Ложная сходимость. На практике для нелинейных уравнений и схем редко удается строго доказать сходимость; например, сходимость разностных схем для уравнений газодинамики не доказана. Поэтому зачастую пользуются следующими соображениями. Проверим локальную аппроксимацию схемы и затем на численных расчетах со сгущением сеток убедимся, что разностное решение при $h \to 0$ сходится к какой-то предельной функции. Поскольку нет расходимости, то расчет устойчив, а из устойчивости и аппроксимации следует сходимость к решению исходной задачи.

Эти рассуждения справедливы, если точное решение достаточно гладко. Если же решение имеет сильные или слабые разрывы, то локальной аппроксимации в точках разрыва нет и предыдущие рассуждения могут привести к неверному результату.

Пример. Приведем разностную схему, которая сходится, но не к точному решению.

Возьмем схему (53) для уравнения с псевдовязкостью (50) и тем самым для квазилинейного уравнения переноса (44); положим в ней $\varepsilon = 0$, т. е. выбросим псевдовязкость. Тогда схема примет вид

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{1}{h} y_n (y_n - y_{n-1}) = 0, \tag{59}$$

напоминающий явную схему (9) для линейного уравнения переноса. Проведем по схеме (59) расчет движения сильного разрыва (47). Пусть начальные данные таковы, что

$$\begin{aligned} y_n &= a \quad \text{при} \quad n \leqslant n_0 - 1, \\ y_n &= b \quad \text{при} \quad n \geqslant n_0. \end{aligned} \tag{60}$$

Выберем шаг по времени $\tau = h/b$. Подставляя (60) в (59), нетрудно убедиться, что на следующем слое разностное решение будет равно

$$\begin{aligned} \hat{y}_n &= a \quad \text{при} \quad n \leqslant n_0, \\ \hat{y}_n &= b \quad \text{при} \quad n \geqslant n_0 + 1. \end{aligned} \tag{61}$$

Это значит, что разрыв продвинулся за один временной шаг ровно на один интервал сетки и сохранил свою форму. Очевидно, так же он будет двигаться и на всех других шагах по времени.

Таким образом, в этом расчете сильный разрыв будет двигаться без сглаживания, точно сохраняя форму, но с неправильной скоростью

$$D^* = h/\tau = b \neq (a+b)/2.$$

Значит, при $b\tau = h \to 0$ разностное решение (60)—(61) сходится к предельной функции

$$u(x,\ t) = \begin{cases} a & \text{при} \quad x - x_0 < bt, \\ b & \text{при} \quad x - x_0 > bt, \end{cases} \tag{62}$$

которая отлична от точного решения (47).

Таким образом, для задач с разрывными или недостаточно гладкими решениями (а также при разрывных или недостаточно гладких коэффициентах уравнения) визуально наблюдаемая сходимость разностного решения к пределу при $\tau \to 0$, $h \to 0$ может оказаться ложной.

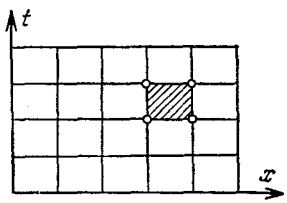


Рис. 74.

5. Консервативные схемы. Ложной сходимости можно избежать, используя консервативные схемы. Эти схемы составляют методом баланса, исходя из физических законов сохранения и соблюдая дополнительное правило, описанное ниже.

Сначала разберем законы сохранения на примере уравнения (44). Запишем ту дивергентную форму этого уравнения (46), которая в п. 1 была условно принята за правильную:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0. \tag{63}$$

Выбирая отдельную ячейку сетки (рис. 74) и интегрируя по ней уравнение (63), получим точное интегральное соотношение

$$\int_{x_{n-1}}^{x_n} (u^{m+1} - u^m)\, dx + \frac{1}{2} \int_{t_m}^{t_{m+1}} (u_n^2 - u_{n-1}^2)\, dt = 0. \tag{64}$$

Уравнение (63) можно проинтегрировать не по отдельной ячейке, а по всей области

$$G = [x_0 \leqslant x \leqslant x_N] \times [t_0 \leqslant t \leqslant t_M]$$

и получить аналогичное интегральное соотношение:

$$\int_{x_0}^{x_N} (u^M - u^0)\,dx + \frac{1}{2}\int_{t_0}^{t_M} (u_N^2 - u_0^2)\,dt = 0. \tag{65}$$

Это соотношение напоминает физические законы сохранения: первый интеграл дает изменение $\int u\,dx$ за истекшее время, а второй есть разность потоков $^1/_2\int u^2\,dt$ через правую и левую границу. Соотношение (65) является законом сохранения для данной задачи.

Очевидно, соотношение (64) является законом сохранения для каждой отдельной ячейки; оно содержит потоки и другие величины на границах этой ячейки. Просуммируем (64) по всем ячейкам области G:

$$\sum_{n=1}^{N}\sum_{m=0}^{M-1}\left[\int_{x_{n-1}}^{x_n} (u^{m+1} - u^m)\,dx + \frac{1}{2}\int_{t_m}^{t_{m+1}} (u_n^2 - u_{n-1}^2)\,dt\right] = 0. \tag{66}$$

Легко видеть, что интегралы по тем границам ячеек, которые лежат внутри G, попарно уничтожаются; остаются только интегралы по наружной границе, и (66) совпадает с (65). Иными словами, *закон сохранения во всей области есть точное следствие закона сохранения в отдельных ячейках.*

Не всякая разностная схема обладает таким свойством. Например, возьмем схему с ложной сходимостью (59), умножим обе части на τh и просуммируем по всем ячейкам:

$$\sum_{n=1}^{N}\sum_{m=0}^{M-1} \tau h\left[\frac{1}{\tau}(y_n^{m+1} - y_n^m) + \frac{1}{h}\,y_n^m (y_n^m - y_{n-1}^m)\right] = 0. \tag{67}$$

Преобразуем второе слагаемое в квадратных скобках:

$$y_n (y_n - y_{n-1}) = \frac{1}{2}(y_n^2 - y_{n-1}^2) + \frac{1}{2}(y_n - y_{n-1})^2.$$

Тогда (67) легко привести к следующему виду:

$$\sum_{n=1}^{N} h(y_n^M - y_n^0) + \frac{1}{2}\sum_{m=0}^{M-1} \tau\left[(y_N^m)^2 - (y_0^m)^2\right] + \Delta = 0, \tag{68}$$

где

$$\Delta = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=0}^{M-1} \tau (y_n^m - y_{n-1}^m)^2 > 0. \tag{69}$$

Первая и вторая суммы в (68) являются разностными аналогами интегралов в (65); они не содержат значений y_n^m во внутренних узлах области G. Но остается еще третья (двойная) сумма (69), содержащая внутренние узлы неустранимым образом и заведомо не равная нулю.

Поэтому при расчете по схеме (59) разностный закон сохранения во всей области G нарушается на величину Δ. Такие схемы называют *неконсервативными*, а величину Δ называют *дисбалансом* схемы.

Построим *консервативную* схему, т. е. такую, у которой дисбаланс равен нулю. Для этого в интегральном соотношении (64) аппроксимируем интегралы линейными квадратурными формулами. Если, для определенности, воспользоваться формулой прямоугольников с теми же узлами, что в предыдущей схеме, то получим явную схему следующего вида:

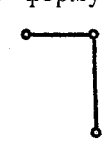


Рис. 75.

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{1}{2h}(y_n^2 - y_{n-1}^2) = 0. \qquad (70)$$

Суммирование (70) по всем ячейкам дает именно две первые суммы в (68), и дисбаланса не возникает.

Выбирая другие шаблоны, можно построить различные консервативные схемы для уравнения (44). Например, если вычислить интегралы в (64) по шаблону рис. 75, то получим неявную схему

$$\frac{1}{\tau}(\hat{y}_n - y_n) + \frac{1}{2h}(\hat{y}_n^2 - \hat{y}_{n-1}^2) = 0. \qquad (71а)$$

Это — схема бегущего счета, и для выполнения вычислений ее удобно переписать в следующем виде:

$$\hat{y}_n = -\frac{h}{\tau} + \sqrt{\frac{h^2}{\tau^2} + \frac{2h}{\tau} y_n + \hat{y}_{n-1}^2}; \qquad (71б)$$

здесь из двух корней квадратного уравнения (71а) согласно условию $u(x, t) > 0$ выбран положительный. Суммируя (71а) по всем ячейкам, получим разностный закон сохранения:

$$\sum_{n=1}^{N} h(y_n^M - y_n^0) + \frac{1}{2} \sum_{m=1}^{M} \tau[(y_N^m)^2 - (y_0^m)^2] = 0. \qquad (72)$$

Вторая сумма немного отличается от второй суммы (68), но это отличие несущественно. Дисбаланс отсутствует, так что схема (71) консервативна.

Схема (71) любопытна во многих отношениях. Она является схемой сквозного счета; хотя ее сходимость строго не доказана, она успешно используется для расчета сильных разрывов даже в отсутствие псевдовязкости (по-види-

мому, это связано с наличием достаточно большой аппроксимационной вязкости схемы). Схема монотонна. Есть обобщения этой схемы, сохраняющие все ее хорошие свойства и существенно повышающие точность расчета [70].

Интерес к таким схемам объясняется тем, что многие изложенные здесь идеи удается перенести на случай газодинамики и других сложных и важных задач.

Консервативные схемы выражают закон сохранения на сетке, т. е. они качественно похожи на исходное интегральное уравнение. Неконсервативные схемы этим свойством не обладают. Поэтому, по сравнению с неконсервативными схемами, консервативные схемы обычно приводят к существенному улучшению точности расчета как разрывных, так и гладких решений.

Построены схемы, одновременно удовлетворяющие большому числу различных физических законов сохранения (см. [34]). Эти схемы, названные *полностью консервативными*, оказались полезными в задачах магнитной газодинамики, физики разреженной плазмы и ряде других.

Таким образом, понятие консервативности широко используется при составлении и исследовании разностных схем.

Заметим, однако, что различные полезные свойства схем (консервативность, монотонность, высокий порядок аппроксимации) нередко противоречат друг другу, так что может не существовать схемы, одновременно удовлетворяющей всем этим требованиям. Кроме того, не для всех классов уравнений консервативность является необходимым условием сходимости, и составлено немало хороших, хотя и неконсервативных схем.

ЗАДАЧИ

1. Получить для схемы (9) априорную оценку точности.

2. Исследовать сходимость схем (10) и (11).

3. Получить невязки схем (10) и (11) и сравнить их между собой и с невязкой (13) схемы (9).

4. Записать схемы (10)—(11) для случая неравномерной сетки.

5. Исследовать устойчивость схемы (25) методом разделения переменных.

6. Показать, что схема (25) имеет аппроксимацию $O(\tau + h^2)$.

7. Проверить аппроксимацию и устойчивость схемы (29) для двумерного уравнения переноса (27а).

8. Составить для двумерного уравнения переноса (27а) явную схему, аналогичную схеме (9), и исследовать ее устойчивость.

9. Составить для двумерного уравнения переноса (27а) симметричную схему, аналогичную схеме (12), и исследовать ее.

10. Показать, что схема (32) для уравнения переноса с поглощением (30) сохраняет положительность решения (т. е. разностное решение положительно, если положительны начальные данные) при любом $b \geqslant 0$, если $c\tau \leqslant h$.

11. Исследовать, монотонна ли схема (10) и при каком условии.

12. Определить скорость ударной волны, соответствующую дивергентной форме (49) записи уравнения (44). Сравнить эту скорость со скоростью (48) и убедиться в правильности замечания 1 к § 2, п. 1.

13. Исследовать квазилинейное уравнение переноса с линейной псевдовязкостью (58); показать, что среди его решений есть сглаженная ударная волна

$$u_\varepsilon(x,\ t)=\frac{a+b\exp\left[\dfrac{a-b}{2\varepsilon}(x-x_0-Dt)\right]}{1+\exp\left[\dfrac{a-b}{2\varepsilon}(x-x_0-Dt)\right]},\qquad D=\frac{1}{2}(a+b).$$

14. Для уравнения с линейной псевдовязкостью (58) составить какую-нибудь разностную схему и исследовать ее устойчивость.

15. Исследовать устойчивость нелинейных схем (70) и (71).

16. Исследовать аппроксимацию схем (70) и (71) на дважды непрерывно дифференцируемых решениях.

17. Доказать монотонность схемы (71).

ГЛАВА XI

ПАРАБОЛИЧЕСКИЕ УРАВНЕНИЯ

Глава XI посвящена численному решению уравнений параболического типа. В § 1 рассмотрены одномерные задачи, начиная от случая простейшего уравнения с постоянными коэффициентами и кончая квазилинейным уравнением с разрывными коэффициентами в криволинейных координатах. Разобраны основные разностные схемы, используемые для решения таких задач.

В § 2 обсуждены принципиальные трудности, возникающие при переходе к случаю многих измерений; изложены продольно-поперечная прогонка, дающая хорошие результаты при решении задач с двумя пространственными переменными, и локально-одномерный метод, пригодный при любом числе измерений.

§ 1. Одномерные уравнения

1. Постановки задач. К параболическим уравнениям приводят задачи теплопроводности, диффузии и ряд других. Типичной полной постановкой одномерной задачи является, например, первая краевая задача для случая линейной теплопроводности в однородной среде:

$$\left.\begin{array}{c} u_t(x,\ t) = k u_{xx}(x,\ t) + f(x,\ t), \\ k = \text{const} > 0,\ 0 < x < a,\ 0 < t \leqslant T, \end{array}\right\} \tag{1а}$$

$$u(x,\ 0) = \mu(x), \qquad 0 \leqslant x \leqslant a, \tag{1б}$$

$$u(0,\ t) = \mu_1(t), \qquad u(a,\ t) = \mu_2(t), \qquad 0 \leqslant t \leqslant T. \tag{2}$$

Она включает в себя задание самого уравнения, начальных данных на некотором отрезке и краевых условий на обоих концах этого отрезка.

Наиболее хорошо изучены линейные задачи, в которых и уравнение и краевые условия линейны. Для таких задач рассматривают три типа краевых условий. Условия первого рода (2) применительно к уравнению теплопроводности означают, что на границах задана зависимость температуры u от времени. Условия второго рода

$$u_x(0,\ t) = \mu_1(t), \qquad u_x(a,\ t) = \mu_2(t) \tag{3}$$

соответствуют заданию тепловых потоков через границы. Условия третьего рода

$$u(0,\ t)+\alpha_1 u_x(0,\ t)=\mu_1(t),\qquad u(a,\ t)+\alpha_2 u_x(a,\ t)=\mu_2(t) \tag{4}$$

возникают, если на границах имеется линейный (ньютоновский) теплообмен с окружающей средой. Для задачи (1) с краевыми условиями (2), (3) или (4) корректность постановки доказана (см., например, [40]).

Часто встречаются и нелинейные задачи. Например, в главе IX было рассмотрено квазилинейное уравнение (9.9), связанное с задачами теплопроводности в плазме. Краевые условия также могут быть нелинейными; так, остывание черного тела за счет излучения с поверхности приводит к краевому условию

$$(u^4+\alpha u_x)_{x-0}=0.$$

В главе IX отмечалась важная качественная особенность решений параболических уравнений: разрывы начальных данных сглаживаются с течением времени.

Другое любопытное свойство следует из вида функции точечного источника на бесконечной прямой для линейного уравнения (1) *):

$$G(x,\ \xi;\ t)=\frac{1}{\sqrt{4\pi kt}}\exp\left[-\frac{(x-\xi)^2}{4kt}\right]. \tag{5}$$

Если $t>0$, то $G(x,\ \xi,\ t)>0$ при сколь угодно больших x и ξ. Следовательно, при $t>0$ температура в каждой точке x зависит от начальных данных во всех точках ξ бесконечной прямой, сколь угодно удаленных от x. Поэтому говорят, что в случае линейной теплопроводности скорость распространения тепла и область влияния бесконечны.

Строго говоря, параболическое уравнение лишь приближенно описывает процесс теплопроводности. На самом деле скорость распространения тепла конечна и не превышает (при молекулярной или электронной теплопроводности) тепловой скорости частиц. Влияние же удаленных точек, как видно из выражения для функции Грина (5), ослабевает очень быстро; отрезку времени Δt соответствует характерная зона влияния $\Delta x\sim\sqrt{k\Delta t}$.

Эти соображения надо учитывать при построении разностных схем, поскольку, как отмечалось в главе X, правильный учет зоны влияния необходим для устойчивости схемы.

2. Семейство неявных схем. Рассмотрим простейшие, но хорошие разностные схемы для уравнения теплопроводности (1)

*) Вывод этой формулы см., например, в [40].

с постоянным коэффициентом:

$$u_t = ku_{xx} + f, \quad 0 < x < a, \quad 0 < t \leqslant T \quad (k = \text{const} > 0).$$

Возьмем в области $G = [0 \leqslant x \leqslant a] \times [0 \leqslant t \leqslant T]$ прямоугольную сетку (рис. 76), для простоты равномерную, с шагами h и τ. Выберем шеститочечный шаблон, изображенный на рисунке жирными линиями, и составим на нем следующую двуслойную схему:

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{k\sigma}{h^2}(\hat{y}_{n-1} - 2\hat{y}_n + \hat{y}_{n+1}) + \frac{k(1-\sigma)}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + \varphi_n,$$
$$1 \leqslant n \leqslant N-1 \quad (\sigma = \text{const}). \tag{6а}$$

Здесь записано меньше уравнений, чем имеется неизвестных $\hat{y}_n$. Недостающие два уравнения находим из краевых условий; например, краевые условия первого рода (2) дают соотношения

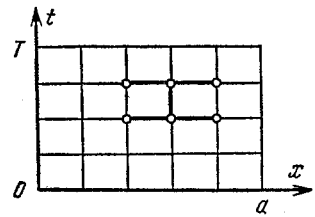


Рис. 76.

$$\hat{y}_0 = \mu_1(t_{m+1}), \quad \hat{y}_N = \mu_2(t_{m+1}). \tag{6б}$$

В качестве правой части φ_n часто выбирают значение $\varphi_n = f(x_n, t_m + \tau/2)$.

Схема (6а, б) содержит параметр σ; он является весовым множителем при пространственной производной с верхнего слоя. Поэтому (6а, б) есть однопараметрическое семейство схем. Меняя вес σ, можно добиться улучшения тех или иных свойств схемы.

Исследуем схему (6а, б).

Существование решения и его вычисление. Если $\sigma = 0$, то схема (6) переходит в рассмотренную ранее явную схему (9.18). Разностное решение при этом легко вычисляется, его существование и единственность очевидны.

Если $\sigma \neq 0$, то схема (6) существенно неявна. Перепишем ее в следующем виде:

$$\hat{y}_{n-1} - \left(2 + \frac{h^2}{k\tau\sigma}\right)\hat{y}_n + \hat{y}_{n+1} = \left(2\frac{1-\sigma}{\sigma} - \frac{h^2}{k\tau\sigma}\right)y_n -$$
$$- \frac{1-\sigma}{\sigma}(y_{n-1} + y_{n+1}) - \frac{h^2}{k\sigma}\varphi_n, \quad 1 \leqslant n \leqslant N-1; \tag{7}$$
$$\hat{y}_0 = \mu_1(t+\tau), \quad \hat{y}_N = \mu_2(t+\tau).$$

На каждом слое уравнения (7) образуют линейную систему с неизвестными $\hat{y}_n$, $0 \leqslant n \leqslant N$. Система (7) имеет трехдиагональную матрицу и решается методом прогонки. При $\sigma > 0$ решение существует и единственно, а прогонка устойчива, ибо диагональный член матрицы (7) преобладает: его модуль больше суммы модулей недиагональных членов.

Таким образом, при $\sigma \geqslant 0$ решение разностной схемы (6) существует и единственно при любых ограниченных начальных и краевых данных и правой части. Это решение легко вычисляется, причем за небольшое число действий.

Замечание 1. При $\sigma = 1$ схема (6) использует только четыре точки шаблона и называется *чисто неявной*. При $\sigma = 1/2$ схему называют схемой *с полусуммой* или *симметричной* (имеется в виду симметрия по времени, ибо схема (6) симметрична по пространству при любом σ).

Аппроксимация. Разложим решение в узлах шаблона рис. 76 по формуле Тейлора, выбирая за центр разложения точку $(x_n,\ t+\tau/2)$ *). Тогда получим

$$\hat{u}_{n+1} = \sum_{p=0}^{\infty} \frac{1}{p!}\left(\frac{\tau}{2}\frac{\partial}{\partial t} + h\frac{\partial}{\partial x}\right)^p u = \bar{u} + \frac{\tau}{2}u_t + hu_x + \\ + \frac{\tau^2}{8}u_{tt} + \frac{\tau h}{2}u_{tx} + \frac{h^2}{2}u_{xx} + \frac{\tau^3}{48}u_{ttt} + \frac{\tau^2 h}{8}u_{ttx} + \\ + \frac{\tau h^2}{4}u_{txx} + \frac{h^3}{6}u_{xxx} + \frac{\tau^4}{384}u_{tttt} + \frac{\tau^3 h}{48}u_{tttx} + \\ + \frac{\tau^2 h^2}{16}u_{ttxx} + \frac{\tau h^3}{12}u_{txxx} + \frac{h^4}{24}u_{xxxx} + \ldots, \qquad (8)$$

где все производные отнесены к центру разложения. Разложение для $\hat{u}_{n-1}$ получается из (8) изменением знака h, разложение для u_{n+1} — изменением знака τ; для определения $\hat{u}_n$ надо в (8) положить $h = 0$ и т. д. Подставляя эти разложения в выражение невязки схемы (6а), получим

$$\psi_n = (u_t - ku_{xx} - f)_{x=x_n}^{t+\tau/2} - \frac{1}{\tau}(\hat{u}_n - u_n) + \\ + \frac{k\sigma}{h^2}(\hat{u}_{n-1} - 2\hat{u}_n + \hat{u}_{n+1}) + \frac{k(1-\sigma)}{h^2}(u_{n-1} - 2u_n + u_{n+1}) + \varphi_n = \\ = k\tau\left(\sigma - \frac{1}{2}\right)u_{txx} + \frac{\tau^2}{8}\left(ku_{ttxx} - \frac{1}{3}u_{ttt}\right) + \\ + \frac{kh^2}{12}u_{xxxx} + \varphi_n - \bar{f}_n + o(\tau^2 + h^2). \qquad (9)$$

Отсюда видно, что если положить $\varphi_n = \bar{f}_n \equiv f(x_n,\ t+\tau/2)$, то при $\sigma \neq 1/2$ схема (6) имеет аппроксимацию $O(\tau + h^2)$; симметричная схема с $\sigma = 1/2$ имеет более хорошую аппроксимацию $O(\tau^2 + h^2)$.

---

*) Для обозначения середины временного шага будем часто употреблять запись $F(x,\ t+\tau/2) = \bar{F}(x)$.

Надо проверить аппроксимацию не только уравнения, но и начальных и краевых условий. Начальное условие (1б) и краевые условия первого рода (2) мы аппроксимировали точно, положив $y_n^0 = \mu(x_n)$, $y_0^m = \mu_1(t_m)$, $y_N^m = \mu_2(t_m)$. Аппроксимация краевых условий второго или третьего рода уже не была бы точной, а содержала бы некоторую погрешность, как это отмечалось в главе IX.

Замечание 2. Для $k = \text{const}$ за счет специального выбора веса и правой части можно построить схемы повышенной точности. Для решения $u(x, t)$ дифференциального уравнения (1а) справедливо соотношение

$$u_{txx} = \frac{\partial^2}{\partial x^2} u_t = k u_{xxxx} + f_{xx}.$$

Подставляя его в (9), преобразуем невязку:

$$\psi_n = \frac{\tau^2}{8}\left(k u_{ttxx} - \frac{1}{3} u_{ttt}\right) + \left[\tau k^2 \left(\sigma - \frac{1}{2}\right) + \frac{kh^2}{12}\right] u_{xxxx} + \\ + \left[k\tau\left(\sigma - \frac{1}{2}\right) f_{xx} + \varphi_n - \bar{f}_n\right] + o(\tau^2 + h^2). \qquad (10)$$

Если положить

$$\sigma = \frac{1}{2} - \frac{h^2}{12k\tau}, \quad \varphi_n = \left(\bar{f} + \frac{h^2}{12}\bar{f}_{xx}\right)_n, \qquad (11)$$

то обе квадратные скобки в (10) обратятся в нуль и погрешность аппроксимации схемы (6), (11) будет равной $O(\tau^2) + o(h^2)$. Удерживая в формуле Тейлора (8) большее число членов, можно показать, что невязка (10) при этом равна $\psi_n = O(\tau^2 + h^4)$.

Замечание 3. Можно заменить f_{xx} в (11) второй пространственной разностью, получая следующее выражение для правой части:

$$\varphi_n = \frac{1}{12}\bar{f}_{n-1} + \frac{5}{6}\bar{f}_n + \frac{1}{12}\bar{f}_{n+1}. \qquad (12)$$

Этот вариант схемы повышенной точности имеет аппроксимацию также $O(\tau^2 + h^4)$.

Замечание 4. Приведенные оценки аппроксимации справедливы, если непрерывны те производные решения $u(x, t)$, которые входят в выражение главного члена невязки.

Устойчивость. Исследуем устойчивость по начальным данным методом разделения переменных. Поскольку схема (6) линейна, то для этого достаточно положить в ней $\varphi_n = 0$ и сделать стандартную подстановку $y_n = \exp(i\pi q x_n/a)$, $\hat{y}_n = \rho_q \times \exp(i\pi q x_n/a)$. Тогда легко получить множитель роста гармоники

$$\rho_q = \left[1 - \frac{4k\tau}{h^2}(1-\sigma)\sin^2\frac{\pi q h}{2a}\right] \Big/ \left(1 + \frac{4k\tau}{h^2}\sigma\sin^2\frac{\pi q h}{2a}\right). \qquad (13)$$

Он вещественный, причем при любом $\sigma \geqslant 0$ справедливо неравенство $\rho_q \leqslant 1$. Следовательно, схема (6) устойчива, если при любом q выполняется условие $-1 \leqslant \rho_q$. Нетрудно проверить, что это справедливо при

$$\sigma \geqslant \frac{1}{2} - \frac{h^2}{4k\tau}. \tag{14}$$

Последнее неравенство является условием равномерной устойчивости схемы (6) по начальным данным (в $\|\cdot\|_{l_2}$).

Примененный здесь простейший вариант метода разделения переменных не является строгим. Однако для схемы на равномерной сетке (6) нетрудно проверить, что функции

$$v_q(x_n) = \sin(\pi q x_n / a), \quad 1 \leqslant q \leqslant N-1, \quad x_n = nh, \tag{15}$$

являются собственными функциями разностной задачи Штурма — Лиувилля для (6). Соответствующие им собственные значения имеют вид (13), причем $q \leqslant N-1$. При их помощи можно получить строгое необходимое и достаточное условие устойчивости, практически не отличающееся от (14).

Дополнительное условие устойчивости по правой части (9.54), как легко видеть из (6), выполняется при любых τ и h. Следовательно, схема устойчива по правой части, если выполнено условие (14) равномерной устойчивости по начальным данным.

Для чисто неявной схемы, симметричной схемы и схемы повышенной точности условие (14) выполняется при любом соотношении шагов τ и h; таким образом, эти схемы безусловно устойчивы. Для явной схемы условие (14) выполняется только при $\tau \leqslant h^2/(2k)$, т. е. схема условно устойчива, что мы уже установили в главе IX.

Замечание 5. Справедливо более сильное утверждение: все эти схемы устойчивы в $\|\cdot\|_c$. В общем случае для доказательства этого утверждения приходится применять более сложную технику. Однако из принципа максимума нетрудно получить достаточное условие устойчивости в норме $\|\cdot\|_c$:

$$\sigma \geqslant 1 - \frac{h^2}{2k\tau}. \tag{16}$$

Оно более жестко, чем условие (14), но в случае явной и чисто неявной схем из него следует сделанное выше утверждение.

Сходимость. Из сказанного выше следует, что на решениях $u(x, t)$, имеющих достаточное число непрерывных производных, семейство схем (6) с весом $0 \leqslant \sigma \leqslant 1$ обеспечивает равномерную сходимость при выполнении условия устойчивости (14).

Для схем с $\sigma \neq 1/2$ погрешность $\|y-u\|_c = O(\tau + h^2)$, т. е. схемы имеют первый порядок точности по времени и второй — по пространству. Для симметричной схемы ($\sigma = 1/2$) погрешность равна $O(\tau^2 + h^2)$, т. е. порядок точности по обеим переменным

второй. Схема повышенной точности с весом (11) и соответственно выбранной φ_n имеет погрешность $O(\tau^2+h^4)$.

Замечание 6. Поскольку схема (6) двуслойная, то она без изменения переносится на неравномерную сетку по t (разумеется, при шаге по времени τ_m надо ставить его индекс). На неравномерную сетку по x эта схема легко обобщается. Достаточно соответствующим образом записать разностный аналог пространственной производной:

$$u_{xx} \approx \frac{2}{x_{n+1}-x_{n-1}}\left(\frac{u_{n+1}-u_n}{x_{n+1}-x_n} - \frac{u_n-u_{n-1}}{x_n-x_{n-1}}\right).$$

В этом случае схема по-прежнему сходится в $\|\cdot\|_c$ с погрешностью $O(h^2)$; однако доказательство этого утверждения значительно сложнее и проводится методом энергетических неравенств (см. [30]).

Подведем итоги. Поскольку погрешность почти для всех значений σ есть $O(h^2)$, то для получения хорошей точности при расчете по схемам (6) надо брать довольно малый шаг h.

В этом случае явная схема устойчива при настолько малом $\tau \leqslant h^2/(2k)$, что для доведения расчета до заданного момента T требуется сделать очень много шагов по времени, т. е. выполнить большой объем вычислений. Поэтому явные схемы для решения параболических уравнений почти никогда не употребляются.

Обычно для расчетов берут двуслойные неявные безусловно устойчивые схемы. Чаще всего используют симметричную схему или схему повышенной точности, обеспечивающие хорошую точность расчета при не слишком малых шагах τ и h. Чисто неявная схема в случае $k=\text{const}$ редко употребляется из-за невысокой точности (хотя при $k=k(u)$ она часто выгодна благодаря своей монотонности).

3. Асимптотическая устойчивость неявной схемы. Исследуем, при каких условиях схема (6) позволяет рассчитывать задачи с нулевыми краевыми значениями для очень больших промежутков времени *), т. е. каковы условия асимптотической устойчивости схемы.

Выход решения параболического уравнения (1) на асимптотику при $t\to\infty$ определяется скоростью затухания начальных данных. Приведенное в главе IX разложение решения $u(x, t)$ в ряд Фурье (9.7):

$$u(x, t) = \sum_{q=1}^{\infty} \alpha_q \exp\left(-k\frac{\pi^2 q^2}{a^2}t\right)\sin\frac{\pi q x}{a},$$

*) Это нужно в задачах фильтрации нефти в пласте при многолетней эксплуатации скважин, задачах прогрева слоя вечной мерзлоты и ряде других.

показывает, что медленнее всего затухает первая гармоника. Ей соответствует множитель роста

$$\bar{\rho}_1 = \exp\left(-k\frac{\pi^2}{a^2}\tau\right) = 1 - \frac{k\pi^2\tau}{a^2} + O(\tau^2). \tag{17}$$

Чтобы схема (6) была асимптотически устойчивой, ее множители роста (13) не должны превосходить по модулю величины $\bar{\rho}_1$, т. е. должно выполняться условие

$$-1 + \frac{k\pi^2\tau}{a^2} \leqslant 1 - \frac{\sin^2\dfrac{\pi qh}{2a}}{\dfrac{h^2}{4k\tau} + \sigma\sin^2\dfrac{\pi qh}{2a}} \leqslant 1 - \frac{k\pi^2\tau}{a^2},\quad 1 \leqslant q \leqslant N-1; \tag{18}$$

здесь мы преобразовали выражение (13) для ρ_q к более удобному виду. Разумеется, достаточно выполнения этих неравенств с точностью до членов $O(\tau^2)$, потому что наличие таких членов приведет к умножению амплитуд гармоник на величину $[1 + O(\tau^2)]^{t/\tau} = 1 + tO(\tau)$, чем при $\tau \to 0$ можно пренебречь, даже если t велико.

Нетрудно проверить, что правое неравенство (18) всегда выполняется. В самом деле, ρ_q монотонно убывает при увеличении $\sin(\pi qh/(2a))$, т. е. при увеличении q. Поэтому наибольшим является ρ_1, которое с учетом малости h равно

$$\rho_1 \approx 1 - \frac{(\pi h/2a)^2}{(h^2/4k\tau) + \sigma(\pi h/2a)^2} \approx 1 - \frac{k\pi^2\tau}{a^2}$$

и совпадает с $\bar{\rho}_1$ с точностью до членов $O(\tau^2)$.

Рассмотрим левое неравенство (18). Величина ρ_q минимальна при $q = N-1$, когда $\sin(\pi qh/2a) \approx 1$. Подстановка этого значения в левое неравенство (18) после несложных выкладок приводит к условию асимптотической устойчивости

$$\sigma \geqslant \frac{1}{2} + \frac{\pi^2 k\tau}{4a^2} - \frac{h^2}{4k\tau}. \tag{19}$$

Оно несколько более жестко, чем условие обычной устойчивости (14). Его можно переписать в следующем виде:

$$\tau^2 - \frac{2a^2}{\pi^2 k}(2\sigma - 1)\tau - \left(\frac{ah}{\pi k}\right)^2 \leqslant 0.$$

Стоящий слева квадратный трехчлен отрицателен, если τ лежит между его корнями:

$$\tau_{1,2} = \frac{a^2}{\pi^2 k}\left[(2\sigma - 1) \pm \sqrt{(2\sigma-1)^2 + \left(\frac{\pi h}{a}\right)^2}\right].$$

Один из корней отрицателен, а другой положителен. Поэтому условие асимптотической устойчивости (19) принимает вид

$$\tau \leqslant \frac{a^2}{\pi^2 k}\left[2\sigma - 1 + \sqrt{(2\sigma-1)^2 + (\pi h/a)^2}\right]. \qquad (20)$$

В частности, симметричная схема асимптотически устойчива не при любом τ, а только при

$$\tau \leqslant \frac{ah}{\pi k} \quad (\sigma = 1/2). \qquad (21)$$

Таким образом, схема (6) при любом σ формально является лишь асимптотически *условно* устойчивой. Однако фактически устойчивость условна только при $\sigma \leqslant 1/2 + O(h)$, когда ограничение на шаги принимает вид $\tau \leqslant h \cdot \text{const}$. Если же $\sigma > 1/2$, то условие (20) требует, чтобы выполнялось неравенство $\tau \leqslant [2a^2(2\sigma-1)/\pi^2 k] = \text{const}$, и по существу схема является асимптотически безусловно устойчивой.

Замечание. При больших t схемы с $\sigma > 1/2$ дают низкую точность. Поэтому для таких расчетов обычно используют схему с $\sigma = 1/2$.

4. Монотонность. Точное решение $u(x, t)$ уравнения $u_t = ku_{xx}$ при определенных условиях сохраняет монотонность. Например, если начальные данные монотонны и температура на концах отрезка постоянна, то профиль температуры будет монотонен в любой момент времени. То же будет при постановке задачи Коши на бесконечной прямой.

Выясним, сохраняет ли схема (6) монотонность решения. Ограничимся задачей на бесконечной прямой, хотя при использовании результатов надо помнить, что краевые условия тоже влияют на монотонность (если разностное краевое условие не точное, то его неудачное написание может привести к немонотонности схемы).

Для случая $\sigma = 0$ результат почти очевиден. Получающаяся при этом явная схема имеет форму (10.34):

$$\hat{y}_n = \sum_{l=-1}^{1} \beta_l y_{n+l}, \quad \beta_0 = 1 - \frac{2k\tau}{h^2}, \quad \beta_1 = \beta_{-1} = \frac{k\tau}{h^2}.$$

Необходимым и достаточным условием монотонности является неотрицательность коэффициентов β_l. Видно, что если выполнено условие устойчивости этой схемы $2k\tau \leqslant h^2$, то коэффициенты неотрицательны и схема монотонна. В противном случае явная схема немонотонна.

При $\sigma \neq 0$ запишем неявную схему (6), полагая $\varphi_n = 0$ и выделяя преобладающий член на новом слое:

$$\hat{y}_n = \left(2 + \frac{h^2}{k\tau\sigma}\right)^{-1} \left[\hat{y}_{n-1} + \hat{y}_{n+1} + \frac{1-\sigma}{\sigma}(y_{n-1} + y_{n+1}) + \right.$$
$$\left. + \left(\frac{h^2}{k\tau\sigma} - 2\frac{1-\sigma}{\sigma}\right) y_n\right]. \qquad (22)$$

Напишем для $\hat{y}_{n-1}$ и $\hat{y}_{n+1}$ аналогичные выражения и подставим их в правую часть (22). При этом появятся другие значения с нового слоя; будем их исключать тем же способом. Коэффициенты при значениях $y_{n \pm l}$ на новом слое в правой части будут при этом убывать в геометрической прогрессии. Поэтому в пределе соотношение (22) перейдет в явную схему вида (10.34) с бесконечной суммой, т. е. с бесконечной зоной влияния.

Очевидно, если выполнено условие

$$\tau \leqslant \frac{h^2}{2k(1-\sigma)}, \qquad (23)$$

то все коэффициенты в (22) неотрицательны. Тогда все коэффициенты в соответствующей явной схеме также будут неотрицательны. Следовательно, неравенство (23) есть достаточное условие монотонности схемы (6).

Можно получить необходимое и достаточное условие монотонности, приведя схему (22) после выполнения громоздких выкладок к явной форме (10.34):

$$\hat{y}_n = \beta_0 y_n + \sum_{l=1}^{\infty} \beta_l (y_{n-l} + y_{n+l}), \qquad (24а)$$

где

$$\left.\begin{aligned} &\beta_0 = 1 - \frac{4k\tau}{\gamma(h+\gamma)}, \quad \beta_1 = \frac{4hk\tau}{\gamma(h+\gamma)^2}, \\ &\beta_l = \beta_{l-1}\frac{4\sigma k\tau}{(h+\gamma)^2} \quad \text{при} \quad l \geqslant 2, \\ &\gamma = \sqrt{h^2 + 4\sigma k\tau}. \end{aligned}\right\} \qquad (24б)$$

Очевидно, $\beta_l \geqslant 0$ при $l \geqslant 1$ и отрицательным может быть только коэффициент β_0. Он неотрицателен, если

$$\tau \leqslant \frac{(2-\sigma)h^2}{4k(1-\sigma)^2}. \qquad (25)$$

Это условие необходимо и достаточно для монотонности схемы (6); оно несколько слабее ограничения (23).

Таким образом, неявные схемы монотонны только при очень малом шаге по времени $\tau \sim h^2$. По абсолютно устойчивым неявным схемам расчеты обычно проводят с шагом $\tau \sim h$, не гаран-

тирующим монотонности. Единственное исключение — чисто неявная схема с $\sigma = 1$, которая монотонна при любых шагах.

Напомним, что достаточно гладкое решение на подробных сетках можно хорошо находить и по немонотонным схемам. На грубых же сетках, особенно при разрывных начальных данных, симметричная схема может привести к «разболтке» счета. Чисто неявная схема даже в этих условиях дает плавно меняющееся разностное решение, хотя точность его невысока.

Замечание. Монотонные схемы для параболического уравнения могут иметь второй порядок точности по пространству. Но, как и для уравнения переноса, для параболического уравнения не известно ни одной монотонной схемы, которая имела бы второй порядок точности по времени (хотя никаких теорем о невозможности построения таких схем не доказано).

5. Явные схемы. Явные схемы имеют важное достоинство: они просто записываются и легко программируются на ЭВМ. Поэтому предпринималось много попыток построить для параболического уравнения $u_t = ku_{xx} + f$ хорошую явную схему. Однако все эти попытки были неудачными.

Например, Ричардсоном была предложена трехслойная схема, использующая шаблон рис. 77 с аппроксимацией производных двусторонними разностями:

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{k}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + f_n. \tag{26}$$

Из симметрии схемы легко усмотреть, что локальная погрешность ее аппроксимации есть $O(\tau^2 + h^2)$. Однако схема Ричардсона непригодна для расчетов, ибо она безусловно неустойчива. В самом деле, используем метод разделения переменных и сделаем подстановку $y_n = \exp(iqx_n)\,\hat{y}_n = \rho_q y_n$; поскольку схема трехслойна, надо дополнительно положить $\check{y}_n = (1/\rho_q)\,y_n$. Тогда для множителя роста получим квадратное уравнение

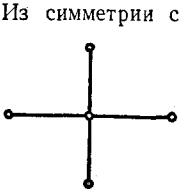


Рис. 77.

$$\rho_q^2 + \frac{8k\tau}{h^2}\rho_q \sin^2\frac{qh}{2} - 1 = 0, \tag{27}$$

один из корней которого при любом $q \neq 0$ по модулю больше единицы на величину $O(k\tau/h^2)$.

Дюфорт и Франкел в 1953 г. видоизменили схему Ричардсона, заменив в правой части (26) величину y_n на $(\hat{y}_n + \check{y}_n)/2$:

$$\frac{1}{2\tau}(\hat{y}_n - \check{y}_n) = \frac{k}{h^2}(y_{n-1} - \hat{y}_n - \check{y}_n + y_{n+1}) + f_n. \tag{28}$$

Эта схема также явно разрешается относительно $\hat{y}_n$. Методом разделения переменных нетрудно показать, что она безусловна устойчива. Однако погрешность аппроксимации схемы (28) равна $O(\tau^2 + h^2 + \tau^2/h^2)$, т. е. аппроксимация условная. Поэтому сходимость имеет место, только если $(\tau/h) \to 0$ при $h \to 0$.

Фактически, чтобы в расчетах по схеме (28) получить точность $O(h^2)$, надо положить $k\tau \sim h^2$, как в явной схеме (6). Правда, коэффициент пропорциональности $\alpha = (k\tau/h^2)$ можно брать любым, ибо его величина влияет только на

точность расчета, а не на устойчивость. Поэтому схема Дюфорта—Франкела удобнее явной схемы (6), но ненамного.

Плохие качества явных схем обусловлены одним принципиальным ограничением: *явная схема для параболического уравнения может сходиться, только если* $(\tau/h)\to 0$ *при* $h\to 0$. В самом деле, пусть решение в точке нового слоя выражается через r точек исходного слоя, т. е. через отрезок длиной rh (рис. 78). Тогда оно выражается через отрезок нулевого слоя длиной $mrh = rth/\tau$; этот отрезок будет зоной влияния. Для точного решения зона влияния бесконечна. Значит, сходимость к точному решению при $\tau\to 0$, $h\to 0$ возможна, только если дополнительно $(rth/\tau)\to\infty$, т. е. $(\tau/h)\to 0$, что и требовалось доказать.

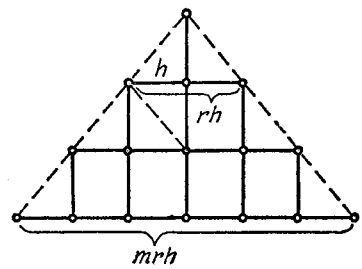


Рис. 78.

Этот результат можно уточнить. В п. 1 отмечалось, что для промежутка времени τ эффективной зоной влияния является отрезок $h\sim\sqrt{k\tau}$. Следовательно, условие сходимости явных схем должно иметь вид $k\tau\lesssim h^2$.

Поэтому для параболического уравнения неявные безусловно устойчивые схемы дают лучшие результаты, чем явные.

Отметим одну любопытную схему для уравнения теплопроводности — схему *бегущего счета* на шаблоне рис. 79. На четных слоях счет идет слева направо (рис. 79, *а*) по формулам

$$\frac{1}{\tau}(y_n-\check{y}_n)=\frac{k}{h^2}(y_{n-1}-y_n-\check{y}_n+\check{y}_{n+1})+f\left(x_n,\ t-\frac{\tau}{2}\right), \tag{29а}$$

а на нечетных слоях — справа налево (рис. 79, *б*) по симметрично преобразованным формулам

$$\frac{1}{\tau}(\hat{y}_n-y_n)=\frac{k}{h^2}(y_{n-1}-y_n-\hat{y}_n+\hat{y}_{n+1})+f\left(x_n,\ t+\frac{\tau}{2}\right). \tag{29б}$$

Организация расчета здесь так же проста, как в явных схемах. В то же время зона влияния бесконечна благодаря наличию двух точек верхнего слоя в каждом уравнении (29); поочередная смена направления расчета обеспечивает бесконечность зоны влияния в обоих направлениях.

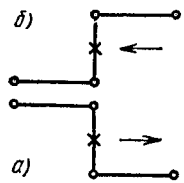


Рис. 79.

Методом разделения переменных нетрудно проверить, что схема (29) безусловно устойчива. Невязка каждого из уравнений (29а) и (29б), вычисленная разложением относительно центров, показанных на шаблонах рис. 79 крестиками, есть $O(\tau^2+\tau h+h^2+\tau/h)$. Если бы расчет производился только по одному из этих уравнений, т. е. использовалась бы односторонняя схема бегущего счета, то именно таким был бы порядок точности.

Однако при сложении погрешностей прямого и обратного хода на последовательных слоях происходит их частичная компенсация. Поэтому двусто-

ронняя схема бегущего счета (29), как показывает более детальный анализ, сходится со скоростью

$$O(\tau^2+h^2+\tau^2/h^2). \tag{30}$$

Тем самым, она по своим свойствам близка к схеме Дюфорта — Франкела (28).

6. **Наилучшая схема.** Рассмотрим, как следует обобщать схему (6) на уравнение теплопроводности с переменным коэффициентом теплопроводности, которое имеет следующий вид:

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left[k(x,\ t)\frac{\partial u}{\partial x}\right]+f(x,\ t). \tag{31}$$

Случай непрерывных и гладких коэффициентов несложен, и отдельно мы его разбирать не будем. Исследуем более общий случай, когда $k(x,\ t)$ и $f(x,\ t)$ — кусочно-непрерывные функции.

Разрывы коэффициентов уравнения (31) возникают, например, на границах областей в задачах для слоистых сред или на ударных волнах в движущейся среде. В точках разрыва коэффициентов решение $u(x,\ t)$ будет иметь особенности, т. е. оно будет обобщенным и, вообще говоря, не единственным.

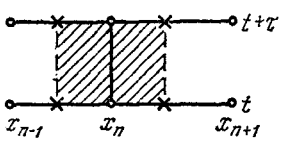


Рис. 80.

Для выделения допустимого решения из множества обобщенных решений надо выяснить, какие величины всюду непрерывны, согласно физическому смыслу задачи. Для теплопроводности, как уже отмечалось в главе VIII, § 2, п. 7, непрерывны температура $u(x,\ t)$ и поток тепла

$$W=-k(\partial u/\partial x).$$

Заметим, что производные этих величин разрывны; u_x имеет разрывы в точках разрыва $k(x,\ t)$, а W_x разрывна в точках разрыва $f(x,\ t)$.

Чтобы получить сходимость к допустимому обобщенному решению, составим методом баланса консервативную разностную схему.

Уравнение (31) записано в дивергентной форме, соответствующей закону сохранения энергии. Удобнее заменить его системой уравнений

$$\frac{\partial u}{\partial t}=-\frac{\partial W}{\partial x}+f,\qquad W=-k\frac{\partial u}{\partial x}. \tag{32}$$

Выберем шаблон и связанную с ним ячейку (рис. 80) и запишем первое уравнение (32) в виде закона сохранения энергии для

этой ячейки:

$$\int_{x_n-1/2}^{x_n+1/2}(\hat{u}-u)\,dx=$$

$$=\int_{t}^{t+\tau}(W_{n-1/2}-W_{n+1/2})\,dt+\int_{t}^{t+\tau}\int_{x_n-1/2}^{x_n+1/2}f(x,\ t)\,dx\,dt. \qquad (33а)$$

Второе уравнение (32) проинтегрируем по интервалу сетки:

$$u_{n+1}-u_n=-\int_{x_n}^{x_{n+1}}\frac{W}{k(x,\ t)}\,dx. \qquad (33б)$$

Справедливость формулы (33б) очевидна; если коэффициент $k(x,\ t)$ непрерывен на интервале сетки; но благодаря аддитивности интегрирования формула остается справедливой при наличии разрывов $k(x,\ t)$ внутри $[x_n,\ x_{n+1}]$.

Припишем значения температуры узлам сетки, а значения тепловых потоков — серединам интервалов (крестики на рис. 80). Аппроксимируем интегралы в (33) квадратурными формулами. При этом $\int W\,dt$ вычислим по двухточечной формуле с весом σ на верхнем слое, а в (33б) вынесем среднее значение потока за знак интеграла:

$$u_{n+1}-u_n\approx -W_{n+1/2}\int_{x_n}^{x_{n+1}}\frac{dx}{k(x,\ t)},$$

что допустимо в силу непрерывности потока. Получим консервативную разностную схему, называемую *наилучшей*:

$$\frac{1}{\tau}(\hat{y}_n-y_n)=\frac{\sigma}{\hbar_n}(\hat{W}_{n-1/2}-\hat{W}_{n+1/2})+$$

$$+\frac{1-\sigma}{\hbar_n}(W_{n-1/2}-W_{n+1/2})+\varphi_n, \qquad (34а)$$

$$W_{n+1/2}=\bar{\varkappa}_{n+1/2}\frac{y_n-y_{n+1}}{h_n},\quad \hat{W}_{n+1/2}=\bar{\varkappa}_{n+1/2}\frac{\hat{y}_n-\hat{y}_{n+1}}{h_n}, \qquad (34б)$$

где

$$h_n=x_{n+1}-x_n,\quad \hbar_n=\frac{1}{2}(h_{n-1}+h_n)=x_{n+1/2}-x_{n-1/2}, \qquad (34в)$$

$$\bar{\varkappa}_{n+1/2}=\left[\frac{1}{h_n}\int_{x_n}^{x_{n+1}}\frac{dx}{k(x,\ \bar{t})}\right]^{-1},\quad \bar{t}=t+\frac{\tau}{2}, \qquad (34г)$$

$$\varphi_n=\frac{1}{\tau\hbar_n}\int_{t}^{t+\tau}dt\int_{x_n-1/2}^{x_n+1/2}dx\,f(x,\ t). \qquad (34д)$$

При вычислениях интегралы (34г), (34д) также аппроксимируют несложными квадратурными формулами. Например, если $k(x, t)$ и $f(x, t)$ непрерывны всюду, за исключением узлов x_n, то можно воспользоваться одной из следующих формул:

$$\bar{\varkappa}_{n+1/2} \approx \bar{k}_{n+1/2} \approx \frac{1}{2}(\bar{k}_n + \bar{k}_{n+1}) \approx \frac{2\bar{k}_n\bar{k}_{n+1}}{\bar{k}_n + \bar{k}_{n+1}} \approx \sqrt{\bar{k}_n\bar{k}_{n+1}}, \quad (35а)$$

$$\varphi_n \approx \frac{x_n - x_{n-1/2}}{h_n}\bar{f}_{n-1/2} + \frac{x_{n+1/2} - x_n}{h_n}\bar{f}_{n+1/2}, \quad (35б)$$

где черта означает, что величина отнесена к моменту времени $\hat{t}$. Под узловыми значениями разрывных величин здесь надо понимать соответствующие односторонние пределы.

Название схемы (34) связано с ее высокой точностью. Например, можно показать, что для однородного стационарного уравнения (31) наилучшая схема является точной, если интегралы (34г) вычисляются точно. Это означает, что разностное решение y_n при любых величинах шагов совпадает с $u(x_n, t)$ (хотя разностные значения $W_{n+1/2}$ могут не совпадать с точными значениями потоков в точках $x_{n+1/2}$).

Исследуем схему (34). Подставляя (34б) в (34а), получим линейную трехточечную (по пространству) схему. Для определения $\hat{y}_n$ надо решить линейную систему с трехдиагональной матрицей, что выполняется методом прогонки. Легко видеть, что диагональные члены матрицы преобладают; это обеспечивает единственность разностного решения и устойчивость прогонки.

Устойчивость по начальным данным исследуем методом операторных неравенств. Ограничимся случаем задачи Коши на бесконечной прямой, когда $u(-\infty, t) = u(+\infty, t) = 0$.

Перепишем двуслойную схему (34) в канонической форме:

$$B\frac{\hat{y} - y}{\tau} + Ay = \varphi, \quad (36а)$$

где

$$By_n = y_n + \sigma\tau Ay_n, \quad (36б)$$

$$Ay_n = -\left(\bar{\varkappa}_{n+1/2}\frac{y_{n+1} - y_n}{h_n h_n} - \bar{\varkappa}_{n-1/2}\frac{y_n - y_{n-1}}{h_n h_{n-1}}\right). \quad (36в)$$

Введем скалярное произведение

$$(v, w) = \sum_{n=-\infty}^{+\infty} h_n v_n w_n. \quad (37)$$

Нетрудно убедиться, что операторы A и B неотрицательные и самосопряженные. В самом деле,

$$(Ay, y) = (y, Ay) =$$

$$= -\sum_{n=-\infty}^{+\infty} y_n\bar{\varkappa}_{n+1/2}\frac{y_{n+1} - y_n}{h_n} + \sum_{n=-\infty}^{+\infty} y_n\bar{\varkappa}_{n-1/2}\frac{y_n - y_{n-1}}{h_{n-1}}.$$

Сдвигая во второй сумме индекс на единицу, получим

$$(Ay,\ y)=\sum_{n=-\infty}^{+\infty}\overline{\varkappa}_{n+1/2}\frac{(y_{n+1}-y_n)^2}{h_n}\geqslant 0. \tag{38}$$

Равенство (36б) означает, что $B=E+\sigma\tau A$; тогда $(By,\ y)=(y,\ By)=(y,\ y)+(y,\ Ay)\geqslant 0$, что доказывает наше утверждение об операторах A и B. Заметим, что из (38) следует оценка

$$\|A\|\leqslant 4\max(\varkappa/h^2). \tag{39}$$

Пусть выполнено условие

$$\sigma\geqslant\sigma_0,\qquad \sigma_0=\frac{1}{2}-\frac{1}{\tau\|A\|}. \tag{40}$$

Учитывая, что $0\leqslant A\leqslant\|A\|E$, т. е. $E\geqslant A/\|A\|$, получим:

$$B-\tfrac{1}{2}\tau A=E+\left(\sigma-\tfrac{1}{2}\right)\tau A\geqslant\left[\frac{1}{\|A\|}+\left(\sigma-\tfrac{1}{2}\right)\tau\right]A=$$
$$=(\sigma-\sigma_0)\,\tau A\geqslant 0.$$

Это означает, что $B\geqslant{}^1/_2\tau A$; следовательно, по теореме из главы IX, § 3, п. 6 схема (34) устойчива в норме $\|\cdot\|_A$. Таким образом, неравенство (40) является достаточным условием устойчивости схемы (34).

Если выполнено условие

$$\delta\geqslant\frac{1}{2}-\frac{1}{4\tau}\min(h^2/\varkappa), \tag{41}$$

то, в силу неравенства (39), условие (40) имеет место. Поэтому неравенство (41) также является достаточным условием устойчивости схемы (34).

Сходимость для своего доказательства требует оценок аппроксимации. Это связано с громоздкими выкладками (см. [30]), поэтому приведем только окончательный результат.

Пусть $k(x,\ t)$ и $f(x,\ t)$ кусочно-непрерывны вместе со своими первыми и вторыми производными, причем разрывы неподвижны (т. е. линии разрыва на плоскости $(x,\ t)$ параллельны оси t). Выберем *специальную* сетку по x, т. е. такую, что все точки разрыва коэффициентов и их указанных производных являются ее узлами; эта сетка будет, вообще говоря, неравномерной. За средний шаг этой сетки примем $h_c=\sqrt{(h,\ h)}$ со скалярным произведением (37).

Тогда *наилучшая схема* (34) *при выполнении условия устойчивости* (41) *равномерно сходится на специальных сетках с точностью* $O(\tau^{\nu}+h_c^2)$, *где* $\nu=2$ *при весе* $\sigma=1/2$ *и* $\nu=1$ *при* $\sigma\neq 1/2$.

Если $k(x, t)$ и $f(x, t)$ дважды непрерывно дифференцируемы, то наилучшая схема при выполнении условия устойчивости равномерно сходится на произвольных (неравномерных) сетках с точностью $O(\tau^{\nu}+h_c^2)$.

Монотонность схемы имеет место при достаточно малом шаге по времени:

$$\tau\leqslant\frac{1}{2(1-\sigma)}\min_n\left(\frac{h_n^2}{\varkappa_n}\right), \tag{42}$$

за одним очевидным исключением: чисто неявная схема с $\sigma=1$ монотонна при любых шагах. Доказательство этого утверждения аналогично доказательству условия (23).

Замечание. Коэффициенты разностной схемы вычисляются с некоторыми ошибками, что может привести к искажению решения. Устойчивость разностного решения относительно изменения коэффициентов называется *коэффициентной устойчивостью* (*ко-устойчивостью*).

Доказано (см. [30]), что наилучшая схема при выполнении условия (41) является ко-устойчивой.

7. Криволинейные координаты. Нередко приходится решать одномерные задачи с цилиндрической или сферической симметрией. Например, цилиндрическая симметрия имеется в задачах об остывании длинного цилиндра или в задачах о шнуровых электрических разрядах, где требуется определить теплопроводность и диффузию магнитного поля. Сферически-симметричными являются задачи о теплоотводе от ядра к поверхности звезды *).

Естественной системой координат в таких задачах является, соответственно, цилиндрическая (r, φ) или сферическая (r, θ, φ). Вследствие одномерности все величины не будут зависеть от углов θ, φ. Тогда параболическое уравнение с переменными коэффициентами в соответствующих координатах примет вид

$$\frac{\partial u}{\partial t}=-\frac{1}{r^{\nu}}\frac{\partial}{\partial r}(r^{\nu}W)+f(r, t),\quad W=-k(r, t)\frac{\partial u}{\partial r}. \tag{43}$$

Здесь ν — показатель симметрии, равный 0, 1, 2 соответственно для плоского, цилиндрического и сферического случаев.

Для уравнения (43) можно построить консервативную схему, являющуюся обобщением наилучшей схемы (34). Для этого проинтегрируем первое уравнение (43) по элементу объема $r^{\nu}dr\,dt$

*) Но в тех слоях звезды, где есть конвекция, перенос тепла описывается не параболическим уравнением.

в пространстве r, t, а второе уравнение — по радиусу:

$$\int_{r_{n-1/2}}^{r_{n+1/2}} (\hat{u}-u)\, r^{\nu}\, dr = \int_{t}^{t+\tau} \left(r^{\nu}_{n-1/2} W_{n-1/2} - r^{\nu}_{n+1/2} W_{n+1/2}\right) dt + \\ + \int_{t}^{t+\tau} dt \int_{r_{n-1/2}}^{r_{n+1/2}} f(r, t)\, r^{\nu}\, dr, \quad (44а)$$

$$u_{n+1} - u_n = -\int_{r_n}^{r_{n+1}} \frac{W}{k(r, t)}\, dr. \quad (44б)$$

Уравнение (44а) есть интегральная запись закона сохранения энергии. Вычислим интеграл в его левой части:

$$\int_{r_{n-1/2}}^{r_{n+1/2}} u r^{\nu}\, dr \approx u_n \int_{r_{n-1/2}}^{r_{n+1/2}} r^{\nu}\, dr = u_n V_n,$$

где V_n есть объем кольцевого или сферического слоя:

$$V_n = \frac{1}{\nu+1}\left(r^{\nu+1}_{n+1/2} - r^{\nu+1}_{n-1/2}\right). \quad (45)$$

Аппроксимируя остальные интегралы так, как в п. 6, получим разностную схему с весами:

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{\sigma}{V_n}\left(r^{\nu}_{n-1/2}\hat{W}_{n-1/2} - r^{\nu}_{n+1/2}\hat{W}_{n+1/2}\right) + \\ + \frac{1-\sigma}{V_n}\left(r^{\nu}_{n-1/2} W_{n-1/2} - r^{\nu}_{n+1/2} W_{n+1/2}\right) + \varphi_n,$$

$$W_{n+1/2} = \overline{\varkappa}_{n+1/2}\, \frac{y_n - y_{n+1}}{h_n}, \qquad h_n = r_{n+1} - r_n, \quad (46)$$

$$\overline{\varkappa}_{n+1/2} = \left[\frac{1}{h_n}\int_{r_n}^{r_{n+1}} \frac{dr}{k(r, t)}\right]^{-1}, \qquad \varphi_n = \frac{1}{\tau V_n}\int_{t}^{t+\tau} dt \int_{r_{n-1/2}}^{r_{n+1/2}} f(r, t)\, r^{\nu}\, dr.$$

Исследование этой схемы проводится аналогично исследованию схемы (34).

Обратим внимание на постановку граничного условия при $r = 0$ для цилиндрического или сферического случаев ($\nu = 1$ или 2). На оси или в центре симметрии естественное граничное условие есть

$$W(0, t) = 0. \quad (47)$$

Для аппроксимации этого условия удобно выбрать пространственную сетку так, чтобы $r_1 = {}^1/_2 h_0$, $r_0 = -{}^1/_2 h_0$; при этом узел r_0

будет фиктивным. Тогда точка $r=0$ является серединой первого интервала, и разностный аналог краевого условия (47) примет вид

$$y_0 = y_1. \tag{48}$$

Замечание. Такой способ выбора сетки нередко применяют на внешней границе, а также в плоском случае, если на границе задано краевое условие второго рода $(u_x)_{\text{гран}} = \mu(t)$.

8. Квазилинейное уравнение. Значительную трудность для численных расчетов представляет случай квазилинейного уравнения теплопроводности, которое мы запишем, для определенности, в плоском случае:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[k(x,\ t,\ u)\frac{\partial u}{\partial x}\right] + f(x,\ t,\ u), \qquad k \geqslant 0. \tag{49}$$

1) В таких задачах коэффициент теплопроводности нередко сильно зависит от температуры *) и при высоких температурах может стать очень большим. Поэтому явные схемы для уравнения (49) совершенно непригодны из-за сильного ограничения на шаг, и расчет надо вести по безусловно устойчивым неявным схемам с весом $\sigma \geqslant 1/2$.

2) У квазилинейного уравнения теплопроводности существуют решения $u(x,\ t)$, производные которых обращаются в отдельных точках в бесконечность. Примером такого решения является рассмотренная в главе IX бегущая тепловая волна (9.12), у которой на фронте $u_x = \infty$. Такие решения близки к разрывным, и при их расчете по немонотонным, хотя и устойчивым схемам легко возникает «разболтка», т. е. пилообразные профили.

Поэтому для численного решения уравнения (49) удобно использовать чисто неявные схемы с весом $\sigma = 1$, которые устойчивы и монотонны при любых шагах. Рассмотрим (ограничиваясь для простоты записи равномерной сеткой) два варианта таких схем, которые будем называть *линейным*:

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{1}{h^2}\left[\varkappa_{n+1/2}(\hat{y}_{n+1} - \hat{y}_n) - \varkappa_{n-1/2}(\hat{y}_n - \hat{y}_{n-1})\right] + \varphi_n, \tag{50}$$

и *нелинейным*:

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{1}{h^2}\left[\hat{\varkappa}_{n+1/2}(\hat{y}_{n+1} - \hat{y}_n) - \hat{\varkappa}_{n-1/2}(\hat{y}_n - \hat{y}_{n-1})\right] + \hat{\varphi}_n. \tag{51}$$

Здесь $\varkappa$ определяется формулами типа (35а), например:

$$\varkappa_{n+1/2} = \frac{1}{2}\left[k(x_n,\ t,\ y_n) + k(x_{n+1},\ t,\ y_{n+1})\right]$$

*) Например, по закону $k(u) \approx u^\alpha$, где $\alpha = 5/2$ для электронной теплопроводности и $\alpha \sim 5-8$ для лучистой теплопроводности.

или

$$\hat{\varkappa}_{n+1/2} = k\left(x_{n+1/2},\ t+\tau,\ \frac{1}{2}(\hat{y}_n + \hat{y}_{n+1})\right);$$

аналогично определяется φ.

Можно показать, что обе схемы абсолютно устойчивы, консервативны, монотонны и на четырежды непрерывно дифференцируемых решениях имеют погрешность аппроксимации $O(\tau + h^2)$. Сравним эти схемы между собой.

Линейный вариант (50) проще. Мы называем его линейным, ибо $\varkappa_{n\pm 1/2}$ зависит только от решения y с известного слоя; поэтому уравнения (50) содержат $\hat{y}_n$ линейно *). Из линейности и преобладания диагональных элементов матрицы следует существование и единственность разностного решения $\hat{y}_n$. Это решение вычисляется прогонкой, так что формулы расчета просты и легко программируются на ЭВМ.

Нелинейный вариант (51) содержит дополнительную зависимость $\hat{\varkappa}(\hat{y})$ от значения $\hat{y}$ на новом слое, благодаря чему алгебраическая система (51) нелинейна относительно $\hat{y}_n$. Очевидно, если $\tau \to 0$, то $\hat{y}_n \to y_n$, поэтому при достаточно малом τ существует вещественное решение системы (51). Но при большом τ система (51) может и не иметь вещественного решения.

Вычислять решение системы (51) можно двумя способами. Первый способ — метод последовательных приближений, в котором значения $\hat{\varkappa}$ и $\hat{\varphi}$ берутся с предыдущей итерации:

$$\frac{1}{\tau}(\hat{y}_n^{(s)} - y_n) =$$

$$= \frac{1}{h^2}\left[\hat{\varkappa}_{n+1/2}^{(s-1)}(\hat{y}_{n+1}^{(s)} - \hat{y}_n^{(s)}) - \hat{\varkappa}_{n-1/2}^{(s-1)}(\hat{y}_n^{(s)} - \hat{y}_{n-1}^{(s)})\right] + \hat{\varphi}_n^{(s-1)}, \quad (52)$$

$$\hat{\varkappa}^{(s-1)} = \hat{\varkappa}(\hat{y}^{(s-1)}), \quad \hat{y}_n^{(0)} = y_n$$

(в качестве нулевого приближения здесь, естественно, берутся значения с известного слоя). Величины $\hat{y}_n^{(s)}$ находятся из (52) прогонкой. Итерации (52) сходятся линейно и обычно не быстро; они могут и расходиться **). В последнем случае можно вести расчет с фиксированным числом итераций, обычно с двумя или тремя итерациями. Отметим, что при одной итерации (52) нелинейная схема (51) совпадает с линейным вариантом (50).

Сложней, но заметно эффективней второй способ решения системы (51) — метод Ньютона. Учитывая, что $\hat{\varkappa}_{n+1/2} = \hat{\varkappa}(\hat{y}_n, \hat{y}_{n+1})$, подставим в уравнения (51) $\hat{y}_n = \hat{y}_n^{(s)} + \delta\hat{y}_n^{(s)}$. Проводя линеариза-

*) Зависимость от y_n остается нелинейной, так что схема (50) в строгом смысле слова нелинейна; это надо учитывать при исследовании ее устойчивости.

**) См. гл. VIII, § 2, п. 5, где рассмотрена сходная система (8.71).

цию, получим довольно громоздкие уравнения, линейные и трехточечные относительно $\delta\hat{y}_n^{(s)}$:

$$\delta\hat{y}_{n+1}\left[\hat{\varkappa}_{n+1/2}+\frac{\partial\hat{\varkappa}_{n+1/2}}{\partial\hat{y}_{n+1}}(\hat{y}_{n+1}-\hat{y}_n)\right]-\delta\hat{y}_n\left[\frac{h^2}{\tau}+\hat{\varkappa}_{n+1/2}+\right.$$
$$\left.+\hat{\varkappa}_{n-1/2}-\frac{\partial\hat{\varkappa}_{n+1/2}}{\partial\hat{y}_n}(\hat{y}_{n+1}-\hat{y}_n)+\frac{\partial\hat{\varkappa}_{n-1/2}}{\partial\hat{y}_n}(\hat{y}_n-\hat{y}_{n-1})-h^2\frac{\partial\hat{\varphi}_n}{\partial\hat{y}_n}\right]+$$
$$+\delta\hat{y}_{n-1}\left[\hat{\varkappa}_{n-1/2}-\frac{\partial\hat{\varkappa}_{n-1/2}}{\partial\hat{y}_{n-1}}(\hat{y}_n-\hat{y}_{n-1})\right]= \tag{53}$$
$$=\frac{h^2}{\tau}(\hat{y}_n-y_n)-\hat{\varkappa}_{n+1/2}(\hat{y}_{n+1}-\hat{y}_n)+\hat{\varkappa}_{n-1/2}(\hat{y}_n-\hat{y}_{n-1})-h^2\hat{\varphi}_n,$$
$$\hat{y}_n^{(s+1)}=\hat{y}_n^{(s)}+\delta\hat{y}_n^{(s)};$$

индекс итерации s в основном уравнении опущен. На каждой итерации уравнения (53) решают прогонкой. Полученный итерационный процесс сходится, если шаг τ не слишком велик, причем вблизи корня сходимость квадратична. Если сходимость недостаточно быстрая (число итераций превышает 5 — 10), то целесообразнее не ограничивать число итераций, а уменьшать шаг τ.

Практика численных расчетов показала, что фактическая точность расчета по нелинейной схеме (51) обычно существенно лучше, чем по линейному варианту (50). Это позволяет вести расчет более крупным шагом τ, так что объем вычислений, требующийся для достижения заданной точности, получается меньше. Поэтому нелинейная схема (51), несмотря на свою сложность, выгоднее линейного варианта (50), особенно при решении *больших задач* *).

Включение точки. В квазилинейных задачах возможно обращение $k(u)$ в нуль при достаточно малых значениях температуры **). Наиболее типичным является случай $k(u)\approx\beta u^\alpha$, когда $k=0$ при $u=0$. При этом надо обращать особое внимание на выбор формулы типа (35а) для вычисления $\varkappa$. Например, полагать

$$\varkappa_{n+1/2}=\sqrt{k(y_n)\,k(y_{n+1})}$$

или

$$\varkappa_{n+1/2}=\frac{2k(y_n)\,k(y_{n+1})}{k(y_n)+k(y_{n+1})}$$

*) Большими задачами называют сложные задачи современной математической физики, описываемые системой большого числа уравнений в частных производных; в число этих уравнений может входить и уравнение теплопроводности.

**) В физических задачах при достаточно малой температуре теплопроводность становится пренебрежимо малой (исключая случаи сверхпроводимости и сверхтекучести); при высоких температурах теплопроводность заведомо не обращается в нуль.

нельзя: если в точке x_n начальная температура $y_n^0 = 0$, то в примыкающих к этой точке интервалах $\varkappa_{n-1/2} = \varkappa_{n+1/2} = 0$, и тепло в эту точку никогда не проникнет (точка «не включится»).

Поэтому надо выбирать такую формулу вычисления $\varkappa_{n+1/2} = \varkappa(y_n, y_{n+1})$, чтобы выполнялось $\varkappa_{n+1/2} \neq 0$, если $k(x, t, u) \neq 0$ хотя бы в некоторой части отрезка $[x_n, x_{n+1}]$. Этому условию удовлетворяют, например, формулы

$$\varkappa_{n+1/2} = k\left(x_{n+1/2},\ t,\ \frac{1}{2}(y_n + y_{n+1})\right) \tag{54а}$$

или

$$\varkappa_{n+1/2} = \frac{1}{2}[k(x_n,\ t,\ y_n) + k(x_{n+1},\ t,\ y_{n+1})]. \tag{54б}$$

При дважды непрерывно дифференцируемом коэффициенте теплопроводности на произвольных сетках, а при кусочно-непрерывном коэффициенте с кусочно-непрерывными вторыми производными — на специальных сетках они имеют аппроксимацию $O(h^2)$.

§ 2. Многомерное уравнение

1. Экономичные схемы. Для уравнения переноса хорошие одномерные схемы — схемы бегущего счета — естественно обобщались на случай многих измерений. Однако попытка обобщить на случай многих измерений хорошие одномерные схемы расчета теплопроводности — неявные схемы типа (6) и (34) — наталкивается на принципиальные трудности.

Рассмотрим их на примере двумерного уравнения теплопроводности с постоянным коэффициентом, для которого задана первая краевая задача в прямоугольной области:

$$u_t = k(u_{x_1x_1} + u_{x_2x_2}) + f(x_1,\ x_2,\ t), \quad k = \text{const} > 0, \tag{55а}$$
$$0 < x_1 < a, \quad 0 < x_2 < b, \quad 0 < t \leqslant T;$$
$$u(0,\ x_2,\ t) = \mu_1(x_2,\ t), \quad u(a,\ x_2,\ t) = \mu_2(x_2,\ t),$$
$$u(x_1,\ 0,\ t) = \mu_3(x_1,\ t), \quad u(x_1,\ b,\ t) = \mu_4(x_1,\ t), \tag{55б}$$
$$u(x_1,\ x_2,\ 0) = \mu(x_1,\ x_2).$$

Введем прямоугольную сетку $\{x_{1n},\ x_{2m},\ 0 \leqslant n \leqslant N,\ 0 \leqslant m \leqslant M\}$ (рис. 81), причем для простоты шаги по каждой переменной h_1, h_2 выберем постоянными. Возьмем изображенный на рис. 82 шаблон, имеющий на каждом слое форму креста, и составим на нем неявную двуслойную схему с весами, являющуюся обобщением схемы (6) на двумерный случай:

$$\frac{1}{\tau}(\hat{y}_{nm} - y_{nm}) = (\Lambda_1 + \Lambda_2)[\sigma\hat{y}_{nm} + (1-\sigma)\,y_{nm}], \tag{56}$$

где

$$\Lambda_1 y_{nm} = \frac{k}{h_1^2}(y_{n+1,\,m} - 2y_{nm} + y_{n-1,\,m}),$$
$$\Lambda_2 y_{nm} = \frac{k}{h_2^2}(y_{n,\,m+1} - 2y_{nm} + y_{n,\,m-1}). \tag{57}$$

Разностная запись первого краевого условия сводится к заданию решения $\hat{y}_{nm}$ в граничных узлах сетки, т. е. при $n=0$, $n=N$, $m=0$ и $m=M$.

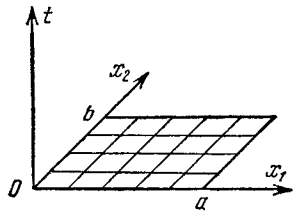


Рис. 81.

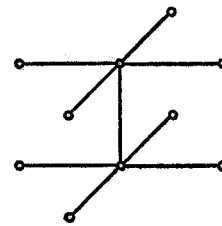


Рис. 82.

Легко проверить, что погрешность аппроксимации этой схемы на решениях с непрерывными четвертыми производными равна $O(\tau^\nu + h_1^\nu + h_2^\nu)$, где $\nu=2$ при $\sigma=1/2$ и $\nu=1$ при $\sigma\neq 1/2$. Методом разделения переменных, подставляя гармоники $y_{nm}=\exp(iqx_{1n}+ + irx_{2m})$ и определяя их множители роста ρ_{qr}, можно получить условие устойчивости схемы (56) в $\|\cdot\|_{l_2}$:

$$\sigma \geqslant \frac{1}{2} - \frac{1}{4k\tau}\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)^{-1}, \tag{58}$$

похожее на одномерное условие (14). При выполнении условия устойчивости (58) схема (56) среднеквадратично сходится с точностью $O(\tau^\nu + h_1^\nu + h_2^\nu)$.

Нетрудно написать обобщение схемы (56) и условия устойчивости (58) на любое число измерений p. Оценим число действий, требующихся для выполнения расчета до момента времени T по такой схеме в случае p измерений.

При $\sigma=0$ схема (56) становится явной и значение $\hat{y}_{nm}$ непосредственно вычисляется по значениям с предыдущего слоя. Поэтому общее число действий для перехода со слоя на слой пропорционально числу узлов сетки; оно $\sim N^p$, если число узлов по каждой пространственной переменной равно N. Но явная схема устойчива только при

$$2k\tau \leqslant (h_1^{-2} + h_2^{-2})^{-1} \sim N^{-2}.$$

Значит, для расчета до момента времени T надо сделать $\sim N^2$ шагов по времени и полный расчет потребует $\sim N^{p+2}$ действий.

Если вести расчет по абсолютно устойчивому варианту схемы ($\sigma \geqslant 1/2$), то можно брать $\tau \sim h$. Но тогда на каждом слое надо решать линейную систему N^p уравнений. Даже с учетом того, что ее матрица ленточная с шириной ленты $2N^{p-1}$, решение этой системы методом Гаусса требует $\sim N^{3p-2}$ действий. Поскольку для расчета до момента T теперь надо делать N шагов по времени, то полный расчет требует $\sim N^{3p-1}$ действий.

Значит, для двумерной задачи ($p=2$) неявная схема (56) и явная схема приводят примерно к одинаковому объему вычислений, а в § 1 мы видели, что явная схема обладает плохими свойствами и невыгодна для расчетов. При $p \geqslant 2$ неявная схема (56) даже невыгодней явной.

Однако для многомерного параболического уравнения построены абсолютно устойчивые схемы, позволяющие вести расчет шагом $\tau \sim h$ и требующие только $\sim N^p$ действий для перехода со слоя на слой (т. е. число действий в расчете на одну точку сетки не зависит от шагов h_α). Такие схемы называются *экономичными*. Подавляющее большинство многомерных расчетов проводится по таким схемам. В следующих пунктах мы рассмотрим два основных вида экономичных схем для параболического уравнения — продольно-поперечную и локально-одномерную схемы.

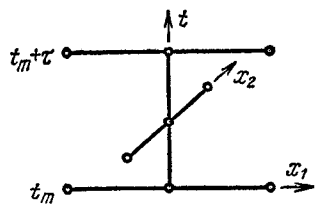


Рис. 83.

2. Продольно-поперечная схема, называемая также схемой переменных направлений, является одной из лучших двумерных экономичных схем. Выберем изображенный на рис. 83 шаблон, содержащий полуцелый слой $\bar{t} = t + \tau/2$, и составим на нем эту схему:

$$\frac{1}{0{,}5\tau}(\bar{y}_{nm} - y_{nm}) = \Lambda_1 \bar{y}_{nm} + \Lambda_2 y_{nm} + \bar{f}_{nm}, \tag{59а}$$

$$\frac{1}{0{,}5\tau}(\hat{y}_{nm} - \bar{y}_{nm}) = \Lambda_1 \bar{y}_{nm} + \Lambda_2 \hat{y}_{nm} + \bar{f}_{nm}, \tag{59б}$$

где разностные операторы Λ_α определены формулами (57). Как обычно, под $\bar{y}$ подразумевается значение на полуцелом слое $\bar{t}$.

Исследуем продольно-поперечную схему.

Вычисление разностного решения. Переход на полуцелый слой делается при помощи уравнений (59а). Согласно определению оператора Λ_1 каждое такое уравнение содержит три неизвестных значения: $\bar{y}_{n-1,m}$, $\bar{y}_{nm}$ и $\bar{y}_{n+1,m}$; остальные значения y берутся с исходного слоя. Иными словами, при таком переходе схема неявна по направлению x_1 и явна по направлению x_2. При любом фиксированном индексе m уравнения (59а) образуют относительно неизвестных $\bar{y}_{nm}$ линейную систему с трехдиагональной

матрицей. Поэтому значения $\bar{y}_{nm}$ легко вычисляются одномерной прогонкой по индексу n, т. е. по направлению x_1.

Наоборот, при переходе при помощи уравнений (59б) с полуцелого слоя на целый схема явна по направлению x_1 и неявна по x_2. Поэтому решение $\hat{y}_{nm}$ на целом слое вычисляется тоже одномерной прогонкой, но в поперечном направлении x_2. Нетрудно подсчитать, что для перехода с целого на целый слой нужно всего $20-30$ действий на каждую точку сетки независимо от величин шагов, так что схема экономична.

Как и в одномерной схеме (6), диагональные матричные элементы в уравнениях (59) преобладают; следовательно, прогонка устойчива, а разностное решение существует и единственно.

Устойчивость продольно-поперечной схемы исследуем методом разделения переменных. Множители роста гармоники на первом и втором полушаге по времени могут быть различными. Поэтому положим

$$y_{nm}=\exp(iqx_{1n}+irx_{2m}),\quad \bar{y}=\rho'_{qr}y,\quad \hat{y}=\rho''_{qr}\bar{y}. \tag{60}$$

Подставляя соотношения (60) в схему (59), получим множители роста

$$\rho'_{qr}=\left(1-\frac{2k\tau}{h_2^2}\sin^2\frac{rh_2}{2}\right)\Big/\left(1+\frac{2k\tau}{h_1^2}\sin^2\frac{qh_1}{2}\right), \tag{61а}$$

$$\rho''_{qr}=\left(1-\frac{2k\tau}{h_1^2}\sin^2\frac{qh_1}{2}\right)\Big/\left(1+\frac{2k\tau}{h_2^2}\sin^2\frac{rh_2}{2}\right). \tag{61б}$$

Нетрудно заметить, что для всех гармоник при любых шагах выполняется неравенство $|\rho'_{qr}\rho''_{qr}|\leqslant 1$. Таким образом, при переходе с одного целого слоя на следующий целый слой ошибки начальных данных не нарастают, и *схема* (59) *равномерно и безусловно устойчива по начальным данным*.

Нетрудно проверить, что дополнительный признак устойчивости по правой части (9.54) выполняется на каждом полушаге по времени. Следовательно, схема (59) устойчива по правой части.

Замечание 1. Если $k\tau>h_1^2$, то существуют такие гармоники, которые усиливаются при переходе с целого слоя на полуцелый; например, $|\rho'_{\text{ом}}|>1$. Зато при переходе с полуцелого на следующий целый слой эти гармоники настолько затухают, что в целом усиления не происходит. Аналогично, при $k\tau>h_2^2$ есть гармоники, усиливающиеся при переходе с полуцелого слоя на целый.

Замечание 2. Суммарный множитель роста $\rho_{qr}=\rho'_{qr}\rho''_{qr}$ таков, что $|\rho_{qr}|=1$ только при $q=r=0$; для всех остальных гармоник $|\rho_{qr}|<1$. Следовательно, продольно-поперечная схема обладает аппроксимационной вязкостью и расчет по ней должен приводить к сглаживанию разрывов.

Аппроксимация. При переходе с целого на полуцелый слой каждая пространственная разность вычисляется несимметрично по времени и погрешность равна $O(\tau+h^2)$. Но ошибка на второй половине слоя компенсирует первую, и в итоге при переходе с целого слоя на целый погрешность локальной аппроксимации на равномерных сетках есть $O(\tau^2+h_1^2+h_2^2)$.

В этом легко убедиться при помощи следующего преобразования. Вычитая уравнение (59б) из (59а), получим

$$\bar{y}_{nm}=\frac{1}{2}(\hat{y}_{nm}+y_{nm})-\frac{\tau}{4}\Lambda_2(\hat{y}_{nm}-y_{nm}). \tag{62}$$

Сложим уравнения (59) и подставим в них полученное значение $\bar{y}_{nm}$, исключив тем самым полуцелый слой:

$$\frac{1}{\tau}(\hat{y}_{nm}-y_{nm})=\frac{1}{2}(\Lambda_1+\Lambda_2)(\hat{y}_{nm}+y_{nm})-\frac{\tau}{4}\Lambda_1\Lambda_2(\hat{y}_{nm}-y_{nm})+\bar{f}_{nm}. \tag{63}$$

Предпоследний член справа есть $\tau^2 u_{x_1^2x_2^2t}/4=O(\tau^2)$, а остальные члены в (63) совпадают с симметричным вариантом схемы (56), который соответствует $\sigma=1/2$ и имеет аппроксимацию $O(\tau^2+h_1^2+h_2^2)$. Поскольку продольно-поперечная схема отличается от этого варианта на член $O(\tau^2)$, она также имеет второй порядок аппроксимации по всем переменным.

Остановимся на аппроксимации краевых условий (55б). На целом слое в уравнения продольно-поперечной схемы (59) входят значения решения $\hat{y}_{nm}$ на сторонах прямоугольника $x_2=0$ и $x_2=b$; очевидно, надо положить

$$\hat{y}_{n0}=\mu_3(x_{1n},\hat{t}),\quad \hat{y}_{nM}=\mu_4(x_{1n},\hat{t}),\qquad 1\leqslant n\leqslant N-1. \tag{64а}$$

Для полуцелого слоя требуются значения $\bar{y}_{nm}$ на сторонах $x_1=0$ и $x_1=a$. Полагать $(\bar{y}-\bar{\mu}_\alpha)_{\text{гран}}=0$ невыгодно, ибо полуцелый слой не вполне соответствует моменту $\hat{t}$ и такая аппроксимация внесла бы погрешность $O(\tau)$. Следует воспользоваться уравнением (62), отнесенным к стороне $x_1=0$:

$$\bar{y}_{0m}=\frac{1}{2}(\hat{\mu}_{1m}+\mu_{1m})-\frac{\tau}{4}\Lambda_2(\hat{\mu}_{1m}-\mu_{1m}),\quad 1\leqslant m\leqslant M-1; \tag{64б}$$

аналогичное условие записывается для стороны $x_1=a$. Граничные условия (64) обеспечивают погрешность аппроксимации $O(\tau^2)$.

Сходимость. Проведенное исследование аппроксимации и устойчивости показывает, что схема (59) безусловно сходится в $\|\cdot\|_{l_2}$, причем в прямоугольной области на равномерной сетке и при краевом условии (64) она имеет точность $O(\tau^2+h_1^2+h_2^2)$ на решениях с непрерывными пятыми производными.

Более сложными методами можно доказать равномерную сходимость со вторым порядком точности.

Отметим некоторые усложнения исходной задачи (55).

Произвольная область. Пусть для уравнения (55а) в области произвольной формы заданы краевые условия первого рода

$$[u-\mu(x,\ t)]_\Gamma=0.$$

Тогда разностное краевое условие (64б) не удается применить, ибо неясно, как вычислять $\Lambda_2\mu$. Приходится ограничиться условиями

$$\begin{aligned}[\bar{y}-\mu(x,\ \bar{t})]_\gamma&=0,\\ [\hat{y}-\mu(x,\ \hat{t})]_\gamma&=0,\end{aligned}\tag{65}$$

где γ — множество граничных узлов. Погрешность аппроксимации условия (65) на полуцелом слое равна $O(\tau)$. Поэтому в произвольной области схема (59) сходится с точностью $O(\tau+h_1^2+h_2^2)$.

Если область ступенчатая, т. е. составлена из прямоугольников со сторонами, параллельными осям координат, то в ней можно написать краевое условие повышенной точности (64б). В этом случае схема (59) имеет второй порядок точности.

Переменный коэффициент теплопроводности. Для уравнения теплопроводности с переменным коэффициентом можно составить два варианта продольно-поперечной схемы, являющихся обобщением наилучшей схемы (34). В первом варианте на всех слоях — исходном, полуцелом и новом целом — разностный коэффициент теплопроводности $\varkappa$ приписывают полуцелому слою $\bar{t}$; во втором варианте на этих слоях берут соответственно $\varkappa(t)$, $\varkappa(\bar{t})$ и $\varkappa(\hat{t})$.

Оба варианта успешно применяются на практике. Второй вариант лучше исследован теоретически; для него доказана безусловная сходимость в прямоугольной области с точностью $O(\tau^2+h_1^2+h_2^2)$, если коэффициент $k(x,\ t)$ непрерывен со своими вторыми производными.

Анизотропная теплопроводность в простейшем случае приводит к тому, что по каждому направлению имеется свой коэффициент $k_\alpha(x,\ t)$. В этом случае уравнение теплопроводности принимает вид

$$\frac{\partial u}{\partial t}=\sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}\left[k_\alpha(x,\ t)\frac{\partial u}{\partial x_\alpha}\right]+f(x,\ t).\tag{66}$$

Продольно-поперечная схема, ее обобщения и все теоретические обоснования переносятся на этот случай практически без изменений.

3. Локально-одномерный метод. Продольно-поперечная схема на задачи с числом измерений $p\geqslant 3$ непосредственно не обобщается. В самом деле, введем $p-1$ промежуточный слой и на

каждом слое составим схему типа (59), неявную по одному направлению и явную по остальным. Во-первых, такая схема несимметрична и имеет аппроксимацию лишь $O(\tau)$. Во-вторых, она оказывается условно устойчивой при $k\tau \lesssim h^2$ и, тем самым, неэкономичной.

Экономичные многомерные разностные схемы можно строить локально-одномерным методом, также используя промежуточные слои. Эти схемы имеют лишь суммарную аппроксимацию. На промежуточных слоях они вообще не аппроксимируют исходное дифференциальное уравнение; но погрешности аппроксимации промежуточных слоев при суммировании гасят друг друга так, что на целом слое аппроксимация есть. При этом разностное решение следует сравнивать с точным только на целых слоях, не придавая промежуточным слоям самостоятельного смысла.

Рассмотрим многомерное параболическое уравнение (66); для простоты ограничимся случаем анизотропной теплопроводности с постоянными коэффициентами:

$$\frac{\partial u}{\partial t} = \sum_{\alpha=1}^{p} A_\alpha u, \quad A_\alpha = k_\alpha \frac{\partial^2}{\partial x_\alpha^2}, \quad k_\alpha = \text{const}, \tag{67}$$

$$x = \{x_1, x_2, \ldots, x_p\}.$$

Аппроксимируем это уравнение симметричной неявной схемой, которую назовем *исходной*:

$$\frac{1}{\tau}(\hat{y} - y) = \frac{1}{2} \sum_{\alpha=1}^{p} \Lambda_\alpha (\hat{y} + y), \tag{68}$$

где Λ_α — разностные операторы, аппроксимирующие A_α с погрешностью $O(h^r)$; обычно для них используют формулы (57), соответствующие $r = 2$. Благодаря симметричной форме исходная схема имеет погрешность

$$O\left(\tau^2 + \sum_\alpha h_\alpha^r\right).$$

Однако эта схема неэкономична, потому что не найдено хорошего алгоритма вычисления $\hat{y}$.

Наряду с исходной схемой построим *локально-одномерную* схему. Введем промежуточные слои и на каждом слое в правой части (68) вместо $\sum_\alpha \Lambda_\alpha$ возьмем $p\Lambda_\alpha$; в левой части поставим шаг τ/p. Обозначим решение на промежуточных шагах через w_α ($\alpha = 1, 2, \ldots, p$). Тогда функции w_α будут удовлетворять разно-

стным уравнениям и начальным условиям следующего вида:

$$\frac{1}{\tau}(\hat{w}_\alpha - w_\alpha) = \frac{1}{2}\Lambda_\alpha(\hat{w}_\alpha + w_\alpha), \quad \alpha = 1, 2, \ldots, p; \tag{69а}$$

$$w_1 = y, \quad w_2 = \hat{w}_1, \quad w_3 = \hat{w}_2, \quad \ldots, \quad w_p = \hat{w}_{p-1}, \quad \hat{y} = \hat{w}_p. \tag{69б}$$

Поскольку Λ_α — одномерные операторы, то каждая $\hat{w}_\alpha$ является решением одномерной разностной схемы; поэтому схему (69) называют локально-одномерной. Исследуем ее.

Устойчивость. Каждое уравнение (69а) является одномерной неявной симметричной схемой типа схемы (6) при $\sigma = 1/2$. Последняя схема безусловно устойчива, так что ошибка начальных данных не возрастает ни на одном промежуточном слое. Следовательно, схема (69) также безусловно устойчива и позволяет вести расчет с шагом $\tau \sim h$.

Вычисление разностного решения несложно. Каждое уравнение (69а) решается одномерной прогонкой. По тем же причинам, что и в случае схемы (6), прогонка устойчива, а разностное решение $\hat{y}$ существует и единственно.

Для нахождения решения на новом целом слое надо выполнить прогонки по всем p направлениям. Это требует $\sim 10p$ действий на каждую точку сетки независимо от величин шагов h_α. Таким образом, локально-одномерная схема экономична.

Аппроксимацию исследуем, сравнивая схему (69) с исходной. Для этого перепишем (69) в следующем виде:

$$\left(E - \frac{1}{2}\tau\Lambda_\alpha\right)\hat{w}_\alpha = \left(E + \frac{1}{2}\tau\Lambda_\alpha\right)w_\alpha, \qquad w_\alpha = \hat{w}_{\alpha-1}. \tag{70}$$

Операторы A_α попарно перестановочны; операторы Λ_α получаются тоже попарно перестановочными. Последовательно применяя (70) и используя перестановочность операторов, нетрудно установить следующее равенство:

$$\left\{\prod_{\alpha=1}^{p}\left(E - \frac{1}{2}\tau\Lambda_\alpha\right)\right\}\hat{w}_p = \left\{\prod_{\alpha=1}^{p}\left(E + \frac{1}{2}\tau\Lambda_\alpha\right)\right\}w_1.$$

Раскроем произведения операторов и положим $w_1 = y$, $\hat{w}_p = \hat{y}$. Пренебрегая членами высокого порядка по τ, получим запись схемы (69):

$$\left\{E - \frac{\tau}{2}\sum_{\alpha}\Lambda_\alpha + \frac{\tau^2}{4}\sum_{\alpha,\,\beta}\Lambda_\alpha\Lambda_\beta + O(\tau^3)\right\}\hat{y} =$$

$$= \left\{E + \frac{\tau}{2}\sum_{\alpha}\Lambda_\alpha + \frac{\tau^2}{4}\sum_{\alpha,\,\beta}\Lambda_\alpha\Lambda_\beta + O(\tau^3)\right\}y,$$

или

$$\frac{1}{\tau}(\hat{y}-y)=\frac{1}{2}\sum_{\alpha=1}^{p}\Lambda_\alpha(\hat{y}+y)-\frac{\tau}{4}\sum_{\alpha,\,\beta}\Lambda_\alpha\Lambda_\beta(\hat{y}-y)+O(\tau^2). \qquad (71)$$

На решениях с непрерывными пятыми производными двойная сумма в (71) есть $O(\tau^2)$, поэтому (71) отличается от исходной схемы только членами $O(\tau^2)$. Но погрешность аппроксимации исходной схемы равна

$$O\left(\tau^2+\sum_{\alpha=1}^{p}h_\alpha^2\right).$$

Следовательно, погрешность аппроксимации симметричной локально-одномерной схемы (69) на целых слоях есть

$$O\left(\tau^2+\sum_{\alpha=1}^{p}h_\alpha^2\right).$$

Заметим, что для получения погрешности аппроксимации $O(\tau^2)$ в граничных условиях надо к естественным граничным условиям добавлять поправки типа (64б).

Сходимость схемы (69), как следует из сказанного выше, является безусловной с погрешностью $O\left(\tau^2+\sum_\alpha h_\alpha^2\right)$.

Замечание. В некоторых случаях расщепление многомерной задачи на последовательность одномерных бывает точным. Например, многомерный перенос по характеристике точно эквивалентен последовательности одномерных переносов по проекциям этой характеристики на координатные плоскости.

Остановимся на некоторых усложнениях задачи (67).

Переменные коэффициенты $k_\alpha(\boldsymbol{x}, t)$ приводят к тому, что операторы A_α становятся неперестановочными и Λ_α — тоже. В этом случае погрешность аппроксимации схемы (69) возрастает до $O\left(\tau+\sum_\alpha h_\alpha^2\right)$. Поэтому для уравнения

$$\frac{\partial u}{\partial t}=\sum_{\alpha=1}^{p}A_\alpha u+f(\boldsymbol{x}, t), \qquad A_\alpha u=\frac{\partial}{\partial x_\alpha}\left[k_\alpha(\boldsymbol{x}, t)\frac{\partial u}{\partial x_\alpha}\right], \qquad (72)$$

нередко ограничиваются чисто неявной локально-одномерной схемой

$$\frac{1}{\tau}(\hat{w}_\alpha-w_\alpha)=\Lambda_\alpha\hat{w}_\alpha+\varphi_\alpha, \qquad w_\alpha=\hat{w}_{\alpha-1}, \qquad \sum_{\alpha=1}^{p}\varphi_\alpha=f \qquad (73а)$$

с естественными граничными условиями

$$[\hat{w}_\alpha-\mu(\boldsymbol{x}, \hat{t}]_\gamma=0; \qquad (73б)$$

здесь операторы Λ_α построены по образцу одномерной наилучшей схемы (34). Схема (73) безусловно устойчива и имеет точность

$$O\left(\tau+\sum_\alpha h_\alpha^2\right)$$

в норме $\|\cdot\|_c$.

Для уравнения (72) можно добиться точности $O(\tau^2)$, строя симметричный по времени алгоритм. Введем полуцелый слой $\bar{t}$ и перейдем на него по симметричной локально-одномерной схеме (69) в прямом порядке $\alpha=1,2,\ldots,p$. Переход с полуцелого на новый слой $\hat{t}$ совершим по той же схеме, но в обратном порядке $\alpha=p,\ p-1,\ \ldots,\ 1$. При этом в естественные краевые условия надо вносить поправки, аналогичные (64б).

Квазилинейное уравнение с $k_\alpha(x,\ t,\ u)$. Чисто неявная локально-одномерная схема (73) естественно обобщается на этот случай. Аналогично § 1, п. 8, можно на промежуточном слое либо полагать $\varkappa_\alpha=\varkappa_\alpha(x,\ t,\ w_\alpha)$ и обходиться однократной прогонкой по данному направлению, либо полагать $\varkappa_\alpha=\varkappa_\alpha(x,\ \hat{t},\ \hat{w}_\alpha)$ и решать одномерную схему (73а) прогонкой с итерациями.

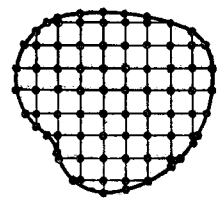


Рис. 84.

Произвольная область $G(x)$ с криволинейной границей. Покроем эту область прямоугольной сеткой, равномерной по каждой переменной (двумерный случай изображен на рис. 84). Точки пересечения линий сетки с границей также возьмем в качестве узлов сетки и запишем в них естественное разностное краевое условие (73б). Во внутренних узлах аппроксимируем дифференциальное уравнение (72) чисто неявной локально-одномерной схемой (73а).

Пусть граничные значения $\mu(x,\ t)$ и коэффициенты уравнения (72) достаточно гладки, так что точное решение $u(x,\ t)$ непрерывно вместе со своими четвертыми производными всюду в $G(x)$, включая границу области. Тогда построенная указанным образом схема безусловно устойчива и равномерно сходится с точностью

$$O\left(\tau+\sum_\alpha h_\alpha^2\right)$$

(доказательство см. в [30]).

В областях специальной формы — сфере или цилиндре — удобнее пользоваться не декартовыми координатами, а соответствующими криволинейными. Это позволяет получить более хорошую аппроксимацию вблизи границы и повышает фактическую точность расчета. Но при этом есть тонкости в аппроксимации вблизи центра или оси, на которых мы не останавливаемся.

4. Метод Монте-Карло. Этот метод можно применять к задачам, которые обычно формулируют в терминах уравнений с частными производными. Рассмотрим его на несложном примере.

Пусть частицы блуждают по узлам двумерной пространственной сетки (рис. 84) так, что за один шаг по времени частица может перейти с вероятностью $^1/_4$ в любой из четырех соседних узлов. Тогда, если на данном шаге в узле есть y_{nm} частиц, то на следующем шаге все они уйдут в соседние узлы. Но зато из каждого соседнего узла примерно четверть бывших там частиц придет в этот узел, так что

$$\hat{y}_{nm} = \frac{1}{4}(y_{n+1,\,m} + y_{n-1,\,m} + y_{n,\,m+1} + y_{n,\,m-1}).$$

Вычитая из обеих частей y_{nm}, запишем

$$\hat{y}_{nm} - y_{nm} = \frac{1}{4}[(y_{n+1,\,m} - 2y_{nm} + y_{n-1,\,m}) + (y_{n,\,m+1} - 2y_{nm} + y_{n,\,m-1})]. \qquad (74)$$

Уравнение (74) совпадает с явным вариантом разностной схемы (56) для уравнения теплопроводности, если в этой схеме положить $\sigma = 0$ и выбрать шаги специальным образом:

$$4k\tau = h_1^2 = h_2^2.$$

Поэтому вместо решения разностных уравнений можно разыграть случайный процесс. Поместим в каждый узел сетки число частиц, пропорциональное начальному значению y_{nm}^0. На каждом шаге для каждой частицы будем разыгрывать переход в один из соседних узлов. Перераспределение частиц будет соответствовать изменению решения со временем.

Вопрос о границе и условиях на ней довольно сложен и здесь не рассматривается.

В обычных задачах теплопроводности этот метод гораздо менее точен, чем локально-одномерные методы. Но в очень сложных задачах, где число измерений велико и написать разностную схему трудно, метод Монте-Карло может оказаться более простым и быстрым способом решения.

ЗАДАЧИ

1. Найти невязку схемы (6) с весом и правой частью (11).

2. Найти невязку схемы (6) с весом (11) и правой частью (12).

3. Записать схему (6) на неравномерной сетке и найти ее погрешность локальной аппроксимации: а) на произвольной неравномерной сетке, б) на квазиравномерной сетке.

4. При каком соотношении шагов τ и h будет асимптотически устойчива схема повышенной точности (6) с весом (11)?

5. Исследовать аппроксимацию схемы Ричардсона (26).

6. Доказать безусловную устойчивость схемы Дюфорта — Франкела (28).

7. Исследовать аппроксимацию схемы Дюфорта — Франкела (28).

8. Доказать безусловную устойчивость схемы (29).

9. Найти невязки схем (29а) и (29б) и определить суммарную невязку схемы (29).

10. Для уравнения $u_t = ku_{xx}$ построить схему на шаблоне рис. 85 и доказать, что она устойчива при $2k\tau \geqslant h^2$.

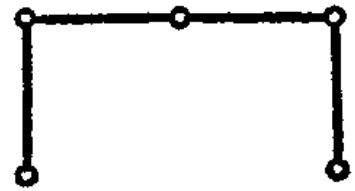


Рис. 85.

11. Доказать, что наилучшая схема (34) монотонна при выполнении условия (42).

12. Исследовать устойчивость схемы (46) для параболического уравнения в криволинейных координатах.

13. Исследовать аппроксимацию схемы (56) для двумерного уравнения (55).

14. Доказать, что двумерная схема (56) устойчива при выполнении условия (58).

15. Разобрать структуру матрицы линейной системы (56). Как изменится эта структура при обобщении схемы (56) на случай трех измерений?

ГЛАВА XII

ЭЛЛИПТИЧЕСКИЕ УРАВНЕНИЯ

Глава XII посвящена методам решения краевых задач для эллиптических уравнений. В § 1 решение таких задач сводится к решению эволюционных задач для параболических уравнений до выхода на стационарный режим; последнее выполняется при помощи многомерных разностных схем, изложенных в гл. XI, § 2. Обсужден выбор оптимального шага по времени (или набора переменных шагов) в таких расчетах.

В § 2 рассмотрены вариационные методы решения эллиптических уравнений и вариационные способы составления стационарных (не эволюционных) разностных схем. В последнем случае указаны прямые и итерационные методы вычисления разностного решения.

§ 1. Счет на установление

1. Стационарные решения эволюционных задач. К эллиптическим уравнениям приводит ряд физических задач: определение прогиба нагруженной мембраны, давления газа в неоднородном силовом поле, стационарного (не зависящего от времени) распределения тепла в теле и т. д. Все эти задачи имеют общее свойство: предполагается, что внешние воздействия не зависят от времени, а начальные условия были заданы достаточно давно, так что физическая система успела выйти на стационарное решение $u(\boldsymbol{r})$, не зависящее от времени.

Примером полной математической постановки является задача с краевыми условиями первого рода, называемая задачей Дирихле; требуется найти непрерывное решение задачи

$$\Delta u(\boldsymbol{r}) = -f(\boldsymbol{r}), \quad \boldsymbol{r} \in G, \qquad u_\Gamma(\boldsymbol{r}) = \mu(\boldsymbol{r}), \tag{1}$$

где $G(\boldsymbol{r})$ есть многомерная замкнутая область с границей Г. В отличие от эволюционных задач, разобранных в предыдущих главах, постановка (1) не содержит начальных условий. Обобщением задачи (1) является следующая задача:

$$\operatorname{div}[k(\boldsymbol{r}) \operatorname{grad} u(\boldsymbol{r})] = -f(\boldsymbol{r}), \quad \boldsymbol{r} \in G, \qquad u_\Gamma(\boldsymbol{r}) = \mu(\boldsymbol{r}), \quad k(\boldsymbol{r}) > 0. \tag{2}$$

Задачи с другими краевыми условиями мы не будем рассматривать.

Задачу (2) будем называть *стационарной*. Наряду с ней рассмотрим *эволюционную* задачу для параболического уравнения с теми же граничными условиями и произвольно выбранными начальными данными:

$$\frac{\partial v(\boldsymbol{r}, t)}{\partial t} = \operatorname{div}[k(\boldsymbol{r}) \operatorname{grad} v(\boldsymbol{r}, t)] + f(\boldsymbol{r}), \quad \boldsymbol{r} \in G, \quad 0 \leqslant t < +\infty, \tag{3}$$
$$v_\Gamma(\boldsymbol{r}, t) = \mu(\boldsymbol{r}), \quad v(\boldsymbol{r}, 0) = v_0(\boldsymbol{r}).$$

Исследуем, насколько решение $v(\boldsymbol{r}, t)$ эволюционной задачи отличается от решения $u(\boldsymbol{r})$ стационарной задачи. Вычитая (2) из (3) и учитывая, что $\partial u(\boldsymbol{r})/\partial t = 0$, найдем, что разность $w(\boldsymbol{r}, t) = v(\boldsymbol{r}, t) - u(\boldsymbol{r})$ удовлетворяет однородному параболическому уравнению с однородными краевыми условиями:

$$\frac{\partial}{\partial t} w(\boldsymbol{r}, t) = \operatorname{div}[k(\boldsymbol{r}) \operatorname{grad} w], \quad \boldsymbol{r} \in G, \quad 0 < t < +\infty, \tag{4}$$
$$w_\Gamma(\boldsymbol{r}, t) = 0, \quad w(\boldsymbol{r}, 0) = w_0(\boldsymbol{r}) \equiv v_0(\boldsymbol{r}) - u(\boldsymbol{r}).$$

Поскольку начальные данные в (3) были выбраны произвольно, то без ограничения общности можно считать, что начальные данные задачи (4) также выбраны произвольно.

В курсах математической физики показано (см., например, [40]), что при помощи метода разделения переменных решение задачи (4) можно представить в следующем виде:

$$w(\boldsymbol{r}, t) = \sum_{q=1}^{\infty} c_q e^{-\lambda_q t} w_q(\boldsymbol{r}). \tag{5}$$

Здесь $w_q(\boldsymbol{r})$ и λ_q — собственные функции и собственные значения многомерной задачи Штурма — Лиувилля:

$$\operatorname{div}[k(\boldsymbol{r}) \operatorname{grad} w_q] + \lambda_q w_q(\boldsymbol{r}) = 0, \quad \boldsymbol{r} \in G, \quad w_q(\boldsymbol{r})_\Gamma = 0, \tag{6}$$

а

$$c_q = \int_G w(\boldsymbol{r}, 0)\, w_q(\boldsymbol{r})\, d\boldsymbol{r}$$

являются коэффициентами Фурье начальных данных (4) по системе функций $w_q(\boldsymbol{r})$. Собственные значения задачи (6) положительны и образуют неубывающую последовательность

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \lambda_3 \leqslant \ldots, \tag{7}$$

а собственные функции $w_q(\boldsymbol{r})$ образуют полную ортонормированную систему в $G(\boldsymbol{r})$.

Из (5) и (7) нетрудно получить неравенство

$$\| w(\boldsymbol{r}, t) \|_{L_2} = \left(\sum_{q=1}^{\infty} c_q^2 e^{-2\lambda_q t} \right)^{1/2} \leqslant e^{-\lambda_1 t} \left(\sum_{q=1}^{\infty} c_q^2 \right)^{1/2} = e^{-\lambda_1 t} \| w_0(\boldsymbol{r}) \|_{L_2}. \tag{8}$$

Оно означает, что разность $w(\boldsymbol{r}, t) = v(\boldsymbol{r}, t) - u(\boldsymbol{r})$ при $t \to \infty$ экспоненциально стремится к нулю по норме $\| \cdot \|_{L_2}$, так что *решение* $v(\boldsymbol{r}, t)$ *эволюционной задачи* (3) *среднеквадратично сходится к решению* $u(\boldsymbol{r})$ *стационарной задачи* (2) *при* $t \to \infty$.

Замечание 1. Пусть граничные и начальные условия таковы, что решения задач (2) и (3) имеют в $G(\boldsymbol{r})$ непрерывные производные, ограниченные равномерно по t. Тогда сходимость $v(\boldsymbol{r}, t)$ к $u(\boldsymbol{r})$ будет равномерной.

Таким образом, вместо задачи (2) для эллиптического уравнения можно взять эволюционную задачу (3) для параболического уравнения с тем же пространственным оператором, произвольно выбрать начальные данные и вычислить решение $v(\boldsymbol{r}, t)$ при достаточно большом t. Стационарный (не зависящий от времени) предел $u(\boldsymbol{r})$, к которому стремится $v(\boldsymbol{r}, t)$ при $t \to \infty$, и будет решением стационарной задачи (2).

Этот способ называется *счетом на установление*. Он позволяет осуществить численное решение эллиптических задач хорошо разработанными методами решения параболических задач, например, продольно-поперечной схемой для двумерных задач и локально-одномерными схемами в случае большего числа измерений.

Установление стационарного решения происходит довольно быстро благодаря экспоненциальному характеру затухания начальных данных. Из (8) видно, что если нужна точность $\sim \varepsilon$, то надо вести вычисления до момента

$$T \approx \frac{1}{\lambda_1} \ln \frac{1}{\varepsilon}, \tag{9}$$

где λ_1 есть наименьшее собственное значение соответствующей задачи Штурма — Лиувилля (6).

Замечание 2. На стационарное решение выходят не только решения параболических задач. То же происходит при других диссипативных процессах со стационарными граничными условиями, например при колебаниях с вязким трением, описываемых уравнением

$$v_{tt} + \beta v_t = \operatorname{div}(k \operatorname{grad} v) + f, \quad v_\Gamma = \mu(\boldsymbol{r}), \quad \beta > 0. \tag{10}$$

Можно формулировать эволюционную задачу для этого уравнения; однако это менее удобно.

Замечание 3. Можно составить разностную схему, непосредственно аппроксимирующую исходную задачу (2). Но в § 2

мы увидим, что вычислять разностное решение при этом обычно приходится итерационными методами. Оказывается, что соответствующие итерационные алгоритмы можно интерпретировать как некоторые разностные схемы для эволюционной задачи (3).

2. Оптимальный шаг. Для расчета эволюционной p-мерной задачи (3) до момента T используют экономичные разностные схемы. При этом шаги τ и h_α $(1 \leqslant \alpha \leqslant p)$ выбирают достаточно малыми, чтобы обеспечить требуемую близость разностного решения y к точному решению $v(\boldsymbol{r}, t)$ эволюционной задачи.

Однако если шаг τ выбран слишком малым, то расчет до момента T потребует большого числа шагов, что неоправданно увеличит объем вычислений. Очевидно, должен существовать оптимальный шаг τ_0; рассмотрим, как его найти.

Для простоты ограничимся двумерной задачей Дирихле в прямоугольнике:

$$\begin{gathered} u_{x_1x_1} + u_{x_2x_2} = -f(\boldsymbol{x}), \quad 0 < x_1 < a, \quad 0 < x_2 < b, \\ u_\Gamma(\boldsymbol{x}) = \mu(\boldsymbol{x}), \quad \boldsymbol{x} = \{x_1, x_2\}. \end{gathered} \tag{11}$$

Ей соответствует эволюционная задача для уравнения

$$v_t = v_{x_1x_1} + v_{x_2x_2} + f(\boldsymbol{x}), \tag{12}$$

которую будем решать на равномерной сетке $\{x_{1n} = nh_1, \; x_{2m} = mh_2, \; 0 \leqslant n \leqslant N, \; 0 \leqslant m \leqslant M\}$ с шагами $h_1 = a/N$, $h_2 = b/M$.

Продольно-поперечная схема. Для исследования этой схемы возьмем ее запись (11.63) в двуслойной форме:

$$\frac{1}{\tau}(\hat{y} - y) = \frac{1}{2}(\Lambda_1 + \Lambda_2)(\hat{y} + y) - \frac{\tau}{4}\Lambda_1\Lambda_2(\hat{y} - y) + \bar{f},$$

и преобразуем ее к канонической форме:

$$B\frac{\hat{y} - y}{\tau} + Ay = \varphi, \qquad \varphi = \bar{f}, \tag{13а}$$

где

$$A = -(\Lambda_1 + \Lambda_2), \quad B = \left(E - \frac{\tau}{2}\Lambda_1\right)\left(E - \frac{\tau}{2}\Lambda_2\right). \tag{13б}$$

Поскольку в уравнении (12) коэффициент теплопроводности $k = 1$, а сетка равномерна, то

$$\begin{aligned} \Lambda_1 y_{nm} &= \frac{1}{h_1^2}(y_{n+1,\,m} - 2y_{nm} + y_{n-1,\,m}), \\ \Lambda_2 y_{nm} &= \frac{1}{h_2^2}(y_{n,\,m+1} - 2y_{nm} + y_{n,\,m-1}). \end{aligned} \tag{13в}$$

Если численный расчет доведен до выхода на стационарное решение, то $\hat{y} \approx y$. Тогда схема (13) в пределе переходит в не-

эволюционную (не содержащую времени) разностную схему

$$Ay=\varphi, \quad A=-(\Lambda_1+\Lambda_2), \quad \varphi=f(x), \tag{14}$$

которая, как нетрудно заметить, аппроксимирует стационарную задачу (11). Очевидно, в этом случае оптимальным будет тот шаг τ_0, при котором разностное решение выйдет на стационарное за наименьшее число шагов. Для этого надо, чтобы начальные данные за один шаг затухали возможно сильнее.

Затухание начальных данных можно исследовать методом разделения переменных, взятым в строгой форме (поскольку нас интересуют точные значения границ спектра оператора). Собственные функции разностного оператора $-(\Lambda_1+\Lambda_2)$ в прямоугольнике на равномерной сетке равны, как нетрудно проверить,

$$w_{qr}(x)=\sin\frac{\pi q x_1}{a}\sin\frac{\pi r x_2}{b}, \quad 1\leqslant q\leqslant N-1, \quad 1\leqslant r\leqslant M-1. \tag{15}$$

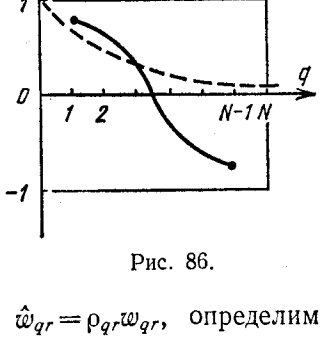


Рис. 86.

Подставляя их в схему (13) и полагая $\hat{w}_{qr}=\rho_{qr}w_{qr}$, определим множители роста гармоник:

$$\rho_{qr}=\frac{(1-\alpha_q)(1-\beta_r)}{(1+\alpha_q)(1+\beta_r)}, \qquad \alpha_q=\frac{2\tau}{h_1^2}\sin^2\frac{\pi q h_1}{2a}, \quad \beta_r=\frac{2\tau}{h_2^2}\sin^2\frac{\pi r h_2}{2b}. \tag{16}$$

Очевидно, все $|\rho_{qr}|<1$, т. е. все гармоники затухают; это означает, что схема (13) обладает аппроксимационной вязкостью.

Какие гармоники затухают наиболее медленно и, тем самым, сильнее всего препятствуют выходу на стационарный режим? Нетрудно заметить, что входящий в ρ_{qr} сомножитель $\gamma_q=(1-\alpha_q)/(1+\alpha_q)$ заключен в пределах $(-1, +1)$ и монотонно убывает при увеличении номера q (рис. 86, жирная линия). Наибольшим по модулю может быть множитель либо с $q=1$, либо с $q=N-1$. Считая N достаточно большим, можно положить

$$\sin\frac{\pi h_1}{2a}=\sin\frac{\pi}{2N}\approx\frac{\pi}{2N}, \quad \sin\frac{\pi(N-1)h_1}{2a}=\sin\frac{\pi(N-1)}{2N}\approx 1$$

и представить экстремальные множители (при $\tau\sim h$) в виде

$$\gamma_1\approx 1-\frac{\pi^2\tau}{a^2}, \quad \gamma_{N-1}\approx-\left(1-\frac{a^2}{\tau N^2}\right). \tag{17}$$

Аналогично, второй сомножитель $(1-\beta_r)/(1+\beta_r)$ максимален по модулю либо при $r=1$, либо при $r=M-1$. Поэтому $|\rho_{qr}|$ максимален либо при $q=r=1$, либо при $q=N-1$, $r=M-1$, причем

$$\rho_{11}\approx 1-\pi^2\tau\left(\frac{1}{a^2}+\frac{1}{b^2}\right), \quad \rho_{N-1,\,M-1}\approx 1-\frac{a^2}{\tau N^2}-\frac{b^2}{\tau M^2}. \tag{18}$$

Чем больше шаг τ, тем меньше ρ_{11} и больше $\rho_{N-1,\,M-1}$, причем оба они близки к 1 *); это значит, что первая и последняя гармоники затухают медленно, причем при малом шаге τ быстрей затухает последняя, а при большом — первая гармоника. Выберем шаг τ_0 так, чтобы $\rho_{11}(\tau_0)=\rho_{N-1,\,M-1}(\tau_0)$; из (18) видно, что

$$\tau_0\approx\frac{1}{\pi}\left(\frac{a^2}{N^2}+\frac{b^2}{M^2}\right)^{1/2}\left(\frac{1}{a^2}+\frac{1}{b^2}\right)^{-1/2}. \tag{19}$$

Если изменить шаг по сравнению с τ_0, то либо первая, либо последняя гармоника будет затухать медленнее, чем при $\tau=\tau_0$. Следовательно, τ_0 есть оптимальный шаг.

Число шагов $K(\varepsilon)$, нужное для достижения заданной точности ε, определяется условием $(\max|\rho_{qr}|)^K=\varepsilon$. При оптимальном шаге наибольшие множители роста равны

$$\rho_{11}(\tau_0)=\rho_{N-1,\,M-1}(\tau_0)\approx\exp\left[-\pi\left(\frac{a^2}{N^2}+\frac{b^2}{M^2}\right)^{1/2}\left(\frac{1}{a^2}+\frac{1}{b^2}\right)^{1/2}\right]. \tag{20}$$

Поэтому минимально необходимое число шагов есть

$$K(\tau_0)=\frac{1}{\pi}\left(\frac{a^2}{N^2}+\frac{b^2}{M^2}\right)^{-1/2}\left(\frac{1}{a^2}+\frac{1}{b^2}\right)^{-1/2}\ln\frac{1}{\varepsilon}. \tag{21}$$

Сравнивая время счета на установление (9) и величину оптимального шага (19), нетрудно убедиться, что $K(\tau_0)=T/\tau_0$.

Отметим, что при $a=b$, $M=N$ выполняется $\tau_0\sim a^2/N$ и $K(\tau_0)\sim N$.

В дифференциальном уравнении (12) установление происходит за достаточно большой промежуток времени. Почему же не взять для разностной схемы очень большой шаг по времени, если устойчивость это позволяет? Казалось бы, тогда мы быстрей добьемся установления. Но это не так. Спектр дифференциального оператора таков, что гармоники затухают тем быстрей, чем больше их номер, причем $|\rho_q|\to 0$ при $q\to\infty$; соответствующая кривая показана пунктиром на рис. 86. А затухание собственных функций разностного оператора имеет, вообще говоря, другой качественный характер, как видно из того же рисунка.

*) При нечетном числе измерений p для последней гармоники $\rho\to -1$ и вместо него во всех выкладках надо использовать $|\rho|=-\rho$.

Локально-одномерная схема (11.69) с полусуммой по времени в двумерном случае может быть записана в виде

$$\left(E-\frac{\tau}{2}\Lambda_1\right)\bar{y}=\left(E+\frac{\tau}{2}\Lambda_1\right)y+\tau\varphi_1, \tag{22а}$$

$$\left(E-\frac{\tau}{2}\Lambda_2\right)\hat{y}=\left(E+\frac{\tau}{2}\Lambda_2\right)\bar{y}+\tau\varphi_2, \tag{22б}$$

где операторы Λ_α имеют вид (13в) и коммутируют друг с другом. Умножая уравнение (22а) слева на $(E+{}^1/_2\tau\Lambda_2)$, а уравнение (22б) — на $(E-{}^1/_2\tau\Lambda_1)$, исключим $\bar{y}$ и запишем (22) в виде двуслойной схемы:

$$\left(E-\frac{\tau}{2}\Lambda_1\right)\left(E-\frac{\tau}{2}\Lambda_2\right)\hat{y}=$$
$$=\left(E+\frac{\tau}{2}\Lambda_1\right)\left(E+\frac{\tau}{2}\Lambda_2\right)y+\tau(\varphi_1+\varphi_2)+\frac{\tau^2}{2}(\Lambda_2\varphi_1-\Lambda_1\varphi_2).$$

Преобразуем ее к канонической форме:

$$\left(E-\frac{\tau}{2}\Lambda_1\right)\left(E-\frac{\tau}{2}\Lambda_2\right)\frac{\hat{y}-y}{\tau}-(\Lambda_1+\Lambda_2)y=$$
$$=\varphi_1+\varphi_2+\frac{\tau}{2}(\Lambda_2\varphi_1-\Lambda_1\varphi_2). \tag{23}$$

Видно, что левая часть (23) совпадает с левой частью продольно-поперечной схемы (13). Поэтому шаг τ_0, обеспечивающий наиболее быстрое затухание начальных данных, для схемы (23) определяется также формулой (19).

Нетрудно понять, как обобщить выражения оптимального шага (19) и минимального числа шагов (21) на случай произвольного числа измерений для локально-одномерной схемы с полусуммой. Запишем эти выражения в простейшем случае, когда задача Дирихле поставлена в p-мерном кубе со стороной a и по каждой стороне взято N узлов сетки:

$$\tau_0=\frac{a^2}{\pi N},\quad K(\tau_0)=\frac{N}{\pi p}\ln\frac{1}{\varepsilon}. \tag{24}$$

Однако если положить $\varphi_1+\varphi_2=f(x)$, то правая часть (23) будет отличаться от $f(x)$ на величину $O(\tau)$. Поэтому установившееся разностное решение будет отличаться от $u(x)$ на $O(\tau+h^2)$, и, тем самым, общая точность расчета будет хуже, чем по продольно-поперечной схеме.

Замечание. Для улучшения точности приравняем $f(x)$ правой части (23). Для этого достаточно произвольно выбрать $\varphi_1(x)$, а $\varphi_2(x)$ определить из уравнения

$$\varphi_2-\frac{\tau}{2}\Lambda_1\varphi_2=f-\varphi_1-\frac{\tau}{2}\Lambda_2\varphi_1. \tag{25}$$

Это линейное уравнение с трехдиагональной матрицей; оно легко решается одномерной прогонкой по направлению x_1.

Произвольная область. Выбрать оптимальный шаг удается только в простейших задачах, когда точно известны границы спектра разностного оператора. В областях сложной формы мы можем, подставляя в формулу (19) характерные размеры области и число узлов сетки, определить лишь порядок величины τ_0. Поэтому надо представлять, как зависит число шагов $K(\tau)$, требуемое для установления с заданной точностью, от величины шага τ.

При $\tau \leqslant \tau_0$ затухание начальных данных определяется множителем ρ_{11}, так что число шагов $K_\varepsilon(\tau)$ находится из условия $\rho_{11}^K = \varepsilon$. Из формулы (18) с учетом малости τ следует, что $\rho_{11}(\tau) \approx \exp(-\text{const} \cdot \tau)$; тем самым,

$$\rho_{11}(\tau) \approx [\rho_{11}(\tau_0)]^{\tau/\tau_0}.$$

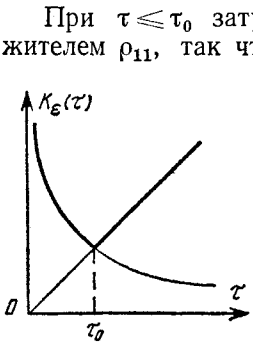


Рис. 87.

Отсюда нетрудно получить, что

$$K_\varepsilon(\tau) = \frac{\tau_0}{\tau} K(\tau_0) \quad \text{при} \quad \tau \leqslant \tau_0. \quad (26а)$$

Аналогично находим

$$K_\varepsilon(\tau) = \frac{\tau}{\tau_0} K(\tau_0) \quad \text{при} \quad \tau_0 \leqslant \tau. \quad (26б)$$

Кривая $K_\varepsilon(\tau)$ изображена на рис. 87 жирной линией. Ее минимум соответствует оптимальному шагу. Видно, что в низшей точке кривая имеет разрыв производной. Значит, умеренное отклонение величины шага τ от оптимума может заметно увеличить требуемое число шагов (во столько раз, во сколько τ отличается от τ_0).

Критерии установления. Из сказанного выше следует, что для задач в достаточно общей постановке (2), (3) заранее неизвестно, какое число шагов надо сделать до установления. Поэтому на практике вычисления прекращают при выполнении какого-нибудь правдоподобного, хотя и нестрогого критерия.

Нередко пользуются простейшим критерием

$$\| \hat{y} - y \| \leqslant \varepsilon; \quad (27а)$$

однако он недостаточно надежен, поскольку разностное решение устанавливается медленно. Если учесть, что установление происходит почти по геометрической прогрессии, то нетрудно получить более надежный критерий:

$$\| \hat{y} - y \| \leqslant \varepsilon(1 - \nu), \quad \text{где} \quad \nu = \frac{\| \hat{y} - y \|}{\| y - \check{y} \|}. \quad (27б)$$

Для схемы типа (13) расчет иногда оканчивают по условию малости невязки:

$$\| Ay - f \| \leqslant \varepsilon. \tag{27в}$$

Комплексная организация расчета, описанная в гл. VIII, § 2, п. 5, очень полезна даже в одномерных задачах. С увеличением числа измерений ее эффективность быстро возрастает. Напомним ее.

В области $G(x)$ строится последовательность сгущающихся вдвое сеток. На самой грубой сетке начальные условия выбираются произвольно; поскольку число узлов этой сетки невелико, то объем вычислений тоже невелик. После установления решение интерполируется на более подробной сетке и выбирается на ней в качестве начальных условий; это в несколько раз уменьшает требующееся для установления число шагов. Затем решение уточняется по способу Рунге с использованием всех сеток.

3. Чебышевский набор шагов. Счет на установление можно проводить с переменным шагом по времени. Для тех задач, в которых известны границы спектра разностных операторов, построены специальные наборы шагов τ_k $(1 \leqslant k \leqslant K)$, обеспечивающие гораздо более быстрое затухание начальных данных, чем при расчете с постоянным оптимальным шагом.

Чебышевский набор. Пусть разностная схема счета на установление приведена к двуслойной канонической форме:

$$B \frac{y^k - y^{k-1}}{\tau_k} + Ay^{k-1} = \varphi. \tag{28}$$

Будем предполагать, что A и B — самосопряженные положительно определенные операторы, удовлетворяющие неравенствам

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad 0 < \gamma_1 \leqslant \gamma_2. \tag{29}$$

Затухание начальных данных z^0 определяется однородным уравнением (28), которое можно записать в следующей форме:

$$\zeta^k = (E - \tau_k C)\, \zeta^{k-1}, \tag{30а}$$

где

$$C = B^{-1/2} A B^{-1/2}, \quad \zeta^k = B^{1/2} z^k. \tag{30б}$$

Отсюда, вытекает, что

$$\zeta^K = P_K(C)\, \zeta^0, \quad P_K(C) = \prod_{k=1}^{K} (E - \tau_k C). \tag{31}$$

Для наиболее быстрого затухания начальных данных последовательность шагов τ_k надо выбрать так, чтобы $\| P_K(C) \|$ была минимальна при заданном числе шагов K.

Поскольку A и B — самосопряженные операторы, то оператор C тоже самосопряженный, причем из (29) следует неравенство

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E. \tag{32}$$

В этом случае норму операторного многочлена $P_K(C)$ можно оценить по формуле *)

$$\| P_K(C) \| \leqslant \max_{\gamma_1 \leqslant \eta \leqslant \gamma_2} | P_K(\eta) |,$$

где $P_K(\eta)$ — алгебраический многочлен.

Для того чтобы максимум модуля алгебраического многочлена был минимален на отрезке $[\gamma_1, \gamma_2]$, этот многочлен с точностью до множителя должен совпадать с многочленом Чебышева первого рода для этого отрезка **). Используя данные в Приложении корни многочленов Чебышева и учитывая, что корнями $P_K(\eta)$ являются величины $1/\tau_k$, получим чебышевский набор шагов:

$$\tau_k = 2\left[(\gamma_2+\gamma_1)+(\gamma_2-\gamma_1)\cos\frac{\pi(2k-1)}{2K}\right]^{-1}, \quad 1 \leqslant k \leqslant K. \tag{33}$$

Определяя множитель, отличающий $P_K(\eta)$ от многочлена Чебышева на отрезке $[\gamma_1, \gamma_2]$, можно найти коэффициент затухания начальных данных после расчета с набором шагов (33):

$$\max_{\gamma_1 \leqslant \eta \leqslant \gamma_2} | P_K(\eta) | = \frac{2\rho^K}{1+\rho^{2K}}, \qquad \rho = \frac{\sqrt{\gamma_2}-\sqrt{\gamma_1}}{\sqrt{\gamma_2}+\sqrt{\gamma_1}}. \tag{34}$$

Если необходимая точность расчета $\varepsilon \ll 1$, то в оценке (34) можно пренебречь членом ρ^{2K}. Тогда требуемое число шагов равно

$$K(\varepsilon) \approx \frac{1}{\ln(1/\rho)} \ln\frac{2}{\varepsilon}. \tag{35}$$

Заметим, что сначала надо найти требуемое число шагов по формуле (35); только после этого можно вычислить искомый набор шагов (33).

Замечание. В случае области сложной формы или задачи с переменными коэффициентами (2) точные границы энергетической эквивалентности операторов (29) установить обычно не удается. Приходится оценивать их, занижая γ_1 и завышая γ_2 (в неизвестные числа раз ν_1 и ν_2). Поскольку всегда $\gamma_2 \gg \gamma_1$, то $[\ln(1/\rho)]^{-1} \approx 1/2\sqrt{\gamma_2/\gamma_1}$. Отсюда видно, что требуемое число шагов (35) возрастает в $\sqrt{\nu_1\nu_2}$ раз по сравнению со случаем, когда границы спектра известны точно.

---

*) Доказательство этого неравенства имеется, например, в [15].

**) О многочленах Чебышева и их свойствах см., например, в [9, 24].

Постоянный шаг. Оптимальный постоянный шаг выбирается так, чтобы начальные данные наиболее сильно затухали за один шаг. Там самым, он является частным случаем чебышевского набора, соответствующим $K=1$. Формулы (33), (34) принимают при этом вид

$$\tau_0=\frac{2}{\gamma_2+\gamma_1},\quad \max_{[\gamma_1,\ \gamma_2]}|P_1(\eta)|=\frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}. \tag{36}$$

Продольно-поперечная схема (13) или локально-одномерная схема с полусуммой (22) не подходят, строго говоря, под разобранный выше случай, потому что они содержат оператор

$$B_k=\left(E-\frac{1}{2}\tau_k\Lambda_1\right)\left(E-\frac{1}{2}\tau_k\Lambda_2\right),$$

явно зависящий от номера шага.

Однако для этих схем в п. 2 были определены оптимальный шаг (19) и соответствующая ему скорость затухания начальных данных. Ограничиваясь задачей Дирихле в p-мерном кубе со стороной a и одинаковым числом узлов N по каждой координате, запишем:

$$\tau_0=\frac{a^2}{\pi N},\quad \max|\lambda(\tau_0)|\approx 1-\frac{\pi p}{N}.$$

Сравнивая эти выражения с (36) и учитывая, что $\gamma_2\gg\gamma_1$, получим нестрогую, но удовлетворительную оценку границ спектра:

$$\gamma_1\sim\frac{\pi^2 p}{a^2},\quad \gamma_2\sim\frac{2\pi N}{a^2}. \tag{37}$$

Подставляя оценку (37) в (34) и (35), получим, что для рассмотренных схем

$$K\sim\sqrt{\frac{N}{2\pi p}}\ln\frac{2}{\varepsilon}. \tag{38}$$

Таким образом, счет на установление по экономичным схемам с чебышевским набором шагов требует всего $K\sim\sqrt{N}$ шагов, в то время как расчет с постоянным оптимальным шагом требует существенно большего числа шагов: $K(\tau_0)\sim N$.

Используя в разностной схеме переменный оператор B_k, можно найти другие наборы шагов, обеспечивающие еще более быстрое установление. Например, для продольно-поперечной схемы в случае задачи Дирихле (1) в прямоугольнике построен жорданов набор шагов (см. [81]), при котором

$$K=\frac{\ln N}{5}\ln\frac{4}{\varepsilon}.$$

Однако для более сложных задач наборы шагов с подобными характеристиками найти пока не удалось.

Порядок шагов. Чебышевский набор шагов позволяет проводить экономичный расчет на установление даже по явной схеме (11.56) с $\sigma=0$. Запишем эту схему в виде

$$E\frac{y^k-y^{k-1}}{\tau_k}+\sum_{\alpha=1}^{p}\Lambda_\alpha y^{k-1}=f. \tag{39}$$

Операторы схемы (39) постоянны, и нетрудно показать, что для задачи Дирихле в кубе $\gamma_1=\pi^2 p/a^2$, $\gamma_2=4pN^2/a^2$. Поэтому для расчета по схеме (39) с чебышевским набором шагов требуется число шагов

$$K=\frac{N}{\pi}\ln\frac{2}{\varepsilon}, \tag{40}$$

что по объему вычислений эквивалентно экономичным схемам с постоянным оптимальным шагом.

Заметим, что схема (39) устойчива только при $\tau\leqslant\tau_0=h^2/(2p)$. Среди шагов чебышевского набора (33) есть такие, которые больше τ_0 и меньше τ_0. Большие шаги вызывают рост погрешностей, а малые — затухание. В целом их действие таково, что если выполнить все $K(\varepsilon)$ шагов, то ошибка затухает в ε^{-1} раз.

Слово «если» употреблено не случайно. Нередко ошибки на промежуточных шагах возрастают в $10^{10}-10^{30}$ раз по сравнению с начальными, выходят за допустимые на ЭВМ пределы, и расчет не удается довести до конца. Поэтому, хотя чебышевский набор шагов для схемы (39) был найден более полувека назад, в практических вычислениях его долго не могли использовать.

Однако если шаги выполнять в определенном порядке, то расчет становится возможным. Идея упорядочения заключается в том, что сразу вслед за шагом, увеличивающим ошибку, надо выполнять шаг, уменьшающий ее. Правило перестановки особенно просто, если число шагов равно 2^r. Тогда надо расположить шаги в естественном порядке и сгруппировать парами: первый — последний, второй — предпоследний и т. д. Затем пары так же группируются в четверки: первая — последняя. Аналогично группируются четверки, восьмерки и т. д. Например, для 16 шагов окончательный порядок такой:

1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11.

При использовании упорядоченного чебышевского набора шагов ошибка на отдельных шагах может нарастать, но никогда в ходе расчета не превзойдет начальной ошибки, а в конце расчета будет соответствовать оценке $\|P_K\|$.

§ 2. Вариационные и вариационно-разностные методы

1. Метод Ритца. Вариационные методы применяются к эллиптическим уравнениям в частных производных независимо от числа измерений. Рассмотрим, например, задачу:

$$\operatorname{div}[k(x)\operatorname{grad} u]-\rho(x)u(x)=-f(x), \quad x\in G, \tag{41а}$$

$$u_\Gamma=\mu(x). \tag{41б}$$

Дифференциальный оператор $A=-\operatorname{div}[k\operatorname{grad}(\cdot)]+\rho$ является самосопряженным. Поэтому задача (41) эквивалентна задаче на минимум функционала $\Phi[u]=(Au-2f, u)$, которую при помощи формул векторного анализа можно записать в виде

$$\int_G [k(x)(\operatorname{grad} u)^2+\rho(x)u^2(x)-2f(x)u(x)]\,dx=\min, \quad u_\Gamma=\mu(x). \tag{42}$$

Возьмем некоторую функцию $\varphi_0(x)$, удовлетворяющую граничному условию (41б), и полную систему функций $\varphi_l(x)$, $l=1, 2, \ldots$, обращающихся в нуль на границе. Будем искать приближенное решение задачи (42) в следующем виде:

$$u(x)\approx y_n(x)=\varphi_0(x)+\sum_{l=1}^{n} c_l\varphi_l(x). \tag{43}$$

Подставляя (43) в (42), получим задачу на минимум квадратичной функции неизвестных коэффициентов c_l; для простоты ограничимся случаем $\varphi_0(x)\equiv 0$, соответствующим $u_\Gamma=0$:

$$\int_G\left[\sum_{r=1}^{n}\sum_{l=1}^{n} c_rc_l(k\operatorname{grad}\varphi_r\operatorname{grad}\varphi_l+\rho\varphi_r\varphi_l)-2f\sum_{r=1}^{n}c_r\varphi_r\right]dx=\min. \tag{44}$$

Приравнивая нулю производные по коэффициентам, получим для определения c_l систему линейных уравнений

$$\sum_{l=1}^{n} c_l\int_G(k\operatorname{grad}\varphi_r\operatorname{grad}\varphi_l+\rho\varphi_r\varphi_l)\,dx=\int_G f\varphi_r\,dx, \quad 1\leqslant r\leqslant n. \tag{45}$$

Обоснование сходимости метода Ритца при $n\to\infty$ рассматривалось в главе VII. При практическом применении метода Ритца успех сильно зависит от выбора системы функций $\varphi_l(x)$. При неудачном выборе этой системы для получения удовлетворительной точности может потребоваться очень много членов ряда (43).

Если область имеет несложную форму, то нередко выбирают систему с разделяющимися переменными; например, для прямоугольника полагают $\varphi_{lm}(x)=\xi_l(x_1)\eta_m(x_2)$, а для круга $\varphi_{lm}(x)=\xi_l(r)\eta_m(\theta)$. Отметим, что если в одномерной задаче для полу-

чения удовлетворительной точности требовалось $\sim n$ членов ряда, то в аналогичной p-мерной задаче обычно надо брать $\sim n^p$ членов.

Ограничиваясь малым числом членов, можно легко получить грубую оценку решения.

Замечание 1. Метод Ритца применим к многомерной задаче Штурма — Лиувилля (задаче на собственные значения).

Замечание 2. Если оператор в задаче типа (41) не самосопряженный, то вместо метода Ритца применяют метод Галеркина.

2. Стационарные разностные схемы. Такие схемы можно составлять, непосредственно аппроксимируя производные разностями, или при помощи интегро-интерполяционного метода. Например, для многомерного уравнения $\Delta u = -f(x)$ простейшая разностная замена производных приводит к схеме

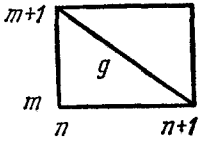


Рис. 88.

$$\sum_{\alpha=1}^{p} \Lambda_\alpha y = -f. \tag{46}$$

Составлять разностные схемы можно также вариационными методами. Для этого специальным образом выбирают пробные функции $y_n(x)$, например, считая их сплайнами, построенными по узловым значениям y.

Пример. Рассмотрим решение двумерного уравнения $\Delta u = -f$ на прямоугольной сетке с шагами h_1, h_2. Эквивалентная задача на минимум в этом случае имеет вид

$$\Phi[u] = \int_G [(\operatorname{grad} u)^2 - 2f(x)\,u(x)]\,dx_1\,dx_2 = \min \tag{47}$$

(для простоты мы опускаем краевые условия). Разобьем каждую прямоугольную ячейку на две треугольных (рис. 88) и в треугольных ячейках аппроксимируем $u(x)$ линейными функциями; например, в нижнем треугольнике $g(x)$

$$u(x) \approx y(x) = y_{nm} + \frac{1}{h_1}(x_1 - x_{1n})(y_{n+1,\,m} - y_{nm}) + \\ + \frac{1}{h_2}(x_2 - x_{2m})(y_{n,\,m+1} - y_{nm}), \qquad x \in g. \tag{48}$$

Совокупность этих функций образует линейный сплайн. Очевидно,

$$(\operatorname{grad} y)^2 = \frac{1}{h_1^2}(y_{n+1,\,m} - y_{nm})^2 + \frac{1}{h_2^2}(y_{n,\,m+1} - y_{nm})^2, \qquad x \in g.$$

Аппроксимируя правую часть $f(x)$ в ячейке $g(x)$, например,

константой f_{nm}, легко вычисляем интеграл по этой ячейке:

$$\int_g [(\operatorname{grad} y)^2 - 2fy]\, dx = \frac{h_1 h_2}{2}\left[\frac{1}{h_1^2}(y_{n+1,\,m} - y_{nm})^2 + \right.$$
$$\left. + \frac{1}{h_2^2}(y_{n,\,m+1} - y_{nm})^2 - \frac{2}{3} f_{nm}(y_{nm} + y_{n+1,\,m} + y_{n,\,m+1})\right].$$

Аналогично вычисляется интеграл по верхней треугольной ячейке. Суммируя эти интегралы, получим

$$\Phi[y] = \frac{1}{2} h_1 h_2 \sum_{n,\,m}\left[\frac{1}{h_1^2}(y_{n+1,\,m} - y_{nm})^2 + \frac{1}{h_1^2}(y_{n+1,\,m+1} - y_{n,\,m+1})^2 + \right.$$
$$+ \frac{1}{h_2^2}(y_{n,\,m+1} - y_{nm})^2 + \frac{1}{h_2^2}(y_{n+1,\,m+1} - y_{n+1,\,m})^2 - \frac{2}{3} f_{nm}(y_{nm} +$$
$$\left. + y_{n+1,\,m} + y_{n,\,m+1}) - \frac{2}{3} f_{n+1,\,m+1}(y_{n+1,\,m+1} + y_{n+1,\,m} + y_{n,\,m+1})\right] =$$
$$= \min. \quad (49)$$

Функционал $\Phi[y] = F(y_{nm})$ является квадратичной функцией узловых значений. Приравнивая нулю производные функционала по y_{nm} и учитывая, что эта величина входит в четыре члена двойной суммы (49), получим разностную схему

$$\frac{1}{h_1^2}(y_{n+1,\,m} - 2y_{nm} + y_{n-1,\,m}) + \frac{1}{h_2^2}(y_{n,\,m+1} - 2y_{nm} + y_{n,\,m-1}) +$$
$$+ \frac{1}{6}(2f_{nm} + f_{n+1,\,m} + f_{n-1,\,m} + f_{n,\,m+1} + f_{n,\,m-1}) = 0. \quad (50)$$

Это — стационарная схема; легко видеть, что она аппроксимирует непосредственно исходное дифференциальное уравнение.

Замечание. При помощи вариационного метода удобно составлять разностные схемы высокой точности. Для этого решение $u(x)$ и правую часть $f(x)$ аппроксимируют сплайнами более высокого порядка, обычно кубическими (такая аппроксимация обсуждалась в гл. VII, § 4, п. 4).

3. Прямые методы решения. Для стационарных схем типа (47) наиболее сложным является вопрос о фактическом вычислении разностного решения.

В самом деле, p-мерная схема (46) является линейной алгебраической системой с N^p неизвестными (если по каждой координате взято N интервалов). Матрица этой системы в двумерном случае изображена на рис. 89. В общем случае эта матрица ленточная, причем лента слабо заполнена и имеет полуширину N^{p-1}.

Вычисление разностного решения методом исключения Гаусса (который не может использовать слабое заполнение ленты) тре-

бует $\sim N^{3p-2}$ действий, т. е. $\sim N^{2p-2}$ операций на узел сетки *). При большом N это число действий неприемлемо велико; кроме того, лента матрицы не помещается в оперативной памяти ЭВМ. Поэтому прямое решение линейной системы (46) методом Гаусса возможно только в двумерных расчетах, и то при небольшом $N \lesssim 50$.

Замечание. Если строить схемы высокого порядка точности (например, сплайновые) или использовать последовательность сгущающихся сеток (обычно при $N=4, 8, 16, 32$) с уточнением по способу Рунге, то даже при небольшом числе узлов удается получить удовлетворительную точность расчета.

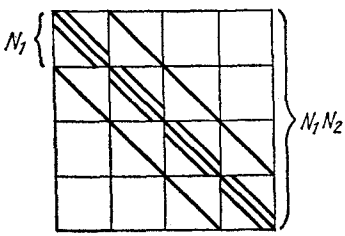


Рис. 89.

Для некоторых важных частных случаев эллиптических задач разработаны очень быстрые прямые методы; перечислим их (они подробно изложены, например, в [3, 6, 31]).

Быстрое преобразование Фурье применимо к задаче Дирихле в прямоугольнике. Оно основано на том, что если число интервалов по каждой переменной N_α разбивается на множители, то вычислять коэффициенты дискретного преобразования Фурье можно не по формулам Бесселя типа (2.44), а по более экономичным рекуррентным формулам. Если $N_\alpha = 2^{r_\alpha}$, то метод является особенно быстрым и требует всего $4\log_2 N$ действий на каждый узел сетки.

Рассмотрим этот метод сначала на примере одномерной задачи для уравнения с постоянными коэффициентами $u'' - \mu u = -f(x)$ и краевыми условиями первого рода (без ограничения общности их можно взять периодическими). Составим разностную схему на равномерной сетке $\{x_n = nh,\ 0 \leqslant n \leqslant N\}$:

$$\frac{1}{h^2}(y_{n-1} - 2y_n + y_{n+1}) - \mu y_n = -\varphi_n, \quad 1 \leqslant n \leqslant N-1, \qquad y_0 = \varphi_0, \quad y_N = \varphi_0. \tag{51}$$

Будем искать разностное решение в виде разложения Фурье:

$$y_n = \sum_{q=0}^{N-1} a_q w^{nq}, \quad \text{где} \quad w = \exp(2\pi i/N) \tag{52}$$

*) Аналогичная ситуация возникла в неявной схеме (11.56) для многомерного уравнения теплопроводности.

Подставим разложение (52) в соотношение (51), умножим на $w^{-np} = \exp(-2\pi inp/N)$ и просуммируем по n от 0 до $N-1$. Замечая, что

$$w^{(n-1)q} - 2w^{nq} + w^{(n+1)q} = -4w^{nq}\sin^2\frac{\pi q}{N},$$

и учитывая условие ортогональности гармоник (см. гл. II, § 2, п. 4), найдем

$$a_p = b_p \Big/ \left(\frac{4}{h^2}\sin^2\frac{\pi p}{N} + \mu\right), \tag{53}$$

где

$$b_p = \frac{1}{N}\sum_{n=0}^{N-1} \varphi_n w^{-np}, \quad 0 \leqslant p \leqslant N-1, \tag{54}$$

являются дискретными коэффициентами Фурье правой части уравнения. Формулы (53), (54) позволяют найти искомое разностное решение.

Однако эти формулы неэкономичны. Необходимо вычислить N коэффициентов b_p, причем нахождение каждого коэффициента по формуле (54) требует примерно $2N$ операций. Следовательно, задача (51) решается за $2N^2$ операций, т. е. много медленнее, чем в методе прогонки.

Если число интервалов сетки составное, $N = KL$, то формулу (54) можно преобразовать так, что требуемое количество операций уменьшится. Представим индексы n и p в следующем виде:

$$n = l_1 + Ll_2, \quad 0 \leqslant l_1 \leqslant L-1, \quad 0 \leqslant l_2 \leqslant K-1,$$
$$p = p_1 + Kp_2, \quad 0 \leqslant p_1 \leqslant K-1, \quad 0 \leqslant p_2 \leqslant L-1.$$

Запишем формулу (54) в виде двойной суммы:

$$b_p = \frac{1}{KL}\sum_{l_1=0}^{L-1}\sum_{l_2=0}^{K-1} \varphi_{l_1+Ll_2} w^{-p_1 l_1 - Lp_1 l_2 - Kl_1 p_2 - LKl_2 p_2}.$$

Отбросим в показателе степени последнее слагаемое, ибо $w^{LK} = 1$, и получим следующее выражение коэффициентов Фурье:

$$b(p) \equiv b_p = \frac{1}{L}\sum_{l_1=0}^{L-1} b(l_1, p_1)\, w^{-pl_1}, \quad 0 \leqslant p \equiv p_1 + Kp_2 \leqslant N-1, \tag{55}$$

где

$$b(l_1, p_1) = \frac{1}{K}\sum_{l_2=0}^{K-1} \varphi_{l_1+Ll_2} w^{-Lp_1 l_2}, \quad 0 \leqslant l_1 \leqslant L-1, \quad 0 \leqslant p_1 \leqslant K-1. \tag{56}$$

Вычисление N коэффициентов $b(p)$ по формуле (55) требует $2NL$ операций; вычисление $LK=N$ вспомогательных коэффициентов $b(l_1, p_1)$ по формуле (56) производится еще за $2NK$ операций. Следовательно, число операций, необходимое для нахождения коэффициентов Фурье по формулам (55), (56), равно $2N(L+K)$; оно существенно меньше, чем $2N^2$ (например, при $K=L=\sqrt{N}$ меньше в $\sqrt{N}/2$ раз).

Если K в свою очередь разбивается на множители, то формулу (56) следует преобразовать аналогичным образом. Это позволяет еще уменьшить объем вычислений.

Приведем без вывода рекуррентные формулы вычисления коэффициентов Фурье для случая $N=L^r$:

$$b(p)=\frac{1}{L}\sum_{l_1=0}^{L-1} b(l_1, p_1)\,\omega^{-pl_1},$$

$$b(l_1, l_2, \ldots, l_k, p_k)=$$

$$=\frac{1}{L}\sum_{l_{k+1}=0}^{L-1} b(l_1, l_2, \ldots, l_{k+1}, p_{k+1})\,\omega^{-L^k l_{k+1} p_k},\quad 1\leqslant k\leqslant r-2, \tag{57}$$

$$b(l_1, l_2, \ldots, l_{r-1}, p_{r-1})=$$

$$=\frac{1}{L}\sum_{l_r=0}^{L-1}\Phi_{l_1+Ll_2+L^2l_3+\ldots+L^{r-1}l_r}\,\omega^{-L^{r-1}l_r p_{r-1}},$$

причем

$$0\leqslant l_k\leqslant L-1,\quad 0\leqslant p_k\leqslant L^{r-k}-1.$$

Число вспомогательных коэффициентов k-го ранга $b(l_1, \ldots, l_k, p_k)$ равно N, поэтому для вычисления коэффициентов всех рангов по формулам (57) требуется около $2NLr$ операций.

Если учесть, что $L=N^{1/r}$, то нетрудно найти оптимальное число сомножителей $r_{\text{опт}}\approx\ln N$ и оптимальное значение $L_{\text{опт}}\approx e\approx 3$. Но для программирования считается более удобным, если $N=2^r$ и $L=2$; в последнем случае требуемое число операций равно $4N\log_2 N$, что мало отличается от оптимального случая и почти не уступает по скорости методу прогонки.

Обобщение этого метода на случай многих измерений очевидно. Пусть, например, для уравнения с постоянными коэффициентами

$$u_{x_1x_1}+u_{x_2x_2}-\mu u=-f(x_1, x_2)$$

поставлена первая краевая задача в прямоугольной области. Введем равномерную сетку $\{x_{1n}=nh_1,\ x_{2m}=mh_2,\ 0\leqslant n\leqslant N,\ 0\leqslant$

$\leqslant m \leqslant M\}$ и составим разностную схему

$$\frac{1}{h_1^2}(y_{n-1,\,m} - 2y_{nm} + y_{n+1,\,m}) + \\ + \frac{1}{h_2^2}(y_{n,\,m-1} - 2y_{nm} + y_{n,\,m+1}) - \mu y_{nm} = -\varphi_{nm}. \qquad (58)$$

Будем искать разностное решение в виде разложения Фурье

$$y_{nm} = \sum_{p=0}^{N-1}\sum_{q=0}^{M-1} a_{pq}\omega_1^{np}\omega_2^{mq}, \qquad (59)$$

$$\omega_1 = \exp(2\pi i/N), \quad \omega_2 = \exp(2\pi i/M).$$

Аналогично одномерному случаю, получим следующие выражения для коэффициентов Фурье:

$$a_{pq} = b_{pq} \Big/ \left(\frac{4}{h_1^2}\sin^2\frac{\pi p}{N} + \frac{4}{h_2^2}\sin^2\frac{\pi q}{M} + \mu\right), \qquad (60)$$

где

$$b_{pq} = \frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\varphi_{nm}\omega_1^{-np}\omega_2^{-mq}. \qquad (61)$$

Запишем последнюю формулу в следующем виде:

$$b_{pq} = \frac{1}{N}\sum_{n=0}^{N-1}\beta_{nq}\omega_1^{-np}, \\ \beta_{nq} = \frac{1}{M}\sum_{m=0}^{M-1}\varphi_{nm}\omega_2^{-mq}. \qquad (62)$$

Каждая сумма в формулах (62) имеет тот же вид, что и в формуле (54). Поэтому, если N и M разлагаются на множители, то каждую сумму можно вычислить по рекуррентным формулам типа (57). Если при этом $N = L_1^{r_1}$ и $M = L_2^{r_2}$, то число операций на каждый узел сетки, аналогично одномерному случаю, есть $O(r_1L_1 + r_2L_2) = O(\log(NM))$. Следовательно, быстрое преобразование Фурье даже в многомерном случае по экономичности мало уступает самому быстрому одномерному методу — прогонке.

Метод декомпозиции, или нечетно-четной редукции, применим для той же задачи, что и быстрое преобразование Фурье. Он использует исключение всех нечетных точек из системы уравнений типа (46). При $N_\alpha = 2^{r_\alpha}$ исключение выполняет-

ся рекуррентно и число действий на узел сетки составляет $O(\log_2 N)$.

Матричная прогонка применима даже для случая областей сложной формы. Число действий на узел сетки в этом методе есть $O(N^2)$. Но если требуется решить на данной сетке большую серию задач с различными правыми частями и граничными значениями, то, сохраняя и используя результаты промежуточных вычислений, можно сократить это число действий до $O(N)$.

Быстрые прямые методы обобщены в настоящее время на задачи в круге и области ступенчатой формы (в этих случаях их скорость падает). Однако для областей произвольной формы, а также для уравнения достаточно общего вида (2) удовлетворительных прямых методов пока не найдено.

4. Итерационные методы. В случае сложных задач неэволюционные разностные схемы $Ay = -f$ решают итерационными методами. Простейшим из них является метод Якоби (5.51), относящийся к методам последовательного приближения. Для двумерного уравнения (46) он имеет вид

$$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y_{nm}^k=\frac{1}{h_1^2}\left(y_{n+1,\,m}^{k-1}+y_{n-1,\,m}^{k-1}\right)+\frac{1}{h_2^2}\left(y_{n,\,m-1}^{k-1}+y_{n,\,m+1}^{k-1}\right)+f_{nm},$$

где k — номер итерации. Это выражение можно формально переписать следующим образом:

$$E\frac{y^k-y^{k-1}}{\tau}+\sum_{\alpha=1}^{2}\Lambda_\alpha y^{k-1}=-f,\qquad \tau=\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)^{-1}.$$

Большинство итерационных методов можно символически записать в аналогичной форме:

$$B_k\frac{y^k-y^{k-1}}{\tau_k}+Ay^{k-1}=-f; \tag{63}$$

если B и τ не зависят от номера итерации k, то процесс называют *стационарным*.

Итерационный процесс (63) можно рассматривать как разностную схему расчета некоторой эволюционной задачи. Эту задачу можно найти, определяя нулевое или первое дифференциальное приближение разностной схемы (63). Физический смысл найденной задачи не имеет значения; важно только, чтобы она соответствовала диссипативному процессу, т. е. обеспечивала бы установление стационарного решения. Тем самым, *несущественно, что именно аппроксимирует оператор* B; он должен только: а) *обеспечивать возможно более быстрое затухание начальных данных* и б) *легко обращаться*, чтобы решение y^k вычислялось за малое число действий.

Продольно-поперечная и локально-одномерная схемы, которые можно формально рассматривать как итерационные процессы (63) для решения системы $Ay=-f$, удовлетворяют этим двум требованиям. Однако этим требованиям удовлетворяют также некоторые схемы, невыгодные для расчета параболических задач. Одной из таких схем является

Попеременно-треугольная схема, которую мы рассмотрим на примере двумерного уравнения

$$k_1 u_{x_1x_1} + k_2 u_{x_2x_2} = -f(x_1, x_2).$$

Запишем вспомогательное параболическое уравнение:

$$v_t = k_1 v_{x_1x_1} + k_2 v_{x_2x_2} + f(x_1, x_2).$$

Выберем шаблон, показанный на рис. 90, и составим на нем чисто неявную разностную схему

$$\frac{1}{\tau}(\hat{y}-y) = (\Lambda_1+\Lambda_2)\hat{y}+\varphi,$$

которую можно переписать в канонической форме:

$$(E-\tau\Lambda_1-\tau\Lambda_2)\frac{\hat{y}-y}{\tau} - (\Lambda_1+\Lambda_2)y = \varphi. \tag{64}$$

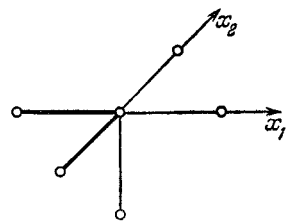


Рис. 90.

Как отмечалось выше, эта схема неэкономична, поскольку обращение оператора $E-\tau\Lambda_1-\tau\Lambda_2$ требует в общем случае большого числа операций на каждый узел сетки.

Введем *треугольные* операторы

$$\begin{aligned} R_1 y_{nm} &= \frac{k_1}{h_1^2}(y_{nm}-y_{n-1,\,m}) + \frac{k_2}{h_2^2}(y_{nm}-y_{n,\,m-1}), \\ R_2 y_{nm} &= \frac{k_1}{h_1^2}(y_{nm}-y_{n+1,\,m}) + \frac{k_2}{h_2^2}(y_{nm}-y_{n,\,m+1}), \end{aligned} \tag{65}$$

определенные на треугольных шаблонах (шаблон для R_1 показан на рис. 90 жирными линиями). Нетрудно заметить, что $\Lambda_1+\Lambda_2 = -(R_1+R_2)$, что позволяет записать схему (64) следующим образом:

$$(E+\tau R_1+\tau R_2)\frac{\hat{y}-y}{\tau} - (\Lambda_1+\Lambda_2)y = \varphi. \tag{66}$$

Слегка изменим схему (66), добавляя в левую часть член $\tau^2 R_1 R_2(\hat{y}-y)/\tau$, имеющий порядок малости $O(\tau^2)$. Возникающий при этом оператор $E+\tau R_1+\tau R_2+\tau^2 R_1 R_2$ *факторизуется*, т. е. представляется в виде произведения операторов $E+\tau R_1$ и $E+\tau R_2$.

Полученную схему называют *попеременно-треугольной*:

$$(E+\tau R_1)(E+\tau R_2)\frac{\hat{y}-y}{\tau}-(\Lambda_1+\Lambda_2)y=\varphi. \tag{67}$$

Операторы $E+\tau R_\alpha$ легко обращаются, так что алгоритм вычисления разностного решения в этой схеме несложен и требует небольшого числа операций на каждый узел сетки. В самом деле, такие операторы уже встречались при составлении схемы бегущего счета (10.29) для многомерного уравнения переноса; организация вычислений в этом случае была подробно разобрана в гл. X, § 1, п. 4. Схема (67) также решается посредством бегущего счета. На каждом слое сначала обращают оператор $E+\tau R_1$; вычисления при этом начинают с узла (x_{10}, x_{20}) и ведут, например, по направлениям x_1, доходя в конечном итоге до узла (x_{1N}, x_{2M}). Затем обращают оператор $E+\tau R_2$, начиная вычисления с узла (x_{1N}, x_{2M}) и ведя их в обратном порядке.

Попеременно-треугольная схема естественно переносится на случай любого числа измерений. Она легко обобщается на дифференциальные уравнения с переменными или разрывными коэффициентами и области $G(x)$ сложной формы. При этом схема для исходной задачи $Ay=-f$ записывается в виде

$$B_k=(D+\tau_k R_1)D^{-1}(D+\tau_k R_2),\quad B_k=\frac{y^k-y^{k-1}}{\tau_k}+Ay^{k-1}=f, \tag{68}$$

где $D=D^H>0$ — диагональный оператор, выбираемый так, чтобы возможно сильнее уменьшить отношение γ_2/γ_1 границ эквивалентности (29) операторов A и B_k, а треугольные операторы R_α выбраны так, чтобы выполнялось $R_1+R_2=A$, $R_1=R_2^H$ (нетрудно заметить, что в схеме (67) эти условия выполнены, причем $D=E$). Если используется чебышевский набор шагов, то процесс (68) сходится за $K=O\left(\sqrt{N}\ln\frac{1}{\varepsilon}\right)$ итераций.

Градиентные методы. Можно заменить линейную задачу $Ay=-f$ задачей на минимум квадратичной функции $F(y)$. Если матрица A положительно определенная, то удобно взять задачу

$$F(y)\equiv(Ay,\ y)+2(f,\ y)=\min. \tag{69}$$

Для произвольной матрицы A (которая встречается в задачах со смешанными производными) можно положить

$$F(y)\equiv(Ay+f,\ Ay+f)=\min. \tag{70}$$

Задачу на минимум можно решать методом наискорейшего спуска, что для случая (69) выполняется по формулам (6.22) — (6.26).

Скорость сходимости метода наискорейшего спуска, согласно оценке (6.27), такая же, как и у экономичных схем с постоянным *оптимальным шагом*, т. е. $K(\varepsilon) = O(N \ln(1/\varepsilon))$. Она меньше, чем у схем с чебышевским набором шагов. Достоинством метода является то, что для его применения не надо знать границы спектра оператора A.

ЗАДАЧИ

1. Найти время, необходимое для установления стационарного режима в эволюционной задаче (10), и исследовать характер установления.

2. Найти оптимальный шаг для счета на установление по локально-одномерной схеме типа (22) в случае задачи Дирихле (1) в трехмерном параллелепипеде.

3. Найти оптимальный шаг и необходимое число шагов $K(\varepsilon)$ для счета на установление по явной схеме (39) в случае задачи Дирихле в p-мерном параллелепипеде, когда сетки равномерны, а число узлов по каждой переменной N_α свое.

4. Для условий задачи 3 построить упорядоченный чебышевский набор шагов при $K = 64$.

5. Обосновать критерий установления (27б).

6. Для решения задачи (41) методом Ритца написать аналог системы (45) при $\varphi_0(x) \neq 0$.

7. Составить вариационным методом разностную схему, аналогичную (50), используя для $f(x)$ сплайновую аппроксимацию типа (48).

8. Составить формулы наискорейшего спуска для задачи (70).

9. Доказать справедливость рекуррентных формул (57). Указание: полагая последовательно $N = LN_1$, $N_1 = LN_2$ и т. д., использовать для индексов следующую замену переменных:

$$p = p_1 + N_1 p_1', \quad p_1 = p_2 + N_2 p_2', \quad p_2 = p_3 + N_3 p_3', \quad \ldots,$$
$$n = l_1 + L l_2', \quad l_2' = l_2 + L l_3', \quad l_3' = l_3 + L l_4', \quad \ldots$$

ГЛАВА XIII

ГИПЕРБОЛИЧЕСКИЕ УРАВНЕНИЯ

Глава XIII посвящена разностным схемам для уравнений в частных производных гиперболического типа. В § 1 рассмотрено гиперболическое уравнение второго порядка — волновое уравнение, которое можно заменить эквивалентной системой двух уравнений первого порядка. На примере одномерной задачи подробно разобраны явные и неявные разностные схемы решения этих уравнений. Дано обобщение этих схем на случай любого числа измерений.

В § 2 рассмотрены одномерные уравнения газодинамики, являющиеся гиперболической системой квазилинейных уравнений первого порядка. Построены две однородные разностные схемы («крест» и неявная консервативная схема), дающие хорошие результаты при решении многих прикладных задач. Приведен вид псевдовязкости, используемый в этих схемах.

§ 1. Волновое уравнение

1. Схема «крест». К гиперболическим уравнениям приводят задачи колебания струны, движения сжимаемого газа, распространения возмущений электромагнитных полей и многие другие.

Типичным примером одномерной задачи является задача малых колебаний натянутой струны с распределенной по длине нагрузкой $f(x,\ t)$:

$$u_{tt} = c^2 u_{xx} + f(x,\ t), \quad 0 < x < a, \quad 0 < t \leqslant T; \tag{1а}$$

$$u(x,\ 0) = \mu_1(x), \quad u_t(x,\ 0) = \mu_2(x), \quad 0 < x < a; \tag{1б}$$

$$u(0,\ t) = \mu_3(t), \quad u(a,\ t) = \mu_4(t), \quad 0 \leqslant t \leqslant T \tag{1в}$$

(это же уравнение описывает плоские акустические волны в газе при наличии внешнего силового поля f). Краевые условия первого рода (1в) соответствуют заданному закону движения концов струны; возможны и другие типы краевых условий.

Заметим, что, в отличие от параболической задачи (11.1), гиперболическая задача (1) требует постановки двух начальных условий: не только начального смещения от положения равновесия u, но и начальной скорости вещества u_t.

Составим несложную и эффективную разностную схему для задачи (1). Выберем по x, t прямоугольную сетку, для простоты

равномерную, и возьмем изображенный на рис. 91 шаблон. Аппроксимируя прозводные разностями, получим трехслойную схему

$$\frac{1}{\tau^2}(\hat{y}_n - 2y_n + \check{y}_n) = \frac{c^2}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + f_n, \quad 1 \leqslant n \leqslant N-1, \quad (2а)$$

с граничными условиями

$$y_0 = \mu_3(t), \qquad y_N = \mu_4(t). \quad (2б)$$

По форме шаблона эту схему называют «крест». Исследуем ее.

Вычисление решения. На нулевом слое решение известно из начального условия

$$y_n^0 = \mu_1(x_n), \qquad 0 \leqslant n \leqslant N. \quad (3)$$

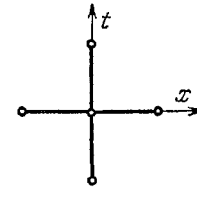


Рис. 91.

На первом слое решение также можно вычислить по начальным данным. Простейший способ состоит в том, что полагают

$$\frac{1}{\tau}(y_n^1 - y_n^0) \approx u_t(x_n, 0) = \mu_2(x_n), \quad 1 \leqslant n \leqslant N-1. \quad (4а)$$

Более хорошие результаты дает использование следующего члена разложения:

$$\frac{1}{\tau}(y_n^1 - y_n^0) \approx u_t(x_n, 0) + \frac{\tau}{2} u_{tt}(x_n, 0);$$

выражение для u_{tt} в это соотношение надо подставить из уравнения (1а). Окончательно получим

$$y_n^1 = y_n^0 + \tau\mu_2(x_n) + \frac{\tau^2}{2}\left[c^2 \frac{d^2\mu_1(x_n)}{dx^2} + f(x_n, 0)\right], \quad 1 \leqslant n \leqslant N-1, \quad (4б)$$

где μ_{1xx} можно заменить второй разностью.

Схема (2а) явная и позволяет выразить $\hat{y}_n$ через значения y с двух предыдущих слоев. Поэтому, начиная со второго слоя, разностное решение вычисляется по этой схеме.

Описанный алгоритм показывает, что, после того как выбрана одна из начальных формул (4а, б), разностное решение существует и единственно.

Аппроксимация. Разложим точное решение по формуле Тейлора с центром в узле (x_n, t_m), предполагая наличие непрерывных четвертых производных:

$$u_{n\pm1}^m = u \pm hu_x + \frac{h^2}{2}u_{xx} \pm \frac{h^3}{6}u_{xxx} + \frac{h^4}{24}u_{xxxx},$$

$$u_n^{m\pm1} = u \pm \tau u_t + \frac{\tau^2}{2}u_{tt} \pm \frac{\tau^3}{6}u_{ttt} + \frac{\tau^4}{24}u_{tttt}.$$

Используя эти разложения, легко найдем невязку схемы (2а):

$$\psi = u_{tt} - c^2 u_{xx} - \frac{1}{\tau^2}\left(u_n^{m+1} - 2u_n^m + u_n^{m-1}\right) +$$

$$+ \frac{c^2}{h^2}\left(u_{n+1}^m - 2u_n^m + u_{n-1}^m\right) = -\frac{\tau^2}{12} u_{tttt} + \frac{c^2 h^2}{12} u_{xxxx} = O(\tau^2 + h^2), \quad (5)$$

и невязку начального условия (4а):

$$\psi = \mu_1(x_n) - \frac{1}{\tau}\left(u_n^1 - u_n^0\right) = -\frac{\tau}{2} u_{tt} = O(\tau), \quad (6а)$$

или начального условия (4б):

$$\psi = \mu_2(x_n) + \frac{\tau}{2}\left(c^2 \frac{\partial^2 \mu_1(x_n)}{\partial x^2} + f_n\right) - \frac{1}{\tau}\left(u_n^1 - u_n^0\right) = -\frac{1}{6}\tau^2 u_{ttt} = O(\tau^2). \quad (6б)$$

Начальное условие (3) и краевые условия (2б) аппроксимируются точно.

Таким образом, схема (2) — (3) с начальным условием (4б) имеет аппроксимацию $O(\tau^2 + h^2)$. Использование начального условия (4а) ухудшает аппроксимацию до $O(\tau + h^2)$.

Устойчивость исследуем методом разделения переменных, полагая в схеме (2а)

$$f = 0, \quad y_n = e^{iqx_n}, \quad \hat{y} = \rho_q y, \quad y = \rho_q \check{y}. \quad (7)$$

Для множителя роста гармоники получим квадратное уравнение

$$\rho_q^2 - 2\rho_q\left(1 - 2\frac{c^2\tau^2}{h^2}\sin^2\frac{qh}{2}\right) + 1 = 0. \quad (8)$$

По теореме Виета произведение его корней $\rho_q' \rho_q'' = 1$. Значит, условие устойчивости $|\rho_q| \leqslant 1$ может быть выполнено, если $|\rho_q'| = |\rho_q''| = 1$. Для уравнения с действительными коэффициентами (8) это означает, что корни образуют комплексно сопряженную пару; для этого дискриминант уравнения не должен быть положительным:

$$\left|1 - 2\left(\frac{c\tau}{h}\sin\frac{qh}{2}\right)^2\right| \leqslant 1.$$

Чтобы это неравенство выполнялось для любых гармоник, необходимо и достаточно соблюдение *условия Куранта*:

$$c\tau < h. \quad (9)$$

Таким образом, схема «крест» условно устойчива.

Замечание 1. Если $c\tau = h$, то для некоторых гармоник λ_q становится кратным корнем уравнения (8). Это приводит к слабой неустойчивости счета: амплитуда этих гармоник при $\tau \to 0$

растет, как $m=(t/\tau)$. Поэтому в условии Куранта (9) стоит строгое неравенство.

Сходимость. Из сказанного выше следует, что схема (2) с начальными условиями (3), (4б) при выполнении условия Куранта (9) сходится со скоростью $O(\tau^2+h^2)$.

Из наших рассуждений вытекает сходимость схемы в $\|\cdot\|_{l_2}$; но методом энергетических неравенств можно доказать, что сходимость равномерная.

Схема (2) обеспечивает хорошую точность расчета решений $u(x, t)$, имеющих непрерывные четвертые производные. Она позволяет рассчитывать менее гладкие и даже разрывные решения, хотя в последнем случае точность расчетов невелика и обычно возникает легкая «разболтка», связанная с немонотонностью схемы. Условие устойчивости (9) естественное, поскольку для получения хорошей точности тоже надо полагать $c\tau \sim h$. Поэтому схему «крест» часто используют для практических расчетов.

Замечание 2. Схема (2) написана для случая постоянных шагов h и τ. Если шаги переменные, то надо заменить производные по пространству и времени соответствующими выражениями (3.2), которые обеспечивают локальную аппроксимацию $O(\tau^2+h^2)$ только в случае квазиравномерных сеток по x и t.

Поэтому для трехслойных схем, в отличие от двуслойных, резкие смены шага τ_m в ходе расчета опасны: это может привести к ухудшению точности.

Замечание 3. Для задач с краевыми условиями первого рода $u_\Gamma=\mu(t)$ удобно выбирать сетку так, чтобы узлы x_0 и x_N были концами отрезка $[0, a]$. Если же на одном из концов задано краевое условие второго рода

$$u_x(a, t)=\mu(t), \tag{10}$$

то целесообразно полагать $x_{N-1}=a-{}^1/_2h$, $x_N=a+{}^1/_2h$, чтобы граница была полуцелой точкой. Тогда естественное разностное краевое условие

$$\frac{1}{h}(y_N-y_{N-1})=\mu(t) \tag{11}$$

обеспечивает аппроксимацию $O(h^2)$. Такой выбор сетки полезен и для других типов уравнений.

2. Неявная схема. Если схема условно устойчива, то случайное небольшое нарушение условия устойчивости может привести к быстрому нарастанию погрешностей, вплоть до «авостов» при расчетах на ЭВМ. Поэтому многие вычислители предпочитают использовать безусловно устойчивые неявные схемы.

Построим хорошую неявную схему для задачи (1). Возьмем изображенный на рис. 92 шаблон и составим схему с весами при

пространственных производных на разных слоях:

$$\frac{1}{\tau^2}(\hat{y}-2y+\check{y})=\Lambda[\sigma\hat{y}+(1-2\sigma)\,y+\sigma\check{y}]+f,$$
$$\Lambda y_n=\frac{c^2}{h^2}(y_{n+1}-2y_n+y_{n-1}); \tag{12}$$

чтобы все веса были неотрицательны, следует брать $0\leqslant\sigma\leqslant{}^1/_2$.

В граничных узлах решение определяется из краевых условий (2б).

Исследуем построенную схему. Значения решения на нулевом и первом слоях вычисляют, как и в п. 1, по формулам (3) и (4б). На остальных слоях схема (12) с краевыми условиями (2б) образует относительно $\hat{y}$ линейную систему уравнений с трехдиагональной матрицей, в которой диагональные элементы преобладают; решение этой системы существует, единственно и вычисляется методом прогонки.

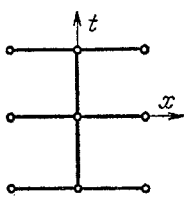


Рис. 92.

Разложением решения по формуле Тейлора нетрудно установить, что на решениях с непрерывными четвертыми производными схема (12) аппроксимирует уравнение (1а) с погрешностью $O(\tau^2+h^2)$ при любом σ.

Устойчивость проверяется методом разделения переменных. Делая подстановку (7), получим для множителя роста квадратное уравнение

$$\rho_q^2-2\alpha_q\rho_q+1=0,$$
$$\alpha_q=\frac{1-2(1-2\sigma)\,\beta_q^2}{1+4\sigma\beta_q^2},\qquad \beta_q=\frac{c\tau}{h}\sin\frac{qh}{2}. \tag{13}$$

На основании тех же рассуждений, что и в п. 1, можно сделать следующий вывод: устойчивость будет только при комплексно сопряженных корнях, т. е. при $|\alpha_q|\leqslant 1$. Отсюда вытекает условие устойчивости схемы:

$$\left(\frac{c\tau}{h}\right)^2(1-4\sigma)\leqslant 1. \tag{14}$$

Из неравенства (14) видно, что при $\sigma\geqslant{}^1/_4$ схема (12) безусловно устойчива. Если $\sigma<{}^1/_4$, то схема условно устойчива при $c\tau\leqslant h(1-4\sigma)^{-1/2}$.

Таким образом, при выборе веса ${}^1/_4\leqslant\sigma\leqslant{}^1/_2$ неявная схема (12) безусловно сходится с точностью $O(\tau^2+h^2)$.

Замечание 1. Схема (12) при $\sigma=0$ переходит в схему «крест», а условие устойчивости (14) — в условие Куранта (9).

Замечание 2. Обобщим неявную схему (12) на случай задачи с переменной скоростью звука:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(k(x,\ t)\frac{\partial u}{\partial x}\right) + f(x,\ t), \quad k(x,\ t) \equiv c^2(x,\ t) > 0, \quad (15)$$

где коэффициенты $k(x, t)$, $f(x, t)$ переменны и кусочно-непрерывны вместе со своими вторыми производными, причем разрывы неподвижны (т. е. лежат на линиях $x = \text{const}$). Предполагается, что на этих разрывах выполняются условия сопряжения $[u] = 0$, $[ku_x] = 0$.

Выберем по t равномерную сетку, а по x — специальную неравномерную сетку (у которой все точки разрыва коэффициентов являются узлами). Построим аналог наилучшей консервативной схемы (11.34), используя во всех пространственных операторах значения $k(x, t)$ со среднего слоя:

$$\frac{1}{\tau^2}(\hat{y} - 2y + \check{y}) = \Lambda[\sigma\hat{y} + (1 - 2\sigma)y + \sigma\check{y}] + \varphi,$$

$$\Lambda y_n(t) = \frac{1}{\hbar_n}\left[\varkappa_{n+1/2}(t_m)\frac{y_{n+1}(t) - y_n(t)}{h_n} - \varkappa_{n-1/2}(t_m)\frac{y_n(t) - y_{n-1}(t)}{h_{n-1}}\right],$$

$$t = t_{m-1},\ t_m,\ t_{m+1}, \quad h_n = x_{n+1} - x_n, \quad \hbar_n = \frac{1}{2}(h_n + h_{n-1}),$$

$$\varkappa_{n+1/2}(t_m) = \left[\frac{1}{h_n}\int_{x_n}^{x_{n+1}} \frac{dx}{k(x,\ t_m)}\right]^{-1}, \quad (16)$$

$$\varphi_n = \frac{1}{2\tau\hbar_n}\int_{t_{m-1}}^{t_{m+1}} dt \int_{x_n - 1/2}^{x_n + 1/2} f(x,\ t)\,dx.$$

Известно, (см. [30]), что при сделанных предположениях (и достаточно гладких начальных и граничных данных) эта схема равномерно сходится со скоростью $O(\tau^2 + \max h_n^2)$, если выполнено условие устойчивости (14).

Из схемы (16) нетрудно получить схемы для гладких и для постоянных коэффициентов на произвольных неравномерных по x сетках. В случае $k = \text{const}$ и равномерной сетки схема (16) совпадает со схемой (12).

3. **Двуслойная акустическая схема.** Уравнение второго порядка (1а) можно заменить эквивалентной ему парой уравнений первого порядка. Для этого введем потенциалы скоростей и правой части:

$$v(x,\ t) = \int_0^x u_t(\xi,\ t)\,d\xi, \quad F(x,\ t) = \int_0^x f(\xi,\ t)\,d\xi. \quad (17)$$

Функции $u(x, t)$, $v(x, t)$ удовлетворяют системе уравнений акустики

$$u_t = v_x, \quad v_t = c^2 u_x + F(x, t). \tag{18а}$$

Начальные условия (1б) с учетом (17) примут вид

$$u(x, 0) = \mu_1(x), \quad v(x, 0) = \int_0^x \mu_2(\xi)\, d\xi, \tag{18б}$$

а граничные условия (1в) останутся без изменения:

$$u(0, t) = \mu_3(t), \quad u(a, t) = \mu_4(t). \tag{18в}$$

Задача акустики (18) нередко оказывается более удобной для численного решения, чем волновое уравнение (1); в частности, она позволяет построить двуслойные разностные схемы, допускающие неравномерную сетку по t.

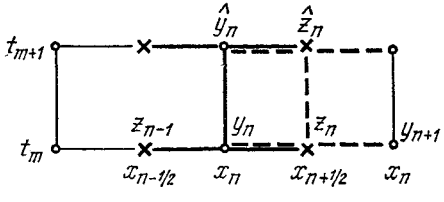


Рис. 93.

Неявная схема. Будем рассматривать в узлах неравномерной пространственной сетки величины $y_n \approx u(x_n, t)$, $0 \leqslant n \leqslant N$, а в серединах интервалов — величины $z_n \approx v(x_{n+1/2}, t)$, $0 \leqslant n \leqslant N-1$. Возьмем шаблон, изображенный на рис. 93, и составим на нем схему с весами

$$\frac{1}{\tau}(\hat{z}_n - z_n) = \frac{c^2}{h_n}[\sigma_1(\hat{y}_{n+1} - \hat{y}_n) + (1-\sigma_1)(y_{n+1} - y_n)] + \varphi_n, \tag{19а}$$

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{1}{\hbar_n}[\sigma_2(\hat{z}_n - \hat{z}_{n-1}) + (1-\sigma_2)(z_n - z_{n-1})], \tag{19б}$$

$$h_n = x_{n+1} - x_n, \quad \hbar_n = \frac{1}{2}(h_n + h_{n-1});$$

подразумевается, что $0 \leqslant \sigma_1 \leqslant 1$, $0 \leqslant \sigma_2 \leqslant 1$. Исследуем эту схему, подробно останавливаясь только на наиболее важных деталях и для простоты ограничиваясь равномерной по x сеткой.

Схема (19) составлена симметрично по переменной x, если положить $\varphi_n = F(x_{n+1/2}, t)$; поэтому нетрудно сообразить, что она имеет аппроксимацию $O(\tau + h^2)$. Если взять $\sigma_1 = \sigma_2 = 1/2$ и $\varphi_n = F(x_{n+1/2}, t_m + \tau/2)$, то схема становится вдобавок симметричной по времени и приобретает аппроксимацию $O(\tau^2 + h^2)$.

Устойчивость исследуем методом разделения переменных, рассматривая возмущения функций y и z в виде гармоник

$$y_n = \alpha e^{iqx_n}, \quad z_n = \beta e^{iqx_n}, \quad \hat{y}_n = \rho y_n, \quad \hat{z}_n = \rho z_n, \tag{20}$$

с одной и той же частотой и множителем роста, но с разными амплитудами α и β. Подставляя (20) в (19) и полагая $\varphi_n = 0$, получим для амплитуд систему линейных однородных уравнений

$$\begin{aligned} &\frac{1}{\tau}(\rho - 1)\alpha - \frac{1}{h}(\sigma_2\rho + 1 - \sigma_2)(1 - e^{-iqh})\beta = 0, \\ &\frac{c^2}{h}(\sigma_1\rho + 1 - \sigma_1)(e^{iqh} - 1)\alpha - \frac{1}{\tau}(\rho - 1)\beta = 0. \end{aligned} \tag{21}$$

Чтобы она имела нетривиальное решение, ее определитель должен обращаться в нуль. Это дает квадратное уравнение для нахождения множителей роста ρ:

$$\varepsilon\rho^2 - 2\mu\rho + \nu = 0; \tag{22а}$$

$$\varepsilon = 1 + 2\gamma_q\sigma_1\sigma_2 \geqslant 1, \qquad \mu = 1 - \gamma_q(\sigma_1 + \sigma_2 - 2\sigma_1\sigma_2),$$
$$\nu = 1 + 2\gamma_q(1 - \sigma_1)(1 - \sigma_2) \geqslant 1, \quad \gamma_q = 2\left(\frac{c\tau}{h}\sin\frac{qh}{2}\right)^2 \geqslant 0. \tag{22б}$$

Оба корня уравнения (22а) меньше единицы по модулю тогда и только тогда, если

$$\nu \leqslant \varepsilon, \quad 2|\mu| \leqslant \varepsilon + \nu. \tag{23}$$

Первое из этих неравенств очевидно, поскольку по теореме Виета $\rho_q'\rho_q'' = \nu/\varepsilon$; второе доказывается несложными, но громоздкими выкладками. Неравенства (23) выполняются для всех гармоник только в том случае, если

$$\sigma_1 + \sigma_2 \geqslant 1, \quad \left(\frac{c\tau}{h}\right)^2(2\sigma_1 - 1)(2\sigma_2 - 1) \geqslant -1, \tag{24}$$

что является условием устойчивости схемы (19). При выполнении этого условия схема сходится со скоростью, соответствующей порядку аппроксимации.

Из неравенств (24) вытекает, что если $\sigma_1 \geqslant 1/2$ и $\sigma_2 \geqslant 1/2$, то схема (19) безусловно устойчива. Если $\sigma_1 + \sigma_2 \geqslant 1$, но один из весов меньше $1/2$, то схема условно устойчива при

$$c\tau \leqslant h/\sqrt{(2\sigma_1 - 1)(1 - 2\sigma_2)}. \tag{25}$$

Если $\sigma_1 + \sigma_2 < 1$, то схема безусловно неустойчива.

Рассмотрим два частных случая схемы (19).

Явная схема. Положим $\sigma_1 = 0$, $\sigma_2 = 1$; тогда схема (19) принимает вид

$$\frac{1}{\tau}(\hat{z}_n - z_n) = \frac{c^2}{h}(y_{n+1} - y_n) + F_{n+1/2}^{m+1/2}, \quad 0 \leqslant n \leqslant N - 1, \tag{26а}$$

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{1}{h}(\hat{z}_n - \hat{z}_{n-1}), \qquad 1 \leqslant n \leqslant N - 1, \tag{26б}$$

и становится явной. В самом деле, величины $\hat{z}_n$ явно выражаются из уравнения (26а) через значения величин на исходном слое.

После того, как вычислены все значения $\hat{z}_n$, можно найти $\hat{y}^n$ также по явным формулам (26б).

Из (25) следует, что схема (26) устойчива при выполнении условия Куранта $c\tau < h$.

Заметим, что схема (26) является схемой типа «крест». В самом деле, будем считать, что величина $z_n \approx v(x_{n+1/2}, t_m)$ сдвинута на полшага по x, а величина $y_n \approx u(x_n, t_{m+1/2})$ — на полшага по t относительно узлов сетки (рис. 94). Тогда этой схеме соответствует шаблон из двух крестов, показанных на рисунке жирными линиями.

Зададим согласованные с этим шаблоном граничные данные:

$$y_0^m = \mu_3(t_{m+1/2}), \quad y_N^m = \mu_4(t_{m+1/2}), \tag{26в}$$

и начальные данные, уточненные аналогично (4б), где надо вместо τ взять $\tau/2$:

$$y_n^0 = \mu_1(x_n) + \frac{\tau}{2}\mu_2(x_n) + \frac{\tau^2}{8}\left[c^2\frac{d^2\mu_1(x_n)}{dx^2} + f(x_n, 0)\right],$$
$$z_n^0 = \int_{x_0}^{x_{n+1/2}} \mu_2(\xi)\,d\xi. \tag{26г}$$

Тогда схема (26) при выполнении условия Куранта $c\tau < h$ сходится со скоростью $O(\tau^2 + h^2)$.

Симметричная схема. Положим $\sigma_1 = \sigma_2 = 1/2$; тогда схема (19) является безусловно устойчивой и сходится со скоростью $O(\tau^2 + h^2)$. Эта схема двуслойна, поэтому она позволяет произвольно менять шаг τ в ходе расчета, обеспечивая при этом точность $O(\max \tau_m^2)$. Кроме того, поскольку значения y^0 и z^0 соответствуют моменту $t = 0$, начальные данные для расчета берут непосредственно из постановки задачи (18):

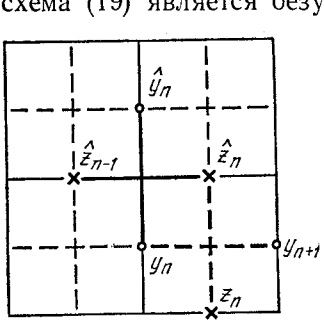


Рис. 94.

$$y_n^0 = \mu_1(x_n), \quad z_n^0 = \int_{x_0}^{x_{n+1/2}} \mu_2(\xi)\,d\xi, \tag{27}$$

без каких-либо сдвигов по времени; такая аппроксимация начальных условий является точной.

Однако при любых значениях весов σ_1, σ_2 (если только один из них не равен нулю) схема (19) неявна. Рассмотрим, как целесообразно вычислять разностное решение в этом случае. Определим $\hat{z}_n$ из уравнения (19а) и напишем аналогичное выражение для $\hat{z}_{n-1}$. Подставляя эти выражения в (19б) и полагая

для простоты $h=\text{const}$, получим

$$\frac{1}{\tau}(\hat{y}_n - y_n) = \frac{\tau c^2}{h^2}\sigma_2[\sigma_1(\hat{y}_{n+1} - 2\hat{y}_n + \hat{y}_{n-1}) + (1-\sigma_1)(y_{n+1} - 2y_n + y_{n-1})] + \frac{1}{h}(z_n - z_{n-1}) + \frac{\tau}{h}\sigma_2(\varphi_n - \varphi_{n-1}), \quad 1 \leqslant n \leqslant N-1 \quad (\sigma_1 = \sigma_2 = 1/2). \tag{28}$$

Это линейная система относительно неизвестных $\hat{y}$, имеющая трехдиагональную матрицу с преобладанием диагональных элементов; ее решение легко вычисляется прогонкой. Найдя $\hat{y}$, нетрудно определить $\hat{z}$ по явным формулам (19а).

Таким образом, симметричная схема (19) приводит к несложному вычислительному алгоритму, безусловно устойчива и имеет хорошую точность. Она является одной из лучших схем для расчета задач акустики. По аналогии с ней строятся надежные однородные схемы расчета газодинамических и других сложных задач.

Замечание 1. Разностное решение схемы (19) можно, вообще говоря, вычислять методом последовательных приближений:

$$\hat{z}_n^{(s)} = z_n + \frac{c^2\tau}{h}[\sigma_1(\hat{y}_{n+1}^{(s)} - \hat{y}_n^{(s)}) + (1-\sigma_1)(y_{n+1} - y_n)] + \varphi_n,$$
$$\hat{y}_n^{(s+1)} = y_n + \frac{\tau}{h}[\sigma_2(\hat{z}_n^{(s)} - \hat{z}_{n-1}^{(s)}) + (1-\sigma_2)(z_n - z_{n-1})]. \tag{29}$$

Однако это эквивалентно применению последовательных приближений к решению системы (28), когда в левой ее части берется $\hat{y}^{(s+1)}$, а в правой — $\hat{y}^{(s)}$. Этот метод, записанный в форме $\hat{y}^{(s+1)} = C\hat{y}^{(s)} + \psi$, сходится при $\|C\| < 1$. Выбирая одну из норм:

$$\|C\| = \max_n \left(\sum_k |C_{nk}|\right) = 4\frac{c^2\tau^2}{h^2}\sigma_1\sigma_2,$$

получим для сходимости итераций условие типа Куранта:

$$c\tau \leqslant \frac{h}{2\sqrt{\sigma_1\sigma_2}}. \tag{30}$$

Поэтому метод последовательных приближений невыгодно применять к вычислению разностного решения безусловно устойчивых схем.

Замечание 2. Для задач с разрывными или недостаточно гладкими решениями нередко используют чисто неявную схему (19) при $\sigma_1 = \sigma_2 = 1$, поскольку она подавляет «разболтку» счета. Однако на достаточно гладких решениях эта схема существенно уступает по точности симметричной схеме.

4. Инварианты. Рассмотрим запись системы уравнений акустики через *инварианты*:

$$r(x, t) = v - cu, \quad s(x, t) = v + cu. \tag{31}$$

Умножая первое из уравнений (18а) на c, прибавим его ко второму уравнению (18а) и вычтем. Получим систему уравнений, которым удовлетворяют инварианты:

$$r_t + cr_x = F(x, t), \quad s_t - cs_x = F(x, t). \tag{32а}$$

Из соотношений (18б) нетрудно получить для инвариантов начальные условия:

$$r(x, 0) = \int_0^x \mu_2(\xi)\, d\xi - c\mu_1(x), \quad s(x, 0) = \int_0^x \mu_2(\xi)\, d\xi + c\mu_1(x), \tag{32б}$$

а из соотношений (18в) — краевые условия:

$$(s - r)_{x=0} = \frac{c}{2}\mu_3(t), \quad (s - r)_{x=a} = \frac{c}{2}\mu_4(t). \tag{32в}$$

Видно, что инвариант $r(x, t)$ удовлетворяет уравнению переноса вправо (т. е. с положительной скоростью), а инвариант $s(x, t)$ — уравнению переноса влево. В случае однородной задачи ($F = \mu_3 = \mu_4 = 0$) величины r, s переносятся по соответствующим характеристикам без изменения, с чем и связано их название.

Для инвариантов можно составить разностные схемы, аналогичные схемам бегущего счета для уравнения переноса. Шаблон каждой схемы должен учитывать направление характеристики соответствующего уравнения. Простейшей будет явная схема:

$$\frac{1}{\tau}(\hat{r}_n - r_n) + \frac{c}{h}(r_n - r_{n-1}) = F_n, \quad \frac{1}{\tau}(\hat{s}_n - s_n) - \frac{c}{h}(s_{n+1} - s_n) = F_n. \tag{33}$$

Она действительно является схемой бегущего счета, и организация вычислений здесь почти такая же, как для одномерного уравнения переноса. Нетрудно показать, что при выполнении условия $c\tau \leqslant h$ эта схема устойчива, монотонна и равномерно сходится с порядком точности $O(\tau + h)$ на дважды непрерывно дифференцируемых решениях.

Счет по неявным схемам типа (10.10)—(10.12) уже не будет бегущим: для развязки счета надо знать граничное значение инварианта на новом слое, а оно выражается через то значение другого инварианта, которое считается последним. Поэтому для определения инвариантов получается линейная система с матрицей специального вида, схематически изображенного на рис. 95. Такая система решается методом исключения; экономные формулы исключения для этого случая называются формулами циклической прогонки (см. [83] и дополнение к [30]).

Схемы для инвариантов можно переписать в терминах исходных переменных. Так, складывая и вычитая уравнения (33), получим для внутренних точек области

$$\begin{aligned}&\frac{1}{\tau}(\hat{y}_n - y_n) - \frac{1}{2h}(z_{n+1} - z_{n-1}) = \frac{c}{2h}(y_{n+1} - 2y_n + y_{n-1}),\\ &\frac{1}{\tau}(\hat{z}_n - z_n) - \frac{c^2}{2h}(y_{n+1} - y_{n-1}) = \frac{c}{2h}(z_{n+1} - 2z_n + z_{n-1}) + F_n.\end{aligned} \tag{34}$$

Каждое из уравнений (34) содержит члены, соответствующие явной схеме (6) для уравнения теплопроводности с коэффициентом $k = ch/2$. Отсюда понятно, что исходная схема (33) будет хорошо сглаживать разрывы начальных данных, т. е. иметь аппроксимационную вязкость. Условие устойчивости явной схемы (6) $2k\tau \leqslant h^2$ совпадает с условием устойчивости исходной схемы.

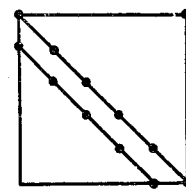


Рис. 95.

Схемы в инвариантах обладают многими достоинствами. Однако широкого распространения они не получили, потому что их нелегко обобщить на нелинейные задачи.

5. Явная многомерная схема. Волновое уравнение в p-мерной изотропной среде (либо в неизотропной среде, если у тензора упругости отличны от нуля только диагональные компоненты) имеет вид

$$\begin{aligned}&\frac{\partial^2 u}{\partial t^2} = \sum_{\alpha=1}^{p} A_\alpha u + f(\boldsymbol{x}, t),\\ &A_\alpha = \frac{\partial}{\partial x_\alpha}\left(k_\alpha(\boldsymbol{x}, t)\frac{\partial}{\partial x_\alpha}\right), \quad \boldsymbol{x} = \{x_1, x_2, \ldots, x_p\} \in G.\end{aligned} \tag{35а}$$

Рассмотрим задачу нахождения решения уравнения (35а) с начальными условиями и с краевыми условиями первого рода:

$$\begin{aligned}&u(\boldsymbol{x}, 0) = \mu_1(\boldsymbol{x}), \quad u_t(\boldsymbol{x}, 0) = \mu_2(\boldsymbol{x}), \quad \boldsymbol{x} \in G;\\ &u(\boldsymbol{x}, t)_{\Gamma(G)} = \mu_3(\boldsymbol{x}, t).\end{aligned} \tag{35б}$$

Обычно многомерные разностные схемы составляют непосредственно для задачи (35). В принципе, можно заменить уравнение (35а) системой уравнений первого порядка; однако это менее выгодно, чем в одномерном случае.

Схема «крест» строится аналогично одномерной схеме (2) на шаблоне, вид которого для двух измерений показан на рис. 96. При произвольном числе измерений эта схема для уравнения (35а) имеет вид

$$\frac{1}{\tau^2}(\hat{y} - 2y + \check{y}) = \sum_{\alpha=1}^{p} \Lambda_\alpha y + f; \tag{36}$$

в случае переменных коэффициентов операторы Λ_α выбираются аналогично наилучшей консервативной схеме (16).

Схема (36) — явная трехслойная; организация вычислений по ней одинаково проста при любом числе измерений. Нетрудно проверить, что она имеет аппроксимацию $O\left(\tau^2 + \sum_{\alpha=1}^{p} h_\alpha^2\right)$. Ее устойчивость можно исследовать методом разделения переменных, подставляя в (36) многомерную гармонику:

$$y = \exp\left(i \sum_{\alpha=1}^{p} q_\alpha x_\alpha\right), \quad \hat{y} = \rho y, \quad y = \rho \check{y}. \tag{37}$$

Учитывая, что

$$\Lambda_\alpha y \to -4 \frac{k_\alpha}{h_\alpha^2} \sin^2 \frac{q_\alpha h_\alpha}{2}, \tag{38}$$

получим для множителя роста квадратное уравнение

$$\rho^2 - 2(1 - 2\gamma)\rho + 1 = 0, \quad \gamma = \tau^2 \sum_{\alpha=1}^{p} \frac{k_\alpha}{h_\alpha^2} \sin^2 \frac{q_\alpha h_\alpha}{2}.$$

Это уравнение аналогично одномерному уравнению (8); анализ его корней показывает, что схема (36) устойчива при выполнении условия

$$\tau < \left(\sum_{\alpha=1}^{p} \frac{k_\alpha}{h_\alpha^2}\right)^{-1/2} \sim \frac{h}{\sqrt{pk(x, t)}}, \tag{39}$$

являющегося обобщением условия Куранта (9). Это естественное условие, а точность схемы неплохая. Поэтому схему «крест» используют в расчетах, если не требуется особенно высокой надежности («безавостности») вычислений.

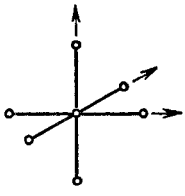


Рис. 96

Таким образом, численный расчет многомерных задач акустики не вызывает принципиальных затруднений.

6. Факторизованные схемы. В «больших задачах», где небольшое нарушение условия устойчивости любого из разностных уравнений в ходе расчета легко приводит к «авостам» ЭВМ, целесообразно использовать безусловно устойчивые многомерные экономичные схемы, несмотря на то что они сложнее явных схем.

Для гиперболических уравнений локально-одномерные схемы имеют сравнительно громоздкий и искусственный вид. Более удобны в данном случае факторизованные схемы (схемы с расщеплением); рассмотрим их.

Исходная схема. Для многомерной задачи (35) рассмотрим аналог неявной схемы с весами (12), который будем называть *исходной* схемой:

$$\frac{1}{\tau^2}(\hat{y}-2y+\check{y})=\sum_{\alpha=1}^{p}\Lambda_\alpha[\sigma\hat{y}+(1-2\sigma)y+\sigma\check{y}]+f, \quad 0\leqslant\sigma\leqslant\frac{1}{2}; \tag{40}$$

операторы Λ_α — трехточечные и вычисляются по формуле (16). Эта схема симметрична по пространству и времени, поэтому легко видеть, что она имеет аппроксимацию $O\left(\tau^2+\sum_{\alpha=1}^{p}h_\alpha^2\right)$ при любых значениях веса σ. Методом разделения переменных можно показать, что при $\sigma\geqslant{}^1/_4$ схема (40) безусловно устойчива.

Однако исходная схема, которую можно переписать в виде

$$\left(E-\tau^2\sigma\sum_{\alpha=1}^{p}\Lambda_\alpha\right)\hat{y}=$$
$$=\left[2E+\tau^2(1-2\sigma)\sum_{\alpha=1}^{p}\Lambda_\alpha\right]y-\left(E-\tau^2\sigma\sum_{\alpha=1}^{p}\Lambda_\alpha\right)\check{y}+\tau^2f, \tag{41}$$

содержит на верхнем слое выражение

$$B\hat{y}, \quad \text{где} \quad B=E-\tau^2\sigma\sum_{\alpha=1}^{p}\Lambda_\alpha. \tag{42}$$

Оператору B, встречавшемуся (почти в той же форме) в схеме (11.68) для многомерного уравнения теплопроводности, соответствует ленточная матрица типа, изображенного на рис. 89 (гл. XII, § 2, п. 3). Решение линейной системы (41) не сводится к одномерным прогонкам, и оператор B оказывается труднообратимым. Поэтому исходная схема (40) неэкономична.

Факторизованная схема. Оператор (42) можно приближенно заменить факторизованным оператором

$$C\equiv\prod_{\alpha=1}^{p}(E-\tau^2\sigma\Lambda_\alpha)=$$
$$=E-\tau^2\sigma\sum_{\alpha=1}^{p}\Lambda_\alpha+\tau^4\sigma^2\sum_{\alpha=1}^{p-1}\sum_{\beta=1+\alpha}^{p}\Lambda_\alpha\Lambda_\beta+\ldots=B+O(\tau^4), \tag{43}$$

т. е. приближенно расщепить B на произведение одномерных операторов. Заметим, что перестановочности операторов Λ_α для этого не требуется. Заменяя в исходной схеме (41) оператор B

на C, получим *факторизованную схему*:

$$\prod_{\alpha=1}^{p}(E-\tau^2\sigma\Lambda_\alpha)\,\hat{y}=$$

$$=\left[2E+\tau^2(1-2\sigma)\sum_{\alpha=1}^{p}\Lambda_\alpha\right]y-\left(E-\tau^2\sigma\sum_{\alpha=1}^{p}\Lambda_\alpha\right)\check{y}+\tau^2 f, \quad (44)$$

отличающуюся от исходной. Исследуем ее.

Аппроксимация. Преобразуя факторизованную схему (44) к форме типа (40) и учитывая соотношение (43), получим

$$\frac{1}{\tau^2}(\hat{y}-2y+\check{y})=$$

$$=\sum_{\alpha=1}^{p}\Lambda_\alpha[\sigma\hat{y}+(1-2\sigma)\,y+\sigma\check{y}]+f-\tau^2\sigma^2\sum_{\alpha=1}^{p-1}\sum_{\beta=1+\alpha}^{p}\Lambda_\alpha\Lambda_\beta\hat{y}+\ldots,$$

что отличается от схемы (40) на члены $O(\tau^2)$. Поскольку исходная схема (40) имеет второй порядок аппроксимации, то факторизованная схема (44) также имеет аппроксимацию $O(\tau^2+\sum h_\alpha^2)$.

Устойчивость исследуем методом разделения переменных, подставляя в (44) многомерную гармонику (37). С учетом соотношения (38) получим для множителя роста квадратное уравнение

$$\varepsilon\rho^2-2\mu\rho+\nu=0; \quad (45а)$$

$$\varepsilon=\prod_\alpha(1+2\sigma\gamma_\alpha)\geqslant 1, \quad \mu=1-(1-2\sigma)\sum_\alpha\gamma_\alpha,$$

$$\nu=1+2\sigma\sum_\alpha\gamma_\alpha\geqslant 1, \quad \gamma_\alpha=2\frac{k_\alpha\tau^2}{h_\alpha^2}\left(\sin\frac{q_\alpha h_\alpha}{2}\right)^2\geqslant 0. \quad (45б)$$

Уравнение (45а) аналогично уравнению (22а); поэтому оба его корня не превышают единицы по модулю только в том случае, если выполняются неравенства (23):

$$\nu\leqslant\varepsilon, \quad 2|\mu|\leqslant\varepsilon+\nu.$$

Первое из этих неравенств для коэффициентов (45б) всегда справедливо. Второе неравенство заменим несколько более жестким требованием $|\mu|\leqslant\nu$; нетрудно проверить, что оно выполняется при

$$\tau^2(1-4\sigma)\sum_{\alpha=1}^{p}\frac{k_\alpha}{h_\alpha^2}\leqslant 1. \quad (46)$$

Это и есть достаточное условие устойчивости схемы (44). В частности, если $\sigma\geqslant 1/4$, то условие (46) выполняется при любых шагах τ, h_α и схема является безусловно устойчивой.

Безусловная сходимость факторизованной схемы (44) со скоростью $O(\tau^2+\sum h_\alpha^2)$ при $^1/_4 \leqslant \sigma \leqslant {}^1/_2$ следует из сказанного выше.

Вычисление разностного решения сводится к последовательности одномерных прогонок по всем направлениям x_α. В самом деле, факторизованный оператор C есть произведение одномерных трехточечных операторов $E-\tau^2\sigma\Lambda_\alpha$, а каждый такой оператор обращается одномерной прогонкой. Тем самым, схема (44) экономична.

Таким образом, для многомерных задач акустики факторизацией удается построить безусловно устойчивые экономичные схемы, сходящиеся со скоростью $O(\tau^2+\sum h_\alpha^2)$.

§ 2. Одномерные уравнения газодинамики

1. Лагранжева форма записи. Одномерные уравнения газодинамики являются хорошим приближением для описания ряда интересных задач: плоского течения сжимаемого газа в трубе, взрыва сферического или длинного цилиндрического заряда в газе, кумулятивных эффектов в мишенях при управляемом термоядерном синтезе (в последней задаче существенна также теплопроводность и другие эффекты) и т. д. Мы рассмотрим простые, но эффективные разностные схемы решения уравнений газодинамики без теплопроводности *).

Уравнения газодинамики могут записываться в различных формах — эйлеровой и лагранжевой. В эйлеровой форме производные по времени выражают изменение величин в данной точке пространства, а в лагранжевой — изменение характеристик данной материальной точки. Эти производные связаны соотношением

$$\left(\frac{\partial}{\partial t}\right)_л = \left(\frac{\partial}{\partial t}\right)_э + (v\nabla).$$

Если нас интересуют параметры потока в заданной пространственной области (течение газа в трубе), то естественно выбрать эйлеровы координаты. Если нам нужно исследовать поведение некоторой массы вещества, то целесообразно применение лагранжевых координат. Особенно выгодны лагранжевы координаты для задач в слоистых средах, потому что они позволяют легко следить за границами раздела различных сред.

Большинство одномерных задач относится ко второму типу (в многомерном случае это не так). Поэтому здесь мы рассмотрим только уравнения газодинамики в лагранжевых координатах.

*) Уравнения газодинамики и исследование простейших газодинамических течений приведены, например, в [11, 19], а более подробное изложение методов решения — в [27, 28, 34].

Сначала запишем их в такой форме, когда производная по времени лагранжева, а пространственные координаты — обычные:

$$\frac{\partial \rho}{\partial t} + \rho \operatorname{div} \boldsymbol{v} = 0, \tag{47}$$

$$\rho \frac{\partial \boldsymbol{v}}{\partial t} + \operatorname{grad} p = 0, \tag{48}$$

$$\rho \frac{\partial \varepsilon}{\partial t} + p \operatorname{div} \boldsymbol{v} = 0; \tag{49}$$

здесь ρ — плотность, $\boldsymbol{v}$ — скорость, p — давление и $\varepsilon(p, \rho)$ — внутренняя энергия единицы массы, зависимость которой от давления и плотности считается известной.

Уравнение (48) выражает закон сохранения импульса и во всех численных расчетах используется именно в такой форме. Уравнение изменения энергии (49) обычно удобнее преобразовать. Если подставить в него $\operatorname{div} \boldsymbol{v}$, определенную из уравнения (47), то получим особенно простую форму записи:

$$\frac{\partial \varepsilon}{\partial t} + p \frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) = 0. \tag{50а}$$

Если умножить уравнение (48) на $\boldsymbol{v}$ и прибавить к уравнению (49), то получим другую форму — закон сохранения полной энергии:

$$\rho \frac{\partial}{\partial t}\left(\varepsilon + \frac{1}{2}\boldsymbol{v}^2\right) + \operatorname{div}(p\boldsymbol{v}) = 0. \tag{50б}$$

Уравнение неразрывности (47) выражает закон сохранения массы. Его тоже обычно преобразуют, но уже после приведения уравнений к одномерной записи.

В том, что указанные уравнения являются законами сохранения, нетрудно убедиться. Например, проинтегрируем (50б) по объему dV, занятому некоторой массой вещества, второй интеграл преобразуем к поверхностному, а в первом интеграле заменим $\rho\, dV$ на dm, после чего производную по времени можно вынести за знак интеграла. Тогда получим

$$\frac{\partial}{\partial t}\int \left(\varepsilon + \frac{1}{2}\boldsymbol{v}^2\right) dm + \oint p\boldsymbol{v}\, d\boldsymbol{S} = 0.$$

Здесь первый интеграл есть полная (внутренняя и кинетическая) энергия данной массы газа; второй равен работе в единицу времени сил давления на поверхности, ограничивающей данную массу газа. Действительно, это закон сохранения энергии.

Одномерные задачи бывают трех типов: с плоской, цилиндрической или сферической симметрией. Введем показатель симметрии ν, равный для этих случаев соответственно 0, 1 или 2. Масса слоя толщиной dr в этих случаях равна (рис. 97)

$$dm = \rho r^{\nu}\, dr \tag{51}$$

с точностью до численного множителя, равного 1, 2π или 4π. При помощи соотношения (51) введем массовую координату данной материальной точки:

$$m(r) = \int_{r_0}^{r} \rho(\xi)\, \xi^{\nu}\, d\xi. \tag{52}$$

По закону сохранения вещества массовая координата материальной точки не меняется со временем; поэтому такая координата позволяет легко следить за каждой частицей вещества и, в частности, за границей раздела слоев.

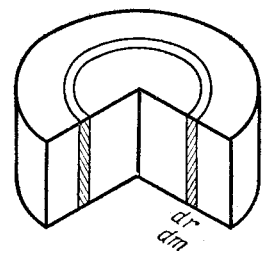


Рис. 97.

Преобразуем уравнения газодинамики в одномерном случае к лагранжевой форме. В качестве первого уравнения возьмем определение скорости:

$$\frac{\partial r}{\partial t} = v. \tag{53}$$

Уравнение неразрывности (47), выражающее закон сохранения массы, заменим имеющим тот же смысл соотношением (51), записывая следующим образом:

$$\frac{\partial}{\partial m}\left(r^{\nu+1}\right) = \frac{\nu+1}{\rho}. \tag{54}$$

В уравнениях импульса (48) и энергии (50б) перейдем к производной по массовой координате, то есть в одномерных выражениях

$$\operatorname{grad} = \frac{\partial}{\partial r},$$

$$\operatorname{div} = r^{-\nu}\frac{\partial}{\partial r}\left(r^{\nu}\ldots\right)$$

заменим ∂r на ∂m при помощи соотношения (51), и получим

$$\frac{\partial v}{\partial t} + r^{\nu}\frac{\partial p}{\partial m} = 0, \tag{55}$$

$$\frac{\partial}{\partial t}\left(\varepsilon + \frac{1}{2}v^2\right) + \frac{\partial}{\partial m}\left(r^{\nu} v p\right) = 0. \tag{56а}$$

Уравнение энергии можно взять также в форме (50а):

$$\frac{\partial \varepsilon}{\partial t} + p\frac{\partial}{\partial t}\left(\frac{1}{\rho}\right) = 0. \tag{56б}$$

Система (53)—(56) является лагранжевой формой записи уравнений одномерной газодинамики. В большинстве численных расчетов используется эта именно форма.

2. Псевдовязкость. Уравнения (53)—(56) составляют гиперболическую квазилинейную систему. Из курса газодинамики известно, что среди ее решений есть сильные разрывы — ударные волны. В главе IX мы видели, что для разностного расчета таких решений надо изменять уравнения, вводя в них специально подобранные диссипативные члены.

В газодинамике такие члены удается найти из физических соображений. Дело в том, что уравнение газодинамики сравнительно грубо описывают поведение газа. Эти уравнения выводятся из кинетического уравнения Больцмана для функции распределения молекул. Если при выводе учесть эффекты диффузии молекул, то в уравнениях газодинамики появятся так называемые вязкие члены. Например, уравнение импульса (48) примет вид (см. [19])

$$\rho \frac{\partial \boldsymbol{v}}{\partial t} = -\operatorname{grad} p + \eta \Delta \boldsymbol{v} + \zeta \operatorname{grad} \operatorname{div} \boldsymbol{v}, \qquad \zeta > \frac{1}{3}\eta > 0, \tag{57}$$

где η и ζ являются коэффициентами физической вязкости. Учет физической вязкости приводит к изменению качественного характера решения: плоские ударные волны превращаются в аналитические решения, в которых скачки сглажены и имеют эффективную ширину порядка длины свободного пробега молекул. Качественно это легко понять на примере плоского течения, где уравнение (57) принимает форму

$$\rho \frac{\partial v}{\partial t} = (\eta + \zeta) \frac{\partial^2 v}{\partial x^2} - \frac{\partial p}{\partial x}, \tag{58}$$

напоминающую уравнение теплопроводности; видно, что вязкий член должен сглаживать разрывы решения.

Обычно в численных расчетах довольствуются только вторым вязким членом в уравнении (57) и считают коэффициент ζ слабо меняющимся. Тогда этот член можно объединить с давлением:

$$-\operatorname{grad} p + \zeta \operatorname{grad} \operatorname{div} \boldsymbol{v} \approx -\operatorname{grad} (p - \zeta \operatorname{div} \boldsymbol{v}), \tag{59}$$

и рассматривать величину

$$w_1 = -\zeta \operatorname{div} \boldsymbol{v} \tag{60}$$

как вязкое давление. При этом в уравнение энергии вместо обычного давления также ставят величину $p + w_1$, называя ее полным давлением.

Вязкость w_1 называется *линейной*. Она приводит к «размазыванию» ударной волны со скачком скорости δv на интервал

$$\delta r \approx \frac{8\zeta}{(\gamma+1)\rho\, \delta v}, \tag{61}$$

где γ — показатель политропы вещества. Физический коэффициент вязкости очень мал и дает ничтожно малое сглаживание. Для расчетов по разностным схемам необходимо сглаживание на несколько интервалов сетки. Поэтому в численных расчетах величину ζ приходится увеличивать на много порядков по сравнению с ее физическим значением.

Для численных расчетов необходимо введение вязкости лишь в окрестности ударных волн. Но вязкий член в (59) присутствует во всех точках пространства. Поскольку в численных расчетах коэффициент ζ выбирается много больше физического коэффициента, то наличие псевдовязкости, помимо положительного эффекта — сглаживания разрывов, — приводит к отрицательному — внесению заметной погрешности.

Чтобы уменьшить эту погрешность, Нейман и Рихтмайер [72] предложили выбирать коэффициент псевдовязкости большим в окрестности скачков скорости δv и малым в зонах гладких течений, где скорости соседних точек близки. Для этого они положили

$$\zeta = \zeta_0 |\operatorname{div} \boldsymbol{v}|, \tag{62}$$

где ζ_0 — коэффициент, небольшой по величине. Такая псевдовязкость называется *квадратичной*, потому что она приводит к вязкому давлению:

$$w_2 = -\zeta_0 (\operatorname{sign} \operatorname{div} \boldsymbol{v}) \cdot (\operatorname{div} \boldsymbol{v})^2. \tag{63}$$

Переписывая (62) в виде $\zeta \approx \zeta_0 (\delta v/\delta r)$ и подставляя в (61), нетрудно убедиться, что квадратичная псевдовязкость сглаживает скачок δv любой интенсивности на один и тот же интервал:

$$\delta r \approx \sqrt{\frac{8\zeta_0}{(\gamma+1)\rho}}. \tag{64}$$

Обычно коэффициент псевдовязкости ζ_0 выбирают так, чтобы δr равнялось 2—3 шагам разностной сетки.

В главе X, § 2 было проведено строгое исследование квадратичной псевдовязкости на примере простейшего квазилинейного уравнения (10.44); при этом для зоны сглаживания сильного разрыва было получено выражение (10.51)—(10.52), сходное с (64).

Линейная вязкость приводит к монотонным (или почти монотонным) разностным решениям, так как ей соответствуют аналитические точные решения, которые хорошо аппроксимируются разностными схемами; зато фронты скачков при этом сильно сглажены. Квадратичная вязкость приводит к более крутым фронтам; но ей соответствуют точные решения с разрывами первой или второй производной, поэтому разностное решение немонотонно вблизи слабых и сильных разрывов. Нередко используют комби-

нацию линейной и квадратичной вязкости $w=\alpha_1 w_1+\alpha_2 w_2$ с экспериментально подобранными коэффициентами.

Поскольку $\operatorname{div} \boldsymbol{v}=-\rho^{-1}(\partial\rho/\partial t)$, то вязкое давление положительно при сжатии и отрицательно при разрежении вещества. Сильными разрывами являются только ударные волны, поэтому для сглаживания разрывов можно вводить вязкость только при сжатиях. При $(\partial\rho/\partial t)<0$ присутствие псевдовязкости не обязательно и даже уменьшает точность расчета. Поэтому обычно полагают

$$w=\begin{cases}-\zeta_1 \operatorname{div}\boldsymbol{v}+\zeta_0(\operatorname{div}\boldsymbol{v})^2 & \text{при } \operatorname{div}\boldsymbol{v}<0,\\ 0 & \text{при } \operatorname{div}\boldsymbol{v}\geqslant 0.\end{cases} \tag{65}$$

Этот вид псевдовязкости независимо предложен рядом авторов (см. [27]).

3. Схема «крест». Это наиболее простая и довольно точная однородная разностная схема счета газодинамики. Ее шаблон приведен на рис. 98; значения радиусов приписываются узлам сетки, значения скорости — границам пространственных интервалов на полуцелых слоях, а значения плотности, давления и внутренней энергии — серединам интервалов на целых слоях.

Построение схемы напоминает акустический «крест». Для простоты записи выберем равномерные по массе и времени шаги m и τ и аппроксимируем систему (53)—(56б) следующими разностными уравнениями:

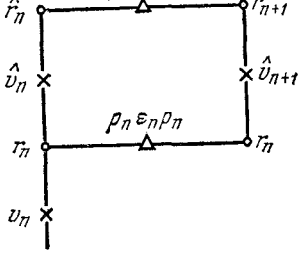


Рис. 98.

$$\hat{v}_n=v_n+\frac{\tau}{m}(g_{n-1}-g_n)\,r_n^{\nu}, \tag{66а}$$

$$g=p+w,$$

$$\hat{r}_n=r_n+\tau\hat{v}_n, \tag{66б}$$

$$\hat{\rho}_n=\frac{(\nu+1)\,m}{\hat{r}_{n+1}^{\nu+1}-\hat{r}_n^{\nu+1}}, \tag{66в}$$

$$\hat{\varepsilon}_n=\varepsilon_n+\frac{1}{2}(\hat{g}_n+g_n)\left(\frac{1}{\rho_n}-\frac{1}{\hat{\rho}_n}\right). \tag{66г}$$

Эти уравнения записаны в том порядке, который удобен для вычислений.

Обсудим разностное выражение для вязкого давления (65). Чтобы выполнить предельный переход от разностной схемы к уравнениям газодинамики, надо сначала устремить τ и m к нулю при фиксированном коэффициенте вязкости, а затем построить серию таких предельных решений для неограниченно уменьшающихся значений ζ. Но это очень трудоемко. Поэтому на практике объединяют эти предельные переходы в один общий, полагая $\zeta_0=\mu_0\rho(\delta r)^2$ и $\zeta_1=\mu_1\rho c\delta r$, хотя законность такой процедуры не

доказана (плотность введена в формулу для того, чтобы коэффициенты μ_k были безразмерны). Таким образом, вязкое давление (65) принимает вид

$$\omega_n = \mu_0 \rho_n (v_{n+1} - v_n)^2 - \mu_1 c \rho_n (v_{n+1} - v_n) \quad \text{при} \quad v_{n+1} - v_n < 0, \qquad (67а)$$

$$\omega_n = 0 \quad \text{при} \quad v_{n+1} - v_n \geqslant 0. \qquad (67б)$$

где $c \approx \sqrt{\partial p / \partial \rho}$ — скорость звука. Выражение (67) написано для плоского случая; но обычно им пользуются при любой симметрии задачи.

Аппроксимация. Из вида шаблона на рис. 98 и симметричного написания схемы (66) нетрудно заметить, что на течениях без сжатий, когда псевдовязкость (67) обращается в нуль, схема «крест» имеет локальную аппроксимацию $O(\tau^2 + h^2)$.

На течениях со сжатиями (в том числе — с ударными волнами) псевдовязкость отлична от нуля. Правда, квадратичный член в (67а) имеет величину $O(h^2)$; но линейный член имеет величину $O(h)$ и, тем самым, ухудшает порядок аппроксимации. Кроме того, вязкие члены записываются не вполне симметрично по времени. В итоге аппроксимация ухудшается до $O(\tau + h)$.

Нахождение разностного решения. Схема (66) — явная; вычисления по ней проводятся следующим образом. Пусть все величины на исходном слое известны. Тогда из разностного уравнения импульса (66а) находим $\hat{v}_n$ во всех интервалах; затем из второго уравнения (66б) определяем $\hat{r}_n$, а из уравнения (66в) — $\hat{\rho}_n$.

Последним решается уравнение энергии (66г). Формально оно является неявным алгебраическим уравнением для определения $\hat{\varepsilon}_n(\hat{p}_n, \hat{\rho}_n)$ в данном интервале. Но при каждом значении индекса n уравнения (66г) решаются независимо, не образуя связанной системы уравнений, так что разностная схема, по существу, остается явной.

Замечание 1. Уравнение энергии в (66) можно сделать явным, используя в нем только значение g_n с исходного слоя:

$$\hat{\varepsilon}_n = \varepsilon_n + g_n \left(\frac{1}{\rho_n} - \frac{1}{\hat{\rho}_n} \right). \qquad (68)$$

Это несколько упрощает расчет, не влияет на устойчивость, но заметно ухудшает точность, так как погрешность аппроксимации становится $O(\tau + h^2)$ даже на гладких течениях. Такой вариант используется редко.

Устойчивость схемы можно исследовать методом разделения переменных, линеаризируя схему и замораживая коэффициенты. Громоздкие выкладки приводят к условию устойчивости типа Куранта. Например, на гладких течениях с нулевой вязко-

стью схема устойчива при

$$\frac{\tau}{\Delta r} \leqslant \rho \sqrt{\left(\frac{\partial \varepsilon}{\partial p}\right)_\rho \Big/ \left[p + \left(\frac{\partial \varepsilon}{\partial V}\right)_p\right]}, \quad V = \frac{1}{\rho}. \tag{69}$$

Для идеального газа $\varepsilon = pV/(\gamma - 1)$ и условие (69) принимает вид $c\tau \leqslant \Delta r$, где $c = \sqrt{\gamma p/\rho}$ есть адиабатическая скорость звука. На течениях с ненулевой вязкостью ограничение на шаг несколько более сильное; при квадратичной вязкости условие устойчивости принимает вид

$$\frac{\tau}{\Delta r} \leqslant \rho \sqrt{\frac{\partial \varepsilon}{\partial p} \Big/ \left(p + w + \frac{\partial \varepsilon}{\partial V}\right)} \left(\sqrt{1+\theta} + \sqrt{\theta}\right)^{-1},$$
$$\theta = 4\,(\mu_0 \rho\, \Delta v)^2 \left(\frac{\partial \varepsilon}{\partial p}\right) \Big/ \left(p + w + \frac{\partial \varepsilon}{\partial V}\right) \approx (2\mu_0\, \Delta v/c)^2, \tag{70}$$

где Δv — скачок скорости на ударной волне. Хотя это исследование не является строгим, тем не менее данное условие устойчивости хорошо подтверждается на практике.

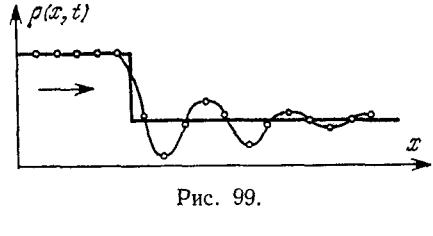


Рис. 99.

Таким образом, «крест» — условно устойчивая схема. Отметим любопытное обстоятельство. Для расчета гладких течений вязкость не нужна. А если рассчитать без вязкости ударную волну (выбирая небольшое $\tau/\Delta r$, удовлетворяющее условию (70)), то получим «разболтку», изображенную на рис. 99. Этот расчет устойчив, поскольку амплитуда колебаний не возрастает со временем. Но сходимости к физически правильному решению при $\tau \to 0$, $h \to 0$ нет, так как на разрыве потеряна аппроксимация.

Сходимость газодинамической схемы «крест» не доказана. Однако эта схема успешно используется в расчетах примерно с 1950 г. и проверена на многих трудных задачах с известными точными решениями. При стремлении шагов к нулю наблюдалась сходимость к правильному решению, если шаги удовлетворяли условию устойчивости.

Замечание 2. Схема (66) неконсервативна; однако ее дисбаланс стремится к нулю при $\tau = \text{const} \cdot h \to 0$.

Замечание 3. Газодинамические задачи с очень тонкими слоями особенно трудны для расчета. В самом деле, если $r_{n+1} \approx r_n$, то для вычисления ρ_n с удовлетворительной точностью по формуле (66в) надо знать радиусы с очень высокой точностью, сравнимой с ошибками округления на ЭВМ. В подобных задачах иногда приходится вести расчет с двойным числом знаков или специально видоизменять разностную схему.

4. Неявная консервативная схема. Есть ряд задач, в которых локальная скорость звука в некоторых участках много больше скорости наиболее важных физических процессов. В таких задачах условие Куранта слишком сильно ограничивает шаг и выгоднее использовать абсолютно устойчивые схемы.

Составим неявную схему. Припишем все сеточные величины целым слоям по времени и выберем шаблон, изображенный на рис. 100. Аппроксимируем консервативную систему (53)—(56а) следующими разностными уравнениями:

$$\hat{v}_n = v_n + \frac{\tau}{m}\hat{r}_n^{\nu}(\hat{g}_{n-1} - \hat{g}_n), \qquad g = p + w; \tag{71а}$$

$$\hat{r}_n = r_n + \tau\hat{v}_n; \tag{71б}$$

$$\hat{\rho}_n = (\nu+1)\, m/(\hat{r}_{n+1}^{\nu+1} - \hat{r}_n^{\nu+1}); \tag{71в}$$

$$\hat{w}_n = \mu_0\hat{\rho}_n(\hat{v}_{n+1} - \hat{v}_n)^2 \text{ при } \hat{v}_{n+1} < \hat{v}_n, \text{ иначе } \hat{w}_n = 0; \tag{71г}$$

$$\hat{\varepsilon}_n + \frac{1}{4}(\hat{v}_{n+1}^2 + \hat{v}_n^2) = \varepsilon_n + \frac{1}{4}(v_{n+1}^2 + v_n^2) + \\ + \frac{\tau}{2m}\left[\hat{r}_n^{\nu}\hat{v}_n(\hat{g}_{n-1} + \hat{g}_n) - \hat{r}_{n+1}^{\nu}\hat{v}_{n+1}(\hat{g}_n + \hat{g}_{n+1})\right]. \tag{71д}$$

Это — консервативная схема. Первые два уравнения взяты чисто неявными для хорошего подавления «разболтки» счета. Уравнение энергии симметрично по времени; чисто неявным его брать невыгодно, поскольку при этом точность расчета заметно ухудшается.

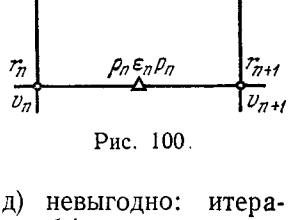


Рис. 100.

Вычисление разностного решения здесь существенно сложней, чем для явной схемы (66). Аналогично задачам акустики (§ 1, п. 3, замечание 1) можно показать, что применять метод последовательных приближений для решения всей цепочки уравнений (71а)—(71д) невыгодно: итерации сходятся при выполнении условия $c\tau \lesssim \Delta r$, что лишает неявную схему всех ее преимуществ.

Поэтому систему (71) линеаризируют и, как в задачах акустики, преобразуют к форме, решаемой прогонкой. Рассмотрим ход решения в случае разных режимов газодинамических течений, для простоты ограничиваясь плоским случаем ($\nu = 0$).

Изотермический случай. Если температура вещества постоянна *), то давление $p = p(T, \rho)$ зависит только от плотности. При этом уравнение энергии (56) становится излишним,

*) Это приближение справедливо в случае очень высоких температур, когда тепловые потоки настолько велики, что практически мгновенно выравнивают температуру во всех точках пространства.

поскольку система (53)—(55) при заданной зависимости $p(\rho)$ полностью определяет решение. Соответственно в численном расчете следует ограничиться уравнениями (71а)—(71г).

Положим $\hat{v}_n = \hat{v}_n^{(s)} + \delta\hat{v}_n$. Подставляя это выражение в уравнение (71а) и линеаризируя это уравнение относительно приращений всех величин на новом слое, получим

$$\hat{v}_n^{(s)} + \delta\hat{v}_n = v_n + \frac{\tau}{m}\left(\hat{g}_{n-1}^{(s)} - \hat{g}_n^{(s)}\right) + \frac{\tau}{m}\left[\left(\frac{\partial\hat{g}}{\partial\hat{\rho}}\right)_{n-1}\delta\hat{\rho}_{n-1} - \left(\frac{\partial\hat{g}}{\partial\hat{\rho}}\right)_n\delta\hat{\rho}_n\right]. \quad (72)$$

Из уравнений (71в) и (71б) найдем вариации

$$\delta\hat{\rho}_n = -m\,(\delta\hat{r}_{n+1} - \delta\hat{r}_n)/(\hat{r}_{n+1} - \hat{r}_n)^2, \qquad \delta\hat{r}_n = \tau\,\delta\hat{v}_n. \quad (73)$$

Подставляя их в (72), получим для определения $\delta\hat{v}$ линейную систему с трехдиагональной матрицей:

$$\varkappa_{n-1}\delta\hat{v}_{n-1} - (1 + \varkappa_{n-1} + \varkappa_n)\,\delta\hat{v}_n + \varkappa_n\delta\hat{v}_{n+1} = \\ = \hat{v}_n^{(s)} - v_n - \frac{\tau}{m}\left(\hat{g}_{n-1}^{(s)} - \hat{g}_n^{(s)}\right), \quad (74)$$

$$\varkappa_n = \left(\frac{\tau}{\hat{r}_{n+1}^{(s)} - \hat{r}_n^{(s)}}\right)^2\left(\frac{\partial\hat{g}}{\partial\hat{\rho}}\right)_n^{(s)} > 0,$$

решаемую прогонкой.

Пренебрегая пока вязкостью (т. е. полагая $g = p$), организуем вычисления следующим образом. Выберем в качестве нулевого приближения

$$\hat{r}_n^{(0)} = r_n, \quad \hat{v}_n^{(0)} = v_n, \quad \hat{\rho}_n^{(0)} = \rho_n, \quad \hat{g}_n^{(0)} = p_n. \quad (75)$$

Затем определим из уравнений (74) значения $\delta\hat{v}$, а по ним при помощи уравнений (71в), (71б) найдем

$$\hat{v}_n^{(s+1)} = \hat{v}_n^{(s)} + \delta\hat{v}_n, \quad \hat{r}_n^{(s+1)} = r_n + \tau\hat{v}_n^{(s+1)}, \quad \hat{\rho}_n^{(s+1)} = \frac{m}{\hat{r}_{n+1}^{(s+1)} - \hat{r}_n^{(s+1)}}. \quad (76)$$

Это позволяет вычислить $\hat{g}_n^{(s+1)} = p\left(\hat{\rho}_n^{(s+1)}\right)$ и выполнить следующую итерацию.

Сходимость итерационного процесса (74), (76) исследована в [34]. Этот процесс является ньютоновским; поэтому он сходится, если начальное приближение (75) недалеко отстоит от корня, т. е. если шаг τ не слишком велик. Это приводит к некоторому ограничению на τ; однако, как показано в [34], такое ограничение несравненно слабее, чем условие Куранта. Имеются примеры успешных численных расчетов задач с тонкими слоями, в которых шаг τ в 10^5 раз превышал значение, допускаемое локальным критерием Куранта (69).

Включение вязкости (71г) можно провести двумя способами. В первом способе линеаризация выполняется так, как описано выше, а к давлению добавляется вязкий член, взятый с предыдущей итерации:

$$\hat{g}_n^{(s)} = p(\hat{\rho}_n^{(s)}) + \mu_0 \hat{\rho}_n^{(s)} (\hat{v}_{n+1}^{(s)} - \hat{v}_n^{(s)})^2. \qquad (77)$$

Это означает, что вязкость включена в итерационный процесс методом последовательных приближений. Такой способ прост, но ухудшает сходимость итераций: уменьшает скорость сходимости и усиливает ограничение на шаг τ, хотя не слишком сильно.

Второй способ — полная линеаризация — сложнее, но надежнее. Линеаризируя уравнение (71а), учтем зависимость $\hat{g}$ не только от $\hat{\rho}$, но и непосредственно от $\hat{v}$ через вязкость (71г):

$$\delta\hat{g}_n = \frac{\partial(\hat{p}_n + \hat{\omega}_n)}{\partial\hat{\rho}_n}\delta\hat{\rho}_n + \frac{\partial\hat{\omega}_n}{\partial\hat{v}_{n+1}}\delta\hat{v}_{n+1} + \frac{\partial\hat{\omega}_n}{\partial\hat{v}_n}\delta\hat{v}_n. \qquad (78)$$

При этом вместо (72) и (74) получаются более громоздкие выражения, которые мы не приводим. Однако такой процесс является чисто ньютоновским и хорошо сходится.

Неизотермический случай требует включения в итерационный процесс уравнения энергии (71д), что часто делают способом *двухкруговых итераций* (последовательных прогонок).

Сначала считаем энергию $\hat{\varepsilon}_n$ (или температуру) известной во всех точках нового слоя. Тогда в каждой точке $p(\hat{\varepsilon}_n, \hat{\rho}_n) = p_n(\hat{\rho}_n)$, т. е. применимы формулы изотермического случая (74), (76); по ним проводят *первый малый круг* итераций.

Когда эти итерации сойдутся, полученные значения $\hat{r}$, $\hat{v}$, $\hat{\rho}$ подставляют в уравнение энергии (71д). Неизвестными в нем остаются значения $\hat{\varepsilon}$; их можно определить, линеаризируя уравнение (71д) с учетом зависимости $p(\varepsilon)$:

$$\hat{v}_{n+1}\theta_{n+1}\delta\hat{\varepsilon}_{n+1} + (1 + \hat{v}_{n+1}\theta_n - \hat{v}_n\theta_n)\delta\hat{\varepsilon}_n - \hat{v}_n\theta_{n-1}\delta\hat{\varepsilon}_{n-1} =$$
$$= \varepsilon_n - \hat{\varepsilon}_n^{(s)} + \frac{1}{4}(v_{n+1}^2 - v_n^2) - \frac{1}{4}(\hat{v}_{n+1}^2 - \hat{v}_n^2) +$$
$$+ \frac{\tau}{2m}\left[\hat{v}_n(\hat{g}_{n-1}^{(s)} + \hat{g}_n^{(s)}) - \hat{v}_{n+1}(\hat{g}_n^{(s)} + \hat{g}_{n+1}^{(s)})\right], \quad \theta_n = \frac{\tau}{2m}\left(\frac{\partial\hat{p}}{\partial\hat{\varepsilon}}\right)_n^{(s)}. \qquad (79)$$

Итерации (79) образуют *второй малый круг*. На каждой итерации трехточечное уравнение (79) решается прогонкой.

Найденные значения $\hat{\varepsilon}_n$ передают в уравнения (74), (76) и снова проводят первый малый круг итераций и т. д. Это взаимное согласование уравнений импульса и энергии составляет *большой круг итераций*.

Обычно считают нормальным, если малые круги сходятся за 3—5 итераций, а большой круг — за 2—3. Большее число итераций указывает на целесообразность уменьшения шага τ.

Замечание. Можно провести итерации в один круг, если полностью линеаризировать систему (71), считая $\hat{p} = p(\hat{\varepsilon}, \hat{\rho})$. Однако при этом получаются существенно более громоздкие уравнения в вариациях, для решения которых надо применять матричную прогонку (см. дополнение к [30]).

Устойчивость. Методом разделения переменных в линейном приближении можно показать, что схема (71) безусловно устойчива. Таким образом, шаг τ ограничивает только условие сходимости итераций при решении нелинейной системы (71).

Аппроксимация и сходимость. Схема (71) не симметрична по t и поэтому даже на гладких течениях имеет аппроксимацию $O(\tau + h^2)$. Тем самым, на гладких течениях схема «крест» может оказаться более точной.

Однако при расчете течений с ударными волнами и другими особенностями неявная схема дает существенно лучшие результаты, чем схема «крест». Поэтому она широко применяется в практике вычислений, особенно в «больших задачах».

Сходимость схемы (71) строго не доказана, но многократно проверена на сложных задачах-тестах с известными точными решениями.

5. О других схемах. Схемы (66) и (71) являются однородными. Имеется много близких к ним алгоритмов, отличающихся деталями написания отдельных членов разностных схем или другой организацией итерационных процессов решения нелинейных разностных уравнений. Из них следует отметить *полностью консервативные* схемы, в которых автоматически выполняются разностные законы сохранения не только массы, импульса и полной энергии, но также законы сохранения энтропии и внутренней энергии. В настоящее время построены полностью консервативные схемы для задач одномерной газодинамики в лагранжевых и эйлеровых переменных, задач магнитной газодинамики и двумерных газодинамических течений (подробнее см. в [34]).

Есть иначе построенные однородные схемы. Из них отметим схему *распада разрыва*. Она составлена так, что в акустическом приближении *) переходит в явную схему бегущего счета для инвариантов (33), обладающую хорошей аппроксимационной вязкостью. Благодаря этому схема позволяет рассчитывать любые разрывы без введения псевдовязкости.

В акустическом приближении схема распада разрыва монотонна; в газодинамике на сильных ударных волнах возможна немонотонность, хотя фактически она невелика. Схема имеет аппроксимацию $O(\tau + h)$, поэтому для расчета гладких течений она невыгодна. Но фронты ударных волн она воспроизводит хорошо, с малым сглаживанием.

Схема распада разрыва — явная и имеет ограничение на шаг типа $D\tau \leqslant \Delta r$, где D — скорость ударной волны. Это ограничение, а также громоздкость схемы препятствуют широкому ее применению.

---

*) Если p, ρ, ε лишь слабо колеблются около равновесных значений, то уравнения газодинамики переходят в уравнения акустики (см., например, [40]).

Помимо однородных схем существуют схемы с явным выделением особенностей, в которых точно прослеживается движение всех сильных и слабых разрывов. Одна такая схема предложена и подробно описана в [87]. Но такие схемы очень сложны, и их применяют только в тех случаях, когда требуется особенно высокая точность расчета.

ЗАДАЧИ

1. Составить схему «крест» для задачи (1) при неравномерных сетках по x и t и исследовать аппроксимацию схемы.

2. Найти невязку схемы (12).

3. Для волнового уравнения (1) составить схему с весом σ на шаблоне рис. 101 и провести исследование этой схемы; показать, что при $\sigma < 1/2$ схема безусловно неустойчива, при $\sigma \geqslant 1/2$ — безусловно устойчива и при $\sigma > 1/2$ обладает аппроксимационной вязкостью.

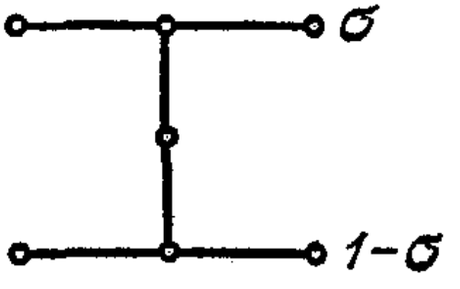


Рис. 101.

4. Установить аппроксимацию схемы (19).

5. Проверить исследование устойчивости схемы (19), данное в § 1, п. 3.

6. Доказать, что схема (26) имеет аппроксимацию $O(\tau^2+h^2)$.

7. Составить схему типа «крест» для задачи (18), приписывая значения y_n узлам сетки, а z_n — центрам ячеек $x_{n+1/2}$, $t_{m+1/2}$; написать для нее начальные данные точности $O(\tau^2+h^2)$

8. Провести полное исследование схемы (33).

9. Рассмотреть, как в схеме (33) вычисляется разностное решение в граничных узлах.

10. Вывести формулы циклической прогонки для случая матрицы, изображенной на рис. 95.

11. Исследовать устойчивость многомерной схемы с весами (40).

12. Провести полную линеаризацию системы (71а) — (71г) для случая изотермической газодинамики с учетом вязкости и свести задачу к решению трехточечного уравнения относительно $\delta\hat{v}$.

ГЛАВА XIV

ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ

В главе XIV рассмотрены простейшие методы решения интегральных уравнений. Корректно поставленным задачам посвящен § 1. В нем изложены некоторые типичные постановки задач и даны методы их решения: разностный метод и некоторые приближенные методы.

В § 2 рассмотрены некорректно поставленные задачи для линейных интегральных уравнений первого рода. Изложена теория построения регуляризирующих алгоритмов по А. Н. Тихонову. Для некоторых некорректных задач, возникших в предыдущих главах, даны алгоритмы решения, доведенные до схем численного расчета.

§ 1. Корректно поставленные задачи

1. Постановки задач. Интегральным называют уравнение, в котором неизвестная функция $u(x)$ стоит под знаком интеграла. Одномерное нелинейное интегральное уравнение имеет вид

$$\int_a^b K(x,\ \xi,\ u(\xi))\,d\xi = F(x,\ u(x)), \qquad a \leqslant x \leqslant b, \tag{1}$$

где ядро $K(x, \xi, u)$ и правая часть $F(x, u)$ — заданные функции.

К интегральным уравнениям приводят многие физические задачи. Так, задача восстановления переданного радиосигнала $u(t)$ по принятому сигналу $f(t)$ сводится к решению интегрального уравнения типа свертки:

$$\int_0^t K(t-\tau)\,u(\tau)\,d\tau = f(t), \tag{2}$$

где ядро $K(\xi)$ зависит от свойств приемной аппаратуры и среды, через которую проходит сигнал.

Заметим, что даже для задач, записанных в терминах уравнений в частных производных, первичной обычно является формулировка в виде интегральных законов сохранения, т. е. интегральных уравнений. В предыдущих главах такие формулировки использовались, например, для построения консервативных разностных схем.

Интегральные уравнения в некоторых отношениях удобнее дифференциальных. Во-первых, интегральное уравнение содержит в себе полную постановку задачи. Например, интегральное уравнение

$$u(x) = u_0 + \int_{x_0}^{x} f(\xi,\ u(\xi))\, d\xi \tag{3}$$

эквивалентно задаче Коши для дифференциального уравнения

$$\frac{du(x)}{dx} = f(x,\ u), \quad u(x_0) = u_0. \tag{4}$$

Тем самым, для уравнения (3) не требуется задавать никаких дополнительных условий, начальных или граничных (см. также задачу 1).

Во-вторых, в интегральных уравнениях переход от одной переменной ко многим является естественным. Так, многомерным аналогом (1) является уравнение

$$\int_G K(\boldsymbol{x},\ \xi,\ u(\xi))\, d\xi = F(\boldsymbol{x},\ u(\boldsymbol{x})),$$
$$\boldsymbol{x} = \{x_1,\ x_2,\ \ldots,\ x_p\} \in G(\boldsymbol{x}), \tag{5}$$

отличающееся от (1) только тем, что интегрирование проводится по многомерной области G. Поскольку оба уравнения не требуют дополнительных условий и полностью определяют задачу, аналогия является полной. Тем самым, теоретическое обоснование постановок и методов решения одномерных задач непосредственно обобщается на случай многих измерений.

Наоборот, в дифференциальных уравнениях переход от одной переменной к нескольким, т. е. от обыкновенных дифференциальных уравнений к уравнениям в частных производных, является принципиальным усложнением, приводит к новым постановкам задач и требует новых методов для их обоснования.

Далее мы ограничимся рассмотрением одномерного уравнения (1) и некоторых его частных случаев.

Линейные задачи. Лучше всего изучены уравнения, в которые неизвестная функция $u(x)$ входит линейно (см. [23]). Их можно записать в виде

$$u(x) - \lambda \int_a^b K(x,\ \xi)\, u(\xi)\, d\xi = f(x), \qquad a \leqslant x \leqslant b. \tag{6}$$

Это уравнение называют *уравнением Фредгольма второго рода*: ядро $K(x,\ \xi)$ этого уравнения определено на квадрате $a \leqslant x \leqslant b$, $a \leqslant \xi \leqslant b$.

Если ядро $K(x,\ \xi)$ отлично от нуля только на треугольнике $a \leqslant \xi \leqslant x \leqslant b$ (т. е. $K(x,\ \xi) = 0$ при $x < \xi$), то уравнение (6)

переходит в *уравнение Вольтерра второго рода*:

$$u(x)-\lambda\int_a^x K(x,\ \xi)\,u(\xi)\,d\xi=f(x),\qquad a\leqslant x\leqslant b. \tag{7}$$

Это уравнение теоретически исследовать или численно решить много проще, чем уравнение Фредгольма.

Если в уравнениях (6) и (7) отбросить член $u(x)$, оставив только $u(\xi)$ под знаком интеграла, то получим уравнения Фредгольма и Вольтерра *первого рода*. Задачи для уравнений первого рода являются некорректно поставленными и будут рассмотрены в § 2. Для уравнений второго рода задачи корректно поставлены; остановимся на этих задачах.

Для однородного уравнения Фредгольма второго рода (6) ставится задача на собственные значения:

$$u(x)=\lambda\int_a^b K(x,\ \xi)\,u(\xi)\,d\xi,\qquad a\leqslant x\leqslant b. \tag{8}$$

Требуется найти такие значения параметра $\lambda=\lambda_i$, при которых уравнение (8) имеет нетривиальные решения $u=\varphi_i(x)$; λ_i называют собственными значениями ядра $K(x,\ \xi)$, а $\varphi_i(x)$ — собственными функциями.

Если ядро вещественное и симметричное, $K(x,\ \xi)=K(\xi,\ x)=$ $=K^*(x,\ \xi)$, то оно имеет по меньшей мере одно собственное значение. Все собственные значения такого ядра вещественны, а его собственные функции ортогональны друг другу. Заметим, однако, что система собственных функций $\varphi_i(x)$ может быть неполной и даже конечной.

Неоднородное уравнение Фредгольма (6) при значении параметра λ, не равном ни одному из собственных значений λ_i ядра, имеет решение $u(x)$, притом единственное.

Если ядро $K(x,\ \xi)$ и правая часть $f(x)$ непрерывны вместе со своими p-ми производными, то решение также p раз непрерывно дифференцируемо. В этом легко убедиться, продифференцировав (6) p раз:

$$u^{(p)}(x)=f^{(p)}(x)+\lambda\int_a^b\frac{\partial^p K(x,\ \xi)}{\partial x^p}\,u(\xi)\,d\xi.$$

При сделанных предположениях правая часть этого равенства непрерывно зависит от x, что доказывает наше утверждение.

Для симметричного ядра решение неоднородного уравнения (6) представляется в виде разложения Шмидта:

$$u(x)=f(x)+\sum_{i\geqslant 1}\frac{\lambda}{\lambda_i-\lambda}\,\varphi_i(x)\int_a^b f(\xi)\,\varphi_i(\xi)\,d\xi; \tag{9}$$

если ядро $K(x, \xi)$ и правая часть $f(x)$ интегрируемы с квадратом, то этот ряд сходится абсолютно и равномерно. В данном случае из формулы (9) непосредственно видно, что при $\lambda \neq \lambda_i$ решение $u(x)$ существует, единственно и непрерывно зависит от $f(x)$, что означает корректность задачи (6).

Пусть параметр λ равен одному из собственных значений λ_i ядра $K(x, \xi)$. Тогда неоднородное уравнение Фредгольма (6) при произвольной правой части $f(x)$, вообще говоря, не имеет решения. Однако при некоторых правых частях $f(x)$ оно может иметь решение, притом не единственное (соответствующие примеры будут рассмотрены в п. 4). Таким образом, при $\lambda = \lambda_i$ в классе непрерывных или даже достаточно гладких правых частей $f(x)$ задача (6) является некорректно поставленной.

Уравнение Вольтерра не имеет собственных значений: если в уравнении (7) положить $f(x) = 0$, то оно будет иметь только тривиальное решение $u(x) = 0$. Поэтому неоднородное уравнение (7) всегда имеет решение, притом единственное.

2. Разностный метод. Это простейший численный метод, позволяющий получать решение одномерных задач с хорошей точностью, а двумерных — с удовлетворительной. Он рассчитан на применение ЭВМ, хотя оценки с небольшим числом узлов сетки можно производить вручную.

Рассмотрим одномерное нелинейное уравнение (1). Возьмем на $[a, b]$ какую-нибудь квадратную формулу, например линейную формулу с узлами x_n и весами c_n:

$$\int_a^b \Phi(\xi)\, d\xi \approx \sum_{n=1}^{N} c_n \Phi(x_n) \tag{10}$$

(нелинейные квадратурные формулы почти никогда не используются). Введем в квадрате $[a \leqslant x \leqslant b,\ a \leqslant \xi \leqslant b]$ сетку x_n, ξ_m, где x_n и ξ_m являются узлами формулы (10). Заменяя интеграл в уравнении (1) суммой (10), получим систему алгебраических уравнений для определения приближенных значений в узлах $y_n \approx u(x_n)$:

$$\sum_{m=1}^{N} c_m K(x_n, x_m, y_m) = F(x_n, y_n), \qquad 1 \leqslant n \leqslant N. \tag{11}$$

Эту систему целесообразно решать методом Ньютона. На вопрос о сходимости y_n к $u(x_n)$ при заданном типе квадратурной формулы и $N \to \infty$ в настолько общей постановке трудно ответить.

Рассмотрим линейные задачи. Для них обоснование сходимости (при использовании линейных квадратурных формул) фактически содержится в теории Фредгольма. Это обоснование громоздко и здесь не приводится (см., например, [23]).

Однородное уравнение Фредгольма (8) линейно, поэтому для него система (11) также линейна. Запишем ее в следующем виде:

$$\sum_{m=1}^{N} c_m K_{nm} y_m = \frac{1}{\lambda} y_n, \quad 1 \leqslant n \leqslant N, \quad K_{nm} = K(x_n, x_m). \tag{12}$$

Система (12) представляет собой задачу на определение собственных значений матрицы K' порядка N с элементами $K'_{nm} = K_{nm} c_m$. Эта матрица имеет N собственных значений $\lambda_i^{(N)}$, $1 \leqslant i \leqslant N$, которые являются приближением к первым собственным значениям λ_i ядра $K(x, \xi)$.

Разностное решение (12) вычисляют методами, описанными в главе VI. Матрица K' является, вообще говоря, плотно заполненной и неэрмитовой; поэтому фактически вычислить разностное решение удается только при небольших $N \lesssim 50$. Получить в этом случае хорошую точность можно лишь для нескольких первых собственных значений, причем ядро и правая часть должны быть достаточно гладкими и не быстропеременными.

Замечание 1. Пусть ядро, правая часть и искомое решение достаточно гладки и квадратурная формула (10) имеет на них аппроксимацию $O(h^p)$. Поскольку алгоритм сходится, то он устойчив. Задача (8) — линейная, поэтому из аппроксимации и устойчивости следует сходимость со скоростью $O(h^p)$.

Сходимость можно исследовать численно, проводя расчеты на последовательности сгущающихся сеток и устанавливая стремление y_n и некоторой предельной функции при $h \to 0$.

Неоднородное уравнение Фредгольма (6) приводит к линейной неоднородной алгебраической системе

$$y_n - \lambda \sum_{m=1}^{N} c_m K_{nm} y_m = f_n, \quad 1 \leqslant n \leqslant N, \quad f_n = f(x_n). \tag{13}$$

Разностное решение легко вычисляется методом исключения Гаусса; на ЭВМ типа БЭСМ-6 скорость и оперативная память позволяют использовать в расчете до $N \approx 150$ узлов. Таким образом, в этой задаче нетрудно получить более высокую точность расчета, чем в задаче на собственные значения.

Линейная система (13) имеет единственное решение, если $\lambda \neq \lambda_i^{(N)}$. Но $\lambda_i^{(N)} \approx \lambda_i$, причем при большом N разница между ними невелика. Следовательно, описанный алгоритм хорошо обусловлен, если параметр λ не лежит в малой окрестности одного из собственных значений λ_i ядра.

Если $\lambda \approx \lambda_i$, то система (13) становится плохо обусловленной. При некоторых числах узлов N возможен сбой алгоритма: если слу-

чайно значение $\lambda_i^{(N)}$ близко подходит к λ, то разностное решение y_n на этой сетке может сильно отличаться от $u(x)$.

Обычно нам неизвестны собственные значения ядра. Поэтому для обнаружения и исключения последнего случая все расчеты надо проводить на последовательности сгущающихся сеток. Если при сгущении сетки y_n сходится к некоторой предельной функции $u(x)$, то эта функция есть искомое решение (см. замечание 1). Если расчет на одной из сеток выпадает из общей закономерности, то имело место случайное совпадение $\lambda \approx \lambda_i^{(N)}$. Если на всех сетках y_n не стремится к пределу при $h \to 0$, то $\lambda \approx \lambda_i$.

Уравнение Вольтерра (7) получают из уравнения Фредгольма (6), полагая $K(x, \xi) = 0$ при $x < \xi$. Алгебраическая система (13) становится при этом треугольной:

$$y_n - \lambda \sum_{m=1}^{n} c_m K_{nm} y_m = f_n, \qquad 1 \leqslant n \leqslant N, \tag{14}$$

и решается обратным ходом метода Гаусса всего за $^3/_2 N^2$ действий. Поэтому здесь объем вычислений остается умеренным даже при $N \approx 1000$, что позволяет проводить расчеты с очень высокой точностью.

Выбор квадратурной формулы. Большинство задач приходится решать, используя сравнительно небольшое число узлов N. Поэтому для получения хорошей точности целесообразно выбирать квадратурные формулы высокого порядка точности, разумеется, если $K(x, \xi)$ и $f(x)$ имеют достаточное число непрерывных производных.

Обычно наилучшие результаты для достаточно гладких решений дают квадратурные формулы Гаусса или Гаусса — Кристоффеля; при числе узлов k их порядок точности $p = 2k$. Можно также использовать простейшую формулу трапеций, последовательно сгущая сетки вдвое от $N_1 = 2$ до $N_k = 2^k$ и уточняя решение способом Рунге; это также дает результат с порядком точности $p = 2k$ *), но требует использования существенно большего числа узлов, чем в формулах Гаусса.

Нередко ядро $K(x, \xi)$ или правая часть $f(x)$ недостаточно гладки и даже имеют разрывы. Наиболее типичен разрыв ядра или его производных при $x = \xi$ (на диагонали квадрата $a \leqslant x \leqslant b$, $a \leqslant \xi \leqslant b$); встречаются особенности и на других линиях в плоскости (x, ξ). В этих случаях использовать формулы Гаусса нецелесообразно. Удобнее построить специальную сетку x_n так, чтобы особые линии пересекали линии сетки $x = x_n$ только в узлах

*) Каждая лишняя сетка позволяет повысить порядок точности на 2, поскольку погрешность формулы трапеций разлагается по четным степеням шага h.

$\xi = x_m$ (рис. 102). Затем в качестве (10) выбирают обобщенную формулу трапеций (4.7), причем в интервалах, примыкающих к особой линии, используют соответствующие односторонние пределы функций.

Если вне особых линий все функции непрерывны вместе с достаточным числом своих производных, то при сгущении специальной сетки можно уточнять решение способом Рунге.

Полезно предварительно так преобразовать исходное уравнение, чтобы гладкость решения повысилась. Например, если ядро непрерывно, а $f(x)$ разрывна, то $u(x)$ тоже разрывна. Полагая $z(x) = u(x) - f(x)$, получим вместо (6) уравнение

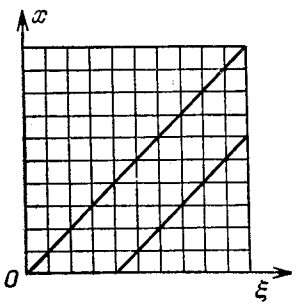


Рис. 102.

$$z(x) - \lambda \int_a^b K(x, \xi)\, z(\xi)\, d\xi = \varphi(x),$$
$$\varphi(x) = \lambda \int_a^b K(x, \xi)\, f(\xi)\, d\xi. \qquad (15)$$

В уравнении (15) правая часть уже непрерывно зависит от x, так что его решение $z(x)$ непрерывно. Поэтому численно решать уравнение (15) проще, чем исходное уравнение (6).

Замечание 2. Уравнение Вольтерра (7) формально сводится к уравнению Фредгольма (6), но ядро при этом имеет особенность (обычно разрыв) на диагонали $x = \xi$. Поэтому для уравнения Вольтерра следует выбирать обобщенную формулу трапеций и проводить уточнение способом Рунге.

Многомерные задачи допускают, в принципе, применение описанного метода; надо только в (10) и других формулах под x_n подразумевать узлы многомерной кубатурной формулы $\boldsymbol{x}_n$. Однако получить удовлетворительную точность при умеренном объеме расчетов удается лишь для достаточно гладких $K(\boldsymbol{x}, \xi)$ и $f(\boldsymbol{x})$, когда можно использовать кубатурные формулы высокого порядка точности (например, произведение одномерных формул Гаусса с небольшим числом узлов k по каждой переменной).

В более сложных случаях развивают специальные методы; многие из них используют симметрию задачи и слабую зависимость решения от части переменных.

3. Метод последовательных приближений. Это простейший приближенный метод. Запишем для неоднородного уравнения Фредгольма (6) итерационный процесс:

$$u_0(x) = 0, \qquad u_{n+1}(x) = f(x) + \lambda \int_a^b K(x, \xi)\, u_n(\xi)\, d\xi. \qquad (16)$$

Нетрудно показать, что при ограниченном ядре и достаточно малом значении $|\lambda|$ этот процесс сходится к решению уравнения (6).

Доказательство. Обозначим погрешность n-й итерации через $z_n(x) = u_n(x) - u(x)$. Вычитая (6) из (16), получим

$$z_{n+1}(x) = \lambda \int_a^b K(x, \xi)\, z_n(\xi)\, d\xi. \tag{17}$$

Отсюда следует неравенство

$$\|z_{n+1}\|_C \leqslant |\lambda|\,(b-a)\,\|K(x, \xi)\|_C\,\|z_n(x)\|_C. \tag{18}$$

Тем самым, если выполнено условие

$$q = |\lambda|\,(b-a)\,\|K(x, \xi)\|_C < 1, \tag{19}$$

то итерации (16) сходятся равномерно по x, причем сходимость линейная. При достаточно малом $|\lambda|$ условие (19) выполняется.

В практических вычислениях квадратуры, возникающие в этом методе, редко удается выразить через элементарные функции. Поэтому обычно ограничиваются нахождением первых приближений.

Замечание 1. Для уравнения Вольтерра (7) метод последовательных приближений сходится равномерно по x при любых значениях λ. Действительно, в этом случае вместо (17) справедливо соотношение

$$z_{n+1}(x) = \lambda \int_a^x K(x, \xi)\, z_n(\xi)\, d\xi. \tag{20}$$

Выкладки, полностью аналогичные доказательству сходимости метода Пикара (гл. VIII, § 1, п. 3), приводят к оценке

$$\|z_n(x)\|_C \leqslant \frac{1}{n!}\{|\lambda|\,(b-a)\|K(x, \xi)\|_C\}^n \|z_0(x)\|_C. \tag{21}$$

При $n \to \infty$ правая часть этого неравенства стремится к нулю при любых значениях λ, что доказывает наше утверждение.

Замечание 2. Оценку (18) можно переписать в следующем виде:

$$\|z_n(x)\|_C \leqslant q^n \|z_0(x)\|_C, \qquad q^n \sim \lambda^n. \tag{22}$$

Отсюда видно, что метод последовательных приближений для уравнения Фредгольма эквивалентен разложению в ряд по степеням параметра λ. Это можно строго показать, выражая $u_n(x)$ через $u_1(x) = f(x)$ при помощи рекуррентного соотношения (16).

Пример. Рассмотрим уравнение

$$u(x) - \lambda \int_0^\infty e^{-(x+\xi)} u(\xi)\, d\xi = x. \tag{23}$$

Применяя процесс (16), получим

$$u_0(x) = 0, \quad u_1(x) = x, \quad u_2(x) = x + \lambda e^{-x},$$

$$u_3(x) = x + \lambda e^{-x} + \frac{1}{2}\lambda^2 e^{-x}, \quad u_4(x) = x + \left(\lambda + \frac{1}{2}\lambda^2 + \frac{1}{4}\lambda^3\right) e^{-x}$$

и т. д. В этом случае удается найти точное решение

$$u(x) = x + \left(\lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{4} + \ldots\right) e^{-x} = x + \frac{2\lambda}{2-\lambda} e^{-x}. \tag{24}$$

Нетрудно заметить, что последовательные приближения здесь сходятся только при $|\lambda| < 2$.

4. Замена ядра вырожденным. Ядро уравнения Фредгольма называется *вырожденным*, если оно является суммой конечного числа членов вида

$$\bar{K}(x, \xi) = \sum_{n=1}^{N} A_n(x)\, B_n(\xi) \tag{25}$$

(для уравнения Вольтерра ядро не может быть вырожденным, иначе оно тождественно равнялось бы нулю). Решение уравнения с вырожденным ядром находится за конечное число действий.

В самом деле, подставляя ядро (25) в неоднородное уравнение (6), представим решение в виде суммы конечного числа членов:

$$u(x) = f(x) + \lambda \sum_{n=1}^{N} \alpha_n A_n(x), \tag{26а}$$

$$\alpha_n = \int_a^b B_n(\xi)\, u(\xi)\, d\xi. \tag{26б}$$

Подставляя (26а) в (26б), получим линейную систему для нахождения коэффициентов α_n:

$$\sum_{m=1}^{N} \left[\delta_{nm} - \lambda \int_a^b B_n(\xi)\, A_m(\xi)\, d\xi\right] \alpha_m = \int_a^b B_n(\xi) f(\xi)\, d\xi, \quad 1 \leqslant n \leqslant N. \tag{27}$$

Решая эту систему и подставляя найденные значения α_n в (26а), найдем искомое решение.

Для однородного уравнения Фредгольма (8) надо положить во всех формулах $f(x) = 0$. Тогда система (27) становится однородной и представляет собой задачу на нахождение собственных

значений матрицы N-го порядка. Отсюда видно, что вырожденное ядро (25) имеет ровно N собственных значений λ_i.

Произвольное ядро нередко удается хорошо аппроксимировать вырожденным ядром. Например, разложим $K(x, \xi)$ в ряд Фурье по некоторой полной ортонормированной системе функций $B_n(\xi)$; коэффициенты этого разложения будут функциями от x:

$$K(x, \xi) = \sum_{n=1}^{\infty} A_n(x) B_n(\xi), \quad A_n(x) = \int_a^b B_n^*(\xi) K(x, \xi) d\xi. \tag{28}$$

В качестве (25) можно взять отрезок разложения (28). Тогда формулы (25)—(27) позволяют найти приближенное решение. Оценки точности таких приближений мы не рассматриваем, поскольку они громоздки и неудобны в практических вычислениях.

Замечание. Пусть в неоднородном уравнении (6) с вырожденным ядром (25) правая часть $f(x) \not\equiv 0$ и такова, что выполняется

$$\int_a^b f(\xi) B_n(\xi) d\xi = 0, \quad 1 \leqslant n \leqslant N. \tag{29}$$

Тогда при λ, равном одному из собственных значений ядра λ_i, система (27) имеет нетривиальное решение, причем не единственное. Тем самым, в данном случае существует решение $u(x)$ уравнения (6).

Пример. Рассмотренное в п. 3 уравнение (23) имеет вырожденное ядро $K(x, \xi) = e^{-x} e^{-\xi}$. У него должно быть ровно одно собственное значение; определим его. Полагая в (27) $N = 1$ и $f(x) = 0$, легко получим

$$\left(1 - \lambda_1 \int_0^{\infty} e^{-2\xi} d\xi\right) \alpha_1 = 0,$$

откуда $\lambda_1 = 2$. Заметим, что точное решение (24) неоднородного уравнения (23) при $\lambda = \lambda_1$ не существует.

5. Метод Галеркина (который для интегральных уравнений обычно называют *методом моментов*). Будем искать решение в виде разложения по полной системе функций $\psi_k(x)$:

$$u(x) \approx f(x) + \lambda \sum_{k=1}^{N} \alpha_k \psi_k(x); \tag{30}$$

поскольку от $u(x)$ не надо специально требовать удовлетворения каким-либо краевым условиям, то от системы $\psi_k(x)$ ничего, кроме полноты, требовать не надо.

Подставляя разложение (30) в неоднородное уравнение Фредгольма (6) и требуя ортогональности невязки ко всем функциям

$\psi_k(x)$, $1 \leqslant k \leqslant N$, получим линейную алгебраическую систему уравнений для нахождения α_k:

$$\sum_{k=1}^{N} a_{mk}\alpha_k = b_m, \qquad 1 \leqslant m \leqslant N,$$

$$a_{mk} = \int_a^b \psi_m(x)\,\psi_k(x)\,dx - \lambda \int_a^b \int_a^b K(x, \xi)\,\psi_m(x)\,\psi_k(\xi)\,dx\,d\xi, \qquad (31)$$

$$b_m = \int_a^b \int_a^b K(x, \xi)\,\psi_m(x)\,f(\xi)\,dx\,d\xi.$$

В случае задачи на собственные значения (8) надо полагать в (30) и (31) $f(x) = 0$. Метод применим и к нелинейному уравнению (1), но тогда он приводит к нелинейной алгебраической системе.

Основной трудностью, препятствующей применению метода моментов, является сложность вычисления двукратных интегралов, входящих в (31). Поэтому обычно ограничиваются малым числом членов суммы (30).

Замечание. Если система $\psi_k(x)$ ортогональна, то метод моментов эквивалентен приближенной замене ядра на специальное вырожденное ядро:

$$\bar{K}(x, \xi) = \sum_{k=1}^{N} \psi_k(x)\,\Psi_k(\xi),$$
$$\Psi_k(\xi) = \int_a^b K(x, \xi)\,\psi_k(x)\,dx. \qquad (32)$$

§ 2. Некорректные задачи

1. Регуляризация. Если в интегральном уравнении (1) правая часть $F(x, u(x))$ не зависит от решения, т. е. $u(x)$ входит только под знак интеграла, то задача оказывается некорректно поставленной. Классическими примерами некорректных задач являются уравнение Фредгольма первого рода:

$$\int_a^b K(x, \xi)\,u(\xi)\,d\xi = f(x), \qquad c \leqslant x \leqslant d, \qquad (33)$$

и уравнение Вольтерра первого рода:

$$\int_a^x K(x, \xi)\,u(\xi)\,d\xi = f(x), \qquad c \leqslant x \leqslant d. \qquad (34)$$

В отличие от уравнений второго рода, ядро уравнения Фредгольма (33) задано на прямоугольнике $[c \leqslant x \leqslant d,\ a \leqslant \xi \leqslant b]$, а в урав-

нении Вольтерра (34) — на трапеции $[c \leqslant x \leqslant d,\ a \leqslant \xi \leqslant x]$*), причем функции $u(\xi)$ и $f(x)$ определены на разных отрезках и принадлежат разным классам функций U и F.

Покажем, что задача (33) неустойчива по правой части и, тем самым, некорректна. Для этого рассмотрим высокочастотное возмущение с конечной амплитудой $\delta u(\xi) = \exp(i\omega\xi)$, $\omega \gg 1$. Ему соответствует возмущение правой части

$$\delta f(x) = \int_a^b K(x,\ \xi)\,\delta u(\xi)\,d\xi = \int_a^b K(x,\ \xi)\,e^{i\omega\xi}\,d\xi.$$

Интегрируя по частям, получим

$$\delta f(x) = \frac{1}{i\omega}\,e^{i\omega\xi}\,K(x,\ \xi)\Big|_{\xi=a}^{\xi=b} - \frac{1}{i\omega}\int_a^b \frac{\partial K(x,\ \xi)}{\partial\xi}\,e^{i\omega\xi}\,d\xi = O\left(\frac{1}{\omega}\right). \quad (35)$$

Это означает, что для достаточно больших частот величина $\| \delta f \|_C = O(1/\omega)$ оказывается сколь угодно малой. Следовательно, существуют такие сколь угодно малые возмущения правой части $\delta f(x)$, которым соответствуют большие возмущения решения $\delta u(\xi)$, т. е. задача (33) неустойчива.

Для уравнения Вольтерра (34) справедливы те же рассуждения. Напомним, что в главе III мы уже сталкивались с некорректностью задачи численного дифференцирования функции $f(x)$; эта задача сводится к решению уравнения

$$\int_a^x u(\xi)\ d\xi = f(x), \quad (36)$$

т. е. является частным случаем уравнения Вольтерра первого рода, с ядром $K(x,\ \xi) = 1$ (при $\xi \leqslant x$).

Кроме того, задачи (33), (34) имеют решение не при любых непрерывных правых частях $f(x)$. Так, задача (36) имеет решение только для дифференцируемых $f(x)$. Другим примером служит уравнение (33) с вырожденным ядром; подставляя в это уравнение выражение для ядра (25), получим

$$\sum_{n=1}^{N} \beta_n A_n(x) = f(x), \qquad \beta_n = \int_a^b B_n(\xi)\ u(\xi)\ d\xi. \quad (37)$$

Это равенство выполнимо для таких $f(x)$, которые представимы в виде линейной комбинации функций $A_n(x)$; для других правых частей задача (33) не имеет решения.

В обоих этих примерах, даже если при некоторой $f(x) = \bar{f}(x)$ существует решение, имеются такие малые изменения правой части $\delta f(x)$, при которых решение не существует.

---

*) При $c < a$ эта трапеция превращается в два треугольника.

Очевидно, непосредственно решать некорректные задачи при неточно заданной правой части бессмысленно. Если $\tilde{f}(x)$ задана с погрешностью $\delta f(x)$, то соответствующее решение $u_\delta(\xi)$ или не существует, или отличается от искомого решения $\bar{u}(\xi)$ на величину $\delta u(\xi)$, которая может быть большой.

Даже если $f(x)$ задана точно, но отыскание решения выполняется численными методами, то неизбежно вносятся погрешности метода и округления. Это снова приводит к большой погрешности решения $\delta u(\xi)$.

Регуляризирующий алгоритм. Пусть требуется найти решение $u(\xi)$ некорректно поставленной задачи:

$$A[x,\ u(\xi)]=f(x),\quad u(\xi)\in U,\quad f(x)\in F. \tag{38}$$

Здесь A — некоторый оператор, не обязательно интегральный, а U и F — нормированные пространства. Предполагается, что для произвольной $f(x)\in F$ решение задачи (38) может не существовать; однако имеются некоторые $\tilde{f}(x)\in F$, для которых существуют решения $\bar{u}(\xi)\in U$.

Ранее, изучая разрывные решения квазилинейных уравнений, мы вводили в исследуемое уравнение дополнительные члены, изменяющие свойства решений в нужную нам сторону. Попробуем и здесь изменить уравнение (38), введя в него дополнительные члены с малым положительным *параметром регуляризации* α. Символически запишем измененную задачу:

$$A_\alpha[x,\ u_\alpha(\xi)]=\tilde{f}(x), \tag{39}$$

а ее решение обозначим через $u_\alpha(\xi)$.

Определение. *Оператор A_α называют регуляризирующим, если* а) *задача* (39) *является корректно поставленной в классе правых частей F при любом (не слишком большом) $\alpha>0$ и* б) *существуют такие функции $\alpha(\delta)$ и $\delta(\varepsilon)$, что если $\|f-\tilde{f}\|_F\leqslant\delta(\varepsilon)$, то $\|u_{\alpha(\delta)}-\bar{u}\|_U\leqslant\varepsilon$.*

Замечание. Функции $\alpha(\delta)$ и $\delta(\varepsilon)$ зависят также от $\tilde{f}(x)$.

Таким образом, если найден регуляризирующий оператор A_α, то задача (39) имеет решение при любых $\tilde{f}(x)\in F$, в том числе отличающихся от $\tilde{f}(x)$ на любого вида погрешность $\delta f(x)$; эта задача устойчива, так что ее можно решать обычными численными методами. При правильно подобранном параметре α ее решение $u_\alpha(\xi)$ достаточно мало отличается от нужного нам решения $\bar{u}(\xi)$ исходной задачи (38).

Для одной и той же задачи можно построить много различных регуляризирующих алгоритмов. Кроме того, при заданном пространстве F разные алгоритмы могут давать решения $u_\alpha(\xi)$, принадлежащим различным пространствам U. Различают регуля-

ризацию *слабую* (U есть гильбертово пространство), *сильную* (чебышевское пространство) и p-го порядка гладкости (пространство $C^{(p)}$ *)).

Можно формально превратить задачу (38) в корректно поставленную, если ограничиться рассмотрением правых частей $f(x)$, принадлежащих некоторому более узкому классу F_0. Например, для задачи численного дифференцирования (36) в качестве F_0 возьмем пространство $C^{(1)}$. Малость $\|\delta f\|_{C^{(1)}}$ означает, что $\max |\delta f'(x)|$ невелик; поэтому такой вариации правой части соответствует малая вариация $\|\delta u(\xi)\|_C$.

Однако такой подход не конструктивен. Зачастую $f(x)$ содержит заметную погрешность, например, она может быть экспериментально определяемой величиной. Поэтому постановки большинства прикладных задач таковы, что в качестве F приходится выбирать чебышевское или даже гильбертово пространство, причем решение $u(\xi)$ необходимо получить в чебышевском пространстве.

2. Вариационный метод регуляризации. Рассмотрим уравнение Фредгольма первого рода (33). Будем считать, что его ядро непрерывно и таково, что в случае $f(x) \equiv 0$ уравнение имеет только тривиальное решение $u(\xi) \equiv 0$. Тогда при любой правой части $f(x) \in F$ решение либо единственное, либо не существует; тем самым, интегральный оператор

$$A[x,\ u(\xi)] = \int_a^b K(x,\ \xi)\, u(\xi)\, d\xi \tag{40}$$

отображает U в F взаимно однозначно.

Исходную задачу (33) можно записать в вариационной форме:

$$\int_c^d \{A[x,\ u(\xi)] - f(x)\}^2\, dx = \min, \tag{41}$$

где оператор A определен формулой (40). Рассмотрим измененную задачу:

$$M[\alpha, f(x), u(\xi)] \equiv \int_c^d \{A[x, u(\xi)] - f(x)\}^2\, dx + \alpha\Omega_n[u(\xi)] = \min, \tag{42а}$$

где так называемый *тихоновский стабилизатор n-го порядка* Ω_n равен

$$\Omega_n[u(\xi)] = \int_a^b d\xi \sum_{k=0}^{n} p_k(\xi) \left(\frac{d^k u(\xi)}{d\xi^k}\right)^2, \tag{42б}$$

а весовые функции $p_k(\xi)$ непрерывны и неотрицательны, причем

*) Это пространство функций $u(\xi)$, $a \leqslant \xi \leqslant b$, непрерывных и ограниченных вместе со своими p-ми производными, причем $\|u\|_{C^{(p)}} = \max\{|u|, |u'|, \ldots, |u^{(p)}|\}$.

$p_n(\xi) > 0$ (если нет специальных оснований для их выбора, то обычно полагают $p_k(\xi) \equiv 1$).

Введем в множестве функций U норму $\| u \|_U = \Omega_n[u]$; полученное пространство называют пространством Соболева W_2^n. Множество правых частей F будем считать гильбертовым пространством. Докажем методами функционального анализа, что алгоритм (42) является регуляризирующим (другое доказательство см. в п. 3).

Теорема 1. *Задача* (42) *имеет решение* $u_\alpha(\xi)$ *при любых* $f(x) \in F$ *и* $\alpha > 0$.

Доказательство. При $\alpha > 0$ функционал $M[\alpha, f, u]$ ограничен снизу. Тем самым, при данных α и $f(x)$ он имеет точную нижнюю грань. Выберем некоторую минимизирующую последовательность $u_i(\xi)$, $i = 0, 1, 2, \ldots$, так, что

$$\lim_{i \to \infty} M_i = \bar{M}, \quad M_i = M[\alpha, f, u_i], \quad \bar{M} = \inf_{u \in U} M[\alpha, f, u].$$

Упорядочим эту последовательность так, чтобы M_i не возрастали. Тогда

$$\Omega_n[u_i] \leqslant \frac{1}{\alpha} M_i \leqslant \frac{1}{\alpha} M_0 = \text{const}. \tag{43}$$

Таким образом, последовательность $u_i(\xi)$ принадлежит множеству $u(\xi)$, для которых $\Omega_n[u] \leqslant \text{const}$. Такое множество является компактом в U. Поэтому из последовательности $u_i(\xi)$ можно выделить подпоследовательность $u_{i(k)}(\xi)$, сходящуюся по норме к некоторой $u_\alpha(\xi) \in U$. В силу непрерывности функционал $M[\alpha, f, u]$ на этой функции $u_\alpha(\xi)$ достигает своей точной нижней грани. Тем самым, $u_\alpha(\xi) \in U$ есть решение задачи (42), что доказывает теорему.

Теорема 2. *Алгоритм* (42) *является регуляризирующим для задачи* (41).

Доказательство. Используем следующие обозначения: $\bar{u}(\xi)$ — решение исходной задачи (41) с правой частью $\bar{f}(x)$; $\tilde{u}_\alpha(\xi)$ — решение измененной задачи (42) с приближенной правой частью $\tilde{f}(x)$; введем также функцию $f_\alpha(x) = A[x, u_\alpha(\xi)]$.

Поскольку функционал $M[\alpha, \tilde{f}, u]$ достигает минимума на $\tilde{u}_\alpha$, то $M[\alpha, \tilde{f}, \tilde{u}_\alpha] \leqslant M[\alpha, \tilde{f}, \bar{u}]$. Отсюда, используя определение функционала (42а), получим

$$\alpha\Omega_n[\tilde{u}_\alpha] \leqslant M[\alpha, \tilde{f}, \tilde{u}_\alpha] \leqslant M[\alpha, \tilde{f}, \bar{u}] =$$

$$= \int_c^d \{A[x, \bar{u}] - \tilde{f}(x)\}^2 dx + \alpha\Omega_n[\bar{u}] = \int_c^d \{\bar{f}(x) - \tilde{f}(x)\}^2 dx +$$

$$+ \alpha\Omega_n[\bar{u}] = \| \bar{f} - \tilde{f} \|_{L_2}^2 + \alpha\Omega_n[\bar{u}]. \tag{44}$$

Пусть приближенные правые части удовлетворяют условию

$$\| \tilde{f} - \bar{f} \|_{L_2} \leqslant C \sqrt{\alpha}, \quad C = \text{const}. \tag{45}$$

Тогда из (44) следует

$$\Omega_n [\tilde{u}_\alpha] \leqslant C^2 + \bar{\Omega}_n = \text{const}, \quad \bar{\Omega}_n = \Omega_n [\bar{u}]. \tag{46}$$

Значит, решения $\tilde{u}_\alpha$ принадлежат компактному множеству U_0 функций из U. Заметим, что $\bar{u}$ также принадлежит U_0.

Множество F_0 функций $f_\alpha(x)$ есть образ множества U_0 при отображении A. Интегральный оператор A непрерывен и таков (по предположению), что обратное отображение единственно. Поэтому обратное отображение F_0 в компактное множество U_0 при помощи нерегуляризованного оператора A^{-1} непрерывно в $\| \cdot \|_U$. Следовательно, по заданному $\varepsilon > 0$ всегда найдется такое $\beta(\varepsilon)$, что если $\| f_\alpha - \bar{f} \| \leqslant \beta(\varepsilon)$, то $\| \tilde{u}_\alpha - \bar{u} \| \leqslant \varepsilon$.

Заметим, что

$$\| f_\alpha - \tilde{f} \|^2 = \int_c^d (f_\alpha - \tilde{f})^2 \, dx = \int_c^d \{A[x, \tilde{u}_\alpha] - \tilde{f}\}^2 \, dx \leqslant$$

$$\leqslant M[\alpha, \tilde{f}, \tilde{u}_\alpha] \leqslant \alpha (C^2 + \bar{\Omega}_n).$$

Отсюда с учетом (45) следует

$$\| f_\alpha - \bar{f} \| \leqslant \| f_\alpha - \tilde{f} \| + \| \tilde{f} - \bar{f} \| \leqslant \sqrt{\alpha} \left(C + \sqrt{C^2 + \bar{\Omega}_n} \right). \tag{47}$$

Выберем α так, чтобы выполнялось

$$\alpha \leqslant \alpha_0(\varepsilon), \quad \alpha_0(\varepsilon) \equiv \left[\beta(\varepsilon) / \left(C + \sqrt{C^2 + \bar{\Omega}_n} \right) \right]^2. \tag{48}$$

Тогда правая часть неравенства (47) будет меньше $\beta(\varepsilon)$, откуда следует $\| \tilde{u}_\alpha - \bar{u} \| \leqslant \varepsilon$.

Таким образом, по заданному ε нашлись такое $\alpha_0(\varepsilon)$ и такое $\delta(\alpha) = C\sqrt{\alpha}$, что если $\alpha \leqslant \alpha_0(\varepsilon)$ и $\| \tilde{f} - \bar{f} \| \leqslant \delta(\alpha)$, то $\| \tilde{u}_\alpha - \bar{u} \| \leqslant \varepsilon$, что и требовалось доказать.

Следствие. *Задача* (42) *корректно поставлена.*

В самом деле, подставим в теорему 2 всюду вместо A регуляризирующий алгоритм (42). Тогда малость $\| \tilde{u}_\alpha - \bar{u} \|$ означает, что регуляризованное решение $\tilde{u}_\alpha$ непрерывно зависит от $\tilde{f}$.

Замечание 1. Теоремы 1 и 2 справедливы не только для линейных интегральных операторов (40), но вообще для непрерывного оператора A, при котором решение задачи $A[u] = f$ единственно (если существует). Соответственно от стабилизатора Ω достаточно требовать, чтобы множество функций u, для которых $\Omega[u] \leqslant \text{const}$, было компактно в U.

Замечание 2. Сходимость в пространстве W_2^n означает, что n-я производная сходится среднеквадратично, а сама функция

и ее производные вплоть до $(n-1)$-й — равномерно. Таким образом, использование стабилизатора (42б) обеспечивает слабую регуляризацию при $n=0$, сильную при $n=1$ и $(n-1)$-го порядка гладкости при $n>1$.

Выбор α. В ряде прикладных задач известно, что правая часть имеет характерную погрешность $\|\tilde{f}-\bar{f}\|\approx\delta$. Если при этом выбрать α настолько малым, что нарушится критерий (45), то устойчивость расчета станет недостаточной, так что регуляризованное решение $\tilde{u}_\alpha$ будет заметно «разболтанным». Если α настолько велико, что не соблюден критерий (48), то регуляризованное решение $\tilde{u}_\alpha$ чрезмерно сглажено, что также нежелательно; например, если точное решение $\bar{u}$ имеет узкие максимумы (типа резонансных пиков в физических задачах), то у $\tilde{u}_\alpha$ они могут отсутствовать или иметь существенно меньшую высоту.

Вдобавок непосредственно проверить выполнение критериев (45) и (48) не удается, поскольку функция $\beta(\varepsilon)$ неизвестна (и, вообще говоря, зависит от C и $\bar{f}$). Поэтому оптимальный выбор параметра регуляризации α является сложной проблемой.

Обычно на практике проводят расчеты с несколькими значениями параметра, составляющими геометрическую прогрессию (например, $\alpha=10^{-1}$, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}). Из полученных результатов выбирают наилучший либо визуальным контролем, либо по какому-нибудь правдоподобному критерию.

Примером такого критерия является требование, чтобы невязка, полученная при подстановке найденного $\tilde{u}_\alpha$ в исходное уравнение, была сравнима с погрешностью правой части:

$$r\approx\delta,\quad r=\left(\int_c^d\{A[x,\tilde{u}_\alpha(\xi)]-\tilde{f}(x)\}^2\,dx\right)^{1/2}. \tag{49}$$

Очевидно, воспроизводить правую часть с точностью много выше δ бессмысленно; поэтому, если в расчете получено $r\ll\delta$, то следует увеличить α. Наоборот, погрешность много больше δ недопустима, так что если $r\gg\delta$, то надо уменьшить α.

Визуальный контроль заключается в том, что выбирают наименьшее значение α, при котором еще не наблюдается заметной «разболтки» регуляризованного решения $\tilde{u}_\alpha$.

Выбор n. При чрезмерно большом n регуляризованное решение сильно сглаживается. Значение $n=0$ обеспечивает лишь среднеквадратичную сходимость $\tilde{u}_\alpha(\xi)$ к $\bar{u}(\xi)$. Поэтому наиболее часто используют $n=1$.

Помимо вариационного способа регуляризации существует ряд других: метод подбора, метод квазиобращения, методы с использованием преобразований Лапласа и Меллина и т. д. Они рассмотрены в [39] и цитированных там работах.

3. Уравнение Эйлера. Учитывая явный вид (40) оператора A, перепишем задачу (42) следующим образом:

$$\alpha \sum_{k=0}^{n} \int_a^b p_k(\xi)\,[u^{(k)}(\xi)]^2\,d\xi + \int_c^d dx \left\{ \int_a^b K(x, \xi)\,u(\xi)\,d\xi - f(x) \right\}^2 = \min. \quad (50)$$

Составим для этой вариационной задачи уравнение Эйлера. Для этого приравняем нулю вариацию левой части по $u(\xi)$:

$$\alpha \sum_{k=0}^{n} \int_a^b p_k(\xi)\,u^{(k)}(\xi)\,\delta u^{(k)}(\xi)\,d\xi + \\ + \int_c^d dx \left\{ \int_a^b K(x, \eta)\,u(\eta)\,d\eta - f(x) \right\} \int_a^b K(x, \xi)\,\delta u(\xi)\,d\xi = 0. \quad (51)$$

Интегралы, стоящие под знаком суммы, вычислим последовательным интегрированием по частям:

$$\int_a^b p_k(\xi)\,u^{(k)}(\xi)\,\delta u^{(k)}(\xi)\,d\xi = \\ = \sum_{r=0}^{k-1} (-1)^r\,\delta u^{(k-1-r)}(\xi)\,\frac{d^r}{d\xi^r}[p_k(\xi)\,u^{(k)}(\xi)]\Big|_{\xi=a}^{\xi=b} + \\ + (-1)^k \int_a^b \delta u(\xi)\,\frac{d^k}{d\xi^k}[p_k(\xi)\,u^{(k)}(\xi)]\,d\xi. \quad (52)$$

Подставляя (52) в (51) и меняя порядок суммирования в двойной сумме по краевым вариациям, найдем

$$\alpha \sum_{r=1}^{n} \delta u^{(r)}(\xi) \sum_{k=r}^{n} (-1)^{k-r} \frac{d^{k-r}}{d\xi^{k-r}}[p_k(\xi)\,u^{(k)}(\xi)]\Big|_{\xi=a}^{\xi=b} + \\ + \alpha \sum_{k=0}^{n} (-1)^k \int_a^b \delta u(\xi)\,\frac{d^k}{d\xi^k}[p_k(\xi)\,u^{(k)}(\xi)]\,d\xi + \\ + \int_c^d dx \int_a^b K(x, \eta)\,u(\eta)\,d\eta \int_a^b K(x, \xi)\,\delta u(\xi)\,d\xi = \\ = \int_c^d f(x)\,dx \int_a^b K(x, \xi)\,\delta u(\xi)\,d\xi.$$

Полагая в этом выражении δ-функцию в качестве вариации $\delta u(\xi)$, получим искомое уравнение Эйлера; оно будет интегро-дифференциальным:

$$\alpha \sum_{k=0}^{n} (-1)^k \frac{d^k}{d\xi^k}[p_k(\xi)\,u^{(k)}(\xi)] + \int_a^b Q(\xi, \eta)\,u(\eta)\,d\eta = \Phi(\xi),\ a \leqslant \xi \leqslant b, \quad (53a)$$

с ядром и правой частью

$$Q(\xi, \eta)=\int_c^d K(x, \xi) K(x, \eta)\, dx, \quad \Phi(\xi)=\int_c^d K(x, \xi) f(x)\, dx \qquad (53б)$$

и краевыми условиями

$$q_r[u(a)]=q_r[u(b)]=0, \quad 1 \leqslant r \leqslant n; \quad q_r[u]=\sum_{k=r}^{n}(-1)^k \frac{d^{k-r}}{d\xi^{k-r}}(p_k u^{(k)}). \qquad (53в)$$

Заметим, что ядро $Q(\xi, \eta)$ определено на квадрате $[a, b; a, b]$, симметрично и непрерывно, а правая часть $\Phi(\xi)$ непрерывна.

Формулировка задачи (42) в виде уравнения Эйлера (53) позволяет доказать, не пользуясь аппаратом функционального анализа, что построенный алгоритм является регуляризирующим; при этом для простоты будем полагать $p_k(\xi) \equiv 1$.

Теорема 1. *Задача* (53) *корректно поставлена при любом* $\alpha>0$.

Доказательство. Сначала рассмотрим простейший случай $n=0$. При этом исчезают все краевые условия (53в) и производные в уравнении (53а), и задача (53) превращается в интегральное уравнение Фредгольма второго рода:

$$\alpha u(\xi)+\int_a^b Q(\xi, \eta) u(\eta)\, d\eta=\Phi(\xi) \qquad (54)$$

с ядром и правой частью (53б).

Пусть λ_i, $u_i(\xi)$ — собственные значения и собственные функции ядра $Q(\xi, \eta)$. Поскольку ядро имеет вид (53б), то они удовлетворяют уравнению

$$u_i(\xi)=\lambda_i \int_a^b u_i(\eta)\, d\eta \int_c^d K(x, \xi) K(x, \eta)\, dx.$$

Умножая обе части уравнения на $u_i(\xi)$ и интегрируя, получим

$$0<\int_a^b u_i^2(\xi)\, d\xi=\lambda_i \int_c^d dx\left\{\int_a^b K(x, \xi) u_i(\xi)\, d\xi\right\}^2.$$

Отсюда видно, что все собственные значения ядра $Q(\xi, \eta)$ положительны.

Поэтому, согласно теории интегральных уравнений Фредгольма (см. § 1, п. 1), при любом $\alpha>0$ уравнение (54) имеет решение $u_\alpha(\xi)$, причем это решение единственно и непрерывно зависит от правой части $\Phi(\xi)$ и, тем самым, от $f(x)$. Таким образом, при $n=0$ задача (53) и эквивалентная ей задача (42) корректны.

При $n>0$ задачу (53) также можно свести к интегральному уравнению. Построим функцию Грина $G(\xi, \tau)$ для дифференциального оператора, стоящего в левой части (53а), при краевых условиях (53в). Рассматривая все интегральные члены в (53а) как правую часть дифференциального уравнения, выразим через них решение при помощи функции Грина:

$$\alpha u(\xi)+\int_a^b u(\eta)\,d\eta\int_a^b G(\xi,\tau)\,Q(\tau,\eta)\,d\tau=\int_a^b G(\xi,\tau)\,\Phi(\tau)\,d\tau. \qquad (55)$$

Таким образом, $u(\xi)$ удовлетворяет уравнению Фредгольма второго рода, причем его ядро имеет только положительные собственные значения. Следовательно, задача (53) корректна при любом n, если $\alpha>0$, что и требовалось доказать.

Замечание 1. Интегро-дифференциальное уравнение (53а) содержит производные решения вплоть до порядка $2n$. Поэтому $u_\alpha(\xi)$ имеет $2n$ непрерывных производных.

Теорема 2. *Пусть* $A[x, \bar{u}]=\bar{f}$; *тогда при* $n=1$ *и положительном* $\alpha\to 0$ *решение* $\bar{u}_\alpha(\xi)$ *задачи* (53), *соответствующее правой части* $\bar{f}(x)$, *равномерно сходится к* $\bar{u}(\xi)$.

Доказательство. При $n=1$ решения $u_\alpha(\xi)$ задачи (53) с любой правой частью являются дважды непрерывно дифференцируемыми. Применяя неравенство Коши — Буняковского, найдем

$$\left(\int_\xi^{\xi+\delta}|u'_\alpha(\tau)|\,d\tau\right)^2\leqslant\int_\xi^{\xi+\delta}d\tau\cdot\int_\xi^{\xi+\delta}|u'_\alpha(\tau)|^2\,d\tau\leqslant$$

$$\leqslant\delta\cdot\int_a^b[u'_\alpha(\tau)]^2\,d\tau\leqslant\delta\cdot\Omega_1[u_\alpha]. \qquad (56)$$

Рассмотрим множество решений $\bar{u}_\alpha(\xi)$, соответствующих одной и той же правой части $\bar{f}(x)$, но разным значениям параметра $\alpha>0$. Полагая $\tilde{f}=\bar{f}$ в неравенстве (44), получим

$$\Omega_1[\bar{u}_\alpha]\leqslant\overline{\Omega}_1,\quad \overline{\Omega}_1=\Omega_1[\bar{u}]. \qquad (57)$$

Из неравенств (56) и (57) следует

$$|\bar{u}_\alpha(\xi+\delta)-\bar{u}_\alpha(\xi)|\leqslant\int_\xi^{\xi+\delta}|u'_\alpha(\tau)|\,d\tau\leqslant\sqrt{\delta\overline{\Omega}_1}, \qquad (58)$$

что означает равностепенную непрерывность множества функций $\bar{u}_\alpha(\xi)$. Кроме того, согласно определению функционала Ω_1 при $p_k(\xi)\equiv 1$,

$$(b-a)\min|\bar{u}_\alpha(\xi)|^2\leqslant\Omega_1[\bar{u}_\alpha]\leqslant\overline{\Omega}_1. \qquad (59)$$

Из (59), (57) и (42б) вытекает, что

$$\max|\bar{u}_\alpha(\xi)| \leqslant \min|\bar{u}_\alpha(\xi)| + \int_a^b |\bar{u}'_\alpha(\xi)|\,d\xi \leqslant \left(\sqrt{b-a} + \frac{1}{\sqrt{b-a}}\right)\sqrt{\Omega_1}, \tag{60}$$

т. е. функции $\bar{u}_\alpha(\xi)$ равномерно ограничены.

Теперь предположим, что функции $\bar{u}_\alpha(\xi)$ не сходятся равномерно к $\bar{u}(\xi)$ при $\alpha \to 0$, т. е. для некоторого $\varepsilon > 0$ найдется такая последовательность $\alpha_k \to 0$, что $\|\bar{u}_{\alpha_k}(\xi) - \bar{u}(\xi)\|_C \geqslant \varepsilon$.

Построим на отрезке $a \leqslant \xi \leqslant b$ последовательность сгущающихся вдвое сеток. Узлы этих сеток образуют счетное множество точек. Перенумеруем эти узлы, как указано на рис. 103. Тогда для отрезка этого множества, состоящего из первых N узлов, длина интервала между соседними узлами не превышает

$$\delta = 2(b-a)/N.$$

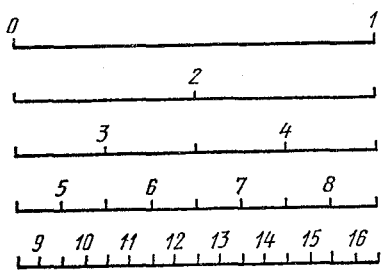


Рис. 103.

Из последовательности ограниченных в совокупности функций $\bar{u}_{\alpha_k}(\xi)$ можно выбрать подпоследовательность, сходящуюся в узле ξ_1. Из этой подпоследовательности выберем подпоследовательность, сходящуюся в узле ξ_2, и т. д. В итоге построим подпоследовательность $\hat{u}_{\alpha_k}(\xi)$, сходящуюся в каждом узле ξ_i к некоторому пределу $\hat{u}(\xi_i)$.

Выберем сколь угодно малое $\varepsilon > 0$ и положим $N = 18(b-a) \times \times \Omega_1 \varepsilon^{-2}$. Возьмем настолько малое $\alpha_0(\varepsilon)$, чтобы при $\alpha < \alpha_0(\varepsilon)$ во всех узлах ξ_i с номерами $i \leqslant N$ выполнялось неравенство $|\hat{u}_{\alpha_k}(\xi_i) - \hat{u}(\xi_i)| \leqslant \varepsilon/3$. Интервал между соседними узлами настолько мал, что в силу (58) значения $\hat{u}_{\alpha_k}(\xi_i)$ в соседних узлах будут различаться меньше, чем на $\varepsilon/3$. Тогда значения $\hat{u}(\xi_i)$ в соседних узлах с номерами $i \leqslant N$ будут различаться меньше чем на ε.

Отсюда, во-первых, следует, что функцию $\hat{u}(\xi)$ можно доопределить во всех точках отрезка $a \leqslant \xi \leqslant b$ так, что она будет непрерывной. Во-вторых, подпоследовательность $\hat{u}_{\alpha_k}(\xi)$ равномерно сходится к доопределенной функции $\hat{u}(\xi)$.

Функции $\hat{u}_{\alpha_k}(\xi)$ являются решением задачи (42) с правой частью $\bar{f}(x)$. Подставляя их в эту задачу и переходя к пределу при $\alpha_k \to 0$, мы убеждаемся, что $\hat{u}(\xi)$ является решением этой задачи при $\alpha = 0$, т. е. решением задачи (41). Поскольку реше-

ние последней задачи единственно, то $\hat{u}(\xi)=\bar{u}(\xi)$, что противоречит сделанному в ходе доказательства предположению. Это противоречие доказывает теорему.

Теорема 3. *Алгоритм* (42) *при* $n=1$ *обеспечивает сильную регуляризацию.*

Доказательство. Пусть точной правой части $\bar{f}(x)$ соответствуют точное решение $\bar{u}(\xi)$ и регуляризованное решение $\bar{u}_\alpha(\xi)$, а приближенной правой части $\tilde{f}(x)$ соответствует регуляризованное решение $\tilde{u}_\alpha(\xi)$.

Зададим сколь угодно малое $\varepsilon>0$. По теореме 2 найдется такое $\alpha_0(\varepsilon)$, что $\|\bar{u}_\alpha-\bar{u}\|_C\leqslant\varepsilon/2$ при $\alpha\leqslant\alpha_0(\varepsilon)$.

Согласно теореме 1 задача (42) корректна, так что при любом заданном $\alpha>0$ найдется такое $\delta(\alpha)$, что если $\|\bar{f}-\tilde{f}\|\leqslant\delta(\alpha)$, то $\|\bar{u}_\alpha-\tilde{u}_\alpha\|_C\leqslant\varepsilon/2$.

Следовательно, если $\alpha\leqslant\alpha_0(\varepsilon)$ и $\|\bar{f}-\tilde{f}\|\leqslant\delta(\alpha)$, то

$$\|\tilde{u}_\alpha-\bar{u}\|_C\leqslant\|\tilde{u}_\alpha-\bar{u}_\alpha\|_C+\|\bar{u}_\alpha-\bar{u}\|_C\leqslant\varepsilon.$$

Это соответствует определению сильной регуляризации (см. п. 1); теорема доказана.

Замечание 2. Поясним действие регуляризации простыми рассуждениями. Пусть правая часть $\Phi(\xi)$ получила возмущение $\beta e^{i\omega\xi}$; тогда решение получит возмущение $\gamma e^{i\omega\xi}$. Прибавляя эти возмущения в (53а) и оценивая каждое слагаемое по порядку величины, получим

$$\gamma\left(\alpha\sum_{k=0}^{n}\omega^{2k}+\frac{1}{\omega}\right)\sim\beta.$$

Рассмотрим поведение возмущений при больших частотах. Если $\alpha=0$, то $\gamma\sim\omega\beta$, т. е. возмущения решения велики, и расчет неустойчив. Регуляризации нет.

Если $\alpha\neq0$, но $n=0$, то $\gamma\sim\beta/\alpha$, т. е. возмущения решения по порядку величины равны возмущениям правой части, и расчет становится устойчивым. Чем больше α, тем меньше возмущения решения и «разболтка» в численном расчете. Но сдвиги фаз отдельных гармоник приводят к тому, что сходимость будет только среднеквадратичной (слабая регуляризация).

Если $n=1$, то $\gamma\sim\beta/\alpha\omega^2$ и возмущения решения для высоких частот малы. Значит, расчет хорошо устойчив и $\bar{u}_\alpha(\xi)$ равномерно сходится к $\bar{u}(\xi)$ (сильная регуляризация). При $n>1$ амплитуды γ_ω настолько быстро убывают при $\omega\to\infty$, что обеспечивается равномерная сходимость не только регуляризованного решения, но и его $(n-1)$-й производной.

4. Некоторые приложения. Некорректные задачи встречаются в практике вычислений довольно часто. К ним относятся, напри-

мер, сглаживание и дифференцирование экспериментально измеренных функций, суммирование рядов Фурье с неточно заданными коэффициентами, решение плохо обусловленных линейных систем, задачи оптимального управления, аналитическое продолжение функций, линейное программирование (оптимальное планирование), обратные задачи теплопроводности и геологической разведки, восстановление переданного сигнала по принятому при наличии искажений аппаратуры и многие другие.

Некоторые из этих задач встречались в предыдущих главах. Покажем, как они регуляризируются вариационным методом. Для определенности ограничимся сильной регуляризацией, полагая $n=1$ в формулах (42) или (53).

Сглаживание функции. Пусть функция $f(x)$, $a \leqslant x \leqslant b$, измерена экспериментально и содержит заметную случайную погрешность. Тогда математическая задача имеет вид $u(x)=f(x)$; ее можно записать в каноническом виде $A[x, u(\xi)]=f(x)$, полагая $A[x, u(\xi)] \equiv u(x)$. Подставляя последнее выражение в измененную задачу (42), составим уравнение Эйлера (53):

$$\alpha\left[\frac{d}{dx}\left(p_1\frac{du}{dx}\right)-p_0u(x)\right]-u(x)+f(x)=0, \quad u'(a)=u'(b)=0. \qquad (61)$$

Таким образом, сглаженная функция $u(x)$ удовлетворяет линейному обыкновенному дифференциальному уравнению второго порядка, для которого поставлена вторая краевая задача. Методы численного решения этой задачи подробно разобраны в главе VIII.

Замечание 1. Весовые функции $p_0(x)$ и $p_1(x)$ выбирают, исходя из дополнительных сведений о виде функции $f(x)$ и величине погрешности $\delta f(x)$. Например, $p_k(x)$ целесообразно брать большими в тех диапазонах значений x, где погрешность $\delta f(x)$ особенно велика. Если подобных сведений нет, то обычно полагают $p_0(x)=p_1(x)=1$.

Замечание 2. На концах отрезка $[a, b]$ погрешность сглаживания может быть значительна, поскольку краевые условия второго рода в (61) не соответствуют, вообще говоря, истинному поведению функции.

Замечание 3. Можно уменьшить погрешность сглаживания вблизи концов отрезка $[a, b]$, если воспользоваться регуляризацией более высокого порядка (см. задачу 10). Однако, как отмечалось в п. 2, при этом могут исказиться качественные особенности решения (типа, например, узких экстремумов).

Дифференцирование. Задачу дифференцирования $u(x)=f'(x)$, $a \leqslant x \leqslant b$, можно записать в виде уравнения Вольтерра первого рода (36):

$$\int_a^x u(\xi)\,d\xi = f(x)-f(a),$$

или, формально, в виде уравнения Фредгольма первого рода с разрывным ядром:

$$\int_a^b K(x, \xi)\, u(\xi)\, d\xi = f(x) - f(a), \qquad a \leqslant x \leqslant b; \tag{62а}$$

$$\begin{aligned} K(x, \xi) &= 1 \text{ при } a \leqslant \xi \leqslant x \leqslant b, \\ K(x, \xi) &= 0 \text{ при } \xi > x. \end{aligned} \tag{62б}$$

Поскольку требование непрерывности ядра не является существенным, применим к этой задаче алгоритм (53). Легко получим

$$Q(\xi, \eta) = \int_a^b K(x, \xi)\, K(x, \eta)\, dx = b - \max(\xi, \eta),$$

$$\Phi(\xi) = \int_a^b [f(x) - f(a)]\, K(x, \xi)\, dx = \int_\xi^b [f(x) - f(a)]\, dx.$$

Отсюда вытекает, что регуляризованное решение удовлетворяет следующему интегро-дифференциальному уравнению и краевым условиям:

$$-\alpha \left[\frac{d}{d\xi}\left(p_1(\xi)\frac{du}{d\xi}\right) - p_0(\xi)\, u(\xi)\right] + \int_a^\xi (b - \xi)\, u(\eta)\, d\eta + \int_\xi^b (b - \eta)\, u(\eta)\, d\eta = \int_\xi^b [f(x) - f(a)]\, dx, \tag{63}$$

$$u'(a) = 0, \quad u'(b) = 0.$$

К этой задаче также относятся сделанные выше замечания о выборе весовых функций, о значительной погрешности на концах отрезка $[a, b]$ и возможностях ее уменьшения.

Суммирование ряда Фурье. Пусть задана полная ортонормированная система функций $\varphi_s(x)$, которую можно рассматривать как систему собственных функций некоторой задачи Штурма — Лиувилля:

$$-\frac{d}{dx}\left[p_1(x)\frac{d\varphi}{dx}\right] - [p_0(x) + \lambda]\,\varphi(x) = 0, \qquad \varphi'(a) = 0, \quad \varphi'(b) = 0. \tag{64}$$

Требуется просуммировать ряд Фурье

$$f(x) = \sum_{s=1}^{\infty} \beta_s \varphi_s(x), \tag{65}$$

коэффициенты которого β_s заданы приближенно.

Эту задачу можно рассматривать как сглаживание неточно заданной функции $f(x)$. Воспользуемся для ее решения уравнением (61), где в качестве $p_0(x)$ и $p_1(x)$ выбраны веса, входящие

в задачу Штурма — Лиувилля (64). Будем искать регуляризованное решение также в виде ряда Фурье:

$$u(x)=\sum_{s=1}^{\infty}\gamma_s\varphi_s(x). \tag{66}$$

Подставляя (66) и (65) в (61) и учитывая (64), получим

$$\gamma_s=\frac{\beta_s}{1+\alpha\lambda_s}, \tag{67}$$

где $\lambda_s>0$ — собственные значения задачи Штурма — Лиувилля (64). Этот способ регуляризации приводился без доказательства в гл. II, § 2, п. 3.

Плохо обусловленные линейные системы $A\boldsymbol{u}=\boldsymbol{f}$, где $\boldsymbol{u}$ и $\boldsymbol{f}$ — конечномерные векторы, можно регуляризировать, записывая их непосредственно в вариационной форме (42) и выбирая $n=0$:

$$\|A\boldsymbol{u}-\boldsymbol{f}\|^2+\alpha\|\boldsymbol{u}\|^2=\min, \qquad \|\boldsymbol{a}\|^2=(\boldsymbol{a},\ \boldsymbol{a}). \tag{68}$$

Формально $n=0$ соответствует слабой регуляризации. Но в конечномерном пространстве все нормы эквивалентны, поэтому сходимость регуляризованного решения к точному при $\alpha\to 0$ является равномерной.

Уравнение (68) означает, что среди решений, приближенно удовлетворяющих исходной задаче, ищут вектор наименьшей длины. Часто рассматривают более общую постановку:

$$\|A\boldsymbol{u}-\boldsymbol{f}\|^2+\alpha\|\boldsymbol{u}-\boldsymbol{u}_0\|^2=\min, \tag{69}$$

которая определяет *нормальное* решение — приближенное решение, наименее отличающееся от заданного вектора $\boldsymbol{u}_0$. Ее используют, например, в задачах линейного программирования (см. гл. VII, § 3).

Поскольку (69) является квадратичной формой относительно $\boldsymbol{u}$, то нахождение ее минимума сводится к решению линейной алгебраической системы

$$(A^HA+\alpha E)\,\boldsymbol{u}=A^H\boldsymbol{f}+\alpha\boldsymbol{u}_0. \tag{70}$$

Благодаря слагаемому αE эта система хорошо обусловлена, по крайней мере, при не слишком малых $\alpha>0$. Поэтому ее нетрудно решить методом исключения Гаусса.

Описанный алгоритм применяют также для решения систем с вырожденной матрицей A.

5. Разностные схемы. При вариационном методе регуляризации численно решать приходится либо задачу на минимум функционала (42), либо краевую задачу для интегро-дифференциального уравнения Эйлера (53). К этим задачам целесообразно применять разностные методы.

Дадим пример построения разностной схемы, исходя из вариационной формулировки (42). Введем на прямоугольнике $[c \leqslant x \leqslant d,\ a \leqslant \xi \leqslant b]$ сетку $\{x_n,\ \xi_m,\ 0 \leqslant n \leqslant N,\ 0 \leqslant m \leqslant M\}$ так, что $x_0 = c$, $x_N = d$, $\xi_0 = a$, $\xi_M = b$. Для простоты ограничимся случаем равномерных сеток $x_n = c + nh_x$, $\xi_m = a + mh_\xi$, сильной регуляризации и единичных весовых функций $p_0(\xi) = p_1(\xi) = 1$.

Задача (42) при указанных ограничениях принимает вид

$$\int_c^d [\delta(x)]^2\,dx + \alpha \int_a^b \left[u^2(\xi) + \left(\frac{du}{d\xi}\right)^2\right] d\xi = \min, \tag{71а}$$

$$\delta(x) = \int_a^b K(x, \xi)\, u(\xi)\, d\xi - f(x), \tag{71б}$$

где величина, обозначенная через $\delta(x)$, имеет смысл невязки исходной нерегуляризованной системы при подстановке в нее регуляризованного решения. Аппроксимируем входящие в (71) интегралы квадратурными формулами, использующими значения функций в узлах сетки. Для этого $\int (u')^2\,d\xi$ вычислим по формуле средних (4.17), одновременно заменяя производную разностью:

$$\int_{\xi_m}^{\xi_{m+1}} \left(\frac{du}{d\xi}\right)^2 d\xi \approx h_\xi \left(\frac{du}{d\xi}\right)^2_{m+1/2} \approx h_\xi \left(\frac{u_{m+1} - u_m}{h_\xi}\right)^2. \tag{72}$$

Остальные интегралы вычислим по формуле трапеций (4.8):

$$\int_a^b u^2(\xi)\,d\xi \approx h_\xi \sum_{m=0}^{M} c_m u_m^2, \qquad u_m = u(\xi_m); \tag{73}$$

$$\int_a^b K(x_n, \xi)\, u(\xi)\,d\xi \approx h_\xi \sum_{m=0}^{M} c_m K_{nm} u_m, \qquad K_{nm} = K(x_n, \xi_m); \tag{74}$$

$$\int_c^d [\delta(x)]^2\,dx \approx h_x \sum_{n=0}^{N} b_n [\delta(x_n)]^2, \tag{75}$$

где

$$\begin{aligned} &c_m = 1 \text{ при } 1 \leqslant m \leqslant M-1, \quad c_0 = c_M = {}^1/_2, \\ &b_n = 1 \text{ при } 1 \leqslant n \leqslant N-1, \quad b_0 = b_N = {}^1/_2. \end{aligned} \tag{76}$$

Подставляя (72) — (76) в (71) и обозначая разностное решение через y_m, получим вместо (71) алгебраическую задачу

$$h_x \sum_{n=0}^{N} b_n \left\{ h_\xi \sum_{m=0}^{M} c_m K_{nm} y_m - f_n \right\}^2 + \\ + \alpha h_\xi \sum_{m=0}^{M} c_m y_m^2 + \frac{\alpha}{h_\xi} \sum_{m=0}^{M-1} (y_{m+1} - y_m)^2 = \min \tag{77}$$

на минимизацию квадратичной формы.

Для решения этой задачи приравняем нулю производные от левой части (77) по y_m. Получим систему уравнений, линейных относительно y_m:

$$\alpha y_m - \frac{\alpha}{c_m}\Lambda(y_m) + h_\xi \sum_{l=0}^{M} c_l Q_{ml} y_l = \Phi_m, \qquad 0 \leqslant m \leqslant M; \tag{78а}$$

$$\Lambda(y_m) = \frac{1}{h_\xi^2}(y_{m-1} - 2y_m + y_{m+1}) \text{ при } 1 \leqslant m \leqslant M-1,$$
$$\Lambda(y_0) = \frac{1}{h_\xi^2}(y_1 - y_0), \qquad \Lambda(y_M) = \frac{1}{h_\xi^2}(y_{M-1} - y_M); \tag{78б}$$

где

$$Q_{ml} = h_x \sum_{n=0}^{N} b_n K_{nm} K_{nl}, \qquad \Phi_m = h_x \sum_{n=0}^{N} b_n K_{nm} f_n. \tag{78в}$$

Матрица системы (78) является, вообще говоря, плотно заполненной; поэтому обычно эту систему решают методом исключения Гаусса.

На исследовании полученной разностной схемы не будем останавливаться, поскольку сходные вопросы были рассмотрены в главе VII, § 4. Отметим только, что схема (77) или (78) имеет аппроксимацию $O(h_x^2 + h_\xi^2)$, если ядро и правая часть непрерывны со своими вторыми производными.

Замечание 1. Если умножить уравнение (78а) на c_m, то матрица этой линейной системы станет симметричной. Тогда для решения этой системы можно будет применить метод квадратного корня (который вдвое быстрей метода Гаусса).

Замечание 2. Нетрудно видеть, что Q_{ml} и Φ_m являются разностными аналогами ядра и правой части (53б) интегро-дифференциального уравнения Эйлера. Выражение $\Lambda(y_m)$, возникшее при дифференцировании последней суммы в (77), есть разностный аналог дифференциального оператора в уравнении (53а). Поэтому система (78) аппроксимирует также задачу регуляризации в форме уравнения Эйлера (53), причем выражения $\Lambda(y_0)$ и $\Lambda(y_M)$ учитывают краевые условия (53в).

ЗАДАЧИ

1. Показать, что интегральное уравнение

$$(\beta-\alpha)x + (b\alpha - a\beta) - (b-a)u(x) =$$
$$= (x-a)\int_x^b (b-\xi) f(\xi, u(\xi))\, d\xi + (b-x)\int_a^x (\xi - a) f(\xi, u(\xi))\, d\xi \tag{79}$$

эквивалентно краевой задаче для дифференциального уравнения

$$u''(x) = f(x, u), \quad u(a) = \alpha, \quad u(b) = \beta.$$

2. Записать уравнение (79) в каноническом виде (1); найти выражение для ядра $K(x, \xi, u)$.

3. Для уравнения Вольтерра (7) составить разностную схему и полный алгоритм вычисления разностного решения, используя формулу трапеций с равномерным шагом.

4. Для двумерного уравнения Фредгольма (6) составить разностную схему, используя в качестве кубатурной формулы произведение одномерных формул Гаусса.

5. В методе последовательных приближений для уравнения (6) выразить $u_n(x)$ через $u_1(x)$ при помощи рекуррентного соотношения (16).

6. Доказать, что из соотношения (20) следует оценка (21).

7. Учитывая, что уравнение (23) имеет вырожденное ядро, а) найти его точное решение; б) сделать то же для $f(x) = \sin x$.

8. В уравнении (23) так подобрать правую часть $f(x)$, чтобы при $\lambda = 2$ существовало решение.

9. Доказать утверждение, сформулированное в § 1, п. 5, замечании 1.

10. Для задачи сглаживания функции $u(x) = f(x)$ написать уравнение и краевые условия вариационной регуляризации с $n = 2$. Обсудить влияние n на погрешность сглаживания вблизи границ, для простоты полагая $p_n(x) = 1$ и $p_k(x) = 0$ при $k < n$.

11. Регуляризировать задачу p-кратного дифференцирования $u(x) = f^{(p)}(x)$, используя запись этой задачи в виде интегрального уравнения

$$\frac{1}{(p-1)!} \int_a^x (x-\xi)^{p-1} u(\xi)\, d\xi = f(x). \tag{80}$$

12. Аппроксимировать разностной схемой краевую задачу для уравнения Эйлера (53); сравнить ее с разностной схемой (78).

13. Составить разностную схему для регуляризации однократного дифференцирования, если $f(x)$ задана а) на равномерной сетке, б) на неравномерной сетке.

ГЛАВА XV

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТА

В главе XV рассмотрены основные вопросы статистической обработки результатов эксперимента: определение наиболее достоверного значения измеряемой величины и погрешности этого значения по нескольким измерениям, оценка достоверности различия двух близких величин, установление достоверной функциональной зависимости между двумя величинами и аппроксимация этой зависимости.

Глава носит вспомогательный характер. Материал в ней изложен в справочной форме, без доказательств. Обоснование и более подробное изложение приведенных методов имеется, например, в [7, 26, 43].

1. Ошибки эксперимента. Численные методы часто применяют при математическом моделировании физических и других процессов. Результаты расчетов в этом случае сравнивают с экспериментальными данными и по степени их согласованности судят о качестве выбранной математической модели. Чтобы обоснованно сделать заключение о соответствии или несоответствии, вычислитель должен знать, что такое погрешность эксперимента и как с ней обращаются, а также уметь в случае необходимости провести статистическую обработку первичных данных эксперимента.

Кроме того, задача статистической обработки эксперимента представляет самостоятельный интерес, поскольку она очень важна в тех приложениях, когда или требуется особенно высокая точность (например, уравнивание триангуляционных сетей в геодезии), или разброс отдельных измерений превосходит исследуемый эффект (что нередко встречается в физике элементарных частиц, химии сложных соединений, испытании сельскохозяйственных сортов, медицине и т. д.).

Обычно, чем точнее эксперимент, тем более сложной аппаратуры он требует и дороже обходится. Однако хорошо продуманная математическая обработка результатов в ряде случаев позволяет выявить и частично исключить ошибки измерений; это может оказаться не менее эффективным, чем использование более дорогой и точной аппаратуры. В этой главе будет рассмотрена статистическая обработка, позволяющая существенно уменьшить и аккуратно оценить случайную ошибку измерений.

Ошибки эксперимента условно разбивают на систематические, случайные и грубые; рассмотрим их подробнее.

Систематические ошибки — это те, которые не меняются при многократном повторении данного эксперимента. Примерами таких ошибок являются пренебрежение выталкивающим действием воздуха при точном взвешивании или измерение тока гальванометром, нуль которого неправильно установлен. Различают три вида систематических ошибок.

а) Ошибки известной природы, величину которых можно определить; их называют *поправками*. Так, при точном взвешивании рассчитывают поправку на выталкивающее действие воздуха и прибавляют ее к измеренной величине. Внесение поправок позволяет существенно уменьшить (или даже практически исключить) ошибки такого рода.

Заметим, что иногда расчет поправок бывает самостоятельной сложной математической задачей. Например, некорректно поставленная задача (14.2) о восстановлении переданного радиосигнала по принятому является, по существу, нахождением поправки на искажение принимающей аппаратуры.

б) Ошибки известного происхождения, но неизвестной величины. К ним относится погрешность измерительных приборов, определяемая их классом точности. Для таких ошибок обычно известна только верхняя граница, а как поправки их учесть нельзя.

в) Ошибки, о существовании которых мы не знаем; например, используется прибор со скрытым дефектом или изношенный, фактическая точность которого существенно хуже, чем обозначено в техническом паспорте.

Для выявления систематических ошибок всех видов обычно заранее отлаживают аппаратуру на эталонных объектах с хорошо известными свойствами.

Случайные ошибки вызываются большим числом факторов, которые при повторении одного и того же эксперимента могут действовать по-разному, причем учесть их влияние практически невозможно. Например, при измерении длины предмета линейка может быть неточно приложена, взгляд наблюдателя может падать не перпендикулярно шкале и т. д.

При многократном повторении эксперимента результат вследствие случайной ошибки будет различным. Однако такое повторение и соответствующая статистическая обработка позволяют, во-первых, определить величину случайной ошибки и, во-вторых, уменьшить ее. Повторяя измерение достаточное число раз, можно уменьшить случайную ошибку до требуемой величины (целесообразно уменьшать ее до величины 50 — 100% от систематической ошибки).

Грубые ошибки — это результат невнимательности наблюдателя, который может записать одну цифру вместо другой. При

единичном измерении грубую ошибку не всегда можно опознать. Но если измерение повторено несколько раз, то при статистической обработке выясняют вероятные пределы случайной ошибки. Измерение, существенно выходящее за полученные пределы, считается грубо ошибочным и не учитывается при окончательной обработке результатов.

Таким образом, если измерение повторено достаточно много раз, то можно практически исключить грубые и случайные ошибки, так что точность ответа будет определяться только систематической ошибкой. Однако во многих приложениях это требуемое число раз оказывается неприемлемо большим, а при реально осуществимом числе повторений случайная ошибка может быть определяющей.

2. Величина и доверительный интервал. Пусть измерение проводят несколько раз, причем условия эксперимента поддерживают, насколько возможно, неизменными. Поскольку строго соблюдать неизменность условий невозможно, результаты отдельных измерений x_i будут несколько различаться. Их можно рассматривать как значения случайной величины ξ, распределенной по некоторому закону, заранее нам неизвестному.

Очевидно, *математическое ожидание* $\mathsf{M}\xi$ *равно точному значению измеряемой величины* x (строго говоря, точному значению плюс систематическая ошибка).

Обработка измерений основана на центральной предельной теореме теории вероятностей: *если* ξ *есть случайная величина, распределенная по любому закону, то*

$$\eta_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

есть также случайная величина, причем

$$\mathsf{M}\eta_n = \mathsf{M}\xi, \qquad \mathsf{D}\eta_n = \frac{1}{n}\mathsf{D}\xi, \tag{1}$$

а закон распределения величины η_n *стремится к нормальному (гауссову) при* $n \to \infty$. Поэтому среднеарифметическое нескольких независимых измерений

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathsf{M}\xi \tag{2}$$

является приближенным значением измеряемой величины, причем с тем большей надежностью, чем больше число измерений n.

Однако равенство $\bar{x} \approx \mathsf{M}\xi$ не является точным, и нельзя даже строго указать предел его ошибки; в принципе $\bar{x}$ может сколь угодно сильно отличаться от $\mathsf{M}\xi$, хотя вероятность такого собы-

тия ничтожно мала. Ошибка приближенного равенства (2) носит вероятностный характер и описывается *доверительным интервалом* β, т. е. границей, которую с *доверительной вероятностью* p_0 не превышает разность $|\bar{x} - \mathsf{M}\xi|$. Символически это записывают следующим образом:

$$p\{|\bar{x} - \mathsf{M}\xi| \leqslant \beta\} = p_0. \tag{3}$$

Доверительный интервал зависит от закона распределения ξ (а тем самым — от постановки эксперимента), от числа измерений n, а также от выбранной доверительной вероятности p_0. Из (3) видно, что чем ближе p_0 к единице, тем шире оказывается доверительный интервал.

Доверительную вероятность p_0 выбирают, исходя из практических соображений, связанных с применениями полученных результатов. Например, если мы делаем игрушечный воздушный змей, то вероятность благополучного полета $p_0 = 0{,}8$ нас устроит, а если конструируем самолет, то даже вероятность $p_0 = 0{,}999$ недостаточна. Во многих физических измерениях $p_0 = 0{,}95 \div 0{,}99$ считается достаточной.

Замечание 1. Пусть требуется найти величину z, но измерять удобнее величину x, связанную с ней известным соотношением $z = f(x)$; например, нас интересует джоулево тепло, а измерять легче ток. При этом следует помнить, что

$$\mathsf{M}\zeta = \int f(x)\,\rho(x)\,dx \neq f(\mathsf{M}\xi) = f\left(\int x\rho(x)\,dx\right);$$

так, среднее значение переменного тока равно нулю, а средний джоулев нагрев отличен от нуля. Поэтому, если мы вычислим сначала $\bar{x}$, а затем положим $\bar{z} = f(\bar{x})$, это будет грубая ошибка. Следует по каждому измерению x_i вычислять $z_i = f(x_i)$ и далее обрабатывать полученные значения z_i.

Ширина доверительного интервала. Если известна плотность распределения $\rho_n(y)$ величины η_n, то доверительный интервал можно определить из (3), разрешая уравнение

$$p_0 = \int_{\mathsf{M}\eta - \beta}^{\mathsf{M}\eta + \beta} \rho_n(y)\,dy, \qquad \mathsf{M}\eta_n = \mathsf{M}\xi, \tag{4}$$

относительно β. Выше отмечалось, что при $n \to \infty$ распределение η_n стремится к нормальному *):

$$\rho_n(y) \approx \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left[-\frac{(y - \mathsf{M}\eta)^2}{2\sigma_n^2}\right], \qquad \sigma_n = \sqrt{\mathsf{D}\eta_n}; \tag{5}$$

*) В самом худшем случае, когда ξ есть равномерно распределенная случайная величина, распределение η_n близко к нормальному при $n \sim 30$, а интеграл в (3) близок к интегралу от нормального распределения при существенно меньших n.

здесь $\mathsf{D}\eta_n$ — дисперсия распределения, а величину σ_n называют *стандартным отклонением* или просто *стандартом* *).

Подставляя (5) в (4) и полагая $\beta = t\sigma_n$, т. е. измеряя доверительный интервал в долях стандарта, получим соотношение

$$p_0 = \sqrt{\frac{2}{\pi}} \int_0^t e^{-\tau^2/2}\, d\tau. \tag{6}$$

Интеграл ошибок, стоящий в правой части (6), табулирован, так что из этого соотношения можно определить доверительный интервал $t(p_0)$. Зависимость $t(p_0)$ дается в таблице 23 строкой, соответствующей $n = \infty$.

Из таблицы 23 видно, что доверительный интервал $\beta = 3\sigma_n$ соответствует доверительной вероятности $p_0 = 0{,}997$, так что отклонение $\bar{x}$ от $\mathsf{M}\xi$ более чем на $3\sigma_n$ маловероятно. Но отклонение более чем на σ_n довольно вероятно, поскольку ширине $\beta = \sigma_n$ соответствует $p_0 = 0{,}7$.

Таким образом, если известна дисперсия $\mathsf{D}\xi$, то нетрудно определить стандарт $\sigma_n = \sqrt{\mathsf{D}\xi/n}$ и, тем самым, абсолютную ширину доверительного интервала β. В этом случае даже при выполнении одного измерения можно оценить случайную ошибку **), а увеличение числа измерений позволяет уменьшать доверительный интервал, поскольку $\sigma_n \sim n^{-1/2}$.

Критерий Стьюдента. Чаще всего дисперсия $\mathsf{D}\xi$ неизвестна, поэтому выполнить оценку ошибки указанным выше способом обычно не удается. При этом точность однократного измерения неизвестна. Однако, если измерение повторено несколько раз, можно приближенно найти дисперсию:

$$\mathsf{D}\xi \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathsf{M}\xi)^2 \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{7}$$

Точность этого выражения невелика по двум причинам: во-первых, число членов суммы обычно мало; во-вторых, использование замены $\mathsf{M}\xi \approx \bar{x}$ вносит ошибку $O(1/n)$, значительную при малых n. Более хорошее приближение дает так называемая *несмещенная оценка* дисперсии:

$$\mathsf{D}\xi \approx s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{8}$$

*) Для произвольного закона распределения $\sqrt{\mathsf{D}\eta}$ называют среднеквадратичным отклонением.

**) Однако при $n = 1$ считать распределение $\rho_1(y) = \rho(x)$ нормальным и пользоваться формулой (6), вообще говоря, нельзя. Этот вопрос будет рассмотрен ниже.

где величину s называют *стандартом выборки*. Далее будем пользоваться только оценкой (8).

Оценка (8) также является приближенной, поэтому нельзя пользоваться формулой (6), заменяя в ней σ_n на $s/\sqrt{n}$. Надо вносить в нее поправку, тем бо́льшую, чем меньше n. Если распределение η_n считать нормальным при любых n, то связь доверительного интервала со стандартом выборки устанавливается критерием Стьюдента:

$$\beta = t(p_0, n)\, s_n, \quad s_n = \frac{s}{\sqrt{n}}, \tag{9}$$

где коэффициенты *) Стьюдента $t(p_0, n)$ представлены в таблице 23.

Таблица 23

Коэффициенты Стьюдента $t(p_0, n)$

$n-1$ \ p_0	0,5	0,7	0,8	0,9	0,95	0,98	0,99	0,995	0,998	0,999
1	1,0	2,0	3,1	6,3	13	32	64	127	318	637
2	0,8	1,3	1,9	2,9	4,3	7,0	9,9	14	22	32
3	0,8	1,3	1,6	2,4	3,2	4,5	5,8	7,5	10	13
4	0,7	1,2	1,5	2,1	2,8	3,7	4,6	5,6	7,2	8,6
5	0,7	1,2	1,5	2,0	2,6	3,4	4,0	4,8	5,9	6,9
6	0,7	1,1	1,4	1,9	2,4	3,1	3,7	4,3	5,2	6,0
7	0,7	1,1	1,4	1,9	2,4	3,0	3,5	4,0	4,8	5,4
8	0,7	1,1	1,4	1,9	2,3	2,9	3,4	3,8	4,5	5,0
9	0,7	1,1	1,4	1,8	2,3	2,8	3,3	3,7	4,3	4,8
10	0,7	1,1	1,4	1,8	2,2	2,8	3,2	3,6	4,1	4,6
15	0,7	1,1	1,3	1,8	2,1	2,6	2,9	3,3	3,7	4,1
20	0,7	1,1	1,3	1,7	2,1	2,5	2,8	3,2	3,5	3,8
30	0,7	1,1	1,3	1,7	2,0	2,5	2,8	3,0	3,4	3,7
60	0,7	1,0	1,3	1,7	2,0	2,4	2,7	2,9	3,2	3,5
∞	0,7	1,0	1,3	1,6	2,0	2,3	2,6	2,8	3,1	3,3
Критерий Чебышева	1,4	1,8	2,2	3,2	4,5	7,1	10	14	22	32

Очевидно, при больших n с хорошей точностью выполняется $\sigma_n \approx s_n$. Поэтому при $n \to \infty$ критерий Стьюдента переходит в формулу (6); выше отмечалось, что этой формуле соответствует строка $n = \infty$ таблицы 23. Однако при малых n доверительный интервал (8) оказывается много шире, чем по критерию (6).

Пример 1. Выбрано $p_0 = 0{,}99$ и выполнено 3 измерения; по таблице 23 доверительный интервал равен $\beta = 9{,}9s/\sqrt{3} = 5{,}7s$.

*) Их называют также *квантилями Стьюдента*.

К сожалению, не все физики и инженеры знакомы с понятием доверительного интервала и критерием Стьюдента. Нередко встречаются экспериментальные работы, в которых при малом числе измерений пользуются критерием (6) или даже считают, что значение s_n является погрешностью величины $\bar{x}$, и вдобавок оценивают дисперсию по формуле (7).

Для приведенного выше примера при первой ошибке был бы дан ответ $\beta = 1{,}5s$, при второй — $\beta = 0{,}6s$, а при третьей — $\beta = 0{,}7s$, что сильно отличается от правильного значения.

Замечание 2. Зачастую одна и та же величина x измерена в разных лабораториях на разном оборудовании. Тогда следует найти среднее и стандарт по формулам (2) и (8), где суммирование проводится по всем измерениям во всех лабораториях, и определить доверительный интервал по критерию Стьюдента.

Нередко при этом суммарный стандарт s оказывается больше, чем стандарты s_j, определенные по данным отдельных лабораторий. Это естественно. Каждая лаборатория делает при измерениях систематические ошибки, и часть систематических ошибок в разных лабораториях совпадает, а часть — различается. При совместной обработке различающиеся систематические ошибки переходят в разряд случайных, увеличивая стандарт.

Значит, при совместной обработке разнотипных измерений обычно систематическая ошибка значения $\bar{x}$ будет меньше, а случайная — больше. Но случайную ошибку можно сколь угодно уменьшить, увеличивая число измерений. Поэтому такой способ позволяет получить окончательный результат с большей точностью.

Замечание 3. Если в разных лабораториях используется оборудование разного класса точности, то при такой совместной обработке надо суммировать с весами ρ_i:

$$\bar{x} = \frac{1}{R}\sum_{i=1}^{n} \rho_i x_i, \quad s^2 = \frac{n}{(n-1)\,R}\sum_{i=1}^{n} \rho_i (x_i - \bar{x})^2, \quad R = \sum_{i=1}^{n} \rho_i, \quad (10)$$

где ρ_i относятся, как квадраты точности приборов.

Произвольное распределение. Чаще всего число измерений n невелико и заранее неясно, можно ли считать распределение η_n нормальным и пользоваться приведенными выше критериями.

Для произвольного распределения $\rho(x)$ справедливо неравенство Чебышева

$$p\{|x - \mathrm{M}\xi| \geqslant \beta\} \leqslant \frac{\mathrm{D}\xi}{\beta^2}.$$

Отсюда можно оценить доверительный интервал:

$$\beta \leqslant \frac{1}{\sqrt{1-p_0}}\sigma_n, \quad \sigma_n = \sqrt{\frac{\mathrm{D}\xi}{n}}. \quad (11)$$

Коэффициент $(1-p_0)^{-1/2}$ в этой оценке приведен в дополнительной строке таблицы 23.

Из таблицы видно, что если в качестве доверительной вероятности принять $p_0=0{,}95$, то *для произвольного закона распределения с известной дисперсией доверительный интервал не превышает* $5\sigma_n$. Для симметричного одновершинного распределения аналогичные оценки показывают, что доверительный интервал не превышает $3\sigma_n$; напомним, что для нормального распределения он равен $2\sigma_n$ (при выбранном $p_0=0{,}95$).

Разумеется, если вместо σ_n используют найденное по тем же измерениям значение s_n, то надо строить критерий, аналогичный критерию Стьюдента. Оценки при этом будут существенно хуже приведенных.

Проверка нормальности распределения. Из сравнения критериев (6) и (11) видно, что даже при невысокой доверительной вероятности $p_0 \leqslant 0{,}9$ оценки доверительного интервала при произвольном распределении вдвое хуже, чем при нормальном. Чем ближе p_0 к единице, тем хуже соотношение этих оценок. Поэтому целесообразно проверять, существенно ли отличается распределение $\rho(x)$ от нормального.

Распространенный способ проверки — исследование так называемых *центральных моментов* распределения:

$$m_k = \int_{-\infty}^{+\infty} (x - \mathsf{M}\xi)^k \rho(x)\, dx, \qquad k = 1, 2, \ldots \tag{12}$$

Два первых момента, по определению, равны $m_1=0$, $m_2=\mathsf{D}\xi=\sigma^2$. Для нормального распределения два следующих момента равны $m_3=0$, $m_4=3\sigma^4$. Обычно ограничиваются этими моментами. Вычисляют их фактические значения по проведенным измерениям и проверяют, согласуются ли они со значениями, соответствующими нормальному распределению.

Удобно вычислять не сами моменты, а составленные из них безразмерные комбинации — *асимметрию* $\mathsf{A}=\sigma^{-3}m_3$ и *эксцесс* $\mathsf{E}=\sigma^{-4}m_4-3$; для нормального распределения они обращаются в нуль. Аналогично дисперсии, вычислим их по несмещенным оценкам:

$$\mathsf{A} \approx \frac{1}{s^3(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^3, \quad \mathsf{E} = \frac{1}{s^4(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^4 - 3, \tag{13}$$

где s определяется формулой (8). Собственные дисперсии этих величин известны и зависят только от числа измерений:

$$\mathsf{D}(\mathsf{A}) = \frac{6(n-1)}{(n+1)(n+3)}, \quad \mathsf{D}(\mathsf{E}) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}, \tag{14}$$

причем собственное распределение A является симметричным. Поэтому, если выполняются соотношения

$$|A| \leqslant 3\sqrt{D(A)}, \quad |E| \leqslant 5\sqrt{D(E)}, \tag{15}$$

то по критерию Чебышева (11) отличие A и E от нуля недостоверно, так что можно принять гипотезу о нормальности распределения $\rho(x)$.

Формулы (13) — (15) непосредственно относятся к распределению единичного измерения. На самом деле надо проверить, нормально ли распределение среднеарифметического η_n при выбранном n. Для этого делают большое число измерений $N = rn$, разбивают их на r групп по n измерений в каждой и среднее значение в каждой группе $\bar{x}$ рассматривают как единичное измерение. Тогда проверка выполняется по формулам (13)—(15), где вместо n надо подставить r.

Разумеется, такую тщательную проверку проводят не в каждой измеряемой точке, а лишь во время отработки методики эксперимента.

Замечание 4. Аналогично проверяют любые естественнонаучные гипотезы. Производят большое число экспериментов и выясняют, нет ли среди них событий, маловероятных с точки зрения этой гипотезы. Если найдутся такие события, то гипотезу отвергают, если нет — условно принимают.

Выбор n. За счет увеличения числа измерений n можно неограниченно уменьшать доверительный интервал. Однако систематическая ошибка β_0 при этом не уменьшается, так что суммарная ошибка все равно будет больше β_0. Поэтому целесообразно выбрать n так, чтобы ширина доверительного интервала составляла 50—100% β_0. Дальнейшее увеличение числа измерений бессмысленно.

Чтобы найти удовлетворяющее этому требованию n, надо отдельные точки измерить достаточное число раз, вычислить стандарт s, убедиться в нормальности распределения η_n и на основании критерия Стьюдента (9) подобрать такое n, чтобы выполнялось неравенство

$$t(p_0, n) \leqslant \frac{\sqrt{n}}{s}\beta_0, \tag{16}$$

где коэффициенты Стьюдента $t(p_0, n)$ даются таблицей 23.

Из таблицы 23 видно, что при $n = 2$ доверительный интервал чересчур велик, так что следует производить не менее 3—4 измерений. При дальнейшем увеличении n коэффициенты Стьюдента убывают слабо и доверительный интервал s_n сужается почти пропорционально $n^{-1/2}$, т. е. довольно медленно. Поэтому обычно считают нецелесообразным брать $n > 5 - 10$, так как возрастающая трудоемкость эксперимента не оправдывается достигаемой точностью.

Пример 2. Отношение систематической ошибки к стандарту выборки оказалось $\beta_0/s = 0{,}8$, и принята доверительная вероятность $p_0 = 0{,}95$. Возьмем соответствующий столбец таблицы 23 и будем перебирать по очереди $n = 2, 3, \ldots$, пока не получим $t(p_0, n)/\sqrt{n} \leqslant 0{,}8$; этому условию удовлетворяет $n = 9$.

Обнаружение грубых ошибок. Отличить грубую ошибку от случайной не всегда легко. Если число измерений мало, то широк доверительный интервал и даже значительные отклонения от среднего в него укладываются. Если же n велико, то возрастает вероятность того, что хотя бы одно измерение сильно отклонится от среднего случайно, т. е. на законном основании.

Пусть сделано n измерений и вычислены среднее $\bar{x}$ и стандарт s. Чтобы с вероятностью p_1 ни одно из этих измерений не отличалось от $\mathsf{M}\xi$ более чем на некоторое δ, каждое измерение должно оставаться в указанных пределах с вероятностью $\sqrt[n]{p_1}$, т. е. должно выполняться условие

$$p\{|x_i - \mathsf{M}\xi| \leqslant \delta\} = \sqrt[n]{p_1}. \tag{17}$$

Предполагая, что ξ имеет нормальное распределение, сравнивая (17) с критерием Стьюдента (9) и учитывая, что величина s вычислена по всей выборке, а применяется к отклонению единичного измерения, получим

$$\delta = t\left(\sqrt[n]{p_1},\ n\right) s. \tag{18}$$

Вместо неизвестной величины $\mathsf{M}\xi$ мы вынуждены подставлять в (17) величину $\bar{x}$, имеющую доверительный интервал $\beta = t(p_0, n)\, s/\sqrt{n}$. Сравним неравенства

$$|x_i - \mathsf{M}\xi| \lesssim \delta, \quad |\bar{x} - \mathsf{M}\xi| \lesssim \beta;$$

поскольку они носят вероятностный характер, то к ним надо применять не неравенство треугольника, а суммирование квадратов, что дает

$$|x_i - \bar{x}| \lesssim \sqrt{\delta^2 + \beta^2}. \tag{19}$$

Подставляя сюда найденные δ и β, можно сделать следующий вывод:

Если для всех измеренных величин выполняется оценка

$$|x_i - \bar{x}| \leqslant s\left[t^2\left(\sqrt[n]{p_1},\ n\right) + \frac{1}{n}\, t^2(p_0, n)\right]^{1/2}, \tag{20}$$

то нет оснований считать одну из них грубо ошибочной. Если какое-либо измерение не укладывается в пределы (20), *то его можно считать грубо ошибочным и отбрасывать.*

Общепринятых критериев для выбора вероятности p_1 нет; естественно полагать $p_1 = p_0$.

Пример 3. Пусть проведено $n=10$ измерений и выбрано $p_1=p_0=0{,}9$. Тогда $p_1^{1/n}=0{,}99$ и вычисления по формуле (20) при помощи таблицы 23 дают $|x_i-\bar{x}|\leqslant 3{,}3s$. Если при той же вероятности p_0 взять $n=100$, то получим условие $|x_i-\bar{x}|\leqslant 4{,}8s$.

3. **Сравнение величин.** Сначала рассмотрим задачу сравнения величины x, измеряемой в эксперименте, с константой a. Величину x можно определить лишь приближенно, вычисляя среднее $\bar{x}$ по n измерениям. Надо узнать, выполняется ли соотношение $\mathsf{M}\xi\geqslant a$. В этом случае ставят две задачи, прямую и обратную:

а) *по известной величине $\bar{x}$ найти константу a, которую $\mathsf{M}\xi$ превосходит с заданной вероятностью p_0;*

б) *найти вероятность p_0 того, что* $\mathsf{M}\xi\geqslant a$, где a — заданная константа.

Очевидно, если $\bar{x}<a$, то вероятность того, что $\mathsf{M}\xi\geqslant a$, меньше $^1/_2$. Этот случай не представляет интереса, и далее будем считать, что $\bar{x}\geqslant a$ и $p_0\geqslant {}^1/_2$.

Задача сводится к задачам, разобранным в п. 2. Пусть по n измерениям определены $\bar{x}$ и его стандарт s_n:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\qquad s_n^2=\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{21}$$

Число измерений будем считать не очень малым, так что $\bar{x}$ есть случайная величина с нормальным распределением. Тогда из критерия Стьюдента (9) при учете симметрии нормального распределения следует, что для произвольно выбранной вероятности p_1 выполняется условие

$$p\{\mathsf{M}\xi-\bar{x}\geqslant -t(p_1,\ n)\,s_n\}=\frac{1}{2}(1+p_1). \tag{22}$$

Полагая $p_0={}^1/_2(1+p_1)$, перепишем это выражение в следующем виде:

$$p\{\mathsf{M}\xi\geqslant a\}=p_0,\qquad a=\bar{x}-t(2p_0-1,\ n)\,s_n, \tag{23}$$

где $t(p_1, n)$ — заданные в таблице 23 коэффициенты Стьюдента. Тем самым, прямая задача решена: найдена константа a, которую с вероятностью p_0 превышает $\mathsf{M}\xi$.

Обратная задача решается при помощи прямой. Перепишем формулы (23) следующим образом:

$$t(2p_0-1,\ n)=\frac{\bar{x}-a}{s_n},\qquad p\{\mathsf{M}\xi\geqslant a\}=p_0. \tag{24}$$

Это значит, что надо вычислить t по известным значениям a, $\bar{x}$, s_n, выбрать в таблице 23 строку с данным n и найти по величине t соответствующее значение p_1. Оно определяет искомую вероятность $p_0={}^1/_2(1+p_1)$.

Две случайные величины. Часто требуется установить влияние некоторого фактора на исследуемую величину — например, увеличивает ли (и насколько) прочность металла определенная присадка. Для этого надо измерить прочность исходного металла x и прочность легированного металла y и сравнить эти две величины, т. е. найти $z = y - x$.

Сравниваемые величины являются случайными; так, свойства металла определенной марки меняются от плавки к плавке, поскольку сырье и режим плавки не строго одинаковы. Обозначим эти величины через ξ и η. Величина исследуемого эффекта равна $\mathrm{M}\zeta = \mathrm{M}\eta - \mathrm{M}\xi$, и требуется определить, выполняется ли условие $\mathrm{M}\zeta \geqslant a$.

Таким образом, задача свелась к сравнению случайной величины ζ с константой a, разобранному выше. Прямая и обратная задачи сравнения в этом случае формулируются следующим образом:

а) *по результатам измерений x_i и y_j найти константу a, которую $\mathrm{M}\zeta$ превосходит с заданной вероятностью p_0* (т. е. оценить величину исследуемого эффекта);

б) *определить вероятность p_0 того, что* $\mathrm{M}\zeta \geqslant a$, где a — желательная величина эффекта; при $a = 0$ это означает, что надо определить вероятность, с которой $\mathrm{M}\eta \geqslant \mathrm{M}\xi$.

Для решения этих задач надо вычислить $\bar{z}$ и дисперсию этой величины. Рассмотрим два способа их нахождения.

Независимые измерения. Измерим величину x в n экспериментах, а величину y — в m экспериментах, независимых от первых n экспериментов. Вычислим средние значения по обычным формулам:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{m}\sum_{j=1}^{m} y_j. \tag{25}$$

Эти средние сами являются случайными величинами, причем их стандарты (не путать со стандартами единичных измерений!) приближенно определяются несмещенными оценками:

$$s_{nx}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_{my}^2 = \frac{1}{m(m-1)}\sum_{j=1}^{m}(y_j - \bar{y})^2. \tag{26}$$

Поскольку эксперименты независимы, то случайные величины $\bar{x}$ и $\bar{y}$ также независимы, так что при вычислении $\bar{z}$ их математические ожидания вычитаются, а дисперсии складываются:

$$\bar{z} \approx \mathrm{M}\zeta = \mathrm{M}\eta - \mathrm{M}\xi \approx \bar{y} - \bar{x}, \tag{27}$$

$$\mathrm{D}\bar{z} \approx s_{\bar{z}}^2 = s_{nx}^2 + s_{my}^2. \tag{28а}$$

Несколько более точная оценка дисперсии такова:

$$s_{\bar{z}}^2 = \frac{1}{(n+m-2)}\left(\frac{1}{n}+\frac{1}{m}\right)\left[\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{j=1}^{m}(y_j-\bar{y})^2\right]. \quad (28б)$$

Таким образом, $\bar{z}$ и ее дисперсия найдены, и дальнейшие вычисления производятся по формулам (23) или (24).

Согласованные измерения. Более высокую точность дает другой способ обработки, когда в каждом из n экспериментов одновременно измеряют x и y. Например, после выпуска половины плавки в оставшийся в печи металл добавляют присадку, а затем сравнивают образцы металла из каждой половины плавки.

При этом, по существу, в каждом эксперименте измеряют сразу значение $z_i = y_i - x_i$ одной случайной величины ζ, которую надо сравнить с константой a. Обработка измерений тогда производится по формулам (21)—(24), где вместо x надо всюду подставить z.

Дисперсия при согласованных измерениях будет меньше, чем при независимых, поскольку она обусловлена только частью случайных факторов: те факторы, которые согласованно меняют ξ и η, не влияют на разброс их разности. Поэтому такой способ позволяет получить более достоверные выводы.

Пример. Любопытной иллюстрацией сравнения величин является определение победителя в тех видах спорта, где судейство ведется «на глазок» — гимнастика, фигурное катание и т. д.

Таблица 24

Судейские оценки в баллах

Судья / Всадница	1	2	3	4	5	Среднее	Место
Линзенхофф	254	257	232	245	241	245,8	I
Петушкова	230	243	220	257	235	237,0	II

В таблице 24 приведен протокол соревнований по выездке на Олимпийских играх 1972 г. Видно, что разброс судейских оценок велик, причем ни одну оценку нельзя признать грубо ошибочной и откинуть. На первый взгляд кажется, что достоверность определения победителя невелика.

Рассчитаем, насколько правильно определен победитель, т. е. какова вероятность события $M\zeta > 0$. Поскольку оценки обеим всадницам выставлялись одними и теми же судьями, можно воспользоваться способом согласованных измерений. По таблице 24 вычисляем $\bar{z} = +8{,}8$ и $s_{nz} = 5{,}9$; подставляя в формулу (24) эти значения и $a = 0$, получим $t(p_1, n) = 1{,}5$. Выбирая в таблице 23

строку $n=5$, находим, что этому значению t соответствует $p_1 = 2p_0 - 1 = 0{,}8$. Отсюда $p_0 = 0{,}9$, т. е. с вероятностью 90% золотая медаль присуждена правильно.

Сравнение по способу независимых измерений даст несколько худшую оценку, поскольку оно не использует информацию о том, что оценки выставляли одни и те же судьи.

Сравнение дисперсий. Пусть требуется сравнить две методики эксперимента. Очевидно, точнее та методика, у которой дисперсия σ^2 единичного измерения меньше (разумеется, если при этом не увеличивается систематическая ошибка). Значит, надо установить, выполняется ли неравенство $\sigma_x > \sigma_y$.

О дисперсиях единичных измерений судят по стандартам выборок

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_j - \bar{y})^2, \tag{29}$$

вычисленным соответственно по n и m измерениям. Эти стандарты сами являются случайными величинами. Однако сравнивать их на основании критерия Стьюдента нельзя, поскольку распределение s не гауссово. Нетрудно видеть, что оно является асимметричным: значения $s<0$ невозможны, а сколь угодно большие $s>0$ возможны.

Дисперсии сравнивают по критерию Фишера. Если

$$\frac{s_x^2}{s_y^2} > F(n,\ m;\ p_0), \tag{30}$$

то с вероятностью p_0 первая дисперсия больше второй. Коэффициенты Фишера *) для случаев $n=m$, $n=\infty$, $m=\infty$ приведены в таблице 25. При малых n, m эти коэффициенты довольно велики; поэтому различие дисперсий можно установить только в том случае, если это различие велико или велико число экспериментов.

Замечание. Критерий Фишера позволяет также найти отношение дисперсий. Если выполнено неравенство

$$\frac{s_x^2}{s_y^2} > aF(n,\ m;\ p_0), \tag{31}$$

то с вероятностью p_0 первая дисперсия в a раз больше второй.

Методы, изложенные в пп. 2 и 3, применимы не только к измерениям непрерывных величин, но и для суждения об очень большой партии объектов (*генеральной совокупности*) по небольшой случайной выборке из n объектов. Эти формулы и критерии применяются в статистике, социологии, выборочной оценке больших партий товара и т. д. В статистике и социологии законы распределения величин нередко сильно отличаются от нормального, и выяснение закона распределения играет там большую роль.

*) Их называют также *квантилями* Фишера.

Таблица 25

Коэффициенты Фишера $F(n, m; p_0)$

p_0	0,8	0,95	0,975	0,990	0,995	0,999
$n-1=m-1$	Случай $n=m$					
1	9,5	161	648	4052	16 211	$4{,}1\times10^5$
2	4,0	19	39	99	199	999
3	2,9	9,3	15	29	47	141
4	2,5	6,4	9,6	16	23	53
5	2,2	5,0	7,2	11	15	30
6	2,1	4,3	5,8	8,5	11	20
12	1,7	2,7	3,3	4,2	4,9	7,0
24	1,4	2,0	2,3	2,7	3,0	3,7
∞	1,0	1,0	1,0	1,0	1,0	1,0
$m-1$	Случай $n=\infty$					
1	16	254	1018	6366	25 465	$6{,}4\times10^5$
2	4,5	19	40	100	200	1000
3	3,0	8,5	14	26	42	124
4	2,4	5,6	8,3	13	19	44
5	2,1	4,4	6,0	9,0	12	24
6	2,0	3,7	4,9	6,9	8,9	16
12	1,5	2,3	2,7	3,4	3,9	5,4
24	1,3	1,7	1,9	2,2	2,4	3,0
∞	1,0	1,0	1,0	1,0	1,0	1,0
$n-1$	Случай $m=\infty$					
1	1,6	3,8	5,0	6,6	7,9	11
2	1,6	3,0	3,7	4,6	5,3	6,9
3	1,6	2,6	3,1	3,8	4,3	5,4
4	1,5	2,4	2,8	3,3	3,7	4,6
5	1,5	2,2	2,6	3,0	3,4	4,1
6	1,4	2,1	2,4	2,8	3,1	3,7
12	1,3	1,8	1,9	2,2	2,4	2,7
24	1,2	1,5	1,6	1,8	1,9	2,1
∞	1,0	1,0	1,0	1,0	1,0	1,0

4. Нахождение стохастической зависимости. Пусть требуется исследовать зависимость $z(x)$, причем обе величины z и x измеряются в одних и тех же экспериментах. Для этого проводят серию экспериментов при разных значениях x, стараясь сохранить прочие условия эксперимента неизменными.

Измерение каждой величины содержит случайные ошибки (систематические ошибки здесь рассматривать не будем); следовательно, эти величины являются случайными. Закономерная

связь случайных величин называется *стохастической* *). Будем рассматривать две задачи:

а) установить, существует ли (с определенной вероятностью) зависимость z от x или величина z от x не зависит;

б) если зависимость существует, описать ее количественно.

Первую задачу называют *дисперсионным анализом*, а если рассматривается функция многих переменных $z(x, y, \ldots)$ — то *многофакторным* дисперсионным анализом. Вторую задачу называют *анализом регрессии*. Если случайные ошибки велики, то они могут маскировать искомую зависимость и выявить ее бывает нелегко.

Без ограничения общности можно считать, что величина x измеряется точно. В самом деле, если z от x не зависит, то ошибка δx ни на что не влияет. Если же зависимость существует, то ошибка δx эквивалентна дополнительной ошибке зависимой переменной $\delta z = (dz/dx)\,\delta x$.

Таким образом, достаточно рассмотреть случайную величину $\zeta(x)$, зависящую от x как от параметра. Математическое ожидание этой величины $\mathbf{M}\zeta(x) \equiv z(x)$ зависит от x; эта зависимость является искомой и называется *законом регрессии*.

Дисперсионный анализ. Проведем при каждом значении x_i небольшую серию измерений и определим z_{ij} $(1 \leqslant j \leqslant n_i)$. Рассмотрим два способа обработки этих данных, позволяющих исследовать, имеется ли *значимая* (т. е. с принятой доверительной вероятностью) зависимость z от x.

При первом способе вычисляют стандарты выборки единичного измерения по каждой серии отдельно и по всей совокупности измерений:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2, \quad s^2 = \frac{1}{N-1} \sum_{i,\,j} (z_{ij} - \bar{z})^2, \tag{32}$$

где $N = \sum_i n_i$ — полное число измерений, а

$$\bar{z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}, \quad \bar{z} = \frac{1}{N} \sum_{i,\,j} z_{ij} \tag{33}$$

являются средними значениями соответственно по каждой серии и по всей совокупности измерений.

Сравним дисперсию совокупности измерений $\sigma^2 \approx s^2$ с дисперсиями отдельных серий $\sigma_i^2 \approx s_i^2$. Если окажется, что при выбранном уровне достоверности p_0 можно считать $\sigma > \sigma_i$ для всех i, то зависимость z от x имеется. Если достоверного превышения

*) С такой связью мы уже встречались в стохастических задачах нахождения корня уравнения (гл. V, § 2, п. 4) и минимума функции (гл. VII, § 1, п. 4).

нет, то зависимость не поддается обнаружению (при данной точности эксперимента и принятом способе обработки).

Дисперсии сравнивают по критерию Фишера (30). Поскольку стандарт s определен по полному числу измерений N, которое обычно достаточно велико, то почти всегда можно пользоваться коэффициентами Фишера $F(\infty, m; p_0)$, приведенными в таблице 25.

Второй способ анализа заключается в сравнении средних $\bar{z}_i$ при разных значениях x_i между собой. Величины $\bar{z}_i$ являются случайными и независимыми, причем их собственные стандарты выборки равны

$$s_{ni} = s_i/\sqrt{n_i}.$$

Поэтому их сравнивают по схеме независимых измерений, описанной в п. 3. Если различия $\bar{z}_i$ значимы, т. е. превышают доверительный интервал, то факт зависимости z от x установлен; если различия всех $\bar{z}_i$ незначимы, то зависимость не поддается обнаружению.

Многофакторный анализ имеет некоторые особенности. Величину $z(x, y)$ целесообразно измерять в узлах прямоугольной сетки (x_i, y_k), чтобы удобнее было исследовать зависимость от одного аргумента, фиксируя другой аргумент. Проводить серию измерений в каждом узле многомерной сетки слишком трудоемко. Достаточно провести серии измерений в нескольких узлах сетки, чтобы оценить дисперсию единичного измерения; в остальных узлах можно ограничиться однократными измерениями. Дисперсионный анализ при этом проводят по первому способу.

Замечание 1. Если измерений много, то в обоих способах отдельные измерения или серии могут с заметной вероятностью довольно сильно отклониться от своего математического ожидания. Это надо учитывать, выбирая доверительную вероятность p_0 достаточно близкой к 1 (как это делалось в п.2 при установлении пределов, отделяющих допустимые случайные ошибки от грубых).

Анализ регрессии. Пусть дисперсионный анализ указал, что зависимость z от x есть. Как ее количественно описать?

Для этого аппроксимируем искомую зависимость некоторой функцией $z(x) \approx f(x, \boldsymbol{a})$, $\boldsymbol{a} = \{a_1, a_2, \ldots, a_m\}$. Оптимальные значения параметров a_k найдем методом наименьших квадратов, решая задачу

$$\sum_{i=1}^{N} w(x_i)\,[z_i - f(x_i, \boldsymbol{a})]^2 = \min, \tag{34}$$

где $w(x_i)$ — веса измерений, выбираемые обратно пропорционально квадрату погрешности измерения в данной точке (т. е. $w_i \sim (\mathrm{D}z_i)^{-1}$). Эта задача была разобрана в главе II, § 2. Остановимся здесь лишь на тех особенностях, которые вызваны присутствием больших случайных ошибок.

Вид $f(x, a)$ подбирают либо из теоретических соображений о природе зависимости $z(x)$, либо формально, сравнивая график $z(x)$ с графиками известных функций. Если формула подобрана из теоретических соображений и правильно (с точки зрения теории) передает асимптотику $z(x)$, то обычно она позволяет не только неплохо аппроксимировать совокупность экспериментальных данных, но и экстраполировать найденную зависимость на другие диапазоны значений x. Формально подобранная функция $f(x, a)$ может удовлетворительно описывать эксперимент, но редко пригодна для экстраполяции.

Проще всего решить задачу (34), если $f(x, a)$ является алгебраическим многочленом $\sum a_k x^k$. Однако такой формальный выбор функции редко оказывается удовлетворительным. Обычно хорошие формулы зависят от параметров нелинейно (*трансцедентная* регрессия). Трансцедентную регрессию наиболее удобно строить, подбирая такую *выравнивающую* замену переменных $Z(z)$, $X(x)$, чтобы зависимость $Z(X)$ была почти линейной (см. гл. II, § 1, п. 8). Тогда ее нетрудно аппроксимировать алгебраическим многочленом: $Z \approx P(X, a)$.

Выравнивающую замену переменных ищут, используя теоретические соображения и учитывая асимптотику $z(x)$. Дальше будем считать, что такая замена уже сделана.

Замечание 2. При переходе к новым переменным задача метода наименьших квадратов (34) принимает вид

$$\sum_{i=1}^{N} W_i [Z_i - P(X_i, a)]^2 = \min, \tag{35}$$

где новые веса связаны с исходными соотношениями

$$W_i = \left(\frac{dZ}{dz}\right)_i^{-2} w_i. \tag{36}$$

Поэтому, даже если в исходной постановке (34) все измерения имели одинаковую точность, так что $w_i \equiv 1$, то для выравнивающих переменных веса не будут одинаковыми.

Корреляционный анализ. Надо проверить, действительно ли замена переменных была выравнивающей, т. е. близка ли зависимость $Z(X)$ к линейной. Это можно сделать, вычислив коэффициент парной корреляции

$$\rho = \frac{\mathsf{M}[(\xi - \mathsf{M}\xi)(\zeta - \mathsf{M}\zeta)]}{\sqrt{\mathsf{D}\xi \cdot \mathsf{D}\zeta}} \approx \frac{\sum_{i=1}^{N} W_i (X_i - \bar{X}_i)(Z_i - \bar{Z}_i)}{\sqrt{\sum_{i=1}^{N} W_i (X_i - \bar{X}_i)^2 \cdot \sum_{i=1}^{N} W_i (Z_i - \bar{Z})^2}}. \tag{37}$$

Нетрудно показать, что всегда выполняется соотношение $|\rho| \leqslant 1$.

Если зависимость $Z(X)$ строго линейная (и не содержит случайных ошибок), то $\rho=+1$ или $\rho=-1$ в зависимости от знака наклона прямой. Чем меньше $|\rho|$, тем менее зависимость $Z(X)$ похожа на линейную. Поэтому, если $|\rho|\approx 1$, а число измерений N достаточно велико, то выравнивающие переменные выбраны удовлетворительно.

Подобные заключения о характере зависимости по коэффициентам корреляции называют *корреляционным анализом*.

При корреляционном анализе не требуется, чтобы в каждой точке x_i проводилась серия измерений. Достаточно в каждой точке сделать одно измерение, но зато взять побольше точек на исследуемой кривой, что часто делают в физических экспериментах.

Замечание 3. Существуют критерии близости $|\rho|$ к 1, позволяющие указать, является ли зависимость $Z(X)$ практически линейной. Мы на них не останавливаемся, поскольку далее будет рассмотрен выбор степени аппроксимирующего многочлена.

Замечание 4. Соотношение $|\rho|\approx 0$ указывает на отсутствие линейной зависимости $Z(X)$, но не означает отсутствия какой-либо зависимости. Так, если $Z=X^2$ на отрезке $-1\leqslant X\leqslant 1$, то $\rho=0$.

Оптимальная степень многочлена. Подставим в задачу (35) *аппроксимирующий многочлен степени* m:

$$P(X)=\sum_{k=0}^{m} a_k X^k. \tag{38}$$

Тогда оптимальные значения параметров a_k удовлетворяют системе линейных уравнений (2.43):

$$\begin{gathered}\sum_{l=0}^{m} A_{kl}a_l=B_k,\quad 0\leqslant k\leqslant m,\\ A_{kl}=\sum_{i=1}^{N} W_i X_i^{k+l},\quad B_k=\sum_{i=1}^{N} W_i Z_i X_i^k,\end{gathered} \tag{39}$$

и найти их нетрудно. Но как выбрать степень многочлена?

Для ответа на этот вопрос вернемся к исходным переменным и вычислим дисперсию аппроксимационной формулы с найденными коэффициентами. Несмещенная оценка этой дисперсии такова *):

$$\mathsf{D}_m=\mathsf{D}[f(x,\ a_0,\ a_1,\ \ldots,\ a_m)]\approx$$

$$\approx\frac{N}{N-m-1}\sum_{i=1}^{N} w_i[z_i-f(x_i,\ \boldsymbol{a})]^2\Big/\sum_{i=1}^{N} w_i. \tag{40}$$

*) В формулах типа (32) имелся делитель вида $N-1$; он связан с тем, что при вычислении стандарта выборки используется одна величина $\bar{z}$, определенная по той же выборке. В (40) используется $m+1$ коэффициентов a_k, определенных по выборке, поэтому появляется делитель $N-(m+1)$.

Очевидно, при увеличении степени многочлена m дисперсия (40) будет убывать: чем больше взято коэффициентов, тем точней можно аппроксимировать экспериментальные точки.

Сравним D_m с дисперсиями s_i^2 единичных измерений (32), определенными по небольшим сериям экспериментов хотя бы в нескольких точках x_i. Если $D_m > s_i^2$ для всех i, то погрешность аппроксимации больше погрешности, с которой измерены значения z_i. Надо увеличивать m до тех пор, пока отличие D_m от s_i^2 хотя бы для одного i не перестанет быть значимым по критерию Фишера (30). Наоборот, если $D_m < s_i^2$, то надо уменьшать m.

Если полученная таким образом оптимальная степень m_0 удовлетворяет условию $m_0 \ll N$, то выравнивающие переменные выбраны удачно; если $m_0 \sim N$, то следует подобрать другую замену переменных.

Замечание 5. Описанный способ нахождения оптимального числа параметров m_0 можно применять при произвольном виде функции $f(x, \boldsymbol{a})$; но сами коэффициенты в этом случае вычисляются не по формулам (39).

Точность коэффициентов. Коэффициенты a_k определяются по случайным величинам z_i и поэтому сами являются случайными величинами. Какие их значащие цифры достоверны, а какие можно отбросить?

На первую половину вопроса ответить нетрудно. Проведем математический эксперимент. Зная дисперсию единичных измерений s_i^2, искусственно внесем в величины z_i случайные ошибки δz_i, распределенные по нормальному закону с дисперсиями s_i^2 (это делается методами Монте-Карло), и вычислим соответствующие a_k.

Повторим эту процедуру многократно. Для каждого a_k получим набор случайных значений, по которому вычислим среднее $\bar{a}_k$ и стандарт $s(\bar{a}_k)$. Отсюда по критерию Стьюдента (9) найдем для $\bar{a}_k$ доверительный интервал и, тем самым, выясним, какие значащие цифры коэффициента достоверны.

Однако, вообще говоря, *недостоверные цифры коэффициента a_k нельзя отбрасывать*. Коэффициенты a_k можно округлять только все одновременно, меняя их на взаимно согласованные величины Δa_k. Для такого округления многочлен $P_m(X)$ представляют в виде линейной комбинации:

$$P_m(X) \equiv \sum_{k=0}^{m} a_k X^k = \sum_{l=0}^{m} \gamma_l Q_l(X), \tag{41}$$

где $Q_l(X)$ — алгебраические многочлены, ортогональные на системе точек X_i $(1 \leqslant i \leqslant N)$ с весами W_i *). Коэффициенты γ_l можно округлять независимо друг от друга в пределах их доверительных интервалов.

*) См. Приложение.

ЗАДАЧИ

1. Для примера, приведенного в таблице 24, определить достоверность победителя способом независимых измерений и сравнить результат с результатом, полученным способом согласованных измерений.

2. Учитывая, что веса в формулах (34) и (35) обратно пропорциональны дисперсиям, обосновать выражение веса (36).

3. Показать, что коэффициент парной корреляции (37) всегда по модулю не превышает 1.

4. Найти коэффициент парной корреляции (37) на отрезке $-1 \leqslant x \leqslant 1$ для а) линейной функции $z = ax + b$, б) квадратичной функции $z = ax^2 + b$; предполагается, что $\omega_i \equiv 1$.

5. Установить связь между коэффициентом парной корреляции (37) и наклоном сглаженной кривой (3.28).

ПРИЛОЖЕНИЕ

ОРТОГОНАЛЬНЫЕ МНОГОЧЛЕНЫ

Общие соотношения. Многочлены $P_n(x)$ называются ортогональными с весом $\rho(x)$ на отрезке $[a, b]$, если они удовлетворяют следующим соотношениям:

$$\int_a^b P_n(x)\, P_m(x)\, \rho(x)\, dx = N_n \delta_{nm}, \quad \rho(x) > 0;$$

по традиции наиболее употребительные многочлены нормируют не на единицу. Все корни ортогональных многочленов вещественные, простые и расположены на интервале (a, b); между каждой парой соседних корней многочлена $P_n(x)$ расположен один и только один корень многочлена $P_{n-1}(x)$.

Дальше ограничимся только так называемыми классическими ортогональными многочленами Якоби (и их частными случаями — многочленами Лежандра и Чебышева), Лагерра и Эрмита. Они удовлетворяют дифференциальному уравнению

$$\frac{d}{dx}\left[\sigma(x)\,\rho(x)\,\frac{dP_n}{dx}\right] + A_n \rho(x)\, P_n(x) = 0, \qquad a < x < b.$$

Их удобно вычислять либо по обобщенной формуле Родрига

$$P_n(x) = [B_n/\rho(x)]\,\frac{d^n}{dx^n}\,[\sigma^n(x)\,\rho(x)],$$

либо, зная два первых многочлена, при помощи рекуррентных соотношений

$$a_n P_{n+1}(x) = (b_n x - c_n)\, P_n(x) - d_n P_{n-1}(x).$$

Конкретный вид всех встречающихся в этих формулах функций и значения констант приведены в таблице 26.

Далее приведены некоторые многочлены с небольшими индексами n, их корни $\xi_i^{(n)}$ и соответствующие им веса $\gamma_i^{(n)}$ формулы Гаусса — Кристоффеля.

Многочлены Лежандра. $L_0(x) = 1$; $L_1(x) = x$; $L_2(x) = \frac{1}{2}(3x^2 - 1)$;

$L_3(x) = \frac{1}{2}(5x^3 - 3x)$; $L_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$; $L_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$.

$n = 1$; $\xi_1 = 0$; $\gamma_1 = 2$.

$n = 2$; $-\xi_1 = \xi_2 = \sqrt{1/3}$; $\gamma_1 = \gamma_2 = 1$.

$n = 3$; $-\xi_1 = \xi_3 = \sqrt{3/5}$, $\xi_2 = 0$; $\gamma_1 = \gamma_3 = 5/9$, $\gamma_2 = 8/9$.

Таблица 26

Ортогональные многочлены

Многочлен	Якоби	Лежандра	Чебышева первого рода	Чебышева второго рода	Лагерра	Эрмита
Обозначение	$P_n^{\alpha,\beta}(x)$	$L_n(x)$	$T_n(x)$	$U_n(x)$	$L_n^{(\alpha)}(x)$	$H_n(x)$
$a,\ b$	$-1,\ +1$	$-1,\ +1$	$-1,\ +1$	$-1,\ +1$	$0,\ +\infty$	$-\infty,\ +\infty$
$\rho(x)$	$(1-x)^{\alpha}(1+x)^{\beta};\ \alpha,\ \beta>-1$	1	$\frac{1}{\sqrt{1-x^2}}$	$\sqrt{1-x^2}$	$x^{\alpha}e^{-x};$ $\alpha>-1$	e^{-x^2}
$\sigma(x)$	$1-x^2$	$1-x^2$	$1-x^2$	$1-x^2$	x	1
N_n	$\frac{2^{\alpha+\beta+1}\Gamma(\alpha+n+1)\,\Gamma(\beta+n+1)}{n!\,(\alpha+\beta+2n+1)\,\Gamma(\alpha+\beta+n+1)}$	$\frac{2}{2n+1}$	$\frac{\pi}{2}$ при $n\neq 0$ π при $n=0$		$n!\times$ $\times\Gamma(\alpha+n+1)$	$\sqrt{\pi}\,2^n n!$
A_n	$n(n+\alpha+\beta+1)$	$n(n+1)$	n^2	$n(n+2)$	n	$2n$
B_n	$(-1)^n/(2n\cdot n!)$	$(-1)^n/(2n\cdot n!)$			$(-1)^n$	$(-1)^n$
a_n	$2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)$	$n+1$	1	1	1	1
b_n	$(2n+\alpha+\beta)(2n+\alpha+\beta+1)\times$ $\times(2n+\alpha+\beta+2)$	$2n+1$	2	2	1	2
c_n	$(\beta^2-\alpha^2)(2n+\alpha+\beta+1)$	0	0	0	$2n+\alpha+1$	0
d_n	$2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)$	n	1	1	$n(n+\alpha)$	$2n$

$n=4;\quad -\xi_1=\xi_4=\sqrt{(15+2\sqrt{30})/35},\quad -\xi_2=\xi_3=\sqrt{(15-2\sqrt{30})/35};$

$-\gamma_1=\gamma_4=(18-\sqrt{30})/36,\qquad \gamma_2=\gamma_3=(18+\sqrt{30})/36.$

$n=5;\quad -\xi_1=\xi_5=\sqrt{(35+2\sqrt{70})/63},\quad -\xi_2=\xi_4=\sqrt{(35-2\sqrt{70})/63},\quad \xi_3=0;$

$\gamma_1=\gamma_5=(322-13\sqrt{70})/900,\quad \gamma_2=\gamma_4=(322+13\sqrt{70})/900,$

$\gamma_3=(128/225).$

Многочлены Лагерра ($\alpha=0$). $L_0^{(0)}(x)=1;\quad L_1^{(0)}(x)=x-1;$

$L_2^{(0)}(x)=x^2-4x+2;\quad L_3^{(0)}(x)=x^3-9x^2+18x-6;$

$L_4^{(0)}(x)=x^4-16x^3+72x^2-96x+24.$

$n=1;\quad \xi_1=1;\quad \gamma_1=1.$

$n=2;\quad \xi_1=2-\sqrt{2},\ \xi_2=2+\sqrt{2};\ \gamma_1=(2+\sqrt{2})/4,\ \gamma_2=(2-\sqrt{2})/4.$

Многочлены Эрмита. $H_0(x)=1;\quad H_1(x)=2x;\quad H_2(x)=4x^2-2;$

$H_3(x)=8x^3-12x;\quad H_4(x)=16x^4-48x^2+12;$

$H_5(x)=32x^5-160x^3+120x.$

$n=1;\quad \xi_1=0;\quad \gamma_1=\sqrt{\pi}.$

$n=2;\quad -\xi_1=\xi_2=1/\sqrt{2};\quad \gamma_1=\gamma_2=\sqrt{\pi}/2.$

$n=3;\quad -\xi_1=\xi_3=\sqrt{3/2},\quad \xi_2=0;\ \gamma_1=\gamma_3=\sqrt{\pi}/6,\ \gamma_2=2\sqrt{\pi}/3.$

$n=4;\quad -\xi_1=\xi_4=\sqrt{(3+\sqrt{6})/2},\quad -\xi_2=\xi_3=\sqrt{(3-\sqrt{6})/2};$

$\gamma_1=\gamma_4=\sqrt{\pi}\,(3-\sqrt{6})/12,\quad \gamma_2=\gamma_3=\sqrt{\pi}\,(3+\sqrt{6})/12.$

$n=5;\quad -\xi_1=\xi_5=\sqrt{(5+\sqrt{10})/2},\quad -\xi_2=\xi_4=\sqrt{(5-\sqrt{10})/2},\quad \xi_3=0;$

$\gamma_1=\gamma_5=\sqrt{\pi}\,(7-2\sqrt{10})/60,\ \gamma_2=\gamma_4=\sqrt{\pi}\,(7+2\sqrt{10})/60,\ \gamma_3=(8\sqrt{\pi}/15).$

Многочлены Чебышева первого рода.

$T_0(x)=1;\quad T_1(x)=x;\quad T_2(x)=2x^2-1;$

$T_3(x)=4x^3-3x;\quad T_4(x)=8x^4-8x^2+1;\quad T_5(x)=16x^5-20x^3+5x;$

$T_6(x)=32x^6-48x^4+18x^2-1;$

$T_7(x)=64x^7-112x^5+56x^3-7x.$

$\xi_i^{(n)}=\cos[\pi\,(i-1/2)/n],\quad \gamma_i^{(n)}=\pi/n,\quad 1\leqslant i\leqslant n.$

Многочлены Чебышева второго рода.

$U_0(x)=1;\quad U_1(x)=2x;\quad U_2(x)=4x^2-1;\quad U_3(x)=8x^3-4x;$

$U_4(x)=16x^4-12x^2+1;\quad U_5(x)=32x^5-32x^3+6x.$

$$\xi_m^{(n)}=\cos\frac{\pi m}{n+1},\qquad 1\leqslant m\leqslant n;$$

$n=1;\quad \gamma_1=\pi/2.$

$n=2;\quad \gamma_1=\gamma_2=\pi/4.$

$n=3;\quad \gamma_1=\gamma_3=\pi/8,\quad \gamma_2=\pi/4.$

$n=4;\quad \gamma_1=\gamma_4=\pi\,(5-\sqrt{5})/40,\quad \gamma_2=\gamma_3=\pi\,(5+\sqrt{5})/40.$

$n=5;\quad \gamma_1=\gamma_5=\pi/24,\quad \gamma_2=\gamma_4=\pi/8,\quad \gamma_3=\pi/6.$

Многочлены на системе точек. Многочлены $P_n(x)$ называют ортонормированными на системе точек x_i с весами ρ_i $(1\leqslant i\leqslant N)$, если они

удовлетворяют соотношениям

$$\sum_{i=1}^{N} \rho_i P_n(x_i) P_m(x_i) = \delta_{nm}.$$

Систему таких многочленов можно построить, пользуясь рекуррентным соотношением

$$\lambda_n P_{n+1}(x) = (x - a_n) P_n(x) + b_n P_{n-1}(x),$$

где

$$a_n = \sum_{i=1}^{N} \rho_i x_i P_n^2(x_i), \qquad b_n = \sum_{i=1}^{N} \rho_i x_i P_n(x_i) P_{n-1}(x_i),$$

а λ_n определяется из условия нормировки. Для начала расчета по этим формулам надо положить

$$P_{-1}(x) \equiv 0, \qquad P_0(x) = \left(\sum_{i=1}^{N} \rho_i \right)^{-1/2}.$$

ЛИТЕРАТУРА

Учебники и монографии

1. Айнс Э. Л., Обыкновенные дифференциальные уравнения, Харьков, Гостехиздат, Украина, 1939.
2. Арсенин В. Я., Методы математической физики и специальные функции, «Наука», 1974.
3. Бахвалов Н. С., Численные методы, «Наука», т. I, 1975.
4. Березин И. С., Жидков Н. П., Методы вычислений, ч. I, «Наука», 1966, ч. II, Физматгиз, 1962.
5. Воеводин В. В., Численные методы алгебры; теория и алгорифмы, «Наука», 1966.
6. Вычислительные методы в физике плазмы. Сборник под ред. Б. Олдера, С. Фернбаха, М. Ротенберга, «Мир», 1974.
7. Гнеденко Б. В., Курс теории вероятностей, Гостехиздат, 1950.
8. Годунов С. К., Рябенький В. С., Введение в теорию разностных схем, Физматгиз, 1977.
9. Гончаров В. Л., Теория интерполирования и приближения функций, Гостехиздат, 1954.
10. Дьяченко В. Ф., Основные понятия вычислительной математики, «Наука», 1977.
11. Зельдович Я. Б., Райзер Ю. П., Физика ударных волн и высокотемпературных гидродинамических явлений, Физматгиз, 1963.
12. Ильин В. А., Позняк Э. Г., Аналитическая геометрия, «Наука», 1968.
13. Ильин В. А., Позняк Э. Г., Линейная алгебра, «Наука», 1974.
14. Ильин В. А., Позняк Э. Г., Основы математического анализа, «Наука», ч. I, 1971, ч. II, 1973.
15. Канторович Л. В., Акилов Г. П., Функциональный анализ, «Наука», 1977.
16. Карташев А. П., Рождественский Б. Л., Обыкновенные дифференциальные уравнения и основы вариационного исчисления, «Наука», 1976.
17. Коллатц Л., Задачи на собственные значения, «Наука», 1968.
18. Крылов В. И., Бобков В. В., Монастырный П. И., Вычислительные методы, «Наука», т. I, 1976, т. II, 1977.
19. Ландау Л. Д., Лифшиц Е. М,, Механика сплошных сред, Гостехиздат, 1954.
20. Люстерник Л. А., Соболев В. И., Элементы функционального анализа, «Наука», 1965.
21. Мак-Кракен Д., Дорн У., Численные методы и программирование на ФОРТРАНе, «Мир», 1969.
22. Марчук Г. И., Методы вычислительной математики, «Наука», 1977.

23. Михлин С. Г., Лекции по линейным интегральным уравнениям, Физматгиз, 1959.
24. Никифоров А. Ф., Уваров В. Б., Основы теории специальных функций, «Наука», 1974.
25. Никольский С. М., Квадратурные формулы, Физматгиз, 1974.
26. Пустыльник Е. И., Статистические методы анализа и обработки наблюдений, «Наука», 1968.
27. Рихтмайер Р. Д., Мортон К., Разностные методы решения краевых задач, «Мир», 1972.
28. Рождественский Б. Л., Яненко Н. Н., Системы квазилинейных уравнений и их применение к газовой динамике, «Наука», 1968.
29. Рябенький В. С., Филиппов А. Ф., Об устойчивости разностных уравнений, Гостехиздат, 1956.
30. Самарский А. А., Введение в теорию разностных схем, «Наука», 1971.
31. Самарский А. А., Теория разностных схем, «Наука», 1977.
32. Самарский А. А., Андреев В. Б., Разностные методы для решения эллиптических уравнений, «Наука», 1976.
33. Самарский А. А., Гулин А. В., Устойчивость разностных схем, «Наука», 1973.
34. Самарский А. А., Попов Ю. П., Разностные схемы газовой динамики, «Наука», 1975.
35. Свешников А. Г., Тихонов А. Н., Теория функций комплексной переменной, «Наука», 1974.
36. Седов Л. И., Методы подобия и размерности в механике, Гостехиздат, 1957.
37. Степанов В. В., Курс дифференциальных уравнений, Гостехиздат, 1953.
38. Соболь И. М., Численные методы Монте-Карло, «Наука», 1973.
39. Тихонов А. Н., Арсенин В. Я., Методы решения некорректных задач, «Наука», 1974.
40. Тихонов А. Н., Самарский А. А., Уравнения математической физики, изд. 4-е, «Наука», 1972.
41. Уилкинсон Дж. Х., Алгебраическая проблема собственных значений, «Наука», 1970.
42. Хемминг Р., Численные методы. Для научных работников и инженеров, «Наука», 1972.
43. Худсон Д., Статистика для физиков, «Мир», 1970.

Отдельные выпуски и статьи

44. Тихонов А. Н., Об устойчивых методах суммирования рядов Фурье, ДАН СССР, 1964, **156**, № 2, 268—271.
45. Филон; L. N. G. Filon, Proc. Roy. Soc., Edinb., 1928—1929, **49**.
46. Соболь И. М., Псевдослучайные числа для машины «Стрела». Теория вероятностей и ее применение, 1958, **3**, № 2, 205—211.
47. Роббинс, Монро; H. Robbins, S. Monro, A stochastic approximation method. Annals of Math. Stat., 1951, **22**, 400—407.
48. Вегстейн; J. H. Wegstein, Accelerating convergence of iterative processes. Comm. Assos. Comput. Mach., 1958, **1**, № 6, 9—13.
49. Мюллер; D. E. Muller, A method for solving algebraic equations using an automatic computer. Math. Tables Aids Comput., 1956, **10**, № 56, 208—215.
50. Хаусхолдер; A. S. Householder, Unitary triangularization of a nonsymmetric matrix. J. Assoc. Comput. Machinary, 1958, **5**, № 4, 339—342.
51. Гивенс; W. Givens, Numerical computation of characteristic values of a real symmetric matrix. Oak Ridge National Laboratory, ORNL-1574 (1954).

52. Голдстейн, Меррей и Нейман; H. H. Goldstine, F. J. Murray, J. von Neumann, The Jacobi method for real symmetric matrix. J. Assoc. Comput. Machinary, 1959, **6**, № 1, 59—96.
53. Дервюдье; L. Derwidue, Une methode mecanique de calcul des vecteurs propers d'une matrice quelconque. Bull. Soc. Roy., Liège, 1955, **24**, № 5, 149—171.
54. Кифер; J. Kiefer, Sequential minimax search for a maximum. Proc. Am. Math. Soc., 1953, **4**, № 3, 502—506.
55. Джонсон; S. M. Johnson, Optimal search for a maximum is fibonaccian. RAND corp. report P-856, 1956.
56. Кифер, Вольфовиц; J. Kiefer, J. Wolfowitz, Stochastic estimation of the maximum of a regression function. Annals of Math. Stat., 1952, **23**, 462—466.
57. Гельфанд И. М., Цетлин М. Л., Метод оврагов. УМН, 1962, **17**, № 1, 3—25.
58. Пауэлл; M. J. D. Powell, An efficient method for finding the minimum of a function of several variables without calculating derivatives. Computer Journ., 1964, 7, № 2, 155—162.
59. Соболь И. М., Исследование асимптотического поведения решений линейного дифференциального уравнения второго порядка при помощи полярных координат. Матем. сб., 1951, **28 (70)**, № 3, 707—713.
60. Уваров В. Б., Алдонясов В. И., Фазовый метод определения собственных значений для уравнения Шредингера. ЖВМ и МФ, 1967, **7**, № 2, 436—440.
61. Калиткин Н. Н., Решение задач на собственные значения методом дополненного вектора. ЖВМ и МФ, 1965, **5**, № 6, 1107—1115.
62. Жидков Е. П., Макаренко Г. И., Пузынин И. В., Непрерывный аналог метода Ньютона в нелинейных задачах физики. ЭЧАЯ, 1973, **4**, 127—166.
63. Самарский А. А., Соболь И. М., Примеры численного расчета температурных волн. ЖВМ и МФ, 1963, **3**, № 4, 702—719.
64. Тихонов А. Н., Самарский А. А., Об однородных разностных схемах. ЖВМ и МФ, 1961, **1**, № 1, 5—63.
65. Кузнецов Н. Н., Асимптотика решений конечноразностной задачи Коши. ЖВМ и МФ, 1972, **12**, № 2, 334—351.
66. Волчинская М. И., Гольдин В. Я., Калиткин Н. Н., Сравнительное исследование разностных схем для уравнений акустики. ЖВМ и МФ, 1974, **14**, № 4, 919—927.
67. Карлсон; B. G. Carlson, The S_n method. Los Alamos Report, 1953.
68. Гольдин В. Я., Характеристическая разностная схема для нестационарного кинетического уравнения. ДАН СССР, 1960, **133**, № 4, 748—751.
69. Годунов С. К., Разностный метод численного расчета разрывных решений уравнений гидродинамики. Матем. сб., 1959, **47 (89)**, № 3, 271—306.
70. Гольдин В. Я., Калиткин Н. Н., Шишова Т. В., Нелинейные разностные схемы для гиперболических уравнений. ЖВМ и МФ, 1965, **5**, № 5, 938—944.
71. Жуков А. И., Предельная теорема для разностных операторов. УМН, 1959, **14**, № 3, 129—136.
72. Нейман, Рихтмайер; J. von Neumann, R. D. Richtmyer, A method for the numerical calculations of hydrodynamical shocks. J. Appl. Phys., 1950, **21**, № 2, 232—237.
73. Писмен, Рэчфорд; D. W. Peaceman, H. H. Rachford, The numerical solution of parabolic and elliptic differential equations. J. Soc. Industr. Appl. Math., 1955, **3**, № 1, 28—42.
74. Дуглас; J. Douglas, On the numerical integration of $u_{xx}+u_{yy}=u_t$ by implicit methods. J. Industr. Appl. Math., 1955, **3**, № 1, 42—65.

75. Дьяконов Е. Г., Разностные схемы с расщепляющимся оператором для многомерных нестационарных задач. ЖВМ и МФ, 1962, **2**, № 4, 549—568.
76. Коновалов А. А., Метод дробных шагов решения задачи Коши для многомерного уравнения колебаний. ДАН СССР, 1962, **147**, № 1, 25—27.
77. Самарский А. А., Об одном экономичном разностном методе решения многомерного параболического уравнения в произвольной области. ЖВМ и МФ, 1962, **2**, № 1, 25—56.
78. Яненко Н. Н., Об одном разностном методе счета многомерного уравнения теплопроводности. ДАН СССР, 1959, **125**, № 6, 1207—1210.
79. Яненко Н. Н., Об экономических неявных схемах (метод дробных шагов). ДАН СССР, 1960, **134**, № 5, 1034—1036.
80. Фрязинов И. В., Экономичные симметризованные схемы решения краевых задач для многомерного уравнения параболического типа. ЖВМ и МФ, 1968, **8**, № 2, 436—443.
81. Вашпресс; E. L. Wachspress, Extended application of alternating-direction-implicit iteration model problem theory. J. Soc. Industr. Appl. Math., 1963, **11**, № 3, 994—1016.
82. Лебедев В. И., Финогенов С. А., О порядке выбора итерационных параметров в чебышевском циклическом итерационном методе. ЖВМ и МФ, 1971, **11**, № 2, 425—438.
83. Абрамов А. А., Андреев В. Б., О применении метода прогонки к нахождению периодических решений дифференциальных и разностных уравнений. ЖВМ и МФ, 1963, **3**, № 2, 377—381.
84. Лакс; P. D. Lax, Weak solutions of nonlinear hyperbolic equations and their numerical computation. Comm. Pure Appl. Math., 1954, **7**, № 1, 159—193.
85. Самарский А. А., Арсенин В. Я., О численном решении уравнений газодинамики с различными типами вязкости. ЖВМ и МФ, 1961, **1**, № 2, 357—360.
86. Русанов В. В., Разностные схемы третьего порядка точности для сквозного счета разрывных решений. ДАН СССР, 1968, **180**, № 6, 1303—1305.
87. Алалыкин Г. Б., Годунов С. К., Киреева И. Л., Плинер Л. А., Решение одномерных задач газовой динамики в подвижных сетках, М., «Наука», 1970.
88. Тихонов А. Н., О решении некорректно поставленных задач. ДАН СССР, 1963, **151**, № 3, 501—504.
89. Тихонов А. Н., О регуляризации некорректно поставленных задач. ДАН СССР, 1963, **153**, № 1, 49—52.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Автомодельные решения 294
Адамса метод 250
Анализ регрессии 495, 496
Анизотропная теплопроводность 394, 395
Аппроксимационная вязкость 351
Аппроксимация 308
— абсолютная 310
— безусловная 310
— дробно-линейная 63
— краевых условий 385, 393, 427
— локальная 309
— условная 310
Асимметрия 487

Бегущая температурная волна 295
Бегущий счет 337, 344, 379
Бесселя формулы 62
Большие задачи 388

Включение точки 388
Вольтерра уравнение второго рода 454
— первого рода 462
Выбор веса 60, 486, 497
Выравнивающая замена переменных 42
Вырожденное ядро 460
Вычисление корней многочлена 147, 148
— кратных интегралов методом Монте-Карло 121
— — — — последовательного интегрирования 111
— — — — ячеек 108
— несобственных интегралов 105
— обратной матрицы 131
— определителя 130

Галеркина метод 276, 288, 461
Гарвика прием 146
Геометрическая интерпретация устойчивости 341, 379
Гивенса метод вращений 175
Гильбертово пространство 20

Двухкруговые итерации 449
Дервюдье метод 189
Дирихле задача 401
Дисбаланс 365
Дисперсионный анализ 495
Диссипативные схемы 353
Дифференцирование быстропеременных функций 80
— интерполяционного многочлена Ньютона 70
— — — — —, погрешность 71
— на квазиравномерных сетках 80
— на равномерной сетке 73
Дихотомия 139, 263
Доверительная вероятность 483
Доверительный интервал 483
Допустимое решение 356

Жорданов набор шагов 411
Жорданова подматрица 157
— форма матрицы 157

Замораживание коэффициентов 320
Зейделя метод 155

Инварианты акустические 434
Интегрирование осциллирующих функций 103
— разрывных функций 100
Интегро-интерполяционный метод 304
Интерполяционный многочлен Ньютона 30
— — —, погрешность 32
— — —, —, апостериорная оценка 33
— — Эрмита 36
— — —, погрешность 37
Интерполяция квазилинейная 43
— лагранжева 28
— линейная 28
— многомерная 47
— — на произвольной сетке 50
— — последовательная 49
— — треугольная 49

Интерполяция монотонная 47
— нелинейная 41
— обратная 35
— сплайнами 44
—, сходимость 39
— эрмитова 36

Квадратурные формулы, априорные оценки точности 99
— —, веса 86
— — Гаусса — Кристоффеля 94
— — Маркова 97
— — нелинейные 100
— —, погрешность 86
— — Симпсона 88
— — средних 89
— —, сходимость 98
— — трапеций 86
— — —, погрешность 87
— —, узлы 86
— — Эйлера — Маклорена 91
Комплексная организация расчета 274, 287, 409
Конечные разности 31
Консервативные схемы 365, 447
Корректность 24
Корреляционный анализ 497
Коши задача 238, 291
— — плохо обусловленная 240
Коэффициент парной корреляции 497
— перекоса матрицы 161
Коэффициентная устойчивость 384
Краевые задачи 261, 291
— — нестационарные 291
Критерии установления 408
Куранта условие 338, 436

Лагерра многочлены 503
Лежандра многочлены 501
Линеаризация разностной схемы 321
Линейное программирование 217
Локально-одномерные схемы 396

Матриц виды 132, 158
— нормы 21
Матрица вращения 175
— отражения 170
— сдвинутая 191
Метод баланса 304, 363, 380
— баллистический 262
— вращений итерационный 177
— — —, выбор оптимального элемента 179
— — прямой 175
— выбранных точек 63
— выравнивания 42
— декомпозиции 419
Метод дополненного вектора 286
— золотого сечения 196
— исключения Гаусса, выбор главного элемента 130
— — —, обратный ход 129
— — —, прямой ход 129
— итерированного веса 64, 68
— касательных 143
— квадратного корня 135
— квадрирования 148
— линеаризации 143, 152, 263, 274
— ломаных 243
— малого параметра 242
— моментов 461
— наименьших квадратов 59, 224
— — —, выбор весов 60
— — —, оптимальное число коэффициентов 60
— неопределенных коэффициентов 305
— оврагов 209
— отражений 170
— парабол 146, 198
— последовательных приближений 141, 150, 272, 458
— — —, стохастические задачи 142
— простых итераций 141, 150
— прямых 298
— разностной аппроксимации 303
— секущих 145, 264
— сопряженных направлений 210
— стрельбы 262, 266, 281
— —, линейные задачи 264, 267
— уменьшения невязки 307
— фиктивных точек 306
— штрафных функций 216
Минимизация функционала по аргументу 223
Многочлены обобщенные 28
— ортогональные 501
— — на системе точек 503
Модуль непрерывности 19
Монотонность схем 376, 384

Наилучшая схема 381
Наилучшее приближение 51
— — равномерное 66
— — среднеквадратичное 53
Наискорейший спуск 207
— —, сходимость 208
Направление 299
Невязка 302
Независимые измерения 491
Непрерывный аналог метода Ньютона 288
— функционал 227
Неявные схемы 252, 301
Нормальное распределение 483, 487

Нормальное решение 222, 476
Нормы 19
— векторов 21
— матриц 21
— — подчиненные 22
— — согласованные 22
— негативные 322
— энергетические 308
Ньютона интерполяционный многочлен 30
— метод 143, 152, 263, 274

Обратные итерации 166
— — с переменным сдвигом 192
— — со сдвигом 191
Овраг 203
— разрешимый 203
Однородные схемы 358
Операторов виды 323
— свойства 323
Оптимальное управление 226
Особые точки дифференциальных уравнений 257
Оценки погрешности апостериорные 33, 330
— — априорные 33, 328
Ошибки грубые 481, 489
— систематические 481
— случайные 481

Первое дифференциальное приближение 352
Пикара метод 240
Плохая обусловленность 25, 240
— — линейных алгебраических систем 127, 130, 137, 476
Подобие 296
Погрешность метода 23
— неустранимая 22
— округления 23
Показатель симметрии 384, 440
Полностью консервативные схемы 366, 450
Попеременно-треугольная схема 421
Порядок точности 325, 327
— — не целый 93, 340
Последовательность точек $ЛП_\tau$ 121
— функций минимизирующая 227
Потенциал скоростей 429
Предиктор-корректор 247
Преобладание диагонального элемента 134, 154
Преобразование подобия матриц 158
Признак равномерной устойчивости 314, 316, 319
Принцип максимума 315
Прогонка 132
Прогонка дифференциальная 266
Продольно-поперечная схема 391
Пространство C 19
Псевдовязкость 359
— квадратичная 361, 443
— линейная 362, 442
Псевдослучайные числа 115

Разделенные разности 29
— — с кратными узлами 37
Разрывные коэффициенты 279, 380
Разыгрывание случайной величины 117
— — — многомерной 122
— — — равномерно распределенной 115
Регуляризация дифференцирования по Тихонову 474
— — по шагу 83
— — сглаживанием 83
— линейного программирования 221
— суммирования ряда по Тихонову 58, 475
— — — по числу членов 57
Регуляризирующий оператор 464
Рельеф функции 201
Решение уравнения обратной интерполяцией 35
Ритца метод 230, 413
Рунге — Кутта метод 246
— — —, оценка точности 249
Рунге метод 75, 259, 332
— — рекуррентный 77, 331
Рунге — Ромберга метод 76

Сглаживание функции 60, 62, 474
Сетки квазиравномерные 78
— специальные 279, 383
Сильный разрыв 357
Симплекс-метод 220
Слабый разрыв 355
Слой 299
Случайная величина 114
— —, плотность распределения 114
— —, равномерно распределенная 114
— —, — —, разыгрывание 115
— —, разыгрывание 117
Собственные значения 156, 280
Согласованные измерения 492
Сплайн 46
— многомерный 235
Способ параллельных касательных 211
Спуск по координатам 203
Стандарт 484
— выборки 485
— —, несмещенная оценка 484
Степенной метод 190
Стохастическая зависимость 495

Стохастическая задача нахождения минимума 194
Стьюдента коэффициенты 485
— критерий 485
Субтабулирование 34
Схема двуслойная 313
— —, каноническая форма 318
— «крест» 425, 435, 444
— ломаных 243
— с весами 370
— с выделением особенностей 358, 430
— с полусуммой 371
Сходимость 325
— векторов по направлению 21
— квадратичная 145
— кубическая 145
— линейная 145
— ложная 362
— равномерная 19
— среднеквадратичная 20
Счет на установление 190, 403
— — —, критерий установления 408
— — —, оптимальный шаг 404

Тихоновский стабилизатор 405
Точки повышенной точности численного дифференцирования 72
Треугольный оператор 421

Удаление найденных корней 140
Узлы сетки нерегулярные 300
— — регулярные 300
Уменьшение дисперсии метода Монте-Карло 119
Устойчивость 24, 312
— асимптотическая 314, 374
— безусловная 313
— по начальным данным 313
— — — — равномерная 313
— слабая 25, 314
— собственных значений и векторов матриц 159
— условная 313

Фазовый метод 282
Факторизованные схемы 437
Филона формулы 103
Фишера коэффициенты 494
— критерий 493
Фредгольма уравнение второго рода 453
— — первого рода 462
Фурье преобразование быстрое 416
— — дискретное 62

Характеристический многочлен 156
Хаусхолдера метод отражений 170

Центральные моменты распределения 487
Циклическая прогонка 434

Чебышева критерий 486
— многочлены 503
Чебышевская система функций 28
Чебышевский набор шагов 409
— — — упорядоченный 412
Чисто неявная схема 371

Шаблон 297, 300

Эйлера метод 243
— уравнение 469
Эйткена экстраполяционный процесс 92
Экономичные схемы 391
Экстраполяция 33
— многомерная 48
Эксцесс 487
Эрмита многочлены интерполяционные 36
— — ортогональные 503

Явно-неявная схема 342
Явные схемы 301
Якоби метод вращений 177
— многочлены ортогональные 501