

А. А. САМАРСКИЙ, Е. С. НИКОЛАЕВ

МЕТОДЫ РЕШЕНИЯ СЕТОЧНЫХ УРАВНЕНИЙ

*Допущено Министерством высшего
и среднего специального образования СССР
в качестве учебного пособия для студентов вузов,
обучающихся по специальности «Прикладная математика»*



МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ

1978

С 17

УДК 518.61

Методы решения сеточных уравнений. А. А. Самарский, Е. С. Николаев. Главная редакция физико-математической литературы изд-ва «Наука», М., 1978.

Книга посвящена методам решения алгебраических систем высокого порядка, возникающих при применении метода сеток к задачам математической физики. Наряду с итерационными методами, которые получили наиболее широкое распространение в вычислительной практике при решении указанных задач, излагаются и прямые методы.

Книга рассчитана на студентов и аспирантов факультетов прикладной математики, а также на инженеров и специалистов, работающих в области вычислительной математики.

C 20204—132
053(02)-78 15-78

© Главная редакция
физико-математической литературы
издательства «Наука», 1978

ОГЛАВЛЕНИЕ

Предисловие	8
Введение	11
Глава I. Прямые методы решения разностных уравнений	24
§ 1. Сеточные уравнения. Основные понятия	24
1. Сетки и сеточные функции (24). 2. Разностные производные и некоторые разностные тождества (26). 3. Сеточные и разностные уравнения (30). 4. Задача Коши и краевые задачи для разностных уравнений (33).	
§ 2. Общая теория линейных разностных уравнений	37
1. Свойства решений однородного уравнения (37). 2. Теоремы о решениях линейного уравнения (40). 3. Метод вариации постоянных (41). 4. Примеры (45).	
§ 3. Решение линейных уравнений с постоянными коэффициентами	48
1. Характеристическое уравнение. Случай простых корней (48). 2. Случай кратных корней (49). 3. Примеры (52).	
§ 4. Уравнения второго порядка с постоянными коэффициентами	54
1. Общее решение однородного уравнения (54). 2. Полиномы Чебышева (57). 3. Общее решение неоднородного уравнения (59).	
§ 5. Разностные задачи на собственные значения	63
1. Первая краевая задача на собственные значения (63). 2. Вторая краевая задача (65). 3. Смешанная краевая задача (66). 4. Периодическая краевая задача (68).	
Глава II. Метод прогонки	73
§ 1. Метод прогонки для трехточечных уравнений	73
1. Алгоритм метода (73). 2. Метод встречных прогонок (76). 3. Обоснование метода прогонки (78). 4. Примеры применения метода прогонки (80).	
§ 2. Варианты метода прогонки	84
1. Потоковый вариант метода прогонки (84). 2. Метод циклической прогонки (86). 3. Метод прогонки для сложных систем (90). 4. Метод немонотонной прогонки (93).	
§ 3. Метод прогонки для пятиточечных уравнений	97
1. Алгоритм монотонной прогонки (97). 2. Обоснование метода (100). 3. Вариант немонотонной прогонки (101).	
§ 4. Метод матричной прогонки	103
1. Системы векторных уравнений (103). 2. Прогонка для трехточечных векторных уравнений (106). 3. Прогонка для двухточечных векторных уравнений (109). 4. Ортогональная прогонка для двухточечных векторных уравнений (112). 5. Прогонка для трехточечных уравнений с постоянными коэффициентами (117).	
Глава III. Метод полной редукции.	121
§ 1. Краевые задачи для трехточечных векторных уравнений	121
1. Постановка краевых задач (121). 2. Первая краевая задача (123). 3. Другие краевые задачи для разностных уравнений (125). 4. Разностная задача Дирихле повышенного порядка точности (128).	
§ 2. Метод полной редукции для первой краевой задачи	130
1. Процесс нечетно-четного исключения (130). 2. Преобразование правой части и обращение матриц (133). 3. Алгоритм метода (136). 4. Второй алгоритм метода (139).	

§ 3. Примеры применения метода	144
1. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике (144). 2. Разностная задача Дирихле повышенного порядка точности (146).	
§ 4. Метод полной редукции для других краевых задач	149
1. Вторая краевая задача (149). 2. Периодическая задача (154). 3. Третья краевая задача (157).	
 Г л а в а IV. Метод разделения переменных	164
§ 1. Алгоритм дискретного преобразования Фурье	164
1. Постановка задачи (164). 2. Разложение по синусам и сдвигнутым синусам (168). 3. Разложение по косинусам (175). 4. Преобразование действительной периодической сеточной функции (178). 5. Преобразование комплексной периодической сеточной функции (183).	
§ 2. Решение разностных задач методом Фурье	185
1. Разностные задачи на собственные значения для оператора Лапласа в прямоугольнике (185). 2. Уравнение Пуассона в прямоугольнике. Разложение в двойной ряд (190). 3. Разложение в однократный ряд (194).	
§ 3. Метод неполной редукции	198
1. Комбинация методов Фурье и редукции (198). 2. Решение краевых задач для уравнения Пуассона в прямоугольнике (205). 3. Разностная задача Дирихле повышенного порядка точности в прямоугольнике (208).	
 Г л а в а V. Математический аппарат теории итерационных методов	212
§ 1. Некоторые сведения из функционального анализа	212
1. Линейные пространства (212). 2. Операторы в линейных нормированных пространствах (215). 3. Операторы в гильбертовом пространстве (218). 4. Функции от ограниченного оператора (223). 5. Операторы в конечномерном пространстве (224). 6. Разрешимость операторных уравнений (227).	
§ 2. Разностные схемы как операторные уравнения	230
1. Примеры пространств сеточных функций (230). 2. Некоторые разностные тождества (233). 3. Границы простейших разностных операторов (235). 4. Оценки снизу для некоторых разностных операторов (238). 5. Оценки сверху для разностных операторов (246). 6. Разностные схемы как операторные уравнения в абстрактных пространствах (247). 7. Разностные схемы для эллиптических уравнений с постоянными коэффициентами (251). 8. Уравнения с переменными коэффициентами и со смешанными производными (254).	
§ 3. Основные понятия теории итерационных методов	258
1. Метод установления (258). 2. Итерационные схемы (259). 3. Сходимость и число итераций (261). 4. Классификация итерационных методов (263).	
 Г л а в а VI. Двухслойные итерационные методы	266
§ 1. Постановка задачи о выборе итерационных параметров	266
1. Исходное семейство итерационных схем (266). 2. Задача для погрешности (267). 3. Самосопряженный случай (268).	
§ 2. Чебышевский двухслойный метод	269
1. Построение набора итерационных параметров (269). 2. О неулучшаемости априорной оценки (271). 3. Примеры выбора оператора D (272). 4. О вычислительной устойчивости метода (275). 5. Построение оптимальной последовательности итерационных параметров (280).	
§ 3. Метод простой итерации	284
1. Выбор итерационного параметра (284). 2. Оценка нормы оператора перехода (285).	
§ 4. Несамосопряженный случай. Метод простой итерации	287
1. Постановка задачи (287). 2. Минимизация нормы оператора перехода (288). 3. Минимизация нормы разрешающего оператора (293). 4. Метод симметризации уравнения (297).	
§ 5. Примеры применения итерационных методов	298
1. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике (298). 2. Разностная задача Дирихле для уравнения Пуассона в произвольной области (301). 3. Разностная задача Дирихле для эллиптического уравнения с переменными коэффициентами (307). 4. Разностная задача Дирихле для эллиптического уравнения со смешанной производной (312).	

Г л а в а VII. Трехслойные итерационные методы	351
§ 1. Оценка скорости сходимости	315
1. Исходное семейство итерационных схем (315). 2. Оценка нормы погрешности (316).	
§ 2. Полуитерационный метод Чебышева	318
1. Формулы для итерационных параметров (318). 2. Примеры выбора оператора D (320). 3. Алгоритм метода (321).	
§ 3. Стационарный трехслойный метод	321
1. Выбор итерационных параметров (321). 2. Оценка скорости сходимости (322).	
§ 4. Устойчивость двухслойных и трехслойных методов по априорным данным	324
1. Постановка задачи (324). 2. Оценки скорости сходимости методов (326).	
Г л а в а VIII. Итерационные методы вариационного типа	331
§ 1. Двухслойные градиентные методы	331
1. Постановка задачи о выборе итерационных параметров (331). 2. Формула для итерационных параметров (333). 3. Оценка скорости сходимости (334). 4. Неулучшаемость оценки в самосопряженном случае (336). 5. Асимптотическое свойство градиентных методов в самосопряженном случае (338).	
§ 2. Примеры двухслойных градиентных методов	340
1. Метод скорейшего спуска (340). 2. Метод минимальных невязок (341). 3. Метод минимальных поправок (343). 4. Метод минимальных погрешностей (344). 5. Пример применения двухслойных методов (344).	
§ 3. Трехслойные методы сопряженных направлений	347
1. Постановка задачи о выборе итерационных параметров. Оценка скорости сходимости (347). 2. Формулы для итерационных параметров. Трехслойная итерационная схема (349). 3. Варианты расчетных формул (354).	
§ 4. Примеры трехслойных методов	355
1. Частные случаи методов сопряженных направлений (355). 2. Локально оптимальные трехслойные методы (356).	
§ 5. Ускорение сходимости двухслойных методов в самосопряженном случае	360
1. Алгоритм процесса ускорения (360). 2. Оценка эффективности (361). 3. Пример (363).	
Г л а в а IX. Треугольные итерационные методы	366
§ 1. Метод Зейделя	366
1. Итерационная схема метода (366). 2. Примеры применения метода (369). 3. Достаточные условия сходимости (372).	
§ 2. Метод верхней релаксации	375
1. Итерационная схема. Достаточные условия сходимости (375). 2. Постановка задачи о выборе итерационного параметра (376). 3. Оценка спектрального радиуса (379). 4. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике (380). 5. Разностная задача Дирихле для эллиптического уравнения с переменными коэффициентами (385).	
§ 3. Треугольные методы	387
1. Итерационная схема (387). 2. Оценка скорости сходимости (389). 3. Выбор итерационного параметра (390). 4. Оценка скорости сходимости методов Зейделя и релаксации (391).	
Г л а в а X. Попеременно-треугольный метод	395
§ 1. Общая теория метода	395
1. Итерационная схема (395). 2. Выбор итерационных параметров (397). 3. Метод нахождения исходных величин δ и Δ (400). 4. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике (402).	
§ 2. Разностные краевые задачи для эллиптических уравнений в прямоугольнике	409
1. Задача Дирихле для уравнения с переменными коэффициентами (409). 2. Модифицированный попеременно-треугольный метод (411). 3. Сравнение вариантов метода (417). 4. Третья краевая задача (418). 5. Разностная задача Дирихле для уравнения со смешанными производными (421).	

§ 3. Попеременно-треугольный метод для эллиптических уравнений в произвольной области	423
1. Постановка разностной задачи (423). 2. Построение попеременно-треугольного метода (425). 3. Задача Дирихле для уравнения Пуассона в произвольной области (429).	
Г л а в а XI. Метод переменных направлений	432
§ 1. Метод переменных направлений в коммутативном случае	432
1. Итерационная схема метода (432). 2. Постановка задачи о выборе параметров (434). 3. Дробно-линейное преобразование (436). 4. Оптимальный набор параметров (438).	
§ 2. Примеры применения метода	440
1. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике (440). 2. Третья краевая задача для эллиптического уравнения с разделяющимися переменными (445). 3. Разностная задача Дирихле повышенного порядка точности (449).	
§ 3. Метод переменных направлений в общем случае	453
1. Случай неперестановочных операторов (453). 2. Разностная задача Дирихле для эллиптического уравнения с переменными коэффициентами (455).	
Г л а в а XII. Методы решения уравнений с незнакоопределенными и вырожденными операторами	459
§ 1. Уравнения с действительным незнакоопределенным оператором	459
1. Итерационная схема. Задача выбора итерационных параметров (459). 2. Преобразование оператора в самосопряженном случае (462). 3. Итерационный метод с чебышевскими параметрами (464). 4. Итерационные методы вариационного типа (468). 5. Примеры (469).	
§ 2. Уравнения с комплексным оператором	471
1. Метод простой итерации (471). 2. Метод переменных направлений (475).	
§ 3. Общие итерационные методы для уравнений с вырожденным оператором	478
1. Итерационные схемы в случае невырожденного оператора B (478). 2. Итерационный метод минимальных невязок (482). 3. Метод с чебышевскими параметрами (485).	
§ 4. Специальные методы	489
1. Разностная задача Неймана для уравнения Пуассона в прямоугольнике (489). 2. Прямой метод для задачи Неймана (493). 3. Итерационные схемы с вырожденным оператором B (496).	
Г л а в а XIII. Итерационные методы решения нелинейных уравнений	500
§ 1. Итерационные методы. Общая теория	500
1. Метод простой итерации для уравнений с монотонным оператором (500). 2. Итерационные методы для случая дифференцируемого оператора (503). 3. Метод Ньютона—Канторовича (505). 4. Двухступенчатые итерационные методы (509). 5. Другие итерационные методы (512).	
§ 2. Методы решения нелинейных разностных схем	514
1. Разностная схема для одномерного эллиптического квазилинейного уравнения (514). 2. Метод простой итерации (522). 3. Итерационные методы для разностных квазилинейных эллиптических уравнений в прямоугольнике (524). 4. Итерационные методы для слабонелинейных уравнений (529).	
Г л а в а XIV. Примеры решения сеточных эллиптических уравнений .	532
§ 1. Способы построения неявных итерационных схем	532
1. Принцип регуляризации в общей теории итерационных методов (532). 2. Итерационные схемы с факторизованным оператором (536). 3. Способ неявного обращения оператора B (двухступенчатый метод) (540).	
§ 2. Системы эллиптических уравнений	542
1. Задача Дирихле для системы эллиптических уравнений в p -мерном параллелепипеде (542). 2. Система уравнений теории упругости (547).	

Глава XV. Методы решения эллиптических уравнений в криволинейных ортогональных координатах	550
§ 1. Постановка краевых задач для дифференциальных уравнений	550
1. Эллиптические уравнения в цилиндрической системе координат (550). 2. Краевые задачи для уравнений в цилиндрической системе координат (553).	
§ 2. Решение разностных задач в цилиндрической системе координат	556
1. Разностные схемы без смешанных производных в осесимметрическом случае (556). 2. Прямые методы (560). 3. Метод переменных направлений (561). 4. Решение уравнений, заданных на поверхности цилиндра (565).	
§ 3. Решение разностных задач в полярной системе координат	569
1. Разностные схемы для уравнений в круге и кольце (569). 2. Разрешимость разностных краевых задач (571). 3. Принцип суперпозиции для задачи в круге (574). 4. Прямые методы решения уравнений в круге и кольце (575). 5. Метод переменных направлений (577). 6. Решение разностных задач в кольцевом секторе (580). 7. Общий случай переменных коэффициентов (582).	
Дополнение. Построение полинома, наименее уклоняющегося от нуля	585
Литература	590
Предметный указатель	591

ПРЕДИСЛОВИЕ

Численное решение дифференциальных уравнений математической физики методом конечных разностей проводится в два этапа: 1) разностная аппроксимация дифференциального уравнения на сетке—написание разностной схемы, 2) решение на ЭВМ разностных уравнений, представляющих собой системы линейных алгебраических уравнений высокого порядка специального вида (плохая обусловленность, ленточная структура матрицы системы). Применение общих методов линейной алгебры для таких систем далеко не всегда целесообразно как из-за необходимости хранения большого объема информации, так и из-за большого объема вычислительной работы, требуемой этими методами. Для решения разностных уравнений уже давно разрабатываются специальные методы, которые в той или иной степени учитывают специфику задачи и позволяют найти решение с затратой меньшего числа действий по сравнению с общими методами линейной алгебры.

Данная книга является продолжением книги А. А. Самарского и В. Б. Андреева «Разностные методы решения эллиптических уравнений», в которой изучается круг вопросов, связанных с разностной аппроксимацией, построением разностных операторов и оценкой скорости сходимости разностных схем для типичных краевых задач эллиптического типа.

Здесь мы рассматриваем только методы решения разностных уравнений. Книга фактически состоит из двух частей. Первая часть (гл. I—IV) посвящена применению прямых методов решения разностных уравнений, вторая часть (гл. V—XV)—теории итерационных методов решения сеточных уравнений общего вида и их применению к разностным уравнениям. При использовании прямых методов существенную роль играет специальный вид разностных уравнений. Для решения одномерных трехточечных уравнений рассматриваются различные варианты метода прогонки (монотонная, немонотонная, циклическая, потоковая прогонка и др.).

В главах III и IV излагаются современные экономичные прямые методы решения разностных уравнений Пуассона в прямоугольнике с краевыми условиями различного вида. Это—метод полной редукции и метод разделения переменных, использующий алгоритм быстрого преобразования Фурье, а также комбинированные методы.

При изучении итерационных методов используется трактовка итерационного метода как операторно-разностной схемы, развитая в книгах А. А. Самарского «Введение в теорию разностных схем» (1971) и «Теория разностных схем» (1977). Эта концепция позволяет излагать теорию итерационных методов как раздел общей теории устойчивости операторно-разностных схем, не прибегая к предположениям о структуре матрицы системы (см. также А. А. Самарский и А. В. Гулин «Устойчивость разностных схем» (1973)). Запись итерационных схем в канонической форме позволяет не только выделить операторы, ответственные за сходимость итераций, но и сравнить различные итерационные методы. Основное внимание уделяется изучению скорости сходимости итераций и выбору оптимальных параметров, при которых скорость сходимости максимальна. Наличие оценок скорости сходимости, а также исследование характера вычислительной устойчивости позволяют провести сравнение разных итерационных методов в конкретных ситуациях и сделать выбор. Хотя читатель, вероятно, знаком с основами теории разностных схем и элементами функционального анализа, однако в главе V приводятся используемые в книге сведения об основах математического аппарата теории итерационных схем и показано, как разностные аппроксимации эллиптических уравнений сводятся к операторным уравнениям первого рода $Au = f$ с операторами A в гильбертовом пространстве сеточных функций.

В последующих главах исследуются двухслойная итерационная схема с чебышевским набором параметров, при котором имеет место вычислительная устойчивость метода; трехслойная схема; итерационные методы вариационного типа (методы скорейшего спуска, минимальных невязок, минимальных поправок, сопряженных градиентов и др.); итерационные методы для несамосопряженных уравнений и в случае незнакоопределенного и вырожденного оператора; методы переменных направлений; «треугольные» методы (с алгоритмом обращения треугольной матрицы при определении новой итерации) такие, как метод Зейделя, метод верхней релаксации и др.; итерационные методы решения нелинейных разностных уравнений, решение разностных краевых задач для эллиптических уравнений в криволинейных системах координат и др.

Особое место в книге занимает предложенный и развитый авторами в 1964—1977 гг. универсальный попеременно-треугольный метод, эффективность которого особенно сильно проявляется при решении задачи Дирихле для уравнения Пуассона в произвольной области и задачи Дирихле для уравнения $\operatorname{div}(k \operatorname{grad} u) = -f(x)$, $x = (x_1, x_2)$ с сильно меняющимся коэффициентом $k(x)$.

В книге показано, как переходить от общей теории к конкретным задачам, и приведено большое число итерационных алгоритмов решения разностных уравнений для эллиптических уравнений и систем уравнений. Даны оценки числа итераций

и проведено сравнение различных методов. Так, в частности, показано, что для простейшей задачи прямые методы более экономичны, чем метод переменных направлений. Следует подчеркнуть, что возникающие на практике все более и более сложные задачи линейной алгебры требуют как разработки новых методов, так и расширения области применимости старых методов. При этом происходит переоценка сравнительных характеристик разных методов.

При написании книги авторы использовали материалы лекций, читавшихся ими в период 1961—1977 гг. на механико-математическом факультете и на факультете вычислительной математики и кибернетики МГУ, а также материалы опубликованных работ авторов.

Авторы пользуются возможностью выразить благодарность В. Б. Андрееву, И. В. Фрязинову, М. И. Бакировой, А. Б. Кучерову, Н. Е. Капорину за ряд полезных замечаний по материалу книги.

Авторы благодарны Т. Н. Галишниковой, А. А. Голубевой и особенно В. М. Марченко за помощь при подготовке рукописи к печати.

A. A. Самарский, E. S. Николаев

Москва, декабрь 1977 г.

ВВЕДЕНИЕ

Применение различных численных методов (разностных, вариационно-разностных, проекционно-разностных методов, в том числе метода конечных элементов) для решения дифференциальных уравнений приводит к системе линейных алгебраических уравнений специального вида — к разностным уравнениям. Эта система обладает следующими специфическими чертами: 1) она имеет высокий порядок, равный числу узлов сетки; 2) система плохо обусловлена (отношение максимального собственного значения матрицы системы к минимальному велико; так, для разностного оператора Лапласа это отношение обратно пропорционально квадрату шага сетки); 3) матрица системы является разреженной — в каждой ее строке отлично от нуля несколько элементов, число которых не зависит от числа узлов; 4) ненулевые элементы матрицы расположены специальным образом — матрица является ленточной.

При аппроксимации на сетке интегральных и интегро-дифференциальных уравнений мы получаем систему уравнений относительно функции, заданной на сетке (сеточной функции). Такие уравнения естественно называть сеточными уравнениями:

$$\sum_{\xi \in \omega} a(x, \xi) y(\xi) = f(x), \quad x \in \omega, \quad (1)$$

где суммирование проводится по всем узлам сетки ω , т. е. по дискретному множеству точек. Матрица $(a(x, \xi))$ сеточного уравнения является, в общем случае, заполненной. Если перенумеровать узлы сетки, то сеточное уравнение можно записать в виде

$$\sum_{j=1}^N a_{ij} y_j = f_i, \quad i = 1, 2, \dots, N, \quad (2)$$

где i, j — номера узлов сетки, N — общее число узлов. Обратный ход рассуждений очевиден. Таким образом, линейное сеточное уравнение есть система линейных алгебраических уравнений и, обратно, любую линейную систему алгебраических уравнений можно трактовать как линейное сеточное уравнение относительно сеточной функции, заданной на некоторой сетке с числом узлов, равным порядку системы. Заметим, что вариационные методы (Ритца, Галеркина и др.) численного решения дифференциальных

уравнений приводят обычно к системам с заполненной матрицей.

Разностное уравнение есть частный случай сеточного уравнения, когда матрица (a_{ij}) является разреженной. Так, например, (2) представляет собой разностное уравнение m -го порядка, если в строке номера i имеется лишь $m+1$ отличный от нуля элемент a_{ij} при $j = i, i+1, \dots, i+m$.

Из сказанного ясно, что решение сеточных и, в частности, разностных уравнений является задачей линейной алгебры.

* * *

Для решения задач линейной алгебры существует много различных численных методов, непрерывно ведется работа по их усовершенствованию, проводится переоценка методов, разрабатываются новые методы. В результате оказывается, что значительная часть имеющихся методов имеет право на существование, обладает своей областью применимости. Поэтому для решения конкретной задачи на ЭВМ существует проблема выбора одного метода из множества допустимых методов решения данной задачи. Этот метод должен, очевидно, обладать наилучшими характеристиками (или, как любят говорить, быть оптимальным методом) такими, как минимум времени решения задачи на ЭВМ (или минимум числа арифметических и логических операций при отыскании решения), вычислительная устойчивость, т. е. устойчивость по отношению к ошибкам округления, и др.

Естественно требовать, чтобы любой вычислительный алгоритм для ЭВМ позволял в принципе получить решение данной задачи с любой наперед заданной точностью $\varepsilon > 0$ за конечное число действий $Q(\varepsilon)$. Этому требованию удовлетворяет бесчисленное множество алгоритмов, в котором и следует искать алгоритм с минимумом $Q(\varepsilon)$ для любого $\varepsilon > 0$. Такой алгоритм называется экономичным. Конечно, поиск «оптимального» или «наилучшего» метода, как правило, проводится на множестве известных (а не всех допустимых) методов, и сам термин «оптимальный алгоритм» имеет ограниченный и условный смысл.

* * *

Задача теории численных методов состоит как в отыскании наилучших алгоритмов для данного класса задач, так и в установлении иерархии методов. Само понятие наилучшего алгоритма зависит от цели вычислений.

Возможны две постановки вопроса о выборе наилучшего метода:

- требуется решить одну конкретную систему уравнений $Au = f$, $A = (a_{ij})$ — матрица;
- требуется решить несколько вариантов одной и той же задачи, например, уравнения $Au = f$ с различными правыми частями f .

При многовариантном расчёте можно уменьшить среднее число операций $\bar{Q}(\epsilon)$ для одного варианта, если хранить некоторые величины, а не вычислять их заново для каждого варианта (например, хранить обратную матрицу).

Отсюда ясно, что выбор алгоритма должен зависеть от типа расчёта (одновариантного или многовариантного), от возможности хранения дополнительной информации в памяти ЭВМ, что в свою очередь связано как с типом ЭВМ, так и с порядком системы уравнений. При теоретических оценках качества вычислительного алгоритма обычно ограничиваются подсчетом числа арифметических операций, которые требуются для отыскания решения с заданной точностью; при этом вопрос о параметрах ЭВМ, как правило, не обсуждается.

Бурное развитие в последние годы численных методов решения разностных уравнений, аппроксимирующих дифференциальные уравнения эллиптического типа, и появление новых экономичных алгоритмов привели к необходимости пересмотра представлений об областях применимости существовавших ранее методов.

* * *

Содержание данной книги в значительной степени обусловлено необходимостью дать эффективные методы решения разностных уравнений, соответствующих краевым задачам для уравнений эллиптического типа второго порядка. Классификация разностных краевых задач может быть проведена по следующим признакам:

1) вид дифференциального оператора L в уравнении

$$Lu = f(x), \quad x = (x_1, x_2, \dots, x_p) \in G; \quad (3)$$

2) форма области G , в которой ищется решение;

3) тип краевых условий на границе Γ области G ;

4) сетка ω в области $\bar{G} = G + \Gamma$ и разностная схема

$$\Lambda y = -\varphi(x), \quad x \in \omega, \quad (4)$$

т. е. вид разностного оператора Λ .

Примерами эллиптического оператора второго порядка могут быть

$$Lu = \Delta u = \sum_{\alpha=1}^p \frac{\partial^2 u}{\partial x_\alpha^2} \text{ — оператор Лапласа,} \quad (5)$$

$$Lu = \sum_{\alpha, \beta=1}^p \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}(x) \frac{\partial u}{\partial x_\beta} \right) - q(x) u, \quad (6)$$

причем коэффициенты $k_{\alpha\beta}(x)$ в каждой точке $x = (x_1, x_2, \dots, x_p)$ удовлетворяют условию сильной эллиптичности

$$c_1 \sum_{\alpha=1}^p \xi_\alpha^2 \leq \sum_{\alpha, \beta=1}^p k_{\alpha\beta}(x) \xi_\alpha \xi_\beta \leq c_2 \sum_{\alpha=1}^p \xi_\alpha^2, \quad c_1, c_2 = \text{const} > 0, \quad (7)$$

где $\xi = (\xi_1, \dots, \xi_p)$ — произвольный вектор. Если $\mathbf{u}(x) = (u^1(x), u^2(x), \dots, u^m(x))$ — вектор-функция, то (3) есть система уравнений и

$$(Lu)^i = \sum_{j=1}^m \sum_{\alpha, \beta=1}^p \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}^{ij}(x) \frac{\partial u^j}{\partial x_\beta} \right), \quad i = 1, 2, \dots, m,$$

а условие сильной эллиптичности имеет вид

$$c_1 \sum_{i=1}^m \sum_{\alpha=1}^p (\xi_\alpha^i)^2 \leq \sum_{i, j=1}^m \sum_{\alpha, \beta=1}^p k_{\alpha\beta}^{ij}(x) \xi_\alpha^i \xi_\beta^j \leq c_2 \sum_{i=1}^m \sum_{\alpha=1}^p (\xi_\alpha^i)^2,$$

$$c_1, c_2 = \text{const} > 0.$$

* * *

Форма области сильно влияет на свойства матрицы разностных уравнений. Мы будем выделять области, для которых уравнение $Lu = 0$ с однородными краевыми условиями допускает разделение переменных. Так, например, для уравнения Лапласа в декартовых координатах (x_1, x_2) $Lu = \Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$ метод разделения переменных применим в случае, когда G есть прямоугольник. Аналогичным свойством обладает и разностная схема на прямоугольной сетке, например, схема «крест»; при этом сетка может быть неравномерной по каждому направлению.

При сравнении различных численных методов решения систем алгебраических уравнений будем использовать в качестве *эталонной* или *модельной* задачи следующую разностную краевую задачу:

уравнение Пуассона, область — квадрат, краевые условия первого рода, сетка — квадратная с шагами $h_1 = h$ и $h_2 = h$ по x_1 и x_2 , разностный оператор Λ — пятиточечный.

Вторая группа разностных краевых задач соответствует следующим данным: L — оператор с переменными коэффициентами вида (6): а) без смешанных производных, б) со смешанными производными, область $G = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ — прямоугольник (параллелепипед при $p \geq 3$).

Третья группа задач — область сложной формы, а L либо оператор Лапласа, либо оператор общего вида; здесь степень сложности задачи определяется в первую очередь формой области, выбором сетки и разностного оператора в окрестности границы.

Для второй и третьей групп задач разностный оператор обычно выбирается так, чтобы сохранить основные свойства (самосо-

пряженность, знакоопределенность и др.) исходной задачи и удовлетворить требованию аппроксимации с определенным порядком относительно шага сетки.

* * *

Для решения разностных эллиптических задач применяются прямые и итерационные методы.

Прямые методы применимы в многомерном случае в основном для задач первой группы (L — оператор Лапласа, G — прямоугольник при $p=2$ и параллелепипед при $p \geq 3$, Λ — пяти- или девятиточечная разностная схема при $p=2$). Для одномерных задач, когда разностное уравнение имеет второй порядок (матрица является трехдиагональной), а коэффициенты уравнения могут быть переменными, применим метод прогонки, который является вариантом метода Гаусса (см. гл. II). Существует ряд вариантов метода прогонки: монотонная прогонка, немонотонная прогонка, потоковая прогонка, циклическая прогонка и др. (см. гл. II). Для двумерных задач первой группы (см. выше) эффективен метод полной редукции (гл. III), метод разделения переменных с быстрым преобразованием Фурье, а также комбинация метода неполной редукции с быстрым преобразованием Фурье (гл. IV). Во всех случаях по одному из направлений методом прогонки решается разностное уравнение второго порядка.

Указанные прямые методы в случае разностной задачи Дирихле для уравнения Пуассона в прямоугольнике ($0 \leq x_\alpha \leq l_\alpha$, $\alpha = 1, 2$) на сетке $\bar{\omega} = \{(i_1 h_1, i_2 h_2), i_\alpha = 0, 1, \dots, N_\alpha, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$ требуют $Q = O(N_1 N_2 \log_2 N_2)$ арифметических действий, где $N_2 = 2^n$, $n > 0$ — целое число.

Прямые методы применимы для весьма специального класса задач.

* * *

Разностные эллиптические задачи в случае операторов L общего вида или областей сложной формы решаются в основном при помощи итерационных методов.

Сеточные уравнения можно трактовать как операторные уравнения первого рода

$$Au = f \quad (8)$$

с операторами, заданными на пространствах H сеточных функций. В пространстве H вводятся скалярное произведение $(,)$ и энергетические нормы $\|u\|_D = \sqrt{V(Du, u)}$, $D = D^* > 0$, $D: H \rightarrow H$, где D — некоторый линейный оператор в H .

Итерационные методы решения операторного уравнения $Au = f$ можно трактовать как операторно-разностные (разностные по фиктивному времени или по индексу-номеру итерации) уравнения с операторами в гильбертовом пространстве H . Если новая

итерация y_{k+1} вычисляется через m предыдущих итераций $y_k, y_{k-1}, \dots, y_{k-m+1}$, то итерационный метод (схема) называется $m+1$ -слойным (m -шаговым). Отсюда видна аналогия итерационных схем с разностными схемами для нестационарных задач. Поэтому и теория итерационных методов фактически является специальным разделом общей теории устойчивости операторно-разностных схем. Мы ограничиваемся изучением двухслойных и в меньшей степени трехслойных схем. Переход к многослойным схемам не дает каких-либо преимуществ (как это и следует из общей теории устойчивости, см. [10]).

Важную роль играет запись итерационных методов в единой (канонической) форме, что позволяет выделить оператор (стабилизатор), ответственный за устойчивость и сходимость итераций и сравнить различные итерационные методы с единых позиций.

Любой двухслойный (одношаговый) итерационный метод записывается в следующей канонической форме:

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (9)$$

где $B: H \rightarrow H$ — линейный оператор, имеющий обратный B^{-1} , τ_1, τ_2, \dots — итерационные параметры, k — номер итерации, y_k — итерационное приближение номера k . В общем случае $B=B_{k+1}$ зависит от k . В общей теории мы предполагаем, что B не зависит от k .

Параметры $\{\tau_k\}$ и оператор B произвольны, и их следует выбрать из условия минимума числа итераций n , при котором решение y_n уравнения (9) приближает в H_D точное решение u уравнения $Au=f$ с относительной точностью $\varepsilon > 0$:

$$\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_D. \quad (10)$$

Для излагаемой в книге общей теории итерационных методов не требуется никаких предположений о структуре оператора A (матрицы (a_{ij})). Используются лишь свойства общего вида

$$A = A^* > 0, \quad B = B^* > 0, \quad \gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \quad (11)$$

Операторные неравенства означают, что заданы постоянные γ_1, γ_2 энергетической эквивалентности операторов A и B или границы спектра оператора A в пространстве H_B (γ_1 — минимальное, γ_2 — максимальное собственные значения обобщенной задачи на собственные значения: $Av = \lambda Bv$).

* * *

Решение $\tau_1, \tau_2, \dots, \tau_n$ указанной выше задачи о $\min_{\tau_1, \tau_2, \dots, \tau_n} \pi_0(\varepsilon)$ при заданных γ_1, γ_2 и фиксированном B в случае $D=AB^{-1}A$ выражается через нули полинома Чебышева n -го порядка (чебышевский итерационный метод). При этих оптимальных значениях $\tau_1, \tau_2, \dots, \tau_n$ и заданном произвольном $\varepsilon > 0$ для числа

итераций n , вычисляемых по схеме (9), верна оценка
 $n \geq \frac{\ln(2/\varepsilon)}{\ln((1-\sqrt{\xi})/(1-\sqrt{\xi}))}$ или $n \geq n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{2\sqrt{\xi}}$, $\xi = \gamma_1/\gamma_2$ и выполняется неравенство

$$\|Ay_n - f\|_{B^{-1}} \leq \varepsilon \|Ay_0 - f\|_{B^{-1}}.$$

Вычислительная устойчивость чебышевского метода имеет место при определенном способе нумерации (упорядочивания) нулей чебышевского полинома и параметров $\tau_1^*, \tau_2^*, \dots, \tau_n^*$; этот способ указан в гл. VI.

При $B = E$ (E — единичный оператор) метод (9) называется явным, а при $B \neq E$ — неявным. Если параметр τ_k выбрать постоянным, $\tau_k = \tau_0 = 2/(\gamma_1 + \gamma_2)$, $k = 1, 2, \dots, n$, то получим неявную схему простой итерации, для нее $n \geq n_0(\varepsilon) = \ln\left(\frac{1}{\varepsilon}\right)/(2\xi)$.

Оператор B (стабилизатор) выбирается из условия экономичности, т. е. минимума вычислительной работы при решении уравнения $Bv = F$ с заданной правой частью F , и, как было уже сказано, из условия минимума числа итераций $n_0(\varepsilon)$.

Предположим, что мы умеем экономично решать задачу $Rv = f$ с затратой $Q_R(\varepsilon)$ действий, где

$$R: H \rightarrow H, \quad R = R^* > 0, \quad c_1 R \leq A \leq c_2 R, \quad c_1 > 0. \quad (12)$$

Тогда можно положить $B = R$ и найти решение задачи $Au = f$ по схеме (9) с параметрами $\{\tau_k^*\}$ при $\gamma_1 = c_1$, $\gamma_2 = c_2$ за $Q_A(\varepsilon) \approx \frac{1}{2} \sqrt{c_2/c_1} \ln(2/\varepsilon) Q_R(\varepsilon)$ действий.

Если, например, L — оператор общего вида, G — прямоугольник, то в качестве R можно взять пятиточечный разностный оператор Лапласа и решать уравнение $Rv = f$ прямым методом.

Может оказаться, что уравнение $Rv = f$ выгодно решать не прямым, а итерационным методом, тогда $B \neq R$ и не выписывается в явном виде, а реализуется в результате итерационной процедуры.

* * *

Известные методы Зейделя и верхней релаксации являются неявными и соответствуют треугольным матрицам (операторам) B . Сходимость этих методов доказывается на основе общей теории разностных схем (см. А. А. Самарский, Теория разностных схем, М., 1977 или А. А. Самарский, А. В. Гулин, Устойчивость разностных схем, М., 1973). Однако для методов Зейделя и верхней релаксации оператор B несамосопряжен, и потому нельзя воспользоваться чебышевским методом (9) с оптимальным набором итерационных параметров $\tau_1^*, \tau_2^*, \dots, \tau_n^*$, что позволило бы повысить скорость сходимости итераций. Оператор B можно

сделать самосопряженным, если положить его равным произведению сопряженных друг другу операторов:

$$B = (E + \omega R_1)(E + \omega R_2), \quad R_2^* = R_1, \quad (13)$$

где $\omega > 0$ — параметр. В качестве R_1 и R_2 можно взять операторы, имеющие треугольные (нижнюю R_1 и верхнюю R_2) матрицы, так что $R_1 + R_2 = R: H \rightarrow H$, $R^* = R > 0$. В частности, можно положить

$$R_1 + R_2 = A, \quad R_2^* = R_1. \quad (14)$$

Типичным является предположение

$$R \geq \delta E, \quad R_1 R_2 \leq \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta > 0. \quad (15)$$

Выбирая затем $\omega = 2/\sqrt{\delta\Delta}$ из условия $\min n_0(\varepsilon)$, находим параметры γ_1 , γ_2 и вычисляем параметры $\{\tau_k^*\}$. Определение y_{k+1} через y_k и f сводится к последовательному решению двух систем уравнений с нижней и верхней треугольными матрицами.

Построенный итерационный метод (9) с факторизованным оператором B вида (13) назовём попеременно-треугольным методом (ПТМ). ПТМ, очевидно, является универсальным, так как представление A в виде суммы $R_1 + R_2 = A$, $R_2^* = R_1$ возможно всегда. В случае разностной эллиптической задачи построение R_1 и R_2 не представляет труда. Так, например, $R_1 y \rightarrow \sum_{\alpha=1}^p \frac{y_{x_\alpha}}{h_\alpha}$,

$R_2 y \rightarrow -\sum_{\alpha=1}^p \frac{y_{x_\alpha}}{h_\alpha}$, если Ay — разностный $2p+1$ -точечный оператор

Лапласа, $Ay \rightarrow -\sum_{\alpha=1}^p y_{x_\alpha x_\alpha}$, h_α — шаг сетки по направлению Ox_α .

Этот метод является быстросходящимся. Если взять чебышевский набор $\{\tau_k^*\}$ и учесть (14), (15), то число итераций для ПТМ

$$n_0(\varepsilon) \geq \frac{1}{2\sqrt{2}\sqrt[4]{\eta}} \ln \frac{2}{\varepsilon}, \quad \eta = \frac{\delta}{\Delta}. \quad (16)$$

В частности, для модельной задачи имеем $n \geq n_0(\varepsilon) = 0,3 \ln \frac{2}{\varepsilon} / \sqrt{h}$.

Для случая произвольной области и уравнений с переменными коэффициентами целесообразно пользоваться модифицированным попеременно-треугольным методом (МПТМ), полагая

$$B = (\mathcal{D} + \omega R_1)\mathcal{D}^{-1}(\mathcal{D} + \omega R_2), \quad R_2^* = R_1, \quad \mathcal{D} = \mathcal{D}^* > 0, \quad (17)$$

где \mathcal{D} — произвольный оператор. Если вместо (15) выполнены

$$R \geq \delta \mathcal{D}, \quad R_1 \mathcal{D}^{-1} R_2 \leq \frac{\Delta}{4} \mathcal{D}, \quad \delta > 0, \quad \Delta > 0, \quad (18)$$

то оценка (16) сохраняет силу.

Здесь заданы δ и Δ , а выбираются оператор \mathcal{D} и параметр ω так, чтобы отношение $\xi = \gamma_1/\gamma_2$ было максимальным. На практике в качестве матрицы \mathcal{D} можно брать диагональную матрицу.

Укажем два примера эффективного применения МПТМ.

1) Задача Дирихле для уравнения Пуассона в двумерной области сложной формы; основная решетка в плоскости (x_1, x_2) равномерна с шагом h , схема пятиточечная. МПТМ при соответствующем выборе \mathcal{D} требует лишь на 4—5% больше итераций, чем для той же задачи в квадрате со стороной, равной диаметру области.

2) Для эллиптических уравнений с сильно меняющимися коэффициентами (отношение c_2/c_1 велико) МПТМ с соответствующим образом выбранным \mathcal{D} позволяет ослабить зависимость от c_2/c_1 .

На практике, помимо одношаговых (двухслойных) методов (9), применяются и двухшаговые (трехслойные) итерационные схемы. При оптимальных итерационных параметрах они по числу итераций сравнимы с чебышевской схемой с параметрами $\{\tau_k^*\}$ при $\xi \rightarrow 0$, однако более чувствительны по отношению к ошибкам в определении γ_1 и γ_2 . При условиях (11) целесообразно пользоваться чебышевской схемой (9) с параметрами $\{\tau_k^*\}$.

* * *

Для решения эллиптических задач весьма важную роль сыграл итерационный метод переменных направлений (МПН), развивавшийся, начиная с 1955 г., многими авторами. Однако он оказался экономичным лишь для очень узкого класса задач первой группы, когда выполнены условия $A = A_1 + A_2$, $A_\alpha = A_\alpha^* \geq 0$, $\alpha = 1, 2$, $A = A^* > 0$, $A_1 A_2 = A_2 A_1$. Если A_1 и A_2 перестановочны, то для МПН можно выбрать оптимальные итерационные параметры. Для модельной задачи с такими параметрами число итераций $n_0(\varepsilon) = O\left(\ln \frac{1}{h} \ln \frac{1}{\varepsilon}\right)$, а число действий $Q(\varepsilon) = O\left(\frac{1}{h^2} \ln \frac{1}{h} \ln \frac{1}{\varepsilon}\right)$, в то время как для прямых методов $Q = O\left(\frac{1}{h^2} \ln \frac{1}{h}\right)$. Прямые методы в этом случае более экономичны, чем МПН. Если A_1 и A_2 неперестановочны, то МПН требует $O\left(\frac{1}{h} \ln \frac{1}{\varepsilon}\right)$ итераций, в то время как для ПТМ достаточно $O\left(\frac{1}{\sqrt{h}} \ln \frac{1}{\varepsilon}\right)$ итераций. В случае трехмерных задач, когда $A = A_1 + A_2 + A_3$ даже в предположении попарной перестановочности A_1, A_2, A_3 МПН требует больше операций, чем ПТМ. Таким образом, МПН в значительной степени утратил свое значение.

* * *

Если оператор $A > 0$ не является самосопряженным, то не удается при помощи схемы (9) с набором параметров и самосопряженным оператором $B = B^* > 0$ построить итерационный

процесс с той же скоростью сходимости, что и чебышевский метод при $A = A^* > 0$. Все известные методы обладают меньшей скоростью сходимости. Здесь рассматривается метод простой итерации (гл. VI) с заданием априорной информации двух типов:

а) заданы параметры γ_1, γ_2 , входящие в условия (для простоты считаем $D = B = E$)

$$\gamma_1(x, x) \leq (Ax, x), \quad (Ax, Ax) \leq \gamma_2(Ax, x), \quad \gamma_1 > 0, \quad \gamma_2 > 0; \quad (19)$$

б) заданы три параметра $\gamma_1, \gamma_2, \gamma_3$, где γ_1 и γ_2 (при $D = B = E$) — граници симметричной части оператора A :

$$\gamma_1 E \leq A \leq \gamma_2 E, \quad \|A_1\| \leq \gamma_3, \quad \gamma_1 > 0, \quad \gamma_3 \geq 0, \quad (20)$$

где $A_1 = 0,5(A - A^*)$ — кососимметричная часть A .

Выбирая τ из условия минимума нормы оператора перехода или разрешающего оператора, во всех случаях получаем увеличение числа итераций по сравнению со случаем $A = A^*$.

* * *

Любой двухслойный итерационный метод, построенный на основе схемы (9), характеризуется операторами B и A , энергетическим пространством H_D , в котором доказывается сходимость метода, и набором параметров. Если оператор B фиксирован, то основной задачей является отыскание $\{\tau_k\}$.

При выборе параметров $\{\tau_k\}$ используется априорная информация об операторах схемы. Вид информации определяется свойствами операторов A , B и D . Так, для чебышевской схемы при $D = AB^{-1}A$, когда A и B — самосопряженные операторы, предполагается, что заданы постоянные γ_1, γ_2 в (11). В общем случае, когда $DB^{-1}A$ самосопряжен в H , то вместо (11) достаточно потребовать, чтобы $\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D$, $\gamma_1 > 0$. В несамосопряженном случае, когда $A \neq A^*$, а $B = B^* > 0$, используются либо два числа γ_1, γ_2 , либо три числа γ_1, γ_2 (входящие в (19)) и γ_3 — постоянная, входящая в оценку кососимметричной части оператора A . В ряде случаев нахождение постоянных γ_1, γ_2 и γ_3 с достаточной точностью может оказаться сложной самостоятельной задачей, требующей для своего решения специальных алгоритмов. Если априорная информация может быть получена ценой небольших вычислительных затрат или требуются многовариантные расчеты для уравнения $Au = f$ с разными правыми частями, то целесообразно найти однажды требуемые числа $\gamma_1, \gamma_2, \gamma_3$ и затем воспользоваться чебышевским методом или ПТМ. Если требуется решить лишь одну задачу $Au = f$ или если задано хорошее начальное приближение, а вычисления постоянных γ_1, γ_2 являются трудоемкими, — следует воспользоваться итерационными методами вариационного типа.

Для итерационных методов вариационного типа при вычислении параметров $\{\tau_k\}$ не надо знать γ_1, γ_2 . Эти методы используют лишь информацию общего вида

$$A = A^* > 0, \quad (DB^{-1}A)^* = DB^{-1}A. \quad (21)$$

Для определения y_{k+1} используется та же схема (9), меняется лишь формула для τ_{k+1} . Параметр τ_{k+1} находится из условия минимума в H_D нормы погрешности $z_{k+1} = y_{k+1} - u$, т. е. минимума функционала $I[y] = (D(y-u), y-u)$. Параметр τ_{k+1} вычисляется через y_k . Выбирая $D = A$, получим метод скорейшего спуска, а при $D = A^*A$ — метод минимальных невязок и т. д. Эти методы имеют ту же скорость сходимости, что и метод простой итерации (с точными постоянными γ_1, γ_2). Скорость сходимости итераций можно повысить, если отказаться от локальной (пошаговой) минимизации $\|z_{k+1}\|_D$ и выбирать параметры τ_k из условия минимизации нормы погрешности $\|z_n\|_D$ сразу за n шагов, т. е. при переходе от y_0 к y_n . Такой путь приводит к двухпараметрическим (при каждом k) трехслойным итерационным схемам сопряженных направлений (сопряженных градиентов, невязок, поправок или погрешностей), которые обладают такой же скоростью сходимости, что и чебышевский метод с параметрами $\{\tau_k^*\}$, вычисленными по точным значениям γ_1, γ_2 . Если $A = A^* > 0$, то можно построить процесс ускорения (\approx в 1,5—2 раза) сходимости двухслойных градиентных методов.

* * *

В общей теории итерационных методов не требуется знания конкретной структуры операторов задачи — используется лишь минимум информации общего функционального характера относительно операторов, например, условия (11). Выбор оператора B схемы (9) подчинен требованиям: 1) обеспечения наиболее быстрой сходимости метода (9), 2) экономичности обращения B . При построении B можно исходить из некоторого оператора $R = R^* > 0$ (регуляризатора), энергетически эквивалентного $A = A^* > 0$, $B = B^* > 0$:

$$c_1 R \leqslant A \leqslant c_2 R, \quad c_1 > 0, \quad \dot{\gamma}_1 B \leqslant R \leqslant \dot{\gamma}_2 B, \quad \dot{\gamma}_1 > 0. \quad (22)$$

Так что $\gamma_1 = c_1 \dot{\gamma}_1$, $\gamma_2 = c_2 \dot{\gamma}_2$. Для различных A можно выбрать один и тот же регуляризатор R . Наиболее распространен случай факторизованного оператора B , например,

$$B = (E + \omega R_1)(E + \omega R_2), \quad R_1 + R_2 = R, \quad (23)$$

где

$$R_1^* = R_2 > 0 — \text{для ПТМ}, \quad (24)$$

$$R_1^* = R_1 > 0, \quad R_2^* = R_2 > 0, \quad R_1 R_2 = R_2 R_1 — \text{для МПН}. \quad (25)$$

Чтобы применить теорию, надо найти γ_1 и γ_2 ; параметр $\omega > 0$ находится из условия $\min(\dot{\gamma}_1(\omega)/\dot{\gamma}_2(\omega))$. Если уравнение $Rw = F$ может быть решено экономичным прямым методом, то полагаем $B = R$ (например, в случае когда R — разностный оператор Лапласа, область — прямоугольник). Оператор B может не выписываться явно, а реализовываться в результате итерационного решения уравнения $Rw = r_k$, $r_k = Ay_k - f$ (двуухстушенчательный метод).

* * *

Для уравнений с неизвестноопределенными, вырожденными и комплексными операторами A можно рассматривать те же схемы (9). Однако, выбор оптимальных параметров усложняется, а скорость сходимости итераций уменьшается. Применение общей теории в этих особых случаях требует предварительной «обработки» исходной задачи. Оказывается возможным построить модификации как чебышевского метода, так и методов вариационного типа.

Если A — линейный вырожденный оператор, т. е. однородное уравнение $Au = 0$ имеет нетривиальное решение, то задача (9) при $B = E$ и любых τ_k всегда разрешима. Пусть $H^{(0)}$ — нулевое собственное подпространство оператора A , $H^{(1)}$ — ортогональное дополнение $H^{(0)}$ до H . Любой вектор $y \in H^{(0)}$ удовлетворяет уравнению $Ay = 0$. Если $f \in H^{(1)}$ и $y_0 \in H^{(1)}$, то и все итерации $y_k \in H^{(1)}$. Если выполнены условия

$$\gamma_1(y, y) \leq (Ay, y) \leq \gamma_2(y, y), \quad y \in H^{(1)}, \quad \gamma_1 > 0,$$

то можно пользоваться явной схемой (9) с чебышевскими параметрами $\{\tau_k^*\}$, найденными по γ_1 , γ_2 . При этом y_k сходится к нормальному решению, имеющему минимальную норму.

Если $f = f^{(0)} + f^{(1)}$ и $f^{(0)} \neq 0$, то под обобщенным нормальным решением уравнения $Au = f$ будем понимать решение уравнения $Au^{(1)} = f^{(1)}$, $u^{(1)} \in H^{(1)}$, имеющее минимальную норму. Справедлива оценка

$$\|y_n - u^{(1)}\| \leq \tilde{q}_n \|y_0 - u^{(1)}\|, \quad \tilde{q}_n = q_{n-1} (1 + (n-1)) \sqrt{\frac{1 - q_{n-1}^2}{\xi}},$$

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad y_n, y_0 \in H^{(1)},$$

если $\tau_1^*, \tau_2^*, \dots, \tau_{n-1}^*$ — чебышевские параметры, а $\tau_n^* = -\sum_{j=1}^{n-1} \tau_j^*$. Скорость сходимости понижается по сравнению со случаем невырожденного A с теми же γ_1 , γ_2 . Наряду с указанным модифицированным чебышевским методом возможны также и методы вариационного типа.

Общая теория позволяет исследовать неявную схему простой итерации для случая, когда H —комплексное гильбертово пространство, $A = \tilde{A} + qE$, \tilde{A} —эрмитов оператор, $q = q_1 + iq_2$ —комплексное число, и выбрать оптимальное значение итерационного параметра. Переход к методу переменных направлений также не представляет труда.

* * *

Результаты общей теории нетрудно использовать для решения разностных уравнений, аппроксимирующих краевые задачи для уравнений эллиптического типа. При этом легко формулировать общие правила решения разностных задач. Пусть дано разностное уравнение $Au = f$, где $A: H \rightarrow H$ —разностный оператор, определенный в пространстве H сеточных функций, заданных на сетке ω . Сначала изучаются общие свойства оператора A и устанавливается, например, его самосопряженность и положительность, $A = A^* > 0$, затем строится оператор $B = B^* > 0$ и вычисляются постоянные γ_1, γ_2 и, наконец, находятся $n = n_0(\epsilon)$ и параметры $\{\tau_k^*\}$.

Если речь идет о ПТМ с факторизованным оператором $B = (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2)$, то надо выбрать матрицу \mathcal{D} и постоянные δ, Δ (см. гл. X), зная δ и Δ , определим $\omega, \gamma_1, \gamma_2$ и т. д.

В книге приведено много примеров применения прямых и итерационных методов для решения конкретных разностных уравнений. В главе XV, в частности, рассматриваются методы решения разностных эллиптических уравнений в криволинейных координатах: в цилиндрической (r, z) и в полярной (r, ϕ) системах координат.

В гл. XIV рассматриваются многомерные задачи, схемы для уравнений теории упругости и др.

Важно отметить, что независимо от метода, который будет применен для решения данной разностной краевой задачи, ее предварительная обработка проводится по одному и тому же рецепту: сначала формируется оператор A , затем он изучается как оператор в пространстве H сеточных функций. После того как «сбор» информации о задаче закончен, принимается решение о выборе метода решения задачи с учетом всех обстоятельств, в том числе типа машины, наличия стандартных программ и др.

ГЛАВА I

ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ РАЗНОСТНЫХ УРАВНЕНИЙ

В главе изучаются общая теория линейных разностных уравнений, а также прямые методы решения уравнений с постоянными коэффициентами, дающие решение в замкнутом виде. В § 1 приведены общие понятия о сеточных уравнениях. § 2 посвящен общей теории линейных разностных уравнений m -го порядка. В § 3 рассмотрены методы решения уравнений с постоянными коэффициентами, а в § 4 эти методы используются для решения уравнений второго порядка. Решению сеточных задач на собственные значения для простейшего разностного оператора посвящен § 5.

§ 1. Сеточные уравнения. Основные понятия

1. Сетки и сеточные функции. Значительное число задач физики и техники приводят к дифференциальным уравнениям в частных производных (уравнениям математической физики). Установившиеся процессы различной физической природы описываются уравнениями эллиптического типа.

Точные решения краевых задач для эллиптических уравнений удается получить лишь в частных случаях. Поэтому эти задачи в основном решают приближенно. Одним из наиболее универсальных и эффективных методов, получивших в настоящее время широкое распространение для приближенного решения уравнений математической физики, является метод конечных разностей или метод сеток.

Суть метода состоит в следующем. Область непрерывного изменения аргументов (например, отрезок, прямоугольник и т. д.) заменяется дискретным множеством точек (узлов), которое называется *сеткой* или *решеткой*. Вместо функций непрерывного аргумента рассматриваются функции дискретного аргумента, определенные в узлах сетки и называемые *сеточными функциями*. Производные, входящие в дифференциальное уравнение и граничные условия, заменяются разностными производными; при этом краевая задача для дифференциального уравнения заменяется системой линейных или нелинейных алгебраических уравнений (сеточных или разностных уравнений). Такие системы часто называют *разностными схемами*.

Остановимся более подробно на основных понятиях метода сеток. Рассмотрим сначала простейшие примеры сеток.

Пример 1. Сетки в одномерной области. Пусть область изменения аргумента x есть отрезок $0 \leq x \leq l$. Разобьем этот отрезок на N равных частей длины $h = l/N$ точками $x_i = ih$, $i = 0, 1, \dots, N$. Множество этих точек называется *равномерной сеткой* на отрезке $[0, l]$ и обозначается $\bar{\omega} = \{x_i = ih, i = 0, 1, \dots, N, hN = l\}$, а число h — расстояние между точками (узлами) сетки $\bar{\omega}$ называется *шагом* сетки.

Для выделения части сетки $\bar{\omega}$ мы будем далее использовать следующие обозначения:

$$\begin{aligned}\omega &= \{x_i = ih, i = 1, 2, \dots, N-1, Nh = l\}, \\ \omega^+ &= \{x_i = ih, i = 1, 2, \dots, N, Nh = l\}, \\ \omega^- &= \{x_i = ih, i = 0, 1, \dots, N-1, Nh = l\}, \\ \gamma &= \{x_0 = 0, x_N = l\}.\end{aligned}$$

Отрезок $[0, l]$ можно разбить на N частей, вводя произвольные точки $0 = x_0 < x_1 < \dots < x_i < x_{i+1} < \dots < x_{N-1} < x_N = l$. В этом случае получим сетку $\bar{\omega} = \{x_i, i = 0, 1, \dots, N, x_0 = 0, x_N = l\}$ с шагом $h_i = x_i - x_{i-1}$ в узле x_i , $i = 1, 2, \dots, N$, который зависит от номера i узла x_i , т. е. является сеточной функцией $h_i = h(i)$.

Если $h_i \neq h_{i+1}$ хотя бы для одного номера i , то сетка $\bar{\omega}$ называется *неравномерной*. Если $h_i = h = l/N$, то получим построенную выше равномерную сетку. Для неравномерной сетки вводится средний шаг $\bar{h}_i = \bar{h}(i)$ в узле x_i , $\bar{h}_i = 0,5(h_i + h_{i+1})$, $1 \leq i \leq N-1$, $\bar{h}_0 = 0,5h_1$, $\bar{h}_N = 0,5h_N$. На бесконечной прямой $-\infty < x < \infty$ можно рассматривать сетки $\Omega = \{x_i = a + ih, i = 0, \pm 1, \pm 2, \dots\}$ с началом в любой точке $x = a$ и шагом h , состоящую из бесконечного числа узлов.

Пример 2. Сетка в двумерной области. Пусть область изменения аргументов $x = (x_1, x_2)$ есть прямоугольник $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ с границей Γ . На отрезках $0 \leq x_\alpha \leq l_\alpha$ построим равномерные сетки $\bar{\omega}_\alpha$ с шагами h_α :

$$\begin{aligned}\bar{\omega}_1 &= \{x_1(i) = ih_1, i = 0, 1, \dots, M, h_1M = l_1\}, \\ \bar{\omega}_2 &= \{x_2(j) = jh_2, j = 0, 1, \dots, N, h_2N = l_2\}.\end{aligned}$$

Множество узлов $x_{ij} = (x_1(i), x_2(j))$, имеющих координаты на плоскости $x_1(i)$ и $x_2(j)$, называется *сеткой* в *прямоугольнике* \bar{G} и обозначается $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), i = 0, 1, \dots, M, j = 0, 1, \dots, N, h_1M = l_1, h_2N = l_2\}$.

Сетка $\bar{\omega}$, очевидно, состоит из точек пересечения прямых $x_1 = x_1(i)$ и $x_2 = x_2(j)$.

Построенная сетка $\bar{\omega}$ равномерна по каждому из переменных x_1 и x_2 . Если хотя бы одна из сеток $\bar{\omega}_\alpha$ неравномерна, то сетка $\bar{\omega}$ называется *неравномерной*. Если $h_1 = h_2$, то сетка называется *квадратной*, иначе — *прямоугольной*.

Точки $\bar{\omega}$, принадлежащие Γ , называются *граничными* и их объединение образует границу сетки: $\gamma = \{x_{ij} \in \Gamma\}$.

Чтобы описать структуру сетки $\bar{\omega}$, удобно использовать запись $\bar{\omega} = \bar{\omega}_1 \times \bar{\omega}_2$, т. е. представлять $\bar{\omega}$ как топологическое произведение сеток $\bar{\omega}_1$ и $\bar{\omega}_2$. Используя введенные в примере 1 обозначения ω^+ , ω^- и ω , можно выделить части сетки $\bar{\omega}$ в прямоугольнике, например:

$$\begin{aligned}\omega_1 \times \omega_2^+ &= \{x_{ij} = (ih_1, jh_2), i = 1, 2, \dots, M-1, j = 1, 2, \dots, N\}, \\ \omega_1^- \times \bar{\omega}_2 &= \{x_{ij} = (ih_1, jh_2), i = 0, 1, \dots, M-1, j = 0, 1, \dots, N\}.\end{aligned}$$

Рассмотрим теперь понятие сеточной функции. Пусть $\bar{\omega}$ — сетка, введенная в одномерной области, а x_i — узлы сетки. Функция $y = y(x_i)$ дискретного аргумента x_i называется *сеточной функцией*, определенной на сетке $\bar{\omega}$. Аналогично определяется сеточная функция на любой сетке $\bar{\omega}$, введенной в области изменения непрерывного аргумента. Например, если x_{ij} — узел сетки $\bar{\omega}$ в двумерной области, то $y = y(x_{ij})$. Очевидно, что сеточные функции можно рассматривать и как функции целочисленного аргумента, являющегося номером узла сетки. Так, можно писать $y = y(x_i) = y(i)$, $y = y(x_{ij}) = y(i, j)$. Иногда мы будем использовать для обозначения сеточных функций следующую запись: $y(x_i) = y_i$, $y(x_{ij}) = y_{ij}$.

Сеточную функцию y_i можно представить в виде вектора, рассматривая значения функции как компоненты вектора $\mathbf{Y} = (y_0, y_1, \dots, y_N)$. В этом примере y_i задана на сетке $\bar{\omega} = \{x_i, i = 0, 1, \dots, N\}$, содержащей $N+1$ узел, а вектор \mathbf{Y} имеет размерность $N+1$. Если $\bar{\omega}$ — сетка в прямоугольнике ($\bar{\omega} = \{x_{ij} = (ih_1, jh_2), i = 0, 1, \dots, M, j = 0, 1, \dots, N\}$), то сеточной функции y_{ij} , заданной на $\bar{\omega}$, соответствует вектор $\mathbf{Y} = (y_{00}, \dots, y_{M0}, y_{01}, \dots, y_{M1}, \dots, y_{0N}, \dots, y_{MN})$ размерности $(M+1)(N+1)$. Узлы сетки $\bar{\omega}$ при этом считаются упорядоченными по строкам сетки.

Мы рассмотрели скалярные сеточные функции, т. е. такие функции, значениями которых в каждом узле сетки являются числа. Приведем теперь примеры *векторных сеточных функций*, значениями которых в узле являются векторы. Если в рассматриваемом выше примере обозначить через $\mathbf{Y}(x_2(j)) = \mathbf{Y}_j$ вектор, компонентами которого являются значения сеточной функции y_{ij} в узлах $x_{0j}, x_{1j}, \dots, x_{Mj}$ j -й строки сетки $\bar{\omega}$: $\mathbf{Y}_j = (y_{0j}, y_{1j}, \dots, y_{Mj})$, $j = 0, 1, \dots, N$, то мы получим векторную сеточную функцию \mathbf{Y}_j , определенную на сетке $\bar{\omega}_2 = \{x_2(j) = jh_2, j = 0, 1, \dots, N\}$.

Если функция, заданная на сетке, принимает комплексные значения, то такая сеточная функция называется *комплексной*.

2. Разностные производные и некоторые разностные тождества. Пусть задана сетка $\bar{\omega}$. Множество всех сеточных функций, заданных на $\bar{\omega}$, образует векторное пространство с определенным очевидным образом сложением функций и умножением функции на число. На пространстве сеточных функций можно определить разностные или сеточные операторы. Оператор Λ , преобразующий сеточную функцию y в сеточную функцию $f = \Lambda y$, называется *разностным* или *сеточным* оператором. Множество узлов сетки,

используемое при написании разностного оператора в узле сетки, называется *шаблоном* этого оператора.

Простейшим разностным оператором является оператор разностного дифференцирования сеточной функции, который порождает разностные производные. Определим разностные производные.

Пусть Ω —равномерная сетка с шагом h , введенная на прямой $-\infty < x < \infty$: $\Omega = \{x_i = a + ih, i = 0, \pm 1, \pm 2, \dots\}$. Разностные производные первого порядка для сеточной функции $y_i = y(x_i)$, $x_i \in \Omega$ определяются формулами

$$\Lambda_1 y_i = y_{\bar{x}, i} = \frac{y_i - y_{i-1}}{h}, \quad \Lambda_2 y_i = y_{x, i} = \frac{y_{i+1} - y_i}{h} \quad (1)$$

и называются *левой* и *правой производными* соответственно. Используется также *центральная производная*

$$\Lambda_3 y_i = y_{\hat{x}, i} = \frac{y_{i+1} - y_{i-1}}{2h} = 0,5 (\Lambda_1 + \Lambda_2) y_i. \quad (2)$$

Если сетка неравномерна, то для разностных производных первого порядка применяют следующие обозначения:

$$y_{\bar{x}, i} = \frac{y_i - y_{i-1}}{h_i}, \quad y_{x, i} = \frac{y_{i+1} - y_i}{h_{i+1}}, \quad y_{\hat{x}, i} = \frac{y_{i+1} - y_i}{\tilde{h}_i}, \quad (3)$$

$$y_{\hat{x}, i} = 0,5 (y_{\bar{x}, i} + y_{x, i}), \quad \tilde{h}_i = 0,5 (h_i + h_{i+1}).$$

Из определений (1) и (3) вытекают следующие соотношения:

$$y_{x, i} = y_{\bar{x}, i+1}, \quad (4)$$

$$y_{x, i} = \frac{\tilde{h}_i}{h_{i+1}} y_{\hat{x}, i}, \quad (5)$$

а также равенства

$$y_i = y_{i+1} - h_{i+1} y_{x, i} = y_{i-1} + h_i y_{\bar{x}, i}. \quad (6)$$

Разностные операторы Λ_1 , Λ_2 и Λ_3 имеют шаблоны, состоящие из двух точек, и используются при аппроксимации первой производной $Lu = u'$ функции $u = u(x)$ одного переменного. При этом операторы Λ_1 и Λ_2 аппроксимируют оператор L на гладких функциях с погрешностью $O(h)$, а Λ_3 —с погрешностью $O(h^2)$.

Разностные производные n-го порядка определяются как сеточные функции, получаемые путем вычисления первой разностной производной от функций, являющейся разностной производной $n-1$ -го порядка. Приведем примеры разностных производных второго порядка:

$$y_{\bar{x}\bar{x}, i} = \frac{y_{\bar{x}, i+1} - y_{\bar{x}, i}}{h} = \frac{1}{h^2} (y_{i-1} - 2y_i + y_{i+1}),$$

$$y_{\hat{x}\hat{x}, i} = \frac{y_{\hat{x}, i+1} - y_{\hat{x}, i-1}}{2h} = \frac{1}{4h^2} (y_{i-2} - 2y_i + y_{i+2}),$$

$$y_{\bar{x}\hat{x}, i} = \frac{1}{\tilde{h}_i} (y_{\bar{x}, i+1} - y_{\bar{x}, i}) = \frac{1}{\tilde{h}_i} (y_{x, i} - y_{\bar{x}, i}) = \frac{1}{\tilde{h}_i} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right),$$

которые используются при аппроксимации второй производной $Lu = u''$ функции $u = u(x)$. В случае равномерной сетки погрешность аппроксимации равна $O(h^2)$. Соответствующие разностные операторы имеют трехточечный шаблон. При аппроксимации четвертой производной $Lu = u^{IV}$ используется разностная производная четвертого порядка $y_{xxx}, i = \frac{1}{h^4}(y_{i-2} - 4y_{i-1} + 6y_i - 4y_{i+1} + y_{i+2})$.

Аналогично при аппроксимации производных n -го порядка используются разностные производные n -го порядка.

Не представляет труда определить разностные производные от сеточных функций нескольких переменных.

Для преобразования выражений, содержащих разностные производные сеточных функций, нам потребуются формулы разностного дифференцирования произведения сеточных функций и формулы суммирования по частям. Эти формулы являются аналогом соответствующих формул дифференциального исчисления.

1) Формулы разностного дифференцирования произведения. Используя определения разностных производных (3), нетрудно проверить, что имеют место тождества:

$$\begin{aligned} (uv)_{\bar{x}, i} &= u_{\bar{x}, i}v_{i-1} + u_iv_{\bar{x}, i} = u_{\bar{x}, i}v_i + u_{i-1}v_{\bar{x}, i} = \\ &\quad = u_{\bar{x}, i}v_i + u_iv_{\bar{x}, i} - h_i u_{\bar{x}, i}v_{\bar{x}, i}, \\ (uv)_{x, i} &= u_{x, i}v_{i+1} + u_iv_{x, i} = u_{x, i}v_i + u_{i+1}v_{x, i} = \\ &\quad = u_{x, i}v_i + u_iv_{x, i} + h_{i+1}u_{x, i}v_{x, i}, \\ (uv)_{\hat{x}, i} &= u_{\hat{x}, i}v_{i+1} + u_iv_{\hat{x}, i} = u_{\hat{x}, i}v_i + u_{i+1}v_{\hat{x}, i} = \\ &\quad = u_{\hat{x}, i}v_i + u_iv_{\hat{x}, i} + h_i u_{\hat{x}, i}v_{\hat{x}, i}. \end{aligned}$$

Используя (4), (5), последнее тождество можно записать в виде

$$(uv)_{\hat{x}, i} = u_{\hat{x}, i}v_i + \frac{h_{i+1}}{h_i} u_{i+1}v_{\bar{x}, i+1}. \quad (7)$$

2) Формулы суммирования по частям. Умножая (7) на $\frac{h_i}{h_{i+1}}$ и суммируя получаемое соотношение по i от $m+1$ до $n-1$, находим, что

$$\begin{aligned} \sum_{i=m+1}^{n-1} (uv)_{\hat{x}, i} \frac{h_i}{h_{i+1}} &= u_nv_n - u_{m+1}v_{m+1} = \\ &= \sum_{i=m+1}^{n-1} u_{\hat{x}, i}v_i \frac{h_i}{h_{i+1}} + \sum_{i=m+1}^{n-1} u_{i+1}v_{\bar{x}, i+1} \frac{h_i}{h_{i+1}}. \end{aligned}$$

Используя (6), получим соотношение $v_{m+1} = v_m + h_{m+1}v_{x, m} = v_m + h_{m+1}v_{\bar{x}, m+1}$, которое подставим в найденное выше равенство. В результате будем иметь

$$u_nv_n - u_{m+1}v_m = \sum_{i=m+1}^{n-1} u_{\hat{x}, i}v_i \frac{h_i}{h_{i+1}} + \sum_{i=m}^{n-1} u_{i+1}v_{\bar{x}, i+1} \frac{h_i}{h_{i+1}}.$$

Замена индекса суммирования $i' = i - 1$ во второй сумме правой части дает следующую формулу суммирования по частям:

$$\sum_{i=m+1}^{n-1} u_{\hat{x}, i} v_i \bar{h}_i = - \sum_{i=m+1}^n u_i v_{\hat{x}, i} \bar{h}_i + u_n v_n - u_{m+1} v_m. \quad (8)$$

Используя (6), легко получить из (8) еще одну формулу суммирования по частям

$$\sum_{i=m+1}^{n-1} u_{\hat{x}, i} v_i h_i = - \sum_{i=m}^{n-1} u_i v_{\hat{x}, i} \bar{h}_i + u_{n-1} v_n - u_m v_m. \quad (9)$$

Из формулы (8) следует, что функция u_i должна быть определена для $m+1 \leq i \leq n$, а функция v_i — для $m \leq i \leq n$. Пусть теперь y_i — сеточная функция, заданная для $m \leq i \leq n$. Тогда функция $u_i = y_{\hat{x}, i}$ определена для $m+1 \leq i \leq n$. Подставляя u_i в (8), получим следующее тождество:

$$\sum_{i=m+1}^{n-1} y_{\hat{x}\hat{x}, i} v_i \bar{h}_i = - \sum_{i=m+1}^n y_{\hat{x}, i} v_{\hat{x}, i} \bar{h}_i + y_{\hat{x}, n} v_n - y_{x, m} v_m. \quad (10)$$

Имеет место

Лемма 1. Пусть на произвольной неравномерной сетке $\bar{\omega} = \{x_i, i = 0, 1, \dots, N, x_0 = 0, x_N = l\}$ задана сеточная функция y_i , обращающаяся в нуль при $i = 0, i = N$. Для этой функции имеет место равенство

$$\sum_{i=1}^{N-1} y_{\hat{x}\hat{x}, i} y_i \bar{h}_i = - \sum_{i=1}^N (y_{\hat{x}, i})^2 h_i.$$

Утверждение леммы 1 очевидным образом следует из тождества (10).

Следствие. Если $\bar{\omega}$ — равномерная сетка, $y_0 = y_N = 0$ и $y_i \neq 0$, то $\sum_{i=1}^{N-1} y_{\hat{x}\hat{x}, i} y_i h_i = - \sum_{i=1}^N y_{\hat{x}, i}^2 h_i < 0$.

На этом рассмотрение разностных формул мы заканчиваем. Некоторые другие формулы будут рассмотрены в гл. V.

Полученные тождества используются не только для преобразования разностных выражений. Они часто применяются, например, при вычислении различного вида конечных сумм и рядов.

Приведем пример. Требуется вычислить сумму $S_n = \sum_{i=1}^{n-1} ia^i$, $a \neq 1$. Введем следующие сеточные функции, заданные на равномерной сетке $\bar{\omega} = \{x_i = i, i = 0, 1, \dots, N, h = 1\}$:

$$v_i = i, \quad u_i = (a^i - a^n)/(a - 1). \quad (11)$$

На указанной сетке формула суммирования по частям (8) для любых сеточных функций имеет вид ($m=0$)

$$\sum_{i=1}^{n-1} u_{x,i} v_i = - \sum_{i=1}^n u_i v_{\bar{x},i} + u_n v_n - u_1 v_0.$$

Учитывая, что для функций (11) верны соотношения $v_0 = u_n = 0$, $v_{\bar{x},i} = 1$, $u_{x,i} = a^i$, отсюда получим

$$S_n = \sum_{i=1}^{n-1} i a^i = - \sum_{i=1}^n \frac{a^i - a^n}{a-1} = \frac{a^n (n(a-1) - a) + a}{(a-1)^2}.$$

Искомая сумма найдена.

3. Сеточные и разностные уравнения. Пусть $y_i = y(i)$ сеточная функция дискретного аргумента i . Значения сеточной функции $y(i)$ в свою очередь образуют дискретное множество. На этом множестве можно определять сеточную функцию, приравнивая которую нулю получаем уравнение относительно сеточной функции $y(i)$ — *сеточное уравнение*. Специальным случаем сеточного уравнения является *разностное уравнение*. Именно разностные уравнения будут основным объектом исследования в нашей книге.

Сеточные уравнения получаются при аппроксимации на сетке интегральных и дифференциальных уравнений.

Приведем сначала примеры разностных аппроксимаций обыкновенных дифференциальных уравнений.

Так, дифференциальные уравнения первого порядка $\frac{du}{dx} = f(x)$, $x > 0$ мы заменяем разностным уравнением первого порядка $\frac{y_{i+1} - y_i}{h} = f(x_i)$, $x_i = ih$, $i = 0, 1, \dots$ или $y_{i+1} = y_i + hf(x_i)$, где h — шаг сетки $\omega = \{x_i = ih, i = 0, 1, \dots\}$. Искомой функцией является сеточная функция $y_i = y(i)$.

При разностной аппроксимации уравнения второго порядка $\frac{d^2u}{dx^2} = f(x)$ мы получаем разностное уравнение второго порядка $y_{i+1} - 2y_i + y_{i-1} = \varphi_i$, $\varphi_i = h^2 f_i$, $f_i = f(x_i)$, $x_i = ih$. Если аппроксимировать на трехточечном шаблоне (x_{i-1}, x_i, x_{i+1}) уравнение общего вида $(ku')' + ru' - qu = f(x)$, то получим разностное уравнение второго порядка с переменными коэффициентами вида $a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -\varphi_i$, $i = 0, 1, \dots$, где a_i , c_i , b_i , φ_i — заданные сеточные функции, а y_i — искомая сеточная функция.

Аппроксимация на сетке уравнения четвертого порядка $(ku'')'' = f(x)$ приводит к разностному уравнению четвертого порядка; оно имеет вид

$$a_i^{(2)} y_{i-2} + a_i^{(1)} y_{i-1} + c_i y_i + b_i^{(1)} y_{i+1} + b_i^{(2)} y_{i+2} = \varphi_i.$$

Для разностной аппроксимации производных u' , u'' , u''' можно пользоваться шаблонами с большим числом узлов. Это приводит к разностным уравнениям более высокого порядка.

Линейное уравнение относительно сеточной функции $y(i)$ (функции целочисленного аргумента i)

$$a_0(i)y(i) + a_1(i)y(i+1) + \dots + a_m(i)y(i+m) = f(i), \quad (12)$$

где $a_0(i) \neq 0$ и $a_m(i) \neq 0$, а $f(i)$ — заданная сеточная функция, называется *разностным уравнением m -го порядка*.

Если (12) не содержит $y(i)$, но содержит $y(i+1)$, то замена независимого переменного $i+1$ на i' приводит это уравнение к уравнению порядка $m-1$.

В этом состоит одно из отличий сеточных уравнений от дифференциальных, где замена независимого переменного порядка уравнения не меняет.

Пусть $F(i, y(i), y(i+1), \dots, y(i+m))$ — нелинейная сеточная функция. Тогда $F(i, y(i), y(i+1), \dots, y(i+m)) = 0$ является *нелинейным разностным уравнением m -го порядка*, если F явно зависит от $y(i)$ и $y(i+m)$.

Для удобства сравнения с дифференциальными уравнениями введем *разности (правые)* для *сеточных функций*: $\Delta y_i = y_{i+1} - y_i$, $\Delta^2 y_i = \Delta(\Delta y_i)$, ..., $\Delta^{k+1} y_i = \Delta(\Delta^k y_i)$, $k = 1, 2, \dots$

Тогда (12) можно записать в виде

$$\alpha_0(i)y(i) + \alpha_1(i)\Delta y_i + \dots + \alpha_m(i)\Delta^m y_i = f_i, \quad (12')$$

где $\alpha_m(i) = a_m(i) \neq 0$ и, кроме того, коэффициент a_0 при y_0 также отличен от нуля.

Разностное уравнение (12') является формальным аналогом дифференциального уравнения m -го порядка:

$$\alpha_0 u + \alpha_1 \frac{du}{dx} + \dots + \alpha_{m-1} \frac{d^{m-1}u}{dx^{m-1}} + \alpha_m \frac{d^m u}{dx^m} = f(x),$$

где $\alpha_m \neq 0$, $\alpha_k = \alpha_k(x)$, $k = 0, 1, \dots, m$. Пусть дана сетка $\omega = \{x_i = ih, i = 0, 1, \dots\}$. Если обозначить

$$y_{x,i} = \frac{y_{i+1} - y_i}{h}, \quad y_{xx,i} = (y_x)_{x,i}, \dots, y_x^{(k)} = \underbrace{y_{x\dots x,i}}_{k \text{ раз}},$$

так что $y_x^{(k)} = (y_x^{(k-1)})_x$, $k \geq 1$, $y_{x,i}^{(0)} = y(i)$, то $y(i+k)$ выразится через $y(i)$, $y_x^{(1)}$, ..., $y_x^{(k-1)}$, например, $y(i+3) = y(i) + 3hy_{x,i} + 3h^2y_{xx,i} + h^3y_{xxx,i}$.

Тогда уравнение (12) запишется в виде

$$\bar{\alpha}_0 y(i) + \bar{\alpha}_1 y_x(i) + \dots + \bar{\alpha}_{m-1} y_x^{(m-1)}(i) + \bar{\alpha}_m y_x^{(m)}(i) = f_i,$$

где $\bar{\alpha}_m = a_m \neq 0$ и $a_0 \neq 0$. Здесь аналогия с дифференциальным уравнением m -го порядка очевидна.

Аналогично определяется разностное уравнение относительно сеточной функции $y_{i_1, i_2} = y(i_1, i_2)$ двух дискретных аргументов и вообще любого числа аргументов. Например, пятиточечная разностная схема «крест» для уравнения Пуассона $\Delta u =$

$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2)$ на сетке $\omega = \{x_i = (i_1 h_1, i_2 h_2), i_1, i_2 = 0, 1, \dots\}$ имеет вид

$$\frac{y(i_1-1, i_2) - 2y(i_1, i_2) + y(i_1+1, i_2)}{h_1^2} + \frac{y(i_1, i_2-1) - 2y(i_1, i_2) + y(i_1, i_2+1)}{h_2^2} = f_{i_1 i_2}$$

и представляет собой разностное уравнение второго порядка по каждому из дискретных аргументов i_1 и i_2 .

Сеточное уравнение *общего* вида получается при аппроксимации интегрального уравнения $u(x) = \int_0^1 K(x, s) u(s) ds + f(x)$, $0 \leq x \leq 1$, на сетке $\bar{\omega} = \{x_i = ih, i = 0, 1, \dots, N, hN = 1\}$. Заменим интеграл суммой

$$\int_0^1 K(x, s) u(s) ds \approx h \sum_{j=0}^N \alpha_j K(x, jh) u(jh),$$

где α_j — коэффициент квадратурной формулы, и вместо интегрального уравнения напишем сеточное уравнение

$$y_i = \sum_{j=0}^N \alpha_j K(ih, jh) y_j + f_i, \quad i = 0, 1, \dots, N,$$

где суммирование производится по всем узлам сетки $\bar{\omega}$, а неизвестной является сеточная функция y_i .

Сеточное уравнение можно записать в виде

$$\sum_{j=0}^N c_{ij} y_j = f_i, \quad i = 0, 1, \dots, N. \quad (13)$$

Оно содержит все значения y_0, y_1, \dots, y_N сеточной функции. Его можно трактовать как разностное уравнение порядка N , равного числу узлов сетки минус единица.

Разностное уравнение (12) m -го порядка является специальным видом сеточного уравнения, когда матрица (c_{ij}) имеет отличные от нуля элементы лишь на m диагоналях, параллельных главной диагонали.

В общем случае под i можно понимать не только индекс $i = 0, 1, \dots$, но и мультииндекс, т. е. вектор $i = (i_1, i_2, \dots, i_p)$ с целочисленными компонентами $i_\alpha = 0, 1, 2, \dots, \alpha = 1, 2, \dots, p$, причем $i \in \omega$, где ω — сетка.

Линейное сеточное уравнение имеет вид

$$\sum_{j \in \omega} c_{ij} y_j = f_i, \quad i \in \omega, \quad (14)$$

где суммирование проводится по всем узлам сетки ω , f_i — заданная, y_i — искомая сеточные функции.

Если перенумеровать все узлы сетки, то можно писать $y_i = y(i)$, где i — номер узла, $i = 0, 1, 2, \dots, N$. Тогда сеточное уравнение (14) примет вид (13).

Очевидно, что это система линейных алгебраических уравнений порядка $N+1$ с матрицей (c_{ij}) . Таким образом, любую систему линейных алгебраических уравнений можно трактовать как сеточное уравнение и обратно.

Если $y(i)$ есть векторная сеточная функция, то говорят о сеточном (разностном) векторном уравнении m -го порядка.

Пусть $F(i, y_0, y_1, \dots, y_N)$ — заданная функция (вообще говоря, нелинейная) $N+2$ аргументов i, y_0, y_1, \dots, y_N . Приравнивая ее нулю, получим нелинейное сеточное уравнение $F(i, y_0, y_1, \dots, y_N) = 0, i = 0, 1, \dots, N$, решением которого называется сеточная функция $y(i)$, обращающая это уравнение в тождество.

Рассмотрим сеточную функцию $\mathcal{F}(i) = F(i, y_0, y_1, \dots, y_N), i = 0, 1, \dots, N$. Отсюда видно, что функция F задает некоторый сеточный оператор, который переводит сеточную функцию $y(i)$ в сеточную функцию $\mathcal{F}(i)$.

Если F — линейная функция, то мы получаем уравнение (14), которое, очевидно, можно записать в операторной форме $Ay = f$, где A — линейный оператор с матрицей (c_{ij}) , а y — вектор в пространстве сеточных функций.

Если коэффициенты c_{ij} не зависят от j , то (14) называют сеточным уравнением с постоянными коэффициентами.

Хотя в этой книге основное внимание уделяется численному решению разностных уравнений, получающихся при разностной аппроксимации дифференциальных уравнений эллиптического типа, итерационные методы применимы для любого линейного сеточного уравнения, т. е. для любой системы линейных алгебраических уравнений. Поэтому излагаемая здесь теория итерационных методов носит общий характер. Специфика сеточных уравнений в том, что это система высокого порядка, причем порядок уравнения увеличивается при сгущении сетки (число неизвестных равно числу N узлов сетки, $N = O\left(\frac{1}{h^p}\right)$ в p -мерном случае, h — шаг сетки).

4. Задача Коши и краевые задачи для разностных уравнений. Приведем некоторые дополнительные примеры разностных уравнений и остановимся на постановке задач для разностных уравнений.

Заметим, что простейшими примерами разностных уравнений первого порядка являются формулы для членов арифметической и геометрической прогрессий:

$$y_{i+1} = y_i + d, \quad y_{i+1} = qy_i, \quad i = 0, 1, \dots$$

Решение уравнения первого порядка может быть найдено, если задано начальное условие при $i = 0$ (задача Коши).

Решение $y(i+m)$ разностного уравнения m -го порядка определяется полностью значениями $y(i)$, заданными в m произвольных, но расположенных подряд точках $i_0, i_0+1, \dots, i_0+m-1$. В самом деле, так как $a_m(i) \neq 0$, то из (12) находим $y(i+m) = -b_{m-1}(i)y(i+m-1) + \dots + b_0(i)y(i) + \varphi(i)$. Полагая здесь последовательно $i = i_0, i_0+1, \dots$, найдем значения $y(i)$ при $i \geq i_0$. Аналогично, выражая из (12) $y(i)$ через $y(i+1), \dots, y(i+m)$ и полагая последовательно $i = i_0-1, i_0-2, \dots$, найдем $y(i)$ для $i \leq i_0-1$. Если в уравнении (12) требуется определить $y(i)$ при $i \geq 0$, то достаточно задать значение в m узлах (начальные условия) $y(0) = y_0, y(1) = y_1, \dots, y(m-1) = y_{m-1}$.

Присоединяя эти условия к уравнению (12), получаем задачу Коши или задачу с начальными данными для разностного уравнения m -го порядка.

Для уравнений первого порядка ($m=1$), как мы видели, достаточно задать одно начальное условие.

Нелинейные разностные уравнения получаются при решении нелинейных дифференциальных уравнений. Рассмотрим, например, дифференциальное уравнение

$$\frac{du}{dx} = f(x, u), \quad x > 0, \quad u(0) = \mu_1$$

(задача Коши). Заменяя его схемой Эйлера (явной схемой), получим разностное уравнение первого порядка $y_{i+1} = y_i + hf(x_i, y_i)$, $i \geq 0$, $y_0 = \mu_1$.

Если производную du/dx при $x = x_i = ih$ заменить левым разностным отношением, то получим нелинейное относительно y_i разностное уравнение первого порядка $y_i = y_{i-1} + hf(x_i, y_i)$, $i > 0$, $y_0 = \mu_1$. Для определения y_i надо решить нелинейное уравнение $\varphi(y_i) = y_i - hf(x_i, y_i) = y_{i-1}$.

Рассмотрим теперь пример разностного уравнения второго порядка. Пусть требуется вычислить интегралы

$$I_k(\varphi) = \int_0^\pi \frac{\cos k\psi - \cos k\varphi}{\cos \psi - \cos \varphi} d\psi, \quad k = 0, 1, 2, \dots$$

Прежде всего заметим, что $I_0(\varphi) = 0$, $I_1(\varphi) = \pi$. Преобразуем выражение $[\cos(k+1)\psi - \cos(k+1)\varphi] + [\cos(k-1)\psi - \cos(k-1)\varphi] = 2\cos k\psi \cos \psi - 2\cos k\varphi \cos \varphi = 2(\cos k\psi - \cos k\varphi) \cos \varphi + 2(\cos \psi - \cos \varphi) \cos k\psi$. Используя его, получаем

$$I_{k+1}(\varphi) + I_{k-1}(\varphi) = 2\cos \varphi I_k(\varphi) + 2 \int_0^\pi \cos k\psi d\psi = 2\cos \varphi I_k(\varphi), \quad k \geq 1.$$

Таким образом, вычисление интегралов $I_k(\varphi)$ сводится к решению задачи Коши для разностного уравнения второго порядка

$$I_{k+1}(\varphi) - 2\cos \varphi I_k(\varphi) + I_{k-1}(\varphi) = 0, \quad k \geq 1, \quad I_0(\varphi) = 0, \quad I_1(\varphi) = \pi. \quad (15)$$

Рассмотрим еще один пример. Требуется найти решение краевой задачи для системы обыкновенных дифференциальных уравнений первого порядка

$$\frac{du}{dx} = Au + f(x), \quad 0 < x < l, \quad (16)$$

$Bu = \mu_1$ при $x=0$, $Cu = \mu_2$ при $x=l$. Здесь $u(x) = (u_1(x), u_2(x), \dots, u_M(x))$ — вектор-функция размерности M , $A = A(x)$ — квадратная матрица размером $M \times M$, B и C — прямоугольные матрицы размером $M_1 \times M$ и $M_2 \times M$ соответственно, $M_1 + M_2 = M$. Векторы $f(x)$, μ_1 , μ_2 заданы и имеют размерности M , M_1 и M_2 соответственно.

Вводя на отрезке $0 \leq x \leq l$ равномерную сетку $\bar{\omega} = \{x_i = ih, i = 0, 1, \dots, N, h = l/N\}$ и определяя на ней сеточную вектор-функцию $\bar{Y}_i = (y_1(i), y_2(i), \dots, y_M(i))$, поставим в соответствие задаче (16) простейшую разностную схему

$$\begin{aligned} Y_{i+1} - (E + hA_i) Y_i &= F_i, & 0 \leq i \leq N-1, \\ BY_0 &= \mu_1, \quad CY_N = \mu_2, \end{aligned} \quad (17)$$

где $F_i = hf(x_i)$. Это пример линейного векторного разностного уравнения первого порядка с M_1 условиями при $i=0$ и M_2 условиями при $i=N$. Таким образом, для системы разностных уравнений первого порядка мы имеем краевую задачу.

Для уравнений второго порядка наиболее типичны краевые задачи. Рассмотрим, например, первую краевую задачу

$$\frac{d^2u}{dx^2} - q(x)u = -f(x), \quad 0 < x < l, \quad u(0) = \mu_1, \quad u(l) = \mu_2, \quad q(x) \geq 0. \quad (18)$$

Выберем сетку $\bar{\omega} = \{x_i = ih, i = 0, 1, \dots, N, h = l/N\}$ и поставим задаче (18) в соответствие разностную краевую задачу

$$y_{xx,i} - d_i y_i = -\varphi_i, \quad 0 < i < N, \quad y_0 = \mu_1, \quad y_N = \mu_2, \quad (19)$$

где $d_i = q(x_i)$, $\varphi_i = f(x_i)$ для гладких $q(x)$, $f(x)$. Эта задача является частным случаем краевой задачи для разностного уравнения второго порядка

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = \varphi_i, \quad 1 \leq i \leq N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2 \quad (20)$$

при $a_i = b_i = 1/h^2$, $c_i = d_i + 2/h^2$.

Разностную задачу (20) можно записать в виде

$$\mathcal{A}Y = F, \quad (21)$$

где $Y = (y_1, y_2, \dots, y_{N-1})$ — неизвестный, $F = (\varphi_1 + \frac{1}{h^2}\mu_1, \varphi_2, \dots, \varphi_{N-2}, \varphi_{N-1} + \frac{1}{h^2}\mu_2)$ — известный векторы размерности $N-1$,

\mathcal{A} — квадратная трехдиагональная матрица вида

$$\mathcal{A} = \begin{vmatrix} c_1 & -b_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -a_2 & c_2 & -b_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & -a_3 & c_3 & -b_3 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & c_{N-3} & -b_{N-3} & 0 \\ 0 & 0 & 0 & 0 & \dots & -a_{N-2} & c_{N-2} & -b_{N-2} \\ 0 & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & c_{N-1} \end{vmatrix}. \quad (22)$$

Отсюда видно, что краевая задача для разностного уравнения второго порядка (20) представляет собой систему линейных алгебраических уравнений специального вида. Если задача Коши для разностного уравнения второго порядка разрешима всегда, то первая краевая задача (20) разрешима для любой правой части лишь тогда, когда матрица \mathcal{A} системы (21) не вырождена.

Краевые задачи для разностных уравнений m -го порядка приводят к системам линейных алгебраических уравнений с матрицей, имеющей не более $m+1$ ненулевых элементов в каждой строке.

При аппроксимации уравнений в частных производных мы приходим также к системе разностных или просто алгебраических уравнений со специальной матрицей. Так как число неизвестных в такой системе обычно равно числу узлов сетки, то на практике приходится встречаться с системами очень высокого порядка (десятки и даже сотни тысяч неизвестных). Другими особенностями таких систем являются разреженность матрицы и ленточная структура, т. е. специальное расположение ненулевых элементов. Эти особенности, с одной стороны, облегчают решение указанных задач, а с другой стороны, требуют создания специальных методов решения, которые бы учитывали специфику задачи. Поэтому нет ничего удивительного в том, что классические методы линейной алгебры зачастую оказываются неэффективными при решении разностных уравнений, и не существует универсального метода, позволяющего эффективно решать любое разностное уравнение.

В настоящее время используются два типа методов решения систем линейных алгебраических уравнений: 1) прямые методы; 2) итерационные методы или методы последовательных приближений. Как правило, прямые методы ориентированы на решение довольно узкого класса сеточных уравнений, но они позволяют находить решения с очень малыми затратами вычислительной работы. Итерационные методы позволяют решать более сложные уравнения и часто в качестве основного этапа алгоритма содержат прямые методы решения специальных разностных уравнений. Тот факт, что разностные уравнения являются плохо обусловленными, приводит к необходимости разработки быстросходящихся итерационных процессов и выделению области эффективности каждого метода.

В ряде случаев, например для линейного уравнения с постоянными коэффициентами относительно сеточной функции одного аргумента, решение может быть найдено в замкнутом виде. Такие методы решения сеточных уравнений будут рассмотрены в § 3 данной главы.

§ 2. Общая теория линейных разностных уравнений

1. Свойства решений однородного уравнения. В данном параграфе будет рассмотрена общая теория линейных разностных уравнений m -го порядка с переменными коэффициентами

$$a_m(i)y(i+m) + \dots + a_0(i)y(i) = f_i,$$

где $a_m(i)$ и $a_0(i)$ отличны от нуля для любого i . Займемся сначала исследованием однородного уравнения

$$a_m(i)y(i+m) + \dots + a_0(i)y(i) = \sum_{k=0}^m a_k(i)y(i+k) = 0. \quad (1)$$

Будем считать, что коэффициенты $a_k(i)$, $k = 0, 1, \dots, m$, имеют для всех рассматриваемых значений i конечные значения.

Каждое частное решение уравнения (1) определяется значениями функции $y(i)$ в m произвольных, но расположенных подряд точках $i_0, i_0 + 1, \dots, i_0 + m - 1$.

Теорема 1. Если $v_1(i), v_2(i), \dots, v_p(i)$ — решения уравнения (1), то функция

$$y(i) = c_1v_1(i) + c_2v_2(i) + \dots + c_pv_p(i), \quad (2)$$

где c_1, c_2, \dots, c_p — произвольные постоянные, есть также решение уравнения (1).

Действительно, в силу условия теоремы имеют место равенства

$$\sum_{k=0}^m a_k(i)v_l(i+k) = 0, \quad l = 1, 2, \dots, p. \quad (3)$$

Подставим (2) в (1):

$$\sum_{k=0}^m a_k(i)y(i+k) = \sum_{k=0}^m a_k(i) \sum_{l=1}^p c_l v_l(i+k)$$

и поменяем порядок суммирования в правой части равенства. Используя (3), получим

$$\sum_{k=0}^m a_k(i)y(i+k) = \sum_{l=1}^p c_l \sum_{k=0}^m a_k(i)v_l(i+k) = 0$$

и, следовательно, функция $y(i)$, определяемая (2), также является решением уравнения (1). Теорема доказана.

Введем обозначение $\Delta_i(v_1, \dots, v_p)$ для определителя

$$\Delta_i(v_1, v_2, \dots, v_p) = \begin{vmatrix} v_1(i) & v_1(i+1) \dots v_1(i+p-1) \\ v_2(i) & v_2(i+1) \dots v_2(i+p-1) \\ \vdots & \vdots \\ v_p(i) & v_p(i+1) \dots v_p(i+p-1) \end{vmatrix}.$$

Имеет место

Лемма 2. Пусть $v_1(i), v_2(i), \dots, v_m(i)$ — решения уравнения (1). Определитель $\Delta_i(v_1, \dots, v_m)$ либо равен нулю тождественно по i , либо отличен от нуля для всех допустимых значений i .

Действительно, так как $v_1(i), \dots, v_m(i)$ — решения уравнения (1), то справедливы следующие равенства:

$$\begin{aligned} a_0(i)v_1(i) + a_1(i)v_1(i+1) + \dots + a_{m-1}(i)v_1(i+m-1) &= -a_m(i)v_1(i+m), \\ a_0(i)v_2(i) + a_1(i)v_2(i+1) + \dots + a_{m-1}(i)v_2(i+m-1) &= -a_m(i)v_2(i+m), \\ \vdots &\quad \vdots \\ a_0(i)v_m(i) + a_1(i)v_m(i+1) + \dots + a_{m-1}(i)v_m(i+m-1) &= -a_m(i)v_m(i+m). \end{aligned}$$

Решая эту систему относительно $a_0(i)$ для фиксированного i по правилу Крамера, получим

$$a_0(i)\Delta_i(v_1, \dots, v_m) = -a_m(i) \begin{vmatrix} v_1(i+m) & v_1(i+1) & \dots & v_1(i+m-1) \\ v_2(i+m) & v_2(i+1) & \dots & v_2(i+m-1) \\ \vdots & \vdots & \ddots & \vdots \\ v_m(i+m) & v_m(i+1) & \dots & v_m(i+m-1) \end{vmatrix}.$$

После соответствующей перестановки столбцов определителя правой части полученного равенства будем иметь соотношение $a_0(i)\Delta_i(v_1, \dots, v_m) = (-1)^m a_m(i)\Delta_{i+1}(v_1, \dots, v_m)$. Так как $a_0(i)$ и $a_m(i)$ не равны нулю для допустимых значений i , то отсюда следует утверждение леммы.

Введем теперь понятие линейно независимых решений уравнения (1). Сеточные функции $v_1(i), v_2(i), \dots, v_m(i)$ называются *линейно независимыми решениями уравнения (1)*, если: 1) они принимают конечные значения и удовлетворяют уравнению (1); 2) соотношения

$$c_1v_1(i) + c_2v_2(i) + \dots + c_mv_m(i) = 0 \quad (4)$$

при любых постоянных c_1, c_2, \dots, c_m , одновременно не равных нулю, не выполняются хотя бы для одного i .

Для линейно независимых решений справедлива

Лемма 3. Если $v_1(i), v_2(i), \dots, v_m(i)$ — линейно независимые решения уравнения (1), то определитель $\Delta_i(v_1, \dots, v_m)$ отличен от нуля для всех допустимых значений i . Обратно, если для решений $v_1(i), \dots, v_m(i)$ уравнения (1) определитель $\Delta_i(v_1, \dots, v_m)$ отличен от нуля хотя бы для одного значения i , то $v_1(i), \dots, v_m(i)$ — линейно независимые решения уравнения (1).

В силу леммы 2 определитель $\Delta_i(v_1, \dots, v_m)$ либо равен нулю тождественно, либо отличен от нуля для всех i . Пусть $v_1(i), \dots, v_m(i)$ — линейно независимые решения уравнения (1), и предположим, что $\Delta_i(v_1, \dots, v_m) \equiv 0$. Рассмотрим систему алгебраических уравнений

$$\begin{aligned} c_1v_1(i_0) + c_2v_2(i_0) + \dots + & c_mv_m(i_0) = 0, \\ c_1v_1(i_0+1) + c_2v_2(i_0+1) + \dots + & c_mv_m(i_0+1) = 0. \quad (5) \\ \vdots & \vdots \\ c_1v_1(i_0+m-1) + c_2v_2(i_0+m-1) + \dots + & c_mv_m(i_0+m-1) = 0. \end{aligned}$$

Так как определитель этой системы $\Delta_{i_0}(v_1, \dots, v_m)$, по предложению, равен нулю, то существует отличное от нуля решение этой системы c_1, c_2, \dots, c_m . Следовательно, для найденных c_1, c_2, \dots, c_m имеют место равенства (4) при $i = i_0, i_0+1, \dots, i_0+m-1$. Покажем теперь, что (4) будет иметь место и для $i = i_0+m$. Для этого, взяв уравнение (1) для $l = 1, 2, \dots, m$

$$\sum_{k=0}^m a_k(i_0)v_l(i_0+k) = 0,$$

умножим его на c_l и просуммируем равенства для $l = 1, 2, \dots, m$. Получим с учетом равенств (5)

$$\begin{aligned} 0 = a_m(i_0) \sum_{l=1}^m c_l v_l(i_0+m) + \sum_{k=0}^{m-1} a_k(i_0) \sum_{l=1}^m c_l v_l(i_0+k) = \\ = a_m(i_0) \sum_{l=1}^m c_l v_l(i_0+m). \end{aligned}$$

Таким образом, доказана справедливость равенства (4) для $i = i_0+m$. Идя таким же образом дальше, получим, что для найденных выше c_1, c_2, \dots, c_m соотношение (4) выполняется для всех допустимых $i \geq i_0$. Аналогично доказывается справедливость (4) для $i \leq i_0$. Следовательно, (4) с ненулевыми c_1, c_2, \dots, c_m выполняются для всех i , что противоречит линейной независимости $v_1(i), \dots, v_m(i)$. Поэтому предположение, что определитель $\Delta_i(v_1, \dots, v_m)$ тождественно по i равен нулю, неверно.

Докажем теперь вторую часть леммы 3. Пусть определитель $\Delta_i(v_1, \dots, v_m)$ для некоторого $i = i_0$ отличен от нуля. Тогда предположим, что $v_1(i), v_2(i), \dots, v_m(i)$ — система линейно зависимых решений уравнения (1). Это означает, что найдутся такие постоянные c_1, c_2, \dots, c_m , одновременно не равные нулю, что соотношение (4) является тождеством по i . Тогда запишем (4) для $i = i_0, i_0+1, \dots, i_0+m-1$ в виде системы (5), причем в силу предположения леммы определитель этой системы $\Delta_{i_0}(v_1, \dots, v_m)$ отличен от нуля. Поэтому все c_1, c_2, \dots, c_m должны равняться нулю. Мы пришли к противоречию. Лемма доказана.

2. Теоремы о решениях линейного уравнения. Сначала докажем теорему об общем решении однородного линейного уравнения (1).

Теорема 2. Если $v_1(i), v_2(i), \dots, v_m(i)$ — линейно независимые решения уравнения (1), то общее решение этого уравнения имеет вид

$$y(i) = c_1 v_1(i) + c_2 v_2(i) + \dots + c_m v_m(i), \quad (6)$$

где c_1, c_2, \dots, c_m — произвольные постоянные.

Действительно, в силу теоремы 1 функция $y(i)$, определенная формулой (6), есть решение уравнения (1). Покажем теперь, что все решения уравнения (1) содержатся в совокупности функций $y(i)$. Действительно, пусть $u(i)$ — произвольное решение уравнения (1). Оно вполне определяется заданiem начальных значений в m точках: $u(i_0)$, $u(i_0+1)$, ..., $u(i_0+m-1)$. Выберем из совокупности функций вида (6) такую, которая имеет те же начальные значения. Для этого достаточно найти такие постоянные c_1, c_2, \dots, c_m , чтобы выполнялись m равенств

$$\begin{aligned} c_1v_1(i_0) + c_2v_2(i_0) + \dots + & \quad c_mv_m(i_0) = u(i_0), \\ c_1v_1(i_0+1) + c_2v_2(i_0+1) + \dots + & \quad c_mv_m(i_0+1) = u(i_0+1), \\ \vdots & \quad \vdots \\ c_1v_1(i_0+m-1) + c_2v_2(i_0+m-1) + \dots + & \quad c_mv_m(i_0+m-1) = \\ & \quad u(i_0+m-1). \end{aligned}$$

Так как $v_1(i), v_2(i), \dots, v_m(i)$ — линейно независимые решения (1), то в силу леммы 3 определитель этой системы $\Delta_{i_0}(v_1, \dots, v_m)$ отличен от нуля. Решив эту систему относительно c_1, c_2, \dots, c_m , получим функцию $y(i)$, имеющую те же начальные значения, что и $u(i)$. Но так как начальные значения определяют однозначно решение уравнения (1), то $y(i) \equiv u(i)$. Теорема доказана.

Рассмотрим теперь решение неоднородного уравнения

$$a_m(i)y(i+m) + \dots + a_0(i)y(i) = f(i). \quad (7)$$

Имеет место

Теорема 3. Общее решение уравнения (7) представляется в виде суммы частного его решения и общего решения линейного однородного уравнения (1).

Действительно, покажем, что любое решение уравнения (7) может быть представлено в виде

$$y(i) = \bar{y}(i) + \bar{\bar{y}}(i), \quad (8)$$

где $\bar{y}(i)$ — какое-то решение уравнения (7), а $\hat{y}(i)$ есть общее решение однородного уравнения (1). Пусть

$$a_m(i)\bar{y}(i+m) + \dots + a_0(i)\bar{y}(i) = f(i). \quad (9)$$

Подставляя (8) в (7) и учитывая (9), будем иметь для $\bar{\bar{y}}(i)$ уравнение $a_m(i)\bar{\bar{y}}(i+m) + \dots + a_0\bar{\bar{y}}(i) = 0$. Следовательно, $\bar{\bar{y}}(i)$ есть общее решение однородного уравнения (1). Теорема доказана.

Следствие 1. Из теорем 2 и 3 вытекает, что общее решение неоднородного уравнения (7) имеет вид

$$y(i) = \bar{y}(i) + c_1 v_1(i) + \dots + c_m v_m(i), \quad (10)$$

где $\bar{y}(i)$ — частное решение уравнения (7), а $v_1(i), v_2(i), \dots, v_m(i)$ — линейно независимые решения однородного уравнения (1), c_1, \dots, c_m — произвольные постоянные.

Следствие 2. Используя лемму 3, следствию 1 можно придать иную формулировку: *решение уравнения (7) имеет вид (10), где частные решения $v_1(i), \dots, v_m(i)$ однородного уравнения таковы, что $\Delta_i(v_1, \dots, v_m) \neq 0$ хотя бы для одного значения i .*

Следствие 3. Если правая часть $f(i)$ уравнения (7) есть сумма двух функций $f(i) = f^{(1)}(i) + f^{(2)}(i)$, то частное решение уравнения (7) можно представить в виде $\bar{y}(i) = \bar{y}^{(1)}(i) + \bar{y}^{(2)}(i)$, где $\bar{y}^{(\alpha)}(i)$ есть частное решение уравнения (7) с правой частью $f^{(\alpha)}(i)$, $\alpha = 1, 2$.

3. Метод вариации постоянных. Доказанные выше теоремы дают структуру общего решения линейного неоднородного разностного уравнения (7). Рассмотрим теперь следующие вопросы: 1) как построить линейно независимые решения однородного уравнения; 2) как найти частное решение неоднородного уравнения; 3) каким образом, используя общее решение неоднородного уравнения, найти единственное решение уравнения (7), удовлетворяющее дополнительным условиям.

Изучим сначала один возможный способ построения линейно независимых решений однородного уравнения. Так как частное решение линейного уравнения m -го порядка полностью определяется заданием начальных значений в m точках, например, $i = i_0, i_0 + 1, \dots, i_0 + m - 1$, то в силу леммы 3 искомые решения уравнения (1) можно построить следующим образом. Пусть A — невырожденная матрица

$$A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{vmatrix}.$$

Построим m решений уравнения (1) $v_1(i), v_2(i), \dots, v_m(i)$, определяемых начальными значениями

$$v_l(i_0 + k - 1) = a_{lk}, \quad l, k = 1, 2, \dots, m. \quad (11)$$

Тогда $\Delta_{i_0}(v_1, \dots, v_m) = \det A \neq 0$. Следовательно, задача построения искомых функций $v_1(i), \dots, v_m(i)$ решена.

Рассмотрим теперь вопрос о выделении из семейства решений (10) единственного решения. Из (10) следует, что для этого нужно

задать ровно m условий на функцию $y(i)$, из которых определяются постоянные c_1, c_2, \dots, c_m .

В случае задачи Коши, т. е. когда заданы начальные условия $y(i_0) = b_1, y(i_0 + 1) = b_2, \dots, y(i_0 + m - 1) = b_m$, определение постоянных c_1, c_2, \dots, c_m осуществляется просто. Из (10) имеем систему линейных алгебраических уравнений относительно c_1, c_2, \dots, c_m :

$$\begin{aligned} v_1(i_0)c_1 + v_2(i_0)c_2 + \dots + v_m(i_0)c_m &= b_1 - \bar{y}(i_0), \\ v_1(i_0 + 1)c_1 + v_2(i_0 + 1)c_2 + \dots + v_m(i_0 + 1)c_m &= \\ &= b_2 - \bar{y}(i_0 + 1), \end{aligned} \quad (12)$$

$$\begin{aligned} \dots &\dots \\ v_1(i_0 + m - 1)c_1 + v_2(i_0 + m - 1)c_2 + \dots + v_m(i_0 + m - 1)c_m &= \\ &= b_m - \bar{y}(i_0 + m - 1). \end{aligned}$$

Так как определитель этой системы $\Delta_{i_0}(v_1, \dots, v_m)$ отличен от нуля, то эта система имеет единственное решение c_1, c_2, \dots, c_m , которое полностью определяет единственное решение неоднородного уравнения (7).

В случае краевой задачи, когда m дополнительных условий для $y(i)$ заданы не в подряд расположенных точках, мы снова придем к системе линейных алгебраических уравнений относительно c_1, c_2, \dots, c_m . Но в этом случае решение такой системы будет существовать лишь при дополнительных предположениях относительно коэффициентов разностного уравнения.

Рассмотрим теперь вопрос о решении уравнений (12). Так как в силу (11) матрицей системы (12) является A^T , то, выбирая в качестве A единичную матрицу, получим решение системы (12) в явном виде: $c_l = b_l - \bar{y}(i_0 + l - 1)$, $l = 1, 2, \dots, m$. Очевидно, что среди частных решений неоднородного уравнения (7) целесообразно выбрать такое, для которого $\bar{y}(i_0) = \bar{y}(i_0 + 1) = \dots = \bar{y}(i_0 + m - 1) = 0$. Тогда будем иметь $c_l = b_l$, $l = 1, 2, \dots, m$.

Такому выбору матрицы A соответствуют следующие начальные значения для $v_1(i), \dots, v_m(i)$:

$$\begin{aligned} v_l(i_0 + l - 1) &= 1, \quad v_l(i_0 + k - 1) = 0, \quad k = 1, 2, \dots, m, \quad k \neq l, \\ l &= 1, 2, \dots, m. \end{aligned}$$

Займемся теперь отысканием частных решений неоднородного уравнения, если известны m линейно независимых решений однородного уравнения. Изложим способ нахождения частного решения *вариацией постоянных* в общем решении однородного уравнения.

Ранее было показано, что общее решение однородного уравнения (1) имеет вид $\bar{y}(i) = c_1 v_1(i) + \dots + c_m v_m(i)$, где $v_1(i), \dots, v_m(i)$ — линейно независимые решения уравнения (1), а c_1, c_2, \dots, c_m — произвольные постоянные. Будем теперь считать

c_1, c_2, \dots, c_m функциями i и поставим задачу выбрать их так, чтобы функция

$$\bar{y}(i) = c_1(i)v_1(i) + \dots + c_m(i)v_m(i) \quad (13)$$

оказалась частным решением неоднородного уравнения (7). Заметим, что каждая функция $c_l(i)$ определяется с точностью до постоянной, так как $v_l(i)$ — решение однородного уравнения:

$$a_m(i)v_l(i+m) + \dots + a_0(i)v_l(i) = 0, \quad l = 1, 2, \dots, m. \quad (14)$$

Введем следующее обозначение:

$$d_k(i) = \sum_{l=1}^m [c_l(i+k) - c_l(i)]v_l(i+k), \quad k = 0, 1, \dots, m.$$

Подставляя (13) в (7), выполняя тождественные преобразования в полученном выражении и учитывая (14), будем иметь

$$\begin{aligned} f(i) &= \sum_{k=0}^m a_k(i)\bar{y}(i+k) = \sum_{k=0}^m a_k(i) \sum_{l=1}^m c_l(i+k)v_l(i+k) = \\ &= \sum_{k=0}^m a_k(i)d_k(i) + \sum_{k=0}^m a_k(i) \sum_{l=1}^m c_l(i)v_l(i+k) = \\ &= \sum_{k=0}^m a_k(i)d_k(i) + \sum_{l=1}^m c_l(i) \left[\sum_{k=0}^m a_k(i)v_l(i+k) \right] = \\ &= \sum_{k=0}^m a_k(i)d_k(i) = \sum_{k=1}^m a_k(i)d_k(i), \end{aligned}$$

так как $d_0(i) \equiv 0$. Полученное соотношение будет выполняться для всех i , если положить

$$d_k(i) = 0, \quad k = 1, 2, \dots, m-1, \quad d_m(i) = f(i)/a_m(i). \quad (15)$$

Итак, задача построения функций $c_1(i), c_2(i), \dots, c_m(i)$ сведена к определению их из условий (15), которые должны выполняться тождественно по i .

Преобразуем систему уравнений (15). Обозначим $b_l(i) = c_l(i+1) - c_l(i)$, $l = 1, 2, \dots, m$. Из определения $d_k(i)$ получим для $k = 1, 2, \dots, m$:

$$\begin{aligned} d_k(i) - d_{k-1}(i+1) &= \sum_{l=1}^m [c_l(i+k) - c_l(i)]v_l(i+k) - \\ &- \sum_{l=1}^m [c_l(i+k) - c_l(i+1)]v_l(i+k) = \sum_{l=1}^m b_l(i)v_l(i+k). \end{aligned}$$

Подставляя сюда (15) и учитывая равенство $d_0(i) = 0$, получим систему линейных алгебраических уравнений относительно $b_l(i)$

для фиксированного i :

$$\begin{aligned} b_1(i)v_1(i+1) + b_2(i)v_2(i+1) + \dots + b_m(i)v_m(i+1) &= 0, \\ b_1(i)v_1(i+2) + b_2(i)v_2(i+2) + \dots + b_m(i)v_m(i+2) &= 0, \\ \vdots &\vdots \\ b_1(i)v_1(i+m) + b_2(i)v_2(i+m) + \dots + b_m(i)v_m(i+m) &= \frac{f(i)}{a_m(i)}. \end{aligned} \quad (16)$$

Определитель системы (16) равен $\Delta_{i+1}(v_1, v_2, \dots, v_m)$ и отличен от нуля в силу линейной независимости v_1, v_2, \dots, v_m . Поэтому система (16) имеет единственное решение

$$b_l(i) = c_l(i+1) - c_l(i) = (-1)^{m+l} \frac{f(i)}{a_m(i)} \frac{\mathcal{D}_l(i)}{\mathcal{D}(i)}, \quad l = 1, \dots, m, \quad (17)$$

где $\mathcal{D}(i) = \Delta_{i+1}(v_1, v_2, \dots, v_m)$, а

$$\mathcal{D}_l(i) = \begin{vmatrix} v_1(i+1) & \dots & v_{l-1}(i+1) & v_{l+1}(i+1) & \dots & v_m(i+1) \\ v_1(i+2) & \dots & v_{l-1}(i+2) & v_{l+1}(i+2) & \dots & v_m(i+2) \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ v_1(i+m-1) & \dots & v_{l-1}(i+m-1) & v_{l+1}(i+m-1) & \dots & v_m(i+m-1) \end{vmatrix},$$

т. е. $\mathcal{D}_l(i)$ получается из определителя $\mathcal{D}(i)$ вычеркиванием l -го столбца и последней строки.

Равенства (17) являются разностными уравнениями первого порядка относительно функций $c_l(i)$, $l = 1, 2, \dots, m$. Так как $c_l(i)$ может быть определена с точностью до константы, то из (17) найдем явное представление для $c_l(i)$:

$$c_l(i) = \sum_{j=i_0}^{i-1} (-1)^{m+l} \frac{f(j)}{a_m(j)} \frac{\mathcal{D}_l(j)}{\mathcal{D}(j)}, \quad l = 1, 2, \dots, m.$$

Подставляя это выражение в (13) и меняя порядок суммирования в получаемом представлении, будем иметь для частного решения $\bar{y}(i)$ неоднородного уравнения (7) следующую формулу:

$$\begin{aligned} \bar{y}(i) &= \sum_{l=1}^m c_l(i) v_l(i) = \\ &= \sum_{j=i_0}^{i-1} \left[f(j) \sum_{l=1}^m (-1)^{m+l} \mathcal{D}_l(j) v_l(i) \right] / (\mathcal{D}(j) a_m(j)) = \\ &= \sum_{j=i_0}^{i-1} G(i, j) f(j), \end{aligned}$$

где

$$G(i, j) = \frac{1}{\mathcal{D}(j) a_m(j)} \sum_{k=1}^m (-1)^{m+k} \mathcal{D}_k(j) v_k(i). \quad (18)$$

Заметим, что сумма, стоящая в (18), легко вычисляется

$$\sum_{k=1}^m (-1)^{m+k} \mathcal{D}_k(j) v_k(i) = \begin{vmatrix} v_1(j+1) & v_2(j+1) & \dots & v_m(j+1) \\ v_1(j+2) & v_2(j+2) & \dots & v_m(j+2) \\ \vdots & \vdots & \ddots & \vdots \\ v_1(j+m-1) & v_2(j+m-1) & \dots & v_m(j+m-1) \\ v_1(i) & v_2(i) & \dots & v_m(i) \end{vmatrix}.$$

Эта сумма равна нулю при $j = i - 1, i - 2, \dots, i - m + 1$. Таким образом, частное решение уравнения (7) имеет следующее представление:

$$\bar{y}(i) = \sum_{j=i_0}^{i-m} \frac{\begin{vmatrix} v_1(j+1) & \cdots & v_m(j+1) \\ \vdots & \ddots & \vdots \\ v_1(j+m-1) & \cdots & v_m(j+m-1) \\ v_1(i) & \cdots & v_m(i) \\ \vdots & \ddots & \vdots \\ v_1(j+1) & \cdots & v_m(j+m) \\ \vdots & \ddots & \vdots \\ v_m(j+1) & \cdots & v_m(j+m) \end{vmatrix}}{\begin{vmatrix} v_1(j+1) & \cdots & v_m(j+1) \\ \vdots & \ddots & \vdots \\ v_1(j+m) & \cdots & v_m(j+m) \\ \vdots & \ddots & \vdots \\ v_1(i) & \cdots & v_m(i) \end{vmatrix}} \cdot \frac{f(j)}{a_m(j)}, \quad (19)$$

где i_0 произвольно, а для $i = i_0, i_0 + 1, \dots, i_0 + m - 1$ имеем $\bar{y}(i) = 0$.

Для уравнения первого порядка ($m = 1$) формула (19) принимает следующий вид:

$$\bar{y}(i) = \sum_{j=i_0}^{i-1} \frac{v_1(i)}{v_1(j+1)} \cdot \frac{f(j)}{a_1(j)}, \quad \bar{y}(i_0) = 0. \quad (20)$$

4. Примеры. Рассмотрим некоторые примеры, иллюстрирующие применение общей теории. Пусть требуется найти общее решение уравнения первого порядка

$$y(i+1) - e^{2i}y(i) = 6i^2e^{i^2+i}. \quad (21)$$

Найдем сначала решение однородного уравнения

$$y(i+1) - e^{2i}y(i) = 0. \quad (22)$$

Из (22) последовательно получим

$$y(i+1) = e^{2i}y(i) = e^{2i}e^{2(i-1)}y(i-1) = \dots = e^{2\sum_{k=1}^{i-1} k} y(1) = e^{i(i+1)}y(1).$$

Полагая здесь $y(1) = 1$, найдем частное решение $v_1(i)$ однородного уравнения (22) в виде $v_1(i) = e^{i(i-1)}$. Следовательно, общее решение однородного уравнения имеет вид $\bar{y}(i) = ce^{i(i-1)}$, где c — произвольная постоянная.

Построим теперь частное решение неоднородного уравнения (21), используя формулу (20). Из (20) получим

$$\bar{y}(i) = \sum_{k=i_0}^{i-1} \frac{e^{i(i-1)}}{e^{k(k+1)}} \cdot \frac{6k^2e^{k^2+k}}{1} = 6e^{i(i-1)} \sum_{k=i_0}^{i-1} k^2.$$

Так как i_0 может быть выбрано произвольным, то, полагая здесь $i_0 = 1$, будем иметь $\bar{y}(i) = i(i-1)(2i-1)e^{i(i-1)}$. Далее, в силу теоремы 3 общее решение уравнения (21) записывается в виде

$$y(i) = \bar{y}(i) + \bar{\bar{y}}(i) = [c + i(i-1)(2i-1)]e^{i(i-1)},$$

где c — произвольная постоянная. Задача решена.

Найдем теперь общее решение уравнения второго порядка

$$a_2(i)y(i+2) + a_1(i)y(i+1) + a_0(i)y(i) = f(i), \quad (23)$$

где $i = 0, 1, 2, \dots$,

$$\begin{aligned} a_2(i) &= i^2 - i + 1, \quad a_0(i) = a_2(i+1) = i^2 + i + 1, \\ a_1(i) &= -a_0(i) - a_2(i) = -2(i^2 + 1), \\ f(i) &= 2^i(i^2 - 3i + 1) = 2^i[2a_2(i) - a_0(i)]. \end{aligned} \quad (24)$$

Так как коэффициенты $a_2(i)$ и $a_0(i)$ отличны от нуля, то для нахождения общего решения уравнения (23) можно применить общую теорию.

Сначала построим линейно независимые решения однородного уравнения. Используя (24), его можно записать в следующем виде:

$$a_2(i)y(i+2) - [a_2(i) + a_2(i+1)]y(i+1) + a_2(i+1)y(i) = 0$$

или

$$a_2(i)[y(i+2) - y(i+1)] - a_2(i+1)[y(i+1) - y(i)] = 0. \quad (25)$$

Частные решения $v_1(i)$ и $v_2(i)$ однородного уравнения (25) выделим следующими условиями: $v_1(0) = v_1(1) = 1$, $v_2(0) = 0$, $v_2(1) = 3$. Так как определитель

$$\Delta_0(v_1, v_2) = \begin{vmatrix} v_1(0) & v_1(1) \\ v_2(0) & v_2(1) \end{vmatrix} = 3 \neq 0,$$

то в силу леммы 3 функции $v_1(i)$ и $v_2(i)$ будут линейно независимыми решениями уравнения (25).

Найдем явный вид для $v_1(i)$ и $v_2(i)$. Из (25) сразу следует, что $v_1(i) \equiv 1$. Построим $v_2(i)$. Из (25) последовательно получим

$$\begin{aligned} y(i+2) - y(i+1) &= \frac{a_2(i+1)}{a_2(i)}[y(i+1) - y(i)] = \\ &= \frac{a_2(i+1)}{a_2(i-1)}[y(i) - y(i-1)] = \dots = \frac{a_2(i+1)}{a_2(0)}[y(1) - y(0)]. \end{aligned}$$

Учитывая начальные значения для $v_2(i)$, отсюда найдем

$$v_2(i+1) - v_2(i) = 3a_2(i) = 3(i^2 - i + 1). \quad (26)$$

Суммируя левую и правую части (26) по i от нуля до $k-1$, будем иметь

$$v_2(k) = v_2(0) + 3 \sum_{i=0}^{k-1} (i^2 - i + 1) = k(k^2 - 3k + 5).$$

Итак, частные решения однородного уравнения (25) найдены

$$v_1(k) \equiv 1, \quad v_2(k) = k(k^2 - 3k + 5), \quad (27)$$

и общее решение (25) имеет вид $\bar{y}(k) = c_1 + c_2 k(k^2 - 3k + 5)$.

Построим теперь частное решение неоднородного уравнения (23). Подставляя (24) и (27) в формулу (19), получим

$$\begin{aligned}\bar{y}(i) &= \sum_{k=0}^{i-2} \frac{v_2(i) - v_2(k+1)}{v_2(k+2) - v_2(k+1)} \cdot \frac{f(k)}{a_2(k)} = \\ &= \sum_{k=0}^{i-2} \frac{v_2(i) - v_2(k+1)}{3a_2(k+1) a_2(k)} [2^{k+1} a_2(k) - 2^k a_2(k+1)] = \\ &= \frac{1}{3} \sum_{k=0}^{i-2} [v_2(i) - v_2(k+1)] \left[\frac{2^{k+1}}{a_2(k+1)} - \frac{2^k}{a_2(k)} \right].\end{aligned}\quad (28)$$

Здесь было использовано равенство (26).

Вычислим полученное выражение. Обозначая

$$v(k) = v_2(i) - v_2(k+1), \quad u(k) = \frac{2^k}{a_2(k)},$$

запишем (28) следующим образом:

$$\bar{y}(i) = \frac{1}{3} \sum_{k=0}^{i-2} [u(k+1) - u(k)] v(k).$$

Используем теперь формулу суммирования по частям (см. (8) § 1) для случая равномерной сетки с шагом $h=1$. Это дает

$$\begin{aligned}\bar{y}(i) &= -\frac{1}{3} \sum_{k=0}^{i-1} u(k) [v(k) - v(k-1)] + \\ &\quad + \frac{1}{3} [u(i-1)v(i-1) - u(0)v(-1)].\end{aligned}$$

Так как в силу (26), условия $v_2(0) = 0$ и определения функций $v(k)$ и $u(k)$ имеем

$$\begin{aligned}v(k) - v(k-1) &= v_2(k) - v_2(k+1) = -3a_2(k), \\ v(i-1) &= v_2(i) - v_2(i) = 0, \\ v(-1) &= v_2(i) - v_2(0) = v_2(i),\end{aligned}$$

то

$$\bar{y}(i) = \sum_{k=0}^{i-1} 2^k - \frac{1}{3} v_2(i) = 2^i - 1 - \frac{1}{3} i (i^2 - 3i + 5).$$

Следовательно, частное решение (23) найдено. В силу теоремы 3 общее решение неоднородного уравнения второго порядка (23) имеет вид

$$\begin{aligned}y(i) &= \bar{y}(i) + \bar{\bar{y}}(i) = 2^i - 1 - \frac{1}{3} i (i^2 - 3i + 5) + c_1 + c_2 i (i^2 - 3i + 5) = \\ &= \bar{c}_1 + 2^i + \bar{c}_2 i (i^2 - 3i + 5),\end{aligned}$$

где $\bar{c}_1 = c_1 - 1$, $\bar{c}_2 = c_2 - \frac{1}{3}$ — произвольные постоянные. Задача решена.

§ 3. Решение линейных уравнений с постоянными коэффициентами

1. Характеристическое уравнение. Случай простых корней.

Рассмотрим теперь важный класс разностных уравнений — линейные уравнения с постоянными коэффициентами. Для уравнений этого класса вопрос о нахождении линейно независимых решений соответствующих однородных уравнений может быть решен достаточно просто. А к этому, как было показано выше, сводится задача решения неоднородного разностного уравнения.

Займемся отысканием линейно независимых решений однородного линейного уравнения с постоянными коэффициентами m -го порядка

$$a_m y(i+m) + a_{m-1} y(i+m-1) + \dots + a_0 y(i) = 0. \quad (1)$$

Будем искать частные решения (1) в виде $v(i) = q^i$, где число q подлежит определению. Подставляя $v(i)$ вместо $y(i)$ в (1), получим уравнение

$$q^i (a_m q^m + a_{m-1} q^{m-1} + \dots + a_1 q + a_0) = 0.$$

Так как ищется не равное тождественно нулю решение (1), то, сокращая на q^i , получим отсюда уравнение для q :

$$a_m q^m + a_{m-1} q^{m-1} + \dots + a_1 q + a_0 = 0. \quad (2)$$

Уравнение (2) называется *характеристическим уравнением* для (1). Корни уравнения (2) q_1, q_2, \dots, q_m могут быть как простыми, так и кратными. Рассмотрим отдельно каждый возможный случай.

Пусть корни простые. Покажем, что функции

$$v_1(i) = q_1^i, v_2(i) = q_2^i, \dots, v_m(i) = q_m^i \quad (3)$$

являются линейно независимыми решениями уравнения (1).

Действительно, в силу леммы 3 достаточно показать, что хотя бы для одного i определитель $\Delta_i(v_1, v_2, \dots, v_m) \neq 0$. Полагая $i = 0$, найдем

$$\Delta_0(v_1, \dots, v_m) = \begin{vmatrix} 1 & q_1 & q_1^2 & \dots & q_1^{m-1} \\ 1 & q_2 & q_2^2 & \dots & q_2^{m-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & q_m & q_m^2 & \dots & q_m^{m-1} \end{vmatrix} = \begin{vmatrix} 1 & 1 & \dots & 1 \\ q_1 & q_2 & \dots & q_m \\ q_1^2 & q_2^2 & \dots & q_m^2 \\ \dots & \dots & \dots & \dots \\ q_1^{m-1} & q_2^{m-1} & \dots & q_m^{m-1} \end{vmatrix}$$

и, следовательно, $\Delta_0(v_1, \dots, v_m)$ — определитель Вандермонда. Он отличен от нуля, так как все q_k различны. Таким образом, функции (3) действительно являются линейно независимыми решениями (1), и поэтому общее решение однородного уравнения (1) может быть записано в виде

$$y(i) = c_1 q_1^i + c_2 q_2^i + \dots + c_m q_m^i, \quad (4)$$

где c_1, c_2, \dots, c_m — произвольные постоянные.

Если корни q_1, q_2, \dots, q_m действительные, то действительное решение $y(i)$ выделяется выбором постоянных c_1, c_2, \dots, c_m действительными числами. Рассмотрим теперь вопрос о выделении действительного решения, если среди корней есть комплексные.

Пусть $q_n = \rho(\cos \varphi + i^* \sin \varphi)$, ($i^* = \sqrt{-1}$) — комплексный корень характеристического уравнения (2). Тогда существует сопряженный к q_n корень $q_s = \rho(\cos \varphi - i^* \sin \varphi)$ уравнения (2). Рассмотрим часть общего решения (4), образуемую линейной комбинацией q_n^t и q_s^t :

$$y(i) = c_n q_n^t + c_s q_s^t = \rho^t [(c_n + c_s) \cos i\varphi + i^* (c_n - c_s) \sin i\varphi].$$

Функция $y(i)$ будет иметь действительные значения, если постоянные c_n и c_s будут комплексно сопряженными. Полагая $c_n = -0,5(\bar{c}_n - i^* \bar{c}_s)$, $c_s = 0,5(\bar{c}_n + i^* \bar{c}_s)$, где \bar{c}_n и \bar{c}_s — произвольные действительные числа, получим $y(i) = \rho^t (\bar{c}_n \cos i\varphi + \bar{c}_s \sin i\varphi)$.

2. Случай кратных корней. Пусть теперь характеристическое уравнение (2) имеет корень q_t кратности n_t , q_2 — кратности n_2 и т. д., т. е. q_1, q_2, \dots, q_s — различные корни кратности n_1, n_2, \dots, n_s соответственно, $n_1 + n_2 + \dots + n_s = m$. Построим линейно независимые решения однородного уравнения (1). Нам потребуется

Лемма 4. Если q_t — корень характеристического уравнения (2), имеющий кратность n_t , то справедливы равенства

$$\sum_{k=0}^m a_k k^p q_t^k = 0, \quad p = 0, 1, \dots, n_t - 1. \quad (5)$$

Действительно, так как q_t — корень уравнения (2) кратности n_t , то имеют место равенства

$$\sum_{k=0}^m a_k q_t^k = 0, \quad (6)$$

$$\sum_{k=0}^m k(k-1)\dots(k-s+1) a_k q_t^k = 0, \quad s = 1, 2, \dots, n_t - 1, \quad (7)$$

получаемые из (2) дифференцированием s раз и дополнительным умножением результата на q_t^s . Покажем, что равенство (5) эквивалентно (6), (7). Очевидно, что нужно доказать эквивалентность только (7) и (5) для $p \geq 1$.

Так как $P_s(k) = k(k-1)\dots(k-s+1)$ — полином степени s от k , то, умножая (5) на соответствующий коэффициент полинома $P_s(k)$ для $p = 1, 2, \dots, s$ и складывая получаемые равенства, будем иметь соотношения (7).

Покажем теперь, что из (7) следуют равенства (5) для $p = 1, 2, \dots, n_t - 1$. Используем разложение для k^p :

$$k^p = \sum_{s=1}^p k(k-1)\dots(k-s+1) \alpha_s, \quad 1 \leq p \leq k, \quad (8)$$

где $\alpha_s = \alpha_s(p)$ будет указано ниже. Умножим s -е равенство (7)

на α_s и просуммируем по s от 1 до p . В силу (8) получим

$$0 = \sum_{s=1}^p \alpha_s \left(\sum_{k=0}^m k(k-1)\dots(k-s+1) a_k q_l^k \right) = \\ = \sum_{k=0}^m a_k q_l^k \left(\sum_{s=1}^p k(k-1)\dots(k-s+1) \alpha_s \right) = \sum_{k=0}^m a_k k^p q_l^k.$$

Осталось обосновать разложение (8). Заметим, что слева и справа в (8) стоят полиномы p -й степени от k . Если положить $\alpha_p = 1$, то коэффициенты при старшей степени k слева и справа в (8) будут равны, коэффициенты при младшой степени равны нулю. Найдем $\alpha_1, \alpha_2, \dots, \alpha_{p-1}$, приравнивая значения полиномов в $p-1$ различных точках, например, полагая $k=1, 2, \dots, p-1$. Для $k=1$ это дает $\alpha_1 = 1$. При $k=n$, $2 \leq n \leq p-1$ будем иметь

$$n^p = \sum_{s=1}^p n(n-1)\dots(n-s+1) \alpha_s = \sum_{s=1}^n n(n-1)\dots(n-s+1) \alpha_s = \\ = n! \alpha_n + n! \sum_{s=1}^{n-1} \frac{\alpha_s}{(n-s)!}.$$

Отсюда можно найти α_n , если $\alpha_1, \alpha_2, \dots, \alpha_{n-1}$ уже определены. Таким образом, получаем следующую рекуррентную формулу для нахождения коэффициентов α_n :

$$\alpha_n = \frac{n^p}{n!} - \sum_{s=1}^{n-1} \frac{\alpha_s}{(n-s)!}, \quad n = 2, 3, \dots, p-1, \alpha_1 = 1.$$

Лемма доказана.

Используя лемму 4, найдем m частных решений однородного уравнения (1). Так как справедливо равенство

$$(j+k)^n = \sum_{p=0}^n C_n^p k^p j^{n-p}, \quad C_n^p = \frac{n!}{p!(n-p)!},$$

то, умножая (5) на $C_n^p j^{n-p} q_l^p$ и суммируя по p от нуля до $n \leq n_l - 1$, получим, что для любого j имеют место равенства

$$\sum_{k=0}^m a_k (j+k)^n q_l^{k+l} = 0, \quad n = 0, 1, \dots, n_l - 1.$$

Используя их, легко найдем, что частными решениями однородного уравнения (1) являются сеточные функции

$$v_{n_1+n_2+\dots+n_{l-1}+n_l+1}(j) = j^n q_l^l, \quad 0 \leq n \leq n_l - 1, \quad l = 1, 2, \dots, s, \quad (9)$$

т. е., если q_l — корень характеристического уравнения кратности n_l , то функции

$$q_l^l, jq_l^l, \dots, j^{n_l-1} q_l^l, \quad l = 1, 2, \dots, s$$

суть решения уравнения (1).

Осталось показать, что функции $v_1(j), \dots, v_m(j)$, определенные в (9), являются линейно независимыми решениями. Для этого вычислим определитель $\Delta_0(v_1, \dots, v_m)$, который в данном случае имеет вид

$$\Delta_0(v_1, \dots, v_m) = \begin{vmatrix} 1 & q_1 & q_1^2 & \cdots & q_1^k & \cdots & q_1^{m-1} \\ 0 & q_1 & 2q_1^2 & \cdots & kq_1^k & \cdots & (m-1)q_1^{m-1} \\ 0 & q_1 & 2^2q_1^2 & \cdots & k^2q_1^k & \cdots & (m-1)^2q_1^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & q_s & q_s^2 & \cdots & q_s^k & \cdots & q_s^{m-1} \\ 0 & q_s & 2q_s^2 & \cdots & kq_s^k & \cdots & (m-1)q_s^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & q_s & 2^n s^{-1} q_s^2 & \cdots & k^n s^{-1} q_s^k & \cdots & (m-1)^n s^{-1} q_s^{m-1} \end{vmatrix}.$$

Он может быть непосредственно получен из определителя Вандермонда

$$W(x_1, x_2, \dots, x_m) = \begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m-1} & x_{m-1}^2 & \cdots & x_{m-1}^{m-1} \\ 1 & x_m & x_m^2 & \cdots & x_m^{m-1} \end{vmatrix} = \prod_{i=1}^{m-1} \prod_{j=i+1}^m (x_j - x_i)$$

следующим образом. Возьмем от W первую производную по x_2 и умножим ее на x_2 . Результат обозначим через $W_2 = x_2 \frac{\partial W}{\partial x_2}$. Далее вычислим

$$W_3 = x_3 \frac{\partial}{\partial x_3} \left(x_3 \frac{\partial W_2}{\partial x_3} \right), \quad W_4 = x_4 \frac{\partial}{\partial x_4} \left(x_4 \frac{\partial}{\partial x_4} \left(x_4 \frac{\partial W_3}{\partial x_4} \right) \right), \dots$$

и т. д., пока не получим W_{n_1} . Затем вычислим $W_{n_1+2} = x_{n_1+2} \frac{\partial W_{n_1}}{\partial x_{n_1+2}}$, продолжим процесс дифференцирования, вычисляя $W_{n_1+3} = x_{n_1+3} \frac{\partial}{\partial x_{n_1+3}} \left(x_{n_1+3} \frac{\partial W_{n_1+2}}{\partial x_{n_1+3}} \right)$, пока не получим $W_{n_1+n_2}$, и т. д. В результате получим $W_m = W_m(x_1, x_2, \dots, x_m)$. Положим здесь $x_1 = x_2 = \dots = x_{n_1} = q_1$, $x_{n_1+1} = x_{n_1+2} = \dots = x_{n_1+n_2} = q_2$ и т. д. Легко убедиться в том, что $\Delta_0(v_1, v_2, \dots, v_m) = W_m$, а простые вычисления дают

$$W_m = \prod_{k=1}^s \prod_{m=1}^{n_k-1} m! q_k^m \prod_{i=1}^{s-1} \prod_{j=i+1}^s (q_j - q_i)^{n_i n_j}.$$

Отсюда следует, что $\Delta_0(v_1, \dots, v_m) \neq 0$, так как $q_j \neq q_i$ при $j \neq i$, а поэтому функции $v_1(j), v_2(j), \dots, v_m(j)$, построенные выше, являются линейно независимыми решениями однородного

уравнения (1). При этом общее решение уравнения (1) записывается в виде

$$y(j) = \sum_{l=1}^s \sum_{n=0}^{n_p-1} c_n^{(l)} j^n q_l^l,$$

где $c_n^{(l)}$ — произвольные постоянные.

3. Примеры. Рассмотрим простейшие примеры нахождения общего решения однородного разностного уравнения с постоянными коэффициентами.

1. Требуется найти общее решение уравнения

$$y(j+2) - y(j+1) - 2y(j) = 0. \quad (10)$$

Составляем характеристическое уравнение $q^2 - q - 2 = 0$ и находим его корни $q_1 = 2$, $q_2 = -1$. Так как корни простые, то общее решение уравнения (10) имеет вид

$$y(j) = c_1 2^j + c_2 (-1)^j.$$

2. Найти общее решение уравнения четвертого порядка

$$y(j+4) - 2y(j+3) + 3y(j+2) - 2y(j+1) - 4y(j) = 0. \quad (11)$$

Характеристическое уравнение $q^4 - 2q^3 + 3q^2 - 2q - 4 = 0$ имеет два действительных корня $q_1 = 1$, $q_2 = -1$ и два комплексно сопряженных корня $q_3 = 2 \left(\cos \frac{\pi}{3} + i \sin \frac{\pi}{3} \right)$ и $q_4 = 2 \left(\cos \frac{\pi}{3} - i \sin \frac{\pi}{3} \right)$,

$i = \sqrt{-1}$. Следовательно, общее решение уравнения (11), принимающее действительные значения, имеет вид

$$y(j) = c_1 + c_2 (-1)^j + 2^j \left(c_3 \cos \frac{\pi}{3} j + c_4 \sin \frac{\pi}{3} j \right).$$

3. Найти общее решение уравнения четвертого порядка

$$y(j+4) - 7y(j+3) + 18y(j+2) - 20y(j+1) + 8y(j) = 0. \quad (12)$$

Характеристическое уравнение

$$q^4 - 7q^3 + 18q^2 - 20q + 8 = (q-2)^3 (q-1) = 0$$

имеет корень $q_1 = 2$ кратности 3 и корень $q_2 = 1$ кратности 1. Следовательно, общее решение (12) имеет вид

$$y(j) = c_1 + 2^j (c_2 + c_3 j + c_4 j^2),$$

а частными линейно независимыми решениями (12) являются сеточные функции $v_1(j) = 1$, $v_2(j) = 2^j$, $v_3(j) = j2^j$, $v_4(j) = j^2 2^j$.

4. Найти общее решение уравнения четвертого порядка

$$y(j+4) + 8y(j+2) + 16y(j) = 0. \quad (13)$$

Характеристическое уравнение $q^4 + 8q^3 + 16 = (q^2 + 4)^2 = 0$ имеет комплексный корень $q_1 = 2 \left(\cos \frac{\pi}{2} + i \sin \frac{\pi}{2} \right)$ кратности 2

и сопряженный ему корень $q_2 = 2 \left(\cos \frac{\pi}{2} - i \sin \frac{\pi}{2} \right)$ тоже кратности 2. Поэтому общее решение уравнения (13), которое принимает действительные значения, имеет вид

$$y(j) = (c_1 + c_2 j) 2^j \cos \frac{\pi}{2} j + (c_3 + c_4 j) 2^j \sin \frac{\pi}{2} j.$$

Рассмотрим еще два примера. В одном примере мы найдем решение задачи Коши для неоднородного уравнения первого порядка, в другом — краевой задачи для однородного уравнения четвертого порядка.

5. Найти решение следующей задачи:

$$y(i+1) - ay(i) = f(i), \quad i \geq 0, \quad y(0) = y_0, \quad (14)$$

где $a = \text{const}$. Характеристическое уравнение $q - a = 0$ имеет единственный корень $q_1 = a$. Поэтому общее решение однородного уравнения имеет вид $\bar{y}(i) = ca^i$, $c = \text{const}$. Частное решение неоднородного уравнения (14) найдем, используя метод вариации постоянной. Формула (20) § 2 дает следующее частное решение уравнения (14):

$$\bar{y}(i) = \sum_{k=0}^{i-1} a^{i-k-1} f(k) = \sum_{k=0}^{i-1} a^k f(i-k-1).$$

В силу теоремы 3 общее решение неоднородного уравнения (14) имеет вид

$$y(i) = ca^i + \sum_{k=0}^{i-1} a^k f(i-k-1).$$

Полагая здесь $i = 0$, получим (сумма при этом исчезает) $y_0 = y(0) = c$. Таким образом, решение задачи (14) дается формулой

$$y(i) = y_0 a^i + \sum_{k=0}^{i-1} a^k f(i-k-1), \quad i \geq 0.$$

6. Найдем теперь решение уравнения четвертого порядка

$$y(j+2) - y(j+1) + 2y(j) - y(j-1) + y(j-2) = 0, \quad 2 \leq j \leq N-2, \quad (15)$$

удовлетворяющее следующим краевым условиям:

$$\begin{aligned} 2y(2) - y(1) + y(0) &= 2, \\ y(3) - y(2) + y(1) - y(0) &= 0, \\ y(N-3) - y(N-2) + y(N-1) - y(N) &= 0, \\ 2y(N-2) - y(N-1) + y(N) &= 0. \end{aligned} \quad (16)$$

Характеристическое уравнение

$$q^4 - q^3 + 2q^2 - q + 1 = (q^2 - q + 1)(q^2 + 1) = 0,$$

соответствующее (15), имеет простые комплексные корни $q_1 = -\cos \frac{\pi}{3} + i \sin \frac{\pi}{3}$, $q_2 = \cos \frac{\pi}{3} - i \sin \frac{\pi}{3}$, $q_3 = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2}$, $q_4 = \cos \frac{\pi}{2} - i \sin \frac{\pi}{2}$, $i = \sqrt{-1}$. Следовательно, общее решение однородного уравнения (15), принимающее действительные значения, имеет вид

$$y(j) = c_1 \cos \frac{1}{3} \pi j + c_2 \sin \frac{1}{3} \pi j + c_3 \cos \frac{1}{2} \pi j + c_4 \sin \frac{1}{2} \pi j. \quad (17)$$

Выделим теперь из общего решения (17) решение, которое удовлетворяет краевым условиям (16). Для этого подставим (17) в (16) и получим следующую систему для постоянных c_1 , c_2 , c_3 и c_4 :

$$\begin{aligned} \cos \frac{2\pi}{3} c_1 &+ \sin \frac{2\pi}{3} c_2 &- c_3 - c_4 &= 2, \\ c_1 &+ 0 \cdot c_2 &+ 0 \cdot c_3 + 0 \cdot c_4 &= 0, \\ \cos \frac{N\pi}{3} c_1 &+ \sin \frac{N\pi}{3} c_2 &+ 0 \cdot c_3 + 0 \cdot c_4 &= 0, \\ \cos \frac{(N-2)\pi}{3} c_1 &+ \sin \frac{(N-2)\pi}{3} c_2 &- \left(\cos \frac{\pi N}{2} + \sin \frac{\pi N}{2} \right) c_3 + \\ &&+ \left(\cos \frac{\pi N}{2} - \sin \frac{\pi N}{2} \right) c_4 &= 0. \end{aligned}$$

Определитель этой системы равен $-2 \sin \frac{N\pi}{3} \cos \frac{N\pi}{2}$ и отличен от нуля, если N четно, но не кратно 3.

В этом случае, учитывая четность N , получим $c_1 = c_2 = 0$, $c_3 = c_4 = -1$. Таким образом, если N четно и не кратно 3, то решение краевой задачи (15), (16) существует и дается формулой

$$y(j) = -\cos \frac{\pi j}{2} - \sin \frac{\pi j}{2}, \quad 0 \leq j \leq N.$$

Если N нечетно или кратно 3, то решение задачи (15), (16) либо не существует, либо неединственно. Этот пример иллюстрирует различие между краевыми задачами, решение которых существует не всегда, и задачей Коши, обладающей единственным решением.

§ 4. Уравнения второго порядка с постоянными коэффициентами

1. Общее решение однородного уравнения. Настоящий параграф посвящен разностным уравнениям второго порядка с постоянными коэффициентами

$$a_2 y(j+2) + a_1 y(j+1) + a_0 y(j) = f(j), \quad a_0, a_2 \neq 0. \quad (1)$$

Сначала найдем общее решение соответствующего однородного

уравнения

$$a_2y(j+2) + a_1y(j+1) + a_0y(j) = 0. \quad (2)$$

Характеристическое уравнение $a_2q^2 + a_1q + a_0 = 0$ имеет корни

$$q_1 = \frac{-a_1 + \sqrt{a_1^2 - 4a_0a_2}}{2a_2}, \quad q_2 = \frac{-a_1 - \sqrt{a_1^2 - 4a_0a_2}}{2a_2}.$$

Согласно общей теории разностных уравнений с постоянными коэффициентами, изложенной в § 3, линейно независимыми решениями уравнения (2) являются функции $v_1(j) = q_1^j$, $v_2(j) = q_2^j$, если $a_1^2 \neq 4a_0a_2$, и $v_1(j) = q_1^j$, $v_2(j) = jq_1^j$, если $a_1^2 = 4a_0a_2$. Для дальнейшего нам будет удобно использовать другие линейно независимые решения

$$v_1(j) = \frac{q_2q_1^j - q_1q_2^j}{q_2 - q_1}, \quad v_2(j) = \frac{q_2^j - q_1^j}{q_2 - q_1}, \quad (3)$$

принимающие при $j=0$ и $j=1$ следующие значения:

$$v_1(0) = 1, \quad v_1(1) = 0, \quad v_2(0) = 0, \quad v_2(1) = 1. \quad (4)$$

Очевидно, нужно только показать, что функции (3) в случае $a_1^2 = 4a_0a_2$ являются решениями однородного уравнения. Линейная независимость построенных функций (3) следует из условия $\Delta_0(v_1, v_2) \neq 0$, где

$$\Delta_0(v_1, v_2) = \begin{vmatrix} v_1(0) & v_1(1) \\ v_2(0) & v_2(1) \end{vmatrix}.$$

Переходя к пределу в (3) при q_2 , стремящемся к q_1 , получим функции $v_1(j) = -(j-1)q_1^j$, $v_2(j) = jq_1^{j-1}$, которые действительно являются решениями однородного уравнения (2). Заметим, что функции $v_1(j)$ и $v_2(j)$ из (3) принимают действительные значения и в том случае, когда корни q_1 и q_2 комплексны. Это позволяет не рассматривать отдельно случай комплексных корней. Итак, общее решение однородного уравнения (2) может быть записано в виде

$$\bar{\bar{y}}(j) = c_1v_1(j) + c_2v_2(j) = c_1 \frac{q_2q_1^j - q_1q_2^j}{q_2 - q_1} + c_2 \frac{q_2^j - q_1^j}{q_2 - q_1}, \quad (5)$$

где c_1 и c_2 —произвольные постоянные. Заметим, что в силу (4) будем иметь $\bar{\bar{y}}(0) = c_1$, $\bar{\bar{y}}(1) = c_2$.

Рассмотрим пример. Требуется найти общее решение однородного уравнения

$$y(j+2) - 2xy(j+1) + y(j) = 0, \quad (6)$$

где x —параметр, принимающий любые действительные значения. В этом случае имеем

$$q_1 = x + \sqrt{x^2 - 1}, \quad q_2 = \frac{1}{q_1}, \quad q_2 - q_1 = -2\sqrt{x^2 - 1}. \quad (7)$$

Подставляя (7) в (5), получим общее решение уравнения (6) для любого x в виде

$$y(j) = -\frac{(x + \sqrt{x^2 - 1})^{j-1} - (x + \sqrt{x^2 - 1})^{-(j-1)}}{2\sqrt{x^2 - 1}} y(0) + \\ + \frac{(x + \sqrt{x^2 - 1})^j - (x + \sqrt{x^2 - 1})^{-j}}{2\sqrt{x^2 - 1}} y(1). \quad (8)$$

В частности, если $|x| \leq 1$, то формула (8) может быть записана в виде

$$y(j) = -\frac{\sin(j-1)\arccos x}{\sin \arccos x} y(0) + \frac{\sin j \arccos x}{\sin \arccos x} y(1). \quad (9)$$

(Для получения (9) было использовано тождество $x = \cos(\arccos x)$).

Воспользуемся полученным результатом для решения поставленной в п. 4 § 1 задачи о вычислении интегралов

$$I_k(\varphi) = \int_0^\pi \frac{\cos k\psi - \cos k\varphi}{\cos \psi - \cos \varphi} d\psi, \quad k = 0, 1, \dots$$

Там было показано, что эта задача сводится к решению задачи Коши для уравнения

$$I_{k+1} - 2 \cos \varphi I_k + I_{k-1} = 0, \quad I_0 = 0, \quad I_1 = \pi. \quad (10)$$

Это уравнение есть частный случай (6) с $x = \cos \varphi$. Так как $|x| \leq 1$, то общее решение уравнения (10) дается формулой (9), т. е.

$$I_k = -\frac{\sin(k-1)\varphi}{\sin \varphi} I_0 + \frac{\sin k\varphi}{\sin \varphi} I_1.$$

Подставляя сюда начальные данные для I_k , получим решение поставленной задачи

$$I_k(\varphi) = \pi \frac{\sin k\varphi}{\sin \varphi}.$$

В качестве второго примера рассмотрим решение краевой задачи

$$y(j+1) - y(j) + y(j-1) = 0, \quad 1 \leq j \leq N-1, \\ y(0) = 1, \quad y(N) = 0. \quad (11)$$

Уравнение задачи (11) также есть частный случай (6), соответствующий значению $x = 1/2$. Формула (9) дает следующее общее решение уравнения (11):

$$y(j) = \left(c_1 \sin \frac{(j-1)\pi}{3} + c_2 \sin \frac{j\pi}{3} \right) / \sin \frac{\pi}{3}.$$

Постоянные c_1 и c_2 находятся из краевых условий для $y(j)$. Если N не кратно 3, то $c_1 = -1$, $c_2 = \sin \frac{1}{3}\pi (N-1) / \sin \frac{1}{3}\pi N$ и

решение задачи (11) имеет вид

$$y(j) = \sin \frac{1}{3}(N-j)\pi / \sin \frac{1}{3}N\pi, \quad 0 \leq j \leq N.$$

Если N кратно 3, то решение краевой задачи (11) не существует.

2. Полиномы Чебышева. Вернемся теперь к уравнению (6). Сначала рассмотрим следующую задачу Коши:

$$\begin{aligned} y(n+2) - 2xy(n+1) + y(n) &= 0, & n \geq 0, \\ y(0) = 1, \quad y(1) = x. \end{aligned} \quad (12)$$

Заметим, что из (12) следует

$$y(2) = 2xy(1) - y(0) = 2x^2 - 1,$$

$$y(3) = 2xy(2) - y(1) = 4x^3 - 3x,$$

и вообще $y(n)$ есть полином n -й степени от x . Обозначим этот полином $T_n(x)$. Подставляя $T_n(x)$ вместо $y(n)$ в (12), получим рекуррентное соотношение, которому удовлетворяет этот полином

$$\begin{aligned} T_{n+2}(x) &= 2xT_{n+1}(x) - T_n(x), & n \geq 0, \\ T_0(x) &= 1, \quad T_1(x) = x, \quad -\infty < x < \infty. \end{aligned} \quad (13)$$

С другой стороны, общее решение уравнения (12) дается формулой (8) для любого x . Подставляя в (8) начальные значения для $y(n)$, будем иметь

$$T_n(x) = \frac{(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^{-n}}{2}. \quad (14)$$

В частности, если $|x| \leq 1$, то, полагая здесь $x = \cos(\arccos x)$, получим

$$T_n(x) = \cos(n \arccos x), \quad |x| \leq 1.$$

Итак, решение задачи (12) найдено. Решение есть полином $T_n(x)$, который для любого x определяется формулой (14) или формулой

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leq 1, \\ \frac{1}{2} [(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^{-n}], & |x| \geq 1. \end{cases} \quad (15)$$

Полином $T_n(x)$ называется *полиномом Чебышева первого рода степени n* .

Рассмотрим теперь другую задачу Коши для уравнения (6)

$$\begin{aligned} y(n+2) - 2xy(n+1) + y(n) &= 0, & n \geq 0, \\ y(0) = 1, \quad y(1) = 2x. \end{aligned} \quad (16)$$

Очевидно, что и здесь $y(n)$ — полином n -й степени от x . Обозначим его через $U_n(x)$. Получим явный вид для $U_n(x)$. Подставляя

начальные значения для $y(n)$ в (8), будем иметь для любого x :

$$\begin{aligned} U_n(x) &= \frac{2x(x + \sqrt{x^2 - 1})^n - (x + \sqrt{x^2 - 1})^{n-1}}{2\sqrt{x^2 - 1}} + \\ &+ \frac{(x + \sqrt{x^2 - 1})^{-(n-1)} - 2x(x + \sqrt{x^2 - 1})^{-n}}{2\sqrt{x^2 - 1}} = \\ &= \frac{(x + \sqrt{x^2 - 1})^{n+1} - (x + \sqrt{x^2 - 1})^{-(n+1)}}{2\sqrt{x^2 - 1}}. \quad (17) \end{aligned}$$

В частности, если $|x| \leq 1$, то

$$U_n(x) = \frac{\sin(n+1)\arccos x}{\sin \arccos x}.$$

Полином $U_n(x)$ называется *полиномом Чебышева второго рода степени n* и определяется формулами

$$U_n(x) = \begin{cases} \frac{\sin(n+1)\arccos x}{\sin \arccos x}, & |x| \leq 1, \\ \frac{1}{2\sqrt{x^2 - 1}} [(x + \sqrt{x^2 - 1})^{n+1} - (x + \sqrt{x^2 - 1})^{-(n+1)}], & |x| \geq 1. \end{cases} \quad (18)$$

Из (16) получим для полиномов $U_n(x)$ следующее рекуррентное соотношение:

$$\begin{aligned} U_{n+2}(x) &= 2xU_{n+1}(x) - U_n(x), \quad n \geq 0, \\ U_0(x) &= 1, \quad U_1(x) = 2x. \end{aligned} \quad (19)$$

Формула (17) позволяет получить вместо (8) следующее представление для общего решения уравнения (6):

$$y(n) = -c_1 U_{n-2}(x) + c_2 U_{n-1}(x).$$

Получим еще одно представление для общего решения уравнения (6). Покажем, что функции $v_1(n) = T_n(x)$ и $v_2(n) = U_{n-1}(x)$ являются линейно независимыми решениями однородного уравнения (6). Действительно, нужно показать лишь их линейную независимость. Так как определитель

$$\Delta_0(v_1, v_2) = \begin{vmatrix} T_0(x) & T_1(x) \\ U_{-1}(x) & U_0(x) \end{vmatrix} = \begin{vmatrix} 1 & x \\ 0 & 1 \end{vmatrix} = 1$$

отличен от нуля, то утверждение справедливо. Следовательно, общее решение уравнения (6) можно представить в виде

$$y(n) = c_1 T_n(x) + c_2 U_{n-1}(x), \quad (20)$$

где c_1 и c_2 —произвольные постоянные, а функции $T_n(x)$ и $U_n(x)$ для любых x и n определяются формулами (14) и (17).

В заключение приведем некоторые легко проверяемые соотношения, выражающие связи между полиномами Чебышева

$T_n(x)$ и $U_n(x)$, а также свойства этих полиномов. Имеют место следующие формулы:

$$T_n(x) = T_{-n}(x), \quad U_{-n}(x) = -U_{n-2}(x), \quad n \geq 0, \quad (21)$$

$$T_{in}(x) = T_i(T_n(x)), \quad U_{in-1}(x) = U_{i-1}(T_n(x)), \quad (22)$$

$$T_{2n}(x) = 2(T_n(x))^2 - 1, \quad (23)$$

$$T_{n-1}(x) - xT_n(x) = (1 - x^2)U_{n-1}(x), \quad (24)$$

$$U_{n-1}(x) - xU_n(x) = -T_{n+1}(x), \quad (25)$$

$$U_{n+i}(x) + U_{n-i}(x) = 2T_i(x)U_n(x). \quad (26)$$

Из (26) при соответствующей замене индексов i и n получим

$$U_{n+i-1}(x) + U_{n-i-1}(x) = 2T_i(x)U_{n-1}(x), \quad (27)$$

$$U_{n+i}(x) + U_{n-i-2}(x) = 2T_{i+1}(x)U_{n-1}(x). \quad (28)$$

Полагая в (26) — (28) $i = n$, будем иметь

$$2T_n(x)U_n(x) = U_{2n}(x) + 1, \quad (29)$$

$$2T_n(x)U_{n-1}(x) = U_{2n-1}(x), \quad (30)$$

$$2T_{n+1}(x)U_{n-1}(x) = U_{2n}(x) - 1. \quad (31)$$

Здесь были учтены равенства (21) и $U_0(x) = 1$, $U_{-1}(x) = 0$. Если положить в (26) $n = 0$, то получим

$$2T_n(x) = U_n(x) - U_{n-2}(x). \quad (32)$$

3. Общее решение неоднородного уравнения. Построим теперь общее решение неоднородного уравнения (1)

$$a_2y(n+2) + a_1y(n+1) + a_0y(n) = f(n). \quad (33)$$

В силу теоремы 3 общее решение уравнения (33) есть сумма $y(n) = \bar{y}(n) + \bar{\bar{y}}(n)$, где $\bar{y}(n)$ — общее решение однородного уравнения (2), а $\bar{\bar{y}}(n)$ — частное решение неоднородного уравнения (33).

Выше было показано, что линейно независимыми решениями уравнения (2) являются функции

$$v_1(n) = \frac{q_2 q_1^n - q_1 q_2^n}{q_2 - q_1}, \quad v_2(n) = \frac{q_2^n - q_1^n}{q_2 - q_1}, \quad (34)$$

а решение $\bar{\bar{y}}(n)$ определяется формулой (5):

$$\bar{\bar{y}}(n) = c_1 v_1(n) + c_2 v_2(n).$$

Для нахождения частного решения $\bar{\bar{y}}(n)$ уравнения (33) воспользуемся методом вариации постоянных, изложенным в п. 3 § 2. Формула (19) § 2 дает решение $\bar{\bar{y}}(n)$ в следующем виде:

$$\bar{\bar{y}}(n) = \sum_{k=n_0}^{n-1} \begin{vmatrix} v_1(k+1) & v_2(k+1) \\ v_1(k) & v_2(k) \\ \hline v_1(k+1) & v_1(k+2) \\ v_2(k+1) & v_2(k+2) \end{vmatrix} \cdot \frac{f(k)}{a_2}.$$

В результате несложных вычислений будем иметь

$$\bar{y}(n) = \sum_{k=n_0}^{n-2} \frac{q_2^{n-k-1} - q_1^{n-k-1}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}, \quad n \neq n_0, n_0 + 1$$

и

$$\bar{y}(n_0) = \bar{y}(n_0 + 1) = 0.$$

Следовательно, общее решение неоднородного уравнения (33) имеет вид

$$y(n) = c_1 \frac{q_2^n - q_1^n}{q_2 - q_1} + c_2 \frac{q_2^n - q_1^n}{q_2 - q_1} + \sum_{k=n_0}^{n-2} \frac{q_2^{n-k-1} - q_1^{n-k-1}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}, \quad (35)$$

где c_1 и c_2 — произвольные постоянные.

Если решается задача Коши, т. е. ищется решение уравнения (33), удовлетворяющее условиям

$$y(n_0) = y_0, \quad y(n_0 + 1) = y_1, \quad (36)$$

то из (35) и (36) получим следующее представление для решения этой задачи:

$$y(n) = y_0 \frac{q_2^{n-n_0} - q_1^{n-n_0}}{q_2 - q_1} + y_1 \frac{q_2^{n-n_0} - q_1^{n-n_0}}{q_2 - q_1} + \sum_{k=n_0}^{n-2} \frac{q_2^{n-k-1} - q_1^{n-k-1}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}. \quad (37)$$

Найдем теперь решение первой краевой задачи для разностного уравнения второго порядка с постоянными коэффициентами. Будет удобно записывать такую задачу в следующем виде:

$$a_2 y(n+1) + a_1 y(n) + a_0 y(n-1) = -f(n), \quad 1 \leq n \leq N-1, \quad (38)$$

$$y(0) = \mu_1, \quad y(N) = \mu_2.$$

Эта запись отличается от (33) сдвигом индекса n , поэтому, используя (35), получим следующую формулу для общего решения уравнения (38):

$$y(n) = c_1 \frac{q_2^n - q_1^n}{q_2 - q_1} + c_2 \frac{q_2^n - q_1^n}{q_2 - q_1} - \sum_{k=1}^{n-1} \frac{q_2^{n-k} - q_1^{n-k}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}. \quad (39)$$

Определим постоянные c_1 и c_2 из условия, чтобы решение (39) принимало при $n=0$ и $n=N$ заданные значения $y(0)=\mu_1$ и $y(N)=\mu_2$. Опуская несложные выкладки, получим следующую

формулу для решения краевой задачи (38):

$$y(n) = \frac{(q_1 q_2)^n (q_2^{N-n} - q_1^{N-n})}{q_2^N - q_1^N} \mu_1 + \frac{q_2^n - q_1^n}{q_2^N - q_1^N} \mu_2 + \\ + \sum_{k=1}^{n-1} \frac{(q_1 q_2)^{n-k} (q_2^{N-n} - q_1^{N-n}) (q_2^k - q_1^k)}{(q_2 - q_1) (q_2^N - q_1^N)} \cdot \frac{f(k)}{a_2} + \\ + \sum_{k=n}^{N-1} \frac{(q_2^{N-k} - q_1^{N-k}) (q_2^n - q_1^n)}{(q_2 - q_1) (q_2^N - q_1^N)} \cdot \frac{f(k)}{a_2}. \quad (40)$$

Заметим, что решение краевой задачи (38) не существует лишь в случае, когда $q_1^N = q_2^N$, но $q_1 \neq q_2$.

Рассмотрим теперь частные случаи использования формулы (40). Пусть требуется решить первую краевую задачу для уравнения

$$\begin{aligned} y(n+1) - 2xy(n) + y(n-1) &= -f(n), \quad 1 \leq n \leq N-1, \\ y(0) = \mu_1, \quad y(N) = \mu_2. \end{aligned} \quad (41)$$

Выше были найдены корни q_1 и q_2 соответствующего (41) характеристического уравнения

$$q_1 = x + \sqrt{x^2 - 1}, \quad q_2 = x - \sqrt{x^2 - 1} = 1/q_1.$$

Подставляя эти значения в (40) и учитывая формулу (17) для полинома $U_n(x)$, получим решение задачи (41) в следующем виде:

$$y(n) = \frac{U_{N-n-1}(x)}{U_{N-1}(x)} \left[\mu_1 + \sum_{k=1}^{n-1} U_{k-1}(x) f(k) \right] + \\ + \frac{U_{n-1}(x)}{U_{N-1}(x)} \left[\mu_2 + \sum_{k=n}^{N-1} U_{N-k-1}(x) f(k) \right]. \quad (42)$$

Решение существует и дается формулой (42), если выполнено условие $x \neq \cos \frac{k\pi}{N}$, $k = 1, 2, \dots, N-1$.

Вернемся к уравнению (38). Если $a_0 a_2 > 0$, то решение (40) этой задачи может быть записано в более компактной, чем (40), форме. Действительно, запишем корни

$$q_1 = \frac{1}{2a_2} [-a_1 + \sqrt{a_1^2 - 4a_0 a_2}], \quad q_2 = \frac{1}{2a_2} [-a_1 - \sqrt{a_1^2 - 4a_0 a_2}]$$

характеристического уравнения, соответствующего (38), в следующем виде:

$$q_1 = \rho(x + \sqrt{x^2 - 1}), \quad q_2 = \rho(x - \sqrt{x^2 - 1}), \quad (43)$$

где

$$\rho = \sqrt{\frac{a_0}{a_2}}, \quad x = -\frac{a_1}{2\sqrt{a_0 a_2}}. \quad (44)$$

Подставим (43) в (40) и учтем формулу (17). Получим решение задачи (38) для случая $a_0 a_2 > 0$ в виде

$$y(n) = \frac{U_{N-n-1}(x)}{U_{N-1}(x)} \rho^n \left[\mu_1 + \sum_{k=1}^{n-1} \frac{U_{k-1}(x)}{\rho^{k-1}} \cdot \frac{f(k)}{a_0} \right] + \\ + \frac{U_{n-1}(x)}{U_{N-1}(x)} \cdot \frac{1}{\rho^{N-n}} \left[\mu_2 + \sum_{k=n}^{N-1} \rho^{N-k-1} U_{N-k-1}(x) \frac{f(k)}{a_0} \right],$$

где ρ и x определены в (44). Решение задачи (38) для случая $a_0 a_2 > 0$ существует, если выполнено условие $a_1 + 2\sqrt{a_0 a_2} \cos \frac{k\pi}{N} \neq 0$, $k = 1, 2, \dots, N-1$.

Рассмотрим теперь первую краевую задачу для трехточечного векторного уравнения с постоянными коэффициентами

$$\begin{aligned} Y_{n-1} - C Y_n + Y_{n+1} &= -F_n, \quad 1 \leq n \leq N-1 \\ Y_0 &= F_0, \quad Y_N = F_N, \end{aligned} \quad (45)$$

где Y_n и F_n —векторы, а C —квадратная матрица. Легко проверить, что общее решение неоднородного уравнения (45) имеет вид

$$Y_n = U_{n-2} \left(\frac{1}{2} C \right) C_1 + U_{n-1} \left(\frac{1}{2} C \right) C_2 - \sum_{k=1}^{n-1} U_{n-k-1} \left(\frac{1}{2} C \right) F_k,$$

где C_1 и C_2 —произвольные векторы, а $U_n(X)$ есть матричный полином от матрицы X , определяемый по рекуррентным формулам (19).

Если матрица C такова, что $U_{N-1} \left(\frac{1}{2} C \right)$ невырожденная матрица, то решение краевой задачи (45) определяется формулой, аналогичной формуле (42)

$$Y_n = U_{N-1}^{-1} \left(\frac{1}{2} C \right) U_{N-n-1} \left(\frac{1}{2} C \right) \left[F_0 + \sum_{k=1}^{n-1} U_{k-1} \left(\frac{1}{2} C \right) F_k \right] + \\ + U_{N-1}^{-1} \left(\frac{1}{2} C \right) U_{n-1} \left(\frac{1}{2} C \right) \left[F_N + \sum_{k=n}^{N-1} U_{N-k-1} \left(\frac{1}{2} C \right) F_k \right]. \quad (46)$$

Ниже будет показано, что к задаче (45) сводится разностная задача Дирихле для уравнения Пуассона в прямоугольнике.

В заключение отметим, что условию существования решения задачи (45) можно придать следующую формулировку: решение существует и определяется формулой (46), если числа $\cos \frac{k\pi}{N}$, $k = 1, 2, \dots, N-1$, не являются собственными значениями матрицы C .

§ 5. Разностные задачи на собственные значения

1. Первая краевая задача на собственные значения. В главе IV будет рассмотрен метод разделения переменных, который используется для нахождения решений сеточных краевых задач для эллиптических уравнений в прямоугольнике. В связи с этим возникает необходимость представления искомых сеточных функций в виде разложения по собственным функциям соответствующей разностной задачи. В данном параграфе мы рассмотрим разностные задачи на собственные значения для простейшего разностного оператора второго порядка, заданного на равномерной сетке.

Сформулируем первую краевую задачу. Пусть на отрезке $[0, l]$ введена равномерная сетка $\omega = \{x_i = ih, i = 0, 1, \dots, N, hN = l\}$ с шагом h . Требуется найти такие значения параметра λ (собственные значения), при которых существуют нетривиальные решения $y(x_i)$ (собственные функции) следующей разностной задачи:

$$y_{xx} + \lambda y = 0, \quad x \in \omega, \quad y(0) = y(l) = 0, \quad (1)$$

где

$$y_{xx,i} = \frac{y(i+1) - 2y(i) + y(i-1)}{h^2}, \quad y(i) = y(x_i).$$

Найдем решение задачи (1). Для этого запишем (1) в виде краевой задачи для разностного уравнения второго порядка

$$y(i+1) - 2\left(1 - \frac{h^2\lambda}{2}\right)y(i) + y(i-1) = 0, \quad 1 \leq i \leq N-1, \\ y(0) = y(N) = 0. \quad (2)$$

В п. 1 § 4 было показано, что общее решение уравнения (2) имеет вид (см. формулу (20) § 4) $y(i) = c_1 T_i(z) + c_2 U_{i-1}(z)$, где c_1 и c_2 — произвольные постоянные, а через z здесь обозначено

$$z = 1 - h^2\lambda/2. \quad (3)$$

Постоянные c_1 и c_2 определим из граничных условий

$$y(0) = c_1 = 0, \quad y(N) = c_2 U_{N-1}(z) = 0. \quad (4)$$

Здесь и далее мы используем формулы (15) и (18) § 4, определяющие полиномы Чебышева первого и второго рода, а также формулы (21)–(32) из того же параграфа.

Так как ищется нетривиальное решение задачи (1), то $c_2 \neq 0$, и из (4) будем иметь условие $U_{N-1}(z) = 0$, при выполнении которого решение задачи (1) имеет вид $y_i = c_2 U_{i-1}(z)$.

Так как числа $z_k = \cos \frac{k\pi}{N}$, $k = 1, 2, \dots, N-1$, суть корни полинома $U_{N-1}(z)$, то из (3) находим собственные значения задачи (1)

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi}{2N} = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2l}, \quad k = 1, 2, \dots, N-1. \quad (5)$$

Каждому собственному значению λ_k соответствует ненулевое решение задачи (1)

$$y_k(i) = c_2 U_{i-1}(z_k) = \bar{c}_k \sin \frac{k\pi i}{N} = \bar{c}_k \sin \frac{k\pi x_i}{l},$$

$$0 \leq i \leq N \quad (c_2 = \bar{c}_k \sin \frac{k\pi}{N}). \quad (6)$$

Определим скалярное произведение сеточных функций, заданных на ω , следующим образом:

$$(u, v) = \sum_{i=1}^{N-1} u(i)v(i)h + 0.5h[u(0)v(0) + u(N)v(N)].$$

Определим теперь постоянную \bar{c}_k в (6) так, чтобы функции $y_k(i)$ имели норму, равную единице, т. е. $(y_k, y_k) = 1$.

Несложные выкладки дают $\bar{c}_k = \sqrt{2/l}$. Подставляя найденное значение для \bar{c}_k в (6), получим собственные функции $\mu_k(i)$ задачи (1)

$$\mu_k(i) = \sqrt{\frac{2}{l}} \sin \frac{k\pi i}{N} = \sqrt{\frac{2}{l}} \sin \frac{k\pi x_i}{l}, \quad (7)$$

$$i = 0, 1, \dots, N, \quad k = 1, 2, \dots, N-1.$$

Итак, задача (1) решена и решение дано в (5) и (7).

Перечислим основные свойства собственных функций и собственных значений первой краевой задачи (1).

1) Собственные функции ортонормированы:

$$(\mu_k, \mu_m) = \delta_{km}, \quad \delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m. \end{cases}$$

2) Для любой сеточной функции $f(i)$, заданной во внутренних узлах сетки ω , т. е. для $1 \leq i \leq N-1$, имеет место разложение

$$f(i) = \frac{2}{N} \sum_{k=1}^{N-1} \varphi_k \sin \frac{k\pi i}{N}, \quad i = 1, 2, \dots, N-1, \quad (8)$$

где

$$\varphi_k = \sum_{i=1}^{N-1} f(i) \sin \frac{k\pi i}{N}, \quad k = 1, 2, \dots, N-1. \quad (9)$$

Поясним это утверждение. Пусть $\bar{f}(i)$ — произвольная сеточная функция, заданная на ω (или на ω и обращающаяся в нуль при $i=0$ и $i=N$). Разложим ее по собственным функциям

$$\bar{f}(i) = \sum_{k=1}^{N-1} f_k \mu_k(i) = \sum_{k=1}^{N-1} \sqrt{\frac{2}{l}} \bar{f}_k \sin \frac{k\pi i}{N}, \quad (10)$$

где f_k — коэффициент Фурье функции $f(i)$. Умножая (10) скалярно на $\mu_m(i)$ и используя ортонормированность собственных функций, найдем коэффициенты Фурье

$$f_m = \sum_{k=1}^{N-1} f_k (\mu_k, \mu_m) = (f, \mu_m) = \sum_{i=1}^{N-1} \sqrt{\frac{2}{l}} f(i) \sin \frac{\pi k i}{N} h.$$

Связь полученных формул с (8) — (9) легко устанавливается, если заметить, что $f_m = \frac{\sqrt{2l}}{N} \Phi_m$.

Разложение (8), (9) удобно тем, что для вычисления фурьеобраза функции $f(i)$ и для восстановления исходной функции по ее образу необходимо вычислять однотипную сумму. Алгоритм быстрого вычисления сумм такого вида будет рассмотрен в главе IV.

3) Для собственных значений справедливы неравенства

$$\frac{8}{h^2} \leq \frac{4}{h^2} \sin^2 \frac{\pi}{2N} = \lambda_1 \leq \lambda_k \leq \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi}{2N}, \quad 1 \leq k \leq N-1.$$

2. Вторая краевая задача. Рассмотрим теперь вторую краевую задачу на собственные значения

$$\begin{aligned} y_{xx} + \lambda y &= 0, \quad x \in \omega, \\ \frac{2}{h} y_x + \lambda y &= 0, \quad x = 0, \quad -\frac{2}{h} y_x + \lambda y = 0, \quad x = l. \end{aligned} \quad (11)$$

Найдем решение задачи (11). Расписывая разностные производные в (11) по точкам, получим задачу

$$\begin{aligned} y(i+1) - 2zy(i) + y(i-1) &= 0, \quad 1 \leq i \leq N-1, \\ y(1) - zy(0) &= 0, \quad y(N-1) - zy(N) = 0, \end{aligned} \quad (12)$$

где $z = 1 - \lambda h^2/2$. Из общего решения уравнения (12) $y(i) = c_1 T_i(z) + c_2 U_{i-1}(z)$ выделим решение, удовлетворяющее поставленным краевым условиям. Используя формулу (24) § 4, будем иметь

$$y(1) - zy(0) = c_1 z + c_2 - c_1 z = c_2 = 0, \quad c_2 = 0,$$

а также

$$y(N-1) - zy(N) = c_1 (T_{N-1}(z) - zT_N(z)) = c_1 (1 - z^2) U_{N-1}(z) = 0.$$

Так как $c_1 \neq 0$, то отсюда получим

$$z_k = \cos \frac{k\pi}{N}, \quad k = 0, 1, \dots, N,$$

и, следовательно, собственными значениями задачи (12) являются

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi}{2N} = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2l}, \quad k = 0, 1, \dots, N. \quad (13)$$

При этом каждому λ_k соответствует ненулевое решение задачи (11)

$$y_k(i) = c_k T_i(z_k) = c_k \cos \frac{k\pi i}{N}, \quad 0 \leq i \leq N.$$

Выберем постоянные c_k из условия $(y_k, y_k) = 1$, где скалярное произведение определено выше. Непосредственные вычисления показывают, что

$$c_k = \sqrt{2/l}, \quad k = 1, 2, \dots, N-1, \quad c_k = \sqrt{1/l}, \quad k = 0, N.$$

Таким образом, нормированными собственными функциями задачи (11) являются функции

$$\begin{aligned} \mu_k(i) &= \sqrt{\frac{2}{l}} \cos \frac{k\pi i}{N} = \sqrt{\frac{2}{l}} \cos \frac{k\pi x_i}{l}, \quad 1 \leq k \leq N-1, \\ \mu_0(i) &= \sqrt{\frac{1}{l}} \cos \frac{k\pi i}{N} = \sqrt{\frac{1}{l}} \cos \frac{k\pi x_i}{l}, \quad k = 0, N, \end{aligned} \quad (14)$$

определенные на сетке $\bar{\omega}$. Отметим, что собственной функцией, соответствующей нулевому собственному значению $\lambda_0 = 0$, является постоянная $\mu_0(i) = \sqrt{1/l}$.

Сформулируем свойства собственных функций и собственных значений второй краевой задачи (11).

1) Собственные функции ортонормированы: $(\mu_k, \mu_m) = \delta_{km}$.

2). Для любой сеточной функции $f(i)$, заданной на $\bar{\omega}$, имеет место разложение

$$f(i) = \frac{2}{N} \sum_{k=0}^N \rho_k \varphi_k \cos \frac{k\pi i}{N}, \quad i = 0, 1, \dots, N, \quad (15)$$

где

$$\varphi_k = \sum_{i=0}^N \rho_i f(i) \cos \frac{k\pi i}{N}, \quad k = 0, 1, \dots, N, \quad (16)$$

$$\rho_i = \begin{cases} 1, & 1 \leq i \leq N-1, \\ 0,5, & i = 0, N. \end{cases} \quad (17)$$

Формулы (15) и (16) суть модификация традиционного разложения $f(i)$ по собственным функциям $\mu_k(i)$

$$f(i) = \sum_{k=0}^N f_k \mu_k(i), \quad f_k = (f, \mu_k)$$

путем следующей замены:

$$f_k = \begin{cases} \frac{\sqrt{2/l}}{N} \varphi_k, & 1 \leq k \leq N-1, \\ \frac{1}{N} \sqrt{l} \varphi_k, & k = 0, N. \end{cases}$$

3) Для собственных значений справедливы неравенства

$$0 = \lambda_0 \leq \lambda_k \leq \lambda_N, \quad 0 \leq k \leq N.$$

3. Смешанная краевая задача. Рассмотрим теперь задачу на собственные значения, когда на одной стороне отрезка $[0, l]$ за-

дано краевое условие первого рода, а на другой—второго, например:

$$\begin{aligned} y_{xx} + \lambda y &= 0, \quad x \in \omega, \\ y(0) = 0, \quad -\frac{2}{h} y_x + \lambda y &= 0, \quad x = l. \end{aligned} \quad (18)$$

Такую задачу мы будем называть *смешанной краевой задачей*.

Найдем решение задачи (18). Соответствующая (18) задача для разностного уравнения второго порядка имеет вид

$$\begin{aligned} y(i+1) - 2zy(i) + y(i-1) &= 0, \quad 1 \leq i \leq N-1, \\ y(0) = 0, \quad y(N-1) - zy(N) &= 0, \end{aligned}$$

где $z = 1 - 0,5\lambda h^2$. Выделим из общего решения этого уравнения

$$y(i) = c_1 T_i(z) + c_2 U_{i-1}(z)$$

решение, удовлетворяющее заданным краевым условиям. Используя (25) § 4, получим

$$\begin{aligned} y(0) &= c_1 = 0, \\ y(N-1) - zy(N) &= c_2 (U_{N-2}(z) - zU_{N-1}(z)) = -c_2 T_N(z) = 0. \end{aligned}$$

Так как $c_2 \neq 0$, то отсюда найдем $T_N(z_k) = 0$, где $z_k = \cos \frac{(2k-1)\pi}{2N}$, $k = 1, 2, \dots, N$ и, следовательно, собственными значениями задачи (18) являются числа

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{(2k-1)\pi}{4N} = \frac{4}{h^2} \sin^2 \frac{(2k-1)\pi h}{4l}, \quad k = 1, 2, \dots, N. \quad (19)$$

Нормированными собственными функциями задачи (18), соответствующими собственным значениям λ_k , являются

$$\begin{aligned} \mu_k(i) &= \sqrt{\frac{2}{l}} \sin \frac{(2k-1)\pi i}{2N} = \\ &= \sqrt{\frac{2}{l}} \sin \frac{(2k-1)\pi x_i}{2l}, \quad k = 1, 2, \dots, N. \end{aligned} \quad (20)$$

Сформулируем свойства собственных функций и собственных значений смешанной краевой задачи (18).

1) Собственные функции ортонормированы: $(\mu_k, \mu_m) = \delta_{km}$.

2) Для любой сеточной функции $f(i)$, заданной на $\omega^+ = \{x_i = ih, 1 \leq i \leq N\}$ (или на $\bar{\omega}$, и обращающейся в нуль при $i = 0$) справедливо разложение

$$f(i) = \frac{2}{N} \sum_{k=1}^N \varphi_k \sin \frac{(2k-1)\pi i}{2N}, \quad i = 1, 2, \dots, N, \quad (21)$$

где

$$\varphi_k = \sum_{i=1}^N \rho_i f(i) \sin \frac{(2k-1)\pi i}{2N}, \quad k = 1, 2, \dots, N, \quad (22)$$

а ρ_i определено в (17).

3) Для собственных значений справедливы неравенства

$$\frac{8}{(2 + \sqrt{2})l^2} \leq \frac{4}{h^2} \sin^2 \frac{\pi}{2N} = \lambda_i \leq \lambda_k \leq \lambda_N = \frac{4}{h^2} \cos^2 \frac{\pi}{4N}, \quad 1 \leq k \leq N.$$

Если для уравнения (18) краевое условие первого рода задано на правом конце отрезка $[0, l]$, т. е. дана задача

$$\begin{aligned} y_{xx} + \lambda y &= 0, \quad x \in \omega, \\ \frac{2}{h} y_x + \lambda y &= 0, \quad x = 0; \quad y(l) = 0, \end{aligned} \quad (23)$$

то собственные значения определяются формулой (19), а нормированными собственными функциями являются

$$\mu_k(i) = \sqrt{\frac{2}{l}} \sin \frac{(2k-1)(N-i)\pi}{2N} = \sqrt{\frac{2}{l}} \sin \frac{(2k-1)\pi(l-x_i)}{2l}, \quad k = 1, 2, \dots, N.$$

Имеет место следующее утверждение. Для любой сеточной функции $f(i)$, заданной на $\omega^- = \{x_i = ih, i = 0, 1, \dots, N-1, hN = l\}$ (или на ω и обращающейся в нуль при $i = N$), справедливо разложение

$$f(N-i) = \frac{2}{N} \sum_{k=1}^N \varphi_k \sin \frac{(2k-1)\pi i}{2N}, \quad i = 1, 2, \dots, N, \quad (24)$$

где

$$\varphi_i = \sum_{i=1}^N \rho_{N-i} f(N-i) \sin \frac{(2k-1)\pi i}{2N}, \quad k = 1, 2, \dots, N, \quad (25)$$

а ρ_i определено в (17).

Отметим, что построенные собственные функции задачи (23) также ортонормированы:

$$(\mu_k, \mu_m) = \delta_{km}.$$

4. Периодическая краевая задача. Пусть на сетке $\Omega = \{x_i = ih, i = 0, \pm 1, \pm 2, \dots\}$, введенной на прямой $-\infty < x < \infty$, ищется нетривиальное периодическое с периодом N решение следующей задачи на собственные значения:

$$\begin{aligned} y_{xx} + \lambda y &= 0, \quad x \in \Omega, \\ y(i+N) &= y(i), \quad i = 0, \pm 1, \pm 2, \dots, h = l/N. \end{aligned} \quad (26)$$

Так как решение периодическое, то его достаточно найти при $i=0, 1, \dots, N-1$. Расписывая (26) по точкам $i=0, 1, \dots, N-1$ и учитывая, что $y(-1)=y(N-1)$, $y(0)=y(N)$, получим следующую задачу:

$$\begin{aligned} y(i+1) - 2zy(i) + y(i-1) &= 0, \quad 0 \leq i \leq N-1, \\ y(0) &= y(N), \quad y(-1) = y(N-1), \end{aligned} \quad (27)$$

где $z = 1 - 0,5\lambda h^2$.

Найдем решение задачи (27). Подставим общее решение

$$y(i) = c_1 T_i(z) + c_2 U_{i-1}(z)$$

в краевые условия. Учитывая свойства полиномов Чебышева, получим следующую систему для определения постоянных c_1 и c_2 :

$$\begin{aligned} c_1(1 - T_N(z)) - c_2 U_{N-1}(z) &= 0, \\ c_1(T_{N-1}(z) - z) + c_2(1 + U_{N-2}(z)) &= 0. \end{aligned} \quad (28)$$

Эта система имеет ненулевое решение тогда и только тогда, когда ее определитель равен нулю. Вычислим его, используя для преобразований формулы (25), (29) и (31) § 4. Получим

$$\begin{aligned} (1 - T_N(z))(1 + U_{N-2}(z)) + (T_{N-1}(z) - z)U_{N-1}(z) &= \\ = 1 + U_{N-2}(z) - zU_{N-1}(z) - T_N(z) + T_{N-1}(z)U_{N-1}(z) - & \\ - T_N(z)U_{N-2}(z) &= 2[1 - T_N(z)] = 0. \end{aligned}$$

Отсюда следует, что при $z = z_k$, где

$$z_k = \cos \frac{2k\pi}{N}, \quad k = 0, 1, \dots, N-1, \quad (29)$$

система (28) имеет ненулевое решение. Таким образом, собственными значениями задачи (26) являются

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi}{N} = \frac{4}{h^2} \sin^2 \frac{k\pi h}{l}, \quad k = 0, 1, \dots, N-1. \quad (30)$$

Получим теперь решение системы (28). Так как имеют место равенства

$$\begin{aligned} T_{N-1}(z_k) &= z_k, \quad 0 \leq k \leq N-1, \\ U_{N-2}(z_k) &= \begin{cases} N-1, & k=0, N/2, \\ -1, & k \neq 0, N/2, \end{cases} \\ U_{N-1}(z_k) &= \begin{cases} N, & k=0, \\ -N, & k=N/2, \\ 0, & k \neq 0, N/2, \end{cases} \end{aligned}$$

то, подставляя (29) в (28), найдем следующее решение системы (28):

- а) для $k=0$ и $k=N/2$ имеем $c_2=0$, $c_1=c_1^{(k)} \neq 0$;
- б) для $k \neq 0, k \neq N/2$, $0 < k \leq N-1$, постоянные $c_1=c_1^{(k)}$, $c_2=c_2^{(k)}$ произвольны, но не равны нулю одновременно. Отсюда

получим, что функции

$$\begin{aligned} y_k(i) &= c_1^{(k)} \cos \frac{2k\pi i}{N}, \quad k = 0, N/2, \\ y_k(i) &= c_1^{(k)} \cos \frac{2k\pi i}{N} + c_2^{(k)} \sin \frac{2k\pi i}{N}, \quad 1 \leq k \leq N-1, k \neq 0, \frac{N}{2} \end{aligned} \quad (31)$$

являются решениями задачи (27), соответствующими собственному значению λ_k . Заметим, что в случае $k \neq 0, N/2$ формулы (31) определяют в действительности две линейно независимые функции $c_1^{(k)} \cos \frac{2k\pi i}{N}$ и $c_2^{(k)} \sin \frac{2k\pi i}{N}$, каждая из которых является решением задачи (27) и соответствует собственному значению λ_k .

Построим теперь нормированные собственные функции задачи (26). Отметим, что для периодических сеточных функций введенное ранее скалярное произведение можно записать следующим образом:

$$\begin{aligned} (u, v)_{\bar{\omega}} &= \sum_{i=1}^{N-1} u(i)v(i)h + 0.5h[u(0)v(0) + u(N)v(N)] = \\ &= \sum_{i=0}^{N-1} u(i)v(i)h. \end{aligned}$$

Рассмотрим два случая. Пусть сначала N четное. Из (31) получим, что собственными функциями, соответствующими λ_0 и $\lambda_{N/2}$ являются

$$\mu_k(i) = \sqrt{\frac{1}{l}} \cos \frac{2k\pi i}{N}, \quad k = 0, \frac{N}{2}. \quad (32)$$

Далее отметим, что из (30) следуют равенства

$$\begin{aligned} \lambda_{N-k} &= \frac{4}{h^2} \sin^2 \frac{(N-k)\pi}{N} = \frac{4}{h^2} \sin^2 \frac{k\pi}{N} = \lambda_k, \\ k &= 1, 2, \dots, \frac{N}{2}-1. \end{aligned}$$

Выбирая в качестве собственной функции, соответствующей собственному значению λ_k , функцию

$$\mu_k(i) = \sqrt{\frac{2}{l}} \cos \frac{2k\pi i}{N}, \quad 1 \leq k \leq \frac{N}{2}-1$$

и функцию

$$\mu_{N-k}(i) = \sqrt{\frac{2}{l}} \sin \frac{2k\pi i}{N}, \quad 1 \leq k \leq \frac{N}{2}-1,$$

соответствующую значению $\lambda_{N-k} = \lambda_k$, получим вместе с (32) полную систему собственных функций задачи (26). Итак, собствен-

ными значениями являются λ_k , определенные в (30), а собственные функции задачи (26) задаются формулами

$$\begin{aligned}\mu_k(i) &= \sqrt{\frac{1}{l}} \cos \frac{2k\pi i}{N}, & k = 0, \frac{N}{2}, \\ \mu_k(i) &= \sqrt{\frac{2}{l}} \cos \frac{2k\pi i}{N}, & 1 \leq k \leq \frac{N}{2} - 1, \\ \mu_k(i) &= \sqrt{\frac{2}{l}} \sin \frac{2(N-k)\pi i}{N}, & \frac{N}{2} + 1 \leq k \leq N - 1\end{aligned}\quad (33)$$

для случая четного N .

Отметим основные свойства собственных функций и собственных значений периодической краевой задачи (26).

- 1) Собственные функции ортонормированы.
- 2) Любая периодическая с периодом N сеточная функция $f(i)$, заданная на сетке Ω , может быть представлена в виде

$$f(i) = \frac{2}{N} \sum_{k=0}^{N/2} \rho_k \varphi_k \cos \frac{2k\pi i}{N} + \frac{2}{N} \sum_{k=N/2+1}^{N-1} \varphi_k \sin \frac{2(N-k)\pi i}{N}, \quad (34)$$

где

$$\begin{aligned}\varphi_k &= \sum_{i=0}^{N-1} \rho_k f(i) \cos \frac{2k\pi i}{N}, & 0 \leq k \leq \frac{N}{2}, \\ \varphi_k &= \sum_{i=0}^{N-1} f(i) \sin \frac{2(N-k)\pi i}{N}, & \frac{N}{2} + 1 \leq k \leq N - 1,\end{aligned}\quad (35)$$

$$\rho_k = \begin{cases} 1, & k \neq 0, N/2, \\ 1/\sqrt{2}, & k = 0, N/2. \end{cases} \quad (36)$$

Формулы (34)–(36) следуют из разложения функции $f(i)$ по собственным функциям $\mu_k(i)$:

$$f(i) = \sum_{k=0}^{N-1} f_k \mu_k(i), \quad f_k = (f, \mu_k)$$

при замене $f_k = \frac{\sqrt{2l}}{N} \varphi_k$.

- 3) Для собственных значений справедливы неравенства

$$0 = \lambda_0 \leq \lambda_k \leq \lambda_{N/2} = \frac{4}{h^2}, \quad 0 \leq k \leq N - 1.$$

Рассмотрим теперь случай, когда N нечетное. В этом случае собственные значения задачи (26) определяются формулами (30), причем $\lambda_0 = 0$ и имеет место равенство $\lambda_{N-k} = \lambda_k$, $k = 1, 2, \dots, (N-1)/2$.

Собственные функции, соответствующие собственным значениям λ_k , определяются следующими формулами:

$$\begin{aligned}\mu_0(i) &= \sqrt{\frac{1}{l}}, & k = 0, \\ \mu_k(i) &= \sqrt{\frac{2}{l}} \cos \frac{2k\pi i}{N}, & 1 \leq k \leq \frac{N-1}{2}, \\ \mu_k(i) &= \sqrt{\frac{2}{l}} \sin \frac{2(N-k)\pi i}{N}, & \frac{N+1}{2} \leq k \leq N-1.\end{aligned}\quad (37)$$

Собственные функции (37) ортонормированы, а собственные значения λ_k удовлетворяют неравенствам $0 = \lambda_0 < \lambda_k < \lambda_{\frac{N-1}{2}} = \frac{4}{h^2} \cos^2 \frac{\pi}{2N}$, $0 < k < N-1$. Кроме того, любая периодическая с периодом N (N —нечетно) сеточная функция $f(i)$, заданная на сетке Ω , представима в виде

$$f(i) = \frac{2}{N} \sum_{k=0}^{(N-1)/2} \rho_k \varphi_k \cos \frac{2k\pi i}{N} + \frac{2}{N} \sum_{k=(N+1)/2}^{N-1} \varphi_k \sin \frac{2(N-k)\pi i}{N},$$

где

$$\begin{aligned}\varphi_k &= \sum_{i=0}^{N-1} \rho_k f(i) \cos \frac{2k\pi i}{N}, & 0 \leq k \leq \frac{N-1}{2}, \\ \varphi_k &= \sum_{i=0}^{N-1} f(i) \sin \frac{2(N-k)\pi i}{N}, & \frac{N+1}{2} \leq k \leq N-1,\end{aligned}$$

а ρ_k определено выше.

ГЛАВА II

МЕТОД ПРОГОНКИ

В этой главе изучаются различные варианты прямого метода решения сеточных уравнений — метода прогонки. Рассматривается применение метода для решения как скалярных, так и векторных уравнений.

В § 1 построен и исследован метод прогонки для скалярных трехточечных уравнений. § 2 посвящен различным вариантам метода прогонки, здесь рассмотрены потоковая, циклическая и немонотонная прогонки. В § 3 рассмотрены монотонная и немонотонная прогонки для пятиточечных скалярных уравнений. В § 4 построены алгоритмы матричной прогонки для двух- и трехточечных векторных уравнений, метод ортогональной прогонки для двухточечных уравнений.

§ 1. Метод прогонки для трехточечных уравнений

1. Алгоритм метода. В главе I были изложены методы решения разностных уравнений с постоянными коэффициентами. Настоящая глава посвящена построению прямых методов решения краевых задач для трех- и пятиточечных разностных уравнений с переменными коэффициентами, а также трехточечных векторных уравнений. Здесь будут изучены различные варианты метода прогонки, который представляет собой метод исключения Гаусса, примененный к специальным системам линейных алгебраических уравнений и учитывающий ленточную структуру матрицы системы.

Рассмотрение метода прогонки начнем со случая скалярных уравнений. Пусть требуется найти решение следующей системы трехточечных уравнений:

$$\begin{aligned} c_0 y_0 - b_0 y_1 &= f_0, & i = 0, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, & 1 \leq i \leq N-1, \\ -a_N y_{N-1} + c_N y_N &= f_N, & i = N, \end{aligned} \quad (1)$$

или в векторном виде

$$\mathcal{A}Y = F, \quad (2)$$

где $Y = (y_0, y_1, \dots, y_N)$ — вектор неизвестных, $F = (f_0, f_1, \dots, f_N)$ —

вектор правых частей, а \mathcal{A} — квадратная $(N+1) \times (N+1)$ матрица

$$\mathcal{A} = \begin{vmatrix} c_0 - b_0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -a_1 & c_1 - b_1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & c_2 - b_2 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -a_{N-2} & c_{N-2} - b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & c_{N-1} - b_{N-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -a_N & c_N \end{vmatrix}$$

с действительными или комплексными коэффициентами.

Системы вида (1) возникают при трехточечной аппроксимации краевых задач для обыкновенных дифференциальных уравнений второго порядка с постоянными и переменными коэффициентами, а также при реализации разностных схем для уравнений в частных производных. В последнем случае обычно требуется решить не единичную задачу (1), а серию задач с различными правыми частями, причем число задач в серии может равняться нескольким десяткам и сотням при числе неизвестных в каждой задаче $N \approx 100$. Поэтому необходимо разработать экономичные методы решения задач вида (1), число действий для которых пропорционально числу неизвестных. Для системы (1) таким методом является *метод прогонки*.

Возможность построения экономичного метода заключена в специфике системы (1). Соответствующая (1) матрица \mathcal{A} принадлежит к классу разреженных матриц — из $(N+1)^2$ элементов ненулевыми являются не более $3N+1$ элементов. Кроме того, она имеет ленточную структуру (является трехдиагональной матрицей). Такое регулярное расположение ненулевых элементов матрицы \mathcal{A} позволяет получить очень простые расчетные формулы для вычисления решения.

Переходим к построению алгоритма решения системы (1). Напомним последовательность действий, которые осуществляются в методе исключения Гаусса. На первом шаге из всех уравнений системы (1) для $i = 1, 2, \dots, N$ исключается при помощи первого уравнения (1) неизвестное y_0 , затем из преобразованных уравнений для $i = 2, 3, \dots, N$ при помощи уравнения, соответствующего $i = 1$, исключается неизвестное y_1 и т. д. В результате получим одно уравнение относительно y_N . На этом прямой ход метода заканчивается. На обратном ходе для $i = N-1, N-2, \dots, 0$ находится y_i через уже найденные $y_{i+1}, y_{i+2}, \dots, y_N$ и преобразованные правые части.

Следуя идею метода Гаусса, проведем исключение неизвестных в (1). Введем обозначения, полагая $\alpha_1 = b_0/c_0$, $\beta_1 = f_0/c_0$, и запишем (1) в следующем виде:

$$\begin{aligned} y_0 - \alpha_1 y_1 &= \beta_1, & i = 0, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, & 1 \leq i \leq N-1, \\ -a_N y_{N-1} + c_N y_N &= f_N, & i = N. \end{aligned} \quad (1')$$

Возьмем первые два уравнения системы (1')

$$y_0 - a_1 y_1 = \beta_1, \quad -a_1 y_0 + c_1 y_1 - b_1 y_2 = f_1.$$

Умножим первое уравнение на a_1 и сложим со вторым уравнением. Будем иметь $(c_1 - a_1 \alpha_1) y_1 - b_1 y_2 = f_1 + a_1 \beta_1$ или после деления на $c_1 - a_1 \alpha_1$

$$y_1 - \alpha_2 y_2 = \beta_2, \quad \alpha_2 = \frac{b_1}{c_1 - a_1 \alpha_1}, \quad \beta_2 = \frac{f_1 + a_1 \beta_1}{c_1 - a_1 \alpha_1}.$$

Все остальные уравнения системы (1') y_0 не содержат, поэтому на этом первый шаг процесса исключения заканчивается. В результате получим новую «укороченную» систему

$$\begin{aligned} y_1 - \alpha_2 y_2 &= \beta_2, & i &= 1, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, & 2 \leq i \leq N-1, \\ -a_N y_{N-1} + c_N y_N &= f_N, & i &= N, \end{aligned} \quad (3)$$

которая не содержит неизвестное y_0 и имеет аналогичную (1') структуру. Если эта система будет решена, то неизвестное y_0 найдется по формуле $y_0 = \alpha_1 y_1 + \beta_1$. К системе (3) можно снова применить описанный способ исключения неизвестных. На втором шаге будет исключено неизвестное y_1 , на третьем y_2 и т. д. В результате l -го шага получим систему для неизвестных y_l, y_{l+1}, \dots, y_N

$$\begin{aligned} y_l - \alpha_{l+1} y_{l+1} &= \beta_{l+1}, & i &= l, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, & l+1 \leq i \leq N-1, \\ -a_N y_{N-1} + c_N y_N &= f_N, & i &= N, \end{aligned} \quad (4)$$

и формулы для нахождения y_i с номерами $i \leq l-1$

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = l-1, l-2, \dots, 0. \quad (5)$$

Коэффициенты α_i и β_i , очевидно, находятся по формулам

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, \alpha_1 = \frac{b_0}{c_0}, \quad \beta_1 = \frac{f_0}{c_0}.$$

Полагая в (4) $l = N-1$, получим систему для y_N и y_{N-1}

$$y_{N-1} - \alpha_N y_N = \beta_N, \quad -a_N y_{N-1} + c_N y_N = f_N,$$

из которой найдем $y_N = \beta_{N+1}$, $y_{N-1} = \alpha_N y_N + \beta_N$.

Объединяя эти равенства с (5) ($l = N-1$), получим окончательные формулы для нахождения неизвестных

$$\begin{aligned} y_i &= \alpha_{i+1} y_{i+1} + \beta_{i+1}, & i &= N-1, N-2, \dots, 0, \\ y_N &= \beta_{N+1}, \end{aligned} \quad (6)$$

где α_i и β_i находятся по рекуррентным формулам

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N-1, \quad \alpha_1 = \frac{b_0}{c_0}, \quad (7)$$

$$\beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N, \quad \beta_1 = \frac{f_0}{c_0}. \quad (8)$$

Итак, формулы (6)–(8) описывают метод Гаусса, который в применении к системе (1) получил специальное название — *метод прогонки*. Коэффициенты α_i и β_i называют *прогоночными коэффициентами*, формулы (7), (8) описывают *прямой ход прогонки*, а (6) — *обратный ход*. Так как значения y_i находятся здесь последовательно при переходе от $i+1$ к i , то формулы (6)–(8) называют иногда формулами *правой прогонки*.

Элементарный подсчет арифметических действий в (6)–(8) показывает, что реализация метода прогонки по этим формулам требует выполнения $3N$ умножений, $2N+1$ делений и $3N$ сложений и вычитаний. Если не делать различия между арифметическими операциями, то общее их число для метода прогонки есть $Q = 8N + 1$. Из этого числа $3N - 2$ операции затрачиваются на вычисление α_i и $5N + 3$ операций на вычисление β_i и y_i .

Заметим, что коэффициенты α_i не зависят от правой части системы (1), а определяются только коэффициентами разностных уравнений a_i , b_i , c_i . Поэтому, если требуется решить серию задач (1) с различными правыми частями, но с одной и той же матрицей A , то прогоночные коэффициенты α_i вычисляются только при решении первой задачи из серии. Для каждой последующей задачи определяются только коэффициенты β_i и решение y_i , причем используются найденные ранее α_i . Таким образом, на решение только первой из серии задач тратится число арифметических действий $Q = 8N + 1$, на решение каждой следующей задачи будет затрачиваться уже только $5N + 3$ операции.

В заключение укажем порядок счета по формулам метода прогонки. Начиная с α_1 и β_1 , по формулам (7) и (8) определяются и запоминаются прогоночные коэффициенты α_i и β_i . Затем по формулам (6) находится решение y_i .

2. Метод встречных прогонок. Выше были получены формулы правой прогонки для решения системы (1). Аналогично выводятся формулы *левой прогонки*:

$$\xi_i = \frac{a_i}{c_i - b_i \xi_{i+1}}, \quad i = N-1, N-2, \dots, 1, \quad \xi_N = \frac{a_N}{c_N}, \quad (9)$$

$$\eta_i = \frac{f_i + b_i \eta_{i+1}}{c_i - b_i \xi_{i+1}}, \quad i = N-1, N-2, \dots, 0, \quad \eta_N = \frac{f_N}{c_N}, \quad (10)$$

$$y_{i+1} = \xi_{i+1} y_i + \eta_{i+1}, \quad i = 0, 1, \dots, N-1, \quad y_0 = \eta_0. \quad (11)$$

Здесь значения y_i находятся последовательно при возрастании индекса i (слева направо).

Иногда оказывается удобным комбинировать правую и левую прогонки, получая так называемый *метод встречных прогонок*. Этот метод целесообразно применять, если надо найти только одно неизвестное, например y_m ($0 \leq m \leq N$) или группу идущих подряд неизвестных. Получим формулы метода встречных прогонок. Пусть $1 \leq m \leq N$ и по формулам (7)–(10) найдены $\alpha_1, \alpha_2, \dots, \alpha_m, \beta_1, \beta_2, \dots, \beta_m$ и $\xi_N, \xi_{N-1}, \dots, \xi_m, \eta_N, \eta_{N-1}, \dots, \eta_m$.

Выпишем формулы (6), (11) для обратного хода правой и левой прогонок для $i = m - 1$. будем иметь систему

$$y_{m-1} = \alpha_m y_m + \beta_m, \quad y_m = \xi_m y_{m-1} + \eta_m,$$

из которой найдем y_m :

$$y_m = \frac{\eta_m + \xi_m \beta_m}{1 - \xi_m \alpha_m}.$$

Используя найденное y_m , по формулам (6) для $i = m - 1, m - 2, \dots, 0$ найдем последовательно $y_{m-1}, y_{m-2}, \dots, y_0$, а по формулам (11) для $i = m, m + 1, \dots, N$ вычислим остальные $y_{m+1}, y_{m+2}, \dots, y_N$.

Итак, формулы метода встречных прогонок имеют вид:

$$\begin{aligned} \alpha_{i+1} &= \frac{b_i}{c_i - a_i \alpha_i}, & i = 1, 2, \dots, m-1, & \alpha_1 = \frac{b_0}{c_0}, \\ \beta_{i+1} &= \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, & i = 1, 2, \dots, m-1, & \beta_1 = \frac{f_0}{c_0}, \\ \xi_i &= \frac{a_i}{c_i - b_i \xi_{i+1}}, & i = N-1, N-2, \dots, m, & \xi_N = \frac{a_N}{c_N}, \\ \eta_i &= \frac{f_i + b_i \eta_{i+1}}{c_i - b_i \xi_{i+1}}, & i = N-1, N-2, \dots, m, & \eta_N = \frac{f_N}{c_N} \end{aligned} \quad (12)$$

для вычисления прогоночных коэффициентов и

$$\begin{aligned} y_i &= \alpha_{i+1} y_{i+1} + \beta_{i+1}, & i = m-1, m-2, \dots, 0, \\ y_{i+1} &= \xi_{i+1} y_i + \eta_{i+1}, & i = m, m+1, \dots, N-1, \\ y_m &= \frac{\eta_m + \xi_m \beta_m}{1 - \xi_m \alpha_m} \end{aligned} \quad (13)$$

для определения решения.

Очевидно, что число действий, затрачиваемое на нахождение решения задачи (1) по методу встречных прогонок, такое же, как и для левой или правой прогонок, т. е. $Q \approx 8N$. Заметим, что для частного случая постоянных коэффициентов $a_i = b_i = 1, c_i = c$ для $i = 1, 2, \dots, N-1$ и $b_0 = a_N = 0$ число действий может быть уменьшено, если N — нечетное число, следующим образом. Пусть $N = 2M - 1$. Положим в формулах (12), (13) метода встречных прогонок $m = M$. Тогда $\xi_{N-i+1} = \alpha_i, i = 1, 2, \dots, M$. Следовательно, прогоночный коэффициент ξ_i находить не нужно, и формулы метода встречных прогонок будут иметь вид

$$\begin{aligned} \alpha_{i+1} &= \frac{1}{c - \alpha_i}, & i = 1, 2, \dots, M-1, & \alpha_1 = 0, \\ \beta_{i+1} &= (f_i + \beta_i) \alpha_{i+1}, & i = 1, 2, \dots, M-1, & \beta_1 = \frac{f_0}{c_0}, \\ \eta_i &= (f_i + \eta_{i+1}) \alpha_{N-i+1}, & i = N-1, N-2, \dots, M, & \eta_N = \frac{f_N}{c_N}, \\ y_i &= \alpha_{i+1} y_{i+1} + \beta_{i+1}, & i = M-1, M-2, \dots, 0, \\ y_{i+1} &= \alpha_{N-i} y_i + \eta_{i+1}, & i = M, M+1, \dots, N-1, \end{aligned}$$

где $y_M = (\eta_M + \alpha_M \beta_M) / (1 - \alpha_M^2)$.

3. Обоснование метода прогонки. Выше были получены формулы метода прогонки без каких-либо предположений относительно коэффициентов системы (1). Остановимся здесь на вопросе о том, каким требованиям должны удовлетворять эти коэффициенты, чтобы метод мог быть применен и позволял получить решение задачи с достаточной точностью.

Поясним ситуацию. Так как расчетные формулы (6)–(8) метода прогонки содержат операции деления, то нужно гарантировать не обращение знаменателя $c_i - a_i \alpha_i$ в (7), (8) в нуль. Будем говорить, что алгоритм метода правой прогонки *корректен*, если $c_i - a_i \alpha_i \neq 0$ при $i = 1, 2, \dots, N$. Далее решение y_i находится по рекуррентной формуле (6). Эта формула может порождать накопление ошибок округления результатов арифметических операций. Действительно, пусть прогоночные коэффициенты α_i и β_i найдены точно, а при вычислении y_N допущена ошибка ε_N , т. е. найдено $\tilde{y}_N = y_N + \varepsilon_N$. Так как решение \tilde{y}_i находится по формулам (6) $\tilde{y}_i = \alpha_{i+1} \tilde{y}_{i+1} + \beta_{i+1}$, $i = N-1, N-2, \dots, 0$, то погрешность $\varepsilon_i = \tilde{y}_i - y_i$ будет, очевидно, удовлетворять однородному уравнению $\varepsilon_i = \alpha_{i+1} \varepsilon_{i+1}$, $i = N-1, N-2, \dots, 0$, с заданным ε_N . Отсюда следует, что если все α_i по модулю больше единицы, то может произойти сильное увеличение погрешности ε_0 , и, если N достаточно велико, то полученное реальное решение y_i будет значительно отличаться от искомого решения \tilde{y}_i .

Не имея возможности более детально останавливаться на обсуждении вопросов вычислительной устойчивости метода и механизма образования неустойчивости, сформулируем требование, обычно предъявляемое к алгоритму метода прогонки. Будем требовать, чтобы прогоночные коэффициенты α_i не превосходили по модулю единицы. Это достаточное условие гарантирует невозвратение погрешности ε_i в рассмотренной выше модельной ситуации. Если выполнено условие $|\alpha_i| \leq 1$, то будем говорить, что алгоритм правой прогонки *устойчив*.

Выясним условия корректности и устойчивости алгоритма (6)–(8). Следующая лемма содержит достаточные условия корректности и устойчивости алгоритма правой прогонки.

Лемма 1. Пусть коэффициенты системы (1) действительны и удовлетворяют условиям $|b_0| \geq 0$, $|a_N| \geq 0$, $|c_0| > 0$, $|c_N| > 0$, $|a_i| > 0$, $|b_i| > 0$, $i = 1, 2, \dots, N-1$,

$$|c_i| \geq |a_i| + |b_i|, \quad i = 1, 2, \dots, N-1, \quad (14)$$

$$|c_0| \geq |b_0|, \quad |c_N| \geq |a_N|, \quad (15)$$

причем хотя бы в одном из неравенств (14) или (15) выполняется строгое неравенство, т. е. матрица \mathcal{A} имеет диагональное преобладание. Тогда для алгоритма (6)–(8) метода прогонки имеют место неравенства $c_i - a_i \alpha_i \neq 0$, $|\alpha_i| \leq 1$, $i = 1, 2, \dots, N$, гарантирующие корректность и устойчивость метода.

Доказательство леммы проведем по индукции. Из условий леммы и (7) следует, что

$$0 \leq |\alpha_i| = \frac{|b_0|}{|c_0|} \leq 1. \quad (16)$$

Покажем, что из неравенства $|\alpha_i| \leq 1$ ($i \leq N-1$) и условий леммы следуют неравенства

$$c_i - a_i \alpha_i \neq 0, \quad |\alpha_{i+1}| \leq 1, \quad i \leq N-1. \quad (17)$$

Тогда, учитывая (16), получим, что имеют место неравенства $|\alpha_i| \leq 1$ для $i=1, 2, \dots, N$ и $c_i - a_i \alpha_i \neq 0$ для $i=1, 2, \dots, N-1$. Для завершения доказательства леммы останется доказать неравенство $c_N - a_N \alpha_N \neq 0$. Итак, сначала установим (17). Пусть $|\alpha_i| \leq 1$, $i \leq N-1$. Тогда из (14)

$$|c_i - a_i \alpha_i| \geq |c_i| - |a_i| |\alpha_i| \geq |b_i| + |a_i| (1 - |\alpha_i|) \geq |b_i| > 0, \quad (18)$$

и, следовательно, $c_i - a_i \alpha_i \neq 0$. Далее из (7) и (18) получим

$$|\alpha_{i+1}| = \frac{|b_i|}{|c_i - a_i \alpha_i|} \leq \frac{|b_i|}{|b_i|} = 1,$$

что и требовалось доказать.

Осталось показать, что $c_N - a_N \alpha_N \neq 0$. Для этого используем предположение, что хотя бы в одном из неравенств (14) или (15) имеет место строгое неравенство. Здесь возможны несколько случаев. Если $|c_N| > |a_N|$, то в силу доказанного $|\alpha_N| \leq 1$ и, следовательно, $c_N - a_N \alpha_N \neq 0$. Если строгое неравенство достигается в (14) для некоторого i_0 , $1 \leq i_0 \leq N-1$, то из (18) получим, что $|c_{i_0} - a_{i_0} \alpha_{i_0}| > |b_{i_0}|$, и, следовательно, имеет место неравенство $|\alpha_{i_0+1}| < 1$. По индукции далее легко устанавливается неравенство $|\alpha_i| < 1$ для $i \geq i_0 + 1$. Следовательно, в этом случае будем иметь $|\alpha_N| < 1$, и поэтому $c_N - a_N \alpha_N \neq 0$. Если $|c_0| > |b_0|$, то неравенство $|\alpha_i| < 1$ имеет место, начиная с $i=1$. Поэтому снова получим $|\alpha_N| < 1$ и $c_N - a_N \alpha_N \neq 0$. Лемма доказана.

Замечание 1. Условия корректности и устойчивости алгоритма (6)–(8), сформулированные в лемме 1, являются лишь достаточными условиями. Эти условия можно ослабить, разрешив некоторым из коэффициентов a_i или b_i обращаться в нуль. Так, например, если при некотором $1 \leq m \leq N-1$ окажется, что $a_m = 0$, то система (1) распадается на две системы:

$$\begin{aligned} c_m y_m - b_m y_{m+1} &= f_m, & i = m, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, & m+1 \leq i \leq N-1, \\ -a_N y_{N-1} + c_N y_N &= f_N, & i = N \end{aligned}$$

для неизвестных y_m, y_{m+1}, \dots, y_N и

$$\begin{aligned} c_0 y_0 - b_0 y_1 &= f_0, & i = 0, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, & 1 \leq i \leq m-2, \\ -a_{m-1} y_{m-2} + c_{m-1} y_{m-1} &= f_{m-1} + b_{m-1} y_m \end{aligned}$$

для неизвестных y_0, y_1, \dots, y_{m-1} . К каждой из этих систем можно применить алгоритм (6)–(8), если для них выполнены условия леммы 1. Но в этом случае формулы (6)–(8) можно использовать для нахождения решения сразу всей распадающейся системы (1), причем алгоритм будет корректен и устойчив.

Замечание 2. Условия леммы 1 обеспечивают корректность и устойчивость алгоритмов левой и встречных прогонок. Эти условия сохраняются и для случая системы (1) с комплексными коэффициентами a_i, b_i и c_i .

Покажем теперь, что при выполнении условий леммы 1 система (1) имеет единственное решение при любой правой части. Действительно, учитывая соотношения (7), непосредственным перемножением матриц можно показать, что матрица \mathcal{A} системы (1) представляется в виде произведения двух треугольных матриц L и U

$$\mathcal{A} = LU,$$

где

$$L = \begin{vmatrix} c_0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -a_1 & \Delta_1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & \Delta_2 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_3 & \Delta_3 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \Delta_{N-3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & -a_{N-2} & \Delta_{N-2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & \Delta_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -a_N & \Delta_N \end{vmatrix},$$

$$U = \begin{vmatrix} 1 & -\alpha_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -\alpha_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & -\alpha_3 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -\alpha_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -\alpha_N \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{vmatrix}$$

и $\Delta_i = c_i - a_i \alpha_i$, $i = 1, 2, \dots, N$. Так как

$$\det \mathcal{A} = \det L \cdot \det U = c_0 \prod_{i=1}^N \Delta_i,$$

а в силу леммы 1 $c_0 \neq 0$ и $\Delta_i \neq 0$ для $i = 1, 2, \dots, N$, то $\det \mathcal{A} \neq 0$. Поэтому система (1) в случае выполнения условий леммы 1 имеет единственное решение, и это решение может быть найдено по методу прогонки (6)–(8).

4. Примеры применения метода прогонки. Рассмотрим некоторые примеры применения изложенного выше метода прогонки.

Пример 1. *Первая краевая задача.* Пусть требуется решить следующую задачу:

$$(k(x)u'(x))' - q(x)u(x) = -f(x), \quad 0 < x < l, \\ u(0) = \mu_1, \quad u(l) = \mu_2, \quad k(x) \geq c_1 > 0, \quad q(x) \geq 0. \quad (19)$$

На отрезке $0 \leq x \leq l$ построим произвольную неравномерную сетку $\bar{\omega} = \{x_i \in [0, l], i = 0, 1, \dots, N, x_0 = 0, x_N = l\}$ с шагами $h_i = x_i - x_{i-1}$, $i = 1, 2, \dots, N$ и заменим (19) следующей разностной задачей:

$$(ay_x)_{\hat{x}, i} - d_i y_i = -\varphi_i, \quad 1 \leq i \leq N-1, \\ y_0 = \mu_1, \quad y_N = \mu_2, \quad (20)$$

где $d_i = q(x_i)$, $\varphi_i = f(x_i)$, а для a_i используем простейшую аппроксимацию коэффициента $k(x)$: $a_i = k(x_i - 0,5h_i)$. Расписывая разностную производную, входящую в (20), по точкам

$$(ay_x)_{\hat{x}, i} = \frac{1}{\hat{h}_i} \left(a_{i+1} \frac{y_{i+1} - y_i}{h_{i+1}} - a_i \frac{y_i - y_{i-1}}{h_i} \right),$$

где $\hat{h}_i = 0,5(h_i + h_{i+1})$ — средний шаг в точке x_i , получим, что задача (20) записывается в виде системы

$$\begin{aligned} C_0 y_0 - B_0 y_1 &= f_0, & i &= 0, \\ -A_i y_{i-1} + C_i y_i - B_i y_{i+1} &= f_i, & 1 \leq i \leq N-1, \\ -A_N y_{N-1} + C_N y_N &= f_N, & i &= N. \end{aligned} \quad (1'')$$

Здесь

$$B_0 = A_N = 0, \quad C_0 = C_N = 1, \quad f_0 = \mu_1, \quad f_N = \mu_2, \quad f_i = \varphi_i, \\ A_i = \frac{a_i}{\hat{h}_i h_i}, \quad B_i = \frac{a_{i+1}}{\hat{h}_i h_{i+1}}, \quad C_i = A_i + B_i + d_i, \quad 1 \leq i \leq N-1. \quad (21)$$

В силу построения разностной схемы (20) для коэффициентов a_i и d_i выполнены следующие условия: $a_i \geq c_1 > 0$, $d_i \geq 0$. Поэтому из (21) следует, что для (1'') условия леммы 1 выполнены, и эта задача может быть решена методом прогонки.

Пример 2. *Третья краевая задача.* Рассмотрим теперь случай краевых условий третьего рода:

$$(k(x)u'(x))' - q(x)u(x) = -f(x), \quad 0 < x < l, \\ k(0)u'(0) = \kappa_1 u(0) - \mu_1, \\ -k(l)u'(l) = \kappa_2 u(l) - \mu_2. \quad (22)$$

Будем считать выполненными следующие условия: $k(x) \geq c_1 > 0$, $q(x) \geq 0$, $\kappa_1 \geq 0$, $\kappa_2 \geq 0$, причем, если $q(x) \equiv 0$, то $\kappa_1^2 + \kappa_2^2 \neq 0$.

На введенной выше неравномерной сетке задача (22) аппроксимируется следующей разностной схемой:

$$\begin{aligned} (ay_{\bar{x}})_{\hat{x}, i} - d_i y_i &= -\varphi_i, \quad 1 \leq i \leq N-1, \\ \frac{2}{h_1} a_1 y_{x, 0} &= \left(d_0 + \frac{2}{h_1} \kappa_1 \right) y_0 - \left(\varphi_0 + \frac{2}{h_1} \mu_1 \right), \quad i = 0, \\ -\frac{2}{h_N} a_N y_{\bar{x}, N} &= \left(d_N + \frac{2}{h_N} \kappa_2 \right) y_N - \left(\varphi_N + \frac{2}{h_N} \mu_2 \right), \quad i = N, \end{aligned} \quad (23)$$

где коэффициенты a_i , d_i и φ_i выбраны указанным в примере 1 способом. Расписывая вторую разностную производную $(ay_{\bar{x}})_{\hat{x}}$ по точкам, а также первые производные

$$y_{x, i} = \frac{y_{i+1} - y_i}{h_{i+1}}, \quad y_{\bar{x}, i} = \frac{y_i - y_{i-1}}{h_i},$$

сведем (23) к виду (1''), где

$$\begin{aligned} B_0 &= \frac{2a_1}{h_1^2}, \quad A_N = \frac{2a_N}{h_N^2}, \quad C_0 = B_0 + d_0 + \frac{2}{h_1} \kappa_1, \\ C_N &= A_N + d_N + \frac{2}{h_N} \kappa_2, \quad f_0 = \varphi_0 + \frac{2}{h_1} \mu_1, \quad f_N = \varphi_N + \frac{2}{h_N} \mu_2, \\ A_i &= \frac{a_i}{h_i h_{i+1}}, \quad B_i = \frac{a_{i+1}}{h_i h_{i+1}}, \quad C_i = A_i + B_i + d_i, \quad f_i = \varphi_i, \quad 1 \leq i \leq N-1. \end{aligned}$$

Легко проверить, что и в этом случае условия леммы 1 также выполнены.

Пример 3. Разностные схемы для уравнения теплопроводности. Рассмотрим первую краевую задачу для уравнения теплопроводности:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0, \\ u(0, t) &= \mu_1(t), \quad u(l, t) = \mu_2(t), \\ u(x, 0) &= u_0(x). \end{aligned} \quad (24)$$

На плоскости (x, t) введем сетку $\bar{\omega} = \{(x_i, t_n), x_i = ih, i = 0, 1, \dots, N, h = l/N, t_n = n\tau, n = 0, 1, \dots\}$ с шагом h по пространству и τ по времени. Аппроксимируем (24) разностной схемой

$$\begin{aligned} y_{t, i} &= \sigma y_{xx, i}^{n+1} + (1 - \sigma) y_{xx, i}^n, \quad 1 \leq i \leq N-1, \\ y_0^n &= \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad y_t^0 = u_0(x_i), \quad n = 0, 1, \dots, \end{aligned} \quad (25)$$

где σ — вещественный параметр, $y_t^n = y(x_i, t_n)$,

$$y_{xx, i} = \frac{1}{h^2} (y_{i+1} - 2y_i + y_{i-1}), \quad y_{t, i} = \frac{1}{\tau} (y_i^{n+1} - y_i^n). \quad (26)$$

Известно (см., например, [9]), что схема (25) имеет аппроксимацию $O(\tau + h^2)$ при любом σ , $O(\tau^2 + h^2)$ при $\sigma = 0,5$ и аппроксимацию $O(\tau^2 + h^4)$ при $\sigma = 1/2 - h^2/(12\tau)$. Условие устойчивости

схемы (25) по начальным данным имеет вид

$$\sigma \geq 1/2 - h^2/(4\tau). \quad (27)$$

Обратимся теперь к методу решения уравнений (25) относительно y_i^{n+1} . Считая y_i^n уже известным, запишем (25) в следующем виде:

$$\begin{aligned} \frac{1}{\sigma\tau} y_i^{n+1} - y_{xx, i}^n &= \varphi_i^n, \quad 1 \leq i \leq N-1, \\ y_0^{n+1} &= \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}), \end{aligned}$$

где $\varphi_i^n = \frac{1}{\sigma\tau} y_i^n + \left(\frac{1}{\sigma} - 1\right) y_{xx, i}^n$, если $\sigma \neq 0$. Используя (26), сведем эту схему к виду (1''), где $B_0 = A_N = 0$, $C_0 = C_N = 1$, $f_0 = \mu_1(t_{n+1})$, $f_N = \mu_2(t_{n+1})$, $A_i = B_i = \frac{1}{h^2}$, $C_i = A_i + B_i + \frac{1}{\sigma\tau}$, $f_i = \varphi_i^n$, $1 \leq i \leq N-1$. Найдем условия, при которых построенную систему (1'') можно будет решать методом прогонки. Из леммы 1 следует, что должно быть выполнено условие $|2/h^2 + 1/(\sigma\tau)| \geq 2/h^2$. Решая это неравенство, найдем достаточное условие применимости прогонки $\sigma \geq -h^2/(4\tau)$. Сравнивая это неравенство с (27), получим, что если для схемы (25) выполнено условие устойчивости (27), то для нахождения решения на верхнем слое можно использовать метод прогонки.

Пример 4. Нестационарное уравнение Шредингера. Рассмотрим нестационарное уравнение Шредингера $i \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$, $0 < x < l$, $t > 0$, $u(0, t) = u(l, t) = 0$, $u(0, x) = u_0(x)$, $i = \sqrt{-1}$.

Для этого уравнения, так же как и для уравнения теплопроводности (24), можно построить двухслойную схему с весами

$$\begin{aligned} iy_{t, k} &= \sigma y_{xx, k}^{n+1} + (1-\sigma) y_{xx, k}^n, \quad 1 \leq k \leq N-1, \\ y_0^n &= y_N^n = 0, \quad y_k^0 = u_0(x_k), \end{aligned} \quad (28)$$

где параметр $\sigma = \sigma_0 + i\sigma_1$ может принимать значения в комплексной плоскости. Схема (28) имеет погрешность аппроксимации $O(\tau + h^2)$ при любом σ , при $\sigma = 0,5$ она равна $O(\tau^2 + h^2)$ и при $\sigma = 1/2 - h^2 i/(12\tau)$ погрешность аппроксимации равна $O(\tau^2 + h^4)$. Условие устойчивости по начальным данным имеет вид

$$\sigma_0 = \operatorname{Re} \sigma \geq 0,5. \quad (29)$$

Схема (28) обычным образом сводится к системе (1''), и условия леммы 1 принимают следующий вид: $|2/h^2 + i/(\sigma\tau)| \geq 2/h^2$. Решая это неравенство, получим, что метод прогонки нахождения решения схемы (28) на верхнем слое при выполнении условия $\sigma_1 = \operatorname{Im} \sigma \geq -h^2/(4\tau)$ будет корректен.

Таким образом, для рассматриваемого примера условие применимости метода прогонки не совпадает с условием устойчивости самой разностной схемы по начальным данным.

§ 2. Варианты метода прогонки

1. Потоковый вариант метода прогонки. Рассмотрим вариант метода прогонки, применяемый при решении разностных задач с сильно меняющимися коэффициентами. Примерами таких задач являются задачи гидродинамики с теплопроводностью и магнитной гидродинамики, где коэффициенты теплопроводности, электропроводности зависят от термодинамических параметров среды. В случае тепловых задач могут иметь место адиабатические участки, где теплопроводность отсутствует, а также изотермические участки, где теплопроводность бесконечно велика. В магнитных задачах — соответственно идеально проводящие и неэлектропроводные участки.

Часто в таких задачах, помимо самого решения, требуется найти еще и поток тепла (тепловая задача). При решении разностных уравнений второго порядка, к которым сводятся разностные схемы для этих задач, по формулам обычной прогонки часто происходит значительная потеря точности. Последующее использование численного дифференцирования для вычисления потока приводит к неудовлетворительному результату. Избавиться от этого недостатка удается путем перехода к так называемому *потоковому варианту метода прогонки*. Формулы для этого варианта прогонки можно получить в результате преобразования формул обычной прогонки.

Итак, рассмотрим разностную краевую задачу

$$\begin{aligned} -a_i y_{i-1} + c_i y_i - a_{i+1} y_{i+1} &= f_i, \quad 1 \leq i \leq N-1, \\ y_0 - \kappa_1 y_1 &= \mu_1, \quad y_N - \kappa_2 y_{N-1} = \mu_2, \end{aligned} \quad (1)$$

где

$$c_i = a_i + a_{i+1} + d_i, \quad 0 < a_i < \infty, \quad (2)$$

$$d_i > 0, \quad i = 1, 2, \dots, N-1, \quad |\kappa_1| \leq 1, \quad |\kappa_2| \leq 1. \quad (3)$$

Формулы правой прогонки (см. (6)–(8) § 1) для задачи (1) с учетом (2) принимают вид

$$y_i = \bar{\alpha}_{i+1} y_{i+1} + \bar{\beta}_{i+1}, \quad i = N-1, N-2, \dots, 0, \quad y_N = \frac{\mu_2 + \kappa_2 \bar{\beta}_N}{1 - \kappa_2 \bar{\alpha}_N}, \quad (4)$$

$$\bar{\alpha}_{i+1} = \frac{a_{i+1}}{a_{i+1} + d_i + a_i (1 - \bar{\alpha}_i)}, \quad i = 1, 2, \dots, N-1, \quad \bar{\alpha}_1 = \kappa_1, \quad (5)$$

$$\bar{\beta}_{i+1} = (f_i + a_i \bar{\beta}_i) \frac{\bar{\alpha}_{i+1}}{a_{i+1}}, \quad i = 1, 2, \dots, N-1, \quad \bar{\beta}_1 = \mu_1. \quad (6)$$

Введем новую неизвестную сеточную функцию (поток) по формуле

$$w_i = -a_i (y_i - y_{i-1}), \quad i = 1, 2, \dots, N, \quad (7)$$

и перепишем (1) в виде

$$\begin{aligned} w_{i+1} - w_i + d_i y_i &= f_i, \quad 1 \leq i \leq N-1, \\ y_0 - \kappa_1 y_1 &= \mu_1, \quad i=0, \\ -\kappa_2 w_N + a_N(1-\kappa_2) y_N &= a_N \mu_2, \quad i=N. \end{aligned} \tag{8}$$

Из (7) найдем

$$y_i = y_{i+1} + \frac{1}{a_{i+1}} w_{i+1}, \quad i=0, 1, \dots, N-1,$$

и подставим это выражение в (4). В результате найдем соотношение, связывающее y_{i+1} и w_{i+1} :

$$w_{i+1} + a_{i+1}(1-\bar{\alpha}_{i+1}) y_{i+1} = a_{i+1} \bar{\beta}_{i+1}, \quad i=0, 1, \dots, N-1.$$

Вводя обозначения

$$\alpha_i = a_i(1-\bar{\alpha}_i), \quad \beta_i = \alpha_i \bar{\beta}_i, \quad i=1, 2, \dots, N,$$

перепишем это соотношение в виде

$$w_i + \alpha_i y_i = \beta_i, \quad i=1, 2, \dots, N. \tag{9}$$

Заметим, что уравнения (8), (9) образуют алгебраическую систему, содержащую $2N+1$ уравнение относительно $2N+1$ неизвестных y_0, y_1, \dots, y_N и w_1, w_2, \dots, w_N . Структура этой системы такова, что она распадается на две независимые системы для неизвестных y_0, y_1, \dots, y_N и w_1, w_2, \dots, w_N . Построим эти системы.

Выразим из (9) y_i : $y_i = (\beta_i - w_i)/\alpha_i$, $i=1, 2, \dots, N$ и подставим в уравнения системы (8) для $i=1, 2, \dots, N$. В результате получим уравнения

$$\begin{aligned} w_i &= \frac{\alpha_i}{\alpha_i + d_i} w_{i+1} + \frac{d_i \beta_i - \alpha_i f_i}{\alpha_i + d_i}, \quad i=N-1, N-2, \dots, 1, \\ w_N &= \frac{a_N [(1-\kappa_2) \beta_N - \alpha_N \mu_2]}{(1-\kappa_2) a_N + \alpha_N \kappa_2}, \end{aligned} \tag{10}$$

решая которые последовательно найдем все w_i .

Получим теперь уравнения для y_i . Для этого выразим w_i из (9): $w_i = -\alpha_i y_i + \beta_i$, $i=1, 2, \dots, N$ и подставим в (8) для $i=1, 2, \dots, N$. В результате получим уравнения

$$\begin{aligned} y_i &= \frac{\alpha_{i+1}}{\alpha_i + d_i} y_{i+1} + \frac{f_i - \beta_{i+1} + \beta_i}{\alpha_i + d_i}, \quad i=N-1, N-2, \dots, 1, \\ y_0 &= \kappa_1 y_1 + \mu_1, \\ y_N &= \frac{\kappa_2 \beta_N + a_N \mu_2}{(1-\kappa_2) a_N + \alpha_N \kappa_2} \end{aligned} \tag{11}$$

для последовательного вычисления y_i .

Напишем рекуррентные формулы для определения α_i и β_i . Используя (5) и (6), найдем

$$\alpha_{i+1} = a_{i+1}(1 - \bar{\alpha}_{i+1}) = \frac{a_{i+1} [a_i(1 - \bar{\alpha}_i) + d_i]}{a_{i+1} + d_i + a_i(1 - \bar{\alpha}_i)} = \frac{a_{i+1}(\alpha_i + d_i)}{a_{i+1} + \alpha_i + d_i}, \quad (12)$$

$$i = 1, 2, \dots, N-1, \alpha_1 = a_1(1 - \kappa_1),$$

$$\beta_{i+1} = a_{i+1}\bar{\beta}_{i+1} = \frac{a_{i+1}(f_i + \beta_i)}{a_{i+1} + \alpha_i + d_i}, \quad i = 1, 2, \dots, N-1, \beta_1 = a_1\mu_1. \quad (13)$$

Из условий (2), (3) и формул (12) следует, что $\alpha_i \geq 0$. Тогда коэффициент $\alpha_i/(\alpha_i + d_i)$ в формуле (10) не превосходит единицу, что обеспечивает устойчивость алгоритма при вычислении w_i . Далее, так как из условий $\alpha_i \geq 0$ и $d_i > 0$ следует, что $a_{i+1} < a_{i+1} + \alpha_i + d_i$, то в силу (12) справедливо неравенство $\alpha_{i+1} < \alpha_i + d_i$. Поэтому коэффициент $\alpha_{i+1}/(\alpha_i + d_i)$ в формуле (11) всегда меньше единицы, что обеспечивает устойчивость при вычислении y_i . Отметим, что знаменатель в выражениях для w_N и y_N всегда больше нуля.

Итак, алгоритм метода потоковой прогонки описывается формулами (10)–(13). Отметим, что указанными рекуррентными формулами для α_i и β_i , а также выражениями для y_N и w_N целесообразно пользоваться, если $a_{i+1} < 1$. Если $a_{i+1} \geq 1$, то рекомендуется использовать следующие формулы, получаемые из (10)–(13) делением числителя и знаменателя дробей на a_{i+1} :

$$\alpha_{i+1} = \frac{\alpha_i + d_i}{1 + (\alpha_i + d_i)/a_{i+1}}, \quad \beta_{i+1} = \frac{f_i + \beta_i}{1 + (\alpha_i + d_i)/a_{i+1}},$$

$$y_N = \frac{\kappa_2 \beta_N / a_N + \mu_2}{1 - \kappa_2 + \kappa_2 \alpha_N / a_N}, \quad w_N = \frac{(1 - \kappa_2) \beta_N - \alpha_N \mu_2}{1 - \kappa_2 + \kappa_2 \alpha_N / a_N}.$$

Подсчитаем число арифметических действий, которое необходимо затратить для реализации (10)–(13). При разумной организации вычислений, когда общие для нескольких формул выражения вычисляются один раз, а общие множители при нескольких слагаемых выносятся за скобку, число действий для (10)–(13) составляет $Q = 21N + 1$ операций. Это примерно в 2 раза больше того числа действий, которое нужно было бы затратить, чтобы по формулам обычной прогонки найти решение y_i задачи (1), а затем по формуле (7) найти поток w_i .

2. Метод циклической прогонки. Рассмотрим теперь следующую систему:

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad i = 0, \pm 1, \pm 2, \dots, \quad (14)$$

коэффициенты и правая часть которой периодичны с периодом N :

$$a_i = a_{i+N}, \quad b_i = b_{i+N}, \quad c_i = c_{i+N}, \quad f_i = f_{i+N}. \quad (15)$$

К системам типа (14), (15) мы приходим, например, при рассмотрении трехточечных разностных схем, предназначенных для отыскания периодических решений обыкновенных дифференциаль-

ных уравнений второго порядка, а также при приближенном решении уравнений с частными производными в цилиндрических и сферических координатах.

При выполнении условий (15) решение системы (14), если оно существует, тоже будет периодическим с периодом N , т. е.

$$y_i = y_{i+N}. \quad (16)$$

Поэтому достаточно найти решение y_i , например, при $i = 0, 1, \dots, N-1$. В этом случае задачу (14)–(16) можно записать так:

$$-a_0 y_{N-1} + c_0 y_0 - b_0 y_1 = f_0, \quad i = 0, \quad (17)$$

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad 1 \leq i \leq N-1, \quad (17)$$

$$y_N = y_0. \quad (18)$$

Условие (18) мы добавили к системе (17), чтобы из уравнения системы для $i = N-1$ не исключать y_N , заменяя его на y_0 . Это позволяет сохранить единый вид для уравнений (17) при $i = 1, 2, \dots, N-1$.

Если ввести векторы неизвестных $\mathbf{Y} = (y_0, y_1, \dots, y_{N-1})$ и правой части $\mathbf{F} = (f_0, f_1, \dots, f_{N-1})$, то (17), (18) можно записать в векторном виде $\mathcal{A}\mathbf{Y} = \mathbf{F}$, где

$$\mathcal{A} = \begin{vmatrix} c_0 & -b_0 & 0 & 0 & \dots & 0 & 0 & -a_0 \\ -a_1 & c_1 & -b_1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -a_2 & c_2 & -b_2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & c_{N-3} & -b_{N-3} & 0 \\ 0 & 0 & 0 & 0 & \dots & -a_{N-2} & c_{N-2} & -b_{N-2} \\ -b_{N-1} & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & c_{N-1} \end{vmatrix}$$

— матрица системы (17), (18). Присутствие отличных от нуля коэффициентов a_0 и b_{N-1} в (17) не позволяет решать эту систему методом прогонки, описанным в § 1. Для нахождения решения системы (17), (18) построим вариант метода прогонки, который называется *методом циклической прогонки*.

Решение задачи (17), (18) будем искать в виде линейной комбинации сеточных функций u_i и v_i

$$y_i = u_i + y_0 v_i, \quad 0 \leq i \leq N, \quad (19)$$

где u_i есть решение неоднородной трехточечной краевой задачи

$$\begin{aligned} -a_i u_{i-1} + c_i u_i - b_i u_{i+1} &= f_i, \quad 1 \leq i \leq N-1, \\ u_0 &= 0, \quad u_N = 0 \end{aligned} \quad (20)$$

с однородными краевыми условиями, а v_i — решение однородной трехточечной краевой задачи

$$\begin{aligned} -a_i v_{i-1} + c_i v_i - b_i v_{i+1} &= 0, \quad 1 \leq i \leq N-1, \\ v_0 &= 1, \quad v_N = 1 \end{aligned} \quad (21)$$

с неоднородными краевыми условиями.

Найдем, при каком условии y_i из (19) есть искомое решение. Умножая (21) на y_0 , складывая с (20) и учитывая (19), получим, что уравнения системы (17) для $i = 1, 2, \dots, N-1$ будут выполнены. Из краевых условий для u_i и v_i следует, что будет выполнено соотношение (18). Таким образом, если y_i , определяемое по формуле (19), будет удовлетворять оставшемуся неиспользованным уравнению системы (17) для $i=0$, то задача будет решена. Подставляя (19) в это уравнение, получим

$$-a_0 u_{N-1} - a_0 y_0 v_{N-1} + c_0 y_0 - b_0 u_1 - b_0 y_0 v_1 = f_0. \quad (22)$$

Таким образом, если выбрать y_0 по формуле

$$y_0 = \frac{f_0 + a_0 u_{N-1} + b_0 u_1}{c_0 - a_0 v_{N-1} - b_0 v_1}, \quad (23)$$

то равенство (22) будет выполнено, и следовательно, решение задачи (17), (18) можно найти по формуле (19).

Остановимся теперь на решении систем (20) и (21). Они являются частными случаями трехточечных систем уравнений, для которых в § 1 был построен метод прогонки. Для (20) и (21) формулы прогонки принимают следующий вид:

$$\begin{aligned} u_i &= \alpha_{i+1} u_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \dots, 1, \quad u_N = 0, \\ v_i &= \alpha_{i+1} v_{i+1} + \gamma_{i+1}, \quad i = N-1, N-2, \dots, 1, \quad v_N = 1, \end{aligned} \quad (24)$$

где прогоночные коэффициенты α_i , β_i и γ_i находятся по следующим формулам:

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N, \quad \alpha_1 = 0, \quad (25)$$

$$\beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N, \quad \beta_1 = 0, \quad (26)$$

$$\gamma_{i+1} = \frac{a_i \gamma_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \dots, N, \quad \gamma_1 = 1. \quad (27)$$

Преобразуем (23). Из (24) получим $u_{N-1} = \alpha_N u_N + \beta_N = \beta_N$, $v_{N-1} = \gamma_N + \alpha_N$. Подставим эти выражения в (23) и учтем условия (15), (25)–(27):

$$y_0 = \frac{f_N + a_N \beta_N + \beta_N u_1}{c_N - a_N \alpha_N - a_N \gamma_N - b_N v_1} = \frac{\beta_{N+1} + \alpha_{N+1} u_1}{1 - \gamma_{N+1} - \alpha_{N+1} v_1}.$$

Мы построили алгоритм решения задачи (17), (18), который носит название метода циклической прогонки:

$$\begin{aligned} \alpha_2 &= b_1/c_1, \quad \beta_2 = f_1/c_1, \quad \gamma_2 = a_1/c_1, \\ \alpha_{i+1} &= \frac{b_i}{c_i - a_i \alpha_i}, \quad \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad \gamma_{i+1} = \frac{a_i \gamma_i}{c_i - a_i \alpha_i}, \\ &\quad i = 2, 3, \dots, N; \\ u_{N-1} &= \beta_N, \quad v_{N-1} = \alpha_N + \gamma_N, \\ u_i &= \alpha_{i+1} u_{i+1} + \beta_{i+1}, \quad v_i = \alpha_{i+1} v_{i+1} + \gamma_{i+1}, \\ &\quad i = N-2, N-3, \dots, 1; \\ y_0 &= \frac{\beta_{N+1} + \alpha_{N+1} u_1}{1 - \gamma_{N+1} - \alpha_{N+1} v_1}, \quad y_i = u_i + y_0 v_i, \quad i = 1, 2, \dots, N-1. \end{aligned} \quad (28)$$

Элементарный подсчет показывает, что для его реализации требуется $6(N-1)$ умножений, $5N-3$ сложений и вычитаний и $3N+1$ делений. Если не делать различия между арифметическими операциями, то общее их число есть $Q = 14N-8$.

Исследуем вопрос о применимости и устойчивости алгоритма (28). Имеет место

Лемма 2. Пусть коэффициенты системы (14), (15) удовлетворяют условиям

$$|a_i| > 0, \quad |b_i| > 0, \quad |c_i| \geq |a_i| + |b_i|, \quad i = 1, 2, \dots, N, \quad (29)$$

и существует такое $1 \leq i_0 \leq N$, что $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$. Тогда

$$c_i - a_i \alpha_i \neq 0, \quad |\alpha_i| \leq 1, \quad |\alpha_i| + |\gamma_i| \leq 1, \quad i = 2, 3, \dots, N,$$

$$1 - \gamma_{N+1} - \alpha_{N+1} v_1 \neq 0.$$

Действительно, так как α_i , β_i и γ_i есть прогоночные коэффициенты метода правой прогонки, примененного к решению задач (20) и (21), а в силу (29), условия леммы 1 выполнены, то из леммы 1 следует справедливость неравенств

$$c_i - a_i \alpha_i \neq 0, \quad |\alpha_i| \leq 1, \quad i = 2, 3, \dots, N,$$

$$|c_i - a_i \alpha_i| \geq |c_i| - |a_i| |\alpha_i| \geq |b_i| > 0. \quad (30)$$

Далее, на основании условий леммы 2, $|a_1| + |b_1| \leq |c_1|$ и, следовательно, $|\alpha_2| + |\gamma_2| \leq 1$. Отсюда методом индукции получим неравенства

$$|\alpha_i| + |\gamma_i| \leq 1, \quad i = 2, 3, \dots, N, \quad (31)$$

так как

$$|\alpha_{i+1}| + |\gamma_{i+1}| = \frac{|b_i| + |a_i| |\gamma_i|}{|c_i - a_i \alpha_i|} \leq \frac{|a_i| + |b_i| - |a_i| (1 - |\gamma_i|)}{|c_i| - |a_i| |\alpha_i|} \leq$$

$$\leq \frac{|a_i| + |b_i| - |a_i| |\alpha_i|}{|c_i| - |a_i| |\alpha_i|} \leq 1$$

и имеет место (30). Заметим, что $|c_i| > |a_i| + |b_i|$ для $i = i_0$ и, следовательно, $|\alpha_{i_0+1}| + |\gamma_{i_0+1}| < 1$. Отсюда следует, что для $i \geq i_0 + 1$ имеет место строгое неравенство $|\alpha_i| + |\gamma_i| < 1$. Так как $1 \leq i_0 \leq N$, то $|\alpha_{N+1}| + |\gamma_{N+1}| < 1$.

Нам осталось показать, что $1 - \gamma_{N+1} - \alpha_{N+1} v_1 \neq 0$. На основании (28) и (31) получим

$$|v_{N-1}| \leq |\alpha_N| + |\gamma_N| \leq 1,$$

а далее методом индукции докажем неравенства $|v_i| \leq 1$, $1 \leq i \leq N-1$, так как в силу (31)

$$|v_i| \leq |\alpha_{i+1}| |v_{i+1}| + |\gamma_{i+1}| \leq |\alpha_{i+1}| + |\gamma_{i+1}| \leq 1.$$

В частности, $|v_1| \leq 1$. Отсюда, с учетом доказанного неравенства $|\alpha_{N+1}| + |\gamma_{N+1}| < 1$, делаем вывод, что

$$|1 - \gamma_{N+1} - \alpha_{N+1} v_1| \geq 1 - |\gamma_{N+1}| - |\alpha_{N+1}| |v_1| \geq 1 - |\alpha_{N+1}| - |\gamma_{N+1}| > 0.$$

Лемма 2 полностью доказана.

В заключение заметим, что от правой части f_i зависит прогоночный коэффициент β_i и, следовательно, u_i и y_i . Прогоночные коэффициенты α_i и γ_i , а также v_i не зависят от f_i и при решении лишь первой задачи из серии вычисляются и запоминаются. Это позволяет вторую и каждую следующую задачу из серии решить за $Q = 9N - 4$ действий.

3. Метод прогонки для сложных систем. Продолжим построение вариантов метода прогонки для решения систем разностных уравнений с матрицами, отличными от трехдиагональных. В п. 2 метод циклической прогонки применялся для решения систем, матрицы которых содержали вне главных диагоналей только два ненулевых элемента. Рассмотрим теперь более общий случай.

Пусть требуется решить следующую систему уравнений:

$$\begin{aligned} c_0 y_0 - \sum_{j=1}^{N-1} d_j y_j - \psi_0 y_N &= f_0, & i = 0, \\ -\varphi_i y_0 - a_i y_{i-1} + c_i y_i - b_i y_{i+1} - \psi_i y_N &= f_i, & 1 \leq i \leq N-1, \\ -\varphi_N y_0 - \sum_{j=1}^{N-1} g_j y_j + c_N y_N &= f_N, & i = N. \end{aligned} \quad (32)$$

Система вида (32) возникает при аппроксимации обыкновенных дифференциальных уравнений второго порядка в случае связанных краевых условий, при нахождении решений, удовлетворяющих дополнительным условиям интегрального типа, и в ряде других случаев. В частности, в таком виде могут быть записаны все рассмотренные выше системы разностных уравнений. Например, если в (32) положить

$$\begin{aligned} d_1 &= b_0, & d_{N-1} &= a_0, & d_i &= 0, & 2 \leq i \leq N-2, \\ \varphi_i &= \psi_i = g_i = 0, & 1 \leq i \leq N-1, \\ \psi_0 &= 0, & \varphi_N &= c_N = 1, & f_N &= 0, \end{aligned}$$

то мы получим задачу (17), (18).

Если ввести векторы $\mathbf{Y} = (y_0, y_1, \dots, y_N)$ и $\mathbf{F} = (f_0, \dots, f_N)$, то (32) можно записать в векторном виде $\mathcal{A}\mathbf{Y} = \mathbf{F}$, где

$\mathcal{A} =$

$$\left(\begin{array}{ccccccccc} c_0 & -d_1 & -d_2 & -d_3 & \dots & -d_{N-3} & -d_{N-2} & -d_{N-1} & -\psi_0 \\ -\varphi_1 - a_1 & c_1 & -b_1 & 0 & \dots & 0 & 0 & 0 & -\psi_1 \\ -\varphi_2 & -a_2 & c_2 & -b_2 & \dots & 0 & 0 & 0 & -\psi_2 \\ -\varphi_3 & 0 & -a_3 & c_3 & \dots & 0 & 0 & 0 & -\psi_3 \\ \dots & \dots \\ -\varphi_{N-3} & 0 & 0 & 0 & \dots & c_{N-3} & -b_{N-3} & 0 & -\psi_{N-3} \\ -\varphi_{N-2} & 0 & 0 & 0 & \dots & -a_{N-2} & c_{N-2} & -b_{N-2} & -\psi_{N-2} \\ -\varphi_{N-1} & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & c_{N-1} & -b_{N-1} - \psi_{N-1} \\ -\varphi_N & -g_1 & -g_2 & -g_3 & \dots & -g_{N-3} & -g_{N-2} & -g_{N-1} & c_N \end{array} \right)$$

Видно, что матрица \mathcal{A} получена окаймлением трехдиагональной матрицы при помощи столбцов и строк со всех четырех сторон. Заметим, что при другом упорядочении неизвестных $\mathbf{Y}^* = (y_1, y_2, \dots, y_N, y_0)$ система (32) запишется в виде $\mathcal{A}^* \mathbf{Y}^* = \mathbf{F}^*$, где матрица \mathcal{A}^* получается окаймлением той же трехдиагональной матрицы, но только при помощи двух столбцов справа и двух строк снизу.

Переходим к построению метода решения задачи (32). Решение задачи (32) будем искать в виде линейной комбинации трех сеточных функций u_i , v_i и w_i :

$$y_i = u_i + y_0 v_i + y_N w_i, \quad 0 \leq i \leq N, \quad (33)$$

где u_i , v_i и w_i есть решения следующих трехточечных краевых задач:

$$\begin{cases} -a_i u_{i-1} + c_i u_i - b_i u_{i+1} = f_i, & 1 \leq i \leq N-1, \\ u_0 = 0, \quad u_N = 0; \end{cases} \quad (34)$$

$$\begin{cases} -a_i v_{i-1} + c_i v_i - b_i v_{i+1} = \varphi_i, & 1 \leq i \leq N-1, \\ v_0 = 1, \quad v_N = 0; \end{cases} \quad (35)$$

$$\begin{cases} -a_i w_{i-1} + c_i w_i - b_i w_{i+1} = \psi_i, & 1 \leq i \leq N-1, \\ w_0 = 0, \quad w_N = 1. \end{cases} \quad (36)$$

Из (33) — (36) видно, что для $1 \leq i \leq N-1$ уравнения системы (32) выполняются. Краевые условия для u_i , v_i и w_i обеспечивают превращение (33) в тождество при $i=0$ и $i=N$. Таким образом, если будут решены задачи (34) — (36) и найдены y_0 и y_N , то формула (33) будет определять решение исходной задачи (32). Найдем сначала y_0 и y_N .

Значения для y_0 и y_N найдем, используя уравнения системы (32) при $i=0$ и $i=N$. Подставляя в эти уравнения y_i из (33), получим систему из двух уравнений для y_0 и y_N :

$$\begin{aligned} \left(c_0 - \sum_{j=1}^{N-1} d_j v_j \right) y_0 - \left(\psi_0 + \sum_{j=1}^{N-1} d_j w_j \right) y_N &= f_0 + \sum_{j=1}^{N-1} d_j u_j, \\ - \left(\varphi_N + \sum_{j=1}^{N-1} g_j v_j \right) y_0 + \left(c_N - \sum_{j=1}^{N-1} g_j w_j \right) y_N &= f_N + \sum_{j=1}^{N-1} g_j u_j. \end{aligned}$$

Если детерминант этой системы

$$\Delta = \left(c_0 - \sum_{j=1}^{N-1} d_j v_j \right) \left(c_N - \sum_{j=1}^{N-1} g_j w_j \right) - \left(\psi_0 + \sum_{j=1}^{N-1} d_j w_j \right) \left(\varphi_N + \sum_{j=1}^{N-1} g_j v_j \right) \quad (37)$$

отличен от нуля, то она имеет единственное решение

$$\begin{aligned} y_0 &= \frac{1}{\Delta} \left[\left(c_N - \sum_{j=1}^{N-1} g_j w_j \right) \left(f_0 + \sum_{j=1}^{N-1} d_j u_j \right) + \right. \\ &\quad \left. + \left(\psi_0 + \sum_{j=1}^{N-1} d_j w_j \right) \left(f_N + \sum_{j=1}^{N-1} g_j u_j \right) \right], \end{aligned} \quad (38)$$

$$\begin{aligned} y_N &= \frac{1}{\Delta} \left[\left(\varphi_N + \sum_{j=1}^{N-1} g_j v_j \right) \left(f_0 + \sum_{j=1}^{N-1} d_j u_j \right) + \right. \\ &\quad \left. + \left(c_0 - \sum_{j=1}^{N-1} d_j v_j \right) \left(f_N + \sum_{j=1}^{N-1} g_j u_j \right) \right]. \end{aligned} \quad (39)$$

Рассмотрим теперь метод решения вспомогательных задач (34) — (36). Так как здесь мы имеем дело с обычными краевыми задачами для трехточ-

ческих уравнений, то можно использовать метод прогонки, описанный в § I. Для (34) — (36) формулы алгоритма правой прогонки принимают следующий вид:

$$\begin{aligned} u_i &= \alpha_{i+1} u_{i+1} + \beta_{i+1}, \quad i = N-1, \dots, 0, \quad u_N = 0, \\ v_i &= \alpha_{i+1} v_{i+1} + \gamma_{i+1}, \quad i = N-1, \dots, 0, \quad v_N = 0, \\ w_i &= \alpha_{i+1} w_{i+1} + \delta_{i+1}, \quad i = N-1, \dots, 0, \quad w_N = 1, \end{aligned} \quad (40)$$

где прогоночные коэффициенты α_i , β_i , γ_i и δ_i определяются по формулам

$$\begin{aligned} \alpha_{i+1} &= \frac{b_i}{c_i - a_i \alpha_i}, \quad \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \\ i &= 1, 2, \dots, N-1, \quad \alpha_1 = 0, \quad \beta_1 = 0, \\ \gamma_{i+1} &= \frac{\varphi_i + a_i \gamma_i}{c_i - a_i \alpha_i}, \quad \delta_{i+1} = \frac{\psi_i + a_i \delta_i}{c_i - a_i \alpha_i}, \\ i &= 1, 2, \dots, N-1, \quad \gamma_1 = 1, \quad \delta_1 = 0. \end{aligned} \quad (41)$$

Таким образом, для задачи (32) метод прогонки описывается формулами (33), (37) — (41).

Рассмотрим теперь вопрос об устойчивости и корректности предложенного алгоритма. В силу леммы 1 условия

$$|a_i| > 0, \quad |b_i| > 0, \quad |c_i| \geq |a_i| + |b_i|, \quad 1 \leq i \leq N-1 \quad (42)$$

достаточны для устойчивости и корректности метода прогонки (40) — (41) решения вспомогательных задач (34) — (36). Можно показать, что если исходная система (32) имеет единственное решение, то детерминант Δ , определенный формулой (37), отличен от нуля. В этом случае формулы (38) и (39) для вычисления y_0 и y_N будут корректны. Сформулируем полученный результат в виде леммы.

Лемма 3. Если система (32) имеет единственное решение и выполнены условия (42), то алгоритм (33), (37) — (41) метода прогонки для задачи (32) корректен и устойчив.

Заметим, что формулировка простых и в то же время не слишком ограничительных достаточных условий разрешимости системы (32) является сложной задачей. Приведем пример условий, которые обеспечивают корректность и устойчивость предложенного алгоритма. Пусть матрица системы (32) имеет диагональное преобладание, т. е. выполнены условия

$$|c_i| \geq |a_i| + |b_i| + |\varphi_i| + |\psi_i|, \quad 1 \leq i \leq N-1, \quad (43)$$

$$|c_0| \geq |\varphi_0| + \sum_{j=1}^{N-1} |d_j|, \quad |c_N| \geq |\psi_N| + \sum_{j=1}^{N-1} |g_j|, \quad (44)$$

$$|a_i| > 0, \quad |b_i| > 0, \quad 1 \leq i \leq N-1, \quad |c_0| > 0, \quad |c_N| > 0,$$

причем, хотя бы в одном из неравенств (43) или (44) выполняется строгое неравенство.

Укажем основные этапы доказательства. Сначала доказывается, что имеют место неравенства $|\alpha_i| + |\gamma_i| + |\delta_i| \leq 1$, $1 \leq i \leq N$. Далее доказываются неравенства $|v_i| + |w_i| \leq 1$ для $1 \leq i \leq N$, причем, если в (43) хотя бы для одного i выполняется строгое неравенство, то для всех $1 \leq i \leq N$ верны неравенства $|v_i| + |w_i| < 1$. Далее имеем

$$\begin{aligned} \left| c_0 - \sum_{j=1}^{N-1} d_j v_j \right| &\geq |c_0| - \sum_{j=1}^{N-1} |d_j| |v_j| \geq |\varphi_0| + \\ &+ \sum_{j=1}^{N-1} (1 - |v_j|) |d_j| \geq |\varphi_0| + \sum_{j=1}^{N-1} |w_j| |d_j| \geq \left| \varphi_0 + \sum_{j=1}^{N-1} w_j d_j \right| \end{aligned}$$

и аналогично

$$\left| c_N - \sum_{j=1}^{N-1} g_j w_j \right| \geq \left| \varphi_N + \sum_{j=1}^{N-1} g_j v_j \right|,$$

причем хотя бы в одном из этих неравенств достигается строгое неравенство. Отсюда следует, что детерминант Δ , определенный в (37), отличен от нуля. Устойчивость и корректность метода прогонки для решения вспомогательных задач (34) — (36) следуют из (43).

В качестве примера задачи, сводящейся к (32), рассмотрим схему с весами

$$\begin{aligned} y_t, i = \sigma y_{xx, i}^{n+1} + (1-\sigma) y_{xx, i}^n, & \quad 1 \leq i \leq N-1, \\ y_0^n - y_k^n = \mu_1(t_n), & \quad y_N^n - y_k^n = \mu_2(t_n), \\ y_i^0 = u_0(x_i), & \quad n = 0, 1, \dots, 1 \leq k \leq N-1, \end{aligned} \quad (45)$$

аппроксимирующую уравнение теплопроводности со связанными (нелокальными) краевыми условиями

$$\begin{aligned} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, & \quad 0 < x < l, \quad t > 0, \\ u(0, t) - u(v(t), t) = \mu_1(t), & \\ u(l, t) - u(v(t), t) = \mu_2(t), & \quad u(x, 0) = u_0(x), \end{aligned}$$

где функция $x = v(t)$ принимает значения от 0 до l . Отметим, что в схеме (45) кривая $x = v(t)$ аппроксимирована ломаной $x_k = v(t_n)$, так что точки (x_k, t_n) являются узловыми точками сетки.

Разностная схема (45) записывается в виде системы (32) относительно $y_t = y_i^{n+1}$ при следующих значениях коэффициентов и правой части ($\sigma \neq 0$):

$$\begin{aligned} c_0 = 1, \quad d_k = 1, \quad f_0 = \mu_1(t_{n+1}), \quad \varphi_0 = 0, \quad d_j = 0, \quad j \neq k, \\ c_N = 1, \quad q_k = 1, \quad f_N = \mu_2(t_{n+1}), \quad \varphi_N = 0, \quad g_j = 0, \quad j \neq k, \\ \varphi_i = \psi_i = 0, \quad a_i = b_i = 1/h^2, \quad c_i = a_i + b_i + 1/(\sigma t), \\ f_i = \frac{1}{\sigma t} y_i^n + \left(\frac{1}{\sigma} - 1 \right) y_{xx, i}^n, \quad i = 1, 2, \dots, N-1. \end{aligned}$$

Отсюда получим, что требование $|2/h^2 + 1/(\sigma t)| > 2/h^2$ обеспечивает выполнение условий (43), (44). Следовательно, при $\sigma > -h^2/(4t)$ для нахождения решения схемы (45) на верхнем слое можно использовать описанный здесь вариант метода прогонки, который будет устойчив и корректен.

4. Метод немонотонной прогонки. Вернемся снова к методу прогонки решения трехточечных уравнений:

$$\begin{aligned} c_0 y_0 - b_0 y_1 = f_0, \quad i = 0, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad i = 1, 2, \dots, N-1, \\ -a_N y_{N-1} + c_N y_N = f_N, \quad i = N, \end{aligned} \quad (46)$$

который был построен в § 1. Напомним, что в алгоритме правой (левой) прогонки неизвестные y_i находятся последовательно при движении в сторону уменьшения (увеличения) индекса i . При этом y_i выражается только через соседнее неизвестное. Такая

структурой алгоритма дает основание назвать построенный метод *методом монотонной прогонки*.

Монотонный порядок определения неизвестных y_i на обратном ходе метода порожден естественным порядком исключения неизвестных из уравнений на прямом ходе. Таким образом, метод монотонной прогонки есть метод исключения Гаусса без выбора главного элемента, примененный к специальной системе линейных алгебраических уравнений (46) с трехдиагональной матрицей. Известно, что такой вариант метода исключения Гаусса корректен для случая систем уравнений с матрицами, имеющими диагональное преобладание. Для системы (46) это утверждение доказано в лемме 1.

Остановимся на этом более подробно. Напомним, что в § 1, п. 1 на l -м шаге процесса исключения неизвестных в системе (46) была получена «укороченная» система

$$\begin{aligned} (c_l - a_l \alpha_l) y_l - b_l y_{l+1} &= f_l + a_l \beta_l, \quad l = l, \\ -a_l y_{l-1} + c_l y_l - b_l y_{l+1} &= f_l, \quad l+1 \leq i \leq N-1, \\ -a_N y_{N-1} + c_N y_N &= f_N \end{aligned} \quad (47)$$

для неизвестных y_l, y_{l+1}, \dots, y_N . Предполагая, что $c_l - a_l \alpha_l$ отлично от нуля, мы преобразовывали первое уравнение системы (47) к виду

$$y_l = \alpha_{l+1} y_{l+1} + \beta_{l+1}, \quad \alpha_{l+1} = b_l / (c_l - a_l \alpha_l) \quad (48)$$

и использовали его для исключения y_l из уравнения (47) при $i = l+1$. Лемма 1 утверждает, что если матрица \mathcal{A} системы (46) имеет диагональное преобладание, то справедливо неравенство $|c_l - a_l \alpha_l| \geq |b_l|$. Следовательно, в первом уравнении системы (47) коэффициент при y_l по модулю больше коэффициента при y_{l+1} . Поэтому выбор главного элемента по строке осуществлять не надо, переход к виду (48) корректен и условие устойчивости $|\alpha_{l+1}| \leq 1$ автоматически выполняется.

Если же диагональное преобладание не имеет места, то гарантировать отличие от нуля величины $c_l - a_l \alpha_l$, равно как и неравенство $|\alpha_{l+1}| \leq 1$, невозможно. В этом случае алгоритм монотонной прогонки может породить деление на нуль или сильную чувствительность к ошибкам округления, и следовательно, необходимо модифицировать такой алгоритм. Построение корректного алгоритма метода прогонки для системы (46), имеющей единственное решение, базируется на использовании выбора главного элемента по строкам в методе исключения Гаусса. В таком алгоритме монотонный порядок определения неизвестных y_i может быть нарушен, и поэтому этот метод будем называть *методом немонотонной прогонки*.

Переходим к описанию алгоритма немонотонной прогонки. Пусть в результате l -го шага процесса исключения Гаусса с выбором главного элемента по строке, примененного к системе (46),

получена следующая «укороченная» система:

$$Cy_{m_l} - b_l y_{l+1} = F, \quad i = l, \quad (49)$$

$$-Ay_{m_l} + c_{l+1}y_{l+1} - b_{l+1}y_{l+2} = \Phi, \quad i = l+1, \quad (50)$$

$$-a_{l+2}y_{l+1} + c_{l+2}y_{l+2} - b_{l+2}y_{l+3} = f_{l+2}, \quad i = l+2, \quad (51)$$

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad l+3 \leq i \leq N-1, \quad (52)$$

$$-a_N y_{N-1} + c_N y_N = f_N, \quad i = N, \quad (53)$$

где $m_l \leq l$. (При $l=0$ в (49)–(53) следует положить $C=c_0$, $A=a_1$, $F=f_0$, $\Phi=f_1$ и $m_0=0$).

Опишем $(l+1)$ -й шаг процесса исключения. Стратегия выбора главного элемента по строке приводит нас к двум случаям:

a) Если $|C| \geq |b_l|$, то уравнение (49) преобразуется к виду

$$y_{m_l} - \alpha_{l+1} y_{l+1} = \beta_{l+1}, \quad \alpha_{l+1} = b_l/C, \quad \beta_{l+1} = F/C,$$

причем $|\alpha_{l+1}| \leq 1$, неизвестное с индексом m_l находится через неизвестное с индексом $l+1$. Далее, при помощи полученного уравнения исключается y_{m_l} из (50). Это дает следующее уравнение:

$$Cy_{m_{l+1}} - b_{l+1}y_{l+2} = F, \quad i = l+1, \quad (54)$$

где обозначено $m_{l+1} = l+1$, $C = c_{l+1} - A\alpha_{l+1}$, $F = \Phi + A\beta_{l+1}$. Уравнение (51) не преобразуется, так как оно не содержит y_{m_l} , но переписывается в виде

$$-Ay_{m_{l+1}} + C_{l+2}y_{l+2} - b_{l+2}y_{l+3} = \Phi, \quad i = l+2, \quad (55)$$

где полагается $A = a_{l+2}$, $\Phi = f_{l+2}$. Сбъединяя (54) и (55) с (52), (53), получим новую «укороченную» систему вида (49)–(53), в которой l заменено на $l+1$. На этом $(l+1)$ -й шаг заканчивается.

b) Если $|C| < |b_l|$, то (49) преобразуется к виду

$$y_{l+1} - \alpha_{l+1} y_{m_l} = \beta_{l+1}, \quad \alpha_{l+1} = C/b_l, \quad \beta_{l+1} = -F/b_l,$$

где снова $|\alpha_{l+1}| \leq 1$, но на этот раз неизвестное с индексом $l+1$ вычисляется через неизвестное с индексом m_l . Полученное уравнение используется для исключения y_{l+1} из (50) и (51). При этом уравнение (50) будет преобразовано к виду (54), где $m_{l+1} = m_l$, $C = c_{l+1}\alpha_{l+1} - A$, $F = \Phi - c_{l+1}\beta_{l+1}$, а уравнение (51) — к виду (55), где величины A и Φ переопределяются по формулам $A = a_{l+2}\alpha_{l+1}$, $\Phi = f_{l+2} + a_{l+2}\beta_{l+1}$. Уравнения (52), (53) не преобразовываются, так как не содержат y_{l+1} . Снова мы получаем систему вида (49)–(53). Она отличается от полученной в первом случае коэффициентами C и A и правыми частями F и Φ , вычисленными по другим формулам.

Итак, описан один шаг процесса исключения с выбором главного элемента. Заметим, что если исходная система не вырождена, то в уравнении (49) коэффициенты C и b_i одновременно в нуль обратиться не могут. Это обеспечивает корректность формул для прогоночных коэффициентов α_{i+1} и β_{i+1} . Так как все вычисляемые α_{i+1} по модулю не превосходят единицы, то процесс вычисления неизвестных y_i на обратном ходе метода будет устойчив по отношению к ошибкам округления.

Для предлагаемого алгоритма порядок вычисления неизвестных может иметь немонотонный характер. Это требует хранения информации о том, какое неизвестное вычисляется через какое, уже найденное на предыдущих шагах неизвестное, при помощи прогоночных коэффициентов α_{i+1} и β_{i+1} . Эту информацию можно хранить в виде двух целочисленных множеств индексов θ и κ : $\theta = \{\theta_i, 1 \leq i \leq N\}$, $\kappa = \{\kappa_i, 1 \leq i \leq N\}$, так что неизвестные находятся по формулам $y_m = \alpha_{i+1}y_n + \beta_{i+1}$, $m = \theta_{i+1}$, $n = \kappa_{i+1}$, $i = -N-1, -N-2, \dots, 0$. Множества θ и κ строятся на прямом ходе метода.

Полностью алгоритм метода немонотонной прогонки можно определить следующим образом.

1) Задаются начальные значения для C , A , F и Φ : $C = c_0$, $A = a_1$, $F = f_0$, $\Phi = f_1$, и формально полагается $\kappa_0 = 0$.

2) Последовательно для $i = 0, 1, \dots, N-1$ выполняются в зависимости от ситуации действия, описанные в пп. а) или б):

а) если $|C| \geq |b_i|$, то $\alpha_{i+1} = b_i/C$, $\beta_{i+1} = F/C$, $C = c_{i+1} - A\alpha_{i+1}$, $F = \Phi + A\beta_{i+1}$, $\theta_{i+1} = \kappa_i$, $\kappa_{i+1} = i+1$, $A = a_{i+2}$, $\Phi = f_{i+2}$;

б) если $|C| < |b_i|$, то $\alpha_{i+1} = C/b_i$, $\beta_{i+1} = -F/b_i$, $C = c_{i+1}\alpha_{i+1} - A$, $F = \Phi - c_{i+1}\beta_{i+1}$, $\theta_{i+1} = i+1$, $\kappa_{i+1} = \kappa_i$, $A = a_{i+2}\alpha_{i+1}$, $\Phi = f_{i+2} + a_{i+2}\beta_{i+1}$.

Замечание. Для $i = N-1$ переопределение A и Φ в пп. а) или б) проводить не нужно.

3) Вычисляется сначала неизвестное y_n , где $n = \kappa_N$ по формуле $y_n = F/C$, а затем последовательно для $i = N-1, N-2, \dots, 0$ вычисляются остальные неизвестные $y_m = \alpha_{i+1}y_n + \beta_{i+1}$, $m = \theta_{i+1}$, $n = \kappa_{i+1}$.

Отметим, что предлагаемый здесь алгоритм переходит в обычный алгоритм правой прогонки, если выполнены условия леммы 1.

Элементарный подсчет числа арифметических действий для алгоритма метода немонотонной прогонки показывает, что в худшем случае, когда для любого i вычисления ведутся по формулам п. б), требуется $Q = 12N$ действий. Это в 1,5 раза больше, чем в алгоритме монотонной прогонки.

Рассмотрим пример применения метода немонотонной прогонки. Пусть требуется решить следующую разностную задачу:

$$\begin{aligned} -y_{i-1} + y_i - y_{i+1} &= 0, & 1 \leq i \leq N-1, \\ y_0 &= 1, \quad y_N = 0. \end{aligned} \tag{56}$$

Задача (56) есть частный случай системы (46), в которой $f_0 = 1$, $b_0 = a_N = 0$, $c_0 = c_N = 1$, $f_N = 0$, $c_i = a_i = b_i = 1$, $f_i = 0$, $1 \leq i \leq N-1$. Если N не кратно 3, то решение задачи (56) существует и имеет вид (см. п. 1 § 4 гл. I)

$$y_i = \sin \frac{(N-i)\pi}{3} / \sin \frac{N\pi}{3}, \quad 0 \leq i \leq N. \quad (57)$$

Алгоритмы правой и левой прогонок для (56) некорректны, так как при вычислении прогоночных коэффициентов α_3 для правой прогонки и ξ_{N-2} для левой прогонки необходимо будет делить на нулевой знаменатель $c_2 - a_2\alpha_2$ или $c_{N-2} - b_{N-2}\xi_{N-1}$. Алгоритм немонотонной прогонки позволяет получить точное решение (57). Приведем для иллюстрации (табл. 1) значения коэффициентов α_i , β_i , а также θ_i и κ_i для $N=11$.

Таблица 1

$i \backslash$	0	1	2	3	4	5	6	7	8	9	10	11
α_i	0	1	0	-1	1	0	-1	1	0	-1	1	
β_i	1	1	-1	-1	-1	1	1	1	-1	-1	-1	
θ_i	0	1	3	2	4	6	5	7	9	8	10	
κ_i	1	2	2	4	5	5	7	8	8	10	11	
y_i	1	1	0	-1	-1	0	1	1	0	-1	-1	0

§ 3. Метод прогонки для пятиточечных уравнений

1. Алгоритм монотонной прогонки. Выше мы рассмотрели различные варианты метода прогонки, который применяется для нахождения решения трехточечных разностных уравнений. Как было отмечено ранее, такие разностные уравнения возникают при аппроксимации краевых задач для обыкновенных дифференциальных уравнений второго порядка.

При нахождении решения краевых задач для уравнений более высокого порядка можно использовать два способа. Первый способ состоит в переходе к системе дифференциальных уравнений первого порядка и построении соответствующей разностной схемы. В этом случае мы получим краевую задачу для двухточечных векторных уравнений. Методы решения таких разностных задач мы рассмотрим в § 4.

Второй способ заключается в непосредственной аппроксимации исходной дифференциальной задачи. В этом случае мы приходим к многоточечным разностным уравнениям. Наиболее часто встречаются системы пятиточечных уравнений следующего

вида:

$$c_0y_0 - d_0y_1 + e_0y_2 = f_0, \quad i = 0, \quad (1)$$

$$-b_1y_0 + c_1y_1 - d_1y_2 + e_1y_3 = f_1, \quad i = 1, \quad (2)$$

$$a_iy_{i-2} - b_iy_{i-1} + c_iy_i - d_iy_{i+1} + e_iy_{i+2} = f_i, \quad 2 \leq i \leq N-2, \quad (3)$$

$$a_{N-1}y_{N-3} - b_{N-1}y_{N-2} + c_{N-1}y_{N-1} - d_{N-1}y_N = f_{N-1}, \quad i = N-1, \quad (4)$$

$$a_Ny_{N-2} - b_Ny_{N-1} + c_Ny_N = f_N, \quad i = N. \quad (5)$$

Такого вида системы возникают при аппроксимации краевых задач для обыкновенных дифференциальных уравнений четвертого порядка, а также при реализации разностных схем для уравнений в частных производных. Матрица \mathcal{A} системы (1)–(5) является пятидиагональной квадратной матрицей размерности $(N+1) \times (N+1)$ и имеет не более $5N-1$ ненулевых элементов.

Для решения системы (1)–(5) используем метод исключения Гаусса. Учитывая структуру системы (1)–(5), легко получим, что обратный ход метода Гаусса должен осуществляться по формулам

$$y_i = \alpha_{i+1}y_{i+1} - \beta_{i+1}y_{i+2} + \gamma_{i+1}, \quad 0 \leq i \leq N-2, \quad (6)$$

$$y_{N-1} = \alpha_Ny_N + \gamma_N, \quad i = N-1. \quad (7)$$

Для реализации (6), (7) необходимо задать y_N , а также определить коэффициенты α_i , β_i , γ_i .

Сначала найдем формулы для α_i , β_i и γ_i . Используя (6), выразим y_{i-1} и y_{i-2} через y_i и y_{i+1} . Получим

$$y_{i-1} = \alpha_iy_i - \beta_iy_{i+1} + \gamma_i, \quad 1 \leq i \leq N-1, \quad (8)$$

$$y_{i-2} = (\alpha_i\alpha_{i-1} - \beta_{i-1})y_i - \beta_i\alpha_{i-1}y_{i+1} + \alpha_{i-1}\gamma_i + \gamma_{i-1} \quad (9)$$

для $2 \leq i \leq N-1$.

Подставляя (8) и (9) в (3), получим

$$[c_i - a_i\beta_{i-1} + \alpha_i(a_i\alpha_{i-1} - b_i)]y_i = [d_i + \beta_i(a_i\alpha_{i-1} - b_i)]y_{i+1} - e_iy_{i+2} + [f_i - a_i\gamma_{i-1} - \gamma_i(a_i\alpha_{i-1} - b_i)], \quad 2 \leq i \leq N-2.$$

Сравнивая это выражение с (6), видим, что если положить

$$\alpha_{i+1} = \frac{1}{\Delta_i} [d_i + \beta_i(a_i\alpha_{i-1} - b_i)], \quad \beta_{i+1} = \frac{e_i}{\Delta_i}, \quad (10)$$

$$\gamma_{i+1} = \frac{1}{\Delta_i} [f_i - a_i\gamma_{i-1} - \gamma_i(a_i\alpha_{i-1} - b_i)],$$

где обозначено $\Delta_i = c_i - a_i\beta_{i-1} + \alpha_i(a_i\alpha_{i-1} - b_i)$, то уравнения системы (1)–(5) для $2 \leq i \leq N-2$ будут удовлетворены.

Рекуррентные соотношения (10) связывают α_{i+1} , β_{i+1} и γ_{i+1} с α_i , α_{i-1} , β_i , β_{i-1} , γ_i и γ_{i-1} . Поэтому, если будут заданы α_i , β_i и γ_i для $i=1, 2$, то по формулам (10) последовательно можно найти коэффициенты α_i , β_i и γ_i для $3 \leq i \leq N-1$.

Найдем α_i , β_i и γ_i для $i = 1, 2$. Из (1) и формулы (6) для $i = 0$ непосредственно получим

$$\alpha_1 = d_0/c_0, \quad \beta_1 = e_0/c_0, \quad \gamma_1 = f_0/c_0. \quad (11)$$

Далее, подставляя значение (8) при $i = 1$ в (2), получим

$$(c_1 - b_1 \alpha_1) y_1 = (d_1 - b_1 \beta_1) y_2 - e_1 y_3 + f_1 + b_1 \gamma_1.$$

Следовательно, (2) будет выполнено, если положить

$$\alpha_2 = \frac{d_1 - b_1 \beta_1}{c_1 - b_1 \alpha_1}, \quad \beta_2 = \frac{e_1}{c_1 - b_1 \alpha_1}, \quad \gamma_2 = \frac{f_1 + b_1 \gamma_1}{c_1 - b_1 \alpha_1}. \quad (12)$$

Итак, используя (10)–(12), можно найти α_i , β_i и γ_i для $1 \leq i \leq N-1$. Осталось определить α_N , γ_N и y_N , входящие в формулу (7).

Воспользуемся для этого уравнениями (4) и (5). Подставляя (8) и (9) при $i = N-1$ в (4) и сравнивая полученное выражение с (7), найдем, что α_N и γ_N определяются по формулам (10) для $i = N-1$. Найдем теперь y_N . Для этого подставим (6) при $i = N-2$ и (7) в уравнение (5). Получим

$$[c_N - a_N \beta_{N-1} + \alpha_N (a_N \alpha_{N-1} - b_N)] y_N = f_N - a_N \gamma_{N-1} - \gamma_N (a_N \alpha_{N-1} - b_N)$$

или

$$y_N = \gamma_{N+1},$$

где γ_{N+1} определяется по формуле (10) при $i = N$.

Объединяя полученные выше формулы, запишем алгоритм правой прогонки для системы (1)–(5) в следующем виде:

1) по формулам

$$\alpha_{i+1} = \frac{1}{\Delta_i} [d_i + \beta_i (a_i \alpha_{i-1} - b_i)], \quad i = 2, 3, \dots, N-1, \quad (13)$$

$$\alpha_1 = \frac{d_0}{c_0}, \quad \alpha_2 = \frac{1}{\Delta_1} (d_1 - b_1 \beta_1),$$

$$\gamma_{i+1} = \frac{1}{\Delta_i} [f_i - a_i \gamma_{i-1} - \gamma_i (a_i \alpha_{i-1} - b_i)], \quad i = 2, 3, \dots, N, \quad (14)$$

$$\gamma_1 = \frac{f_0}{c_0}, \quad \gamma_2 = \frac{1}{\Delta_1} (f_1 + b_1 \gamma_1),$$

$$\beta_{i+1} = e_i / \Delta_i, \quad i = 1, 2, \dots, N-2, \quad \beta_1 = e_0 / c_0, \quad (15)$$

где

$$\Delta_i = c_i - a_i \beta_{i-1} + \alpha_i (a_i \alpha_{i-1} - b_i), \quad 2 \leq i \leq N, \quad \Delta_1 = c_1 - b_1 \alpha_1, \quad (16)$$

находятся прогоночные коэффициенты α_i , β_i и γ_i ;

2) неизвестные y_i находятся последовательно по формулам

$$y_i = \alpha_{i+1} y_{i+1} - \beta_{i+1} y_{i+2} + \gamma_{i+1}, \quad i = N-2, N-3, \dots, 0, \quad (17)$$

$$y_{N-1} = \alpha_N y_N + \gamma_N, \quad y_N = \gamma_{N+1}.$$

Построенный алгоритм будем также называть алгоритмом *мопотонной прогонки*.

Замечание. Не представляет труда построить алгоритм левой прогонки, а также алгоритм встречных прогонок для системы (1)–(5).

Подсчитаем число арифметических действий для алгоритма (13)–(17). Для реализации (13)–(17) потребуется: $8N - 5$ операций сложения и вычитания, $8N - 5$ операций умножения и $3N$ операций деления. Если не делать различий между временем выполнения арифметических операций на ЭВМ, то общее число действий для предложенного алгоритма $Q = 19N - 10$.

2. Обоснование метода. Построенный выше алгоритм прогонки (13)–(17) будем называть *корректным*, если для любого $2 \leq i \leq N$ будет верно неравенство

$$\Delta_i = c_i - a_i \beta_{i-1} + \alpha_i (a_i \alpha_{i-1} - b_i) \neq 0, \quad \Delta_1 = c_1 - \alpha_1 b_1 \neq 0.$$

Следующая лемма дает достаточные условия корректности алгоритма (13)–(17).

Лемма 4. Пусть коэффициенты системы (1)–(5) удовлетворяют условиям

$$|a_i| > 0, \quad 2 \leq i \leq N, \quad |b_i| > 0, \quad 1 \leq i \leq N, \\ |d_i| > 0, \quad 0 \leq i \leq N-1, \quad |e_i| > 0, \quad 0 \leq i \leq N-2,$$

и условиям

$$|c_0| \geq |d_0| + |e_0|, \quad |c_1| \geq |b_1| + |d_1| + |e_1|, \\ |c_N| \geq |a_N| + |b_N|, \quad |c_{N-1}| \geq |a_{N-1}| + |b_{N-1}| + |d_{N-1}|, \quad (18) \\ |c_i| \geq |a_i| + |b_i| + |d_i| + |e_i|, \quad 2 \leq i \leq N-2,$$

причем хотя бы в одном из неравенств (18) достигается строгое неравенство. Тогда алгоритм (13)–(17) корректен и, кроме того, имеют место неравенства

$$|\alpha_i| + |\beta_i| \leq 1, \quad 1 \leq i \leq N-1, \quad |\alpha_N| \leq 1.$$

Действительно, в силу условий леммы из (13) и (15) получим

$$|\alpha_1| + |\beta_1| = \frac{|d_0| + |e_0|}{|c_0|} \leq 1.$$

Далее, используя полученное неравенство $1 - |\alpha_1| \geq |\beta_1|$, найдем
 $|c_1 - b_1 \alpha_1| \geq |c_1| - |b_1| |\alpha_1| \geq |b_1| (1 - |\alpha_1|) + |d_1| + |e_1| \geq$
 $\geq |b_1| |\beta_1| + |d_1| + |e_1| \geq |d_1 - b_1 \beta_1| + |e_1| > 0.$

Отсюда и из (13)–(15) следует оценка

$$|\alpha_2| + |\beta_2| = \frac{|d_1 - b_1 \beta_1| + |e_1|}{|c_1 - b_1 \alpha_1|} \leq 1.$$

Далее доказательство проведем по индукции. Пусть выполнены неравенства

$$|\alpha_{i-1}| + |\beta_{i-1}| \leq 1, \quad |\alpha_i| + |\beta_i| \leq 1. \quad (19)$$

Покажем, что тогда будут справедливы неравенства

$$\Delta_i = c_i - a_i \beta_{i-1} + \alpha_i (a_i \alpha_{i-1} - b_i) \neq 0, \quad |\alpha_{i+1}| + |\beta_{i+1}| \leq 1.$$

Действительно, из (18) и (19) получим

$$\begin{aligned} |\Delta_i| &\geq |c_i| - |a_i||\beta_{i-1}| - |\alpha_i||\alpha_{i-1}||a_i| - |\alpha_i||b_i| \geq \\ &\geq |a_i|(1 - |\beta_{i-1}|) + |b_i|(1 - |\alpha_i|) - |\alpha_i||\alpha_{i-1}||a_i| + |d_i| + |e_i| \geq \\ &\geq |a_i||\alpha_{i-1}| + |b_i||\beta_i| - |\alpha_i||\alpha_{i-1}||a_i| + |d_i| + |e_i| \geq \\ &\geq |a_i||\alpha_{i-1}|(1 - |\alpha_i|) + |d_i| - b_i \beta_i + |e_i| \geq \\ &\geq |a_i||\alpha_{i-1}||\beta_i| + |d_i| - b_i \beta_i + |e_i| \geq \\ &\geq |d_i| + \beta_i(a_i \alpha_{i-1} - b_i) + |e_i| > 0, \quad i \leq N-2. \end{aligned} \quad (20)$$

Отсюда и из (13), (15) найдем

$$|\alpha_{i+1}| + |\beta_{i+1}| = \frac{|d_i + \beta_i(a_i \alpha_{i-1} - b_i)| + |e_i|}{|\Delta_i|} \leq 1, \quad i \leq N-2.$$

Далее, для $i = N-1$ будем иметь вместо (20) оценку

$$|\Delta_{N-1}| \geq |a_{N-1}||\alpha_{N-2}||\beta_{N-1}| + |b_{N-1}||\beta_{N-1}| + |d_{N-1}| > 0.$$

Кроме того, отсюда получим

$$|\Delta_{N-1}| \geq |d_{N-1} + \beta_{N-1}(a_{N-1} \alpha_{N-2} - b_{N-1})|,$$

и, следовательно,

$$|\alpha_N| = \frac{1}{|\Delta_{N-1}|} |d_{N-1} + \beta_{N-1}(a_{N-1} \alpha_{N-2} - b_{N-1})| \leq 1.$$

Осталось показать, что $\Delta_N \neq 0$. Будем иметь

$$\begin{aligned} |\Delta_N| &\geq |c_N| - |a_N||\beta_{N-1}| - |\alpha_N||\alpha_{N-1}||a_N| - |\alpha_N||b_N| = \\ &= |c_N| - |a_N| - |b_N| + |a_N|(1 - |\beta_{N-1}|) + |b_N|(1 - |\alpha_N|) - \\ &\quad - |\alpha_N||\alpha_{N-1}||a_N| \geq |c_N| - |a_N| - |b_N| + \\ &\quad + (1 - |\alpha_N|)(1 - |\beta_{N-1}|)|a_N| + |b_N|(1 - |\alpha_N|). \end{aligned}$$

В силу предположений леммы легко получить, что хотя бы в одном из неравенств $|c_N| \geq |a_N| + |b_N|$, $|\alpha_N| \leq 1$, достигается строгое неравенство. Отсюда следует, что $\Delta_N \neq 0$. Лемма доказана.

Замечание. Из указанных в лемме 4 оценок $|\alpha_i| + |\beta_i| \leq 1$ следует, что если при вычислении y_N допущена погрешность, то она не будет расти при счете по формулам (17).

3. Вариант немонотонной прогонки. Приведем теперь алгоритм метода прогонки, который получается, если искать решение системы (1)–(5) по методу Гаусса с выбором главного элемента по строке. Такой алгоритм будет корректен при единственном условии невырожденности матрицы \mathcal{A} системы

(1)–(5). Так как прием построения алгоритма аналогичен рассмотренному в п. 4 § 2, то мы ограничимся приведением окончательной формы алгоритма.

1) Задаются начальные значения: $C=c_0$, $D=d_0$, $B=b_1$, $Q=c_1$, $S=a_2$, $T=b_2$, $R=0$, $A=a_3$, $F=f_0$, $\Phi=f_1$, $G=f_2$, $H=f_3$ и полагается $\kappa_0=0$, $\eta_0=1$.

2) Последовательно для $i=0, 1, \dots, N-2$ в зависимости от ситуации выполняются действия, описанные в пп. а), б) или в):

а) если $|C| \geq |D|$ и $|C| \geq |e_i|$, то

$$\begin{aligned} \alpha_{i+1} &= D/C, \quad \beta_{i+1} = e_i/C, \quad \gamma_{i+1} = F/C, \\ C &= Q - B\alpha_{i+1}, \quad D = d_{i+1} - B\beta_{i+1}, \quad F = \Phi + B\gamma_{i+1}, \\ B &= T - S\alpha_{i+1}, \quad Q = c_{i+2} - S\beta_{i+1}, \quad \Phi = G - S\gamma_{i+1}, \\ S &= A - R\alpha_{i+1}, \quad T = b_{i+3} - R\beta_{i+1}, \quad G = H + R\gamma_{i+1}, \end{aligned} \quad (21)$$

$$\left. \begin{aligned} R &= 0, \quad A = a_{i+4}, \quad H = f_{i+4}, \\ \theta_{i+1} &= \kappa_i, \quad \kappa_{i+1} = \eta_i, \quad \eta_{i+1} = i+2; \end{aligned} \right\} \quad (22)$$

б) если $|D| > |C|$ и $|D| \geq |e_i|$, то

$$\begin{aligned} \alpha_{i+1} &= C/D, \quad \beta_{i+1} = -e_i/D, \quad \gamma_{i+1} = -F/D, \\ C &= Q\alpha_{i+1} - B, \quad D = Q\beta_{i+1} + d_{i+1}, \quad F = \Phi - Q\gamma_{i+1}, \\ B &= T\alpha_{i+1} - S, \quad Q = T\beta_{i+1} + c_{i+2}, \quad \Phi = T\gamma_{i+1} + G, \\ S &= A\alpha_{i+1} - R, \quad T = A\beta_{i+1} + b_{i+3}, \quad G = H - A\gamma_{i+1}, \end{aligned} \quad (23)$$

$$\left. \begin{aligned} R &= 0, \quad A = a_{i+4}, \quad H = f_{i+4}, \\ \theta_{i+1} &= \eta_i, \quad \kappa_{i+1} = \kappa_i, \quad \eta_{i+1} = i+2; \end{aligned} \right\} \quad (24)$$

в) если $|e_i| > C$ и $|e_i| > |D|$, то

$$\begin{aligned} \alpha_{i+1} &= D/e_i, \quad \beta_{i+1} = C/e_i, \quad \gamma_{i+1} = F/e_i, \\ C &= Q - d_{i+1}\alpha_{i+1}, \quad D = B - d_{i+1}\beta_{i+1}, \quad F = \Phi + d_{i+1}\gamma_{i+1}, \\ B &= T - c_{i+2}\alpha_{i+1}, \quad Q = S - c_{i+2}\beta_{i+1}, \quad \Phi = G - c_{i+2}\gamma_{i+1}, \\ S &= A - b_{i+3}\alpha_{i+1}, \quad T = R - b_{i+3}\beta_{i+1}, \quad G = H + b_{i+3}\gamma_{i+1}, \end{aligned} \quad (25)$$

$$\left. \begin{aligned} R &= -a_{i+4}\alpha_{i+1}, \quad A = -a_{i+4}\beta_{i+1}, \quad H = f_{i+4} - a_{i+4}\gamma_{i+1}, \\ \theta_{i+1} &= i+2, \quad \kappa_{i+1} = \eta_i, \quad \eta_{i+1} = \kappa_i. \end{aligned} \right\} \quad (26)$$

Замечание. Для $i \geq N-3$ вычислений по формулам (22), (24) или (26) проводить не нужно, а для $i=N-2$ не проводятся вычисления по формулам (21), (23), (25).

3) Если $|C| \geq |D|$, то $\alpha_N = D/C$, $\gamma_N = F/C$, $\eta_{N+1} = (\Phi + B\gamma_N)/(Q - B\alpha_N)$, $\theta_N = \kappa_{N-1}$, $\kappa_N = \eta_{N-1}$. Если $|D| > |C|$, то $\alpha_N = C/D$, $\gamma_N = -F/D$, $\eta_{N+1} = -(\Phi - Q\gamma_N)/(Q\alpha_N - B)$, $\theta_N = \eta_{N-1}$, $\kappa_N = \kappa_{N-1}$.

4) Вычисляются неизвестные $y_n = \gamma_{N+1}$, $y_m = \alpha_N y_n + \gamma_N$, $m = \theta_N$, $n = \kappa_N$, а затем последовательно для $i=N-2, N-3, \dots, 0$ определяются остальные неизвестные $y_m = \alpha_{i+1}y_n - \beta_{i+1}y_k + \gamma_{i+1}$, $m = \theta_{i+1}$, $n = \kappa_{i+1}$, $k = \eta_{i+1}$.

Рассмотрим пример применения метода немонотонной прогонки. В п. 3 § 3 гл. I была решена следующая разностная краевая задача:

$$\begin{aligned} y_0 - y_1 + 2y_2 &= 2, & i=0, \\ -y_0 + y_1 - y_2 + y_3 &= 0, & i=1, \\ y_{i-2} - y_{i-1} + 2y_i - y_{i+1} + y_{i+2} &= 0, & 2 \leq i \leq N-2, \\ y_{N-3} - y_{N-2} + y_{N-1} - y_N &= 0, & i=N-1, \\ 2y_{N-2} - y_{N-1} + y_N &= 0, & i=N. \end{aligned} \quad (27)$$

Если N четно и не кратно 3, то система (27) имеет единственное решение

$$y_i = -\cos \frac{i\pi}{2} - \sin \frac{i\pi}{2}, \quad 0 \leq i \leq N. \quad (28)$$

Несложно убедиться в том, что алгоритм монотонной прогонки для системы (27) некорректен — при вычислении прогоночных коэффициентов α_2 , β_2 и γ_2 нужно будет делить на нуль. Алгоритм немонотонной прогонки позволяет

получить точное решение (28). Приведем для иллюстрации этого алгоритма (табл. 2) значения прогоночных коэффициентов α_i , β_i и γ_i , а также θ_i , χ_i и η_i для $N=10$.

Таблица 2

$i \backslash$	0	1	2	3	4	5	6	7	8	9	10	11
α_i		$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	$-\frac{1}{3}$	$-\frac{1}{3}$	0	1	
β_i		$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	1	-1	1	$-\frac{2}{3}$	$-\frac{2}{3}$	1		
γ_i		1	1	-1	-2	-2	2	$\frac{4}{3}$	$-\frac{2}{3}$	0	-2	1
θ_i		2	3	4	0	5	6	7	9	8	1	
χ_i		1	0	1	1	1	1	1	8	1	10	
η_i		0	1	0	5	6	7	8	1	10		
y_i	-1	-1	1	1	-1	-1	1	1	-1	-1	1	

Из таблицы видно, что неизвестные y_i определяются в следующем порядке: $y_{10}, y_1, y_8, y_9, y_7, y_6, y_5, y_0, y_4, y_3, y_2$, т. е. в немонотонном порядке.

§ 4. Метод матричной прогонки

1. Системы векторных уравнений. Выше было отмечено, что одним из способов решения краевых задач для обыкновенных дифференциальных уравнений высокого порядка является сведение к системе уравнений первого порядка с последующей аппроксимацией этой системы разностной схемой. В результате мы получим двухточечную векторную систему уравнений с краевыми условиями первого рода. В общем виде она записывается следующим образом:

$$\begin{aligned} P_{i+1}V_{i+1} - Q_iV_i &= F_{i+1}, \quad 0 \leq i \leq N-1, \\ P_0V_0 &= F_0, \quad Q_NV_N = F_{N+1}, \end{aligned} \quad (1)$$

где V_i — вектор неизвестных размерности M , P_{i+1} и Q_i для $0 \leq i \leq N-1$ — квадратные матрицы $M \times M$, P_0 и Q_N — прямоугольные матрицы соответственно размеров $M_1 \times M$ и $M_2 \times M$, $M_1 + M_2 = M$. Вектор F_{i+1} для $0 \leq i \leq N-1$ имеет размерность M , а векторы F_0 и F_{N+1} — M_1 и M_2 соответственно.

Заметим, что другим способом решения указанных дифференциальных уравнений является непосредственная аппроксимация этих уравнений разностными схемами. При этом мы получим систему многоточечных скалярных уравнений. Методы решения трех- и пятиточечных скалярных уравнений были изучены нами в § 1—3. Если же аппроксимируется система обыкновенных дифференциальных уравнений высокого порядка, то возникает

система многоточечных векторных уравнений. Однако как скалярные, так и векторные системы многоточечных уравнений могут быть сведены к системам вида (1). При этом любому методу решения (1) будет соответствовать некоторый метод решения исходной многоточечной системы. Идею указанного преобразования поясним на примере системы пятиточечных уравнений, рассмотренной в § 3 (см. (1)–(5)). Если обозначить

$$\begin{aligned} \mathbf{Y}_i &= (y_{i+1}, y_i, y_{i-1}, y_{i-2}), & 2 \leq i \leq N-1, \\ \mathbf{F}_{i+1} &= (f_i, 0, 0, 0), & 2 \leq i \leq N-2, \\ \mathbf{F}_2 &= (f_0, f_1), \quad \mathbf{F}_N = (f_{N-1}, f_N), \end{aligned}$$

то с учетом тождественных соотношений между Y_{i+1} и Y_i указанная система из § 3 запишется в виде

$$\begin{aligned} P_{i+1} Y_{i+1} - Q_i Y_i &= \mathbf{F}_{i+1}, & 2 \leq i \leq N-2, \\ P_2 Y_2 &= \mathbf{F}_2, \quad Q_{N-1} Y_{N-1} &= \mathbf{F}_N, \end{aligned} \tag{2}$$

где

$$\begin{aligned} P_{i+1} &= \begin{vmatrix} e_i & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}, \quad Q_i = \begin{vmatrix} d_i - c_i & b_i - a_i \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{vmatrix}, \quad 2 \leq i \leq N-2, \\ P_2 &= \begin{vmatrix} 0 & e_0 - d_0 & c_0 \\ e_1 - d_1 & c_1 - b_1 \end{vmatrix}, \quad Q_{N-1} = \begin{vmatrix} -d_{N-1} & c_{N-1} - b_{N-1} & a_{N-1} \\ c_N & -b_N & a_N \\ 0 & 0 & 0 \end{vmatrix}. \end{aligned}$$

В данном случае $M_1 = M_2 = 2$, $M = 4$.

Несмотря на то, что многоточечные векторные уравнения можно свести к виду (1) и ограничиться построением метода решения только системы (1), мы рассмотрим отдельно класс *трехточечных векторных уравнений*. Более того, в п. 3 мы сведем (1) к системе трехточечных векторных уравнений и получим метод решения системы (1) как вариант метода решения трехточечных уравнений.

Прежде чем описывать общий вид трехточечных уравнений, рассмотрим пример. Мы покажем, как разностная задача для простейшего эллиптического уравнения сводится к системе трехточечных уравнений специального вида.

Пусть на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq M, 0 \leq j \leq N, l_1 = Mh_1, l_2 = Nh_2\}$ с границей γ , введенной в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, требуется найти решение *разностной задачи Дирихле для уравнения Пуассона*

$$\begin{aligned} y_{x_1 x_1}^- + y_{x_2 x_2}^- &= -\varphi(x), & x \in \omega, \\ y(x) &= g(x), & x \in \gamma, \end{aligned} \tag{3}$$

где

$$y_{x_1 x_1}^- = \frac{1}{h_1^2} [y(i+1, j) - 2y(i, j) + y(i-1, j)],$$

$$y_{x_2 x_2}^- = \frac{1}{h_2^2} [y(i, j+1) - 2y(i, j) + y(i, j-1)], \quad y(i, j) = y(x_{ij}).$$

Преобразуем схему (3). Для этого умножим (3) на $(-h_2^2)$ и распишем по точкам разностную производную $y_{x_1 x_1}^-$. При $1 \leq j \leq N-1$ будем иметь:

для $2 \leq i \leq M-2$

$$-y(i, j-1) + [2y(i, j) - h_2^2 y_{x_1 x_1}^-(i, j)] - y(i, j+1) = h_2^2 \varphi(i, j);$$

для $i=1$

$$\begin{aligned} -y(i, j-1) + \left[2y(i, j) - \frac{h_2^2}{h_1^2} (y(i+1, j) - 2y(i, j)) \right] - y(i, j+1) = \\ = h_2^2 \bar{\varphi}(i, j); \end{aligned}$$

для $i=M-1$

$$\begin{aligned} -y(i, j-1) + \left[2y(i, j) - \frac{h_2^2}{h_1^2} (y(i-1, j) - 2y(i, j)) \right] - y(i, j+1) = \\ = h_2^2 \bar{\varphi}(i, j), \end{aligned}$$

где

$$\begin{aligned} \bar{\varphi}(1, j) &= \varphi(1, j) + \frac{1}{h_1^2} g(0, j), \\ \bar{\varphi}(M-1, j) &= \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j). \end{aligned}$$

Кроме того, для $j=0, N$ имеем

$$y(i, 0) = g(i, 0), \quad y(i, N) = g(i, N), \quad 1 \leq i \leq M-1.$$

Обозначим теперь через \mathbf{Y}_j вектор размерности $M-1$, компонентами которого являются значения сеточной функции $y(i, j)$ во внутренних узлах сетки $\bar{\omega}$ на j -й строке:

$$\mathbf{Y}_j = (y(1, j), y(2, j), \dots, y(M-1, j)), \quad 0 \leq j \leq N,$$

а через \mathbf{F}_j — вектор размерности $M-1$

$$\mathbf{F}_j = (h_2^2 \bar{\varphi}(1, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(M-2, j), h_2^2 \bar{\varphi}(M-1, j)),$$

$$1 \leq j \leq N-1,$$

$$\mathbf{F}_j = (g(1, j), g(2, j), \dots, g(M-1, j)), \quad j=0, N.$$

Определим также квадратную матрицу C размером $(M-1) \times (M-1)$ следующим образом:

$$\begin{aligned} C \mathbf{V} &= (\Lambda v(1), \Lambda v(2), \dots, \Lambda v(M-1)), \\ \mathbf{V} &= (v(1), v(2), \dots, v(M-1)), \end{aligned}$$

где разностный оператор Λ есть

$$\begin{aligned} \Lambda v(i) &= 2v(i) - h_2^2 v_{x_1 x_1}^-(i), \quad 1 \leq i \leq M-1, \\ v(0) &= v(M) = 0. \end{aligned}$$

Легко видеть, что C есть трехдиагональная матрица вида

$$C = \begin{vmatrix} 2(1+\alpha) & -\alpha & 0 & \dots & 0 & 0 & 0 \\ -\alpha & 2(1+\alpha) & -\alpha & \dots & 0 & 0 & 0 \\ 0 & -\alpha & 2(1+\alpha) & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2(1+\alpha) & -\alpha & 0 \\ 0 & 0 & 0 & \dots & -\alpha & 2(1+\alpha) & -\alpha \\ 0 & 0 & 0 & \dots & 0 & -\alpha & 2(1+\alpha) \end{vmatrix}, \quad (4)$$

где $\alpha = h_2^2/h_1^2$, причем C является матрицей с диагональным преобладанием, так как $|1+\alpha| > |\alpha|$, $\alpha > 0$, и следовательно, не вырождена.

Используя введенные обозначения, полученные выше соотношения можно записать в виде следующей системы трехточечных векторных уравнений:

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, & 1 \leq j \leq N-1, \\ Y_0 &= F_0, & Y_N &= F_N. \end{aligned} \quad (5)$$

Это и есть искомая трехточечная система специального вида с постоянными коэффициентами.

Задача (5) является частным случаем следующей общей задачи: найти векторы Y_j ($0 \leq j \leq N$), удовлетворяющие следующей системе:

$$\begin{aligned} C_0 Y_0 - B_0 Y_1 &= F_0, & j &= 0, \\ -A_j Y_{j-1} + C_j Y_j - B_j Y_{j+1} &= F_j, & 1 \leq j \leq N-1, \\ -A_N Y_{N-1} + C_N Y_N &= F_N, & j &= N, \end{aligned} \quad (6)$$

где Y_j и F_j — векторы размерности M_j , C_j — квадратная матрица $M_j \times M_j$, A_j и B_j — прямоугольные матрицы соответственно размеров $M_j \times M_{j-1}$ и $M_j \times M_{j+1}$.

К системам вида (6) сводятся разностные схемы для эллиптических уравнений второго порядка с переменными коэффициентами в произвольной области любого числа измерений. В двумерном случае, как и в разобранном выше примере, вектор Y_j образуют неизвестные на j -ой строке сетки ω , а в случае трех измерений — неизвестные на j -м слое сетки ω . В последнем случае C_j — блочно-трехдиагональные матрицы с трехдиагональными матрицами на главной диагонали.

Для решения системы (6) мы рассмотрим *метод матричной прогонки*, который аналогичен методу прогонки для скалярных трехточечных уравнений.

2. Прогонка для трехточечных векторных уравнений. Построим метод решения системы трехточечных векторных уравнений (6). Эта система родственна системе скалярных трехточечных уравнений, метод решения которой был изучен нами в § 1. Поэтому решение задачи (6) будем искать в виде

$$Y_j = \alpha_{j+1} Y_{j+1} + \beta_{j+1}, \quad j = N-1, N-2, \dots, 0, \quad (7)$$

где α_{j+1} — неопределенная пока прямоугольная матрица размеров $M_j \times M_{j+1}$, а β_{j+1} — вектор размерности M_j . Из формулы (7) и уравнений системы (6) для $1 \leq j \leq N-1$ находятся (как и в случае обычной прогонки) рекуррентные соотношения для вычисления матриц α_j и векторов β_j . Из (7) и уравнений (6) для $j=0, N$, находятся начальные значения α_1, β_1 и Y_N , позволяющие начать счет по рекуррентным соотношениям. Выпишем окончательные формулы алгоритма предлагаемого метода, который будем называть *методом матричной прогонки*:

$$\alpha_{j+1} = (C_j - A_j \alpha_j)^{-1} B_j, \quad j = 1, 2, \dots, N-1, \quad \alpha_1 = C_0^{-1} B_0, \quad (8)$$

$$\beta_{j+1} = (C_j - A_j \alpha_j)^{-1} (F_j + A_j \beta_j), \quad j = 1, 2, \dots, N, \quad \beta_1 = C_0^{-1} F_0, \quad (9)$$

$$Y_j = \alpha_{j+1} Y_{j+1} + \beta_{j+1}, \quad j = N-1, N-2, \dots, 0, \quad Y_N = \beta_{N+1}. \quad (10)$$

Будем говорить, что алгоритм (8)–(10) *корректен*, если матрицы C_0 и $C_j - A_j \alpha_j$ для $1 \leq j \leq N$ не вырождены. Прежде чем дать определение устойчивости алгоритма (8)–(10), напомним некоторые сведения из линейной алгебры.

Пусть A — произвольная прямоугольная $m \times n$ матрица.

Пусть $\|x\|_n$ есть норма вектора x в n -мерном пространстве H_n . Тогда норма A определяется равенством

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_m}{\|x\|_n}.$$

Очевидно, что норма A определяется как самой матрицей A , так и теми векторными нормами, которые введены в H_n и H_m .

Для случая евклидовых норм в H_n и H_m ($\|x\|_n^2 = \sum_{i=1}^n x_i^2$) имеем $\|A\| = \sqrt{\rho}$, где ρ — максимальное по модулю собственное значение матрицы $A^* A$.

Из определения нормы следует очевидное соотношение $\|Ax\|_m \leq \|A\| \|x\|_n$.

Далее, пусть заданы матрицы A и B соответственно размеров $m \times n$ и $n \times k$. Вводя в H_m, H_k и H_n векторные нормы и определяя с их помощью нормы матриц A, B и AB , получим неравенства $\|AB\| \leq \|A\| \|B\|$.

Будем говорить, что алгоритм *устойчив*, если выполняется оценка $\|\alpha_j\| \leq 1$ для $1 \leq j \leq N$ (предполагается, что в конечномерных пространствах H_m , которым принадлежат векторы Y_j , введена однотипная норма, например евклидова).

Лемма 5. Если C_j для $0 \leq j \leq N$ — невырожденные матрицы, а A_j и B_j — ненулевые матрицы для $1 \leq j \leq N-1$ и выполнены условия

$$\|C_0^{-1} B_0\| \leq 1, \quad \|C_N^{-1} A_N\| \leq 1, \quad \|C_j^{-1} A_j\| + \|C_j^{-1} B_j\| \leq 1, \quad 1 \leq j \leq N-1,$$

причем хотя бы в одном из неравенств имеет место строгое неравенство, то алгоритм метода матричной прогонки (8) — (10) устойчив и корректен.

Приведем только основной этап, оставляя читателю завершение доказательства леммы. Доказательство использует известное утверждение: если для квадратной матрицы S имеет место оценка $\|S\| \leq q < 1$, то существует обратная к $E - S$ матрица, причем $\|(E - S)^{-1}\| \leq 1/(1 - q)$.

Предположим теперь, что $\|\alpha_j\| \leq 1$. Отсюда и из условий леммы будем иметь

$$\|C_j^{-1}A_j\alpha_j\| \leq \|C_j^{-1}A_j\| \leq 1 - \|C_j^{-1}B_j\| < 1.$$

Так как $C_j^{-1}A_j\alpha_j$ квадратная матрица, то, следовательно, существуют обратные к $E - C_j^{-1}A_j\alpha_j$ и к $C_j - A_j\alpha_j$ матрицы, причем $\|(E - C_j^{-1}A_j\alpha_j)^{-1}\| \leq 1/\|C_j^{-1}B_j\|$. Отсюда и из (8) сразу получим

$$\|\alpha_{j+1}\| \leq \|(E - C_j^{-1}A_j\alpha_j)^{-1}C_j^{-1}B_j\| \leq \|(E - C_j^{-1}A_j\alpha_j)^{-1}\| \|C_j^{-1}B_j\| \leq 1.$$

Доказательство завершается по индукции.

Применим лемму 5 к системе трехточечных векторных уравнений (5), полученных из разностной задачи Дирихле для уравнения Пуассона в прямоугольнике. Система (5) есть частный случай (6), где $C_j = C$, $B_j = A_j = E$, $1 \leq j \leq N-1$, $C_0 = C_N = E$, $B_0 = A_N = 0$, а квадратная матрица C задана в (4). Условия леммы 5 для рассматриваемого примера принимают вид $\|C^{-1}\| \leq 0,5$. Для случая евклидовой нормы в силу симметрии C имеем

$$\|C^{-1}\| = \max_k |\lambda_k(C^{-1})| = \frac{1}{\min_k |\lambda_k(C)|},$$

где $\lambda_k(C)$ — собственное значение матрицы C . Из определения C получим, что $\lambda_k(C)$ есть собственное значение определенного выше оператора Λ

$$\begin{aligned} \Lambda v(i) - \lambda_k v(i) &= (2 - \lambda_k)v(i) - h_2^2 v_{x_1 x_1}(i) = 0, \\ v(0) = v(M) &= 0, \quad 1 \leq i \leq M-1. \end{aligned}$$

Эта задача сводится заменой $\lambda_k = 2 + h_2^2 \mu_k$ к рассмотренной в п. 1 § 5 гл. I задаче на собственные значения для простейшего разностного оператора: $v_{x_1 x_1}^- + \mu_k v = 0$, $1 \leq i \leq M-1$, $v(0) = v(M) = 0$. Так как эта задача имеет решение, равное

$$\mu_k = \frac{4}{h_1^2} \sin^2 \frac{k\pi h_1}{2l_1} > 0, \quad k = 1, 2, \dots, M-1,$$

то $\lambda_k = \lambda_k(C) = 2 + h_2^2 \mu_k > 2$. Следовательно, условие $\|C^{-1}\| \leq 0,5$ выполнено. Алгоритм (8) — (10) в применении к решению системы (5) корректен и устойчив.

Рассмотрим теперь вопрос об объеме запоминаемой промежуточной информации и об оценке числа арифметических действий для алгоритма (8) — (10), считая для простоты, что в системе (6) все матрицы квадратные

и имеют размер $M \times M$, а все векторы \mathbf{Y}_j и \mathbf{F}_j имеют размерность M . В этом случае прогоночные коэффициенты α_j будут квадратными матрицами размера $M \times M$, а векторы β_j будут иметь размерность M .

Для реализации (8)–(10) необходимо хранить все матрицы α_j для $1 \leq j \leq N$, все векторы β_j для $1 \leq j \leq N+1$ и матрицу $(C_N - A_N \alpha_N)^{-1}$, используемую для вычисления β_{N+1} . Векторы β_j могут быть расположены на месте, отведенном для векторов неизвестных Y_{j-1} . Для хранения же всех матриц α_j и матрицы $(C_N - A_N \alpha_N)^{-1}$ нужно запоминать $M^2(N+1)$ элементов этих матриц, так как в общем случае матрицы α_j являются полными и несимметричными. Объем дополнительно запоминаемой информации при этом в M раз превышает общее число неизвестных в задаче, которое равно $M(N+1)$.

Оценим теперь число арифметических действий для алгоритма (8)–(10), имея в виду, что для решения серии задач (6) с различными правыми частями \mathbf{F}_j прогоночные матрицы α_j и матрица $(C_N - A_N \alpha_N)^{-1}$ могут быть сосчитаны только один раз, в то время как векторы β_j и Y_j для каждой новой задачи пересчитываются.

В общем случае матрицы $C_j - A_j \alpha_j$ являются для любого j полными матрицами. Поэтому для их обращения потребуется $O(M^3)$ арифметических действий. Далее, умножение $(C_j - A_j \alpha_j)^{-1}$ на матрицу B_j потребует не более $O(M^3)$ операций. Поэтому для вычисления α_{j+1} при заданном α_j по формуле (8) потребуется $O(M^3)$ арифметических действий. Для вычисления всех α_j и матрицы $(C_N - A_N \alpha_N)^{-1}$ потребуется $O(M^3N)$ действий.

Если матрица A_j является полной, то на вычисление β_{j+1} при заданном β_j и вычисленной $(C_j - A_j \alpha_j)^{-1}$ потребуется: $2M^2$ операций умножения и $2M^2 - M$ операций сложения. Если A_j является диагональной матрицей, то это число сокращается — потребуется $M^2 + M$ умножений и $2M^2 - M$ сложений. Следовательно, на вычисление β_j для $2 \leq j \leq N+1$ потребуется в общем случае $2M^2N$ умножений и $(2M^2 - M)N$ сложений. Добавляя сюда операции, затрачиваемые на вычисление β_1 при заданной C_0^{-1} (M^2 умножений и $M^2 - M$ сложений), окончательно получим $M^2(2N+1)$ умножений и $M^2(2N+1) - M(N+1)$ сложений.

Для нахождения всех Y_j для $0 \leq j \leq N-1$ при заданном Y_N потребуется M^2N умножений и M^2N сложений. Итак, для вычисления β_j и Y_j потребуется $M^2(3N+1)$ операций умножения и $M^2(3N+1) - M(N+1)$ операций сложения. Если не делать различия между этими операциями, то это составит $Q \approx 6M^3N$ действий. Именно такое количество арифметических операций нужно затратить, чтобы найти решение каждой новой задачи из серии. Для решения единичной задачи вида (6), когда необходимо вычислять и прогоночные матрицы α_j , потребуется $Q = O(M^3N + M^2N)$ действий.

Пусть серия состоит из n задач вида (6). Тогда потребуется затратить $Q_n = O(M^3N) + 6nM^2N$ действий. При этом общее число неизвестных в серии равно $nM(N+1)$. Отсюда следует, что на нахождение одного неизвестного потребуется $q \approx O\left(\frac{M^2}{n}\right) + 6M$ арифметических операций. Таким образом, с увеличением n относительное число операций на одно неизвестное уменьшается, однако оно всегда больше $6M$. В этом заключается существенное отличие метода матричной прогонки от метода скалярной прогонки, где относительное число действий на одно неизвестное есть конечная величина, не зависящая от числа неизвестных.

3. Прогонка для двухточечных векторных уравнений. Рассмотрим теперь метод решения двухточечных векторных уравнений

$$\begin{aligned} P_{i+1}V_{i+1} - Q_i V_i &= F_{i+1}, & 0 \leq i \leq N-1, \\ P_0 V_0 &= F_0, \quad Q_N V_N = F_{N+1}, \end{aligned} \tag{11}$$

где V_i — вектор размерности M , P_{i+1} и Q_i для $0 \leq i \leq N-1$ квадратные матрицы $M \times M$, P_0 и Q_N — прямоугольные матрицы

соответственно размеров $M_1 \times M$ и $M_2 \times M$, $M_1 + M_2 = M$. Вектор \mathbf{F}_{i+i} для $0 \leq i \leq N-1$ имеет размерность M , а \mathbf{F}_0 и $\mathbf{F}_{N+i} - M_i$ и M_2 соответственно.

Сначала сведем систему (11) к виду (6). Для этого представим матрицы, входящие в (11), следующим образом:

$$\begin{aligned} P_0 &= \|P_0^{11}|P_0^{12}\|, \quad Q_N = \|Q_N^{21}|Q_N^{22}\|, \\ P_{i+i} &= \begin{cases} \|P_{i+1}^{11}|P_{i+1}^{12}\|, \\ \|P_{i+1}^{21}|P_{i+1}^{22}\|, \end{cases}, \quad Q_i = \begin{cases} \|Q_i^{11}|Q_i^{12}\|, \\ \|Q_i^{21}|Q_i^{22}\|, \end{cases}, \end{aligned} \quad (12)$$

где P_i^{kl} и Q_i^{kl} для $0 \leq i \leq N$ — матрицы размеров $M_k \times M_l$, $k, l = 1, 2$. В соответствии с представлением (12) положим

$$\begin{aligned} \mathbf{V}_i &= \begin{pmatrix} \mathbf{v}_i^1 \\ \mathbf{v}_i^2 \end{pmatrix}, \quad 0 \leq i \leq N, \quad \mathbf{F}_{i+i} = \begin{pmatrix} \mathbf{f}_{i+1}^1 \\ \mathbf{f}_{i+1}^2 \end{pmatrix}, \quad 0 \leq i \leq N-1, \\ \mathbf{F}_0 &= \mathbf{f}_0^1, \quad \mathbf{F}_{N+i} = \mathbf{f}_{N+i}, \end{aligned} \quad (13)$$

где \mathbf{v}_i^k и \mathbf{f}_i^k — векторы размерности M_k , $k = 1, 2$. Используя (12) и (13), запишем систему (11) в следующем виде:

$$\left. \begin{aligned} P_0^{11}\mathbf{v}_0^1 - P_0^{12}\mathbf{v}_0^2 &= \mathbf{f}_0^1, \\ -Q_i^{11}\mathbf{v}_i^1 + Q_i^{12}\mathbf{v}_i^2 + P_{i+1}^{11}\mathbf{v}_{i+1}^1 - P_{i+1}^{12}\mathbf{v}_{i+1}^2 &= \mathbf{f}_{i+1}^1, \\ -Q_i^{21}\mathbf{v}_i^1 + Q_i^{22}\mathbf{v}_i^2 + P_{i+1}^{21}\mathbf{v}_{i+1}^1 - P_{i+1}^{22}\mathbf{v}_{i+1}^2 &= \mathbf{f}_{i+1}^2, \\ -Q_N^{21}\mathbf{v}_N^1 + Q_N^{22}\mathbf{v}_N^2 &= \mathbf{f}_{N+i}^2. \end{aligned} \right\} 0 \leq i \leq N-1, \quad (14)$$

Введем теперь новые векторы неизвестных, полагая

$$\mathbf{Y}_0 = \mathbf{v}_0^1, \quad \mathbf{Y}_{N+i} = \mathbf{v}_N^2, \quad \mathbf{Y}_{i+i} = \begin{pmatrix} \mathbf{v}_i^2 \\ \mathbf{v}_{i+1}^1 \end{pmatrix}, \quad 0 \leq i \leq N-1,$$

и матрицы

$$\begin{aligned} C_0 &= P_0^{11}, \quad B_0 = \|P_0^{12}|0^{11}\|, \quad C_{N+i} = Q_N^{22}, \quad A_{N+i} = \|0^{22}|Q_N^{21}\|, \\ A_1 &= \begin{cases} \|Q_0^{11}\|, \\ \|Q_0^{21}\|, \end{cases}, \quad B_N = \begin{cases} \|P_N^{12}\|, \\ \|P_N^{22}\|, \end{cases}, \quad A_{i+1} = \begin{cases} \|0^{12}|Q_i^{11}\|, \\ \|0^{22}|Q_i^{21}\|, \end{cases}, \quad 1 \leq i \leq N-1, \\ B_{i+i} &= \begin{cases} \|P_{i+1}^{12}|0^{11}\|, \\ \|P_{i+1}^{22}|0^{21}\|, \end{cases}, \quad 0 \leq i \leq N-2, \quad C_{i+i} = \begin{cases} \|Q_i^{12}|P_{i+1}^{11}\|, \\ \|Q_i^{22}|P_{i+1}^{21}\|, \end{cases}, \quad 0 \leq i \leq N-1, \end{aligned}$$

где 0^{kl} — нулевая матрица размеров $M_k \times M_l$, $k, l = 1, 2$.

В этих обозначениях система (14) будет иметь вид

$$\begin{aligned} C_0 \mathbf{Y}_0 - B_0 \mathbf{Y}_1 &= \mathbf{F}_0, & i = 0, \\ -A_i \mathbf{Y}_{i-1} + C_i \mathbf{Y}_i - B_i \mathbf{Y}_{i+1} &= \mathbf{F}_i, & 1 \leq i \leq N, \\ -A_{N+i} \mathbf{Y}_N + C_{N+i} \mathbf{Y}_{N+i} &= \mathbf{F}_{N+i}, & i = N+1. \end{aligned} \quad (15)$$

Итак, система двухточечных векторных уравнений (11) сведена к системе трехточечных векторных уравнений вида (15), метод матричной прогонки для которой построен в п. 2. Для (15)

алгоритм матричной прогонки имеет следующий вид:

$$\alpha_{i+1} = (C_i - A_i \alpha_i)^{-1} B_i, \quad i = 1, 2, \dots, N, \quad \alpha_1 = C_0^{-1} B_0, \quad (16)$$

$$\beta_{i+1} = (C_i - A_i \alpha_i)^{-1} (F_i + A_i \beta_i), \quad i = 1, 2, \dots, N+1, \quad \beta_1 = C_0^{-1} F_0, \quad (17)$$

$$Y_i = \alpha_{i+1} Y_{i+1} + \beta_{i+1}, \quad i = N, N-1, \dots, 0, \quad Y_{N+1} = \beta_{N+2}, \quad (18)$$

причем матрицы α_1 и α_{N+1} имеют размер $M_1 \times M$ и $M \times M_2$ соответственно, а α_i для $2 \leq i \leq N$ являются квадратными матрицами размера $M \times M$. Векторы β_i для $2 \leq i \leq N+1$ имеют размерность M , а β_1 и β_{N+2} — размерность M_1 и M_2 .

Преобразуем формулы (16)–(18). Учитывая структуру матриц B_i , находим, что матрицы α_i имеют вид

$$\alpha_1 = \left\| \alpha_1^{12} \mid 0^{11} \right\|, \quad \alpha_{N+1} = \begin{pmatrix} \alpha_{N+1}^{22} \\ \alpha_{N+1}^{12} \end{pmatrix}, \quad \alpha_i = \begin{pmatrix} \alpha_i^{22} \mid 0^{21} \\ \alpha_i^{12} \mid 0^{11} \end{pmatrix}, \quad 2 \leq i \leq N. \quad (19)$$

Подставляя (19) в (16) и учитывая определение матриц A_i , B_i и C_i , получим формулы для вычисления α_i^{12} и α_i^{22}

$$\begin{pmatrix} \alpha_{i+1}^{22} \\ \alpha_{i+1}^{12} \end{pmatrix} = \begin{pmatrix} Q_{i-1}^{12} - Q_{i-1}^{11} \alpha_i^{12} \mid P_i^{11} \\ Q_{i-1}^{22} - Q_{i-1}^{21} \alpha_i^{12} \mid P_i^{21} \end{pmatrix}^{-1} \begin{pmatrix} P_i^{12} \\ P_i^{22} \end{pmatrix}, \quad 1 \leq i \leq N, \quad (20)$$

где $\alpha_1^{12} = (P_0^{11})^{-1} P_0^{12}$. Далее, представляя вектор β_i в виде

$$\beta_i = \beta_1^1, \quad \beta_{N+2} = \beta_{N+2}^2, \quad \beta_i = \begin{pmatrix} \beta_i^2 \\ \beta_i^1 \end{pmatrix}, \quad 2 \leq i \leq N+1 \quad (21)$$

и подставляя это выражение в (17), получим

$$\begin{pmatrix} \beta_{i+1}^2 \\ \beta_{i+1}^1 \end{pmatrix} = \begin{pmatrix} Q_{i-1}^{12} - Q_{i-1}^{11} \alpha_i^{12} \mid P_i^{11} \\ Q_{i-1}^{22} - Q_{i-1}^{21} \alpha_i^{12} \mid P_i^{21} \end{pmatrix}^{-1} \begin{pmatrix} f_i^1 + Q_{i-1}^{11} \beta_i^1 \\ f_i^2 + Q_{i-1}^{21} \beta_i^1 \end{pmatrix}, \quad 1 \leq i \leq N, \quad (22)$$

$$\beta_{N+2}^2 = Q_N^{22} - Q_N^{21} \alpha_{N+1}^{12} \mid P_{N+1}^{21} \mid^{-1} (f_{N+1}^2 + Q_N^{21} \beta_{N+1}^1), \quad (23)$$

где $\beta_1^1 = \|P_0^{11}\|^{-1} f_0^1$.

Подставим теперь (19) и (21) в (18) и используем введенные обозначения для Y_i . В результате получим следующие формулы для вычисления компонент вектора неизвестных:

$$\begin{aligned} v_{i-1}^2 &= \alpha_{i+1}^{22} v_i^2 + \beta_{i+1}^2, & i = N, N-1, \dots, 1, & v_N^2 = \beta_{N+2}^2, \\ v_i^1 &= \alpha_{i+1}^{12} v_i^2 + \beta_{i+1}^1, & i = N, N-1, \dots, 0. \end{aligned} \quad (24)$$

Итак, алгоритм метода матричной прогонки для системы двухточечных векторных уравнений (11) описывается формулами (20), (22)–(24).

Так как эти формулы являются следствием из алгоритма прогонки решения системы (15), к которой мы свели исходную систему двухточечных векторных уравнений (11), то достаточные условия корректности и устойчивости полученного алгоритма сформулированы в лемме 5, где нужно заменить N на $N+1$, а матрицы C_i , A_i и B_i определены выше.

Используя алгоритм встречных прогонок для системы (15), можно построить соответствующий алгоритм для исходной системы двухточечных векторных уравнений (11).

4. Ортогональная прогонка для двухточечных векторных уравнений. Рассмотрим еще один метод решения системы двухточечных уравнений (11), известный под названием *метода ортогональной прогонки*. Этот метод содержит обращение матриц P_i для $1 \leq i \leq N$ и ортогонализацию вспомогательных прямоугольных матриц.

Будем искать решение системы (11) в следующем виде:

$$V_i = B_i \beta_i + Y_i, \quad 0 \leq i \leq N, \quad (25)$$

где B_i для любого i — прямоугольная матрица размера $M \times M_2$, а β_i и Y_i — векторы размерности M_2 и M соответственно.

Определяя B_0 и Y_0 из условия $P_0 B_0 = 0^{12}$, $P_0 Y_0 = F_0$, где 0^{12} — нулевая матрица размера $M_1 \times M_2$, получим, что V_0 удовлетворяет условию $P_0 V_0 = F_0$. Найдем теперь рекуррентные формулы для последовательного построения, исходя из B_0 и Y_0 , матриц B_i и векторов Y_i .

Подставим (25) в (11). Если P_{i+1} — невырожденные матрицы, то будем иметь

$$B_{i+1} \beta_{i+1} + Y_{i+1} - P_{i+1}^{-1} Q_i B_i \beta_i = P_{i+1}^{-1} (F_{i+1} + Q_i Y_i), \quad 0 \leq i \leq N-1,$$

или

$$B_{i+1} \beta_{i+1} + Y_{i+1} - A_{i+1} \beta_i = X_{i+1}, \quad 0 \leq i \leq N-1, \quad (26)$$

где $A_{i+1} = P_{i+1}^{-1} Q_i B_i$, $X_{i+1} = P_{i+1}^{-1} (F_{i+1} + Q_i Y_i)$. Матрица A_{i+1} имеет размер $M \times M_2$, а вектор X_{i+1} — размерность M .

Определим B_{i+1} и Y_{i+1} следующим образом:

$$A_{i+1} = B_{i+1} \Omega_{i+1}, \quad Y_{i+1} = X_{i+1} - B_{i+1} \Phi_{i+1}, \quad (27)$$

где Ω_{i+1} и Φ_{i+1} — неопределенные пока квадратная матрица $M_2 \times M_2$ и вектор размерности M_2 . Подставляя (27) в (26), получим соотношение $B_{i+1} (\beta_{i+1} - \Omega_{i+1} \beta_i) = B_{i+1} \Phi_{i+1}$, которое превращается в тождество, если положить

$$\Omega_{i+1} \beta_i = \beta_{i+1} - \Phi_{i+1}, \quad 0 \leq i \leq N-1. \quad (28)$$

Итак, если заданы невырожденные матрицы Ω_i для $1 \leq i \leq N$ и векторы Φ_i для тех же i , то по формулам (27) можно найти, исходя из B_0 и Y_0 , все необходимые матрицы B_i и векторы Y_i для $1 \leq i \leq N$.

Осталось определить векторы β_i . Из (25) при $i = N$ и системы (11) получим два соотношения $V_N = B_N \beta_N + Y_N$, $Q_N V_N = F_{N+1}$ с известными B_N и Y_N . Отсюда для β_N найдем уравнение $Q_N B_N \beta_N = F_{N+1} - Q_N Y_N$ с квадратной матрицей $Q_N B_N$ размера $M_2 \times M_2$. Это соотношение можно записать в виде (28)

$$Q_{N+1} \beta_N = \beta_{N+1} - \Phi_{N+1}, \quad (29)$$

где $\beta_{N+1} = F_{N+1}$, $\Phi_{N+1} = Q_N Y_N$, $\Omega_{N+1} = Q_N B_N$.

Если матрица Ω_{N+1} не вырождена, по формулам (28), (29) последовательно, начиная с β_{N+1} , найдем все β_i для $0 \leq i \leq N$. Решение системы (11) может быть тогда найдено по формулам (25).

Так как имеется произвол в выборе матриц Ω_i и векторов Φ_i , то приведенные выше формулы описывают скорее принцип построения методов решения системы (11), нежели конкретный алгоритм. Выбор определенных Ω_i и Φ_i порождает некоторый метод для системы (11). Такие методы мы будем называть по-прежнему прогонкой, на прямом ходе которой вычисляются B_i и Y_i , а на обратном — β_i и решение V_i .

Остановимся теперь на одном способе выбора Ω_i и Φ_i . Так как формулы (27) и (28) предполагают обращение матрицы Ω_{i+1} , то она должна быть достаточно легко обратима.

В рассматриваемом методе ортогональной прогонки матрица Ω_{i+1} и вектор Φ_{i+1} порождаются требованиями: 1) матрица B_{i+1} строится путем ортонормирования столбцов матрицы A_{i+1} ; 2) вектор Y_{i+1} должен быть ортогонален столбцам матрицы B_{i+1} .

Следствием этих требований являются равенства

$$B_{i+1}^* B_{i+1} = E^{22}, \quad B_{i+1}^* Y_{i+1} = 0, \quad (29')$$

где B_{i+1}^* — транспонированная к B_{i+1} матрица, а E^{22} — единичная матрица размера $M_2 \times M_2$.

Найдем сначала выражение для Φ_{i+1} . Из (27) и (29') получим $0 = B_{i+1}^* Y_{i+1} = B_{i+1}^* X_{i+1} - B_{i+1}^* B_{i+1} \Phi_{i+1} = B_{i+1}^* X_{i+1} - \Phi_{i+1}$. Итак, вектор Φ_{i+1} определен: $\Phi_{i+1} = B_{i+1}^* X_{i+1}$.

Построим теперь матрицы Ω_{i+1} и B_{i+1} . Существует несколько способов ортонормирования столбцов матрицы A_{i+1} . Мы рассмотрим метод Грама — Шмидта.

Пусть матрица A_{i+1} имеет ранг M_2 . Обозначим через a_k и b_k — k -е столбцы матриц A_{i+1} и B_{i+1} соответственно, а через (\cdot, \cdot) — скалярное произведение векторов. В качестве b_1 возьмем нормированный столбец a_1

$$b_1 = a_1 / \omega_{11}, \quad \omega_{11} = \sqrt{(a_1, a_1)}. \quad (30)$$

Далее будем искать столбец b_k в виде

$$b_k = \frac{1}{\omega_{kk}} \left(a_k - \sum_{n=1}^{k-1} \omega_{nk} b_n \right), \quad 2 \leq k \leq M_2, \quad (31)$$

где коэффициенты ω_{nk} находятся из условия ортогональности вектора b_k векторам b_1, b_2, \dots, b_{k-1} , а ω_{kk} — из условия нормировки b_k :

$$\omega_{nk} = (b_n, a_k), \quad n = 1, 2, \dots, k-1, \quad \omega_{kk} = \sqrt{(a_k, a_k) - \sum_{n=1}^{k-1} \omega_{nk}^2}. \quad (32)$$

В силу сделанного предположения о ранге матрицы A_{i+1} столбцы a_k для $1 \leq k \leq M_2$ линейно независимы, и процесс ортонормирования протекает без особенностей.

Из (30) — (32) следует, что матрицы A_{i+1} и B_{i+1} связаны соотношением $A_{i+1} = B_{i+1}\Omega_{i+1}$, где Ω_{i+1} — квадратная верхнетреугольная матрица размера $M_2 \times M_2$ с элементами ω_{nk} для $1 \leq n \leq M_2$, $n \leq k \leq M_2$, определенными в (30) и (32), и $\omega_{nk} = 0$ для $k < n$.

Итак, формулы (30) — (32) определяют матрицы B_{i+1} и Ω_{i+1} . Несложный подсчет показывает, что построение матриц B_{i+1} и Ω_{i+1} можно осуществить, затратив: $MM_2^2 + 0,5(M_2^2 - M_2)$ операций умножения, $MM_2^2 - M_2$ операций сложения и вычитания, MM_2 операций деления и M_2 извлечений квадратного корня. Всю указанную процедуру ортонормирования необходимо осуществить N раз на прямом ходе метода прогонки. Это потребует $O(MNM_2^2)$ арифметических операций и NM_2 операций извлечения квадратного корня.

Нам осталось указать, как находятся матрица B_0 и вектор \mathbf{Y}_0 . Будем предполагать, что матрицы P_{i+1} и Q_i для $0 \leq i \leq N-1$ не вырождены. Кроме того, пусть не вырождена матрица P_0^{11} , а матрица Q_N имеет ранг M_2 .

Построим B_0 и \mathbf{Y}_0 . Пусть

$$A_0 = \left\| \frac{(P_0^{11})^{-1} P_0^{12}}{E^{22}} \right\|, \quad \mathbf{X}_0 = \begin{pmatrix} (P_0^{11})^{-1} \mathbf{F}_0 \\ 0 \end{pmatrix}$$

— прямоугольная матрица размера $M \times M_2$ и вектор размерности M . Так как размерность квадратной единичной матрицы E^{22} есть $M_2 \times M_2$, то ранг A_0 равен M_2 . Матрица B_0 строится, исходя из A_0 , при помощи процесса ортонормирования (30) — (32), а \mathbf{Y}_0 выбирается по формуле $\mathbf{Y}_0 = \mathbf{X}_0 - B_0 \Phi_0$ из условия ортогональности столбцов матрицы B_0 , что дает $\Phi_0 = B_0^* \mathbf{X}_0$. Так как

$$B_0 = A_0 \Omega_0^{-1}, \quad P_0 A_0 = \|P_0^{11}| - P_0^{12}\| \left\| \frac{(P_0^{11})^{-1} P_0^{12}}{E^{22}} \right\| = \|0^{12}\|,$$

то $P_0 B_0 = 0^{12}$. Далее имеем

$$P_0 \mathbf{Y}_0 = P_0 \mathbf{X}_0 - P_0 B_0 \Phi_0 = P_0 \mathbf{X}_0 = \mathbf{F}_0.$$

Таким образом, построенные B_0 и \mathbf{Y}_0 удовлетворяют требуемым соотношениям: $P_0 B_0 = 0^{12}$ и $P_0 \mathbf{Y}_0 = \mathbf{F}_0$.

Заметим, что в силу невырожденности P_{i+1} и Q_i ранг матрицы A_{i+1} совпадает с рангом B_i . Кроме того, в силу невырожденности Ω_0 ранг B_0 совпадает с рангом A_0 и равен M_2 . Поэтому процесс ортонормирования (30) — (32) будет протекать без осложнений. Далее, так как ранги матриц Q_N и B_N равны M_2 , то квадратная матрица $\Omega_{N+1} = Q_N B_N$ будет невырожденной, что позволит найти вектор Φ_N .

Таким образом, алгоритм метода ортогональной прогонки имеет следующий вид:

$$1) \quad B_i \Omega_i = A_i, \quad i = 0, 1, 2, \dots, N,$$

$$A_i = P_i^{-1} Q_{i-1} B_{i-1}, \quad 1 \leq i \leq N, \quad A_0 = \left\| \frac{(P_0^{11})^{-1} P_0^{12}}{E^{22}} \right\|. \quad (33)$$

Матрицы B_i и Ω_i для $0 \leq i \leq N$ вычисляются по формулам (30) — (32) и запоминаются. Полагается $\Omega_{N+1} = Q_N B_N$.

$$2) \quad Y_i = X_i - B_i \Phi_i, \quad \Phi_i = B_i^* X_i, \quad i = 0, 1, \dots, N,$$

$$X_i = P_i^{-1} (F_i + Q_{i-1} Y_{i-1}), \quad 1 \leq i \leq N, \quad X_0 = \begin{pmatrix} (P_0^{11})^{-1} F_0 \\ 0 \end{pmatrix}. \quad (34)$$

Вычисляются и запоминаются векторы Y_i для $0 \leq i \leq N$ и Φ_i для $1 \leq i \leq N$. Полагается $\Phi_{N+1} = Q_N Y_N$.

$$3) \quad \Omega_{i+1} \beta_i = \beta_{i+1} - \Phi_{i+1}, \quad i = N, N-1, \dots, 0, \quad \beta_{N+1} = F_{N+1}, \\ V_i = B_i \beta_i + Y_i, \quad 0 \leq i \leq N. \quad (35)$$

Замечание. Так как матрицы Ω_i для $1 \leq i \leq N$ являются верхнетреугольными матрицами размера $M_2 \times M_2$, то для нахождения β_i по заданным β_{i+1} и Φ_{i+1} требуется $O(M_2^2)$ действий.

Для иллюстрации предложенного алгоритма рассмотрим пример. Пусть требуется решить следующую трехточечную разностную задачу:

$$\begin{aligned} -y_{i-1} + y_i - y_{i+1} &= 0, & 1 \leq i \leq N-1, \\ y_0 &= 1, \quad y_N &= 0. \end{aligned} \quad (36)$$

Эта задача рассматривалась нами ранее в п. 4. § 2, где методом немонотонной трехточечной прогонки было найдено ее решение для N , не кратных 3, а именно,

$$y_i = \frac{\sin \frac{(N-i)\pi}{3}}{\sin \frac{N\pi}{3}}, \quad 0 \leq i \leq N.$$

Сведем систему (36) к системе двухточечных векторных уравнений вида (11), полагая

$$V_i = \begin{pmatrix} y_i \\ y_{i+1} \end{pmatrix}, \quad 0 \leq i \leq N-1.$$

Несложно видеть, что (36) эквивалентна следующей системе:

$$\begin{aligned} V_{i+1} - Q V_i &= 0, & 0 \leq i \leq N-2, \\ P_0 V_0 &= 1, \quad Q_{N-1} V_{N-1} &= 0, \end{aligned} \quad (37)$$

где $P_0 = \|1|0\|$, $Q_{N-1} = \|0|1\|$, $Q = \begin{vmatrix} 0 & 1 \\ -1 & 1 \end{vmatrix}$. Система (37) есть частный случай (11) с $M_1 = M_2 = 1$, $M = 2$.

Для решения (37) используем алгоритм ортогональной прогонки (33) — (35). Для рассматриваемого примера матрицы B_i имеют размерность 2×1 , Ω_i — размерность 1×1 , векторы Y_i будут иметь размерность 2, а векторы β_i и Φ_i — размерность 1.

В табл. 3 приведены матрицы B_i и Ω_i , а также векторы Y_i , Φ_i и β_i для $N = 11$. Применяемый метод ортогональной прогонки позволяет получить точное решение y_i задачи (36).

Таблица 3

	6	0	1	2	3	4	5	6	7	8	9	10	11	
Ω_i	1	$\sqrt{2}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$-\frac{1}{\sqrt{2}}$	
Ψ_i	0	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	1	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	$-\frac{1}{2}$	1	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	1	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	
β_i	1	$\frac{1}{\sqrt{2}}$	0	1	$-\frac{1}{\sqrt{2}}$	0	$-\frac{1}{\sqrt{2}}$	0	$\frac{1}{\sqrt{2}}$	0	0	$\frac{1}{\sqrt{2}}$	0	
B_i	(0)	$\left(\begin{array}{cc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{array}\right)$	$\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)$	(0)	$\left(\begin{array}{cc} -1 & \frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} \end{array}\right)$	$\left(\begin{array}{cc} -1 & 0 \\ 0 & 1 \end{array}\right)$	$\left(\begin{array}{cc} 0 & \frac{1}{\sqrt{2}} \\ 1 & -\frac{1}{\sqrt{2}} \end{array}\right)$	(0)	$\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)$	$\left(\begin{array}{cc} 0 & \frac{1}{\sqrt{2}} \\ 1 & -\frac{1}{\sqrt{2}} \end{array}\right)$	$\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)$	$\left(\begin{array}{cc} -1 & 0 \\ 0 & 1 \end{array}\right)$	$\left(\begin{array}{cc} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{array}\right)$	
Y_i	1	0	$\left(\begin{array}{cc} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{array}\right)$	$\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)$	$\left(\begin{array}{cc} 0 & -1 \\ -1 & 0 \end{array}\right)$	$\left(\begin{array}{cc} -1 & 0 \\ 0 & 1 \end{array}\right)$	$\left(\begin{array}{cc} -\frac{1}{2} & 1 \\ 1 & \frac{1}{2} \end{array}\right)$	$\left(\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right)$	$\left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right)$	$\left(\begin{array}{cc} 0 & -1 \\ -1 & 0 \end{array}\right)$	$\left(\begin{array}{cc} -1 & 0 \\ 0 & 1 \end{array}\right)$	$\left(\begin{array}{cc} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{array}\right)$	0	
y_i	1	1	0	-1	-1	-1	0	1	1	0	-1	-1	0	

5. Прогонка для трехточечных уравнений с постоянными коэффициентами. Обратимся снова к методу матричной прогонки для трехточечных уравнений и рассмотрим частный случай таких уравнений, а именно:

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N-1, \\ Y_0 &= F_0, \quad Y_N = F_N, \end{aligned} \quad (38)$$

где C — квадратная матрица размера $M \times M$, а Y_j и F_j — искомый и заданный векторы размерности M .

В п. 1 было показано, что к системе трехточечных уравнений вида (38) сводится разностная задача Дирихле для уравнения Пуассона на прямоугольной сетке, заданной в прямоугольнике, причем матрица C будет симметричной и трехдиагональной. Далее, в п. 2 § 4 было показано, что метод матричной прогонки, имеющий для (38) вид

$$\alpha_{j+1} = (C - \alpha_j)^{-1}, \quad j = 1, 2, \dots, N-1, \quad \alpha_1 = 0, \quad (39)$$

$$\beta_{j+1} = \alpha_{j+1} (F_j + \beta_j), \quad j = 1, 2, \dots, N-1, \quad \beta_1 = F_0, \quad (40)$$

$$Y_j = \alpha_{j+1} Y_{j+1} + \beta_{j+1}, \quad j = N-1, N-2, \dots, 1, \quad Y_N = F_N, \quad (41)$$

является корректным и устойчивым. Там же было показано, что собственные значения матрицы C больше 2:

$$\lambda_k = \lambda_k(C) = 2 + 4 \frac{h_2^2}{h_1^2} \sin^2 \frac{k\pi h_1}{2l_1} > 2. \quad (42)$$

Напомним, что в случае общих трехточечных векторных уравнений для алгоритма матричной прогонки требуется $O(M^3N)$ арифметических действий для вычисления матриц α_j и $O(M^2N)$ действий для вычисления прогоночных векторов β_j и решения Y_j . Для хранения полных и, вообще говоря, несимметричных матриц α_j необходимо запомнить $M^2(N+1)$ элементов этих матриц. Уменьшаются ли эти величины, если методом матричной прогонки решать специальную трехточечную векторную систему (38) с постоянными коэффициентами?

Для рассматриваемого примера все матрицы α_j будут симметричными в силу симметрии матрицы C , но хотя C есть трехдиагональная матрица, все матрицы α_j , $j \geq 2$, будут полными. Следовательно, можно уменьшить, учитывая симметрию матриц α_j , только объем промежуточной запоминаемой информации, но не более чем вдвое. Порядок числа арифметических действий по M и N не изменится.

Построим теперь модификацию алгоритма (39) — (41), которая не требует дополнительной памяти для хранения промежуточной информации и реализуется с затратой $O(MN^2)$ арифметических действий, если решается задача (38) с трехдиагональной матрицей C .

Сначала найдем явный вид прогоночных матриц α_j для любого j . Для этого, используя (39), выразим α_j через матрицу C .

Замечая, что

$$\alpha_1 = 0, \quad \alpha_2 = C^{-1}, \quad \alpha_3 = (C^2 - E)^{-1}C, \quad (43)$$

будем искать решение нелинейного разностного уравнения (39) в виде

$$\alpha_j = P_{j-1}^{-1}(C)P_{j-2}(C), \quad j \geq 2, \quad (44)$$

где $P_j(C)$ — полином от C степени j . Перепишем (39) в виде

$$\alpha_{j+1}(C - \alpha_j) = E, \quad j \geq 2,$$

и подставим сюда (44). Получим рекуррентное соотношение $P_j(C) = CP_{j-1}(C) - P_{j-2}(C)$, $j \geq 2$, или после сдвига индекса на единицу и учета (43)

$$\begin{aligned} P_{j+1}(C) &= CP_j(C) - P_{j-1}(C), & j \geq 1, \\ P_0(C) &= E, \quad P_1(C) = C. \end{aligned} \quad (45)$$

Итак, формулы (45) полностью определяют полином $P_j(C)$ для любого $j \geq 0$.

Найдем решение (45). Соответствующий алгебраический полином удовлетворяет соотношениям

$$\begin{aligned} P_{j+1}(t) &= tP_j(t) - P_{j-1}(t), & j \geq 1, \\ P_0(t) &= 1, \quad P_1(t) = t, \end{aligned}$$

которые представляют собой задачу Коши для трехточечного разностного уравнения с постоянными коэффициентами. В п. 2 § 4 гл. I было найдено решение этой задачи $P_j(t) = U_j\left(\frac{t}{2}\right)$, $j \geq 0$, где $U_j(x)$ — полином Чебышева второго рода степени j

$$U_j(x) = \begin{cases} \frac{\sin((j+1)\arccos x)}{\sin \arccos x}, & |x| \leq 1, \\ \frac{\operatorname{sh}((j+1)\operatorname{Arch} x)}{\operatorname{sh} \operatorname{Arch} x}, & |x| \geq 1. \end{cases}$$

Таким образом, явное выражение для прогоночных матриц α_j найдено:

$$\alpha_j = U_{j-1}^{-1}\left(\frac{C}{2}\right)U_{j-2}\left(\frac{C}{2}\right), \quad j \geq 2, \quad \alpha_1 = 0. \quad (46)$$

Это избавляет нас от необходимости проводить вычисления по формуле (39) прогоночных матриц α_j , на что требуется основной объем вычислительной работы в алгоритме (39) — (41). Кроме того, матрицы α_j нет необходимости запоминать.

Рассмотрим теперь формулы (40) и (41). Они содержат умножение матрицы α_{j+1} на векторы $F_j + \beta_j$ и Y_{j+1} . Покажем сейчас, как можно, не вычисляя α_j по формуле (46), определить произведение матрицы α_j на вектор. Для этого нам потребуется лемма 6, которую мы приведем без доказательства.

Лемма 6. Пусть многочлен $f_n(x)$ степени n имеет простые корни. Отношение многочлена $g_m(x)$ степени m к многочлену $f_n(x)$ степени $n > m$ без общих корней может быть представлено в виде суммы n элементарных дробей

$$\frac{g_m(x)}{f_n(x)} = \sum_{l=1}^n \frac{a_l}{x-x_l}, \quad a_l = \frac{g_m(x_l)}{f'_n(x_l)},$$

где x_l — корни $f_n(x)$, a_l — производная полинома $f_n(x)$.

Используя лемму 6, найдем разложение на простые дроби отношения $\varphi(x) = \frac{U_{j-2}(x)}{U_{j-1}(x)}$, $j \geq 2$. Так как корни $U_{j-1}(x)$ есть

$$x_k = \cos \frac{k\pi}{j}, \quad k = 1, 2, \dots, j-1,$$

а

$$U_{j-2}(x_k) = (-1)^{k-1}, \quad \frac{d}{dx}[U_{j-1}(x_k)] = \frac{j(-1)^{k-1}}{\sin^2 \frac{k\pi}{j}},$$

то в силу леммы 6 имеем следующее разложение для $\varphi(x)$:

$$\varphi(x) = \frac{U_{j-2}(x)}{U_{j-1}(x)} = \sum_{k=1}^{j-1} \frac{\sin^2 \frac{k\pi}{j}}{j} \left(x - \cos \frac{k\pi}{j} \right)^{-1}. \quad (47)$$

Из (46) и (47) следует еще одно представление для матриц α_j , которым мы и будем пользоваться

$$\alpha_j = \sum_{k=1}^{j-1} a_{kj} \left(C - 2 \cos \frac{k\pi}{j} E \right)^{-1}, \quad a_{kj} = \frac{2 \sin^2 \frac{k\pi}{j}}{j}, \quad j \geq 2. \quad (48)$$

Используя (48), умножение матрицы α_j на вектор Y можно осуществить по следующему алгоритму: для $k = 1, 2, \dots, j-1$ решаются уравнения

$$\left(C - 2 \cos \frac{k\pi}{j} E \right) V_k = a_{kj} Y, \quad (49)$$

где a_{kj} определено в (48), а результат $\alpha_j Y$ получается последовательным суммированием векторов V_k

$$\alpha_j Y = \sum_{k=1}^{j-1} V_k. \quad (50)$$

Заметим, что в силу (42) матрица $C - 2 \cos \frac{k\pi}{j} E$ является невырожденной и, кроме того, трехдиагональной, если таковой была матрица C . В этом случае каждое из уравнений (49) решается за $O(M)$ арифметических действий методом скалярной

трехточечной прогонки, описанным в § 1. Следовательно, на решение всех задач (49), а также на вычисление суммы (50) потребуется $O(Mj)$ действий. Так как в (40) и (41) умножение матрицы α_j на векторы осуществляется для $j=2, 3, \dots, N$, то модифицированный метод матричной прогонки (40), (41) и (49), (50) требует $O(MN^2)$ арифметических действий.

Итак, построен *модифицированный метод матричной прогонки*, позволяющий найти решение разностной задачи Дирихле для уравнения Пуассона в прямоугольнике с затратой $O(MN^2)$ арифметических действий. Уменьшение числа действий по сравнению с исходным алгоритмом (39) — (41) достигнуто за счет учета специфики решаемой задачи.

В последующих двух главах мы рассмотрим другие прямые методы решения указанной задачи и ей подобных разностных задач, которые будут требовать еще меньшего числа действий, чем построенный здесь метод.

ГЛАВА III

МЕТОД ПОЛНОЙ РЕДУКЦИИ

В данной главе изучается метод решения специальных сеточных эллиптических уравнений — метод полной редукции. Этот прямой метод позволяет найти решение разностной задачи Дирихле для уравнения Пуассона в прямоугольнике за $O(N^2 \log_2 N)$ арифметических действий, где N — число узлов сетки по каждому направлению.

В § 1 дана постановка краевых задач для разностных уравнений, для решения которых можно использовать метод редукции. В § 2 изложен алгоритм метода для случая первой краевой задачи, а в § 3 рассмотрены примеры применения метода. В § 4 дано обобщение метода на случай общих краевых условий.

§ 1. Краевые задачи для трехточечных векторных уравнений

1. Постановка краевых задач. В главе II для решения трехточечных скалярных и векторных уравнений были построены методы скалярной и матричной прогонок. Метод матричной прогонки для уравнения с переменными коэффициентами реализуется с затратой $O(M^3N)$ арифметических действий, где N — число уравнений, а M — размерность векторов неизвестных (число неизвестных в задаче равно MN). Для специальных классов векторных уравнений, соответствующих, например, разностной задаче Дирихле для уравнения Пуассона в прямоугольнике, был предложен модифицированный алгоритм метода матричной прогонки. Этот алгоритм позволяет сократить число действий до $O(MN^2)$.

Данная глава посвящена дальнейшему изучению прямых методов решения специальных векторных уравнений, к которым сводятся разностные схемы для простейших эллиптических уравнений. Будет построен метод *полной редукции*, позволяющий решать основные краевые задачи с затратой $O(MN \log_2 N)$ арифметических действий. Если не учитывать слабую логарифмическую зависимость от N , то число действий для этого метода пропорционально числу неизвестных MN . Создание этого метода является существенным шагом в развитии как прямых, так и итерационных методов решения сеточных уравнений.

Сформулируем краевые задачи для трехточечных векторных уравнений, решение которых можно найти по методу полной редукции. Мы будем рассматривать следующие задачи:

1) *Первая краевая задача.* Требуется найти решение уравнения

$$-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leq j \leq N-1, \quad (1)$$

удовлетворяющее заданным значениям при $j=0$ и $j=N$

$$Y_0 = F_0, \quad Y_N = F_N. \quad (2)$$

Здесь Y_j —вектор неизвестных номера j , F_j —заданная правая часть, а C —заданная квадратная матрица.

2) *Вторая и третья краевые задачи.* Ищется решение уравнения (1), удовлетворяющее следующим краевым условиям при $j=0$ и $j=N$:

$$\begin{aligned} (C + 2\alpha E) Y_0 - 2Y_1 &= F_0, & j = 0, \\ -2Y_{N-1} + (C + 2\beta E) Y_N &= F_N, & j = N, \end{aligned} \quad (3)$$

где $\alpha \geq 0$, $\beta \geq 0$. При $\alpha = \beta = 0$ формулы (3) задают краевые условия второго рода. Мы будем также рассматривать комбинации краевых условий, например, когда при $j=0$ задано краевое условие первого рода, а при $j=N$ —третьего или второго рода.

3) *Периодическая краевая задача.* Требуется найти решение уравнения $-Y_{j-1} + CY_j - Y_{j+1} = F_j$, которое является периодическим, $Y_{N+j} = Y_j$. Предполагается, что правая часть F_j также периодична, $F_{N+j} = F_j$. Эта задача формулируется в следующей эквивалентной форме: найти решение уравнения

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, & 1 \leq j \leq N-1, \\ -Y_{N-1} + CY_0 - Y_1 &= F_0, \quad Y_N = Y_0. \end{aligned} \quad (4)$$

К такого рода уравнениям сводятся разностные схемы для эллиптических уравнений в криволинейных ортогональных системах координат: в цилиндрической, полярной и сферической системах.

Помимо основного векторного уравнения (1), содержащего одну матрицу C , мы будем рассматривать первую краевую задачу для более общего уравнения

$$\begin{aligned} -BY_{j-1} + AY_j - BY_{j+1} &= F_j, & 1 \leq j \leq N-1, \\ Y_0 = F_0, \quad Y_N &= F_N \end{aligned} \quad (5)$$

с квадратными матрицами A и B . Подобного рода задачи возникают при решении разностной задачи Дирихле повышенного порядка точности для уравнения Пуассона в прямоугольнике.

Сформулируем требования на матрицы C , A и B , которые обеспечивают возможность применения метода полной редукции для решения поставленных задач (1)–(5). Для задач (1)–(4) будем предполагать, что для любого вектора Y справедливо неравенство $(CY, Y) \geq 2(Y, Y)$, а для задачи (5)—неравенство $(AY, Y) \geq 2(BY, Y) > 0$. Здесь используется обычное скалярное произведение векторов.

2. Первая краевая задача. Изучение метода полной редукции начнем с описания сеточных краевых задач для эллиптических уравнений, которые могут быть записаны в виде специальных векторных уравнений (1)–(5). Пусть на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq M, 0 \leq j \leq N, h_1 = l_1/M, h_2 = l_2/N\}$ с границей γ , введенной в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, требуется найти решение разностной задачи Дирихле для уравнения Пуассона

$$\begin{aligned} y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} &= -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma. \end{aligned} \quad (6)$$

В § 4 гл. II было показано, что задача (6) может быть записана в виде (1), (2), где \mathbf{Y}_j —вектор размерности $M-1$, компонентами которого являются значения сеточной функции $y(i, j) = y(x_{ij})$ во внутренних узлах j -й строки сетки $\bar{\omega}$:

$$\mathbf{Y}_j = (y(1, j), y(2, j), \dots, y(M-1, j)), \quad 0 \leq j \leq N.$$

C —квадратная матрица размерности $(M-1) \times (M-1)$, которая соответствует разностному оператору Λ , где

$$\begin{aligned} \Lambda y &= 2y - h_2^2 y_{\bar{x}_1 x_1}, \quad h_1 \leq x_1 \leq l_1 - h_1, \\ y &= 0, \quad x_1 = 0, \quad l_1. \end{aligned} \quad (7)$$

Правая часть \mathbf{F}_j —вектор размерности $M-1$, определяемый следующим образом:

1) для $j = 1, 2, \dots, N-1$

$$\mathbf{F}_j = (h_2^2 \bar{\varphi}(1, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(M-2, j), h_2^2 \bar{\varphi}(M-1, j)), \quad (8)$$

где

$$\bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} g(0, j),$$

$$\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j);$$

2) для $j = 0, N$

$$\mathbf{F}_j = (g(1, j), g(2, j), \dots, g(M-1, j)). \quad (9)$$

Из (7) следует, что для рассматриваемого примера матрица C является трехдиагональной симметричной матрицей.

Рассмотрим более сложную разностную задачу, которая также записывается в виде уравнений (1), (2). Пусть на сетке $\bar{\omega}$ требуется найти решение разностного уравнения Пуассона

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi(x), \quad x \in \omega, \quad (10)$$

удовлетворяющее на сторонах $x_1=0$ и $x_1=l_1$ краевым условиям третьего или второго рода

$$\frac{2}{h_1} y_{x_1} + y_{\bar{x}_1 x_2} = \frac{2}{h_1} \kappa_{-1} y - \bar{\varphi}, \quad x_1 = 0, \quad (11)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} + y_{x_2 x_2} = \frac{2}{h_1} \kappa_{+1} y - \bar{\varphi}, \quad x_1 = l_1, \quad (12)$$

$$h_2 \leq x_2 \leq l_2 - h_2$$

и краевым условиям первого рода на сторонах $x_2=0$, $x_2=l_2$: $y(x)=g(x)$, $x_2=0$, l_2 , $0 \leq x_1 \leq l_1$. Для того чтобы поставленная задача могла быть записана в виде (1), (2) с матрицей C , не зависящей от j , необходимо предположить выполнение условия $\kappa_{\pm 1} = \text{const}$.

Сведем эту задачу к (1), (2). Для этого умножим (10)–(12) на $(-h_2^2)$ и распишем разностную производную $y_{\bar{x}_1 x_2}$ по точкам для всех $j=1, 2, \dots, N-1$. Получим следующие уравнения:

1) для $i=0$

$$-y(0, j-1) + 2 \left[\left(1 + \frac{h_2^2}{h_1} \kappa_{-1} \right) y(0, j) - \frac{h_2^2}{h_1} y_{x_1}(0, j) \right] - y(0, j+1) = h_2^2 \bar{\varphi}(0, j);$$

2) для $i=1, 2, \dots, M-1$

$$-y(i, j-1) + [2y(i, j) - h_2^2 y_{\bar{x}_1 x_1}(i, j)] - y(i, j+1) = h_2^2 \varphi(i, j);$$

3) для $i=M$

$$-y(M, j-1) + 2 \left[\left(1 + \frac{h_2^2}{h_1} \kappa_{+1} \right) y(M, j) + \frac{h_2^2}{h_1} y_{\bar{x}_1}(M, j) \right] - y(M, j+1) = h_2^2 \bar{\varphi}(M, j).$$

Обозначим

$$\mathbf{Y}_j = (y(0, j), y(1, j), \dots, y(M, j)), \quad 0 \leq j \leq N,$$

$$\mathbf{F}_j = (h_2^2 \bar{\varphi}(0, j), h_2^2 \varphi(1, j), \dots, h_2^2 \varphi(M-1, j), h_2^2 \bar{\varphi}(M, j)), \quad 1 \leq j \leq N-1,$$

$$\mathbf{G}_j = (g(0, j), g(1, j), \dots, g(M, j)), \quad j=0, N.$$

В этих обозначениях полученные уравнения записываются в виде (1), (2), где квадратная матрица C размерности $(M+1) \times (M+1)$ соответствует разностному оператору Λ :

$$\Lambda \mathbf{y} = \begin{cases} 2 \left(1 + \frac{h_2^2}{h_1} \kappa_{-1} \right) y - \frac{2h_2^2}{h_1} y_{x_1}, & x_1 = 0, \\ 2y - h_2^2 y_{\bar{x}_1 x_1}, & h_2 \leq x_2 \leq l_2 - h_2, \\ 2 \left(1 + \frac{h_2^2}{h_1} \kappa_{+1} \right) y + \frac{2h_2^2}{h_1} y_{\bar{x}_1}, & x_1 = l_1. \end{cases} \quad (14)$$

Здесь мы снова имеем дело со случаем, когда C есть трехдиагональная матрица. Задание на сторонах $x_1=0$, l_1 краевых усло-

вий третьего рода (11), (12) вместо условий первого рода приводит лишь к другому определению оператора Λ — вместо (7) мы имеем (14). Вид уравнений (1) и краевых условий (2) при этом не меняется. Если при $x_1 = 0$ вместо условия (11) задано краевое условие первого рода $y(x) = g(x)$, а при $x_1 = l_1$ по-прежнему задано условие (12), то такая разностная задача также сводится к (1), (2). В этом случае

$$Y_j = (y(1, j), y(2, j), \dots, y(M, j)), \quad 0 \leq j \leq N,$$

$$F_j = (h_2^2 \bar{\varphi}(1, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(M-1, j), h_2^2 \bar{\varphi}(M, j)), \\ 1 \leq j \leq N-1,$$

где $\bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} g(0, j)$, $\bar{\varphi}(M, j)$ — значение в соответствующей точке правой части $\bar{\varphi}$ из (12), а квадратная матрица C соответствует разностному оператору Λ , где

$$\Lambda y = \begin{cases} 2y - h_2^2 y_{x_1}, & h_1 \leq x_1 \leq l_1 - h_1, \\ 2 \left(1 + \frac{h_2^2}{h_1} \kappa_{+1} \right) y + \frac{2h_2^2}{h_1} y_{x_1}, & x_1 = l_1 \end{cases} \quad (15)$$

и $y = 0$ при $x_1 = 0$.

Если краевое условие первого рода задано при $x_1 = l_1$, а краевое условие третьего рода (11) задано при $x_1 = 0$, то в (1), (2)

$$Y_j = (y(0, j), y(1, j), \dots, y(M-1, j)), \quad 0 \leq j \leq N,$$

$$F_j = (h_2^2 \bar{\varphi}(0, j), h_2^2 \varphi(1, j), \dots, h_2^2 \varphi(M-2, j), h_2^2 \bar{\varphi}(M-1, j)), \\ 1 \leq j \leq N-1,$$

где $\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j)$, а матрица C соответствует разностному оператору Λ , где

$$\Lambda y = \begin{cases} 2 \left(1 + \frac{h_2^2}{h_1} \kappa_{-1} \right) y - \frac{2h_2^2}{h_1} y_{x_1}, & x_1 = 0, \\ 2y - h_2^2 y_{x_1}, & h_1 \leq x_1 \leq l_1 - h_1 \end{cases} \quad (16)$$

и $y = 0$ при $x_1 = l_1$.

Итак, мы показали, что если по направлению x_2 заданы краевые условия первого рода, а по направлению x_1 — любые комбинации краевых условий первого, второго или третьего рода, то разностные схемы для уравнения Пуассона в прямоугольнике записываются в виде первой краевой задачи для трехточечных векторных уравнений (1), (2). Матрица C определяется при помощи разностного оператора Λ , который в зависимости от типа краевого условия на сторонах $x_1 = 0$ и $x_1 = l_1$ задается формулами (7), (14) — (16).

3. Другие краевые задачи для разностных уравнений. Тип краевых условий для уравнения (1) полностью определяется типом граничных условий для разностного уравнения (10) на

сторонах прямоугольника $x_2 = 0$ и $x_2 = l_2$. Мы рассмотрели случай, когда на этих сторонах были заданы краевые условия первого рода.

Рассмотрим теперь другие краевые задачи для уравнения (10), которые сводятся к векторным уравнениям (1), (3). Пусть на прямоугольной сетке $\bar{\omega}$, определенной выше, требуется найти решение *третьей краевой задачи* для разностного уравнения Пуассона. Разностная схема имеет следующий вид:

$$y_{\bar{x}_1 \bar{x}_1} + y_{\bar{x}_2 \bar{x}_2} = -\varphi(x), \quad x \in \omega, \quad (17)$$

$$\frac{2}{h_1} y_{x_1} + y_{\bar{x}_2 \bar{x}_2} = \frac{2}{h_1} \kappa_{-1} y - \bar{\varphi}, \quad x_1 = 0, \quad (18)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} + y_{\bar{x}_2 \bar{x}_2} = \frac{2}{h_1} \kappa_{+1} y - \bar{\varphi}, \quad x_1 = l_1, \quad h_2 \leq x_2 \leq l_2 - h_2, \quad (19)$$

$$y_{\bar{x}_1 \bar{x}_1} + \frac{2}{h_2} y_{x_2} = \frac{2}{h_2} \kappa_{-2} y - \bar{\varphi}, \quad x_2 = 0, \quad (19)$$

$$y_{\bar{x}_1 \bar{x}_1} - \frac{2}{h_2} y_{\bar{x}_2} = \frac{2}{h_2} \kappa_{+2} y - \bar{\varphi}, \quad x_2 = l_2, \quad h_1 \leq x_1 \leq l_1 - h_1. \quad (20)$$

Аппроксимация в уголках сетки имеет специальный вид:

$$\frac{2}{h_1} y_{x_1} + \frac{2}{h_2} y_{x_2} = \left(\frac{2}{h_1} \kappa_{-1} + \frac{2}{h_2} \kappa_{-2} \right) y - \bar{\varphi}, \quad x_1 = 0, \quad x_2 = 0, \quad (21)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} + \frac{2}{h_2} y_{x_2} = \left(\frac{2}{h_1} \kappa_{+1} + \frac{2}{h_2} \kappa_{-2} \right) y - \bar{\varphi}, \quad x_1 = l_1, \quad x_2 = 0, \quad (22)$$

$$\frac{2}{h_1} y_{x_1} - \frac{2}{h_2} y_{\bar{x}_2} = \left(\frac{2}{h_1} \kappa_{-1} + \frac{2}{h_2} \kappa_{+2} \right) y - \bar{\varphi}, \quad x_1 = 0, \quad x_2 = l_2, \quad (23)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} - \frac{2}{h_2} y_{\bar{x}_2} = \left(\frac{2}{h_1} \kappa_{+1} + \frac{2}{h_2} \kappa_{+2} \right) y - \bar{\varphi}, \quad x_1 = l_1, \quad x_2 = l_2. \quad (24)$$

Здесь предполагается, что выполнены условия $\kappa_{\pm\alpha} = \text{const}$, $\alpha = 1, 2$.

Покажем, что задача (17)–(24) сводится к (1), (3). Действительно, обозначая через \mathbf{Y}_j вектор размерности $M+1$

$$\mathbf{Y}_j = (y(0, j), y(1, j), \dots, y(M, j)), \quad 0 \leq j \leq N$$

и определяя правую часть \mathbf{F}_j для $j = 1, 2, \dots, N-1$ по формулам (13), получим из (17) и (18), как и в предыдущем пункте, уравнения (1) с матрицей C , соответствующей Λ из (14). Осталось показать, что условия (19)–(24) могут быть записаны в виде краевых условий (3).

Умножим (19), (21) и (22) на $(-h_2^2)$ и распишем входящую в них разностную производную y_{x_2} по точкам. Получим:

1) для $i = 0$

$$2 \left[\left(1 + \frac{h_2^2}{h_1} \kappa_{-1} \right) y(0, 0) - \frac{h_2^2}{h_1} y_{x_1}(0, 0) \right] + \\ + 2h_2 \kappa_{-2} y(0, 0) - 2y(0, 1) = h_2^2 \bar{\varphi}(0, 0),$$

2) для $i = 1, 2, \dots, M-1$

$$[2y(i, 0) - h_2^2 y_{x_1 x_1}(i, 0)] + 2h_2 \kappa_{-2} y(i, 0) - 2y(i, 1) = h_2^2 \varphi(i, 0),$$

3) для $i = M$

$$2 \left[\left(1 + \frac{h_2^2}{h_1} \kappa_{+1} \right) y(M, 0) + \frac{h_2^2}{h_1} y_{x_1}(M, 0) \right] + \\ + 2h_2 \kappa_{-2} y(M, 0) - 2y(M, 1) = h_2^2 \bar{\varphi}(M, 0).$$

Если обозначить $\alpha = h_2 \kappa_{-2}$, то эти равенства могут быть записаны в векторном виде

$$(C + 2\alpha E) Y_0 - 2Y_1 = F_0, \quad (25)$$

где $F_0 = (h_2^2 \bar{\varphi}(0, 0), h_2^2 \bar{\varphi}(1, 0), \dots, h_2^2 \bar{\varphi}(M, 0))$.

Аналогично из (20), (23) и (24) находится уравнение

$$-2Y_{N-1} + (C + 2\beta E) Y_N = F_N,$$

где обозначено $\beta = h_2 \kappa_{+2}$ и $F_N = (h_2^2 \bar{\varphi}(0, N), h_2^2 \bar{\varphi}(1, N), \dots, h_2^2 \bar{\varphi}(M, N))$. Итак, разностная схема (17) — (24) сведена к задаче (1), (3).

Рассмотрим теперь случай задания некоторых комбинаций краевых условий на сторонах прямоугольника \bar{G} . Как было отмечено выше, задание отличных от (18) краевых условий на сторонах $x_1 = 0$ и $x_1 = l_1$ влияет лишь на определение матрицы C . Если при $x_2 = 0$ задано краевое условие первого рода, т. е. вместо (19), (21) и (22) задано $y(x) = g(x)$, $x_2 = 0$, то условие (25) должно быть заменено на условие $Y_0 = F_0$, где $F_0 = (g(0, 0), \dots, g(M, 0))$. В этом случае трехточечная векторная краевая задача имеет вид

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N-1, \\ Y_0 &= F_0, \\ -2Y_{N-1} + (C + 2\beta E) Y_N &= F_N. \end{aligned} \quad (26)$$

К аналогичной системе мы приходим и в случае, когда на стороне $x_2 = l_2$ задано краевое условие первого рода, а на стороне $x_2 = 0$ — краевое условие третьего рода. В этом случае векторная краевая задача имеет вид

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N-1, \\ (C + 2\alpha E) Y_0 - 2Y_1 &= F_0, \quad Y_N = F_N. \end{aligned} \quad (27)$$

Мы рассмотрели примеры краевых задач для разностного уравнения Пуассона в прямоугольнике и показали, что им соответствуют векторные краевые задачи (1), (2) или (1), (3), или (26), (27) с соответствующей трехдиагональной матрицей C .

К указанным векторным краевым задачам сводятся и разностные схемы для более сложных эллиптических уравнений как в декартовой, так и в криволинейных ортогональных системах

координат. Приведем примеры. В декартовой системе это основные краевые задачи для эллиптического уравнения

$$\frac{\partial}{\partial x_1} \left(k_1(x_1) \frac{\partial u}{\partial x_1} \right) + k_2(x_1) \frac{\partial^2 u}{\partial x_2^2} - q(x_1) u = -f(x), \quad x \in G,$$

коэффициенты которого зависят только от одной переменной. В этом случае в прямоугольнике \bar{G} можно вводить прямоугольную сетку $\bar{\omega}$ с равномерным шагом h_2 по направлению x_2 и произвольными неравномерными шагами по направлению x_1 .

В цилиндрической системе координат такими примерами являются краевые задачи для уравнения Пуассона в конечном круговом цилиндре или трубе при наличии осевой симметрии:

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{\partial^2 u}{\partial z^2} = -f(r, z), \\ 0 \leq r_0 < r < R, \quad 0 < z < l.$$

В этом случае по направлению r можно вводить произвольную неравномерную сетку, а по направлению z — сетку с постоянным шагом h_2 .

Если для уравнения Пуассона ставится задача отыскания решения на поверхности цилиндра, т. е.

$$\frac{1}{R^2} \frac{\partial^2 u}{\partial \varphi^2} + \frac{\partial^2 u}{\partial z^2} = -f(\varphi, z), \quad 0 \leq \varphi \leq 2\pi, \quad 0 < z < l,$$

то соответствующая разностная задача сводится к периодической векторной краевой задаче (4), причем по направлению z допускается произвольная неравномерная сетка.

В полярной системе координат допустимыми являются разностные схемы для уравнения Пуассона в круге, кольце и круговом или кольцевом секторах

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \varphi^2} = -f(r, \varphi), \quad (r, \varphi) \in G.$$

Для круга и кольца разностная схема сводится к периодической задаче (4), а для секторов — к задачам (1), (2) или (1), (3). Здесь можно ввести неравномерную сетку по направлению r .

К периодической краевой задаче (4) сводится разностная схема для уравнения Пуассона, заданного на поверхности сферы радиуса R :

$$\frac{1}{R^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{R^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \varphi^2} = -f(\varphi, \theta).$$

4. Разностная задача Дирихле повышенного порядка точности. Рассмотрим теперь пример разностной схемы, которая приводится к более общему, чем (1), векторному уравнению (5). Запишем на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, \quad 0 \leq i \leq M,$

$0 \leq j \leq N$, $h_1 M = l_1$, $h_2 N = l_2$ } разностную задачу Дирихле для уравнения Пуассона повышенного порядка точности

$$\begin{aligned} & y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 x_1 \bar{x}_2 x_2} = -\varphi(x), \quad x \in \omega, \\ & y(x) = g(x), \quad x \in \gamma. \end{aligned} \quad (28)$$

Решение разностной схемы (28) при соответствующем выборе правой части $\varphi(x)$ сходится со скоростью $O(h_1^4 + h_2^4)$ к достаточно гладкому решению дифференциальной задачи, если $h_1 \neq h_2$, и со скоростью $O(h^6)$, если $h_1 = h_2 = h$.

Сведем (28) к краевой задаче для векторного трехточечного уравнения

$$\begin{aligned} & -B Y_{j-1} + A Y_j - B Y_{j+1} = F_j, \quad 1 \leq j \leq N-1, \\ & Y_0 = F_0, \quad Y_N = F_N. \end{aligned} \quad (29)$$

Для этого нужно умножить (28) на $(-h_2^2)$, расписать разностную производную $\left(y + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 x_1}\right)_{\bar{x}_2 x_2}$ по точкам и использовать обозначения

$$\begin{aligned} & Y_j = (y(1, j), y(2, j), \dots, y(M-1, j)), \\ & F_j = (h_2^2 \bar{\varphi}(1, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(M-2, j), h_2^2 \bar{\varphi}(M-1, j)), \\ & \quad 1 \leq j \leq N-1, \end{aligned}$$

где

$$\begin{aligned} & \bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} \left(g(0, j) + \frac{h_1^2 + h_2^2}{12} g_{\bar{x}_2 x_2}(0, j) \right), \\ & \bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} \left(g(M, j) + \frac{h_1^2 + h_2^2}{12} g_{\bar{x}_2 x_2}(M, j) \right) \end{aligned}$$

и

$$F_j = (g(1, j), g(2, j), \dots, g(M-1, j)), \quad j = 0, N.$$

В этом случае матрицы B и A соответствуют разностным операторам Λ_1 и Λ , где

$$\begin{aligned} & \Lambda_1 y = y + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 x_1}, \quad h_1 \leq x_1 \leq l_1 - h_1, \\ & \Lambda y = 2y - \frac{5h_2^2 - h_1^2}{6} y_{\bar{x}_1 x_1}, \quad h_1 \leq x_1 \leq l_1 - h_1, \end{aligned}$$

и $y = 0$ для $x_1 = 0$ и $x_1 = l_1$. Эти матрицы являются трехдиагональными и, как нетрудно проверить, перестановочными.

Краевую задачу (29) можно свести к задаче (1), (2). Для этого каждое из уравнений (29) нужно умножить слева на B^{-1} , если существует обратная к B матрица. Найдем достаточное условие

существования B^{-1} . Очевидно, что обратная к B матрица будет существовать, если система линейных алгебраических уравнений

$$BY = F \quad (30)$$

имеет единственное решение для любой правой части F .

В силу определения матрицы B , (30) может быть записано в виде разностной схемы

$$\begin{aligned} \Lambda_1 y &= y + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1, x_1} = f, \quad h_1 \leq x_1 \leq l_1 - h_1, \\ y(0) &= y(l_1) = 0. \end{aligned} \quad (31)$$

В § 1 гл. II было показано, что если для схемы (31) выполнены достаточные условия устойчивости метода прогонки, то решение уравнения (31) существует и единственno при любой правой части f , причем оно может быть найдено методом прогонки. Расписывая разностную производную $y_{\bar{x}_1, x_1}$ по точкам, запишем (31) в виде скалярных трехточечных уравнений

$$\begin{aligned} -A_i y_{i-1} + C_i y_i - B_i y_{i+1} &= F_i, \quad 1 \leq i \leq M-1, \\ y_0 &= 0, \quad y_M = 0, \end{aligned} \quad (32)$$

$$\text{где } A_i = B_i = \frac{h_1^2 + h_2^2}{12h_1^2}, \quad C_i = \frac{h_1^2 + h_2^2}{6h_1^2} - 1.$$

Напомним, что для (32) достаточные условия устойчивости метода прогонки имеют вид $|C_i| \geq |A_i| + |B_i|$, $i = 1, 2, \dots, M-1$. Из этих условий найдем, что матрица B имеет обратную, если шаги сетки $\bar{\omega}$ удовлетворяют ограничению $h_2 \leq \sqrt{2}h_1$. При выполнении этого условия задача (29) может быть сведена к задаче (1), (2) с $C = B^{-1}A$.

§ 2. Метод полной редукции для первой краевой задачи

1. Процесс нечетно-четного исключения. Переходим теперь к описанию метода полной редукции. Начнем с первой краевой задачи для трехточечных векторных уравнений

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N-1, \\ Y_0 &= F_0, \quad Y_N = F_N. \end{aligned} \quad (1)$$

Идея метода полной редукции решения задачи (1) состоит в последовательном исключении из уравнений (1) неизвестных Y_j , сначала с нечетными номерами j , затем из оставшихся уравнений с номерами j , кратными 2, затем 4 и т. д. Каждый шаг процесса исключения уменьшает число неизвестных, и если N есть степень 2, т. е. $N = 2^n$, то в результате процесса исключения останется одно уравнение, из которого можно найти $Y_{N/2}$. Обратный ход метода заключается в последовательном нахожде-

ции неизвестных \mathbf{Y}_j спачала с номерами j , кратными $N/4$, затем $N/8$, $N/16$ и т. д.

Очевидно, что метод полной редукции есть модификация метода исключения Гаусса, примененного к задаче (1), в котором исключение неизвестных происходит в специальном порядке. Напомним, что, в отличие от этого метода, в методе матричной прогонки исключение неизвестных происходит в естественном порядке.

Итак, пусть $N = 2^n$, $n > 0$. Для удобства введем следующие обозначения: $C^{(0)} = C$, $\mathbf{F}_j^{(0)} = \mathbf{F}_j$, $j = 1, 2, \dots, N-1$, используя которые запишем (1) в виде

$$-\mathbf{Y}_{j-1} + C^{(0)}\mathbf{Y}_j - \mathbf{Y}_{j+1} = \mathbf{F}_j^{(0)}, \quad 1 \leq j \leq N-1, \quad N = 2^n, \\ \mathbf{Y}_0 = \mathbf{F}_0, \quad \mathbf{Y}_N = \mathbf{F}_N. \quad (1')$$

Рассмотрим первый шаг процесса исключения. На этом шаге из уравнений системы (1') для j , кратных 2, исключим неизвестные \mathbf{Y}_j с нечетными номерами j . Для этого выпишем три идущие подряд уравнения (1'):

$$-\mathbf{Y}_{j-2} + C^{(0)}\mathbf{Y}_{j-1} - \mathbf{Y}_j = \mathbf{F}_{j-1}^{(0)}, \\ -\mathbf{Y}_{j-1} + C^{(0)}\mathbf{Y}_j - \mathbf{Y}_{j+1} = \mathbf{F}_j^{(0)}, \\ -\mathbf{Y}_j + C^{(0)}\mathbf{Y}_{j+1} - \mathbf{Y}_{j+2} = \mathbf{F}_{j+1}^{(0)}, \quad j = 2, 4, 6, \dots, N-2.$$

Умножим второе уравнение слева на $C^{(0)}$ и сложим все три получившиеся уравнения. В результате будем иметь

$$-\mathbf{Y}_{j-2} + C^{(1)}\mathbf{Y}_j - \mathbf{Y}_{j+2} = \mathbf{F}_j^{(1)}, \quad j = 2, 4, 6, \dots, N-2, \\ \mathbf{Y}_0 = \mathbf{F}_0, \quad \mathbf{Y}_N = \mathbf{F}_N, \quad (2)$$

где

$$C^{(1)} = [C^{(0)}]^2 - 2E, \\ \mathbf{F}_j^{(1)} = \mathbf{F}_{j-1}^{(0)} + C^{(0)}\mathbf{F}_j^{(0)} + \mathbf{F}_{j+1}^{(0)}, \quad j = 2, 4, 6, \dots, N-2.$$

Система (2) содержит неизвестные \mathbf{Y}_j только с четными номерами j , число неизвестных в (2) равно $N/2-1$, и если эта система будет решена, то неизвестные \mathbf{Y}_j с нечетными номерами в силу (1') могут быть найдены из уравнений

$$C^{(0)}\mathbf{Y}_j = \mathbf{F}_j^{(0)} + \mathbf{Y}_{j-1} + \mathbf{Y}_{j+1}, \quad j = 1, 3, 5, \dots, N-1 \quad (3)$$

с уже известными правыми частями.

Итак, исходная задача (1') эквивалентна системе (2) и уравнениям (3), причем по структуре система (2) аналогична исходной системе.

На втором шаге процесса исключения из уравнений «укороченной» системы (2) для j , кратных 4, исключаются неизвестные с номерами j , кратными 2, но не кратными 4. По аналогии

с первым шагом берутся три уравнения системы (2):

$$\begin{aligned} -Y_{j-4} + C^{(1)} Y_{j-2} - Y_j &= \mathbf{F}_{j-2}^{(1)}, \\ -Y_{j-2} + C^{(1)} Y_j - Y_{j+2} &= \mathbf{F}_j^{(1)}, \\ -Y_j + C^{(1)} Y_{j+2} - Y_{j+4} &= \mathbf{F}_{j+2}^{(1)}, \quad j = 4, 8, 12, \dots, N-4, \end{aligned}$$

второе уравнение умножается на $C^{(1)}$ слева, и все три уравнения складываются. В результате получаем систему из $N/4-1$ уравнений, содержащую неизвестные Y_j с номерами, кратными 4:

$$\begin{aligned} -Y_{j-4} + C^{(2)} Y_j - Y_{j+4} &= \mathbf{F}_j^{(2)}, \quad j = 4, 8, 12, \dots, N-4, \\ Y_0 &= \mathbf{F}_0, \quad Y_N = \mathbf{F}_N; \end{aligned}$$

уравнения $C^{(1)} Y_j = \mathbf{F}_j^{(1)} + Y_{j-2} + Y_{j+2}$, $j = 2, 6, 10, \dots, N-2$ для нахождения неизвестных с номерами, кратными 2, но не кратными 4, и уравнения (3) для неизвестных с нечетными номерами. При этом матрица $C^{(2)}$ и правые части $\mathbf{F}_j^{(2)}$ определяются по формулам

$$\begin{aligned} C^{(2)} &= [C^{(1)}]^2 - 2E, \\ \mathbf{F}_j^{(2)} &= \mathbf{F}_{j-2}^{(1)} + C^{(1)} \mathbf{F}_j^{(1)} + \mathbf{F}_{j+2}^{(1)}, \quad j = 4, 8, 12, \dots, N-4. \end{aligned}$$

Этот процесс исключения может быть продолжен. В результате l -го шага получим редуцированную систему для неизвестных с номерами, кратными 2^l :

$$\begin{aligned} -Y_{j-2^l} + C^{(l)} Y_j - Y_{j+2^l} &= \mathbf{F}_j^{(l)}, \quad j = 2^l, 2 \cdot 2^l, 3 \cdot 2^l, \dots, N-2^l, \\ Y_0 &= \mathbf{F}_0, \quad Y_N = \mathbf{F}_N, \end{aligned} \tag{4}$$

и группы уравнений

$$\begin{aligned} C^{(k-1)} Y_j &= \mathbf{F}_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ j &= 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N-2^{k-1}, \end{aligned} \tag{5}$$

решая которые последовательно для $k = l, l-1, \dots, 1$, найдем оставшиеся неизвестные. Матрицы $C^{(k)}$ и правые части $\mathbf{F}_j^{(k)}$ находятся по рекуррентным формулам

$$\begin{aligned} C^{(k)} &= [C^{(k-1)}]^2 - 2E, \\ \mathbf{F}_j^{(k)} &= \mathbf{F}_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_j^{(k-1)} + \mathbf{F}_{j+2^{k-1}}^{(k-1)}, \\ j &= 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N-2^k, \end{aligned} \tag{6}$$

для $k = 1, 2, \dots$

Из (4) следует, что после $(n-1)$ -го шага исключения ($l = n-1$) останется одно уравнение для $Y_{2^{n-1}} = Y_{N/2}$:

$$\begin{aligned} C^{(n-1)} Y_j &= \mathbf{F}_j^{(n-1)} + Y_{j-2^{n-1}} + Y_{j+2^{n-1}} = \mathbf{F}_j^{(n-1)} + Y_0 + Y_N, \quad j = 2^{n-1}, \\ Y_0 &= \mathbf{F}_0, \quad Y_N = \mathbf{F}_N \end{aligned}$$

с известной правой частью. Объединяя это уравнение с (5), получим, что все неизвестные находятся последовательно из

уравнений

$$\begin{aligned} C^{(k-1)} \mathbf{Y}_j &= \mathbf{F}_j^{(k-1)} + \mathbf{Y}_{j-2^k-1} + \mathbf{Y}_{j+2^k-1}, \quad \mathbf{Y}_0 = \mathbf{F}_0, \quad \mathbf{Y}_N = \mathbf{F}_N, \\ j &= 2^k-1, 3 \cdot 2^k-1, 5 \cdot 2^k-1, \dots, N-2^k-1, \quad k=n, n-1, \dots, 1. \end{aligned} \quad (7)$$

Итак, формулы (6) и (7) полностью описывают метод полной редукции. По формулам (6) преобразуются правые части, а из уравнений (7) находится решение исходной задачи (1).

Описанный метод мы назовем методом полной редукции, так как здесь последовательное уменьшение числа уравнений в системе осуществляется до конца, пока не останется одно уравнение для $\mathbf{Y}_{N/2}$. В методе неполной редукции, который будет рассмотрен в главе IV, осуществляется лишь частичное понижение порядка системы и «укороченная» система решается специальным методом.

2. Преобразование правой части и обращение матриц. Вычисление правой части $\mathbf{F}_j^{(k)}$ по рекуррентным формулам (6) может привести к накоплению погрешностей вычислений, если норма матрицы $C^{(k-1)}$ будет больше единицы. Кроме того, матрицы $C^{(k)}$ являются, вообще говоря, полными матрицами, даже если исходная матрица $C^{(0)} = C$ была трехдиагональной. А это существенным образом влияет на увеличение объема вычислительной работы при вычислении $\mathbf{F}_j^{(k)}$ по формулам (6). Для рассмотренных в § 1 примеров норма матрицы действительно будет значительно превышать единицу, и такой алгоритм метода будет вычислительно неустойчив.

Чтобы обойти эту трудность, будем вместо векторов $\mathbf{F}_j^{(k)}$ вычислять векторы $\mathbf{p}_j^{(k)}$, которые связаны с $\mathbf{F}_j^{(k)}$ следующим соотношением:

$$\mathbf{F}_j^{(k)} = \prod_{l=0}^{k-1} C^{(l)} \mathbf{p}_j^{(k)} 2^k, \quad (8)$$

причем формально положим $\prod_{l=0}^{-1} C^{(l)} = E$, так что $\mathbf{p}_j^{(0)} = \mathbf{F}_j^{(0)} = F_j$.

Найдем рекуррентные соотношения, которым удовлетворяют $\mathbf{p}_j^{(k)}$. Для этого подставим (8) в (6). Считая, что $C^{(l)}$ — невырожденная матрица для любого l , из (6) получим

$$2 \prod_{l=0}^{k-1} C^{(l)} \mathbf{p}_j^{(k)} = \prod_{l=0}^{k-2} C^{(l)} [\mathbf{p}_{j-2^k-1}^{(k-1)} + C^{(k-1)} \mathbf{p}_j^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}]$$

или

$$2C^{(k-1)} \mathbf{p}_j^{(k)} = \mathbf{p}_{j-2^k-1}^{(k-1)} + C^{(k-1)} \mathbf{p}_j^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}. \quad (9)$$

Обозначая $S_j^{(k-1)} = 2\mathbf{p}_j^{(k)} - \mathbf{p}_j^{(k-1)}$, из (9) получим, что $\mathbf{p}_j^{(k)}$ могут быть последовательно найдены по следующим формулам:

$$\begin{aligned} C^{(k-1)} S_j^{(k-1)} &= \mathbf{p}_{j-2^k-1}^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}, \quad \mathbf{p}_j^{(k)} = 0,5 (\mathbf{p}_j^{(k-1)} + S_j^{(k-1)}), \\ j &= 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N-2^k, \quad k=1, 2, \dots, n-1, \quad \mathbf{p}_j^{(0)} = \mathbf{F}_j. \end{aligned} \quad (10)$$

Рекуррентные соотношения (10) содержат сложение векторов, умножение вектора на число и обращение матриц $C^{(k-1)}$.

Осталось теперь исключить $F_j^{(k-1)}$ из уравнений (7). Подставляя (8) в (7), получим

$$\begin{aligned} C^{(k-1)}Y_j &= 2^{k-1} \prod_{l=0}^{k-2} C^{(l)} p_l^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ Y_0 &= F_0, \quad Y_N = F_N, \\ j &= 2^{k-1}, 3 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \quad k = n, n-1, \dots, 1. \end{aligned} \quad (11)$$

Здесь тоже необходимо обращать матрицы $C^{(k-1)}$, но, кроме того, в правой части (11) появилось умножение матрицы на вектор. В рассмотренном ниже алгоритме используемый способ обращения матрицы $C^{(k-1)}$ позволяет избежать нежелательной операции умножения матрицы на вектор и реализацию (11) свести к обращению матриц и сложению векторов.

Рассмотрим теперь вопрос об обращении матриц $C^{(k-1)}$, определяемых по рекуррентным формулам (6)

$$C^{(k)} = [C^{(k-1)}]^2 - 2E, \quad k = 1, 2, \dots, C^{(0)} = C. \quad (12)$$

Из (12) следует, что $C^{(k)}$ есть матричный полином степени 2^k относительно C с единичным коэффициентом при старшей степени. Этот полином через известные полиномы Чебышева выражается следующим образом:

$$C^{(k)} = 2T_{2^k} \left(\frac{1}{2}C \right), \quad k = 0, 1, \dots, \quad (13)$$

где $T_n(x)$ — полином Чебышева n -й степени первого рода (см. п. 2 § 4 гл. 1):

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leq 1, \\ \frac{1}{2} [(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^{-n}], & |x| \geq 1. \end{cases}$$

Действительно, в силу свойств полинома $T_n(x)$

$$T_{2n}(x) = 2[T_n(x)]^2 - 1, \quad T_1(x) = x,$$

из (12) очевидным образом следует (13).

Далее, используя соотношение

$$\prod_{l=0}^{k-2} 2T_{2^l}(x) = U_{2^{k-1}-1}(x),$$

связывающее полиномы Чебышева первого рода с полиномом второго рода $U_n(x)$, где

$$U_n(x) = \begin{cases} \frac{\sin((n+1)\arccos x)}{\sin(\arccos x)}, & |x| \leq 1, \\ \frac{1}{2\sqrt{x^2-1}} [(x + \sqrt{x^2-1})^{n+1} - (x - \sqrt{x^2-1})^{-(n+1)}], & |x| \geq 1, \end{cases}$$

ЛЮТКО ВЫЧИСЛИТЬ произведение полиномов $C^{(l)}$

$$\prod_{l=0}^{k-2} C^{(l)} = U_{2^{k-1}-1} \left(\frac{1}{2} C \right). \quad (14)$$

Итак, явное выражение для $C^{(k)}$ и $\prod_{l=0}^{k-1} C^{(l)}$ получено.

Для дальнейшего нам потребуется лемма 6 (см. п. 5 § 4 гл. II). Согласно лемме 6 любое отношение $g_m(x)/f_n(x)$ многочленов без общих корней в случае $n > m$ и простых корней $f_n(x)$ разлагается следующим образом на элементарные дроби:

$$\frac{g_m(x)}{f_n(x)} = \sum_{l=1}^n \frac{a_l}{x-x_l}, \quad a_l = \frac{g_m(x_l)}{f'_n(x_l)},$$

где x_l — корни полинома $f_n(x)$.

Используем лемму 6 для разложения отношений $1/T_n(x)$ и $U_{n-1}(x)/T_n(x)$ на элементарные дроби. Корни полинома $T_n(x)$ известны:

$$x_l = \cos \frac{(2l-1)}{2n} \pi, \quad l = 1, 2, \dots, n, \quad (15)$$

и в этих точках полином $U_{n-1}(x)$ принимает отличные от нуля значения

$$U_{n-1}(x_l) = \frac{\sin(n \arccos x_l)}{\sin(\arccos x_l)} = \frac{(-1)^{l+1}}{\sin \frac{(2l-1)}{2n} \pi}, \quad l = 1, 2, \dots, n.$$

Поэтому, используя соотношение $T'_n(x) = nU_{n-1}(x)$, из леммы 6 получим следующие разложения:

$$\frac{1}{T_n(x)} = \sum_{l=1}^n \frac{(-1)^{l+1} \sin \frac{(2l-1)\pi}{2n}}{n(x-x_l)}, \quad (16)$$

$$\frac{U_{n-1}(x)}{T_n(x)} = \sum_{l=1}^n \frac{1}{n(x-x_l)}, \quad (17)$$

где x_l определено в (15). Необходимые разложения найдены.

Получим теперь выражения для матриц $[C^{(k-1)}]^{-1}$ и $[C^{(k-1)}]^{-1} \prod_{l=0}^{k-2} C^{(l)}$ через матрицу C . Из (13) и (14) с учетом разложений алгебраических полиномов (16), (17) получим

$$[C^{(k-1)}]^{-1} = \sum_{l=1}^{2^{k-1}} \alpha_{l, k-1} \left(C - 2 \cos \frac{(2l-1)\pi}{2^k} E \right)^{-1},$$

$$[C^{(k-1)}]^{-1} \prod_{l=0}^{2^{k-1}} C^{(l)} = \frac{1}{2^{k-1}} \sum_{l=1}^{2^{k-1}} \left(C - 2 \cos \frac{(2l-1)\pi}{2^k} E \right)^{-1}.$$

Найденные соотношения позволяют записать в следующем виде как формулы (10):

$$\begin{aligned} \mathbf{S}_j^{(k-1)} &= \prod_{l=1}^{2^{k-1}} \alpha_{l, k-1} C_{l, k-1}^{-1} (\mathbf{p}_{j-2^{k-1}}^{(k-1)} + \mathbf{p}_{j+2^{k-1}}^{(k-1)}), \\ \mathbf{p}_j^{(k)} &= 0,5 (\mathbf{p}_j^{(k-1)} + \mathbf{S}_j^{(k-1)}), \\ \mathbf{p}_j^{(0)} &\equiv \mathbf{F}_j, \\ j &= 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, \quad k = 1, 2, \dots, n-1, \end{aligned} \quad (18)$$

так и формулы (11):

$$\begin{aligned} \mathbf{Y}_j &= \sum_{l=1}^{2^{k-1}} C_{l, k-1}^{-1} [\mathbf{p}_j^{(k-1)} + \alpha_{l, k-1} (\mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}})], \\ \mathbf{Y}_0 &= \mathbf{F}_0, \quad \mathbf{Y}_N = \mathbf{F}_N, \\ j &= 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \\ k &= n, n-1, \dots, 1, \end{aligned} \quad (19)$$

где обозначено

$$C_{l, k-1} = C - 2 \cos \frac{(2l-1)\pi}{2^k} E, \quad \alpha_{l, k-1} = \frac{(-1)^{l+1}}{2^{k-1}} \sin \frac{(2l-1)\pi}{2^k}. \quad (20)$$

Итак, получены преобразованные формулы (18), (19), описывающие метод полной редукции решения задачи (1). Эти формулы содержат только операции сложения векторов, умножения вектора на число и обращения матриц.

Заметим, что если C — трехдиагональная матрица, то трехдиагональной будет и любая матрица $C_{l, k-1}$. Задача обращения таких матриц была решена в главе II. Далее, если для матрицы C выполняется условие $(CY, Y) \geq 2(Y, Y)$, то из (20) следует, что матрицы $C_{l, k}$ будут положительно определенными, и, следовательно, будут иметь ограниченные обратные. Тогда из разложения $[C^{(k-1)}]^{-1}$ получим, что для любого $k \geq 1$ матрицы $C^{(k-1)}$ не вырождены. Напомним, что это предположение использовалось при получении формул (10).

§ 3. Алгоритм метода. Полученные выше формулы (18), (19) служат основой для первого алгоритма метода. Рассмотрим прежде всего, какие промежуточные величины и на каком этапе должны вычисляться и запоминаться для последующего использования.

Анализ формул (19) показывает, что при фиксированном k для вычисления \mathbf{Y}_j используются векторы $\mathbf{p}_j^{(k-1)}$ с номерами $j = 2^{k-1}, 3 \cdot 2^{k-1}, \dots, N - 2^{k-1}$. Любой вектор $\mathbf{p}_j^{(l)}$ с тем же номером j , но меньшим, чем $k-1$, номером l , является вспомогательным и запоминается временно. Поэтому определяемые на k -м шаге по (18) векторы $\mathbf{p}_j^{(k)}$ могут размещаться на месте $\mathbf{p}_j^{(k-1)}$, равно как и неизвестные \mathbf{Y}_j , вычисляемые по (19). Метод не требует дополнительной памяти ЭВМ — все векторы $\mathbf{p}_j^{(k)}$ размещаются на том месте, где затем будут размещаться \mathbf{Y}_j .

Проиллюстрируем организацию вычислений в рассматриваемом алгоритме на примере. Пусть $N=16$ ($n=4$). На рис. 1 показана последовательность вычисления и запоминания векторов $\mathbf{p}_j^{(k)}$. Заштрихованный квадрат означает, что для указанного

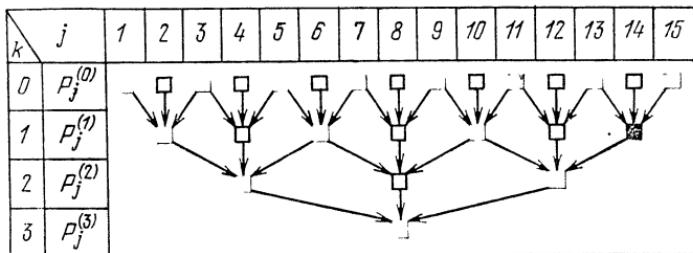


Рис. 1.

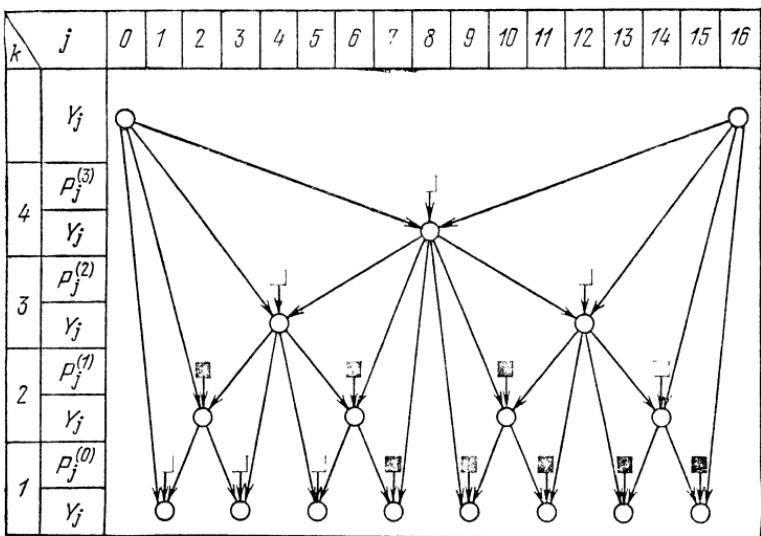


Рис. 2.

значения индекса k запоминается для последующего использования вектор $\mathbf{p}_j^{(k)}$ с соответствующим номером j . Соответственно, незаштрихованный квадрат означает, что $\mathbf{p}_j^{(k)}$ является вспомогательным и запоминается на указанном месте временно. Стрелки указывают, какие векторы $\mathbf{p}_i^{(k-1)}$ используются при вычислении $\mathbf{p}_j^{(k)}$.

В результате прямого хода метода будут запомнены следующие векторы $\mathbf{p}_j^{(k)}$:

$\mathbf{p}_1^{(0)}, \mathbf{p}_2^{(1)}, \mathbf{p}_3^{(0)}, \mathbf{p}_4^{(2)}, \mathbf{p}_5^{(0)}, \mathbf{p}_6^{(1)}, \mathbf{p}_7^{(0)}, \mathbf{p}_8^{(3)}, \mathbf{p}_9^{(0)}, \mathbf{p}_{10}^{(1)}, \mathbf{p}_{11}^{(0)}, \mathbf{p}_{12}^{(2)}, \mathbf{p}_{13}^{(0)}, \mathbf{p}_{14}^{(1)}, \mathbf{p}_{15}^{(0)}$.

Они используются для вычисления \mathbf{Y}_j на обратном ходе метода.

На рис. 2 показана последовательность вычисления неизвестных \mathbf{Y}_j (символическое обозначение о). Стрелками указано,

какие \mathbf{Y}_j , найденные на предыдущих шагах, и какие $\mathbf{p}_j^{(k-1)}$ (символическое обозначение \llcorner) используются для вычисления \mathbf{Y}_j при заданном k .

Переходим теперь к описанию алгоритма метода полной редукции. Прямой ход метода, согласно (18), реализуется следующим образом:

1) Задаются значения для $\mathbf{p}_j^{(0)} = \mathbf{F}_j$, $j = 1, 2, \dots, N-1$.

2) Для каждого фиксированного $k = 1, 2, \dots, n-1$ при фиксированном $j = 2^k, 2 \cdot 2^k, \dots, N - 2^k$ сначала вычисляются и запоминаются векторы

$$\Phi = \mathbf{p}_{j-2^k-1}^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}. \quad (21)$$

Затем для $l = 1, 2, \dots, 2^{k-1}$ решаются уравнения

$$C_{l, k-1} \mathbf{v}_l = \alpha_{l, k-1} \Phi. \quad (22)$$

В результате постепенным накоплением результата на месте $\mathbf{p}_j^{(k-1)}$ находится $\mathbf{p}_j^{(k)}$

$$\mathbf{p}_j^{(k)} = 0,5 (\mathbf{p}_j^{(k-1)} + \mathbf{v}_1 + \mathbf{v}_2 + \dots + \mathbf{v}_{2^{k-1}}). \quad (23)$$

Обратный ход метода, согласно (19), реализуется следующим образом:

1) Задаются значения для \mathbf{Y}_0 и \mathbf{Y}_N : $\mathbf{Y}_0 = \mathbf{F}_0$, $\mathbf{Y}_N = \mathbf{F}_N$.

2) Для каждого фиксированного $k = n, n-1, \dots, 1$ при фиксированном $j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}$ вычисляются и запоминаются векторы

$$\Psi = \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \quad \Psi = \mathbf{p}_j^{(k-1)}. \quad (24)$$

Затем для $l = 1, 2, \dots, 2^{k-1}$ решаются уравнения

$$C_{l, k-1} \mathbf{v}_l = \Psi + \alpha_{l, k-1} \Phi. \quad (25)$$

В результате постепенным накоплением значений на месте $\mathbf{p}_j^{(k-1)}$ находится вектор неизвестных \mathbf{Y}_j

$$\mathbf{Y}_j = \mathbf{v}_1 + \mathbf{v}_2 + \dots + \mathbf{v}_{2^{k-1}}. \quad (26)$$

Подсчитаем теперь число арифметических действий, затрачиваемых на реализацию описанного алгоритма. Пусть размерность вектора неизвестных \mathbf{Y}_j есть M , а через q обозначено число действий, требуемых для решения уравнения вида (22) или (25) при заданной правой части. Будем считать, что величины $\alpha_{l, k}$ заранее найдены.

Подсчитаем сначала число действий Q_1 , затрачиваемых на прямом ходе. При фиксированных k и j на вычисление вектора Φ по формулам (21) потребуется M операций. Далее, для каждого l на вычисление правой части в (22) и на решение уравнения (22) потребуется $M + q$ операций. Поэтому нахождение всех \mathbf{v}_l потребуется $2^{k-1}(M + q)$ действий. Вычисление $\mathbf{p}_j^{(k)}$ по формуле (23) осуществляется с затратой $2^{k-1}M + M$ действий.

Нтак, для вычисления $\mathbf{p}_j^{(k)}$ для одного k и j нужно затратить $M + 2^{k-1}(2M + \dot{q})$ операций.

Далее, при каждом фиксированном k нужно вычислять $N/2^k - 1$ различных $\mathbf{p}_j^{(k)}$. Следовательно, общее количество действий Q_1 , затрачиваемых на реализацию прямого хода, равно

$$Q_1 = \sum_{k=1}^{n-1} [M + (2M + \dot{q}) 2^{k-1}] \left(\frac{N}{2^k} - 1 \right) = \\ = (M + 0,5\dot{q}) Nn - (M + \dot{q}) N - M(n-1) + \dot{q}. \quad (27)$$

Подсчитаем теперь число действий Q_2 , затрачиваемых на обратном ходе. При фиксированных k и j на вычисление по формулам (24) потребуется M действий, на нахождение всех v_i в (25) — $(2M + \dot{q}) 2^{k-1}$ действий и на вычисление Y_j по формуле (26) — $(2^{k-1} - 1) M$ действий. Так как число различных значений j , для которых при фиксированном k проводятся указанные вычисления, равно $N/2^k$, то Q_2 равно

$$Q_2 = \sum_{k=1}^n [M + (2M + \dot{q}) 2^{k-1} + (2^{k-1} - 1) M] \frac{N}{2^k} = \\ = (1,5M + 0,5\dot{q}) Nn. \quad (28)$$

Складывая (27) и (28) и учитывая, что $n = \log_2 N$, получим следующую оценку для числа действий метода полной редукции, реализуемого по приведенному выше алгоритму

$$Q = Q_1 + Q_2 = (2,5M + \dot{q}) N \log_2 N - (M + \dot{q}) N - M(n-1) + \dot{q}. \quad (29)$$

Из (29) следует, что если $\dot{q} = O(M)$, то $Q = O(MN \log_2 N)$.

4. Второй алгоритм метода. Главным достоинством построенного алгоритма является минимальное требование к памяти ЭВМ — он не требует дополнительной памяти для хранения вспомогательной информации. Это качество достигается ценой некоторого увеличения объема вычислительной работы, которая затрачивается на повторное вычисление промежуточных величин. Рассмотрим еще один алгоритм метода, который характеризуется меньшим объемом вычислительной работы, но который требует дополнительную память, сравнимую по величине с общим числом неизвестных в задаче.

Для построения второго алгоритма вернемся к формулам (6), (7), описывающим метод полной редукции:

$$C^{(k)} = [C^{(k-1)}]^2 - 2E, \\ \mathbf{F}_j^{(k)} = \mathbf{F}_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_j^{(k-1)} + \mathbf{F}_{j+2^{k-1}}^{(k-1)}, \quad (6') \\ j = 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, k = 1, 2, \dots, n-1,$$

$$C^{(k-1)} Y_j = \mathbf{F}_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ Y_0 = F_0, \quad Y_N = Y_N, \quad (7') \\ j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}, k = n, n-1, \dots, 1.$$

Здесь, как и в первом алгоритме, векторы $\mathbf{F}_j^{(k)}$ непосредственно не вычисляются, а вместо них определяются векторы $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$, связанные с $\mathbf{F}_j^{(k)}$ следующим соотношением:

$$\mathbf{F}_j^{(k)} = C^{(k)} \mathbf{p}_j^{(k)} + \mathbf{q}_j^{(k)}, \quad j = 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, \quad k = 0, 1, \dots, n - 1. \quad (30)$$

Найдем рекуррентные формулы для вычисления векторов $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$. Так как вместо одного вектора $\mathbf{F}_j^{(k)}$ мы ввели два вектора, то имеется определенный произвол в определении $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$. Выберем $\mathbf{p}_j^{(0)}$ и $\mathbf{q}_j^{(0)}$ так, чтобы удовлетворить начальному условию $\mathbf{F}_j^{(0)} \equiv \mathbf{F}_j$. Для этого положим

$$\mathbf{p}_j^{(0)} = 0, \quad \mathbf{q}_j^{(0)} = \mathbf{F}_j, \quad j = 1, 2, \dots, N - 1. \quad (31)$$

Далее, подставляя (30) в (6'), получим

$$C^{(k)} \mathbf{p}_j^{(k)} + \mathbf{q}_j^{(k)} = C^{(k-1)} [\mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^k-1}^{(k-1)} + C^{(k-1)} \mathbf{p}_j^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}] + \\ + \mathbf{q}_{j-2^k-1}^{(k-1)} + \mathbf{q}_{j+2^k-1}^{(k-1)}, \quad j = 2^k, 2 \cdot 2^k, \dots, N - 2^k, k = 1, 2, \dots, n - 1.$$

Выбирая

$$\mathbf{q}_j^{(k)} = 2\mathbf{p}_j^{(k)} + \mathbf{q}_{j-2^k-1}^{(k-1)} + \mathbf{q}_{j+2^k-1}^{(k-1)} \quad (32)$$

и учитывая, что $C^{(k)} + 2E = [C^{(k-1)}]^2$, найдем отсюда

$$C^{(k-1)} \mathbf{p}_j^{(k)} = \mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^k-1}^{(k-1)} + C^{(k-1)} \mathbf{p}_j^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}. \quad (33)$$

Здесь мы снова предполагаем, что $C^{(l)}$ для любого l есть невырожденная матрица.

Обозначая $\mathbf{S}_j^{(k-1)} = \mathbf{p}_j^{(k)} - \mathbf{p}_j^{(k-1)}$, получим из (31)–(33) следующие рекуррентные формулы для вычисления векторов $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$:

$$C^{(k-1)} \mathbf{S}_j^{(k-1)} = \mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^k-1}^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}, \\ \mathbf{p}_j^{(k)} = \mathbf{p}_j^{(k-1)} + \mathbf{S}_j^{(k-1)}, \quad (34)$$

$$\mathbf{q}_j^{(k)} = 2\mathbf{p}_j^{(k)} + \mathbf{q}_{j-2^k-1}^{(k-1)} + \mathbf{q}_{j+2^k-1}^{(k-1)},$$

$$\mathbf{q}_j^{(0)} \equiv \mathbf{F}_j, \quad \mathbf{p}_j^{(0)} \equiv 0,$$

$$j = 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, \quad k = 1, 2, \dots, n - 1.$$

Осталось исключить $\mathbf{F}_j^{(k-1)}$ из формулы (7'). Подставляя (30) в (7') и обозначая $\mathbf{t}_j^{(k-1)} = \mathbf{Y}_j - \mathbf{p}_j^{(k-1)}$, получим следующие формулы для вычисления \mathbf{Y}_j :

$$C^{(k-1)} \mathbf{t}_j^{(k-1)} = \mathbf{q}_j^{(k-1)} + \mathbf{Y}_{j-2^k-1} + \mathbf{Y}_{j+2^k-1}, \\ \mathbf{Y}_j = \mathbf{p}_j^{(k-1)} + \mathbf{t}_j^{(k-1)}, \\ \mathbf{Y}_0 = \mathbf{F}_0, \quad \mathbf{Y}_N = \mathbf{F}_N, \quad (35)$$

$$j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \quad k = n, n - 1, \dots, 1.$$

Итак, мы получили формулы (34), (35), на которых основывается второй алгоритм метода полной редукции. Эти формулы содержат операции сложения векторов и обращения матриц $C^{(k-1)}$.

Остановимся теперь на вопросе обращения матриц $C^{(k-1)}$. Как было показано выше, матрица $C^{(k)}$ есть полином 2^k степени относительно исходной матрицы C и определяется по формуле (13) через полином Чебышева первого рода $T_n(x)$:

$$C^{(k)} = 2T_{2^k} \left(\frac{1}{2} C \right),$$

причем коэффициент при старшей степени равен единице. Так как корни полинома $T_n(x)$ известны (см. (15)), то $C^{(k)}$ можно представить в следующем факторизованном виде:

$$C^{(k)} = \prod_{l=1}^{2^k} \left(C - 2 \cos \frac{(2l-1)\pi}{2^{k+1}} E \right), \quad k = 0, 1, \dots$$

Используя обозначения (20), матрицу $C^{(k-1)}$ можно записать в следующем виде:

$$C^{(k-1)} = \prod_{l=1}^{2^{k-1}} C_{l, k-1}, \quad C_{l, k-1} = C - 2 \cos \frac{(2l-1)\pi}{2^k} E. \quad (36)$$

Факторизация (36) позволяет легко решать уравнения вида $C^{(k-1)} \mathbf{v} = \varphi$ с заданной правой частью φ . Следующий алгоритм дает решение этой задачи путем последовательного обращения множителей в (36):

$$\mathbf{v}_0 = \varphi, \quad C_{l, k-1} \mathbf{v}_l = \mathbf{v}_{l-1}, \quad l = 1, 2, \dots, 2^{k-1},$$

причем $\mathbf{v} = \mathbf{v}_{2^{k-1}}$. Этот алгоритм мы будем использовать для обращения матриц $C^{(k-1)}$.

Опишем теперь второй алгоритм метода полной редукции. Прямой ход метода реализуется на основании (34) следующим образом:

1) Задаются значения для $\mathbf{q}_j^{(0)}$: $\mathbf{q}_j^{(0)} = \mathbf{F}_j, j = 1, 2, \dots, N-1$.

2) Первый шаг для $k=1$ осуществляется отдельно по формулам, учитывающим начальные данные $\mathbf{p}_j^{(0)} \equiv 0$. Решаются уравнения для $\mathbf{p}_j^{(1)}$ и вычисляются $\mathbf{q}_j^{(1)}$:

$$\begin{aligned} C \mathbf{p}_j^{(1)} &= \mathbf{q}_j^{(0)}, \\ \mathbf{q}_j^{(1)} &= 2\mathbf{p}_j^{(1)} + \mathbf{q}_{j-1}^{(0)} + \mathbf{q}_{j+1}^{(0)}, \quad j = 2, 4, 6, \dots, N-2. \end{aligned} \quad (37)$$

3) Для каждого фиксированного $k=2, 3, \dots, n-1$ вычисляются и запоминаются векторы

$$\mathbf{v}_j^{(0)} = \mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^{k-1}}^{(k-1)} + \mathbf{p}_{j+2^{k-1}}^{(k-1)}, \quad j = 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N-2^k. \quad (38)$$

Затем при фиксированном $l=1, 2, 3, \dots, 2^{k-1}$ для каждого $j = 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N-2^k$ решаются уравнения

$$C_{l, k-1} \mathbf{v}_j^{(l)} = \mathbf{v}_j^{(l-1)} \quad (39)$$

с одной и той же матрицей, но разными правыми частями. В результате будут найдены векторы $\mathbf{v}_j^{(2^{k-1})}$ (в формулах (34) этим векторам соответствуют $\mathbf{S}_j^{(k-1)}$). Векторы $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$ вычисляются по формулам

$$\begin{aligned}\mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} + \mathbf{v}_j^{(2^{k-1})}, \\ \mathbf{q}_j^{(k)} &= 2\mathbf{p}_j^{(k)} + \mathbf{q}_{j-2^{k-1}}^{(k-1)} + \mathbf{q}_{j+2^{k-1}}^{(k-1)}, \\ j &= 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k.\end{aligned}\quad (40)$$

Обратный ход метода реализуется согласно (35):

- 1) Задаются значения для \mathbf{Y}_0 и \mathbf{Y}_N : $\mathbf{Y}_0 = \mathbf{F}_0$, $\mathbf{Y}_N = \mathbf{F}_N$.
- 2) Для каждого фиксированного $k = n, n-1, \dots, 2$ вычисляются и запоминаются векторы

$$\begin{aligned}\mathbf{v}_j^{(0)} &= \mathbf{q}_j^{(k-1)} + \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \\ j &= 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}.\end{aligned}\quad (41)$$

Затем при фиксированном $l = 1, 2, \dots, 2^{k-1}$ для каждого $j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}$ решаются уравнения

$$C_{l, k-1} \mathbf{v}_j^{(l)} = \mathbf{v}_j^{(l-1)}. \quad (42)$$

В результате находятся векторы $\mathbf{v}_j^{(2^{k-1})}$ (в (35) им соответствуют векторы $\mathbf{t}_j^{(k-1)}$). Далее вычисляется \mathbf{Y}_j по формуле

$$\mathbf{Y}_j = \mathbf{p}_j^{(k-1)} + \mathbf{v}_j^{(2^{k-1})}, \quad j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}. \quad (43)$$

- 3) Последний шаг обратного хода для $k = 1$ осуществляется решением уравнения

$$CY_j = \mathbf{q}_j^{(0)} + \mathbf{Y}_{j-1} + \mathbf{Y}_{j+1}, \quad j = 1, 3, 5, \dots, N-1. \quad (44)$$

Замечание к алгоритму. Все вновь определяемые по формулам (37) и (40) векторы $\mathbf{p}_j^{(k)}$ размещаются на месте $\mathbf{p}_j^{(k-1)}$. Все векторы $\mathbf{v}_j^{(l)}$ в формулах (38), (39), (41), (42), вновь определяемые по формулам (37), (40) векторы $\mathbf{q}_j^{(k)}$, а также решение \mathbf{Y}_j из (43) и (44) размещаются на месте $\mathbf{q}_j^{(k-1)}$. Следовательно, этот алгоритм требует память ЭВМ в 1,5 раза больше, чем число неизвестных в задаче.

Уменьшение объема вычислительной работы в данном алгоритме по сравнению с первым алгоритмом основано на том, что при решении серий задач (39) и (42) для различных j с одинаковыми матрицами $C_{l, k-1}$ полный объем работы затрачивается при решении лишь первой задачи из серии, а при решении каждой последующей задачи потребуется уже значительно меньше арифметических действий. Приведем число действий для второго алгоритма, обозначая, как и раньше через \dot{q} — число действий, затрачиваемых на решение уравнения вида (39) или (42) при заданной правой части, а через \bar{q} — число действий для решения того же уравнения, но с другой правой частью ($\bar{q} < \dot{q}$).

Количество действий, затрачиваемых на реализацию прямого хода, равно

$$Q_1 = \sum_{k=1}^{n-1} \left\{ 6M \left(\frac{N}{2^k} - 1 \right) + \left[\dot{q} + \bar{q} \left(\frac{N}{2^k} - 2 \right) \right] 2^{k-1} \right\} - 3M \left(\frac{N}{2} - 1 \right) = \\ = 0,5\bar{q}Nn + (0,5\dot{q} - 1,5\bar{q} + 4,5M)N - 6Mn - (\dot{q} - 2\bar{q} + 3M),$$

а обратного хода

$$Q_2 = \sum_{k=1}^n \left\{ 3M \frac{N}{2^k} + \left[\dot{q} + \left(\frac{N}{2^k} - 1 \right) \bar{q} \right] 2^{k-1} \right\} - \frac{MN}{2} = \\ = 0,5\bar{q}Nn + (\dot{q} - \bar{q} + 2,5M)N - \dot{q} + \bar{q} - 3M.$$

Общее число действий для второго алгоритма равно

$$Q = Q_1 + Q_2 =$$

$$= \bar{q}N \log_2 N + (1,5\dot{q} - 2,5\bar{q} + 7M)N - 6Mn - 2\dot{q} + 3\bar{q} - 6M. \quad (45)$$

Из оценки (45) следует, что если $\dot{q} = O(M)$, то $\bar{q} = O(M)$ и $Q = O(MN \log_2 N)$, причем здесь коэффициент при главном члене $MN \log_2 N$ меньше, чем в оценке (29), так как $\bar{q} < \dot{q}$.

Кратко остановимся еще на одной особенности второго алгоритма. Если в первом алгоритме обращение матриц $C^{(k-1)}$ осуществлялось обращением множителей $C_{l, k-1}$ и последующим суммированием результатов, то во втором алгоритме происходит последовательное обращение множителей и результат получается после обращения последнего множителя. С точки зрения реального вычислительного процесса, который учитывает погрешности округления, порядок обращения множителей $C_{l, k-1}$ во втором алгоритме является существенным. С аналогичной ситуацией мы встретимся в главе VI при изучении чебышевского итерационного метода.

Можно рекомендовать следующий порядок обращения матриц $C_{l, k-1}$. Матрице $C^{(k-1)}$ поставим в соответствие вектор $\theta_{2^{k-1}}$ размерности 2^{k-1} , компонентами которого являются целые числа от 1 до 2^{k-1} . Пусть

$$\theta_{2^{k-1}} = \{\theta_{2^{k-1}}(1), \theta_{2^{k-1}}(2), \dots, \theta_{2^{k-1}}(2^{k-1})\},$$

т. е. l -й элемент вектора $\theta_{2^{k-1}}$ обозначен через $\theta_{2^{k-1}}(l)$. Число $\theta_{2^{k-1}}(l)$ определяет очередь обращения матрицы $C_{l, k-1}$.

Вектор $\theta_{2^{k-1}}$ строится рекуррентно. Пусть $\theta_2 = \{2, 1\}$. Тогда процесс удвоения размерности вектора описывается следующими формулами:

$$\theta_{2m} = \{\theta_{2m}(4i-3) = \theta_m(2i-1), \theta_{2m}(4i-2) = \theta_m(2i-1) + m, \\ \theta_{2m}(4i-1) = \theta_m(2i) + m, \theta_{2m}(4i) = \theta_m(2i), \\ i = 1, 2, \dots, m/2\}, \quad m = 2, 4, 8, \dots$$

Пример: $\theta_{16} = \{2, 10, 14, 6, 8, 16, 12, 4, 3, 11, 15, 7, 5, 13, 9, 1\}$ и, следовательно, матрица $C_{6,16}$ будет обращаться шестнадцатой, а матрица $C_{12,16}$ — седьмой.

§ 3. Примеры применения метода

1. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Рассмотрим применение построенного выше метода полной редукции к нахождению решения разностной задачи Дирихле для уравнения Пуассона в прямоугольнике. Как было показано ранее, разностная задача

$$\begin{aligned} y_{x_1 x_1} + y_{x_2 x_2} &= -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned}$$

заданная на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leq i \leq M, 0 \leq j \leq N, h_1 M = l_1, h_2 N = l_2\}$, записывается в виде первой краевой задачи для векторных трехточечных уравнений

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N-1, \\ Y_0 &= F_0, \quad Y_N = F_N. \end{aligned} \quad (1)$$

Здесь

$$Y_j = (y(1, j), y(2, j), \dots, y(M-1, j)), \quad 0 \leq j \leq N,$$

— вектор неизвестных, компонентами которого являются значения сеточной функции $y(i, j)$ на j -й строке сетки,

$$F_j = (h_2^2 \bar{\varphi}(1, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(M-2, j), h_2^2 \bar{\varphi}(M-1, j)), \quad 1 \leq j \leq N-1,$$

$$F_j = (g(1, j), g(2, j), \dots, g(M-1, j)), \quad j = 0, N,$$

где

$$\bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} g(0, j),$$

$$\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j).$$

Квадратная матрица C соответствует разностному оператору Λ , где

$$\begin{aligned} \Lambda y &= 2y - h_2^2 y_{x_1 x_1}, \quad h_1 \leq x_1 \leq l_1 - h_1, \\ y &= 0, \quad x_1 = 0, l_1, \end{aligned}$$

так что

$$CY_j = (\Lambda y(1, j), \Lambda y(2, j), \dots, \Lambda y(M-1, j)).$$

Задача (1) может быть решена любым из двух приведенных выше алгоритмов метода полной редукции. Основным этапом этих алгоритмов является решение уравнений вида

$$C_{l,k-1} V = F, \quad C_{l,k-1} = C - 2 \cos \frac{(2l-1)\pi}{2^k} E \quad (2)$$

с заданной правой частью \mathbf{F} . Здесь \mathbf{V} —вектор неизвестных, $\mathbf{V} = (v(1), v(2), \dots, v(M-1))$ размерности $M-1$ (для упрощения записи индекс у \mathbf{V} и \mathbf{F} опущен).

Напомним, что число действий, затрачиваемых на решение задачи (1) по первому алгоритму, определяется числом действий \dot{q} , требуемых для решения уравнения (2) (см. (29) п. 3 § 2), а по второму алгоритму—в основном числом дополнительных действий \bar{q} , которое требуется затратить на решение уравнения (2), но с другой правой частью (см. (45) п. 4 § 2).

Для рассматриваемого примера приведем способ решения уравнения (2) и оценим \dot{q} и \bar{q} . Из определения матрицы C следует, что решение уравнения (2) эквивалентно нахождению решения следующей разностной задачи:

$$\begin{aligned} 2 \left(1 - \cos \frac{(2l-1)\pi}{2^k} \right) v - h_2^2 v_{x_1 x_1} &= f(i), \quad 1 \leq i \leq M-1, \\ v(0) = v(M) &= 0, \end{aligned} \quad (3)$$

где $f(i) = f_i$ — i -я компонента вектора \mathbf{F} . Расписывая разностную производную $v_{x_1 x_1}$ по точкам, запишем (3) в виде обычного трехточечного разностного уравнения для скалярных неизвестных $v(i) = v_i$:

$$\begin{aligned} -v_{i-1} + av_i - v_{i+1} &= bf_i, \quad 1 \leq i \leq M-1, \\ v_0 = v_M &= 0, \end{aligned} \quad (4)$$

где $a = 2 \left[1 + b \left(1 - \cos \frac{(2l-1)\pi}{2^k} \right) \right]$, $b = \frac{h_1^2}{h_2^2}$. Задача (4) является специальным случаем трехточечных краевых задач, методы решения которых были изучены в главе II. Было показано, что эффективным методом решения задач вида (4) является метод прогонки. Приведем расчетные формулы метода прогонки для задачи (4):

$$\begin{aligned} \alpha_{i+1} &= 1/(a - \alpha_i), & i = 1, 2, \dots, M-1, \alpha_1 &= 0, \\ \beta_{i+1} &= (bf_i + \beta_i) \alpha_{i+1}, & i = 1, 2, \dots, M-1, \beta_1 &= 0, \\ v_i &= \alpha_{i+1} v_{i+1} + \beta_{i+1}, & i = M-1, M-2, \dots, 1, v_M &= 0. \end{aligned}$$

Из этих формул следует, что задача (4), следовательно, и уравнение (2), при заданных a и b могут быть решены с затратой $\dot{q} = 7(M-1)$ действий. Для решения уравнения (2) с другой правой частью \mathbf{F} прогоночные коэффициенты α_i пересчитывать не нужно, и поэтому дополнительное число действий \bar{q} равно $\bar{q} = 5(M-1)$. Эти действия будут затрачены на вычисление β_i и на нахождение решения v_i . Отметим, что метод прогонки для (4) будет численно устойчив, так как достаточное условие устойчивости метода к ошибкам округления, имеющее в данном случае вид $a \geq 2$, выполнено.

Подставляя \dot{q} в оценку (29) п. 3 § 2 для числа действий первого алгоритма, получим, удерживая главные члены, $Q^{(1)} \approx 9,5MN \log_2 N - 8MN$. Для второго алгоритма из оценки (45) п. 4 § 2 получим следующую оценку для числа действий: $Q^{(2)} \approx 5MN \log_2 N + 5MN$. Итак, для каждого из рассмотренных алгоритмов число действий метода полной редукции, применяемого для решения разностной задачи Дирихле для уравнения Пуассона в прямоугольнике, есть величина порядка $O(MN \log_2 N)$, причем для второго алгоритма требуется меньше арифметических действий. Например, для $M=N=64$ получим $Q^{(1)} \approx 1,4Q^{(2)}$ и для $M=N=128$ соответственно $Q^{(1)} \approx 1,46Q^{(2)}$.

Мы не будем приводить расчетные формулы для алгоритмов решения указанной разностной задачи, так как на векторном уровне они подробно описаны в § 2.

В п. 2 § 1 были приведены примеры других разностных краевых задач, которые сводятся к задаче (1). Они отличаются от рассмотренной задачи Дирихле типом краевых условий на сторонах прямоугольника при $x_1=0$ и $x_1=l_1$, что приводит к различным матрицам C . Так для задачи (10)–(12) п. 2 § 1 с краевыми условиями третьего или второго рода при $x_1=0$, l_1 уравнение (2) эквивалентно разностной задаче

$$\begin{aligned} 2\left(1-\cos\frac{(2l-1)\pi}{2^k}\right)v-h_2^2v_{x_1,x_1} &= f, & 1 \leq i \leq M-1, \\ 2\left(1+\frac{h_2^2}{h_1}\kappa_{-1}-\cos\frac{(2l-1)\pi}{2^k}\right)v-\frac{2h_2^2}{h_1}v_{x_1} &= f, & i = 0, \\ 2\left(1+\frac{h_2^2}{h_1}\kappa_{+1}-\cos\frac{(2l-1)\pi}{2^k}\right)v+\frac{2h_2^2}{h_1}v_{x_1} &= f, & i = M. \end{aligned}$$

Эта задача в обычной трехточечной форме имеет вид

$$\begin{aligned} -v_{i-1}+av_i-v_{i+1} &= bf_i, & 1 \leq i \leq M-1, \\ v_0 &= \bar{\kappa}_1 v_1 + \mu_1, \\ v_M &= \bar{\kappa}_2 v_{M-1} + \mu_2, \end{aligned} \tag{5}$$

где

$$\bar{\kappa}_1 = \frac{2}{a+2h_1\kappa_{-1}}, \quad \bar{\kappa}_2 = \frac{2}{a+2h_1\kappa_{+1}}, \quad \mu_1 = \frac{bf_0}{a+2h_1\kappa_{-1}}, \quad \mu_2 = \frac{bf_M}{a+2h_1\kappa_{+1}},$$

а a и b определены выше.

Так как $a > 2$ и $\kappa_{\pm 1} \geq 0$, то $0 < \bar{\kappa}_1 < 1$ и $0 < \bar{\kappa}_2 < 1$, и метод прогонки решения задачи (5) будет также устойчив, а алгоритмы метода полной редукции и в этом случае будут требовать $O(MN \log_2 N)$ арифметических действий.

2. Разностная задача Дирихле повышенного порядка точности. В п. 4 § 1 разностная задача Дирихле для уравнения Пуассона

повышенного порядка точности

$$y_{\bar{x}_1 \bar{x}_1} + y_{\bar{x}_2 \bar{x}_2} + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 \bar{x}_1 \bar{x}_2 \bar{x}_2} = -\varphi(x), \quad x \in \omega, \\ y(x) = g(x), \quad x \in \gamma$$

была сведена к первой краевой задаче для неприведенного трехточечного векторного уравнения

$$-BY_{j-1} + AY_j - BY_{j+1} = F_j, \quad 1 \leq j \leq N-1, \\ Y_0 = F_0, \quad Y_N = F_N. \quad (6)$$

Квадратные матрицы B и A размерности $(M-1) \times (M-1)$ соответствуют разностным операторам Λ_1 и Λ , где

$$\Lambda_1 y = y + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 \bar{x}_1}, \quad h_1 \leq x_1 \leq l_1 - h_1, \\ \Lambda y = 2y - \frac{5h_2^2 - h_1^2}{6} y_{\bar{x}_1 \bar{x}_1}, \quad h_1 \leq x_1 \leq l_1 - h_1$$

и $y=0$ для $x_1=0$ и $x_1=l_1$.

Было показано, что если выполнено условие $h_2 \leq \sqrt{2}h_1$, то уравнения (6) приводятся к стандартному виду

$$-Y_{j-1} + CY_j - Y_{j+1} = \Phi_j, \quad 1 \leq j \leq N-1, \\ Y_0 = \Phi_0, \quad Y_N = \Phi_N, \quad (7)$$

где $C = B^{-1}A$, $\Phi_j = B^{-1}F_j$, $1 \leq j \leq N-1$ и $\Phi_j = F_j$ для $j=0, N$. Кроме того, было отмечено, что матрицы A и B перестановочны.

Для решения (7) используем первый алгоритм метода. Так как матрицу $C_{l,k-1}$ можно записать в виде

$$C_{l,k-1} = C - 2 \cos \frac{(2l-1)\pi}{2^k} E = B^{-1} \left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right),$$

то формулы (18), (19) § 2, определяющие первый алгоритм, принимают следующий вид:

$$S_i^{(k-1)} = \sum_{l=1}^{2^{k-1}} \alpha_{l,k-1} \left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right)^{-1} B \left(p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)} \right), \\ p_j^{(k)} = 0,5 (p_j^{(k-1)} + S_i^{(k-1)}), \\ j = 2^k, 2 \cdot 2^k, \dots, N - 2^k, k = 1, 2, \dots, n-1, \\ Bp_j^{(0)} \equiv F_j, \\ Y_j = \sum_{l=1}^{2^{k-1}} \left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right)^{-1} B [p_j^{(k-1)} + \\ + \alpha_{l,k-1} (Y_{j-2^{k-1}} + Y_{j+2^{k-1}})], \\ Y_0 = F_0, \quad Y_N = F_N, \quad j = 2^{k-1}, 3 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \\ k = n, n-1, \dots, 1.$$

Чтобы избежать обращения матриц B при задании $\mathbf{p}_j^{(0)}$ и умножения $\bar{\mathbf{p}}_j^{(k-1)}$ на матрицу B при вычислении \mathbf{Y}_j , сделаем замены, полагая $\bar{\mathbf{p}}_j^{(k)} = B\mathbf{p}_j^{(k)}$, $\bar{\mathbf{S}}_j^{(k)} = BS_j^{(k)}$. Тогда с учетом перестановочности матриц A и B , а следовательно, и матриц $(A - 2 \cos \frac{(2l-1)\pi}{2^k} B)^{-1}$ и B , написанные выше формулы примут вид (чертка сверху у $\bar{\mathbf{p}}_j^{(k)}$ и $\bar{\mathbf{S}}_j^{(k)}$ опущена):

$$\mathbf{S}_j^{(k-1)} = \sum_{l=1}^{2^{k-1}} \alpha_{l, k-1} \left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right)^{-1} B (\mathbf{p}_{j-2^{k-1}}^{(k-1)} + \mathbf{p}_{j+2^{k-1}}^{(k-1)}),$$

$$\mathbf{p}_j^{(k)} = 0,5 (\mathbf{p}_j^{(k-1)} + \mathbf{S}_j^{(k-1)}), \quad j = 2^k, 2 \cdot 2^k, \dots, N - 2^k, \quad k = 1, 2, \dots, n-1,$$

$$\mathbf{p}_j^{(0)} \equiv \mathbf{F}_j,$$

$$\mathbf{Y}_j = \sum_{l=1}^{2^{k-1}} \left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right)^{-1} [\mathbf{p}_j^{(k-1)} + \alpha_{l, k-1} B (\mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}})],$$

$\mathbf{Y}_0 = \mathbf{F}_0$, $\mathbf{Y}_N = \mathbf{F}_N$, $j = 2^{k-1}, 3 \cdot 2^{k-1}, \dots, N - 2^{k-1}$, $k = n, n-1, \dots, 1$. Полученные формулы порождают следующие изменения в первом алгоритме: формула (21) § 2 заменяется на

$$\Phi = B (\mathbf{p}_{j-2^{k-1}}^{(k-1)} + \mathbf{p}_{j+2^{k-1}}^{(k-1)}),$$

а вместо уравнений (22) решаются уравнения

$$\left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right) \mathbf{v}_l = \alpha_{l, k-1} \Phi$$

с вычисленным Φ . Аналогично (24) заменяется на

$$\Phi = B (\mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}), \quad \Psi = \mathbf{p}_j^{(k-1)},$$

а вместо (25) решаются уравнения

$$\left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right) \mathbf{v}_l = \Psi + \alpha_{l, k-1} \Phi.$$

Следовательно, для рассматриваемой задачи основным этапом алгоритма является решение уравнений вида

$$\left(A - 2 \cos \frac{(2l-1)\pi}{2^k} B \right) \mathbf{V} = \mathbf{F} \tag{8}$$

с заданной правой частью \mathbf{F} . Используя определение матриц A и B при помощи разностных операторов Λ и Λ_1 , получим, что (8) эквивалентно нахождению решения следующей разностной задачи:

$$2 \left(1 - \cos \frac{(2l-1)\pi}{2^k} \right) v - \left(\frac{5h_2^2 - h_1^2}{6} + \frac{h_1^2 + h_2^2}{6} \cos \frac{(2l-1)\pi}{2^k} \right) v_{x_1 x_1} = f, \tag{9}$$

$1 \leq i \leq M-1, \quad v_0 = v_M = 0.$

Расписывая это уравнение по точкам, получим первую краевую задачу для скалярного трехточечного уравнения

$$\begin{aligned} -v_{i-1} + av_i - v_{i+1} &= bf_i, \quad 1 \leq i \leq M-1, \\ v_0 = v_M &= 0, \end{aligned} \quad (10)$$

где

$$a = 2 \left[1 + b \left(1 - \cos \frac{(2l-1)\pi}{2^k} \right) \right],$$

$$b = \frac{6h_1^2}{5h_2^2 - h_1^2 + (h_1^2 + h_2^2) \cos \frac{(2l-1)\pi}{2^k}}.$$

Разностная задача (10) может быть решена методом прогонки, который будет численно устойчив, если выполнено достаточное условие $|a| \geq 2$. Покажем, что для любых h_1 и h_2 это условие выполнено. Действительно, если h_1 и h_2 таковы, что выполняется неравенство

$$\frac{h_2^2}{h_1^2} \geq \frac{1 - \cos \frac{(2l-1)\pi}{2^k}}{5 + \cos \frac{(2l-1)\pi}{2^k}}, \quad (11)$$

то $0 < b \leq \infty$ и, следовательно, $a > 2$. Заметим, что при равенстве в (11), коэффициент при $v_{x_1 x_1}$ в (9) обращается в нуль, и v может быть найдено из (9) по явной формуле.

Если (11) не выполнено, то для b верна оценка

$$b < -6 \left(1 - \cos \frac{(2l-1)\pi}{2^k} \right),$$

и, следовательно, $a < -10$. Утверждение доказано.

Итак, для решения разностной задачи Дирихле повышенного порядка точности можно применить метод полной редукции с оценкой $O(MN \log_2 N)$ арифметических действий.

§ 4. Метод полной редукции для других краевых задач

1. Вторая краевая задача. Выше был изучен метод полной редукции решения первой краевой задачи для трехточечных векторных уравнений. Изучение метода для более сложных краевых условий начнем с рассмотрения *второй краевой задачи*. Пусть требуется найти решение следующей задачи:

$$\begin{aligned} CY_0 - 2Y_1 &= F_0, & j &= 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, & 1 \leq j \leq N-1, \\ -2Y_{N-1} + CY_N &= F_N, & j &= N, \end{aligned} \quad (1)$$

где $N = 2^n$, $n > 0$.

Процесс последовательного исключения неизвестных в (1) осуществляется так же, как и в случае краевых условий первого рода. Именно, для четных j будем иметь уравнения

$$-Y_{j-2} + C^{(1)} Y_j - Y_{j+2} = F_j^{(1)}, \quad j = 2, 4, 6, \dots, N-2, \quad (2)$$

а для нечетных j — уравнения

$$C^{(0)} Y_j = F_j^{(0)} + Y_{j-1} + Y_{j+1}, \quad j = 1, 3, 5, \dots, N-1, \quad (3)$$

где, как и раньше, используются обозначения

$$\begin{aligned} F_j^{(1)} &= F_{j-1}^{(0)} + C^{(0)} F_j^{(0)} + F_{j+1}^{(0)}, & C^{(1)} &= [C^{(0)}]^2 - 2E, \\ C^{(0)} &= C, & F_j^{(0)} &\equiv F_j. \end{aligned}$$

Непреобразованными остались уравнения системы (1) лишь для $j=0$ и $j=N$. Исключим из указанных уравнений неизвестные Y_j с нечетными номерами j . Для этого используем два соседних уравнения. Выпишем уравнения для $j=0$ и $j=1$:

$$C^{(0)} Y_0 - 2Y_1 = F_0^{(0)}, \quad -Y_0 + C^{(0)} Y_1 - Y_2 = F_1^{(0)}.$$

Умножим первое уравнение слева на $C^{(0)}$, а второе — на 2, сложим получающиеся уравнения и найдем

$$C^{(1)} Y_0 - 2Y_2 = F_0^{(1)}, \quad (4)$$

где $F_0^{(1)} = C^{(0)} F_0^{(0)} + 2F_1^{(0)}$. Аналогично получим уравнение

$$-2Y_{N-2} + C^{(1)} Y_N = F_N^{(1)}, \quad (5)$$

где $F_N^{(1)} = 2F_{N-1}^{(0)} + C^{(0)} F_N^{(0)}$.

Объединяя (2), (4) и (5), получим «укороченную» полную систему уравнений для неизвестных с четными номерами j , имеющую аналогичную (1) структуру:

$$\begin{aligned} C^{(1)} Y_0 - 2Y_2 &= F_0^{(1)}, & j &= 0, \\ -Y_{j-2} + C^{(1)} Y_j - Y_{j+2} &= F_j^{(1)}, & j &= 2, 4, 6, \dots, N-2, \\ -Y_{N-2} + C^{(1)} Y_N &= F_N^{(1)}, & j &= N, \end{aligned}$$

и группу уравнений (3) для неизвестных с нечетными номерами j .

Продолжая описанный процесс исключения неизвестных дальше, после n -го шага исключения получим систему для Y_0 и Y_N :

$$C^{(n)} Y_0 - 2Y_N = F_0^{(n)}, \quad -2Y_0 + C^{(n)} Y_N = F_N^{(n)} \quad (6)$$

и уравнения для определения остальных неизвестных:

$$\begin{aligned} C^{(k-1)} Y_j &= F_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, & (7) \\ j &= 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N-2^{k-1}, & k &= n, n-1, \dots, 1, \end{aligned}$$

где $\mathbf{F}_i^{(k)}$ и $C^{(k)}$ определяются рекуррентно для $k = 1, 2, \dots, n$:

$$\begin{aligned}\mathbf{F}_0^{(k)} &= C^{(k-1)} \mathbf{F}_0^{(k-1)} + 2 \mathbf{F}_{2^{k-1}}^{(k-1)}, \\ \mathbf{F}_j^{(k)} &= \mathbf{F}_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} Y_j + \mathbf{F}_{j+2^{k-1}}^{(k-1)}, \\ j &= 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, \\ \mathbf{F}_N^{(k)} &= 2 \mathbf{F}_{N-2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_N^{(k-1)}, \\ C^{(k)} &= [C^{(k-1)}]^2 - 2E.\end{aligned}\tag{8}$$

Итак, нужно решить систему (6) и затем последовательно из уравнений (7) найти все остальные неизвестные.

Здесь, как и во втором алгоритме метода полной редукции, применяемого в случае первой краевой задачи, вместо векторов $\mathbf{F}_j^{(k)}$ будем определять векторы $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$, связанные с $\mathbf{F}_j^{(k)}$ соотношением

$$\begin{aligned}\mathbf{F}_j^{(k)} &= C^{(k)} \mathbf{p}_j^{(k)} + \mathbf{q}_j^{(k)}, \\ j &= 0, 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, N, k = 0, 1, \dots, n.\end{aligned}\tag{9}$$

Из (8) найдем, как и раньше, что $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$ для $j \neq 0, N$ могут быть найдены по формулам

$$\begin{aligned}C^{(k-1)} \mathbf{S}_j^{(k-1)} &= \mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^{k-1}}^{(k-1)} + \mathbf{p}_{j+2^{k-1}}^{(k-1)}, \\ \mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} + \mathbf{S}_j^{(k-1)}, \\ \mathbf{q}_j^{(k)} &= 2 \mathbf{p}_j^{(k)} + \mathbf{q}_{j-2^{k-1}}^{(k-1)} + \mathbf{q}_{j+2^{k-1}}^{(k-1)}, \\ j &= 2^k, 2 \cdot 2^k, \dots, N - 2^k, k = 1, 2, \dots, n-1, \\ \mathbf{q}_j^{(0)} &\equiv \mathbf{F}_j, \quad \mathbf{p}_j^{(0)} \equiv 0.\end{aligned}\tag{10}$$

Найдем теперь формулы для $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$ при $j = 0, N$. Подставляя (9) при $j = 0$ в (8) для $\mathbf{F}_0^{(k)}$, получим

$$C^{(k)} \mathbf{p}_0^{(k)} + \mathbf{q}_0^{(k)} = C^{(k-1)} [\mathbf{q}_0^{(k-1)} + 2 \mathbf{p}_{2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{p}_0^{(k-1)}] + 2 \mathbf{q}_{2^{k-1}}^{(k-1)}.$$

Выбирая $\mathbf{q}_0^{(k)} = 2 \mathbf{p}_0^{(k)} + 2 \mathbf{q}_{2^{k-1}}^{(k-1)}$ и учитывая равенство (12) п. 1 § 2, найдем уравнение для $\mathbf{p}_0^{(k)}$

$$C^{(k-1)} \mathbf{p}_0^{(k)} = C^{(k-1)} \mathbf{p}_0^{(k-1)} + \mathbf{q}_0^{(k-1)} + 2 \mathbf{p}_{2^{k-1}}^{(k-1)}.$$

Итак, векторы $\mathbf{p}_0^{(k)}$ и $\mathbf{q}_0^{(k)}$ могут быть найдены по следующим рекуррентным формулам:

$$\begin{aligned}C^{(k-1)} \mathbf{S}_0^{(k-1)} &= \mathbf{q}_0^{(k-1)} + 2 \mathbf{p}_{2^{k-1}}^{(k-1)}, \\ \mathbf{p}_0^{(k)} &= \mathbf{p}_0^{(k-1)} + \mathbf{S}_0^{(k-1)}, \\ \mathbf{q}_0^{(k)} &= 2 \mathbf{p}_0^{(k)} + 2 \mathbf{q}_{2^{k-1}}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_0^{(0)} &= \mathbf{F}_0, \quad \mathbf{p}_0^{(0)} = 0.\end{aligned}\tag{11}$$

Формулы для $\mathbf{p}_N^{(k)}$ и $\mathbf{q}_N^{(k)}$ получаются аналогично:

$$\begin{aligned} C^{(k-1)} \mathbf{S}_N^{(k-1)} &= \mathbf{q}_N^{(k-1)} + 2\mathbf{p}_{N-2^{k-1}}^{(k-1)}, \\ \mathbf{p}_N^{(k)} &= \mathbf{p}_N^{(k-1)} + \mathbf{S}_N^{(k-1)}, \\ \mathbf{q}_N^{(k)} &= 2\mathbf{p}_N^{(k)} + 2\mathbf{q}_{N-2^{k-1}}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_N^{(0)} &= \mathbf{F}_N, \quad \mathbf{p}_N^{(0)} = 0. \end{aligned} \tag{12}$$

Итак, формулы (10) – (12) позволяют полностью найти все необходимые нам векторы $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$. Осталось исключить $\mathbf{F}_j^{(k)}$ из (6) и (7). Подставляя (9) в (7), получим следующие формулы для вычисления \mathbf{Y}_j :

$$\begin{aligned} C^{(k-1)} \mathbf{t}_j^{(k-1)} &= \mathbf{q}_j^{(k-1)} + \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \\ \mathbf{Y}_j &= \mathbf{p}_j^{(k-1)} + \mathbf{t}_j^{(k-1)}. \end{aligned} \tag{13}$$

$j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}$, $k = n, n-1, \dots, 1$.

Осталось найти \mathbf{Y}_0 и \mathbf{Y}_N из (6). Но прежде заметим, что из (11) и (12) при $k=n$ следуют равенства

$$\mathbf{q}_0^{(n)} = 2\mathbf{p}_0^{(n)} + 2\mathbf{q}_{2^n-1}^{(n-1)}, \quad \mathbf{q}_N^{(n)} = 2\mathbf{p}_N^{(n)} + 2\mathbf{q}_{2^n-1}^{(n-1)},$$

т. е.

$$\mathbf{q}_0^{(n)} - \mathbf{q}_N^{(n)} = 2(\mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}). \tag{14}$$

Далее из (9) и (14) получим, что

$$\mathbf{F}_0^{(n)} - \mathbf{F}_N^{(n)} = C^{(n)}(\mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}) + \mathbf{q}_0^{(n)} - \mathbf{q}_N^{(n)} = (C^{(n)} + 2E)(\mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}).$$

Учитывая формулу (12) п.1 § 2, окончательно будем иметь:

$$\mathbf{F}_0^{(n)} - \mathbf{F}_N^{(n)} = [C^{(n-1)}]^2(\mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}). \tag{15}$$

Воспользуемся полученным соотношением для нахождения \mathbf{Y}_0 и \mathbf{Y}_N из (6). Вычитая из первого уравнения системы (6) второе, учитывая (15) и равенство (12) п.1 § 2, получим, что

$$\begin{aligned} (C^{(n)} + 2E)(\mathbf{Y}_0 - \mathbf{Y}_N) &= [C^{(n-1)}]^2(\mathbf{Y}_0 - \mathbf{Y}_N) = \mathbf{F}_0^{(n)} - \mathbf{F}_N^{(n)} = \\ &= [C^{(n-1)}]^2(\mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}). \end{aligned}$$

Считая, что $C^{(n-1)}$ — невырожденная матрица, отсюда найдем

$$\mathbf{Y}_0 = \mathbf{Y}_N + \mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}. \tag{16}$$

Подставляя найденное \mathbf{Y}_0 во второе уравнение системы (9), получим уравнение для нахождения \mathbf{Y}_N :

$$B^{(n)} \mathbf{Y}_N = \mathbf{F}_N^{(n)} + 2(\mathbf{p}_0^{(n)} - \mathbf{p}_N^{(n)}) = B^{(n)} \mathbf{p}_N^{(n)} + \mathbf{q}_N^{(n)} + 2\mathbf{p}_0^{(n)},$$

где $B^{(n)} = C^{(n)} - 2E$. Следовательно, если обозначить $\mathbf{t}^{(n)} = \mathbf{Y}_N - \mathbf{p}_N^{(n)}$, то \mathbf{Y}_N можно найти, решая уравнение

$$B^{(n)} \mathbf{t}^{(n)} = \mathbf{q}_N^{(n)} + 2\mathbf{p}_0^{(n)}, \quad (\mathbf{Y}_N = \mathbf{p}_N^{(n)} + \mathbf{t}^{(n)}). \tag{17}$$

Из (16) получим, что \mathbf{Y}_0 можно найти по формуле

$$\mathbf{Y}_0 = \mathbf{p}_0^{(n)} + \mathbf{t}^{(n)}, \quad (18)$$

где $\mathbf{t}^{(n)}$ найдено выше.

Итак, формулы (10)–(12), (17) и (18) описывают метод полной редукции решения второй краевой задачи для трехточечных векторных уравнений (1).

Замечание. Если \mathbf{Y}_0 задано, т. е. вместо задачи (1) решается задача

$$\begin{aligned} -\mathbf{Y}_{j-1} + C\mathbf{Y}_j - \mathbf{Y}_{j+1} &= \mathbf{F}_j, \quad 1 \leq j \leq N-1, \\ -2\mathbf{Y}_{N-1} + C\mathbf{Y}_N &= \mathbf{F}_N, \quad j=N, \quad \mathbf{Y}_0 = \mathbf{F}_0, \end{aligned}$$

то векторы $\mathbf{p}_0^{(k)}$ и $\mathbf{q}_0^{(k)}$ считать не нужно, а \mathbf{Y}_N , как это следует из (6) и (9), находится решением уравнения

$$C^{(n)}\mathbf{t}_N^{(n)} = \mathbf{q}_N^{(n)} + 2\mathbf{Y}_0, \quad (\mathbf{Y}_N = \mathbf{p}_N^{(n)} + \mathbf{t}_N^{(n)}).$$

Аналогично, если задано \mathbf{Y}_N , то вычислять векторы $\mathbf{p}_N^{(k)}$ и $\mathbf{q}_N^{(k)}$ не нужно, а \mathbf{Y}_0 определяется из уравнения $C^{(n)}\mathbf{t}_0^{(n)} = \mathbf{q}_0^{(n)} + 2\mathbf{Y}_N$, $\mathbf{Y}_0 = \mathbf{p}_0^{(n)} + \mathbf{t}_0^{(n)}$.

Для завершения описания метода редукции нужно указать способы обращения матриц $C^{(k)}$ и $B^{(n)} = C^{(n)} - 2E$. Для обращения матриц $C^{(k-1)}$ используется полученная выше (см. (36) § 2) факторизация

$$C^{(k-1)} = \prod_{l=1}^{2^{k-1}} C_{l, k-1}, \quad C_{l, k-1} = C - 2 \cos \frac{(2l-1)\pi}{2^k} E. \quad (19)$$

Заметим, что при выполнении условия $(CY, Y) \geq 2(Y, Y)$, все матрицы $C_{l, k-1}$ не вырождены, и, следовательно, не вырождена матрица $C^{(k-1)}$. Остановимся более подробно на вопросе обращения матрицы $B^{(n)}$.

Из определения $B^{(n)}$ и соотношения (12) п.1 § 2 получим

$$\begin{aligned} B^{(n)} &= C^{(n)} - 2E = [C^{(n-1)}]^2 - 4E = (C^{(n-1)} + 2E)(C^{(n-1)} - 2E) = \\ &= [C^{(n-2)}]^2 [C^{(n-1)} - 2E] = \dots = [C^{(n-2)} C^{(n-3)} \dots C^{(0)}]^2 (C^{(1)} - 2E) = \\ &= [C^{(n-2)} C^{(n-3)} \dots C^{(0)}]^2 [C^{(0)} - 2E] (C^{(0)} + 2E) = \\ &= \left[\prod_{k=1}^{n-1} C^{(k-1)} \right]^2 (C - 2E)(C + 2E). \end{aligned}$$

Подставляя сюда (19), найдем следующее представление для матрицы:

$$B^{(n)} = \left[\prod_{k=1}^{n-1} \prod_{l=1}^{2^{k-1}} C_{l, k-1} \right]^2 (C - 2E)(C + 2E). \quad (20)$$

Итак, матрица $B^{(n)}$ факторизована и обращение $B^{(n)}$ может быть осуществлено последовательным обращением множителей.

З а м е ч а н и е 1. Можно получить более компактную запись (20):

$$B^{(n)} = \prod_{l=1}^n \left(C - 2 \cos \frac{l\pi}{2^{n-1}} E \right).$$

З а м е ч а н и е 2. Из (20) следует, что матрица $B^{(n)}$ будет невырожденной, если выполнено условие $(CY, Y) > 2(Y, Y)$. Если же существует такой вектор $Y^* \neq 0$, для которого $CY^* = 2Y^*$, то $B^{(n)}$ вырождена и непосредственное применение метода редукции невозможно. Это является следствием вырожденности матрицы системы (1) в рассматриваемом случае. Действительно, в этом случае однородная система (1) имеет ненулевое решение $Y_j = Y^*$, и поэтому система (1) разрешима не для любой правой части. Если для данной правой части решение существует, то оно не единственное, а определяется с точностью до слагаемого Y^* . Одно из возможных решений выделяется на этапе обращения вырожденной матрицы $B^{(n)}$. Указанная ситуация имеет место при решении задачи Неймана для уравнения Пуассона в прямоугольнике. Более подробно указанные вопросы будут рассмотрены в главе XII, посвященной решению вырожденных сеточных уравнений.

2. Периодическая задача. Периодические трехточечные векторные задачи возникают при решении разностными методами эллиптических уравнений в криволинейных ортогональных системах координат—цилиндрической, полярной и сферической системах. В п. 3 § 1 приведены примеры дифференциальных задач, разностные схемы для которых могут быть сведены к следующей задаче: найти решение уравнений

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N-1, \\ -Y_{N-1} + CY_0 - Y_1 &= F_0, \quad j=0, \quad Y_N = Y_0. \end{aligned} \quad (21)$$

Задача (21) также может быть решена методом полной редукции. Рассмотрим первый шаг процесса исключения неизвестных. Как и раньше, из уравнений системы (21) для $j=2, 4, 6, \dots, N-2$ исключим при помощи двух соседних уравнений неизвестные Y_j с нечетными номерами j . Получим

$$-Y_{j-2} + C^{(1)}Y_j - Y_{j+2} = F_j^{(1)}, \quad j=2, 4, 6, \dots, N-2. \quad (22)$$

Осталось исключить Y_1 и Y_{N-1} из уравнения (21) для $j=0$. Для этого выпишем следующие три уравнения системы (21):

$$\begin{aligned} -Y_0 + CY_1 - Y_2 &= F_1, & j=1, \\ -Y_{N-1} + CY_0 - Y_1 &= F_0, & j=0, \\ -Y_{N-2} + CY_{N-1} - Y_N &= F_{N-1}, & j=N-1, \end{aligned}$$

умножим второе уравнение слева на C , сложим все три уравнения и учтем, что $Y_N = Y_0$. В результате получим уравнение

$$-Y_{N-2} + C^{(1)}Y_0 - Y_2 = F_0^{(1)}, \quad Y_N = Y_0. \quad (23)$$

где

$$\mathbf{F}_0^{(0)} = \mathbf{F}_1^{(0)} + C^{(0)} \mathbf{F}_0^{(0)} + \mathbf{F}_{N-1}^{(0)}, \quad C^{(0)} = C, \quad \mathbf{F}_j^{(0)} \equiv \mathbf{F}_j.$$

Объединяя (22) и (23), получим полную систему для неизвестных \mathbf{Y}_j с четными номерами j , имеющую аналогичную (21) структуру. Неизвестные \mathbf{Y}_j с нечетными номерами j находятся из обычных уравнений

$$C^{(0)} \mathbf{Y}_j = \mathbf{F}_j^{(0)} + \mathbf{Y}_{j-1} + \mathbf{Y}_{j+1}, \quad j = 1, 3, 5, \dots, N-1.$$

Процесс исключения может быть продолжен дальше. После l -го шага процесса исключения получим систему для неизвестных \mathbf{Y}_j с номерами j , кратными 2^l :

$$\begin{aligned} -\mathbf{Y}_{j-2^l} + C^{(l)} \mathbf{Y}_j - \mathbf{Y}_{j+2^l} &= \mathbf{F}_j^{(l)}, \quad j = 2^l, 2 \cdot 2^l, 3 \cdot 2^l, \dots, N-2^l, \\ -\mathbf{Y}_{N-2^l} + C^{(l)} \mathbf{Y}_0 - \mathbf{Y}_{2^l} &= \mathbf{F}_0^{(l)}, \quad j = 0, \quad \mathbf{Y}_N = \mathbf{Y}_0, \end{aligned}$$

и группу уравнений

$$\begin{aligned} C^{(k-1)} \mathbf{Y}_j &= \mathbf{F}_j^{(k-1)} + \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \\ j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N-2^{k-1}, k = l, l-1, \dots, 1 \end{aligned} \quad (24)$$

для последовательного нахождения остальных неизвестных. Правые части $\mathbf{F}_j^{(k)}$ определяются рекуррентно для $k = 1, 2, \dots, n-1$:

$$\begin{aligned} \mathbf{F}_j^{(k)} &= \mathbf{F}_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_j^{(k-1)} + \mathbf{F}_{j+2^{k-1}}^{(k-1)}, \\ j = 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N-2^k, \\ \mathbf{F}_0^{(k)} &= \mathbf{F}_{2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_0^{(k-1)} + \mathbf{F}_{N-2^{k-1}}^{(k-1)}, \quad \mathbf{F}_j^{(0)} \equiv \mathbf{F}_j. \end{aligned} \quad (25)$$

В результате $(n-1)$ -го шага процесса исключения получим систему относительно \mathbf{Y}_0 и $\mathbf{Y}_{2^{n-1}}$ ($\mathbf{Y}_N = \mathbf{Y}_0$):

$$\begin{aligned} C^{(n-1)} \mathbf{Y}_0 - 2 \mathbf{Y}_{2^{n-1}} &= \mathbf{F}_0^{(n-1)}, \\ -2 \mathbf{Y}_0 + C^{(n-1)} \mathbf{Y}_{2^{n-1}} &= \mathbf{F}_{2^{n-1}}^{(n-1)}. \end{aligned} \quad (26)$$

Решив эту систему, найдем \mathbf{Y}_0 , $\mathbf{Y}_{2^{n-1}}$ и $\mathbf{Y}_N = \mathbf{Y}_0$, а остальные неизвестные в силу (24) будут найдены как решения уравнений

$$\begin{aligned} C^{(k-1)} \mathbf{Y}_j &= \mathbf{F}_j^{(k-1)} + \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \\ j = 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N-2^{k-1}, k = n-1, n-2, \dots, 1. \end{aligned}$$

Прежде чем решать (26), найдем рекуррентные формулы для векторов $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$, связанных с $\mathbf{F}_j^{(k)}$ следующим соотношением:

$$\mathbf{F}_j^{(k)} = C^{(k)} \mathbf{p}_j^{(k)} + \mathbf{q}_j^{(k)}, \quad j = 0, 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N-2^k.$$

Используя рекуррентные формулы (25) для $\mathbf{F}_j^{(k)}$, получим

$$\begin{aligned} C^{(k-1)} \mathbf{S}_j^{(k-1)} &= \mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^{k-1}}^{(k-1)} + \mathbf{p}_{j+2^{k-1}}^{(k-1)}, \\ \mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} + \mathbf{S}_j^{(k-1)}, \\ \mathbf{q}_j^{(k)} &= 2\mathbf{p}_j^{(k)} + \mathbf{q}_{j-2^{k-1}}^{(k-1)} + \mathbf{q}_{j+2^{k-1}}^{(k-1)}, \\ j &= 2^k, 2 \cdot 2^k, 3 \cdot 2^k, \dots, N - 2^k, k = 1, 2, \dots, n-1, \\ \mathbf{q}_j^{(0)} &\equiv \mathbf{F}_j, \mathbf{p}_j^{(0)} \equiv 0, j = 1, 2, \dots, N-1, \end{aligned} \quad (27)$$

из которых находятся $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$ для $j \neq 0$, и формулы

$$\begin{aligned} C^{(k-1)} \mathbf{S}_0^{(k-1)} &= \mathbf{q}_0^{(k-1)} + \mathbf{p}_{2^k-1}^{(k-1)} + \mathbf{p}_{N-2^k-1}^{(k-1)}, \\ \mathbf{p}_0^{(k)} &= \mathbf{p}_0^{(k-1)} + \mathbf{S}_0^{(k-1)}, \\ \mathbf{q}_0^{(k)} &= 2\mathbf{p}_0^{(k)} + \mathbf{q}_{2^k-1}^{(k-1)} + \mathbf{q}_{N-2^k-1}^{(k-1)}, k = 1, 2, \dots, n-1, \\ \mathbf{q}_0^{(0)} &= \mathbf{F}_0, \mathbf{p}_0^{(0)} = 0 \end{aligned} \quad (28)$$

для нахождения $\mathbf{p}_0^{(k)}$ и $\mathbf{q}_0^{(k)}$.

Обратимся теперь к решению системы (26). Из (27) и (28) для $k = n-1$ получим соотношения

$$\begin{aligned} \mathbf{q}_{2^{n-1}}^{(n-1)} &= 2\mathbf{p}_{2^{n-1}}^{(n-1)} + \mathbf{q}_{2^{n-2}}^{(n-2)} + \mathbf{q}_{3 \cdot 2^{n-2}}^{(n-2)}, \\ \mathbf{q}_0^{(n-1)} &= 2\mathbf{p}_0^{(n-1)} + \mathbf{q}_{2^{n-2}}^{(n-2)} + \mathbf{q}_{3 \cdot 2^{n-2}}^{(n-2)}, \end{aligned}$$

из которых найдем

$$\mathbf{q}_0^{(n-1)} - \mathbf{q}_{2^{n-1}}^{(n-1)} = 2(\mathbf{p}_0^{(n-1)} - \mathbf{p}_{2^{n-1}}^{(n-1)}). \quad (29)$$

Вычтем теперь из первого уравнения системы (26) второе. Получим с учетом (29) и равенства (12) п.1 § 2

$$(C^{(n-1)} + 2E)(Y_0 - Y_{2^{n-1}}) = [C^{(n-2)}]^2(Y_0 - Y_{2^{n-1}}) = \mathbf{F}_0^{(n-1)} - \mathbf{F}_{2^{n-1}}^{(n-1)} = C^{(n-1)}(\mathbf{p}_0^{(n-1)} - \mathbf{p}_{2^{n-1}}^{(n-1)}) + \mathbf{q}_0^{(n-1)} - \mathbf{q}_{2^{n-1}}^{(n-1)} = [C^{(n-2)}]^2(\mathbf{p}_0^{(n-1)} - \mathbf{p}_{2^{n-1}}^{(n-1)}).$$

Предполагая, что $C^{(n-2)}$ невырожденная матрица, отсюда получим соотношение

$$Y_{2^{n-1}} = Y_0 - \mathbf{p}_0^{(n-1)} + \mathbf{p}_{2^{n-1}}^{(n-1)}. \quad (30)$$

Подставляя (30) в первое уравнение системы (26), получим

$$\begin{aligned} (C^{(n-1)} - 2E)Y_0 &= \mathbf{F}_0^{(n-1)} - 2(\mathbf{p}_0^{(n-1)} - \mathbf{p}_{2^{n-1}}^{(n-1)}) = \\ &= (C^{(n-1)} - 2E)\mathbf{p}_0^{(n-1)} + \mathbf{q}_0^{(n-1)} + 2\mathbf{p}_{2^{n-1}}^{(n-1)}. \end{aligned}$$

Следовательно, Y_0 можно найти по формулам

$$\begin{aligned} B^{(n-1)} \mathbf{t}^{(n-1)} &= \mathbf{q}_0^{(n-1)} + 2\mathbf{p}_{2^{n-1}}^{(n-1)}, \quad B^{(n-1)} = C^{(n-1)} - 2E, \\ Y_0 &= \mathbf{p}_0^{(n-1)} + \mathbf{t}^{(n-1)}, \end{aligned} \quad (31)$$

а $Y_{2^{n-1}}$ в силу (30) найдется тогда из соотношения

$$Y_{2^{n-1}} = \mathbf{p}_{2^{n-1}}^{(n-1)} + \mathbf{t}^{(n-1)}. \quad (32)$$

Остальные неизвестные найдутся последовательно по формулам

$$\begin{aligned} Y_N &= Y_0, \\ C^{(k-1)} t_j^{(k-1)} &= q_j^{(k-1)} + Y_{j-2^k-1} + Y_{j+2^k-1}, \\ Y_j &= p_j^{(k-1)} + t_j^{(k-1)}, \\ j &= 2^{k-1}, 3 \cdot 2^{k-1}, 5 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \\ k &= n-1, n-2, \dots, 1. \end{aligned} \quad (33)$$

Итак, формулы (27), (28), (31)–(33) описывают метод полной редукции решения периодической задачи (21). Для обращения матриц $C^{(k-1)}$ и $B^{(n-1)}$ используются факторизации (19), (20), причем в (20) нужно только заменить n на $n-1$.

Приведем оценку для числа арифметических действий Q , которые требуются для реализации метода полной редукции в случае периодической задачи. Обозначим, как и раньше, через \bar{q} число действий, затрачиваемых на решение уравнения $C_{l-k-1}V = F$, а через \bar{q} – число дополнительных действий для решения того же уравнения, но с другой правой частью F . Оценка дается формулой

$$Q = \bar{q}N \log_2 N + (1,5\bar{q} - 2\bar{q} + 7M)N - 2\bar{q} + 2\bar{q} - 14M.$$

Сравнение этой оценки с оценкой (45) § 2, полученной для случая первой краевой задачи, показывает, что затраты на решение периодической задачи практически равны затратам на решение первой краевой задачи.

3. Третья краевая задача.

3.1. Процесс исключения неизвестных. Рассмотрим теперь метод полной редукции решения третьей краевой задачи для трехточечных векторных уравнений

$$\begin{aligned} (C + 2\alpha E)Y_0 - 2Y_1 &= F_0, & j = 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, & 1 \leq j \leq N-1, \\ -2Y_{N-1} + (C + 2\beta E)Y_N &= F_N, & j = N. \end{aligned} \quad (34)$$

Предполагая, что выполнены условия $\alpha \geq 0$, $\beta \geq 0$, $\alpha^2 + \beta^2 \neq 0$, введем следующие обозначения:

$$C^{(0)} = C, \quad C_1^{(0)} = C + 2\alpha E, \quad C_2^{(0)} = C + 2\beta E, \quad F_j^{(0)} = F_j,$$

используя которые запишем (34) в виде

$$\begin{aligned} C_1^{(0)}Y_0 - 2Y_1 &= F_0^{(0)}, & j = 0, \\ -Y_{j-1} + C^{(0)}Y_j - Y_{j+1} &= F_j^{(0)}, & 1 \leq j \leq N-1, \\ -2Y_{N-1} + C_2^{(0)}Y_N &= F_N^{(0)}, & j = N. \end{aligned} \quad (34')$$

Пусть $N = 2^n$. Процесс исключения неизвестных для (34') осуществляется так же, как и для системы (1), которая соответствует случаю $C_1^{(0)} = C_2^{(0)} = C^{(0)}$ ($\alpha = \beta = 0$).

Выпишем редуцированную систему, получаемую в результате n -го шага процесса исключения неизвестных

$$C_1^{(n)} \mathbf{Y}_0 - 2 \mathbf{Y}_N = \mathbf{F}_0^{(n)}, \quad -2 \mathbf{Y}_0 + C_2^{(n)} \mathbf{Y}_N = \mathbf{F}_N^{(n)}, \quad (6')$$

и группы уравнений

$$\begin{aligned} & C^{(k-1)} \mathbf{Y}_j = \mathbf{F}_j^{(k-1)} + \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \\ & j = 2^{k-1}, 3 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \quad k = n, n-1, \dots, 1 \end{aligned} \quad (35)$$

для последовательного нахождения неизвестных \mathbf{Y}_j . Здесь правые части $\mathbf{F}_j^{(k)}$ определяются по рекуррентным формулам:

$$\mathbf{F}_j^{(k)} = \mathbf{F}_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_j^{(k-1)} + \mathbf{F}_{j+2^{k-1}}^{(k-1)}, \quad (36)$$

$$j = 2^k, 2 \cdot 2^k, \dots, N - 2^k, \quad k = 1, 2, \dots, n-1,$$

$$\mathbf{F}_0^{(k)} = C^{(k-1)} \mathbf{F}_0^{(k-1)} + 2 \mathbf{F}_{2^{k-1}}^{(k-1)}, \quad k = 1, 2, \dots, n, \quad (37)$$

$$\mathbf{F}_N^{(k)} = 2 \mathbf{F}_{N-2^{k-1}}^{(k-1)} + C^{(k-1)} \mathbf{F}_N^{(k-1)}, \quad k = 1, 2, \dots, n, \quad (38)$$

а матрицы $C_1^{(k)}$, $C_2^{(k)}$ и $C^{(k)}$ — по формулам:

$$C^{(k)} = [C^{(k-1)}]^2 - 2E, \quad k = 1, 2, \dots, n-1, \quad C^{(0)} = C,$$

$$C_1^{(k)} = C^{(k-1)} C_1^{(k-1)} - 2E, \quad k = 1, 2, \dots, n, \quad C_1^{(0)} = C + 2\alpha E, \quad (39)$$

$$C_2^{(k)} = C^{(k-1)} C_2^{(k-1)} - 2E, \quad k = 1, 2, \dots, n, \quad C_2^{(0)} = C + 2\beta E.$$

Из системы (6') получим уравнения для определения \mathbf{Y}_0 и \mathbf{Y}_N . Из (39) можно получить, что $C_1^{(k)}$, $C_2^{(k)}$ и $C^{(k)}$ суть матричные полиномы степени 2^k относительно одной и той же матрицы C . Следовательно, они перестановочны. Поэтому из (6') получим уравнения

$$\mathcal{D}^{(n+1)} \mathbf{Y}_0 = \mathbf{F}_0^{(n+1)}, \quad C_2^{(n)} \mathbf{Y}_N = \mathbf{F}_N^{(n)} + 2 \mathbf{Y}_0 \quad (40)$$

и эквивалентные им уравнения

$$\mathcal{D}^{(n+1)} \mathbf{Y}_N = \mathbf{F}_N^{(n+1)}, \quad C_1^{(n)} \mathbf{Y}_0 = \mathbf{F}_0^{(n)} + 2 \mathbf{Y}_N, \quad (40')$$

где обозначено

$$\mathbf{F}_0^{(n+1)} = C_2^{(n)} \mathbf{F}_0^{(n)} + 2 \mathbf{F}_N^{(n)}, \quad (41)$$

$$\mathbf{F}_N^{(n+1)} = 2 \mathbf{F}_0^{(n)} + C_1^{(n)} \mathbf{F}_N^{(n)}, \quad (42)$$

$$\mathcal{D}^{(n+1)} = C_1^{(n)} C_2^{(n)} - 4E = C_2^{(n)} C_1^{(n)} - 4E. \quad (43)$$

Итак, для нахождения \mathbf{Y}_0 и \mathbf{Y}_N можно воспользоваться уравнениями (40) или (40'). Будем использовать (40).

Вместо векторов $\mathbf{F}_j^{(k)}$ будем определять векторы $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$, связанные с $\mathbf{F}_j^{(k)}$ следующими соотношениями:

$$\mathbf{F}_0^{(k)} = C_1^{(k)} \mathbf{p}_0^{(k)} + \mathbf{q}_0^{(k)}, \quad (44)$$

$$\mathbf{F}_N^{(k)} = C_2^{(k)} \mathbf{p}_N^{(k)} + \mathbf{q}_N^{(k)}, \quad k = 0, 1, \dots, n, \quad (45)$$

$$\mathbf{F}_0^{(n+1)} = \mathcal{D}^{(n+1)} \mathbf{p}_0^{(n+1)} + \mathbf{q}_0^{(n+1)}, \quad (46)$$

$$\mathbf{F}_j^{(k)} = C^{(k)} \mathbf{p}_j^{(k)} + \mathbf{q}_j^{(k)}, \quad (47)$$

$$j = 2^k, 2 \cdot 2^k, \dots, N - 2^k, \quad k = 0, 1, 2, \dots, n-1.$$

Получим рекуррентные формулы для $\mathbf{p}_j^{(k)}$ и $\mathbf{q}_j^{(k)}$. Если $j \neq 0, N$, то из (36), (39) и (47) получим, предполагая, как и раньше, невырожденность матриц $C^{(k-1)}$, следующие формулы:

$$\begin{aligned} C^{(k-1)} S_j^{(k-1)} &= \mathbf{q}_j^{(k-1)} + \mathbf{p}_{j-2^k-1}^{(k-1)} + \mathbf{p}_{j+2^k-1}^{(k-1)}, \\ \mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} + S_j^{(k-1)}, \\ \mathbf{q}_j^{(k)} &= 2\mathbf{p}_j^{(k-1)} + \mathbf{q}_{j-2^k-1}^{(k-1)} + \mathbf{q}_{j+2^k-1}^{(k-1)}, \\ j &= 2^k, 2 \cdot 2^k, \dots, N - 2^k, k = 1, 2, \dots, n-1, \\ \mathbf{q}_j^{(0)} &\equiv \mathbf{F}_j, \quad \mathbf{p}_j^{(0)} \equiv 0. \end{aligned} \quad (48)$$

Найдем формулы для $\mathbf{p}_0^{(k)}$ и $\mathbf{q}_0^{(k)}$ при $k = 0, 1, \dots, n+1$. Подставляя (44) и (47) в (37), а (44)–(46) в (41), получим для $k = 1, 2, \dots, n$

$$C_1^{(k)} \mathbf{p}_0^{(k)} + \mathbf{q}_0^{(k)} =$$

$$= C^{(k-1)} (C_1^{(k-1)} \mathbf{p}_0^{(k-1)} + \mathbf{q}_0^{(k-1)} + 2\mathbf{p}_{2^k-1}^{(k-1)}) + 2\mathbf{q}_{2^k-1}^{(k-1)} \quad (49)$$

и для $k = n+1$

$$\mathcal{D}^{(n+1)} \mathbf{p}_0^{(n+1)} + \mathbf{q}_0^{(n+1)} = C_2^{(n)} (C_1^{(n)} \mathbf{p}_0^{(n)} + \mathbf{q}_0^{(n)} + 2\mathbf{p}_N^{(n)}) + 2\mathbf{q}_N^{(n)}. \quad (50)$$

Выберем $\mathbf{q}_0^{(k)}$ и $\mathbf{q}_0^{(n+1)}$ по формулам

$$\begin{aligned} \mathbf{q}_0^{(k)} &= 2\mathbf{p}_0^{(k)} + 2\mathbf{q}_{2^k-1}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_0^{(n+1)} &= 4\mathbf{p}_0^{(n+1)} + 2\mathbf{q}_N^{(n)} \end{aligned} \quad (51)$$

и используем вытекающие из (39) и (43) равенства

$$C_1^{(k)} + 2E = C^{(k-1)} C_1^{(k-1)}, \quad \mathcal{D}^{(n+1)} + 4E = C_2^{(n)} C_1^{(n)}.$$

Тогда (49) и (50) при условии невырожденности $C^{(k-1)}$ и $C_2^{(n)}$ можно записать в виде единого уравнения

$$\begin{aligned} C_1^{(k-1)} \mathbf{p}_0^{(k)} &= C_1^{(k-1)} \mathbf{p}_0^{(k-1)} + \mathbf{q}_0^{(k-1)} + 2\mathbf{p}_{2^k-1}^{(k-1)}, \\ k &= 1, 2, \dots, n+1. \end{aligned}$$

Объединяя эти уравнения с (51), получим окончательные формулы вычисления $\mathbf{p}_0^{(k)}$ и $\mathbf{q}_0^{(k)}$:

$$\begin{aligned} C_1^{(k-1)} S_0^{(k-1)} &= \mathbf{q}_0^{(k-1)} + 2\mathbf{p}_{2^k-1}^{(k-1)}, \\ \mathbf{p}_0^{(k)} &= \mathbf{p}_0^{(k-1)} + S_0^{(k-1)}, \quad k = 1, 2, \dots, n+1, \\ \mathbf{q}_0^{(k)} &= 2\mathbf{p}_0^{(k)} + 2\mathbf{q}_{2^k-1}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_0^{(n+1)} &= 4\mathbf{p}_0^{(n+1)} + 2\mathbf{q}_N^{(n)}, \\ \mathbf{q}_0^{(0)} &= \mathbf{F}_0, \quad \mathbf{p}_0^{(0)} = 0. \end{aligned} \quad (52)$$

Аналогично, используя (45), (47), рекуррентные соотношения (38) и (39), получим формулы для вычисления $\mathbf{p}_N^{(k)}$ и $\mathbf{q}_N^{(k)}$:

$$\begin{aligned} C_2^{(k-1)} S_N^{(k-1)} &= \mathbf{q}_N^{(k-1)} + 2\mathbf{p}_{N-2^k-1}^{(k-1)}, \\ \mathbf{p}_N^{(k)} &= \mathbf{p}_N^{(k-1)} + S_N^{(k-1)}, \\ \mathbf{q}_N^{(k)} &= 2\mathbf{p}_N^{(k)} + 2\mathbf{q}_{N-2^k-1}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_N^{(0)} &= \mathbf{F}_N, \quad \mathbf{p}_N^{(0)} = 0. \end{aligned} \quad (53)$$

Осталось исключить $\mathbf{F}_j^{(k)}$ из (35) и (40). Подставляя (47) в (35), а (45) и (46) в (40), получим следующие формулы для нахождения \mathbf{Y}_j :

$$\mathcal{D}^{(n+1)} \mathbf{S}_0^{(n+1)} = \mathbf{q}_0^{(n+1)}, \quad \mathbf{Y}_0 = \mathbf{p}_0^{(n+1)} + \mathbf{S}_0^{(n+1)}, \quad (54)$$

$$C_2^{(n)} \mathbf{S}_N^{(n)} = \mathbf{q}_N^{(n)} + 2\mathbf{Y}_0, \quad \mathbf{Y}_N = \mathbf{p}_N^{(n)} + \mathbf{S}_N^{(n)}, \quad (55)$$

$$C^{(k-1)} \mathbf{S}_j^{(k-1)} = \mathbf{q}_j^{(k-1)} + \mathbf{Y}_{j-2^{k-1}} + \mathbf{Y}_{j+2^{k-1}}, \quad (56)$$

$$\mathbf{Y}_j = \mathbf{p}_j^{(k-1)} + \mathbf{S}_j^{(k-1)},$$

$$j = 2^{k-1}, 3 \cdot 2^{k-1}, \dots, N - 2^{k-1}, \quad k = n, n-1, \dots, 1.$$

Итак, формулы (48), (52)–(56) описывают метод полной редукции решения третьей краевой задачи (34).

Замечание 1. Если для нахождения \mathbf{Y}_0 и \mathbf{Y}_N использовать уравнения (40'), то, вводя вместо $\mathbf{p}_0^{(n+1)}$ и $\mathbf{q}_0^{(n+1)}$ векторы $\mathbf{p}_N^{(n+1)}$ и $\mathbf{q}_N^{(n+1)}$, связанные с $\mathbf{F}_N^{(n+1)}$ соотношением

$$\mathbf{F}_N^{(n+1)} = \mathcal{D}^{(n+1)} \mathbf{p}_N^{(n+1)} + \mathbf{q}_N^{(n+1)},$$

получим из (38), (42), (44) и (47) следующие формулы для нахождения $\mathbf{p}_N^{(k)}$ и $\mathbf{q}_N^{(k)}$:

$$\begin{aligned} C_2^{(k-1)} \mathbf{S}_N^{(k-1)} &= \mathbf{q}_N^{(k-1)} + 2\mathbf{p}_{N-2^{k-1}}^{(k-1)}, \\ \mathbf{p}_N^{(k)} &= \mathbf{p}_N^{(k-1)} + \mathbf{S}_N^{(k-1)}, \quad k = 1, 2, \dots, n+1, \\ \mathbf{q}_N^{(k)} &= 2\mathbf{p}_N^{(k)} + 2\mathbf{q}_{N-2^{k-1}}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_N^{(n+1)} &= 4\mathbf{p}_N^{(n+1)} + 2\mathbf{q}_0^{(n)}, \\ \mathbf{q}_N^{(0)} &= \mathbf{F}_N, \quad \mathbf{p}_N^{(0)} = 0. \end{aligned} \quad (53')$$

Формулы (53') заменяют (53). Так как в этом случае вектор $\mathbf{F}_0^{(n+1)}$ и, следовательно, векторы $\mathbf{p}_0^{(n+1)}$ и $\mathbf{q}_0^{(n+1)}$ вычислять не нужно, то формулы (52) заменяются на следующие:

$$\begin{aligned} C_1^{(k-1)} \mathbf{S}_0^{(k-1)} &= \mathbf{q}_0^{(k-1)} + 2\mathbf{p}_{2^{k-1}}^{(k-1)}, \quad \mathbf{p}_0^{(k)} = \mathbf{p}_0^{(k-1)} + \mathbf{S}_0^{(k-1)}, \\ \mathbf{q}_0^{(k)} &= 2\mathbf{p}_0^{(k)} + 2\mathbf{q}_{2^{k-1}}^{(k-1)}, \quad k = 1, 2, \dots, n, \\ \mathbf{q}_0^{(0)} &= \mathbf{F}_0, \quad \mathbf{p}_0^{(0)} = 0. \end{aligned} \quad (52')$$

Из (35) и (40') получим формулы для нахождения \mathbf{Y}_0 и \mathbf{Y}_N :

$$\mathcal{D}^{(n+1)} \mathbf{S}_N^{(n+1)} = \mathbf{q}_N^{(n+1)}, \quad \mathbf{Y}_N = \mathbf{p}_N^{(n+1)} + \mathbf{S}_N^{(n+1)}, \quad (55')$$

$$C_1^{(n)} \mathbf{S}_0^{(n)} = \mathbf{q}_0^{(n)} + 2\mathbf{Y}_N, \quad \mathbf{Y}_0 = \mathbf{p}_0^{(n)} + \mathbf{S}_0^{(n)}. \quad (54')$$

Остальные неизвестные находятся согласно (56). Таким образом, формулы (48), (52')–(55') и (56) также можно использовать для решения задачи (34).

Замечание 2. Если \mathbf{Y}_N задано, т. е. вместо (34) нужно решить краевую задачу

$$\begin{aligned} (\mathcal{C} + 2\alpha E) \mathbf{Y}_0 - 2\mathbf{Y}_1 &= \mathbf{F}_0, & j = 0, \\ -\mathbf{Y}_{j-1} + C\mathbf{Y}_j - \mathbf{Y}_{j+1} &= \mathbf{F}_j, & 1 \leq j \leq N-1, \\ \mathbf{Y}_N &= \mathbf{F}_N, & j = N, \end{aligned}$$

то метод полной редукции в этом случае описывается формулами (48), (52'), (54') и (56). Если же задано \mathbf{Y}_0 , т. е. решается задача

$$\begin{aligned} -\mathbf{Y}_{j-1} + C\mathbf{Y}_j - \mathbf{Y}_{j+1} &= \mathbf{F}_j, & 1 \leq j \leq N-1, \\ -2\mathbf{Y}_{N-1} + (C + 2\beta E)\mathbf{Y}_N &= \mathbf{F}_N, & j = N, \quad \mathbf{Y}_0 = \mathbf{F}_0, \end{aligned}$$

то метод описывается формулами (48), (53), (55) и (56).

3.2. Факторизация матриц. Из (39) и (43) следует, что $C_1^{(k)}$, $C_2^{(k)}$ и $C^{(k)}$ являются матричными полиномами степени 2^k , а $\mathcal{D}^{(n+1)}$ — степени 2^{n+1} относительно матрицы C с коэффициентом, равным 1 при старшей степени. Имея в виду необходимость обращения этих матриц, факторизуем их. Для этого получим явное представление для этих полиномов через известные полиномы и изучим вопрос о нахождении корней указанных полиномов.

В п. 2 § 2 было показано, что $C^{(k)}$ выражаются через полином Чебышева первого рода следующим образом:

$$C^{(k)} = 2T_{2^k} \left(\frac{1}{2} C \right), \quad k = 0, 1, \dots \quad (57)$$

Далее, из соотношений (39) найдем

$$\begin{aligned} C_1^{(k)} - C^{(k)} &= C^{(k-1)} [C_1^{(k-1)} - C^{(k-1)}] = \dots \\ \dots &= \prod_{l=0}^{k-1} C^{(l)} [C_1^{(0)} - C^{(0)}] = 2\alpha \prod_{l=0}^{k-1} C^{(l)}. \end{aligned} \quad (58)$$

Так как имеет место равенство

$$\prod_{l=0}^{k-1} C^{(l)} = \prod_{l=0}^{k-1} 2T_{2^l} \left(\frac{1}{2} C \right) = U_{2^k-1} \left(\frac{1}{2} C \right),$$

где $U_n(x)$ — полином Чебышева второго рода, то из (58) получим следующее представление для $C_1^{(k)}$:

$$C_1^{(k)} = 2T_{2^k} \left(\frac{1}{2} C \right) + 2\alpha U_{2^k-1} \left(\frac{1}{2} C \right), \quad k = 0, 1, \dots \quad (59)$$

Аналогично получим представление для $C_2^{(k)}$:

$$C_2^{(k)} = 2T_{2^k} \left(\frac{1}{2} C \right) + 2\beta U_{2^k-1} \left(\frac{1}{2} C \right), \quad k = 0, 1, \dots \quad (60)$$

Далее, подставляя (59) и (60) в (43), будем иметь

$$\begin{aligned} \mathcal{D}^{(n+1)} &= 4 \left[T_{2^k} \left(\frac{1}{2} C \right) \right]^2 - 4E + \\ &+ 4(\alpha + \beta) T_{2^k} \left(\frac{1}{2} C \right) U_{2^k-1} \left(\frac{1}{2} C \right) + 4\alpha\beta \left[U_{2^k-1} \left(\frac{1}{2} C \right) \right]^2. \end{aligned} \quad (61)$$

Так как имеет место равенство

$$1 - T_n(x) = U_{n-1}(x)(1 - x^2), \quad (62)$$

то из (61) получим

$$\begin{aligned}\mathcal{D}^{(n+1)} = & U_{2^n-1} \left(\frac{1}{2} C \right) \left[(C^2 + 4\alpha\beta E - 4E) U_{2^n-1} \left(\frac{1}{2} C \right) + \right. \\ & \left. + 4(\alpha + \beta) T_{2^n} \left(\frac{1}{2} C \right) \right].\end{aligned}$$

Итак, представление для $C^{(k)}$, $C_1^{(k)}$, $C_2^{(k)}$ и $\mathcal{D}^{(n+1)}$ через известные полиномы получено. Так как корни полиномов Чебышева первого и второго рода известны, то из (57) и (62) получим

$$\begin{aligned}C^{(k)} = & \prod_{l=1}^{2^k} \left(C - 2 \cos \frac{(2l-1)\pi}{2^{k+1}} E \right), \\ \mathcal{D}^{(n+1)} = & \prod_{l=1}^{2^n-1} \left(C - 2 \cos \frac{l\pi}{2^n} E \right) \left[(C^2 + 4\alpha\beta E - 4E) U_{2^n-1} \left(\frac{1}{2} C \right) + \right. \\ & \left. + 4(\alpha + \beta) T_{2^n} \left(\frac{1}{2} C \right) \right].\end{aligned}$$

Поэтому отсюда и из (59), (60) следует, что нам осталось найти корни полиномов

$$\begin{aligned}P_m(t) &= 2T_m \left(\frac{t}{2} \right) + 2\alpha U_{m-1} \left(\frac{t}{2} \right), \\ Q_m(t) &= 2T_m \left(\frac{t}{2} \right) + 2\beta U_{m-1} \left(\frac{t}{2} \right), \\ m &= 2^k, \quad k = 0, 1, \dots, n-1,\end{aligned}\tag{63}$$

которые соответствуют матричным полиномам $C_1^{(k)}$ и $C_2^{(k)}$, и корни полинома

$$R_{2^n+1}(t) = (t^2 + 4\alpha\beta - 4) U_{2^n-1} \left(\frac{t}{2} \right) + 4(\alpha + \beta) T_{2^n} \left(\frac{t}{2} \right),\tag{64}$$

который порождает полином $\mathcal{D}^{(n+1)}$.

Эта задача может быть решена двумя способами. Первый путь состоит в использовании одного из методов приближенного нахождения корней полинома, второй путь — в сведении этой задачи к проблеме нахождения всех собственных значений некоторых трехдиагональных матриц. Остановимся подробнее на втором способе.

Обозначим через $S_k(\lambda)$ следующий определитель k -го порядка:

$$S_k(\lambda) = \begin{vmatrix} \lambda + 2\alpha & 2 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & \lambda & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & \lambda & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & \lambda & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & \lambda \end{vmatrix}, \quad k \geq 2$$

и положим $S_1(\lambda) = \lambda + 2\alpha$. Из определения и структуры соответствующей $S_k(\lambda)$ матрицы найдем рекуррентные соотношения для $S_k(\lambda)$:

$$\begin{aligned}S_{k+1}(\lambda) &= \lambda S_k(\lambda) - S_{k-1}(\lambda), \quad k \geq 2, \\ S_2(\lambda) &= \lambda S_1(\lambda) - 2, \quad S_1(\lambda) = \lambda + 2\alpha.\end{aligned}\tag{65}$$

Используя рекуррентные соотношения для полиномов Чебышева (см. п. 2 гл. I)

$$\begin{aligned} T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \quad T_1(x) = x, \quad T_0(x) = 1, \\ U_{n+1}(x) &= 2xU_n(x) - U_{n-1}(x), \quad U_1(x) = 2x, \quad U_0(x) = 1 \end{aligned}$$

и соотношения (65), получим представление $S_m(\lambda)$ через полиномы Чебышева:

$S_m(\lambda) = 2T_m\left(\frac{\lambda}{2}\right) + 2\alpha U_{m-1}\left(\frac{\lambda}{2}\right)$, $m \geq 1$. Сравнивая это выражение с (63), находим, что корни полинома $P_m(t)$ совпадают с корнями определителя $S_m(\lambda)$, зависящего от λ как от параметра.

Задача нахождения корней $S_m(\lambda)$ эквивалентна задаче нахождения таких значений параметра λ , при которых система алгебраических уравнений

$$\begin{aligned} y_{i-1} + \lambda y_i + y_{i+1} &= 0, \quad 1 \leq i \leq m-1, \\ (\lambda + 2\alpha)y_0 + 2y_1 &= 0, \quad i = 0, \\ y_m &= 0 \end{aligned} \tag{66}$$

имеет ненулевое решение. Дадим другую запись для (66). Используя обозначение для второй разностной производной

$$y_{\bar{xx}, i} = \frac{1}{h} (y_{x, i} - y_{\bar{x}, i}) = \frac{1}{h^2} (y_{i+1} - 2y_i + y_{i-1}),$$

перепишем (66) в следующем виде:

$$\begin{aligned} y_{\bar{xx}} + \mu y &= 0, \quad 1 \leq i \leq m-1, \\ \frac{2}{h} y_{x, i} + \frac{2\alpha}{h^2} y + \mu y &= 0, \quad i = 0, \quad y_m = 0, \end{aligned} \tag{66'}$$

где λ и μ связаны соотношением $\lambda = \mu h^2 - 2$. Итак, для нахождения корней полинома $C_1^{(k)}$ достаточно решить задачу (66') для $m = 2^k$, $k = 0, 1, \dots$

По аналогии с вышеизложенным можно показать, что корни полинома $Q_m(t)$ находятся из решения задачи

$$\begin{aligned} y_{\bar{xx}} + \mu y &= 0, \quad 1 \leq i \leq m-1, \\ -\frac{2}{h} y_{\bar{x}, i} + \frac{2\beta}{h^2} y + \mu y &= 0, \quad i = m, \quad y_0 = 0, \end{aligned} \tag{67}$$

причем соотношение $\lambda = \mu h^2 - 2$ определяет эти корни.

Для нахождения корней полинома $R_{2^n+1}(t)$, определенного в (64), нужно решить следующую задачу на собственные значения:

$$\begin{aligned} y_{\bar{xx}} + \mu y &= 0, \quad 1 \leq i \leq 2^n-1, \\ \frac{2}{h} y_{x, i} + \frac{2\alpha}{h^2} y + \mu y &= 0, \quad i = 0, \\ -\frac{2}{h} y_{\bar{x}, i} + \frac{2\beta}{h^2} y + \mu y &= 0, \quad i = 2^n, \end{aligned} \tag{68}$$

а корни найти из равенства $\lambda = \mu h^2 - 2$.

Отметим, что для решения задач (66) — (68) можно использовать известный QR-алгоритм решения полной проблемы собственных значений.

ГЛАВА IV

МЕТОД РАЗДЕЛЕНИЯ ПЕРЕМЕННЫХ

В главе изучаются варианты метода разделения переменных, который применяется для нахождения решения простейших сеточных эллиптических уравнений в прямоугольнике. В § 1 излагается алгоритм быстрого дискретного преобразования Фурье действительных и комплексных функций. В § 2 рассмотрен классический вариант метода разделения переменных, использующий алгоритм преобразования Фурье. В § 3 построен комбинированный метод, включающий в себя неполную редукцию и разделение переменных. Рассмотрено применение этого метода к решению разностных краевых задач для уравнения Пуассона второго и четвертого порядков точности.

§ 1. Алгоритм дискретного преобразования Фурье

1. Постановка задачи. Одним из методов отыскания решений сеточных многомерных задач, допускающих разделение переменных, является разложение искомого решения в конечную сумму Фурье по собственным функциям соответствующих сеточных операторов. Эффективность этого метода существенно зависит от того, как быстро можно вычислить коэффициенты Фурье заданной сеточной функции и восстановить искомую функцию по заданным коэффициентам Фурье.

Если, например, на сетке $\bar{\omega} = \{x_i = ih, 0 \leq i \leq N, hN = l\}$, содержащей $N+1$ узел, заданы функция $f(i)$ и система ортонормированных функций $\mu_k(i)$, $k = 0, 1, \dots, N$, а коэффициенты Фурье функции $f(i)$ вычисляются по формулам

$$\varphi_k = \sum_{i=0}^N f(i) \mu_k(i) h, \quad k = 0, 1, \dots, N, \quad (1)$$

то для вычисления всех коэффициентов φ_k достаточно $(N+1)(N+2)$ операций умножения и $N(N+1)$ операций сложения.

В общем случае произвольной системы функций $\{\mu_k(i)\}$ это есть минимально необходимое количество арифметических операций. В ряде специальных случаев, когда ортонормированная система функций имеет специальный вид, общее число арифметических действий, необходимое для вычисления сумм вида (1), может быть значительно сокращено. Мы рассмотрим эти случаи и приведем алгоритмы, позволяющие вычислять все коэффициенты Фурье и восстанавливать функцию по заданным коэффициентам Фурье с затратой $O(N \ln N)$ арифметических действий.

Переходим к описанию отмеченных случаев.

Задача 1. Разложение по синусам. Пусть на отрезке $0 \leq x \leq l$ введена равномерная с шагом h сетка $\omega = \{x_j = jh, 0 \leq j \leq N, hN = l\}$. Обозначим через $\underline{\omega} = \{x_j = jh, 1 \leq j \leq N-1\}$ множество внутренних узлов сетки ω .

Пусть на ω задана действительная сеточная функция $f(j)$ (или $f(j)$ задана на $\bar{\omega}$, причем $f(0) = f(N) = 0$).

В § 5 гл. I было показано, что функция $f(j)$ может быть представлена в виде разложения

$$f(j) = \frac{2}{N} \sum_{k=1}^{N-1} \varphi_k \sin \frac{k\pi j}{N}, \quad j = 1, 2, \dots, N-1, \quad (2)$$

где коэффициенты φ_k определяются формулой

$$\varphi_k = \sum_{j=1}^{N-1} f(j) \sin \frac{k\pi j}{N}, \quad k = 1, 2, \dots, N-1. \quad (3)$$

Сравнивая (2) и (3), находим, что задачи вычисления коэффициентов φ_k заданной функции $f(j)$ и восстановления этой функции по заданным $\{\varphi_k\}$ сводятся к вычислению $N-1$ суммы вида

$$y_k = \sum_{j=1}^{N-1} a_j \sin \frac{k\pi j}{N}, \quad k = 1, 2, \dots, N-1. \quad (4)$$

Формула (4) описывает правило преобразования сеточной функции a_j , $1 \leq j \leq N-1$, заданной на сетке ω , в сеточную функцию y_j , $1 \leq j \leq N-1$. Алгебраическая трактовка (4) такова: если обозначить через $\alpha = (a_1, a_2, \dots, a_{N-1})$ вектор размерности $N-1$, то (4) описывает преобразование вектора α при переходе от естественного базиса к базису, образованному системой ортогональных векторов

$$z_k = (z_k(1), z_k(2), \dots, z_k(N-1)), z_k(j) = \sin \frac{k\pi j}{N}.$$

Задача 2. Разложение по сдвигнутым синусам. Пусть сеточная функция $f(j)$, принимающая действительные значения, задана на множестве $\omega^+ = \{x_j = jh, 1 \leq j \leq N\}$ (или на $\bar{\omega}$, причем $f(0) = 0$). В § 5 гл. I было показано, что такая функция $f(j)$ может быть представлена в виде

$$f(j) = \frac{2}{N} \sum_{k=1}^N \varphi_k \sin \frac{(2k-1)\pi j}{2N}, \quad j = 1, 2, \dots, N, \quad (5)$$

где коэффициенты φ_k определяются формулой

$$\varphi_k = \sum_{j=1}^N \rho_j f(j) \sin \frac{(2k-1)\pi j}{2N}, \quad k = 1, 2, \dots, N, \quad (6)$$

а

$$\rho_j = \begin{cases} 1, & j \neq 0, N, \\ 0,5, & j = 0, N. \end{cases} \quad (7)$$

Если функция $f(j)$ задана на множестве $\omega^- = \{x_j = jh, 0 \leq j \leq N-1\}$ (или на $\bar{\omega}$, причем $f(N) = 0$), то аналогичное (5) и (6) разложение имеет вид

$$f(N-j) = \frac{2}{N} \sum_{k=1}^N \varphi_k \sin \frac{(2k-1)\pi j}{2N}, \quad j = 1, 2, \dots, N, \quad (8)$$

$$\varphi_k = \sum_{j=1}^N \rho_{N-j} f(N-j) \sin \frac{(2k-1)\pi j}{2N}, \quad k = 1, 2, \dots, N, \quad (9)$$

где функция ρ_j определена в (7).

Из (5), (6), (8) и (9) следует, что здесь возникают задачи вычисления сумм вида

$$y_k = \sum_{j=1}^N a_j \sin \frac{(2k-1)\pi j}{2N}, \quad k = 1, 2, \dots, N, \quad (10)$$

$$y_j = \sum_{k=1}^N a_k \sin \frac{(2k-1)\pi j}{2N}, \quad j = 1, 2, \dots, N. \quad (10')$$

Задача 3. *Разложение по косинусам.* Пусть действительная сеточная функция $f(j)$ задана на сетке $\bar{\omega}$. Тогда для функции $f(j)$ имеет место разложение

$$f(j) = \frac{2}{N} \sum_{k=0}^N \rho_k \varphi_k \cos \frac{k\pi j}{N}, \quad j = 0, 1, \dots, N, \quad (11)$$

где

$$\varphi_k = \sum_{j=0}^N \rho_j f(j) \cos \frac{k\pi j}{N}, \quad k = 0, 1, \dots, N, \quad (12)$$

а ρ_j определено в (7). Из формул (11) и (12) следует задача вычисления сумм вида

$$y_k = \sum_{j=0}^N a_j \cos \frac{k\pi j}{N}, \quad k = 0, 1, \dots, N. \quad (13)$$

Задача 4. *Преобразование действительной периодической сеточной функции.* Пусть на оси $-\infty < x < \infty$ задана равномерная с шагом h сетка $\Omega = \{x_j = jh, j = 0, \pm 1, \pm 2, \dots, Nh = l\}$. Пусть на сетке Ω задана периодическая с периодом N сеточная функция

$$f(j) = f(j + N), \quad j = 0, \pm 1, \dots,$$

принимающая действительные значения. В § 5 гл. I было показано, что функция $f(j)$ для $0 \leq j \leq N-1$ представима в виде (при четном N)

$$f(j) = \frac{2}{N} \left[\sum_{k=0}^{N/2} \rho_k \varphi_k \cos \frac{2k\pi j}{N} + \sum_{k=1}^{N/2-1} \bar{\varphi}_k \sin \frac{2k\pi j}{N} \right], \quad j = 0, 1, \dots, N-1, \quad (14)$$

где коэффициенты φ_k и $\bar{\varphi}_k$ определяются формулами

$$\varphi_k = \sum_{j=0}^{N-1} f(j) \cos \frac{2k\pi j}{N}, \quad k = 0, 1, \dots, \frac{N}{2}, \quad (15)$$

$$\bar{\varphi}_k = \sum_{j=1}^{N-1} f(j) \sin \frac{2k\pi j}{N}, \quad k = 1, 2, \dots, \frac{N}{2}-1, \quad (16)$$

а функция ρ_k есть

$$\rho_k = \begin{cases} 1, & j \neq 0, N/2, \\ 0, 5, & j = 0, N/2. \end{cases}$$

Формулы (14)–(16) приводят нас к задаче вычисления сумм трех видов:

$$y_k = \sum_{j=0}^{N/2} a_j \cos \frac{2k\pi j}{N} + \sum_{j=1}^{N/2-1} \bar{a}_j \sin \frac{2k\pi j}{N}, \quad k = 0, 1, \dots, N-1, \quad (17)$$

$$\left. \begin{aligned} y_k &= \sum_{j=0}^{N-1} a_j \cos \frac{2k\pi j}{N}, \quad k = 0, 1, \dots, N/2, \\ \bar{y}_k &= \sum_{j=1}^{N-1} a_j \sin \frac{2k\pi j}{N}, \quad k = 1, 2, \dots, N/2-1, \end{aligned} \right\} \quad (18)$$

причем в суммах (18) коэффициенты a_j одни и те же.

Задача 5. Преобразование комплексной периодической сеточной функции. Пусть периодическая с периодом N сеточная функция $f(j)$, заданная на сетке Ω , принимает теперь комплексные значения. Тогда функция $f(j)$ для $0 \leq j \leq N-1$ может быть представлена в виде

$$f(j) = \frac{1}{N} \sum_{k=0}^{N-1} \varphi_k e^{\frac{-2k\pi j}{N} i}, \quad j = 0, 1, \dots, N-1, \quad i = \sqrt{-1}, \quad (19)$$

где комплексные коэффициенты φ_k определены формулой

$$\varphi_k = \sum_{j=0}^{N-1} f(j) e^{\frac{-2k\pi j}{N} i}, \quad k = 0, 1, \dots, N-1. \quad (20)$$

Заметим, что $\varphi_0 = \varphi_N$ и, кроме того,

$$\varphi_{N-k} = \sum_{j=0}^{N-1} f(j) e^{-\frac{2k\pi j}{N}}, \quad k = 0, 1, \dots, N-1.$$

Поэтому вычисление коэффициентов φ_k и восстановление функции $f(j)$ сводятся к вычислению суммы вида

$$y_k = \sum_{j=0}^{N-1} a_j e^{-\frac{2k\pi j}{N}}, \quad k = 0, 1, \dots, N-1 \quad (21)$$

с комплексными a_j .

Итак, нам необходимо построить алгоритмы для вычисления сумм вида (4), (10), (13), (17), (18) и (21), требующие меньшего чем $O(N^2)$ количества арифметических действий. Наиболее просто конструируются алгоритмы для случая, когда N есть степень 2: $N = 2^n$, и мы ограничимся только этим случаем.

2. Разложение по синусам и сдвигнутым синусам. Рассмотрим подробно алгоритм вычисления сумм (4), предполагая, что $N = 2^n$. В этом случае (4) имеет вид

$$y_k = \sum_{j=1}^{2^n-1} a_j^{(0)} \sin \frac{k\pi j}{2^n}, \quad k = 1, 2, \dots, 2^n-1, \quad (22)$$

где введено обозначение $a_j^{(0)} = a_j$.

Идея метода состоит в том, что в сумме (22) члены с общим множителем группируются прежде, чем выполняется умножение. На первом шаге алгоритма члены сумм (22) группируются с индексами j и $2^n - j$ для $j = 1, 2, \dots, 2^n-1$, причем используется равенство

$$\sin \frac{k\pi (2^n - j)}{2^n} = (-1)^{k-1} \sin \frac{k\pi j}{2^n}. \quad (23)$$

Для этого запишем (22) в виде трех слагаемых

$$y_k = \sum_{j=1}^{2^n-1} a_j^{(0)} \sin \frac{k\pi j}{2^n} + \sum_{j=2^n-1+1}^{2^n-1} a_j^{(0)} \sin \frac{k\pi j}{2^n} + a_{2^n-1}^{(0)} \sin \frac{k\pi}{2}$$

и совершим замену $j' = 2^n - j$ во второй сумме. Учитывая (23), получим

$$y_k = \sum_{j=1}^{2^n-1-1} [a_j^{(0)} + (-1)^{k-1} a_{2^n-j}^{(0)}] \sin \frac{k\pi j}{2^n} + a_{2^n-1}^{(0)} \sin \frac{k\pi}{2}. \quad (24)$$

Если обозначить

$$\begin{aligned} a_j^{(1)} &= a_j^{(0)} - a_{2^n-j}^{(0)}, \\ a_{2^n-j}^{(1)} &= a_j^{(0)} + a_{2^n-j}^{(0)}, \quad j = 1, 2, \dots, 2^n-1-1, \\ a_{2^n-1}^{(1)} &= a_{2^n-1}^{(0)}, \end{aligned}$$

то из (24) будем иметь

$$y_{2k-1} = \sum_{j=1}^{2^{n-1}} a_{2^n-j}^{(1)} \sin \frac{(2k-1)\pi j}{2^n}, \quad k = 1, 2, \dots, 2^{n-1}, \quad (25)$$

$$y_{2k} = \sum_{j=1}^{2^{n-1}-1} a_j^{(1)} \sin \frac{k\pi j}{2^{n-1}}, \quad k = 1, 2, \dots, 2^{n-1}-1. \quad (26)$$

Итак, в результате первого шага имеем две суммы вида (25) и (26), каждая из которых содержит примерно в два раза меньше слагаемых, чем исходная сумма (22). Кроме того, суммы вида (26) и исходная сумма имеют аналогичную структуру. Поэтому к (26) можно применить описанный выше способ группировки слагаемых.

На втором шаге, как и выше, при помощи разбиения суммы (26) на три слагаемых и учета равенства (23), где n заменено на $n-1$, группируются члены суммы (26) с индексами j и $2^{n-1}-j$ для $j = 1, 2, \dots, 2^{n-2}-1$. В результате второго шага вместо (26) получим

$$y_{2(2k-1)} = \sum_{j=1}^{2^{n-2}} a_{2^{n-1}-j}^{(2)} \sin \frac{(2k-1)\pi j}{2^{n-1}}, \quad k = 1, 2, \dots, 2^{n-2}, \quad (27)$$

$$y_{2^{2k}} = \sum_{j=1}^{2^{n-2}-1} a_j^{(2)} \sin \frac{k\pi j}{2^{n-2}}, \quad k = 1, 2, \dots, 2^{n-2}-1, \quad (28)$$

где

$$\begin{aligned} a_j^{(2)} &= a_j^{(1)} - a_{2^{n-1}-j}^{(1)}, \\ a_{2^{n-1}-j}^{(2)} &= a_j^{(1)} + a_{2^{n-1}-j}^{(1)}, \quad j = 1, 2, \dots, 2^{n-2}-1, \\ a_{2^{n-2}}^{(2)} &= a_{2^{n-2}}^{(1)}. \end{aligned}$$

Таким образом, исходная задача (22) эквивалентна вычислению сумм (25), (27), (28). Формула (28) позволяет вычислить y_k для k , кратных 4, (27) — для k , кратных 2, но некратных 4 и формула (25) используется для вычисления y_k с нечетным k .

Продолжая процесс преобразования возникающих сумм, получим в результате p -го шага

$$y_{2^{s-1}(2k-1)} = \sum_{j=1}^{2^{n-s}} a_{2^{n-s+1}-j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s+1}}, \quad k = 1, 2, \dots, 2^{n-s}, \quad s = 1, 2, \dots, p, \quad (29)$$

$$y_{2^{p_k}} = \sum_{j=1}^{2^{n-p-1}} a_j^{(p)} \sin \frac{k\pi j}{2^{n-p}}, \quad k = 1, 2, \dots, 2^{n-p}-1,$$

где $p = 1, 2, \dots, n-1$, а коэффициенты $a_j^{(p)}$ определяются рекуррентно

$$\begin{aligned} a_j^{(p)} &= a_j^{(p-1)} - a_{2^n-p+1-j}^{(p-1)}, \\ a_{2^n-p+1-j}^{(p)} &= a_j^{(p-1)} + a_{2^n-p+1-j}^{(p-1)}, \quad j = 1, 2, \dots, 2^{n-p}-1, \\ a_{2^n-p}^{(p)} &= a_{2^n-p}^{(p-1)}, \quad p = 1, 2, \dots, n-1. \end{aligned} \quad (30)$$

Полагая в (29) $p = n-1$, найдем

$$\begin{aligned} y_{2^n-1} &= \sum_{j=1}^{2^{n-s}} a_j^{(n-1)} \sin \frac{\pi j}{2} = a_1^{(n-1)}, \\ y_{2^{s-1}(2k-1)} &= \sum_{j=1}^{2^{n-s}} a_{2^n-s+1-j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s+1}}, \quad k = 1, 2, \dots, 2^{n-s} \end{aligned} \quad (31)$$

для $s = 1, 2, \dots, n-1$.

Итак, исходная задача (22) сведена к вычислению $(n-1)$ -й группы сумм (31). Необходимое для этого преобразование коэффициентов $a_j^{(0)}$ описывается формулами (30).

Второй этап алгоритма состоит в преобразовании сумм (31), которые после замены для каждого фиксированного s

$$\begin{aligned} z_k^{(0)}(1) &= y_{2^{s-1}(2k-1)}, \quad k = 1, 2, \dots, 2^{n-s}, \\ b_j^{(0)}(1) &= a_{2^n-s+1-j}^{(s)}, \quad j = 1, 2, \dots, 2^{n-s}, \\ l &= n-s, \quad s = 1, 2, \dots, n-1, \end{aligned}$$

записываются в следующем виде:

$$z_k^{(0)}(1) = \sum_{j=1}^{2^l} b_j^{(0)}(1) \sin \frac{(2k-1)\pi j}{2^{l+1}}, \quad k = 1, 2, \dots, 2^l, \quad (32)$$

где $l = 1, 2, \dots, n-1$. Здесь коэффициенты $b_j^{(0)}(1)$ и функции $z_k^{(0)}(1)$ зависят от индекса l , но так как мы будем излагать способ вычисления суммы (32) для фиксированного l , то этот индекс всюду опущен.

Займемся преобразованием суммы (32). Представим ее в виде двух слагаемых, разделив члены с четными и нечетными индексами j :

$$\begin{aligned} z_k^{(0)}(1) &= \sum_{i=1}^{2^{l-1}} b_{2i}^{(0)}(1) \sin \frac{(2k-1)\pi i}{2^l} + \\ &\quad + \sum_{j=1}^{2^{l-1}} b_{2j-1}^{(0)}(1) \sin \frac{(2k-1)\pi(2j-1)}{2^{l+1}}. \end{aligned} \quad (33)$$

Используя равенство

$$\begin{aligned} \sin \frac{(2k-1)(2j-2)\pi}{2^{l+1}} + \sin \frac{(2k-1)2j\pi}{2^{l+1}} &= \\ &= 2 \cos \frac{(2k-1)\pi}{2^{l+1}} \sin \frac{(2k-1)(2j-1)\pi}{2^{l+1}}, \end{aligned}$$

пишем второе слагаемое в виде двух сумм:

$$\begin{aligned}
 & \sum_{j=1}^{2^l-1} b_{2j-1}^{(0)}(1) \sin \frac{\pi(2k-1)(2j-1)}{2^{l+1}} = \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l+1}}} \times \\
 & \times \left[\sum_{j=1}^{2^l-1} b_{2j-1}^{(0)}(1) \sin \frac{(2k-1)\pi j}{2^l} + \sum_{j=1}^{2^l-1} b_{2j-1}^{(0)}(1) \sin \frac{(2k-1)\pi(j-1)}{2^l} \right] = \\
 & = \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l+1}}} \left(b_{2^l-1}^{(0)}(1) \sin \frac{(2k-1)\pi}{2} + \sum_{j=1}^{2^l-1-1} \left(b_{2j+1}^{(0)}(1) + \right. \right. \\
 & \quad \left. \left. + b_{2j-1}^{(0)}(1) \right) \sin \frac{(2k-1)\pi j}{2^l} \right). \quad (34)
 \end{aligned}$$

Поясним, что во второй сумме, стоящей в квадратных скобках, была сделана замена индекса $j = j' + 1$.

Обозначим

$$\begin{aligned}
 b_j^{(1)}(1) &= b_{2j-1}^{(0)}(1) + b_{2j+1}^{(0)}(1), \quad j = 1, 2, \dots, 2^{l-1}-1, \\
 b_{2^l-1}^{(1)}(1) &= b_{2^l-1}^{(0)}(1), \\
 b_j^{(1)}(2) &= b_{2j}^{(0)}(1), \quad j = 1, 2, \dots, 2^{l-1}
 \end{aligned}$$

и подставим (34) в (33). Получим выражение

$$\begin{aligned}
 z_k^{(0)}(1) &= \sum_{j=1}^{2^l-1} b_j^{(1)}(2) \sin \frac{(2k-1)\pi j}{2^l} + \\
 & + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l+1}}} \sum_{j=1}^{2^l-1} b_j^{(1)}(1) \sin \frac{(2k-1)\pi j}{2^l},
 \end{aligned}$$

справедливо для $k = 1, 2, \dots, 2^l$. Подставляя сюда вместо k индекс $2^l - k + 1$, получим

$$\begin{aligned}
 z_{2^l-k+1}^{(0)}(1) &= - \sum_{j=1}^{2^l-1} b_j^{(1)}(2) \sin \frac{(2k-1)\pi j}{2^l} + \\
 & + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l+1}}} \sum_{j=1}^{2^l-1} b_j^{(1)}(1) \sin \frac{(2k-1)\pi j}{2^l}.
 \end{aligned}$$

Следовательно, если обозначить

$$\begin{aligned}
 z_k^{(1)}(s) &= \sum_{j=1}^{2^l-1} b_j^{(1)}(s) \sin \frac{(2k-1)\pi j}{2^l}, \\
 k &= 1, 2, \dots, 2^{l-1}, \quad s = 1, 2,
 \end{aligned}$$

то исходная сумма $z_k^{(0)}(1)$ может быть вычислена по формулам

$$z_k^{(0)}(1) = z_k^{(1)}(2) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{\ell+1}}} z_k^{(1)}(1),$$

$$z_{2^{\ell-k+1}}^{(0)}(1) = -z_k^{(1)}(2) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{\ell+1}}} z_k^{(1)}(1), \quad k = 1, 2, \dots, 2^{\ell-1}.$$

Итак, первый шаг привел к возникновению сумм $z_k^{(1)}(1)$ и $z_k^{(1)}(2)$, каждая из которых содержит в два раза меньше слагаемых, чем исходная сумма $z_k^{(0)}(1)$, но имеет ту же структуру, что и $z_k^{(0)}(1)$. В силу этого описанный выше процесс преобразования исходной суммы может быть применен отдельно к суммам $z_k^{(1)}(1)$ и $z_k^{(1)}(2)$. В результате возникнут суммы $z_k^{(2)}(s)$, $s = 1, 2, 3, 4$, сохраняющие структуру исходной суммы. Продолжая процесс преобразований, на m -м шаге получим суммы

$$z_k^{(m)}(s) = \sum_{j=1}^{2^{\ell-m}} b_j^{(m)}(s) \sin \frac{(2k-1)\pi j}{2^{\ell-m+1}}, \quad (35)$$

$$k = 1, 2, \dots, 2^{\ell-m}, \quad s = 1, 2, \dots, 2^m$$

для каждого $m = 0, 1, \dots, l$, где коэффициенты $b_j^{(m)}(s)$ определяются рекуррентно для $s = 1, 2, \dots, 2^{m-1}$ по формулам

$$b_j^{(m)}(2s-1) = b_{2j-1}^{(m-1)}(s) + b_{2j+1}^{(m-1)}(s),$$

$$j = 1, 2, \dots, 2^{\ell-m}-1, \quad m = 1, 2, \dots, l-1,$$

$$b_{2^{\ell-m}}^{(m)}(2s-1) = b_{2^{\ell-m}+1-1}^{(m-1)}(s), \quad m = 1, 2, \dots, l,$$

$$b_j^{(m)}(2s) = b_{2j}^{(m-1)}(s), \quad j = 1, 2, \dots, 2^{\ell-m},$$

$$m = 1, 2, \dots, l. \quad (36)$$

При этом суммы m -го шага связаны с суммами, полученными на $(m-1)$ -м шаге, следующими формулами:

$$z_k^{(m-1)}(s) = z_k^{(m)}(2s) + \frac{1}{2 \cos \frac{\pi(2k-1)}{2^{\ell-m+2}}} z_k^{(m)}(2s-1),$$

$$z_{2^{\ell-m+1}-k+1}^{(m-1)}(s) = -z_k^{(m)}(2s) + \frac{1}{2 \cos \frac{\pi(2k-1)}{2^{\ell-m+2}}} z_k^{(m)}(2s-1), \quad (37)$$

$$k = 1, 2, \dots, 2^{\ell-m}, \quad s = 1, 2, \dots, 2^{m-1}, \quad m = 1, 2, \dots, l.$$

Полагая в (35) $m = l$, получим

$$z_1^{(l)}(s) = b_1^{(l)}(s), \quad s = 1, 2, \dots, 2^l. \quad (38)$$

Итак, суммы $z_k^{(0)}(1)$ вычисляются следующим образом. Исходя из заданных коэффициентов $b_j^{(0)}(1)$, $1 \leq j \leq 2^l$, по формулам (36) вычисляются в итоге коэффициенты $b_1^{(l)}(s)$, $1 \leq s \leq 2^l$. Они используются затем в силу (38) в качестве начальных данных для ре-

рекуррентных соотношений (37). Полагая в (37) последовательно $m = l, l-1, \dots, 1$, получим в результате $z_k^{(0)}(1)$ и, следовательно, $\frac{1}{2} z^{l-1}(2k-1)$.

Таким образом, алгоритм вычисления сумм (22) описывается формулами (30), (36), (38).

Замечание. В рекуррентных соотношениях (37) можно избежать деления на $2 \cos \frac{(2k-1)\pi}{2^{l-m+2}}$ при помощи замены

$$z_k^{(m)}(s) = \sin \frac{(2k-1)\pi}{2^{l-m+1}} w_k^{(m)}(s).$$

При этом формулы для вычисления $w_k^{(m)}(s)$ принимают вид

$$w_k^{(m-1)}(s) = 2 \cos \frac{\pi(2k-1)}{2^{l-m+2}} w_k^{(m)}(2s) + w_k^{(m)}(2s-1),$$

$$w_{2^{l-m+1}-k+1}^{(m-1)}(s) = -2 \cos \frac{\pi(2k-1)}{2^{l-m+2}} w_k^{(m)}(2s) + w_k^{(m)}(2s-1), \quad (39)$$

$k = 1, 2, \dots, 2^l-m$, $s = 1, 2, \dots, 2^{m-1}$, $m = l, l-1, \dots, 1$,

причем $w_1^{(l)}(s) = b_1^{(l)}(s)$, $s = 1, 2, \dots, 2^l$ и

$$z_k^{(0)}(1) = \sin \frac{(2k-1)\pi}{2^{l+1}} w_k^{(0)}(1), \quad k = 1, 2, \dots, 2^l. \quad (40)$$

Подсчитаем теперь число арифметических действий, которое нужно выполнить для реализации алгоритма (30), (36)–(38). Будем предполагать, что значения тригонометрических функций вычислены заранее.

Элементарный подсчет дает:

1) на реализацию (30) требуется

$$Q_1 = \sum_{p=1}^{n-1} 2(2^{n-p}-1) = 2 \cdot 2^n - 2(n+1)$$

операций сложения и вычитания;

2) для фиксированного l на реализацию (36) требуется

$$\bar{q}_l = \sum_{m=1}^{l-1} (2^{l-m}-1) \cdot 2^{m-1} = (l-2) 2^{l-1} + 1$$

операций сложения, а на реализацию (37) требуется

$$\bar{\bar{q}}_l = \sum_{m=1}^l 2 \cdot 2^{l-m} \cdot 2^{m-1} = 2l \cdot 2^{l-1}$$

операций сложения и

$$q_l^* = \sum_{m=1}^l 2^{l-m} \cdot 2^{m-1} = l \cdot 2^{l-1} \quad (41)$$

операций умножения. Всего же формулы (36) и (37) требуют при фиксированном l

$$q_l = \bar{q}_l + \bar{\bar{q}}_l = (3l - 2) \cdot 2^{l-1} + 1 \quad (42)$$

операций сложения и q_l^* умножений. Для всех $l = 1, 2, \dots, n-1$ затраты составят

$$Q_2 = \sum_{l=1}^{n-1} q_l = \sum_{l=1}^{n-1} [(3l - 2) \cdot 2^{l-1} + 1] = \frac{3}{2} n2^n - 4 \cdot 2^n + n + 4$$

сложений и

$$Q_3 = \sum_{l=1}^{n-1} q_l^* = \sum_{l=1}^{n-1} l2^{l-1} = \frac{n}{2} 2^n - 2^n + 1$$

умножений.

Таким образом, алгоритм (30), (36)–(38) характеризуется следующими оценками числа арифметических операций: $Q_+ = Q_1 + Q_2 = (3n/2 - 2)2^n - n + 2$ сложений и $Q_* = (n/2 - 1)2^n + 1$ умножений. Если не делать различия между операциями сложения и умножения, то общее число действий есть

$$Q = Q_1 + Q_2 + Q_3 = (2 \log_2 N - 3)N - \log_2 N + 3, \quad N = 2^n.$$

Для сравнения приведем оценку числа действий, которое нужно выполнить, чтобы вычислить все суммы (22) непосредственным суммированием. Будем иметь $(2^n - 1)^2$ операций умножения и $(2^n - 2)(2^n - 1)$ операций сложения, а всего $\tilde{Q} = (N - 1)(2N - 3)$. Например, для $N = 128$ ($n = 7$) получим $Q = 1404$ операции (из них 321 операция умножения) для построенного алгоритма и $\tilde{Q} = 32\,131$ операция (из них 15 873 операции умножения) для алгоритма непосредственного суммирования.

Отметим, что использование в алгоритме вместо (37) и (38) соотношений (39) и (40) приводит к следующим оценкам числа действий: $Q_+ = \left(\frac{3}{2}n - 2\right)2^n - n + 2$ сложений и $Q_* = \frac{n}{2}2^n - 1$ умножений, а всего $Q = (2 \log_2 N - 2)N - \log_2 N + 1$, $N = 2^n$, что несколько больше, чем в алгоритме (30), (36)–(38).

Итак, поставленная выше задача 1 решена. Рассмотрим теперь задачу 2 о разложении по сдвинутым синусам. Предполагая, что $N = 2^n$, запишем сумму, фигурирующую в задаче 2, в следующем виде:

$$y_k = \sum_{j=1}^{2^n} a_j \sin \frac{(2k-1)\pi j}{2^{n+1}}, \quad k = 1, 2, \dots, 2^n. \quad (43)$$

Сравнивая (43) с (32), находим, что вычисление сумм (43) по сдвинутым синусам является вторым этапом изложенного выше

алгоритма вычисления сумм (22), если в (32) положить $l = n$. Следовательно, если обозначить

$$\begin{aligned} z_k^{(0)}(1) &= y_k, \quad k = 1, 2, \dots, 2^n, \\ b_j^{(0)}(1) &= a_j, \quad j = 1, 2, \dots, 2^n, \end{aligned}$$

то формулы (36) — (38) при $l = n$ описывают алгоритм вычисления сумм (43). Полагая в формулах (41) и (42) $l = n$, получим следующие оценки для построенного алгоритма: $Q_+ = q_n = = \left(\frac{3}{2}n - 1\right)2^n + 1$ операций сложения и $Q_* = q_n^* = \frac{n}{2}2^n$ операций умножения, а всего $Q = (2 \log_2 N - 1)N + 1$, $N = 2^n$. Таким образом, суммы (43) вычисляются примерно с такими же затратами арифметических действий, как и суммы (22).

Напомним, что суммы (43) используются для вычисления коэффициентов Фурье сеточной функции a_i , заданной при $i = 1, 2, \dots, N$. Для восстановления функции по заданным коэффициентам Фурье необходимо вычислить суммы

$$y_j = \sum_{k=1}^{2^n} a_k \sin \frac{(2k-1)\pi j}{2^{n+1}}, \quad j = 1, 2, \dots, 2^n. \quad (43')$$

Используя для $j \neq 2^n$ соотношение

$$\sin \frac{(2k-1)\pi j}{2^{n+1}} = \frac{1}{2 \cos \frac{\pi j}{2^{n+1}}} \left[\sin \frac{(k-1)\pi j}{2^n} + \sin \frac{k\pi j}{2^n} \right],$$

получим

$$\begin{aligned} y_j &= \frac{1}{2 \cos \frac{\pi j}{2^{n+1}}} \left[\sum_{k=1}^{2^n} a_k \sin \frac{(k-1)\pi j}{2^n} + \sum_{k=1}^{2^n} a_k \sin \frac{k\pi j}{2^n} \right] = \\ &= \frac{1}{2 \cos \frac{\pi j}{2^{n+1}}} \sum_{k=1}^{2^n-1} a_k^{(0)} \sin \frac{k\pi j}{2^n}, \quad j = 1, 2, \dots, 2^{n-1}, \end{aligned}$$

где $a_k^{(0)}$ вычисляются по формуле $a_k^{(0)} = a_k + a_{k+1}$, $k = 1, 2, \dots, 2^n - 1$. Сравнивая полученную сумму с (22), находим, что задача свелась к решенной ранее задаче 1.

Для вычисления y_{2^n} получим формулу

$$y_{2^n} = \sum_{k=1}^{2^n} a_k (-1)^{k-1} = \sum_{k=1}^{2^{n-1}} (a_{2k-1} - a_{2k}).$$

Здесь суммирование ведется непосредственно.

Для числа операций изложенного алгоритма верна оценка $Q = 2N \log_2 N - \log_2 N$.

3. Разложение по косинусам. Рассмотрим теперь алгоритм решения задачи 3, состоящей в вычислении сумм (13), при $N = 2^n$.

Имеем

$$y_k = \sum_{j=0}^{2^n} a_j^{(0)} \cos \frac{k\pi j}{2^n}, \quad k = 0, 1, \dots, 2^n, \quad (44)$$

где введено обозначение $a_j^{(0)} = a_j$.

Принцип построения алгоритма совершенно такой же, как и при разложении по синусам, и состоит из двух этапов. На первом этапе группируются слагаемые сумм сначала с индексами j и $2^n - j$ для $j = 0, 1, \dots, 2^{n-1} - 1$, затем с индексами j и $2^{n-1} - j$, $j = 0, 1, \dots, 2^{n-2} - 1$ и т.д.

В результате p -го шага будем иметь

$$\begin{aligned} y_{2^{s-1}(2k-1)} &= \sum_{j=0}^{2^{n-s}-1} a_{2^n-s+1-j}^{(s)} \cos \frac{(2k-1)\pi j}{2^{n-s+1}}, \\ k &= 1, 2, \dots, 2^{n-s}, \quad s = 1, 2, \dots, p, \\ y_{2^p k} &= \sum_{j=0}^{2^{n-p}} a_j^{(p)} \cos \frac{k\pi j}{2^{n-p}}, \quad k = 0, 1, \dots, 2^{n-p}. \end{aligned} \quad (45)$$

Эти формулы справедливы для $p = 1, 2, \dots, n$. Коэффициенты $a_j^{(p)}$ определяются рекуррентно

$$\begin{aligned} a_j^{(p)} &= a_j^{(p-1)} + a_{2^n-p+1-j}^{(p-1)}, \\ a_{2^n-p+1-j}^{(p)} &= a_j^{(p-1)} - a_{2^n-p+1-j}^{(p-1)}, \quad j = 0, 1, \dots, 2^{n-p}-1, \\ a_{2^n-p}^{(p)} &= a_{2^n-p}^{(p-1)}, \quad p = 1, 2, \dots, n. \end{aligned} \quad (46)$$

Полагая в (45) $s = p = n$, найдем

$$y_0 = a_0^{(n)} + a_1^{(n)}, \quad y_{2^n} = a_0^{(n)} - a_1^{(n)}, \quad y_{2^{n-1}} = a_2^{(n)}, \quad (47)$$

а остальные y_k находятся по формулам

$$\begin{aligned} y_{2^{s-1}(2k-1)} &= \sum_{j=0}^{2^{n-s}-1} a_{2^n-s+1-j}^{(s)} \cos \frac{(2k-1)\pi j}{2^{n-s+1}}, \\ k &= 1, 2, \dots, 2^{n-s}, \quad s = 1, 2, \dots, n-1. \end{aligned}$$

Замены для каждого фиксированного s

$$\begin{aligned} z_k^{(0)}(1) &= y_{2^{s-1}(2k-1)}, \quad k = 1, 2, \dots, 2^{n-s}, \\ b_j^{(0)}(1) &= a_{2^n-s+1-j}^{(s)}, \quad j = 0, 1, \dots, 2^{n-s}-1, \\ l &= n-s, \quad s = 1, 2, \dots, n-1 \end{aligned}$$

приводят нас к вычислению следующих сумм:

$$\begin{aligned} z_k^{(0)}(1) &= \sum_{j=0}^{2^l-1} b_j^{(0)}(1) \cos \frac{(2k-1)\pi j}{2^{l+1}}, \quad k = 1, 2, \dots, 2^l, \\ l &= 1, 2, \dots, n-1. \end{aligned} \quad (48)$$

Второй этап алгоритма состоит в вычислении сумм (48). Как и раньше, последовательно разделяя слагаемые с четными и нечетными индексами j , будем иметь следующие рекуррентные соотношения:

$$\begin{aligned} z_k^{(m-1)}(s) &= z_k^{(m)}(2s) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+2}}} z_k^{(m)}(2s-1), \\ z_{2^{l-m+1}-k+1}^{(m-1)}(s) &= z_k^{(m)}(2s) - \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+2}}} z_k^{(m)}(2s-1), \end{aligned} \quad (49)$$

$$k = 1, 2, \dots, 2^{l-m}, \quad s = 1, 2, \dots, 2^{m-1}, \quad m = 1, 2, \dots, l$$

для вычисления

$$z_k^{(m)}(s) = \sum_{j=0}^{2^{l-m}-1} b_j^{(m)}(s) \cos \frac{(2k-1)\pi j}{2^{l-m+1}}, \quad (50)$$

$$k = 1, 2, \dots, 2^{l-m}, \quad s = 1, 2, \dots, 2^m$$

при $m = 0, 1, \dots, l$. Коэффициенты $b_j^{(m)}(s)$ также определяются рекуррентно для $s = 1, 2, \dots, 2^{m-1}$, начиная с $b_j^{(0)}(1)$, по формулам

$$\begin{aligned} b_j^{(m)}(2s-1) &= b_{2j-1}^{(m-1)}(s) + b_{2j+1}^{(m-1)}(s), \\ j &= 1, 2, \dots, 2^{l-m}-1, \quad m = 1, 2, \dots, l-1, \\ b_0^{(m)}(2s-1) &= b_1^{(m-1)}(s), \quad m = 1, 2, \dots, l, \\ b_j^{(m)}(2s) &= b_{2j}^{(m-1)}(s), \\ j &= 0, 1, \dots, 2^{l-m}-1, \quad m = 1, 2, \dots, l. \end{aligned} \quad (51)$$

Полагая в (50) $m = l$, найдем начальные данные для соотношений (49)

$$z_1^{(l)}(s) = b_0^{(l)}(s), \quad s = 1, 2, \dots, 2^l. \quad (52)$$

Таким образом, алгоритм вычисления сумм (44) описывается формулами (46), (47), (49), (51) и (52).

Элементарный подсчет числа арифметических действий для построенного алгоритма дает: $Q_+ = (3/2)n - 2$ $2^n + n + 2$ операций сложения и $Q_* = (n/2 - 1)2^n + 1$ операций умножения, а всего

$$Q = Q_+ + Q_* = (2 \log_2 N - 3)N + \log_2 N + 3, \quad N = 2^n.$$

Заметим, что, как и в предыдущем алгоритме, здесь в соотношениях (49) возможна замена

$$z_k^{(m)}(s) = \sin \frac{(2k-1)\pi}{2^{l-m+1}} w_k^{(m)}(s);$$

при этом из (52) следует $w_1^{(l)}(s) = b_0^{(l)}(s)$, $s = 1, 2, \dots, 2^l$.

Рекуррентные формулы для $w_k^{(m)}(s)$ имеют вид

$$w_k^{(m-1)}(s) = 2 \cos \frac{(2k-1)\pi}{2^l-m+2} w_k^{(m)}(2s) + w_k^{(m)}(2s-1),$$

$$w_{2^l-m+1-k+1}^{(m-1)}(s) = 2 \cos \frac{(2k-1)\pi}{2^l-m+2} w_k^{(m)}(s) - w_k^{(m)}(2s-1),$$

$$k = 1, 2, \dots, 2^l-m, s = 1, 2, \dots, 2^{m-1}, m = 1, 2, \dots, l.$$

4. Преобразование действительной периодической сеточной функции. Задача 4 о преобразовании действительной периодической сеточной функции состоит в восстановлении функции по формулам (17) при заданных коэффициентах Фурье a_j и \bar{a}_j и в нахождении коэффициентов для заданной функции по формулам (18).

Пусть $N = 2^n$ и заданы коэффициенты Фурье. Тогда необходимо вычислить суммы

$$y_k = \sum_{j=0}^{2^n-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n} + \sum_{j=1}^{2^n-1-1} \bar{a}_j^{(0)} \sin \frac{2k\pi j}{2^n}, \quad k = 0, 1, \dots, 2^n-1. \quad (53)$$

Построим соответствующий алгоритм. Для этого заменим в (53) индекс k на 2^n-k . Учитывая равенства

$$\cos \frac{2(2^n-k)\pi j}{2^n} = \cos \frac{2k\pi j}{2^n}, \quad \sin \frac{2(2^n-k)\pi j}{2^n} = -\sin \frac{2k\pi j}{2^n},$$

получим, что y_k можно вычислить по формулам

$$\begin{aligned} y_k &= \bar{y}_k + \bar{\bar{y}}_k, \\ y_{2^n-k} &= \bar{y}_k - \bar{\bar{y}}_k, \quad k = 1, 2, \dots, 2^{n-1}-1, \\ y_0 &= \bar{y}_0, \quad y_{2^n-1} = \bar{y}_{2^n-1}, \end{aligned} \quad (54)$$

где

$$\bar{y}_k = \sum_{j=0}^{2^n-1} a_j^{(0)} \cos \frac{k\pi j}{2^{n-1}}, \quad k = 0, 1, \dots, 2^{n-1}, \quad (55)$$

$$\bar{\bar{y}}_k = \sum_{j=1}^{2^n-1-1} \bar{a}_j^{(0)} \sin \frac{k\pi j}{2^{n-1}}, \quad k = 1, 2, \dots, 2^{n-1}-1. \quad (56)$$

Итак, вычисление сумм (53) сводится к вычислению сумм (55) и (56) и последующему использованию формул (54).

Сравнивая формулы (55) и (56) с формулами (44) и (22), находим, что суммы (55) и (56) можно вычислить по алгоритмам пп. 2 и 3, заменив в них n на $n-1$.

Подсчитаем теперь число арифметических операций, необходимых для вычисления сумм (53) указанным способом. Из оценок числа операций, найденных для алгоритма п. 2, получим, что суммы (56) вычисляются с затратой $Q_+ = (3n/4 - 7/4)2^n - n + 3$ операций сложения и $Q_* = (n/4 - 3/4)2^n + 1$ операций умножения.

Оценки алгоритма п. 3 дают для сумм (55) следующие значения: $Q_+ = (3n/4 - 7/4) 2^n + n + 1$ операций сложения и $Q_* = (n/4 - 3/4) \times 2^n + 1$ операций умножения. Добавляя сюда $Q_+ = 2^n - 2$ операций сложения, затрачиваемых на реализацию (54), получим для построенного алгоритма $Q_+ = (3n/2 - 5/2) 2^n + 2$ операций сложения и $Q_* = (n/2 - 3/2) 2^n + 2$ операций умножения, а всего $Q = (2 \log_2 N - 4) N + 4$, $N = 2^n$.

Обратимся теперь к вычислению коэффициентов Фурье действительной периодической сеточной функции. Задача состоит в нахождении сумм

$$y_k = \sum_{j=0}^{2^n-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n}, \quad k = 0, 1, \dots, 2^{n-1}, \quad (57)$$

$$\bar{y}_k = \sum_{j=1}^{2^n-1} a_j^{(0)} \sin \frac{2k\pi j}{2^n}, \quad k = 1, 2, \dots, 2^{n-1}-1, \quad (58)$$

где $a_j^{(0)}$ — заданная функция.

Алгоритм вычисления (57) и (58) родствен алгоритмам пп. 2 и 3, но отличается некоторыми деталями. Здесь на первом этапе группируются члены сумм (57) и (58) сначала с индексами j и $2^{n-1} + j$ для $j = 0, 1, \dots, 2^{n-1}-1$, затем с индексами j и $2^{n-2} + j$ для $j = 0, 1, \dots, 2^{n-2}-1$ и т. д. Приведем более подробно первый шаг процесса последовательной группировки слагаемых на примере суммы (57). Преобразование суммы (58) осуществляется аналогично.

Итак, представим (57) в следующем виде:

$$y_k = \sum_{j=0}^{2^{n-1}-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n} + \sum_{j=2^{n-1}}^{2^n-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n}$$

и совершим во второй сумме замену, полагая $j = 2^{n-1} + j'$. Это дает

$$y_k = \sum_{j=0}^{2^{n-1}-1} [a_j^{(0)} + (-1)^k a_{2^{n-1}+j}^{(0)}] \cos \frac{2k\pi j}{2^n}, \quad k = 0, 1, \dots, 2^{n-1}.$$

Обозначая

$$\begin{aligned} a_j^{(1)} &= a_j^{(0)} + a_{2^{n-1}+j}^{(0)}, \\ a_{2^{n-1}+j}^{(1)} &= a_j^{(0)} - a_{2^{n-1}+j}^{(0)}, \quad j = 0, 1, \dots, 2^{n-1}-1, \end{aligned} \quad (59)$$

получим вместо (57) следующие суммы:

$$y_{2k-1} = \sum_{j=0}^{2^{n-1}-1} a_{2^{n-1}+j}^{(1)} \cos \frac{(2k-1)\pi j}{2^{n-1}}, \quad k = 1, 2, \dots, 2^{n-2}, \quad (60)$$

$$y_{2k} = \sum_{j=0}^{2^{n-1}-1} a_j^{(1)} \cos \frac{2k\pi j}{2^{n-1}}, \quad k = 0, 1, \dots, 2^{n-2}.$$

Аналогично вместо (58) получим суммы

$$\begin{aligned}\bar{y}_{2k-1} &= \sum_{j=1}^{2^{n-1}-1} a_{2^{n-1}+j}^{(1)} \sin \frac{(2k-1)\pi j}{2^{n-1}}, \quad k=1, 2, \dots, 2^{n-2}, \\ \bar{y}_{2k} &= \sum_{j=1}^{2^{n-1}-1} a_j^{(1)} \sin \frac{2k\pi j}{2^{n-1}}, \quad k=1, 2, \dots, 2^{n-2}-1,\end{aligned}\tag{61}$$

где $a_j^{(1)}$ определены в (59). На этом первый шаг закончен. На втором шаге описанным способом преобразуются суммы (60) и (61). В результате p -го шага будем иметь

$$\begin{aligned}\bar{y}_{2^{s-1}(2k-1)} &= \sum_{j=0}^{2^{n-s-1}-1} a_{2^{n-s}+j}^{(s)} \cos \frac{(2k-1)\pi j}{2^{n-s}}, \\ k &= 1, 2, \dots, 2^{n-s-1}, \quad s=1, 2, \dots, p, \\ \bar{y}_{2^p k} &= \sum_{j=0}^{2^{n-p-1}-1} a_j^{(p)} \cos \frac{2k\pi j}{2^{n-p}}, \quad k=0, 1, \dots, 2^{n-p-1},\end{aligned}\tag{62}$$

где $p=1, 2, \dots, n-1$ и

$$\begin{aligned}\bar{y}_{2^{s-1}(2k-1)} &= \sum_{j=1}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s}}, \\ k &= 1, 2, \dots, 2^{n-s-1}, \quad s=1, 2, \dots, p, \\ \bar{y}_{2^p k} &= \sum_{j=1}^{2^{n-p}-1} a_j^{(p)} \sin \frac{2k\pi j}{2^{n-p}}, \quad k=1, 2, \dots, 2^{n-p-1}-1,\end{aligned}\tag{63}$$

где $p=1, 2, \dots, n-2$. Коэффициенты $a_j^{(p)}$ находятся рекуррентно по формулам

$$\begin{aligned}a_j^{(p)} &= a_j^{(p-1)} + a_{2^{n-p}+j}^{(p-1)}, \quad j=0, 1, \dots, 2^{n-p}-1, \\ a_{2^{n-p}+j}^{(p)} &= a_j^{(p-1)} - a_{2^{n-p}+j}^{(p-1)}, \quad p=1, 2, \dots, n.\end{aligned}\tag{64}$$

Полагая в (62) $p=n-1$ и $s=p=n-1$, получим

$$\begin{aligned}y_0 &= a_0^{(n-1)} + a_1^{(n-1)} = a_0^{(n)}, \\ y_{2^{n-1}} &= a_0^{(n-1)} - a_1^{(n-1)} = a_1^{(n)}, \\ y_{2^{n-2}} &= a_2^{(n-1)},\end{aligned}\tag{65}$$

и (63) при $p = n - 2$ найдем

$$\bar{y}_{2^{n-2}} = a_1^{(n-2)} - a_3^{(n-2)} = a_3^{(n-1)}. \quad (66)$$

Остальные y_k и \bar{y}_k находятся по формулам

$$y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a_{2^n-s+j}^{(s)} \cos \frac{(2k-1)\pi j}{2^{n-s}},$$

$$\bar{y}_{2^{s-1}(2k-1)} = \sum_{j=1}^{2^{n-s}-1} a_{2^n-s+j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s}},$$

$$k = 1, 2, \dots, 2^{n-s-1}, \quad s = 1, 2, \dots, n-2.$$

Совершим здесь замены для фиксированного s :

$$z_k^{(0)}(1) = y_{2^{s-1}(2k-1)}, \quad \bar{z}_k^{(0)}(1) = \bar{y}_{2^{s-1}(2k-1)},$$

$$k = 1, 2, \dots, 2^{n-s-1}, \quad b_j^{(0)}(1) = a_{2^n-s+j}^{(s)}, \quad j = 0, 1, \dots, 2^{n-s}-1,$$

$$l = n-s, \quad s = 1, 2, \dots, n-2.$$

Это приводит нас к вычислению сумм

$$z_k^{(0)}(1) = \sum_{j=0}^{2^l-1} b_j^{(0)}(1) \cos \frac{(2k-1)\pi j}{2^l},$$

$$\bar{z}_k^{(0)}(1) = \sum_{j=1}^{2^l-1} b_j^{(0)}(1) \sin \frac{(2k-1)\pi j}{2^l},$$

$$k = 1, 2, \dots, 2^{l-1}, \quad l = 2, 3, \dots, n-1. \quad (67)$$

На втором этапе алгоритма вычисляются суммы (67). Здесь, как и в алгоритме п. 2, эти суммы преобразуются путем разделения слагаемых с четными и нечетными индексами j и использования равенств

$$\begin{aligned} \sin \frac{(2k-1)(2j-2)\pi}{2^{l-m+1}} + \sin \frac{(2k-1)2j\pi}{2^{l-m+1}} &= \\ &= 2 \cos \frac{(2k-1)\pi}{2^{l-m+1}} \sin \frac{(2k-1)(2j-1)\pi}{2^{l-m+1}}, \\ \cos \frac{(2k-1)(2j-2)\pi}{2^{l-m+1}} + \cos \frac{(2k-1)2j\pi}{2^{l-m+1}} &= \\ &= 2 \cos \frac{(2k-1)\pi}{2^{l-m+1}} \cos \frac{(2k-1)(2j-1)\pi}{2^{l-m+1}} \end{aligned}$$

для $m = 1, 2, \dots$ Это дает следующие рекуррентные формулы:

$$\begin{aligned} z_k^{(m-1)}(s) &= z_k^{(m)}(2s) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}} z_k^{(m)}(2s-1), \\ z_{2^{l-m}-k+1}^{(m-1)}(s) &= z_k^{(m)}(2s) - \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}} z_k^{(m)}(2s-1), \\ \bar{z}_k^{(m-1)}(s) &= \bar{z}_k^{(m)}(2s) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}} \bar{z}_k^{(m)}(2s-1), \\ \bar{z}_{2^{l-m}-k+1}^{(m-1)}(s) &= -\bar{z}_k^{(m)}(2s) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}} \bar{z}_k^{(m)}(2s-1), \end{aligned} \quad (68)$$

$k = 1, 2, \dots, 2^{l-m-1}$, $s = 1, 2, \dots, 2^{m-1}$, $m = 1, 2, \dots, l-1$

для последовательного вычисления сумм

$$\begin{aligned} z_k^{(m)}(s) &= \sum_{j=0}^{2^{l-m}-1} b_j^{(m)}(s) \cos \frac{(2k-1)\pi j}{2^{l-m}}, \\ \bar{z}_k^{(m)}(s) &= \sum_{j=1}^{2^{l-m}-1} b_j^{(m)}(s) \sin \frac{(2k-1)\pi j}{2^{l-m}}, \\ k &= 1, 2, \dots, 2^{l-m-1}, \quad s = 1, 2, \dots, 2^m \end{aligned} \quad (69)$$

при $m = 0, 1, \dots, l-1$.

Коэффициенты $b_j^{(m)}(s)$ также определяются рекуррентно для $s = 1, 2, \dots, 2^{m-1}$, начиная с заданных $b_j^{(0)}(1)$, по формулам

$$\begin{aligned} b_j^{(m)}(2s-1) &= b_{2j-1}^{(m-1)}(s) + b_{2j+1}^{(m-1)}(s), \quad j = 1, 2, \dots, 2^{l-m}-1, \\ b_0^{(m)}(2s-1) &= b_1^{(m-1)}(s) - b_{2^{l-m}+1-1}^{(m-1)}(s), \\ b_j^{(m)}(2s) &= b_{2j}^{(m-1)}(s), \quad j = 0, 1, \dots, 2^{l-m}-1, \\ s &= 1, 2, \dots, 2^{m-1}, \quad m = 1, 2, \dots, l-1. \end{aligned} \quad (70)$$

Полагая в (69) $m = l-1$, получим начальные значения для соотношений (68)

$$z_1^{(l-1)}(s) = b_0^{(l-1)}(s), \quad \bar{z}_1^{(l-1)}(s) = b_1^{(l-1)}(s), \quad s = 1, 2, \dots, 2^{l-1}. \quad (71)$$

Итак, алгоритм одновременного вычисления сумм (57) и (58) описывается формулами (64)–(66), (68), (70) и (71). Заметим, что, как и в алгоритмах пп.2 и 3, здесь в соотношениях (68) возможны замены

$$\begin{aligned} z_k^{(m)}(s) &= \sin \frac{(2k-1)\pi}{2^{l-m}} w_k^{(m)}(s), \\ \bar{z}_k^{(m)}(s) &= \sin \frac{(2k-1)\pi}{2^{l-m}} \bar{w}_k^{(m)}(s), \end{aligned}$$

которые позволяют избежать деления на $2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}$.

Элементарный подсчет числа арифметических операций для построенного алгоритма дает: $Q_+ = 3n/2 \cdot 2^n - 1$ операций сложения и $Q_* = (n/2 - 3/2) 2^n + 2$ операций умножения, а всего $Q = (2 \log_2 N - 3/2) N + 1$, $N = 2^n$.

Таким образом, вычисление коэффициентов Фурье и восстановление действительной периодической сеточной функции по предложенному алгоритму требуют $O(N \ln N)$ арифметических действий.

5. Преобразование комплексной периодической сеточной функции. Рассмотрим теперь задачу 5 о вычислении коэффициентов Фурье и восстановлении комплексной периодической сеточной функции. В п. 1 было показано, что эта задача сводится к вычислению сумм (21), которые в случае $N = 2^n$ имеют вид

$$y_k = \sum_{j=0}^{2^n-1} a_j^{(0)} e^{\frac{2k\pi j}{2^n} i}, \quad k = 0, 1, \dots, 2^n - 1, \quad (72)$$

где $a_j^{(0)}$ — комплексные числа.

Алгоритм для вычисления сумм (72) строится так же, как и алгоритм вычисления коэффициентов Фурье действительной периодической функции. На первом этапе группируются члены сумм (72) сначала с индексами j и $2^{n-1} + j$ для $j = 0, 1, \dots, 2^{n-1} - 1$, затем с индексами j и $2^{n-2} + j$ для $j = 0, 1, \dots, 2^{n-2} - 1$ и т. д. Учитывая равенство $e^{\pi k i} = (-1)^k$, получим в результате p -го шага следующие суммы:

$$\begin{aligned} y_{2^{s-1}(2k-1)} &= \sum_{j=0}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} e^{\frac{(2k-1)\pi j}{2^{n-s}} i}, \\ k &= 1, 2, \dots, 2^{n-s}, \quad s = 1, 2, \dots, p, \\ y_{2^p k} &= \sum_{j=0}^{2^{n-p}-1} a_j^{(p)} e^{\frac{2k\pi j}{2^{n-p}} i}, \quad k = 0, 1, \dots, 2^{n-p} - 1, \end{aligned} \quad (73)$$

где коэффициенты $a_j^{(p)}$ находятся по рекуррентным формулам (64).

Полагая в (73) $s = p = n$, будем иметь

$$y_0 = a_0^{(n)}, \quad y_{2^n-1} = a_1^{(n)}, \quad (74)$$

а остальные y_k находятся по формулам

$$\begin{aligned} y_{2^{s-1}(2k-1)} &= \sum_{j=0}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} e^{\frac{(2k-1)\pi j}{2^{n-s}} i}, \\ k &= 1, 2, \dots, 2^{n-s}, \quad s = 1, 2, \dots, n-1. \end{aligned}$$

Совершим здесь для фиксированного j замены, полагая

$$z_k^{(0)}(1) = y_{2^{s-1}(2k-1)}, \quad k = 1, 2, \dots, 2^{n-s},$$

$$b_j^{(0)}(1) = a_{2^{n-s}+j}^{(s)}, \quad j = 0, 1, \dots, 2^{n-s} - 1,$$

$$l = n - s, \quad s = 1, 2, \dots, n - 1,$$

перейдем к вычислению сумм

$$z_k^{(0)}(1) = \sum_{j=0}^{2^l-1} b_j^{(0)}(1) e^{-\frac{(2k-1)\pi j}{2^l}}, \quad k=1, 2, \dots, 2^l \quad (75)$$

для $l=1, 2, \dots, n-1$.

Второй этап алгоритма, заключающийся в вычислении сумм (75), строится, как и ранее, путем разделения слагаемых с четными и нечетными индексами j при использовании равенств

$$e^{-\frac{(2k-1)(2j-2)\pi}{2^{l-m+1}}} + e^{-\frac{(2k-1)2j\pi}{2^{l-m+1}}} = 2 \cos \frac{(2k-1)\pi}{2^{l-m+1}} e^{-\frac{(2k-1)(2j-1)\pi}{2^{l-m+1}}}.$$

Будем иметь рекуррентные формулы

$$\begin{aligned} z_k^{(m-1)}(s) &= z_k^{(m)}(2s) + \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}} z_k^{(m)}(2s-1), \\ z_{2^{l-m}+k}^{(m-1)}(s) &= z_k^{(m)}(2s) - \frac{1}{2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}} z_k^{(m)}(2s-1), \end{aligned} \quad (76)$$

$$k=1, 2, \dots, 2^{l-m}, s=1, 2, \dots, 2^{m-1}, m=1, 2, \dots, l-1$$

для вычисления сумм

$$z_k^{(m)}(s) = \sum_{j=0}^{2^{l-m}-1} b_j^{(m)}(s) e^{-\frac{(2k-1)\pi j}{2^{l-m}}}, \quad (77)$$

$$k=1, 2, \dots, 2^{l-m}, s=1, 2, \dots, 2^m$$

при $m=0, 1, \dots, l-1$. Коэффициенты $b_j^{(m)}$ вычисляются по рекуррентным формулам (70). Осталось указать начальные значения для (76). Полагая в (77) $m=l-1$, получим

$$\begin{aligned} z_1^{(l-1)}(s) &= b_0^{(l-1)}(s) + i b_1^{(l-1)}(s), \\ z_{2^{l-1}}^{(l-1)}(s) &= b_0^{(l-1)}(s) - i b_1^{(l-1)}(s), \quad s=1, 2, \dots, 2^{l-1}. \end{aligned} \quad (78)$$

Итак, алгоритм вычисления сумм (72) описывается формулами (64), (70), (74), (76) и (78). Отметим, что построенный алгоритм не содержит (за исключением простейшей формулы (78)) операций умножения комплексных чисел. Поэтому в приведенных формулах легко выделить действительную и мнимую части вычисляемых величин. Это удобно для реализации алгоритма на ЭВМ, не имеющей комплексной арифметики. Далее, в соотношениях (76) может оказаться полезной замена

$$z_k^{(m)}(s) = \sin \frac{(2k-1)\pi}{2^{l-m}} w_k^{(m)}(s).$$

Подсчитаем теперь число арифметических операций для построенного алгоритма. Получим $Q_+ = (3n/2 - 1/2)2^n$ операций сло-

жения комплексных чисел и $Q_* = (n/2 - 3/2) 2^n$ операций умножения комплексного числа на действительное число. Если выразить эти значения в терминах числа операций над действительными числами, то получим $Q_+ = (3n - 1) 2^n$ действительных операций сложения и $Q_* = (n - 3) 2^n$ действительных операций умножения, а всего $Q = (4 \log_2 N - 4) N$, $N = 2^n$ операций над действительными числами. Эта оценка в два раза превосходит полученную в п. 4 оценку для случая действительной периодической сеточной функции, что является естественным, поскольку в рассматриваемом комплексном случае обрабатывается в два раза больше действительных чисел.

На этом мы заканчиваем рассмотрение алгоритмов быстрого дискретного преобразования Фурье и переходим к использованию их для решения сеточных эллиптических уравнений.

§ 2. Решение разностных задач методом Фурье

1. Разностные задачи на собственные значения для оператора Лапласа в прямоугольнике. В § 5 гл. I были рассмотрены краевые задачи на собственные значения для оператора второй разностной производной, заданного на равномерной сетке на отрезке. В двумерном случае аналогами этих задач являются задачи на собственные значения для разностного оператора Лапласа, заданного на равномерной прямоугольной сетке в прямоугольнике. Воспользуемся методом разделения переменных для отыскания собственных значений λ_k и собственных функций $\mu_k(i, j)$ разностного оператора Лапласа

$$\Lambda = \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2.$$

Пусть в прямоугольнике $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ задана прямоугольная равномерная сетка $\bar{\omega}$ с шагами h_1 и h_2 : $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$. Через ω обозначим, как обычно, внутренние, а через γ — граничные узлы сетки $\bar{\omega}$.

Простейшая задача на собственные значения для оператора Лапласа в случае условий Дирихле ставится так: найти такие значения параметра λ , при которых существуют нетривиальные решения $y(x)$ следующей задачи:

$$\begin{aligned} \Lambda y(x) + \lambda y(x) &= 0, & x \in \omega, \\ y(x) &= 0, & x \in \gamma. \end{aligned} \tag{1}$$

Будем искать собственную функцию $\mu_k(i, j)$ задачи (1) соответствующую собственному значению λ_k , в виде

$$\mu_k(i, j) = \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \quad k = (k_1, k_2). \tag{2}$$

Подставим в (1) вместо $y(x_{ij}) = y(i, j)$ функцию $\mu_k(i, j)$. Так как

$$\Lambda_1 y(i, j) = \frac{1}{h_1^2} [y(i+1, j) - 2y(i, j) + y(i-1, j)],$$

то оператор Λ_1 действует лишь на сеточную функцию, зависящую от аргумента i . Аналогично оператор Λ_2 действует на функцию, зависящую от аргумента j . Поэтому после подстановки (2) в (1) будем иметь

$$\mu_{k_1}^{(2)}(j) \Lambda_1 \mu_{k_1}^{(1)}(i) + \mu_{k_1}^{(1)}(i) \Lambda_2 \mu_{k_2}^{(2)}(j) + \lambda_k \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j) = 0 \quad (3)$$

для $1 \leq i \leq N_1 - 1$, $1 \leq j \leq N_2 - 1$, а также

$$\mu_{k_1}^{(1)}(0) = \mu_{k_1}^{(1)}(N_1) = 0, \quad \mu_{k_2}^{(2)}(0) = \mu_{k_2}^{(2)}(N_2) = 0. \quad (4)$$

Из (3) находим, что

$$\frac{\Lambda_1 \mu_{k_1}^{(1)}(i)}{\mu_{k_1}^{(1)}(i)} = -\frac{\Lambda_2 \mu_{k_2}^{(2)}(j)}{\mu_{k_2}^{(2)}(j)} - \lambda_k. \quad (5)$$

Так как левая часть не зависит от j , то не зависит от j и правая часть. С другой стороны, так как правая часть не зависит от i , то не зависит от i и левая часть. Тем самым, и левая и правая части в (5) есть постоянные. Положим

$$\frac{\Lambda_1 \mu_{k_1}^{(1)}(i)}{\mu_{k_1}^{(1)}(i)} = -\lambda_{k_1}^{(1)}, \quad \frac{\Lambda_2 \mu_{k_2}^{(2)}(j)}{\mu_{k_2}^{(2)}(j)} = -\lambda_{k_2}^{(2)}, \quad \lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)} \quad (6)$$

и добавим сюда краевые условия (4). В результате получим одномерные сеточные задачи на собственные значения

$$\begin{aligned} \Lambda_1 \mu_{k_1}^{(1)} + \lambda_{k_1}^{(1)} \mu_{k_1}^{(1)} &= 0, \quad 1 \leq i \leq N_1 - 1, \\ \mu_{k_1}^{(1)}(0) &= \mu_{k_1}^{(1)}(N_1) = 0 \end{aligned} \quad (7)$$

и

$$\begin{aligned} \Lambda_2 \mu_{k_2}^{(2)} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} &= 0, \quad 1 \leq j \leq N_2 - 1, \\ \mu_{k_2}^{(2)}(0) &= \mu_{k_2}^{(2)}(N_2) = 0. \end{aligned} \quad (8)$$

Решения задач (7) и (8) были нами найдены ранее в § 5 гл. I:

$$\lambda_{k_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi}{2N_\alpha} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi h_\alpha}{2l_\alpha}, \quad k_\alpha = 1, 2, \dots, N_\alpha - 1,$$

$$\mu_{k_1}^{(1)}(i) = \sqrt{\frac{2}{l_1}} \sin \frac{k_1 \pi i}{N_1}, \quad k_1 = 1, 2, \dots, N_1 - 1,$$

$$\mu_{k_2}^{(2)}(j) = \sqrt{\frac{2}{l_2}} \sin \frac{k_2 \pi j}{N_2}, \quad k_2 = 1, 2, \dots, N_2 - 1.$$

Итак, собственные функции и собственные значения разностного оператора Лапласа Λ для случая граничных условий Дирихле найдены

$$\begin{aligned}\mu_k(i, j) &= \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j) = \frac{2}{\sqrt{l_1 l_2}} \sin \frac{k_1 \pi i}{N_1} \sin \frac{k_2 \pi j}{N_2}, \\ 0 &\leq i \leq N_1, \quad 0 \leq j \leq N_2, \\ \lambda_k &= \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)} = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi h_\alpha}{2l_\alpha},\end{aligned}\quad (9)$$

где $k_\alpha = 1, 2, \dots, N_\alpha - 1$, $\alpha = 1, 2$.

Отметим основные свойства найденных собственных функций и собственных значений (9). Введем скалярное произведение сеточных функций, заданных на сетке $\bar{\omega}$, следующим образом:

$$\begin{aligned}(u, v) &= \sum_{x \in \bar{\omega}} u(x)v(x)\bar{h}_1(x_1)\bar{h}_2(x_2), \\ \bar{h}_\alpha(x_\alpha) &= \begin{cases} 0,5h_\alpha, & x_\alpha = 0, l_\alpha, \\ h_\alpha, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha. \end{cases}\end{aligned}$$

Если обозначить

$$(u, v)_{\bar{\omega}_\alpha} = \sum_{x_\alpha \in \bar{\omega}_\alpha} u(x)v(x)\bar{h}_\alpha(x_\alpha), \quad \alpha = 1, 2, \quad (10)$$

где

$\bar{\omega}_1 = \{x_1(i) = ih_1, 0 \leq i \leq N_1\}$, $\bar{\omega}_2 = \{x_2(j) = jh_2, 0 \leq j \leq N_2\}$,
то, очевидно, что $\bar{\omega} = \bar{\omega}_1 \times \bar{\omega}_2$ и $x_{ij} = (x_1(i), x_2(j))$, кроме того,

$$(u, v) = ((u, v)_{\bar{\omega}_1}, 1)_{\bar{\omega}_2} = ((u, v)_{\bar{\omega}_2}, 1)_{\bar{\omega}_1}. \quad (11)$$

Напомним, что в § 5 гл. I было отмечено, что сеточные функции $\mu_{k_1}^{(1)}(i)$ и $\mu_{k_2}^{(2)}(j)$ ортонормированы в смысле скалярного произведения (10), т. е.

$$(\mu_{k_\alpha}^{(\alpha)}, \mu_{m_\alpha}^{(\alpha)})_{\bar{\omega}_\alpha} = \delta_{k_\alpha, m_\alpha} = \begin{cases} 1, & k_\alpha = m_\alpha, \\ 0, & k_\alpha \neq m_\alpha. \end{cases}$$

Поэтому отсюда и из (11) следует ортонормированность заданной формулами (9) системы собственных функций $\mu_k(i, j)$:

$$(\mu_k, \mu_m) = \delta_{k, m} = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \quad k = (k_1, k_2), \quad m = (m_1, m_2). \end{cases}$$

Так как число собственных функций $\mu_k(i, j) = \mu_{k_1, k_2}(i, j)$ равно $(N_1 - 1)(N_2 - 1)$ и совпадает с числом внутренних узлов сетки ω , то любую сеточную функцию $f(i, j)$, заданную на ω (или на

$\bar{\omega}$ и обращающуюся в нуль на γ), можно представлять в следующем виде:

$$f(i, j) = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} f_{k_1 k_2} \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \quad (12)$$

$$1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1,$$

где коэффициенты Фурье $f_{k_1 k_2}$ определяются следующим образом:

$$f_k = f_{k_1 k_2} = (f, \mu_k) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} f(i, j) \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j) h_1 h_2, \quad (13)$$

$$k_1 = 1, 2, \dots, N_1 - 1, \quad k_2 = 1, 2, \dots, N_2 - 1.$$

Для собственных значений λ_k справедлива оценка

$$\lambda_{\min} = \lambda_1^{(1)} + \lambda_1^{(2)} \leq \lambda_k = \lambda_{k_1} + \lambda_{k_2} \leq \lambda_{N_1-1}^{(1)} + \lambda_{N_2-1}^{(2)} = \lambda_{\max},$$

где

$$\lambda_{\min} = \sum_{\alpha=1}^2 \frac{4}{h_{\alpha}^2} \sin^2 \frac{\pi h_{\alpha}}{2l_{\alpha}} \geq 8 \left(\frac{1}{l_1^2} + \frac{1}{l_2^2} \right) > 0,$$

$$\lambda_{\max} = \sum_{\alpha=1}^2 \frac{4}{h_{\alpha}^2} \cos^2 \frac{\pi h_{\alpha}}{2l_{\alpha}} < 4 \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right).$$

Рассмотрим теперь пример более сложной задачи на собственные значения для разностного оператора Лапласа. Пусть на сторонах прямоугольника при $x_1 = 0$ и $x_1 = l_1$ по-прежнему заданы условия Дирихле, а при $x_2 = 0$ и $x_2 = l_2$ — условия Неймана, т. е. поставлена следующая задача на собственные значения:

$$\begin{aligned} \Lambda y(x) + \lambda y(x) &= 0, \quad x \in \omega_1 \times \bar{\omega}_2, \\ y(x) &= 0, \quad x_1 = 0, \quad x_1 = l_1. \end{aligned} \quad (14)$$

Здесь $\Lambda = \Lambda_1 + \Lambda_2$, оператор Λ_1 определен ранее, а

$$\Lambda_2 y = \begin{cases} \frac{2}{h_2} y_{x_2}, & x_2 = 0, \\ y_{x_2 x_2}, & h_2 \leq x_2 \leq l_2 - h_2, \\ -\frac{2}{h_2} y_{x_2}, & x_2 = l_2. \end{cases} \quad (15)$$

Используя определение операторов Λ_1 и Λ_2 , задачу (14) можно записать в следующем виде:

$$\begin{aligned} y_{x_1 x_1} + y_{x_2 x_2} + \lambda y &= 0, \quad x \in \omega, \\ y_{x_1 x_1} + \frac{2}{h_2} y_{x_2} + \lambda y &= 0, \quad x_2 = 0, \\ y_{x_1 x_1} - \frac{2}{h_2} y_{x_2} + \lambda y &= 0, \quad x_2 = l_2, \\ y(0, x_2) &= y(l_1, x_2) = 0, \quad 0 \leq x_2 \leq l_2. \end{aligned} \quad \left. \right\} h_1 \leq x_1 \leq l_1 - h_1,$$

Решение задачи (14) находится методом разделения переменных. Подставляя в (14) вместо y сеточную функцию $\mu_k(i, j)$ из (2), получим для $\mu_{k_1}^{(1)}(i)$ задачу (7), а для $\mu_{k_2}^{(2)}(j)$ будем иметь следующую краевую задачу:

$$\Lambda_2 \mu_{k_2}^{(2)} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} = 0, \quad 0 \leq j \leq N_2$$

или в силу (15)

$$\begin{aligned} (\mu_{k_2}^{(2)})_{x_2 x_2} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} &= 0, \quad 1 \leq j \leq N_2 - 1, \\ \frac{2}{h_2} (\mu_{k_2}^{(2)})_{x_2} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} &= 0, \quad j = 0, \\ -\frac{2}{h_2} (\mu_{k_2}^{(2)})_{x_2} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} &= 0, \quad j = N_2. \end{aligned} \quad (16)$$

Задача (16) также была решена нами ранее в § 5 гл. I. Решение имеет вид

$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi h_2}{2l_2}, \quad k_2 = 0, 1, \dots, N_2,$$

$$\mu_{k_2}^{(2)}(j) = \begin{cases} \sqrt{\frac{2}{l_2}} \cos \frac{k_2 \pi j}{N_2}, & 1 \leq k_2 \leq N_2 - 1, \\ \sqrt{\frac{1}{l_2}} \cos \frac{k_2 \pi j}{N_2}, & k_2 = 0, N_2. \end{cases} \quad (17)$$

Итак, решение задачи (14), (15) найдено:

$$\begin{aligned} \mu_k(i, j) &= \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leq i \leq N_1, \quad 0 \leq j \leq N_2, \\ \lambda_k &= \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad 1 \leq k_1 \leq N_1 - 1, \quad 0 \leq k_2 \leq N_2, \end{aligned}$$

где $\lambda_{k_1}^{(1)}$ и $\mu_{k_1}^{(1)}(i)$ определены выше, а $\lambda_{k_2}^{(2)}$ и $\mu_{k_2}^{(2)}(j)$ определены в (17).

Аналогично решаются задачи на собственные значения для разностного оператора Лапласа в прямоугольнике и в случае других комбинаций краевых условий на сторонах прямоугольника G . Метод разделения переменных позволяет свести их к одномерным задачам, решения которых получены в § 5 гл. I. Обобщение на многомерный случай очевидно. Напомним, что аналитическое решение соответствующих одномерных задач в виде синусов и косинусов было получено в § 5 гл. I лишь для краевых условий первого и второго рода, их комбинаций, а также для случая периодической краевой задачи. Поэтому, если на сторонах прямоугольника (на гранях прямоугольного параллелепипеда в трехмерном случае) заданы перечисленные краевые условия, то собственные функции разностного оператора Лапласа представляются в виде произведения синусов и косинусов.

2. Уравнение Пуассона в прямоугольнике. Разложение в двойной ряд. Рассмотрим теперь метод разделения переменных применительно к решению разностной задачи Дирихле для уравнения Пуассона на равномерной сетке в прямоугольнике:

$$\begin{aligned} Ly &= -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \\ \Lambda &= \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha y = y_{x_\alpha x_\alpha}, \quad \alpha = 1, 2. \end{aligned} \quad (18)$$

Сначала сведем задачу (18) к задаче с однородным граничным условием путем изменения правой части уравнения в приграничных узлах. Стандартный прием такого преобразования состоит в перенесении известных величин в правую часть уравнения, записанного в приграничном узле. Например, если $x = (h_1, h_2) \in \omega$, то уравнение Пуассона в этой точке записывается в следующем виде:

$$\begin{aligned} \frac{1}{h_1^2} [y(0, h_2) - 2y(h_1, h_2) + y(2h_1, h_2)] + \\ + \frac{1}{h_2^2} [y(h_1, 0) - 2y(h_1, h_2) + y(h_1, 2h_2)] = -\varphi(h_1, h_2). \end{aligned}$$

Так как $y(0, h_2) = g(0, h_2)$, $y(h_1, 0) = g(h_1, 0)$, то перенося эти величины из левой в правую часть уравнения, будем иметь

$$\begin{aligned} \frac{1}{h_1^2} [-2y(h_1, h_2) + y(2h_1, h_2)] + \frac{1}{h_2^2} [-2y(h_1, h_2) + y(h_1, 2h_2)] = \\ = -[\varphi(h_1, h_2) + \frac{1}{h_1^2} g(0, h_2) + \frac{1}{h_2^2} g(h_1, 0)]. \end{aligned}$$

Проведя подобное преобразование для каждой приграничной точки, получим разностные уравнения, не содержащие значений $y(x)$ на γ в левой части. Правые части уравнений для приграничных узлов отличаются от правой части $\varphi(x)$. Если обозначить через $f(x)$ построенную правую часть, то она определяется формулами

$$f(x) = \varphi(x) + \frac{1}{h_1^2} \varphi_1(x) + \frac{1}{h_2^2} \varphi_2(x), \quad x \in \omega, \quad (19)$$

где

$$\varphi_1(x) = \begin{cases} g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leq x_1 \leq l_1 - 2h_1, \\ g(l_1, x_2), & x_1 = l_2, \end{cases} \quad \varphi_2(x) = \begin{cases} g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leq x_2 \leq l_2 - 2h_2, \\ g(x_1, l_2), & x_2 = l_2. \end{cases}$$

Левая часть преобразованных уравнений отличается для приграничных узлов от записи разностного оператора Лапласа. Однако

если положить $y(x) = u(x)$, $x \in \omega$, $u(x) = 0$, $x \in \gamma$, то уравнения во всех узлах сетки ω будут записываться одинаково:

$$\Lambda u = -f(x), \quad x \in \omega, \quad u(x) = 0, \quad x \in \gamma. \quad (20)$$

Так как $u(x)$ совпадает с $y(x)$ для $x \in \omega$, то достаточно найти решение задачи (20).

Найдем решение задачи (20). Так как функция $u(x)$ обращается в нуль на γ , то в силу сказанного выше она может быть представлена в виде разложения по собственным функциям $\mu_k(i, j)$ оператора Лапласа

$$u(i, j) = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} u_{k_1 k_2} \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \quad (21)$$

что справедливо для $0 \leq i \leq N_1$, $0 \leq j \leq N_2$. Далее, сеточная функция $f(x)$, заданная на ω , также допускает представление

$$f(i, j) = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} f_{k_1 k_2} \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j) \quad (22)$$

для $1 \leq i \leq N_1 - 1$, $1 \leq j \leq N_2 - 1$, где коэффициенты Фурье $f_{k_1 k_2}$ определены в (13). Так как $\mu_k(i, j) = \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j)$ есть собственная функция оператора Лапласа, соответствующая собственному значению λ_k , т. е.

$$\Lambda \mu_k + \lambda_k \mu_k = 0, \quad x \in \omega, \quad \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)} = \lambda_k,$$

то после подстановки (21) и (22) в уравнение (20) будем иметь

$$\begin{aligned} \Lambda u & \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} (\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}) u_{k_1 k_2} \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j) = \\ & = -f(i, j) = -\sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} f_{k_1 k_2} \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \\ & 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1. \end{aligned}$$

Используя ортонормированность собственных функций $\mu_k(i, j)$, отсюда получим следующие равенства:

$$u_{k_1 k_2} = \frac{f_{k_1 k_2}}{\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}}, \quad 1 \leq k_1 \leq N_1 - 1, \quad 1 \leq k_2 \leq N_2 - 1.$$

Подставляя это выражение в (21), получим для решения задачи (20) следующее представление:

$$\begin{aligned} u(i, j) & = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} \frac{f_{k_1 k_2}}{\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}} \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \\ & 0 \leq i \leq N_1, \quad 0 \leq j \leq N_2. \end{aligned} \quad (23)$$

Итак, формулы (13) и (23) дают решение задачи (20). Проанализируем их с вычислительной точки зрения. При вычислении

решения $u(i, j)$ по формулам (13) и (20), где $\mu_k(i, j) = \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j)$ и $\lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}$ определены в (9), целесообразно ввести три вспомогательные величины: $\varphi_{k_2}(i)$, $\varphi_{k_1 k_2}$ и $u_{k_2}(i)$. Тогда вычислительный процесс можно организовать следующим образом:

$$\varphi_{k_2}(i) = \sum_{j=1}^{N_2-1} f(i, j) \sin \frac{k_2 \pi j}{N_2}, \quad (24)$$

$$\varphi_{k_1 k_2} = \sum_{i=1}^{N_1-1} \varphi_{k_2}(i) \sin \frac{k_1 \pi i}{N_1}, \quad (25)$$

$$u_{k_2}(i) = \sum_{k_1=1}^{N_1-1} \frac{\varphi_{k_1 k_2}}{\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}} \sin \frac{k_1 \pi i}{N_1}, \quad (26)$$

$$u(i, j) = \frac{4}{N_1 N_2} \sum_{k_2=1}^{N_2-1} u_{k_2}(i) \sin \frac{k_2 \pi j}{N_2}, \quad (27)$$

$$1 \leqslant j \leqslant N_2 - 1, \quad 1 \leqslant i \leqslant N_1 - 1.$$

Подсчитаем число арифметических действий для алгоритма (24)–(27), предполагая, что величины $(\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)})^{-1}$ заданы, а суммы (24)–(27) вычисляются с использованием алгоритма быстрого преобразования Фурье, изложенного в п. 2 § 1. Для того чтобы применить указанный алгоритм, нужно предположить, что N_1 и N_2 есть степени 2: $N_1 = 2^n$, $N_2 = 2^m$.

Напомним, что суммы вида

$$y_k = \sum_{j=1}^{2^n-1} a_j \sin \frac{k \pi j}{2^n}, \quad k = 1, 2, \dots, 2^n - 1,$$

вычисляются с затратой $Q_+ = (3/2n - 2) 2^n - n + 2$ операций сложения и вычитания и $Q_* = (n/2 - 1) 2^n + 1$ операций умножения, если используется алгоритм из п. 2 § 1.

Элементарный подсчет дает следующие затраты арифметических действий для вычисления решения $u(i, j)$ по формулам (24)–(27):

$$Q_+ = (N_1 N_2 - N_1 - N_2) [3 \log_2 (N_1 N_2) - 8] + (N_1 + 2) \log_2 N_2 + (N_2 + 2) \log_2 N_1 - 8$$

операций сложения и вычитания и

$$Q_* = (N_1 N_2 - N_1 - N_2) [\log_2 (N_1 N_2) - 2] + N_1 \log_2 N_2 + N_1 \log_2 N_1 - 2$$

операций умножения. Если не делать различия между арифметическими операциями, то при $N_1 = N_2 = N = 2^n$ общее число действий для алгоритма (24)–(27) составит

$$Q = (N^2 - 1,5N)(8 \log_2 N - 10) + 5N + 4 \log_2 N - 10.$$

Итак, описанный метод решения задачи (20) может быть реализован с затратой $O(N^2 \log_2 N)$ арифметических действий. Такого типа оценку для числа действий имеет и рассмотренный в главе III метод полной редукции. Сравнение этих оценок показывает, что данный алгоритм метода разделения переменных требует примерно в 1,5 раз больше действий, чем метод полной редукции.

Отметим, что можно построить алгоритм, аналогичный предложенному выше, и для случая, когда на сторонах прямоугольника задается любая комбинация из краевых условий первого или второго рода и условий периодичности, при которых разностная задача не вырождена. Необходимо лишь подставить в (13) и (23) соответствующие собственные функции и собственные значения, согласовать с типом краевых условий пределы суммирования и использовать соответствующий алгоритм быстрого преобразования Фурье из § 1 для вычисления возникающих при этом сумм. Оценка числа действий будет такого же вида, как и для рассмотренного выше случая задачи Дирихле.

Мы описали простейший вариант метода разделения переменных. Если же требуется решить более общую разностную краевую задачу, например уравнение Пуассона в полярной или цилиндрической системах координат с краевыми условиями, допускающими разделение переменных, то снова можно использовать разложения (21) и (22). Но в этом случае по крайней мере одна из собственных функций $\mu_{k_1}^{(1)}(i)$ или $\mu_{k_2}^{(2)}(j)$ отлична от синуса или косинуса. Это не позволяет воспользоваться алгоритмом быстрого преобразования Фурье при вычислении всех необходимых сумм. Поэтому для таких задач число арифметических действий будет того же порядка, что и в случае непосредственного вычисления сумм без учета вида собственных функций $\mu_{k_1}^{(1)}(i)$ и $\mu_{k_2}^{(2)}(j)$, т. е. $O(N^3)$.

Следовательно, необходимо модифицировать построенный метод, чтобы в случае, когда хотя бы одна из функций $\mu_{k_1}^{(1)}(i)$ или $\mu_{k_2}^{(2)}(j)$ есть синус или косинус, число арифметических действий оставалось величиной порядка $O(N^2 \log_2 N)$. Разумеется, что рассмотренные в этом пункте задачи могут быть решены и модифицированным методом, и, как окажется ниже, с меньшим числом арифметических действий. Этот метод—разложения в однократный ряд—будет построен в п. 3. С вычислительной точки зрения он отличается от построенного здесь метода тем, что две суммы из (24)–(27) не вычисляются, а вместо них решается серия краевых задач для трехточечных разностных уравнений.

3. Разложение в однократный ряд. Вернемся к задаче (20):

$$\begin{aligned} \Lambda u &= -f(x), \quad x \in \omega, \quad u(x) = 0, \quad x \in \gamma, \\ \Lambda &= \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha u = u_{x_\alpha x_\alpha}, \quad \alpha = 1, 2. \end{aligned} \quad (28)$$

Будем рассматривать исковую функцию $u(x_{ij}) = u(i, j)$ и заданную $f(i, j)$ при фиксированном i , $0 \leq i \leq N_1$ как сеточные функции аргумента j . Так как $u(i, j)$ обращается в нуль при $j = 0$ и $j = N_2$, а $f(i, j)$ задана для $1 \leq j \leq N_2 - 1$, то они могут быть представлены в виде сумм по собственным функциям $\mu_{k_2}^{(2)}(j)$ разностного оператора Λ_2 :

$$u(i, j) = \sum_{k_2=1}^{N_2-1} u_{k_2}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leq j \leq N_2, \quad 0 \leq i \leq N_1, \quad (29)$$

$$f(i, j) = \sum_{k_2=1}^{N_2-1} f_{k_2}(i) \mu_{k_2}^{(2)}(j), \quad 1 \leq j \leq N_2 - 1, \quad 1 \leq i \leq N_1 - 1, \quad (30)$$

где

$$\mu_{k_2}^{(2)}(j) = \sqrt{\frac{2}{l_2}} \sin \frac{k_2 \pi j}{N_2}, \quad k_2 = 1, 2, \dots, N_2 - 1. \quad (31)$$

Подставим выражения (29) и (30) в (28) и учтем равенства

$$\begin{aligned} \Lambda_2 \mu_{k_2}^{(2)} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} &= 0, \quad 1 \leq j \leq N_2 - 1, \\ \mu_{k_2}^{(2)}(0) &= \mu_{k_2}^{(2)}(N_2) = 0. \end{aligned} \quad (32)$$

В результате получим

$$\sum_{k_2=1}^{N_2-1} [\Lambda_1 u_{k_2}(i) - \lambda_{k_2}^{(2)} u_{k_2}(i) + f_{k_2}(i)] \mu_{k_2}^{(2)}(j) = 0$$

для $1 \leq i \leq N_1 - 1$, $1 \leq j \leq N_2 - 1$, а также $u_{k_2}(0) = u_{k_2}(N_1) = 0$, $k_2 = 1, 2, \dots, N_2 - 1$.

Отсюда, в силу ортогональности системы собственных функций $\mu_{k_2}^{(2)}(j)$, получим серию краевых задач для определения функций $u_{k_2}(i)$, $k_2 = 1, 2, \dots, N_2 - 1$:

$$\begin{aligned} \Lambda_1 u_{k_2}(i) - \lambda_{k_2}^{(2)} u_{k_2}(i) &= -f_{k_2}(i), \quad 1 \leq i \leq N_1 - 1, \\ u_{k_2}(0) &= u_{k_2}(N_1) = 0. \end{aligned} \quad (33)$$

Собственные значения $\lambda_{k_2}^{(2)}$ задачи (32) известны

$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2}, \quad k_2 = 1, 2, \dots, N_2 - 1, \quad (34)$$

а коэффициенты Фурье $f_{k_2}(i)$ для каждого $1 \leq i \leq N_1 - 1$ вычисляются по формулам

$$f_{k_2}(i) = (f, \mu_{k_2}^{(2)})_{\bar{\omega}_2} = \sum_{j=1}^{N_2-1} h_2 f(i, j) \mu_{k_2}^{(2)}(j), \quad 1 \leq k_2 \leq N_2 - 1. \quad (35)$$

Итак, найденные формулы (29), (31) и (33)–(35) полностью описывают метод решения задачи (20). По формулам (35) на-

ходятся для $1 \leq i \leq N_1 - 1$ функции $f_{k_2}(i)$, затем для $1 \leq k_2 \leq N_2 - 1$ решаются задачи (33) для определения функций $u_{k_2}(i)$, а по формулам (29) вычисляется искомое решение $u(i, j)$.

Рассмотрим теперь алгоритм, реализующий указанный метод. Вместо $u_{k_2}(i)$ и $f_{k_2}(i)$ удобно ввести новые вспомогательные функции $v_{k_2}(i)$ и $\Phi_{k_2}(i)$ по формулам

$$u_{k_2}(i) = \frac{\sqrt{2l_2}}{N_2} v_{k_2}(i), \quad f_{k_2}(i) = \frac{\sqrt{2l_2}}{N_2} \Phi_{k_2}(i). \quad (36)$$

Подставим (31) и (36) в (29), (33) и (35), учтем, что $h_2 N_2 = l_2$, и распишем разностный оператор Λ_1 по точкам. В результате получим

$$\Phi_{k_2}(i) = \sum_{j=1}^{N_2-1} f(i, j) \sin \frac{k_2 \pi j}{N_2}, \quad \begin{cases} 1 \leq k_2 \leq N_2 - 1, \\ 1 \leq i \leq N_1 - 1, \end{cases} \quad (37)$$

$$\begin{aligned} & -v_{k_2}(i-1) + (2 + h_1^2 \lambda_{k_2}^{(2)}) v_{k_2}(i) - v_{k_2}(i+1) = h_1^2 \Phi_{k_2}(i), \\ & 1 \leq i \leq N_1 - 1, \quad v_{k_2}(0) = v_{k_2}(N_1) = 0, \quad 1 \leq k_2 \leq N_2 - 1, \end{aligned} \quad (38)$$

$$u(i, j) = \frac{2}{N_2} \sum_{k_2=1}^{N_2-1} v_{k_2}(i) \sin \frac{k_2 \pi j}{N_2}, \quad \begin{cases} 1 \leq j \leq N_2 - 1, \\ 1 \leq i \leq N_1 - 1, \end{cases} \quad (39)$$

где $\lambda_{k_2}^{(2)}$ определено в (34).

Суммы (37) и (39), очевидно, следует вычислять, используя алгоритм быстрого дискретного преобразования Фурье, который изложен в п.2 § 1. Для решения трехточечных краевых задач (38) целесообразно использовать алгоритм прогонки, построенный в § 1 гл. II. Для задачи (38) алгоритм прогонки описывается формулами

$$\begin{aligned} \alpha_{i+1} &= \frac{1}{c_{k_2} - \alpha_i}, & 1 \leq i \leq N_1 - 1, \quad \alpha_1 = 0, \\ \beta_{i+1} &= [h_1^2 \Phi_{k_2}(i) + \beta_i] \alpha_{i+1}, & 1 \leq i \leq N_1 - 1, \quad \beta_1 = 0, \\ v_{k_2}(i) &= \alpha_{i+1} v_{k_2}(i+1) + \beta_{i+1}, & 1 \leq i \leq N_1 - 1, \quad v_{k_2}(N_1) = 0, \end{aligned} \quad (40)$$

где $c_{k_2} = 2 + h_1^2 \lambda_{k_2}^{(2)}$ и $k_2 = 1, 2, \dots, N_2 - 1$.

Сравним формулы (37), (39) и (40) с полученными ранее (24)–(27) для метода разложения в двойной ряд. Здесь вместо вычисления двух сумм (25) и (26) мы решаем серию краевых задач (38) методом прогонки (40). Поэтому на вычисление сумм (37) и (39) будет затрачиваться арифметических действий примерно в два раза меньше, чем для алгоритма (24)–(27). Дополнительные затраты на решение задач (38) составят, очевидно, $O(N_1 N_2)$ действий, что не повлияет на главный член в оценке числа арифметических действий алгоритма (37), (39), (40). Приведем точные оценки для числа действий для этого алгоритма. Имеем (для $N_2 = 2^m$) $Q_{\pm} = [(3 \log_2 N_2 - 1) N_2 - 2 \log_2 N_2 + 1](N_1 - 1)$ операций сложения и вычитания, $Q_* = -[(\log_2 N_2 + 2) N_2 - 2](N_1 - 1)$ операций умножения и $Q_1 =$

$= (N_1 - 1)(N_2 - 1)$ операций деления, а всего при $N_1 = N_2 = N = 2^n$ число операций равно

$$Q = (N^2 - 1,5N)(4 \log_2 N + 2) - N + 2 \log_2 N + 2.$$

Мы рассмотрели метод разложения в однократный ряд на примере разностной задачи Дирихле для уравнения Пуассона. Существенным моментом является то, что собственные функции разностного оператора Λ_2 допускают использование алгоритма быстрого преобразования Фурье для вычисления соответствующих сумм. Такая возможность будет иметь место и для случая, когда на сторонах $x_2 = 0$ и $x_2 = l_2$ прямоугольника \bar{G} заданы вместо краевых условий первого рода условия второго рода или комбинация условий первого и второго рода, а также для случая периодических условий.

Рассмотрим для примера следующую краевую задачу для уравнения Пуассона:

$$\begin{aligned} u_{\bar{x}_1 x_1} + u_{x_2 \bar{x}_2} &= -\varphi(x), \quad x \in \omega, \\ u(x) &= 0, \quad x_1 = 0, \quad l_1, \quad 0 \leq x_2 \leq l_2, \\ u_{\bar{x}_1 x_1} + \frac{2}{h_2} u_{x_2} &= -\varphi(x) - \frac{2}{h_2} g_{-2}(x), \quad x_2 = 0, \\ u_{\bar{x}_1 x_1} - \frac{2}{h_2} u_{\bar{x}_2} &= -\varphi(x) - \frac{2}{h_2} g_{+2}(x), \quad x_2 = l_2, \\ h_1 \leq x_1 \leq l_2 - h_1. \end{aligned} \quad (41)$$

Схема (41) есть разностная аппроксимация для задачи

$$\begin{aligned} \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} &= -\varphi(x), \quad x \in G, \\ u(x) &= 0, \quad x_1 = 0, \quad l_1, \quad 0 \leq x_2 \leq l_2, \\ \frac{\partial u}{\partial x_2} &= -g_{-2}(x), \quad x_2 = 0, \\ -\frac{\partial u}{\partial x_2} &= -g_{+2}(x), \quad x_2 = l_2, \quad 0 \leq x_1 \leq l_1. \end{aligned}$$

Запишем задачу (41) в другом виде, вводя обозначения:

$$\Lambda_2 u = \begin{cases} \frac{2}{h_2} u_{x_2}, & x_2 = 0, \\ u_{\bar{x}_2 x_2}, & h_2 \leq x_2 \leq l_2 - h_2, \\ -\frac{2}{h_2} u_{\bar{x}_2}, & x_2 = l_2, \end{cases}$$

$$\varphi_2(x) = \begin{cases} \frac{2}{h_2} g_{-2}(x), & x_2 = 0, \\ 0, & h_2 \leq x_2 \leq l_2 - h_2, \\ \frac{2}{h_2} g_{+2}(x), & x_2 = l_2, \end{cases}$$

$$f(x) = \varphi(x) + \varphi_2(x), \quad \Lambda_1 u = u_{\bar{x}_1 x_1}$$

для $h_1 \leq x_1 \leq l_1 - h_1, 0 \leq x_2 \leq l_2$.

В новых обозначениях задача (41) запишется в виде

$$\begin{aligned} \Lambda u &= (\Lambda_1 + \Lambda_2) u = -f(x), \quad h_1 \leq x_1 \leq l_1 - h_1, \quad 0 \leq x_2 \leq l_2, \\ u(x) &= 0, \quad x_1 = 0, \quad l_1, \quad 0 \leq x_2 \leq l_2. \end{aligned} \quad (42)$$

Разлагая $u(i, j)$ и $f(i, j)$ в суммы по собственным функциям оператора Λ_2 , будем иметь

$$\begin{aligned} u(i, j) &= \sum_{k_2=0}^{N_2} u_{k_2}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leq j \leq N_2, \quad 0 \leq i \leq N_1, \\ f(i, j) &= \sum_{k_2=0}^{N_2} f_{k_2}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leq j \leq N_2, \quad 1 \leq i \leq N_1 - 1, \end{aligned} \quad (43)$$

где

$$\mu_{k_2}^{(2)}(j) = \begin{cases} \sqrt{\frac{1}{l_2}} \cos \frac{k_2 \pi j}{N_2}, & k_2 = 0, \quad N_2, \\ \sqrt{\frac{2}{l_2}} \cos \frac{k_2 \pi j}{N_2}, & 1 \leq k_2 \leq N_2 - 1 \end{cases}$$

есть собственная функция оператора Λ_2 , соответствующая собственному значению

$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2}, \quad k_2 = 0, 1, \dots, N_2. \quad (44)$$

Коэффициент Фурье $f_{k_2}(i)$ для каждого $1 \leq i \leq N_1 - 1$ вычисляется по формулам

$$f_{k_2}(i) = \sum_{j=1}^{N_2-1} h_2 f(i, j) \mu_{k_2}^{(2)}(j) + 0,5 h_2 [f(i, 0) \mu_{k_2}^{(2)}(0) + f(i, N_2) \mu_{k_2}^{(2)}(N_2)].$$

Подставляя (43) в (42), получим для рассматриваемой задачи (42) следующий аналог формул (37)–(39):

$$\begin{aligned} \varphi_{k_2}(i) &= \sum_{j=0}^{N_2} \rho_j f(i, j) \cos \frac{k_2 \pi j}{N_2}, \\ 0 \leq k_2 \leq N_2, \quad 1 \leq i \leq N_1 - 1, \\ -v_{k_2}(i-1) + (2 + h_1^2 \lambda_{k_2}^{(2)}) v_{k_2}(i) - v_{k_2}(i+1) &= h_1^2 \varphi_{k_2}(i), \\ 1 \leq i \leq N_1 - 1, \quad v_{k_2}(0) = v_{k_2}(N_1) = 0, \quad 0 \leq k_2 \leq N_2, \\ u(i, j) &= \frac{2}{N_2} \sum_{k_2=0}^{N_2} \rho_{k_2} v_{k_2}(i) \cos \frac{k_2 \pi j}{N_2}, \\ 0 \leq j \leq N_2, \quad 1 \leq i \leq N_1 - 1, \end{aligned}$$

где $\lambda_{k_2}^{(2)}$ определено в (44), а

$$\rho_j = \begin{cases} 0,5, & j=0, \quad N_2, \\ 1, & 1 \leq j \leq N_2 - 1. \end{cases}$$

Приведем оценку числа действий для построенного алгоритма при $N_1 = N_2 = N = 2^n$: $Q_{\pm} = [(3 \log_2 N_2 - 1) N_2 + 2 \log_2 N_2 + 7] \times \times (N_1 - 1)$ операций сложения и вычитания, $Q_* = [(\log_2 N_2 + 2) N_2 + 10] (N_1 - 1)$ операций умножения и $Q_r = (N_2 + 1)(N_1 - 1)$ операций деления, а всего

$$Q = \left(N^2 - \frac{N}{2} \right) (4 \log_2 N + 2) + 17N - 2 \log_2 N - 18.$$

Далее, так как в методе разложения в однократный ряд собственные функции разностного оператора Λ_i не используются и единственное требование к Λ_i состоит в возможности разделять переменные, то в качестве Λ_i можно взять более общий, чем мы рассмотрели, оператор. Если ограничиться эллиптическими уравнениями второго порядка, то наиболее общему случаю выбора оператора Λ_i соответствует разностная аппроксимация для дифференциального оператора

$$L_1 u = \frac{1}{k_2(x_1)} \frac{\partial}{\partial x_1} \left(k_1(x_1) \frac{\partial u}{\partial x_1} \right) + r(x_1) \frac{\partial u}{\partial x_1} - q(x_1) u,$$

коэффициенты которого зависят лишь от x_1 . Краевые же условия на сторонах $x_1 = 0$ и $x_1 = l_2$ прямоугольника \bar{G} могут быть любой комбинацией краевых условий первого, второго или третьего рода (коэффициенты в краевом условии третьего рода должны быть константами). Это позволяет решать краевые задачи для уравнения Пуассона в цилиндрической, сферической и полярной системах координат.

§ 3. Метод неполной редукции

1. Комбинация методов Фурье и редукции. Построенный в п. 3 § 2 метод разложения в однократный ряд позволил ограничиться вычислением только двух сумм Фурье с затратой $O(N_1 N_2 \log_2 N_2)$ действий и решением серии трехточечных краевых задач за $O(N_1 N_2)$ действий. Очевидно, дальнейшее усовершенствование метода разделения переменных возможно на пути уменьшения числа слагаемых в вычисляемых суммах при сохранении возможности использовать алгоритм быстрого преобразования Фурье.

Мы достигнем этой цели, комбинируя метод разложения в однократный ряд с изученным в главе III методом редукции. Построим сначала такой комбинированный метод для простейшей задачи Дирихле

$$\begin{aligned} \Lambda u &= -f(x), \quad x \in \omega, \quad u(x) = 0, \quad x \in \gamma, \\ \Lambda = \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha u &= u_{x_\alpha x_\alpha}, \quad \alpha = 1, 2 \end{aligned} \tag{1}$$

на прямоугольной сетке $\bar{\omega}$.

Для упрощения описания метода перейдем от точечной (скалярной) записи задачи (1) к векторной.

Введем вектор неизвестных \mathbf{U}_j следующим образом:

$$\mathbf{U}_j = (u(1, j), u(2, j), \dots, u(N_1 - 1, j)), \quad 0 \leq j \leq N_2,$$

и определим вектор правых частей \mathbf{F}_j формулой

$$\mathbf{F}_j = (h_2^2 f(1, j), h_2^2 f(2, j), \dots, h_2^2 f(N_1 - 1, j)), \quad 1 \leq j \leq N_2 - 1.$$

Тогда разностную задачу (1) можно записать (см. гл. III, § 1) в виде следующей системы векторных уравнений:

$$\begin{aligned} -\mathbf{U}_{j-1} + C\mathbf{U}_j - \mathbf{U}_{j+1} &= \mathbf{F}_j, \quad 1 \leq j \leq N_2 - 1, \\ \mathbf{U}_0 &= \mathbf{U}_{N_2} = 0, \end{aligned} \quad (2)$$

где квадратная трехдиагональная матрица C определяется равенствами

$$\begin{aligned} C\mathbf{U}_j &= ((2E - h_2^2 \Lambda_1) u(1, j), \dots, (2E - h_2^2 \Lambda_1) u(N_1 - 1, j)), \\ \Lambda_1 u &= u_{x_1 x_1}, \quad u(0, j) = u(N_1, j) = 0. \end{aligned}$$

Пусть N_2 есть степень 2: $N_2 = 2^m$. Напомним, что первый шаг процесса исключения в методе полной редукции состоит (см. гл. III, § 2) в выделении из (2) «укороченной» системы для неизвестных \mathbf{U}_j с четными номерами j

$$\begin{aligned} -\mathbf{U}_{j-2} + C^{(1)}\mathbf{U}_j - \mathbf{U}_{j+2} &= \mathbf{F}_j^{(1)}, \quad j = 2, 4, 6, \dots, N_2 - 2, \\ \mathbf{U}_0 &= \mathbf{U}_{N_2} = 0 \end{aligned} \quad (3)$$

и уравнений

$$C\mathbf{U}_j = \mathbf{F}_j + \mathbf{U}_{j-1} + \mathbf{U}_{j+1}, \quad j = 1, 3, 5, \dots, N_2 - 1 \quad (4)$$

для определения неизвестных с нечетными номерами j . Здесь обозначено

$$\mathbf{F}_j^{(1)} = \mathbf{F}_{j-1} + C\mathbf{F}_j + \mathbf{F}_{j+1}, \quad j = 2, 4, 6, \dots, N_2 - 2, \quad (5)$$

$$C^{(1)} = [C]^2 - 2E. \quad (6)$$

Займемся системой (3). Введем обозначения

$$\mathbf{V}_j = (v(1, j), v(2, j), \dots, v(N_1 - 1, j)),$$

$$\Phi_j = (h_2^2 \varphi(1, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(N_1 - 1, j))$$

и положим

$$\begin{aligned} \mathbf{V}_j &= \mathbf{U}_{2j}, \quad 0 \leq j \leq N_2/2, \quad \Phi_j = \mathbf{F}_{2j}^{(1)}, \quad 1 \leq j \leq N_2/2 - 1, \\ v(0, j) &= v(N_1, j) = 0, \quad 0 \leq j \leq N_2/2. \end{aligned}$$

Эти обозначения позволяют записать систему (3) в виде

$$\begin{aligned} -\mathbf{V}_{j-1} + C^{(1)}\mathbf{V}_j - \mathbf{V}_{j+1} &= \Phi_j, \quad j = 1, 2, \dots, M_2 - 1, \\ \mathbf{V}_0 &= \mathbf{V}_{M_2} = 0, \end{aligned} \quad (7)$$

где $2M_2 = N_2$ и в силу (5)

$$\Phi_j = \mathbf{F}_{2j-1} + C\mathbf{F}_{2j} + \mathbf{F}_{2j+1}, \quad j = 1, 2, \dots, M_2 - 1. \quad (8)$$

Заметим теперь, что сеточная функция $v(i, j)$ определена для $0 \leq i \leq N_1$ и $0 \leq j \leq M_2$ и обращается в нуль при $j = 0$ и $j = M_2$. Функция $\varphi(i, j)$ определена для $1 \leq i \leq N_1 - 1$ и $1 \leq j \leq M_2 - 1$. Поэтому эти функции можно представить в виде однократных рядов Фурье

$$\begin{aligned} v(i, j) &= \sum_{k_2=1}^{M_2-1} y_{k_2}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leq i \leq N_1, \quad 0 \leq j \leq M_2, \\ \varphi(i, j) &= \sum_{k_2=1}^{M_2-1} z_{k_2}(i) \mu_{k_2}^{(2)}(j), \\ &\quad 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq M_2 - 1, \end{aligned} \quad (9)$$

где функции

$$\mu_{k_2}^{(2)}(j) = \frac{2}{\sqrt{l_2}} \sin \frac{k_2 \pi j}{M_2}, \quad k_2 = 1, 2, \dots, M_2 - 1 \quad (10)$$

образуют ортонормированную систему на сетке $\bar{\omega}$ в смысле скалярного произведения

$$(u, v) = \sum_{j=1}^{M_2-1} u(j) v(j) h_2.$$

Коэффициенты Фурье $z_{k_2}(i)$ функции $\varphi(i, j)$ находятся по формулам

$$z_{k_2}(i) = (\varphi, \mu_{k_2}^{(2)}) = \sum_{j=1}^{M_2-1} h_2 \varphi(i, j) \mu_{k_2}^{(2)}(j), \quad (11)$$

$$1 \leq k_2 \leq M_2 - 1, \quad 1 \leq i \leq N_1 - 1.$$

Из (9) получим для векторов V_j и Φ_j следующие разложения:

$$\begin{aligned} V_j &= \sum_{k_2=1}^{M_2-1} Y_{k_2} \mu_{k_2}^{(2)}(j), \quad 0 \leq j \leq M_2, \\ \Phi_j &= \sum_{k_2=1}^{M_2-1} h_2^2 Z_{k_2} \mu_{k_2}^{(2)}(j), \quad 1 \leq j \leq M_2 - 1, \end{aligned} \quad (12)$$

где

$$\begin{aligned} Y_{k_2} &= (y_{k_2}(1), y_{k_2}(2), \dots, y_{k_2}(N_1 - 1)), \\ Z_{k_2} &= (z_{k_2}(1), z_{k_2}(2), \dots, z_{k_2}(N_1 - 1)). \end{aligned}$$

Подставим (12) в (7) и учтем равенства

$$\mu_{k_2}^{(2)}(j-1) + \mu_{k_2}^{(2)}(j+1) = 2 \cos \frac{k_2 \pi}{M_2} \mu_{k_2}^{(2)}(j), \quad 1 \leq k_2 \leq M_2 - 1.$$

Получим

$$\sum_{k_2=1}^{M_2-1} \left(C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E \right) Y_{k_2} \mu_{k_2}^{(2)}(j) = \sum_{k_2=1}^{M_2-1} h_2^2 Z_{k_2} \mu_{k_2}^{(2)}(j),$$

откуда в силу ортонормированности системы (10) будем иметь

$$\left(C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E \right) Y_{k_2} = h_2^2 Z_{k_2}, \quad 1 \leq k_2 \leq M_2 - 1. \quad (13)$$

Используем соотношение (6) и получим

$$\begin{aligned} C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E &= [C]^2 - 2 \left(1 + \cos \frac{k_2 \pi}{M_2} \right) E = \\ &= \left(C - 2 \cos \frac{k_2 \pi}{2M_2} E \right) \left(C + 2 \cos \frac{k_2 \pi}{2M_2} E \right). \end{aligned}$$

Так как матрица $C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E$ факторизована, то для решения уравнения (13) можно использовать алгоритм

$$\begin{aligned} \left(C - 2 \cos \frac{k_2 \pi}{2M_2} E \right) W_{k_2} &= h_2^2 Z_{k_2}, \\ \left(C + 2 \cos \frac{k_2 \pi}{2M_2} E \right) Y_{k_2} &= W_{k_2}, \quad 1 \leq k_2 \leq M_2 - 1, \end{aligned} \quad (14)$$

где вспомогательный вектор W_{k_2} имеет компоненты $w_{k_2}(i)$:

$$\begin{aligned} W_{k_2} &= (w_{k_2}(1), w_{k_2}(2), \dots, w_{k_2}(N_1 - 1)), \\ w_{k_2}(0) &= w_{k_2}(N_1) = 0. \end{aligned}$$

Необходимые формулы получены. Переходя в (4), (8) и (14) от векторной записи к скалярной и используя соотношение $u(i, 2j) = v(i, j)$, вытекающее из определения V_j , получим следующие формулы для построенного метода:

$$\begin{aligned} \varphi(i, j) &= f(i, 2j - 1) + 2f(i, 2j) + f(i, 2j + 1) - h_2^2 \Lambda_1 f(i, 2j), \quad (15) \\ 1 \leq j &\leq N_2/2 - 1, \quad 1 \leq i \leq N_1 - 1, \quad f(0, 2j) = f(N_1, 2j) = 0 \end{aligned}$$

для вычисления функции $\varphi(i, j)$; уравнения

$$\begin{aligned} 2 \left(1 - \cos \frac{k_2 \pi}{2M_2} \right) w_{k_2}(i) - h_2^2 \Lambda_1 w_{k_2}(i) &= h_2^2 z_{k_2}(i), \\ 1 \leq i &\leq N_1 - 1, \\ w_{k_2}(0) &= w_{k_2}(N_1) = 0, \\ 2 \left(1 + \cos \frac{k_2 \pi}{2M_2} \right) y_{k_2}(i) - h_2^2 \Lambda_1 y_{k_2}(i) &= w_{k_2}(i), \\ 1 \leq i &\leq N_1 - 1, \\ y_{k_2}(0) &= y_{k_2}(N_1) = 0 \end{aligned} \quad (16)$$

для определения $y_{k_2}(i)$ при $k_2 = 1, 2, \dots, M_2 - 1$ и уравнения

$$\begin{aligned} 2u(i, 2j - 1) - h_2^2 \Lambda_1 u(i, 2j - 1) &= \\ &= h_2^2 f(i, 2j - 1) + u(i, 2j - 2) + u(i, 2j), \quad (17) \\ 1 \leq i &\leq N_1 - 1, \quad u(0, 2j - 1) = u(N_1, 2j - 1) = 0 \end{aligned}$$

для нахождения решения при $j = 1, 2, \dots, M_2$. Для коэффициентов Фурье $z_{k_2}(i)$ имеем формулу (11), а из (9) получим

$$u(i, 2j) = \sum_{k_2=1}^{M_2-1} y_{k_2}(i) \mu_{k_2}^{(2)}(j), \quad 1 \leq j \leq M_2 - 1, \quad 1 \leq i \leq N_1 - 1. \quad (18)$$

Итак, формулы (10), (11), (15)–(18) полностью описывают метод решения задачи (1), который является комбинацией методов разложения в однократный ряд Фурье и редукции.

Переходим теперь к построению алгоритма метода. В формулах (9), (16) и (18) сделаем замену $y_{k_2}(i) = a\bar{y}_{k_2}(i)$, $w_{k_2}(i) = a\bar{w}_{k_2}(i)$, $z_{k_2}(i) = az_{k_2}(i)$, где $a = 2\sqrt{l_2}/N_2$, а в полученных формулах знак тильда опустим. Эта замена позволяет избавиться от нормирующего множителя $2\sqrt{l_2}$, стоящего при собственной функции $\mu_{k_2}^{(2)}(j)$ в суммах (11) и (18). Далее, задачи (16) и (17) будем решать методом прогонки. Легко убедиться в том, что здесь условия корректности и устойчивости обычного метода прогонки выполнены. Отметим особенность задачи (17). Так как коэффициенты уравнения (17) не зависят от j , то прогоночные коэффициенты α_i следует вычислить один раз при решении задачи (17) для $j = 1$ и далее использовать при решении уравнений (17) для остальных j .

Приведем сводку расчетных формул. Сначала вычисляются

$$\begin{aligned} \varphi(i, j) &= f(i, 2j-1) + f(i, 2j+1) + 2 \left(1 + \frac{h_2^2}{h_1^2} \right) f(i, 2j) - \\ &\quad - \frac{h_2^2}{h_1^2} [f(i-1, 2j) + f(i+1, 2j)], \\ 1 \leq j &\leq M_2 - 1, \quad 1 \leq i \leq N_1 - 1, \end{aligned} \quad (19)$$

где $f(0, 2j) = f(N_1, 2j) = 0$. Значения $\varphi(i, j)$ можно разместить на месте $f(i, 2j)$. Суммы

$$z_{k_2}(i) = \sum_{j=1}^{M_2-1} \varphi(i, j) \sin \frac{k_2 \pi j}{M_2}, \quad 1 \leq k_2 \leq M_2 - 1 \quad (20)$$

для $1 \leq i \leq N_1 - 1$ вычисляются по алгоритму быстрого дискретного преобразования Фурье, и $z_{k_2}(i)$ размещается на месте $\varphi(i, k_2)$. Методом прогонки

$$\begin{aligned} \alpha_{i+1} &= 1/(c_{k_2} - \alpha_i), \quad \beta_{i+1} = [h_1^2 z_{k_2}(i) + \beta_i] \alpha_{i+1}, \\ i &= 1, 2, \dots, N_1 - 1, \quad \alpha_1 = \beta_1 = 0, \\ w_{k_2}(i) &= \alpha_{i+1} w_{k_2}(i+1) + \beta_{i+1}, \quad i = N_1 - 1, N_1 - 2, \dots, 1, \quad (21) \\ w_{k_2}(N_1) &= 0, \quad c_{k_2} = 2 + 2 \frac{h_1^2}{h_2^2} - 2 \frac{h_1^2}{h_2^2} \cos \frac{k_2 \pi}{N_2} \end{aligned}$$

решается первое из уравнений (16), и аналогично по формулам

$$\begin{aligned} \alpha_{i+1} &= \frac{1}{c_{k_2} - \alpha_i}, \quad \beta_{i+1} = \left[\frac{h_1^2}{h_2^2} w_{k_2}(i) + \beta_i \right] \alpha_{i+1}, \\ i &= 1, 2, \dots, N_1 - 1, \quad \alpha_1 = \beta_1 = 0, \\ y_{k_2}(i) &= \alpha_{i+1} y_{k_2}(i+1) + \beta_{i+1}, \quad i = N_1 - 1, N_2 - 1, \dots, 1, \quad (22) \\ y_{k_2}(N_1) &= 0, \quad c_{k_2} = 2 + 2 \frac{h_1^2}{h_2^2} + 2 \frac{h_1^2}{h_2^2} \cos \frac{k_2 \pi}{N_2} \end{aligned}$$

решается второе из уравнений (16). Здесь вычисления проводятся последовательно для $k_2 = 1, 2, \dots, M_2 - 1$ и результаты $w_{k_2}(i)$ и $y_{k_2}(i)$ размещаются последовательно на месте $z_{k_2}(i)$.

Для вычисления сумм

$$u(i, 2j) = \frac{4}{N_2} \sum_{k_2=1}^{M_2-1} y_{k_2}(i) \sin \frac{k_2 \pi j}{M_2}, \quad 1 \leq j \leq M_2 - 1, \quad (23)$$

для $1 \leq i \leq N_1 - 1$ снова используем алгоритм быстрого преобразования Фурье. Задачи (17) решаются методом прогонки с учетом отмеченной особенности этих уравнений:

$$\begin{aligned} \alpha_{i+1} &= 1/(c - \alpha_i), \quad i = 1, 2, \dots, N_1 - 1, \quad \alpha_1 = 0, \\ \beta_{i+1} &= \left[h_1^2 f(i, 2j-1) + \frac{h_1^2}{h_2^2} (u(i, 2j-2) + u(i, 2j)) + \beta_i \right] \alpha_{i+1}, \\ i &= 1, 2, \dots, N_1 - 1, \quad \beta_1 = 0, \\ u(i, 2j-1) &= \alpha_{i+1} u(i+1, 2j-1) + \beta_{i+1}, \\ i &= N_1 - 1, N_1 - 2, \dots, 1, \quad u(N_1, 2j-1) = 0, \\ c &= 2(1 + h_1^2/h_2^2) \end{aligned} \quad (24)$$

для $1 \leq j \leq M_2$. Решение $u(i, j)$ размещается на месте $f(i, j)$, и, следовательно, алгоритм не требует дополнительной памяти для промежуточной информации.

Подсчитаем число арифметических действий для алгоритма (19) — (24). Для вычисления по формулам (19), (21), (22) и (24) требуется $Q_{\pm} = (6,5N_2 - 9)(N_1 - 1)$ операций сложения и вычитания, $Q_* = (6N_2 - 8)(N_1 - 1)$ операций умножения и $Q_{\div} = (N_2 - 1)(N_1 - 1)$ операций деления. Для вычисления сумм (20) и (23) потребуется

$$Q_{\pm} = \left[\left(\frac{3}{2} \log_2 N_2 - \frac{7}{2} \right) N_2 - 2 \log_2 N_2 + 6 \right] (N_1 - 1)$$

операций сложения и вычитания и

$$Q_* = \left[\left(\frac{1}{2} \log_2 N_2 - 1 \right) N_2 + 1 \right] (N_1 - 1)$$

операций умножения. Всего же алгоритм (19) — (24) требует при $N_1 = N_2 = N = 2^n$

$$Q = (N^2 - 2N)(2 \log_2 N + 9) - 2N + 2 \log_2 N + 11 \quad (25)$$

арифметических операций.

Для сравнения приведем число операций метода разложения в однократный ряд (см. п. 3 § 2):

$$Q = \left(N^2 - \frac{3}{2}N \right) (4 \log_2 N + 2) - N + 2 \log_2 N + 2, \quad (26)$$

метода разложения в двойной ряд (см. п. 2 § 2):

$$Q = \left(N^2 - \frac{3}{2}N \right) (8 \log_2 N - 10) + 5N + 4 \log_2 N - 10, \quad (27)$$

а также число операций для второго алгоритма метода полной редукции (см. гл. III, § 2, п. 4):

$$Q = \left(N^2 - \frac{11}{5}N \right) (5 \log_2 N + 5) + N + 6 \log_2 N + 5. \quad (28)$$

Если сравнить в оценках (25) — (28) константы при главном члене $N^2 \log_2 N$, то получим, что комбинированный метод требует примерно в 4 раза меньше арифметических операций, чем метод разложения в двойной ряд. Этот вывод верен при больших N . Для получения реальных соотношений между рассматриваемыми методами при допустимых N приведем таблицу, содержащую значения Q для этих методов.

Таблица 4

$N \backslash$ Оценка	(25)	(26)	(27)	(28)
32	18 383	21 496	29 510	28 541
64	83 601	104 950	152 334	138 537
128	371 515	485 708	745 582	643 921

Итак, комбинация методов Фурье и редукции позволяет уменьшить число операций по сравнению с исходным методом разложения в однократный ряд. Обобщим этот комбинированный метод, включив в него l шагов исключения метода редукции перед выполнением разложения в однократный ряд. Тогда метод из п. 3 § 2 можно трактовать как частный случай такого обобщенного метода с $l=0$, а построенный в этом пункте метод соответствует $l=1$. Метод полной редукции можно рассматривать как метод с $l=\log_2 N_2$.

Данные табл. 4 показывают, что существует оптимальный с точки зрения затрат арифметических операций обобщенный метод с $1 \leq l < \log_2 N_2$. Анализ оценок для числа действий в методе, содержащем l шагов редукции, дает оптимальное значение $l=1$ или $l=2$. При этом незначительное преимущество в числе дей-

ствий метода для $l=2$ может быть утрачено из-за возросшей сложности алгоритма.

2. Решение краевых задач для уравнения Пуассона в прямоугольнике. Рассмотрим теперь применение построенного в п. 1 метода к нахождению решения краевых задач для уравнения Пуассона в прямоугольнике. Пусть в области $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ требуется найти решение уравнения

$$\frac{\partial^2 v}{\partial x_1^2} + \frac{\partial^2 v}{\partial x_2^2} = -\varphi(x), \quad x \in G, \quad (29)$$

удовлетворяющее следующим краевым условиям на границе Γ прямоугольника \bar{G} :

$$\begin{aligned} \frac{\partial v}{\partial x_1} &= \kappa_{-1}v - g_{-1}(x_2), & x_1 = 0, \\ -\frac{\partial v}{\partial x_1} &= \kappa_{+1}v - g_{+1}(x_2), & x_1 = l_1, \quad 0 \leq x_2 \leq l_2, \\ \frac{\partial v}{\partial x_2} &= -g_{-2}(x_1), & x_2 = 0, \\ -\frac{\partial v}{\partial x_2} &= -g_{+2}(x_1), & x_2 = l_2, \quad 0 \leq x_1 \leq l_1, \end{aligned} \quad (30)$$

где $\kappa_{+1} \geq 0$, $\kappa_{-1} \geq 0$, $\kappa_{+1}^2 + \kappa_{-1}^2 > 0$.

Будем предполагать, что в условиях (30) κ_{-1} и κ_{+1} — постоянные. При этом предположении переменные в задаче (29), (30) разделяются.

На прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ задаче (29) — (30) соответствует разностная схема

$$\Lambda u = (\Lambda_1 + \Lambda_2) u = -f(x), \quad x \in \bar{\omega}, \quad (31)$$

где $f(x) = \varphi(x) + \varphi_1(x) + \varphi_2(x)$,

$$\begin{aligned} \Lambda_1 u &= \begin{cases} \frac{2}{h_1} (u_{x_1} - \kappa_{-1} u), & x_1 = 0, \\ u_{\bar{x}_1 x_1}, & h_1 \leq x_1 \leq l_1 - h_1, \\ \frac{2}{h_1} (-u_{\bar{x}_1} - \kappa_{+1} u), & x_1 = l_1; \end{cases} \\ \Lambda_2 u &= \begin{cases} \frac{2}{h_2} u_{x_2}, & x_2 = 0, \\ u_{\bar{x}_2 x_2}, & h_2 \leq x_2 \leq l_2 - h_2, \\ -\frac{2}{h_2} u_{\bar{x}_2}, & x_2 = l_2, \end{cases} \end{aligned}$$

а функции $\varphi_\alpha(x)$ определяются соотношением

$$\varphi_\alpha(x) = \begin{cases} \frac{2}{h_\alpha} g_{-\alpha}(x_\beta), & x_\alpha = 0, \\ 0, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \beta = 3 - \alpha, \alpha = 1, 2, \\ \frac{2}{h_\alpha} g_{+\alpha}(x_\beta), & x_\alpha = l_\alpha. \end{cases}$$

В главе III было показано, что схема (31) в векторном виде имеет следующую запись:

$$\begin{aligned} CU_0 - 2U_1 &= \mathbf{F}_0, \\ -U_{j-1} + CU_j - U_{j+1} &= \mathbf{F}_j, \quad 1 \leq j \leq N_2 - 1, \\ -2U_{N_2-1} + CU_{N_2} &= \mathbf{F}_{N_2}, \end{aligned} \quad (32)$$

где

$$\begin{aligned} \mathbf{U}_j &= (u(0, j), u(1, j), \dots, u(N_1, j)), \\ \mathbf{F}_j &= (h_2^2 f(0, j), h_2^2 f(1, j), \dots, h_2^2 f(N_1, j)), \\ CU_j &= ((2E - h_2^2 \Lambda_1) u(0, j), \dots, (2E - h_2^2 \Lambda_1) u(N_1, j)), \quad 0 \leq j \leq N_2. \end{aligned}$$

Векторная система (32) отличается от рассмотренной ранее системы (2) краевыми условиями и определением матрицы C . Тем не менее построить аналог метода п. 1 для задачи (32) не представляет труда. Поскольку вывод основных формул для этого метода лишь в деталях отличается от приведенного в п. 2, то мы ограничимся сводкой главных промежуточных и окончательных формул. Для метода полной редукции необходимые формулы описаны в § 4 гл. III.

Итак, для векторов $\mathbf{V}_j = \mathbf{U}_{2j}$, $0 \leq j \leq M_2$, где $2M_2 = N_2$, после шага исключения будем иметь задачу

$$\begin{aligned} C^{(1)} \mathbf{V}_0 - 2\mathbf{V}_1 &= \Phi_0, \\ -\mathbf{V}_{j-1} + C^{(1)} \mathbf{V}_j - \mathbf{V}_{j+1} &= \Phi_j \quad 1 \leq j \leq M_2 - 1, \\ -2\mathbf{V}_{M_2-1} + C^{(1)} \mathbf{V}_{M_2} &= \Phi_{M_2}, \end{aligned} \quad (33)$$

где правая часть $\Phi_j = \mathbf{F}_{2j}^{(1)}$, $0 \leq j \leq M_2$ определяется по формулам

$$\Phi_j = \begin{cases} C\mathbf{F}_0 + 2\mathbf{F}_1, & j = 0, \\ \mathbf{F}_{2j-1} + C\mathbf{F}_{2j} + \mathbf{F}_{2j+1}, & 1 \leq j \leq M_2 - 1, \\ C\mathbf{F}_{N_2} + 2\mathbf{F}_{N_2-1}, & j = M_2. \end{cases}$$

Для векторов \mathbf{V}_j и Φ_j имеем разложения

$$\mathbf{V}_j = \sum_{k_2=0}^{M_2} Y_{k_2} \mu_{k_2}^{(2)}(j), \quad \Phi_j = \sum_{k_2=0}^{M_2} h_2^2 \mathbf{Z}_{k_2} \mu_{k_2}^{(2)}(j), \quad 0 \leq j \leq M_2,$$

где

$$\mu_{k_2}^{(2)}(j) = \begin{cases} \frac{2}{\sqrt{l_2}} \cos \frac{k_2 \pi j}{M_2}, & 1 \leq k_2 \leq M_2 - 1, \\ \sqrt{\frac{1}{l_2}} \cos \frac{k_2 \pi j}{M_2}, & k_2 = 0, M_2. \end{cases}$$

Коэффициенты Фурье векторов \mathbf{V}_j и Φ_j в силу (33) связаны соотношением

$$\left(C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E \right) Y_{k_2} = h_2^2 \mathbf{Z}_{k_2}, \quad 0 \leq k_2 \leq M_2,$$

а компоненты вектора \mathbf{Z}_{k_2} выражаются через компоненты вектора Φ_j следующим образом:

$$\mathbf{z}_{k_2}(i) = \sum_{j=1}^{M_2-1} h_2 \varphi(i, j) \mu_{k_2}^{(2)}(j) + 0,5 h_2 [\varphi(i, 0) \mu_{k_2}^{(2)}(0) + \varphi(i, M_2) \mu_{k_2}^{(2)}(M_2)], \quad 0 \leq i \leq N_1.$$

Неизвестные U_j с нечетными номерами j , как и раньше, определяются из уравнений (4).

В полученных формулах осталось перейти к скалярной записи и к ненормированной собственной функции $\bar{\mu}_{k_2}^{(2)}(j) = \cos \frac{k_2 \pi j}{M_2}$.

В результате получим следующие формулы для метода решения задачи (31): для каждого $0 \leq i \leq N_1$ вычисляются

$$\varphi(i, j) = \begin{cases} 2[f(i, 0) + f(i, 1)] - h_2^2 \Lambda_1 f(i, 0), & j = 0, \\ f(i, 2j-1) + f(i, 2j+1) + 2f(i, 2j) - h_2^2 \Lambda_1 f(i, 2j), & 1 \leq j \leq M_2-1, \\ 2[f(i, N_2) + f(i, N_2-1)] - h_2^2 \Lambda_1 f(i, N_2), & j = M_2, \end{cases}$$

решаются уравнения

$$4 \sin^2 \frac{k_2 \pi}{2N_2} w_{k_2}(i) - h_2^2 \Lambda_1 w_{k_2}(i) = h_2^2 z_{k_2}(i), \quad 0 \leq i \leq N_1,$$

$$4 \cos^2 \frac{k_2 \pi}{2N_2} y_{k_2}(i) - h_2^2 \Lambda_1 y_{k_2}(i) = w_{k_2}(i), \quad 0 \leq i \leq N_1$$

для $0 \leq k_2 \leq M_2$, где

$$z_{k_2}(i) = \sum_{j=0}^{M_2} \rho_j \varphi(i, j) \cos \frac{k_2 \pi j}{M_2},$$

$$0 \leq k_2 \leq M_2, \quad 0 \leq i \leq N_1.$$

Решение $u(i, j)$ задачи (31) определяется по формулам

$$u(i, 2j) = \sum_{k_2=0}^{M_2} \rho_{k_2} y_{k_2}(i) \cos \frac{k_2 \pi j}{M_2}, \quad 0 \leq j \leq M_2, \quad 0 \leq i \leq N_1$$

и из уравнений

$$2u(i, 2j-1) - h_2^2 \Lambda_1 u(i, 2j-1) =$$

$$= h_2^2 f(i, 2j-1) + u(i, 2j-2) + u(i, 2j),$$

$$1 \leq j \leq M_2, \quad 0 \leq i \leq N_1.$$

Здесь использованы обозначения

$$\rho_j = \begin{cases} 1, & 1 \leq j \leq M_2-1, \\ 0,5, & j = 0, M_2, \quad M_2 = 0,5N_2, \end{cases}$$

а оператор Λ_i определен выше. Для нахождения $w_{k_1}(i)$, $y_{k_2}(i)$ и $u(i, 2j-1)$ здесь мы имеем трехточечные уравнения с краевыми условиями третьего рода, которые решаются методом прогонки.

Заметим, что приведенные формулы нисколько не изменяются, если сетка по направлению x_1 будет неравномерной. Изменится лишь вид оператора Λ_1 —это будет разностный аналог второй производной и краевых условий третьего рода на неравномерной сетке.

Вообще следует отметить, что можно построить соответствующий вариант метода разделения переменных с оценкой числа действий $O(N^2 \log_2 N)$ во всех, за исключением одного, случаях, в которых можно использовать метод полной редукции. Исключение составляет тот случай, в котором по направлению исключения неизвестных задано краевое условие третьего рода хотя бы на одной из сторон прямоугольника.

3. Разностная задача Дирихле повышенного порядка точности в прямоугольнике. Рассмотрим еще один пример применения метода разделения переменных. Пусть на прямоугольной сетке ω требуется найти решение *разностной задачи Дирихле повышенного порядка точности для уравнения Пуассона*

$$\Lambda u = \left(\Lambda_1 + \Lambda_2 + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \Lambda_2 \right) u = -f(x), \quad x \in \omega, \quad (34)$$

$$u(x) = 0, \quad x \in \gamma,$$

где $\Lambda_\alpha u = u_{x_\alpha x_\alpha}$, $\alpha = 1, 2$.

Краевое условие задано однородным для простоты—задача с неоднородным краевым условием сводится к (34) путем поправки правой части уравнения в приграничных узлах.

В п. 4 § 1 гл. III была получена векторная запись задачи (34) в следующем виде:

$$\begin{aligned} -BU_{j-1} + AU_j - BU_{j+1} &= F_j, \quad 1 \leq j \leq N_2 - 1, \\ U_0 = U_{N_2} &= 0, \end{aligned} \quad (35)$$

где

$$U_j = (u(1, j), u(2, j), \dots, u(N_1 - 1, j)), \quad 0 \leq j \leq N_2,$$

$$F_j = (h_2^2 f(1, j), h_2^2 f(2, j), \dots, h_2^2 f(N_1 - 1, j)), \quad 1 \leq j \leq N_2 - 1,$$

а матрицы B и A определяются соотношениями

$$\begin{aligned} BU_j &= \left(\left(E + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \right) u(1, j), \dots \right. \\ &\quad \left. \dots, \left(E + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \right) u(N_1 - 1, j) \right), \end{aligned}$$

$$\begin{aligned} AU_j &= \left(\left(2E - \frac{5h_2^2 - h_1^2}{6} \Lambda_1 \right) u(1, j), \dots \right. \\ &\quad \left. \dots, \left(2E - \frac{5h_2^2 - h_1^2}{6} \Lambda_1 \right) u(N_1 - 1, j) \right). \end{aligned}$$

Матрицы A и B перестановочные, т. е. $AB = BA$.

Построим комбинированный метод разделения переменных для задачи (34). Сначала совершим первый шаг исключения метода редукции для системы (35). Дадим независимое от изложения главы III описание этого шага. Выпишем три подряд идущих уравнения системы (35) для $j = 2, 4, 6, \dots, N_2 - 2$:

$$\begin{aligned} -BU_{j-2} + AU_{j-1} - BU_j &= F_{j-1}, \\ -BU_{j-1} + AU_j - BU_{j+1} &= F_j, \\ -BU_j + AU_{j+1} - BU_{j+2} &= F_{j+1}, \end{aligned}$$

умножим слева первое и третье уравнения на B , а среднее — на A , и сложим их. В силу перестановочности A и B получим

$$\begin{aligned} -B^2U_{j-2} + (A^2 - 2B^2)U_j - B^2U_{j+2} &= F_j^{(1)}, \\ j = 2, 4, 6, \dots, N_2 - 2, \\ U_0 = U_{N_2} &= 0, \end{aligned}$$

где $F_j^{(1)} = B(F_{j-1} + F_{j+1}) + AF_j$, $j = 2, 4, 6, \dots, N_2 - 2$. Обозначая, как обычно, $V_j = U_{2j}$, $0 \leq j \leq M_2$, $\Phi_j = F_{2j}^{(1)}$, $1 \leq j \leq M_2 - 1$, где $2M_2 = N_2$, запишем эту систему в виде

$$\begin{aligned} -B^2V_{j-1} + (A^2 - 2B^2)V_j - B^2V_{j+1} &= \Phi_j, \quad 1 \leq j \leq M_2 - 1, \\ V_0 = V_{M_2} &= 0, \end{aligned} \quad (36)$$

при этом

$$\Phi_j = B(F_{2j-1} + F_{2j+1}) + AF_{2j}, \quad 1 \leq j \leq M_2 - 1. \quad (37)$$

Остальные неизвестные векторы находятся из уравнений

$$AU_{2j-1} = F_{2j-1} + B(U_{2j-2} + U_{2j}), \quad 1 \leq j \leq M_2. \quad (38)$$

«Укороченную» систему (36) будем, как и раньше, решать методом Фурье. Подставим разложения (12) в (36), где $\mu_{k_2}^{(2)}(j)$ определены (10). В результате для коэффициентов Фурье Y_{k_2} и Z_{k_2} векторов V_j и Φ_j получим соотношение

$$\left(A^2 - 4 \cos^2 \frac{k_2 \pi}{2M_2} B^2 \right) Y_{k_2} = h_2^2 Z_{k_2}, \quad 1 \leq k_2 \leq M_2 - 1, \quad (39)$$

являющееся аналогом соотношения (13), причем компоненты векторов Z_{k_2} и Φ_j связаны формулой (11). Для решения уравнения (39) можно использовать алгоритм

$$\begin{aligned} \left(A - 2 \cos \frac{k_2 \pi}{2M_2} B \right) W_{k_2} &= h_2^2 Z_{k_2}, \\ \left(A + 2 \cos \frac{k_2 \pi}{2M_2} B \right) Y_{k_2} &= W_{k_2}, \quad 1 \leq k_2 \leq M_2 - 1. \end{aligned} \quad (40)$$

Итак, метод решения задачи (34) в векторной форме описывается формулами (37), (11), (40), (12) и (38). Переходя к скалярной записи и к ненормированной собственной функции $\bar{\mu}_{k_2}^{(2)}(j) =$

$= \sin \frac{k_2 \pi j}{M_2}$ при помощи замены из п. 1, получим следующие формулы:

$$\begin{aligned} \varphi(i, j) = & \left(E + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \right) [f(i, 2j-1) + f(i, 2j+1) + 2f(i, 2j)] - \\ & - h_2^2 \Lambda_1 f(i, 2j), \quad 1 \leq j \leq M_2 - 1, \quad 1 \leq i \leq N_1 - 1, \\ f(0, j) = & 0, \quad 1 \leq j \leq N_1 - 1. \end{aligned} \quad (41)$$

для вычисления $\varphi(i, j)$; уравнения

$$4 \sin^2 \frac{k_2 \pi}{2N_2} w_{k_2}(i) - h_2^2 \left(1 - \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2} \cdot \frac{h_1^2 + h_2^2}{12} \right) \Lambda_1 w_{k_2}(i) = h_2^2 z_{k_2}(i), \quad (42)$$

$$1 \leq i \leq N_1 - 1, \quad w_{k_2}(0) = w_{k_2}(N_1) = 0$$

для вычисления $w_{k_2}(i)$ и

$$4 \cos^2 \frac{k_2 \pi}{2N_2} y_{k_2}(i) - h_2^2 \left(1 - \frac{4}{h_2^2} \cos^2 \frac{k_2 \pi}{2N_2} \cdot \frac{h_1^2 + h_2^2}{12} \right) \Lambda_1 y_{k_2}(i) = w_{k_2}(i), \quad (43)$$

$$1 \leq i \leq N_1 - 1, \quad y_{k_2}(0) = y_{k_2}(N_1) = 0$$

для вычисления $y_{k_2}(i)$, которые решаются при $1 \leq k_2 \leq M_2 - 1$, где

$$z_{k_2}(i) = \sum_{j=1}^{M_2-1} \varphi(i, j) \sin \frac{k_2 \pi j}{M_2}, \quad 1 \leq k_2 \leq M_2 - 1, \quad 1 \leq i \leq N_1 - 1. \quad (44)$$

Решение $u(i, j)$ задачи (34) определяется по формулам

$$u(i, 2j) = \frac{4}{N_2} \sum_{k_2=1}^{M_2-1} y_{k_2}(i) \sin \frac{k_2 \pi j}{M_2}, \quad 1 \leq j \leq M_2 - 1, \quad 1 \leq i \leq N_1 - 1, \quad (45)$$

и из уравнений

$$\begin{aligned} 2u(i, 2j-1) - \frac{5h_2^2 - h_1^2}{6} \Lambda_1 u(i, 2j-1) = & h_2^2 f(i, 2j-1) + \\ & + \left(E + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \right) [u(i, 2j-2) + u(i, 2j)], \\ 1 \leq i \leq N_1 - 1, \quad u(0, 2j-1) = & u(N_1, 2j-1) = 0, \quad 1 \leq j \leq M_2. \end{aligned} \quad (46)$$

Нам осталось показать, что трехточечные уравнения (42), (43) и (46) разрешимы. Тогда для нахождения решения можно воспользоваться методом обычной прогонки или методом немонотонной прогонки.

Достаточно показать, что для $1 \leq k_2 \leq N_2 - 1$ собственные значения разностного оператора

$$\mathcal{R} = \lambda_{k_2}^{(2)} E - \left(1 - \frac{h_1^2 + h_2^2}{12} \lambda_{k_2}^{(2)} \right) \Lambda_1, \quad \lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2}$$

отличны от нуля. Действительно, при $1 \leq k_2 \leq N_2/2 - 1$ оператор $h_2^2 \mathcal{R}$ совпадает с оператором задачи (42), а при $k_2 = N_2/2$ — с оператором задачи (46). Если $N_2/2 + 1 \leq k_2 \leq N_2 - 1$, то оператор $h_2^2 \mathcal{R}$ имеет вид

$$h_2^2 \mathcal{R} = 4 \sin^2 \frac{k_2 \pi}{2N_2} - h_2^2 \left(1 - \frac{h_1^2 + h_2^2}{12} \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2} \right) \Lambda_1.$$

Замена $k_2 = N_2 - k'_2$ дает

$$h_2^2 \mathcal{R} = 4 \cos^2 \frac{k'_2 \pi}{2N_2} - h_2^2 \left(1 - \frac{h_1^2 + h_2^2}{12} \frac{4}{h_2^2} \cos^2 \frac{k'_2 \pi}{2N_2} \right) \Lambda_1,$$

где $1 \leq k'_2 \leq N_2/2 - 1$, т. е. в этом случае оператор $h_2^2 \mathcal{R}$ совпадает с оператором задачи (43).

Найдем теперь собственные значения оператора \mathcal{R} для фиксированного значения k_2 . Так как собственными значениями оператора Λ_1 в случае краевых условий первого рода являются (см. § 5 гл. I)

$$\lambda_{k_1}^{(1)} = \frac{4}{h_1^2} \sin^2 \frac{k_1 \pi}{2N_1}, \quad k_1 = 1, 2, \dots, N_1 - 1,$$

то собственные значения λ оператора \mathcal{R} есть

$$\lambda_{k_1 k_2} = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)} - \frac{h_1^2 + h_2^2}{12} \lambda_{k_1}^{(1)} \lambda_{k_2}^{(2)}, \quad 1 \leq k_1 \leq N_1 - 1, \quad 1 \leq k_2 \leq N_2 - 1.$$

Так как имеют место следующие оценки для собственных значений $\lambda_{k_1}^{(1)}$ и $\lambda_{k_2}^{(2)}$:

$$0 < \lambda_{k\alpha}^{(\alpha)} < \frac{4}{h_\alpha^2}, \quad \alpha = 1, 2,$$

то легко получим для любых k_1 и k_2

$$\lambda_{k_1 k_2} = \lambda_{k_1}^{(1)} \left(1 - \frac{h_2^2}{12} \lambda_{k_2}^{(2)} \right) + \lambda_{k_2}^{(2)} \left(1 - \frac{h_1^2}{12} \lambda_{k_1}^{(1)} \right) > \frac{2}{3} (\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}) > 0,$$

что и требовалось доказать.

Несложно найти, что для задачи (42) достаточное условие применимости метода обычной прогонки имеет вид

$$1 + \frac{2h_1^2 - h_2^2}{3h_2^2} \sin^2 \frac{k_2 \pi}{2N_2} \geq 0 \tag{47}$$

и, очевидно, выполнено для любого k_2 . Для задачи (43) аналогичное условие имеет вид

$$1 + \frac{2h_1^2 - h_2^2}{3h_2^2} \cos^2 \frac{k_2 \pi}{2N_2} \geq 0$$

и то же выполнено для всех k_2 . Задаче (46) соответствует условие (47) с $k_2 = 0,5N_2$. Следовательно, задачи (42), (43) и (46) можно решить методом обычной прогонки.

ГЛАВА V

МАТЕМАТИЧЕСКИЙ АППАРАТ ТЕОРИИ ИТЕРАЦИОННЫХ МЕТОДОВ

Настоящая глава содержит сведения и основные понятия теории итерационных методов, которые излагаются в последующих главах. В § 1 изложены простейшие понятия функционального анализа, приведены основные свойства линейных и нелинейных операторов в гильбертовом пространстве, а также некоторые теоремы о разрешимости операторных уравнений. В § 2 проводится систематическая трактовка разностных схем как операторных уравнений в абстрактном пространстве и указываются свойства соответствующих операторов. В § 3 даны основные определения и понятия теории итерационных процессов, рассмотрен канонический вид итерационных схем, даны понятия сходимости и числа итераций.

§ 1. Некоторые сведения из функционального анализа

1. Линейные пространства. В предыдущих главах были изучены основные прямые методы решения простейших разностных уравнений. Построенные методы характеризуются тем, что с их помощью принципиально возможно, проделав конечное число действий, получить точное решение разностной задачи. При этом, естественно, предполагается, что входная информация задана точно и все вычисления проводятся без округления.

Эффективность этих методов достаточно высокая, что достигается учетом структуры матрицы решаемой системы. Требование выполнения специальных свойств матрицы сужает область применимости этих методов, ограничивая ее простейшими задачами.

Для решения сложных и, в частности, нелинейных разностных задач наибольшее распространение получили итерационные методы. Суть итерационных методов состоит в построении тем или иным способом сходящейся к решению последовательности приближений, начиная с некоторого начального приближения. При этом за приближенное решение задачи принимают решение, полученное после конечного числа итераций.

Универсальность итерационных методов заключается прежде всего в том, что они позволяют решать не одну конкретную задачу, а класс задач, обладающих определенными свойствами. Эти свойства определяются не структурой сеточных уравнений, а общими функциональными свойствами. Поскольку в большин-

стве итерационных методов конкретная структура уравнений не используется, то теорию итерационных методов можно строить с единой точки зрения, рассматривая в качестве исходного объекта исследований операторное уравнение первого рода

$$Au = f,$$

где A — оператор, f — заданный, а u — искомый элементы некоторого пространства H .

Прежде чем переходить к построению и исследованию итерационных методов, дадим краткий перечень сведений из функционального анализа (без доказательств).

Линейным пространством над полем K действительных или комплексных чисел называется множество H , для элементов которого определены операции сложения элементов и умножения элемента на число из поля K , причем выполняются следующие аксиомы (x, y, z — элементы из H , λ и μ — числа из K):

1) обе операции не выводят из H ;

2) $x + y = y + x$, $x + (y + z) = (x + y) + z$ (коммутативность и ассоциативность сложения);

3) $\lambda(\mu x) = (\lambda\mu)x$ (ассоциативность умножения);

4) $\lambda(x + y) = \lambda x + \lambda y$, $(\lambda + \mu)x = \lambda x + \mu x$ (дистрибутивность умножения относительно сложения);

5) существует однозначно определенный элемент 0 такой, что $x + 0 = x$ для любого $x \in H$;

6) для каждого $x \in H$ существует однозначно определенный элемент $(-x) \in H$ такой, что $x + (-x) = 0$;

7) $1 \cdot x = x$.

В зависимости от того, на какие числа, вещественные или комплексные, допускается умножение элементов H , мы получаем *вещественное* или *комплексное линейное пространство*.

В линейных пространствах можно ввести понятие линейной зависимости и линейной независимости элементов. Элементы x_1, x_2, \dots, x_n линейного пространства H называются *линейно независимыми*, если из равенства

$$\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n = 0 \quad (1)$$

следует, что $\lambda_1 = \lambda_2 = \dots = \lambda_n = 0$. Если, наоборот, найдутся не все равные нулю $\lambda_1, \lambda_2, \dots, \lambda_n$ такие, что имеет место (1), то элементы x_1, x_2, \dots, x_n называются *линейно зависимыми*.

Пространство H называется *n-мерным*, если в H существуют n линейно независимых элементов, а всякий $(n+1)$ -й элемент линейно зависим.

Непустое замкнутое множество H_1 элементов линейного пространства H называется *подпространством*, если вместе с элементами x_1, x_2, \dots, x_n множество H_1 содержит любую линейную комбинацию $\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n$ этих элементов.

Сумма конечного числа подпространств H_1, H_2, \dots, H_n есть множество элементов вида

$$x = x_1 + x_2 + \dots + x_n, \quad x_i \in H_i, \quad i = 1, 2, \dots, n. \quad (2)$$

Пусть H_1, H_2, \dots, H_n — подпространства, принадлежащие линейному пространству H . Если каждый элемент $x \in H$ однозначно представим в виде (2), то говорят, что H есть *прямая сумма подпространств H_1, H_2, \dots, H_n* , а выражение (2) называется *разложением элемента x по элементам из H_1, H_2, \dots, H_n* .

Будем иметь в этом случае

$$H = H_1 \oplus H_2 \oplus \dots \oplus H_n.$$

Нетрудно показать, что если $H = H_1 \oplus H_2$, то H_1 и H_2 имеют общим лишь нулевой элемент пространства. Обратно, если любой элемент $x \in H$ может быть представлен в виде $x = x_1 + x_2$, $x_1 \in H_1$, $x_2 \in H_2$ и $H_1 \cap H_2 = 0$, то $H = H_1 \oplus H_2$.

Линейное пространство H называется *нормированным*, если для каждого элемента $x \in H$ определено вещественное число $\|x\|$, называемое *нормой*, которое удовлетворяет условиям:

- 1) $\|x\| \geqslant 0$, причем $\|x\| = 0$, если $x = 0$;
- 2) $\|x + y\| \leqslant \|x\| + \|y\|$ (неравенство треугольника);
- 3) $\|\lambda x\| = |\lambda| \|x\|$, λ — число.

Последовательность $\{x_n\}$ элементов линейного нормированного пространства H называется *сходящейся* к элементу $x \in H$, если $\|x - x_n\| \rightarrow 0$ при $n \rightarrow \infty$. Если $\|x_n - x_m\| \rightarrow 0$ при $n, m \rightarrow \infty$, то последовательность $\{x_n\}$ называется *фундаментальной*.

Линейное нормированное пространство H называется *полным*, если всякая фундаментальная последовательность $\{x_n\}$ из этого пространства сходится к некоторому элементу $x \in H$. Полные линейные нормированные пространства называются *банаховыми* пространствами. Всякое конечномерное линейное нормированное пространство полно. Подпространства нормированного пространства нормированы естественным образом.

Одно и то же линейное пространство можно нормировать бесконечным множеством способов. Пусть в линейном пространстве двумя различными способами введены нормы $\|x\|_1$ и $\|x\|_2$. Если существуют такие постоянные $0 < m \leqslant M$, что для любого $x \in H$ верны неравенства

$$m \|x\|_1 \leqslant \|x\|_2 \leqslant M \|x\|_1,$$

то нормы называются *эквивалентными*. Отметим, что в конечномерном пространстве любые две нормы эквивалентны.

Если в линейном пространстве введены две эквивалентные нормы, то из сходимости некоторой последовательности $\{x_n\}$ в одной норме следует сходимость и в другой.

Пусть H — линейное вещественное (комплексное) пространство, и пусть любым двум элементам x, y из H сопоставлено вещественное (комплексное) число (x, y) такое, что:

- 1) $(x, y) = (\overline{y}, \overline{x})$ (симметрия);
- 2) $(x+y, z) = (x, z) + (y, z)$ (дистрибутивность);
- 3) $(\lambda x, y) = \lambda(x, y)$ (однородность);
- 4) $(x, x) \geq 0$ для любого $x \in H$, причем $(x, x) = 0$ тогда и только тогда, когда $x = 0$.

Число (x, y) называется *скалярным произведением* элементов x и y . Черта сверху означает переход к комплексно сопряженному числу.

Линейное нормированное пространство H , в котором норма порождена скалярным произведением $\|x\| = \sqrt{(x, x)}$, называется *унитарным пространством* H . Полное унитарное пространство называется *гильбертовым*. Конечномерное унитарное пространство является полным.

Для скалярного произведения справедливо неравенство Коши—Буняковского $|(x, y)| \leq \|x\| \|y\|$. Элементы x и y унитарного пространства называются *взаимно ортогональными*, если $(x, y) = 0$. Элемент $x \in H$ называется *ортогональным подпространству* H_1 *пространства* H , если x ортогонален любому элементу $y \in H_1$. Множество H_2 всех элементов $x \in H$, ортогональных подпространству H_1 пространства H , называется *ортогональным дополнением* подпространства H_1 . Заметим, что ортогональное дополнение само является подпространством пространства H .

Пусть H_1 —произвольное подпространство пространства H , а H_2 —ортогональное дополнение. Тогда H есть прямая сумма H_1 и H_2 , $H = H_1 \bigoplus H_2$. Следовательно, каждый элемент $x \in H$ представляется единственным образом в виде $x = x_1 + x_2$, $x_\alpha \in H_\alpha$, $\alpha = 1, 2$, причем $(x_1, x_2) = 0$.

Система $x_1, x_2, \dots, x_n, \dots$ элементов пространства H называется *ортогональной системой*, если $(x_m, x_n) = \delta_{mn}$, $m, n = 1, 2, \dots$, где δ_{mn} —символ Кронекера, равный единице при $m = n$ и нулю при $m \neq n$.

Если не существует элемента $x \in H$, отличного от нулевого и ортогонального всем элементам ортонормированной системы $\{x_n\}$, то эта система называется *полной*. Ряд Фурье $\sum_{k=1}^{\infty} c_k x_k$, где $c_k = (x, x_k)$, $k = 1, 2, \dots$, построенный для любого $x \in H$ по полной ортонормированной системе $\{x_n\}$, сходится к этому элементу, и для любого $x \in H$ имеет место равенство

$$\|x\|^2 = (x, x) = \sum_{k=1}^{\infty} c_k^2.$$

2. Операторы в линейных нормированных пространствах. Пусть X и Y —линейные нормированные пространства. Говорят, что на множестве $\mathcal{D} \subset X$ задан оператор A со значениями в Y (оператор, действующий из \mathcal{D} в Y), если каждому элементу $x \in \mathcal{D}$ поставлен в соответствие элемент $y = Ax \in Y$. Множество \mathcal{D}

называется *областью определения оператора* A и обозначается через $\mathcal{D}(A)$. Совокупность всех элементов $y \in Y$, представимых в виде $y = Ax$ ($x \in \mathcal{D}(A)$), называется *областью значений оператора* A и обозначается $\text{im } A$. Если $\mathcal{D}(A) = X$, $\text{im } A \subset X$, т. е. оператор A отображает X в себя, то говорят, что A —оператор в X . Если $\mathcal{D}(A) = X$, $\text{im } A = X$, т. е. оператор A отображает X на себя, то говорят, что A —оператор на X .

Оператор A называется *линейным*, если $\mathcal{D}(A)$ —линейное многообразие в X и для любых $x_1, x_2 \in \mathcal{D}(A)$

$$A(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 Ax_1 + \lambda_2 Ax_2,$$

где λ_1 и λ_2 —числа из поля K .

Линейный оператор A называется *ограниченным*, если существует такая постоянная $M > 0$, что для любых $x \in \mathcal{D}(A)$

$$\|Ax\|_2 \leq M \|x\|_1, \quad (3)$$

где $\|\cdot\|_1$ норма в X , $\|\cdot\|_2$ —норма в Y . Произвольный нелинейный оператор A называется *ограниченным* на $\mathcal{D}(A)$, если

$$\sup_{x \in \mathcal{D}(A)} \|Ax\|_2 < \infty.$$

Для линейного оператора A наименьшая из постоянных M , удовлетворяющих условию (3), называется *нормой* оператора и обозначается $\|A\|$. Из определения нормы следует, что

$$\|A\| = \sup_{\|x\|_1=1} \|Ax\|_2 \quad \text{или} \quad \|A\| = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_1}.$$

Отметим, что в конечномерном пространстве любой линейный оператор ограничен. Пусть A —произвольный оператор, действующий из X в Y . Оператор A называется *непрерывным в точке* $x \in X$, если из условия $\|x_n - x\|_1 \rightarrow 0$ ($x_n \in X$) следует, что $\|Ax_n - Ax\|_2 \rightarrow 0$ при $n \rightarrow \infty$. Линейный ограниченный оператор непрерывен.

Произвольный оператор A удовлетворяет *условию Липшица с постоянной* q , если

$$\|Ax_1 - Ax_2\|_2 \leq q \|x_1 - x_2\|_1, \quad x_1, x_2 \in \mathcal{D}(A). \quad (4)$$

Любой линейный ограниченный оператор A удовлетворяет условию Липшица (4) с $q = \|A\|$.

Пусть A —произвольный оператор, действующий из X в Y . Линейный ограниченный оператор $A'(x)$ называется *производной Гато оператора* A в точке x пространства X , если для любого $z \in X$

$$\lim_{t \rightarrow 0} \left\| \frac{A(x + tz) - Ax}{t} - A'(x) z \right\|_2 = 0.$$

При этом область значений оператора A' принадлежит Y .

Если оператор A имеет производную Гато в каждой точке пространства X , то для любых $x_1, x_2 \in X$ справедливо неравен-

ство (4), где $q = \sup_{0 \leq t \leq 1} \|A'(x_1 + t(x_2 - x_1))\|$. Если A — линейный оператор, то $A' = A$.

Всевозможные линейные ограниченные операторы, действующие из X в Y , образуют *линейное нормированное пространство*, так как норма $\|A\|$ оператора A удовлетворяет всем аксиомам нормы. Рассмотрим множество линейных ограниченных операторов, действующих из X в X . На этом множестве можно ввести произведение AB операторов A и B следующим образом: $(AB)x = A(Bx)$. Очевидно, что AB — линейный ограниченный оператор: $\|AB\| \leq \|A\|\|B\|$.

Если $(AB)x = (BA)x$ для всех $x \in X$, то операторы A и B называются *перестановочными* или *коммутативными*; в этом случае пишут $AB = BA$.

В связи с решением уравнений вида $Ax = y$ вводится понятие *обратного* оператора A^{-1} . Пусть A — оператор из X на Y . Если каждому $y \in Y$ соответствует только один $x \in X$, для которого $Ax = y$, то этим соотвествием определяется оператор A^{-1} , называемый *обратным* для A и имеющий область определения Y , а область значений X .

Для любых $x \in X$ и $y \in Y$ имеем тождества $A^{-1}(Ax) = x$, $A(A^{-1}y) = y$. Нетрудно показать, что если A линеен, то и A^{-1} (если он существует) также линеен.

Лемма 1. Для того чтобы линейный оператор A , отображающий X на Y , имел обратный, необходимо и достаточно, чтобы $Ax = 0$ только при $x = 0$.

Теорема 1. Пусть A — линейный оператор из X на Y . Для того чтобы обратный оператор A^{-1} существовал и был ограниченным (как оператор из Y на X), необходимо и достаточно, чтобы существовала такая постоянная $\delta > 0$, что для всех $x \in X$

$$\|Ax\|_2 \geq \delta \|x\|_1.$$

При этом справедлива оценка $\|A^{-1}\| \leq 1/\delta$. Здесь $\|\cdot\|_1$ — норма в X , а $\|\cdot\|_2$ — норма в Y .

Иными словами, для существования обратного оператора A^{-1} необходимо и достаточно, чтобы однородное уравнение $Ax = 0$ имело только тривиальное решение.

Пусть A и B линейные ограниченные операторы, действующие в X и имеющие обратные. Тогда $(AB)^{-1} = B^{-1}A^{-1}$.

Если оператор A обратим, то имеют смысл степени A^k с любыми (а не только неотрицательными) целыми показателями. Именно, по определению $A^{-k} = (A^{-1})^k$, $k = 1, 2, \dots$. Степени одного и того же оператора коммутируют.

Введем понятие *ядра* линейного оператора A . Ядром линейного оператора A называется множество всех тех элементов x пространства X , для которых $Ax = 0$. Ядро линейного оператора A обозначается символом $\ker A$.

Условие $\ker A = 0$ является необходимым и достаточным для того, чтобы оператор A имел обратный.

Подпространство X_1 пространства X называется *инвариантным подпространством* оператора A , действующего в X , если A не выводит элементы из X_1 , т. е. $Ax \in X_1$, если $x \in X_1$.

Если подпространство X_1 инвариантно относительно обратного оператора A , то оно инвариантно относительно оператора A^{-1} .

Примерами инвариантных подпространств оператора A могут служить $\ker A$ и $\text{im } A$. Заметим, что если операторы A и B коммутируют, то подпространства $\ker B$ и $\text{im } B$ инвариантны относительно оператора A .

Число

$$\rho(A) = \lim_{k \rightarrow \infty} \sqrt[k]{\|A^k\|}$$

называется *спектральным радиусом линейного оператора A* . Оно не зависит от определения нормы, причем $\rho(A) = \inf_{\|\cdot\|} \|A\|$.

Для любого линейного ограниченного оператора A справедливы неравенства

$$\rho(A) \leq \|A\|, \quad \rho(A) \leq \sqrt[k]{\|A^k\|}, \quad k = 2, 3, \dots$$

Лемма 2. Для того чтобы $\|A\| = \rho(A)$, необходимо и достаточно, чтобы $\|A^k\| = \|A\|^k$, $k = 2, 3, \dots$

Отметим еще одно свойство спектрального радиуса. Если операторы A и B коммутируют, то

$$\rho(AB) \leq \rho(A)\rho(B), \quad \rho(A+B) \leq \rho(A)+\rho(B).$$

3. Операторы в гильбертовом пространстве. Пусть линейный ограниченный оператор A действует в унитарном пространстве H . Согласно общему определению нормы оператора имеем

$$\|A\| = \sup_{\|x\|=1} \|Ax\| = \sup_{x \in H} \sqrt{\frac{(Ax, Ax)}{(x, x)}}$$

и, следовательно, для любого $x \in H$ верно неравенство

$$(Ax, Ax) \leq \|A\|^2 (x, x).$$

Используя неравенство Коши—Буняковского, отсюда получим

$$|(Ax, x)| \leq \|Ax\| \|x\| \leq \|A\| (x, x). \quad (5)$$

Далее будем рассматривать только ограниченные операторы.

Оператор A^* называется *сопряженным* оператору A , если для любых $x, y \in H$ выполнено тождество

$$(Ax, y) = (x, A^*y).$$

Для любого линейного ограниченного оператора A с областью определения $\mathcal{D}(A) = H$ существует, причем единственный, оператор A^* с областью определения $\mathcal{D}(A^*) = H$. Оператор A^* линеен и ограничен, $\|A^*\| = \|A\|$.

Приведем основные свойства операции сопряжения: $(A^*)^* = A$, $(A + B)^* = A^* + B^*$, $(AB)^* = B^*A^*$, $(\lambda A)^* = \bar{\lambda}A^*$. Если операторы A и B коммутируют, то коммутируют и сопряженные операторы A^* и B^* . Если A имеет обратный, то $(A^{-1})^* = (A^*)^{-1}$, т. е. операции взятия обратного оператора и сопряжения перестановочны.

Лемма 3. Пусть A — линейный оператор в H . Пространство H представимо в виде прямых сумм ортогональных подпространств

$$H = \ker A \bigoplus \text{im } A^*, \quad H = \ker A^* \bigoplus \text{im } A.$$

Действительно, пусть H_1 — ортогональное дополнение $\text{im } A^*$ до пространства H , т. е.

$$H = H_1 \bigoplus \text{im } A^*, \quad (x_1, x_2) = 0, \quad x_1 \in H_1, \quad x_2 \in \text{im } A^*.$$

Покажем, что $H_1 = \ker A$. Пусть $x_1 \in \ker A$, тогда для любого $x \in H$ имеем $A^*x \in \text{im } A^*$ и

$$(x_1, A^*x) = (Ax_1, x) = 0.$$

Следовательно, x_1 ортогонально $\text{im } A^*$, и поэтому $x_1 \in H_1$. С другой стороны, пусть $x_1 \in H_1$ (следовательно, x_1 ортогонален $\text{im } A^*$). Тогда для любого $x \in H$

$$0 = (x_1, A^*x) = (Ax_1, x).$$

Так как x — любой элемент H , то $Ax_1 = 0$ и, следовательно, $x_1 \in \ker A$. Первое утверждение леммы доказано. Аналогично доказывается и второе.

Линейный оператор A называется *самосопряженным* в H , если $A = A^*$. Для самосопряженного оператора $(Ax, y) = (x, Ay)$ для всех $x, y \in H$.

Оператор A называется *нормальным*, если он коммутирует со своим сопряженным, $A^*A = AA^*$, и *кососимметричным*, если $A^* = -A$. Самосопряженные и кососимметричные операторы нормальны.

Известно, что если A и B — самосопряженные операторы, то оператор AB является самосопряженным тогда и только тогда, когда A и B перестановочны.

Если A — линейный оператор, то A^*A и AA^* самосопряженные операторы, причем $\|A^*A\| = \|AA^*\| = \|A\|^2$ и

$$\begin{aligned} \ker A^*A &= \ker A, \quad \text{im } A^*A = \text{im } A^*, \\ \ker AA^* &= \ker A^*, \quad \text{im } AA^* = \text{im } A. \end{aligned}$$

Любой оператор A можно представить в виде суммы самосопряженного A_0 и кососимметричного A_1 операторов

$$A = A_0 + A_1,$$

где $A_0 = 0,5(A + A^*)$, $A_1 = 0,5(A - A^*)$. Если H —вещественное пространство, то отсюда вытекают равенства

$$(Ax, x) = (A_0x, x), \quad (A_1x, x) = 0.$$

В комплексном пространстве H имеет место декартово представление оператора A :

$$A = A_0 + iA_1,$$

где $A_0 = \operatorname{Re} A = \frac{1}{2}(A + A^*)$, $A_1 = \operatorname{Im} A = \frac{1}{2i}(A - A^*)$ —самосопряженные в H операторы. При этом для любых $x \in H$ справедливы тождества

$$\operatorname{Re}(Ax, x) = (A_0x, x), \quad \operatorname{Im}(Ax, x) = (A_1x, x).$$

Если A —самосопряженный в H оператор, то имеет место формула

$$\|A\| = \sup_{x \neq 0} \frac{|(Ax, x)|}{(x, x)}, \quad x \in H.$$

Лемма 4. Если A —самосопряженный ограниченный в H оператор, то при любом целом $n > 0$ верно равенство $\|A^n\| = \|A\|^n$.

Лемма 4 остается справедливой и для нормального оператора.

Из лемм 2 и 4 следует, что для нормального (в частности, для самосопряженного) оператора A имеет место равенство $\rho(A) = \|A\|$.

Лемма 5. Пусть в линейном пространстве H двумя способами введено скалярное произведение элементов x и y : $(x, y)_1$ и $(x, y)_2$. Если оператор A самосопряжен в смысле каждого скалярного произведения, то $\|A\|_1 = \|A\|_2 = \rho(A)$.

Спектральный радиус дает оценку снизу для любой нормы оператора. Введем числовой радиус оператора, позволяющий получить двусторонние оценки для нормы.

Числовой радиус оператора A , действующего в комплексном пространстве H , определим следующим образом:

$$\bar{\rho}(A) = \sup_{\|x\|=1} |(Ax, x)|, \quad x \in H.$$

Для любого линейного ограниченного оператора A справедливы неравенства: $\mu(A)\|A\| \leq \bar{\rho}(A) \leq \|A\|$, $\mu(A) \geq 1/2$ и, кроме того, $\bar{\rho}(A^n) \leq [\bar{\rho}(A)]^n$ для любого натурального n . Если оператор A самосопряжен, то $\bar{\rho}(A) = \|A\|$. Отметим еще ряд интересных свойств числового радиуса. Так, например, $\bar{\rho}(A^*) = \bar{\rho}(A)$, $\bar{\rho}(A^*A) = \|A\|^2$. Кроме того, $\rho(A) \leq \bar{\rho}(A)$, где $\rho(A)$ —введенный ранее спектральный радиус оператора.

Линейный оператор A , действующий в гильбертовом пространстве H , называется *положительным* ($A > 0$), если $(Ax, x) > 0$ для всех $x \in H$, кроме $x = 0$. В случае комплексного простран-

ства H определение положительности вводится только для самосопряженных операторов, так как из положительности оператора в этом случае уже следует его самосопряженность.

Аналогично вводится определение *неотрицательности* оператора A (для всех $x \in H$ $(Ax, x) \geq 0$) и *положительной определенности* (для всех $x \in H$ $(Ax, x) \geq \delta(x, x)$, где $\delta > 0$).

Нелинейный оператор A , действующий в H , называется *монотонным*, если

$$(Ax - Ay, x - y) \geq 0, \quad x, y \in H,$$

строго монотонным, если

$$(Ax - Ay, x - y) > 0, \quad x, y \in H, \quad x \neq y,$$

сильно монотонным, если для всех $x, y \in H$ имеет место неравенство

$$(Ax - Ay, x - y) \geq \delta \|x - y\|^2, \quad \delta > 0.$$

Теорема 2. Пусть нелинейный оператор A имеет непрерывную в каждой точке $x \in H$ производную Гато. Тогда оператор A сильно монотонен на H в том и только в том случае, когда существует такое $\delta > 0$, что

$$(A'(x)y, y) \geq \delta(y, y), \quad y \in H.$$

Пусть A — неотрицательный линейный оператор. Число (Ax, x) назовем *энергией оператора*. Будем сравнивать операторы A и B по энергии. Если $((A - B)x, x) \geq 0$ для всех $x \in H$, то будем писать $A \geq B$.

Если существуют такие постоянные $\gamma_2 \geq \gamma_1 > 0$, что для линейных операторов A и B верны неравенства $\gamma_1 B \leq A \leq \gamma_2 B$, то такие операторы будем называть *энергетически эквивалентными* (эн. эк.), а γ_1 и γ_2 — постоянными энергетической эквивалентности операторов A и B . Пусть

$$\delta = \inf_{\|x\|=1} (Ax, x) \quad \text{и} \quad \Delta = \sup_{\|x\|=1} (Ax, x).$$

Числа δ и Δ называются *границами* оператора A (самосопряженного в случае комплексного H). Очевидно, что верны неравенства

$$\delta(x, x) \leq (Ax, x) \leq \Delta(x, x), \quad x \in H$$

или

$$\delta E \leq A \leq \Delta E,$$

где E — тождественный оператор, $Ex = x$.

Нетрудно убедиться в том, что введенное на множестве линейных операторов, действующих в H , отношение неравенства обладает следующими свойствами:

- 1) из $A \geq B$ и $C \geq D$ следует $A + C \geq B + D$,
- 2) из $A \geq 0$ и $\lambda \geq 0$ следует $\lambda A \geq 0$,

- 3) из $A \geqslant B$ и $B \geqslant C$ следует $A \geqslant C$,
 4) если $A > 0$ и A^{-1} существует, то $A^{-1} > 0$.

Далее очевидно, что A^*A и AA^* —неотрицательные операторы для любого линейного оператора A . Эти операторы будут положительны, если A —положительный оператор.

Теорема 3. *Произведение AB двух перестановочных неотрицательных операторов A и B , один из которых самосопряжен, есть также неотрицательный оператор.*

Для любого самосопряженного неотрицательного оператора A имеет место обобщенное неравенство Коши—Буняковского

$$|(Ax, y)| \leqslant \sqrt{(Ax, x)} \sqrt{(Ay, y)}, \quad x, y \in H.$$

Пусть D —самосопряженный положительный оператор, действующий в H . Тогда можно ввести *энергетическое пространство* H_D , состоящее из элементов H , со скалярным произведением $(x, y)_D = (Dx, y)$ и нормой

$$\|x\|_D = \sqrt{(Dx, x)}.$$

Отметим, что если D —самосопряженный положительно определенный и ограниченный в H оператор, то для любого $x \in H$ в силу неравенства Коши—Буняковского справедливы оценки $\delta(x, x) \leqslant (Dx, x) \leqslant \|Dx\| \|x\| \leqslant \Delta(x, x)$, $\Delta = \|D\|$, $\delta > 0$.

Эти неравенства можно записать в виде

$$\sqrt{\delta} \|x\| \leqslant \|x\|_D \leqslant \sqrt{\Delta} \|x\|,$$

откуда следует, что обычная норма $\|\cdot\|$ и энергетическая норма $\|\cdot\|_D$ эквивалентны.

Заметим, что унитарное энергетическое пространство H_D можно построить, исходя из несамосопряженного положительного оператора D . Для этого скалярное произведение в H_D определим следующим образом:

$$(x, y)_D = (D_0 x, y), \text{ где } D_0 = 0,5 (D + D^*).$$

Приведем ряд лемм, содержащих основные неравенства, необходимые нам для дальнейшего.

Лемма 6. *Пусть для линейного оператора выполнено условие $A \geqslant \delta E$, $\delta > 0$. Тогда для любого $x \in H$ имеет место неравенство*

$$(Ax, Ax) \geqslant \delta (Ax, x).$$

Если для неотрицательного самосопряженного оператора выполнено условие $A \leqslant \Delta E$, то для любого $x \in H$ имеет место неравенство

$$(Ax, Ax) \leqslant \Delta (Ax, x).$$

Лемма 7. *Из условия $(Ax, Ax) \leqslant \Delta (Ax, x)$, $x \in H$, $\Delta > 0$ для неотрицательного оператора A следует неравенство*

$$A \leqslant \Delta E,$$

а из условия $(Ax, Ax) \geq \delta (Ax, x)$, $\delta > 0$, **для неотрицательного самосопряженного оператора** A **следует неравенство**

$$A \geq \delta E.$$

Следствие 1. *Из лемм 6 и 7 вытекает, что для самосопряженного положительно определенного оператора A неравенства*

$$\delta E \leq A \leq \Delta E, \quad \delta > 0,$$

и

$$\delta (Ax, x) \leq (Ax, Ax) \leq \Delta (Ax, x), \quad \delta > 0,$$

эквивалентны.

Следствие 2. *Из (5) и леммы 6 следует оценка* $(Ax, Ax) \leq \|A\|(Ax, x)$, $x \in H$, *для неотрицательного самосопряженного в H оператора A .*

Лемма 8. *Пусть A —положительный ограниченный в H самосопряженный оператор $A > 0$, $\|Ax\| \leq \Delta \|x\|$. Тогда обратный оператор A^{-1} является положительно определенным $A^{-1} \geq \frac{1}{\Delta} E$.*

Лемма 9. *Пусть A и B —самосопряженные положительно определенные в H операторы. Тогда неравенства*

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_2 \geq \gamma_1 > 0$$

и

$$\gamma_1 A^{-1} \leq B^{-1} \leq \gamma_2 A^{-1}, \quad \gamma_2 \geq \gamma_1 > 0$$

эквивалентны.

Лемма 10. *Если A —положительно определенный оператор $A \geq \delta E$, $\delta > 0$, то существует обратный оператор A^{-1} и $\|A^{-1}\| \leq 1/\delta$.*

Доказательство следует из неравенства

$$\delta \|x\|^2 \leq (Ax, x) \leq \|Ax\| \|x\|, \quad \delta > 0$$

и из теоремы 1.

Замечание. Если A —положительный оператор, то A^{-1} существует. В случае комплексного пространства H для существования оператора A^{-1} достаточно положительности действительной составляющей $A_0 = 0,5 (A + A^*)$ или положительности мнимой составляющей $A_1 = \frac{1}{2i} (A - A^*)$ оператора A .

4. Функции от ограниченного оператора. В теории итерационных методов нам придется иметь дело с функциями от оператора. Пусть A —ограниченный линейный оператор, действующий в нормированном пространстве X . Если $f(\lambda)$ —целая аналитическая функция переменного λ , разлагающаяся в ряд $\sum_{k=0}^{\infty} a_k \lambda^k$, то можно определить **функцию $f(A)$ от оператора A с помощью формулы**

$f(A) = \sum_{k=0}^{\infty} a_k A^k$. Оператор $f(A)$ будет также линейным и ограниченным. В качестве примера приведем экспоненциальную функцию оператора $e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$. Введенное определение функции от оператора можно распространить на более широкий класс функций и построить операторное исчисление для ограниченных операторов. Мы дадим более общее определение лишь для самосопряженных ограниченных операторов в гильбертовом пространстве.

Пусть δ и Δ —нижняя и верхняя границы самосопряженного в H оператора A . Пусть $f(\lambda)$ —непрерывная на отрезке $[\delta, \Delta]$ функция. Оператор $f(A)$ называется *функцией самосопряженного оператора A* .

Соответствие между функциями вещественной переменной и функциями от оператора обладает следующими свойствами:

- 1) Если $f(\lambda) = \alpha f_1(\lambda) + \beta f_2(\lambda)$, то $f(A) = \alpha f_1(A) + \beta f_2(A)$.
- 2) Если $f(\lambda) = f_1(\lambda) f_2(\lambda)$, то $f(A) = f_1(A) f_2(A)$.
- 3) Из $AB = BA$ следует $f(A)B = Bf(A)$ для любого ограниченного линейного оператора B .

4) Если $f_1(\lambda) \leq f(\lambda) \leq f_2(\lambda)$ для всех $\lambda \in [\delta, \Delta]$, то $f_1(A) \leq f(A) \leq f_2(A)$.

5) $\|f(A)\| \leq \max_{\delta < \lambda < \Delta} |f(\lambda)|$.

6) $\bar{f}(A) = [f(A)]^*$, где черта над функцией означает переход к комплексно сопряженной функции. Если $f(\lambda)$ —вещественная функция, то отсюда следует, что оператор $f(A)$ самосопряжен в H .

Из свойства 4) следует, что если $f(\lambda) \geq 0$ на $[\delta, \Delta]$, то $f(A)$ —неотрицательный оператор.

Важным примером функции от оператора является корень квадратный из оператора. Оператор B называется *квадратным корнем из оператора A* , если $B^2 = A$.

Теорема 4. Существует единственный неотрицательный самосопряженный квадратный корень из любого неотрицательного самосопряженного оператора A , перестановочный со всяким оператором, перестановочным с A .

Квадратный корень из оператора A будем обозначать $A^{1/2}$. Отметим следующее свойство: $\|A\| = \|A^{1/2}\|^2$, если $A = A^* \geq 0$.

Теорема 5. Если A —самосопряженный положительно определенный оператор, $A = A^* \geq \delta E$, $\delta > 0$, то существует ограниченный самосопряженный оператор $A^{-1/2}$, $\|A^{-1/2}\| \leq 1/\sqrt{\delta}$.

Доказательство следует из неравенства

$$\delta(x, x) \leq (Ax, x) = (A^{1/2}x, A^{1/2}x) = \|A^{1/2}x\|^2$$

и из теоремы 1.

5. Операторы в конечномерном пространстве. Рассмотрим n -мерное унитарное пространство H . Пусть элементы x_1, x_2, \dots, x_n

образуют ортонормированный базис в H . По определению конечномерного пространства любой элемент $x \in H$ можно единственным образом представить в виде линейной комбинации

$$x = c_1 x_1 + c_2 x_2 + \dots + c_n x_n. \quad (6)$$

Из ортонормированности системы x_1, x_2, \dots, x_n следует, что $c_k = (x, x_k)$.

Таким образом, каждому элементу $x \in H$ можно поставить в соответствие вектор $c = (c_1, c_2, \dots, c_n)$, компонентами которого являются коэффициенты c_k из разложения (6).

Пусть A — линейный оператор, заданный на H . В базисе x_1, x_2, \dots, x_n ему соответствует матрица $\mathcal{A} = (a_{ik})$ размера $n \times n$, где $a_{ik} = (\bar{A}x_k, x_i)$. Обратно, всякая матрица \mathcal{A} размера $n \times n$ определяет линейный оператор в H . При этом элементу $\bar{A}x$ ставится в соответствие вектор $\left(\sum_{k=1}^n a_{1k} c_k, \sum_{k=1}^n a_{2k} c_k, \dots, \sum_{k=1}^n a_{nk} c_k \right)$, т. е. вектор $\mathcal{A}c$.

Если оператор A самосопряжен в H , то соответствующая ему матрица \mathcal{A} симметрична в любом ортонормированном базисе. Отметим, что в неортонормированном базисе самосопряженному оператору A соответствует несимметричная матрица.

Остановимся на свойствах собственных значений и собственных элементов линейного оператора A . Число λ называется *собственным значением оператора A*, если уравнение

$$Ax = \lambda x \quad (7)$$

имеет ненулевые решения. Элемент $x \neq 0$, удовлетворяющий (7), называется *собственным элементом оператора A*, соответствующим собственному значению λ . Иначе, собственные значения оператора A — это те значения λ , для которых $\ker(A - \lambda E) \neq 0$; собственные элементы, соответствующие собственному значению λ , — это отличные от нуля элементы подпространства $\ker(A - \lambda E)$. Само это подпространство называется *собственным подпространством*, соответствующим собственному значению λ .

Множество $\sigma(A)$ собственных значений оператора A называется *спектром оператора A*.

1. Самосопряженный оператор A имеет n ортонормированных собственных элементов x_1, x_2, \dots, x_n . Соответствующие собственные значения λ_k , $k = 1, 2, \dots, n$ вещественны. Если все собственные значения различны, то A называется *оператором с простым спектром*.

2. Для самосопряженного оператора A имеют место равенства

$$\|A\| = \rho(A) = \max_{1 \leq k \leq n} |\lambda_k|,$$

где $\rho(A)$ — *спектральный радиус оператора A*. Эти равенства сохраняются и для нормального оператора A .

3. Если $A = A^* \geqslant 0$, то все собственные значения оператора A неотрицательны. При этом для любого $x \in H$

$$\delta(x, x) \leqslant (Ax, x) \leqslant \Delta(x, x),$$

где $0 \leqslant \delta = \min_k \lambda_k$, $\Delta = \max_k \lambda_k$. Для самосопряженного оператора A отношением Релея называют выражение $(Ax, x)/(x, x)$.

Наибольшее и наименьшее собственные значения оператора A определяются с помощью отношения Релея следующим образом:

$$\delta = \min_{x \neq 0} \frac{(Ax, x)}{(x, x)}, \quad \Delta = \max_{x \neq 0} \frac{(Ax, x)}{(x, x)}.$$

4. Будем обозначать через $\lambda(A)$ собственные значения оператора A . Пусть $f(A)$ — функция от самосопряженного оператора A . Тогда $\lambda(f(A)) = f(\lambda(A))$ (теорема об отображении спектров).

5. Если самосопряженные операторы A и B перестановочны, $A = A^*$, $B = B^*$, $AB = BA$, то они имеют общую систему собственных элементов. При этом операторы AB и $A + B$ имеют ту же систему собственных элементов, что и операторы A и B , и собственные значения

$$\lambda(AB) = \lambda(A)\lambda(B), \quad \lambda(A+B) = \lambda(A) + \lambda(B).$$

6. Произвольный элемент $x \in H$ можно разложить по собственным элементам самосопряженного оператора A

$$x = \sum_{k=1}^n c_k x_k, \quad c_k = (x, x_k), \quad \text{причем } \|x\|^2 = \sum_{k=1}^n c_k^2.$$

Число λ называется *собственным значением оператора A относительно оператора B*, если уравнение

$$Ax = \lambda Bx \tag{8}$$

имеет ненулевые решения. Элемент $x \neq 0$, удовлетворяющий уравнению (8), называется *собственным элементом оператора A относительно оператора B*, соответствующим числу λ .

7. Если операторы A и B самосопряжены в H , а оператор B , кроме того, положительно определен, то существует n собственных элементов x_1, x_2, \dots, x_n , ортонормированных в энергетическом пространстве H_B : $(x_k, x_i)_B = \delta_{ki}$, $k, i = 1, 2, \dots, n$. Соответствующие собственные значения вещественны и имеют место неравенства

$$\gamma_1(Bx, x) \leqslant (Ax, x) \leqslant \gamma_2(Bx, x),$$

где

$$\gamma_1 = \min_k \lambda_k = \min_{x \neq 0} \frac{(Ax, x)}{(Bx, x)},$$

$$\gamma_2 = \max_k \lambda_k = \max_{x \neq 0} \frac{(Ax, x)}{(Bx, x)}.$$

Следовательно, постоянные эн. эк. самосопряженных операторов A и B в случае положительно определенного оператора B совпадают с минимальным и максимальным собственными значениями обобщенной задачи (8).

6. Разрешимость сператорных уравнений. Пусть требуется найти решение операторного уравнения первого рода

$$Au = f, \quad (9)$$

где A — линейный ограниченный оператор в гильбертовом пространстве H , f — заданный, а u — искомый элементы H . Будем предполагать, что H конечномерно. Нас будет интересовать вопрос о разрешимости уравнения (9). Имеет место

Теорема 6. Для того чтобы уравнение (9) было разрешимо при любой правой части f , необходимо и достаточно, чтобы соответствующее однородное уравнение $Au = 0$ имело только триivialное решение $u = 0$. При этом решение уравнения (9) единственno.

Доказательство теоремы основано на лемме 1.

Формулировка теоремы можно придать иной вид: уравнение (9) однозначно разрешимо при любой $f \in H$ тогда и только тогда, когда $\ker A = 0$ (см. п. 2).

Если $\ker A \neq 0$, то уравнение разрешимо лишь при дополнительном ограничении на f . Напомним, что в силу леммы 3 пространство H есть прямая сумма ортогональных подпространств: $H = \ker A \bigoplus \text{im } A^*$, $H = \ker A^* \bigoplus \text{im } A$.

Теорема 7. Для разрешимости неоднородного уравнения (9) необходимо и достаточно, чтобы правая часть f была ортогональна подпространству $\ker A^*$. В этом случае решение не единствено и определяется с точностью до произвольного элемента, принадлежащего $\ker A$:

$$u = \tilde{u} + \bar{u}, \quad \tilde{u} \in \ker A, \quad \bar{u} \in \text{im } A^*.$$

Пусть f ортогонально $\ker A^*$. Нормальным решением уравнения (9) называется решение, имеющее минимальную норму.

Лемма 11. Нормальное решение единственно и принадлежит подпространству $\text{im } A^*$ (т. е. ортогонально $\ker A$).

Действительно, пусть $u = \tilde{u} + \bar{u}$, $\tilde{u} \in \ker A$, $\bar{u} \in \text{im } A^*$. Тогда $\|u\|^2 = (u, u) = \|\tilde{u}\|^2 + \|\bar{u}\|^2 \geq \|\bar{u}\|^2$, так как \tilde{u} — произвольный элемент подпространства $\ker A$. Следовательно, норма $\|u\|$ будет минимальной, если $u = \bar{u} \in \text{im } A^*$.

Пусть условие ортогональности f подпространству $\ker A^*$ не выполнено. Тогда решение уравнения (9) в классическом смысле не существует. Пусть

$$f = \tilde{f} + \bar{f}, \quad \tilde{f} \in \ker A^*, \quad \bar{f} \in \text{im } A.$$

Обобщенным решением уравнения (9) называется элемент $u \in H$, для которого $Au = \bar{f}$; обобщенное решение доставляет минимум

функционалу $\|Au - f\|$. Действительно, так как $(Au - \tilde{f}) \in \text{im } A$ для любого $u \in H$, то

$$\|Au - f\|^2 = \|Au - f\|^2 + \|\tilde{f}\|^2 \geq \|\tilde{f}\|^2,$$

причем равенство достигается, если u — обобщенное решение.

Обобщенное решение определяется с точностью до произвольного элемента из подпространства $\text{ker } A$. Назовем обобщенным нормальным решением уравнения (9) обобщенное решение, имеющее минимальную норму. Нормальное решение единственно и принадлежит $\text{im } A^*$.

Введенное здесь понятие нормального решения, очевидно, полностью согласуется с данным выше. Отметим, что если существует классическое нормальное решение, то оно совпадает с обобщенным нормальным решением.

Рассмотрим теперь уравнение (9) с произвольным нелинейным оператором A , действующим в гильбертовом пространстве H . В этом случае для доказательства существования и единственности решения уравнения (9) часто используют принцип сжатых отображений С. Банаха.

Теорема 8. Пусть в гильбертовом пространстве H задан оператор B , отображающий замкнутое множество T пространства H в себя. Пусть, кроме того, оператор B является равномерно сжимающим, т. е. удовлетворяет условию Липшица

$$\|Bx - By\| \leq q \|x - y\|, \quad x, y \in T,$$

где $q < 1$ и не зависит от x и y . Тогда существует одна и только одна точка $x_* \in T$ такая, что $x_* = Bx_*$.

Точка x_* называется неподвижной точкой оператора B .

Следствие 1. Если оператор B имеет производную Гато в H , которая удовлетворяет условию $\|B'(x)\| \leq q < 1$ для любого $x \in H$, то уравнение $x = Bx$ имеет в H единственное решение.

Следствие 2. Пусть оператор C отображает замкнутое множество T в себя и коммутирует с оператором B , удовлетворяющим условиям принципа сжатых отображений. Тогда неподвижная точка оператора B является неподвижной точкой (возможно неединственной) оператора C . В частности, если некоторая итерация B^n оператора B удовлетворяет принципу сжатых отображений, то неподвижная точка оператора B^n является неподвижной точкой (единственной) и оператора B .

Вернемся теперь к решению уравнения (9) с нелинейным оператором A . Имеет место

Теорема 9. Пусть оператор A имеет в каждой точке $x \in H$ производную Гато $A'(x)$ и существует $\tau \neq 0$ такое, что для всех $x \in H$ выполнена оценка $\|E - \tau A'(x)\| \leq q < 1$. Тогда уравнение (9) имеет в H единственное решение.

Действительно, уравнение (9) можно записать в следующем виде:

$$u = u - \tau Au + \tau f, \quad \tau \neq 0. \tag{10}$$

Определим оператор B : $Bx = x - \tau Ax + \tau f$. Очевидно, что оператор B имеет производную Гато, равную $B'(x) = E - \tau A'(x)$. В силу условий теоремы имеем $\|B'(x)\| \leq q < 1$ для любого $x \in H$. Поэтому из следствия 1 теоремы 8 вытекает существование и единственность решения уравнения (10) и, следовательно, уравнения (9). Теорема доказана.

Отметим, что в главе VI будут рассмотрены некоторые способы получения оценок для норм линейных операторов вида $E - \tau C$, где τ — число.

Принципом сжатых отображений не исчерпываются все случаи, когда решение нелинейного уравнения существует. При доказательстве разрешимости операторного уравнения (9) можно использовать один из вариантов теоремы о неподвижной точке — *принцип Браудера*.

Теорема 10. Пусть в конечномерном гильбертовом пространстве H непрерывный монотонный (строго монотонный) оператор B удовлетворяет условию

$$(Bx, x) \geq 0 \quad \text{для } \|x\| = \rho > 0.$$

Тогда уравнение $Bx = 0$ имеет в шаре $\|x\| \leq \rho$ по крайней мере одно (соответственно единственное) решение

Воспользуемся этой теоремой и сформулируем условия, при выполнении которых операторное уравнение (9) однозначно разрешимо при любой правой части f .

Теорема 11. Пусть в конечномерном гильбертовом пространстве H задано уравнение (9) с непрерывным и сильно монотонным оператором A ,

$$(Ax - Ay, x - y) \geq \delta \|x - y\|^2, \quad \delta > 0, \quad x, y \in H.$$

Тогда в шаре $\|u\| \leq \frac{1}{\delta} \|A0 - f\|$ уравнение (9) имеет единственное решение.

Действительно, запишем уравнение (4) в следующем виде:

$$Bu = Au - f = 0.$$

Видно, что оператор B непрерывный и сильно монотонный. Используя условие теоремы и неравенство Коши — Буняковского, получим

$$\begin{aligned} (Bx, x) &= (Ax - f, x) = (Ax - A0, x - 0) - (f - A0, x) \geq \\ &\geq \delta \|x\|^2 - \|f - A0\| \|x\| = (\delta \|x\| - \|A0 - f\|) \|x\|. \end{aligned}$$

Отсюда следует, что на сфере $\|x\| = \frac{1}{\delta} \|A0 - f\|$ оператор B удовлетворяет условию $(Bx, x) \geq 0$. Поэтому в силу теоремы 10 уравнение $Bu = 0$ (а вместе с ним и уравнение (9)) имеет единственное решение в указанном шаре. Теорема 11 доказана.

Следствие 1. Если оператор A имеет в H производную Гато, являющуюся положительно определенным в H оператором, то условия теоремы 11 выполнены.

Действительно, так как в конечномерном пространстве линейный оператор ограничен, то производная Гато является непрерывным ограниченным и положительно определенным в H оператором. Из теоремы 2 следует, что A —сильно монотонный оператор. Кроме того, из ограниченности производной Гато вытекает, что оператор A удовлетворяет условию Липшица и поэтому непрерывен.

§ 2. Разностные схемы как операторные уравнения

1. Примеры пространств сеточных функций. В § 1 гл. I были введены основные понятия теории разностных схем: сетки, сеточные уравнения, сеточные функции, разностные производные и т. д. Теория формулирует общие принципы и правила построения разностных схем заданного качества. Характерной чертой этой теории является возможность сопоставить каждому дифференциальному уравнению целый класс разностных схем с требуемыми свойствами. При построении общей теории естественно освободиться от конкретной структуры и явного вида разностных уравнений. Это приводит к определению разностных схем как операторных уравнений с операторами, действующими в некотором функциональном пространстве, а именно, в пространстве сеточных функций.

Под *пространством сеточных функций* понимается множество функций, заданных на некоторой сетке. Так как каждой сеточной функции можно поставить в соответствие вектор, координатами которого являются значения сеточной функции в узлах сетки, то операции сложения функций и умножения функции на число определяются так же, как и для векторов.

Пространство сеточных функций линейно, и если сетка содержит конечное число узлов, то пространство конечномерно. Размерность его равна числу узлов сетки.

В пространстве сеточных функций можно ввести скалярное произведение функций, превратив это пространство в гильбертово. Различные пространства сеточных функций могут отличаться одно от другого выбором сетки и нормировкой. Приведем некоторые примеры.

Пример 1. Пусть на отрезке $0 \leqslant x \leqslant l$ введена равномерная сетка $\bar{\omega} = \{x_i = ih, 0 \leqslant i \leqslant N, hN = l\}$ с шагом h . Через ω , ω^+ и ω^- обозначим следующие части сетки $\bar{\omega}$:

$$\omega = \{x_i \in \bar{\omega}, 1 \leqslant i \leqslant N-1\},$$

$$\omega^+ = \{x_i \in \bar{\omega}, 1 \leqslant i \leqslant N\},$$

$$\omega^- = \{x_i \in \bar{\omega}, 0 \leqslant i \leqslant N-1\}.$$

На множестве H сеточных функций, заданных на $\bar{\omega}$ и принимающих вещественные значения, определим скалярное произведение и норму следующим образом:

$$(u, v) = (u, v)_{\bar{\omega}} = \sum_{i=1}^{N-1} u_i v_i h + 0,5h(u_0 v_0 + u_N v_N), \quad (1)$$

$$\|u\| = \sqrt{(u, u)}, \quad u_i = u(x_i), \quad v_i = v(x_i).$$

Если u_i и v_i рассматривать как значения на сетке $\bar{\omega}$ функций $u(x)$ и $v(x)$ непрерывного аргумента $x \in [0, l]$, то скалярное произведение (1) представляет собой квадратурную формулу трапеций для интеграла $\int_0^l u(x)v(x)dx$. Если сеточные функции заданы на ω , ω^+ или ω^- , то скалярное произведение вещественных сеточных функций определяется соответственно по формулам

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h, \quad u, v \in H(\omega),$$

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0,5h u_N v_N, \quad u, v \in H(\omega^+),$$

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0,5h u_0 v_0, \quad u, v \in H(\omega^-).$$

Легко проверить, что введенные скалярные произведения удовлетворяют всем аксиомам скалярного произведения, и поэтому построенные пространства являются гильбертовыми.

Пример 2. Пусть теперь на отрезке $0 \leq x \leq l$ введена произвольная неравномерная сетка

$$\bar{\omega} = \{x_i \in [0, l], \quad x_i = x_{i-1} + h_i, \quad 1 \leq i \leq N, \quad x_0 = 0, \quad x_N = l\}. \quad (2)$$

Напомним определение среднего шага \bar{h}_i в узле x_i :

$$\bar{h}_i = 0,5(h_i + h_{i+1}), \quad 1 \leq i \leq N-1, \quad \bar{h}_0 = 0,5h_1, \quad \bar{h}_N = 0,5h_N. \quad (3)$$

Отметим, что равномерная сетка есть частный случай неравномерной сетки (2) при $h_i \equiv h$. При этом имеем $\bar{h}_i = h$, $1 \leq i \leq N-1$, $\bar{h}_0 = \bar{h}_N = 0,5h$.

Обозначим, как и выше, через ω , ω^+ и ω^- соответствующие части сетки $\bar{\omega}$. По аналогии с примером 1 определим в вещественных пространствах сеточных функций, заданных на

указанных сетках, скалярное произведение по формулам:

$$(u, v) = \sum_{i=0}^N u_i v_i \hbar_i, \quad u, v \in H(\bar{\omega}), \quad (4)$$

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i \hbar_i, \quad u, v \in H(\omega), \quad (5)$$

$$(u, v) = \sum_{i=1}^N u_i v_i \hbar_i, \quad u, v \in H(\omega^+),$$

$$(u, v) = \sum_{i=0}^{N-1} u_i v_i \hbar_i, \quad u, v \in H(\omega^-).$$

Построенные пространства сеточных функций являются *гильбертовыми* и имеют *конечную размерность*, равную числу узлов соответствующей сетки.

Введенные скалярные произведения удобно записать в виде

$$(u, v) = \sum_{x_i \in \Omega} u(x_i) v(x_i) \hbar(x_i), \quad u, v \in H(\Omega),$$

где под Ω понимается либо $\bar{\omega}$, либо ω , ω^+ или ω^- . Помимо указанных скалярных произведений часто встречаются суммы вида

$$(u, v)_{\omega^+} = \sum_{i=1}^N u_i v_i h_i, \quad (u, v)_{\omega^-} = \sum_{i=0}^{N-1} u_i v_i h_{i+1}, \quad (6)$$

которые можно использовать в качестве скалярных произведений в пространствах $H(\omega^+)$ и $H(\omega^-)$. Видно, что для скалярного произведения (4) в пространстве $H(\bar{\omega})$ верно равенство

$$(u, v) = 0,5[(u, v)_{\omega^+} + (u, v)_{\omega^-}], \quad u, v \in H(\bar{\omega}).$$

Пример 3. Пусть в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ введена произвольная прямоугольная неравномерная сетка $\bar{\omega} = \bar{\omega}_1 \times \bar{\omega}_2$, где

$$\begin{aligned} \bar{\omega}_\alpha &= \{x_\alpha(i_\alpha) \in [0, l_\alpha], \quad x_\alpha(i_\alpha) = x_\alpha(i_\alpha - 1) + h_\alpha(i_\alpha), \quad 1 \leq i_\alpha \leq N_\alpha, \\ &\quad x_\alpha(0) = 0, \quad x_\alpha(N_\alpha) = l_\alpha\}, \quad \alpha = 1, 2. \end{aligned}$$

Пусть $\hbar_\alpha(i_\alpha)$, $0 \leq i_\alpha \leq N_\alpha$ — средний шаг в узле $x_\alpha(i_\alpha)$ по направлению x_α :

$$\begin{aligned} \hbar_\alpha(i_\alpha) &= 0,5[h_\alpha(i_\alpha) + h_\alpha(i_\alpha + 1)], \quad 1 \leq i_\alpha \leq N_\alpha - 1, \\ \hbar_\alpha(0) &= 0,5h_\alpha(1), \quad \hbar_\alpha(N_\alpha) = 0,5h_\alpha(N_\alpha), \quad \alpha = 1, 2. \end{aligned}$$

В пространстве $H(\Omega)$ сеточных функций, заданных на Ω , где Ω — любая часть сетки $\bar{\omega}$, скалярное произведение определим по формуле

$$(u, v) = \sum_{x_i \in \Omega} u(x_i) v(x_i) \hbar_1 \hbar_2, \quad x_i = (x_1(i_1), x_2(i_2)).$$

В частности, если сетка равномерна по каждому направлению, $h_\alpha = h_\alpha$, $\alpha = 1, 2$, и сеточные функции заданы на ω (во внутренних узлах сетки $\bar{\omega}$), то введенное скалярное произведение записывается в виде

$$(u, v) = \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} u(i_1, i_2) v(i_1, i_2) h_1 h_2, \quad u, v \in H(\omega).$$

Мы ограничимся здесь приведенными примерами, другие более сложные примеры будут рассмотрены в последующих главах при изучении конкретных разностных задач.

2. Некоторые разностные тождества. Переходим теперь к выводу основных формул, при помощи которых преобразуются выражения, содержащие сеточные функции. Мы приведем эти формулы для случая, когда сеточные функции заданы на неравномерной сетке, определенной в (2).

Напомним определение основных разностных производных сеточной функции:

$$\begin{aligned} y_{\bar{x}, i} &= \frac{y_i - y_{i-1}}{h_i}, \quad y_{x, i} = y_{\bar{x}, i+1} = \frac{y_{i+1} - y_i}{h_{i+1}}, \quad y_{\hat{x}, i} = \frac{y_i - y_{i-1}}{\hat{h}_i}, \\ y_{\hat{x}, i} &= \frac{y_{i+1} - y_i}{\hat{h}_i}, \quad y_{\bar{x}\bar{x}, i} = y_{\bar{x}\hat{x}, i} = \frac{1}{\hat{h}_i} (y_{x, i} - y_{\bar{x}, i}). \end{aligned}$$

В п. 2 § 1 гл. I были получены две формулы суммирования по частям:

$$\sum_{i=m+1}^{n-1} u_{\hat{x}, i} v_i \hat{h}_i = - \sum_{i=m+1}^n u_i v_{\bar{x}, i} h_i + u_n v_n - u_{m+1} v_m, \quad (7)$$

$$\sum_{i=m+1}^{n-1} u_{\bar{x}, i} v_i h_i = - \sum_{i=m}^{n-1} u_i v_{\hat{x}, i} \hat{h}_i + u_{n-1} v_n - u_m v_m. \quad (8)$$

Подставляя в эти формулы соотношения

$$h_i u_{\bar{x}, i} = \hat{h}_i u_{\hat{x}, i}, \quad \hat{h}_i u_{\hat{x}, i} = h_{i+1} u_{x, i},$$

после несложных преобразований получим формулы

$$\sum_{i=m+1}^{n-1} u_{\hat{x}, i} v_i \hat{h}_i = - \sum_{i=m}^{n-1} u_i v_{x, i} h_{i+1} + u_{n-1} v_n - u_m v_m, \quad (9)$$

$$\sum_{i=m+1}^{n-1} u_{x, i} v_i h_{i+1} = - \sum_{i=m+1}^n u_i v_{\hat{x}, i} \hat{h}_i + u_n v_n - u_{m+1} v_m, \quad (10)$$

$$\sum_{i=m+1}^n u_{\bar{x}, i} v_i h_i = - \sum_{i=m}^{n-1} u_i v_{x, i} h_{i+1} + u_n v_n - u_m v_m. \quad (11)$$

Подставим в формулы (7), (9), (11) $m = 0$ и $n = N$ и учтем определение (5) для скалярного произведения в $H(\omega)$, а также

обозначение (6). Получим тождества

$$(u_x, v) = -(u, v_x)_{\omega+} + u_N v_N - u_0 v_0, \quad (7')$$

$$(u_{\bar{x}}, v) = -(u, v_x)_{\omega-} + u_{N-1} v_N - u_0 v_0, \quad (9')$$

$$(u_{\bar{x}}, v)_{\omega+} = -(u, v_x)_{\omega-} + u_N v_N - u_0 v_0 \quad (11')$$

для сеточных функций u_i и v_i , заданных на сетке $\bar{\omega}$. Если в (7') положить $u_i = a_i y_{x,i}$ для $1 \leq i \leq N$, то получим первую разностную формулу Грина

$$((ay_x)_{\hat{x}}, v) = -(ay_{\bar{x}}, v_{\bar{x}})_{\omega+} + a_N y_{\bar{x}, N} v_N - a_1 y_{x,0} v_0. \quad (12)$$

Аналогично, полагая в (9) $u_i = a_i y_{x,i}$ для $0 \leq i \leq N-1$, получим

$$((ay_x)_{\hat{x}}, v) = -(ay_x, v_x)_{\omega-} + a_{N-1} y_{\bar{x}, N} v_N - a_0 y_{x,0} v_0.$$

Если из (12) вычесть равенство

$$((y, (av_x)_{\hat{x}}) = -(ay_{\bar{x}}, v_{\bar{x}})_{\omega+} + a_N y_{\bar{x}, N} v_N - a_1 y_{x,0} v_0,$$

то получим вторую разностную формулу Грина

$$((ay_{\bar{x}})_{\hat{x}}, v) - (y, (av_x)_{\hat{x}}) = a_N (y_{\bar{x}} v - v_{\bar{x}} y)_N - a_1 (y_x v - v_x y)_0. \quad (13)$$

Отметим, что для функций y_i и v_i , обращающихся в нуль при $i=0$ и $i=N$ ($y_0 = y_N = 0$, $v_0 = v_N = 0$), формула (12) имеет вид

$$((ay_{\bar{x}})_{\hat{x}}, v) = -(ay_{\bar{x}}, v_{\bar{x}})_{\omega+},$$

а вторая формула Грина (13) — вид

$$((ay_{\bar{x}})_{\hat{x}}, v) = (y, (av_x)_{\hat{x}}).$$

В общем случае произвольных сеточных функций, заданных на $\bar{\omega}$, формулы (12) и (13) можно записать в виде

$$(\Lambda y, v) = -(ay_{\bar{x}}, v_{\bar{x}})_{\omega+}, \quad (\Lambda y, v) - (y, \Lambda v) = 0, \quad (14)$$

где разностный оператор Λ , отображающий $H(\bar{\omega})$ на $H(\bar{\omega})$, определяется следующим образом:

$$\Lambda y_i = \begin{cases} \frac{1}{h_0} a_1 y_{x,0}, & i = 0, \\ (ay_{\bar{x}})_{\hat{x},i}, & 1 \leq i \leq N-1, \\ -\frac{1}{h_N} a_N y_{\bar{x},N}, & i = N. \end{cases}$$

Здесь скалярное произведение в $H(\bar{\omega})$ задано формулой (4). Отметим, что равенство (14) выражает самосопряженность оператора Λ в пространстве $H(\bar{\omega})$.

Мы рассмотрели случай, когда сеточные функции принимают на сетке вещественные значения. Если они принимают на $\bar{\omega}$ комплексные значения, то вводится комплексное гильбертово пространство $H(\bar{\omega})$ со скалярным произведением

$$(u, v) = \sum_{i=0}^N u_i \bar{v}_i \bar{h}_i, \quad u, v \in H(\bar{\omega}), \quad (15)$$

где \bar{v}_i — число, комплексно сопряженное v_i . Аналогично определяется скалярное произведение в $H(\omega)$

$$(u, v) = \sum_{i=1}^{N-1} u_i \bar{v}_i \bar{h}_i, \quad u, v \in H(\omega), \quad (16)$$

а также в $H(\omega^+)$ и $H(\omega^-)$. При этом формулы суммирования по частям (7'), (9'), (11') принимают вид

$$\begin{aligned} (u_{\hat{x}}, v) &= -(u, v_{\hat{x}})_{\omega^+} + u_N \bar{v}_N - u_0 \bar{v}_0, \\ (u_{\hat{x}}, v) &= -(u, v_x)_{\omega^-} + u_{N-1} \bar{v}_N - u_0 \bar{v}_0, \\ (u_{\hat{x}}, v)_{\omega^+} &= -(u, v_x)_{\omega^-} + u_N \bar{v}_N - u_0 \bar{v}_0, \end{aligned}$$

а разностные формулы Грина — вид:

$$\begin{aligned} ((ay_{\hat{x}})_{\hat{x}}, v) &= -(ay_{\hat{x}}, v_{\hat{x}})_{\omega^+} + a_N y_{\hat{x}, N} \bar{v}_N - a_1 y_{x, 0} \bar{v}_0, \\ ((ay_{\hat{x}})_{\hat{x}}, v) - (y, (av_{\hat{x}})_{\hat{x}}) &= \\ &= ((\bar{a} - a) y_{\hat{x}}, v_{\hat{x}})_{\omega^+} + (ay_{\hat{x}} \bar{v} - \bar{a} y_{\hat{x}} \bar{v})_N - (a_1 y_{x, 0} \bar{v}_0 - \bar{a}_1 y_0 \bar{v}_{x, 0}). \end{aligned}$$

Здесь использовано обозначение (16).

Используя введенный выше оператор Λ и обозначение (15) для скалярного произведения в $H(\bar{\omega})$, вторую разностную формулу Грина можно записать в виде

$$(\Lambda y, v) - (y, \Lambda v) = ((\bar{a} - a) y_{\hat{x}}, v_{\hat{x}})_{\omega^+}.$$

Отсюда следует, что в комплексном гильбертовом пространстве $H(\bar{\omega})$ оператор Λ самосопряжен, если все a_i вещественны.

Соотношения, аналогичные первой и второй разностным формулам Грина (12), (13), имеют место и для разностного оператора $(ay_{\hat{x}\hat{x}})_{\hat{x}\hat{x}}$. Приведем, например, аналог формулы (12)

$$\begin{aligned} \sum_{i=2}^{N-2} (ay_{\hat{x}\hat{x}})_{\hat{x}\hat{x}, i} v_i \bar{h}_i &= \sum_{i=1}^{N-1} a_i y_{\hat{x}\hat{x}, i} v_{\hat{x}\hat{x}, i} \bar{h}_i + \\ &\quad + [(ay_{\hat{x}\hat{x}})_x v - ay_{\hat{x}\hat{x}} v_x]_{N-1} - [(ay_{\hat{x}\hat{x}})_x v - ay_{\hat{x}\hat{x}} v_x]. \end{aligned}$$

3. Границы простейших разностных операторов. При изучении свойств разностных операторов нам понадобятся неравенства, дающие оценки для границ операторов и для постоянных энергетической эквивалентности двух операторов, действующих в пространстве сеточных функций \tilde{H} .

Рассмотрим сначала разностные операторы, заданные на множестве сеточных функций одного аргумента, определенные на равномерной сетке $\bar{\omega} = \{x_i = ih \in [0, l], 0 \leq i \leq N, hN = l\}$. Ниже будут использованы обозначения

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0,5h(u_0 v_0 + u_N v_N), \quad (u, v)_{\omega+} = \sum_{i=1}^N u_i v_i h.$$

Имеет место

Лемма 12. Для всякой функции $y_i = y(x_i)$, заданной на равномерной сетке $\bar{\omega}$ и обращающейся в нуль при $i=0$ и $i=N$, справедливы неравенства

$$\gamma_1(y, y) \leq (y_x^2, 1)_{\omega+} \leq \gamma_2(y, y), \quad (17)$$

где

$$\gamma_1 = \frac{4}{h^2} \sin^2 \frac{\pi}{2N} \geq \frac{8}{l^2}, \quad \gamma_2 = \frac{4}{h^2} \cos^2 \frac{\pi}{2N} < \frac{4}{h^2}.$$

Действительно, пусть $\mu_k(i)$ — ортонормированная собственная функция задачи

$$(\mu_k)_{xx} + \lambda_k \mu_k = 0, \quad 1 \leq i \leq N-1, \\ \mu_k(0) = \mu_k(N) = 0. \quad (18)$$

В п. 1 § 5 гл. I было отмечено, что сеточная функция y_i , удовлетворяющая условиям леммы, может быть представлена в виде суммы

$$y_i = \sum_{k=1}^{N-1} c_k \mu_k(i), \quad c_k = (y, \mu_k). \quad (19)$$

Из (18) и (19) найдем

$$y_{xx, i} = \sum_{k=1}^{N-1} c_k (\mu_k)_{xx, i} = - \sum_{k=1}^{N-1} \lambda_k c_k \mu_k(i), \quad 1 \leq i \leq N-1.$$

Используя ортогональность собственных функций μ_k , получим

$$(y, y) = \sum_{k=1}^{N-1} c_k^2, \quad -(y_{xx}, y) = \sum_{k=1}^{N-1} \lambda_k c_k^2. \quad (20)$$

В силу первой разностной формулы Грина (12) будем иметь

$$-(y_{xx}, y) = (y_x^2, 1)_{\omega+}. \quad (21)$$

Собственные значения λ_k задачи (18) были найдены в п. 1 § 5 гл. I:

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2l} = \frac{4}{h^2} \sin^2 \frac{k\pi}{2N}, \quad 1 \leq k \leq N-1,$$

причем

$$\gamma_1 = \min_k \lambda_k = \lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi}{2N},$$

$$\gamma_2 = \max_k \lambda_k = \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi}{2N}.$$

Отсюда и из (20), (21) следуют оценки (17) леммы 12.

Замечание 1. Оценки (17) точны в том смысле, что они переходят в равенства, если в качестве y_i взять $\mu_1(i)$ и $\mu_{N-1}(i)$. Отметим, что $\gamma_1 = 8/l^2$, если $h = l/2$, т. е. при $N = 2$. При $N = 4$ имеем $\gamma_1 = 32/(l^2(2 + \sqrt{2})) > 8/l^2$.

Замечание 2. Если y_i обращается в нуль лишь при $i = 0$ или $i = N$, то в (17) имеем

$$\gamma_1 = \frac{4}{h^2} \sin^2 \frac{\pi}{4N} \geq \frac{8}{l^2(2 + \sqrt{2})}, \quad \gamma_2 = \frac{4}{h^2} \cos^2 \frac{\pi}{4N} < \frac{4}{h^2}.$$

Если же y_i — произвольная на $\bar{\omega}$ сеточная функция, то в (17) имеем $\gamma_1 = 0$ и $\gamma_2 = 4/h^2$. Для доказательства этих утверждений следует рассмотреть вместо задачи (18) соответствующую задачу на собственные значения, изученную в § 5 гл. I.

Неравенства (17) можно записать в виде

$$\gamma_1(y, y) \leq (-\Lambda y, y) \leq \gamma_2(y, y), \quad (22)$$

если ввести разностный оператор Λ по формуле $\Lambda y_i = y_{\bar{x}x, i}$, $1 \leq i \leq N-1$, на функциях y_i , удовлетворяющих условиям $y_0 = y_N = 0$. Если сеточная функция y_i обращается в нуль лишь на одном конце сетки $\bar{\omega}$, то оператор Λ следует определить по формулам

$$\Lambda y_i = \begin{cases} y_{\bar{x}x, i}, & 1 \leq i \leq N-1, \\ -\frac{2}{h} y_{\bar{x}x, i}, & i = N, \text{ если } y_0 = 0, \end{cases} \quad (23)$$

или

$$\Lambda y_i = \begin{cases} \frac{2}{h} y_{\bar{x}x, i}, & i = 0, \\ y_{\bar{x}x, i}, & 1 \leq i \leq N-1, \text{ если } y_N = 0. \end{cases}$$

Учитывая, что в каждом из этих случаев из первой разностной формулы Грина следуют равенства $(\Lambda y, y) = (y_x^2, 1)_{\omega+}$, получим неравенства (22), где γ_1 и γ_2 указаны в замечании 2, а y_i обращается в нуль на соответствующем конце сетки $\bar{\omega}$.

Если y_i — произвольная сеточная функция, то оператор Λ следует определить так:

$$\Lambda y_i = \begin{cases} \frac{2}{h} y_{x,0}, & i = 0, \\ y_{xx,i}, & 1 \leq i \leq N-1, \\ -\frac{2}{h} y_{x,N}, & i = N. \end{cases}$$

В этом случае также верны неравенства (22) и

$$(-\Lambda y, y) = -(y_{xx}, y) + y_{x,N} y_N - y_{x,0} y_0 = (y_x^2, 1)_{\omega+}.$$

Постоянные γ_1 и γ_2 указаны в замечании 2.

Итак, мы нашли границы для простейших разностных операторов. Покажем теперь, что для всех введенных в этом пункте операторов Λ справедливо неравенство

$$|(-\Lambda u, v)| \leq (-\Lambda u, u)^{1/2} (-\Lambda v, v)^{1/2}. \quad (24)$$

Идею получения неравенства (24) проиллюстрируем на примере оператора $\Lambda y = y_{xx}$. Введем пространство $H(\omega)$ сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{i=1}^{N-1} u_i v_i h$, $u, v \in H(\omega)$. Разностному оператору Λ в пространстве $H(\omega)$ соответствует линейный оператор A , определяемый равенством

$$Ay_i = -\Lambda \dot{y}_i, \quad 1 \leq i \leq N-1,$$

где $y \in H(\omega)$, $y_i = \dot{y}_i$ для $1 \leq i \leq N-1$ и $\dot{y}_0 = \dot{y}_N = 0$. Оператор A отображает $H(\omega)$ на $H(\omega)$.

В силу равенства $(u, v) = (\dot{u}, \dot{v})$, имеем $(Au, v) = -(\Lambda \dot{u}, \dot{v})$, где $\dot{u}_0 = \dot{u}_N = 0$, $\dot{v}_0 = \dot{v}_N = 0$. Из (22) следует, что $(Au, u) \geq \gamma_1(u, u)$, $\gamma_1 > 0$. Таким образом, оператор A положительно определен в $H(\omega)$.

Докажем, что он самосопряжен в $H(\omega)$. Действительно, из второй разностной формулы Грина (13) будем иметь

$$(Au, v) = -(\Lambda \dot{u}, \dot{v}) = -(\dot{u}_{xx}, \dot{v}) = -(\dot{u}, \dot{v}_{xx}) = (u, Av).$$

Так как для неотрицательного самосопряженного оператора справедливо обобщенное неравенство Коши—Буняковского $|(Au, v)| \leq (Au, u)^{1/2} (Av, v)^{1/2}$, то отсюда получим

$$|(-\Lambda \dot{u}, \dot{v})| \leq (-\Lambda \dot{u}, \dot{u})^{1/2} (-\Lambda \dot{v}, \dot{v})^{1/2},$$

что и требовалось доказать.

4. Оценки снизу для некоторых разностных операторов. В лемме 12 фактически найдены постоянные энергетической эквивалентности единичного оператора E и оператора A , кото-

рый соответствует разностному оператору $-\Lambda y = -y_{xx}$ на функциях, обращающихся в нуль на концах сетки $\bar{\omega}$, т. е. γ_1 и γ_2 из неравенств $\gamma_1 E \leq A \leq \gamma_2 E$.

Получим теперь неравенство, связывающее операторы A и D , где $Dy_i = \rho_i y_i$, $1 \leq i \leq N-1$ и $\rho_i \geq 0$. Для этого нам необходимо определить разностную функцию Грина оператора Λ .

Пусть на сетке $\bar{\omega}$, введенной выше, требуется найти решение разностной задачи

$$\begin{aligned} \Lambda v_i &= v_{xx, i} = -f_i, \quad 1 \leq i \leq N-1, \\ v_0 &= v_N = 0. \end{aligned} \tag{25}$$

Сеточную функцию G_{ik} , которая при фиксированном $k = 1, 2, \dots, N-1$ удовлетворяет условиям

$$\begin{aligned} \Lambda G_{ik} &= G_{xx, ik} = -\frac{1}{h} \delta_{ik}, \quad 1 \leq i \leq N-1, \\ G_{0k} &= G_{Nk} = 0, \end{aligned}$$

где δ_{ik} — символ Кронекера:

$$\delta_{ik} = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases}$$

назовем функцией Грина разностного оператора Λ .

Приведем основные свойства функции Грина:

1) функция Грина симметрична, $G_{ik} = G_{ki}$ и, кроме того, G_{ik} как функция k при фиксированном $i = 1, 2, \dots, N-1$ удовлетворяет условиям

$$\begin{aligned} \Lambda G_{ik} &= G_{xx, ik} = -\frac{1}{h} \delta_{ik}, \quad 1 \leq k \leq N-1, \\ G_{i0} &= G_{iN} = 0. \end{aligned}$$

2) функция Грина положительна, $G_{ik} > 0$ при $i, k \neq 0, N$.

3) для любой сеточной функции y_i , удовлетворяющей условиям $y_0 = y_N = 0$, верно представление

$$y_i = - \sum_{k=1}^{N-1} G_{ik} \Lambda y_k h, \tag{26}$$

так что решение задачи (25) представимо в виде

$$v_i = \sum_{k=1}^{N-1} G_{ik} f_k h, \quad 0 \leq i \leq N.$$

Это утверждение доказывается при помощи второй разностной формулы Грина (13) и свойства 1).

Лемма 13. Пусть $\rho_i \geq 0$ — сеточная функция, заданная на ω и не равная тождественно нулю. Для всякой сеточной функци-

ции y_i , заданной на $\bar{\omega}$ и удовлетворяющей условиям $y_0 = y_N = 0$, верна оценка

$$\gamma_1(\rho y, y) \leq (y_x^2, 1)_{\omega+}, \quad (27)$$

где $1/\gamma_1 = \max_{1 \leq i \leq N-1} v_i$, а v_i есть решение краевой задачи

$$\begin{aligned} \Lambda v_i &= v_{xx,i} = -\rho_i, \quad 1 \leq i \leq N-1, \\ v_0 &= v_N = 0. \end{aligned} \quad (28)$$

Действительно, пусть $y_0 = y_N = 0$. Используя (26), получим

$$\begin{aligned} (\rho y, y) &= \sum_{i=1}^{N-1} \rho_i y_i^2 h = - \sum_{i=1}^{N-1} \rho_i y_i h \left(\sum_{k=1}^{N-1} G_{ik} \Lambda y_k h \right) = \\ &= - \sum_{k=1}^{N-1} h \Lambda y_k \left(\sum_{i=1}^{N-1} \rho_i y_i G_{ik} h \right) = - (\Lambda y, w), \end{aligned}$$

где обозначено $w_k = \sum_{i=1}^{N-1} \rho_i y_i G_{ik} h$, $0 \leq k \leq N$. Применяя неравенство (24), отсюда найдем

$$(\rho y, y) \leq (-\Lambda y, y)^{1/2} (-\Lambda w, w)^{1/2}$$

или в силу (21)

$$(\rho y, y)^2 \leq (y_x^2, 1)_{\omega+} (-\Lambda w, w). \quad (29)$$

Воспользуемся свойством 1) функции Грина G_{ik} . Получим

$$-\Lambda w_k = - \sum_{i=1}^{N-1} h \rho_i y_i \Lambda G_{ik} = \sum_{i=1}^{N-1} \rho_i y_i \delta_{ik} = \rho_k y_k$$

и, следовательно,

$$(-\Lambda w, w) = \sum_{k=1}^{N-1} h \rho_k y_k \left(\sum_{i=1}^{N-1} h \rho_i y_i G_{ik} \right) = \sum_{i=1}^{N-1} \sum_{k=1}^{N-1} a_{ik} y_i y_k,$$

где обозначено $a_{ik} = h^2 \rho_i \rho_k G_{ik}$, $1 \leq i, k \leq N-1$. Используя неравенство $2y_i y_k \leq y_i^2 + y_k^2$, а также симметрию и положительность функции Грина G_{ik} , отсюда найдем

$$\begin{aligned} (-\Lambda w, w) &\leq \sum_{i=1}^{N-1} 0,5 y_i^2 \sum_{k=1}^{N-1} a_{ik} + \sum_{k=1}^{N-1} 0,5 y_k^2 \sum_{i=1}^{N-1} a_{ki} = \\ &= \sum_{i=1}^{N-1} y_i^2 \sum_{k=1}^{N-1} a_{ik} = \sum_{i=1}^{N-1} \rho_i y_i^2 h \left(\sum_{k=1}^{N-1} \rho_k G_{ik} h \right). \end{aligned}$$

В силу свойства 3) решение задачи (28) записывается в виде

$$v_i = \sum_{k=1}^{N-1} \rho_k G_{ik} h > 0, \quad 1 \leq i \leq N-1.$$

Следовательно,

$$(-\Lambda w, w) = \sum_{i=1}^{N-1} \rho_i y_i^2 v_i h \leq \max_{1 \leq i \leq N-1} v_i (\rho y, y) = \frac{1}{\gamma_1} (\rho y, y).$$

Отсюда и из (29) следует оценка (27) леммы.

Замечание 1. Можно показать, что функция $v_i = 0,5x_i(l-x_i)$, где $x_i = ih \in [0, l]$, есть решение задачи (28) при $\rho_i = 1$. Отсюда следует оценка

$$\gamma_1(y, y) \leq (y_x^2, 1)_{\omega^+}, \quad \gamma_1 = 8/l^2, \quad y_0 = y_N = 0. \quad (30)$$

Замечание 2. Лемма 13 обобщается на случай, когда y_i обращается в нуль лишь на одном конце сетки $\bar{\omega}$. Например, если $y_0 = 0$, то в (27) имеем $1/\gamma_1 = \max_{1 \leq i \leq N} v_i$, где v_i — решение задачи $\Lambda v_i = -\rho_i$, $1 \leq i \leq N$, $v_0 = 0$ с разностным оператором Λ , определенным в (23).

Лемма 14. Пусть $\rho_i \geq 0$, $d_i \geq 0$ заданы на ω , а функция $a_i \geq c_1 > 0$ задана на ω^+ . Для всякой функции y_i , заданной на ω и удовлетворяющей условиям $y_0 = y_N = 0$, верна оценка

$$\gamma_1(\rho y, y) \leq (ay_x^2, 1)_{\omega^+} + (dy, y), \quad 1/\gamma_1 = \max_{1 \leq i \leq N-1} v_i,$$

где v_i — решение краевой задачи

$$\Lambda v_i = (av_x)_x|_i - d_i v_i = -\rho_i, \quad 1 \leq i \leq N-1, \quad v_0 = v_N = 0.$$

Замечание 1. Если y_i обращается в нуль лишь на одном конце сетки $\bar{\omega}$, например $y_N = 0$, то верна оценка

$$\gamma_1(\rho y, y) \leq (ay_x^2, 1)_{\omega^+} + (dy, y) + \kappa_0 y_0^2, \quad (31)$$

где $1/\gamma_1 = \max_{0 \leq i \leq N-1} v_i$, а функция v_i есть решение задачи

$$\begin{aligned} \Lambda v_i &= -\rho_i, \quad 0 \leq i \leq N-1, \quad v_N = 0, \\ \Lambda y_i &= \begin{cases} \frac{2}{h}(a_1 y_{x,0} - \kappa_0 y_0) - d_0 y_0, & i = 0, \\ (ay_x)_x|_i - d_i y_i, & 1 \leq i \leq N-1, \quad \kappa_0 \geq 0. \end{cases} \end{aligned} \quad (32)$$

Замечание 2. Для произвольной сеточной функции y_i , заданной на $\bar{\omega}$, можно получить оценку

$$\gamma_1(\rho y, y) \leq (ay_x^2, 1)_{\omega^+} + (dy, y) + \kappa_0 y_0^2 + \kappa_1 y_N^2, \quad (33)$$

где $\kappa_0 \geq 0$, $\kappa_1 \geq 0$, $\kappa_0 + \kappa_1 + (d, 1) > 0$, а сеточные функции $\rho_i \geq 0$, $d_i \geq 0$ заданы на ω . Здесь $1/\gamma_1 = \max_{0 \leq i \leq N} v_i$, где v_i — решение

краевой задачи

$$\Lambda y_i = \begin{cases} \Lambda v_i = -\rho_i, & 0 \leq i \leq N, \\ \frac{2}{h} (a_1 y_{x,0} - \kappa_0 y_0) - d_0 y_0, & i = 0, \\ (ay_x)_x, i - d_i y_i, & 1 \leq i \leq N-1, \\ -\frac{2}{h} (a_N y_{x,N} + \kappa_1 y_N) - d_N y_N, & i = N. \end{cases} \quad (34)$$

Доказательство леммы 14 и замечаний 1 и 2 проводится так же, как и леммы 13. Здесь используется функция Грина указанных разностных операторов Λ , которая удовлетворяет перечисленным выше свойствам 1)–4).

Лемма 15. Для сеточной функции y_i , обращающейся в нуль при $i=N$, верна оценка

$$y_0^2 \leq \text{th}(\varepsilon l) \left[\varepsilon(y, y) + \frac{1}{\varepsilon} (y_x^2, 1)_{\omega+} \right], \quad \varepsilon \geq 0. \quad (35)$$

Аналогичная оценка

$$y_N^2 \leq \text{th}(\varepsilon l) \left[\varepsilon(y, y) + \frac{1}{\varepsilon} (y_x^2, 1)_{\omega+} \right], \quad \varepsilon \geq 0,$$

верна для случая, когда $y_0=0$. Для произвольной сеточной функции y_i , заданной на сетке $\bar{\omega}$, имеет место оценка

$$y_0^2 + y_N^2 \leq \frac{8 + \varepsilon^2 l^2}{\varepsilon l \sqrt{16 + \varepsilon^2 l^2}} \left[\varepsilon(y, y) + \frac{1}{\varepsilon} (y_x^2, 1)_{\omega+} \right], \quad \varepsilon > 0. \quad (36)$$

Сначала докажем справедливость оценки (35). Для этого воспользуемся замечанием 1 к лемме 14. Положим в (32) $a_i \equiv 1/\varepsilon$, $d_i \equiv \varepsilon$, $\kappa_0 = 0$ и $\rho_0 = 2/h$, $\rho_i = 0$, $1 \leq i \leq N-1$. Тогда из (31) получим оценку

$$y_0^2 \leq \max_{0 \leq i \leq N-1} v_i \left[\varepsilon(y, y) + \frac{1}{\varepsilon} (y_x^2, 1)_{\omega+} \right],$$

где v_i — решение следующей вспомогательной задачи:

$$\begin{aligned} \Lambda v_i &= \frac{1}{\varepsilon} v_{xx, i} - \varepsilon v_i = 0, \quad 1 \leq i \leq N-1, \\ \Lambda v_0 &= \frac{2}{\varepsilon h} v_{x,0} - \varepsilon v_0 = -\frac{2}{h}, \quad v_N = 0. \end{aligned} \quad (37)$$

Запишем (37) по точкам

$$\begin{aligned} v_{i-1} - 2\alpha v_i + v_{i+1} &= 0, \quad 1 \leq i \leq N-1, \\ v_1 - \alpha v_0 &= -eh, \quad v_N = 0, \end{aligned} \quad (38)$$

где $\alpha = 1 + 0,5\varepsilon^2 h^2 \geq 1$.

Мы получили краевую задачу для разностного уравнения второго порядка с постоянными коэффициентами.

Используя общую теорию, развитую в п. 1 § 4 гл. I, а также свойства полиномов Чебышева (см. п. 2 там же), найдем, что функция

$$v_i = \frac{\varepsilon h U_{N-i-1}(\alpha)}{T_N(\alpha)}, \quad 0 \leq i \leq N,$$

является решением задачи (38). Здесь

$$T_n(\alpha) = \operatorname{ch}(n \operatorname{Arch} \alpha), \quad U_n(\alpha) = \frac{\operatorname{sh}((n+1) \operatorname{Arch} \alpha)}{\operatorname{sh}(\operatorname{Arch} \alpha)}, \quad |\alpha| \geq 1,$$

— полиномы Чебышева степени n первого и второго рода.

Так как $\alpha \geq 1$, то

$$\max_{0 \leq i \leq N-1} v_i = v_0 = \frac{\varepsilon h U_{N-1}(\alpha)}{T_N(\alpha)}.$$

Итак, получена оценка

$$y_0^2 \leq v_0 \left[\varepsilon(y, y) + \frac{1}{\varepsilon} \left(\frac{y^2}{x}, 1 \right)_{\omega^+} \right]$$

для сеточной функции y_i , удовлетворяющей условию $y_N = 0$. Эта оценка точна в том смысле, что она переходит в равенство, если в качестве y_i взять функцию v_i .

Оценим теперь v_0 сверху для любого h . Если обозначить $\operatorname{ch} 2z = \alpha$, то $z \geq 0$ и

$$\begin{aligned} \varepsilon h &= 2 \operatorname{sh} z, \quad N = l/h = \varepsilon l/(2 \operatorname{sh} z), \\ T_N(\alpha) &= \operatorname{ch} 2Nz = \operatorname{ch} w(z), \\ U_{N-1}(\alpha) &= \frac{\operatorname{sh} 2Nz}{\operatorname{sh} 2z} = \frac{\operatorname{sh} w(z)}{2 \operatorname{sh} z \operatorname{ch} z}, \quad w(z) = \frac{\varepsilon l z}{\operatorname{sh} z}. \end{aligned} \tag{39}$$

Поэтому

$$v_0 = \frac{\operatorname{sh} w(z)}{\operatorname{ch} z \operatorname{ch} w(z)}.$$

Так как при фиксированном ε

$$\frac{dw}{dz} = \frac{\varepsilon l (\operatorname{sh} z - z \operatorname{ch} z)}{\operatorname{sh}^2 z} \leq 0,$$

то

$$\frac{dv_0}{dz} = \frac{\operatorname{ch} z \frac{dw}{dz} - \operatorname{sh} z \operatorname{sh} w \operatorname{ch} w}{\operatorname{ch}^2 z \operatorname{ch}^2 w} \leq 0.$$

Следовательно, v_0 максимально при $z = 0$. Это дает оценку $v_0 \leq \operatorname{th}(\varepsilon l)$. Неравенство (35) доказано.

Пусть теперь y_i — произвольная сеточная функция. Из замечания 2 к лемме 14 при $\alpha_i = 1/\varepsilon$, $d_i = \varepsilon$, $x_0 = x_1 = 0$, $\rho_0 = \rho_N = 2/h$, $\rho_i = 0$ для $1 \leq i \leq N-1$ получим оценку

$$y_0^2 + y_N^2 \leq \max_{0 \leq i \leq N} v_i \left[\varepsilon(y, y) + \frac{1}{\varepsilon} \left(\frac{y^2}{x}, 1 \right)_{\omega^+} \right],$$

где v_i — решение краевой задачи

$$\begin{aligned} \frac{1}{\varepsilon} v_{xx, i} - \varepsilon v_i &= 0, \quad 1 \leq i \leq N-1, \\ \frac{2}{\varepsilon h} v_{x, 0} - \varepsilon v_0 &= -\frac{2}{h}, \quad -\frac{2}{\varepsilon h} v_{x, N} - \varepsilon v_N = -\frac{2}{h}. \end{aligned} \tag{40}$$

Решением задачи (40) является функция

$$v_i = \frac{\varepsilon h [T_{N-i}(\alpha) + T_i(\alpha)]}{(\alpha^2 - 1) U_{N-1}(\alpha)}, \quad 0 \leq i \leq N,$$

где α определено выше.

Отсюда находим, что

$$\max_{0 \leq i \leq N} v_i = v_0 = v_N = \frac{\varepsilon h (1 + T_N(\alpha))}{(\alpha^2 - 1) U_{N-1}(\alpha)}. \quad (41)$$

Оценим это выражение сверху для любого h . Используя (39), получим

$$v_0 = \frac{1 + \operatorname{ch} w(z)}{\operatorname{ch} z \operatorname{sh} w(z)} = \frac{\operatorname{ch} \frac{1}{2} w(z)}{\operatorname{ch} z \operatorname{sh} \frac{1}{2} w(z)} \leq \frac{\operatorname{ch} \frac{1}{2} w(z)}{\operatorname{sh} \frac{1}{2} w(z)} = \varphi(z).$$

Так как

$$\frac{d\varphi}{dz} = -\frac{1}{\operatorname{sh}^2 0,5w} \frac{\partial w}{\partial z} > 0,$$

то функция $\varphi(z)$ максимальна при максимальном $z = z_0$, которое находится из соотношения $\operatorname{ch} 2z_0 = 1 + \varepsilon^2 l^2 / 8$ ($h \leq l/2$). Из (39) получим, что $w(z_0) = 4z_0$.

Следовательно,

$$\varphi(z_0) = \frac{\operatorname{ch} 2z_0}{\operatorname{sh} 2z_0} = \frac{1 + \varepsilon^2 l^2 / 8}{\sqrt{\varepsilon^2 l^2 / 8 + \varepsilon^4 l^4 / 64}} = \frac{8 + \varepsilon^2 l^2}{\varepsilon l \sqrt{16 + \varepsilon^2 l^2}}.$$

Оценка (36) получена.

Леммы 13 и 14 без затруднения обобщаются на случай произвольной неравномерной сетки $\bar{\omega}$. В этом случае для скалярных произведений используются обозначения (4), (6), а разностные операторы Λ заменяются соответствующими операторами на неравномерной сетке.

Лемма 16. Пусть $\rho_i \geq 0$, $d_i \geq 0$ заданы на произвольной неравномерной сетке $\bar{\omega}$, $\rho_i \neq 0$ и $a_i \geq c_1 > 0$ задано на ω^+ . Пусть $x_0 \geq 0$, $x_1 \geq 0$ — произвольные числа и выполнено условие $x_0 + x_1 + (d, 1) > 0$. Для любой сеточной функции y_i , заданной на $\bar{\omega}$, справедливо неравенство (33), где $1/\gamma_1 = \max_{0 \leq i \leq N} v_i$, а v_i — решение задачи $\Lambda v_i = -\rho_i$, $0 \leq i \leq N$. Здесь оператор Λ определяется формулами

$$\Lambda y_i = \begin{cases} \frac{1}{\bar{h}_0} (a_1 y_{x_0} - x_0 y_0) - d_0 y_0, & i = 0, \\ (ay_{\bar{x}})_{\bar{x}, i} - d_i y_i, & 1 \leq i \leq N-1, \\ -\frac{1}{\bar{h}_N} (a_N y_{\bar{x}, N} + x_1 y_N) - d_N y_N, & i = N. \end{cases} \quad (42)$$

Лемма 16 доказывается так же, как и предыдущие леммы.

Замечание 1. Если $a_i \equiv 1$, $d_i \equiv 0$, $\rho_i \equiv 1$, то неравенство (33) принимает вид

$$\gamma_1(y, y) \leq \left(y_{\bar{x}}^2, 1 \right)_{\omega^+} + x_0 y_0^2 + x_1 y_N^2, \quad (43)$$

где

$$\gamma_1 = \frac{8 (x_0 + x_1 + lx_0 x_1)^2}{l (2 + lx_0) (2 + lx_1) (2x_0 + 2x_1 + lx_0 x_1)}.$$

Если, кроме того, $y_0 = y_N = 0$, то неравенство (43) переходит в неравенство (30). Если y_i обращается в нуль лишь на одном конце, например при $i = N$, то, полагая в (43) $y_N = 0$ и переходя к пределу при $\kappa_1 \rightarrow \infty$, получим оценку

$$v_1(y, y) \leq \left(\frac{y^2}{x}, 1 \right)_{\omega+} + \kappa_0 y_0^2, \quad v_1 = \frac{8(1+\kappa_0)^2}{l^2(2+\kappa_0)^2}.$$

Замечание 2. Из определения (42) разностного оператора Λ и первой разностной формулы Грина следует, что

$$(-\Lambda y, y) = \left(ay_x^2, 1 \right)_{\omega+} + (dy, y) + \kappa_0 y_0^2 + \kappa_1 y_N^2.$$

Поэтому неравенство (33) леммы 16 может быть записано в виде

$$v_1(p y, y) \leq -(\Lambda y, y).$$

Перейдем к выводу оценки (43). Найдем решение задачи $\Lambda v_i = -p_i$, $0 \leq i \leq N$, при указанных в замечании 1 предположениях. Имеем разностную краевую задачу

$$v_{xx, i} = -1, \quad 1 \leq i \leq N-1, \quad (44)$$

$$v_{x, 0} = \kappa_0 v_0 - \hbar_0, \quad i = 0, \quad (45)$$

$$-v_{x, N} = \kappa_1 v_N - \hbar_N, \quad i = N. \quad (46)$$

Умножим уравнение (44) на \hbar_i , просуммируем по i от j до $N-1$ и учтем краевое условие (46). Получим

$$\begin{aligned} \sum_{i=j}^{N-1} v_{xx, i} \hbar_i &= \sum_{i=j}^{N-1} (v_{x, i+1} - v_{x, i}) = v_{x, N} - v_{x, j} = \\ &= -\kappa_1 v_N + \hbar_N - v_{x, j} = -\sum_{i=j}^{N-1} \hbar_i = x_j - 0.5h_j - l + \hbar_N. \end{aligned}$$

Отсюда следует, что

$$v_{x, j} = l - \kappa_1 v_N + 0.5h_j - x_j, \quad 1 \leq j \leq N. \quad (47)$$

Полагая в (47) $j = 1$ и учитывая равенства $\hbar_0 = 0.5h_1$, $v_{x, 1} = v_{x, 0} = \kappa_0 v_0 - \hbar_0$, получим соотношение между v_0 и v_N

$$\kappa_0 v_0 + \kappa_1 v_N = l. \quad (48)$$

Умножая (47) на h_j и суммируя по j от 1 до i , найдем

$$\sum_{j=1}^i v_{x, j} h_j = v_i - v_0 = (l - \kappa_1 v_N) \sum_{j=1}^i h_j - \sum_{j=1}^i (x_j - 0.5h_j) h_j.$$

Так как $h_j = x_j - x_{j-1}$, $x_j - 0.5h_j = 0.5(x_j + x_{j-1})$, то

$$\sum_{j=1}^i h_j = x_i, \quad \sum_{j=1}^i (x_j - 0.5h_j) h_j = 0.5 \sum_{j=1}^i (x_j^2 - x_{j-1}^2) = 0.5x_i^2.$$

Таким образом, имеем

$$\begin{aligned} v_i &= v_0 + x_i(l - \kappa_1 v_N) - 0.5x_i^2 = \\ &= v_0 + 0.5(l - \kappa_1 v_N)^2 - 0.5(x_i - l + \kappa_1 v_N)^2, \quad 0 \leq i \leq N. \end{aligned} \quad (49)$$

Полагая здесь $i = N$, найдем второе соотношение для v_0 и v_N

$$v_N = v_0 + l(l - \kappa_1 v_N) - 0,5l^2. \quad (50)$$

Из (48), (50) получим

$$v_0 = \frac{l(2 + l\kappa_1)}{2(\kappa_0 + \kappa_1 + \kappa_0\kappa_1 l)}, \quad v_N = \frac{l(2 + l\kappa_0)}{2(\kappa_0 + \kappa_1 + \kappa_0\kappa_1 l)}. \quad (51)$$

Так как $0 \leq l - \kappa_1 v_N < l$, то из (49), (51) найдем, что

$$\max_{0 \leq i \leq N} v_i \leq v_0 + 0,5(l - \kappa_1 v_N)^2 = \frac{l(2 + l\kappa_0)(2 + l\kappa_1)(2\kappa_0 + 2\kappa_1 + l\kappa_0\kappa_1)}{8(\kappa_0 + \kappa_1 + l\kappa_0\kappa_1)^2}.$$

Отсюда и из леммы 16 следует оценка (43). Если $y_0 = y_N = 0$, то, полагая в (33) $a_i = 1$, $d_i = 0$, $\rho_i = 1$ и переходя в (43) к пределу при $\kappa_0 \rightarrow \infty$ и $\kappa_1 \rightarrow \infty$, получим оценку (30) с $\gamma_1 = 8/l^2$.

5. Оценки сверху для разностных операторов. Получим теперь оценки сверху для некоторых разностных операторов.

Лемма 17. Для произвольной сеточной функции y_i , заданной на неравномерной сетке $\bar{\omega}$, справедлива оценка

$$(ay_x^2, 1)_{\omega+} \leq \gamma_2(y, y), \quad (52)$$

где

$$\gamma_2 = \max \left[\frac{4a_1}{h_1^2}, \frac{4a_N}{h_N^2}, \max_{1 \leq i \leq N-1} \frac{2}{h_i} \left(\frac{a_i}{h_i} + \frac{a_{i+1}}{h_{i+1}} \right) \right].$$

Если сетка равномерна, то

$$\gamma_2 = \frac{4}{h^2} \max \left[a_1, a_N, \max_{1 \leq i \leq N-1} \left(\frac{a_i + a_{i+1}}{2} \right) \right].$$

Если $y_0 = y_N = 0$, то $\gamma_2 = \max_{1 \leq i \leq N-1} \frac{2}{h_i} \left(\frac{a_i}{h_i} + \frac{a_{i+1}}{h_{i+1}} \right)$.

Действительно, имеем

$$\begin{aligned} (ay_x^2, 1)_{\omega+} &= \sum_{i=1}^N \frac{a_i (y_i - y_{i-1})^2}{h_i} = \\ &= \sum_{i=1}^N \frac{a_i}{h_i} y_i^2 + \sum_{i=0}^{N-1} \frac{a_{i+1}}{h_{i+1}} y_i^2 - 2 \sum_{i=1}^N \frac{a_i}{h_i} y_i y_{i-1}. \end{aligned}$$

Используя неравенство $2y_i y_{i-1} \leq y_i^2 + y_{i-1}^2$, получим при $a_i > 0$, что

$$\begin{aligned} (ay_x^2, 1)_{\omega+} &\leq \sum_{i=1}^N \frac{2a_i}{h_i} y_i^2 + \sum_{i=0}^{N-1} \frac{2a_{i+1}}{h_{i+1}} y_i^2 = \\ &= \frac{2a_1}{h_1 h_0} y_0^2 + \frac{2a_N}{h_N h_N} y_N^2 + \sum_{i=1}^{N-1} \frac{2}{h_i} \left(\frac{a_i}{h_i} + \frac{a_{i+1}}{h_{i+1}} \right) y_i^2 h_i. \end{aligned}$$

Так как $\tilde{h}_0 = 0,5h_1$, $\tilde{h}_N = 0,5h_N$ и $(y, y) = \sum_{i=0}^N i^2 \tilde{h}_i y_i$, то отсюда следует оценка (52) с указанным значением для γ_2 . Лемма 17 доказана.

Лемма 18. Пусть $a_i > 0$, $b_i \geq 0$, а σ_0 и σ_1 — неотрицательны, причем $(b, 1) + \sigma_0 + \sigma_1 \neq 0$. Для произвольной сеточной функции y_i , заданной на неравномерной сетке $\bar{\omega}$, справедлива оценка

$$(ay_x^2, 1)_{\omega+} + (by, y) + \sigma_0 y_0^2 + \sigma_1 y_N^2 \leq \bar{\gamma}_2 (y, y), \quad (53)$$

где $\bar{\gamma}_2 = \gamma_2 + (1 + \gamma_2) \max_{0 \leq i \leq N} v_i$, γ_2 определено в лемме (17), а v — решение краевой задачи

$$\begin{aligned} (av_x)_{\hat{x}, i} - v_i &= -b_i, \quad 1 \leq i \leq N-1, \\ \frac{a_1}{\tilde{h}_0} v_{x, 0} - v_0 &= -b_0 - \frac{\sigma_0}{\tilde{h}_0}, \quad i = 0, \\ -\frac{a_N}{\tilde{h}_N} v_{x, N} - v_N &= -b_N - \frac{\sigma_1}{\tilde{h}_N}, \quad i = N. \end{aligned} \quad (54)$$

Действительно, из леммы 16 при $\rho_i = b_i$ для $1 \leq i \leq N-1$, $\rho_0 = b_0 + \sigma_0/\tilde{h}_0$, $\rho_N = b_N + \sigma_1/\tilde{h}_N$ и $x_0 = x_1 = 0$, $d_i \equiv 1$ получим оценку

$$(by, y) + \sigma_0 y_0^2 + \sigma_1 y_N^2 = (\rho y, y) \leq \max_{0 \leq i \leq N} v_i [(ay_x^2, 1)_{\omega+} + (y, y)],$$

где v_i — решение вспомогательной задачи (54). Используя лемму 17, будем иметь

$$\begin{aligned} (ay_x^2, 1)_{\omega+} + (by, y) + \sigma_0 y_0^2 + \sigma_1 y_N^2 &\leq (1+c)(ay_x^2, 1)_{\omega+} + \\ &+ c(y, y) \leq [\gamma_2 + (1+\gamma_2)c](y, y), \quad c = \max_{0 \leq i \leq N} v_i. \end{aligned}$$

Лемма 18 доказана.

6. Разностные схемы как операторные уравнения в абстрактных пространствах. После замены производных, входящих в дифференциальное уравнение и краевые условия, разностными отношениями на некоторой сетке $\bar{\omega}$ мы получаем разностную схему. Разностные уравнения, связывая искомые значения сеточной функции в узлах $\bar{\omega}$, образуют систему алгебраических уравнений. Эта система линейная, если исходная задача была линейной.

Разностная схема определяется разностным оператором, задающим структуру разностных уравнений в узлах сетки, где ищется искомое решение, и красными условиями в граничных узлах. Разностный оператор действует в пространстве сеточных функций, заданных на $\bar{\omega}$.

Рассмотрим пример. Пусть на отрезке $0 \leq x \leq l$ требуется найти решение задачи

$$\begin{aligned} u'' &= -\varphi(x), \quad 0 < x < l, \\ u'(0) &= \kappa_0 u(0) - \mu_1, \quad u(l) = \mu_2, \quad \kappa_0 \geq 0. \end{aligned} \quad (55)$$

На равномерной сетке $\bar{\omega} = \{x_i = ih, i = 0, 1, \dots, N, hN = l\}$ задаче (55) поставим в соответствие разностную схему

$$\begin{aligned} \Lambda y_i &= y_{xx,i} = -\varphi_i, \quad 1 \leq i \leq N-1, \\ \Lambda y_0 &= \frac{2}{h}(y_{x,0} - \kappa_0 y_0) = -\left(\varphi_0 + \frac{2}{h}\mu_1\right), \\ y_N &= \mu_2. \end{aligned} \quad (56)$$

Разностный оператор Λ определен на $(N+1)$ -мерном множестве сеточных функций, заданных на $\bar{\omega}$, и отображает его на N -мерное множество функций, заданных на $\omega^- = \{x_i \in \bar{\omega}, i = 0, 1, \dots, N-1\}$. Видно, что область определения и область значений оператора Λ не совпадают.

Рассмотрим теперь пространство $H(\omega^-)$ сеточных функций, заданных на ω^- . Скалярное произведение в $H(\omega^-)$ определим, как в примере 1 из п.1 § 2:

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0.5 h u_0 v_0, \quad u, v \in H(\omega^-).$$

Определим теперь линейный оператор A следующим образом: $Ay_i = -\dot{\Lambda}y_i$, $0 \leq i \leq N-1$, где $y \in H(\omega^-)$, $\dot{y}_i = y_i$ для $0 \leq i \leq N-1$ и $\dot{y}_N = 0$. Используя это определение, дадим подробную запись оператора A :

$$Ay_i = \begin{cases} -\frac{2}{h}(y_{x,0} - \kappa_0 y_0), & i = 0, \\ -y_{xx,i}, & 1 \leq i \leq N-2, \\ \frac{1}{h^2}(2y_{N-1} - y_{N-2}), & i = N-1. \end{cases} \quad (57)$$

Оператор A отображает $H(\omega^-)$ на $H(\omega^-)$ и является линейным.

Преобразуем разностную схему (2). Учитывая условие $y_N = \mu_2$, запишем (56) в виде

$$\begin{aligned} -\frac{2}{h}(y_{x,0} - \kappa_0 y_0) &= f_0 = \left(\varphi_0 + \frac{2}{h}\mu_1\right), \\ -y_{xx,i} &= f_i = \varphi_i, \quad 1 \leq i \leq N-2, \\ \frac{1}{h^2}(2y_{N-1} - y_{N-2}) &= f_{N-1} = \left(\varphi_{N-1} + \frac{1}{h^2}\mu_2\right). \end{aligned} \quad (58)$$

Сравнивая (57) и (58), найдем, что разностная схема (56) записывается в виде операторного уравнения первого рода

$$Ay = f, \quad (59)$$

где y — искомый, f — заданный элементы пространства $H(\omega^-)$, а A — оператор, действующий в $H(\omega^-)$, определен выше.

Укажем основные свойства оператора A .

Оператор A самосопряжен в $H(\omega^-)$, т. е.

$$(Au, v) = (u, Av), \quad u, v \in H(\omega^-).$$

Действительно, $(Au, v) = -(\Lambda \dot{u}, \dot{v})$, причем $\dot{u}_N = \dot{v}_N = 0$. Пользуясь второй разностной формулой Грина (13), получим

$$\begin{aligned} (\Lambda \dot{u}, \dot{v}) &= \sum_{i=1}^{N-1} \dot{u}_{\bar{x}x, i} \dot{v}_{\bar{x}i} h + (\dot{u}_{x, 0} - \kappa_0 \dot{u}_0) \dot{v}_0 = \\ &= \sum_{i=1}^{N-1} \dot{u}_{\bar{x}} \dot{v}_{\bar{x}x, i} h + (\dot{u}_{\bar{x}} \dot{v} - \dot{v}_{\bar{x}} \dot{u})_N - (\dot{u}_x \dot{v} - \dot{v}_x \dot{u})_0 + \\ &\quad + (\dot{u}_x \dot{v} - \kappa_0 \dot{u} \dot{v})_0 = \sum_{i=1}^{N-1} \dot{u}_{\bar{x}} \dot{v}_{\bar{x}x, i} h + (\dot{v}_{\bar{x}} \dot{u} - \kappa_0 \dot{u} \dot{v})_0 = (\dot{u}, \Lambda \dot{v}). \end{aligned}$$

Утверждение доказано.

Оператор A положительно определен, т. е.

$$(Au, u) \geq \gamma_1 (u, u), \quad u \in H(\omega^-),$$

где $\gamma_1 = \frac{8(1+\kappa_0)^2}{l^2(2+\kappa_0)^2} \geq \frac{2}{l^2} > 0$. Это утверждение следует из замечаний 1 и 2 к лемме 16. Оператор A в силу леммы 10 имеет ограниченный обратный оператор A^{-1} . Поэтому решение уравнения (59) существует и единствено.

Для оператора A имеет место оценка сверху

$$(Au, u) \leq \gamma_2 (u, u), \quad u \in H(\omega^-),$$

где $\gamma_2 = \frac{4}{h^2} \left(1 + \kappa_0 \frac{h}{2}\right)$, так как $y_N = 0$ и

$$(Ay, y) = (y_{\bar{x}}^2, 1)_{\omega+} + \kappa_0 y_0^2,$$

$$y_0^2 \leq \frac{2}{h} (y, y), \quad (y_{\bar{x}}^2, 1)_{\omega+} \leq \frac{4}{h^2}.$$

Последнее неравенство следует из леммы 17.

В качестве второго примера рассмотрим на неравномерной сетке $\bar{\omega} = \{x_i \in [0, l], x_i = x_{i-1} + h_i, 1 \leq i \leq N, x_0 = 0, x_N = l\}$ разностную схему

$$\Lambda y_i = (ay_{\bar{x}})_{\bar{x}, i} - d_i y_i = -\varphi_i, \quad 1 \leq i \leq N-1,$$

$$\Lambda y_0 = \frac{1}{h_0} (a_1 y_{x, 0} - \kappa_0 y_0) - d_0 y_0 = -\left(\varphi_0 + \frac{1}{h_0} \mu_1\right), \quad i = 0, \quad (60)$$

$$\Lambda y_N = -\frac{1}{h_N} (a_N y_{x, N} + \kappa_1 y_N) - d_N y_N = -\left(\varphi_N + \frac{1}{h_N} \mu_2\right), \quad i = N.$$

Схема (60) аппроксимирует третью краевую задачу для уравнения с переменными коэффициентами

$$\begin{aligned} (ku')' - qu &= -\varphi(x), \quad 0 < x < l, \\ ku' &= \kappa_0 u - \mu_1, \quad x = 0, \\ -ku' &= \kappa_1 u - \mu_2, \quad x = l \end{aligned}$$

при соответствующем выборе коэффициентов a_i и d_i , например при $a_i = k(x_i - 0,5h_i)$ и $d_i = q(x_i)$.

Если в пространстве $H(\omega^-)$ сеточных функций, заданных на $\bar{\omega}$, со скалярным произведением

$$(u, v) = \sum_{i=0}^N u_i v_i \bar{h}_i, \quad \bar{h}_0 = 0,5h_1, \quad \bar{h}_N = 0,5h_N,$$

определен оператор $A = -\Lambda$ и сеточную функцию $f_i = \varphi_i$, $1 \leq i \leq N-1$, $f_0 = \varphi_0 + \mu_1/\bar{h}_0$, $f_N = \varphi_N + \mu_2/\bar{h}_N$, то разностная схема (60) запишется в виде операторного уравнения (59).

Самосопряженность оператора A , отображающего $H(\bar{\omega})$ на $H(\bar{\omega})$, следует из второй разностной формулы Грина.

Если выполнены условия $a_i \geq c_1 > 0$, $d_i \geq 0$, $\kappa_0 \geq 0$, $\kappa_1 \geq 0$, $\kappa_0 + \kappa_1 + (d, 1) > 0$, то оператор A положительно определен в $H(\omega)$, и верна оценка $(Au, u) \geq \gamma_1(u, u)$, $1/\gamma_1 = \max_{0 \leq i \leq N} v_i$, где v_i — решение задачи $\Lambda v_i = -1$, $0 \leq i \leq N$. Заметим, что положительность v_i следует из принципа максимума, справедливого для оператора Λ при указанных условиях.

Если $d_i \equiv 0$, то грубую оценку для γ_1 можно получить следующим образом. Из первой разностной формулы Грина получим

$$(Ay, y) = (-\Lambda y, y) = (ay_x^2, 1)_{\omega^+} + \kappa_0 y_0^2 + \kappa_1 y_1^2.$$

В силу условий $a_i \geq c_1 > 0$, $1 \leq i \leq N$, отсюда найдем

$$(Ay, y) \geq c_1 [(y_x^2, 1)_{\omega^+} + \bar{\kappa}_0 y_0^2 + \bar{\kappa}_1 y_1^2],$$

где $c_1 \bar{\kappa}_0 = \kappa_0$, $c_1 \bar{\kappa}_1 = \kappa_1$. Так как $\kappa_0 + \kappa_1 > 0$, то из замечания 1 к лемме 16 получим оценку

$$(y_x^2, 1)_{\omega^+} + \bar{\kappa}_0 y_0^2 + \bar{\kappa}_1 y_1^2 \geq \bar{\gamma}_1(y, y),$$

где $\bar{\gamma}_1 = \frac{8(\bar{\kappa}_0 + \bar{\kappa}_1 + l\bar{\kappa}_0\bar{\kappa}_1)^2}{l(2 + l\bar{\kappa}_0)(2 + l\bar{\kappa}_1)(2\bar{\kappa}_0 + 2\bar{\kappa}_1 + l\bar{\kappa}_0\bar{\kappa}_1)}.$

Подставляя сюда $\bar{\kappa}_0$ и $\bar{\kappa}_1$, найдем, что $(Au, u) \geq \gamma_1(u, u)$, где

$$\gamma_1 = c_1 \bar{\gamma}_1 = \frac{8c_1(c_1\kappa_0 + c_1\kappa_1 + l\kappa_0\kappa_1)^2}{l(2c_1 + l\kappa_0)(2c_1 + l\kappa_1)(2c_1\kappa_0 + 2c_1\kappa_1 + l\kappa_0\kappa_1)}.$$

Для оператора A имеет место оценка сверху $(Au, u) \leq \gamma_2(u, u)$, где γ_2 определено в лемме 18, так как

$$(Ay, y) = (ay_x^2, 1)_{\omega^+} + (dy^2, 1) + \kappa_0 y_0^2 + \kappa_1 y_1^2.$$

В рассматриваемом примере оператор A и разностный оператор Λ определены в одном и том же пространстве сеточных функций $H(\bar{\omega})$ и отличаются лишь знаком. В отличие от первого примера, правые части разностной схемы (60) и операторного уравнения (59) совпадают.

Мы ограничились здесь простейшими примерами. В следующем пункте разностные схемы, аппроксимирующие эллиптические краевые задачи в пространстве нескольких измерений, аналогичным способом будут сводиться к операторным уравнениям в соответствующих конечномерных гильбертовых пространствах сеточных функций. Будут также изучены основные свойства таких операторов.

Из приведенных примеров видно, что разностные схемы можно трактовать как операторные уравнения с операторами в линейном нормированном конечномерном пространстве. Для этих операторов характерно то, что они отображают все пространство в себя.

7. Разностные схемы для эллиптических уравнений с постоянными коэффициентами. Пусть $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ — прямоугольник, $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha; \alpha = 1, 2\}$ — сетка в \bar{G} , γ — множество граничных узлов сетки $\bar{\omega}$. Сетка равномерна по каждому направлению x_α с шагом h_α . Обозначим через ω множество внутренних узлов сетки. Введем пространство сеточных функций $H = H(\omega)$, заданных на ω . Определим в H скалярное произведение

$$(u, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} u(i, j) v(i, j) h_1 h_2.$$

Рассмотрим разностную задачу Дирихле для уравнения Пуассона на сетке $\bar{\omega}$

$$\begin{aligned} \Lambda y &= \sum_{\alpha=1}^2 \Lambda_\alpha y = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \tag{61}$$

где $\Lambda_\alpha y = y_{x_\alpha x_\alpha}$, $\alpha = 1, 2$.

Разностную схему (61) можно записать в виде операторного уравнения (59). Для этого определим оператор A по формуле $Ay = -\Lambda \hat{y}$, $x \in \omega$, где $y \in H$, $\hat{y} \in \dot{H}$ и $y(x) = \hat{y}(x)$ для $x \in \omega$. Здесь \dot{H} — множество сеточных функций, заданных на $\bar{\omega}$ и обращающихся в нуль на γ . Правая часть f уравнения (59) отличается от правой части φ разностной схемы (61) лишь в приграничных узлах

$$f = \varphi + \varphi_1/h_1^2 + \varphi_2/h_2^2,$$

где

$$\varphi_1(x) = \begin{cases} g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leqslant x_1 \leqslant l_1 - 2h_1, \\ g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leqslant x_2 \leqslant l_2 - 2h_2, \\ g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

Исследуем свойства оператора A , действующего из $H(\omega)$ в $H(\omega)$.

1. Оператор A самосопряжен:

$$(Au, v) = (u, Av), \quad u, v \in H(\omega). \quad (62)$$

Для доказательства учтем, что

$$(A_1 u, v) = (-\Lambda_1 \dot{u}, \dot{v}) = - \sum_{j=1}^{N_2-1} h_2 \sum_{i=1}^{N_1-1} h_1 (\dot{v} \Lambda_1 \dot{u})_{ij} =$$

$$= - \sum_{j=1}^{N_2-1} h_2 \sum_{i=1}^{N_1-1} h_1 (\dot{u} \Lambda_1 \dot{v})_{ij} = - (\dot{u}, \Lambda_1 \dot{v}) = (u, A_1 v),$$

так как разностный оператор Λ_1 в силу второй разностной формулы Грина на сетке $\omega_1 = \{x_1(i) = ih_1, 0 \leqslant i \leqslant N_1, h_1 N_1 = l_1\}$ удовлетворяет равенству

$$\sum_{i=1}^{N_1-1} h_1 (\dot{v} \Lambda_1 \dot{u})_{ij} = \sum_{i=1}^{N_1-1} h_1 (\dot{u} \Lambda_1 \dot{v})_{ij}$$

и, кроме того, можно менять порядок суммирования по i и j .

Аналогично находим, что $(A_2 u, v) = (u, A_2 v)$. Отсюда следует (62).

2. Оператор A положительно определен, и для него справедливы оценки

$$\delta E \leqslant A \leqslant \Delta E, \quad \delta > 0, \quad (63)$$

где

$$\delta = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{\pi}{2N_\alpha} \geqslant \sum_{\alpha=1}^2 \frac{8}{l_\alpha^2}, \quad \Delta = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \cos^2 \frac{\pi}{2N_\alpha} \leqslant \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2}. \quad (64)$$

Заметим, что δ и Δ являются минимальным и максимальным собственными значениями разностного оператора Лапласа Λ (см. п. 1, § 2 гл. IV).

Это утверждение доказывается так же, как и лемма 12. Таким образом, мы установили, что в $H = H(\omega)$

$$A = A^*, \quad \delta E \leqslant A \leqslant \Delta E, \quad \delta > 0.$$

Если на части γ_0 сеточной границы γ задано краевое условие первого рода $y(x) = g(x)$, $x \in \gamma_0$, а на остальной части — краевые

условия второго или третьего рода, то оператор A определяется описанным выше методом, причем \dot{H} — множество функций, обращающихся в нуль лишь на γ_0 , а $H = H(\omega_0)$ — пространство сечеточных функций, заданных на $\omega_0 = \omega \setminus (\gamma \setminus \gamma_0)$. Например, пусть $\gamma_0 = \{x_{ij} \in \omega, i=0, 0 \leq j \leq N_2\}$, а на $\gamma \setminus \gamma_0$ заданы краевые условия второго рода. Тогда разностная схема записывается в виде

$$\begin{aligned}\Lambda y = (\Lambda_1 + \Lambda_2) y &= -\varphi(x), & x \in \omega_0, \\ y(x) &= g(x), & x \in \gamma_0.\end{aligned}$$

Здесь

$$\Lambda_2 y = \begin{cases} \frac{2}{h_2} y_{x_2}, & x_2 = 0, \\ \frac{y_{\bar{x}_1 x_2}}{h_2}, & h_2 \leq x_2 \leq l_2 - h_2, \\ -\frac{2}{h_2} y_{\bar{x}_2}, & x_2 = l_2, h_1 \leq x_1 \leq l_1, \end{cases}$$

а оператор Λ_1 задается формулами

$$\Lambda_1 y = \begin{cases} y_{\bar{x}_1 x_1}, & h_1 \leq x_1 \leq l_1 - h_1, \\ -\frac{2}{h_1} y_{\bar{x}_1}, & x_1 = l_1, 0 \leq x_2 \leq l_2. \end{cases}$$

Скалярное произведение в пространстве $H = H(\omega_0)$ определяется по формуле

$$(u, v) = \sum_{i=1}^{N_1} \sum_{j=0}^{N_2} u(i, j) v(i, j) \tilde{h}_1(i) \tilde{h}_2(j),$$

где

$$\begin{aligned}\tilde{h}_1(i) &= \begin{cases} h_1, & 1 \leq i \leq N_1 - 1, \\ 0,5h_1, & i = N_1, \end{cases} \\ \tilde{h}_2(j) &= \begin{cases} h_2, & 1 \leq j \leq N_2 - 1, \\ 0,5h_2, & j = 0, N_2. \end{cases}\end{aligned}$$

Можно показать, что оператор $A = A_1 + A_2$, соответствующий разностному оператору Λ , самосопряжен в H , и для него верны оценки (63) с $\delta = \delta_1 + \delta_2$, $\Delta = \Delta_1 + \Delta_2$, $\delta_1 = \frac{4}{h_1^2} \sin^2 \frac{\pi}{4N_1}$, $\Delta_1 = \frac{4}{h_1^2} \times \cos^2 \frac{\pi}{4N_1}$, $\delta_2 = 0$, $\Delta_2 = \frac{4}{h_2^2}$. Здесь δ_α и Δ_α — минимальное и максимальное собственные значения разностного оператора Λ_α , $\alpha = 1, 2$.

Заметим, что операторы A_1 и A_2 являются перестановочными как для первой, так и для второй краевых задач. Поэтому, в силу общей теории (см. п. 5 § 1 гл. V), собственные значения оператора A являются суммой собственных значений операторов A_1 и A_2 : $\lambda(A) = \lambda(A_1) + \lambda(A_2)$.

8. Уравнения с переменными коэффициентами и со смешанными производными. Рассмотрим задачу Дирихле для эллиптического уравнения с переменными коэффициентами в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$:

$$\begin{aligned} Lu &= \sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x) \frac{\partial u}{\partial x_\alpha} \right) - q(x) u = -\varphi(x), \quad x \in G, \\ u(x) &= g(x), \quad x \in \Gamma, \end{aligned} \quad (65)$$

где $k_\alpha(x)$ и $q(x)$ — достаточно гладкие функции, удовлетворяющие условиям $0 < c_1 \leq k_\alpha(x) \leq c_2$, $0 \leq d_1 \leq q(x) \leq d_2$. Обозначим через $\bar{\omega} = \omega + \gamma$ сетку с шагами h_1 и h_2 , введенную в п. 7.

Задаче (65) поставим в соответствие разностную задачу Дирихле на сетке $\bar{\omega}$:

$$\begin{aligned} Ly &= (\Lambda_1 + \Lambda_2)y - dy = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \quad (66)$$

где $\Lambda_\alpha y = (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}$, $\alpha = 1, 2$, а $a_\alpha(x)$ и $d(x)$ выбираются, например, так:

$$\begin{aligned} a_1(x_1, x_2) &= k_1(x_1 - 0,5h_1, x_2), \\ a_2(x_1, x_2) &= k_2(x_1, x_2 - 0,5h_2), \quad d(x) = q(x). \end{aligned}$$

Тогда коэффициенты разностной схемы удовлетворяют условиям

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad 0 \leq d_1 \leq d \leq d_2. \quad (67)$$

Обозначим через $H = H(\omega)$ пространство сеточных функций, введенное в предыдущем пункте, а через \dot{H} — множество сеточных функций, обращающихся в нуль на γ .

Запишем разностную схему (66) в виде операторного уравнения (59), где оператор A определим обычным образом: $Ay = -Ly$, где $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$ для $x \in \omega$.

Обозначим через $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2$, где $\mathcal{R}_\alpha y = y_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$, разностный оператор Лапласа и определим соответствующий ему в H оператор R : $Ry = -\mathcal{R}\dot{y}$, $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$ для $x \in \omega$.

Лемма 19. *Оператор A самосопряжен в H , и для него верны оценки*

$$(c_1 + d_1/\Delta)(Ru, u) \leq (Au, u) \leq (c_2 + d_2/\delta)(Ru, u), \quad (68)$$

$$(c_1\delta + d_1)(u, u) \leq (Au, u) \leq (c_2\Delta + d_2)(u, u), \quad (69)$$

где δ и Δ определены в (64).

В самом деле, из условий (67) и оценок, полученных в предыдущем пункте

$$\delta E \leq R \leq \Delta E, \quad (70)$$

следует, что для любого $u \in H$ верны неравенства

$$\frac{d_1}{\Delta} (Ru, u) \leq d_1(u, u) \leq (du, u) \leq d_2(u, u) \leq \frac{d_2}{\delta} (Ru, u). \quad (71)$$

Далее, первая разностная формула Грина дает

$$(A_1 u, u) = -(\Lambda_1 \dot{u}, \dot{u}) = \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} (a_1 \dot{u}_{x_i}^2)_{ij} h_1 h_2,$$

$$(R_1 u, u) = -(\mathcal{R}_1 \dot{u}, \dot{u}) = \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} (\dot{u}_{x_i}^2)_{ij} h_1 h_2.$$

В силу (67) отсюда получим неравенства

$$c_1 (R_1 u, u) \leq (A_1 u, u) \leq c_2 (R_1 u, u).$$

Аналогично найдем, что

$$c_1 (R_2 u, u) \leq (A_2 u, u) \leq c_2 (R_2 u, u).$$

Отсюда и из (70) вытекают неравенства

$$c_1 \delta(u, u) \leq c_1 (Ru, u) \leq ((A_1 + A_2) u, u) \leq c_2 (Ru, u) \leq c_2 \Delta(u, u),$$

складывая которые с неравенствами (71) будем иметь (68) и (69).

Самосопряженность оператора A доказывается по аналогии с предыдущим пунктом.

Отметим, что в неравенствах (68) указаны постоянные энергетической эквивалентности операторов R и A , причем, так как $d_1 \geq 0$ и $\delta \geq 8/l_1^2 + 8/l_2^2$, то эти операторы эквивалентны с постоянными, не зависящими от числа узлов сетки.

Рассмотрим теперь задачу Дирихле для эллиптического уравнения, содержащего смешанные производные

$$Lu = \sum_{\alpha, \beta=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}(x) \frac{\partial u}{\partial x_\beta} \right) = -\varphi(x), \quad x \in \bar{G}, \quad (72)$$

$$u(x) = g(x), \quad x \in \Gamma.$$

Предполагается, что выполнены условия эллиптичности

$$c_1 \sum_{\alpha=1}^2 \xi_\alpha^2 \leq \sum_{\alpha, \beta=1}^2 k_{\alpha\beta}(x) \xi_\alpha \xi_\beta \leq c_2 \sum_{\alpha=1}^2 \xi_\alpha^2, \quad x \in \bar{G}, \quad (73)$$

где $c_2 \geq c_1 > 0$, а $\xi = (\xi_1, \xi_2)$ — произвольный вектор.

На прямоугольной сетке $\bar{\omega}$ задаче (72) можно поставить в соответствие разностную схему

$$\Lambda y = 0,5 \sum_{\alpha, \beta=1}^2 [(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}] = -\varphi(x), \quad x \in \omega,$$

$$y(x) = g(x), \quad x \in \gamma. \quad (74)$$

Запишем (74) в виде операторного уравнения (59), определяя

оператор A обычным образом: $Ay = -\Lambda \dot{y}$, где $y \in H(\omega)$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$ для $x \in \omega$. При этом правая часть f отличается от правой части φ уравнения (74) лишь в приграничных узлах. Для нахождения явного вида f следует записать разностное уравнение в приграничном узле, воспользоваться краевым условием и перенести в правую часть уравнения известные значения $y(x)$ на γ .

Покажем теперь, что при выполнении условия симметрии $k_{12}(x) = k_{21}(x)$ оператор A самосопряжен в пространстве $H = H(\omega)$, определенном выше. Для этого запишем оператор Λ в виде суммы $\Lambda = (\Lambda_1 + \Lambda_2)/2$, где

$$\Lambda_\alpha y = (k_{\alpha\beta} y_{\bar{x}_\alpha} + k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\alpha} + k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha},$$

$$\beta = 3 - \alpha, \alpha = 1, 2.$$

Используя формулы суммирования по частям (7') и (9'), получим для любых $u, v \in \dot{H}$

$$(\Lambda_1 \dot{u}, \dot{v}) = - \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} [(k_{11} \dot{u}_{\bar{x}_i} + k_{12} \dot{u}_{\bar{x}_j}) \dot{v}_{\bar{x}_i}]_{ij} h_1 h_2 -$$

$$- \sum_{j=1}^{N_2-1} \sum_{i=0}^{N_1-1} [(k_{11} \dot{u}_{x_i} + k_{12} \dot{u}_{x_j}) \dot{v}_{x_i}]_{ij} h_1 h_2.$$

Учитывая, что $\dot{v}_{\bar{x}_i}$ и \dot{v}_{x_i} равны нулю соответственно при $j = N_2$ и $j = 0$, полученное равенство можно записать в виде

$$(\Lambda_1 \dot{u}, \dot{v}) = - \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} [(k_{11} \dot{u}_{\bar{x}_i} + k_{12} \dot{u}_{\bar{x}_j}) \dot{v}_{\bar{x}_i}]_{ij} h_1 h_2 -$$

$$- \sum_{j=0}^{N_2-1} \sum_{i=0}^{N_1-1} [(k_{11} \dot{u}_{x_i} + k_{12} \dot{u}_{x_j}) \dot{v}_{x_i}]_{ij} h_1 h_2. \quad (75)$$

Аналогично найдем

$$(\Lambda_2 \dot{u}, \dot{v}) = - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [(k_{22} \dot{u}_{\bar{x}_j} + k_{21} \dot{u}_{\bar{x}_i}) \dot{v}_{\bar{x}_j}]_{ij} h_1 h_2 -$$

$$- \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} [(k_{22} \dot{u}_{x_j} + k_{21} \dot{u}_{x_i}) \dot{v}_{x_j}]_{ij} h_1 h_2. \quad (76)$$

Складывая (75) и (76), получаем

$$(\Lambda \dot{u}, \dot{v}) = -0,5 \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} h_1 h_2 \left(\sum_{\alpha, \beta=1}^2 k_{\alpha\beta} \dot{u}_{\bar{x}_\alpha} \dot{v}_{\bar{x}_\beta} \right)_{ij} -$$

$$-0,5 \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} h_1 h_2 \left(\sum_{\alpha, \beta=1}^2 k_{\alpha\beta} \dot{u}_{x_\alpha} \dot{v}_{x_\beta} \right)_{ij}. \quad (77)$$

Отсюда следует, что при условии $k_{12} = k_{21}$ выполняется равенство

$$(\Lambda \dot{u}, \dot{v}) = (\dot{u}, \Lambda \dot{v}).$$

В силу равенства $(Au, v) = -(\Lambda \dot{u}, \dot{v})$ оператор A самосопряжен в H .

Найдем границы оператора A . Подставим в (77) вместо \dot{v} сечочную функцию \dot{u} , учтем условия эллиптичности (73) и условие $\dot{u}(x) = 0$ для $x \in \gamma$. Получим

$$\begin{aligned} -(\Lambda \dot{u}, \dot{u}) &\geq 0,5c_1 \left\{ \sum_{j=1}^{N_2-1} h_2 \left[\sum_{i=1}^{N_1} (\dot{u}_{\bar{x}_1})_{ij}^2 h_1 + \sum_{i=0}^{N_1-1} (\dot{u}_{x_1})_{ij}^2 h_1 \right] + \right. \\ &\quad \left. + \sum_{i=1}^{N_1-1} h_1 \left[\sum_{j=1}^{N_2} (\dot{u}_{\bar{x}_2})_{ij}^2 h_2 + \sum_{j=0}^{N_2-1} (\dot{u}_{x_2})_{ij}^2 h_2 \right] \right\} = \\ &= c_1 \left[\sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} (\dot{u}_{\bar{x}_1})_{ij}^2 h_1 h_2 + \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2} (\dot{u}_{\bar{x}_2})_{ij}^2 h_1 h_2 \right] = c_1 (-\mathcal{R}\dot{u}, \dot{u}), \end{aligned}$$

где \mathcal{R} — разностный оператор Лапласа. Аналогично найдем

$$-(\Lambda \dot{u}, \dot{u}) \leq c_2 (-\mathcal{R}\dot{u}, \dot{u}).$$

Учитывая оценку (70), получаем следующие неравенства для оператора A :

$$\begin{aligned} c_1 (Ru, u) &\leq (Au, u) \leq c_2 (Ru, u), \\ c_1 \delta(u, u) &\leq (Au, u) \leq c_2 \Delta(u, u), \end{aligned} \tag{78}$$

где δ и Δ определены в (64). Следовательно, оператор A , соответствующий разностному эллиптическому оператору со смешанными производными, и оператор R , соответствующий разностному оператору Лапласа, энергетически эквивалентны с постоянными c_1 и c_2 , не зависящими от числа узлов сетки. Оператор A имеет границы $c_1 \delta = O(1)$ и $c_2 \Delta = O(1/h^2)$ ($h^2 = h_1^2 + h_2^2$), и если число узлов сетки велико, то оператор A плохо обусловлен.

Отметим, что неравенства (78) остаются верными и в случае, когда для аппроксимации дифференциального оператора L используются разностные операторы

$$\Lambda y = \frac{1}{2} \sum_{\alpha=1}^2 [(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \frac{1}{2} \sum_{\alpha \neq \beta}^{1 \div 2} [(k_{\alpha\beta} y_{x_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha}]$$

или

$$\begin{aligned} \Lambda y &= \frac{1}{2} \sum_{\alpha=1}^2 [(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \\ &\quad + \frac{1}{4} \sum_{\alpha \neq \beta}^{1 \div 2} [(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha}]. \end{aligned}$$

§ 3. Основные понятия теории итерационных методов

1. Метод установления. Выше было показано, что разностные схемы для эллиптических уравнений естественным образом записываются в виде операторного уравнения первого рода

$$Au = f \quad (1)$$

с оператором A , действующим в гильбертовом пространстве H конечной размерности. Линейным эллиптическим уравнениям соответствуют линейные операторы A , а квазилинейным — нелинейные операторы A .

Теория итерационных методов для операторного уравнения (1) может быть изложена как один из разделов общей теории устойчивости разностных схем. Итерационные схемы можно трактовать как методы установления для соответствующего нестационарного уравнения. Поясним это на примере уравнения с самосопряженным положительно определенным и ограниченным оператором A , $A = A^* \geqslant \delta E$, $\delta > 0$.

Пусть $v = v(t)$ абстрактная функция t со значениями в H , т. е. $v(t)$ при каждом фиксированном t есть элемент пространства H . Рассмотрим абстрактную задачу Коши:

$$\frac{dv}{dt} + Av = f, \quad t > 0, \quad v(0) = v_0 \in H. \quad (2)$$

Покажем, что $\lim_{t \rightarrow \infty} \|v(t) - u\| = 0$, где u — решение уравнения (1), т. е. решение $v(t)$ нестационарного уравнения (2) с ростом t стремится к решению u стационарного (не зависящего от t) уравнения (1) (имеет место «установление» или «выход на стационарный режим»). Для погрешности $z(t) = v(t) - u$ имеем однородное уравнение

$$\frac{dz}{dt} + Az = 0, \quad t > 0, \quad z(0) = v(0) - u.$$

Умножая это уравнение скалярно на z : $\left(\frac{dz}{dt}, z\right) + (Az, z) = 0$ и учитывая, что

$$\left(\frac{dz}{dt}, z\right) = \frac{1}{2} \frac{d}{dt} (z, z) = \frac{1}{2} \frac{d}{dt} \|z\|^2, \quad (Az, z) \geqslant \delta \|z\|^2,$$

получим

$$\frac{d}{dt} \|z(t)\|^2 + 2\delta \|z(t)\|^2 \leqslant 0.$$

После умножения этого неравенства на $e^{2\delta t} > 0$ имеем

$$\frac{d}{dt} e^{2\delta t} \|z(t)\|^2 \leqslant 0,$$

откуда следует $e^{2\delta t} \|z(t)\|^2 \leqslant \|z(0)\|^2$ или

$$\|v(t) - u\| \leqslant e^{-\delta t} \|v(0) - u\| \rightarrow 0 \quad \text{при } t \rightarrow \infty.$$

Таким образом, решая уравнение (2) с любым $v_0 \in H$, мы при достаточно большом t получаем приближенное решение исходного уравнения (1) с любой заданной точностью. Такой метод получения решения называют *методом установления*. Аналогичным свойством затухания начальных данных обладают и разностные аналоги уравнения (2).

2. Итерационные схемы. Остановимся сначала на общей характеристике понятия итерационной схемы. Пусть требуется найти решение уравнения (1). Будем сначала предполагать, что A — линейный оператор, заданный в H .

В любом итерационном методе решения уравнения (1) исходят из некоторого начального приближения $y_0 \in H$ и последовательно определяют приближенные решения $y_1, y_2, \dots, y_k, y_{k+1}, \dots$, где k — номер итерации. Приближение y_{k+1} выражается через известные предыдущие приближения по рекуррентной формуле

$$y_{k+1} = F_k(y_0, y_1, \dots, y_k),$$

где F_k — некоторая функция, зависящая, вообще говоря, от оператора A , правой части f , номера итерации k .

Говорят, что итерационный метод имеет порядок m , если каждое последующее приближение зависит лишь от m предыдущих, т. е.

$$y_{k+1} = F_k(y_{k-m+1}, y_{k-m+2}, \dots, y_k).$$

Итерационные схемы высокого порядка при своей реализации требуют запоминания большого объема промежуточной информации и поэтому на практике обычно ограничиваются значениями $m=1$ или $m=2$.

От выбора функции F_k зависит структура итерационной схемы. Если функция линейная, то итерационный метод тоже называется линейным. Если F_k не зависит от номера итерации k , то итерационный метод называется стационарным.

Рассмотрим общий вид линейной итерационной схемы первого порядка. Любая такая схема, в соответствии с определением, может быть записана в виде

$$y_{k+1} = S_{k+1}y_k + \tau_{k+1}\Phi_{k+1}, \quad k = 0, 1, \dots, \quad (3)$$

где S_k — линейные операторы, заданные на H , τ_k — некоторые числовые параметры.

Обычно к итерационным схемам предъявляется естественное требование: решение $u = A^{-1}f \in H$ уравнения (1) для любого f должно быть неподвижной точкой процесса последовательных приближений (3), т. е.

$$A^{-1}f = S_{k+1}A^{-1}f + \tau_{k+1}\Phi_{k+1}. \quad (4)$$

Отсюда следует, что если положить

$$S_{k+1} = E - \tau_{k+1}B_{k+1}^{-1}A, \quad \Phi_{k+1} = B_{k+1}^{-1}f, \quad (5)$$

где B_{k+1} — линейный обратимый оператор, действующий в H , то условие (4) будет выполнено. Подставляя (5) в (3), получим в результате несложных преобразований

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H. \quad (6)$$

Сохраняя терминологию теории разностных схем (см. А. А. Са-марский, Теория разностных схем, 1977, гл. V), назовем (6) каноническим видом двухслойной итерационной схемы. Итак, любой линейный итерационный процесс первого порядка может быть записан в виде (6). Если $B_{k+1} \equiv E$, то итерационная схема называется явной, так как в этом случае приближение y_{k+1} находится по явной формуле

$$y_{k+1} = y_k - \tau_{k+1} (Ay_k - f), \quad k = 0, 1, \dots$$

Если B_k хотя бы для одного k отлично от единичного оператора, то схема называется неявной. Числа τ_k называются итерационными параметрами. Если τ_{k+1} зависит от итерационного приближения y_k , то итерационный процесс будет нелинейным. Очевидно, что в стационарном итерационном процессе операторы B_k и параметры τ_k (точнее, B_k/τ_{k+1}) не должны зависеть от номера итераций k .

Отметим, что схему (6) можно трактовать как неявную двухслойную схему для нестационарного уравнения

$$B(t) \frac{dv}{dt} + Av = f, \quad t > 0, v(0) = y_0,$$

более общего, чем рассмотренное выше уравнение (2). При этом параметр τ_{k+1} можно рассматривать как шаг по фиктивному времени.

Различие между итерационными схемами и схемами для нестационарных задач вида (2) заключается в следующем:

1) при любых B_{k+1} и τ_{k+1} решение исходного уравнения (1) удовлетворяет (6);

2) выбор параметров τ_{k+1} и операторов B_{k+1} следует подчинить лишь требованиям сходимости итераций и минимума арифметических действий, необходимых для нахождения решения уравнения (1) с заданной точностью (для нестационарных задач выбор шага подчинен прежде всего требованию аппроксимации).

Выше предполагалось, что оператор A линеен. Очевидно, схема (6) может быть использована для нахождения приближенного решения уравнения (1) и в случае нелинейного оператора A . При этом обычно оператор B_{k+1} выбирается линейным.

Двухслойные итерационные схемы (6) являются наиболее употребительными. Однако при решении уравнения (1) используют и трехслойные схемы, которые описывают итерационные

процессы второго порядка. Наиболее исследованными являются трехслойные схемы «стандартного» типа. Они записываются в виде

$$B_{k+1}y_{k+1} = \alpha_{k+1}(B_{k+1} - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})B_{k+1}y_{k-1} + \alpha_{k+1}\tau_{k+1}f \quad (7)$$

для $k = 1, 2, \dots$. Здесь используются две последовательности итерационных параметров $\{\tau_k\}$ и $\{\alpha_k\}$. Для реализации схемы (7) необходимо, помимо начального приближения y_0 , задать еще приближение y_1 . Обычно оно находится по y_0 с использованием двухслойной схемы (6), т. е.

$$B_1y_1 = (B_1 - \tau_1A)y_0 + \tau_1f, \quad y_0 \in H. \quad (8)$$

Можно показать, что для (7), (8) решение u уравнения (1) является неподвижной точкой.

Если $B_k \equiv E$ для всех $k = 1, 2, \dots$, то схема (7) называется явной:

$$y_{k+1} = \alpha_{k+1}(E - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})y_{k-1} + \alpha_{k+1}\tau_{k+1}f.$$

В противном случае схема (7) неявная.

3. Сходимость и число итераций. Основное отличие итерационных методов от прямых заключается в том, что итерационные методы дают точное решение уравнения (1) лишь как предел последовательности итерационных приближений $\{y_k\}$ при $k \rightarrow \infty$. Исключение составляют методы «конечных» итераций, к которым относятся методы сопряженных направлений, теоретически позволяющие найти точное решение при любом начальном приближении за конечное число действий, если A — линейный оператор в конечномерном пространстве.

Для характеристики отклонения итерационного приближения y_k от точного решения u задачи (1) вводится погрешность $z_k = y_k - u$. Итерационный процесс называется сходящимся в энергетическом пространстве H_D , если $\|z_k\|_D \rightarrow 0$ при $k \rightarrow \infty$. Здесь H_D — пространство, порожденное самосопряженным положительно определенным в H оператором D .

Смысъ введения энергетического пространства H_D заключается в следующем. Как мы знаем, последовательность элементов H , сходящаяся в одной норме, сходится и в эквивалентной норме. Поэтому при исследовании конкретной итерационной схемы удобно выбрать такое энергетическое пространство H_D , в котором операторы итерационной схемы A и B_k обладали бы заданными свойствами, например были самосопряжены и положительно определены.

Одной из важных количественных характеристик итерационного метода является число итераций. Обычно задается некоторая точность $\varepsilon > 0$, с которой надо найти приближенное решение уравнения (1). Если $\|u\|_D = O(1)$, то требуется, чтобы выполнялось условие

$$\|y_n - u\|_D \leq \varepsilon \quad \text{при } n \geq n_0(\varepsilon). \quad (9)$$

Здесь $n_0(\varepsilon)$ — минимальное число итераций, гарантирующее заданную точность ε . Это число зависит от того, какое взято начальное приближение. Условием (9) можно пользоваться для определения момента окончания итераций, если указанная норма может быть эффективно вычислена в процессе итераций. Например, если оператор A невырожден и положительно определен, то, выбирая в качестве D оператор A^*A , получим из (9)

$$\|y_n - u\|_D = \|Ay_n - f\| \leq \varepsilon,$$

так как

$$\begin{aligned} (y_n - u, y_n - u)_D &= (A^*A(y_n - u), y_n - u) = \\ &= (Ay_n - Au, Ay_n - Au) = \|Ay_n - f\|^2. \end{aligned}$$

Для сравнения качества различных методов в общем случае используется число итераций, определяемое из условия

$$\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_D \quad \text{при } n \geq n_0(\varepsilon). \quad (10)$$

Это число указывает, сколько итераций достаточно выполнить, чтобы при любом начальном приближении y_0 норма начальной погрешности в H_D была уменьшена в $1/\varepsilon$ раз. Условие (10) также можно использовать в качестве критерия для окончания процесса итераций.

Уравнению (1) можно поставить в соответствие большое число итерационных схем (6) или (7), (8) с любыми B_k и τ_k , α_k . Между тем при решении конкретной задачи возникает проблема выбора одной схемы. С точки зрения вычислительной математики наиболее важным является построение таких итерационных методов, которые позволяют получить решение (1) с заданной точностью за минимальное машинное время. Это требование экономичности метода является естественным. При теоретических оценках качества метода оно часто заменяется требованием минимума числа арифметических действий $Q(\varepsilon)$, достаточных для получения решения с заданной точностью.

Общий объем вычислений $Q(\varepsilon)$ равен $Q(\varepsilon) = \sum_{k=1}^n q_k$, где q_k — число действий для вычисления итерации номера k , а n — число итераций, $n \geq n_0(\varepsilon)$. Задача построения итерационного метода ставится так (для двухслойной схемы (6)): оператор A фиксирован, а параметры $\{\tau_k, k = 1, 2, \dots, n\}$ и операторы B_k нужно выбрать из условия минимума $Q(\varepsilon)$.

В такой общей постановке эта задача вряд ли имеет решение. Обычно набор операторов B_k задается априори, и если число действий, необходимое для обращения оператора B_k , не зависит от k , то $q_k \equiv q$ и $Q(\varepsilon) = qn_0(\varepsilon)$. В этом случае задача о минимуме $Q(\varepsilon)$ сводится к задаче выбора итерационных параметров τ_k из условия минимума числа итераций $n_0(\varepsilon)$.

Чтобы установить иерархию методов, надо сравнить их по каким-либо характеристикам. Иногда используются асимптотические оценки для числа действий или для числа итераций при стремлении числа неизвестных в разностной схеме к бесконечности. Однако существует фактическое ограничение на число неизвестных при решении эллиптических многомерных уравнений методом сеток. Так, например, для трехмерного уравнения Пуассона среднее число узлов по каждому переменному $N \approx 100$ приводит нас к системе линейных алгебраических уравнений с $M = 10^6$ неизвестными. Вряд ли целесообразно увеличение числа узлов. Поэтому надо сравнивать методы прежде всего на реальных сетках.

4. Классификация итерационных методов. Итерационные методы характеризуются структурой итерационной схемы, энергетическим пространством H_D , в котором исследуется сходимость метода, типом итерационного метода, условием окончания процесса итераций, а также алгоритмом реализации одного итерационного шага.

Мы будем рассматривать только двухслойные и трехслойные итерационные схемы, явные и неявные, для которых условием окончания процесса итераций будет условие

$$\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_D, \quad \varepsilon > 0.$$

В общей теории итерационных методов рассматриваются методы двух типов: использующие априорную информацию об операторах итерационной схемы и не использующие (методы вариационного типа). В первом случае итерационные параметры τ_k для схемы (6) и τ_k, α_k для схемы (7), (8) выбираются из условия минимума либо нормы разрешающего оператора (оператора, связывающего начальное и конечное приближение), либо нормы оператора перехода от итерации к итерации. При этом итерационные параметры выбираются так, чтобы обеспечить наивысшую скорость сходимости при самом плохом начальном приближении. В методах этого типа качество начального приближения не используется.

В методах вариационного типа итерационные параметры выбираются из условия минимума некоторых функционалов, связанных с исходным уравнением. Например, в качестве функционала берется энергетическая норма погрешности k -й итерации. В этом случае итерационные параметры зависят от предыдущих итерационных приближений и обладают свойством учитывать качество начального приближения.

В общей теории итерационных методов мы отказываемся от изучения конкретной структуры операторов итерационной схемы — теория использует минимум информации общего функционального характера относительно операторов. Это позволяет достичь главной цели — указать общие принципы конструирования оптимальных итерационных методов в зависимости от

характера и вида априорной информации о задаче, а также от тех требований, которые предъявляются к способу решения этой задачи. Эти дополнительные требования могут, например, состоять в том, что нужно построить метод оптимальный не на одной задаче, а на серии задач с одним и тем же оператором A , но с различными правыми частями.

Несомненно, что учет структуры оператора решаемой задачи позволяет построить специальные итерационные методы, которые обладают более высокой скоростью сходимости, чем методы из общей теории. Это достигается особым выбором операторов B_k и итерационных параметров. Специальные методы имеют узкую область применения.

Остановимся теперь на роли операторов B_k . Для неявных итерационных схем выбор операторов B_k должен быть подчинен двум требованиям: обеспечению наиболее быстрой сходимости метода и требованию простоты и экономичности обращения этих операторов. Эти требования противоречивы. Действительно, если в схеме (6) взять $B_1 = A$ и $\tau_1 = 1$, то при любом начальном приближении решение уравнения (1) может быть получено за одну итерацию. В этом случае скорость сходимости максимальна, однако обращение такого оператора B_1 эквивалентно решению исходной задачи.

Оказывается, и это будет показано ниже, что нет необходимости выбирать оператор B_k равным оператору A . Достаточно, чтобы были близки энергии этих операторов. Это требование открывает возможность выбирать из класса операторов B , близких по энергии к оператору A , легко обратимые операторы.

В настоящее время наиболее часто используется следующий подход при построении неявных итерационных методов. Оператор B_{k+1} задается либо конструктивно в явном виде, либо итерационное приближение y_{k+1} находится в результате некоторой вспомогательной вычислительной процедуры, которую можно трактовать как неявное обращение оператора B_{k+1} .

В первом случае оператор B_{k+1} обычно выбирают в виде произведения некоторого числа легко обратимых операторов так, чтобы оператор B_{k+1} в некотором смысле был близок к оператору A . При этом операторы, входящие в произведение, сами могут зависеть от параметров, которые можно рассматривать как дополнительные итерационные параметры. Например, если $B_k = (E + \omega_k A_1)(E + \omega_k A_2)$, где A_α — операторы, то ω_k — числа, являющиеся параметрами. В этом случае переменность оператора B_k проявляется лишь в зависимости указанных параметров ω_k от номера итерации k . При такой конструкции оператора B_k обеспечивается единообразие вычислительного процесса нахождения приближенного решения на каждой итерации.

Остановимся на двух алгоритмах нахождения нового приближения y_{k+1} в случае, когда оператор B_{k+1} имеет факторизованный вид. Пусть $B_{k+1} = B_{k+1}^1 B_{k+1}^2 \dots B_{k+1}^p$ и y_{k+1} находится

по двухслойной итерационной схеме (6). В первом алгоритме решается последовательность уравнений

$$B_{k+1}^1 v^1 = F_{k+1}, \quad B_{k+1}^\alpha v^\alpha = v^{\alpha-1}, \quad \alpha = 2, 3, \dots, p, \quad (11)$$

где $F_{k+1} = B_{k+1} y_k - \tau_{k+1} (A y_k - f)$. Видно, что $y_{k+1} = v^p$. Каждое из уравнений (11) должно легко решаться. Алгоритм не требует запоминания промежуточной информации, будучи полученной, она тут же используется. Недостаток алгоритма заключается в необходимости вычислять элемент $B_{k+1} y_k$, что может оказаться сложной процедурой.

Второй алгоритм имеет вид схемы с поправкой:

$$\begin{aligned} y_{k+1} &= y_k - \tau_{k+1} v^p, \\ B_{k+1}^1 v^1 &= A y_k - f, \quad B_{k+1}^\alpha v^\alpha = v^{\alpha-1}, \quad \alpha = 2, 3, \dots, p. \end{aligned} \quad (12)$$

В этом случае требуется дополнительно запоминать предыдущее итерационное приближение y_k и хранить его до тех пор, пока не будет найдена поправка v^p .

Во втором способе построения неявного итерационного метода исходят, например, из схемы для поправки (12), а поправку v^p находят как приближенное решение вспомогательного уравнения

$$R_{k+1} v = r_k, \quad r_k = A y_k - f. \quad (13)$$

Пусть (13) решается при помощи какой-либо двухслойной итерационной схемы. Тогда погрешность $z^m = v^m - v$ удовлетворяет однородному уравнению

$$z^{m+1} = S_{m+1} z^m, \quad m = 0, 1, \dots, p-1, \quad z^0 = v^0 - v,$$

где S_{m+1} — оператор перехода от m -й к $(m+1)$ -й итерации. Отсюда найдем

$$z^p = v^p - v = S_p S_{p-1} \dots S_1 z^0 = T_p (v^0 - v), \quad T_p = \prod_{m=1}^p S_m,$$

где T_p — разрешающий оператор. Подставляя сюда $v = R_{k+1}^{-1} r_k$ и выбирая $v^0 = 0$, получим

$$v^p = (E - T_p) R_{k+1}^{-1} r_k \quad \text{или} \quad v^p = B_{k+1}^{-1} r_k, \quad (14)$$

где через B_{k+1} обозначен оператор $R_{k+1} (E - T_p)^{-1}$.

Подставим (14) в (12) и найдем, что y_{k+1} удовлетворяет двухслойной схеме (6) с указанным оператором B_{k+1} . Если норма оператора T_p мала, то оператор B_{k+1} «близок» к оператору R_{k+1} . Поэтому в качестве оператора R_{k+1} естественно выбирать близкий к A оператор.

ГЛАВА VI

ДВУХСЛОЙНЫЕ ИТЕРАЦИОННЫЕ МЕТОДЫ

В главе рассматриваются двухслойные итерационные методы решения операторного уравнения $Au=f$. Итерационные параметры выбираются с использованием априорной информации об операторах итерационной схемы. В § 1 ставится задача о выборе параметров для двухслойной схемы. В §§ 2 и 3 эта задача решается для самосопряженного случая. Здесь построены чебышевский метод и метод простой итерации. В § 4 изучено несколько способов выбора итерационного параметра в несамосопряженном случае в зависимости от объема априорной информации. В § 5 рассмотрены некоторые примеры применения построенных методов для решения сеточных уравнений.

§ 1. Постановка задачи о выборе итерационных параметров

1. Исходное семейство итерационных схем. В главе V было показано, что разностные краевые задачи для эллиптических уравнений представляют собой специальные системы алгебраических уравнений, которые можно трактовать как операторные уравнения первого рода

$$Au = f \quad (1)$$

в вещественном гильбертовом пространстве H . В некоторых частных случаях такие системы могут быть эффективно решены прямыми методами, изученными в главах I—IV. В общем случае одним из приближенных методов решения сеточных эллиптических уравнений является метод итераций. Изучение итерационных методов начнем с простейших двухслойных методов — чебышевского метода и метода простой итерации.

Для приближенного решения уравнения (1) с линейным невырожденным оператором A , заданным в H , рассмотрим *неявную двухслойную итерационную схему*

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots \quad (2)$$

с произвольным начальным приближением $y_0 \in H$. Здесь $\{\tau_k\}$ — последовательность итерационных параметров, а B — произвольный линейный невырожденный оператор, действующий в H . Вопрос о наилучшем выборе оператора B будет изучен отдельно,

здесь же только отметим, что оператор B должен быть легко обратимым.

Сходимость итерационной схемы (2) будем исследовать в энергетическом пространстве H_D , порождаемом произвольным самосопряженным положительно определенным в H оператором D .

Так как оператор B не фиксирован, то (2) порождает семейство итерационных схем, которое будем называть *исходным семейством*.

В главе V было показано, что для изучения сходимости итерационного метода необходимо исследовать поведение нормы в H_D погрешности $z_k = y_k - u$ при $k \rightarrow \infty$, где y_k — итерационное приближение, получаемое по схеме (2), а u — решение уравнения (1). Итерационный метод сходится в H_D , если норма погрешности z_k в H_D стремится к нулю, когда k стремится в бесконечность.

Так как скорость сходимости зависит от выбора итерационных параметров τ_k , то их следует выбирать так, чтобы скорость сходимости была максимальной.

2. Задача для погрешности. Исследуем сначала сходимость двухслойных итерационных схем (2). Для этого получим уравнение, которому удовлетворяет погрешность z_k .

Подставляя $y_k = z_k + u$ для $k = 0, 1, \dots$ в (2) и учитывая уравнение (1), найдем

$$B \frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \dots, z_0 = y_0 - u,$$

т. е. погрешность z_k удовлетворяет однородному уравнению. Разрешая это уравнение относительно z_{k+1} :

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A) z_k$$

и полагая $z_k = D^{-1/2} x_k$, перейдем к уравнению для эквивалентной погрешности x_k , которое будет содержать один оператор. Уравнение для x_k будет иметь вид

$$x_{k+1} = S_{k+1} x_k, \quad S_{k+1} = E - \tau_{k+1} C, \quad k = 0, 1, \dots, \quad (3)$$

где $C = D^{1/2} B^{-1} A D^{-1/2}$. В силу сделанной замены справедливо равенство

$$\|x_k\| = \|D^{1/2} z_k\| = \|z_k\|_D,$$

поэтому задача исследования сходимости итерационного метода (2) в H_D сводится к изучению числовой последовательности $\|x_k\|$, $k = 1, 2, \dots$, где x_k определено в (3).

Найдем решение уравнения (3). Из (3) получим

$$x_k = T_{k, 0} x_0, \quad T_{k, 0} = \prod_{i=1}^k S_i = S_n S_{n-1} \dots S_1.$$

Отсюда вытекает следующая оценка для нормы погрешности z_k в H_D :

$$\|z_k\|_D = \|x_k\| \leq \|T_{k,0}\| \|x_0\| = \|T_{k,0}\| \|z_0\|_D. \quad (4)$$

Оператор $T_{k,0}$ называется *разрешающим оператором* для k -й итерации, а S_k — *оператором перехода от $(k-1)$ -й итерации к k -й итерации*.

Из оценки (4) следует, что итерационный метод (2) сходится в H_D , если норма разрешающего оператора $T_{k,0}$ стремится к нулю, когда k стремится к бесконечности.

Таким образом, задача исследования сходимости итерационной схемы (2) в H_D сведена к изучению поведения нормы разрешающего оператора $T_{k,0}$ в пространстве H в зависимости от номера итерации k .

Разрешающий оператор $T_{k,0}$ определяется оператором C и итерационными параметрами $\tau_1, \tau_2, \dots, \tau_k$.

Считая оператор C фиксированным, поставим задачу выбрать параметры $\{\tau_k\}$ так, чтобы итерационный метод сходился. Среди сходящихся итерационных методов *оптимальным* методом будет, очевидно, тот, параметры τ_k которого обеспечивают достижение заданной точности $\varepsilon > 0$ за минимальное число итераций. Этому требованию в силу оценки (4) можно придать следующую эквивалентную форму: для заданного n построить набор итерационных параметров $\tau_1, \tau_2, \dots, \tau_n$, для которого норма оператора $T_{n,0}$ была бы минимальной.

3. Самосопряженный случай. Дадим теперь строгую постановку задачи о наилучшем выборе итерационных параметров для двухслойной схемы (2). Эта задача будет иметь решение при определенных предположениях относительно операторов A , B и D . Сформулируем эти предположения.

1) Будем предполагать, что операторы A , B и D таковы, что оператор $DB^{-1}A$ самосопряжен в H . Если это предположение выполнено, то будем говорить, что рассматривается самосопряженный случай.

2) Пусть заданы γ_1 и γ_2 — постоянные энергетической эквивалентности операторов D и $DB^{-1}A$, т. е. постоянные из неравенств

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \quad (5)$$

Второе предположение определяет тип априорной информации об операторах итерационной схемы; эта информация используется при построении формул для итерационных параметров в самосопряженном случае. Простейший пример, для которого предположение о самосопряженности оператора $DB^{-1}A$ выполняется, следующий: $A = A^*$, $D = B = E$, т. е. рассматривается явная схема в исходном пространстве H для уравнения (1) с самосопряженным оператором A . В этом случае априорная ин-

формация состоит в задании границ оператора A . Более сложные примеры выбора оператора D будут рассмотрены ниже.

Итак, пусть выполнены условия (5). Из (5) следует, что оператор $C = D^{-1/2} (DB^{-1}A) D^{-1/2}$ самосопряжен в H , а γ_1 и γ_2 — его границы, т. е.

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \gamma_1 > 0, \quad C = C^* = D^{-1/2} (DB^{-1}A) D^{-1/2}. \quad (6)$$

Действительно, полагая в неравенствах

$$\gamma_1 (Dx, x) \leq (DB^{-1}Ax, x) \leq \gamma_2 (Dx, x)$$

$x = D^{-1/2}y$, получим неравенства (6). Таким образом, сделанные выше предположения относительно операторов A , B и D эквивалентны условиям (6).

Сформулируем теперь задачу об оптимальном выборе итерационных параметров для схемы (2). Из определения разрешающего оператора $T_{k,0}$ и условий (6) следует, что оператор $T_{k,0} = T_{k,0}(C)$ самосопряжен в H и норма операторного полинома $T_{n,0}(C)$ оценивается следующим образом:

$$\|T_{n,0}\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} \left| \prod_{k=1}^n (1 - \tau_k t) \right|.$$

Из оценки (4) следует, что в самосопряженном случае итерационные параметры $\tau_1, \tau_2, \dots, \tau_n$ должны быть выбраны так,

чтобы максимум модуля полинома $P_n(t) = \prod_{k=1}^n (1 - \tau_k t)$, построенного по этим параметрам, на отрезке $[\gamma_1, \gamma_2]$ был минимальным, т. е. нужно найти параметры из условия

$$\min_{\{\tau_k\}} \max_{\gamma_1 \leq t \leq \gamma_2} \left| \prod_{k=1}^n (1 - \tau_k t) \right| = \max_{\gamma_1 \leq t \leq \gamma_2} |P_n(t)|.$$

Тогда для погрешности метода (2) будет верна оценка $\|z_n\|_D \leq q_n \|z_0\|_D$, где

$$q_n = \max_{\gamma_1 \leq t \leq \gamma_2} |P_n(t)|.$$

Сформулированная выше задача является классической задачей минимакса. В § 2 будет приведено решение этой задачи и будет построен набор итерационных параметров $\tau_1, \tau_2, \dots, \tau_n$. Итерационный метод с этим набором параметров называется *чебышевским методом*. В литературе этот метод называют также *методом Ричардсона*.

§ 2. Чебышевский двухслойный метод

1. Построение набора итерационных параметров. В § 1 было показано, что построение оптимального набора итерационных параметров $\tau_1, \tau_2, \dots, \tau_n$ сводится к нахождению полинома

$P_n(t)$ вида $P_n(t) = \prod_{k=1}^n (1 - \tau_k t)$, максимум модуля которого на отрезке $[\gamma_1, \gamma_2]$ минимален.

Решим эту задачу. Так как вид полинома определяется условием нормировки $P_n(0) = 1$, то указанная задача формулируется следующим образом: среди всех полиномов степени n , принимающих в точке $t = 0$ значение 1, найти полином, наименее уклоняющийся от нуля на отрезке $[\gamma_1, \gamma_2]$, не содержащем точку 0.

Решение этой задачи было получено русским математиком В. А. Марковым в 1892 г. и приведено в дополнении. Искомый полином $P_n(t)$ имеет вид

$$P_n(t) = q_n T_n\left(\frac{1-\tau_0 t}{\rho_0}\right), \quad q_n = \frac{1}{T_n\left(\frac{1}{\rho_0}\right)}, \quad (1)$$

где $T_n(x)$ — полином Чебышева первого рода степени n ,

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leq 1, \\ \operatorname{ch}(n \operatorname{Arch} x), & |x| \geq 1, \end{cases}$$

$$q_n = \frac{2\rho_0^n}{1+\rho_0^{2n}}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \quad (2)$$

При этом $\max_{\gamma_1 \leq t \leq \gamma_2} |P_n(t)| = q_n$. Отсюда следует оценка для нормы погрешности z_n в H_D :

$$\|z_n\|_D \leq q_n \|z_0\|_D, \quad (3)$$

где q_n определено в (2).

Получим формулы для итерационных параметров. Так как полиномы, стоящие в левой и правой частях (1), принимают при $t = 0$ одно и то же значение, равное 1, то тождество в (1) будет иметь место лишь в том случае, когда множества корней полиномов $P_n(t)$ и $T_n\left(\frac{1-\tau_0 t}{\rho_0}\right)$ совпадают. Полином $P_n(t)$ имеет корни $1/\tau_k$, $k = 1, 2, \dots, n$, а полином $T_n(x)$ имеет корни, равные $-\cos\left(\frac{2i-1}{2n}\pi\right)$, $i = 1, 2, \dots, n$. Если обозначить через \mathfrak{M}_n множество корней полинома Чебышева $T_n(x)$:

$$\mathfrak{M}_n = \left\{ -\cos \frac{2i-1}{2n}\pi, \quad i = 1, 2, \dots, n \right\}, \quad (4)$$

то получим следующую формулу для итерационных параметров:

$$\tau_k = \tau_0 / (1 + \rho_0 \mu_k), \quad \mu_k \in \mathfrak{M}_n, \quad k = 1, 2, \dots, n. \quad (5)$$

Здесь $\mu_k \in \mathfrak{M}_n$ означает, что в качестве μ_k должны выбираться последовательно все элементы множества \mathfrak{M}_n .

Из полученной формулы для параметров τ_k видно, что для вычисления итерационных параметров требуется задать число

итераций n . Поэтому оценим число итераций. Обычно в качестве условия окончания процесса итераций берется неравенство

$$\|z_n\|_D \leq \varepsilon \|z_0\|_D$$

и *числом итераций* называют наименьшее целое n , для которого это неравенство выполняется.

Из (3) следует, что для рассматриваемого метода число итераций находится из неравенства $q_n \leq \varepsilon$. Используя (2), решим это неравенство. Получим

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \ln \left(\frac{1}{\varepsilon} + \sqrt{\frac{1}{\varepsilon^2} - 1} \right) / \ln \frac{1}{\rho_1}.$$

Обычно используют более простую формулу для $n_0(\varepsilon)$

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \ln \frac{2}{\varepsilon} / \ln \frac{1}{\rho_1}. \quad (6)$$

После того как найдено требуемое число итераций n , по формулам (5) можно построить набор итерационных параметров.

Итак, для неявной двухслойной схемы

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (7)$$

доказана

Теорема 1. Пусть выполнены условия

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*, \quad D = D^* > 0. \quad (8)$$

Тогда чебышевский итерационный процесс (7), (4), (5), (2) сходится в H_D и для погрешности z_n имеет место оценка (3). Для числа итераций справедлива оценка (6).

Из полученных оценок следует, что в самосопряженном случае скорость сходимости чебышевского метода зависит от отношения $\xi = \gamma_1/\gamma_2$, причем скорость сходимости будет тем выше, чем больше ξ .

2. О неулучшаемости априорной оценки. Покажем теперь, что на классе произвольных начальных приближений y_0 оценка для погрешности чебышевского метода, полученная в теореме 1, является неулучшаемой в случае конечномерного пространства H . Достаточно указать такое начальное приближение y_0 , при котором для нормы эквивалентной погрешности x_k будет иметь место равенство $\|x_n\| = q_n \|x_0\|$. Мы найдем начальную погрешность x_0 , которая обеспечивает выполнение этого равенства, а начальное приближение y_0 в силу связи между погрешностью z_k и x_k , $z_k = D^{-1/2}x_k$, определится тогда по формуле $y_0 = u + D^{-1/2}x_0$.

Найдем искомое x_0 . Пусть H — конечномерное пространство ($H = H_N$). Так как оператор C самосопряжен в H , то существует полная система собственных функций v_1, v_2, \dots, v_N оператора C . Обозначим через λ_k собственное значение оператора C , соответствующее собственной функции v_k . Пусть собственные значения

упорядочены $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$. Тогда в качестве границ оператора C можно взять $\gamma_1 = \lambda_1$ и $\gamma_2 = \lambda_N$.

В качестве начальной погрешности x_0 возьмем собственную функцию v_1 . Из уравнения для погрешности x_k :

$$x_{k+1} = (E - \tau_{k+1}C)x_k, \quad k = 0, 1, \dots, x_0 = v_1$$

и равенства $Cv_k = \lambda_k v_k$ последовательно получим

$$x_1 = (E - \tau_1 C)x_0 = (1 - \tau_1 \gamma_1)v_1 = (1 - \tau_1 \gamma_1)x_0,$$

$$x_2 = (E - \tau_2 C)x_1 = (1 - \tau_1 \gamma_1)(E - \tau_2 C)x_0 = (1 - \tau_1 \gamma_1)(1 - \tau_2 \gamma_1)x_0,$$

• •

$$x_n = \prod_{k=1}^n (1 - \tau_k \gamma_1)x_0 = P_n(\gamma_1)(x_0).$$

Подставляя в (1) $t = \gamma_1$ и учитывая равенство $1 - \tau_0 \gamma_1 = \rho_0$, вычислим $P_n(\gamma_1) = q_n T_n(1) = q_n$ и, следовательно,

$$x_n = q_n x_0, \quad \|x_n\| = q_n \|x_0\|,$$

что и требовалось доказать.

Итак, показано, что полученная в теореме 1 априорная оценка неулучшаема на классе произвольных начальных приближений.

3. Примеры выбора оператора D . Приведем некоторые примеры выбора оператора D . Напомним, что чебышевский метод рассматривается в предположении самосопряженности оператора $DB^{-1}A$. Ниже будут указаны требования на операторы A и B , при которых это предположение выполняется для выбранного оператора D . Для каждого конкретного выбора оператора D будут приведены неравенства, задающие априорную информацию об операторах итерационной схемы. Эта информация используется для построения набора итерационных параметров в чебышевском методе.

Рассмотрим первый пример. Пусть операторы A и B самосопряжены и положительно определены в H . Тогда в качестве оператора D можно взять один из следующих операторов: A или B . Если, кроме того, оператор B ограничен в H , то можно взять $D = AB^{-1}A$. При этом априорная информация сводится к заданию постоянных энергетической эквивалентности операторов A и B :

$$\gamma_1 \leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \quad B > 0. \quad (9)$$

Действительно, нужно показать, что выполнены следующие условия: выбранный оператор D самосопряжен и положительно определен в H , оператор $DB^{-1}A$ самосопряжен в H , а неравенства (8) и (9) эквивалентны.

Самосопряженность операторов D и $DB^{-1}A$ для всех рассматриваемых случаев следует из самосопряженности операторов A и B . Для случая, когда $D = A$ или $D = B$, положительная определенность D вытекает из положительной определенности

операторов A и B . Покажем теперь, что оператор $D = AB^{-1}A$ также положительно определен в H .

Действительно, пусть выполнены сформулированные выше условия на операторы A и B : $A = A^* \geqslant \alpha E$, $B = B^* \geqslant \beta E$, $\|Bx\| \leqslant M\|x\|$, $\alpha, \beta > 0$, $M < \infty$. Из этих условий и лемм 6 и 8 из § 1 гл. V получим, что $B^{-1} \geqslant \frac{1}{M}E$ и $(Ax, Ax) \geqslant \alpha(Ax, x) \geqslant \geqslant \alpha^2(x, x)$. Отсюда найдем для энергии оператора D оценку снизу

$$\begin{aligned}(Dx, x) &= (AB^{-1}Ax, x) = (B^{-1}Ax, Ax) \geqslant \\ &\geqslant \frac{1}{M}(Ax, Ax) \geqslant \frac{\alpha^2}{M}(x, x), \quad \text{т. е. } D \geqslant \frac{\alpha^2}{M}E.\end{aligned}$$

Следовательно, положительная определенность оператора $D = AB^{-1}A$ доказана.

Покажем теперь, что неравенства (8) и (9) эквивалентны для рассматриваемого примера. Действительно, пусть выполнены неравенства (9):

$$\gamma_1(Bx, x) \leqslant (Ax, x) \leqslant \gamma_2(Bx, x), \quad \gamma_1 > 0. \quad (10)$$

Если $D = B$, то $DB^{-1}A = A$, и следовательно, неравенства (10) и (8) совпадают. Пусть теперь $D = AB^{-1}A$. В этом случае $DB^{-1}A = AB^{-1}AB^{-1}A$ и, полагая в (10) $x = B^{-1}Ay$, получим

$$\gamma_1(AB^{-1}Ay, y) \leqslant (AB^{-1}Ay, B^{-1}Ay) \leqslant \gamma_2(AB^{-1}Ay, y)$$

или

$$\gamma_1(Dy, y) \leqslant (DB^{-1}Ay, y) \leqslant \gamma_2(Dy, y),$$

т. е. получим неравенства (8). Обратный переход от (8) к (10) очевиден.

Пусть $D = A$, тогда $DB^{-1}A = AB^{-1}A$. Из леммы 9 § 1 гл. V следует, что для самосопряженных и положительно определенных операторов A и B неравенства (10) и неравенства

$$\gamma_1(A^{-1}x, x) \leqslant (B^{-1}x, x) \leqslant \gamma_2(A^{-1}x, x), \quad \gamma_1 > 0$$

эквивалентны. Полагая здесь $x = Ay$, получим неравенства (8). Обратный переход очевиден.

Это неравенство позволяет сразу доказать положительную определенность D :

$$(Dx, x) \geqslant \alpha\gamma_1(x, x).$$

В самом деле, $(Dx, x) = (B^{-1}Ax, Ax) \geqslant \gamma_1(A^{-1}Ax, Ax) = \gamma_1(Ax, x) \geqslant \gamma_1\alpha(x, x)$.

Второй пример. Пусть операторы A и B самосопряжены, положительно определены в H и перестановочны: $A = A^* > 0$, $B = B^* > 0$, $AB = BA$. Если в качестве оператора D взять оператор A^2 , то априорная информация может быть задана в виде неравенств (9).

Действительно, самосопряженность и положительная определенность оператора D следуют из самосопряженности и невырожденности оператора A . Далее, $DB^{-1}A = A(AB^{-1})A$, а так как операторы A и B перестановочны, то перестановочные и операторы A и B^{-1} . Отсюда и из самосопряженности операторов A и B следует самосопряженность оператора $DB^{-1}A$.

Неравенства (8) в данном случае имеют вид

$$\gamma_1(Ax, Ax) \leq (AB^{-1}Ax, Ax) \leq \gamma_2(Ax, Ax), \quad \gamma_1 > 0.$$

Полагая здесь $x = A^{-1}B^{1/2}y$ и используя перестановочность корня из оператора B с оператором A , найдем

$$\gamma_1(By, y) \leq (Ay, y) \leq \gamma_2(By, y),$$

т. е. получим неравенство (9). Обратный переход от (9) к (8) очевиден.

Рассмотрим еще один пример. Пусть A и B —произвольные невырожденные операторы, удовлетворяющие условию

$$B^*A = A^*B. \quad (11)$$

Если в качестве D выбрать оператор A^*A , то априорная информация может быть задана в виде неравенств

$$\gamma_1(Bx, Bx) \leq (Ax, Bx) \leq \gamma_2(Bx, Bx), \quad \gamma_1 > 0. \quad (12)$$

Самосопряженность оператора D очевидна, а положительная определенность следует из невырожденности оператора A . Так как оператор B невырожден, то условия (11) могут быть записаны в виде условий $AB^{-1} = (B^*)^{-1}A^*$, которые выражают самосопряженность оператора AB^{-1} . Отсюда получим, что оператор $DB^{-1}A = A^*AB^{-1}A$ самосопряжен в H . Далее, полагая в (12) $x = B^{-1}Ay$, получим

$$\gamma_1(Ay, Ay) \leq (AB^{-1}Ay, Ay) \leq \gamma_2(Ay, Ay)$$

или

$$\gamma_1(Dy, y) \leq (DB^{-1}Ay, y) \leq \gamma_2(Dy, y).$$

Таким образом, из неравенств (12) следуют неравенства (8). Обратный переход от (8) к (12) очевиден.

В заключение отметим, что для случая самосопряженных положительно определенных и ограниченных в H операторов A и B чебышевский итерационный метод сходится в H_D , где $D = A$, B или $AB^{-1}A$ (а если, кроме того, A и B перестановочны, то и для $D = A^2$), с одинаковой скоростью, определяемой отношением постоянных γ_1 и γ_2 из неравенств (9).

Особо отметим случаи $D = AB^{-1}A$ и $D = A^*A$. При таком выборе оператора D норма погрешности в H_D может быть вычислена в процессе итераций. Действительно, для $D = AB^{-1}A$ получим

$$\|z_n\|_D^2 = (Dz_n, z_n) = (B^{-1}Az_n, Az_n) = (B^{-1}r_n, r_n) = (w_n, r_n),$$

а для $D = A^*A$:

$$\|z_n\|_D^2 = (Az_n, Az_n) = (r_n, r_n),$$

где $r_n = Az_n = Ay_n - Au = Ay_n - f$ — невязка n -й итерации, а $w_n = B^{-1}r_n$ — поправка. Эти величины можно найти в процессе итераций.

4. О вычислительной устойчивости метода. При изучении сходимости чебышевского метода предполагалось, что вычислительный процесс является идеальным, т. е. вычисления ведутся с бесконечным числом знаков. В реальном вычислительном процессе все вычисления осуществляются с конечным числом знаков, и на каждом этапе счета появляются ошибки округления. Ошибки округления результатов арифметических операций порождают вычислительную погрешность метода.

В итерационных методах вычислительная погрешность метода образуется из погрешностей, допускаемых на каждой итерации. Если число итераций достаточно велико, а итерационный метод обладает свойством накапливать погрешности округления каждого итерационного шага, то вычислительная погрешность такого метода может оказаться настолько большой, что произойдет полная потеря точности и итерационное приближение y_n будет сильно отличаться от искомого решения. Поэтому для итерационных методов важно изучить механизм возникновения вычислительной погрешности и найти те этапы алгоритма, на которых происходит рост вычислительной погрешности метода. В ряде случаев некоторые изменения в процессе вычислений позволяют существенно уменьшить рост вычислительной погрешности и сделать метод пригодным для практического использования.

Другая особенность реального вычислительного процесса связана с наличием на ЭВМ «машинного нуля» и «машинной бесконечности». Эти понятия характеризуют допустимый порядок чисел, которые могут быть представлены в ЭВМ. Например, в ЭВМ БЭСМ-6 в режиме однократной точности могут быть представлены действительные числа, абсолютная величина которых принадлежит диапазону от 10^{-19} до 10^{19} . Это и есть границы для «машинного нуля» и «машинной бесконечности». Если в результате вычислений на ЭВМ появляется величина, не принадлежащая этому интервалу, то вычисления прекращаются и происходит так называемый «аварийный останов» (авост). Поэтому требование «безаварийности» итерационного процесса является естественным.

Итак, итерационные методы должны быть «безаварийными» и устойчивыми по отношению к ошибкам округления.

В п. 1 § 2 был построен чебышевский двухслойный метод. В теореме 1 доказано, что если выполнить n итераций с параметрами $\tau_k = \tau_0 / (1 + \rho_0 \mu_k)$, $\mu_k \in \mathfrak{M}_n$, $k = 1, 2, \dots, n$, то для погрешности z_n будет справедлива оценка $\|z_n\|_D \leq q_n \|z_0\|_D$. В

качестве μ_k выбираются последовательно все элементы множества \mathfrak{M}_n , причем порядок выбора произвольный.

Изучим вычислительную устойчивость чебышевского метода. Будем для определенности считать, что μ_k есть k -й элемент множества \mathfrak{M}_n . Тогда различные упорядочения множества \mathfrak{M}_n будут порождать различные последовательности $\{\mu_k\}$ и, следовательно, различные последовательности итерационных параметров $\{\tau_k\}$.

С точки зрения идеального вычислительного процесса все последовательности чебышевских итерационных параметров эквивалентны, т. е. каждая последовательность должна обеспечивать получение одного и того же приближения y_n и, следовательно, одной точности после выполнения n итераций. Наличие ошибок округления в реальном вычислительном процессе приводит к неэквивалентности последовательностей итерационных параметров.

Проиллюстрируем это утверждение на примере. Пусть на сетке $\bar{\omega} = \{x_i = ih, 0 \leq i \leq N, h = 1/N\}$ требуется найти решение следующей разностной задачи:

$$\begin{aligned} \Lambda y &= y_{xx} - dy = -\varphi(x), \quad x \in \omega, \\ y(0) &= 0, \quad y(1) = 1, \quad d = \text{const} > 0. \end{aligned}$$

В § 2 гл. V было показано, что разностная задача может быть сведена к операторному уравнению

$$Ay = f, \quad (13)$$

оператор A в котором определяется следующим образом:

$Ay = -\Lambda y$, где $y \in H$, $\dot{y} \in \dot{H}$, $\dot{y}(x) = y(x)$ для $x \in \omega$. Здесь \dot{H} — множество сеточных функций, заданных на $\bar{\omega}$ и обращающихся в нуль при $x = 0$ и $x = 1$, а H — пространство сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{x \in \omega} u(x)v(x)h$.

Правая часть f уравнения (13) отличается от правой части φ разностной схемы лишь в приграничных узлах сетки: $f(x) = \varphi(x)$, $h \leq x \leq 1 - 2h$, $f(1-h) = \varphi(1-h) + 1/h^2$.

Для приближенного решения уравнения (13) рассмотрим явный чебышевский метод

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H. \quad (14)$$

Так как операторы A и $B = E$ самосопряжены в H , то из рассмотренных в п. 3 § 2 примеров следует, что в качестве априорной информации для чебышевского метода (14) достаточно задать границы оператора A : $\gamma_1 E \leq A \leq \gamma_2 E$, $\gamma_1 > 0$, если в качестве оператора D взять оператор $B = E$. Очевидно, что γ_1 и γ_2 совпадают с минимальным и максимальным собственными значениями

разностного оператора Λ , т. е.

$$\gamma_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2} + d, \quad \gamma_2 = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} + d.$$

Итерационные параметры τ_k вычисляются по формулам

$$\begin{aligned} \tau_k &= \tau_0 / (1 + \rho_0 \mu_k), \quad \mu_k \in \mathfrak{M}_n, \quad k = 1, 2, \dots, n, \\ \tau_0 &= 2 / (\gamma_1 + \gamma_2), \quad \rho_0 = (\gamma_2 - \gamma_1) / (\gamma_2 + \gamma_1). \end{aligned} \quad (15)$$

Рассматривались три последовательности итерационных параметров, определяемые следующими упорядочениями \mathfrak{M}_n :

1) «прямая» последовательность

$$\mathfrak{M}_n = \mathfrak{M}_n^{(1)} = \{\sigma_1, \sigma_2, \dots, \sigma_n\}, \quad \text{т. е. } \mu_k = \sigma_k, \quad k = 1, 2, \dots, n;$$

2) «обратная» последовательность

$$\mathfrak{M}_n = \mathfrak{M}_n^{(2)} = \{\sigma_n, \sigma_{n-1}, \dots, \sigma_1\}, \quad \text{т. е. } \mu_k = \sigma_{n-k+1}, \quad k = 1, 2, \dots, n;$$

3) «чередующаяся» последовательность

$$\begin{aligned} \mathfrak{M}_n = \mathfrak{M}_n^{(3)} &= \{\sigma_1, \sigma_n, \sigma_2, \sigma_{n-1}, \dots\}, \quad \text{т. е. } \mu_{2k-1} = \sigma_k, \\ \mu_{2k} &= \sigma_{n-k+1}, \quad k = 1, 2, \dots, n/2. \end{aligned}$$

Здесь обозначено $\sigma_k = -\cos \frac{2k-1}{2n} \pi$.

Вычисления проводились следующим образом: задавалось число итераций n и по схеме (14), (15) для каждой последовательности итерационных параметров проводилось n итераций. Реальная точность, которая достигалась после выполнения n итераций, определялась по формуле

$$\varepsilon_{\text{реал}} = \frac{\|y_n - u\|}{\|y_0 - u\|}.$$

Для сравнения вычислялось значение q_n , где

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

определенное теоретическую точность метода, когда число итераций равно n . Во всех расчетах начальное приближение y_0 бралось равным нулю на ω . Точное решение разностной задачи $y(x) = x$ соответствует правой части $\varphi(x) = dx$. Коэффициент d выбирался так, чтобы γ_1 было равно 0,1:

$$\gamma_1 = 0,1, \quad \gamma_2 = 0,1 + \frac{4}{h^2} \cos \pi h, \quad \frac{1}{\xi} = \frac{40}{h^2} \cos \pi h + 1.$$

Результаты вычислений для $N = 10$ приведены в табл. 5. В этой таблице, помимо указанных последовательностей параметров, приведены результаты для оптимального упорядочения множества \mathfrak{M}_n^* , которое будет описано ниже.

Таблица 5

n	q_n	$\varepsilon_{\text{реал}}$			
		$\mathfrak{M}_n^{(1)}$	$\mathfrak{M}_n^{(2)}$	$\mathfrak{M}_n^{(3)}$	\mathfrak{M}_n^*
16	$8,79 \cdot 10^{-1}$	$8,14 \cdot 10^{-1}$	$8,14 \cdot 10^{-1}$	$8,14 \cdot 10^{-1}$	$8,14 \cdot 10^{-1}$
24	$7,58 \cdot 10^{-1}$	$9,62 \cdot 10^{-1}$	$7,11 \cdot 10^{-1}$	$7,11 \cdot 10^{-1}$	$7,11 \cdot 10^{-1}$
32	$6,30 \cdot 10^{-1}$	$3,38 \cdot 10^3$	$3,55 \cdot 10^2$	$5,63 \cdot 10^{-1}$	$5,63 \cdot 10^{-1}$
40	$5,09 \cdot 10^{-1}$	$3,07 \cdot 10^7$	$2,44 \cdot 10^6$	$5,03 \cdot 10^{-1}$	$4,85 \cdot 10^{-1}$
48	$4,04 \cdot 10^{-1}$	авост	$3,46 \cdot 10^{10}$	$2,47 \cdot 10^0$	$3,64 \cdot 10^{-1}$
56	$3,17 \cdot 10^{-1}$	—	$1,02 \cdot 10^{15}$	$2,29 \cdot 10^2$	$3,10 \cdot 10^{-1}$
64	$2,47 \cdot 10^{-1}$	—	авост	$1,87 \cdot 10^4$	$2,23 \cdot 10^{-1}$
72	$1,92 \cdot 10^{-1}$	—	—	$1,73 \cdot 10^6$	$1,72 \cdot 10^{-1}$
80	$1,49 \cdot 10^{-1}$	—	—	авост	$1,44 \cdot 10^{-1}$
..
256	$4,97 \cdot 10^{-4}$	—	—	—	$4,80 \cdot 10^{-4}$
..
512	$1,23 \cdot 10^{-7}$	—	—	—	$1,15 \cdot 10^{-7}$

Проведенные расчеты показывают, что для реального вычислительного процесса рассмотренные последовательности итерационных параметров действительно не являются эквивалентными. Расчеты продемонстрировали две характерные особенности реального вычислительного процесса: возможность «авоста», вызываемого ростом промежуточных итерационных решений, и возможность потери окончательной точности в безавостной ситуации, вызываемой накоплением погрешностей округления.

Причиной такой вычислительной неустойчивости метода для некоторых последовательностей итерационных параметров является тот факт, что норма оператора перехода от итерации к итерации $S_k = E - \tau_k C$ для некоторых значений k больше единицы.

Действительно, так как S — самосопряженный в H оператор, то $\|S_k\| = \sup_{\|x\|=1} |(S_k x, x)|$. Используя границы γ_1, γ_2 оператора C

$$\gamma_1 E \leq S \leq \gamma_2 E, \quad \gamma_1 > 0,$$

найдем

$$(1 - \tau_k \gamma_2) E \leq S_k \leq (1 - \tau_k \gamma_1) E.$$

Подставим сюда τ_k из (15) и учтем равенства $1 - \rho_0 = \tau_0 \gamma_1$, $1 + \rho_0 = \tau_0 \gamma_2$. Получим

$$-\frac{\rho_0 (1 - \mu_k)}{1 + \rho_0 \mu_k} E \leq S_k \leq \frac{\rho_0 (1 + \mu_k)}{1 + \rho_0 \mu_k} E$$

и, следовательно,

$$\|S_k\| = \begin{cases} \frac{\rho_0 (1 + \mu_k)}{1 + \rho_0 \mu_k} < 1, & \mu_k \geq 0, \\ \frac{\rho_0 (1 - \mu_k)}{1 + \rho_0 \mu_k}, & \mu_k < 0. \end{cases}$$

Отсюда следует, что $\|S_k\| > 1$ при $\mu_k < -(1 - \rho_0)/(2\rho_0)$. Так как $\mu_k \in \mathfrak{M}_n$, то

$$-\cos \frac{\pi}{2n} \leq \mu_k \leq -\cos \frac{2n-1}{2n} \pi = \cos \frac{\pi}{2n}, \quad k = 1, 2, \dots, n,$$

и, следовательно, для большого числа номеров k норма $\|S_k\| > 1$ (число таких номеров k примерно равно $n/2$). Поэтому, если использовать подряд слишком много параметров τ_k , для которых норма оператора S_k больше единицы, то может произойти накопление вычислительной погрешности и рост итерационных приближений, что служит причиной вычислительной неустойчивости метода.

Теорема 1 фактически выражает устойчивость итерационной схемы по начальным данным. В случае реального вычислительного процесса необходимо исследовать устойчивость итерационной схемы и по правой части, поскольку ошибки округления можно трактовать как возмущение правой части итерационной схемы на каждой итерации.

Если учесть погрешность округления, то вместо однородного уравнения для эквивалентной погрешности x_k получим неоднородное уравнение

$$x_{k+1} = S_{k+1}x_k + \tau_{k+1}\varphi_{k+1}, \quad k = 0, 1, \dots \quad (16)$$

Здесь $x_k = D^{1/2}(\bar{y}_k - u)$, где \bar{y}_k — реальное итерационное приближение.

Решая уравнение (16), найдем $x_n = T_{n,0}x_0 + \sum_{j=1}^n \tau_j T_{n,j}\varphi_j$, где $T_{n,j} = \prod_{i=j+1}^n S_i$, $T_{n,n} = E$. Отсюда получим следующую оценку:

$$\|x_n\| \leq \|T_{n,0}\| \|x_0\| + \sum_{j=1}^n \tau_j \|T_{n,j}\| \max_{1 \leq j \leq n} \|\varphi_j\|. \quad (17)$$

Оценка нормы оператора $T_{n,0}$ не зависит от упорядочения множества \mathfrak{M}_n , и для любой последовательности чебышевских параметров τ_k имеем $\|T_{n,0}\| \leq q_n$. Оценка для $\sum_{j=1}^n \tau_j \|T_{n,j}\|$ зависит от упорядочения множества \mathfrak{M}_n . Из (17) следует, что множество \mathfrak{M}_n должно быть упорядочено так, чтобы указанная сумма принимала минимальное значение.

Следующая лемма указывает минимально возможное значение этой суммы.

Лемма 1. *Если γ_1 и γ_2 — точные границы оператора C , то для любого упорядочения множества \mathfrak{M}_n имеет место оценка*

$$\sum_{j=1}^n \tau_j \|T_{n,j}\| \geq \frac{1-q_n}{\gamma_1}.$$

Действительно, из определения оператора $T_{n,j}$ получим

$$\tau_j T_{n,j} = (T_{n,j} - T_{n,j-1}) C^{-1}, \quad \sum_{j=1}^n \tau_j T_{n,j} = (E - T_{n,0}) C^{-1}.$$

Так как

$$\|(E - T_{n,0}) C^{-1}\| = \left\| \sum_{j=1}^n \tau_j T_{n,j} \right\| \leq \sum_{j=1}^n \tau_j \|T_{n,j}\|,$$

то достаточно оценить норму оператора $(E - T_{n,0}) C^{-1}$. Этот оператор самосопряжен в H , и если γ_1 и γ_2 — границы оператора C , то

$$\begin{aligned} \|(E - T_{n,0}) C^{-1}\| &\leq \max_{\gamma_1 \leq t \leq \gamma_2} \left| \frac{1 - q_n T_n \left(\frac{1 - \tau_0 t}{\rho_0} \right)}{t} \right| = \\ &= \frac{1 - q_n T_n \left(\frac{1 - \tau_0 \gamma_1}{\rho_0} \right)}{\gamma_1} = \frac{1 - q_n}{\gamma_1}. \end{aligned}$$

Таким образом, показано, что для любого $x \in H$ имеет место оценка

$$\|(E - T_{n,0}) C^{-1} x\| \leq \frac{1 - q_n}{\gamma_1} \|x\|. \quad (18)$$

Так как γ_1 есть точная граница самосопряженного оператора C , то γ_1 совпадает с минимальным собственным значением оператора C . Подставляя в (18) вместо x собственную функцию, соответствующую минимальному собственному значению оператора C , получим, что в (18) достигается равенство. Следовательно, получена оценка $\|(E - T_{n,0}) C^{-1}\| = (1 - q_n)/\gamma_1$. Лемма доказана.

5. Построение оптимальной последовательности итерационных параметров*).

5.1. Случай $n = 2^p$. Порядок использования итерационных параметров τ_k в чебышевском методе существенно влияет на сходимость метода. Поэтому возникает задача построения наилучшей последовательности итерационных параметров, обеспечивающей минимальное влияние вычислительной погрешности метода. Так как последовательность параметров определяется упорядочением множества \mathfrak{M}_n , то необходимо построить оптимальное упорядочение множества \mathfrak{M}_n .

Приведем решение этой задачи. Пусть сначала число итераций есть степень 2: $n = 2^p$. Обозначим через θ_m множество, состоящее из m целых чисел:

$$\theta_m = \{\theta_1^{(m)}, \theta_2^{(m)}, \dots, \theta_m^{(m)}\}.$$

*) Способ упорядочения итерационных параметров дан по работам: см. Е. С. Николаев, А. А. Самарский (ЖВМ и МФ, 12, № 4, 1972) для любого n и [8] для $n = 2^p$.

Исходя из множества $\theta_1 = \{1\}$, построим множество θ_{2^p} по следующему правилу. Пусть множество θ_m построено. Тогда множество θ_{2m} определим по формулам

$$\theta_{2m} = \{\theta_{2i}^{(2m)} = 4m - \theta_i^{(m)}, \theta_{2i-1}^{(2m)} = \theta_i^{(m)}, i = 1, 2, \dots, m\}, \\ m = 1, 2, 4, \dots, 2^{p-1}. \quad (19)$$

Нетрудно убедиться, что множество θ_{2k} состоит из нечетных чисел от 1 до $2^{k+1}-1$.

Используя построенное множество θ_{2^p} , упорядочим множество \mathfrak{M}_{2^p} следующим образом:

$$\mathfrak{M}_n^* = \left\{ -\cos \beta_i, \beta_i = \frac{\pi}{2^n} \theta_i^{(n)}, i = 1, 2, \dots, n \right\}, n = 2^p. \quad (20)$$

Это и есть искомое упорядочение множества \mathfrak{M}_n в случае, когда $n = 2^p$. Для соответствующей этому упорядочению последовательности итерационных параметров доказана оценка

$$\sum_{j=1}^n \tau_j \|T_{n,j}\| \leq \frac{1-q_n}{\gamma_1}.$$

Сравнивая эту оценку с оценкой леммы 1, убеждаемся, что построенное упорядочение множества \mathfrak{M}_n^* действительно обеспечивает минимальное влияние вычислительной погрешности на сходимость чебышевского метода.

Приведем некоторые примеры построения множества θ_n .

1) $n = 8$.

$$\theta_1 = \{1\}, \theta_2 = \{1, 3\}, \theta_4 = \{1, 7, 3, 5\}, \\ \theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\}.$$

Множество θ_8 построено. По формуле (20) упорядочивается множество \mathfrak{M}_8^* .

2) $n = 16$.

Используя найденное выше множество θ_8 , построим по формулам (19) множество θ_{16} :

$$\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}.$$

3) $n = 32$.

$\theta_{32} = \{1, 63, 31, 33, 15, 49, 17, 47, 7, 57, 25, 39, 9, 55, 23, 41, 3, 61, 29, 35, 13, 51, 19, 45, 5, 59, 27, 37, 11, 53, 21, 43\}$.

Из формул (19) следует простое правило перехода от множества θ_m к множеству θ_{2m} : $\theta_{2i-1}^{(2m)} = \theta_i^{(m)}$ и сумма двух соседних чисел равна $4m$:

$$\theta_{2i-1}^{(2m)} + \theta_{2i}^{(2m)} = 4m, \quad i = 1, 2, \dots, m.$$

Аналогичное правило перехода применяется и в общем случае, к рассмотрению которого мы переходим.

5.2. Общий случай. Пусть число итераций n есть любое целое число. Опишем процесс построения множества θ_n . Элементарными этапами этого процесса являются переходы от

множества θ_m к множеству θ_{2m} и от множества θ_{2m} к множеству θ_{2m+1} , где m — произвольное целое число.

Сформулируем правила перехода от множества к множеству.

1) Переход от θ_{2m} к θ_{2m+1} состоит в добавлении к элементам множества θ_{2m} нечетного числа $2m+1$.

2) Переход от θ_m к θ_{2m} осуществляется следующим образом. Если за этим переходом следует переход от θ_{2m} к θ_{4m} или переход от θ_m к θ_{2m} есть последний шаг в процессе построения множества θ_n , то используются формулы, приведенные выше:

$$\theta_{2i-1}^{(2m)} = \theta_i^{(m)}, \quad \theta_{2i-1}^{(2m)} + \theta_{2i}^{(2m)} = 4m, \quad i = 1, 2, \dots, m. \quad (21)$$

Если за переходом от θ_m к θ_{2m} следует переход от θ_{2m} к θ_{2m+1} , то используются формулы

$$\theta_{2i-1}^{(2m)} = \theta_i^{(m)}, \quad \theta_{2i-1}^{(2m)} + \theta_{2i}^{(2m)} = 4m + 2, \quad i = 1, 2, \dots, m. \quad (22)$$

Используя эти правила и чередуя должным образом переходы от множества с четным числом элементов к множеству с нечетным числом элементов и от множества из m элементов к множеству из $2m$ элементов, можно, исходя из $\theta_1 = \{1\}$, построить множество θ_n для любого n .

Приведем некоторые примеры.

1) $n=15$. В этом случае переход от θ_1 к θ_n осуществляется по следующей цепочке:

$$\theta_1 \rightarrow \theta_2 \rightarrow \theta_3 \rightarrow \theta_6 \rightarrow \theta_7 \rightarrow \theta_{14} \rightarrow \theta_{15}.$$

Согласно изложенным правилам переходы от θ_1 к θ_2 , от θ_3 к θ_6 и от θ_7 к θ_{14} осуществляются по формулам (22), а при переходе от θ_2 к θ_3 , от θ_6 к θ_7 и от θ_{14} к θ_{15} нужно добавить соответствующее нечетное число к исходному множеству. Это дает:

$$\theta_1 = \{1\}, \quad \theta_2 = \{1, 5\}, \quad \theta_3 = \{1, 5, 3\}.$$

$$\theta_6 = \{1, 13, 5, 9, 3, 11\}, \quad \theta_7 = \{1, 13, 5, 9, 3, 11, 7\},$$

$$\theta_{14} = \{1, 29, 13, 17, 5, 25, 9, 21, 3, 27, 11, 19, 7, 23\},$$

$$\theta_{15} = \{1, 29, 13, 17, 5, 25, 9, 21, 3, 27, 11, 19, 7, 23, 15\}.$$

Множество \mathfrak{M}_{15}^* упорядочивается по формуле (20).

2) $n=25$. Этому случаю соответствует цепочка

$$\theta_1 \rightarrow \theta_2 \rightarrow \theta_3 \rightarrow \theta_6 \rightarrow \theta_{12} \rightarrow \theta_{24} \rightarrow \theta_{25},$$

и переходы от θ_1 к θ_2 и от θ_{12} к θ_{24} выполняются по формулам (22), переходы от θ_3 к θ_6 и от θ_6 к θ_{12} — по формулам (21), переходы от θ_2 к θ_3 и от θ_{24} к θ_{25} осуществляются добавлением нечетного числа. Получим

$$\theta_1 = \{1\}, \quad \theta_2 = \{1, 5\}, \quad \theta_3 = \{1, 5, 3\}, \quad \theta_6 = \{1, 11, 5, 7, 3, 9\},$$

$$\theta_{12} = \{1, 23, 11, 13, 5, 19, 7, 17, 3, 21, 9, 15\},$$

$$\theta_{24} = \{1, 49, 23, 27, 11, 39, 13, 37, 5, 45, 19, 31, 7, 43, 17,$$

$$33, 3, 47, 21, 29, 9, 41, 15, 35\},$$

$$\Theta_{25} = \{1, 49, 23, 27, 11, 39, 13, 37, 5, 45, 19, 31, 7, 43, 17, 33, 3, 47, 21, 29, 9, 41, 15, 35, 25\}.$$

Изложенная выше процедура построения множества Θ_n для произвольного n может быть формализована. Для этого представим n в виде разложения по степеням 2 с целыми показателями k_j :

$$n = 2^{k_1} + 2^{k_2} + \dots + 2^{k_s}, \quad k_j \leq k_{j-1} - 1, \quad j = 2, 3, \dots, s.$$

Образуем следующие величины:

$$n_j = \sum_{i=1}^j 2^{k_i - k_j}, \quad j = 1, 2, \dots, s,$$

и положим $n_{s+1} = 2n + 1$. По формулам (23) строим множество Θ_{n_j} :

$$\Theta_{n_j} = \left\{ \theta_i^{(n_j)} = \theta_i^{(n_j-1)}, \theta_{n_j}^{(n_j)} = n_j, i = 1, 2, \dots, n_j - 1 \right\}, \quad (23)$$

для $j = 1$ выбираем $\Theta_1 = \{1\}$. Затем по формуле (24) строятся множества

$$\Theta_{2m} = \{\theta_{2i}^{(2m)} = 4m - \theta_i^{(m)}, \theta_{2i-1}^{(2m)} = \theta_i^{(m)}, i = 1, 2, \dots, m\} \quad (24)$$

для $m = n_j, 2n_j, 4n_j, \dots, [(n_{j+1}-1)/4]$, где $[a]$ —целая часть a . Если $[(n_{j+1}-1)/4] < n_j$, то вычисления по формуле (24) не проводятся, выполняется переход к следующему этапу. Если $j = s$, то необходимое множество Θ_n уже построено. Иначе полагаем $m = (n_{j+1}-1)/2$ и строим множество

$$\Theta_{2m} = \{\theta_{2i}^{(2m)} = 4m + 2 - \theta_i^{(m)}, \theta_{2i-1}^{(2m)} = \theta_i^{(m)}, i = 1, 2, \dots, m\}. \quad (25)$$

Затем j увеличивается на единицу и процесс повторяется, начиная с формулы (23). В результате будет построено множество Θ_n . Множество Θ_n^* упорядочивается согласно формуле (20).

Для случая $n = 2^p$ алгоритм (23)–(25) упрощается и переходит в алгоритм, описываемый формулой (19). Действительно, для $n = 2^p$ получим $s = 1$, $k_1 = p$, $n_1 = 1$, $n_{s+1} = 2^{p+1} - 1$. Следовательно, в алгоритме (23)–(25) j принимает единственное значение, равное единице, и вычисления ведутся по формуле (24) для $m = 1, 2, 4, \dots, 2^{p-1}$.

Проиллюстрируем качество построенного здесь упорядочения множества Θ_n^* на примере, рассмотренном в п. 4 § 2. Задаваемое число итераций n изменялось от 16 до 512 с шагом 8. Для каждого n реальная точность, достигнутая после выполнения n итераций, не превосходила теоретическую точность q_n ($q_{512} = 1,23 \cdot 10^{-7}$), и процесс был «безавостным» (см. табл. 5).

§ 3. Метод простой итерации

1. Выбор итерационного параметра. В § 2 была решена задача о построении оптимального набора итерационных параметров τ_k для двухслойной схемы

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H$$

в предположении, что оператор $DB^{-1}A$ самосопряжен в H и заданы γ_1 и γ_2 — постоянные энергетической эквивалентности операторов D и $DB^{-1}A$:

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0. \quad (1)$$

Получим теперь решение этой задачи при дополнительном ограничении $\tau_k \equiv \tau$, т. е. в предположении, что итерационные параметры τ_k не зависят от номера итерации k . Эта задача возникает при нахождении итерационного параметра τ для стационарной двухслойной схемы

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots \quad (2)$$

Напомним формулировку указанной выше задачи: среди полиномов степени n вида $Q_n(t) = \prod_{j=1}^n (1 - \tau_j t)$ найти полином, наименее уклоняющийся от нуля на отрезке $[\gamma_1, \gamma_2]$. В силу сделанного ограничения полином $P_n(t)$ имеет вид

$$P_n(t) = (1 - \tau t)^n.$$

Поэтому поставленная выше задача эквивалентна следующей: среди полиномов первой степени, принимающих в точке $t=0$ значение единицы, найти полином наименее уклоняющийся от нуля на отрезке $[\gamma_1, \gamma_2]$.

Эта задача является частным случаем рассмотренной в § 2 задачи. В данном случае $n=1$, и из результатов п. 1 § 2 следует, что искомый полином имеет вид

$$Q_1(t) = q_1 T_1 \left(\frac{1 - \tau_0 t}{\rho_0} \right), \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

где

$$q_1 = \frac{2\rho_1}{1 + \rho_1^2} = \rho_0, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Здесь $T_1(x)$ — полином Чебышева первого рода. Так как $T_1(x) = x$, то полином $Q_1(t)$ имеет вид

$$Q_1(t) = 1 - \tau_0 t, \quad \max_{\gamma_1 \leq t \leq \gamma_2} |Q_1(t)| = q_1 = \rho_0,$$

поэтому

$$P_n(t) = (1 - \tau_0 t)^n.$$

Таким образом, оптимальное значение параметра τ для схемы (2) найдено:

$$\tau = \tau_0 = 2/(\gamma_1 + \gamma_2). \quad (3)$$

Так как норма разрешающего оператора $T_{n,0}$ для схемы (2) (см. п. 3 § 1) оценивается следующим образом:

$$\|T_{n,0}\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} |P_n(t)|,$$

то при $\tau = \tau_0$ получим оценку $\|T_{n,0}\| \leq \rho_0^n$. Отсюда следует оценка для погрешности z_n в H_D :

$$\|z_n\|_D \leq \rho_0^n \|z_0\|_D. \quad (4)$$

Итерационный метод (2), (3) называется *методом простой итерации*.

Итак, доказана

Теорема 2. Пусть самосопряженный оператор $DB^{-1}A$ удовлетворяет условиям (1). Метод простой итерации (2), (3) сходится в H_D , и для погрешности имеет место оценка (4). Для числа итераций верна оценка $n \geq n_0(\varepsilon)$, где $n_0(\varepsilon) = \ln \varepsilon / \ln \rho_0$.

Замечание. Как и для чебышевского метода, априорная оценка погрешности для метода простой итерации является неулучшаемой в случае конечномерного пространства.

Сравним число итераций для чебышевского метода и метода простой итерации. Из теоремы 1 в случае малых ξ имеем следующую оценку для числа итераций чебышевского метода:

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln 0,5\varepsilon}{\ln \rho_1} \approx \frac{1}{2\sqrt{\xi}} \ln \frac{2}{\varepsilon}.$$

Из теоремы 2 получим оценку для числа итераций метода простой итерации

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho_0} \approx \frac{1}{2\xi} \ln \frac{1}{\varepsilon}.$$

Из этих оценок следует, что для $\xi \ll 1$ число итераций чебышевского метода существенно меньше числа итераций метода простой итерации. Например, для $\xi = 0,01$ число итераций для метода простой итерации примерно в 10 раз больше, чем для чебышевского метода.

2. Оценка нормы оператора перехода. В п. 1 § 3 была исследована скорость сходимости метода простой итерации. При этом метод простой итерации рассматривался как частный случай чебышевского метода. Из методических соображений будет полезно изучить сходимость метода простой итерации независимо от чебышевского метода.

Итак, пусть для нахождения приближенного решения уравнения

$$Au = f$$

используется двухслойная схема (2)

$$\frac{B^{y_{k+1}-y_k}}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H. \quad (5)$$

Для изучения сходимости схемы (5) перейдем к задаче для эквивалентной погрешности $x_k = D^{1/2}z_k$:

$$x_{k+1} = Sx_k, \quad k = 0, 1, \dots, \quad S = E - \tau C, \quad (6)$$

где $C = D^{1/2}B^{-1}AD^{-1/2}$. Используя (6), найдем явное выражение для x_n через x_0 : $x_n = S^n x_0$, из которого следует оценка для нормы погрешности z_n в H_D

$$\|z_n\|_D = \|x_n\| \leq \|S^n\| \|x_0\| = \|S^n\| \|z_0\|_D. \quad (7)$$

Будем предполагать, что оператор $DB^{-1}A$ самосопряжен в H и заданы постоянные γ_1 и γ_2 в неравенствах (1). При этих предположениях оператор C , а вместе с ним и оператор S , самосопряжены в H и γ_1 и γ_2 — границы оператора C :

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \gamma_1 > 0, \quad C = C^*. \quad (8)$$

В силу самосопряженности оператора S имеет место равенство $\|S^n\| = \|S\|^n$. Поэтому из оценки (7) следует, что итерационный параметр τ нужно выбирать из условия минимума по τ нормы оператора перехода $S = E - \tau C$.

Имеет место

Лемма 2. Пусть $S = E - \tau C$ и выполнены условия (8). Норма оператора S минимальна при $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$, и имеет место оценка

$$\|S\| = \|E - \tau_0 C\| = \rho_0, \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Действительно, так как S самосопряженный в H оператор, то из определения нормы получим

$$\|S\| = \sup_{x \neq 0} \frac{|(Sx, x)|}{(x, x)} = \sup_{x \neq 0} \left| 1 - \tau \frac{(Cx, x)}{(x, x)} \right| = \max_{\gamma_1 \leq t \leq \gamma_2} |1 - \tau t|.$$

Так как $\varphi(t) = 1 - \tau t$ — линейная функция, то максимальное по модулю значение $\varphi(t)$ на отрезке $[\gamma_1, \gamma_2]$ может достигаться лишь на концах отрезка. Непосредственные вычисления дают

$$\|S\| = \max(|1 - \tau\gamma_1|, |1 - \tau\gamma_2|) = \begin{cases} \varphi_1(\tau) = 1 - \tau\gamma_1, & 0 \leq \tau \leq \tau_0, \\ \varphi_2(\tau) = \tau\gamma_2 - 1, & \tau_0 \leq \tau, \end{cases}$$

где τ_0 указано в лемме. Так как функция $\varphi_1(\tau)$ убывает на отрезке $[0, \tau_0]$, а $\varphi_2(\tau)$ возрастает при $\tau \geq \tau_0$, то минимум нормы

оператора S достигается при $\tau = \tau_0$ и равен $\rho_0 = 1 - \tau_0 \gamma_1 = \tau_0 \gamma_2 - 1 = -(1 - \xi)/(1 + \xi)$, $\xi = \gamma_1/\gamma_2$. Лемма доказана.

Из леммы 2 и оценки (7) следует, что при $\tau = \tau_0$ для погрешности итерационной схемы (5) верна оценка

$$\|z_n\|_D \leq \rho_0^n \|z_0\|_D.$$

Таким образом, получено еще одно доказательство сформулированной выше теоремы 2 о сходимости метода простой итерации. Примеры выбора оператора D , для которого выполнено условие самосопряженности оператора $DB^{-1}A$, рассмотрены в п. 3 § 2.

§ 4. Несамосопряженный случай. Метод простой итерации

1. Постановка задачи. В §§ 2, 3 были построены двухслойные итерационные методы для приближенного решения линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A , заданным в вещественном гильбертовом пространстве H . Предполагалось, что операторы A , B и D таковы, что оператор $DB^{-1}A$ самосопряжен в H , и заданы постоянные энергетической эквивалентности γ_1 и γ_2 операторов D и $DB^{-1}A$, причем $\gamma_1 > 0$.

При этих предположениях задача оптимального выбора итерационных параметров была решена и были построены чебышевский метод и метод простой итерации. В п. 3 § 2 были рассмотрены некоторые примеры выбора оператора D и найдены условия самосопряженности оператора $DB^{-1}A$ для каждого конкретного выбора оператора D .

Очевидно, что если заданы операторы A и B , то не всегда можно указать такой оператор D , для которого оператор $DB^{-1}A$ будет самосопряжен в H . Следовательно, необходимо изучить итерационные методы и в несамосопряженном случае.

В данном параграфе изучается метод простой итерации для несамосопряженного случая. Будут рассмотрены некоторые способы выбора итерационного параметра в зависимости от объема априорной информации об операторах итерационной схемы.

Итак, пусть оператор $DB^{-1}A$ несамосопряжен в H . Для приближенного решения уравнения (1) рассмотрим неявную двухслойную итерационную схему

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H. \quad (2)$$

Для исследования сходимости схемы (2), как обычно, перейдем к задаче для эквивалентной погрешности $x_k = D^{1/2}z_k$

$$x_{k+1} = Sx_k, \quad k = 0, 1, \dots, \quad S = E - \tau C, \quad (3)$$

где $C = D^{1/2}B^{-1}AD^{-1/2}$. В силу сделанных выше предположений, оператор C несамосопряжен в H . Из сделанной замены и уравнения (3) получим

$$x_n = S^n x_0, \quad \|x_n\| = \|z_n\|_D \leq \|S^n\| \|x_0\| = \|S^n\| \|z_0\|_D. \quad (4)$$

Следовательно, итерационный параметр τ должен быть выбран из условия минимума по τ нормы разрешающего оператора S^n .

2. Минимизация нормы оператора перехода.

2.1. Первый случай. Получим оценку для нормы оператора S^n . Так как для любого оператора имеет место оценка $\|S^n\| \leq \|S\|^n$, то первый способ выбора параметра τ состоит в нахождении параметра τ из условия минимума нормы оператора перехода S . Получим два типа оценок нормы оператора S в зависимости от объема априорной информации относительно оператора C .

В первом случае предполагается, что априорная информация состоит в задании постоянных γ_1 и γ_2 из неравенств

$$\gamma_1(x, x) \leq (Cx, x), \quad (Cx, Cx) \leq \gamma_2(Cx, x), \quad \gamma_1 > 0. \quad (5)$$

Если $C = C^*$, то γ_1 и γ_2 — границы оператора C .

Лемма 3. Пусть в неравенствах (5) заданы γ_1 и γ_2 , тогда для нормы оператора $S = E - \tau C$ при $\tau = 1/\gamma_2$ справедлива оценка

$$\|S\| \leq \rho, \quad \rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2.$$

Действительно, используя (5), получим

$$\begin{aligned} \|Sx\|^2 &= \|x - \tau Cx\|^2 = (x, x) - 2\tau(Cx, x) + \tau^2(Cx, Cx) \leq \\ &\leq (x, x) - 2\tau(Cx, x) + \tau^2\gamma_2(Cx, x) = \|x\|^2 - \tau(2 - \tau\gamma_2)(Cx, x). \end{aligned}$$

Отсюда следует, что если выполнено условие $\tau(2 - \tau\gamma_2) > 0$, т. е. $0 < \tau < 2/\gamma_2$, то норма оператора S будет меньше единицы. Пусть это условие выполнено, тогда, используя (5), получим

$$\|Sx\|^2 \leq [1 - \tau\gamma_1(2 - \tau\gamma_2)]\|x\|^2$$

и, следовательно,

$$\|S\|^2 = \sup_{x \neq 0} \frac{\|Sx\|^2}{\|x\|^2} \leq 1 - \tau\gamma_1(2 - \tau\gamma_2).$$

Функция $\varphi(\tau) = 1 - \tau\gamma_1(2 - \tau\gamma_2)$ в точке $\tau = 1/\gamma_2$ имеет минимум, равный $\varphi(1/\gamma_2) = 1 - \xi$, где $\xi = \gamma_1/\gamma_2$. Поэтому для указанного значения параметра τ для нормы оператора S справедлива оценка $\|S\| \leq \sqrt{1 - \xi}$. Лемма доказана.

Подставляя в (5) оператор $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$, получим, что неравенства (5) эквивалентны следующим неравенствам:

$$\begin{aligned} \gamma_1(Dx, x) &\leq (DB^{-1}Ax, x), \\ (DB^{-1}Ax, B^{-1}Ax) &\leq \gamma_2(DB^{-1}Ax, x), \quad \gamma_1 > 0. \end{aligned} \quad (6)$$

Подставляя в (4) оценку для нормы оператора S , полученную в лемме 3, найдем

$$\|z_n\|_D \leq \rho^n \|z_0\|_D, \quad \rho = \sqrt{1 - \xi}. \quad (7)$$

Теорема 3. Пусть γ_1 и γ_2 — постоянные из неравенств (6). Метод простой итерации (2) при значении итерационного параметра $\tau = 1/\gamma_2$ сходится в H_D , и для погрешности z_n имеет место оценка (7). Для числа итераций верна оценка $n \geq n_0(\epsilon)$, где $n_0(\epsilon) = \ln \epsilon / \ln \rho$, $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1 / \gamma_2$.

Приведем примеры выбора оператора D и конкретный вид неравенств (6). В табл. 6 приведены: предположения относительно операторов A и B , указан оператор D и вид неравенств (6). При получении конкретного вида неравенств (6) мы исходим как из самих неравенств (6), так и из эквивалентных им неравенств получающихся из (6) при помощи замены $x = A^{-1}By$.

Таблица 6

A и B	D	Неравенства
1) $A = A^* > 0$, $B \neq B^* > 0$	A или $B^*A^{-1}B$ A^2 или B^*B	$\gamma_1(Bx, A^{-1}Bx) \leq (Bx, x)$, $(Ax, x) \leq \gamma_2(Bx, x)$ $\gamma_1(Bx, Bx) \leq (Ax, Bx)$, $(Ax, Ax) \leq \gamma_2(Ax, Bx)$
2) $A \neq A^* > 0$, $B = B^* > 0$	B или $A^*B^{-1}A$ B^2 или A^*A	$\gamma_1(Bx, x) \leq (Ax, x)$, $(Ax, B^{-1}Ax) \leq \gamma_2(Ax, x)$ $\gamma_1(Bx, Bx) \leq (Ax, Bx)$, $(Ax, Ax) \leq \gamma_2(Ax, Bx)$
3) $A \neq A^* > 0$, $B \neq B^* > 0$	A^*A или B^*B	$\gamma_1(Bx, Bx) \leq (Ax, Bx)$, $(Ax, Ax) \leq \gamma_2(Ax, Bx)$
4) $A = A^*$, $B = B^*$, $AB \neq BA$	A^2 или B^2	$\gamma_1(Bx, Bx) \leq (Ax, Bx)$, $(Ax, Ax) \leq \gamma_2(Ax, Bx)$

Отметим неравенства

$$\gamma_1(Bx, Bx) \leq (Ax, Bx), \quad (Ax, Ax) \leq \gamma_2(Ax, Bx), \quad \gamma_1 > 0.$$

Если эти условия выполнены, то для рассмотренных в табл. 6 частных случаев в качестве оператора D можно взять либо оператор A^2 , если $A = A^*$, либо оператор A^*A . При таком выборе оператора D норма погрешности z_n в H_D может быть вычислена в процессе итераций

$$\|z_n\|_D^2 = (Dz_n, z_n) = (Az_n, Az_n) = \|r_n\|^2, \quad r_n = Ay_n - f.$$

Вернемся к оценке нормы оператора S . Если оператор C самосопряжен в H , то в силу (5) он положительно определен, и следовательно, существует корень квадратный из оператора C . Полагая во втором из неравенств (5) $x = C^{-1/2}y$, получим, что неравенства (5) эквивалентны неравенствам

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \gamma_1 > 0.$$

Из леммы 2 при этих предположениях вытекает следующая оценка для нормы оператора S : $\|S\| \leq \rho_0$, $\rho_0 = (1 - \xi)/(1 + \xi)$, $\xi = \gamma_1/\gamma_2$.

Сравнивая эту оценку с полученной в лемме 3, убеждаемся, что оценка леммы 3 является грубой и не переходит в оценку леммы 2, когда оператор C является самосопряженным в H .

2.2. Второй случай. Получим теперь другую оценку для нормы оператора перехода S , которая будет переходить в оценку леммы 2, когда C является самосопряженным в H оператором. Для этого увеличим объем априорной информации относительно оператора C , предполагая, что заданы три числа γ_1 , γ_2 и γ_3 :

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \|C_1\| \leq \gamma_3, \quad \gamma_1 > 0, \quad \gamma_3 \geq 0, \quad (9)$$

где $C_1 = 0,5(C - C^*)$ — несамосопряженная часть оператора C .

Имеет место

Лемма 4. Пусть в неравенствах (9) заданы γ_1 , γ_2 и γ_3 . Тогда для нормы оператора $S = E - \tau C$ при $\tau = \bar{\tau}_0 = \tau_0(1 - \kappa\rho_0)$ справедлива оценка

$$\|S\| \leq \bar{\rho}_0, \quad \bar{\rho}_0 = (1 - \bar{\xi})/(1 + \bar{\xi}), \quad (9')$$

где

$$\bar{\tau}_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \kappa = \frac{\gamma_3}{\sqrt{\gamma_1\gamma_2 + \gamma_3^2}}, \quad \bar{\xi} = \frac{1 - \kappa}{1 + \kappa} \cdot \frac{\gamma_1}{\gamma_2}.$$

Приведем доказательство леммы 4. Пусть θ — произвольное число из интервала $(0, 1)$. Представим оператор S в следующем виде:

$$S = E - \tau C = [\theta E - \tau C_0] + [(1 - \theta)E - \tau C_1],$$

где $C_0 = 0,5(C + C^*)$ — самосопряженная часть оператора C . Используя неравенство треугольника, получим оценку для нормы оператора S :

$$\|S\| \leq \|\theta E - \tau C_0\| + \|(1 - \theta)E - \tau C_1\|. \quad (10)$$

Оценим отдельно норму каждого оператора. Из (9) и равенства $(C_0x, x) = 0,5(Cx, x) + 0,5(C^*x, x) = (Cx, x)$ получим, что γ_1 и γ_2 — границы самосопряженного оператора C_0 :

$$\gamma_1 E \leq C_0 \leq \gamma_2 E, \quad \gamma_1 > 0.$$

По аналогии с леммой 2 получим следующую оценку для нормы оператора $\theta E - \tau C_0$:

$$\|\theta E - \tau C_0\| \leq \max(|\theta - \tau\gamma_1|, |\theta - \tau\gamma_2|) = \begin{cases} \theta - \tau\gamma_1, & 0 \leq \tau \leq \theta\tau_0, \\ \tau\gamma_2 - \theta, & \tau \geq \theta\tau_0. \end{cases}$$

Оценим норму оператора $(1-\theta)E - \tau C_1$. Так как $(C_1x, x) = 0$, то для всех $x \in H$ получим

$$\|((1-\theta)E - \tau C_1)x\|^2 = (1-\theta)^2 \|x\|^2 + \tau^2 \|C_1x\|^2 \leq ((1-\theta)^2 + \tau^2 \|C_1\|^2) \|x\|^2.$$

Отсюда и из (9) следует оценка $\|(1-\theta)E - \tau C_1\| \leq [(1-\theta)^2 + \tau^2 \gamma_3^2]^{1/2}$. Подставляя полученные оценки в (10), будем иметь

$$\|S\| \leq \begin{cases} \varphi_1(\theta, \tau) = \theta - \tau\gamma_1 + \sqrt{(1-\theta)^2 + \tau^2 \gamma_3^2}, & 0 \leq \tau \leq \tau_0 \theta, \\ \varphi_2(\theta, \tau) = \tau\gamma_2 - \theta + \sqrt{(1-\theta)^2 + \tau^2 \gamma_3^2}, & \tau \geq \tau_0 \theta. \end{cases}$$

Выберем теперь параметры τ и θ из условия минимума оценки для нормы оператора S . Заметим, что функция $\varphi_2(\theta, \tau)$ монотонно возрастает по τ . Поэтому для минимизации нормы оператора S достаточно рассмотреть область $0 \leq \tau \leq \tau_0 \theta$, $0 < \theta < 1$. В этой области $\|S\| \leq \varphi_1(\theta, \tau)$.

Исследуем функцию $\varphi_1(\theta, \tau)$. Эта функция монотонно возрастает по θ , следовательно, минимум достигается при $\tau = \tau_0 \theta$. При этом значении параметра τ будем иметь

$$\begin{aligned} \|S\| \leq \varphi(\theta) &= \varphi_1(\theta, \tau_0 \theta) = \\ &= \theta(1 - \tau_0 \gamma_1) + \sqrt{(1-\theta)^2 + \tau_0^2 \gamma_3^2 \theta^2} = \theta \rho_0 + \sqrt{(1-\theta)^2 + \tau_0^2 \gamma_3^2 \theta^2}. \end{aligned}$$

Итак, нужно показать, что $\min_{0 < \theta < 1} \varphi(\theta) = \bar{\rho}_0$. Найдем минимум функции $\varphi(\theta)$. Сделаем замену переменной, полагая

$$\theta = (1-x)/(1+a^2), x \in (-a^2, 1), a^2 = \tau_0^2 \gamma_3^2.$$

Функция $\varphi(\theta)$ запишется в виде

$$\varphi(\theta) = \bar{\varphi}(x) = \frac{1}{\sqrt{1+a^2}} \left(\sqrt{x^2 + a^2} - \frac{\rho_0}{\sqrt{1+a^2}} x \right) + \frac{\rho_0}{1+a^2}. \quad (11)$$

Отсюда видно, что достаточно найти минимум функции

$$v(x) = \sqrt{x^2 + a^2} - \rho_0 x / \sqrt{1+a^2}$$

в области $-a^2 < x < 1$. Вычисляя производные функции $v(x)$

$$v'(x) = \frac{x}{\sqrt{x^2 + a^2}} - \frac{\rho_0}{\sqrt{1+a^2}}, \quad v''(x) = \frac{a^2}{(x^2 + a^2)^{3/2}} > 0,$$

находим, что уравнение $v'(x) = 0$ дает точку минимума функции $v(x)$. Решая уравнение

$$\sqrt{\frac{x^2 + a^2}{1+a^2}} = \frac{x}{\rho_0}, \quad (12)$$

найдем искомую точку минимума функции $v(x)$:

$$x_0 = a\rho_0 / \sqrt{1+a^2 - \rho_0^2} \in (0, 1), \quad \theta_0 = (1-x_0)/(1+a^2).$$

Подставляя (12) в (11), найдем минимальное значение функции $\varphi(\theta)$:

$$\varphi(\theta_0) = \sqrt{\frac{x_0^2 + a^2}{1+a^2}} + \rho_0 \frac{1-x_0}{1+a^2} = \frac{x_0}{\rho_0} + \theta_0 \rho_0. \quad (13)$$

Осталось выразить x_0 и θ_0 через известные величины. Используя обозначения леммы 4, получим

$$1 - \rho_0^2 = \tau_0^2 \gamma_1 \gamma_2, \quad x_0 = \tau_0 \gamma_3 \rho_0 / \sqrt{1 - \rho_0^2 + \tau_0^2 \gamma_3^2} = \kappa \rho_0. \quad (14)$$

Из (12) найдем

$$a^2 = x_0^2(1 - \rho_0^2)/(p_0^2 - x_0^2), \quad 1 + a^2 = \rho_0^2(1 - x_0^2)/(p_0^2 - x_0^2).$$

Поэтому

$$\theta_0 p_0 = \frac{1 - x_0}{1 + a^2} p_0 = \frac{\rho_0^2 - x_0^2}{\rho_0(1 + x_0)} = \frac{\rho_0(1 - \kappa^2)}{1 + \kappa p_0}. \quad (15)$$

Подставим (14) и (15) в (13):

$$\varphi(\theta_0) = \kappa + \frac{\rho_0(1 - \kappa^2)}{1 + \rho_0 \kappa} = \frac{\kappa + \rho_0}{1 + \rho_0 \kappa} = \frac{(1 + \kappa) - \xi(1 - \kappa)}{(1 + \kappa) + \xi(1 - \kappa)} = \frac{1 - \bar{\xi}}{1 + \bar{\xi}} = \bar{\rho}_0. \quad (16)$$

Найдем теперь выражение для параметра $\tau = \tau_0 \theta_0$. Сравнивая (15) и (16), получим

$$\theta_0 p_0 = \bar{\rho}_0 - \kappa. \quad (17)$$

С другой стороны, из (16) можно выразить ρ_0 через $\bar{\rho}_0$ и κ :

$$\rho_0 = (\bar{\rho}_0 - \kappa)/(1 - \kappa \bar{\rho}_0).$$

Подставляя ρ_0 в (17), находим

$$\theta_0 = 1 - \kappa \bar{\rho}_0, \quad \tau = \tau_0 (1 - \kappa \bar{\rho}_0).$$

Лемма доказана.

Неравенства (9) могут быть записаны в следующем виде:

$$\gamma_1(x, x) \leqslant (Cx, x) \leqslant \gamma_2(x, x), \quad (C_1 x, C_1 x) \leqslant \gamma_3^2(x, x) \quad \gamma_1 > 0.$$

Подставляя сюда $C = D^{-1/2} (DB^{-1}A) D^{-1/2}$ и $C_1 = 0,5D^{-1/2} \times (DB^{-1}A - (DB^{-1}A)^*) D^{-1/2}$, получим неравенства

$$\begin{aligned} \gamma_1 D &\leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \\ \left(D^{-1} \frac{DB^{-1}A - (DB^{-1}A)^*}{2} x, \frac{DB^{-1}A - (DB^{-1}A)^*}{2} x \right) &\leqslant \gamma_3^2(Dx, x). \end{aligned} \quad (18)$$

Подставляя в (4) оценку (9') для нормы оператора S , найдем

$$\|z_n\|_D \leqslant \bar{\rho}_0^n \|z_0\|_D. \quad (19)$$

Теорема 4. Пусть γ_1 , γ_2 и γ_3 — постоянные в неравенствах (18). Метод простой итерации (2) при значении итерационного параметра $\tau = \bar{\tau}_0 = \tau_0(1 - \kappa \bar{\rho}_0)$ сходится в H_D , и для погрешности z_n имеет место оценка (19). Для числа итераций верна оценка $n \geqslant n_0(\varepsilon)$, где $n_0(\varepsilon) = \ln \varepsilon / \ln \bar{\rho}_0$,

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \bar{\rho}_0 = \frac{1 - \bar{\xi}}{1 + \bar{\xi}}, \quad \bar{\xi} = \frac{1 - \kappa}{1 + \kappa} \cdot \frac{\gamma_1}{\gamma_2}, \quad \kappa = \frac{\gamma_3}{\sqrt{\gamma_1 \gamma_2 + \gamma_3^2}}.$$

Замечание. Так как число итераций определяется величиной $\bar{\xi}$, которую можно записать в виде

$$\bar{\xi} = (\sqrt{\gamma_1/\gamma_2 + (\gamma_3/\gamma_2)^2} - \gamma_3/\gamma_2)^2,$$

то оператор B следует выбрать так, чтобы отношение $\xi = \gamma_1/\gamma_2$ было максимальным, а γ_3/γ_2 минимальным.

Приведем примеры выбора оператора D . Если в качестве D выбрать оператор A^*A или B^*B , то неравенства (18) можно записать в виде

$$\begin{aligned} \gamma_1(Bx, Bx) &\leq (Ax, Bx) \leq \gamma_2(Bx, Bx), \\ \|0.5(AB^{-1} - (B^*)^{-1}A^*)\| &\leq \gamma_3. \end{aligned} \quad (20)$$

Действительно, для случая $D = B^*B$ это утверждение очевидно, а если $D = A^*A$, то в (18) нужно сделать замену $x = A^{-1}By$ и получить неравенства (20).

Если оператор B является самосопряженным положительно определенным и ограниченным в H , то в качестве оператора D можно взять оператор B или $A^*B^{-1}A$. В этом случае неравенства (18) эквивалентны следующим неравенствам:

$$\begin{aligned} \gamma_1 B &\leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \\ (B^{-1}A_1x, A_1x) &\leq \gamma_3^2 (Bx, x), \quad A_1 = 0.5(A - A^*). \end{aligned} \quad (21)$$

Действительно, для $D = B$ неравенства (18) и (21) совпадают, а для $D = A^*B^{-1}A$ неравенства (21) следуют из неравенств (18) после замены $x = A^{-1}By$ в (18).

3. Минимизация нормы разрешающего оператора.

3.1. Первый случай. В п. 2 § 4 были получены оценки для нормы оператора S^n , основанные на неравенстве $\|S^n\| \leq \|S\|^n$. Рассмотрим теперь другой способ получения оценки для $\|S^n\|$. Этот способ основан на оценке числового радиуса оператора.

Напомним (см. § 1 гл. V), что *числовым радиусом оператора T* , действующего в комплексном гильбертовом пространстве \tilde{H} , называется величина

$$\rho(T) = \sup_{\|z\|=1} |(Tz, z)|, \quad z \in \tilde{H}.$$

Для линейного ограниченного оператора T числовой радиус удовлетворяет неравенствам

$$\mu(T)\|T\| \leq \rho(T) \leq \|T\|, \quad \rho(T^n) \leq [\rho(T)]^n, \quad (22)$$

где n — натуральное число, а $\mu(T) \geq 1/2$.

Используя понятие числового радиуса оператора, получим две оценки для нормы оператора S^n в зависимости от типа априорной информации относительно оператора C .

Рассмотрим случай, когда априорная информация задана в виде постоянных γ_1 , γ_2 и γ_3 :

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \|C_1x\| \leq \gamma_3 \|x\|, \quad \gamma_1 > 0, x \in H. \quad (23)$$

Комплексное пространство \tilde{H} определим следующим образом: оно состоит из элементов вида $z = x + iy$, где $x, y \in H$. Скалярное произведение в \tilde{H} определяется формулой

$$\begin{aligned} (z, w) &= (x, u) + i(y, u) - i(x, v) + (y, v), \\ z &= x + iy, \quad w = u + iv. \end{aligned}$$

Линейный оператор C , заданный на H , определим на \tilde{H} следующим образом: $Cz = Cx + iCy$.

В силу свойств (22) для любого целого n верна оценка

$$\|S^n\| \leq \frac{1}{\mu(S^n)} \rho(S^n) \leq 2 [\rho(S)]^n,$$

поэтому достаточно оценить числовой радиус оператора S .

Имеет место

Лемма 5. Пусть в неравенствах (23) заданы γ_1, γ_2 и γ_3 . Тогда для нормы оператора $S = E - \tau C$ в H при $\tau = \min(\tau_0, \kappa \tau_0)$ справедлива оценка

$$\|S^n\| \leq 2\rho^n,$$

где

$$\rho^2 = \begin{cases} 1 - \kappa(1 - \rho_0), & 0 < \kappa \leq 1, \\ 1 - (2 - 1/\kappa)(1 - \rho_0), & \kappa \geq 1, \end{cases} \quad \kappa = \frac{\gamma_1(\gamma_1 + \gamma_2)}{2(\gamma_1^2 + \gamma_3^2)},$$

$$\tau_0 = 2/(\gamma_1 + \gamma_2), \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Для доказательства леммы представим оператор C в виде суммы $C = C_0 + C_1$, $C_0 = 0,5(C + C^*)$, $C_1 = 0,5(C - C^*)$. Оценим числовой радиус оператора $S = E - \tau C$. Для любого $z \in H$ получим

$$(Sz, z) = (z, z) - \tau(C_0 z, z) - \tau(C_1 z, z).$$

В силу самосопряженности оператора C_0 скалярное произведение $(C_0 z, z)$ есть действительное число. Так как $C_1 = -C_1^*$, то $(C_1 z, z)$ — мнимое число. Поэтому

$$|(Sz, z)|^2 = |(z, z) - \tau(C_0 z, z)|^2 + \tau^2 |(C_1 z, z)|^2. \quad (24)$$

Из неравенств (23) получим

$$\gamma_1(z, z) \leq (C_0 z, z) \leq \gamma_2(z, z), \quad \|C_1 z\| \leq \gamma_3 \|z\|. \quad (25)$$

Пусть $\|z\| = 1$. Из (25) найдем

$$|(z, z) - \tau(C_0 z, z)| \leq \max(|1 - \tau\gamma_1|, |1 - \tau\gamma_2|) = \begin{cases} 1 - \tau\gamma_1, & 0 \leq \tau \leq \tau_0, \\ \tau\gamma_2 - 1, & \tau \geq \tau_0, \end{cases}$$

$$|(C_1 z, z)| \leq \|C_1 z\| \|z\| \leq \gamma_3.$$

Подставляя эти оценки в (24), получим

$$\rho^2(S) = \sup_{\|z\|=1} |(Sz, z)|^2 \leq \begin{cases} \varphi_1(\tau) = (1 - \tau\gamma_1)^2 + \tau^2\gamma_3^2, & 0 \leq \tau \leq \tau_0, \\ \varphi_2(\tau) = (1 - \tau\gamma_2)^2 + \tau^2\gamma_3^2, & \tau \geq \tau_0. \end{cases}$$

Выберем параметр τ из условия минимума оценки для числового радиуса оператора S . Так как функция $\varphi_2(\tau)$ возрастает по τ при $\tau \geq \tau_0$:

$$\varphi'_2(\tau) = 2[\tau(\gamma_2^2 + \gamma_3^2) - \gamma_2] \geq 2 \frac{\gamma_2(\gamma_2 - \gamma_1) + 2\gamma_3^2}{\gamma_1 + \gamma_2} > 0,$$

то минимум $\rho(S)$ по τ следует искать в области $\tau \leq \tau_0$, где для $\rho(S)$ выполняется оценка $\rho^2(S) \leq \varphi_1(\tau)$.

Исследуем функцию $\varphi_1(\tau)$. Так как

$$\varphi''_1(\tau) = 2(\gamma_1^2 + \gamma_3^2) > 0,$$

то, приравнивая производную

$$\varphi'_1(\tau) = 2[\tau(\gamma_1^2 + \gamma_3^2) - \gamma_1]$$

нулю, найдем точку экстремума функции $\varphi_1(\tau)$

$$\tau = \tau_1 = \frac{\gamma_1}{\gamma_1^2 + \gamma_3^2} = \tau_0 \kappa.$$

При $\tau \leq \tau_1$ функция $\varphi_1(\tau)$ убывает, а при $\tau \geq \tau_1$ — возрастает. Поэтому минимальное значение $\varphi_1(\tau)$ достигается в точке $\tau = \tau_1$, если $\tau_1 \leq \tau_0$, и в точке $\tau = \tau_0$, если $\tau_1 \geq \tau_0$. Итак, оптимальное значение параметра τ найдено $\tau = \min(\tau_0, \tau_0 \kappa)$. При этом

$$\min_{\tau} \rho^2(S) \leq \begin{cases} \varphi_1(\tau_0), & \kappa \geq 1, \\ \varphi_1(\tau_1), & 0 \leq \kappa \leq 1. \end{cases}$$

Вычислим $\varphi_1(\tau_0)$ и $\varphi_1(\tau_1)$. Из определения κ и равенства $1 - \tau_0 \gamma_1 = \rho_0$ получим

$$\kappa = \frac{\tau_0 \gamma_1}{\tau_0^2 \gamma_1^2 + \tau_0^2 \gamma_3^2}, \quad \tau_0^2 \gamma_3^2 = \frac{\tau_0 \gamma_1}{\kappa} - 1 - \tau_0^2 \gamma_1^2 = \frac{1 - \rho_0}{\kappa} - (1 - \rho_0)^2.$$

Далее,

$$\begin{aligned} \varphi_1(\tau_0) &= (1 - \tau_0 \gamma_1)^2 + \tau_0^2 \gamma_3^2 = \rho_0^2 + (1 - \rho_0)/\kappa - (1 - \rho_0)^2 = 1 - (2 - 1/\kappa)(1 - \rho_0), \\ \varphi_1(\tau_1) &= (1 - \tau_1 \gamma_1)^2 + \tau_1^2 \gamma_3^2 = 1 - 2\tau_1 \gamma_1 + \tau_1^2 (\gamma_1^2 + \gamma_3^2) = \\ &= 1 - \tau_1 \gamma_1 = 1 - \kappa \tau_0 \gamma_1 = 1 - \kappa (1 - \rho_0). \end{aligned}$$

Итак, числовой радиус оценен. Оценка леммы следует из неравенства $\|S^n\| \leq 2[\rho(S)]^n$. Лемма доказана.

Используя лемму 5, получим оценку для нормы погрешности z_n :

$$\|z_n\|_D \leq 2\rho^n \|z_0\|_D. \quad (26)$$

Теорема 5. Пусть γ_1 , γ_2 и γ_3 — постоянные в неравенствах (18). Метод простой итерации (2) при значении итерационного параметра $\tau = \min(\tau_0, \kappa \tau_0)$ сходится в H_D , и для погрешности z_n имеет место оценка (26). Для числа итераций верна оценка $n \geq n_0(\varepsilon)$, где $n_0(\varepsilon) = \ln(0,5\varepsilon)/\ln \rho$, а κ и ρ определены в лемме 5.

Примеры выбора оператора D и конкретный вид неравенств (18) приведены в п. 2.2.

3.2. Второй случай. Используя понятие числового радиуса оператора, получим еще одну оценку для нормы оператора S^n . Будем предполагать, что априорная информация задана в виде постоянных γ_1 , γ_2 и γ_3 в неравенствах

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad (C_1 x, C_1 x) \leq \gamma_3 (Cx, x), \quad \gamma_1 > 0. \quad (27)$$

Имеет место

Лемма 6. Пусть в неравенствах (27) заданы γ_1 , γ_2 и γ_3 . Тогда для нормы оператора $S = E - \tau C$ в H при $\tau = \min(\tau_0^*, \kappa \tau_0^*)$ справедлива оценка

$$\|S^n\| \leq 2\rho^n,$$

где

$$\rho^2 = \begin{cases} 1 - (2\kappa - 1) \frac{1 - \rho_0}{1 + \rho_0}, & \frac{1}{2} \leq \kappa \leq 1, \\ 1 - \left(2 - \frac{1}{\kappa}\right)^2 \frac{1 - \rho_0}{1 + \rho_0}, & \kappa \geq 1, \end{cases} \quad \kappa = \frac{\gamma_1 + \gamma_2 + \gamma_3}{2(\gamma_1 + \gamma_3)},$$

$$\tau_0^* = 2/(\gamma_1 + \gamma_2 + \gamma_3), \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Действительно, представляя оператор C в виде $C = C_0 + C_1$, где $C_0 = 0,5(C + C^*)$ и $C_1 = 0,5(C - C^*)$, получим

$$|(Sz, z)|^2 = [(z, z) - \tau(C_0 z, z)]^2 + \tau^2 |(C_1 z, z)|^2.$$

Из неравенства Коши—Буняковского и из условий леммы найдем

$$|(C_1 z, z)|^2 \leq (C_1 z, C_1 z) (z, z) \leq \gamma_3 (C_0 z, z) (z, z).$$

Так как для любого $z \in \tilde{H}$ имеют место неравенства

$$\gamma_i(z, z) \leq (C_0 z, z) \leq \gamma_2(z, z), \quad \gamma_i > 0,$$

то из трех предыдущих соотношений получим следующую оценку для числового радиуса оператора S :

$$\rho^2(S) \leq \max_{\gamma_1 \leq t \leq \gamma_2} \varphi(t), \quad \text{где } \varphi(t) = (1 - \tau t)^2 + \tau^2 \gamma_3 t.$$

Исследуем функцию $\varphi(t)$. Эта функция может принимать максимальное значение только на концах отрезка $[\gamma_1, \gamma_2]$. Поэтому

$$\rho^2(S) \leq \begin{cases} \varphi_1(\tau) = (1 - \tau \gamma_1)^2 + \tau^2 \gamma_1 \gamma_3, & 0 \leq \tau \leq \tau_0^*, \\ \varphi_2(\tau) = (1 - \tau \gamma_2)^2 + \tau^2 \gamma_2 \gamma_3, & \tau \geq \tau_0^*. \end{cases}$$

Выберем параметр τ из условия минимума оценки для $\rho(S)$. Так как функция $\varphi_2(\tau)$ возрастает по τ при $\tau \geq \tau_0^*$:

$$\varphi_2'(\tau) = 2\gamma_2 [\tau(\gamma_2 + \gamma_3) - 1] \geq 2\gamma_2 \frac{\gamma_2 - \gamma_1 + \gamma_3}{\gamma_2 + \gamma_1 + \gamma_3} > 0,$$

то минимум $\rho(S)$ следует искать в области $\tau \leq \tau_0^*$, где для $\rho(S)$ выполняется оценка $\rho^2(S) \leq \varphi_1(\tau)$.

Функция $\varphi_1(\tau)$ при $\tau = \tau_1 = 1/(\gamma_1 + \gamma_3) = \kappa \tau_0^*$ достигает минимального значения, причем при $\tau \leq \tau_1$ функция $\varphi_1(\tau)$ убывает, а при $\tau \geq \tau_1$ — возрастает. Поэтому минимальное значение $\varphi_1(\tau)$ на отрезке $[0, \tau_0^*]$ достигается в точке $\tau = \tau_1$, если $\tau_1 \leq \tau_0^*$, и в точке $\tau = \tau_0^*$, если $\tau_1 \geq \tau_0^*$.

Итак, найдено оптимальное значение параметра τ :

$$\tau = \min(\tau_0^*, \kappa \tau_0^*).$$

При этом

$$\min_{\tau} \rho^2(S) \leq \begin{cases} \varphi_1(\tau_0^*), & \kappa \geq 1, \\ \varphi_1(\tau_1), & \kappa \leq 1. \end{cases}$$

Вычислим $\varphi_1(\tau_0^*)$ и $\varphi_1(\tau_1)$. Несложные вычисления дают $\tau_0^* = (2 - 1/\kappa)/\gamma_2$, $\gamma_3 = 1/(\kappa \tau_0^*) - \gamma_1$. Используя эти соотношения, получим

$$\begin{aligned} \varphi_1(\tau_0^*) &= (1 - \tau_0^* \gamma_1)^2 + (\tau_0^*)^2 \gamma_1 \gamma_3 = 1 - 2\tau_0^* \gamma_1 + \tau_0^* \gamma_1 / \kappa = \\ &= 1 - (2 - 1/\kappa)^2 \gamma_1 / \gamma_2 = 1 - (2 - 1/\kappa)^2 (1 - \rho_0) / (1 + \rho_0). \end{aligned}$$

Далее

$$\begin{aligned} \varphi_1(\tau_1) &= (1 - \tau_0^* \kappa \gamma_1)^2 + (\tau_0^*)^2 \kappa^2 \gamma_1 \gamma_3 = 1 - \tau_0^* \kappa \gamma_1 = \\ &= 1 - (2\kappa - 1) \gamma_1 / \gamma_2 = 1 - (2\kappa - 1) (1 - \rho_0) / (1 + \rho_0). \end{aligned}$$

Итак, числовой радиус оценен. Оценка леммы следует из неравенства $\|S^n\| \leq 2 [\rho(S)]^n$. Лемма доказана.

Подставляя в неравенства (27) $C = D^{-1/2} (DB^{-1}A) D^{-1/2}$ и $C_1 = 0,5D^{-1/2} \times (DB^{-1}A - (DB^{-1}A)^*) D^{-1/2}$, получим неравенства

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \tag{28}$$

$$\left(D^{-1} \frac{DB^{-1}A - (DB^{-1}A)^*}{2} x, \frac{DB^{-1}A - (DB^{-1}A)^*}{2} x \right) \leq \gamma_3 (DB^{-1}Ax, x).$$

Теорема 6. Пусть γ_1, γ_2 и γ_3 — постоянные в неравенствах (28). Метод простой итерации (2) при значении итерационного параметра $\tau =$

$\min(\tau_0^*, \kappa\tau_0^*)$ сходится в H_D , и для погрешности z_n имеет место оценка (26). Для числа итераций верна оценка $n \geq n_0(\varepsilon)$, где

$$n_0(\varepsilon) = \ln 0.5\varepsilon / \ln \rho,$$

а κ , ρ и τ_0^* определены в лемме 6.

Приведем вид неравенств (28) для некоторых примеров выбора оператора D . Если в качестве оператора D взять оператор A^*A или B^*B , то неравенства (28) можно записать в следующем виде:

$$\begin{aligned} \gamma_1(Bx, Bx) &\leq (Ax, Bx) \leq \gamma_2(Bx, Bx), \quad \gamma_1 > 0, \\ \|0.5(AB^{-1} - (B^*)^{-1}A^*)x\|^2 &\leq \gamma_3(A^*x, B^{-1}x). \end{aligned} \quad (29)$$

Действительно, неравенства (29) следуют непосредственно из (28) после подстановки $D = B^*B$ в (29). Для случая $D = A^*A$ в (28) достаточно сделать замену $x = A^{-1}By$.

Если оператор B самосопряжен положительно определен в H и ограничен, то в качестве оператора D можно взять операторы B или $A^*B^{-1}A$. В этом случае неравенства (28) будут иметь вид

$$\begin{aligned} \gamma_1 B &\leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \\ (B^{-1}A_1x, A_1x) &\leq \gamma_3(Ax, x), \quad A_1 = 0.5(A - A^*). \end{aligned} \quad (30)$$

Отметим, что в случае $D = A^*B^{-1}A$ неравенства (30) следуют из (28) после указанной выше замены.

4. Метод симметризации уравнения. При решении уравнения $Au = f$ с несамосопряженным оператором A используют хорошо известный прием *симметризации уравнения*. Вместо исходного уравнения рассматривается симметризованное уравнение

$$\tilde{A}u = \tilde{f}, \quad \tilde{A} = A^*A, \quad \tilde{f} = A^*f, \quad (31)$$

которое получается из исходного уравнения умножением слева на сопряженный к A оператор. В алгебре такое преобразование уравнения называется *первой трансформацией Гаусса*.

Для приближенного решения уравнения (31) рассмотрим неявную двухслойную схему

$$\tilde{B} \frac{y_{k+1} - y_k}{\tau_{k+1}} + \tilde{A}y_k = \tilde{f}, \quad k = 0, 1, \dots, y_0 \in H, \quad (32)$$

с самосопряженным положительно определенным оператором \tilde{B} . В качестве оператора D выберем операторы \tilde{B} или $\tilde{A} = A^*A$. В этом случае оператор $D\tilde{B}^{-1}\tilde{A}$ самосопряжен в H , поэтому итерационные параметры τ_k могут быть выбраны по формулам чебышевского метода, исследованного в § 2. Априорная информация для этого метода в случае указанных операторов D имеет вид постоянных энергетической эквивалентности операторов \tilde{B} и $\tilde{A} = A^*A$

$$\gamma_1 \tilde{B} \leq A^*A \leq \gamma_2 \tilde{B}, \quad \gamma_1 > 0.$$

Оценка скорости сходимости чебышевского метода (32) и формулы для итерационных параметров даны в теореме 1.

§ 5. Примеры применения итерационных методов

1. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Для иллюстрации применения построенных в этой главе двухслойных итерационных методов рассмотрим решение разностной задачи Дирихле для линейных эллиптических уравнений второго порядка. Разностную задачу будем трактовать как операторное уравнение

$$Au = f \quad (1)$$

в конечномерном пространстве сеточных функций. Будут рассмотрены явный и неявный чебышевский методы, а также метод простой итерации.

Рассмотрение примеров начнем с задачи Дирихле для уравнения Пуассона в прямоугольнике. Пусть в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ с границей Γ требуется найти решение уравнения Пуассона

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in G, \quad (2)$$

принимающее на границе Γ заданные значения

$$u(x) = g(x), \quad x \in \Gamma. \quad (3)$$

Соответствующая (2), (3) разностная задача Дирихле на прямоугольной сетке

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha / N_\alpha, \alpha = 1, 2\}$$

имеет вид

$$\Lambda y = \sum_{\alpha=1}^2 y_{\bar{x}_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \quad (4)$$

где $\gamma = \{x_{ij} \in \Gamma\}$ — граница сетки $\bar{\omega}$, а

$$y_{\bar{x}_1 x_1} = \frac{1}{h_1^2} (y(i+1, j) - 2y(i, j) + y(i-1, j)),$$

$$y_{\bar{x}_2 x_2} = \frac{1}{h_2^2} (y(i, j+1) - 2y(i, j) + y(i, j-1)),$$

$$y(i, j) = y(x_{ij}).$$

В § 2 гл. V было показано, что разностная задача (4) может быть сведена к операторному уравнению (1), для которого оператор A определяется следующим образом: $Ay = -\Lambda y$, где $y \in H$, $\dot{y} \in \dot{H}$ и $\dot{y}(x) = y(x)$ для $x \in \omega$. Здесь \dot{H} — множество сеточных функций, заданных на ω и обращающихся в нуль на γ , а H — пространство сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2$. Правая часть f

уравнения (1) отличается от правой части φ разностного уравнения (4) лишь в приграничных узлах:

$$f(x) = \varphi(x) + \varphi_1(x)/h_1^2 + \varphi_2(x)/h_2^2,$$

$$\varphi_1(x) = \begin{cases} g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leq x_1 \leq l_1 - 2h_1, \\ g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leq x_2 \leq l_2 - 2h_2, \\ g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

Итак, разностная краевая задача (4) сведена к операторному уравнению (1) в конечномерном гильбертовом пространстве H .

Для приближенного решения уравнения (1) рассмотрим явный чебышевский метод ($B = E$):

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (5)$$

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad i = 1, 2, \dots, n \right\}, \quad (6)$$

$$k = 1, 2, \dots, n,$$

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \ln(0.5\varepsilon)/\ln \rho_1. \quad (7)$$

В § 2 гл. V было показано, что определенный здесь оператор A является самосопряженным в H и его границы γ_1 и γ_2 совпадают с минимальным и максимальным собственными значениями разностного оператора Λ , т. е.

$$\gamma_1 E \leq A \leq \gamma_2 E, \quad \gamma_1 > 0, \quad (8)$$

где

$$\gamma_1 = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \gamma_2 = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}. \quad (9)$$

Операторы A и $B = E$ самосопряжены и положительно определены в H . Поэтому из рассмотренных в п. 3 § 2 примеров следует, что γ_1 , γ_2 из (8) являются постоянными для чебышевского метода (5)–(7), если в качестве D выбран один из операторов E , A или A^2 . Тогда в формулах (6), (7)

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

где γ_1 и γ_2 определены в (9).

Так как $\gamma_1 = O(1)$, а $\gamma_2 = O(1/h_1^2 + 1/h_2^2)$, то $\xi = O(|h|^2)$, где $|h|^2 = h_1^2 + h_2^2$. Следовательно, для рассматриваемого примера асимптотическая оценка числа итераций $n_0(\varepsilon)$ имеет вид

$$n_0(\varepsilon) = O\left(\frac{1}{|h|} \ln \frac{2}{\varepsilon}\right).$$

В частном случае, когда \bar{G} есть квадрат со стороной l ($l_1 = l_2 = l$) и сетка ω квадратная ($h_1 = h_2 = h = l/N$), имеем

$$\begin{aligned}\gamma_1 &= \frac{8}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \gamma_2 = \frac{8}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \xi = \operatorname{tg}^2 \frac{\pi h}{2l}, \\ \tau_0 &= \frac{h^2}{4}, \quad \rho_0 = \cos \frac{\pi h}{l}, \quad \rho_1 = \frac{1 - \sin \frac{\pi h}{l}}{\cos \frac{\pi h}{l}}, \\ n_0(\varepsilon) &\approx \frac{l}{\pi h} \ln \frac{2}{\varepsilon} \approx 0,32N \ln \frac{2}{\varepsilon}. \end{aligned} \quad (10)$$

Таким образом, число итераций n пропорционально числу узлов N по одному направлению. Отметим, что число неизвестных в задаче (4) равно $M = (N-1)^2$, т. е. число итераций пропорционально квадратному корню из числа неизвестных.

Итерационную операторную схему (5) при $B = E$ можно записать, используя определение оператора A и правой части f , в виде следующей разностной схемы:

$$y_{k+1} = y_k + \tau_{k+1} (\Lambda y_k + \varphi), \quad x \in \omega, \quad y_k|_\gamma = g, \quad k = 0, 1, \dots$$

Подставляя сюда (4), получим расчетные формулы

$$\begin{aligned}y_{k+1}(i, j) &= \left(1 - \frac{\tau_{k+1}}{\tau_0}\right) y_k(i, j) + \\ &+ \tau_{k+1} \left[\frac{y_k(i+1, j) + y_k(i-1, j)}{h_1^2} + \frac{y_k(i, j+1) + y_k(i, j-1)}{h_2^2} + \varphi(i, j) \right], \\ 1 \leqslant i &\leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1.\end{aligned}$$

Начальное приближение y_0 есть произвольная сеточная функция на ω , принимающая на границе γ заданные значения $y_0(x) = g(x)$ для $x \in \gamma$.

Оценим число арифметических действий $Q(\varepsilon)$, которое необходимо затратить, чтобы получить приближенное решение разностной задачи (4) с точностью ε по чебышевскому методу (5)–(7).

Считая заданными итерационные параметры τ_k , получим, что для вычисления y_{k+1} в одном узле сетки ω потребуется девять арифметических операций. Так как число внутренних узлов сетки ω равно $M = (N_1-1)(N_2-1)$, то на реализацию одного итерационного шага потребуется $Q_0 \approx 9N_1N_2$ действий. Поэтому $Q(\varepsilon) = nQ_0 \approx 9nN_1N_2$, где n – число итераций.

Для рассмотренного выше частного случая число итераций n определено в (10), и следовательно, для этого примера получим

$$Q(\varepsilon) \approx 2,9N^3 \ln(2/\varepsilon).$$

Для решения уравнения (1) рассмотрим теперь метод простой итерации. Итерационная схема метода простой итера-

ции имеет вид (5), а итерационные параметры τ_k и число итераций n определяются по формулам теоремы 2:

$$\tau_k \equiv \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad n \geq n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho_0}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad (11)$$

где γ_1 и γ_2 заданы в (9). Из (9) и (11) получим асимптотическую по h оценку числа итераций для метода простой итерации $n_0(\varepsilon) = O\left(\frac{1}{|h|^2} \ln \frac{1}{\varepsilon}\right)$. Для рассмотренного выше частного случая найдем

$$n_0(\varepsilon) \approx \frac{2l^2}{\pi^2 h^2} \ln \frac{1}{\varepsilon} \approx 0,2 N^2 \ln \frac{1}{\varepsilon}, \quad (12)$$

т. е. число итераций для метода простой итерации пропорционально квадрату числа узлов N по одному направлению (или пропорционально числу неизвестных в уравнении).

Сравнивая оценки для числа итераций чебышевского метода (10) и метода простой итерации (12), получим, что метод простой итерации требует значительно большего числа итераций, чем чебышевский метод. Для сравнения этих методов на реальных сетках приведем точное значение числа итераций для указанного частного случая в зависимости от числа узлов N по одному направлению для $\varepsilon = 10^{-4}$ (первым указано число итераций для чебышевского метода):

$$\begin{aligned} N &= 32 & n &= 101 & n &= 1909 \\ N &= 64 & n &= 202 & n &= 7642 \\ N &= 128 & n &= 404 & n &= 30577. \end{aligned}$$

Приведем расчетные формулы метода простой итерации для рассматриваемого частного случая:

$$y_{k+1}(i, j) = \frac{1}{4} [y_k(i+1, j) + y_k(i-1, j) + y_k(i, j+1) + y_k(i, j-1)] + \frac{h^2}{4} \Phi(i, j).$$

Очень сильная зависимость числа итераций метода простой итерации от числа узлов сетки N является причиной того, что в настоящее время этот метод почти не используется для решения сеточных эллиптических уравнений.

2. Разностная задача Дирихле для уравнения Пуассона в произвольной области. Пусть в произвольной ограниченной области G с границей Γ требуется найти решение уравнения Пуассона

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in G, \quad (13)$$

принимающее на границе Γ заданные значения

$$u(x) = g(x), \quad x \in \Gamma. \quad (14)$$

Для простоты рассмотрим случай, когда пересечение области G с прямой, проходящей через точку $x \in G$ и параллельной оси координат, состоит из одного интервала.

Покроем плоскость решеткой, образованной пересечением прямых, параллельных осям координат и проведенных на одинаковом расстоянии h друг от друга.

Точки решетки x_{ij} , принадлежащие G , назовем *узлами* сетки $\omega = \{x_{ij} \in G\}$. Через $\Delta_\alpha(x_{ij})$ обозначим интервал, образующийся при пересечении G с прямой, проведенной через точку $x_{ij} \in \omega$ параллельно оси координат Ox_α , $\alpha = 1, 2$. Концы этого интервала назовем *границыми узлами* по направлению x_α . Множество всех границных узлов по направлению x_α обозначим через γ_α , а через $\gamma = \gamma_1 \cup \gamma_2$ обозначим границу сеточной области. Множества внутренних и границных узлов образуют сетку $\bar{\omega} = \omega \cup \gamma$ в области \bar{G} .

Рассмотрим один из интервалов Δ_α . Множество узлов $x_{ij} \in \omega$, лежащих на этом интервале, обозначим через $\omega_\alpha(x_\beta)$, $\beta = 3 - \alpha$, $\alpha = 1, 2$. Через $\omega_\alpha^+(x_\beta)$ обозначим множество, состоящее из узлов $\omega_\alpha(x_\beta)$ и правого конца интервала Δ_α . Определим $\bar{\omega}_\alpha(x_\beta)$ как множество, состоящее из узлов $\omega_\alpha(x_\beta)$ и концов интервала Δ_α . Обозначим через $x_{ij}^{(+1)\alpha}$ и $x_{ij}^{(-1)\alpha}$ узлы, ближайшие к точке $x_{ij} \in \omega_\alpha(x_\beta)$ соответственно справа и слева и принадлежащие $\omega_\alpha(x_\beta)$.

Шагами $h_\alpha^\pm(x_{ij})$ сетки ω в точке $x_{ij} \in \omega$ будем называть расстояние между узлами x_{ij} и $x_{ij}^{(\pm 1)\alpha} \in \bar{\omega}$. Отметим, что если все четыре соседних к x_{ij} узла $x_{ij}^{(\pm 1)\alpha}$ принадлежат ω , то шаги h_α^\pm равны основному шагу решетки h . В приграничных узлах $h_\alpha^\pm \leq h$. Между шагами h_α^+ и h_α^- имеет место соотношение

$$h_\alpha^+(x_{ij}) = h_\alpha^-(x_{ij}^{(+1)\alpha}).$$

Задаче (13), (14) поставим в соответствие на сетке $\bar{\omega}$ разностную краевую задачу

$$\Lambda y = \sum_{\alpha=1}^2 y_{\hat{x}_\alpha \hat{x}_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \quad (15)$$

где

$$\begin{aligned} y_{\hat{x}_\alpha} &= \frac{1}{h_\alpha^-} (y - y^{-1}\alpha), & y_{\hat{x}_\alpha} &= \frac{1}{h_\alpha^+} (y^{+1}\alpha - y), \\ y_{\hat{x}_\alpha \hat{x}_\alpha} &= \frac{1}{h} (y^{+1}\alpha - y), & y_{\hat{x}_\alpha \hat{x}_\alpha} &= \frac{1}{h} \left(\frac{y^{+1}\alpha - y}{h_\alpha^+} - \frac{y - y^{-1}\alpha}{h_\alpha^-} \right), \\ y^{\pm 1}\alpha &= y(x^{(\pm 1)\alpha}), & \alpha &= 1, 2. \end{aligned}$$

Разностная задача (15) сводится к операторному уравнению (1) и оператор A определяется таким же образом, как и в п. 1. Скалярное произведение в H задается следующим образом:

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h^2.$$

Введем некоторые обозначения, которыми будем пользоваться в дальнейшем. Определим для сеточных функций, заданных на ω , скалярные произведения по формулам:

$$(u, v)_{\omega_\alpha} = \sum_{x_\alpha \in \omega_\alpha (x_\beta)} u(x)v(x)h,$$

$$(u, v)_{\omega_\alpha^+} = \sum_{x_\alpha \in \omega_\alpha^+ (x_\beta)} u(x)v(x)h_\alpha^-(x),$$

$$(u, v)_\alpha = ((u, v)_{\omega_\alpha^+}, 1)_{\omega_\beta}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Используя эти обозначения, скалярное произведение в H можно записать в виде

$$(u, v) = ((u, v)_{\omega_1}, 1)_{\omega_2} = ((u, v)_{\omega_2}, 1)_{\omega_1}. \quad (16)$$

Из определения оператора A получим, что

$$(Au, v) = -(\Lambda \dot{u}, \dot{v}) =$$

$$= -((\dot{u}_{\bar{x}_1 \hat{x}_1}, \dot{v})_{\omega_1}, 1)_{\omega_2} - ((\dot{u}_{\bar{x}_2 \hat{x}_2}, \dot{v})_{\omega_2}, 1)_{\omega_1},$$

а так как в силу разностных формул Грина имеет место равенство (см. п. 3 § 2 гл. V)

$$-(\dot{u}_{\bar{x}_\alpha \hat{x}_\alpha}, \dot{v})_{\omega_\alpha} = (\dot{u}_{\bar{x}_\alpha}, \dot{v}_{\bar{x}_\alpha})_{\omega_\alpha^+} = -(\dot{u}, \dot{v}_{\bar{x}_\alpha \hat{x}_\alpha})_{\omega_\alpha},$$

то отсюда вытекают равенства

$$(Au, v) = (u, Av),$$

$$(Au, u) = \sum_{\alpha=1}^2 (\dot{u}_{\bar{x}_\alpha}^2, 1)_\alpha, \quad u \in H, \quad \dot{u} \in \dot{H}. \quad (17)$$

Первое из этих равенств доказывает самосопряженность в H оператора A .

Для приближенного решения уравнения (1) рассмотрим неявный чебышевский метод

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H,$$

где в качестве оператора B возьмем легко обратимый диагональный сператор

$$By = (b_1 + b_2)y, \quad b_\alpha(x) = \frac{1}{h} \left(\frac{1}{h_\alpha^+(x)} + \frac{1}{h_\alpha^-(x)} \right), \quad \alpha = 1, 2. \quad (18)$$

Поясним выбор оператора B . Если трактовать уравнение (1) как систему линейных алгебраических уравнений с матрицей \mathcal{A} , соответствующей оператору A , то матрица \mathcal{B} , соответствующая оператору B , есть диагональная часть матрицы \mathcal{A} .

Так как операторы A и B являются самосопряженными и положительно определенными в H , то γ_1 и γ_2 , входящие в условия (6), (7), являются постоянными энергетической эквивалентности операторов A и B :

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0,$$

если в качестве D выбран один из операторов A , B или $AB^{-1}A$.

Найдем оценки для γ_1 и γ_2 . Сначала покажем, что имеет место равенство

$$\gamma_1 + \gamma_2 = 2. \quad (19)$$

Действительно, пусть $u(x)$ — произвольная сеточная функция из H . Рассмотрим функцию $v(x)$, которую определим следующим образом:

$$v(x_{ij}) = (-1)^{i+j} u(x_{ij}), \quad x_{ij} \in \omega.$$

Вычислим значение разностного оператора Λv в точке x_{ij} . Получим

$$\begin{aligned} \Lambda v(i, j) &= \sum_{\alpha=1}^2 \frac{1}{h} \left(\frac{\overset{\circ}{v}^{+\alpha} - \overset{\circ}{v}}{h_\alpha^+} - \frac{\overset{\circ}{v} - \overset{\circ}{v}^{-\alpha}}{h_\alpha^-} \right)_{x=x_{ij}} = \\ &= -(-1)^{i+j} \sum_{\alpha=1}^2 \frac{1}{h} \left(\frac{\overset{\circ}{u}^{+\alpha} - \overset{\circ}{u}}{h_\alpha^+} - \frac{\overset{\circ}{u} - \overset{\circ}{u}^{-\alpha}}{h_\alpha^-} \right)_{x=x_{ij}} - \\ &\quad - 2(-1)^{i+j} (b_1(x_{ij}) + b_2(x_{ij})) \overset{\circ}{u}(x_{ij}). \end{aligned}$$

Следовательно,

$$Av(i, j) = -\Lambda v(i, j) = (-1)^{i+j} (2B - A) u(i, j).$$

Далее, так как

$$\gamma_1 = \min_{u \neq 0} \frac{(Au, u)}{(Bu, u)}, \quad (Av, v) = 2(Bu, u) - (Au, u), \quad (Bv, v) = (Bu, u),$$

то

$$\gamma_2 = \max_{v \neq 0} \frac{(Av, v)}{(Bv, v)} = 2 - \min_{u \neq 0} \frac{(Au, u)}{(Bu, u)} = 2 - \gamma_1.$$

Утверждение доказано.

Используя соотношение (19), получим, что в формулах (6)

$$\tau_0 = 2/(\gamma_1 + \gamma_2) = 1, \quad \rho_0 = (\gamma_2 - \gamma_1)/(\gamma_2 + \gamma_1) = 1 - \gamma_1.$$

Следовательно, для вычисления итерационных параметров τ_k достаточно найти оценку для γ_1 . Из леммы 13 § 2 гл. V полу-

чим, что для любой сеточной функции $\dot{y} \in \dot{H}$ имеет место неравенство

$$(b_\alpha \dot{y}, \dot{y})_{\omega_\alpha} \leq \kappa_\alpha (\dot{y}_{x_\alpha}^2, 1)_{\omega_\alpha^+}, \quad \alpha = 1, 2, \quad (20)$$

где $\kappa_\alpha = \kappa_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha(x_\beta)} v^\alpha(x)$, а $v^\alpha(x)$ есть решение следующей трехточечной краевой задачи:

$$\begin{aligned} v_{x_\alpha}^\alpha \hat{x}_\alpha &= -b_\alpha(x), & x_\alpha \in \omega_\alpha(x_\beta), \\ v^\alpha(x) &= 0, & x_\alpha \in \gamma_\alpha. \end{aligned} \quad (21)$$

Разделив неравенство (20) на κ_α и суммируя по ω_β , получим

$$\left(\frac{b_\alpha}{\kappa_\alpha} \dot{y}, \dot{y} \right) \leq ((\dot{y}_{x_\alpha}^2, 1)_{\omega_\alpha^+}, 1)_{\omega_\beta} = (\dot{y}_{x_\alpha}^2, 1)_\alpha, \quad \alpha = 1, 2.$$

Складывая эти неравенства, найдем

$$\left(\sum_{\alpha=1}^2 \frac{b_\alpha}{\kappa_\alpha} \dot{y}, \dot{y} \right) \leq \sum_{\alpha=1}^2 (\dot{y}_{x_\alpha}^2, 1)_\alpha. \quad (22)$$

Из (17), (18) и (22) следует, что в качестве γ_1 можно взять величину

$$\gamma_1 = \min_{x \in \omega} \frac{1}{b_1(x) + b_2(x)} \sum_{\alpha=1}^2 \frac{b_\alpha(x)}{\kappa_\alpha(x_\beta)}. \quad (23)$$

Осталось вычислить κ_α . Для этого найдем решение задачи (21).

Пусть концы интервала Δ_α , на котором расположены узлы сетки $\omega_\alpha(x_\beta)$, есть $l_\alpha(x_\beta)$ и $L_\alpha(x_\beta)$. В силу построения решетки на плоскости шаги h_α^\pm отличны от h лишь для приграничных узлов (см. рис. 3).

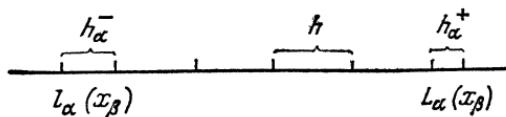


Рис. 3.

Поэтому на сетке $\omega_\alpha(x_\beta)$ разностная производная $v_{x_\alpha}^- \hat{x}_\alpha$ и правая часть уравнения (21) записываются в виде

$$v_{x_\alpha}^- \hat{x}_\alpha = \frac{1}{h} \left(\frac{v^{+\alpha} - v}{h} - \frac{v - v^{-\alpha}}{h_\alpha^-} \right), \quad b_\alpha = \frac{1}{h} \left(\frac{1}{h} + \frac{1}{h_\alpha^-} \right), \quad x_\alpha = l_\alpha + h_\alpha^-,$$

$$v_{x_\alpha}^- \hat{x}_\alpha = \frac{1}{h^2} (v^{+\alpha} - 2v + v^{-\alpha}), \quad b_\alpha = \frac{2}{h^2}, \quad l_\alpha + h_\alpha^- < x_\alpha < L_\alpha - h_\alpha^+,$$

$$v_{x_\alpha}^- \hat{x}_\alpha = \frac{1}{h} \left(\frac{v^{+\alpha} - v}{h_\alpha^+} - \frac{v - v^{-\alpha}}{h} \right), \quad b_\alpha = \frac{1}{h} \left(\frac{1}{h} + \frac{1}{h_\alpha^+} \right), \quad x_\alpha = L_\alpha - h_\alpha^+.$$

Непосредственная проверка показывает, что сеточная функция

$$v^\alpha(x) = \frac{1}{h^2} [(x_\alpha - l_\alpha) \left(L_\alpha - x_\alpha + \frac{(h_\alpha^-)^2 - (h_\alpha^+)^2}{L_\alpha - l_\alpha} \right) + h^2 - (h_\alpha^-)^2]$$

для $x_\alpha \in \omega_\alpha(x_\beta)$ есть решение задачи (21). Так как

$$v^\alpha(x) \leq \frac{1}{h^2} (x_\alpha - l_\alpha)(L_\alpha - x_\alpha) + 1,$$

то

$$\kappa_\alpha = \max_{x_\alpha \in \omega_\alpha(x_\beta)} v^\alpha(x) \leq \frac{1}{h^2} \left(\frac{L_\alpha - l_\alpha}{2} \right)^2 + 1. \quad (24)$$

Подставляя (18) и (24) в (23), найдем оценку для γ_1 .

Грубую оценку для γ_1 можно получить следующим образом. Пусть рассматриваемая область \bar{G} вписана в квадрат со стороной l . Тогда $L_\alpha - l_\alpha \leq l$ для любого α и, следовательно, $\kappa_\alpha \leq \leq l^2/(4h^2) + 1$, $\alpha = 1, 2$. Подставляя эту оценку в (23), получим $\gamma_1 \geq 4h^2/(l^2 + 4h^2)$, т. е. $\gamma_1 \approx 4h^2/l^2$. Так как $\gamma_2 = 2 - \gamma_1$, то $\xi = \gamma_1/\gamma_2 \approx 2h^2/l^2$. Следовательно, из оценки (7) для числа итераций получим

$$n_0(\varepsilon) \approx \frac{l^2}{2\sqrt{2}h} \ln \frac{2}{\varepsilon} \approx 0,35 N \ln \frac{2}{\varepsilon}, \quad (25)$$

где N есть максимальное число узлов по каждому направлению.

Итак, для рассмотренного здесь неявного чебышевского метода число итераций зависит только от основного шага сетки h и не зависит от неравномерных шагов h_α^\pm в приграничных узлах. Сравнивая оценку (25) с полученной ранее оценкой (10), находим, что число итераций для случая произвольной области \bar{G} , вписанной в квадрат со стороной l , такое же, как и для случая, когда область \bar{G} есть указанный квадрат.

Приведем расчетные формулы для чебышевского итерационного метода решения разностной задачи Дирихле для уравнения Пуассона в произвольной области \bar{G} :

$$y_{k+1}(x_{ij}) = (1 - \tau_{k+1}) y_k(x_{ij}) + \\ + \frac{\tau_{k+1}}{b_1(x_{ij}) + b_2(x_{ij})} \left[\frac{1}{h} \left(\frac{y^{+1}_1}{h_1^+} + \frac{y^{-1}_1}{h_1^-} + \frac{y^{+1}_2}{h_2^+} + \frac{y^{-1}_2}{h_2^-} \right) + \varphi \right]_{x=x_{ij}}, \\ x_{ij} \in \omega, \quad y_k(x) = g(x), \quad x \in \gamma.$$

Заметим, что в главе X для рассматриваемой задачи будет построен другой неявный чебышевский метод (попеременно-треугольный итерационный метод), для которого число арифметических действий, затрачиваемых на реализацию одного итерационного шага, несколько больше, чем для рассмотренного здесь метода, а число итераций значительно меньше, что и обеспечивает эффективность указанного метода.

3. Разностная задача Дирихле для эллиптического уравнения с переменными коэффициентами.

3.1. Явный чебышевский метод. Рассмотрим задачу Дирихле для эллиптического уравнения второго порядка с переменными коэффициентами в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$:

$$Lu = \sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x) \frac{\partial u}{\partial x_\alpha} \right) - q(x) u = -f(x), \quad x \in G, \quad (26)$$

$$u(x) = g(x), \quad x \in \Gamma.$$

На прямоугольной сетке

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$$

дифференциальной задаче (26) соответствует разностная задача

$$\Lambda y = \sum_{\alpha=1}^2 (a_\alpha(x) y_{x_\alpha})_{x_\alpha} - d(x) y = -\varphi(x), \quad x \in \omega, \quad (27)$$

$$y(x) = g(x), \quad x \in \gamma.$$

Если коэффициенты $k_\alpha(x)$, $q(x)$ и $f(x)$ являются достаточно гладкими функциями, то коэффициенты $a_\alpha(x)$, $d(x)$ и $\varphi(x)$ разностной схемы (27) можно, например, определить следующим образом:

$$a_1(x_{ij}) = k_1((i-0,5)h_1, jh_2), \quad a_2(x_{ij}) = k_2(ih_1, (j-0,5)h_2),$$

$$d(x_{ij}) = q(x_{ij}), \quad \varphi(x_{ij}) = f(x_{ij}).$$

Будем предполагать, что коэффициенты разностной схемы (27) удовлетворяют условиям

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad x \in \bar{\omega},$$

$$0 \leq d_1 \leq d(x) \leq d_2, \quad x \in \omega, \quad \alpha = 1, 2. \quad (28)$$

Эти условия обеспечивают существование и единственность решения задачи (27).

Разностная схема (27) сводится к операторному уравнению (1) обычным образом: $Ay = -\Lambda y$, где $y \in H$, $\dot{y} \in \dot{H}$, а H — пространство сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2,$$

правая часть f уравнения (1) отличается от правой части фокусной (27) лишь в приграничных узлах.

Для приближенного решения уравнения (1) применим явный чебышевский метод (5) — (7) ($B = E$). В § 2 гл. V было показано, что определенный здесь оператор A является самосопряженным в H . Поэтому априорная информация для чебышевского метода

имеет вид постоянных γ_1 и γ_2 из неравенств $\gamma_1 E \leq A \leq \gamma_2 E$, $\gamma_1 > 0$, если в качестве D выбран один из операторов E , A или A^2 . Найдем эти постоянные. Для этого введем оператор \hat{A} , соответствующий разностному оператору $\hat{\Lambda}$, где $\hat{\Lambda}y = y_{\bar{x}_1x_1} + y_{\bar{x}_2x_2}$, и определим следующие скалярные произведения для сеточных функций, заданных на $\bar{\omega}$:

$$(u, v)_{\omega_\alpha} = \sum_{x_\alpha \in \omega_\alpha} u(x)v(x)h_\alpha, \quad (u, v)_{\omega_\alpha^+} = \sum_{x_\alpha \in \omega_\alpha^+} u(x)v(x)h_\alpha,$$

$$(u, v)_\alpha = ((u, v)_{\omega_\alpha^+}, 1)_{\omega_\beta}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Здесь

$$\omega_\alpha = \{x_\alpha, i = ih_\alpha, 1 \leq i \leq N_\alpha - 1\}, \quad \omega_\alpha^+ = \{x_\alpha, i = ih_\alpha, 1 \leq i \leq N_\alpha\}.$$

Введенные здесь скалярные произведения являются аналогом скалярных произведений, определенных в п. 2.

Из определения операторов A и \hat{A} и разностных формул Грина (см. п. 2 § 2 гл. V) получим

$$(Au, u) = -(\Lambda \dot{u}, \dot{u}) = -((a_1 \dot{u}_{\bar{x}_1 x_1}, \dot{u})_{\omega_1}, 1)_{\omega_2} - ((a_2 \dot{u}_{\bar{x}_2 x_2}, \dot{u})_{\omega_2}, 1)_{\omega_1} + \\ + (du, u) = \sum_{\alpha=1}^2 (a_\alpha \dot{u}_{\bar{x}_\alpha}^2, 1)_\alpha + (du, u), \quad (29)$$

$$(\hat{A}u, u) = -(\hat{\Lambda} \dot{u}, \dot{u}) = \sum_{\alpha=1}^2 (\dot{u}_{\bar{x}_\alpha}^2, 1)_\alpha, \quad u \in H, \quad \dot{u} \in \dot{H}.$$

Учитывая неравенства (28), отсюда получим операторные неравенства вида

$$c_1 \hat{A} + d_1 E \leq A \leq c_2 \hat{A} + d_2 E. \quad (30)$$

В п. 1 § 5 было показано, что оператор \hat{A} имеет границы

$$\dot{\gamma}_1 = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \dot{\gamma}_2 = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha},$$

т. е. имеют место неравенства

$$\dot{\gamma}_1 E \leq \hat{A} \leq \dot{\gamma}_2 E. \quad (31)$$

Из (30) и (31) найдем, что оператор A имеет границы $\gamma_1 = c_1 \dot{\gamma}_1 + d_1$, $\gamma_2 = c_2 \dot{\gamma}_2 + d_2$.

Итак, постоянные γ_1 и γ_2 найдены. Используя их, вычислим по формулам (6) итерационные параметры τ_k , а по формулам (7) найдем оценку для числа итераций n .

Так как $\xi = \dot{\gamma}_1 / \dot{\gamma}_2 = O(|h|^2)$, то $\xi = \gamma_1 / \gamma_2 = O(|h|^2)$ и для числа итераций рассматриваемого метода имеет место следующая асимптотическая оценка:

$$n_0(\varepsilon) = O\left(\frac{1}{|h|} \ln \frac{2}{\varepsilon}\right),$$

а константа в оценке зависит от экстремальных характеристик коэффициентов $a_\alpha(x)$ и $d(x)$, т. е. от c_α и d_α , $\alpha=1, 2$.

В частном случае, когда область \bar{G} есть квадрат со стороной l ($l_1=l_2=l$), сетка ω квадратная ($h_1=h_2=h=l/N$) и $d=0$, получим

$$\gamma_1 = \frac{8c_1}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \gamma_2 = \frac{8c_2}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \xi = \frac{c_1}{c_2} \operatorname{tg}^2 \frac{\pi h}{2l}$$

и, следовательно,

$$n_0(\varepsilon) \approx \sqrt{\frac{c_2}{c_1}} \frac{l}{\pi h} \ln \frac{2}{\varepsilon} \approx 0,32 \sqrt{\frac{c_2}{c_1}} N \ln \frac{2}{\varepsilon}.$$

Сравнивая полученную здесь оценку для числа итераций явного чебышевского метода решения разностного уравнения (27) с переменными коэффициентами с оценкой (10), находим, что для рассматриваемого примера число итераций в $\sqrt{c_2/c_1}$ раз больше числа итераций для случая постоянных коэффициентов.

Приведем расчетные формулы для явного чебышевского метода (5)–(7), используемого для решения разностного уравнения (27). Эти формулы имеют вид:

$$\begin{aligned} y_{k+1}(i, j) = & \\ & = \alpha_{k+1}(i, j) y_k(i, j) + \tau_{k+1} \left\{ \frac{1}{h_1^2} [a_1(i+1, j) y_k(i+1, j) + \right. \\ & \quad \left. + a_1(i, j) y_k(i-1, j)] + \frac{1}{h_2^2} [a_2(i, j+1) y_k(i, j+1) + \right. \\ & \quad \left. + a_2(i, j) y_k(i, j-1)] + \varphi(i, j) \right\}, \\ & 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \end{aligned}$$

где обозначено

$$\begin{aligned} \alpha_{k+1}(i, j) = & \\ & = 1 - \tau_{k+1} \left[\frac{a_1(i+1, j) + a_1(i, j)}{h_1^2} + \frac{a_2(i, j+1) + a_2(i, j)}{h_2^2} + d(i, j) \right], \end{aligned}$$

а начальное приближение y_0 есть произвольная на ω сеточная функция, принимающая на γ заданные значения: $y(x) = g(x)$ для $x \in \gamma$.

3.2. Неявный чебышевский метод. Для приближенного решения построенного в предыдущем пункте уравнения (1), соответствующего разностной схеме (27), рассмотрим теперь простейший неявный чебышевский метод (5)–(7). В качестве оператора B , как и в п. 2, снова возьмем диагональную часть оператора A

$$By = by,$$

$$\begin{aligned} b(i, j) = & \\ & = \frac{1}{h_1^2} [a_1(i+1, j) + a_1(i, j)] + \frac{1}{h_2^2} [a_2(i, j+1) + a_2(i, j)] + \\ & \quad + d(i, j). \end{aligned} \tag{32}$$

Так как операторы A и B самосопряжены в H и положительно определены, то априорная информация для неявного чебышевского метода (5)–(7) имеет вид постоянных энергетической эквивалентности операторов $\gamma_1 B \leq A \leq \gamma_2 B$, $\gamma_1 > 0$, если в качестве D выбран один из операторов A , B или $AB^{-1}A$.

Найдем постоянные γ_1 и γ_2 . Точно так же, как и в п. 2, доказывается, что имеет место равенство $\gamma_1 + \gamma_2 = 2$. Поэтому в формулах (6) для итерационных параметров τ_k имеем $\tau_0 = 2/(\gamma_1 + \gamma_2) = 1$, $\rho_0 = (\gamma_2 - \gamma_1)/(\gamma_2 + \gamma_1) = 1 - \gamma_1$.

Оценим γ_1 . Из леммы 14 § 2 гл. V следует, что для любой сеточной функции $\dot{y} \in \dot{H}$ имеет место неравенство

$$(b\dot{y}, \dot{y})_{\omega_\alpha} \leq \kappa_\alpha \left[\left(a_\alpha \dot{y}_{x_\alpha}^{\frac{2}{\alpha}}, 1 \right)_{\omega_\alpha^+} + \frac{1}{2} (d\dot{y}, \dot{y})_{\omega_\alpha} \right], \quad \alpha = 1, 2, \quad (33)$$

где $\kappa_\alpha = \kappa_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x)$, а $v^\alpha(x)$ есть решение следующей трехточечной краевой задачи:

$$\left(a_\alpha v_{x_\alpha}^\alpha \right)_{x_\alpha} - \frac{1}{2} dv^\alpha = -b(x), \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \quad (34)$$

$$v^\alpha(x) = 0, \quad x_\alpha = 0, \quad l_\alpha, \quad h_\beta \leq x_\beta \leq l_\beta - h_\beta, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Из неравенств (33) делением на κ_α и последующим суммированием по ω_β получим

$$\left(\frac{b}{\kappa_\alpha} \dot{y}, \dot{y} \right) \leq \left(a_\alpha \dot{y}_{x_\alpha}^{\frac{2}{\alpha}}, 1 \right) + \frac{1}{2} (d\dot{y}, \dot{y}), \quad \alpha = 1, 2.$$

Складывая эти неравенства и учитывая (29), получим

$$\left(\sum_{\alpha=1}^2 \frac{1}{\kappa_\alpha} b\dot{y}, \dot{y} \right) \leq \sum_{\alpha=1}^2 \left(a_\alpha \dot{y}_{x_\alpha}^{\frac{2}{\alpha}}, 1 \right) + (d\dot{y}, \dot{y}) = (Ay, y).$$

Следовательно, в качестве γ_1 можно взять

$$\gamma_1 = \min_{x \in \omega} \sum_{\alpha=1}^2 \frac{1}{\kappa_\alpha} = \min_{x_2 \in \omega_2} \frac{1}{\kappa_1(x_2)} + \min_{x_1 \in \omega_1} \frac{1}{\kappa_2(x_1)}. \quad (35)$$

Итак, для нахождения γ_1 нужно решить уравнения (34), найти $\kappa_\alpha(x_\beta)$ и по формуле (35) вычислить γ_1 . Постоянная γ_2 находится по формуле $\gamma_2 = 2 - \gamma_1$.

Получим оценку для числа итераций рассматриваемого неявного чебышевского метода. Из теории разностных схем следует, что разностная схема (34) устойчива по правой части в равномерной метрике, т. е. существует такая постоянная M , не зависящая от шагов сетки h_1 и h_2 , что для решения уравнения (34) имеет место оценка

$$\max_{x_\alpha \in \omega_\alpha} v^\alpha(x) \leq M (b^2, 1)_{\omega_\alpha}^{1/2}.$$

Так как $b(x) = O\left(\frac{1}{h^2}\right)$, $h = \min_{\alpha} h_{\alpha}$ для $x \in \omega$, то отсюда получим, что

$$\kappa_{\alpha} = \max_{x_{\alpha} \in \omega_{\alpha}} v^{\alpha}(x) = O\left(\frac{1}{h^2}\right)$$

и, следовательно, $\gamma_1 = O(h^2)$ и $\gamma_2 = O(1)$. Поэтому $\xi = \gamma_1/\gamma_2 = O(h^2)$, а для числа итераций имеет место такая же асимптотическая по h оценка, как и для явного метода

$$n_0(\varepsilon) = O\left(\frac{1}{h} \ln \frac{2}{\varepsilon}\right).$$

В чем же преимущество неявного итерационного метода по сравнению с рассмотренным выше явным чебышевским методом? Ответ на этот вопрос дает следующая теорема, которую мы приведем без доказательства.

Теорема 7*). Для итерационной схемы (5)–(7) с оператором A , соответствующим разностной схеме (27), наилучшим в классе диагональных операторов B (т. е. для которого отношение ξ максимально) является оператор, определенный формулами (32).

Из теоремы 7 следует, что если в качестве оператора B выбрать диагональную часть оператора A , то отношение $\xi = \gamma_1/\gamma_2$ постоянных энергетической эквивалентности γ_1 и γ_2 операторов A и B будет максимальным и, следовательно, число итераций n – минимальным.

Проиллюстрируем преимущество неявного метода на следующем модельном примере. Пусть разностная схема (27) задана на квадратной сетке в единичном квадрате $h_1 = h_2 = h = 1/N$, $l_1 = l_2 = 1$.

Коэффициенты $a_1(x)$, $a_2(x)$ и $d(x)$ выберем следующим образом:

$$\begin{aligned} a_1(x) &= 1 + c[(x_1 - 0,5)^2 + (x_2 - 0,5)^2], \\ a_2(x) &= 1 + c[0,5 - (x_1 - 0,5)^2 - (x_2 - 0,5)^2], \\ d(x) &\equiv 0, \quad c > 0. \end{aligned}$$

При этом в неравенствах (28) имеем $c_1 = 1$, $c_2 = 1 + 0,5c$, $d_1 = d_2 = 0$. Меняя параметр c , мы будем получать коэффициенты разностной схемы (27) с различными экстремальными характеристиками.

Для явного метода было показано, что число итераций зависит от отношения c_2/c_1 . Для неявного метода число итераций зависит не от максимального и минимального значений коэффициентов $a_{\alpha}(x)$, а от некоторых интегральных характеристик этих коэффициентов.

В табл. 7 приведено число итераций для явного и неявного методов в зависимости от отношения c_2/c_1 и от числа узлов N

*) Эта теорема есть частный случай более общей теоремы, доказанной в работе: G. Forsythe, E. G. Straus, On best conditioned matrices, Proc. Amer. Math. Soc. 6 (1955), 340–345.

по одному направлению. Расчеты проведены для $\epsilon = 10^{-4}$. Для случая, когда параметр $c=0$, т. е. $a_\alpha(x) \equiv 1$, и рассматривается уравнение Пуассона, число итераций для явного и неявного метода одинаково и приведено в п. 1.

Из таблицы следует, что для рассматриваемого примера число итераций для неявного метода значительно меньше, чем для явного метода. Зависимость числа итераций от отношения c_2/c_1 более слабая для неявного метода, чем для явного.

Таблица 7

$\frac{c_2}{c_1}$	$N=32$		$N=64$		$N=128$	
	неявный	явный	неявный	явный	неявный	явный
2	123	143	246	286	494	571
8	149	286	305	571	616	1142
32	175	571	365	1142	749	2283
128	192	1141	409	2283	856	4565
512	202	2281	436	4565	926	9130

В заключение приведем расчетные формулы для неявного чебышевского метода:

$$y_{k+1}(i, j) = (1 - \tau_{k+1}) y_k(i, j) + \\ + \frac{\tau_{k+1}}{b(i, j)} \left\{ \frac{1}{h_1^2} [a_1(i+1, j) y_k(i+1, j) + a_1(i, j) y_k(i-1, j)] + \right. \\ \left. + \frac{1}{h_2^2} [a_2(i, j+1) y_k(i, j+1) + a_2(i, j) y_k(i, j-1)] + \varphi(i, j) \right\}, \\ 1 \leq i \leq N_1 - 1 \quad 1 \leq j \leq N_2 - 1,$$

где $b(i, j)$ определено в (32), а начальное приближение y_0 есть произвольная на ω сеточная функция, принимающая на границе γ заданные значения: $y_0(x) = g(x)$, $x \in \gamma$.

Из сравнения расчетных формул для явного и неявного чебышевских методов следует, что число арифметических действий для вычисления y_{k+1} по заданному y_k для обоих методов практически одинаково. Так как число итераций для неявного метода значительно меньше, чем для явного, то следует отдать предпочтение неявному методу.

4. Разностная задача Дирихле для эллиптического уравнения со смешанной производной. В прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ с границей Γ требуется решить задачу Дирихле для эллиптического уравнения со смешанными производными

$$Lu = \sum_{\alpha, \beta=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}(x) \frac{\partial u}{\partial x_\beta} \right) = -f(x), \quad x \in G, \\ u(x) = g(x), \quad x \in \Gamma.$$

Предполагается, что выполнены условия симметрии

$$k_{12}(x) = k_{21}(x), \quad x \in \bar{G}, \quad (36)$$

и эллиптичности

$$c_1 \sum_{\alpha=1}^2 \xi_{\alpha}^2 \leq \sum_{\alpha, \beta=1}^2 k_{\alpha \beta} \xi_{\alpha} \xi_{\beta} \leq c_2 \sum_{\alpha=1}^2 \xi_{\alpha}^2, \quad c_1 > 0, \quad (37)$$

где $\xi = (\xi_1, \xi_2)$ — произвольный вектор.

На прямоугольной сетке

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, \quad 0 \leq i \leq N_1, \quad 0 \leq j \leq N_2, \\ h_{\alpha} = l_{\alpha}/N_{\alpha}, \quad \alpha = 1, 2\}$$

дифференциальной задаче соответствует разностная задача Дирихле

$$\begin{aligned} \Lambda y = 0.5 \sum_{\alpha, \beta=1}^2 [(k_{\alpha \beta} y_{x_{\beta}})_{x_{\alpha}} + (k_{\alpha \beta} y_{x_{\alpha}})_{x_{\beta}}] = -\varphi(x), \quad x \in \omega, \\ y(x) = g(x), \quad x \in \gamma, \end{aligned} \quad (38)$$

где γ — граница сетки $\bar{\omega}$.

В § 2 гл. V было показано, что разностная задача (38) сводится к операторному уравнению (1) обычным образом: $Ay = -\Lambda y$, где $y \in H$, $\dot{y} \in \dot{H}$ и $\dot{y}(x) = y(x)$ для $x \in \omega$. Здесь \dot{H} — множество сеточных функций, заданных на ω и обращающихся в нуль на γ , а H — пространство сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2.$$

Там же было показано, что при выполнении условия (36) построенный оператор A самосопряжен в H и, если выполнены условия (37), имеет границы γ_1 и γ_2 , равные

$$\gamma_1 = c_1 \sum_{\alpha=1}^2 \frac{4}{h_{\alpha}^2} \sin^2 \frac{\pi h_{\alpha}}{2l_{\alpha}}, \quad \gamma_2 = c_2 \sum_{\alpha=1}^2 \frac{4}{h_{\alpha}^2} \cos^2 \frac{\pi h_{\alpha}}{2l_{\alpha}}, \quad (39)$$

т. е.

$$\gamma_1 E \leq A \leq \gamma_2 E. \quad (40)$$

Для приближенного решения уравнения (1), соответствующего разностной схеме (38), рассмотрим явный чебышевский метод (5) — (7) ($B = E$). Так как операторы A и B самосопряжены и положительно определены в H , то априорная информация имеет вид постоянных γ_1 и γ_2 в неравенствах (40), и метод сходится в H_D , где $D = A$, B или $AB^{-1}A$.

Из (39) получим

$$\gamma_1 = O(c_1), \quad \gamma_2 = O\left(\frac{c_2}{h^2}\right), \quad \xi = \frac{\gamma_1}{\gamma_2} = O\left(\frac{c_1}{c_2} h^2\right), \quad h^2 = h_1^2 + h_2^2.$$

Следовательно, для рассматриваемого примера асимптотическая по h оценка числа итераций $n_0(\varepsilon)$ имеет вид

$$n_0(\varepsilon) = O\left(\sqrt{\frac{c_2}{c_1}} \frac{1}{h} \ln \frac{2}{\varepsilon}\right).$$

В частном случае, когда \bar{G} есть квадрат со стороной l и сетка $\bar{\omega}$ квадратная ($h_1 = h_2 = h = l/N$), получим

$$\gamma_1 = \frac{8c_1}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \gamma_2 = \frac{8c_2}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \xi = \frac{c_1}{c_2} \operatorname{tg}^2 \frac{\pi h}{2l},$$

$$n \geq n_0(\varepsilon) \approx \sqrt{\frac{c_2}{c_1}} \frac{l}{\pi h} \ln \frac{2}{\varepsilon} = 0,32 \sqrt{\frac{c_2}{c_1}} N \ln \frac{2}{\varepsilon},$$

т. е. число итераций так же пропорционально числу узлов N по одному направлению, как и в случае уравнения без смешанных производных.

На этом мы закончим рассмотрение примеров применения двухслойных итерационных методов к решению эллиптических уравнений. Более сложные примеры будут рассмотрены в главе XIV.

ГЛАВА VII

ТРЕХСЛОЙНЫЕ ИТЕРАЦИОННЫЕ МЕТОДЫ

В главе изучаются трехслойные итерационные методы решения операторного уравнения $Au = f$. Итерационные параметры выбираются с учетом априорной информации об операторах схемы. В § 1 дается оценка скорости сходимости для трехслойных схем стандартного типа. В §§ 2, 3 рассмотрены полуитерационный метод Чебышева и стационарный трехслойный метод. § 4 посвящен исследованию устойчивости двухслойных и трехслойных методов относительно возмущения априорных данных.

§ 1. Оценка скорости сходимости

1. Исходное семейство итерационных схем. В главе VI для нахождения приближенного решения линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A , действующим в вещественном гильбертовом пространстве H , были построены двухслойные итерационные методы. В этих методах двухслойная схема связывает два итерационных приближения y_{k+1} и y_k .

В данной главе будут изучены трехслойные итерационные схемы. Трехслойная итерационная схема для уравнения (1) связывает три итерационных приближения y_{k+1} , y_k и y_{k-1} , так что y_{k+1} определяется через y_k и y_{k-1} . Для реализации трехслойной схемы должны быть заданы два начальных приближения y_0 и y_1 . Обычно при произвольном y_0 приближение y_1 находится по двухслойной схеме.

Ограничимся изучением трехслойных схем стандартного типа. Неявная стандартная трехслойная итерационная схема имеет вид

$$\begin{aligned} By_{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \alpha_{k+1}\tau_{k+1}f, \\ By_1 &= (B - \tau_1A)y_0 + \tau_1f, \quad k = 1, 2, \dots, y_0 \in H, \end{aligned} \quad (2)$$

где y_0 — произвольное начальное приближение, B — линейный невырожденный оператор, действующий в H , α_k и τ_k — итерационные параметры. Формулами (2) определяется исходное семейство трехслойных итерационных схем.

Нахождение нового приближения y_{k+1} можно трактовать следующим образом. Пусть \bar{y} — промежуточное итерационное приближение, которое находится по неявной двухслойной схеме

$$B \frac{\bar{y} - y_k}{\tau_{k+1}} + A y_k = f.$$

Тогда из (2) следует, что y_{k+1} есть линейная комбинация приближений \bar{y} и y_{k-1}

$$y_{k+1} = \alpha_{k+1} \bar{y} + (1 - \alpha_{k+1}) y_{k-1}.$$

Таким образом, приближение y_{k+1} есть линейная экстраполяция по приближениям \bar{y} и y_{k-1} .

Если положить в (2) $\alpha_k \equiv 1$, то трехслойная схема (2) перейдет в двухслойную схему, сходимость которой была изучена в главе VI. Поэтому введение итерационных параметров α_k позволяет рассчитывать на то, что сходимость схемы (2) будет не хуже сходимости двухслойной схемы.

Отметим, что, в отличие от двухслойной итерационной схемы, реализация трехслойной схемы требует запоминания не одного, а двух итерационных приближений y_k и y_{k-1} .

2. Оценка нормы погрешности. Займемся теперь исследованием сходимости трехслойной схемы (2) в энергетическом пространстве H_D , порождаемом самосопряженным и положительно определенным в H оператором D . Для этого изучим поведение нормы в H_D погрешности $z_k = y_k - u$ при $k \rightarrow \infty$.

Подставляя $y_k = z_k + u$ для $k = 0, 1, \dots$ в (2) и учитывая уравнение (1), найдем уравнение для погрешности z_k :

$$\begin{aligned} Bz_{k+1} &= \alpha_{k+1} (B - \tau_{k+1} A) z_k + (1 - \alpha_{k+1}) B z_{k-1}, & k = 1, 2, \dots, \\ Bz_1 &= (B - \tau_1 A) z_0, \quad z_0 = y_0 - u. \end{aligned}$$

Разрешим это уравнение относительно z_{k+1} и, полагая $z_k = D^{-1/2} x_k$, перейдем к уравнению для эквивалентной погрешности x_k . Уравнение для x_k будет иметь следующий вид:

$$\begin{aligned} x_{k+1} &= \alpha_{k+1} S_{k+1} x_k + (1 - \alpha_{k+1}) x_{k-1}, & k = 1, 2, \dots, \\ x_1 &= S_1 x_0, \quad S_k = E - \tau_k C, \end{aligned} \tag{3}$$

где $C = D^{1/2} B^{-1} A D^{-1/2}$.

В силу сделанной замены $z_k = D^{-1/2} x_k$ справедливо равенство $\|x_k\| = \|z_k\|_D$, и, следовательно, сходимость схемы (2) в H_D будет иметь место, если $\|x_k\| \rightarrow 0$ при $k \rightarrow \infty$.

Изучим поведение нормы x_k в H при $k \rightarrow \infty$. Для этого найдем явный вид решения уравнения (9). Используя формулы (9),

последовательно получим

$$\begin{aligned}x_1 &= (E - \tau_1 C) x_0 = P_1(C) x_0, \\x_2 &= \alpha_2 (E - \tau_2 C) x_1 + (1 - \alpha_2) x_0 = [\alpha_2 (E - \tau_2 C) P_1(C) + \\&\quad + (1 - \alpha_2) E] x_0 = P_2(C) x_0, \\&\dots \\x_{k+1} &= \alpha_{k+1} (E - \tau_{k+1} C) P_k(C) x_0 + (1 - \alpha_{k+1}) P_{k-1}(C) x_0 = \\&= P_{k+1}(C) x_0\end{aligned}$$

и т. д.

Следовательно, решение уравнения (9) для любого k имеет вид

$$x_k = P_k(C) x_0, \quad k = 0, 1, \dots, \quad (4)$$

где $P_k(C)$ — операторный полином степени k относительно оператора C . В силу произвольности x_0 , соответствующий алгебраический полином $P_k(t)$ удовлетворяет следующим рекуррентным соотношениям:

$$\begin{aligned}P_{k+1}(t) &= \alpha_{k+1} (1 - \tau_{k+1} t) P_k(t) + (1 - \alpha_{k+1}) P_{k-1}(t), \\P_1(t) &= 1 - \tau_1 t, \quad P_0(t) \equiv 1, \quad k = 1, 2, \dots\end{aligned} \quad (5)$$

Из (5) следует, что полином $P_k(t)$ для любого k удовлетворяет условию нормировки $P_k(0) = 1$.

Оценим теперь норму x_k . Из (4) получим

$$\|x_k\| = \|P_k(C) x_0\| \leq \|P_k(C)\| \|x_0\|, \quad k = 0, 1, \dots$$

или, в силу сделанной замены $z_k = D^{-1/2} x_k$,

$$\|z_k\|_D \leq \|P_k(C)\| \|z_0\|_D. \quad (6)$$

Итак, оценка нормы погрешности z_k получена. Из (6) следует, что метод будет обладать наибольшей скоростью сходимости, если норма полинома $P_k(C)$ будет стремиться к нулю при $k \rightarrow \infty$ наиболее быстро. Так как полином $P_k(C)$ есть функция итерационных параметров $\tau_1, \tau_2, \dots, \tau_k$ и $\alpha_2, \alpha_3, \dots, \alpha_k$, то эти параметры должны быть выбраны из условия минимума нормы операторного полинома $P_k(C)$. Другими словами, нужно построить полином степени k , нормированный условием $P_k(0) = E$, который имеет минимальную норму.

В главе VI при изучении чебышевского двухслойного метода эта задача была решена в предположении, что оператор $DB^{-1}A$ самосопряжен в H и заданы постоянные энергетической эквивалентности γ_1 и γ_2 самосопряженных операторов D и $DB^{-1}A$:

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \quad (7)$$

При построении трехслойных итерационных методов будем рассматривать только самосопряженный случай, т. е. будем предполагать, что выполнены условия (7).

Пусть выполнены условия (7). Укажем оптимальный оператор $P_k(C)$ и получим априорную оценку для погрешности z_k .

Так как $C = D^{1/2} B^{-1} A D^{-1/2} = D^{-1/2} (DB^{-1}A) D^{-1/2}$, то из (7) следует, что оператор C самосопряжен в H , а γ_1 и γ_2 —его границы:

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \gamma_1 > 0, \quad C = C^*. \quad (8)$$

Тогда в силу (8) для нормы оператора $P_k(C)$ справедлива оценка

$$\|P_k(C)\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} |P_k(t)|.$$

Следовательно, оптимальный полином $P_k(t)$ выделяется следующим условием: максимум модуля этого полинома на отрезке $[\gamma_1, \gamma_2]$ минимален. Из § 2 гл. VI следует, что при условии нормировки $P_k(0) = 1$ искомый полином имеет вид

$$P_k(t) = q_k T_k \left(\frac{1 - \tau_0 t}{\rho_0} \right), \quad q_k = \frac{1}{T_k \left(\frac{1}{\rho_0} \right)}, \quad (9)$$

где $T_k(x)$ —полином Чебышева 1-го рода степени k :

$$T_k(x) = \begin{cases} \cos(k \arccos x), & |x| \leq 1, \\ \operatorname{ch}(k \operatorname{Arch} x), & |x| \geq 1, \end{cases}$$

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad q_k = \frac{2\rho_0^k}{1 + \rho_0^{2k}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

При этом имеет место оценка

$$\|P_k(C)\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} |P_k(t)| = q_k, \quad k = 0, 1, \dots$$

Подставляя эту оценку в (6), получим

$$\|z_k\|_D \leq q_k \|z_0\|_D.$$

Таким образом, скорость сходимости трехслойного итерационного метода (2), параметры τ_k и α_k которого выбираются из условия минимума нормы разрешающего оператора, равна скорости сходимости чебышевского двухслойного итерационного метода.

Формулы (9) дают решение задачи о построении наиболее быстро сходящегося трехслойного итерационного метода. В § 2 будут получены формулы для итерационных параметров τ_k и α_k этого метода, который называется *полуитерационным методом Чебышева*.

§ 2. Полуитерационный метод Чебышева

1. Формулы для итерационных параметров. Найдем теперь формулы для итерационных параметров α_k и τ_k *полуитерационного метода Чебышева*. В § 1, используя трехслойную

итерационную схему метода

$$\begin{aligned}By_{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \alpha_{k+1}\tau_{k+1}f, \\By_1 &= (B - \tau_1A)y_0 + \tau_1f, \quad k = 1, 2, \dots, \quad y_0 \in H,\end{aligned}\quad (1)$$

мы получили уравнение для эквивалентной погрешности

$$\begin{aligned}x_{k+1} &= \alpha_{k+1}(E - \tau_{k+1}C)x_k + (1 - \alpha_{k+1})x_{k-1}, \quad k = 1, 2, \dots, \\x_1 &= (E - \tau_1C)x_0.\end{aligned}\quad (2)$$

Было показано, что для любого k решение этого уравнения имеет вид

$$x_k = P_k(C)x_0, \quad k = 0, 1, \dots, \quad (3)$$

а оптимальный полином $P_k(C)$ определяется формулами

$$P_k(t) = q_k T_k\left(\frac{1 - \tau_0 t}{\rho_0}\right), \quad q_k = \frac{1}{T_k\left(\frac{1}{\rho_0}\right)} = \frac{2\rho_1^k}{1 + \rho_1^{2k}}. \quad (4)$$

Для того чтобы получить формулы для итерационных параметров α_k и τ_k , найдем рекуррентные соотношения, которым удовлетворяет полином $P_k(t)$.

Известно, что для любого x полиномы Чебышева первого рода $T_k(x)$ удовлетворяют следующим рекуррентным соотношениям (см. § 4 гл. I):

$$\begin{aligned}T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x), \quad k = 1, 2, \dots, \\T_1(x) &= x, \quad T_0(x) \equiv 1.\end{aligned}\quad (5)$$

Используя (4) и (5), получим

$$\frac{P_{k+1}(t)}{q_{k+1}} = 2\left(\frac{1 - \tau_0 t}{\rho_0}\right) \frac{P_k(t)}{q_k} - \frac{P_{k-1}(t)}{q_{k-1}}, \quad k = 1, 2, \dots, \quad (6)$$

$$P_1(t)/q_1 = (1 - \tau_0 t)/\rho_0, \quad P_0(t)/q_0 \equiv 1. \quad (7)$$

Из определения (4) и соотношений (5) следует

$$1/q_{k+1} = 2/(\rho_0 q_k) - 1/q_{k-1}, \quad q_1 = \rho_0, \quad q_0 = 1. \quad (8)$$

Отсюда найдем

$$q_{k+1}/q_{k-1} = 2q_{k+1}/(\rho_0 q_k) - 1, \quad k = 1, 2, \dots \quad (9)$$

Подставляя (8), (9) в (6) и (7), получим рекуррентные формулы для полиномов $P_k(t)$:

$$\begin{aligned}P_{k+1}(t) &= \frac{2}{\rho_0} \frac{q_{k+1}}{q_k} (1 - \tau_0 t) P_k(t) + \left(1 - \frac{2}{\rho_0} \frac{q_{k+1}}{q_k}\right) P_{k-1}(t), \\P_1(t) &= 1 - \tau_0 t, \quad P_0(t) \equiv 1, \quad k = 1, 2, \dots\end{aligned}$$

Отсюда и из (3) следуют рекуррентные соотношения для x_k

$$\begin{aligned}x_{k+1} &= \frac{2}{\rho_0} \frac{q_{k+1}}{q_k} (E - \tau_0 C)x_k + \left(1 - \frac{2}{\rho_0} \frac{q_{k+1}}{q_k}\right) x_{k-1}, \quad k = 1, 2, \dots, \\x_1 &= (E - \tau_0 C)x_0.\end{aligned}$$

Сравнивая эти формулы с (2), получим

$$\alpha_{k+1} = 2q_{k+1}/(\rho_0 q_k), \quad \tau_k \equiv \tau_0 = 2/(\gamma_1 + \gamma_2), \quad k = 1, 2, \dots \quad (10)$$

Итак, формулы для итерационных параметров τ_k и α_k получены. Преобразуем формулу для параметров α_k . Для этого вычислим, используя (8), выражение

$$\begin{aligned} 4 \frac{\alpha_{k+1} - 1}{\alpha_k \alpha_{k+1}} &= \frac{4}{\alpha_k} \left(1 - \frac{1}{\alpha_{k+1}} \right) = \frac{4}{\alpha_k} \left(1 - \frac{\rho_0}{2} \frac{q_k}{q_{k+1}} \right) = \\ &= \frac{4}{\alpha_k} \left[1 - \frac{\rho_0}{2} \left(\frac{2}{\rho_0} - \frac{q_k}{q_{k-1}} \right) \right] = \frac{2\rho_0}{\alpha_k} \frac{q_k}{q_{k-1}} = \rho_0^2. \end{aligned}$$

Отсюда получим $\alpha_{k+1} = 4/(4 - \rho_0^2 \alpha_k)$, $k = 1, 2, \dots$. Полагая в (10) $k = 0$ и учитывая (8), найдем, что $\alpha_1 = 2$.

Итак, доказана

Теорема 1. Пусть выполнены условия

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*.$$

Полуитерационный метод Чебышева (2) с итерационными параметрами

$$\tau_k = 2/(\gamma_1 + \gamma_2), \quad \alpha_{k+1} = 4/(4 - \rho_0^2 \alpha_k), \quad k = 1, 2, \dots, \alpha_1 = 2, \quad (11)$$

сходится в H_D , и для погрешности z_k справедлива оценка

$$\|z_k\|_D \leq q_k \|z_0\|_D.$$

Для числа итераций n имеет место оценка $n \geq n_0(\varepsilon)$, где

$$n_0(\varepsilon) = \frac{\ln 0,5\varepsilon}{\ln \rho_1}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad q_k = \frac{2\rho_1^k}{1+\rho_1^{2k}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Замечание. Сравнение полуитерационного метода Чебышева с чебышевским двухслойным методом показывает, что для этих методов имеет место одна и та же оценка $\|z_n\|_D \leq q_n \|z_0\|_D$, если выполнено n итераций. Однако для двухслойного метода эта оценка верна лишь после выполнения всех итераций, в то время как для трехслойного метода такого вида оценка имеет место для любых промежуточных итераций. В отличие от двухслойного метода в трехслойном методе нормы погрешностей на промежуточных итерациях монотонно убывают, и это обеспечивает вычислительную устойчивость трехслойного метода.

2. Примеры выбора оператора D . Приведем теперь примеры выбора оператора D и требования, налагаемые на операторы A и B , для которых условия теоремы 1 будут выполнены.

В п. 3 § 2 гл. VI были рассмотрены некоторые случаи выбора оператора D в зависимости от свойств операторов A и B . Приведем эти результаты.

1) Если операторы A и B самосопряжены и положительно определены в H , то в качестве D можно выбрать один из сле-

дующих операторов: A , B или $AB^{-1}A$. При этом априорная информация может быть задана в виде

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \quad (12)$$

2) Если операторы A и B самосопряжены положительно определены и перестановочны $A = A^* > 0$, $B = B^* > 0$, $AB = BA$, то в качестве D можно взять оператор A^*A . В этом случае априорная информация имеет вид неравенств (12).

3) Если A и B — невырожденные операторы, удовлетворяющие условию $B^*A = A^*B$, то в качестве D можно также взять оператор A^*A . В этом случае априорная информация задается в виде неравенств

$$\gamma_1 (Bx, Bx) \leq (Ax, Bx) \leq \gamma_2 (Bx, Bx), \quad \gamma_1 > 0.$$

При выполнении этих предположений для трехслойного полуитерационного метода Чебышева верна теорема 1.

3. Алгоритм метода. Рассмотрим вопрос о реализации трехслойной схемы (1). Алгоритм метода может быть описан следующим образом:

1) по значению параметра α_k и заданным приближениям y_{k-1} и y_k находим α_{k+1} и τ_{k+1} по формулам (11) и вычисляем

$$\varphi = B(\alpha_{k+1}y_k + (1 - \alpha_{k+1})y_{k-1}) - \alpha_{k+1}\tau_{k+1}(Ay_k - f).$$

Вычисленное φ может быть размещено на месте y_{k-1} , которое для дальнейших итераций уже не нужно;

2) для нахождения нового приближения y_{k+1} решаем уравнение $By_{k+1} = \varphi$. Приближение y_1 находится из уравнения $By_1 = \varphi$, где $\varphi = By_0 - \tau_1(Ay_0 - f)$. Такой алгоритм решения может быть рекомендован в случае, когда требуется экономить память ЭВМ.

Если на вычисление значения оператора B требуется большое количество арифметических действий, а ограничений на память нет, то целесообразно воспользоваться следующим алгоритмом:

1) По заданному y_k вычисляется невязка $r_k = Ay_k - f$;

2) решается уравнение для поправки w_k : $Bw_k = r_k$;

3) по заданному α_k вычисляется α_{k+1} по формуле (11) и новое приближение находится по формуле

$$y_{k+1} = \alpha_{k+1}y_k + (1 - \alpha_{k+1})y_{k-1} - \alpha_{k+1}\tau_{k+1}w_k,$$

где τ_{k+1} определяется по формуле (11).

Описанный алгоритм не содержит вычисления значения оператора B , но требует дополнительную память для хранения r_k и w_k .

§ 3. Стационарный трехслойный метод

1. Выбор итерационных параметров. Вернемся теперь к формулам для итерационных параметров α_k и τ_k полуитерационного метода Чебышева. В § 1 были получены следующие выражения

для α_{k+1} и τ_{k+1} :

$$\alpha_{k+1} = 2q_{k+1}/(\rho_0 q_k), \quad \tau_k \equiv \tau_0 = 2/(\gamma_1 + \gamma_2), \quad k = 1, 2, \dots, \quad (1)$$

где

$$q_k = \frac{2\rho_1^k}{1+\rho_1^{2k}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \quad (2)$$

Значение итерационного параметра τ_k не зависит от номера итерации k , в то время как параметр α_k изменяется, начиная с $\alpha_1=2$. Найдем предельное значение для α_k , когда k стремится к бесконечности. Из (1), (2) получим

$$\alpha_{k+1} = 2\rho_1(1+\rho_1^{2k})/(\rho_0(1+\rho_1^{2k+2})).$$

Так как $\rho_1 < 1$ и $\rho_0 = q_1 = 2\rho_1/(1+\rho_1^2)$, то $\alpha = \lim_{k \rightarrow \infty} \alpha_k = 1 + \rho_1^2$, и

при достаточно больших k имеем $\alpha_k \approx \alpha$. Поэтому естественно изучить *стационарный итерационный трехслойный метод*

$$\begin{aligned} By_{k+1} &= \alpha(B - \tau A)y_k + (1 - \alpha)By_{k-1} + \alpha\tau f, \quad k = 1, 2, \dots, \\ By_1 &= (B - \tau A)y_0 + \tau f, \quad y_0 \in H \end{aligned} \quad (3)$$

с постоянными (стационарными) параметрами

$$\alpha = 1 + \rho_1^2, \quad \tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad (4)$$

где γ_1 и γ_2 — постоянные энергетической эквивалентности самоспряженных операторов D и $DB^{-1}A$:

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \quad (5)$$

2. Оценка скорости сходимости. Для получения оценки скорости сходимости стационарного трехслойного метода перейдем от (3) к схеме для эквивалентной погрешности $x_k = D^{1/2}z_k$:

$$\begin{aligned} x_{k+1} &= \alpha(E - \tau C)x_k + (1 - \alpha)x_{k-1}, \quad k = 1, 2, \dots, \\ x_1 &= (E - \tau C)x_0, \quad C = D^{1/2}B^{-1}AD^{-1/2}. \end{aligned}$$

Отсюда следует, что x_k для любого $k \geq 0$ выражается через x_0 следующим образом:

$$x_k = P_k(C)x_0, \quad (6)$$

где соответствующий $P_k(C)$ алгебраический полином $P_k(t)$ определяется рекуррентными соотношениями

$$\begin{aligned} P_{k+1}(t) &= \alpha(1 - \tau t)P_k(t) + (1 - \alpha)P_{k-1}(t), \quad k = 1, 2, \dots, \\ P_1(t) &= 1 - \tau t, \quad P_0(t) \equiv 1. \end{aligned} \quad (7)$$

Из (6) следует оценка для нормы погрешности z_k в H_D :

$$\|z_k\|_D = \|x_k\| \leq \|P_k(C)\| \|x_0\| = \|P_k(C)\| \|z_0\|_D. \quad (8)$$

Поэтому необходимо оценить норму операторного полинома $P_k(C)$ для случая, когда параметры α и τ выбраны по формулам (4). Из условий (5) следует, что C —самосопряженный в H оператор, а γ_1 и γ_2 —его граници, следовательно,

$$\|P_k(C)\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} |P_k(t)|.$$

Оценим максимум модуля полинома $P_k(t)$ на отрезке $[\gamma_1, \gamma_2]$. Для этого выразим полином $P_k(t)$ через полиномы Чебышева. Нам будет удобнее рассматривать $P_k(t)$ не на отрезке $[\gamma_1, \gamma_2]$, а на стандартном отрезке $[-1, 1]$. Полагая

$$t = \frac{1 - \rho_0 x}{\tau_0}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

отобразим отрезок $[\gamma_1, \gamma_2]$ на $[-1, 1]$. Тогда

$$P_k(t) = Q_k(x), \quad x \in [-1, 1], \\ \max_{\gamma_1 \leq t \leq \gamma_2} |P_k(t)| = \max_{|x| \leq 1} |Q_k(x)|.$$

Учитывая выбор параметров α и τ согласно (4), из (7) получим следующие рекуррентные соотношения для полиномов $Q_k(x)$:

$$Q_{k+1}(x) = 2\rho_1 x Q_k(x) - \rho_1^2 Q_{k-1}(x), \quad k = 1, 2, \dots, \\ Q_1(x) = \rho_0 x, \quad Q_0(x) \equiv 1.$$

Отсюда при помощи замены

$$Q_k = \rho_1^k R_k(x) \tag{9}$$

легко получим стандартное рекуррентное соотношение

$$R_{k+1}(x) = 2x R_k(x) - R_{k-1}(x), \quad k = 1, 2, \dots, \\ R_1(x) = \rho_0 x / \rho_1, \quad R_0(x) \equiv 1. \tag{10}$$

Этому соотношению удовлетворяют полином Чебышева первого рода $T_k(x)$ с начальными условиями $T_k(x) = x$, $T_0(x) \equiv 1$ и полином Чебышева второго рода $U_k(x)$:

$$U_k(x) = \begin{cases} \frac{\sin((k+1) \arccos x)}{\sin(\arccos x)}, & |x| \leq 1, \\ \frac{\operatorname{sh}((k+1) \operatorname{Arch} x)}{\operatorname{sh}(\operatorname{Arch} x)}, & |x| \geq 1, \end{cases}$$

с начальными условиями $U_1(x) = 2x$, $U_0(x) \equiv 1$. Используя указанные свойства полиномов $T_k(x)$ и $U_k(x)$ и равенство $\rho_0 = q_1 = 2\rho_1/(1 + \rho_1^2)$, из (10) найдем выражение для полинома $R_k(x)$ через полиномы Чебышева

$$R_k(x) = \frac{2\rho_1^2}{1 + \rho_1^2} T_k(x) + \frac{1 - \rho_1^2}{1 + \rho_1^2} U_k(x), \quad k \geq 0.$$

Далее, используя известные оценки

$$\max_{|x| \leq 1} |T_k(x)| = T_k(1) = 1,$$

$$\max_{|x| \leq 1} |U_k(x)| = U_k(1) = k + 1,$$

получим

$$\max_{|x| \leq 1} |R_k(x)| = R_k(1) = 1 + k(1 - \rho_1^2)/(1 + \rho_1^2).$$

Отсюда, учитывая сделанные выше замены, найдем следующую оценку для нормы операторного полинома $P_k(C)$:

$$\|P_k(C)\| \leq \rho_1^k (1 + k(1 - \rho_1^2)/(1 + \rho_1^2)). \quad (11)$$

Подставляя (11) в (8), получим оценку для нормы погрешности z_k в H_D :

$$\|z_k\|_D \leq \bar{q}_k \|z_0\|_D, \quad \bar{q}_k = \rho_1^k (1 + k(1 - \rho_1^2)/(1 + \rho_1^2)),$$

причем $\bar{q}_k \rightarrow 0$ при $k \rightarrow \infty$ и $\bar{q}_{k+1} < \bar{q}_k$. Итак, доказана

Теорема 2. Стационарный трехслойный итерационный метод (3)–(5) сходится в H_D , и для погрешности z_k справедлива оценка

$$\|z_k\|_D \leq \bar{q}_k \|z_0\|_D, \quad \bar{q}_k = \rho_1^k (1 + k(1 - \rho_1^2)/(1 + \rho_1^2)).$$

Замечание. Можно показать, что $\lim_{k \rightarrow \infty} q_k/\bar{q}_k = \lim_{k \rightarrow \infty} \bar{q}_k/\rho_0^k = 0$, где q_k определено в теореме 1. Поэтому стационарный трехслойный метод сходится быстрее метода простой итерации, но медленнее чебышевского двухслойного метода и полуитерационного метода Чебышева.

§ 4. Устойчивость двухслойных и трехслойных методов по априорным данным

1. Постановка задачи. Для приближенного решения операторного уравнения $Au = f$ в главе VI были изучены двухслойные методы простой итерации и чебышевский метод, а в §§ 2, 3 гл. VII построены полуитерационный метод Чебышева и трехслойный стационарный метод.

Напомним, что для вычисления итерационных параметров в этих методах используется определенная априорная информация об операторах итерационной схемы. В случае самосопряженного оператора $DB^{-1}A$ эта информация имеет вид постоянных энергетической эквивалентности γ_1 и γ_2 операторов D и $DB^{-1}A$:

$$\gamma_1(Dx, x) \leq (DB^{-1}Ax, x) \leq \gamma_2(Dx, x), \quad \gamma_1 > 0. \quad (1)$$

В ряде случаев постоянные γ_1 и γ_2 могут быть найдены точно, т. е. существуют такие элементы $x \in H$, для которых в (1) до-

стигается равенство. В других случаях для нахождения γ_1 и γ_2 используются вспомогательные процессы и эти постоянные находятся приближенно.

Использование неточной априорной информации приводит к уменьшению скорости сходимости, а в некоторых случаях и к расходимости метода. Целью настоящего параграфа является изучение влияния неточного задания априорной информации на скорость сходимости перечисленных выше итерационных методов.

Ограничимся рассмотрением самосопряженного случая, т. е. предположим, что оператор $DB^{-1}A$ самосопряжен в H . Пусть вместо точных значений γ_1 и γ_2 в неравенствах (1) заданы некоторые приближенные значения $\tilde{\gamma}_1$ и $\tilde{\gamma}_2$. Рассмотрим двухслойные и трехслойные методы, итерационные параметры для которых будем выбирать по заданным $\tilde{\gamma}_1$ и $\tilde{\gamma}_2$. Напомним формулы для итерационных параметров.

Для двухслойной схемы

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad (2)$$

параметры метода простой итерации определяются по формуле

$$\tau_k = \tilde{\tau}_0 = 2/(\tilde{\gamma}_1 + \tilde{\gamma}_2), \quad k = 1, 2, \dots, \quad (3)$$

а параметры чебышевского метода строятся по формуле

$$\begin{aligned} \tau_k &= \tilde{\tau}_0 / (1 + \tilde{\rho}_0 \mu_k), \quad \mu_k \in \mathfrak{M}_n^*, \quad k = 1, 2, \dots, n, \\ \tilde{\rho}_0 &= (1 - \tilde{\xi}) / (1 + \tilde{\xi}), \quad \tilde{\xi} = \tilde{\gamma}_1 / \tilde{\gamma}_2. \end{aligned} \quad (4)$$

Для трехслойной итерационной схемы

$$By_{k+1} = \alpha_{k+1} (B - \tau_{k+1} A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \alpha_{k+1} \tau_{k+1} f, \quad k = 1, 2, \dots, \quad (5)$$

$$By_1 = (B - \tau_1 A) y_0 + \tau_1 f$$

параметры полуитерационного метода Чебышева определяются по формулам

$$\tau_k = \tilde{\tau}_0, \quad \alpha_{k+1} = 4 / (4 - \tilde{\rho}_0^2 \alpha_k), \quad k = 1, 2, \dots, \quad \alpha_1 = 2, \quad (6)$$

а параметры стационарного трехслойного метода задаются формулами

$$\tau_k = \tilde{\tau}_0, \quad \alpha_k = 1 + \tilde{\rho}_1^2, \quad k = 1, 2, \dots, \quad \tilde{\rho}_1 = (1 - V \tilde{\xi}) / (1 + V \tilde{\xi}). \quad (7)$$

Из общей теории итерационных методов, изложенной выше, следует, что для погрешности $z_k = y_k - u$ рассматриваемых методов справедливы оценки:

1) для метода простой итерации

$$\|z_n\|_D \leq (\max_{\gamma_1 \leq t \leq \gamma_2} |1 - \tilde{\tau}_0 t|)^n \|z_0\|_D; \quad (8)$$

2) для чебышевского двухслойного метода и полуитерационного метода Чебышева

$$\|z_n\|_D \leq \tilde{q}_n \max_{\gamma_1 \leq t \leq \gamma_2} \left| T_n \left(\frac{1-\tilde{\tau}_0 t}{\tilde{\rho}_0} \right) \right| \|z_0\|_D, \quad (9)$$

где $\tilde{q}_n = 2\tilde{\rho}_1^n / (1 + \tilde{\rho}_1^{2n})$;

3) для стационарного трехслойного метода

$$\|z_n\|_D \leq \tilde{\rho}_1^n \max_{\gamma_1 \leq t \leq \gamma_2} \left| \frac{2\tilde{\rho}_1^2}{1 + \tilde{\rho}_1^2} T_n \left(\frac{1-\tilde{\tau}_0 t}{\tilde{\rho}_0} \right) + \frac{1-\tilde{\rho}_1^2}{1 + \tilde{\rho}_1^2} U_n \left(\frac{1-\tilde{\tau}_0 t}{\tilde{\rho}_0} \right) \right| \|z_0\|_D. \quad (10)$$

Здесь $T_n(x)$ и $U_n(x)$ —полиномы Чебышева первого и второго рода, γ_1 и γ_2 —точные значения постоянных из (1).

Приведенные оценки определяют скорость сходимости рассматриваемых методов в случае, когда итерационные параметры вычисляются по неточной априорной информации.

2. Оценки скорости сходимости методов. Оценим теперь максимумы модулей полиномов, входящих в оценки (8)–(10). Для этого сделаем в (8)–(10) замену, полагая $x = (1 - \tilde{\tau}_0 t) / \tilde{\rho}_0$, и обозначим $a = (1 - \tilde{\tau}_0 \gamma_2) / \tilde{\rho}_0$, $b = (1 - \tilde{\tau}_0 \gamma_1) / \tilde{\rho}_0$. Тогда оценки (8)–(10) будут иметь вид

$$\begin{aligned} \|z_n\|_D &\leq \tilde{\rho}_0^n \left(\max_{a \leq x \leq b} |x| \right)^n \|z_0\|_D, \\ \|z_n\|_D &\leq \tilde{q}_n \max_{a \leq x \leq b} |T_n(x)| \|z_0\|_D, \\ \|z_n\|_D &\leq \tilde{\rho}_1^n \max_{a \leq x \leq b} \left| \frac{2\tilde{\rho}_1^2}{1 + \tilde{\rho}_1^2} T_n(x) + \frac{1 - \tilde{\rho}_1^2}{1 + \tilde{\rho}_1^2} U_n(x) \right| \|z_0\|_D. \end{aligned} \quad (11)$$

Рассмотрим сначала случай, когда $\tilde{\gamma}_1$ и $\tilde{\gamma}_2$ являются приближениями для γ_1 и γ_2 соответственно снизу и сверху, т. е.

$$\tilde{\gamma}_1 \leq \gamma_1 \leq \gamma_2 \leq \tilde{\gamma}_2. \quad (12)$$

В этом случае, как легко проверить, будут выполняться неравенства $-1 \leq a \leq b \leq 1$. Из (11) получим, что скорость сходимости метода простой итерации будет определяться величиной $\tilde{\rho}_0^n$, чебышевского двухслойного метода — полуитерационного метода Чебышева — величиной \tilde{q}_n , а стационарного трехслойного метода — величиной $\tilde{\rho}_1^n (1 + n(1 - \tilde{\rho}_1^2)/(1 + \tilde{\rho}_1^2))$. Итерационные методы будут сходиться, но скорость сходимости уменьшится.

Рассмотрим пример, для которого выполняются условия (12). Пусть

$$\tilde{\gamma}_1 = \gamma_1(1 - \alpha), \quad \tilde{\gamma}_2 = \gamma_2, \quad 0 \leq \alpha < 1.$$

В этом случае $a = -1$, $b < 1$. Поэтому для погрешности рассматриваемых методов из (11) получим следующие оценки:

$$\begin{aligned}\|z_n\|_D &\leq \tilde{\rho}_0^n \|z_0\|_D, \\ \|z_n\|_D &\leq \tilde{q}_n \|z_0\|_D, \\ \|z_n\|_D &\leq \tilde{\rho}_1^n (1 + n(1 - \tilde{\rho}_1^2)/(1 + \tilde{\rho}_1^2)) \|z_0\|_D.\end{aligned}$$

Из соответствующих формул для числа итераций получим, что для метода простой итерации в случае неточного задания γ_1 число итераций увеличивается примерно в $1/(1-\alpha)$ раз по сравнению с точным заданием γ_1 , в то время как для чебышевского двухслойного метода и полуитерационного метода Чебышева число итераций увеличивается лишь в $1/\sqrt{1-\alpha}$ раз.

Пусть теперь условия (12) не выполнены. В этом случае $\max(|a|, |b|) > 1$. Введем следующие обозначения:

$$\begin{aligned}\frac{1}{\rho_0^*} &= \max(|a|, |b|), \\ q_n^* &= \frac{1}{T_n\left(\frac{1}{\rho_0^*}\right)} = \frac{2\rho_1^{*n}}{1 + \rho_1^{*2n}}, \quad \rho_1^* = \frac{\rho_0^*}{1 + \sqrt{1 - \rho_0^{*2}}}.\end{aligned}$$

Используя эти обозначения, а также соотношение между полиномами Чебышева первого и второго рода

$$U_n(x) = (T_{n+1}^2(x) - 1)^{1/2} / (T_1^2(x) - 1)^{1/2}, \quad |x| \geq 1,$$

получим

$$\begin{aligned}\max_{a \leq x \leq b} |x| &= \frac{1}{\rho_0^*}, \quad \max_{a \leq x \leq b} |T_n(x)| = T_n\left(\frac{1}{\rho_0^*}\right) = \frac{1}{q_n^*} \leq \frac{1}{\rho_1^{*n}}, \\ \max_{a \leq x \leq b} |U_n(x)| &= U_n\left(\frac{1}{\rho_0^*}\right) = \frac{1 - \rho_1^{*(n+1)}}{\rho_1^{*n}(1 - \rho_1^{*2})} \leq \frac{n+1}{\rho_1^{*n}}.\end{aligned}$$

Подставляя эти оценки в (11), найдем

$$\|z_n\|_D \leq \left(\frac{\tilde{\rho}_0}{\rho_0^*}\right)^n \|z_0\|_D, \tag{13}$$

$$\|z_n\|_D \leq \frac{\tilde{q}_n}{\rho_0^*} \|z_0\|_D, \tag{14}$$

$$\|z_n\|_D \leq (\tilde{\rho}_1/\rho_1^*)^n (1 + n(1 - \tilde{\rho}_1^2)/(1 + \tilde{\rho}_1^2)) \|z_0\|_D. \tag{15}$$

Заметим, что если H — конечномерное пространство, то можно указать такое начальное приближение y_0 , для которого в оценках (13), (14) будут достигаться равенства.

Найдем теперь условие, при выполнении которого можно гарантировать сходимость рассматриваемых итерационных методов, построенных по неточной априорной информации. Так как отношение \tilde{q}_n/q_n^* стремится к нулю при $n \rightarrow \infty$ лишь при условии $\rho_1^* > \tilde{\rho}_1$, а это условие эквивалентно требованию $\rho_0^* > \tilde{\rho}_0$, то из (13)–(15) следует, что итерационные методы будут сходиться, если выполняется неравенство

$$\tilde{\rho}_0 < \rho_0^*. \tag{16}$$

Используя определения ρ_0^* , a и b , получим, что (16) будет иметь место, если $|1 - \tilde{\tau}_0 \gamma_1| < 1$, $|1 - \tilde{\tau}_0 \gamma_2| < 1$.

Решая эти неравенства, найдем

$$\tilde{\gamma}_1 + \tilde{\gamma}_2 > \gamma_2. \quad (17)$$

Итак, если выполнено условие (17), то итерационные методы, построенные по неточной априорной информации, будут сходиться. Из сказанного выше следует, что в случае конечномерного пространства H условие (17) является и необходимым для сходимости методов.

Оценим теперь реальное число итераций, достаточных для достижения заданной точности ε . Будем, как и раньше, через n обозначать число итераций для случая точного задания априорной информации, через \tilde{n} обозначим теоретическое число итераций, вычисленное по формулам соответствующих теорем по неточной априорной информации, а через n^* обозначим реальное число итераций, которое достаточно для достижения точности ε . Из формул (13)–(15) следует, что реальное число итераций n^* должно определяться из условий:

- 1) для метода простой итерации из условия $\tilde{\rho}_0^n \leq \varepsilon \rho_0^{*n}$;
- 2) для чебышевского двухслойного метода и полуитерационного метода Чебышева из условия $\tilde{q}_n \leq \varepsilon q_n^*$.

Легко убедиться в том, что имеют место неравенства $n^* \geq \tilde{n}$, $n^* \geq n$, причем число итераций \tilde{n} может быть больше или меньше n . Так как единственной количественной характеристикой итерационного метода, которая может быть заранее вычислена, является теоретическое число итераций \tilde{n} , то с точки зрения реализации методов важно оценить, во сколько раз реальное число итераций n^* будет больше \tilde{n} . Для теоретического сравнения качества итерационных методов нужно оценить отношение n^*/\tilde{n} .

Получим требуемые оценки для одного примера. Пусть $\tilde{\gamma}_1$ и $\tilde{\gamma}_2$ — приближения для γ_1 и γ_2 соответственно сверху и снизу

$$\tilde{\gamma}_1 = (1 + \alpha) \gamma_1, \quad \tilde{\gamma}_2 = (1 - \alpha) \gamma_2, \quad \alpha \geq 0. \quad (18)$$

Из условия (17) и естественного требования $\tilde{\gamma}_1 \leq \tilde{\gamma}_2$ получим, что методы будут сходиться, если выполнено условие

$$\alpha < \min(\xi/(1-\xi), (1-\xi)/(1+\xi)), \quad \xi = \gamma_1/\gamma_2.$$

Для рассматриваемого примера будут иметь место неравенства $n^* \geq n \geq \tilde{n}$. Действительно, из (18) получим

$$\xi = \frac{\tilde{\gamma}_1}{\tilde{\gamma}_2} = \frac{1 + \alpha}{1 - \alpha} \xi \geq \xi$$

и, следовательно,

$$\tilde{\rho}_0 \leq \rho_0 = (1 - \xi)/(1 + \xi), \quad \tilde{\rho}_1 \leq \rho_1 = (1 - \sqrt{\xi})/(1 + \sqrt{\xi}), \quad \tilde{q}_n \leq q_n.$$

Отсюда следует, что $n \geq \tilde{n}$. Оценим теперь величины, входящие в неравенства (13)–(14). Так как

$$\tilde{\tau} = 2/(\tilde{\gamma}_1 + \tilde{\gamma}_2) = \tau_0/(1 - \alpha \rho_0) < \tau_0, \quad \tau_0 = 2/(\gamma_1 + \gamma_2),$$

то

$$1/\rho_0^* = \max(|a|, |b|) = |a| = (\tilde{\tau}_0 \gamma_2 - 1)/\tilde{\rho}_0.$$

Опуская несложные выкладки, получим

$$\tilde{\rho}_0 = \frac{1 - \xi}{1 + \xi} = \frac{\rho_0 - \alpha}{1 - \alpha \rho_0},$$

$$\frac{\tilde{\rho}_0}{\rho_0^*} = \tilde{\tau}_0 \gamma_2 - 1 = 1 - \frac{1 - \frac{\alpha}{\xi} (1 - \xi)}{1 + \alpha} (1 - \tilde{\rho}_0) = 1 - \frac{1 - \frac{\alpha}{\xi} (1 - \xi)}{1 - \alpha \rho_0} (1 - \rho_0),$$

$$\tilde{\rho}_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}} = \frac{\rho_0 - \alpha}{1 - \alpha \rho_0 + \sqrt{(1 - \alpha^2)(1 + \rho_0^2)}},$$

$$\begin{aligned} \frac{\tilde{\rho}_1}{\rho_1^*} &= 1 - \frac{(1 + \alpha) \sqrt{\xi} + \sqrt{1 - \alpha^2} - \frac{\alpha}{\sqrt{\xi}} - \sqrt{\frac{\alpha}{\xi} [1 - (1 + \alpha) \xi]}}{(1 + \alpha) \sqrt{\xi} + \sqrt{1 - \alpha^2}} (1 - \tilde{\rho}_1) = \\ &= 1 - \frac{\left[(1 + \alpha) \sqrt{\xi} + \sqrt{1 - \alpha^2} - \frac{\alpha}{\sqrt{\xi}} - \sqrt{\frac{\alpha}{\xi} [1 - (1 + \alpha) \xi]} \right] (1 + \sqrt{\xi})}{1 - \alpha + (1 + \alpha) \xi + 2 \sqrt{(1 - \alpha^2) \xi}} (1 - \rho_1). \end{aligned}$$

Рассмотрим сначала метод простой итерации. Из теоремы 2 § 3 гл. VI и из (13) получим для числа итераций n^* , \tilde{n} и n метода простой итерации следующие оценки:

$$\begin{aligned} n &= \frac{\ln \varepsilon}{\ln \rho_0} \approx \frac{\ln(1/\varepsilon)}{1 - \rho_0}, \quad \tilde{n} = \frac{\ln \varepsilon}{\ln \tilde{\rho}_0} \approx \frac{\ln(1/\varepsilon)}{1 - \tilde{\rho}_0}, \\ n^* &= \frac{\ln \varepsilon}{\ln(\tilde{\rho}_0/\rho_0^*)} = \frac{\ln(1/\varepsilon)}{1 - \tilde{\rho}_0/\rho_0^*}. \end{aligned}$$

Подставляя сюда полученные выше выражения, найдем

$$\frac{n^*}{\tilde{n}} \approx \frac{1 + \alpha}{1 - \frac{\alpha}{\xi} (1 - \xi)}, \quad \frac{n^*}{n} \approx \frac{1 - \alpha \rho_0}{1 - \frac{\alpha}{\xi} (1 - \xi)}.$$

Если $\alpha \approx c\xi$, где $c < 1$, то отсюда получим

$$n^* \approx \tilde{n}/(1 - c), \quad n^* \approx n/(1 - c).$$

Итак, если $\alpha \approx c\xi$, то реальное число итераций n^* для метода простой итерации в $1/(1 - c)$ раз больше теоретического числа итераций \tilde{n} , которое вычисляется по неточной априорной информации.

Рассмотрим теперь чебышевский метод и полуитерационный метод Чебышева. Из определения \tilde{q}_n и q_n^* получим

$$\frac{\tilde{q}_n}{q_n^*} = \frac{\tilde{\rho}_1^n}{\rho_1^{*n}} \cdot \frac{1 + \rho_1^{*2n}}{1 + \tilde{\rho}_1^{2n}} \leq \frac{2 (\tilde{\rho}_1/\rho_1^*)^n}{1 + (\tilde{\rho}_1/\rho_1^*)^{2n}}.$$

Поэтому для числа итераций n^* найдем следующую оценку:

$$n^* = \frac{\ln(0.5\varepsilon)}{\ln(\tilde{\rho}_1/\rho_1^*)} \approx \frac{\ln(2/\varepsilon)}{1 - \tilde{\rho}_1/\rho_1^*}.$$

Далее, из теоремы 1 § 2 гл. VI и теоремы 1 § 2 гл. VII получим оценки для n и \tilde{n} :

$$n = \frac{\ln(0,5\epsilon)}{\ln \rho_1} \approx \frac{\ln(2/\epsilon)}{1-\rho_1}, \quad \tilde{n} = \frac{\ln(0,5\epsilon)}{\ln \tilde{\rho}_1} \approx \frac{\ln(2/\epsilon)}{1-\tilde{\rho}_1}.$$

Подставляя сюда полученные выше выражения для отношения $\tilde{\rho}_1/\rho_1^*$ и предполагая, что $\alpha \approx c\xi$, найдем

$$n^*/\tilde{n} \approx 1/(1 - \sqrt{c}), \quad n^*/n \approx 1/(1 - \sqrt{c}).$$

Таким образом, если $\alpha \approx c\xi$, где $c < 1$, то реальное число итераций n^* для чебышевского метода и полуитерационного метода Чебышева примерно в $1/(1 - \sqrt{c})$ раз больше теоретического числа итераций \tilde{n} , вычисленного по неточной априорной информации.

ГЛАВА VIII

ИТЕРАЦИОННЫЕ МЕТОДЫ ВАРИАЦИОННОГО ТИПА

В главе рассматриваются двухслойные и трехслойные итерационные методы вариационного типа. Для реализации этих методов не требуется никакой априорной информации об операторах итерационной схемы. В §§ 1, 2 изучаются двухслойные градиентные методы, в §§ 3, 4—трехслойные методы сопряженных направлений. Ускорению сходимости двухслойных методов в самосопряженном случае посвящен § 5.

§ 1. Двухслойные градиентные методы

1. Постановка задачи о выборе итерационных параметров. Для нахождения приближенного решения линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A , заданным в вещественном гильбертовом пространстве H , рассмотрим неявную двухслойную итерационную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad (2)$$

с произвольным начальным приближением $y_0 \in H$ и невырожденным оператором B .

Итерационная схема (2) изучалась нами ранее в главе VI, где были построены наборы итерационных параметров $\{\tau_k\}$ и даны оценки скорости сходимости соответствующих итерационных методов (чебышевского метода и метода простой итерации).

Любой двухслойный итерационный метод, построенный на основе схемы (2), характеризуется операторами A и B , энергетическим пространством H_D , в котором доказывается сходимость метода, и набором итерационных параметров τ_k . Основным вопросом теории итерационных методов является вопрос об оптимальном выборе параметров τ_k .

В главе VI были построены итерационные методы, параметры τ_k в которых выбирались из условия минимума в H_D либо нормы оператора перехода от итерации к итерации, либо нормы разрешающего оператора. Отличительной чертой построенных на таком принципе итерационных методов является

использование для вычисления параметров τ_k определенной априорной информации об операторах итерационной схемы.

Вид требуемой априорной информации определяется свойствами операторов A , B и D . Так, в случае, когда оператор $DB^{-1}A$ самосопряжен в пространстве H , эта информация означает задание постоянных энергетической эквивалентности операторов D и $DB^{-1}A$, т.е. постоянных γ_1 и γ_2 из неравенств

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad (3)$$

или границ оператора $DB^{-1}A$ в H_D .

В несамосопряженном случае используются либо два числа γ_1 и γ_2 из неравенств

$$\gamma_1 D \leq DB^{-1}A, \quad (DB^{-1}Ax, B^{-1}Ax) \leq \gamma_2 (DB^{-1}Ax, x), \quad \gamma_1 > 0, \quad (4)$$

либо три числа — γ_1 , γ_2 и γ_3 , где γ_1 и γ_2 — постоянные из неравенств (3), а γ_3 — постоянная либо из неравенства

$$\|0,5(DB^{-1}A - A^*(B^*)^{-1}D)x\|_{D^{-1}}^2 \leq \gamma_3^2 (Dx, x), \quad (5)$$

либо из неравенства

$$\|0,5(DB^{-1}A - A^*(B^*)^{-1}D)x\|_{D^{-1}}^2 \leq \gamma_3 (DB^{-1}Ax, x). \quad (6)$$

В ряде случаев нахождение постоянных γ_1 , γ_2 и γ_3 с достаточной точностью может оказаться сложной самостоятельной задачей, требующей для своего решения использования специальных вычислительных методов. Если априорная информация может быть получена ценой небольших вычислительных затрат или если требуется решить серию задач (1) с различными правыми частями, то целесообразно найти однажды необходимую априорную информацию и затем воспользоваться итерационными методами, построенными в главе VI. Такой путь можно рекомендовать, если дополнительное время, затрачиваемое на получение априорной информации, существенно меньше времени решения всей серии задач (1).

В тех случаях, когда требуется решить лишь одну задачу (1), или когда задано хорошее начальное приближение, а вычисление постоянных γ_1 , γ_2 и γ_3 является трудоемким процессом, следует воспользоваться итерационными методами вариационного типа, к рассмотрению которых мы переходим.

В двухслойных итерационных методах вариационного типа для вычисления параметров τ_k не требуется никакой априорной информации об операторах схемы (2) (кроме условий общего вида $A = A^* > 0$, $(DB^{-1}A)^* = DB^{-1}A$ и т. д.), и построение этих методов основано на следующем принципе. Если задано приближение y_k , а y_{k+1} находится по схеме (2), то итерационный параметр τ_{k+1} выбирается из условия минимума в H_D нормы погрешности $z_{k+1} = y_{k+1} - u$, где u — решение уравнения (1).

Название методов связано с тем, что последовательность y_k , построенная по формуле (2), в которой параметры τ_k выбираются из указанного выше условия, является минимизирующей последовательностью для квадратичного функционала

$$I(y) = (D(y-u), y-u).$$

Этот функционал в силу положительной определенности оператора D ограничен снизу и достигает минимума, равного нулю, на решении уравнения (1), т. е. при $y=u$. Выбор параметра τ_{k+1} из указанного условия обеспечивает локальную минимизацию функционала $I(y)$ при переходе от y_k к y_{k+1} , т. е. за один итерационный шаг. В случае явной схемы ($B=E$) переход от y_k к y_{k+1} осуществляется по формуле

$$y_{k+1} = y_k - \tau_{k+1} r_k, \quad r_k = Ay_k - f.$$

Отметим, что для самосопряженного положительно определенного оператора A переход от y_k к y_{k+1} происходит по направлению $-r_k$, которое совпадает с направлением антиградиента для функционала $(A(y-u), y-u)$ в точке y_k . Известно, что по направлению антиградиента происходит наибольшее убывание значения функционала. Поэтому такие методы называют иногда методами градиентного спуска или просто градиентными методами. Мы сохраним это название и для неявных двухслойных методов вариационного типа.

Нашей ближайшей задачей является нахождение параметра τ_{k+1} из условия минимума в H_D нормы погрешности $z_{k+1} = y_{k+1} - u$.

2. Формула для итерационных параметров. Найдем теперь формулу для вычисления итерационного параметра τ_{k+1} , предполагая, что оператор A не вырожден. Выпишем сначала уравнение для погрешности $z_k = y_k - u$, $k=0, 1, \dots$. Подставляя $y_k = z_k + u$ в схему (2), получим

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A) z_k, \quad k=0, 1, \dots, z_0 = y_0 - u.$$

Замена $z_k = D^{-1/2} x_k$ позволяет перейти к уравнению, содержащему только один оператор

$$\begin{aligned} x_{k+1} &= S_{k+1} x_k, \quad S_k = E - \tau_k C, \\ C &= D^{-1/2} (DB^{-1} A) D^{-1/2}. \end{aligned} \tag{7}$$

Используя равенство $\|z_k\|_D = \|x_k\|$, поставленную выше задачу о выборе параметра τ_{k+1} можно сформулировать следующим образом: выбрать параметр τ_{k+1} из условия минимума нормы x_{k+1} в пространстве H .

Решим эту задачу. Вычислим норму x_{k+1} :

$$\begin{aligned} \|x_{k+1}\|^2 &= ((E - \tau_{k+1} C) x_k, (E - \tau_{k+1} C) x_k) = \\ &= \|x_k\|^2 - 2\tau_{k+1} (Cx_k, x_k) + \tau_{k+1}^2 (Cx_k, Cx_k) = \\ &= (Cx_k, Cx_k) \left[\tau_{k+1} - \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)} \right]^2 + \|x_k\|^2 - \frac{(Cx_k, x_k)^2}{(Cx_k, Cx_k)}. \end{aligned} \tag{8}$$

Так как оператор A не вырожден, то не вырожден и оператор C . Поэтому для любого x_k имеем $(Cx_k, Cx_k) > 0$, и минимум нормы x_{k+1} достигается при

$$\tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \quad (9)$$

Подставляя (9) в (8), получим

$$\|x_{k+1}\| = \rho_{k+1} \|x_k\|, \quad (10)$$

где

$$\rho_{k+1}^2 = 1 - \frac{(Cx_k, x_k)^2}{(Cx_k, Cx_k)(x_k, x_k)}. \quad (11)$$

Итак, формула (9) определяет оптимальное значение итерационного параметра τ_{k+1} . Подставляя в (9) $x_k = D^{1/2}z_k$, получим

$$\tau_{k+1} = \frac{(DB^{-1}Az_k, z_k)}{(DB^{-1}Az_k, B^{-1}Az_k)}, \quad k = 0, 1, \dots$$

Учитывая, что $Az_k = Ay_k - Au = Ay_k - f = r_k$ — невязка, а $B^{-1}r_k = w_k$ — поправка, формулу для параметра τ_{k+1} можно записать в следующем виде:

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots, \quad (12)$$

а итерационную схему (2) — в виде явной формулы для вычисления y_{k+1} :

$$y_{k+1} = y_k - \tau_{k+1}w_k, \quad k = 0, 1, \dots \quad (13)$$

Алгоритм, реализующий построенный метод, можно описать следующим образом:

- 1) по заданному y_k вычисляется невязка $r_k = Ay_k - f$,
- 2) решается уравнение для поправки $Bw_k = r_k$,
- 3) по формуле (12) вычисляется параметр τ_{k+1} ,
- 4) по формуле (13) находится новое приближение y_{k+1} .

Формулы (12) еще не пригодны для вычислений, так как наряду с известными в процессе итераций величинами r_k и w_k содержат неизвестную погрешность z_k . В § 2, выбирая конкретный оператор D , мы получим формулы для параметров τ_k , которые будут содержать только известные величины. А сейчас мы переходим к получению оценки скорости сходимости построенного итерационного метода.

3. Оценка скорости сходимости. Оценим теперь скорость сходимости двухслойных градиентных методов. Так как итерационный параметр τ_{k+1} выбирается из условия минимума в H_D нормы погрешности z_{k+1} , которое эквивалентно условию минимума в H нормы x_{k+1} , то из (7) получим

$$\begin{aligned} \|x_{k+1}\| &= \min_{\tau_{k+1}} \|S_{k+1}x_k\| \leq \min_{\tau_{k+1}} \|S_{k+1}\| \|x_k\| = \\ &= \min_{\tau} \|E - \tau C\| \|x_k\| = \rho \|x_k\|, \quad \rho = \min_{\tau} \|E - \tau C\|. \end{aligned}$$

Сравнивая эту оценку с равенством (10), находим

$$\rho_k \leq \rho \leq 1, \quad k = 1, 2, \dots \quad (14)$$

Из (10), (14) следует оценка $\|x_{k+1}\| \leq \rho \|x_k\|$, а в силу сделанной замены $x_k = D^{1/2} z_k$, отсюда вытекает оценка для нормы погрешности z_n в энергетическом пространстве H_D :

$$\|z_n\|_D \leq \rho^n \|z_0\|_D, \quad \rho = \min_{\tau} \|E - \tau C\|. \quad (15)$$

Если выполнено условие $\rho < 1$, то двухслойный градиентный метод сходится в H_D . Из оценки (15) следует, что для уменьшения нормы в H_D начальной погрешности в $1/\varepsilon$ раз достаточно выполнить $n \geq n_0(\varepsilon)$ итераций, где

$$n_0(\varepsilon) = \ln \varepsilon / \ln \rho. \quad (16)$$

Итак, скорость сходимости двухслойного градиентного метода определяется величиной ρ . Напомним, что в главе VI при изучении метода простой итерации при различных предположениях относительно оператора C были получены оценки для ρ . Величина ρ определяет скорость сходимости метода простой итерации. Поэтому из полученной здесь оценки (15) следует, что любой двухслойный градиентный метод сходится не медленнее соответствующего метода простой итерации.

Приведем оценки для ρ , полученные в §§ 3, 4 главы VI при различных предположениях относительно операторов A , B и D .

1. Если оператор $DB^{-1}A$ самосопряжен в H , а γ_1 и γ_2 — постоянные из неравенств (3), то

$$\rho = (1 - \xi) / (1 + \xi), \quad \xi = \gamma_1 / \gamma_2. \quad (17)$$

2. Пусть оператор $DB^{-1}A$ несамосопряжен в H ;

а) если выполнены условия (4), то

$$\rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1 / \gamma_2; \quad (18)$$

б) если выполнены условия (3), (5), то

$$\rho = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{1 - \kappa \gamma_1}{1 + \kappa \gamma_2}, \quad \kappa = \sqrt{\frac{\gamma_3}{\gamma_1 \gamma_2 + \gamma_3}}. \quad (19)$$

Итак, доказана

Теорема 1. Если для схемы (2) метод простой итерации сходится, то сходится двухслойный градиентный метод (2), (12). При этом для погрешности z_n справедлива оценка

$$\|z_n\|_D \leq \rho^n \|z_0\|_D,$$

где ρ определено в (17), если оператор $DB^{-1}A$ самосопряжен в H и выполнены условия (3), ρ определено в (18), если для несамосопряженного оператора $DB^{-1}A$ выполнены условия (4), и в (19), если выполнены условия (3), (15). Оценка для числа итераций дана в (16).

Замечание. Если уравнение (1) рассматривается в комплексном гильбертовом пространстве, то итерационный параметр τ_{k+1} должен быть выбран по формуле

$$\tau_{k+1} = \frac{\operatorname{Re}(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots$$

Теорема 1 сохраняет силу, только условия (3), (4) должны быть заменены на неравенства

$$\begin{aligned}\gamma_1(Dx, x) &\leq \operatorname{Re}(DB^{-1}Ax, x) \leq \gamma_2(Dx, x), \\ \gamma_1(Dx, x) &\leq \operatorname{Re}(DB^{-1}Ax, x), \\ (DB^{-1}Ax, B^{-1}Ax) &\leq \gamma_2 \operatorname{Re}(DB^{-1}Ax, x),\end{aligned}$$

где $\operatorname{Re} z$ — действительная часть комплексного числа z .

4. Неулучшаемость оценки в самосопряженном случае. Покажем, что на классе произвольных начальных приближений y_0 в случае самосопряженного оператора $DB^{-1}A$ в конечномерном пространстве H априорная оценка погрешности итерационного метода (2), (12), полученная в теореме 1, является неулучшаемой. Для этого достаточно указать такое начальное приближение x_0 , при котором для решения уравнения (7) имеет место равенство $\|x_{k+1}\| = \rho \|x_k\|$, где ρ определено в (17).

Найдем искомое начальное приближение x_0 . Пусть H — конечномерное пространство ($H = H_N$). Так как оператор $DB^{-1}A$ самосопряжен в H , то оператор $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ также самосопряжен в H . Следовательно, существует полная система собственных функций v_1, v_2, \dots, v_N оператора C . Обозначим через λ_k собственное значение оператора C , соответствующее собственной функции v_k , так что $Cv_k = \lambda_k v_k$, $k = 1, 2, \dots, N$. Пусть $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$. Так как неравенства (3) эквивалентны неравенствам

$$\gamma_1 E \leq C \leq \gamma_2 E, \quad \gamma_1 > 0,$$

то в (3) в качестве γ_1 и γ_2 можно взять λ_1 и λ_N . При этом ρ , определенное в (17), можно записать в следующем виде:

$\rho = (\lambda_N - \lambda_1)/(\lambda_N + \lambda_1)$. Выберем начальное приближение

$$x_0 = \sqrt{\lambda_N} v_1 + \sqrt{\lambda_1} v_N. \quad (20)$$

Тогда $Cx_0 = \lambda_1 \sqrt{\lambda_N} v_1 + \lambda_N \sqrt{\lambda_1} v_N$. Используя ортонормированность системы собственных функций v_1, v_2, \dots, v_N , получим

$$\begin{aligned}(x_0, x_0) &= \lambda_1 + \lambda_N, \\ (Cx_0, x_0) &= 2\lambda_1 \lambda_N, \\ (Cx_0, Cx_0) &= \lambda_1 \lambda_N (\lambda_1 + \lambda_N).\end{aligned}$$

Подставляя эти значения в (9), (11), получим $\tau_1 = 2/(\lambda_1 + \lambda_N)$, $\rho_1 = (\lambda_N - \lambda_1)/(\lambda_N + \lambda_1) = \rho$. Из (10) следует равенство $\|x_1\| = \rho \|x_0\|$, а из (7) найдем x_1 :

$$x_1 = \rho (\sqrt{\lambda_N} v_1 - \sqrt{\lambda_1} v_N).$$

Дальнейшие вычисления дают

$$\begin{aligned} Cx_1 &= \rho (\lambda_1 \sqrt{\lambda_N} v_1 - \lambda_N \sqrt{\lambda_1} v_N), \\ (x_1, x_1) &= \rho^2 (x_0, x_0), \\ (Cx_1, x_1) &= \rho^2 (Cx_0, x_0), \\ (Cx_1, Cx_1) &= \rho^2 (Cx_0, Cx_0). \end{aligned}$$

Поэтому

$$\begin{aligned} \tau_2 &= \frac{(Cx_1, x_1)}{(Cx_1, Cx_1)} = \frac{(Cx_0, x_0)}{(Cx_0, Cx_0)} = \tau_1, \\ \rho_2^2 &= 1 - \frac{(Cx_1, x_1)^2}{(Cx_1, Cx_1) (x_1, x_1)} = 1 - \frac{(Cx_0, x_0)^2}{(Cx_0, Cx_0) (x_0, x_0)} = \rho_1^2 = \rho^2. \end{aligned}$$

Следовательно, $\|x_2\| = \rho \|x_1\|$. Кроме того, $x_2 = x_1 - \tau_2 Cx_1 = \rho^2 x_0$, т. е. x_2 пропорционально x_0 . Отсюда сразу следует, что $\tau_3 = \tau_2 = \tau_1$, $\rho_3 = \rho$ и $x_3 = \rho^2 x_1$. Поэтому для любого k :

$$\begin{aligned} \tau_k &\equiv 2/(\lambda_1 + \lambda_N), \quad \rho_k \equiv \rho = (\lambda_N - \lambda_1)/(\lambda_N + \lambda_1), \\ \|x_{k+1}\| &= \rho \|x_k\|. \end{aligned}$$

Утверждение доказано.

Итак, мы показали, что если начальное приближение выбрано по формуле (20), то в двухслойном градиентном методе все параметры τ_k одинаковы и совпадают с параметром метода простой итерации (см. § 3 гл. VI), погрешности через одну итерацию пропорциональны, а скорость сходимости самая медленная.

Отметим, что такая медленная сходимость метода имеет место лишь для специального «плохого» начального приближения. В случае же «хорошего» начального приближения скорость сходимости метода может быть значительно большей. Более детальное изучение характера изменения скорости сходимости метода будет проведено в следующем пункте, а здесь мы рассмотрим один пример, иллюстрирующий приведенное выше замечание.

Покажем, что если в качестве начального приближения x_0 взять любую собственную функцию v_m , то двухслойный градиентный метод сойдется за одну итерацию.

Действительно, пусть $x_0 = v_m$. Тогда несложные вычисления дают

$$\begin{aligned} Cx_0 &= \lambda_m v_m = \lambda_m x_0, \quad (Cx_0, x_0) = \lambda_m (x_0, x_0), \\ (Cx_0, Cx_0) &= \lambda_m^2 (x_0, x_0), \quad \tau_1 = 1/\lambda_m, \quad \rho_1 = 0, \end{aligned}$$

т. е. $x_1 = 0$ или $y_1 = u$.

Это качественно новое свойство двухслойных градиентных методов — возможность увеличивать скорость сходимости в случае, когда задано «хорошее» начальное приближение, — отличает эти методы от рассмотренных в главе VI двухслойных итерационных методов, жестко ориентированных на самое плохое начальное приближение.

5. Асимптотическое свойство градиентных методов в самосопряженном случае. Рассмотрим теперь асимптотическое свойство двухслойных градиентных методов, которым они обладают в случае самосопряженного оператора $DB^{-1}A$. Это свойство заключается в том, что последовательность $\{\rho_k\}$, определенная в (11), является возрастающей. Так как величина ρ_k определяет скорость убывания нормы погрешности при переходе от k к $(k+1)$ -й итерации, то наличие указанного свойства приводит к уменьшению скорости убывания нормы погрешностей z_n для больших n по сравнению с началом процесса итераций. Причем для достаточно больших n скорость сходимости градиентных методов становится практически такой же, как и для метода простой итерации.

Будет показано, что для больших номеров итераций погрешности, рассматриваемые через одну итерацию, становятся почти пропорциональными. Используя этот факт, мы построим приближенный метод нахождения постоянных γ_1 и γ_2 для неравенств (3), а в § 5 построим процесс ускорения сходимости двухслойных градиентных методов.

Итак, пусть оператор $DB^{-1}A$, а вместе с ним и оператор C самосопряжены в H . Покажем, что последовательность $\{\rho_k\}$ является возрастающей. Из (10) следуют равенства

$$\|x_{k+2}\| = \rho_{k+2} \|x_{k+1}\|, \quad \|x_{k+1}\| = \rho_{k+1} \|x_k\|.$$

Вычислим норму разности $x_{k+2} - \rho_{k+2} \rho_{k+1} x_k$:

$$\begin{aligned} \|x_{k+2} - \rho_{k+2} \rho_{k+1} x_k\|^2 &= \|x_{k+2}\|^2 - 2\rho_{k+2} \rho_{k+1} (x_{k+2}, x_k) + \\ &\quad + \rho_{k+2}^2 \rho_{k+1}^2 \|x_k\|^2 = 2 (\|x_{k+2}\|^2 - \rho_{k+2} \rho_{k+1} (x_{k+2}, x_k)). \end{aligned} \quad (21)$$

Вычислим отдельно скалярное произведение (x_{k+2}, x_k) . Из уравнения (7) найдем

$$x_{k+2} = x_{k+1} - \tau_{k+2} C x_{k+1}, \quad x_k = x_{k+1} + \tau_{k+1} C x_k. \quad (22)$$

Умножая последнее равенство скалярно на $C x_k$ и учитывая (9), получим

$$(C x_k, x_k) = (x_{k+1}, C x_k) + \tau_{k+1} (C x_k, C x_k) = (x_{k+1}, C x_k) + (C x_k, x_k).$$

Следовательно, для любого k имеет место равенство

$$(x_{k+1}, C x_k) = 0, \quad (23)$$

а в силу самосопряженности оператора C —равенство $(C x_{k+1}, x_k) = 0$.

Из (22) и (23) получим

$$\begin{aligned} (x_{k+2}, x_k) &= (x_{k+1} - \tau_{k+2} C x_{k+1}, x_k) = (x_{k+1}, x_k) = \\ &= (x_{k+1}, x_{k+1} + \tau_{k+1} C x_k) = \|x_{k+1}\|^2. \end{aligned}$$

Подставляя полученное равенство в (21), найдем

$$\|x_{k+2} - \rho_{k+2} \rho_{k+1} x_k\|^2 = 2 \left(1 - \frac{\rho_{k+1}}{\rho_{k+2}}\right) \|x_{k+2}\|^2. \quad (24)$$

Из (24) следует, что либо $\rho_{k+2} > \rho_{k+1}$, либо $\rho_{k+1} = \rho_{k+2} = \bar{\rho}$ и $x_{k+2} = \bar{\rho}^2 x_k$. В последнем случае, очевидно, что для всех $n \geq k$ будут выполняться равенства

$$\rho_{n+1} = \bar{\rho}, \quad x_{n+2} = \bar{\rho}^2 x_n, \quad (25)$$

т. е. последовательность ρ_k выходит на предельное значение.

Итак, показано, что последовательность $\{\rho_k\}$ действительно является возрастающей. В п. 3 этого параграфа было показано, что эта последовательность ограничена сверху и, следовательно, имеет предел. Поэтому для достаточно больших номеров k будет иметь место приближенное равенство $\rho_{k+1} \approx \rho_{k+2}$ и, следовательно, $x_{k+2} \approx \rho_{k+2} \rho_{k+1} x_k$, т. е. погрешности будут почти пропорциональны через одну итерацию.

Рассмотрим, что следует из выхода последовательности ρ_k на предельное значение. В этом случае выполняются равенства (25), т. е. $x_{n+2} = \bar{\rho}^2 x_n$. Пусть пространство H конечномерно, а v_1, v_2, \dots, v_N — система собственных функций оператора C . Разложим x_n по собственным функциям

$$x_n = \sum_{k=1}^N \alpha_k^{(n)} v_k. \quad (26)$$

Из уравнения (7) получим

$$\begin{aligned} x_{n+2} &= (E - \tau_{n+2} C)(E - \tau_{n+1} C)x_n = \\ &= \sum_{k=1}^N (1 - \tau_{n+2} \lambda_k)(1 - \tau_{n+1} \lambda_k) \alpha_k^{(n)} v_k. \end{aligned}$$

Так как $x_{n+2} = \bar{\rho}^2 x_n$, то это означает, что для всех номеров k , для которых $\alpha_k^{(n)} \neq 0$, должно выполняться равенство

$$(1 - \tau_{n+2} \lambda_k)(1 - \tau_{n+1} \lambda_k) = \bar{\rho}^2.$$

Отсюда следует, что в разложении (26) присутствуют собственные функции, соответствующие только двум различным (каждое из которых может быть кратным) собственным значениям. Пусть это будут λ_i и λ_j . Тогда λ_i и λ_j есть корни уравнения

$$(1 - \tau_{n+2} \lambda_i)(1 - \tau_{n+1} \lambda_i) = \bar{\rho}^2. \quad (27)$$

Зная τ_{n+1} , τ_{n+2} и $\bar{\rho}$, из этого уравнения можно найти собственные значения λ_i и λ_j .

Не останавливаясь на деталях, отметим, что если в разложении начальной погрешности x_0 присутствуют собственные функции, соответствующие минимальному собственному значению λ_1 оператора C и максимальному — λ_N , то в случае выхода на предельное значение последовательности ρ_k в разложении (26) останутся только эти собственные функции. Поэтому, решая уравнение (27), мы найдем λ_1 и λ_N .

Выход последовательности $\{\rho_k\}$ на предельное значение при конечном n является исключительным случаем. В общем же

случае можно лишь утверждать, что при достаточно большом n $\rho_{n+1} \approx \rho_{n+2}$ и $x_{n+2} \approx \rho_{n+2}\rho_{n+1}x_n$.

Наличие приближенного равенства позволяет рассчитывать на то, что для достаточно большого n корни уравнения

$$(1 - \tau_{n+2}\lambda)(1 - \tau_{n+1}\lambda) = \rho_{n+2}\rho_{n+1} \quad (28)$$

будут являться хорошими приближениями для λ_1 и λ_N , а следовательно, и для γ_1 и γ_2 из неравенств (3).

Опишем этот метод нахождения приближенных значений для γ_1 и γ_2 . По итерационной схеме (2) с $f=0$ проводится $n+2$ итерации с параметрами τ_{k+1} , определенными в (12). Так как при $f=0$ решение уравнения (1) есть нуль ($u=0$), то $z_k=y_k$ и, следовательно, ρ_{k+1} можно найти по формуле

$$\rho_{k+1} = \frac{\|z_{k+1}\|_D}{\|z_k\|_D} = \frac{\|y_{k+1}\|_D}{\|y_k\|_D}.$$

Вычислив τ_{n+1} , τ_{n+2} , ρ_{n+1} и ρ_{n+2} , решают уравнение (28). Корни этого уравнения являются приближениями для γ_1 сверху и для γ_2 снизу.

В § 5 будет приведен пример, иллюстрирующий предложенный метод нахождения γ_1 и γ_2 .

§ 2. Примеры двухслойных градиентных методов

1. Метод скорейшего спуска. В § 1 были изучены общие свойства двухслойных итерационных методов вариационного типа, используемых для нахождения приближенного решения линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A . Итерационные приближения вычисляются по двухслойной схеме

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, x_0 \in H, \quad (2)$$

а итерационные параметры τ_k находятся по формуле

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots, \quad (3)$$

где $w_k = B^{-1}r_k$ — поправка, $r_k = Ay_k - f$ — невязка, а $z_k = y_k - u$ — погрешность. Выбор параметра τ_{k+1} по формуле (3) обеспечивает минимум нормы погрешности z_{k+1} в H_D при переходе от y_k к y_{k+1} .

Рассмотрим теперь частные случаи двухслойных градиентных методов. Каждый конкретный метод определяется выбором оператора D и имеет свою область применимости. Оператор D будет выбираться так, чтобы в формулу (3) для итерационного параметра τ_{k+1} входили только известные в процессе итераций величины.

Рассмотрение примеров начнем с метода скорейшего спуска. Этот метод можно применять лишь в случае самосопряженного и положительно определенного оператора A .

Пусть оператор A самосопряжен и положительно определен в H . Метод скорейшего спуска характеризуется следующим выбором оператора D : $D = A$. Оператор B должен быть положительно определен в H . Учитывая соотношения $Az_k = Ay_k - f = r_k$ и $A = A^*$, из (3) получим формулу для итерационного параметра τ_{k+1} в неявном методе скорейшего спуска

$$\tau_{k+1} = \frac{(r_k, w_k)}{(Aw_k, w_k)}, \quad k = 0, 1, \dots$$

Для случая явной двухслойной схемы (2) ($B = E$) получим $w_k = -B^{-1}r_k = r_k$, и формула для τ_{k+1} принимает вид

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \quad k = 0, 1, \dots$$

В методе скорейшего спуска минимизируется норма погрешности z_{k+1} в энергетическом пространстве H_A : $\|z_k\|_A = (Az_k, z_k)^{1/2}$. Условия сходимости метода сформулированы в теореме 1, из которой следуют оценки

$$\|z_n\|_A \leq \rho^n \|z_0\|_A, \quad n \geq n_0(\varepsilon) = \ln \varepsilon / \ln \rho.$$

Значение величины ρ определяется свойствами операторов A и B и объемом априорной информации относительно их. Заметим, что требование самосопряженности оператора $DB^{-1}A = AB^{-1}A$ для данного метода эквивалентно требованию самосопряженности оператора B . Поэтому

1) если $B = B^*$ и выполнены условия (3) § 1 или эквивалентные им условия (см. гл. VI, § 2, п. 3)

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0,$$

то

$$\rho = (1 - \xi) / (1 + \xi), \quad \xi = \gamma_1 / \gamma_2;$$

2) если $B \neq B^*$ и выполнены условия (4) § 1 или эквивалентные им условия (см. гл. VI, § 4, п. 2)

$$\gamma_1 (Bx, A^{-1}Bx) \leq (Bx, x), \quad (Ax, x) \leq \gamma_2 (Bx, x), \quad \gamma_1 > 0,$$

то

$$\rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1 / \gamma_2.$$

Отметим, что если $B = B^*$, то метод скорейшего спуска обладает асимптотическим свойством.

2. Метод минимальных невязок. Этот метод можно применять в случае любого несамосопряженного невырожденного оператора A . Положительная определенность каждого в отдельности

операторов A и B не предполагается, требуется лишь положительная определенность оператора B^*A . Метод минимальных невязок определяется следующим выбором оператора D : $D = A^*A$.

Формула (3) для итерационного параметра τ_{k+1} в методе минимальных невязок имеет вид

$$\tau_{k+1} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}, \quad k = 0, 1, \dots$$

В случае явной схемы (2) ($B = E$) требуется положительная определенность оператора A , а формула для τ_{k+1} имеет вид

$$\tau_{k+1} = \frac{(Ar_k, r_k)}{(Ar_k, Ar_k)}, \quad k = 0, 1, \dots$$

Название метода связано с тем, что в нем минимизируется норма невязки. Действительно, для указанного оператора D имеем

$$\|z_k\|_D^2 = (Dz_k, z_k) = (A^*Az_k, z_k) = \|Az_k\|^2 = \|r_k\|^2.$$

Следовательно, для рассматриваемого метода норма погрешности в H_D равна норме невязки, которую можно вычислить в процессе итераций и использовать для контроля окончания итераций.

Из теоремы 1 следуют оценки сходимости метода

$$\|r_n\| \leq \rho^n \|r_0\|, \quad n \geq n_0(\varepsilon) = \ln \varepsilon / \ln \rho.$$

Оператор $DB^{-1}A = A^*AB^{-1}A$ будет самосопряжен в H , если самосопряжен оператор AB^{-1} , что эквивалентно требованию самосопряженности оператора B^*A . Если это требование выполнено, то из условий (3) § 1, которые в данном случае имеют вид

$$\gamma_1(Ay, Ay) \leq (AB^{-1}Ay, Ay) \leq \gamma_2(Ay, Ay), \quad \gamma_1 > 0,$$

или после замены $y = A^{-1}Bx$

$$\gamma_1(Bx, Bx) \leq (Ax, Bx) \leq \gamma_2(Bx, Bx), \quad \gamma_1 > 0, \quad (4)$$

и теоремы 1 следует, что

$$\rho = (1 - \xi) / (1 + \xi), \quad \xi = \gamma_1 / \gamma_2.$$

Отметим, что условие $\gamma_1 > 0$ будет выполнено, если выполнено указанное выше требование положительной определенности оператора B^*A . Условия самосопряженности и положительной определенности оператора B^*A будут выполнены, например, при следующих предположениях: $A = A^* > 0$, $B = B^* > 0$, $AB = BA$.

В этом случае неравенства (4) эквивалентны более простым. Действительно, полагая в (4) $x = B^{-1/2}y$ и используя перестановочность операторов A и B , получим

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \quad (5)$$

Условия самосопряженности и положительной определенности оператора B^*A будут автоматически выполняться и в том случае, когда оператор B имеет вид $B = (A^*)^{-1}B_0$, где B_0 — самосопряженный

женный и положительно определенный оператор. В этом случае вместо неравенств (5) нужно использовать неравенства

$$\gamma_1 B_0 \leq A^* A \leq \gamma_2 B_0, \quad \gamma_1 > 0, \quad (6)$$

а в формуле для параметра τ_{k+1} поправку w_k можно находить из уравнения $B_0 w_k = A^* r_k$.

Если оператор $B^* A$ несамосопряжен в H , то из условий (4) § 1 или эквивалентных им условий

$$\gamma_1 (Bx, Bx) \leq (Ax, Bx), \quad (Ax, Ax) \leq \gamma_2 (Ax, Bx), \quad \gamma_1 > 0$$

и теоремы 1 следует, что $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1 / \gamma_2$.

3. Метод минимальных поправок. Этот метод можно применять для решения уравнения (1) с несамосопряженным, но положительно определенным оператором A . Требуется, чтобы оператор B был самосопряженным положительно определенным и ограниченным оператором. *Метод минимальных поправок* определяется следующим выбором оператора D : $D = A^* B^{-1} A$.

Формула (3) для итерационного параметра τ_{k+1} в методе минимальных поправок имеет вид

$$\tau_{k+1} = \frac{(Aw_k, w_k)}{(B^{-1} Aw_k, Aw_k)}, \quad k = 0, 1, \dots$$

В случае явной схемы (2) ($B = E$) методы минимальных поправок и минимальных невязок совпадают.

В методе минимальных поправок минимизируется норма поправки в H_B . Действительно, для выбранного оператора D получаем

$$\|z_k\|_D^2 = (Dz_k, z_k) = (A^* B^{-1} A z_k, z_k) = (w_k, r_k) = (Bw_k, w_k) = \|w_k\|_B^2.$$

Норма поправки в H_B может быть вычислена в процессе итераций и использована для контроля окончания итераций.

Из теоремы 1 следуют оценки сходимости метода

$$\|w_n\|_B \leq \rho^n \|w_0\|_B, \quad n \geq n_0(\varepsilon) = \ln \varepsilon / \ln \rho.$$

Оператор $DB^{-1}A = A^* B^{-1} AB^{-1} A$ самосопряжен в H одновременно с оператором A . Поэтому:

1) если $A = A^*$ и выполнены условия (3) § 1 или эквивалентные им условия (см. гл. VI, § 2, п. 3)

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0,$$

то

$$\rho = (1 - \xi) / (1 + \xi), \quad \xi = \gamma_1 / \gamma_2;$$

2) если $A \neq A^*$ и выполнены условия (4) § 1 или эквивалентные им условия (см. гл. VI, § 4, п. 2)

$$\gamma_1 B \leq A, \quad (Ax, B^{-1} Ax) \leq \gamma_2 (Ax, x), \quad \gamma_1 > 0,$$

то

$$\rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1 / \gamma_2.$$

Отметим, что по сравнению с методами скорейшего спуска и минимальных невязок в методе минимальных поправок требуется не один раз обращать оператор B , а дважды, сначала для вычисления поправки w_k , а затем для вычисления $B^{-1}Aw_k$.

Отметим также, что если $A = A^*$, то метод минимальных поправок обладает асимптотическим свойством.

4. Метод минимальных погрешностей. Этот метод можно применять, как и метод минимальных невязок, в случае любого несамосопряженного и невырожденного оператора A . *Метод минимальных погрешностей* определяется следующим выбором операторов B и D :

$$B = (A^*)^{-1}B_0, \quad D = B_0,$$

где B_0 — самосопряженный положительно определенный в H оператор.

Подставляя в формулу (3) для итерационного параметра τ_{k+1} выбранный оператор D и учитывая, что $w_k = B^{-1}r_k = B_0^{-1}A^*r_k$, получим формулу для τ_{k+1} в методе минимальных погрешностей

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Aw_k, r_k)}, \quad k = 0, 1, \dots$$

Поправка w_k находится из уравнения $B_0w_k = A^*r_k$.

В случае явной схемы ($B_0 = E$) формула для τ_{k+1} имеет вид

$$\tau_{k+1} = \frac{(r_k, r_k)}{(A^*r_k, A^*r_k)}, \quad k = 0, 1, \dots$$

В методе минимальных погрешностей минимизируется норма погрешности в H_{B_0} . Для этого метода оператор $DB^{-1}A = A^*A$ является самосопряженным в H , а условия (3) § 1 принимают вид неравенств (6). Из теоремы 1 следует оценка сходимости метода

$$\|z_n\|_{B_0} \leq \rho^n \|z_0\|_{B_0}, \quad n \geq n_0(\varepsilon) = \ln \varepsilon / \ln \rho,$$

где $\rho = (1 - \xi) / (1 + \xi)$, $\xi = \gamma_1 / \gamma_2$, а γ_1 и γ_2 определены в (6).

Метод минимальных погрешностей всегда обладает асимптотическим свойством.

5. Пример применения двухслойных методов. Для иллюстрации применения построенных двухслойных градиентных методов рассмотрим решение модельной задачи явным методом скорейшего спуска. В качестве примера возьмем разностную задачу Дирихле для уравнения Пуассона на квадратной сетке $\omega = \{x_{ij} = (ih, jh), 0 \leq i \leq N, 0 \leq j \leq N, h = 1/N\}$ в единичном квадрате

$$\Delta u = u_{x_1 x_1} + u_{x_2 x_2} = -\varphi, \quad x \in \omega, \quad u|_{\gamma} = g. \quad (7)$$

Введем пространство H , состоящее из сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{x \in \omega} u(x)v(x)h^2$.

Оператор A на H определим следующим образом: $Ay = -\Lambda v$, $y \in H$, где $v(x) = y(x)$ для $x \in \omega$ и $v|_{\gamma} = 0$. Задачу (7) запишем в виде операторного уравнения

$$Au = f, \quad (8)$$

где f отличается от φ лишь в приграничных узлах

$$f = \varphi + \frac{\varphi_1}{h^2} + \frac{\varphi_2}{h^2},$$

$$\varphi_1 = \begin{cases} g(0, x_2), & x_1 = h, \\ 0, & 2h \leqslant x_1 \leqslant 1 - 2h, \\ g(1, x_2), & x_1 = 1 - h, \end{cases} \quad \varphi_2 = \begin{cases} g(x_1, 0), & x_2 = h, \\ 0, & 2h \leqslant x_2 \leqslant 1 - 2h, \\ g(x_1, 1), & x_2 = 1 - h. \end{cases}$$

Оператор A является самосопряженным и положительно определенным в H . Поэтому для решения уравнения (8) можно применить метод скорейшего спуска. Явная итерационная схема имеет вид

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f \quad \text{или} \quad y_{k+1} = y_k - \tau_{k+1}r_k, \quad k = 0, 1, \dots,$$

а итерационные параметры τ_k находятся по формуле

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \quad r_k = Ay_k - f, \quad k = 0, 1, \dots$$

Приведем расчетные формулы и подсчитаем число арифметических действий, затрачиваемых на одну итерацию.

Учитывая определение оператора A и правой части f , расчетные формулы можно записать в следующем виде:

$$1) \quad r_k(x_{ij}) = -(y_k)_{\bar{x}_1 x_1} - (y_k)_{\bar{x}_2 x_2} - \varphi(x_{ij}), \quad 1 \leqslant i, j \leqslant N-1, \\ y_k|_{\gamma} = g;$$

$$2) \quad (r_k, r_k) = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} r_k^2(x_{ij})h^2,$$

$$(Ar_k, r_k) = - \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} r_k(x_{ij}) [(r_k)_{\bar{x}_1 x_1} + (r_k)_{\bar{x}_2 x_2}]h^2, \quad r_k|_{\gamma} = 0, \\ \tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)};$$

$$3) \quad y_{k+1}(x_{ij}) = y_k(x_{ij}) - \tau_{k+1}r_k(x_{ij}), \quad 1 \leqslant i, j \leqslant N-1.$$

Начальное приближение y_0 есть произвольная сеточная функция в ω , принимающая на γ заданные значения $y_0|_{\gamma} = g$.

Подсчитаем число арифметических действий. Если вычисление разностных производных проводить по формуле

$$u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2} = \frac{1}{h^2} (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij}),$$

то для вычисления r_k потребуется $6(N-1)^2$ сложений и $2(N-1)^2$ умножений и делений. Для вычисления $(r_k, r_k) - (N-1)^2$ сложений и $(N-1)^2$ умножений, $(Ar_k, r_k) - 6(N-1)^2$ сложений и $2(N-1)^2$ умножений, $y_{k+1} - (N-1)^2$ сложений и $(N-1)^2$ умножений. Всего потребуется $14(N-1)^2$ сложений и $6(N-1)^2$ умножений и делений. Ровно половина от этого общего числа действий требуется для вычисления скалярных произведений, т. е. для вычисления итерационного параметра τ_{k+1} . Следовательно, один шаг метода скорейшего спуска примерно вдвое более трудоемок по сравнению с шагом метода простой итерации или чебышевского метода, где параметры τ_{k+1} известны априори. Для неявных методов это различие будет меньшим, так как для вычисления скалярных произведений потребуется такое же число действий, как и для явного метода, а к общему числу действий добавятся арифметические операции, затрачиваемые на обращение оператора B .

Подсчитаем теперь общее число арифметических действий $Q(\varepsilon)$, которое нужно выполнить, чтобы получить относительную точность ε . Для этого нужно оценить число итераций $n_0(\varepsilon)$. В п. 1 была получена следующая оценка:

$$n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho}, \quad \rho = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

где γ_1 и γ_2 в случае явной схемы есть границы оператора A : $\gamma_1 E \leqslant A \leqslant \gamma_2 E$.

Для рассматриваемого примера γ_1 и γ_2 совпадают с минимальным δ и максимальным Δ собственными значениями разностного оператора Лапласа Λ . Известно, что

$$\delta = \frac{8}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \frac{8}{h^2} \cos^2 \frac{\pi h}{2}.$$

Поэтому

$$\rho = \frac{1-\xi}{1+\xi} = 1 - 2 \sin^2 \frac{\pi h}{2}, \quad \xi = \frac{\delta}{\Delta} = \operatorname{tg}^2 \frac{\pi h}{2},$$

и, следовательно, если $h \ll 1$, то

$$n_0(\varepsilon) \approx \frac{2 \ln \frac{1}{\varepsilon}}{\pi^2 h^2} \approx 0,2 N^2 \ln \frac{1}{\varepsilon}.$$

Если считать эквивалентными операции сложения, умножения и деления, то на одну итерацию затрачивается примерно $20N^2$ действий. Поэтому для общего числа арифметических операций будет справедлива оценка $Q(\varepsilon) \approx 4N^4 \ln \frac{1}{\varepsilon}$.

§ 3. Трехслойные методы сопряженных направлений

1. Постановка задачи о выборе итерационных параметров.

Оценка скорости сходимости. Для нахождения приближенного решения линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A в § 1 мы рассмотрели двухслойные итерационные методы вариационного типа. Итерационная схема для этих методов имеет вид

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (2)$$

а итерационные параметры τ_{k+1} выбираются из условия минимума нормы погрешности z_{k+1} в энергетическом пространстве H_D . Напомним, что на последовательности y_k , построенной по формуле (2), осуществляется пошаговая минимизация функционала $I(y) = (D(y - u), y - u)$, минимум которого достигается на решении уравнения (1), т. е. при $y = u$.

Такая стратегия локальной минимизации, однако, не является оптимальной, так как в конечном счете нас интересует глобальный минимум функционала $I(y)$, и, если задано некоторое значение этого функционала, достичь искомого минимума мы должны за минимальное число итераций. Локальная же минимизация на каждом итерационном шаге приводит к решению этой задачи не самым коротким путем.

Естественно попытаться выбрать параметры τ_k из условия минимума нормы погрешности z_n в H_D сразу за n шагов, т. е. при переходе от y_0 к y_n . С аналогичной ситуацией мы уже встречались в главе VI при изучении чебышевского метода и метода простой итерации. Оказалось, что более быстро сходится тот метод, итерационные параметры для которого выбираются из условия минимума нормы разрешающего оператора, а не оператора перехода от итерации к итерации. Это свойство имеет место и для итерационных методов вариационного типа. Будет показано, что рассматриваемые в этом параграфе итерационные методы, параметры τ_k для которых выбираются из указанного выше условия, сходятся значительно быстрее, чем двухслойные градиентные методы. Более того, в случае конечномерного пространства H эти методы являются методами конечных итераций при любом начальном приближении, т. е. точное решение уравнения (1) может быть получено за конечное число итераций.

Переходим к построению *метода сопряженных направлений*. Будем предполагать, что оператор $DB^{-1}A$ самосопряжен и положительно определен в H . Выполним по схеме (2) n итераций. Переходя от задачи для погрешности $z_k = y_k - u$ к задаче для

$x_k = D^{1/2}z_k$, получим, как и раньше,

$$\begin{aligned} x_{k+1} &= S_{k+1}x_k, \quad k = 0, 1, \dots, n-1, \\ S_k &= E - \tau_k C, \quad C = D^{1/2}B^{-1}AD^{-1/2}. \end{aligned}$$

Отсюда найдем

$$x_n = T_n x_0, \quad T_n = \prod_{j=1}^n (E - \tau_j C). \quad (3)$$

Разрешающий оператор T_n представляет собой операторный полином степени n относительно оператора C с коэффициентами, зависящими от параметров $\tau_1, \tau_2, \dots, \tau_n$

$$T_n = P_n(C) = E + \sum_{j=1}^n a_j^{(n)} C^j, \quad a_n^{(n)} \neq 0. \quad (4)$$

В силу равенства $\|x_n\| = \|z_n\|_D$ поставленная выше задача о выборе итерационных параметров τ_k формулируется следующим образом: среди всех полиномов вида (4) выбрать тот, для которого норма $x_n = P_n(C)x_0$ минимальна, другими словами — выбрать коэффициенты $a_1^{(n)}, a_2^{(n)}, \dots, a_n^{(n)}$ полинома $P_n(C)$ из условия минимума нормы x_n в H .

Эта задача будет решена в следующем пункте, а сначала получим оценку скорости сходимости метода сопряженных направлений, построенного на основе сформулированного выше принципа выбора параметров. Эту оценку получим, используя априорную информацию об операторах схемы в виде γ_1 и γ_2 — постоянных энергетической эквивалентности самосопряженных операторов D и $DB^{-1}A$:

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \quad (5)$$

Пусть $P_n(C)$ является искомым полиномом. Тогда из (3), (4) следует оценка для x_n :

$$\|x_n\| = \|P_n(C)x_0\| = \min_{\{Q_n\}} \|Q_n(C)x_0\| \leq \min_{\{Q_n\}} \|Q_n(C)\| \|x_0\|,$$

где минимум ищется среди полиномов $Q_n(C)$, нормированных в силу (4) условием $Q_n(0) = E$.

Оценим минимум нормы полинома $Q_n(C)$. Из (5) следует, что оператор $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ самосопряжен в H , а γ_1 и γ_2 — его границы: $C = C^*$, $\gamma_1 E \leq C \leq \gamma_2 E$, $\gamma_1 > 0$. Поэтому имеет место оценка

$$\min_{\{Q_n\}} \|Q_n(C)\| \leq \min_{\{Q_n\}} \max_{\gamma_1 \leq t \leq \gamma_2} |Q_n(t)|.$$

Из результатов § 2 гл. VI следует, что задачу о построении нормированного условием $Q_n(0) = 1$ полинома, максимум модуля

которого на отрезке $[\gamma_1, \gamma_2]$ минимален, решает полином Чебышева 1-го рода, для которого

$$\max_{\gamma_1 \leq t \leq \gamma_2} |Q_n(t)| = q_n, \quad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Следовательно, для x_n имеет место оценка $\|x_n\| \leq q_n \|x_0\|$.

Итак, доказана

Теорема 2. Если выполнены условия (5), то итерационный метод сопряженных направлений сходится в H_D , и для погрешности z_n при любом n справедлива оценка $\|z_n\|_D \leq q_n \|z_0\|_D$. При этом оценка для числа итераций имеет вид

$$n \geq n_0(\epsilon) = \ln(0.5\epsilon)/\ln \rho_1,$$

где $\rho_1 = (1 - \sqrt{\xi})/(1 + \sqrt{\xi})$, $\xi = \gamma_1/\gamma_2$.

2. Формулы для итерационных параметров. Трехслойная итерационная схема. Переходим теперь к построению полинома $P_n(C)$. Используя (3) и (4), вычислим норму x_n :

$$\|x_n\|^2 = (P_n(C)x_0, P_n(C)x_0) =$$

$$= \|x_0\|^2 + 2 \sum_{j=1}^n a_j^{(n)} (C^j x_0, x_0) + \sum_{j=1}^n \sum_{i=1}^n a_j^{(n)} a_i^{(n)} (C^j x_0, C^i x_0).$$

Норма x_n есть функция параметров $a_1^{(n)}, a_2^{(n)}, \dots, a_n^{(n)}$. Прививая частные производные от $\|x_n\|^2$ по $a_j^{(n)}$

$$\frac{\partial \|x_n\|^2}{\partial a_j^{(n)}} = 2 \sum_{i=1}^n a_i^{(n)} (C^j x_0, C^i x_0) + 2 (C^j x_0, x_0), \quad j = 1, 2, \dots, n,$$

нулю, получим систему линейных алгебраических уравнений

$$\sum_{i=1}^n a_i^{(n)} (C^j x_0, C^i x_0) + (C^j x_0, x_0) = 0, \quad j = 1, 2, \dots, n. \quad (6)$$

Для самосопряженного и положительно определенного в H оператора C система (6) дает условия минимума нормы x_n в H .

Итак, задача построения оптимального полинома $P_n(C)$ в принципе решена. Коэффициенты полинома $a_1^{(n)}, a_2^{(n)}, \dots, a_n^{(n)}$ найдем, решая систему (6). Но сначала построим формулы для вычисления итерационного приближения y_n . Первый путь состоит в использовании итерационной схемы (2). Однако для этого потребуется найти корни полинома $P_n(t)$ и затем в качестве τ_k взять обратные к корням значения. Такой способ не является экономичным.

Второй путь заключается в использовании для вычисления y_n коэффициентов полинома. Из (3), (4) и замены $x_k = D^{1/2} z_k$, где $z_k = y_k - u$, получим

$$y_n - u = D^{-1/2} P_n(C) D^{1/2} (y_0 - u). \quad (7)$$

Используя (4) и равенство $D^{-1/2}C^jD^{1/2} = (B^{-1}A)^j$, найдем

$$D^{-1/2}P_n(C)D^{1/2} = E + \sum_{j=1}^n a_j^{(n)} (B^{-1}A)^j.$$

Подставляя это равенство в (7), получим

$$y_n = y_0 + \sum_{j=1}^n a_j^{(n)} (B^{-1}A)^j (y_0 - u) = y_0 + \sum_{j=1}^n a_j^{(n)} (B^{-1}A)^{j-1} w_0, \quad (8)$$

где w_0 — поправка, $w_0 = B^{-1}A(y_0 - u) = B^{-1}r_0$, $r_0 = Ay_0 - f$.

Этот путь также является не оптимальным. Для каждого нового n приходится заново решать систему (6).

Сейчас мы покажем, что последовательность $y_1, y_2, \dots, y_k, \dots$, построенная согласно (6), (8) для $n = 1, 2, \dots$, может быть найдена по следующей трехслойной схеме:

$$\begin{aligned} By_{k+1} &= \alpha_{k+1} (B - \tau_{k+1} A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \alpha_{k+1} \tau_{k+1} f, \\ k &= 1, 2, \dots, \end{aligned} \quad (9)$$

$$By_1 = (B - \tau_1 A) y_0 + \tau_1 f, \quad y_0 \in H.$$

Для этого нужно указать набор параметров $\{\tau_k\}$ и $\{\alpha_k\}$, при котором норма эквивалентной погрешности x_k была бы минимальной для любого k . Действительно, из уравнения для погрешности x_k в случае схемы (9)

$$\begin{aligned} x_{k+1} &= \alpha_{k+1} (E - \tau_{k+1} C) x_k + (1 - \alpha_{k+1}) x_{k-1}, \quad k = 1, 2, \dots, \\ x_1 &= (E - \tau_1 C) x_0, \end{aligned} \quad (10)$$

получим, что $x_k = P_k(C) x_0$, где полином $P_k(C)$ имеет вид (4) ($n = k$). Поэтому, если параметры $\{\tau_k\}$ и $\{\alpha_k\}$ в (9) будут выбраны так, чтобы для любого $n = 1, 2, \dots$ выполнялись условия (6), то построенные согласно (9) итерационные приближения y_n будут совпадать с приближениями, построенными по формулам (6), (8) для любого n .

Построим искомый набор параметров $\{\tau_k\}$ и $\{\alpha_k\}$. Для этого нам потребуется

Лемма. Необходимыми и достаточными условиями минимума нормы x_n в H для любого $n \geq 1$ являются условия

$$(Cx_j, x_n) = 0, \quad j = 0, 1, \dots, n-1. \quad (11)$$

Действительно, из (4), (6) следует, что условия (6), являющиеся условиями минимума нормы x_n , эквивалентны следующим:

$$(C^j x_0, x_n) = 0, \quad j = 1, 2, \dots, n, \quad (12)$$

для любого $n = 1, 2, \dots$. Отсюда получим для $j \leq n-1$

$$(Cx_0, x_n) + \sum_{i=2}^{j+1} a_i^{(n)} (C^i x_0, x_n) = (Cx_j, x_n) = 0,$$

т. е. условия (11) необходимы.

Докажем теперь достаточность условий (11). Пусть выполнены условия (11). Покажем, что тогда выполнены и условия (12). Из (11) при $j=0$ получаем, что равенства (12) верны для $j=1$. Справедливость (12) для $j \geq 2$ докажем по индукции. Пусть для $j \leq k$ условия (12) уже выполнены, т. е. $(C^j x_0, x_n) = 0$, $j=1, 2, \dots, k$. Покажем, что они выполнены и при $j=k+1$, если выполнены условия (11).

Действительно, из (11) при $j=k$ получим

$$\begin{aligned} 0 &= (Cx_k, x_n) = (CP_k(C)x_0, x_n) = \\ &= (Cx_0, x_n) + \sum_{j=1}^k a_j^{(k)} (C^{j+1}x_0, x_n) = a_k^{(k)} (C^{k+1}x_0, x_n). \end{aligned}$$

Следовательно, $(C^{k+1}x_0, x_n) = 0$. Лемма доказана.

Воспользуемся теперь леммой для построения набора параметров $\{\tau_k\}$ и $\{\alpha_k\}$ для схемы (9). Для сокращения выкладок будем считать, что y_1 в схеме (9) находится по общей формуле (9) при $\alpha_1 = 1$.

Рассмотрим схему (10). Так как x_1 находится по двухслойной схеме, то из § 1 следует выбор оптимального параметра τ_1 по формуле

$$\tau_1 = \frac{(Cx_0, x_0)}{(Cx_0, Cx_0)}.$$

Построение параметров τ_2, τ_3, \dots и $\alpha_2, \alpha_3, \dots$ будем осуществлять постепенно. Пусть итерационные параметры $\tau_1, \tau_2, \dots, \tau_k$ и $\alpha_1, \alpha_2, \dots, \alpha_k$ уже выбраны оптимальным образом. Так как эти параметры определяют приближения y_1, y_2, \dots, y_k , то из леммы следует, что выполнены условия

$$(Cx_j, x_i) = 0, \quad j = 0, 1, \dots, i-1, \quad i = 1, 2, \dots, k. \quad (13)$$

Выберем теперь параметры τ_{k+1} и α_{k+1} , определяющие приближение y_{k+1} . Из леммы следует, что норма x_{k+1} будет минимальна, если выполнены условия

$$(Cx_j, x_{k+1}) = 0, \quad j = 0, 1, \dots, k. \quad (14)$$

Из этих условий найдем параметры τ_{k+1} и α_{k+1} . Покажем сначала, что из (13) следует выполнение условий (14) для $j \leq k-2$, а затем из оставшихся двух условий (14) для $j=k-1$ и $j=k$ получим формулы для τ_{k+1} и α_{k+1} .

Итак, пусть $j \leq k-2$. Из (10) и (13) найдем

$$\begin{aligned} (x_{k+1}, Cx_j) &= \alpha_{k+1}(x_k, Cx_j) - \alpha_{k+1}\tau_{k+1}(Cx_k, Cx_j) + \\ &\quad + (1-\alpha_{k+1})(x_{k-1}, Cx_j) = -\alpha_{k+1}\tau_{k+1}(Cx_k, Cx_j). \end{aligned}$$

Покажем, что $(Cx_k, Cx_j) = 0$ для $j \leq k-2$. Действительно, из (10)

при $k = j$ получим

$$Cx_j = \frac{1}{\tau_{j+1}} x_j - \frac{1}{\tau_{j+1}\alpha_{j+1}} [x_{j+1} - (1 - \alpha_{j+1}) x_{j-1}], \quad j \geq 0. \quad (15)$$

Используя самосопряженность оператора C и условия (13), отсюда получим для $j \leq k-2$

$$\begin{aligned} (Cx_k, Cx_j) &= \\ &= \frac{1}{\tau_{j+1}} (Cx_j, x_k) - \frac{1}{\tau_{j+1}\alpha_{j+1}} [(Cx_{j+1}, x_k) - (1 - \alpha_{j+1})(Cx_{j-1}, x_k)] = 0. \end{aligned}$$

Следовательно, $(x_{k+1}, Cx_j) = 0$ для $j \leq k-2$.

Найдем теперь τ_{k+1} и α_{k+1} . Полагая в (14) $j = k-1$ и $j = k$, получим из (10) и (13)

$$\begin{aligned} 0 &= (Cx_{k-1}, x_{k+1}) = \\ &= -\alpha_{k+1}\tau_{k+1}(Cx_k, Cx_{k-1}) + (1 - \alpha_{k+1})(Cx_{k-1}, x_{k-1}), \\ 0 &= (Cx_k, x_{k+1}) = \alpha_{k+1}[(Cx_k, x_k) - \tau_{k+1}(Cx_k, Cx_k)]. \end{aligned} \quad (16)$$

Из второго уравнения сразу найдем параметр τ_{k+1} :

$$\tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \quad (17)$$

Из первого уравнения исключим выражение (Cx_k, Cx_{k-1}) . Для этого положим в (15) $j = k-1$ и умножим скалярно левую и правую части (15) на Cx_k .

Учитывая самосопряженность оператора C из условия (13), получим

$$\begin{aligned} (Cx_k, Cx_{k-1}) &= \frac{1}{\tau_k} (Cx_{k-1}, x_k) - \frac{1}{\tau_k\alpha_k} (Cx_k, x_k) + \frac{1 - \alpha_k}{\tau_k\alpha_k} (Cx_{k-2}, x_k) = \\ &= -\frac{1}{\tau_k\alpha_k} (Cx_k, x_k). \end{aligned}$$

Подставляя это выражение в (16), получим

$$\frac{\alpha_{k+1}\tau_{k+1}}{\alpha_k\tau_k} \frac{(Cx_k, x_k)}{(Cx_{k-1}, x_{k-1})} + (1 - \alpha_{k+1}) = 0.$$

Из этого равенства найдем рекуррентную формулу для параметра α_{k+1} :

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Cx_k, x_k)}{(Cx_{k-1}, x_{k-1})} \frac{1}{\alpha_k} \right)^{-1}. \quad (18)$$

Итак, предполагая, что итерационные параметры $\tau_1, \tau_2, \dots, \tau_k$ и $\alpha_1, \alpha_2, \dots, \alpha_k$ были уже выбраны ранее, мы получим формулы для параметров τ_{k+1} и α_{k+1} . Так как $\alpha_1 = 1$ и $\tau_1 = \frac{(Cx_0, x_0)}{(Cx_0, Cx_0)}$, то формулы (17), (18) определяют параметры τ_{k+1} и α_{k+1} для любого k .

Подставляя $x_k = D^{1/2}z_k$ в (17) и (18) и учитывая, что $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ и $Az_k = r_k$, $B^{-1}r_k = w_k$,

получим следующие формулы для итерационных параметров τ_{k+1} и α_{k+1} :

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots, \quad (19)$$

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Dw_k, z_k)}{(Dw_{k-1}, z_{k-1})} \frac{1}{\alpha_k} \right)^{-1}, \\ k = 1, 2, \dots, \quad \alpha_1 = 1. \quad (20)$$

Итак, метод сопряженных направлений описывается трехслойной схемой (9), итерационные параметры τ_{k+1} и α_{k+1} для которой выбираются по формулам (19), (20). Для этого метода справедлива доказанная ранее теорема 2.

Из формул (19), (20) следует, что итерационные параметры τ_{k+1} в методе сопряженных направлений и в двухслойных градиентных методах выбираются по одним и тем же формулам, а для вычисления параметров α_{k+1} никаких дополнительных скалярных произведений считать не нужно. Поэтому на вычисление итерационных параметров в двухслойных и трехслойных методах вариационного типа затрачивается практически одинаковое число арифметических операций. В то же время из теорем 1 и 2 следует, что методы сопряженных направлений сходятся существенно быстрее, чем градиентные методы.

Покажем теперь, что если H — конечномерное пространство ($H = H_N$), то методы сопряженных направлений сходятся за конечное число итераций, не превышающее размерность пространства. Действительно, из леммы следует, что для эквивалентных погрешностей x_k метода сопряженных направлений должны выполняться равенства $(Cx_j, x_n) = (x_j, x_n)_C = 0$, $j = 0, 1, \dots, n$. Следовательно, система векторов x_0, x_1, \dots, x_n для любого n должна быть ортогональной системой в H_C . А так как в H_N нельзя построить больше чем N ортогональных векторов, то отсюда следует, что $x_N = 0$ и $z_N = y_N - u = 0$. Таким образом, на классе произвольных начальных приближений y_0 методы сопряженных направлений сходятся за N итераций к точному решению уравнения (1).

Для специальных начальных приближений y_0 эти методы сходятся за меньшее число итераций. Действительно, пусть y_0 таково, что в разложении x_0 по собственным функциям оператора C присутствуют $N_0 < N$ функций, т. е. x_0 принадлежит инвариантному относительно оператора C подпространству H_{N_0} . Тогда очевидно, что и все $x_k \in H_{N_0}$. Поэтому в этом случае итерационный процесс сойдется за N_0 итераций.

Из сказанного выше не следует, что оценка сходимости метода, полученная в теореме 2, является очень грубой и равенство $\|z_n\|_D = q_n \|z_0\|_D$ никогда не достигается. Можно построить пример уравнения (1) и указать для любого $n < N$ такое начальное приближение y_0 , что указанное равенство будет выполняться.

3. Варианты расчетных формул. Приведем теперь некоторые способы реализации трехслойных методов сопряженных направлений. Из (9), (19) и (20) получим следующий алгоритм:

- 1) по заданному y_0 вычисляется невязка $r_0 = Ay_0 - f$;
- 2) решается уравнение для поправки $Bw_0 = r_0$;
- 3) вычисляется параметр τ_1 по формуле (19);
- 4) приближение y_1 находится по формуле $y_1 = y_0 - \tau_1 w_0$.

Далее для $k = 1, 2, \dots$ последовательно выполняются следующие действия:

- 5) вычисляется невязка $r_k = Ay_k - f$ и решается уравнение для поправки $Bw_k = r_k$;
- 6) по формулам (19), (20) вычисляются параметры τ_{k+1} и α_{k+1} ;
- 7) приближение y_{k+1} находится по формуле

$$y_{k+1} = \alpha_{k+1} y_k + (1 - \alpha_{k+1}) y_{k-1} - \alpha_{k+1} \tau_{k+1} w_k.$$

Таким образом, в описанном алгоритме для нахождения y_{k+1} используются y_{k-1} , y_k и w_k , которые необходимо запоминать. Ниже будет указан вид формул (19) и (20) для некоторых конкретных выборов оператора D . Здесь мы ограничимся замечанием, что в эти формулы, помимо хранимой величины w_k , может входить невязка r_k , которую мы не запоминаем. Для ее вычисления можно воспользоваться либо равенством $r_k = Bw_k$, если вычисление значения Bw_k не является трудоемким процессом, либо определением невязки $r_k = Ay_k - f$.

На практике встречаются и другие алгоритмы реализации метода сопряженных направлений. Приведем один из них. Для этого будем трактовать схему (9) как схему с поправкой. Из (9) получим

$$y_{k+1} = y_k - a_{k+1} s_k, \quad s_{k+1} = w_{k+1} + b_{k+1} s_k, \quad k = 0, 1, \dots, s_0 = w_0, \quad (21)$$

где $w_k = B^{-1} r_k$, $r_k = Ay_k - f$, а параметры a_{k+1} и b_k связаны с α_{k+1} и τ_{k+1} следующими формулами:

$$a_{k+1} = \alpha_{k+1} \tau_{k+1}, \quad b_k = (\alpha_{k+1} - 1) \alpha_k \tau_k / (\alpha_{k+1} \tau_{k+1}).$$

Получим выражения для b_k и a_{k+1} . Из (19), (20) найдем

$$b_k = (Dw_k, z_k) / (Dw_{k-1}, z_{k-1}), \quad k = 1, 2, \dots \quad (22)$$

Для a_{k+1} из этих же формул легко получить рекуррентные соотношения, однако можно найти явное выражение для a_{k+1} :

$$a_{k+1} = \frac{(Cx_k, x_k)}{(p_k, p_k)} = \frac{(Dw_k, z_k)}{(Ds_k, s_k)}, \quad k = 0, 1, \dots \quad (23)$$

Формулы (21), (22) и (23) описывают второй алгоритм метода сопряженных направлений. Здесь вычисления проводятся в следующем порядке:

- 1) по заданному y_0 вычисляется невязка $r_0 = Ay_0 - f$, решается уравнение $Bw_0 = r_0$ для поправки w_0 и полагается $s_0 = w_0$;

Далее для $k = 1, 2, \dots$ последовательно выполняются действия:

- 2) по формуле (23) находится параметр a_1 и вычисляется $y_1 = y_0 - a_1 s_0$;
- 3) вычисляется невязка $r_k = Ay_k - f$ и решается уравнение для поправки $Bw_k = r_k$;

4) по формуле (22) вычисляется параметр b_k и находится s_k по формуле $s_k = w_k + b_k s_{k-1}$;

5) по формуле (23) определяется параметр a_{k+1} и приближение y_{k+1} вычисляется по формуле

$$y_{k+1} = y_k - a_{k+1} s_k.$$

Отметим, что в приведенном алгоритме необходимо хранить y_k , w_k и s_k , т. е. такой же объем промежуточной информации, как и в первом алгоритме.

§ 4. Примеры трехслойных методов

1. Частные случаи методов сопряженных направлений. В § 3 были построены трехслойные итерационные методы сопряженных направлений, используемые для решения линейного уравнения

$$Au = f. \quad (1)$$

Итерационные приближения вычисляются по трехслойной схеме

$$\begin{aligned} By_{k+1} &= \alpha_{k+1} (B - \tau_{k+1} A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \alpha_{k+1} \tau_{k+1} f, \\ k &= 1, 2, \dots, \\ By_1 &= (B - \tau_1 A) y_0 + \tau_1 f, \quad y_0 \in H, \end{aligned} \quad (2)$$

а итерационные параметры α_{k+1} и τ_{k+1} находятся по формулам

$$\begin{aligned} \tau_{k+1} &= \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots, \\ \alpha_{k+1} &= \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Dw_k, z_k)}{(Dw_{k-1}, z_{k-1})} \cdot \frac{1}{\alpha_k} \right)^{-1}, \quad k = 1, 2, \dots, \alpha_1 = 1, \end{aligned} \quad (3)$$

где $w_k = B^{-1}r_k$ — поправка, $r_k = Ay_k - f$ — невязка, $z_k = \hat{y}_k - u$ — погрешность.

Выбор параметров α_k и τ_k по формулам (3) обеспечивает в случае самосопряженного и положительно определенного оператора $DB^{-1}A$ минимум для любого n нормы погрешности z_n в H_D при переходе от y_0 к y_n .

Рассмотрим теперь частные случаи методов сопряженных направлений, определяемые выбором оператора D . В § 2 были рассмотрены четыре примера двухслойных градиентных методов. Каждому из этих двухслойных методов соответствует определенный трехслойный метод сопряженных направлений. Мы перечислим эти методы с указанием условий на операторы A и B , которые обеспечивают самосопряженность оператора $DB^{-1}A$. Для этих методов имеет место теорема 2, а вид неравенств, которые определяют постоянные γ_1 и γ_2 , будет указан в описании соответствующего метода.

1) *Метод сопряженных градиентов.*

Оператор D : $D = A$.

Условия: $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$, $A = A^* > 0$, $B = B^* > 0$.

Формулы для итерационных параметров:

$$\tau_{k+1} = \frac{(r_k, w_k)}{(Aw_k, w_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, w_k)}{(r_{k-1}, w_{k-1})} \frac{1}{\alpha_k} \right)^{-1}.$$

2) *Метод сопряженных невязок.*

Оператор D : $D = A^*A$.

Условия: $\gamma_1 (Bx, Bx) \leqslant (Ax, Bx) \leqslant \gamma_2 (Bx, Bx)$, $\gamma_1 > 0$, $B^*A = A^*B$.

Если выполнены предположения $A = A^* > 0$, $B = B^* > 0$, $AB = BA$, то условия имеют вид

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0.$$

Формулы для итерационных параметров:

$$\tau_{k+1} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Aw_k, r_k)}{(Aw_{k-1}, r_{k-1})} \cdot \frac{1}{\alpha_k} \right)^{-1}.$$

3) *Метод сопряженных поправок.*

Оператор D : $D = AB^{-1}A$.

Условия: $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$, $A = A^* > 0$, $B = B^* > 0$.

Формулы для итерационных параметров:

$$\tau_{k+1} = \frac{(Aw_k, w_k)}{(B^{-1}Aw_k, Aw_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Aw_k, w_k)}{(Aw_{k-1}, w_{k-1})} \cdot \frac{1}{\alpha_k} \right)^{-1}.$$

4) *Метод сопряженных погрешностей.*

Оператор D : $D = B_0$.

Условия: $B = (A^*)^{-1}B_0$, $\gamma_1 B_0 \leqslant A^*A \leqslant \gamma_2 B_0$, $B_0 = B_0^* > 0$.

Формулы для итерационных параметров:

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Aw_k, r_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \cdot \frac{1}{\alpha_k} \right)^{-1}.$$

2. Локально оптимальные трехслойные методы. Вернемся теперь к рассмотренному в § 3 способу построения итерационных параметров α_{k+1} и τ_{k+1} для трехслойной схемы метода сопряженных направлений. Напомним, что параметры α_{k+1} и τ_{k+1} были выбраны из условий $(Cx_{k-1}, x_{k+1}) = 0$ и $(Cx_k, x_{k+1}) = 0$ в предположении, что итерационные приближения y_1, y_2, \dots, y_k обеспечивают выполнение условий

$$(Cx_j, x_i) = 0, \quad j = 0, 1, \dots, i-1, \quad i = 1, 2, \dots, k. \quad (4)$$

Для идеального вычислительного процесса условия (4) выполнены, поэтому выбор параметров α_{k+1} и τ_{k+1} по полученным в § 3 формулам действительно обеспечивает минимум нормы погрешности z_{k+1} в H_D при переходе от y_0 к y_{k+1} . В реальном вычислительном процессе, который учитывает наличие ошибок округления, итерационные приближения y_1, y_2, \dots, y_k будут вычислены неточно, и, следовательно, условия (4) не будут выполняться. В ряде случаев это может привести к уменьшению скорости сходимости метода, а иногда и к его расходимости.

Построим сейчас одну модификацию метода сопряженных направлений, не обладающую указанным недостатком. Для приближенного решения уравнения $Au = f$ рассмотрим трехслойную итерационную схему

$$By_{k+1} = \alpha_{k+1} (B - \tau_{k+1} A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \alpha_{k+1} \tau_{k+1} f, \quad k = 1, 2, \dots \quad (5)$$

с произвольными приближениями y_0 и $y_1 \in H$. Считая y_k и y_{k+1} заданными, выберем параметры α_{k+1} и τ_{k+1} из условия минимума нормы погрешности z_{k+1} в H_D , т. е. из условия локальной оптимизации за один шаг по трехслойной схеме.

Эту задачу решим при единственном предположении о положительной определенности оператора $DB^{-1}A$. Для этого перейдем к уравнению для эквивалентной погрешности $x_k = D^{1/2}z_k$:
 $x_{k+1} = \alpha_{k+1}(E - \tau_{k+1}C)x_k + (1 - \alpha_{k+1})x_{k-1}, \quad C = D^{1/2}B^{-1}AD^{-1/2}. \quad (6)$
Для сокращения выкладок введем обозначения

$$1 - \alpha_{k+1} = a, \quad \tau_{k+1}\alpha_{k+1} = b \quad (7)$$

и перепишем (6) в следующем виде:

$$x_{k+1} = x_k - a(x_k - x_{k-1}) - bCx_k. \quad (8)$$

Задача ставится так: выбрать a и b из условия минимума нормы x_{k+1} в H . Вычислим норму x_{k+1} . Из (8) получим

$$\|x_{k+1}\|^2 = \|x_k\|^2 + a^2\|x_k - x_{k-1}\|^2 + b^2\|Cx_k\|^2 - 2a(x_k, x_k - x_{k-1}) - 2b(Cx_k, x_k) + 2ab(Cx_k, x_k - x_{k-1}).$$

Приравнивая частные производные по a и b нулю, получим систему относительно параметров a и b

$$\begin{aligned} \|x_k - x_{k-1}\|^2 a + (Cx_k, x_k - x_{k-1}) b &= (x_k, x_k - x_{k-1}), \\ (Cx_k, x_k - x_{k-1}) a + \|Cx_k\|^2 b &= (Cx_k, x_k). \end{aligned} \quad (9)$$

Определитель системы равен $\|x_k - x_{k-1}\|^2 \|Cx_k\|^2 - (Cx_k, x_k - x_{k-1})^2$ и в силу неравенства Коши—Буняковского обращается в нуль лишь тогда, когда $x_k - x_{k-1}$ пропорционально Cx_k : $x_k - x_{k-1} = dCx_k$. В этом случае уравнения системы пропорциональны, и она сводится к одному уравнению

$$(b + ad)\|Cx_k\|^2 = (Cx_k, x_k). \quad (10)$$

Так как при этом (8) имеет вид $x_{k+1} = x_k - (b + ad)Cx_k$, то, полагая в (10) $a = 0$, получим из (7), (10)

$$\alpha_{k+1} = 1, \quad \tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \quad (11)$$

Если определитель не равен нулю, то, решая систему (9), получим

$$\begin{aligned} a &= \frac{\|Cx_k\|^2(x_k, x_k - x_{k-1}) - (Cx_k, x_k)(Cx_k, x_k - x_{k-1})}{\|x_k - x_{k-1}\|^2\|Cx_k\|^2 - (Cx_k, x_k - x_{k-1})^2}, \\ b &= \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}(1 - a) + \frac{(Cx_k, x_{k-1})}{(Cx_k, Cx_k)}a. \end{aligned}$$

Отсюда, используя обозначения (7), найдем формулы для параметров α_{k+1} и τ_{k+1} :

$$\begin{aligned} \alpha_{k+1} &= \frac{(Cx_k, x_k - x_{k-1})(Cx_k, x_{k-1}) - (x_{k-1}, x_k - x_{k-1})(Cx_k, Cx_k)}{(Cx_k, Cx_k)(x_k - x_{k-1}, x_k - x_{k-1}) - (Cx_k, x_k - x_{k-1})^2}, \\ \tau_{k+1} &= \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)} + \frac{1 - \alpha_{k+1}}{\alpha_{k+1}} \frac{(Cx_k, x_{k-1})}{(Cx_k, Cx_k)}, \quad k = 1, 2, \dots \end{aligned} \quad (12)$$

Полученные ранее формулы (11) можно рассматривать как частный случай общих формул (12), полагая $\alpha_{k+1} = 1$, если знаменатель в выражении для α_{k+1} обращается в нуль.

Формулы (12) сложнее формул для параметров α_{k+1} и τ_{k+1} метода сопряженных направлений, полученных в § 3. Здесь требуется вычислять дополнительные скалярные произведения. Однако построенный здесь итерационный процесс (5), (12) менее подвержен влиянию ошибок округления, погрешности, допущенные на предыдущих шагах, затухают.

Связь между локально оптимальными трехслойными методами и методами сопряженных направлений устанавливает

Теорема 3. *Если для метода (5), (12) начальное приближение y_1 выбрано следующим образом:*

$$By_1 = (B - \tau_1 A)y_0 + \tau_1 f, \quad \tau_1 = \frac{(Dw_0, z_0)}{(Dw_0, w_0)}, \quad (13)$$

то в случае самосопряженного оператора $DB^{-1}A$ метод (5), (12) совпадает с методом сопряженных направлений.

Доказательство проведем по индукции. Из условия теоремы следует, что приближения y_1 , полученные здесь и в методе сопряженных направлений, совпадают. Пусть совпадают приближения y_1, y_2, \dots, y_k . Докажем, что y_{k+1} , построенное по формулам (5), (12), совпадает с приближением y_{k+1} метода сопряженных направлений.

Из сделанных предположений следует, что итерационные параметры $\tau_1, \tau_2, \dots, \tau_k$ и $\alpha_2, \alpha_3, \dots, \alpha_k$ обоих методов также совпадают. Если будет показано, что совпадают и параметры τ_{k+1} и α_{k+1} в этих методах, то утверждение теоремы 3 будет доказано.

Так как y_1, y_2, \dots, y_k — итерационные приближения метода сопряженных направлений, то в силу леммы выполнены условия

$$(Cx_j, x_i) = 0, \quad j = 0, 1, \dots, i-1, \quad i = 1, 2, \dots, k. \quad (14)$$

Подставляя (14) с $j = k-1$ и $i = k$ в (12) и используя самосопряженность оператора C , получим

$$\alpha_{k+1} = \frac{(x_{k-1}, x_k - x_{k-1})(Cx_k, Cx_k)}{(Cx_k, x_k)^2 - \|Cx_k\|^2 \|x_k - x_{k-1}\|^2}, \quad \tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \quad (15)$$

Итак, параметры τ_{k+1} локально оптимального метода и метода сопряженных направлений совпадают. Осталось показать, что совпадают параметры α_{k+1} .

Из (6) и (13) получим

$$x_k - x_{k-1} = (\alpha_k - 1)(x_{k-1} - x_{k-2}) - \alpha_k \tau_k Cx_{k-1}, \quad k = 2, 3, \dots, \quad (16)$$

$$x_1 - x_0 = -\tau_1 Cx_0.$$

Из (16) следует, что разность $x_k - x_{k-1}$ есть линейная комбинация $Cx_0, Cx_1, \dots, Cx_{k-1}$ и имеет следующий вид:

$$\begin{aligned} x_k - x_{k-1} &= -\alpha_k \tau_k Cx_{k-1} + \sum_{j=0}^{k-2} \beta_j Cx_j, \quad k \geq 2, \\ x_1 - x_0 &= -\tau_1 Cx_0, \end{aligned} \quad (17)$$

где коэффициенты β_j выражаются через $\tau_1, \tau_2, \dots, \tau_{k-1}$ и $\alpha_2, \alpha_3, \dots, \alpha_{k-1}$. Умножая левую и правую части (17) скалярно на x_{k-1} и $x_k - x_{k-1}$ и учитывая (14), получим

$$\begin{aligned} (x_{k-1}, x_k - x_{k-1}) &= -\alpha_k \tau_k (Cx_{k-1}, x_{k-1}), \\ \|x_k - x_{k-1}\|^2 &= \alpha_k \tau_k (Cx_{k-1}, x_{k-1}). \end{aligned} \quad (18)$$

Подставляя (18) в выражение (15) для α_{k+1} и учитывая формулу для параметра τ_{k+1} , получим

$$\begin{aligned} \alpha_{k+1} &= \frac{\alpha_k \tau_k (Cx_{k-1}, x_{k-1}) (Cx_k, Cx_k)}{\alpha_k \tau_k (Cx_{k-1}, x_{k-1}) (Cx_k, Cx_k) - (Cx_k, x_k)^2} = \\ &= \left(1 - \frac{\tau_{k+1} (Cx_k, x_k)}{\tau_k (Cx_{k-1}, x_{k-1})} \cdot \frac{1}{\alpha_k} \right)^{-1}, \end{aligned}$$

что совпадает с формулой для параметра α_{k+1} в методе сопряженных направлений. Теорема доказана.

Подставляя $x_k = D^{1/2} z_k$ и $C = D^{-1/2} (DB^{-1} A) D^{-1/2}$ в (12), получим следующий вид формул для параметров α_{k+1} и τ_{k+1} :

$$\begin{aligned} \alpha_{k+1} &= \frac{(Dw_k, z_k - z_{k-1}) (Dw_k, z_{k-1}) - (Dz_{k-1}, y_k - y_{k-1}) (Dw_k, w_k)}{(Dw_k, w_k) (D(z_k - z_{k-1}), y_k - y_{k-1}) - (Dw_k, z_k - z_{k-1})^2}, \\ \tau_{k+1} &= \frac{(Dw_k, z_k)}{(Dw_k, w_k)} + \frac{1 - \alpha_{k+1}}{\alpha_{k+1}} \frac{(Dw_k, z_{k-1})}{(Dw_k, w_k)}. \end{aligned} \quad (19)$$

Если ввести обозначения для скалярных произведений

$$\begin{aligned} a_k &= (Dw_k, z_k), \quad b_k = (Dw_k, z_{k-1}), \quad c_k = (Dz_k, y_k - y_{k-1}), \\ d_k &= (Dz_{k-1}, y_k - y_{k-1}), \quad e_k = (Dw_k, w_k), \end{aligned}$$

то формулы (19) перепишутся в виде

$$\begin{aligned} \alpha_{k+1} &= \frac{(a_k - b_k) b_k - d_k e_k}{(c_k - d_k) e_k - (a_k - b_k)^2}, \quad k = 1, 2, \dots, \alpha_1 = 1, \\ \tau_{k+1} &= \frac{a_k}{e_k} + \frac{1 - \alpha_{k+1}}{\alpha_{k+1}} \frac{b_k}{e_k}, \quad k = 0, 1, \dots \end{aligned}$$

Приведем выражения для a_k, b_k, c_k, d_k и e_k для конкретных выборов оператора D :

1) $D = A, A = A^*$.

$$\begin{aligned} a_k &= (w_k, r_k), \quad b_k = (w_k, r_{k-1}), \quad c_k = (r_k, y_k - y_{k-1}), \\ d_k &= (r_{k-1}, y_k - y_{k-1}), \quad e_k = (Aw_k, w_k). \end{aligned}$$

2) $D = A^* A$.

$$\begin{aligned} a_k &= (Aw_k, r_k), \quad b_k = (Aw_k, r_{k-1}), \quad c_k = (r_k, r_k - r_{k-1}), \\ d_k &= (r_{k-1}, r_k - r_{k-1}), \quad e_k = (Aw_k, Aw_k). \end{aligned}$$

3) $D = A^*B^{-1}A$, $B = B^*$.

$$a_k = (Aw_k, w_k), b_k = (Aw_k, w_{k-1}), c_k = (w_k, r_k - r_{k-1}), \\ d_k = (w_{k-1}, r_k - r_{k-1}), e_k = (B^{-1}Aw_k, Aw_k).$$

§ 5. Ускорение сходимости двухслойных методов в самосопряженном случае

1. Алгоритм процесса ускорения. В п. 5 § 1 было установлено, что в случае самосопряженного оператора $DB^{-1}A$ двухслойные градиентные методы обладают асимптотическим свойством. Оно проявляется в том, что для больших номеров итераций скорость сходимости метода существенно понижается по сравнению с началом итераций. Было также показано, что для больших номеров итераций погрешности, рассматриваемые через одну итерацию, становятся почти пропорциональными.

Используя это свойство, построим сейчас процесс ускорения сходимости двухслойных градиентных методов.

Для решения уравнения

$$Au = f \quad (1)$$

рассмотрим двухслойный градиентный итерационный метод

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (2)$$

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots \quad (3)$$

Пусть оператор $DB^{-1}A$ самосопряжен в H . Тогда итерационный метод обладает асимптотическим свойством, и при достаточно большом номере итераций k имеет место приближенное равенство

$$z_{k+2} \approx \rho^2 z_k, \quad z_k = y_k - u. \quad (4)$$

Рассмотрим сначала случай, когда в (4) выполняется строгое равенство, т. е. $z_{k+2} = \rho^2 z_k$. Построим по найденным уже приближениям y_k и y_{k+2} новое приближение по формуле

$$y = \alpha y_{k+2} + (1 - \alpha) y_k, \quad \alpha = 1/(1 - \rho^2). \quad (5)$$

Для погрешности $z = y - u$ получим

$$z = \alpha z_{k+2} + (1 - \alpha) z_k = (\alpha \rho^2 + 1 - \alpha) z_k = [1 - \alpha(1 - \rho^2)] z_k = 0.$$

Следовательно, в случае выполнения строгого равенства (4) линейная комбинация (5) приближений y_k и y_{k+2} дает точное решение уравнения (1).

Как было отмечено в § 1, выполнение точного равенства в (4) является исключительным случаем, имеющим место лишь для специального начального приближения. В общем случае имеет место приближенное равенство (4), а приведенные выше рас-

суждения позволяют надеяться, что некоторая линейная комбинация y_k и y_{k+2} будет давать хорошее приближение к решению исходной задачи.

Найдем наилучшую среди таких линейных комбинаций. Пусть y_k , y_{k+1} и y_{k+2} — итерационные приближения, полученные по формулам (2), (3). Будем искать новое приближение y по формуле

$$y = \alpha y_{k+2} + (1 - \alpha) y_k. \quad (6)$$

Поставим задачу выбрать параметр α так, чтобы норма погрешности $z = y - u$ в H_D была минимальной.

Сначала, используя схему (2), исключим из (6) y_{k+2} . Получим

$$By_{k+2} = (B - \tau_{k+2} A) y_{k+1} + \tau_{k+2} f$$

и после подстановки y_{k+2} в (6) будем иметь

$$By = \alpha (B - \tau_{k+2} A) y_{k+1} + (1 - \alpha) By_k + \alpha \tau_{k+2} f, \quad (7)$$

где y_{k+1} находится по двухслойной схеме

$$By_{k+1} = (B - \tau_{k+1} A) y_k + \tau_{k+1} f. \quad (8)$$

Если считать, что y_k — заданное начальное приближение, то схема (7), (8) совпадает со схемой метода сопряженных направлений, причем параметры τ_{k+1} и τ_{k+2} совпадают с такими же параметрами метода сопряженных направлений. Из теории этого метода следует (п. 1 § 4, формула (3)), что оптимальное значение параметра α определяется формулой

$$\alpha = \frac{1}{1 - \frac{\tau_{k+2}}{\tau_{k+1}} \frac{(Dw_{k+1}, z_{k+1})}{(Dw_k, z_k)}}. \quad (9)$$

Итак, поставленная задача о наилучшем выборе параметра α решена. Формулы (6), (9) определяют ускоряющую процедуру.

Заметим, что для нахождения y можно пользоваться не формулой (6), а вычислять y по следующей двухслойной схеме:

$$\begin{aligned} \bar{By}_{k+1} &= (B - \bar{\tau}_{k+1} A) y_k + \bar{\tau}_{k+1} f, \\ \bar{By} &= (B - \bar{\tau}_{k+2} A) \bar{y}_{k+1} + \bar{\tau}_{k+2} f, \end{aligned} \quad (10)$$

где $\bar{\tau}_{k+1}$ и $\bar{\tau}_{k+2}$ — корни уравнения

$$\tau^2 - \alpha(\tau_{k+1} + \tau_{k+2})\tau + \alpha\tau_{k+1}\tau_{k+2} = 0.$$

В качестве $\bar{\tau}_{k+1}$ следует взять минимальный корень.

Использование (10) вместо (6) позволяет не увеличивать объем запоминаемой промежуточной информации.

2. Оценка эффективности. Оценим теперь эффективность способа ускорения. Прежде чем вычислять норму погрешности $z = y - u$ в H_D , преобразуем выражение (9) для α .

Замена $z_k = D^{-1/2}x_k$ в (9) дает

$$\alpha = \left(1 - \frac{\tau_{k+2}}{\tau_{k+1}} \frac{(Cx_{k+1}, x_{k+1})}{(Cx_k, x_k)} \right)^{-1}, \quad C = D^{1/2} B^{-1} A D^{-1/2}. \quad (11)$$

Из (10) и (11) § 1 имеем

$$\|x_{k+1}\| = \rho_{k+1} \|x_k\|, \quad \rho_{k+1}^2 = 1 - \frac{(Cx_k, x_k)^2}{(Cx_k, Cx_k) \|x_k\|^2}. \quad (12)$$

Из формулы (9) § 1 получим

$$\tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \quad (13)$$

Используя (12) и (13), найдем

$$\frac{\tau_{k+2}}{\tau_{k+1}} \frac{(Cx_{k+1}, x_{k+1})}{(Cx_k, x_k)} = \frac{1 - \rho_{k+2}^2}{1 - \rho_{k+1}^2} \rho_{k+1}^2.$$

Подставляя это выражение в (11), получим

$$\alpha = \frac{1 - \rho_{k+1}^2}{1 - 2\rho_{k+1}^2 + \rho_{k+1}^2 \rho_{k+2}^2}. \quad (14)$$

Вычислим теперь норму погрешности $z = y - u$ в H_D . Из (6) получим

$$z = \alpha z_{k+2} + (1 - \alpha) z_k.$$

Отсюда для эквивалентной погрешности $z_k = D^{1/2} x_k$ и $x = D^{1/2} z$ будем иметь $x = \alpha x_{k+2} + (1 - \alpha) x_k$. Вычислим норму x в H . Получим

$$\|x\|^2 = \alpha^2 \|x_{k+2}\|^2 + 2\alpha(1 - \alpha)(x_{k+2}, x_k) + (1 - \alpha)^2 \|x_k\|^2.$$

Из доказанного в п. 5 § 1 равенства $(x_{k+2}, x_k) = \|x_{k+1}\|^2$ следует

$$\|x\|^2 = \alpha^2 \|x_{k+2}\|^2 + 2\alpha(1 - \alpha) \|x_{k+1}\|^2 + (1 - \alpha)^2 \|x_k\|^2.$$

Подставляя сюда выражение (14) для α и используя (12), получим

$$\|x\|^2 = \frac{\rho_{k+2}^2 - \rho_{k+1}^2}{\rho_{k+1}^2 (1 - 2\rho_{k+1}^2 + \rho_{k+1}^2 \rho_{k+2}^2)} \|x_{k+2}\|^2 < \|x_{k+2}\|^2. \quad (15)$$

Так как $\rho_{k+1} \leq \rho_{k+2} \leq \rho < 1$, то

$$1 - 2\rho_{k+1}^2 + \rho_{k+1}^2 \rho_{k+2}^2 \geq (1 - \rho^2)^2,$$

следовательно, для нормы x имеет место оценка

$$\|x\|^2 \leq \left(\frac{\rho_{k+2}^2}{\rho_{k+1}^2} - 1 \right) \frac{\|x_{k+2}\|^2}{(1 - \rho^2)^2}.$$

В силу асимптотического свойства для достаточно больших номеров k имеем $\rho_{k+1} \approx \rho_{k+2}$, поэтому эффект ускоряющей процедуры будет значителен.

Отметим, что хотя эффективное ускорение сходимости имеет место для больших номеров итераций k , этим способом можно пользоваться и для любого номера итераций. Рекомендуется время от времени прерывать процесс итераций, прово-

димый по двухслойной схеме (2), (3), и вычислять новое приближение по предлагаемому способу. Процесс итераций можно закончить на вычислении такого приближения, если для найденного y_{k+2} будет выполняться неравенство

$$\frac{\rho_{k+2}^2 - \rho_{k+1}^2}{\rho_{k+1}^2 (1 - 2\rho_{k+1}^2 + \rho_{k+1}^2 \rho_{k+2}^2)} \|z_{k+2}\|_D^2 \leq \varepsilon^2 \|z_0\|_D^2.$$

Действительно, в этом случае получим в силу (15) $\|y - u\|_D \leq \varepsilon \|y_0 - u\|_D$, т. е. требуемая точность ε будет достигнута.

3. Пример. Для иллюстрации эффективности предлагаемого способа ускорения сходимости двухслойных градиентных методов рассмотрим решение модельной задачи неявным методом скорейшего спуска. В качестве примера возьмем разностную задачу Дирихле для уравнения Лапласа на квадратной сетке $\omega = \{x_{ij} = (ih, jh), 0 \leq i \leq N, 0 \leq j \leq N, h = 1/N\}$ в единичном квадрате

$$\begin{aligned} \Lambda u &= \Lambda_1 u + \Lambda_2 u = 0, \quad x \in \omega, \quad u|_{\gamma} = 0, \\ \Lambda_\alpha u &= u_{x_\alpha x_\alpha}, \quad \alpha = 1, 2. \end{aligned} \quad (16)$$

Введем пространство H , состоящее из сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h^2.$$

Оператор A определим следующим образом: $A = A_1 + A_2$, $A_\alpha y = -\Lambda_\alpha v$, $y \in H$, где $v(x) = y(x)$ для $x \in \omega$ и $v|_{\gamma} = 0$.

Задачу (16) запишем в виде операторного уравнения

$$Au = f, \quad f = 0. \quad (17)$$

В качестве оператора B выберем следующий факторизованный оператор: $B = (E + \omega A_1)(E + \omega A_2)$, $\omega > 0$, где ω — итерационный параметр.

Так как операторы A_1 и A_2 самосопряжены и перестановочны в H , то операторы A и B являются самосопряженными в H . Кроме того, легко показать, что операторы A и B являются положительно определенными в H . Следовательно, для решения уравнения (17) можно использовать неявный метод скорейшего спуска

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad \tau_{k+1} = \frac{(\omega_k, r_k)}{(Aw_k, \omega_k)}, \quad k = 0, 1, \dots \quad (18)$$

В этом методе $D = A$ и $DB^{-1}A = AB^{-1}A$. Так как оператор $DB^{-1}A$ самосопряжен в H , то для рассматриваемого метода имеет место асимптотическое свойство. Из теории метода скорейшего спуска (см. п. 1 § 2) следует, что скорость сходимости метода в данном случае определяется отношением $\xi = \gamma_1/\gamma_2$, где γ_1 и

γ_2 — постоянные энергетической эквивалентности операторов A и B : $\gamma_1 B \leq A \leq \gamma_2 B$, $\gamma_1 > 0$.

Поэтому итерационный параметр ω выбирается из условия максимальности ξ . В § 2 гл. XI будет показано, что оптимальное значение ω определяется по формуле

$$\omega = \frac{1}{V\delta\Delta}, \quad \delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2},$$

при этом

$$\gamma_1 = \frac{2\delta}{(1 + V\eta)^2}, \quad \gamma_2 = \frac{(\Delta + \delta)V\eta}{(1 + V\eta)^2}, \quad \xi = \frac{2V\eta}{1 + \eta}, \quad \eta = \frac{\delta}{\Delta}.$$

Для рассматриваемого примера

$$\omega = \frac{h^2}{2 \sin \pi h}, \quad \gamma_1 = \frac{2}{h^2} \frac{\sin \pi h}{1 + \sin \pi h}, \quad \gamma_2 = \frac{2}{h^2} \frac{\sin \pi h}{1 + \sin \pi h}, \quad \xi = \sin \pi h.$$

Приведем результаты расчетов, когда начальное приближение y_0 выбиралось равным $y_0(x) = e^{(x_1 - x_2)}$ для $x \in \omega$, $y_0|_{\gamma} = 0$. Требуемая точность ε бралась равной 10^{-4} , $N = 40$.

В табл. 8 для некоторых номеров итераций k приводятся: $\|z_k\|_D / \|z_0\|_D$ — относительная точность k -й итерации, $\rho_k = \|z_k\|_D / \|z_{k-1}\|_D$ — величина, характеризующая уменьшение нормы погрешности при переходе от $(k-1)$ -й итерации к k -й итерации, $\gamma_1^{(k)}$ и $\gamma_2^{(k)}$ — приближения для γ_1 и γ_2 , которые находятся как корни квадратного уравнения

$$(1 - \tau_k \gamma)(1 - \tau_{k-1} \gamma) = \rho_k \rho_{k-1}, \quad k = 2, 3, \dots,$$

и итерационные параметры τ_k .

Таблица 8

k	$\ z_k\ _D / \ z_0\ _D$	ρ_k	$\gamma_1^{(k)}$	$\gamma_2^{(k)}$	τ_k
1	$3,6 \cdot 10^{-1}$	0,36203	—	—	$5,392 \cdot 10^{-3}$
2	$2,3 \cdot 10^{-1}$	0,63810	77,31858	236,1883	$7,809 \cdot 10^{-3}$
3	$1,8 \cdot 10^{-1}$	0,76998	40,59796	232,1435	$6,911 \cdot 10^{-3}$
4	$1,4 \cdot 10^{-1}$	0,81178	26,87824	233,4976	$8,644 \cdot 10^{-3}$
...
26	$3,9 \cdot 10^{-3}$	0,85175	18,27141	230,5962	$8,876 \cdot 10^{-3}$
27	$3,4 \cdot 10^{-3}$	0,85178	18,26983	230,6607	$7,338 \cdot 10^{-3}$
28	$2,9 \cdot 10^{-3}$	0,85183	18,27026	230,7191	$8,872 \cdot 10^{-3}$
29	$2,4 \cdot 10^{-3}$	0,85186	18,26895	230,7771	$7,335 \cdot 10^{-3}$
...
46	$1,6 \cdot 10^{-4}$	0,85226	18,26677	231,4121	$8,845 \cdot 10^{-3}$
47	$1,4 \cdot 10^{-4}$	0,85227	18,26632	231,4375	$7,318 \cdot 10^{-3}$
48	$1,2 \cdot 10^{-4}$	0,85229	18,26664	231,4612	$8,843 \cdot 10^{-3}$
49	$9,9 \cdot 10^{-5}$	0,85230	18,26623	231,4849	$7,317 \cdot 10^{-3}$

Заданная точность ϵ была достигнута после выполнения 49 итераций по схеме (18). Для $\epsilon = 10^{-4}$ теоретическое число итераций равно 59. Приведенные в таблице значения для ρ_k хорошо иллюстрируют асимптотическое свойство метода. Видно, что с ростом номера итераций происходит замедление скорости сходимости метода. Точность $4 \cdot 10^{-3}$ была достигнута за 26 итераций, а на увеличение точности еще в 40 раз потребовалось провести дополнительные 23 итерации. Величина ρ_k монотонно возрастает и для $k = 26$ имеем $\rho_{k+1} - \rho_k \approx 3 \cdot 10^{-5}$. Итерационные параметры τ_k и τ_{k+2} становятся почти равными.

Для сравнения приближенных значений $\gamma_1^{(k)}$ и $\gamma_2^{(k)}$ с точными приведем γ_1 и γ_2 :

$$\gamma_1 = 18,26556, \quad \gamma_2 = 232,8036.$$

После выполнения 49 итераций γ_1 было найдено с точностью 0,004%, а γ_2 — с точностью 0,6%.

Для ускорения сходимости метода была применена описанная в п. 1 процедура ускорения. По приближениям y_{26} и y_{28} , найденным по схеме (18), было построено по формулам (6), (9) новое приближение y . Заданная точность $\epsilon = 10^{-4}$ была достигнута. Применение построенного в этом параграфе способа ускорения сходимости двухслойных градиентных методов позволило уменьшить число требуемых итераций для рассматриваемого примера приблизительно в 1,8 раза.

ГЛАВА IX

ТРЕУГОЛЬНЫЕ ИТЕРАЦИОННЫЕ МЕТОДЫ

В главе изучаются неявные двухслойные итерационные методы, оператору B в которых соответствуют треугольные матрицы. В § 1 рассматривается метод Зейделя, для которого формулируются достаточные условия сходимости. В § 2 исследуется метод верхней релаксации. Здесь дан выбор итерационного параметра и получена оценка для спектрального радиуса оператора перехода. В § 3 рассмотрена общая итерационная схема треугольных методов, указан выбор итерационного параметра и доказана сходимость метода в норме H_A .

§ 1. Метод Зейделя

1. Итерационная схема метода. В предыдущих главах была изложена общая теория двухслойных и трехслойных итерационных методов, применяемых для нахождения приближенного решения линейного операторного уравнения первого рода

$$Au = f. \quad (1)$$

Эта теория указывает выбор итерационных параметров и дает оценку для числа итераций соответствующих методов, причем теория использует минимум информации общего характера относительно операторов итерационной схемы. Отказ от изучения конкретной структуры операторов итерационной схемы позволяет развивать теорию с единой точки зрения и конструировать неявные итерационные методы, оптимальные на классе операторов B .

В общей теории итерационных методов было показано, что эффективность метода существенным образом зависит от выбора оператора B . От выбора оператора B зависят как число итераций, которое нужно выполнить для достижения заданной точности ε , так и число арифметических действий, требующихся на реализацию одного итерационного шага. Каждая из этих величин в отдельности не может служить критерием эффективности итерационного метода. Поясним это утверждение. Пусть операторы A и B самосопряжены и положительно определены в H . Из теории итерационных методов следует, что если в качестве оператора D взять один из операторов \bar{A} , \bar{B} или $\bar{AB}^{-1}A$, то число итераций для рассмотренных в гл. VI—VIII итерационных методов (чебышевского, простой итерации, методов вариационного типа и т. д.) определяется отношением $\xi = \gamma_1/\gamma_2$, где γ_1 и γ_2 —постоянные энергетической эквивалентности операторов A и B : $\gamma_1 B \leq A \leq \gamma_2 B$.

Поэтому, если выбрать $B = A$, то получим максимально возможное значение $\xi = 1$, и итерационные методы дадут точное решение уравнения (1) за одну итерацию при любом начальном приближении. Следовательно, указанный выбор оператора B минимизирует число итераций. Однако для реализации этого единственного итерационного шага требуется обратить оператор B , т. е. оператор A . Очевидно, что при этом число арифметических действий будет максимальным.

С другой стороны, для явных схем с $B = E$ требуется минимальное число арифметических действий на одну итерацию, но при этом число итераций становится слишком большим.

Итак, возникает задача оптимального выбора оператора B из условия минимизации общего объема вычислительной работы, которая должна быть выполнена для получения решения с заданной точностью.

Естественно, что в такой общей постановке эта задача не может быть решена. В настоящее время развитие итерационных методов идет по пути конструирования легко обратимых операторов B , среди которых выбирают операторы с наилучшим отношением γ_1/γ_2 . К легко обратимым или экономичным операторам обычно относят такие операторы, обращение которых осуществляется за число арифметических действий, пропорциональное или почти пропорциональное числу неизвестных. Примерами таких операторов являются операторы, которым соответствуют диагональная, трехдиагональная, треугольные матрицы, а также их произведения. В качестве более сложного примера приведем разностный оператор Лапласа в прямоугольнике, который, как было показано в главе IV, можно обратить прямыми методами с малыми затратами арифметических операций.

Следует отметить, что использование в качестве оператора B диагональных операторов позволяет уменьшить число итераций по сравнению со случаем явной итерационной схемы. Однако асимптотический порядок зависимости числа итераций от числа неизвестных задачи остается таким же, как и для явной схемы. Более перспективным направлением является использование треугольных операторов B .

В настоящей главе и главе X будут рассмотрены универсальные двухслойные неявные итерационные методы, оператору B в которых соответствуют треугольные матрицы (треугольные методы) или произведение треугольных матриц (попеременно-треугольный метод).

Рассмотрение этих методов начнем с простейшего — с метода Зейделя.

Рассмотрим систему линейных алгебраических уравнений (1) или в развернутом виде

$$\sum_{j=1}^M a_{ij} u_j = f_i, \quad i = 1, 2, \dots, M.$$

В данном случае мы имеем дело с уравнением (1), заданным в конечномерном пространстве $H = H_M$.

Итерационный метод Зейделя в предположении, что диагональные элементы матрицы $A = (a_{ij})$ отличны от нуля ($a_{ii} \neq 0$), записывается в следующем виде:

$$\sum_{j=1}^t a_{ij} y_j^{(k+1)} + \sum_{j=i+1}^M a_{ij} y_j^{(k)} = f_i, \quad i = 1, 2, \dots, M, \quad (2)$$

где $y_j^{(k)}$ — j -я компонента итерационного приближения номера k . В качестве начального приближения выбирается произвольный вектор.

Определение $(k+1)$ -й итерации начинаем с $i = 1$:

$$a_{11} y_1^{(k+1)} = - \sum_{j=2}^M a_{1j} y_j^{(k)} + f_1.$$

Так как $a_{11} \neq 0$, то отсюда найдем $y_1^{(k+1)}$. Для $i = 2$ получим

$$a_{22} y_2^{(k+1)} = - a_{21} y_1^{(k+1)} - \sum_{j=3}^M a_{2j} y_j^{(k)} + f_2.$$

Пусть уже найдены $y_1^{(k+1)}, y_2^{(k+1)}, \dots, y_{i-1}^{(k+1)}$. Тогда $y_i^{(k+1)}$ находится из уравнения

$$a_{ii} y_i^{(k+1)} = - \sum_{j=1}^{i-1} a_{ij} y_j^{(k+1)} - \sum_{j=i+1}^M a_{ij} y_j^{(k)} + f_i. \quad (3)$$

Из формулы (3) видно, что алгоритм метода Зейделя является чрезвычайно простым. Найденное по формуле (3) значение $y_i^{(k+1)}$ размещается на месте $y_i^{(k)}$.

Оценим число арифметических действий, которое требуется для реализации одного итерационного шага. Если все a_{ij} не равны нулю, то вычисления по формуле (3) требуют $M-1$ операций сложения, $M-1$ операций умножения и одного деления. Поэтому реализация одного итерационного шага осуществляется за $2M^2-M$ арифметических действий.

Если в каждой строке матрицы A отлично от нуля лишь m элементов, а именно эта ситуация имеет место для сеточных эллиптических уравнений, то на реализацию итерационного шага потребуется $2mM-M$ действий, т. е. число действий, пропорциональное числу неизвестных M .

Запишем теперь итерационный метод Зейделя (2) в матричной форме. Для этого представим матрицу A в виде суммы диагональной, нижней треугольной и верхней треугольной матриц

$$A = \mathcal{D} + L + U, \quad (4)$$

где

$$L = \begin{vmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \cdot & \cdot & \cdot & \ddots & \\ \cdot & \cdot & \cdot & \cdots & \\ a_{M1} & a_{M2} & \cdots & a_{MM-1} & 0 \end{vmatrix}, \quad U = \begin{vmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1M} \\ 0 & a_{23} & \cdots & a_{2M} & \\ \cdot & \cdot & \ddots & \cdot & \\ 0 & a_{M-1M} & & & \\ 0 & 0 & & & \end{vmatrix},$$

$$\mathcal{D} = \begin{vmatrix} a_{11} & & & & \\ & a_{22} & & & \\ & & \ddots & & \\ 0 & & & & a_{MM} \end{vmatrix}.$$

Обозначим через $y_k = (y_1^{(k)}, y_2^{(k)}, \dots, y_M^{(k)})$ — вектор k -го итерационного приближения.

Пользуясь этими обозначениями, запишем метод Зейделя в виде

$$(\mathcal{D} + L)y_{k+1} + Uy_k = f, \quad k = 0, 1, \dots$$

Приведем эту итерационную схему к каноническому виду двухслойных схем

$$(\mathcal{D} + L)(y_{k+1} - y_k) + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H. \quad (5)$$

Сравнив (5) с каноническим видом

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H,$$

находим, что $B = \mathcal{D} + L$, $\tau_k \equiv 1$. Схема (5) неявная, оператор B является треугольной матрицей и, следовательно, несамосопряжен в H .

Мы рассмотрели так называемый точечный или скалярный метод Зейделя, предполагая, что элементы a_{ij} матрицы A есть числа. Аналогично строится блочный или векторный метод Зейделя для случая, когда a_{ii} есть квадратные матрицы, вообще говоря, различной размерности, а a_{ij} для $i \neq j$ — прямоугольные матрицы. В этом случае y_i и f_i есть векторы, размерность которых соответствует размерности матрицы a_{ii} .

В предположении невырожденности матриц a_{ii} блочный метод Зейделя записывается в виде (2) или в каноническом виде (5).

2. Примеры применения метода. Рассмотрим применение метода Зейделя для нахождения приближенного решения разностной задачи Дирихле для уравнения Пуассона и эллиптического уравнения с переменными коэффициентами в прямоугольнике.

Пусть на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$, введенной в прямоуголь-

нике $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$, требуется найти решение разностной задачи Дирихле для уравнения Пуассона

$$\Lambda y = \sum_{\alpha=1}^2 y_{x_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \quad (6)$$

где $\gamma = \{x_{ij} \in \Gamma\}$ — граница сетки $\bar{\omega}$.

В данном примере неизвестными являются $y(i, j) = y(x_{ij})$ во внутренних узлах сетки. Если упорядочить неизвестные естественным образом по строкам сетки ω , начиная с нижней строки, то разностная схема (6) может быть записана в виде следующей системы алгебраических уравнений:

$$-\frac{1}{h_1^2} y(i-1, j) - \frac{1}{h_2^2} y(i, j-1) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y(i, j) - \\ - \frac{1}{h_1^2} y(i+1, j) - \frac{1}{h_2^2} y(i, j+1) = \varphi(i, j)$$

для $i=1, 2, \dots, N_1-1, j=1, 2, \dots, N_2-1$ и $y(x) = g(x)$ для $x \in \gamma$. При этом неизвестные $y(i-1, j)$ и $y(i, j-1)$ предшествуют $y(i, j)$, а $y(i+1, j)$ и $y(i, j+1)$ следуют за $y(i, j)$. Так как в каждом уравнении связаны не более пяти неизвестных, то в каждой строке матрицы A отличны от нуля не более пяти элементов.

Для рассматриваемой системы точечный метод Зейделя будет иметь следующий вид:

$$\left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{k+1}(i, j) = \frac{1}{h_1^2} y_{k+1}(i-1, j) + \frac{1}{h_2^2} y_{k+1}(i, j-1) + \\ + \frac{1}{h_1^2} y_k(i+1, j) + \frac{1}{h_2^2} y_k(i, j+1) + \varphi(i, j), \\ 1 \leqslant i \leqslant N_1-1, \quad 1 \leqslant j \leqslant N_2-1,$$

причем $y_k(x) = g(x)$, $x \in \gamma$ для любого $k \geqslant 0$.

Вычисления начинаются с точки $i=1, j=1$ и продолжаются либо по строкам, либо по столбцам сетки ω . Для нахождения $y_{k+1}(i, j)$ требуется 7 арифметических операций, а всего на реализацию итерационного шага потребуется $7M$ операций, где $M = (N_1-1)(N_2-1)$ — число неизвестных в задаче.

Для рассматриваемого примера оператор B в конечномерном пространстве H сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2$, $u, v \in H$ определяется следующим образом:

$$By = (\mathcal{D} + L)y = \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) \dot{y}(i, j) - \frac{1}{h_1^2} \dot{y}(i-1, j) - \frac{1}{h_2^2} \dot{y}(i, j-1) = \\ = \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right) \dot{y} + \sum_{\alpha=1}^2 \frac{1}{h_\alpha} \dot{y}_{x_\alpha},$$

где $y \in H$, $\dot{y} \in \dot{H}$ и $\ddot{y}(x) = y(x)$ для $x \in \omega$. Здесь \dot{H} — множество се-точных функций, заданных на ω и обращающихся в нуль на γ .

Рассмотрим теперь *блочный метод Зейделя*. Если обозначить через $\mathbf{Y}_j = (y(1, j), y(2, j), \dots, y(N_1 - 1, j))$ вектор, состоящий из неизвестных на j строке сетки, то, как было показано в § 1 гл. I, разностная задача (6) может быть записана в виде трехточечной системы векторных уравнений:

$$\begin{aligned} -\mathbf{Y}_{j-1} + C\mathbf{Y}_j - \mathbf{Y}_{j+1} &= \mathbf{F}_j, \quad j = 1, 2, \dots, N_2 - 1, \\ \mathbf{Y}_0 &= \mathbf{F}_0, \quad \mathbf{Y}_{N_2} = \mathbf{F}_{N_2}, \end{aligned} \quad (7)$$

где C — квадратная трехдиагональная матрица размерности $(N_1 - 1) \times (N_1 - 1)$, определяемая следующим образом:

$$(C\mathbf{Y}_j)_i = (2y - h_2^2 y_{\bar{x}_1 x_1})_{ij}, \quad y_{0j} = y_{N_1 j} = 0.$$

Правые части \mathbf{F}_j определяются формулами

$$\begin{aligned} \mathbf{F}_j &= (h_2^2 \varphi(1, j) + \frac{h_2^2}{h_1^2} g(0, j), h_2^2 \varphi(2, j), \dots, h_2^2 \varphi(N_1 - 2, j), \\ &\quad h_2^2 \varphi(N_1 - 1, j) + \frac{h_2^2}{h_1^2} g(N_1, j)) \text{ для } j = 1, 2, \dots, N_2 - 1, \\ \mathbf{F}_j &= (g(1, j), g(2, j), \dots, g(N_1 - 1, j)) \text{ для } j = 0, N_2. \end{aligned}$$

Блочный метод Зейделя для системы (7) имеет вид

$$\begin{aligned} C\mathbf{Y}_j^{(k+1)} &= \mathbf{Y}_{j-1}^{(k+1)} + \mathbf{Y}_{j+1}^{(k)} + \mathbf{F}_j, \quad j = 1, 2, \dots, N_2 - 1, \\ \mathbf{Y}_0^{(k)} &= \mathbf{F}_0, \quad \mathbf{Y}_{N_2}^{(k)} = \mathbf{F}_{N_2}, \quad k = 0, 1, \dots, \end{aligned} \quad (8)$$

и для нахождения $\mathbf{Y}_j^{(k+1)}$ требуется обращать трехдиагональную матрицу C .

Если расписать схему (8) по точкам сетки, то получим следующие формулы:

$$\begin{aligned} -\frac{1}{h_1^2} y_{k+1}(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{k+1}(i, j) - \frac{1}{h_2^2} y_{k+1}(i+1, j) &= \\ = \frac{1}{h_2^2} y_{k+1}(i, j-1) + \frac{1}{h_2^2} y_k(i, j+1) + \varphi(i, j), \quad (9) \\ 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \end{aligned}$$

причем $y_k(x) = g(x)$, $x \in \gamma$ для любого $k \geq 0$. Для нахождения y_{k+1} на j -й строке нужно решить трехточечную краевую задачу (9) с известной правой частью, например методом прогонки, и полученное решение разместить на месте y_k в j -й строке.

Блочному методу Зейделя соответствует следующий оператор B :

$$By = \frac{1}{h_2^2} \dot{y} + \frac{1}{h_2} \dot{y}_{\bar{x}_2} + \dot{y}_{\bar{x}_1 x_1}, \quad y \in H, \quad \dot{y} \in \dot{H}.$$

Пусть теперь на сетке $\bar{\omega}$ требуется найти решение разностной задачи Дирихле для эллиптического уравнения с переменными коэффициентами

$$\begin{aligned} \Lambda y &= \sum_{\alpha=1}^2 \left(a_\alpha(x) y_{x_\alpha} \right)_{x_\alpha} - d(x) y = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \quad (10)$$

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad x \in \bar{\omega}, \quad \alpha = 1, 2, \quad 0 \leq d_1 \leq d(x) \leq d_2, \quad x \in \omega.$$

Для рассматриваемой задачи точечный метод Зейделя при упорядочении неизвестных по строкам сетки будет иметь следующий вид:

$$\begin{aligned} \left(\frac{a_1(i+1, j) + a_1(i, j)}{h_1^2} + \frac{a_2(i, j+1) + a_2(i, j)}{h_2^2} + d(i, j) \right) y_{k+1}(i, j) &= \\ = \frac{a_1(i, j)}{h_1^2} y_{k+1}(i-1, j) + \frac{a_2(i, j)}{h_2^2} y_{k+1}(i, j-1) &+ \\ + \frac{a_1(i+1, j)}{h_1^2} y_k(i+1, j) + \frac{a_2(i, j+1)}{h_2^2} y_k(i, j+1) &+ \varphi(i, j) \end{aligned}$$

для $i = 1, 2, \dots, N_1 - 1$ и $j = 1, 2, \dots, N_2 - 1$, причем $y_k(x) = g(x)$ при $x \in \gamma$ для любого $k \geq 0$.

Оператор B в канонической форме итерационной схемы для данного примера определяется следующим образом:

$$\begin{aligned} By(x_{ij}) &= \left(\frac{a_1(i+1, j)}{h_1^2} + \frac{a_2(i, j+1)}{h_2^2} + d(i, j) \right) \dot{y}(i, j) + \\ &+ \frac{a_1(i, j)}{h_1^2} \dot{y}_{x_1} + \frac{a_2(i, j)}{h_2^2} \dot{y}_{x_2}, \quad y \in H, \quad \dot{y} \in \dot{H}, \end{aligned}$$

где пространство H и множество \dot{H} определены выше.

3. Достаточные условия сходимости. Сформулируем теперь некоторые достаточные условия сходимости метода Зейделя. Нам потребуется следующая

Теорема 1. Пусть в уравнении (1) оператор A самосопряжен и положительно определен в H . Тогда двухслойный итерационный процесс

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad \tau > 0, \quad (11)$$

сходится в H_A , если оператор $B - 0,5\tau A$ положительно определен в H , т. е. выполнено условие

$$B > \frac{\tau}{2} A. \quad (12)$$

Действительно, из (11) получим для погрешности $z_k = y_k - u$ следующую задачу:

$$B \frac{z_{k+1} - z_k}{\tau} + Az_k = 0, \quad k = 0, 1, \dots, z_0 = y_0 - u. \quad (13)$$

Установим для z_k основное энергетическое тождество. Подставим в (13) z_k в виде $z_k = \frac{1}{2}(z_{k+1} + z_k) - \frac{\tau}{2}\left(\frac{z_{k+1} - z_k}{\tau}\right)$ и получим

$$\left(B - \frac{\tau}{2}A\right) \frac{z_{k+1} - z_k}{\tau} + \frac{1}{2}A(z_{k+1} + z_k) = 0.$$

Умножим левую и правую части этого равенства скалярно на $2(z_{k+1} - z_k)$ и учтем, что для самосопряженного оператора A имеет место равенство $(A(z_{k+1} + z_k), z_{k+1} - z_k) = (Az_{k+1}, z_{k+1}) - -(Az_k, z_k)$. В результате получим основное энергетическое тождество

$$2\tau \left(\left(B - \frac{\tau}{2}A \right) \frac{z_{k+1} - z_k}{\tau}, \frac{z_{k+1} - z_k}{\tau} \right) + \|z_{k+1}\|_A^2 - \|z_k\|_A^2 = 0.$$

Отсюда и из неравенств $B - 0,5\tau A > 0$, $\tau > 0$ вытекает, что $\|z_{k+1}\|_A^2 \leq \|z_k\|_A^2$, т. е. последовательность $\{\|z_k\|_A^2\}$ не возрастает, ограничена снизу нулем и является сходящейся. Тогда из энергетического тождества следует, что

$$\lim_{k \rightarrow \infty} \left(\left(B - \frac{\tau}{2}A \right) \frac{z_{k+1} - z_k}{\tau}, \frac{z_{k+1} - z_k}{\tau} \right) = 0. \quad (14)$$

Далее, из неравенства $B - 0,5\tau A > 0$ вытекает, что $\|z_{k+1} - z_k\| \rightarrow 0$ при $k \rightarrow \infty$. Замечая из (13), что $A^{1/2}z_k = -A^{-1/2}B(z_{k+1} - z_k)/\tau$, получим $\|z_k\|_A^2 \leq \|A^{-1}\| \|B\|^2 \|z_{k+1} - z_k\|^2 / \tau^2 \rightarrow 0$ при $k \rightarrow \infty$.

Сформулируем достаточное условие сходимости метода Зейделя.

Теорема 2. Если оператор A самосопряжен и положительно определен в H , то метод Зейделя (4), (5) сходится в H_A .

Действительно, из (5) и теоремы 1 следует, что достаточно проверить неравенство $\mathcal{D} + L - 0,5A > 0$. Так как $A = A^*$, то в (4) имеем $U = L^*$ и

$$((\mathcal{D} + L - 0,5A)x, x) = 0,5((\mathcal{D} + L - U)x, x) = 0,5(\mathcal{D}x, x).$$

Так как A положительно определенный оператор, то для точечного метода Зейделя имеем $a_{ii} > 0$, $1 \leq i \leq M$, а для блочного метода Зейделя матрицы $a_{ii} = a_{ii}^* > 0$. Следовательно, $\mathcal{D} = \mathcal{D}^* > 0$. Таким образом, $\mathcal{D} + L - 0,5A > 0$.

Приведем без доказательства еще одно условие сходимости метода Зейделя.

Теорема 3. Если оператор A самосопряжен и не вырожден, а все $a_{ii} > 0$, то метод Зейделя сходится при любом начальном приближении тогда и только тогда, когда A — положительно определенный оператор.

Чтобы оценить скорость сходимости метода Зейделя, используют различного рода предположения.

Например, если выполнено условие

$$\sum_{j \neq i} |a_{ij}| \leq q |a_{ii}|, \quad i = 1, 2, \dots, M, \quad q < 1, \quad (15)$$

то метод Зейделя сходится со скоростью геометрической прогрессии со знаменателем q , и для погрешности z_n имеет место оценка $\|z_n\| \leq q^n \|z_0\|$, где $\|z_n\| = \max_{1 \leq i \leq M} |y_i^{(n)} - u_i|$.

Действительно, из (3) получим для погрешности $z_i^{(k)} = y_i^{(k)} - u_i$ однородное уравнение

$$a_{ii} z_i^{(k+1)} = - \sum_{j=1}^{i-1} a_{ij} z_j^{(k+1)} - \sum_{j=i+1}^M a_{ij} z_j^{(k)}.$$

Отсюда найдем

$$\begin{aligned} |z_i^{(k+1)}| &\leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| |z_j^{(k+1)}| + \sum_{j=i+1}^M \left| \frac{a_{ij}}{a_{ii}} \right| |z_j^{(k)}| \leq \\ &\leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \|z_{k+1}\| + \sum_{j=i+1}^M \left| \frac{a_{ij}}{a_{ii}} \right| \|z_k\|. \end{aligned} \quad (16)$$

Из (15) получим

$$\sum_{j=i+1}^M \left| \frac{a_{ij}}{a_{ii}} \right| \leq q - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \leq q \left(1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right).$$

Подставляя эту оценку в (16), получим следующее неравенство:

$$|z_i^{(k+1)}| \leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \|z_{k+1}\| + q \left(1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right) \|z_k\|. \quad (17)$$

Пусть $\max_i |z_i^{(k+1)}|$ достигается при некотором $i = i_0$, так что $\|z_{k+1}\| = |z_{i_0}^{(k+1)}|$. Из (17) при $i = i_0$ получим

$$\left(1 - \sum_{j=1}^{i_0-1} \left| \frac{a_{i_0 j}}{a_{i_0 i_0}} \right| \right) \|z_{k+1}\| \leq q \left(1 - \sum_{j=1}^{i_0-1} \left| \frac{a_{i_0 j}}{a_{i_0 i_0}} \right| \right) \|z_k\|;$$

отсюда следует оценка $\|z_{k+1}\| \leq q \|z_k\| \leq \dots \leq q^{k+1} \|z_0\|$. Утверждение доказано.

Условие (15) означает, что A является матрицей с диагональным преобладанием. Для рассмотренных в п. 2 примеров применения метода Зейделя условие (15) не выполняется ($q = 1$). В этих примерах оператор A самосопряжен и положительно определен в H . Поэтому в силу теоремы 2 можно лишь утверждать, что метод сходится в H_A . Оценка скорости сходимости в H_A будет дана ниже после рассмотрения общей схемы треугольных итерационных методов.

§ 2. Метод верхней релаксации

1. Итерационная схема. Достаточные условия сходимости.

Для ускорения сходимости метода Зейделя его модифицируют, вводя в итерационную схему итерационный параметр ω , так что

$$(\mathcal{D} + \omega L) \frac{y_{k+1} - y_k}{\omega} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (1)$$

где, как и раньше, матрица A представлена в виде суммы

$$A = \mathcal{D} + L + U. \quad (2)$$

Метод Зейделя соответствует значению $\omega = 1$.

Сравнивая (1) с каноническим видом двухслойных итерационных схем, находим, что

$$B = \mathcal{D} + \omega L, \quad \tau_k \equiv \omega.$$

Как и для метода Зейделя, для рассматриваемого метода оператору B соответствует нижняя треугольная матрица, так что введение параметра ω не выводит нас из класса треугольных итерационных методов. Новым является вопрос о выборе параметра ω .

Если расписать итерационную схему (1) по компонентам вектора y_{k+1} , то получим следующие формулы:

$$a_{ii}y_i^{(k+1)} = (1 - \omega)a_{ii}y_i^{(k)} - \omega \sum_{j=1}^{i-1} a_{ij}y_j^{(k+1)} - \omega \sum_{j=i+1}^M a_{ij}y_j^{(k)} + \omega f_i \quad (3)$$

для $i = 1, 2, \dots, M$ (найденное $y_i^{(k+1)}$ размещается на месте $y_i^{(k)}$). Реализация одного итерационного шага осуществляется примерно с такими же затратами арифметических действий, как и в методе Зейделя.

Итерационный метод (1) при $\omega > 1$ называется *методом верхней релаксации*, при $\omega = 1$ — *полней релаксации* и при $\omega < 1$ — *нижней релаксации*.

В § 1 было доказано, что метод Зейделя сходится в H_A для случая самосопряженного и положительно определенного оператора A . Для сходимости метода релаксации, помимо этих требований, требуется дополнительное условие на итерационный параметр ω . Сформулируем достаточные условия сходимости метода релаксации.

Теорема 4. *Если оператор A самосопряжен и положительно определен в H , а параметр ω удовлетворяет условию $0 < \omega < 2$, то метод релаксации (1) сходится в H_A .*

Действительно, из теоремы 1 следует, что достаточно проверить выполнение неравенства $\mathcal{D} + \omega L > 0, 5\omega A$ при $\omega > 0$. Так как $A = A^* > 0$, то оператор \mathcal{D} самосопряжен и положительно

определен в H и $U = L^*$. Поэтому, используя равенство (14) § 1, получим

$$((\mathcal{D} + \omega L)x, x) = (1 - 0,5\omega)(\mathcal{D}x, x) + 0,5\omega((\mathcal{D} + 2L)x, x) = \\ = (1 - 0,5\omega)(\mathcal{D}x, x) + 0,5\omega(Ax, x).$$

При $\omega < 2$ отсюда следует утверждение теоремы.

Замечание. Теорема 4 справедлива как для точечного метода релаксации, когда в (3) a_{ij} — числа, так и для блочного или векторного метода релаксации, когда в (3) a_{ij} — матрицы соответствующей размерности.

2. Постановка задачи о выборе итерационного параметра.

Теорема 4 дает достаточные условия сходимости метода релаксации, оставляя открытым вопрос об оптимальном выборе параметра ω . Особенность рассматриваемого итерационного процесса (1) состоит в том, что итерационный параметр ω входит в оператор $B = \mathcal{D} + \omega L$, который является несамосопряженным в H оператором. С несамосопряженным случаем мы уже имели дело в § 4 гл. VI, где был рассмотрен метод простой итерации, итерационный параметр для которого выбирался из различных условий, например из условия минимума нормы оператора перехода от итерации к итерации. Здесь же необходимо учесть указанную выше особенность итерационной схемы. Выбор параметра ω из условия минимума нормы в H_A оператора перехода от итерации к итерации будет сделан в § 3 этой главы, где будет рассмотрена общая схема треугольных итерационных методов. В данном пункте параметр ω для метода релаксации будет выбираться из условия минимума спектрального радиуса оператора перехода от итерации к итерации.

Напомним определение спектрального радиуса оператора

$$\rho(S) = \lim_{n \rightarrow \infty} \sqrt[n]{\|S^n\|} = \max_k |\lambda_k|, \quad (4)$$

где λ_k — собственные значения оператора S . Спектральный радиус обладает следующими свойствами:

$$\rho(S^n) = \rho^n(S), \quad \rho(S) \leq \|S\| \quad (5)$$

и $\rho(S) = \|S\|$, если S — самосопряженный в H оператор. Из (5) для произвольного оператора S получим $\rho^n(S) = \rho(S^n) \leq \|S^n\|$. С другой стороны, из (4) при достаточно большом n будем иметь $\rho^n(S) \approx \|S^n\|$.

Переходим теперь к постановке задачи об оптимальном выборе параметра ω для итерационной схемы (1). Получим сначала задачу для погрешности $z_k = y_k - u$. Из (1) найдем

$$(\mathcal{D} + \omega L) \frac{z_{k+1} - z_k}{\omega} + Az_k = 0, \quad k = 0, 1, \dots, z_0 = y_0 - u$$

или

$$z_{k+1} = Sz_k, \quad k = 0, 1, \dots, S = E - \omega(\mathcal{D} + \omega L)^{-1}A. \quad (6)$$

Используя (6), выразим z_n через z_0 :

$$z_n = S^n z_0, \quad \|z_n\| \leq \|S^n\| \|z_0\|. \quad (7)$$

Оператор S является несамосопряженным в H оператором, зависящим от параметра ω . Задачу об оптимальном выборе параметра ω сформулируем следующим образом: найти ω из условия минимума спектрального радиуса оператора S .

Следует отметить, что мы не минимизируем норму разрешающего оператора S^n , как это следовало бы делать в силу оценки (7), а минимизируем спектральный радиус $\rho(S)$ оператора перехода S , для которого имеет место оценка $\rho^n(S) \leq \|S^n\|$. Однако в силу приближенного равенства $\rho^n(S) \approx \|S^n\|$ можно ожидать, что для достаточно большого n указанный способ выбора ω окажется удачным.

Решение сформулированной выше задачи является сложной проблемой, однако при некоторых дополнительных предположениях относительно оператора A эта задача может быть успешно решена.

Предположение 1. Оператор A самосопряжен и положительно определен в H ($U = L^*$, $\mathcal{D} = \mathcal{D}^* > 0$).

Предположение 2. Оператор A такой, что для любого комплексного $z \neq 0$ собственные значения μ обобщенной задачи на собственные значения $(zL + \frac{1}{z}U)x - \mu \mathcal{D}x = 0$ не зависят от z .

Используя эти предположения, докажем следующее утверждение, которое нам понадобится в дальнейшем.

Лемма 1. Если оператор A удовлетворяет предположениям 1 и 2, то все собственные значения задачи

$$Ax - \lambda \mathcal{D}x = 0 \quad (8)$$

действительны, положительны и, если λ — собственное значение, то $2 - \lambda$ — тоже собственное значение.

В самом деле, положительность и вещественность собственных значений λ следует из самосопряженности и положительной определенности оператора A . Далее, пусть λ — собственное значение задачи (8), т. е.

$$Ax - \lambda \mathcal{D}x = (L + U)x - (\lambda - 1)\mathcal{D}x = 0, \quad x \neq 0.$$

В силу предположения 2 будет иметь место равенство

$$(-L - U)y - (\lambda - 1)\mathcal{D}y = 0 \text{ или } Ay - (2 - \lambda)\mathcal{D}y = 0.$$

Отсюда следует утверждение леммы.

Переходим теперь к решению задачи об оптимальном выборе параметра ω . Для этого необходимо оценить спектральный радиус оператора перехода $S = E - \omega(\mathcal{D} + \omega L)^{-1}A$, т. е. оценить собственные значения μ оператора S :

$$Sx - \mu x = 0. \quad (9)$$

Будем считать, что предположения 1 и 2 выполнены. Следующая лемма устанавливает соотношение между собственными значениями μ задачи (9) и собственными значениями λ задачи (8).

Лемма 2. Для $\omega \neq 1$ собственные значения задач (8) и (9) связаны соотношением

$$(\mu + \omega - 1)^2 = \omega^2 \mu (1 - \lambda)^2. \quad (10)$$

Действительно, пусть μ и λ — собственные значения задач (9) и (8). Из определения оператора S и разложения $A = \mathcal{D} + L + U$ следует, что (9) может быть записано в виде

$$\frac{1 - \mu - \omega}{\omega} \mathcal{D}x - (\mu L + U)x = 0, \quad x \neq 0. \quad (11)$$

Покажем сначала, что при $\omega \neq 1$ все μ отличны от нуля. В самом деле, предположим, что $\mu = 0$. Тогда (11) принимает вид

$$\frac{1 - \omega}{\omega} \mathcal{D}x - Ux = 0.$$

Так как U — верхняя треугольная матрица, а \mathcal{D} — диагональная (блочно-диагональная) матрица, которая положительно определена в силу предположения 1, то последнее равенство может иметь место для $x \neq 0$ тогда и только тогда, когда $\omega = 1$. Следовательно, мы пришли к противоречию, предположив, что при $\omega \neq 1$ имеем $\mu = 0$.

Разделив левую и правую части (11) на $\sqrt{\mu}$, получим

$$\frac{1 - \mu - \omega}{\omega \sqrt{\mu}} \mathcal{D}x - \left(\sqrt{\mu} L + \frac{1}{\sqrt{\mu}} U \right) x = 0.$$

Отсюда в силу предположения 2 находим

$$\frac{1 - \mu - \omega}{\omega \sqrt{\mu}} \mathcal{D}y - (L + U)y = 0$$

или

$$Ay - \left(1 + \frac{1 - \mu - \omega}{\omega \sqrt{\mu}} \right) \mathcal{D}y = 0.$$

Сравнивая это равенство с (8), получим соотношение

$$\frac{\mu + \omega - 1}{\omega \sqrt{\mu}} = 1 - \lambda.$$

Этим доказательство леммы 2 заканчивается.

Замечание. При доказательстве леммы 2 самосопряженность оператора A не использовалась. Соотношение (10) имеет место и для случая любого несамосопряженного оператора A в предположении невырожденности оператора \mathcal{D} .

Из леммы 1 следует, что собственные значения λ расположены на действительной оси симметрично относительно точки $\lambda = 1$, причем $\lambda \in [\lambda_{\min}, 2 - \lambda_{\min}]$, $\lambda_{\min} > 0$. Поэтому из леммы 2 получим, что при $\omega \neq 1$ каждому $\lambda_i = 1$ соответствует $\mu_i = 1 - \omega$, каждой паре λ_i и $2 - \lambda_i$ соответствует пара ненулевых μ_i , получаемых решением уравнения (10) с $\lambda = \lambda_i$. Следовательно, все μ_i могут быть найдены как корни квадратного уравнения (10), в котором в качестве λ берутся все λ_i , расположенные на отрезке $[\lambda_{\min}, 1]$.

3. Оценка спектрального радиуса. Найдем теперь оптимальное значение параметра ω и оценим спектральный радиус оператора S . Для этого исследуем уравнение (10):

$$\mu^2 + [2(\omega - 1) - \omega^2(1 - \lambda)^2]\mu + (\omega - 1)^2 = 0, \quad (12)$$

где $\lambda_{\min} \leq \lambda \leq 1$ и $0 < \omega < 2$.

Решая уравнение (12), найдем два корня

$$\begin{aligned}\mu_1(\lambda, \omega) &= \left(\frac{\omega(1-\lambda) + \sqrt{\omega^2(1-\lambda)^2 - 4(\omega-1)}}{2} \right)^2, \\ \mu_2(\lambda, \omega) &= \left(\frac{\omega(1-\lambda) - \sqrt{\omega^2(1-\lambda)^2 - 4(\omega-1)}}{2} \right)^2.\end{aligned}\quad (13)$$

Исследование дискриминанта уравнения (12) дает, что при $\omega > \omega_0 > 1$, где

$$\omega_0 = \frac{2}{1 + \sqrt{\lambda_{\min}(2 - \lambda_{\min})}} \in (1, 2), \quad (14)$$

корни μ_1 и μ_2 для любого $\lambda \in [\lambda_{\min}, 1]$ являются комплексными, причем $|\mu_1| = |\mu_2| = \omega - 1$. Поэтому спектральный радиус оператора S при $\omega > \omega_0$ равен $\rho(S) = \omega - 1$ и возрастает по ω . Если $\omega = \omega_0$, то

$$\mu_1(\lambda_{\min}, \omega_0) = \mu_2(\lambda_{\min}, \omega_0) = \omega_0 - 1,$$

и для $\lambda_{\min} < \lambda \leq 1$ корни μ_1 и μ_2 снова будут комплексными и $|\mu_1| = |\mu_2| = \omega_0 - 1$. Следовательно, в области $\omega \geq \omega_0$ оптимальным является значение $\omega = \omega_0$, которому соответствует $\rho(S) = \omega_0 - 1$.

Пусть теперь $1 < \omega < \omega_0$. Исследуем поведение корней μ_1 и μ_2 , определяемых формулой (13), как функций переменной λ при фиксированном ω .

Если λ принадлежит отрезку $[\lambda_{\min}, \lambda_0]$,

$$\lambda_{\min} \leq \lambda \leq \lambda_0 = 1 - 2 \frac{\sqrt{\omega-1}}{\omega} < 1,$$

то дискриминант $\omega^2(1-\lambda)^2 - 4(\omega-1)^2$ неотрицателен и, следовательно, корни μ_1 и μ_2 действительны, причем максимальным является корень μ_1 .

Покажем, что $\mu_1(\lambda, \omega)$ есть убывающая функция λ на отрезке $[\lambda_{\min}, \lambda_0]$. Действительно, дифференцируя (12) по λ и учитывая (13), получим

$$\frac{\partial \mu_1}{\partial \lambda} = - \frac{2\omega\mu_1}{\sqrt{\omega^2(1-\lambda)^2 - 4(\omega-1)}} < 0.$$

Следовательно, корень $\mu_1(\lambda, \omega)$ при $1 < \omega < \omega_0$ убывает при изменении λ от λ_{\min} до λ_0 , принимая следующие значения:

$$\begin{aligned}\mu_1(\lambda_{\min}, \omega) &= \left(\frac{\omega(1-\lambda_{\min}) + \sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}}{2} \right)^2, \\ \mu_1(\lambda_0, \omega) &= \omega - 1.\end{aligned}$$

Далее, если λ меняется от λ_0 до 1, то корни μ_1 и μ_2 комплексны и равны по модулю: $|\mu_1| = |\mu_2| = \omega - 1$. Следовательно, если $1 < \omega < \omega_0$, то

$$\rho(S) = \mu_1(\lambda_{\min}, \omega) = \left(\frac{\omega(1-\lambda_{\min}) + \sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}}{2} \right)^2. \quad (15)$$

Если $\omega < 1$, то все корни уравнения (12) действительны, максимальным является корень μ_1 , значения которого убывают при изменении λ от λ_{\min}

до 1. Следовательно, при $\omega < 1$ спектральный радиус оператора S определяется формулой (15). Так как при $\omega = 1$ ненулевые μ_k удовлетворяют уравнению (12), то (15) имеет место и при $\omega = 1$.

Итак, если $0 < \omega < \omega_0$, то спектральный радиус оператора S определяется формулой (15). Покажем, что $\mu_1(\lambda_{\min}, \omega)$ убывает по ω в интервале $0 < \omega < \omega_0$.

Действительно, так как при $\omega < \omega_0$ корень μ_1 убывает по λ для $\lambda \leq \lambda_0$, а $\mu_1(0, \omega) = 1$, то $\mu_1(\lambda_{\min}, \omega) < 1$.

Далее, из (15) получим

$$\begin{aligned} \frac{\partial \mu_1(\lambda_{\min}, \omega)}{\partial \omega} &= V \overline{\mu_1} \left(1 - \lambda_{\min} + \frac{\omega (1 - \lambda_{\min})^2 - 2}{V \omega^2 (1 - \lambda_{\min})^2 - 4 (\omega - 1)} \right) = \\ &= \frac{V \overline{\mu_1}}{\omega} \frac{[\omega^2 (1 - \lambda_{\min})^2 - 2 (\omega - 1) + (1 - \lambda_{\min}) \omega V \overline{\mu_1} (1 - \lambda_{\min})^2 - 4 (\omega - 1) - 2]}{V \omega^2 (1 - \lambda_{\min})^2 - 4 (\omega - 1)}. \end{aligned}$$

Подставляя сюда (13), окончательно найдем

$$\frac{\partial \mu_1}{\partial \omega} = \frac{2 V \overline{\mu_1} (\mu_1 - 1)}{\omega V \omega^2 (1 - \lambda_{\min})^2 - 4 (\omega - 1)} < 0.$$

Утверждение доказано. Следовательно, в области $\omega \leq \omega_0$ оптимальным является значение $\omega = \omega_0$, которому соответствует

$$(\rho S) = \omega_0 - 1 = \frac{1 - V \overline{\lambda_{\min}} (2 - \lambda_{\min})}{1 + V \overline{\lambda_{\min}} (2 - \lambda_{\min})} = \left(\frac{1 - V \overline{\eta}}{1 + V \overline{\eta}} \right)^2, \quad \eta = \frac{\lambda_{\min}}{2 - \lambda_{\min}}.$$

Заметим, что из проведенных выше исследований вытекает, что если δ — оценка для λ_{\min} снизу, т. е. $\delta \leq \lambda_{\min}$, а ω выбрано по формуле (14) при замене λ_{\min} на δ , то $\omega_0 \leq \omega$, $\rho(S) \leq \left(\frac{1 - V \overline{\eta}}{1 + V \overline{\eta}} \right)^2$, $\eta = \frac{\delta}{2 - \delta}$.

Итак, доказана следующая

Теорема 5. Пусть выполнены предположения 1 и 2 и δ — постоянная из неравенства

$$\delta D \leq A, \quad \delta > 0. \quad (16)$$

Тогда для спектрального радиуса оператора перехода S итерационной схемы (1) при оптимальном значении параметра ω ,

$$\omega = \omega_0 = \frac{2}{1 + V \delta (2 - \delta)}, \quad (17)$$

справедлива оценка

$$\rho(S) \leq \left(\frac{1 - V \overline{\eta}}{1 + V \overline{\eta}} \right)^2, \quad \eta = \frac{\delta}{2 - \delta}, \quad (18)$$

причем, если в (16) достигается равенство, то равенство имеет место и в формуле (18).

Итерационный метод (1), (17) является методом верхней релаксации, так как $\omega_0 > 1$.

4. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Рассмотрим применение метода верхней релаксации для нахождения приближенного решения разностной задачи Дирихле для уравнения Пуассона, заданной на прямом

угольной сетке $\bar{\omega} = \{x_{i,j} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$ в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$:

$$\Lambda y = \sum_{\alpha=1}^2 y_{x_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma. \quad (19)$$

Оператор A в пространстве H сеточных функций, заданных на ω со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2$$

определяется обычным способом: $Ay = -\Lambda \dot{y}$, $y \in H$, $\dot{y} \in \dot{H}$. Как мы уже знаем, оператор A , соответствующий задаче (19), является самосопряженным и положительно определенным в H . Следовательно, предположение 1 выполняется.

Рассмотрим сначала *точечный метод верхней релаксации*. Если неизвестные упорядочены по строкам сетки ω , то разностная схема (19) может быть записана в виде следующей системы алгебраических уравнений:

$$-\frac{1}{h_1^2} y(i-1, j) - \frac{1}{h_2^2} y(i, j-1) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y(i, j) - \\ - \frac{1}{h_1^2} y(i+1, j) - \frac{1}{h_2^2} y(i, j+1) = \varphi(i, j)$$

для $i = 1, 2, \dots, N_1 - 1$, $j = 1, 2, \dots, N_2 - 1$ и $y(x) = g(x)$, $x \in \gamma$.

Такой записи оператора A соответствует представление A в виде суммы $A = \mathcal{D} + L + U$, где

$$\begin{aligned} \mathcal{D}y &= \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y, \\ Ly(i, j) &= -\frac{1}{h_1^2} \dot{y}(i-1, j) - \frac{1}{h_2^2} \dot{y}(i, j-1), \\ Uy(i, j) &= -\frac{1}{h_1^2} \dot{y}(i+1, j) - \frac{1}{h_2^2} \dot{y}(i, j+1). \end{aligned}$$

Для рассматриваемой системы точечный метод верхней релаксации в соответствии с формулой (3) будет иметь следующий вид:

$$\begin{aligned} \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{k+1}(i, j) &= (1 - \omega) \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_k(i, j) + \omega \left[\frac{1}{h_1^2} y_{k+1}(i-1, j) + \right. \\ &\quad \left. + \frac{1}{h_2^2} y_{k+1}(i, j-1) + \frac{1}{h_1^2} y_k(i+1, j) + \frac{1}{h_2^2} y_k(i, j+1) + \varphi(i, j) \right] \end{aligned}$$

для $i = 1, 2, \dots, N_1 - 1$, $j = 1, 2, \dots, N_2 - 1$, причем $y_k(x) = g(x)$ при $x \in \gamma$ для любого $k \geq 0$.

Вычисления, как и в методе Зейделя, начинаются с точки $i=1, j=1$ и продолжаются либо по строкам, либо по столбцам сетки ω . Найденное $y_{k+1}(i, j)$ размещается на месте $y_k(i, j)$.

Докажем теперь, что для рассматриваемого примера предположение 2 выполняется. Для этого нужно показать, что для любого комплексного $z \neq 0$ собственные значения μ задачи

$$z \left(\frac{1}{h_1^2} y(i-1, j) + \frac{1}{h_2^2} y(i, j-1) \right) + \frac{1}{z} \left(\frac{1}{h_1^2} y(i+1, j) + \frac{1}{h_2^2} y(i, j+1) \right) + \\ + \mu \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y(i, j) = 0, \quad 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \\ y(x) = 0, \quad x \in \gamma$$

не зависят от z .

Действительно, полагая здесь

$$y(i, j) = z^{i+j} v(i, j), \quad 0 \leq i \leq N_1, \quad 0 \leq j \leq N_2,$$

получим

$$\frac{1}{h_1^2} v(i-1, j) + \frac{1}{h_2^2} v(i, j-1) + \frac{1}{h_1^2} v(i+1, j) + \frac{1}{h_2^2} v(i, j+1) + \\ + \mu \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) v(i, j) = 0, \\ 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \quad v(x) = 0, \quad x \in \gamma.$$

Следовательно, μ не зависит от z .

Осталось найти оптимальное значение параметра ω . Для этого необходимо найти или оценить снизу минимальное собственное значение задачи (8), которая для данного случая записывается в виде

$$y_{x_1 x_1}^- + y_{x_2 x_2}^- + \lambda \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y = 0, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma.$$

Так как собственные значения разностного оператора Лапласа $\hat{\Delta} y = y_{x_1 x_1}^- + y_{x_2 x_2}^-$ известны

$$\hat{\lambda}_k = \frac{4}{h_1^2} \sin^2 \frac{k_1 \pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi h_2}{2l_2}, \quad k_\alpha = 1, 2, \dots, N_\alpha - 1,$$

то

$$\lambda_k = \hat{\lambda}_k / \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) = \frac{2h_2^2}{h_1^2 + h_2^2} \sin^2 \frac{k_1 \pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin^2 \frac{k_2 \pi h_2}{2l_2}.$$

Следовательно,

$$\lambda_{\min} = \frac{2h_2^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_2}{2l_2},$$

и параметр ω_0 находится по формуле (14). В частном случае, когда \bar{G} — квадрат со стороной l ($l_1 = l_2 = l$) и сетка квадратная ($N_1 = N_2 = N$), имеем

$$\lambda_{\min} = 2 \sin^2 \frac{\pi}{2N}, \quad \omega_0 = \frac{2}{1 + \sin \frac{\pi}{N}}, \quad \eta = \operatorname{tg}^2 \frac{\pi}{2N},$$

$$\rho(S) = \frac{1 - \sin \frac{\pi}{N}}{1 + \sin \frac{\pi}{N}} \approx 1 - \frac{2\pi}{N}.$$

Заметим, что спектральный радиус оператора перехода соответствующего точечного метода Зейделя оценивается по формуле (15), в которой следует положить $\omega = 1$. Это дает $\rho(S) = (1 - \lambda_{\min})^2 = \cos^2 \frac{\pi}{N}$, что значительно хуже, чем для метода верхней релаксации.

Рассмотрим теперь блочный метод верхней релаксации. Если в блок объединить неизвестные $y(i, j)$ на j -й строке сетки, то блочной записи оператора A соответствует следующее представление $A = \mathcal{D} + L + U$, где

$$\begin{aligned} \mathcal{D}y &= -\frac{1}{h_1^2} \dot{y}(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) \dot{y}(i, j) - \frac{1}{h_1^2} \dot{y}(i+1, j), \\ Ly(i, j) &= -\frac{1}{h_2^2} \dot{y}(i, j-1), \quad Uy(i, j) = -\frac{1}{h_2^2} \dot{y}(i, j+1). \end{aligned}$$

Расчетные формулы для блочного метода верхней релаксации имеют вид

$$\begin{aligned} -\frac{1}{h_1^2} y_{k+1}(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{k+1}(i, j) - \frac{1}{h_2^2} y_{k+1}(i+1, j) &= \\ = (1-\omega) \left(-\frac{1}{h_1^2} y_k(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_k(i, j) - \frac{1}{h_2^2} y_k(i+1, j) \right) &+ \\ + \omega \left(\frac{1}{h_2^2} y_{k+1}(i, j-1) + \frac{1}{h_2^2} y_k(i, j+1) + \varphi(i, j) \right), & \\ 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, & \end{aligned}$$

причем $y_k(x) = g(x)$, $x \in \gamma$ для всех $k \geq 0$. Для нахождения y_{k+1} на j -й строке необходимо решать, например, методом прогонки трехточечную краевую задачу.

Покажем, что для рассматриваемого примера предположение 2 выполнено, т. е. собственные значения μ задачи

$$\begin{aligned} z \frac{1}{h_2^2} y(i, j-1) + \frac{1}{z} \frac{1}{h_2^2} y(i, j+1) + \mu \left(-\frac{1}{h_1^2} y(i-1, j) + \right. \\ \left. + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y(i, j) - \frac{1}{h_2^2} y(i+1, j) \right) &= 0, \\ 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \quad y(x) = 0, \quad x \in \gamma & \end{aligned}$$

не зависят от z . Это легко устанавливается при помощи замены $y(i, j) = z^j v(i, j)$, $0 \leq i \leq N_1$, $0 \leq j \leq N_2$.

Найдем теперь оптимальное значение параметра ω . Соответствующая задача (8) имеет вид

$$\begin{aligned} y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \lambda \left(\frac{2}{h_2^2} y - y_{\bar{x}_1 x_1} \right) &= 0, \quad x \in \omega, \\ y(x) &= 0, \quad x \in \gamma. \end{aligned} \quad (20)$$

Несложно проверить, что собственными функциями задачи (20) являются

$$y_k(x) = \sin \frac{k_1 \pi x_1}{l_1} \sin \frac{k_2 \pi x_2}{l_2}. \quad (21)$$

Подставляя (21) в (20), найдем

$$\lambda_k = \frac{\lambda_{k_1} + \lambda_{k_2}}{\frac{2}{h_2^2} + \lambda_{k_1}}, \quad k_\alpha = 1, 2, \dots, N_\alpha - 1, \quad k = (k_1, k_2),$$

где

$$\lambda_{k_\alpha} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi h_\alpha}{2l_\alpha}, \quad k_\alpha = 1, 2, \dots, N_\alpha - 1, \quad \alpha = 1, 2.$$

Отсюда получим

$$\lambda_{\min} = \frac{2h_2^2 \sin^2 \frac{\pi h_1}{2l_1} + 2h_1^2 \sin^2 \frac{\pi h_2}{2l_2}}{2h_2^2 \sin^2 \frac{\pi h_1}{2l_1} + h_1^2}.$$

Для рассмотренного выше частного случая будем иметь

$$\lambda_{\min} = \frac{4 \sin^2 \frac{\pi}{2N}}{1 + 2 \sin^2 \frac{\pi}{2N}}, \quad \omega_0 = \frac{2 + 4 \sin^2 \frac{\pi}{2N}}{\left(1 + \sqrt{2} \sin \frac{\pi}{2N}\right)^2},$$

$$\eta = 2 \sin^2 \frac{\pi}{2N}, \quad \rho(S) = \left(\frac{1 - \sqrt{2} \sin \frac{\pi}{2N}}{1 + \sqrt{2} \sin \frac{\pi}{2N}} \right)^2 \approx 1 - 2\sqrt{2} \frac{\pi}{N}.$$

Сравнивая оценки спектрального радиуса блочного и точечного методов верхней релаксации, находим, что блочный метод будет сходиться в $\sqrt{2}$ раз быстрее, чем точечный метод. С другой стороны, блочный метод требует большего числа арифметических действий, затрачиваемых на реализацию одного итерационного шага, чем точечный метод.

В заключение приведем число итераций для точечного метода верхней релаксации в зависимости от числа узлов N по одному направлению для $\epsilon = 10^{-4}$. В качестве модельной задачи возьмем

разностную схему (19) на квадратной сетке с $N_1 = N_2 = N$ и $\varphi(x) \equiv 0$, $g(x) \equiv 0$. Начальное приближение $y_0(x)$ выберем следующим образом: $y_0(x) = 1$, $x \in \omega$, $y_0(x) = 0$, $x \in \gamma$.

Процесс итераций будем оканчивать, если выполняется условие

$$\|z_n\|_A \leq \varepsilon \|z_0\|_A. \quad (22)$$

Из теории метода следует, что для погрешности z_n имеет место оценка $\|z_n\|_A \leq \|S^n\|_A \|z_0\|_A$, и так как спектральный радиус оператора меньше либо равен любой нормы оператора, то $\rho^n(S) \leq \|S^n\|_A$. Поэтому условие $\rho^n(S) \leq \varepsilon$ нельзя использовать для оценки требуемого числа итераций.

Приведем число итераций n , определяемое из условия (22), и для сравнения найдем число итераций n^* , которое следует из неравенства $\rho^n(S) \leq \varepsilon$:

$$N = 32 \quad n = 65 \quad n^* = 47$$

$$N = 64 \quad n = 128 \quad n^* = 94$$

$$N = 128 \quad n = 257 \quad n^* = 187$$

Сравнение числа итераций для метода верхней релаксации и явного чебышевского метода, рассмотренного для задачи (19) в п. 1 § 5 гл. VI, показывает, что метод верхней релаксации требует примерно в 1,6 раза меньше итераций, чем явный чебышевский метод. Число арифметических действий, затрачиваемых на одну итерацию, в этих методах практически одинаково.

5. Разностная задача Дирихле для эллиптического уравнения с переменными коэффициентами. Рассмотрим теперь применение метода верхней релаксации для нахождения приближенного решения разностной задачи Дирихле для уравнения с переменными коэффициентами в прямоугольнике

$$\begin{aligned} \Lambda y &= \sum_{\alpha=1}^2 (a_\alpha(x) y_{x_\alpha})_{x_\alpha} - d(x) y = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \quad (23)$$

считая, что выполнены следующие условия:

$$\begin{aligned} 0 < c_1 &\leq a_\alpha(x) \leq c_2, \quad x \in \bar{\omega}, \quad \alpha = 1, 2, \\ 0 &\leq d_1 \leq d(x) \leq d_2, \quad x \in \omega. \end{aligned} \quad (24)$$

Для задачи (23) точечный метод верхней релаксации при упорядочении неизвестных по строкам сетки ω описывается формулой

$$\begin{aligned} b(i, j) y_{k+1}(i, j) &= (1 - \omega) b(i, j) y_k(i, j) + \\ &+ \omega \left[\frac{a_1(i, j)}{h_1^2} y_{k+1}(i-1, j) + \frac{a_2(i, j)}{h_2^2} y_{k+1}(i, j-1) + \right. \\ &+ \left. \frac{a_1(i+1, j)}{h_1^2} y_k(i+1, j) + \frac{a_2(i, j+1)}{h_2^2} y_k(i, j+1) + \varphi(i, j) \right], \\ 1 &\leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \end{aligned} \quad (25)$$

где

$$b(i, j) = \frac{a_1(i, j) + a_1(i+1, j)}{h_1^2} + \frac{a_2(i, j) + a_2(i, j+1)}{h_2^2} + d(i, j)$$

и $y_k(x) = g(x)$, $x \in \gamma$ для любого $k \geq 0$.

Для рассматриваемого примера операторы \mathcal{D} , L и U определяются следующим образом:

$$\mathcal{D}y = by,$$

$$Ly(i, j) = -\frac{a_1(i, j)}{h_1^2} \dot{y}(i-1, j) - \frac{a_2(i, j)}{h_2^2} \dot{y}(i, j-1),$$

$$Uy(i, j) = -\frac{a_1(i+1, j)}{h_1^2} \dot{y}(i+1, j) - \frac{a_2(i, j+1)}{h_2^2} \dot{y}(i, j+1).$$

Предположения 1 и 2 выполняются, что доказывается так же, как и для примера из п. 4.

Для того чтобы найти параметр ω , необходимо оценить постоянную δ в неравенстве $A \geq \delta \mathcal{D}$. Эта задача была решена ранее в п. 3 § 5 гл. VI, где рассматривался простейший неявный чебышевский метод для разностной задачи (23). Приведем оценку для δ :

$$\delta = \min_{0 < x_2 < l_2} \frac{1}{\kappa_1(x_2)} + \min_{0 < x_1 < l_1} \frac{1}{\kappa_2(x_1)},$$

где $\kappa_\alpha(x_\beta) = \max_{0 < x_\alpha < l_\alpha} v^\alpha(x)$, $\beta = 3 - \alpha$, $\alpha = 1, 2$, а $v^\alpha(x)$ есть решение следующей трехточечной краевой задачи:

$$\begin{aligned} (a_\alpha v_{x_\alpha}^\alpha)_{x_\alpha} - \frac{1}{2} dv^\alpha &= -b(x), \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ v^\alpha(x) &= 0, \quad x_\alpha = 0, \quad l_\alpha, \\ h_\beta \leq x_\beta \leq l_\beta - h_\beta, \quad \beta &= 3 - \alpha, \quad \alpha = 1, 2. \end{aligned}$$

Итерационный параметр ω находится по формуле (17):

$$\omega = \omega_0 = \frac{2}{1 + \sqrt{\delta(2-\delta)}}.$$

Для сравнения описанного метода верхней релаксации с простейшим неявным чебышевским методом, рассмотренным в п. 3 § 5 гл. VI, приведем число итераций метода верхней релаксации для следующего модельного примера. Пусть разностная схема (23) задана на квадратной сетке с $N_1 = N_2 = N$ и $\varphi(x) = 0$, $g(x) = 0$. Коэффициенты $a_1(x)$, $a_2(x)$ и $d(x)$ выберем следующим образом:

$$a_1(x) = 1 + c[(x_1 - 0,5)^2 + (x_2 - 0,5)^2],$$

$$a_2(x) = 1 + c[0,5 - (x_1 - 0,5)^2 - (x_2 - 0,5)^2],$$

$$d(x) \equiv 0, \quad c > 0.$$

При этом в неравенствах (24) $c_1 = 1$, $c_2 = 1 + 0,5c$, $d_1 = d_2 = 0$. Начальное приближение для итерационного метода верхней релаксации (25) выберем следующим образом: $y_0(x) = 1$, $x \in \omega$, $y_0(x) = 0$, $x \in \gamma$, и процесс итераций будем оканчивать при выполнении условия (22).

В табл. 9 приведено число итераций для метода релаксации в зависимости от отношения c_2/c_1 и от числа узлов N по одному направлению для $\varepsilon = 10^{-4}$. Для случая, когда $a_\alpha(x) \equiv 1$ и $d(x) \equiv 0$, число итераций метода верхней релаксации приведено в п. 4 настоящего параграфа.

Т а б л и ц а 9

c_1/c_2	2	8	32	128	512
$N = 32$	65	81	95	96	98
$N = 64$	129	164	192	193	195

Из табл. 9 следует, что число итераций метода верхней релаксации для модельного примера примерно в два раза меньше числа итераций простейшего неявного чебышевского метода. Так как число арифметических действий, затрачиваемых на реализацию одного итерационного шага, для указанных методов одинаково, то метод верхней релаксации примерно в два раза эффективнее простейшего неявного чебышевского метода.

§ 3. Треугольные методы

1. Итерационная схема. В §§ 1, 2 были изучены два метода — метод Зейделя и метод релаксации. Эти методы принадлежат классу неявных двухслойных методов, оператору B в которых соответствует треугольная или блочно треугольная матрица. В каноническом виде итерационная схема методов имеет следующий вид:

$$(\mathcal{D} + \omega L) \frac{y_{k+1} - y_k}{\omega} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (1)$$

где \mathcal{D} и L — операторы из разложения A на сумму диагональной, нижней и верхней треугольных матриц

$$A = \mathcal{D} + L + U. \quad (2)$$

Методу Зейделя соответствует значение параметра $\omega = 1$.

Для случая самосопряженного и положительно определенного в H оператора A достаточное условие сходимости в H_A итерационного метода (1) имеет вид

$$0 < \omega < 2. \quad (3)$$

В § 2 мы рассмотрели вопрос об оптимальном выборе итерационного параметра ω . Считая, что выполнены предположения 1 и 2 и априорная информация задана в виде постоянной δ из неравенства

$$\delta \mathcal{D} \leq A, \quad \delta > 0, \quad (4)$$

мы доказали, что оптимальное значение ω , при котором минимизируется спектральный радиус оператора перехода S схемы (1), определяется формулой

$$\omega = \omega_0 = \frac{2}{1 + \sqrt{\delta(2-\delta)}}. \quad (5)$$

В пп. 4, 5 § 2 были рассмотрены примеры задач, для которых предположения 1 и 2 выполнены. Эти предположения выполнены и для более сложных задач, например для пятиточечной разностной схемы, аппроксимирующей на неравномерной сетке в произвольной области задачу Дирихле для эллиптического уравнения с переменными коэффициентами.

Существуют, однако, примеры задач, для которых предположение 2 не выполняется. К ним относятся разностная задача Дирихле для эллиптического уравнения со смешанными производными, разностная задача Дирихле повышенного порядка точности и другие.

Неуниверсальность способа выбора итерационного параметра ω и отсутствие оценок скорости сходимости метода в какой-либо норме являются основными недостатками теории, развитой в § 2.

В настоящем параграфе будет рассмотрена общая схема треугольных итерационных методов, для которых итерационный параметр ω выбирается из условия минимизации в H_A нормы оператора перехода. Здесь же будет найдена оценка скорости сходимости метода в H_A в предположении самосопряженности и положительной определенности оператора A .

Рассмотрение треугольных методов начнем с преобразования итерационной схемы (1). Введем операторы R_1 и R_2 следующим образом:

$$R_1 = \frac{1}{2} \mathcal{D} + L, \quad R_2 = \frac{1}{2} \mathcal{D} + U.$$

Тогда разложение (2) будет иметь вид

$$A = R_1 + R_2, \quad (6)$$

и если A — самосопряженный в H оператор, то операторы R_1 и R_2 сопряжены друг другу

$$R_1 = R_2^*, \quad (7)$$

Подставляя $L = R_1 - \frac{1}{2} \mathcal{D}$ в (1) и обозначая

$$\tau = 2\omega/(2-\omega), \quad (8)$$

запишем итерационную схему (1) в эквивалентной форме

$$(\mathcal{D} + \tau R_1) \frac{y_{k+1} - y_k}{\tau} + A y_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (9)$$

причем в силу (3), (8) $\tau > 0$.

Схему (9) можно рассматривать независимо от схемы (1). Именно, пусть самосопряженный в H оператор A представлен по формуле (6) в виде суммы сопряженных друг другу операторов R_1 и R_2 , а \mathcal{D} —произвольный самосопряженный положительно определенный в H оператор. Итерационную схему (9) будем называть *каноническим видом треугольных итерационных методов*. Мы сохраняем название треугольные методы и в том случае, когда матрицы, соответствующие операторам R_1 и R_2 , не являются треугольными, а матрица, соответствующая оператору \mathcal{D} , не есть диагональная матрица.

Из теоремы 1 следует, что для положительно определенного оператора A итерационный метод (9) при $\tau > 0$ сходится в H_A . Действительно, для этого достаточно установить справедливость неравенства $\mathcal{D} + \tau R_1 > 0,5\tau A$. Из (7) получим

$$(Ax, x) = (R_1 x, x) + (R_2 x, x) = 2(R_1 x, x) = 2(R_2 x, x) \quad (10)$$

и, следовательно,

$$((\mathcal{D} + \tau R_1)x, x) = (\mathcal{D}x, x) + 0,5\tau(Ax, x) > 0,5\tau(Ax, x),$$

что и требовалось доказать.

В заключение отметим, что методу Зейделя в схеме (9) соответствует значение $\tau = 2$, а методу верхней релаксации— $\tau = 2/\sqrt{\delta(2-\delta)}$.

2. Оценка скорости сходимости. Оценим теперь скорость сходимости итерационной схемы (9) в H_A , предполагая, что A —самосопряженный положительно определенный в H оператор.

Переход в (9) к погрешности $z_k = y_k - u$ дает однородную схему для z_k

$$B \frac{z_{k+1} - z_k}{\tau} + Az_k = 0, \quad k = 0, 1, \dots, \quad z_0 = y_0 - u, \quad B = \mathcal{D} + \tau R_1,$$

откуда получим

$$\begin{aligned} z_{k+1} &= Sz_k, \quad k = 0, 1, \dots, \quad S = E - \tau B^{-1} A, \\ \|z_{k+1}\|_A &\leq \|S\|_A \|z_k\|_A. \end{aligned} \quad (11)$$

Оценим норму оператора перехода S в H_A . Из определения нормы оператора получим

$$\begin{aligned} \|S\|_A^2 &= \sup_{x \neq 0} \frac{(ASx, Sx)}{(Ax, x)} = \\ &= \sup_{x \neq 0} \left[1 - 2\tau \frac{(B^{-1}Ax, Ax)}{(Ax, x)} + \tau^2 \frac{(AB^{-1}Ax, B^{-1}Ax)}{(Ax, x)} \right]. \end{aligned} \quad (12)$$

Преобразуем выражение, стоящее в квадратных скобках. Используя (10) и определение оператора B , получим

$$(By, y) = (\mathcal{D}y, y) + \tau(R_1y, y) = (\mathcal{D}y, y) + 0,5\tau(Ay, y).$$

Отсюда найдем $\tau^2(Ay, y) = 2\tau(By, y) - 2\tau(\mathcal{D}y, y)$ или после замены $y = B^{-1}Ax$

$$\tau^2(AB^{-1}Ax, B^{-1}Ax) = 2\tau(B^{-1}Ax, Ax) - 2\tau(\mathcal{D}B^{-1}Ax, B^{-1}Ax).$$

Подставляя это выражение в (12), будем иметь

$$\|S\|_A^2 = \sup_{x \neq 0} \left[1 - 2\tau \frac{(\mathcal{D}B^{-1}Ax, B^{-1}Ax)}{(Ax, x)} \right].$$

Проведем дальнейшие преобразования. Полагая $x = (B^*)^{-1}\mathcal{D}^{1/2}y$, получим

$$\frac{(\mathcal{D}B^{-1}Ax, B^{-1}Ax)}{(Ax, x)} = \frac{(Cy, Cy)}{(Cy, y)}, \quad C = \mathcal{D}^{1/2}B^{-1}A(B^*)^{-1}\mathcal{D}^{1/2}.$$

Так как оператор C самосопряжен и положительно определен в H , то, полагая $y = C^{-1/2}\mathcal{D}^{-1/2}B^*v$, найдем

$$\frac{(\mathcal{D}B^{-1}Ax, B^{-1}Ax)}{(Ax, x)} = \frac{(Av, v)}{(B\mathcal{D}^{-1}B^*v, v)}.$$

Итак, окончательно будем иметь

$$\|S\|_A^2 = \sup_{v \neq 0} \left[1 - 2\tau \frac{(Av, v)}{(B\mathcal{D}^{-1}B^*v, v)} \right].$$

Отсюда получим, если γ_1 — величина из неравенства

$$\gamma_1 B\mathcal{D}^{-1}B^* \leq A, \quad (13)$$

то

$$\|S\|_A \leq (1 - 2\tau\gamma_1)^{1/2}. \quad (14)$$

Так как γ_1 зависит от параметра τ , то оптимальное значение для τ можно будет найти, получив при некоторых дополнительных предположениях относительно операторов \mathcal{D} , R_1 и R_2 выражение для γ_1 .

3. Выбор итерационного параметра. Выберем теперь параметр τ . Нам потребуется

Лемма 3. Пусть δ и Δ — постоянные в неравенствах

$$\delta\mathcal{D} \leq A, \quad R_1\mathcal{D}^{-1}R_2 \leq \frac{\Delta}{4}A, \quad \delta > 0. \quad (15)$$

Тогда в неравенстве (13)

$$\gamma_1 = \delta / \left(1 + \tau\delta + \tau^2 \frac{\delta\Delta}{4} \right). \quad (16)$$

Действительно, так как $B^* = \mathcal{D} + \tau R_2$, то

$$B\mathcal{D}^{-1}B^* = (\mathcal{D} + \tau R_1)\mathcal{D}^{-1}(\mathcal{D} + \tau R_2) = \mathcal{D} + \tau(R_1 + R_2) + \tau^2 R_1 \mathcal{D}^{-1} R_2 = \\ = \mathcal{D} + \tau A + \tau^2 R_1 \mathcal{D}^{-1} R_2.$$

Используя предположения (15), отсюда получим

$$B\mathcal{D}^{-1}B^* \leq (1/\delta + \tau + \tau^2 \Delta/4) A.$$

Лемма доказана.

Итак, если априорная информация имеет вид постоянных δ и Δ в неравенствах (15), то γ_1 оценивается формулой (16).

Подставляя (16) в (14), получим

$$\|S\|_A^2 \leq \varphi(\tau) = 1 - 2\tau\delta/(1 + \tau\delta + \tau^2 \frac{\delta\Delta}{4}).$$

Осталось минимизировать функцию $\varphi(\tau)$. Приравнивая производную $\varphi'(\tau)$ нулю, найдем

$$\varphi'(\tau) = \frac{2\delta \left(\tau^2 \frac{\delta\Delta}{4} - 1 \right)}{\left(1 + \tau\delta + \tau^2 \frac{\delta\Delta}{4} \right)^2} = 0, \quad \tau_0 = \frac{2}{\sqrt{\delta\Delta}}.$$

Так как при $\tau < \tau_0$ производная $\varphi'(\tau) < 0$, а при $\tau > \tau_0$ производная $\varphi'(\tau) > 0$, то при $\tau = \tau_0$ функция $\varphi(\tau)$ достигает минимума, равного $\varphi(\tau_0) = (1 - \sqrt{\eta})/(1 + \sqrt{\eta})$, $\eta = \delta/\Delta$. Итак, доказана

Теорема 6. Пусть A и \mathcal{D} —самосопряженные положительные определенные в H операторы, а δ и Δ —постоянные в (15). Треугольный итерационный метод (9), (6) при $\tau = \tau_0 = 2/\sqrt{\delta\Delta}$ сходится в H_A , и для погрешности z_n справедлива оценка $\|z_n\|_A \leq \rho^n \|z_0\|_A$. Для числа итераций n справедлива оценка $n \geq n_0(\varepsilon)$,

$$n_0(\varepsilon) = \ln \varepsilon / \ln \rho,$$

$$\text{где } \rho = \left(\frac{1 - \sqrt{\eta}}{1 + \sqrt{\eta}} \right)^{1/2}, \quad \eta = \frac{\delta}{\Delta}.$$

4. Оценка скорости сходимости методов Зейделя и релаксации.

Доказанная теорема 6 позволяет получить оценки скорости сходимости в H_A рассмотренных ранее методов Зейделя и верхней релаксации. В п. 2 § 1 и п. 4 § 2 указанные методы были применены для нахождения приближенного решения разностной задачи Дирихле для уравнения Пуассона на прямоугольной сетке

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$$

$$\begin{aligned} \Lambda y &= y_{\bar{x}_1, \bar{x}_1} + y_{\bar{x}_2, \bar{x}_2} = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma. \end{aligned}$$

Итерационная схема этих методов имела вид (1), где

$$\begin{aligned}\mathcal{D}y &= \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) \dot{y}, \\ Ly(i, j) &= -\frac{1}{h_1^2} \dot{y}(i-1, j) - \frac{1}{h_2^2} \dot{y}(i, j-1), \\ Uy(i, j) &= -\frac{1}{h_1^2} \dot{y}(i+1, j) - \frac{1}{h_2^2} \dot{y}(i, j+1).\end{aligned}$$

Для метода Зейделя $\omega = 1$, а для метода верхней релаксации ω находилось по формуле (5), где δ из неравенства (4) оценено следующим образом:

$$\delta = \frac{2h_2^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_2}{2l_2}. \quad (17)$$

Приведем схему (1) для рассматриваемого примера к виду (9). Для этого определим операторы R_1 и R_2 :

$$\begin{aligned}R_1y &= \left(\frac{1}{2} \mathcal{D} + L \right) y = \frac{1}{h_1} \dot{y}_{x_1} + \frac{1}{h_2} \dot{y}_{x_2}, \\ R_2y &= \left(\frac{1}{2} \mathcal{D} + U \right) y = -\frac{1}{h_1} \dot{y}_{x_1} - \frac{1}{h_2} \dot{y}_{x_2}.\end{aligned}$$

Очевидно, что

$$(R_1 + R_2)y = Ay = -\Lambda \dot{y} = -\dot{y}_{x_1} - \dot{y}_{x_2}.$$

Сопряженность операторов R_1 и R_2 друг другу легко устанавливается при помощи разностной формулы Грина. Как было отмечено выше, методу Зейделя в схеме (9) соответствует значение $\tau = 2$, а методу верхней релаксации — значение $\tau = 2/\sqrt{\delta(2-\delta)}$, где δ определено в (17).

Из (11), (14) и леммы 3 следует, что для получения оценок скорости сходимости этих методов в H_A требуется найти δ и Δ из неравенств (15). Постоянная δ уже найдена. Найдем Δ . Из определения операторов \mathcal{D} , R_1 и R_2 получим

$$(R_1 \mathcal{D}^{-1} R_2 y, y) = 0,5 \frac{h_1^2 h_2^2}{h_1^2 + h_2^2} (R_2 y, R_2 y). \quad (18)$$

Далее,

$$\begin{aligned}(R_2 y, R_2 y) &= \frac{1}{h_1^2} (\dot{y}_{x_1}^2, 1) - \frac{2}{h_1 h_2} (\dot{y}_{x_1}, \dot{y}_{x_2}) + \frac{1}{h_2^2} (\dot{y}_{x_2}^2, 1) \leqslant \\ &\leqslant \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right) [(\dot{y}_{x_1}^2, 1) + (\dot{y}_{x_2}^2, 1)] \leqslant \frac{h_1^2 + h_2^2}{h_1^2 h_2^2} (Ay, y).\end{aligned}$$

Подставляя эту оценку в (18), получим

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leqslant \frac{1}{2} (Ay, y)$$

и, следовательно, в неравенстве (15) $\Delta = 2$.

Оценим теперь скорость сходимости метода Зейделя и метода верхней релаксации.

Из (11) получим

$$\|z_n\|_A \leq \|S\|_A^n \|z_0\|_A$$

и, следовательно, для достижения точности ε достаточно выполнить $n \geq n_0(\varepsilon)$ итераций, где $n_0(\varepsilon) = \ln \varepsilon / \ln \|S\|_A$. Из (14) найдем

$$n_0(\varepsilon) = 2 \ln \varepsilon / \ln \|S\|_A^2 \geq \ln \frac{1}{\varepsilon} / (\tau \gamma_1). \quad (19)$$

Для метода Зейделя из (16) получим ($\tau = 2$)

$$\tau \gamma_1 = 2\delta / (1 + 4\delta) \quad (20)$$

и в частном случае, когда $N_1 = N_2 = N$, $l_1 = l_2 = l$, будем иметь из (17), (19) и (20)

$$\delta = 2 \sin^2 \frac{\pi}{2N}, \quad \tau \gamma_1 \approx 4 \sin^2 \frac{\pi}{2N} \approx \frac{\pi^2}{N^2},$$

$$n_0(\varepsilon) \approx \frac{N^2}{\pi^2} \ln \frac{1}{\varepsilon} \approx 0,1 N^2 \ln \frac{1}{\varepsilon}.$$

В п. 1 § 5 гл. VI для явного метода простой итерации, примененного для рассматриваемого частного случая, была получена следующая оценка числа итераций: $n_0(\varepsilon) \approx 0,2N^2 \ln \frac{1}{\varepsilon}$.

Сравнивая эти оценки, находим, что для метода Зейделя требуется примерно в два раза меньше итераций, чем для метода простой итерации. Характер зависимости числа итераций от числа узлов N по одному направлению для этих методов одинаков — число итераций пропорционально N^2 .

Рассмотрим теперь метод верхней релаксации. Подставляя в (16)

$$\delta = 2 \sin^2 \frac{\pi}{2N}, \quad \Delta = 2 \quad \text{и} \quad \tau = \frac{2}{\sqrt{\delta(2-\delta)}} = \frac{2}{\sin \frac{\pi}{N}},$$

получим

$$\tau \gamma_1 = \frac{2 \operatorname{tg} \frac{\pi}{2N}}{2 + 2 \operatorname{tg} \frac{\pi}{2N} + \operatorname{tg}^2 \frac{\pi}{2N}} \approx \operatorname{tg} \frac{\pi}{2N} \approx \frac{\pi}{2N}.$$

Из (19) найдем следующую оценку числа итераций для метода верхней релаксации:

$$n_0(\varepsilon) \approx \frac{2N}{\pi} \ln \frac{1}{\varepsilon} \approx 0,64N \ln \frac{1}{\varepsilon}, \quad (21)$$

т. е. число итераций для метода верхней релаксации пропорционально числу узлов N по одному направлению.

В заключение приведем оценку для числа итераций, которая следует из теоремы 6. При значении параметра τ

$$\tau = \tau_0 = \frac{2}{\sqrt{\delta \Delta}} = \frac{1}{\sin \frac{\pi}{2N}}$$

для числа итераций будет справедлива оценка

$$n \geq n_0(\varepsilon) = \ln \frac{1}{\varepsilon} / \sin \frac{\pi}{2N} \approx 0,64N \ln \frac{1}{\varepsilon}.$$

Заметим, что оценка (21) несколько завышена. Для того чтобы убедиться в этом, нужно сравнить значения $n_0(\varepsilon)$, вычисляемые по формуле (21), с числом итераций, приведенным в п. 4 § 2. Это связано с тем, что здесь число итераций оценивалось из неравенства $\|S\|_A^n \leq \varepsilon$, а в п. 4 § 2 итерации проводились до выполнения условия $\|S^n\|_A \leq \varepsilon$.

ГЛАВА X

ПОПЕРЕМЕННО-ТРЕУГОЛЬНЫЙ МЕТОД

В главе изучается попеременно-треугольный итерационный метод *) решения операторного уравнения с самосопряженным оператором. В § 1 излагается общая теория метода, описана конструкция итерационной схемы и указан набор итерационных параметров. Иллюстрируется метод на примере разностной задачи Дирихле для уравнения Пуассона в прямоугольнике. В § 2 этот метод применяется к решению разностных эллиптических уравнений с переменными коэффициентами и смешанными производными в прямоугольнике. Для решения эллиптического уравнения с переменными коэффициентами на неравномерной сетке в произвольной области в § 3 построен вариант попеременно-треугольного метода.

§ 1. Общая теория метода

1. Итерационная схема. В § 3 гл. IX был изучен треугольный итерационный метод решения уравнения

$$Au = f. \quad (1)$$

Итерационная схема этого метода имеет вид

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (2)$$

где $\tau_k \equiv \tau$, а оператор $B = B_1 = \mathcal{D} + \tau R_1$ определяется следующим разложением оператора A на сумму операторов:

$$A = R_1 + R_2, \quad R_1 = R_2^*, \quad A = A^* > 0. \quad (3)$$

Относительно оператора \mathcal{D} предполагается, что он самосопряжен и положительно определен в H , т. е.

$$\mathcal{D} = \mathcal{D}^* > 0. \quad (4)$$

Треугольный итерационный метод относится к классу методов, итерационные параметры для которых выбираются с учетом априорной информации об операторах итерационной схемы. Для

*) Метод предложен А. А. Самарским в 1964 г. (см. ЖВМ и МФ, 4, № 3, 1964) и усовершенствован в [8].

треугольного метода первичная информация состоит в задании постоянных δ и Δ из неравенств

$$\delta \mathcal{D} \leq A, \quad R_1 \mathcal{D}^{-1} R_2 \leq \frac{\Delta}{4} A, \quad \delta > 0. \quad (5)$$

Найденный в п. 3 § 3 гл. IX параметр τ позволяет получить точность ε за $n_0 = O(\ln(1/\varepsilon) \sqrt{\eta})$ итераций, где $\eta = \delta/\Delta$.

Отметим, что несамосопряженность оператора B не позволяет использовать в итерационной схеме (2) набор параметров τ_k и увеличить тем самым скорость сходимости метода. Однако простота конструкции оператора B и возможность разложения (3) для операторов A любой структуры стимулировали изучение возможных видоизменений треугольного метода. В результате был построен попаременно-треугольный метод, сочетающий универсальность построения оператора B с возможностью выбора в схеме (2) набора параметров τ_k .

Приступаем к изучению попаременно-треугольного метода. Итерационная схема метода имеет вид (2), где оператор B определяется следующим образом:

$$B = (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2), \quad \omega > 0. \quad (6)$$

Здесь ω — итерационный параметр, подлежащий определению. Будем далее предполагать, что для схемы (2), (6) выполнены условия (3), (4) и заданы δ и Δ в неравенствах (5).

Отметим некоторые свойства оператора B , определяемого соотношением (6). Если оператору $\mathcal{D} + \omega R_1$ соответствует треугольная матрица, а \mathcal{D} — диагональная, то B соответствует произведение двух треугольных и диагональной матриц. В этом случае сращение оператора B не является сложной задачей.

Покажем, что оператор B самосопряжен в H , и если оператор \mathcal{D} ограничен, то B положительно определен. Действительно, в силу (3) имеем равенства

$$(Au, u) = 2(R_1 u, u) = 2(R_2 u, u) > 0.$$

Отсюда и из (4) следует, что операторы $B_1 = \mathcal{D} + \omega R_1$ и $B_2 = \mathcal{D} + \omega R_2$ являются сопряженными и положительно определенными: $B_1^* = (\mathcal{D} + \omega R_1)^* = \mathcal{D} + \omega R_2 = B_2$, $B_\alpha > \mathcal{D} > 0$, $\alpha = 1, 2$, поэтому

$$B^* = (B_1 \mathcal{D}^{-1} B_2)^* = B_2^* \mathcal{D}^{-1} B_1^* = B_1 \mathcal{D}^{-1} B_2 = B.$$

Далее, так как \mathcal{D} есть самосопряженный ограниченный и положительно определенный оператор, то обратный оператор \mathcal{D}^{-1} будет положительно определен в H . Следовательно, используя неравенство $(\mathcal{D}^{-1}x, x) \geq d(x, x)$, $d > 0$, выражающее положительную определенность оператора \mathcal{D}^{-1} , имеем

$$(Bu, u) = (\mathcal{D}^{-1} B_2 u, B_2 u) \geq d \|B_2 u\|^2 > 0.$$

Из (2), (6) видно, что для определения y_{k+1} при заданном y_k надо решать уравнение

$$(\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2) y_{k+1} = \varphi_k, \quad k = 0, 1, \dots,$$

где $\varphi_k = By_k - \tau_{k+1}(Ay_k - f)$. Оно сводится к решению двух уравнений

$$(\mathcal{D} + \omega R_1) v = \varphi_k, \quad (\mathcal{D} + \omega R_2) y_{k+1} = \mathcal{D}v.$$

Возможен второй алгоритм реализации схемы (2), (6), основанный на записи ее в виде схемы с поправкой

$$y_{k+1} = y_k - \tau_{k+1} w_k, \quad Bw_k = r_k,$$

где $r_k = Ay_k - f$ — невязка. Поправка w_k находится решением двух уравнений

$$(\mathcal{D} + \omega R_1) \bar{w}_k = r_k, \quad (\mathcal{D} + \omega R_2) w_k = \mathcal{D}\bar{w}_k.$$

В этом алгоритме мы избавлены от необходимости вычислять By_k , но обязаны хранить одновременно и y_k , и промежуточные величины r_k , w_k , \bar{w}_k .

2. Выбор итерационных параметров. Займемся теперь исследованием сходимости итерационной схемы (2), (6). Так как операторы A и B самосопряжены и положительно определены в H , то можно изучать сходимость в H_D , где в качестве D взят один из операторов A , B или $AB^{-1}A$ (в последнем случае B должен быть ограниченным оператором). Для указанного оператора D оператор $DB^{-1}A$ будет, очевидно, самосопряжен в H и, следовательно, согласно классификации главы VI, мы имеем итерационную схему в самосопряженном случае.

Используя результаты § 2 гл. VI, мы можем сразу указать для схемы (2), (6) оптимальный набор итерационных параметров τ_k . Пусть γ_1 и γ_2 взяты из неравенств

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \quad (7)$$

Тогда чебышевский набор параметров $\{\tau_k\}$ определяется формулами

$$\begin{aligned} \tau_k &= \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ \cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \quad 1 \leq k \leq n, \\ \tau_0 &= \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad n \geq n_0(\varepsilon) = \frac{\ln(0.5\varepsilon)}{\ln \rho_1}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \end{aligned} \quad (8)$$

и для погрешности $z_n = y_n - u$ итерационного метода (2), (6), (8) имеет место оценка

$$\|z_n\|_D \leq q_n \|z_0\|_D, \quad q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}} \leq \varepsilon, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}. \quad (9)$$

Это результат общей теории итерационных двухслойных методов. Для схемы (2), (6) априорной информацией являются

постоянные δ и Δ в неравенствах (5). Поэтому одной из задач является получение выражений для γ_1 и γ_2 через δ и Δ . Далее, поскольку оператор B зависит от итерационного параметра ω , то γ_1 и γ_2 являются функциями ω : $\gamma_1 = \gamma_1(\omega)$, $\gamma_2 = \gamma_2(\omega)$. Так как из оценки (9) следует, что максимальная скорость сходимости будет тогда, когда отношение $\xi = \gamma_1/\gamma_2$ максимально, то мы приходим к задаче о выборе параметра ω из условия максимума ξ . Сформулированные задачи решает

Лемма 1. *Пусть выполнены условия (3), (4), оператор B определен по формуле (6) и в неравенствах (5) заданы постоянные δ и Δ . Тогда в неравенствах (7) имеем*

$$\gamma_1 = \delta / (1 + \omega\delta + \frac{1}{4}\omega^2\delta\Delta), \quad \gamma_2 = 1/(2\omega). \quad (10)$$

Отношение $\xi = \gamma_1/\gamma_2$ максимально, если

$$\omega = \omega_0 = 2/\sqrt{\delta\Delta}, \quad (11)$$

при этом

$$\gamma_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \gamma_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \xi = \frac{2\sqrt{\eta}}{1 + \sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}. \quad (12)$$

Действительно, запишем оператор B в виде

$$B = (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2) = \mathcal{D} + \omega (R_1 + R_2) + \omega^2 R_1 \mathcal{D}^{-1} R_2. \quad (13)$$

Учитывая, что $A = R_1 + R_2 \geq \delta \mathcal{D}$ или $\mathcal{D} \leq \frac{1}{\delta} A$, получаем для B оценку сверху

$$B \leq \left(\frac{1}{\delta} + \omega + \frac{1}{4}\omega^2\Delta \right) A = \frac{1}{\gamma_1} A,$$

т. е. $A \geq \gamma_1 B$, где γ_1 определено в (10).

Преобразуем теперь формулу (13):

$$B = \mathcal{D} - \omega (R_1 + R_2) + \omega^2 R_1 \mathcal{D}^{-1} R_2 + 2\omega (R_1 + R_2) = \\ = (\mathcal{D} - \omega R_1) \mathcal{D}^{-1} (\mathcal{D} - \omega R_2) + 2\omega A.$$

Отсюда следует

$$(By, y) = 2\omega (Ay, y) + (\mathcal{D}^{-1}(\mathcal{D} - \omega R_2)y, (\mathcal{D} - \omega R_2)y).$$

Используя положительную определенность оператора \mathcal{D}^{-1} , получаем $(By, y) \geq 2\omega (Ay, y)$, т. е. $A \leq \gamma_2 B$. Итак, γ_1 и γ_2 найдены. Рассмотрим далее отношение

$$\xi = \xi(\omega) = \gamma_1/\gamma_2 = 2\omega\delta / \left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4} \right).$$

Приравнивая производную

$$\xi'(\omega) = \frac{2\delta(1 - \omega^2\delta\Delta/4)}{\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right)^2}$$

нулю, находим $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$. В этой точке достигается максимум $\xi(\omega)$, так как $\xi''(\omega_0) < 0$. Подставляя найденное ω в (10), получаем (12). Покажем, что $\delta \leq \Delta$, $\eta \leq 1$. Действительно, используя равенство $(Ax, x) = 2(R_2x, x)$ и неравенство Коши—Буняковского, из (5) получим

$$\begin{aligned} \delta(\mathcal{D}x, x) &\leq (Ax, x) = \frac{(Ax, x)^2}{(Ax, x)} = 4 \frac{(R_2x, x)^2}{(Ax, x)} = \\ &= 4 \frac{(\mathcal{D}^{-1/2}R_2x, \mathcal{D}^{1/2}x)^2}{(Ax, x)} \leq 4 \frac{(\mathcal{D}^{-1/2}R_2x, \mathcal{D}^{-1/2}R_2x)}{(Ax, x)} (\mathcal{D}^{1/2}x, \mathcal{D}^{1/2}x) = \\ &= 4 \frac{(R_1\mathcal{D}^{-1}R_2x, x)}{(Ax, x)} (\mathcal{D}x, x) \leq \Delta(\mathcal{D}x, x), \end{aligned}$$

что и требовалось доказать. Лемма доказана.

Теорема 1. Пусть выполнены условия леммы 1. Тогда для попаременно-треугольного метода (2), (6), (11) с чебышевскими параметрами τ_k , определяемыми формулами (8) и (12), справедлива оценка (9). Для выполнения неравенства $\|z_n\|_D \leq \varepsilon \|z_0\|_D$ достаточно n итераций, где $n \geq n_0(\varepsilon)$, $n_0(\varepsilon) = \ln \frac{2}{\varepsilon} / (2\sqrt{2}\sqrt[4]{\eta})$, $\eta = \delta/\Delta$. Здесь $D = A$, B или $AB^{-1}A$.

Для доказательства теоремы следует использовать лемму 1 и формулы (8) для итерационных параметров и числа итераций.

Рассмотрим теперь один прием, который используется при построении неявных итерационных схем. Пусть в H задан самосопряженный и положительно определенный оператор R , энергетически эквивалентный оператору A с постоянными c_1 и c_2 :

$$c_1R \leq A \leq c_2R, \quad c_1 > 0, \quad (14)$$

и оператору B с постоянными $\dot{\gamma}_1$ и $\dot{\gamma}_2$:

$$\dot{\gamma}_1B \leq R \leq \dot{\gamma}_2B, \quad \dot{\gamma}_1 > 0. \quad (15)$$

Предположим, что операторы A и B самосопряжены. Из (14) и (15) для них получим следующие неравенства: $\dot{\gamma}_1B \leq A \leq \dot{\gamma}_2B$, $\gamma_1 = c_1\dot{\gamma}_1$, $\gamma_2 = c_2\dot{\gamma}_2$. Изложенный прием позволяет при построении оператора B исходить не из разложения (3) оператора A , а из разложения оператора R , который может быть выбран для большого класса различных операторов A одним и тем же. При этом постоянные $\dot{\gamma}_1$ и $\dot{\gamma}_2$ в (15) могут быть найдены один раз, и вся задача получения априорной информации для метода сводится к нахождению c_1 и c_2 в (14).

Итак, пусть оператор R представлен в виде суммы сопряженных операторов R_1 и R_2 :

$$R = R^* > 0, \quad R = R_1 + R_2, \quad R_1 = R_2^*, \quad (16)$$

и вместо (5) имеют место неравенства

$$\delta \mathcal{D} \leq R, \quad R_1 \mathcal{D}^{-1} R_2 \leq \frac{\Delta}{4} R, \quad \delta > 0. \quad (17)$$

Оператор B для схемы (2) построим по формуле (6). Тогда в силу леммы 1 при $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$ в неравенствах (15) имеем

$$\dot{\gamma}_1 = \frac{\delta}{2(1+\sqrt{\eta})}, \quad \dot{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \dot{\xi} = \frac{\dot{\gamma}_1}{\dot{\gamma}_2} = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}. \quad (18)$$

Отсюда следует

Теорема 2. Пусть $A = A^* > 0$, $\mathcal{D} = \mathcal{D}^* > 0$, выполнены условия (16) и заданы c_1 и c_2 в (14) и δ , Δ в (17). Тогда для попеременно-треугольного метода (2), (6), (8), (11) с чебышевскими параметрами τ_k , где $\gamma_1 = c_1 \dot{\gamma}_1$ и $\gamma_2 = c_2 \dot{\gamma}_2$, а γ_1 и γ_2 определены в (18), справедлива оценка (9). Для выполнения неравенства $\|z_n\|_{\mathcal{D}} \leq \varepsilon \|z_0\|_{\mathcal{D}}$ достаточно n итераций, где

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}} \sqrt{\frac{c_2}{c_1}}, \quad \eta = \frac{\delta}{\Delta}.$$

3. Метод нахождения исходных величин δ и Δ . Из теорем 1 и 2 следует, что для применения попеременно-треугольного метода требуется задать два числа δ и Δ в неравенствах (5) или (17). В рассмотренных ниже примерах сеточных эллиптических уравнений эти постоянные будут найдены в явном виде или будут указаны алгоритмы для их вычисления. При этом, естественно, используется структура операторов A , R_1 , R_2 и \mathcal{D} . Для общей теории итерационных методов, которая не учитывает конкретную структуру операторов, необходимо предложить общий способ нахождения априорной информации, требуемой для реализации метода.

Этот способ может быть основан на использовании асимптотического свойства итерационных методов вариационного типа (см. п. 5 § 1 гл. VIII). Пусть операторы A и B самосопряжены и положительно определены в H . Если в итерационной схеме

$$B \frac{v_{k+1} - v_k}{\tau_{k+1}} + Av_k = 0, \quad k = 0, 1, \dots, \quad v_0 \neq 0 \quad (19)$$

выбрать параметры τ_{k+1} по формуле метода скорейшего спуска

$$\tau_{k+1} = \frac{(w_k, r_k)}{(Aw_k, w_k)}, \quad k = 0, 1, \dots, \quad r_k = Av_k, \quad Bw_k = r_k, \quad (20)$$

и для достаточно большого номера итераций n найти корни $x_1 \leq x_2$ уравнения

$$(1 - \tau_n x)(1 - \tau_{n-1} x) = \rho_n \sigma_{n-1}, \quad \rho_n = \frac{\|v_n\|_A}{\|v_{n-1}\|_A}, \quad (21)$$

где $\|\cdot\|_A$ — норма в H_A , то x_1 и x_2 будут приближениями к γ_1 и γ_2 в неравенствах (7) соответственно сверху и снизу.

Воспользуемся описанным способом. Рассмотрим итерационную схему (2), (3), (6). Заметим, что в силу леммы 1 для неравенства (7) $\gamma_2 = 1/(2\omega)$, а от априорных данных зависит лишь γ_1 . Мы попытаемся, не находя отдельно δ и Δ для неравенств (5), сразу найти выражение для γ_1 как функции итерационного параметра ω . Этому выражению мы придадим вид (10), указав соответствующие δ и Δ значения. Тогда из леммы 1 найдем ω_0 по формуле (11) и γ_1 и γ_2 , соответствующие ω_0 , по формуле (12). Для набора параметров τ_k используем (8).

Найдем искомое выражение для γ_1 . Возьмем $\omega = 0$ и по методу (19)–(21) найдем x_1 . Считая, что в (19), (20) проделано достаточное количество итераций, и учитывая, что при $\omega = 0$ оператор $B = \mathcal{D}$, получим приближенное неравенство

$$x_1 \mathcal{D} \leq A, \quad x_1 > 0. \quad (22)$$

Далее, возьмем $\omega = \omega_1 > 0$ и по методу (19)–(21) найдем \bar{x}_1 , так что $\bar{x}_1 > 0$ и

$$\bar{x}_1 B \leq A \quad \text{или} \quad \bar{x}_1 (\mathcal{D} + \omega_1 A + \omega_1^2 R_1 \mathcal{D}^{-1} R_2) \leq A, \quad (23)$$

причем видно, что $\bar{x}_1 \omega_1 < 1$. Запишем (23) в виде

$$\bar{x}_1 \mathcal{D} + \bar{x}_1 \omega_1^2 R_1 \mathcal{D}^{-1} R_2 \leq (1 - \bar{x}_1 \omega_1) A$$

и сложим с (22), которое предварительно умножим на некоторый неопределенный пока коэффициент $\alpha > 0$. Получим

$$(\alpha x_1 + \bar{x}_1) \mathcal{D} + \bar{x}_1 \omega_1^2 R_1 \mathcal{D}^{-1} R_2 \leq (1 - \bar{x}_1 \omega_1 + \alpha) A. \quad (24)$$

Разделим это неравенство на $\alpha x_1 + \bar{x}_1$, добавим к правой и левой частям слагаемое ωA и выберем α из условия

$$\bar{x}_1 \omega_1^2 = \omega^2 (\alpha x_1 + \bar{x}_1); \quad (25)$$

тогда преобразованное неравенство будет иметь вид

$$\mathcal{D} + \omega A + \omega^2 R_1 \mathcal{D}^{-1} R_2 = B \leq \frac{1}{\gamma_1} A,$$

где

$$\frac{1}{\gamma_1} = \frac{1}{\gamma_1(\omega)} = \omega + \frac{1 - \bar{x}_1 \omega_1 + \alpha}{\alpha x_1 + \bar{x}_1}. \quad (26)$$

Из (25) найдем α :

$$\alpha = \bar{x}_1 (\omega_1^2 - \omega^2) / (\omega^2 x_1).$$

Так как α должно быть положительным, то выражение (26) будет иметь место для $0 < \omega < \omega_1$. Подставляя найденное α в (26), получим

$$\frac{1}{\gamma_1} = \frac{1}{x_1} + \omega + \frac{x_1 - \bar{x}_1 - x_1 \bar{x}_1 \omega_1}{x_1 \bar{x}_1 \omega_1^2} \omega^2.$$

Сравнивая это выражение с (10), получим, что в качестве δ и Δ можно взять

$$\delta = x_1, \quad \Delta = 4 \frac{x_1 - \bar{x}_1 - x_1 \bar{x}_1 \omega_1}{x_1 \bar{x}_1 \omega_1^2}.$$

Отметим, что ω_0 , найденное по указанным δ и Δ согласно (11), будет принадлежать интервалу $(0, \omega_1)$, если выполнено неравенство $2\bar{x}_1 \leq x_1(1 - \bar{x}_1 \omega_1)$. Если это неравенство не выполняется, то следует увеличить ω_1 и провести указанные расчеты заново (рекомендуется брать $\omega_1 = 2/x_1$).

4. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Проиллюстрируем попеременно-треугольный метод на примере разностной задачи Дирихле для уравнения Пуассона в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$:

$$\begin{aligned} \Lambda y &= y_{\bar{x}_1, x_1} + y_{\bar{x}_2, x_2} = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma \end{aligned} \tag{27}$$

на сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$ с границей γ .

Для данного примера H — пространство сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2.$$

Оператор A определяется равенством $Ay = -\Lambda y$, где $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$, $x \in \omega$, а $\dot{y}(x) = 0$ для $x \in \gamma$. Правую часть f определим обычным образом: $f(x) = \varphi(x) + \frac{1}{h_1^2}\varphi_1(x) + \frac{1}{h_2^2}\varphi_2(x)$, где

$$\begin{aligned} \varphi_1(x) &= \begin{cases} g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leq x_1 \leq l_1 - 2h_1, \\ g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases} \\ \varphi_2(x) &= \begin{cases} g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leq x_2 \leq l_2 - 2h_2, \\ g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases} \end{aligned}$$

Тогда задача (27) записывается в виде уравнения (1).

Оператор A самосопряжен и положительно определен в H , так как он соответствует разностному оператору Лапласа при краевых условиях Дирихле.

Займемся теперь конструированием оператора B . Будем рассматривать классический вариант попаременно-треугольного метода, для которого в (6) положим

$$\mathcal{D} = E. \quad (28)$$

Определим теперь разностные операторы \mathcal{R}_1 и \mathcal{R}_2 , которые действуют на сеточные функции, заданные на $\bar{\omega}$, следующим образом:

$$\mathcal{R}_1 y = - \sum_{\alpha=1}^2 \frac{1}{h_\alpha} y_{\bar{x}_\alpha}, \quad \mathcal{R}_2 y = \sum_{\alpha=1}^2 \frac{1}{h_\alpha} y_{x_\alpha}, \quad x \in \omega.$$

Очевидно, что $\mathcal{R}_1 + \mathcal{R}_2 = \Lambda$. Используя разностные формулы Грина, легко получим, что для сеточных функций $y(x) \in \dot{H}$, $\dot{u}(x) \in \dot{H}$, т. е. заданных на $\bar{\omega}$ и обращающихся в нуль на γ , имеет место равенство

$$(\mathcal{R}_1 \dot{y}, \dot{u}) = (\dot{y}, \mathcal{R}_2 \dot{u}). \quad (29)$$

Определим на H операторы R_1 и R_2 следующим образом: $R_\alpha y = -\mathcal{R}_\alpha \dot{y}$, $\alpha = 1, 2$, где $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$, $x \in \omega$. Тогда в силу определения разностных операторов \mathcal{R}_α и равенства (29) выполнены условия (3), т. е. $A = R_1 + R_2$, $R_1 = R_2^*$. Учитывая (28), из (6) получим следующий вид оператора B :

$$B = (E + \omega R_1)(E + \omega R_2).$$

Найдем теперь необходимую для реализации попаременно-треугольного метода априорную информацию. В данном случае она имеет вид постоянных δ и Δ в неравенствах $\delta E \leq A$, $R_1 R_2 \leq (\Delta/4)A$. Очевидно, что в качестве δ можно взять минимальное собственное значение разностного оператора Лапласа

$$\delta = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}.$$

Оценка для Δ найдена в п. 4 § 3 гл. IX (операторы R_1 и R_2 , определенные здесь и там, совпадают). Имеем $\Delta = 4/h_1^2 + 4/h_2^2$.

Итак, необходимая информация о δ и Δ получена. Из леммы 1 найдем оптимальное значение для параметра ω_0 , а также γ_1 и γ_2 . Итерационные параметры τ_k вычисляются по формулам (8). В частном случае, когда $N_1 = N_2 = N$, $l_1 = l_2 = l$, получим

$$\delta = \frac{8}{h^2} \sin^2 \frac{\pi}{2N}, \quad \Delta = \frac{8}{h^2}, \quad \eta = \frac{\delta}{\Delta} = \sin^2 \frac{\pi}{2N},$$

$$\xi = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}} \approx 2\sqrt{\eta} = 2 \sin \frac{\pi}{2N} \approx \frac{\pi}{N}, \quad \omega_0 = \frac{h^2}{4 \sin \frac{\pi h}{2l}}.$$

Из теоремы 1 для числа итераций n в этом случае будем иметь $n \geq n_0(\varepsilon)$, где

$$n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{2\sqrt[4]{\frac{2}{\pi}}} \approx \frac{\sqrt{N}}{2\sqrt{\pi}} \ln \frac{2}{\varepsilon} \approx 0,28\sqrt{N} \ln \frac{2}{\varepsilon},$$

т. е. число итераций пропорционально корню четвертой степени от числа неизвестных в задаче.

В п. 4 § 3 гл. IX для метода верхней релаксации, примененного к решению разностной задачи (27), была получена следующая оценка числа итераций:

$$n \geq n_0(\varepsilon) \approx 0,64N \ln(1/\varepsilon), \quad (N = l/h).$$

Сравнение метода релаксации с попеременно-треугольным методом показывает явное преимущество последнего. Хотя на реализацию одного итерационного шага в попеременно-треугольном методе нужно затратить вдвое больше арифметических действий, чем в методе релаксации, он имеет существенный выигрыш в числе итераций, что и обеспечивает общую эффективность этого метода.

Приведем теперь число итераций для попеременно-треугольного метода с чебышевскими параметрами, рассмотренного здесь для разностной задачи (27), в зависимости от числа узлов N по одному направлению квадратной сетки $\bar{\omega}$ для $\varepsilon = 10^{-4}$:

$$\begin{aligned} N &= 32 & n &= 16 \\ N &= 64 & n &= 23 \\ N &= 128 & n &= 32 \end{aligned}$$

Сравнение с числом итераций метода верхней релаксации, которое приведено в п. 4 § 2 гл. IX, показывает, что метод релаксации требует примерно в 3,5—7,5 раза больше итераций, чем попеременно-треугольный метод.

Замечание 1. Если вместо прямоугольника \bar{G} рассмотреть p -мерный параллелепипед $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2, \dots, p\}$, а в нем разностную задачу Дирихле для уравнения Пуассона

$$\begin{aligned} \Lambda y &= \sum_{\alpha=1}^p y_{x_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned}$$

на прямоугольной сетке $\bar{\omega} = \{x_i = (i_1 h_1, i_2 h_2, \dots, i_p h_p) \in \bar{G}, 0 \leq i_\alpha \leq N_\alpha, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2, \dots, p\}$, то разностные операторы \mathcal{R}_1 и \mathcal{R}_2 определяются следующим образом:

$$\mathcal{R}_1 y = - \sum_{\alpha=1}^p \frac{1}{h_\alpha} y_{x_\alpha}, \quad \mathcal{R}_2 y = \sum_{\alpha=1}^p \frac{1}{h_\alpha} y_{x_\alpha}, \quad x \in \omega.$$

В этом случае в неравенствах (5) для $\mathcal{D} = E$ следует положить

$$\delta = \sum_{\alpha=1}^p \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \sum_{\alpha=1}^p \frac{4}{h_\alpha^2},$$

так как

$$\begin{aligned} \|\mathcal{R}_2 \dot{y}\|^2 &= (\mathcal{R}_1 \mathcal{R}_2 \dot{y}, \dot{y}) = \left\| \sum_{\alpha=1}^p \frac{1}{h_\alpha} \dot{y}_{x_\alpha} \right\|^2 \leqslant \\ &\leqslant \sum_{\alpha=1}^p \frac{1}{h_\alpha^2} \sum_{\alpha=1}^p \|\dot{y}_{x_\alpha}\|^2 \leqslant \left(\sum_{\alpha=1}^p \frac{1}{h_\alpha^2} \right) (A \dot{y}, \dot{y}). \end{aligned}$$

При этом в случае $N_1 = N_2 = \dots = N_p = N$, $l_1 = l_2 = \dots = l_p = l$ для числа итераций имеем оценку

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) \approx \frac{\sqrt{N}}{2\sqrt{\pi}} \ln \frac{2}{\varepsilon} \approx 0,28\sqrt{N} \ln \frac{2}{\varepsilon},$$

которая не зависит от числа измерений p .

Рассмотрим теперь вопросы, связанные с реализацией попарно-треугольного метода для задачи (27). В п. 1 были приведены два алгоритма для нахождения y_{k+1} по заданному y_k для итерационной схемы метода. Рассмотрим сначала второй алгоритм. Для $\mathcal{D} = E$ он имеет следующий вид:

$$\begin{aligned} r_k &= Ay_k - f, \\ (E + \omega_0 R_1) \bar{w}_k &= r_k, \quad (E + \omega_0 R_2) w_k = \bar{w}_k, \\ y_{k+1} &= y_k - \tau_{k+1} w_k, \quad k = 0, 1, \dots \end{aligned} \tag{30}$$

Этим алгоритмом следует пользоваться, когда параметры τ_k выбираются не по формулам (8), а по формулам итерационных методов вариационного типа.

Используя определение операторов A , R_1 и R_2 посредством разностных операторов Λ , \mathcal{R}_1 и \mathcal{R}_2 , формулы (30) можно записать в следующем виде:

$$\begin{aligned} r_k(i, j) &= \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_k(i, j) - \frac{1}{h_1^2} [y_k(i-1, j) + y_k(i+1, j)] - \\ &\quad - \frac{1}{h_2^2} [y_k(i, j-1) + y_k(i, j+1)] - \varphi(i, j), \end{aligned} \tag{31}$$

$$1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \quad y_k|_\gamma = g.$$

$$\begin{aligned} \bar{w}_k(i, j) &= \alpha \bar{w}_k(i-1, j) + \beta \bar{w}_k(i, j-1) + \kappa r_k(i, j), \\ i &= 1, 2, \dots, N_1 - 1, \quad j = 1, 2, \dots, N_2 - 1, \end{aligned} \tag{32}$$

$$\bar{w}_k(0, j) = 0, \quad 1 \leq j \leq N_2 - 1, \quad \bar{w}_k(i, 0) = 0, \quad 1 \leq i \leq N_1 - 1.$$

Здесь счет ведется, начиная с точки $i = 1, j = 1$ либо по строкам сетки ω , т. е. при возрастании i для фиксированного j , либо по столбцам при возрастании j для фиксированного i .

$$\begin{aligned} w_k(i, j) &= \alpha w_k(i+1, j) + \beta w_k(i, j+1) + \kappa \bar{w}_k(i, j), \\ i &= N_1 - 1, N_1 - 2, \dots, 1, \quad j = N_2 - 1, N_2 - 2, \dots, 1, \quad (33) \\ w_k(N_1, j) &= 0, \quad 1 \leq j \leq N_2 - 1, \quad w_k(i, N_2) = 0, \quad 1 \leq i \leq N_1 - 1. \end{aligned}$$

Здесь счет ведется, начиная с точки $i = N_1 - 1, j = N_2 - 1$, либо по строкам, либо по столбцам сетки при убывании соответствующего индекса i или j . В результате y_{k+1} определяется по формулам

$$\begin{aligned} y_{k+1}(i, j) &= y_k(i, j) - \tau_{k+1} w_k(i, j), \\ 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1, \\ y_{k+1}|_{\gamma} &= g. \end{aligned} \quad (34)$$

Здесь использованы следующие обозначения:

$$\begin{aligned} \alpha &= \frac{\omega_0 h_2^2}{h_1^2 h_2^2 + \omega_0 (h_1^2 + h_2^2)}, \quad \beta = \frac{\omega_0 h_1^2}{h_1^2 h_2^2 + \omega_0 (h_1^2 + h_2^2)}, \\ \kappa &= \frac{h_1^2 h_2^2}{h_1^2 h_2^2 + \omega_0 (h_1^2 + h_2^2)}. \end{aligned} \quad (35)$$

Так как $\alpha, \beta, \kappa > 0$ и $\alpha + \beta + \kappa = 1$, то счет по формулам (32) и (33) устойчив. Элементарный подсчет арифметических операций для алгоритма (31)–(35) дает $Q_+ = 10(N_1 - 1)(N_2 - 1)$ операций сложения и вычитания и $Q_* = Q_+$ операций умножения, а всего $Q = 20(N_1 - 1)(N_2 - 1)$.

Учитывая найденную ранее оценку для числа итераций, получим, что для вычисления решения разностной задачи (27) по алгоритму (31)–(35) с точностью ε следует затратить в случае $N_1 = N_2 = N, l_1 = l_2 = l$

$$Q(\varepsilon) \approx 5,6N^2 \sqrt{N} \ln(2/\varepsilon)$$

арифметических действий.

Рассмотрим теперь первый алгоритм, который имеет в данном случае вид

$$\begin{aligned} \varphi_k &= (E + \omega_0 R_1)(E + \omega_0 R_2)y_k - \tau_{k+1}(Ay_k - f), \\ (E + \omega_0 R_1)v &= \varphi_k, \quad (E + \omega_0 R_2)y_{k+1} = v. \end{aligned} \quad (36)$$

В этом алгоритме при переходе к разностной его записи нам будет удобно работать с сеточными функциями, заданными на ω и обращающимися в нуль на γ . Эти функции совпадают с y_k , v и y_{k+1} на ω и, как обычно, обозначаются \dot{y}_k , \dot{v} и \dot{y}_{k+1} . Чтобы получить приближение к решению задачи (27), его следует определить так: $y_k(x) = \dot{y}_k(x)$ для $x \in \omega$ и $y_k(x) = g(x)$, $x \in \gamma$.

Для того чтобы осуществить в (36) переход к разностной (поточечной) записи, необходимо определить разностный оператор $\bar{\mathcal{R}}$, который соответствует произведению операторов $R_1 R_2$. Заметим, что в силу определения операторы R_1 и R_2 записываются следующим образом:

$$R_1 y = \begin{cases} \frac{1}{h_1} y_{\bar{x}_1} + \frac{1}{h_2} y_{\bar{x}_2}, & 2 \leq i \leq N_1 - 1, 2 \leq j \leq N_2 - 1, \\ \frac{1}{h_1^2} y + \frac{1}{h_2} y_{\bar{x}_2}, & i = 1, 2 \leq j \leq N_2 - 1, \\ \frac{1}{h_1} y_{\bar{x}_1} + \frac{1}{h_2^2} y, & 2 \leq i \leq N_1 - 1, j = 1, \\ \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right) y, & i = j = 1. \end{cases}$$

$$R_2 y = \begin{cases} -\frac{1}{h_1} y_{x_1} - \frac{1}{h_2} y_{x_2}, & 1 \leq i \leq N_1 - 2, 1 \leq j \leq N_2 - 2, \\ \frac{1}{h_1^2} y - \frac{1}{h_2} y_{x_2}, & i = N_1 - 1, 1 \leq j \leq N_2 - 2, \\ -\frac{1}{h_1} y_{x_1} + \frac{1}{h_2^2} y, & 1 \leq i \leq N_1 - 2, j = N_2 - 1, \\ \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right) y, & i = N_1 - 1, j = N_2 - 1. \end{cases}$$

Выкладки показывают, что если оператор $\bar{\mathcal{R}}$ определить следующим образом:

$$\bar{\mathcal{R}}y = \sum_{\alpha=1}^2 \frac{1}{h_\alpha^2} y_{\bar{x}_\alpha x_\alpha} + \frac{1}{h_1 h_2} (y_{x_2 \bar{x}_1} + y_{\bar{x}_1 x_2}) + qy, \quad x \in \omega,$$

где

$$q(i, j) = \begin{cases} 0, & 2 \leq i \leq N_1 - 1, 2 \leq j \leq N_2 - 1, \\ \frac{1}{h_1^4}, & i = 1, 2 \leq j \leq N_2 - 1, \\ \frac{1}{h_2^4}, & 2 \leq i \leq N_1 - 1, j = 1, \\ \frac{1}{h_1^4} + \frac{1}{h_2^4}, & i = 1, j = 1, \end{cases}$$

то $R_1 R_2 y = -\bar{\mathcal{R}}y$, где $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$ для $x \in \omega$.

Используем определение операторов A , R_1 и R_2 , а также полученное выражение для $R_1 R_2$ и запишем алгоритм (36) в виде

$$\begin{aligned} \varphi_k(i, j) = & [d_{k+1} - q(i, j)\omega_0^2] \ddot{y}_k(i, j) + a_{k+1} [\ddot{y}_k(i+1, j) + \\ & + \ddot{y}_k(i-1, j)] + b_{k+1} [\ddot{y}_k(i, j+1) + \dot{y}_k(i, j-1)] + \\ & + c [\ddot{y}_k(i-1, j+1) + \dot{y}_k(i+1, j-1)] + \tau_{k+1} f(i, j), \end{aligned} \quad (37)$$

где обозначено

$$a_{k+1} = \frac{\tau_{k+1}}{h_1^2} - \frac{\omega_0}{h_1^2} \left(1 + \frac{\omega_0}{h_1^2} + \frac{\omega_0}{h_2^2} \right),$$

$$b_{k+1} = \frac{\tau_{k+1}}{h_2^2} - \frac{\omega_0}{h_2^2} \left(1 + \frac{\omega_0}{h_1^2} + \frac{\omega_0}{h_2^2} \right),$$

$$c = \frac{\omega_0^2}{h_1^2 h_2^2}, \quad d_{k+1} = 1 - 2(a_{k+1} + b_{k+1} + c).$$

Далее,

$$v(i, j) = \alpha v(i-1, j) + \beta v(i, j-1) + \kappa \varphi_k(i, j), \\ i = 1, 2, \dots, N_1 - 1, \quad j = 1, 2, \dots, N_2 - 1, \quad (38)$$

$$v(0, j) = 0, \quad 1 \leq j \leq N_2 - 1, \quad v(i, 0) = 0, \quad 1 \leq i \leq N_1 - 1,$$

$$\dot{y}_{k+1}(i, j) = \alpha \dot{y}_{k+1}(i+1, j) + \beta \dot{y}_{k+1}(i, j+1) + \kappa v(i, j), \\ i = N_1 - 1, N_1 - 2, \dots, 1, \quad j = N_2 - 1, N_2 - 2, \dots, 1, \quad (39)$$

$$\dot{y}_{k+1}|_{\gamma} = 0,$$

где α, β и κ определены в (35). Подсчет числа арифметических действий дает $Q_+ = 11N_1N_2 - 10(N_1 + N_2) + 10$ операций сложения и вычитания и $Q_* = Q_+$ операций умножения, а всего $Q = 22N_1N_2 - 20(N_1 + N_2) + 20$. Это примерно в 1,1 раза больше, чем в алгоритме (31)–(34). Преимущество же алгоритма (37)–(39) заключается в том, что здесь не требуется дополнительная память для хранения промежуточной информации $\varphi_k(i, j)$, $v(i, j)$, и вновь определяемое $\dot{y}_{k+1}(i, j)$ располагаются последовательно на месте, которое занимало $y_k(i, j)$.

Замечание 2. В случае p измерений оператор $\bar{\mathcal{R}}$ имеет вид

$$\bar{\mathcal{R}}y = \sum_{\alpha=1}^p \frac{1}{h_{\alpha}^2} y_{x_{\alpha} x_{\alpha}} + \sum_{\alpha=1}^p \sum_{\beta \neq \alpha}^{1+p} \frac{1}{h_{\alpha}^2 h_{\beta}^2} y_{x_{\beta} \bar{x}_{\alpha}} + qy,$$

где

$$q(i_1, i_2, \dots, i_p) = \sum_{\alpha=1}^p \frac{\delta_{i_{\alpha}, 1}}{h_{\alpha}^4}, \quad \delta_{i, j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Замечание 3. Блочному попеременно-треугольному методу соответствует следующее определение разностных операторов \mathcal{R}_1 и \mathcal{R}_2 :

$$\mathcal{R}_1 y = -\frac{1}{2} y_{\bar{x}_1 x_1} - \frac{1}{h_2} y_{\bar{x}_2}, \quad \mathcal{R}_2 y = -\frac{1}{2} y_{x_1 \bar{x}_1} + \frac{1}{h_2} y_{x_2}.$$

В этом случае для обращения оператора B необходимо использовать метод трехточечной прогонки. Это приводит к увеличению вычислительной работы на одной итерации, что не компенсируется небольшим уменьшением числа итераций (примерно в 1,2 раза).

§ 2. Разностные краевые задачи для эллиптических уравнений в прямоугольнике

1. Задача Дирихле для уравнения с переменными коэффициентами. Рассмотрим теперь применение попеременно-треугольного метода к нахождению решения разностной задачи Дирихле для эллиптического уравнения без смешанных производных

$$\begin{aligned} \Lambda y &= \sum_{\alpha=1}^2 (a_\alpha(x) y_{x_\alpha})_{x_\alpha} = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \quad (1)$$

в прямоугольнике, где $\bar{\omega} = \omega \cup \gamma$ — прямоугольная равномерная сетка с шагами h_1 и h_2 : $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$. Будем предполагать, что коэффициенты $a_\alpha(x)$ удовлетворяют условиям

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad \alpha = 1, 2. \quad (2)$$

Потребуем также, чтобы при фиксированном j , $1 \leq j \leq N_2 - 1$, число узлов сетки ω , в которых $(a_1)_{x_1} = O(h_1^{-1})$, было конечным и не зависело от h_1 . Это означает, что соответствующий коэффициент в дифференциальном уравнении при каждом фиксированном x_2 имеет конечное число точек разрыва по направлению x_1 . Аналогичное требование должно быть выполнено и для $(a_2)_{x_2}$.

Разностная задача (1) сводится к операторному уравнению

$$Au = f \quad (3)$$

обычным образом. Здесь H — пространство сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2,$$

$Ay = -\Lambda \dot{y}$, $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \dot{y}(x)$ для $x \in \omega$;

$$f(x) = \varphi(x) + \frac{1}{h_1^2} \varphi_1(x) + \frac{1}{h_2^2} \varphi_2(x),$$

где

$$\varphi_1(x) = \begin{cases} a_1(h_1, x_2)g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leq x_1 \leq l_1 - 2h_1, \\ a_1(l_1, x_2)g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} a_2(x_1, h_2)g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leq x_2 \leq l_2 - 2h_2, \\ a_2(x_1, l_2)g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

Используя разностные формулы Грина, найдем, что оператор A самосопряжен в H и имеет место равенство

$$(Ay, y) = -(\Lambda \dot{y}, \dot{y}) = \sum_{\alpha=1}^2 \left(a_\alpha \dot{y}_{x_\alpha}^2, 1 \right)_\alpha, \quad (4)$$

где

$$(u, v)_\alpha = \sum_{x_\alpha = h_\alpha}^{l_\alpha} \sum_{x_\beta = h_\beta}^{l_\beta - h_\beta} u(x)v(x)h_\alpha h_\beta, \quad \beta = 3 - \alpha, \alpha = 1, 2.$$

Для приближенного решения уравнения рассмотрим попеременно-треугольный метод, построенный на использовании регуляризатора $R \neq A$:

$$\begin{aligned} B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k &= f, \quad k = 0, 1, \dots, y_0 \in H, \\ B &= (E + \omega R_1)(E + \omega R_2), \quad R_1 = R_2^*, \quad R = R_1 + R_2. \end{aligned} \quad (5)$$

Регуляризатор R выберем следующим образом:

$$Ry = -\mathcal{R}\dot{y}, \quad \mathcal{R}y = y_{x_1 x_1} - y_{x_2 x_2}, \quad \dot{y} \in \dot{H}, \quad (6)$$

а операторы R_1 и R_2 определим по формулам

$$R_\alpha y = -\mathcal{R}_\alpha \dot{y}, \quad \mathcal{R}_1 y = -\sum_{\alpha=1}^2 \frac{1}{h_\alpha} y_{x_\alpha}, \quad \mathcal{R}_2 y = \sum_{\alpha=1}^2 \frac{1}{h_\alpha} y_{x_\alpha}. \quad (7)$$

В п. 4 § 1 было показано, что для определенных здесь операторов R_1 и R_2 имеют место неравенства

$$\delta E \leq R, \quad R_1 R_2 \leq (\Delta/4)R,$$

$$\delta = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2}.$$

Далее, используя разностные формулы Грина, получим

$$(Ry, y) = -(\mathcal{R}\dot{y}, \dot{y}) = \sum_{\alpha=1}^2 \left(\dot{y}_{x_\alpha}^2, 1 \right)_\alpha. \quad (8)$$

Следовательно, из (2), (4) и (8) вытекают неравенства $c_1 R \leq A \leq c_2 R$, $c_1 > 0$. Так как оператор R самосопряжен и положительно определен в H , то для рассматриваемого метода имеет место теорема 2 с $\mathcal{D} = E$, в которой указан выбор итерационных параметров ω и $\{\tau_k\}$. Из этой теоремы получим оценку для погрешности

$$\|z_n\|_D \leq q_n \|z_0\|_D, \quad D = A, B \text{ или } AB^{-1}A,$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{c_1}{c_2} \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}.$$

При малом η получим оценку для числа итераций:

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \sqrt{\frac{c_2}{c_1}} \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}}, \quad \eta = \frac{\pi^2}{4N^2}.$$

Отсюда следует, что число итераций пропорционально $\sqrt{c_2/c_1}$ и методом (5), (7) целесообразно пользоваться, когда это отношение не слишком велико.

2. Модифицированный попеременно-треугольный метод*). Продолжим изучение попеременно-треугольного метода для разностной задачи (1) в случае, когда коэффициенты $a_\alpha(x)$ сильно меняются, т. е. отношение c_2/c_1 велико.

Рассмотрим теперь для уравнения (3) модифицированный вариант попеременно-треугольного метода

$$\begin{aligned} B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k &= f, \quad k = 0, 1, \dots, \\ B = (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2), \quad R_1 &= R_2^*, \quad R_1 + R_2 = A, \end{aligned} \quad (9)$$

где положим $\mathcal{D}y = d(x)y$, $x \in \omega$. Здесь $d(x)$ — некоторая положительная на ω сеточная функция, подлежащая определению. В этом случае \mathcal{D} — самосопряженный и положительно определенный в H оператор. Сеточная функция $d(x)$ играет в (9) роль дополнительного итерационного параметра и позволяет учесть особенности оператора A в каждом узле x сетки ω .

Определим теперь операторы R_α следующим образом: $R_\alpha y = -\mathcal{R}_\alpha y$, $y \in H$ и $y \in \dot{H}$, где

$$\begin{aligned} \mathcal{R}_1 y &= -\sum_{\alpha=1}^2 \left(\frac{a_\alpha}{h_\alpha} y_{x_\alpha} + \frac{a_{\alpha x_\alpha}}{2h_\alpha} y \right), \\ \mathcal{R}_2 y &= \sum_{\alpha=1}^2 \left(\frac{a_\alpha^{+1}}{h_\alpha} y_{x_\alpha} + \frac{a_{\alpha x_\alpha}}{2h_\alpha} y \right), \quad x \in \omega, \end{aligned} \quad (10)$$

и $a_1^{\pm 1}(x) = a_1(x_1 \pm h_1, x_2)$, $a_2^{\pm 1}(x) = a_2(x_1, x_2 \pm h_2)$.

Покажем, что операторы R_1 и R_2 сопряжены в H . Для этого достаточно показать, что имеет место равенство $(\mathcal{R}_1 \dot{y}, \dot{v}) = (\dot{y}, \mathcal{R}_2 \dot{v})$, $\dot{y} \in \dot{H}$, $\dot{v} \in \dot{H}$. Из разностных формул Грина для функций, обращающихся в нуль на γ , и формулы разностного дифференци-

*) См. А. Б. Кучеров и Е. С. Николаев (ЖВМ и МФ, **16**, № 5, 1976; **17**, № 3, 1977).

рования произведения сеточных функций $(yv)_{x_\alpha} = y^{+1}v_{x_\alpha} + y_{x_\alpha}v$ следует, что

$$\begin{aligned} (\mathcal{R}_1 \dot{y}, \dot{v}) &= - \sum_{\alpha=1}^2 \frac{1}{h_\alpha} (a_\alpha \dot{y}_{x_\alpha}, \dot{v}) - \sum_{\alpha=1}^2 \frac{1}{2h_\alpha} (a_{\alpha x_\alpha} \dot{y}, \dot{v}) = \\ &= \sum_{\alpha=1}^2 \left[\frac{1}{h_\alpha} (\dot{y}, (a_\alpha \dot{v})_{x_\alpha}) - \frac{1}{2h_\alpha} (a_{\alpha x_\alpha} \dot{y}, \dot{v}) \right] = \\ &= \sum_{\alpha=1}^2 \left[\frac{1}{h_\alpha} (\dot{y}, a_\alpha^{+1} \dot{v}_{x_\alpha}) + \frac{1}{2h_\alpha} (\dot{y}, a_{\alpha x_\alpha} \dot{v}) \right] = (\dot{y}, \mathcal{R}_2 \dot{v}). \end{aligned}$$

Утверждение доказано.

Так как $R_1 + R_2 = A$, то в силу теоремы 1 априорная информация для попеременно-треугольного метода (9) имеет вид постоянных δ и Δ из неравенств

$$\delta \mathcal{D} \leq A, \quad R_1 \mathcal{D}^{-1} R_2 \leq \frac{\Delta}{4} A, \quad \delta > 0. \quad (11)$$

Так как отношение $\eta = \delta/\Delta$ определяет число итераций, то сеточная функция $d(x)$ должна быть выбрана из условия максимальности этого отношения.

Займемся теперь выбором функции $d(x)$ и оценками δ и Δ . Докажем сначала одно неравенство.

Лемма 2. Пусть $p_\alpha(x)$, $q_\alpha(x)$, $u_\alpha(x)$ и $v_\alpha(x)$, $\alpha = 1, 2$ — сеточные функции, заданные на ω . Тогда для любого $x \in \omega$ имеет место неравенство

$$\begin{aligned} \left[\sum_{\alpha=1}^2 (p_\alpha u_\alpha + q_\alpha v_\alpha) \right]^2 &\leq \\ &\leq (1 + \varepsilon) (|p_1| + \kappa_1 |q_1|) \left(|p_1| u_1^2 + \frac{|q_1|}{\kappa_1} v_1^2 \right) + \\ &\quad + \frac{1 + \varepsilon}{\varepsilon} (|p_2| + \kappa_2 |q_2|) \left(|p_2| u_2^2 + \frac{|q_2|}{\kappa_2} v_2^2 \right), \quad (12) \end{aligned}$$

где $\varepsilon(x)$, $\kappa_1(x)$ и $\kappa_2(x)$ — произвольные положительные на ω сеточные функции.

Действительно, используя ε — неравенство $2ab \leq \varepsilon a^2 + b^2/\varepsilon$, $\varepsilon > 0$, получим

$$\begin{aligned} \left[\sum_{\alpha=1}^2 (p_\alpha u_\alpha + q_\alpha v_\alpha) \right]^2 &= \\ &= (p_1 u_1 + q_1 v_1)^2 + 2(p_1 u_1 + q_1 v_1)(p_2 u_2 + q_2 v_2) + (p_2 u_2 + q_2 v_2)^2 \leq \\ &\leq (1 + \varepsilon) (p_1 u_1 + q_1 v_1)^2 + \frac{1 + \varepsilon}{\varepsilon} (p_2 u_2 + q_2 v_2)^2. \quad (13) \end{aligned}$$

Снова пользуясь указанным неравенством, найдем

$$\begin{aligned} (p_\alpha u_\alpha + q_\alpha v_\alpha)^2 &= p_\alpha^2 u_\alpha^2 + 2p_\alpha q_\alpha u_\alpha v_\alpha + q_\alpha^2 v_\alpha^2 \leq \\ &\leq p_\alpha^2 u_\alpha^2 + |p_\alpha| |q_\alpha| \left(u_\alpha u_\alpha^2 + \frac{1}{\kappa_\alpha} v_\alpha^2 \right) + q_\alpha^2 v_\alpha^2 = \\ &= (|p_\alpha| + \kappa_\alpha |q_\alpha|) \left(|p_\alpha| u_\alpha^2 + \frac{|q_\alpha|}{\kappa_\alpha} v_\alpha^2 \right), \quad \kappa_\alpha > 0, \alpha = 1, 2. \end{aligned}$$

Подставляя полученное неравенство в (13), будем иметь (12). Лемма доказана.

Воспользуемся неравенством (12), а также определением операторов R_1 и R_2 и найдем, что

$$\begin{aligned} (R_1 \mathcal{D}^{-1} R_2 y, y) &= (\mathcal{D}^{-1} R_2 \dot{y}, R_2 \dot{y}) = \\ &= \left(\frac{1}{d} \sum_{\alpha=1}^2 \left(\frac{a_\alpha^{+1}}{h_\alpha} \dot{y}_{x_\alpha} + \frac{a_{\alpha x_\alpha}}{2h_\alpha} \ddot{y} \right)^2, 1 \right) \leq \\ &\leq \left(\frac{(1+\varepsilon)}{dh_1^2} (a_1^{+1} + 0,5h_1 \kappa_1 |a_{1x_1}|) \left(a_1^{+1} \dot{y}_{x_1}^2 + \frac{0,5h_1 |a_{1x_1}|}{\kappa_1 h_1^2} \dot{y}^2 \right), 1 \right) + \\ &\quad + \left(\frac{(1+\varepsilon)}{dh_2^2} (a_2^{+1} + 0,5h_2 \kappa_2 |a_{2x_2}|) \left(a_2^{+1} \dot{y}_{x_2}^2 + \frac{0,5h_2 |a_{2x_2}|}{\kappa_2 h_2^2} \dot{y}^2 \right), 1 \right). \end{aligned}$$

Отметим, что в (12) вместо p_α , q_α , u_α и v_α мы подставили $p_\alpha = \frac{a_\alpha^{+1}}{h_\alpha}$, $q_\alpha = 0,5a_{\alpha x_\alpha}$, $u_\alpha = \dot{y}_{x_\alpha}$, $v_\alpha = \frac{1}{h_\alpha} \dot{y}$, $\alpha = 1, 2$. Будем требовать, чтобы в полученном неравенстве κ_1 была функцией только x_2 , а κ_2 — только x_1 , т. е. положим

$$\kappa_\alpha = \kappa_\alpha(x_\beta), \quad \beta = 3 - \alpha, \quad \alpha = 1, 2. \quad (14)$$

Положим

$$\varepsilon = \varepsilon(x) = \frac{a_2^{+1} + 0,5h_2 \kappa_2 |a_{2x_2}|}{a_1^{+1} + 0,5h_1 \kappa_1 |a_{1x_1}|} \cdot \frac{h_1^2 \theta_2(x_1)}{h_2^2 \theta_1(x_2)} \quad (15)$$

и определим $d(x)$ следующим образом:

$$d(x) = \sum_{\alpha=1}^2 (a_\alpha^{+1} + 0,5h_\alpha \kappa_\alpha |a_{\alpha x_\alpha}|) \frac{\theta_\alpha}{h_\alpha^2}, \quad (16)$$

где $\theta_\alpha = \theta_\alpha(x_\beta)$, $\beta = 3 - \alpha$, $\alpha = 1, 2$, — положительные на ω сеточные функции, подлежащие определению.

Подставляя (15) и (16) в полученное ранее неравенство, будем иметь

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leq \sum_{\alpha=1}^2 \left(\frac{a_\alpha^{+1}}{\theta_\alpha} \dot{y}_{x_\alpha}^2, 1 \right) + \sum_{\alpha=1}^2 \left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha \theta_\alpha \kappa_\alpha} \dot{y}^2, 1 \right).$$

Так как θ_α не зависит от x_α , то, используя введенное ранее скалярное произведение $(\cdot, \cdot)_\alpha$, получим, что

$$\left(\frac{a_\alpha^{+1}}{\theta_\alpha} \dot{y}_{x_\alpha}^2, 1 \right) \leq \left(\frac{a_\alpha}{\theta_\alpha} \dot{y}_{x_\alpha}^2, 1 \right)_\alpha, \quad \alpha = 1, 2.$$

Следовательно,

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leq \sum_{\alpha=1}^2 \left(\frac{a_\alpha}{\theta_\alpha} \dot{y}_{x_\alpha}^2, 1 \right)_\alpha + \sum_{\alpha=1}^2 \left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha \theta_\alpha \kappa_\alpha} \dot{y}^2, 1 \right). \quad (17)$$

Выберем теперь θ_α и κ_α . Обозначим

$$\begin{aligned} \omega_1 &= \{x_1 = ih_1, 1 \leq i \leq N_1 - 1, h_1 N_1 = l_1\}, \\ \omega_1^+ &= \{x_1 = ih_1, 1 \leq i \leq N_1, h_1 N_1 = l_1\} \end{aligned}$$

и определим

$$\begin{aligned} (u, v)_{\omega_1} &= \sum_{x_1 \in \omega_1} u(x)v(x)h_1, \\ (u, v)_{\omega_1^+} &= \sum_{x_1 \in \omega_1^+} u(x)v(x)h_1. \end{aligned}$$

Аналогично вводятся ω_2 и ω_2^+ , а также $(u, v)_{\omega_2}$ и $(u, v)_{\omega_2^+}$. Тогда легко видеть, что имеют место соотношения

$$\begin{aligned} (u, v) &= ((u, v)_{\omega_1}, 1)_{\omega_2} = ((u, v)_{\omega_2}, 1)_{\omega_1}, \\ (u, v)_\alpha &= ((u, v)_{\omega_\alpha^+}, 1)_{\omega_\beta}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2. \end{aligned} \quad (18)$$

Пусть теперь $b_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x)$, $\alpha = 1, 2$, $x_\beta \in \omega_\beta$, где $v^\alpha(x)$ для фиксированного x_β есть решение следующей трехточечной краевой задачи:

$$\begin{aligned} (a_\alpha v_{x_\alpha}^\alpha)_{x_\alpha} &= -\frac{a_\alpha^{+1}}{h_\alpha^2}, \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ v^\alpha(x) &= 0, \quad x_\alpha = 0, \quad l_\alpha, \quad x_\beta \in \omega_\beta. \end{aligned} \quad (19)$$

Тогда в силу леммы 13 из п. 4 § 2 гл. V получим

$$\left(\frac{a_\alpha^{+1}}{h_\alpha^2} \dot{y}^2, 1 \right)_{\omega_\alpha} \leq b_\alpha(x_\beta) (a_\alpha \dot{y}_{x_\alpha}^2, 1)_{\omega_\alpha^+}, \quad \alpha = 1, 2.$$

Умножим это неравенство на $\theta_\alpha(x_\beta)$ и просуммируем скалярно по ω_β . Тогда в силу (18) будем иметь

$$\left(\frac{\theta_\alpha a_\alpha^{+1}}{h_\alpha^2} \dot{y}^2, 1 \right) \leq (b_\alpha a_\alpha \theta_\alpha \dot{y}_{x_\alpha}^2, 1)_\alpha, \quad \alpha = 1, 2. \quad (20)$$

Пусть $c_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} w^\alpha(x)$, $\alpha = 1, 2$, $x_\beta \in \omega_\beta$, где $w^\alpha(x)$ для фиксированного x_β есть решение следующей трехточечной краевой

задачи:

$$(a_\alpha w_{x_\alpha}^\alpha)_{x_\alpha} = -\frac{|a_{\alpha x_\alpha}|}{2h_\alpha}, \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ w^\alpha(x) = 0, \quad x_\alpha = 0, \quad l_\alpha, \quad x_\beta \in \omega_\beta. \quad (21)$$

Аналогично тому, как были получены неравенства (20), в силу (14) будем иметь следующие неравенства:

$$\left(\frac{\kappa_\alpha \theta_\alpha |a_{\alpha x_\alpha}|}{2h_\alpha} \dot{y}^2, 1 \right) \leq (\kappa_\alpha \theta_\alpha c_\alpha a_\alpha \dot{y}_{x_\alpha}^2, 1)_\alpha, \quad \alpha = 1, 2, \quad (22)$$

$$\left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha \theta_\alpha \kappa_\alpha} \dot{y}^2, 1 \right) \leq \left(\frac{c_\alpha}{\kappa_\alpha \theta_\alpha} a_\alpha \dot{y}_{x_\alpha}^2, 1 \right)_\alpha, \quad \alpha = 1, 2. \quad (23)$$

Сложим теперь неравенства (20) и (22) и просуммируем их по α . Тогда в силу (16) получим

$$(d\dot{y}^2, 1) = (\mathcal{D}y, y) \leq ((\kappa_\alpha c_\alpha + b_\alpha) \theta_\alpha a_\alpha \dot{y}_{x_\alpha}^2, 1)_\alpha.$$

Выбирая θ_α по формуле

$$\theta_\alpha(x_\beta) = \frac{1}{b_\alpha(x_\beta) + c_\alpha(x_\beta) \kappa_\alpha(x_\beta)}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2, \quad (24)$$

и учитывая (4), найдем отсюда, что $(\mathcal{D}y, y) \leq (Ay, y)$. Следовательно, в (11) можно положить $\delta = 1$.

Оценим теперь Δ . Для этого подставим (23) в (17) и учтем выбор θ_α по формуле (24). В результате получим следующую оценку:

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leq \sum_{\alpha=1}^2 ((1 + c_\alpha/\kappa_\alpha) (b_\alpha + c_\alpha \kappa_\alpha) a_\alpha \dot{y}_{x_\alpha}^2, 1)_\alpha.$$

Выберем теперь оптимальное κ_α из условия минимума выражения $(1 + c_\alpha/\kappa_\alpha) (b_\alpha + c_\alpha \kappa_\alpha)$ по κ_α . Получим $\kappa_\alpha(x_\beta) = \sqrt{b_\alpha(x_\beta)}$, $\beta = 3 - \alpha$, $\alpha = 1, 2$, и при этом

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leq \sum_{\alpha=1}^2 ((c_\alpha + \sqrt{b_\alpha})^2 a_\alpha \dot{y}_{x_\alpha}^2, 1).$$

Сравнивая эту оценку с (4), найдем, что в неравенствах (11) можно положить

$$\Delta = 4 \max_{\alpha=1, 2} \left(\max_{x_\beta \in \omega_\beta} (c_\alpha(x_\beta) + \sqrt{b_\alpha(x_\beta)})^2 \right), \quad \beta = 3 - \alpha. \quad (25)$$

Подставляя в (16) найденные выражения для κ_α и θ_α , получим для функции $d(x)$ представление вида

$$d(x) = \sum_{\alpha=1}^2 \left(\frac{a_\alpha^{+1}}{h_\alpha^2 \sqrt{b_\alpha}} + \frac{|a_{\alpha x_\alpha}|}{2h_\alpha} \right) \frac{1}{c_\alpha + \sqrt{b_\alpha}}, \quad x \in \omega. \quad (26)$$

Итак, функция $d(x)$ и постоянные δ и Δ найдены. Теперь остается применить теорему 1. Отметим, что так как $\delta=1$, то $\omega_0=2/\sqrt{\Delta}$, и для числа итераций верна оценка

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\sqrt[4]{\Delta} \ln(2/\varepsilon)}{2 \sqrt[4]{2}}.$$

Далее, в силу условий (2) из (19) и (21) получим, что $b_\alpha = O(1/h_\alpha^2)$ и $c_\alpha = O(1/h_\alpha)$, если число точек, в которых $a_{\alpha x_\alpha} = O(h_\alpha^{-1})$, конечно. Отсюда следует, что $n_0(\varepsilon) = O(\sqrt{N} \ln(2/\varepsilon))$.

Остановимся теперь на реализации построенного варианта попаременно-треугольного метода (9). Сначала для фиксированного x_β , $h_\beta \leq x_\beta \leq l_\beta - h_\beta$ решаются методом прогонки трехточечные краевые задачи (19) и (21) и находятся значения $b_\alpha(x_\beta)$ и $c_\alpha(x_\beta)$, $\alpha = 1, 2$. Эти четыре одномерные сеточные функции запоминаются и используются в процессе итераций для вычисления $d(x)$ по формуле (26). Простота формулы (26) позволяет не хранить двумерную сеточную функцию $d(x)$, а вычислять ее по мере необходимости заново.

Далее по формуле (25) находится Δ и полагается $\delta=1$. Значения итерационных параметров ω и τ_k для схемы (9) определяются согласно теореме 1.

Для нахождения y_{k+1} по заданному y_k используется первый из алгоритмов, описанных для попаременно-треугольного метода в п. 1 § 1:

$$\begin{aligned} (\mathcal{D} + \omega_0 R_1) v &= \varphi_k, & (\mathcal{D} + \omega_0 R_2) y_{k+1} &= \mathcal{D} v, \\ \varphi_k &= (\mathcal{D} + \omega_0 R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega_0 R_2) y_k - \tau_{k+1} (A y_k - f). \end{aligned} \quad (27)$$

Не останавливаясь на деталях, приведем разностный вид алгоритма (27):

$$v(i, j) = \alpha_1(i, j) v(i-1, j) + \beta_1(i, j) v(i, j-1) + \kappa(i, j) \varphi_k(i, j), \quad i = 1, 2, \dots, N_1-1, \quad j = 1, 2, \dots, N_2-1, \quad (28)$$

$$v(0, j) = 0, \quad 1 \leq j \leq N_2-1, \quad v(i, 0) = 0, \quad 1 \leq i \leq N_1-1.$$

$$\begin{aligned} \dot{y}_{k+1}(i, j) &= \alpha_2(i, j) \dot{y}_{k+1}(i+1, j) + \beta_2(i, j) \dot{y}_{k+1}(i, j+1) + \\ &+ \kappa(i, j) d(i, j) v(i, j), \quad i = N_1-1, \dots, 1, \quad j = N_2-1, \dots, 1, \quad (29) \end{aligned}$$

где

$$\begin{aligned} \alpha_1 &= \frac{\omega_0 a_1 \kappa}{h_1^2}, & \beta_1 &= \frac{\omega_0 a_2 \kappa}{h_2^2}, & \alpha_2 &= \frac{\omega_0 a_1^{+1} \kappa}{h_1^2}, & \beta_2 &= \frac{\omega_0 a_2^{+1} \kappa}{h_2^2}, \\ \frac{1}{\kappa} &= d + \omega_0 \left[\frac{a_1^{+1} + a_1}{2h_1^2} + \frac{a_2^{+1} + a_2}{2h_2^2} \right]. \end{aligned}$$

Правая часть $\varphi_k(i, j)$ вычисляется по формулам

$$\begin{aligned}\varphi_k(i, j) = & [P(i-1, j) + Q(i, j-1) + S(i, j)] \dot{y}_k(i, j) + \\ & + R_1(i, j) \dot{y}_k(i+1, j) + R_1(i-1, j) \dot{y}_k(i-1, j) + \\ & + R_2(i, j) \dot{y}_k(i, j+1) + R_2(i, j-1) \dot{y}_k(i, j-1) + \\ & + G(i-1, j) \dot{y}_k(i-1, j+1) + G(i, j-1) \dot{y}_k(i+1, j-1) + \\ & + \tau_{k+1} f(i, j), \\ 1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1,\end{aligned}\quad (30)$$

где

$$G = \frac{\omega_0^2 a_1^{+1} a_2^{+1}}{h_1^2 h_2^2 d}, \quad R_\alpha = \left(\tau_{k+1} - \frac{\omega_0}{dx} \right) \frac{a_\alpha^{+1}}{h_\alpha^2}, \quad \alpha = 1, 2,$$

$$S = \frac{1}{\omega_0 \kappa} \left[\frac{\omega_0}{dx} - 2\tau_{k+1}(1-\kappa d) \right], \quad P = \frac{\omega_0^2 (a_1^{+1})^2}{h_1^4 d}, \quad Q = \frac{\omega_0^2 (a_2^{+1})^2}{h_2^4 d},$$

причем $P(0, j) = 0, 1 \leq j \leq N_2 - 1, Q(i, 0) = 0, 1 \leq i \leq N_1 - 1$.

Заметим, что в силу (25) и (26) верны оценки

$$c_\alpha + V \bar{b}_\alpha \leq \frac{V \bar{\Delta}}{2} = \frac{1}{\omega_0}, \quad d \geq \omega_0 \sum_{\alpha=1}^2 \frac{|a_{\alpha x_\alpha}|}{2h_\alpha}.$$

Отсюда получим

$$\begin{aligned}\frac{1}{\kappa} = d + \omega_0 \sum_{\alpha=1}^2 \frac{a_\alpha^{+1} + a_\alpha}{2h_\alpha^2} \geq \omega_0 \sum_{\alpha=1}^2 \left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha} + \frac{a_\alpha^{+1} + a_\alpha}{2h_\alpha^2} \right) = \\ = \omega_0 \left(\max \left(\frac{a_1^{+1}}{h_1^2}, \frac{a_1}{h_1^2} \right) + \max \left(\frac{a_2^{+1}}{h_2^2}, \frac{a_2}{h_2^2} \right) \right)\end{aligned}$$

или

$$\max(\alpha_1, \alpha_2) + \max(\beta_1, \beta_2) \leq 1.$$

Отсюда следует, что $\alpha_1 + \beta_1 \leq 1$ и $\alpha_2 + \beta_2 \leq 1$. Поэтому счет по формулам (28), (30) устойчив.

3. Сравнение вариантов метода. Выше для решения разностной задачи (1) были построены два варианта попеременно-треугольного метода. Вариант (5), (7) построен на основе регуляризатора R , а вариант (9), (10) использует оператор \mathcal{D} , который выбирается специальным образом. Эти варианты характеризуются одной и той же асимптотической зависимостью числа итераций от числа узлов сетки. Однако оценка числа итераций для первого варианта зависит от экстремальных характеристик коэффициентов $a_\alpha(x), \alpha = 1, 2$, разностного уравнения (1), тогда как для второго варианта она определяется их интегральными характеристиками.

Сравним эти варианты метода на следующем традиционном модельном примере. Пусть на квадратной сетке с $N_1 = N_2 = N$,

введенной в единичном квадрате ($l_1 = l_2 = 1$), задано разностное уравнение (1), в котором

$$\begin{aligned} a_1(x) &= 1 + c [(x_1 - 0,5)^2 + (x_2 - 0,5)^2], \\ a_2(x) &= 1 + c [0,5 - (x_1 - 0,5)^2 - (x_2 - 0,5)^2], \quad x \in \bar{\omega}. \end{aligned}$$

Тогда в неравенствах (2) имеем $c_1 = 1$, $c_2 = 1 + 0,5c$. Меняя параметр c , будем получать коэффициенты $a_\alpha(x)$ с различными экстремальными свойствами.

Т а б л и ц а 10

c_2/c_1	$N=32$		$N=64$		$N=128$	
	(5), (7)	(9), (10)	(5), (7)	(9), (10)	(5), (7)	(9), (10)
2	23	18	32	26	45	36
8	46	21	64	30	90	43
32	92	23	128	34	180	49
128	184	24	256	36	360	53
512	367	24	512	36	720	54

В табл. 10 приведено число итераций для указанных вариантов в зависимости от числа узлов N по одному направлению и от отношения c_2/c_1 для $\varepsilon = 10^{-4}$. Видно, что для случая больших значений c_2/c_1 модифицированный попеременно-треугольный метод требует меньшего числа итераций, причем число итераций слабо зависит от этого отношения.

4. Третья краевая задача. Рассмотрим попеременно-треугольный метод решения третьей краевой задачи для эллиптического уравнения в прямоугольнике $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$:

$$\begin{aligned} \sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x) \frac{\partial u}{\partial x_\alpha} \right) &= -\varphi(x), \quad x \in G, \\ k_\alpha \frac{\partial u}{\partial x_\alpha} &= \kappa_{-\alpha}(x) u - g_{-\alpha}(x), \quad x_\alpha = 0, \\ -k_\alpha \frac{\partial u}{\partial x_\alpha} &= \kappa_{+\alpha}(x) u - g_{+\alpha}(x), \quad x_\alpha = l_\alpha, \end{aligned} \quad (31)$$

На прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ задаче (31) соответствует разностная задача

$$\begin{aligned} \Lambda y &= -f(x), \quad x \in \bar{\omega}, \\ \Lambda = \Lambda_1 + \Lambda_2, \quad f(x) &= \varphi(x) + \frac{2}{h_1} \varphi_1(x) + \frac{2}{h_2} \varphi_2(x), \end{aligned} \quad (32)$$

где

$$\Lambda_\alpha y = \begin{cases} \frac{2}{h_\alpha} (a_\alpha^{+1} y_{x_\alpha} - \kappa_{-\alpha} y), & x_\alpha = 0, \\ (a_\alpha y_{x_\alpha})_{x_\alpha}, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} (-a_\alpha y_{x_\alpha} - \kappa_{+\alpha} y), & x_\alpha = l_\alpha, \end{cases}$$

$$\Phi_\alpha(x) = \begin{cases} g_{-\alpha}(x_\alpha), & x_\alpha = 0, \\ 0, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ g_{+\alpha}(x_\alpha), & x_\alpha = l_\alpha. \end{cases}$$

Будем предполагать, что коэффициенты $a_\alpha(x)$ удовлетворяют условиям (2) и имеют конечное число точек, в которых $a_{\alpha x_\alpha} = O(h_\alpha^{-1})$. Также будем считать, что $\kappa_{-\alpha}(x_\beta)$ и $\kappa_{+\alpha}(x_\beta)$ для каждого фиксированного x_β одновременно в нуль не обращаются ($\kappa_{-\alpha} \geq 0$, $\kappa_{+\alpha} \geq 0$, $\kappa_{-\alpha} + \kappa_{+\alpha} > 0$).

Разностную задачу (32) удобно предварительно свести к задаче Дирихле в расширенной области $\bar{\omega}^* = \{x_{ij} = (ih_1, jh_2), -1 \leq i \leq N_1 + 1, -1 \leq j \leq N_2 + 1\}$, для которой сетка $\bar{\omega}$ является внутренней. Обозначим через γ^* границу сетки $\bar{\omega}^*$ и определим сеточную функцию $y(x)$ нулем на γ^* . Если обозначить

$$\bar{a}_\alpha(x) = \begin{cases} \rho(x_\beta) h_\alpha \kappa_{-\alpha}(x_\beta), & x_\alpha = 0, \\ \rho(x_\beta) a_\alpha(x), & h_\alpha \leq x_\alpha \leq l_\alpha, \\ \rho(x_\beta) h_\alpha \kappa_{+\alpha}(x_\beta), & x_\alpha = l_\alpha + h_\alpha, 0 \leq x_\beta \leq l_\beta, \end{cases}$$

$$\bar{f}(x) = \rho(x_1) \rho(x_2) f(x), \quad x \in \bar{\omega},$$

$$\rho(x_\beta) = \begin{cases} 0,5, & x_\beta = 0, l_\beta, \\ 1, & h_\beta \leq x_\beta \leq l_\beta - h_\beta, \end{cases}$$

$$\beta = 3 - \alpha, \quad \alpha = 1, 2,$$

то задача (32) может быть записана в виде

$$\bar{\Lambda}y = \sum_{\alpha=1}^2 (\bar{a}_\alpha y_{x_\alpha})_{x_\alpha} = -\bar{f}(x), \quad x \in \bar{\omega}, \quad (33)$$

$$y(x) = 0, \quad x \in \gamma^*.$$

Напомним, что для разностной задачи вида (33) в п. 2 был построен модифицированный попеременно-треугольный метод (9)–(10). Следовательно, в формулах п. 2 необходимо лишь заменить $a_\alpha(x)$ на $\bar{a}_\alpha(x)$, и мы получим метод решения третьей краевой задачи для эллиптического уравнения в прямоугольнике.

Для рассматриваемого случая трехточечная краевая задача (19) записывается в виде

$$\begin{aligned} (\bar{a}_\alpha v_{\bar{x}_\alpha}^\alpha)_{x_\alpha} &= -\frac{\bar{a}_\alpha^{+1}}{h_\alpha^2}, \quad 0 \leq x_\alpha \leq l_\alpha, \\ v^\alpha(x) &= 0, \quad x_\alpha = -h_\alpha, \quad l_\alpha + h_\alpha. \end{aligned} \quad (34)$$

Используя введенные выше обозначения для \bar{a}_α , получим, что (34) можно придать другой вид

$$\begin{aligned} (a_\alpha v_{\bar{x}_\alpha}^\alpha)_{x_\alpha} &= -\frac{a_\alpha^{+1}}{h_\alpha^2}, \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ a_\alpha^{+1} v_{\bar{x}_\alpha}^\alpha - \kappa_{-\alpha} v^\alpha &= -\frac{a_\alpha^{+1}}{h_\alpha}, \quad x_\alpha = 0, \\ -a_\alpha v_{\bar{x}_\alpha}^\alpha - \kappa_{+\alpha} v^\alpha &= -\kappa_{+\alpha}, \quad x_\alpha = l_\alpha. \end{aligned} \quad (35)$$

В силу сделанных предположений относительно $a_\alpha(x)$, $\kappa_{-\alpha}$ и $\kappa_{+\alpha}$, разностная задача разрешима. При этом

$$b_\alpha(x_\beta) = \max_{0 \leq x_\alpha \leq l_\alpha} v^\alpha(x) = O\left(\frac{1}{h_\alpha^2}\right), \quad 0 \leq x_\beta \leq l_\beta.$$

Аналогично задача (21), которая в данном случае имеет вид

$$\begin{aligned} (\bar{a}_\alpha w_{\bar{x}_\alpha}^\alpha)_{x_\alpha} &= -\frac{|\bar{a}_{\alpha x_\alpha}|}{2h_\alpha}, \quad 0 \leq x_\alpha \leq l_\alpha, \\ w^\alpha(x) &= 0, \quad x_\alpha = -h_\alpha, \quad l_\alpha + h_\alpha, \end{aligned}$$

в силу обозначений сводится к третьей краевой задаче

$$\begin{aligned} (a_\alpha w_{\bar{x}_\alpha}^\alpha)_{x_\alpha} &= -\frac{|a_{\alpha x_\alpha}|}{2h_\alpha}, \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ a_\alpha^{+1} w_{\bar{x}_\alpha}^\alpha - \kappa_{-\alpha} w^\alpha &= -\frac{|a_\alpha^{+1} - h_\alpha \kappa_{-\alpha}|}{2h_\alpha}, \quad x_\alpha = 0, \\ -a_\alpha w_{\bar{x}_\alpha}^\alpha - \kappa_{+\alpha} w^\alpha &= -\frac{|a_\alpha - h_\alpha \kappa_{+\alpha}|}{2h_\alpha}, \quad x_\alpha = l_\alpha. \end{aligned} \quad (36)$$

Отсюда получим, что

$$c_\alpha(x_\beta) = \max_{0 \leq x_\alpha \leq l_\alpha} w^\alpha(x) = O\left(\frac{1}{h_\alpha^2}\right),$$

и следовательно,

$$\Delta = 4 \max_{\alpha=1,2} \left(\max_{0 \leq x_\beta \leq l_\beta} (c_\alpha(x_\beta) + \sqrt{b_\alpha(x_\beta)})^2 \right) = O\left(\frac{1}{|h|^2}\right).$$

Поэтому для модифицированного попеременно-треугольного метода, примененного к нахождению решения третьей краевой за-

дачи (32), число итераций зависит от числа узлов так же, как и в случае первой краевой задачи.

Очевидно, что описанный выше прием сведения к задаче Дирихле можно использовать и в том случае, когда на каждой стороне прямоугольника задано одно из краевых условий первого, второго или третьего рода.

5. Разностная задача Дирихле для уравнения со смешанными производными. Пусть в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha=1, 2\}$ с границей Γ требуется найти решение задачи Дирихле для уравнения эллиптического типа со смешанными производными

$$\begin{aligned} Lu &= \sum_{\alpha, \beta=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}(x) \frac{\partial u}{\partial x_\beta} \right) = -\varphi(x), \quad x \in G, \\ u(x) &= g(x), \quad x \in \Gamma, \quad k_{\alpha\beta}(x) = k_{\beta\alpha}(x). \end{aligned} \quad (37)$$

Будем требовать выполнения условий

$$k_{\alpha\alpha}(x) \geq c > 0, \quad k_{12}^2(x) \leq \rho^2 k_{11}(x) k_{22}(x), \quad x \in \bar{G}, \quad (38)$$

$$0 \leq \rho < 1.$$

Отметим, что условия (38) обеспечивают равномерную эллиптичность уравнения (37). В самом деле, рассмотрим для фиксированного $x \in G$ задачу на собственные значения для пучка матриц

$$\begin{vmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{vmatrix} - \lambda \begin{vmatrix} k_{11} & 0 \\ 0 & k_{22} \end{vmatrix} = 0.$$

Для λ имеем квадратное уравнение $(1-\lambda)^2 k_{11} k_{22} - k_{12}^2 = 0$. Отсюда найдем

$$|1-\lambda| = \frac{|k_{12}|}{\sqrt{k_{11} k_{22}}} \leq \rho, \quad 1-\rho \leq \lambda \leq 1+\rho.$$

Следовательно, имеет место неравенство

$$\begin{aligned} c_1 \sum_{\alpha=1}^2 k_{\alpha\alpha}(x) \xi_\alpha^2 &\leq \sum_{\alpha, \beta=1}^2 k_{\alpha\beta}(x) \xi_\alpha \xi_\beta \leq c_2 \sum_{\alpha=1}^2 k_{\alpha\alpha}(x) \xi_\alpha^2, \\ c_1 = 1-\rho, \quad c_2 = 1+\rho, \end{aligned} \quad (39)$$

где $\xi = (\xi_1, \xi_2)$ — произвольный вектор. Отсюда в силу условия $k_{\alpha\alpha}(x) \geq c > 0$ следует равномерная эллиптичность оператора L .

На прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ задаче (37), (38) поставим в соответствие разностную задачу Дирихле

$$\begin{aligned} \Lambda y &= \frac{1}{2} \sum_{\alpha, \beta=1}^2 [(k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha} + (k_{\alpha\beta} y_{\bar{x}_\alpha})_{\bar{x}_\beta}] = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma. \end{aligned} \quad (40)$$

В пространстве H сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2$$

определим оператор A следующим образом: $Ay = -\Lambda \dot{y}$, $y \in H$ и $y(x) = \dot{y}(x)$ для $x \in \omega$, $\dot{y}(x) = 0$ для $x \in \gamma$, а также оператор $R: Ry = -\mathcal{R}y$, где

$$\mathcal{R}y = \sum_{\alpha=1}^2 (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, \quad x \in \omega; \quad a_\alpha(x) = \frac{k_{\alpha\alpha} + k_{\alpha\alpha}^{-1}}{2}.$$

Тогда задачу (40) можно записать в виде уравнения (3), где $f(x)$ отличается от $\varphi(x)$ лишь в приграничных узлах. Так как $k_{\alpha\beta}(x) = k_{\beta\alpha}(x)$, то операторы A и R самосопряжены. Покажем, что имеют место неравенства

$$c_1 R \leq A \leq c_2 R, \quad c_1 = 1 - \rho, \quad c_2 = 1 + \rho, \quad (41)$$

где ρ задано в (38). Действительно, из разностных формул Грина получим

$$(Ay, y) = -(\Lambda \dot{y}, \dot{y}) = \sum_{\alpha, \beta=1}^2 \frac{1}{2} [(k_{\alpha\beta} \dot{y}_{\bar{x}_\beta}, \dot{y}_{\bar{x}_\alpha})_\alpha + (k_{\alpha\beta} \dot{y}_{x_\beta}, \dot{y}_{x_\alpha})],$$

где скалярное произведение (u, v) определено в п. 1 § 2, а

$${}_\alpha(u, v) = \sum_{x_\alpha=0}^{l_\alpha-h_\alpha} \sum_{x_\beta=h_\beta}^{l_\beta-h_\beta} u(x)v(x)h_1h_2, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Заметим, что если одна из функций $u(x)$ или $v(x)$ обращается в нуль при $x_\beta = 0$ (или при $x_\beta = l_\beta$), то

$$\begin{aligned} {}_\alpha(u, v) &= [u, v] = \sum_{x_1=0}^{l_1-h_1} \sum_{x_2=0}^{l_2-h_2} u(x)v(x)h_1h_2, \\ (u, v)_\alpha &= (u, v) = \sum_{x_1=h_1}^{l_1} \sum_{x_2=h_2}^{l_2} u(x)v(x)h_1h_2, \quad \alpha = 1, 2. \end{aligned}$$

Это сразу дает

$$(Ay, y) = \sum_{\alpha, \beta=1}^2 \frac{1}{2} \left\{ (k_{\alpha\beta} \dot{y}_{\bar{x}_\beta}, \dot{y}_{\bar{x}_\alpha}) + (k_{\alpha\beta} \dot{y}_{x_\beta}, \dot{y}_{x_\alpha}) \right\}. \quad (42)$$

Далее, так как $\mathcal{R}y$ можно записать в виде

$$\mathcal{R}y = \frac{1}{2} \sum_{\alpha=1}^2 \left\{ (k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha} \right\},$$

то

$$\begin{aligned} (Ry, \dot{y}) &= -(\mathcal{R}\dot{y}, \dot{y}) = \sum_{\alpha=1}^2 \frac{1}{2} \left\{ (k_{\alpha\alpha} \dot{y}_{x_\alpha}, \dot{y}_{x_\alpha})_\alpha + \alpha (k_{\alpha\alpha} \dot{y}_{x_\alpha}, \dot{y}_{x_\alpha}) \right\} = \\ &= \sum_{\alpha=1}^2 \frac{1}{2} \left\{ (k_{\alpha\alpha} \dot{y}_{x_\alpha}^2, 1) + [k_{\alpha\alpha} \dot{y}_{x_\alpha}^2, 1] \right\}. \end{aligned} \quad (43)$$

Из (42), (43) и неравенств (39) получим

$$c_1 \left[\sum_{\alpha=1}^2 k_{\alpha\alpha} \dot{y}_{x_\alpha}^2, 1 \right] \leq \left[\sum_{\alpha, \beta=1}^2 k_{\alpha\beta} \dot{y}_{x_\beta} \dot{y}_{x_\alpha}, 1 \right] \leq c_2 \left[\sum_{\alpha=1}^2 k_{\alpha\alpha} \dot{y}_{x_\alpha}^2, 1 \right]$$

и аналогично

$$c_1 \left[\sum_{\alpha=1}^2 k_{\alpha\alpha} \dot{y}_{x_\alpha}^2, 1 \right] \leq \left[\sum_{\alpha, \beta=1}^2 k_{\alpha\beta} \dot{y}_{x_\beta} \dot{y}_{x_\alpha}, 1 \right] \leq c_2 \left[\sum_{\alpha=1}^2 k_{\alpha\alpha} \dot{y}_{x_\alpha}^2, 1 \right],$$

и следовательно, оценки (41) доказаны.

Таким образом, оператор R , спределенный выше, можно использовать в качестве регуляризатора в попеременно-треугольном методе

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \dots,$$

$$B = (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2), \quad R_1 = R_1^*, \quad R_1 + R_2 = R,$$

где операторы R_1 , R_2 и \mathcal{D} определены в п. 2 § 2. Там же были найдены постоянные δ и Δ для неравенств $\delta \mathcal{D} \leq R$, $R_1 \mathcal{D}^{-1} R_2 \leq \frac{\Delta}{4} R$, $\delta > 0$. Применение теоремы 2 завершает построение попеременно-треугольного метода для разностной задачи (40).

§ 3. Попеременно-треугольный метод для эллиптических уравнений в произвольной области

1. Постановка разностной задачи. Построим модифицированный попеременно-треугольный метод для решения задачи Дирихле в произвольной ограниченной области \bar{G} с границей Γ в случае эллиптического уравнения с переменными коэффициентами

$$\begin{aligned} \sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x) \cdot \frac{\partial u}{\partial x_\alpha} \right) &= -\varphi(x), \quad x \in G, \\ u(x) &= g(x), \quad x \in \Gamma, \quad k_\alpha(x) \geq c_1 > 0, \quad \alpha = 1, 2. \end{aligned} \quad (1)$$

Предположим, что граница Γ достаточно гладкая. Кроме того, для простоты изложения будем считать, что пересечение области с прямой, проходящей через любую точку $x \in G$ параллельно оси координат Ox_α , $\alpha = 1, 2$, состоит из одного интервала.

В области \bar{G} построим неравномерную сетку $\bar{\omega}$ следующим образом. Проведем семейство прямых $x_\alpha = x_\alpha(i_\alpha)$, $i_\alpha = 0, \pm 1, \pm 2, \dots$, $\alpha = 1, 2$. Тогда точки $x_i = (x_1(i_1), x_2(i_2))$, $i = (i_1, i_2)$ образуют основную решетку на плоскости. Точку x_i решетки, принадлежащую G , назовем внутренним узлом сетки ω . Множество всех внутренних узлов, как сбычно, обозначим через ω .

Пересечением любой прямой, проведенной через точку $x_i \in \omega$ параллельно оси Ox_α , с областью G является интервал $\Delta_\alpha(x_i)$. Концы этого интервала назовем граничными узлами по направлению x_α . Множество всех граничных узлов по x_α обозначим через γ_α . Граница сетки $\bar{\omega}$ есть $\gamma = \gamma_1 \cup \gamma_2$, так что $\bar{\omega} = \omega \cup \gamma$. Сетка $\bar{\omega}$ построена.

Введем ряд обозначений. Обозначим через $\omega_\alpha(x_\beta)$, $\beta = 3 - \alpha$, $\alpha = 1, 2$ — множество узлов сетки ω , лежащих на интервале Δ_α ; $\omega_\alpha^+(x_\beta)$ — множество, состоящее из $\omega_\alpha(x_\beta)$ и правого конца интервала Δ_α ; $\bar{\omega}_\alpha(x_\beta)$ состоит из $\omega_\alpha(x_\beta)$ и концов интервала Δ_α .

Обозначим $x^{(+1)\alpha}$ и $x^{(-1)\alpha}$ узлы, соседние с $x \in \omega_\alpha(x_\beta)$ справа и слева и принадлежащие $\bar{\omega}_\alpha(x_\beta)$. Заметим, что если, например, $x^{(+1)\alpha} \in \gamma_\alpha$, то этот узел может не совпадать с узлом основной решетки.

Определим $h_\alpha^+(x) = x^{(+1)\alpha} - x$, $h_\alpha^-(x) = x - x^{(-1)\alpha}$, $x \in \omega_\alpha$, $x^{(\pm 1)\alpha} \in \bar{\omega}_\alpha$. Во всех внутренних узлах сетки ω определим также средние шаги $\tilde{h}_\alpha(x_\alpha) = 0,5(x_\alpha(i_\alpha + 1) - x_\alpha(i_\alpha - 1))$ как расстояние между соответствующими прямыми основной решетки.

Задаче (1) на сетке $\bar{\omega}$ поставим в соответствие разностную задачу

$$\begin{aligned} \Lambda y &= \sum_{\alpha=1}^2 (a_\alpha y_{\bar{x}_\alpha})_{\hat{x}_\alpha} = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma. \end{aligned} \tag{2}$$

Здесь использованы следующие обозначения:

$$\begin{aligned} (a_\alpha y_{\bar{x}_\alpha})_{\hat{x}_\alpha} &= \frac{1}{\tilde{h}_\alpha} (a_\alpha^{+1} y_{x_\alpha} - a_\alpha^- y_{\bar{x}_\alpha}), \quad a_\alpha^{+1} = a_\alpha(x^{(+1)\alpha}), \\ y_{x_\alpha} &= \frac{1}{h_\alpha^+} (y(x^{(+1)\alpha}) - y(x)), \quad y_{\bar{x}_\alpha} = \frac{1}{h_\alpha^-} (y(x) - y(x^{(-1)\alpha})). \end{aligned}$$

Коэффициенты $a_\alpha(x)$ и $\varphi(x)$ выбраны так, чтобы схема (2) на равномерной сетке имела локальный второй порядок аппроксимации.

Введем теперь H — пространство сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{x \in \omega} u(x)v(x) \times \tilde{h}_1(x_1) \tilde{h}_2(x_2)$. Оператор A определим обычным образом: $Ay = -\Lambda y$,

$y \in H$ и $y(x) = \dot{y}(x)$ для $x \in \omega$, $\dot{y}(x) = 0$ для $x \in \gamma$. Тогда разностная задача (2) запишется в виде уравнения

$$Au = f, \quad (3)$$

где $f(x)$ отличается от $\varphi(x)$ лишь в приграничных узлах.

2. Построение попаременно-треугольного метода. Для уравнения (3) рассмотрим модифицированный попаременно-треугольный метод

$$\begin{aligned} B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k &= f, \quad k = 0, 1, \dots, \\ B &= (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2), \quad R_1 = R_1^*, \quad R_1 + R_2 = A. \end{aligned} \quad (4)$$

Определим операторы R_1 , R_2 и \mathcal{D} . Как и в случае прямоугольника, выберем простейший оператор \mathcal{D} :

$$\mathcal{D}y = d(x)y, \quad d(x) > 0 \text{ для } x \in \omega, \quad (5)$$

и положим $R_\alpha y = -\mathcal{R}_\alpha \dot{y}$, $y \in H$ и $y(x) = \dot{y}(x)$, $x \in \omega$, где

$$\begin{aligned} \mathcal{R}_1 y &= -\sum_{\alpha=1}^2 \left[\frac{a_\alpha}{\tilde{h}_\alpha} y_{\tilde{x}_\alpha} + \frac{1}{2\tilde{h}_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha^-}{h_\alpha^-} \right) y \right], \\ \mathcal{R}_2 y &= \sum_{\alpha=1}^2 \left[\frac{a_\alpha^{+1}}{\tilde{h}_\alpha} y_{x_\alpha} + \frac{1}{2\tilde{h}_\alpha} \left(\frac{a_\alpha^+}{h_\alpha^+} - \frac{a_\alpha^-}{h_\alpha^-} \right) y \right]. \end{aligned} \quad (6)$$

Так как $\mathcal{R}_1 + \mathcal{R}_2 = \Lambda$, то получим, что $R_1 + R_2 = A$.

Покажем, что операторы R_1 и R_2 сопряжены. Сначала введем обозначения, которыми будем пользоваться в дальнейшем. Определим следующие суммы:

$$\begin{aligned} (u, v)_{\omega_\alpha(x_\beta)} &= \sum_{x_\alpha \in \omega_\alpha(x_\beta)} u(x)v(x)\tilde{h}_\alpha(x_\alpha), \\ (u, v)_{\omega_\alpha^+(x_\beta)} &= \sum_{x_\alpha \in \omega_\alpha^+(x_\beta)} u(x)v(x)h_\alpha^-(x), \\ (u, v)_\alpha &= ((u, v)_{\omega_\alpha^+(x_\beta)}, 1)_{\omega_\beta(x_\alpha)} = \\ &= \sum_{x_\beta \in \omega_\beta} \sum_{x_\alpha \in \omega_\alpha^+} u(x)v(x)h_\alpha^-(x)\tilde{h}_\beta(x_\beta). \end{aligned}$$

$\beta = 3 - \alpha$, $\alpha = 1, 2$,

Используя введенные обозначения, скалярное произведение в H можно записать следующим образом:

$$(u, v) = ((u, v)_{\omega_1(x_2)}, 1)_{\omega_2(x_1)} = ((u, v)_{\omega_2(x_1)}, 1)_{\omega_1(x_2)}. \quad (7)$$

Сначала докажем одно вспомогательное утверждение. Пусть y_i и v_i — сеточные функции, заданные для $0 \leq i \leq N$, причем

$y_0 = y_N = 0$ и $v_0 = v_N = 0$. Пусть u_i — сеточная функция, заданная для $1 \leq i \leq N$. Тогда имеет место равенство

$$\sum_{i=1}^{N-1} (u_{i+1} - u_i) y_i v_i = - \sum_{i=1}^{N-1} (v_{i+1} - v_i) u_{i+1} y_i - \sum_{i=1}^{N-1} (y_i - y_{i-1}) u_i v_i. \quad (8)$$

Действительно, имеем

$$\begin{aligned} \sum_{i=1}^{N-1} [(u_{i+1} - u_i) y_i v_i + (v_{i+1} - v_i) u_{i+1} y_i + (y_i - y_{i-1}) u_i v_i] &= \\ &= \sum_{i=1}^{N-1} (u_{i+1} v_{i+1} y_i - u_i v_i y_{i-1}) = u_N v_N y_{N-1} - u_1 v_1 y_0 = 0. \end{aligned}$$

Утверждение (8) доказано. Используя (8), легко показать, что для функций $\dot{y}(x)$ и $\dot{v}(x)$, заданных на ω и обращающихся в нуль на γ , имеет место равенство

$$(u_{\hat{x}_\alpha} \dot{y}, \dot{v})_{\omega_\alpha} = - \left(\dot{y}, \frac{h_\alpha^+}{h_\alpha} u^{+1} \alpha \dot{v}_{x_\alpha} \right)_{\omega_\alpha} - \left(\frac{h_\alpha^-}{h_\alpha} u \dot{y}_{x_\alpha}, \dot{v} \right)_{\omega_\alpha}. \quad (9)$$

Здесь использованы обозначения

$$u^{+1} \alpha = u(x^{(+1)\alpha}), \quad u_{\hat{x}_\alpha} = \frac{u^{+1} \alpha - u}{h_\alpha(x_\alpha)}.$$

Подставляя в (9) выражение $u(x) = a_\alpha(x)/h_\alpha^-(x)$ и учитывая равенство $h_\alpha^-(x^{(+1)\alpha}) = h_\alpha^+(x)$, получим

$$\left(\frac{1}{h_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-} \right) \dot{y}, \dot{v} \right)_{\omega_\alpha} = - \left(\dot{y}, \frac{a_\alpha^{+1}}{h_\alpha} \dot{v}_{x_\alpha} \right)_{\omega_\alpha} - \left(\frac{a_\alpha}{h_\alpha} \dot{y}_{x_\alpha}, \dot{v} \right)_{\omega_\alpha}.$$

Умножая это соотношение на $h_\beta(x_\beta)$, суммируя по ω_β и учитывая (7), найдем

$$-\left(\frac{a_\alpha}{h_\alpha} \dot{y}_{x_\alpha}, \dot{v} \right) = \left(\frac{a_\alpha^{+1}}{h_\alpha} \dot{v}_{x_\alpha}, \dot{y} \right) + \left(\frac{1}{h_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-} \right) \dot{y}, \dot{v} \right). \quad (10)$$

Докажем теперь, что операторы R_1 и R_2 сопряжены. Из (6) и (10) получим

$$\begin{aligned} (R_1 y, v) &= -(\mathcal{R}_1 \dot{y}, \dot{v}) = \\ &= \sum_{\alpha=1}^2 \left[\left(\frac{a_\alpha}{h_\alpha} \dot{y}_{x_\alpha}, \dot{v} \right) + \left(\frac{1}{2h_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-} \right) \dot{y}, \dot{v} \right) \right] = \\ &= - \sum_{\alpha=1}^2 \left[\left(\dot{y}, \frac{a_\alpha^{+1}}{h_\alpha} \dot{v}_{x_\alpha} \right) + \left(\frac{1}{2h_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-} \right) \dot{y}, \dot{v} \right) \right] = \\ &= -(\dot{y}, \mathcal{R}_2 \dot{v}) = (y, R_2 v). \end{aligned}$$

Утверждение доказано. Отсюда, кстати, следует и самосопряженность оператора A .

Нам осталось построить функцию $d(x)$, определяющую оператор \mathcal{D} , найти постоянные δ и Δ в неравенствах

$$\delta \mathcal{D} \leq A, \quad R_1 \mathcal{D}^{-1} R_2 \leq \frac{\Delta}{4} A, \quad \delta > 0 \quad (11)$$

и воспользоваться теоремой 1. Все это мы сделаем так же, как и в п. 2 § 2, где была рассмотрена задача Дирихле для эллиптического уравнения в прямоугольнике на равномерной сетке.

Сначала отметим, что в силу разностных формул Грина имеет место равенство

$$(Ay, y) = \sum_{\alpha=1}^2 \left(a_{\alpha} \dot{y}_{x_{\alpha}}, 1 \right)_{\alpha}, \quad \dot{y}(x) = 0, \quad x \in \gamma.$$

Далее, из (5) и (6) найдем

$$\begin{aligned} (R_1 \mathcal{D}^{-1} R_2 y, y) &= (\mathcal{D}^{-1} \mathcal{R}_2 \dot{y}, \mathcal{R}_2 \dot{y}) = \\ &= \left(\frac{1}{d} \sum_{\alpha=1}^2 \left[\frac{a_{\alpha}^{+1}}{h_{\alpha}} \dot{y}_{x_{\alpha}} + \frac{1}{2h_{\alpha}} \left(\frac{a_{\alpha}^{+1}}{h_{\alpha}^{+}} - \frac{a_{\alpha}}{h_{\alpha}^{-}} \right) y \right]^2, 1 \right). \end{aligned}$$

Используем теперь лемму 2, полагая

$$\begin{aligned} p_{\alpha} &= \frac{a_{\alpha}^{+1}}{h_{\alpha}}, \quad q_{\alpha} = \frac{h_{\alpha}^{+1}}{2h_{\alpha}} \left(\frac{a_{\alpha}^{+1}}{h_{\alpha}^{+}} - \frac{a_{\alpha}}{h_{\alpha}^{-}} \right), \\ u_{\alpha} &= \dot{y}_{x_{\alpha}}, \quad v_{\alpha} = \frac{1}{h_{\alpha}^{+}} \dot{y}, \quad \alpha = 1, 2. \end{aligned}$$

В результате получим неравенство

$$\begin{aligned} (R_1 \mathcal{D}^{-1} R_2 y, y) &\leq \left(\frac{(1+\varepsilon)}{d h_1^2} \left[a_1^{+1} + \frac{\kappa_1 h_1^+}{2} \left| \frac{a_1^{+1}}{h_1^+} - \frac{a_1}{h_1^-} \right| \right] \left[a_1^{+1} \dot{y}_{x_1}^2 + \right. \right. \\ &\quad \left. \left. + \frac{1}{2\kappa_1 h_1^+} \left| \frac{a_1^{+1}}{h_1^+} - \frac{a_1}{h_1^-} \right| \dot{y}^2 \right], 1 \right) + \left(\frac{(1+\varepsilon)}{d e h_2^2} \left[a_2^{+1} + \frac{\kappa_2 h_2^+}{2} \left| \frac{a_2^{+1}}{h_2^+} - \frac{a_2}{h_2^-} \right| \right] \times \right. \\ &\quad \left. \left. \times \left[a_2^{+1} \dot{y}_{x_2}^2 + \frac{1}{2\kappa_2 h_2^+} \left| \frac{a_2^{+1}}{h_2^+} - \frac{a_2}{h_2^-} \right| \dot{y}^2 \right] \right], 1. \right) \end{aligned}$$

Положим здесь

$$e = e(x) = \frac{a_2^{+1} + 0,5\kappa_2 h_2^+ \left| \frac{a_2^{+1}}{h_2^+} - \frac{a_2}{h_2^-} \right|}{a_1^{+1} + 0,5\kappa_1 h_1^+ \left| \frac{a_1^{+1}}{h_1^+} - \frac{a_1}{h_1^-} \right|} \frac{\hbar_1 h_1^+}{\hbar_2 h_2^+} \frac{\theta_2}{\theta_1}$$

и определим $d(x)$ следующим образом:

$$d(x) = \sum_{\alpha=1}^2 \left(a_{\alpha}^{+1} + \frac{\kappa_{\alpha} h_{\alpha}^+}{2} \left| \frac{a_{\alpha}^{+1}}{h_{\alpha}^+} - \frac{a_{\alpha}}{h_{\alpha}^-} \right| \right) \frac{\theta_{\alpha}}{\hbar_{\alpha} h_{\alpha}^+}, \quad x \in \Omega.$$

Здесь предполагается, что $\kappa_\alpha = \kappa_\alpha(x_\beta) > 0$, $\theta_\alpha = \theta_\alpha(x_\beta) > 0$, $\beta = 3 - \alpha$, $\alpha = 1, 2$. В результате получим неравенство

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leqslant \sum_{\alpha=1}^2 \left(\frac{a_\alpha^{+1} h_\alpha^+}{\theta_\alpha \tilde{h}_\alpha} \dot{y}_{x_\alpha}^2, 1 \right) + \sum_{\alpha=1}^2 \left(\frac{1}{2\tilde{h}_\alpha \theta_\alpha \kappa_\alpha} \left| \frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha^-}{h_\alpha^-} \right| \dot{y}_{x_\alpha}^2, 1 \right).$$

Так как θ_α не зависит от x_α , то

$$\left(\frac{a_\alpha^{+1}}{\theta_\alpha} \frac{h_\alpha^+}{\tilde{h}_\alpha} \dot{y}_{x_\alpha}^2, 1 \right)_{\omega_\alpha} = \frac{1}{\theta_\alpha} \sum_{x_\alpha \in \omega_\alpha} a_\alpha^{+1} h_\alpha^+ \dot{y}_{x_\alpha}^2 \leqslant \frac{1}{\theta_\alpha} \sum_{x_\alpha \in \omega_\alpha^+} a_\alpha h_\alpha^- \dot{y}_{x_\alpha}^2.$$

Следовательно,

$$\left(\frac{a_\alpha^{+1}}{\theta_\alpha} \frac{h_\alpha^+}{\tilde{h}_\alpha} \dot{y}_{x_\alpha}^2, 1 \right) \leqslant \left(\frac{a_\alpha}{\theta_\alpha} \dot{y}_{x_\alpha}^2, 1 \right)_\alpha,$$

и поэтому окончательно имеем

$$(R_1 \mathcal{D}^{-1} R_2 y, y) \leqslant \sum_{\alpha=1}^2 \left(\frac{a_\alpha}{\theta_\alpha} \dot{y}_{x_\alpha}^2, 1 \right)_\alpha + \sum_{\alpha=1}^2 \left(\frac{1}{2\tilde{h}_\alpha \theta_\alpha \kappa_\alpha} \left| \frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha^-}{h_\alpha^-} \right| \dot{y}_{x_\alpha}^2, 1 \right).$$

Дальнейшие выкладки являются полным аналогом преобразований и оценок, полученных в п. 2 § 2. Приведем итог: в неравенствах (11)

$$\delta = 1, \quad \Delta = 4 \max_{\alpha=1, 2} \left(\max_{x_\beta \in \omega_\beta} (c_\alpha(x_\beta) + \sqrt{b_\alpha(x_\beta)})^2 \right), \quad \beta = 3 - \alpha,$$

где

$$b_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x), \quad c_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} w^\alpha(x), \quad x_\beta \in \omega_\beta;$$

функция $v^\alpha(x)$ есть решение трехточечной краевой задачи

$$\begin{aligned} \left(a_\alpha v_{x_\alpha}^\alpha \right)_{\hat{x}_\alpha} &= -\frac{a_\alpha^{+1}}{\tilde{h}_\alpha h_\alpha^+}, \quad x_\alpha \in \omega_\alpha(x_\beta), \\ v^\alpha(x) &= 0, \quad x_\alpha \in \gamma_\alpha, \end{aligned} \tag{12}$$

а функция $w^\alpha(x)$ — решение задачи

$$\left(a_\alpha w_{x_\alpha}^\alpha \right)_{\hat{x}_\alpha} = -\frac{1}{2\tilde{h}_\alpha} \left| \frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha^-}{h_\alpha^-} \right|, \quad x_\alpha \in \omega_\alpha(x_\beta), \tag{13}$$

$$w^\alpha(x) = 0, \quad x_\alpha \in \gamma_\alpha,$$

Функция $d(x)$ при этом вычисляется по формуле

$$d(x) = \sum_{\alpha=1}^2 \left(\frac{a_\alpha^{+1}}{\tilde{h}_\alpha h_\alpha^+} \sqrt{b_\alpha} + \frac{1}{2\tilde{h}_\alpha} \left| \frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha^-}{h_\alpha^-} \right| \right) \frac{1}{c_\alpha + \sqrt{b_\alpha}}, \quad x \in \omega.$$

Итерационные параметры ω и $\{\tau_k\}$ вычисляются по формулам теоремы 1. Для нахождения y_{k+1} можно воспользоваться алгоритмом

$$\begin{aligned} v(x) &= \alpha_1(x)v^{(-1)} + \beta_1(x)v^{(-1)} + \kappa(x)\varphi_k(x), \quad x \in \omega, \\ v(x) &= 0, \quad x \in \gamma, \\ \dot{y}_{k+1}(x) &= \alpha_2(x)\overset{\circ}{y}_{k+1}^{(+1)} + \beta_2(x)\overset{\circ}{y}_{k+1}^{(+1)} + \kappa(x)d(x)v(x), \quad x \in \omega, \\ \overset{\circ}{y}_{k+1}(x) &= 0, \quad x \in \gamma, \end{aligned}$$

где

$$\begin{aligned} \alpha_1 &= \frac{\omega_0 a_1 \kappa}{\hbar_1 h_1^-}, \quad \beta_1 = \frac{\omega_0 a_2 \kappa}{\hbar_2 h_2^+}, \quad \alpha_2 = \frac{\omega_0 a_1^{+1} \kappa}{\hbar_1 h_1^+}, \quad \beta_2 = \frac{\omega_0 a_2^{+1} \kappa}{\hbar_2 h_2^+}, \\ \frac{1}{\kappa} &= d + \frac{1}{2} \sum_{\alpha=1}^2 \frac{\omega_0}{\hbar_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} + \frac{a_\alpha}{h_\alpha^-} \right), \quad \varphi_k(x) = By_k - \tau_{k+1}(Ay_k - f). \end{aligned}$$

Следует отметить, что в подобных случаях, когда вычисление значения By_k требует больших затрат вычислительной работы, а ограничений на объем запоминаемой промежуточной информации нет, целесообразно использовать второй алгоритм, описанный в п. 1 § 1.

3. Задача Дирихле для уравнения Пуассона в произвольной области. Рассмотрим в качестве примера применения построенного метода задачу Дирихле для уравнения Пуассона

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -\varphi(x), \quad x \in G, \quad u(x) = g(x), \quad x \in \Gamma.$$

Предположим, что решетка квадратная, т. е. $x_\alpha = x_\alpha(i_\alpha)$, $x_\alpha(i_\alpha + 1) = x_\alpha(i_\alpha) + h$, $i_\alpha = 0, \pm 1, \pm 2, \dots$, и
 $\omega = \{x_i = (i_1 h, i_2 h) \in G, \quad i_\alpha = 0, \pm 1, \pm 2, \dots\}$.

При этом $\hbar_\alpha \equiv h$, а шаги $h_\alpha^\pm(x)$ отличны от h только в приграничных узлах сетки ω .

Воспользуемся разностной схемой (2), в которой положим $a_\alpha(x) \equiv 1$ и $\hbar_\alpha \equiv h$. Чтобы применить построенный в п. 2 § 3 попеременно-треугольный метод, надо найти решения одномерных трехточечных краевых задач (12) и (13), которые в данном случае имеют вид

$$\Lambda_\alpha v^\alpha = v_{x_\alpha \hat{x}_\alpha}^\alpha = -\frac{1}{hh_\alpha^+}, \quad x_\alpha \in \omega_\alpha(x_\beta), \quad v^\alpha(x) = 0, \quad x_\alpha \in \gamma_\alpha, \quad (14)$$

$$\begin{aligned} \Lambda_\alpha w^\alpha &= w_{x_\alpha \hat{x}_\alpha}^\alpha = -\frac{1}{2h} \left| \frac{1}{h_\alpha^+} - \frac{1}{h_\alpha^-} \right|, \\ x_\alpha \in \omega_\alpha(x_\beta), \quad w^\alpha(x) &= 0, \quad x_\alpha \in \gamma_\alpha. \end{aligned} \quad (15)$$

Рассмотрим интервал Δ_α , содержащий $\omega_\alpha(x_\beta)$, и обозначим через $l_\alpha(x_\beta)$ и $L_\alpha(x_\beta)$ его левый и правый концы. Тогда $h_\alpha^-(x) \leq h$, если x — ближайший к l_α узел сетки ω_α , и $h_\alpha^+(x) \leq h$, если x —

ближайший к L_α узел сетки ω_α . Все остальные шаги h_α^\pm равны основному шагу решетки h .

При этом оператор Λ_α на сетке ω_α подробно записывается следующим образом:

$$\Lambda_\alpha y = \begin{cases} \frac{1}{h} \left(\frac{y^{+1\alpha} - y}{h} - \frac{y - y^{-1\alpha}}{h_\alpha^-} \right), & x_\alpha = l_\alpha + h_\alpha^-, \\ \frac{1}{h^2} (y^{+1\alpha} - 2y + y^{-1\alpha}), & l_\alpha + h_\alpha^- + h \leq x_\alpha \leq L_\alpha - h_\alpha^+ - h, \\ \frac{1}{h} \left(\frac{y^{+1\alpha} - y}{h_\alpha^+} - \frac{y - y^{-1\alpha}}{h} \right), & x_\alpha = L_\alpha - h_\alpha^+. \end{cases}$$

Решения уравнений (14), (15) можно найти в явном виде. Для этого подставим краевые условия в уравнения, записанные в точках $x_\alpha = l_\alpha + h_\alpha^-$ и $x_\alpha = L_\alpha - h_\alpha^+$. Эти уравнения преобразуются, они станут двухточечными, и их можно рассматривать как новые краевые условия для трехточечных уравнений с постоянными коэффициентами, записанными для $l_\alpha + h_\alpha^- + h \leq x_\alpha \leq L_\alpha - h_\alpha^+ - h$. Следовательно, будем иметь следующие задачи для $v(x)$ и $w(x)$ (верхний индекс α у v и w временно опущен):

$$\begin{aligned} v^{+1\alpha} - 2v + v^{-1\alpha} &= -1, & l_\alpha + h_\alpha^- + h \leq x_\alpha \leq L_\alpha - h_\alpha^+ - h, \\ \left(1 + \frac{h}{h_\alpha^-}\right)v &= v^{+1\alpha} + 1, & x_\alpha = l_\alpha + h_\alpha^-, \\ \left(1 + \frac{h_\alpha^+}{h}\right)v &= \frac{h_\alpha^+}{h} v^{-1\alpha} + 1, & x_\alpha = L_\alpha - h_\alpha^+, \end{aligned} \quad (16)$$

$$\begin{aligned} w^{+1\alpha} - 2w + w^{-1\alpha} &= 0, & l_\alpha + h_\alpha^- + h \leq x_\alpha \leq L_\alpha - h_\alpha^+ - h, \\ \left(1 + \frac{h}{h_\alpha^-}\right)w &= w^{+1\alpha} - \frac{1}{2} \left(1 - \frac{h}{h_\alpha^-}\right), & x_\alpha = l_\alpha + h_\alpha^-, \\ \left(1 + \frac{h_\alpha^+}{h}\right)w &= \frac{h_\alpha^+}{h} w^{-1\alpha} + \frac{1}{2} \left(1 - \frac{h_\alpha^+}{h}\right), & x_\alpha = L_\alpha - h_\alpha^+. \end{aligned} \quad (17)$$

Используя методы решения разностных уравнений с постоянными коэффициентами, изложенные в § 4 гл. I, найдем явный вид решения краевых задач (16) и (17):

$$\begin{aligned} v_\alpha(x) &= \\ &= \frac{1}{2h^2} \left[(x_\alpha - l_\alpha) \left(L_\alpha - x_\alpha + \frac{2h^2 - (h_\alpha^+ + h_\alpha^-)(h_\alpha^+ + h - h_\alpha^-)}{L_\alpha - l_\alpha} \right) + h_\alpha^-(h - h_\alpha^-) \right], \\ w^\alpha(x) &= \frac{1}{2} - \frac{h_\alpha^-(L_\alpha - x_\alpha) + h_\alpha^+(x_\alpha - l_\alpha)}{2h(L_\alpha - l_\alpha)} \end{aligned}$$

для $l_\alpha + h_\alpha^- \leq x_\alpha \leq L_\alpha - h_\alpha^+$. Так как $h_\alpha^\pm \leq h$, то

$$(h_\alpha^+ + h_\alpha^-)(h_\alpha^+ + h - h_\alpha^-) \geq h_\alpha^-(h - h_\alpha^-),$$

поэтому

$$v^\alpha(x) \leq \frac{1}{2h^2} (x_\alpha - l_\alpha)(L_\alpha - x_\alpha) + 1 \leq \frac{1}{2h^2} \left(\frac{L_\alpha - l_\alpha}{2} \right)^2 + 1,$$

$$w^\alpha(x) \leq \frac{1}{2}, \quad \alpha = 1, 2.$$

Следовательно,

$$b_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x) \leq \frac{1}{2h^2} \left(\frac{L_\alpha - l_\alpha}{2} \right)^2 + 1,$$

$$c_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} w^\alpha(x) \leq \frac{1}{2}.$$

Следовательно, $\Delta = O(l_0^2/h^2)$, где l_0 — диаметр области G . Поэтому в силу теоремы 1 для числа итераций верна оценка

$$n \geq n_0(\epsilon) = \frac{\ln(2/\epsilon)}{2\sqrt[4]{2}\sqrt[4]{\eta}} = \frac{\ln(2/\epsilon)}{2\sqrt[4]{2}\sqrt[4]{2}\sqrt{h/l_0}} \approx 0,298\sqrt{N} \ln \frac{2}{\epsilon}, \quad (18)$$

где N — максимальное число узлов по направлению x_1 или x_2 . Таким образом, число итераций для рассматриваемого модельного примера зависит лишь от основного шага h решетки и не зависит от шагов в приграничных узлах сетки ω .

Сравним оценку (18) с оценкой числа итераций для случая задачи Дирихле в квадрате со стороной l_0 и числом узлов N по каждому направлению для квадратной сетки ω . Соответствующая оценка для числа итераций была получена в п. 4 § 1, она имеет вид

$$n \geq n_0(\epsilon) = 0,28\sqrt{N} \ln(2/\epsilon).$$

Отсюда следует, что для произвольной области G число итераций модифицированного попеременно-треугольного метода практически такое же, как и число итераций для той же задачи Дирихле для уравнения Пуассона в квадрате, сторона которого равна диаметру области G .

З а м е ч а н и е 1. Изложенный здесь способ построения попеременно-треугольного метода можно, очевидно, использовать и для случая, когда требуется решить эллиптическое уравнение в прямоугольнике, но на неравномерной сетке.

З а м е ч а н и е 2. Построение метода для случая уравнения со смешанными производными можно осуществить при помощи выбора регуляризатора R , подобно тому, как это было сделано в п. 5 § 2.

Г Л А В А Х I

МЕТОД ПЕРЕМЕННЫХ НАПРАВЛЕНИЙ

В главе рассматриваются специальные итерационные методы решения сеточных эллиптических уравнений $Au=f$, оператор A в которых обладает определенной структурой. В § 1 изучен метод переменных направлений для коммутативного случая; построен оптимальный набор параметров. В § 2 метод иллюстрируется на примерах решения краевых задач для эллиптических уравнений с разделяющимися переменными. § 3 посвящен методу переменных направлений в некоммутативном случае.

§ 1. Метод переменных направлений в коммутативном случае

1. Итерационная схема метода. В главе X был изучен универсальный попеременно-треугольный итерационный метод, оператор B в котором выбирался с учетом разложения оператора A на сумму двух сопряженных друг другу операторов. Наиболее часто используется разложение A на сумму треугольных операторов, при этом B есть произведение треугольных операторов, зависящих от дополнительного итерационного параметра. Учет структуры оператора B позволяет оптимально выбрать итерационные параметры и построить метод, который сходится существенно быстрее явного метода. В применении к решению сеточных эллиптических уравнений этот метод является и экономичным, так как на реализацию одного итерационного шага требуется число арифметических действий, пропорциональное числу неизвестных в задаче.

Как мы знаем, операторы A , соответствующие сеточным эллиптическим уравнениям, имеют специфическую структуру. Поэтому при выборе операторов B в неявных итерационных схемах естественно попытаться использовать эту особенность оператора A . Очевидно, что такие итерационные методы не будут универсальными, однако сужение класса исходных задач требованием определенной структуры оператора A позволяет построить быстросходящиеся итерационные методы, ориентированные на решение именно сеточных уравнений.

В настоящей главе будет изучен специальный метод — итерационный метод переменных направлений. Сначала дается описание этого метода в операторном виде, а затем на примерах будет продемонстрировано применение этого метода к нахожде-

нию приближенного решения различных сеточных эллиптических уравнений.

Описание метода начнем с итерационной схемы. Пусть требуется найти решение линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A , заданным в гильбертовом пространстве H . Пусть оператор A представлен в виде суммы двух операторов A_1 и A_2 , т. е. $A = A_1 + A_2$. Для приближенного решения уравнения (1) рассмотрим неявную двухслойную итерационную схему

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (2)$$

$$B_k = (\omega_k^{(1)} E + A_1)(\omega_k^{(2)} E + A_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}, \quad (3)$$

в которой оператор B_{k+1} на верхнем слое зависит от номера итерации k . Здесь $\omega_k^{(1)}$ и $\omega_k^{(2)}$ — итерационные параметры, также зависящие от номера итерации k и подлежащие определению.

Остановимся сначала на способах нахождения y_{k+1} при заданном y_k . Одним из возможных алгоритмов реализации схемы (2) является следующий:

$$\begin{aligned} (\omega_{k+1}^{(1)} E + A_1) y_{k+1/2} &= (\omega_{k+1}^{(1)} E - A_2) y_k + f, \\ (\omega_{k+1}^{(2)} E + A_2) y_{k+1} &= (\omega_{k+1}^{(2)} E - A_1) y_{k+1/2} + f, \end{aligned} \quad k = 0, 1, \dots, \quad (4)$$

где $y_{k+1/2}$ — промежуточное итерационное приближение.

Покажем, что система (4) алгебраически эквивалентна схеме (2). Для этого исключим $y_{k+1/2}$ из (4). Учитывая, что $A = A_1 + A_2$, перепишем (4) в следующем виде:

$$\begin{aligned} (\omega_{k+1}^{(1)} E + A_1)(y_{k+1/2} - y_k) + Ay_k &= f, \\ (\omega_{k+1}^{(2)} E + A_2)(y_{k+1} - y_k) - (\omega_{k+1}^{(2)} E - A_1)(y_{k+1/2} - y_k) + Ay_k &= f \end{aligned} \quad (5)$$

и из первого равенства вычтем второе. Получим

$$y_{k+1/2} - y_k = (\omega_{k+1}^{(2)} E + A_2) \frac{y_{k+1} - y_k}{\omega_{k+1}^{(1)} + \omega_{k+1}^{(2)}} = (\omega_{k+1}^{(2)} E + A_2) \frac{y_{k+1} - y_k}{\tau_{k+1}}.$$

Подставляя это выражение в (5), получим схему (2). Обратный переход очевиден.

Для нахождения y_{k+1} можно использовать и другой алгоритм, трактуя (2) как схему с поправкой w_k ,

$$\begin{aligned} (\omega_{k+1}^{(1)} E + A_1) v &= r_k, \quad r_k = Ay_k - f, \\ (\omega_{k+1}^{(2)} E + A_2) w_k &= v, \\ y_{k+1} &= y_k - \tau_{k+1} w_k, \quad k = 0, 1, \dots \end{aligned}$$

Этот алгоритм более экономичен по сравнению с (4), однако требует запоминать больше промежуточной информации, т. е. требует дополнительной памяти ЭВМ, что не всегда удобно.

Отметим, что как при построении итерационной схемы (3), так и при конструировании алгоритмов никакие условия, кроме естественного предположения о невырожденности операторов $\omega_k^{(\alpha)}E + A_\alpha$, $\alpha = 1, 2$, на операторы A_1 и A_2 не накладывались. Все дополнительные требования на операторы A_1 и A_2 связаны с задачей об оптимальном выборе параметров $\omega_k^{(1)}$ и $\omega_k^{(2)}$.

2. Постановка задачи о выборе параметров. В методе переменных направлений мы имеем дело с двумя последовательностями параметров $\{\omega_k^{(1)}\}$ и $\{\omega_k^{(2)}\}$, которые будем выбирать из условия минимума нормы разрешающего оператора в исходном пространстве H .

Для решения задачи о выборе итерационных параметров необходимо сделать определенные предположения относительно операторов A_1 и A_2 , которые имеют функциональный характер, а также задать некоторую априорную информацию. Сформулируем эти предположения.

Будем предполагать, что оператор A можно представить в виде суммы двух самосопряженных и перестановочных операторов A_1 и A_2 :

$$A = A_1 + A_2, \quad A_1 = A_1^*, \quad A_2 = A_2^*, \quad A_1 A_2 = A_2 A_1. \quad (6)$$

Пусть априорная информация задана в виде границ δ_α и Δ_α оператора A_α , $\alpha = 1, 2$, т. е.

$$\delta_1 E \leq A_1 \leq \Delta_1 E, \quad \delta_2 E \leq A_2 \leq \Delta_2 E, \quad (7)$$

причем выполнено условие

$$\delta_1 + \delta_2 > 0. \quad (8)$$

Заметим, что из (6)–(8) следует самосопряженность и положительная определенность оператора A .

Если предположение о перестановочности операторов A_1 и A_2 выполнено, то будем говорить, что рассматривается *коммутативный случай*, иначе — *общий случай*. Условия (6) обеспечивают самосопряженность операторов B_k для любого k . Действительно, в силу (6) операторы $\omega_k^{(1)}E + A_1$ и $A_2 + \omega_k^{(2)}E$ самосопряжены и перестановочны, а произведение самосопряженных и перестановочных операторов есть самосопряженный оператор.

Переходим к изучению сходимости итерационной схемы (2). Подставляя $y_k = z_k + u$, где z_k — погрешность, а u — решение уравнения (1), в (2), получим для z_k однородное уравнение

$$z_{k+1} = S_{k+1} z_k, \quad k = 0, 1, \dots, \quad z_0 = y_0 - u, \quad (9)$$

где

$$S_k = E - \tau_k B_k^{-1} A = (\omega_k^{(2)} E + A_2)^{-1} (\omega_k^{(1)} E + A_1)^{-1} (\omega_k^{(2)} E - A_1) (\omega_k^{(1)} E - A_2). \quad (10)$$

Используя (9), выразим z_n через z_0 . Получим

$$z_n = T_{n,0} z_0, \quad T_{n,0} = \prod_{j=1}^n S_j = S_n S_{n-1} \dots S_1, \quad (11)$$

где $T_{n,0}$ — разрешающий оператор. Так как операторы A_1 и A_2 перестановочны, то порядок сомножителей в (10) безразличен, все операторы S_k самосопряжены и попарно перестановочны и, следовательно, оператор $T_{n,0}$ самосопряжен в H : $T_{n,0} = R_n(A_1, A_2)$, где $R_n(x, y)$ есть произведение дробно-рациональных функций от x и y :

$$R_n(x, y) = \prod_{j=1}^n \frac{\omega_j^{(2)} - x}{\omega_j^{(1)} + x} \frac{\omega_j^{(1)} - y}{\omega_j^{(2)} + y}. \quad (12)$$

Из (11) получим

$$\|z_n\| \leq \|T_{n,0}\| \|z_0\|. \quad (13)$$

В силу самосопряженности оператора $T_{n,0}$ имеем $\|T_{n,0}\| = \max_k |\lambda_k(T_{n,0})|$, где $\lambda_k(T_{n,0})$ — собственные значения оператора $T_{n,0}$. Далее, в силу условий (6) (см. п. 5 § 1 гл. V) операторы A_1 , A_2 и $T_{n,0}$ имеют общую систему собственных функций. Поэтому

$$\lambda_k(T_{n,0}) = R_n(\lambda_{k_1}^{(1)}, \lambda_{k_2}^{(2)}),$$

где $\lambda_{k_1}^{(1)}$ и $\lambda_{k_2}^{(2)}$ — собственные значения операторов A_1 и A_2 соответственно, причем в силу (7) имеем $\delta_1 \leq \lambda_{k_1}^{(1)} \leq \Delta_1$, $\delta_2 \leq \lambda_{k_2}^{(2)} \leq \Delta_2$. Следовательно,

$$\|T_{n,0}\| = \max_{k_1, k_2} |R_n(\lambda_{k_1}^{(1)}, \lambda_{k_2}^{(2)})| \leq \max_{\substack{\delta_1 \leq x \leq \Delta_1 \\ \delta_2 \leq y \leq \Delta_2}} |R_n(x, y)|.$$

Подставляя эту оценку в (13), получим

$$\|z_n\|_D \leq \max_{\substack{\delta_1 \leq x \leq \Delta_1 \\ \delta_2 \leq y \leq \Delta_2}} |R_n(x, y)| \|z_0\|_D, \quad (14)$$

где $R_n(x, y)$ определено в (12), а $D = E$. Заметим, что в силу перестановочности операторов A_1 и A_2 оператор $T_{n,0}$ будет самосопряжен и в энергетическом пространстве H_D для $D = A$, A^2 . Поэтому в силу леммы 5 § 1 гл. V будем иметь $\|T_{n,0}\| = \|T_{n,0}\|_A = \|T_{n,0}\|_{A^2}$, и, следовательно, оценка (14) верна для $D = A$, $D = A^2$.

Итак, задача оценки погрешности итерационной схемы (2) сведена к задаче об оценке максимума модуля функции двух

переменных $R_n(x, y)$ в прямоугольнике $G = \{\delta_1 \leq x \leq \Delta_1, \delta_2 \leq y \leq \Delta_2\}$ и выборе итерационных параметров из условия минимума максимума модуля этой функции. Поставленная задача является достаточно сложной и в п. 3 она будет сведена к более простой задаче о нахождении дробно-рациональной функции одной переменной, наименее уклоняющейся от нуля на отрезке.

3. Дробно-линейное преобразование. Изучим функцию $R(x, y)$. При помощи дробно-линейного преобразования неизвестных отобразим прямоугольник G на квадрат $\{\eta \leq u \leq 1, \eta \leq v \leq 1, \eta > 0\}$, причем преобразование выберем таким образом, чтобы оно не меняло вида функции $R_n(x, y)$. Искомое преобразование имеет вид

$$x = \frac{ru - s}{1 - tu}, \quad y = \frac{rv + s}{1 + tv}, \quad \eta \leq u, v \leq 1, \quad (15)$$

где постоянные r, s, t и η подлежат определению.

Подставляя (15) в (12) и вводя новые параметры $\kappa_j^{(1)}$ и $\kappa_j^{(2)}$,

$$\kappa_j^{(1)} = \frac{\omega_j^{(1)} - s}{r - t\omega_j^{(1)}}, \quad \kappa_j^{(2)} = \frac{\omega_j^{(2)} + s}{r + t\omega_j^{(2)}}, \quad j = 1, 2, \dots, n, \quad (16)$$

получим

$$R_n(x, y) = P_n(u, v) = \prod_{j=1}^n \frac{\kappa_j^{(2)} - u}{\kappa_j^{(1)} + u} \frac{\kappa_j^{(1)} - v}{\kappa_j^{(2)} + v}.$$

Из (16) найдем соотношения, при помощи которых параметры $\omega_j^{(1)}$ и $\omega_j^{(2)}$ выражаются через введенные параметры $\kappa_j^{(1)}$ и $\kappa_j^{(2)}$:

$$\omega_j^{(1)} = \frac{r\kappa_j^{(1)} + s}{1 + t\kappa_j^{(1)}}, \quad \omega_j^{(2)} = \frac{r\kappa_j^{(2)} - s}{1 - t\kappa_j^{(2)}}, \quad j = 1, 2, \dots, n. \quad (17)$$

Итак, если будут найдены параметры $\kappa_j^{(1)}$ и $\kappa_j^{(2)}$, то по формулам (17) определяются параметры $\omega_j^{(1)}$ и $\omega_j^{(2)}$.

В силу замены (15) мы приходим к задаче отыскания таких значений параметров $\kappa_j^{(1)}$ и $\kappa_j^{(2)}$, при которых достигается

$$\min_{\kappa^{(1)}, \kappa^{(2)}} \max_{\eta \leq u, v \leq 1} |P_n(u, v)|.$$

Заметим, что если наложить некоторые ограничения на выбор параметров $\kappa_j^{(1)}$ и $\kappa_j^{(2)}$, например $\kappa_j^{(1)} = \kappa_j^{(2)} = \kappa_j$, то, очевидно, что минимум может только увеличиться. Поэтому

$$\begin{aligned} \min_{\kappa^{(1)}, \kappa^{(2)}} \max_{\eta \leq u, v \leq 1} |P_n(u, v)| &\leq \min_{\kappa} \max_{\eta \leq u, v \leq 1} \left| \prod_{j=1}^n \frac{\kappa_j - u}{\kappa_j + u} \frac{\kappa_j - v}{\kappa_j + v} \right| = \\ &= \min_{\kappa} \max_{\eta \leq u \leq 1} |r_n(u, \kappa)|^2, \quad r_n(u, \kappa) = \prod_{j=1}^n \frac{\kappa_j - u}{\kappa_j + u}. \end{aligned}$$

Итак, поставленная выше задача об оптимальном выборе итерационных параметров $\omega_j^{(1)}$ и $\omega_j^{(2)}$ сведена к нахождению дробно-рациональной функции

$r_n(u, \kappa)$, которая наименее уклоняется от нуля на отрезке $[\eta, 1]$. Иначе, нужно найти такие κ^* , при которых

$$\max_{\eta \leq u \leq 1} |r_n(u, \kappa^*)| = \min_{\kappa} \max_{\eta \leq u \leq 1} |r_n(u, \kappa)| = \rho.$$

Если такие параметры найдены, то для погрешности z_n из (14) будет следовать оценка $\|z_n\|_D \leq \rho^2 \|z_0\|_D$, и точность ε будет достигнута, если положить $\rho^2 = \varepsilon$.

Искомый выбор итерационных параметров будет дан в п. 4, а здесь мы найдем постоянные r, s, t и η преобразования (15).

Если $r \neq ts$, то преобразование (15) монотонно по u и v , а, следовательно, обратное преобразование $u = (x+s)/(r+tx)$, $v = (y-s)/(r-ty)$ будет монотонно по x и y . Поэтому для отображения прямоугольника $\{\delta_1 \leq x \leq \Delta_1, \delta_2 \leq y \leq \Delta_2\}$ на квадрат $\{\eta \leq u, v \leq 1\}$ достаточно, чтобы концы отрезка $[\delta_\alpha, \Delta_\alpha]$ переходили в концы отрезка $[\eta, 1]$. Это дает четыре соотношения для определения постоянных преобразования (15):

$$\delta_1 = \frac{r\eta - s}{1 - t\eta}, \quad \delta_2 = \frac{r\eta + s}{1 + t\eta}, \quad \Delta_1 = \frac{r - s}{1 - t}, \quad \Delta_2 = \frac{r + s}{1 + t}. \quad (18)$$

Найдем решение нелинейной системы (18). Заметим сначала, что в силу предположения (8) справедливы неравенства

$$\Delta_2 + \delta_1 \geq \delta_1 + \delta_2 > 0, \quad \Delta_1 + \delta_2 \geq \delta_1 + \delta_2 > 0. \quad (19)$$

Далее, из (18) получим

$$\begin{aligned} \Delta_1 - \delta_1 &= \frac{(1-\eta)(r-st)}{(1-t)(1-t\eta)}, & \Delta_2 - \delta_2 &= \frac{(1-\eta)(r-st)}{(1+t)(1+t\eta)}, \\ \Delta_2 + \delta_1 &= \frac{(1+\eta)(r-st)}{(1+t)(1-t\eta)}, & \Delta_1 + \delta_2 &= \frac{(1+\eta)(r-st)}{(1-t)(1+t\eta)}. \end{aligned} \quad (20)$$

Отсюда найдем

$$\left(\frac{1-\eta}{1+\eta}\right)^2 = \frac{(\Delta_1 - \delta_1)(\Delta_2 - \delta_2)}{(\Delta_1 + \delta_2)(\Delta_2 + \delta_1)} < 1,$$

и так как в силу (19) знаменатель в нуль не обращается, то

$$\eta = \frac{1-a}{1+a}, \quad a = \sqrt{\frac{(\Delta_1 - \delta_1)(\Delta_2 - \delta_2)}{(\Delta_1 + \delta_2)(\Delta_2 + \delta_1)}}, \quad \eta \in [0, 1]. \quad (21)$$

Найдем теперь t . Из (20) получим

$$\frac{\Delta_2 + \delta_1}{\Delta_1 - \delta_1} = \frac{1+\eta}{1-\eta} \frac{1-t}{1+t} = \frac{1}{a} \frac{1-t}{1+t}.$$

Отсюда будем иметь

$$t = \frac{1-b}{1+b}, \quad b = \frac{\Delta_2 + \delta_1}{\Delta_1 - \delta_1} a. \quad (22)$$

Из двух последних уравнений системы (18) найдем

$$r = \frac{1}{2} [\Delta_1(1-t) + \Delta_2(1+t)] = \frac{1+t}{2} [\Delta_2 + \Delta_1 b] = \frac{\Delta_2 + \Delta_1 b}{1+b}, \quad (23)$$

$$s = \frac{1}{2} [\Delta_2(1+t) - \Delta_1(1+t)] = \frac{1+t}{2} [\Delta_2 - \Delta_1 b] = \frac{\Delta_2 - \Delta_1 b}{1+b}. \quad (24)$$

Так как

$$r - st = \frac{2b(\Delta_1 + \Delta_2)}{(1+b)^2} > 0, \quad |t| < 1,$$

то преобразование (15) действительно монотонно. В п. 4 будет показано, что $\eta < \kappa_j = \kappa_j^{(1)} = \kappa_j^{(2)} < 1$. Поэтому в (17) знаменатели в нуль не обращаются.

Рассмотрим некоторые примеры. Пусть $\delta_1 = \delta_2 = \delta$ и $\Delta_1 = \Delta_2 = \Delta$, т. е. границы операторов A_1 и A_2 одинаковы. Тогда $\eta = \delta/\Delta$, $t = s = 0$, $r = \Delta$, $\omega_j^{(1)} = \omega_j^{(2)} = \Delta\kappa_j$. Пусть теперь $\delta_1 = 0$, $\delta_2 = \delta$, $\Delta_1 = \Delta_2 = \Delta$, т. е. оператор A_1 вырожденный. Тогда

$$\eta = \delta/(\Delta + \sqrt{\Delta^2 - \delta^2}), \quad t = \eta, \quad s = \Delta\eta, \quad r = \Delta,$$

$$\omega_j^{(1)} = \frac{\Delta\kappa_j + \Delta\eta}{1 + \eta\kappa_j}, \quad \omega_j^{(2)} = \frac{\Delta\kappa_j - \Delta\eta}{1 - \eta\kappa_j}, \quad j = 1, 2, \dots, n.$$

4. Оптимальный набор параметров. Приведем решение задачи об оптимальном выборе итерационных параметров. В отличие от случая нахождения полинома, наименее уклоняющегося от нуля, который был рассмотрен в § 2 гл. VI, здесь итерационные параметры κ_j выражаются не через тригонометрические функции, а при помощи эллиптических функций Якоби.

Напомним некоторые определения. Определенный интеграл

$$K(k) = \int_0^{\pi/2} \frac{d\varphi}{\sqrt{1 - k^2 \sin^2 \varphi}}$$

называется *полным эллиптическим интегралом первого рода*, число k — *модулем* этого интеграла, а число $k' = \sqrt{1 - k^2}$ — *дополнительным модулем*. Принято обозначать $K(k') = K'(k)$.

Если обозначить через $u(z, k)$ функцию

$$u(z, k) = \int_z^1 \frac{dy}{\sqrt{(1-y^2)(y^2-k^2)}},$$

то функция $z = \operatorname{dn}(u, k')$, обратная к $u(z, k)$, называется *эллиптической функцией Якоби* аргумента u и модуля k' .

Используя эти обозначения, точное решение задачи об оптимальном выборе итерационных интегральных параметров κ_j можно записать в следующем виде:

$$\kappa_j \in \mathfrak{M}_n = \left\{ \mu_i = \operatorname{dn} \left(\frac{2i-1}{2n} K'(\eta), \eta' \right), \quad i = 1, 2, \dots, n \right\}, \quad (25)$$

$$j = 1, 2, \dots, n,$$

где число итераций n , достаточное для достижения точности ε , оценивается по формуле

$$n \geq n_0(\varepsilon) = \frac{1}{4} \frac{K'(\eta)}{K(\eta)} \frac{K'(\varepsilon)}{K(\varepsilon)}. \quad (26)$$

Здесь, как и в чебышевском методе, в качестве κ_j последовательно выбираются все элементы множества \mathfrak{M}_n . Сформулируем полученные результаты для метода переменных направлений в коммутативном случае в виде теоремы.

Теорема 1. Пусть выполнены условия (6)–(8), а параметры $\omega_j^{(1)}$ и $\omega_j^{(2)}$ выбраны по формулам

$$\omega_j^{(1)} = \frac{r\kappa_j + s}{1 + t\kappa_j}, \quad \omega_j^{(2)} = \frac{r\kappa_j - s}{1 - t\kappa_j}, \quad j = 1, 2, \dots, n, \quad (27)$$

где κ_j и n определены в (25), (26), а r, s, t и η – в (21)–(24). Метод переменных направлений (2), (3) сходится в H_D , и после выполнения n итераций для погрешности $z_n = y_n$ – и будет верна оценка $\|z_n\|_D \leq \varepsilon \|z_0\|_D$, где $D = E, A$ или A^2 , а n определяется согласно (26).

Обратимся теперь к вычислительной стороне вопроса реализации метода переменных направлений с оптимальным набором параметров. Найдем приближенные формулы вычисления κ_j и n и укажем порядок, в котором следует выбирать параметры κ_j из множества \mathfrak{M}_n .

Используя асимптотическое представление для полных эллиптических интегралов при малых значениях k :

$$\frac{1}{K(k)} = \frac{2}{\pi} + O(k^2), \quad K'(k) = \ln \frac{4}{k} + O\left(k^2 \ln \frac{1}{k}\right),$$

из (26) получим следующую приближенную формулу для числа итераций n :

$$n \geq n_0(\varepsilon) = \frac{1}{\pi^2} \ln \frac{4}{\eta} \ln \frac{4}{\varepsilon}. \quad (28)$$

Рассмотрим теперь вопрос о вычислении μ_i . Функция $\operatorname{dn}(u, k')$ монотонно убывает по u , принимая следующие значения: $\operatorname{dn}(0, k') = 1$, $\operatorname{dn}(K'(k), k') = K$. Поэтому $\eta < \mu_n < \mu_{n-1} < \dots < \mu_1 < 1$. Далее, из свойства эллиптической функции $\operatorname{dn}(u, k')$:

$$\operatorname{dn}(u, k') = \frac{k}{\operatorname{dn}(K'(k) - u, k')}$$

следует, что имеет место равенство

$$\mu_i = \eta / \mu_{n+1-i}, \quad i = 1, 2, \dots \quad (29)$$

Поэтому достаточно найти половину значений μ_i , а остальные определить из соотношения (29).

Приближенную формулу для μ_i получим, используя разложение функции $\operatorname{dn}(u, k')$ по степеням k . Для этого выразим функцию $\operatorname{dn}(\sigma K'(\eta), \eta')$ через тета-функции Якоби, а эти функции представим рядами. Получим

$$\operatorname{dn}(\sigma K'(\eta), \eta') = \frac{\sqrt{\eta} \theta_3 \left(\frac{i\sigma\pi K'}{K}, \bar{q} \right)}{\theta_2 \left(\frac{i\sigma\pi K'}{K'}, \bar{q} \right)} = \sqrt{\eta} q^{\frac{2\sigma-1}{4}} \frac{\sum_{m=-\infty}^{\infty} \bar{q}^{m(m+\sigma)}}{\sum_{m=-\infty}^{\infty} \bar{q}^{m(m-1+\sigma)}},$$

где $\bar{q} = \exp \left(-\frac{\pi K'(\eta)}{K(\eta)} \right) = \frac{\eta^2}{16} \left(1 + \frac{\eta^2}{2} \right) + O(\eta^6)$.

Отсюда находим

$$\operatorname{dn}(\sigma K'(\eta), \eta') = \sqrt{\eta} q^{\frac{2\sigma-1}{4}} \frac{1 + q^{1-\sigma} + q^{1+\sigma}}{1 + q^\sigma + q^{2-\sigma}} + O(\eta^6), \quad (30)$$

)

где

$$q = \eta^2 (1 + \eta^2/2)/16, \quad v = \begin{cases} 4 + 5\sigma, & 0 < \sigma < 1/2, \\ 8 - 3\sigma, & 1/2 \leq \sigma < 1. \end{cases}$$

При $\sigma \geq 1/2$ порядок остаточного члена в (30) равномерно по σ равен 5, а при $\sigma < 1/2$ порядок равен 4. Поэтому приближенная формула для $\operatorname{dn}(\sigma K'(\eta), \eta')$ будет более точна для $\sigma \geq 1/2$, чем для $\sigma < 1/2$.

Из (25), (29) и (30) получим следующие формулы для вычисления μ_i :

$$\begin{aligned} \mu_i &= \sqrt{\eta} q^{\frac{2\sigma_i-1}{4}} \frac{1+q^{1-\sigma_i}+q^{1+\sigma_i}}{1+q^{\sigma_i}+q^{2-\sigma_i}}, \quad [n/2]+1 \leq i \leq n, \\ \mu_i &= \eta/\mu_{n+1-i}, \quad 1 \leq i \leq [n/2], \quad \sigma_i = (2i-1)/(2n), \\ q &= \eta^2 (1 + \eta^2/2)/16, \end{aligned}$$

где $[a]$ — целая часть a .

Рассмотрим теперь вопрос о порядке выбора x_j из множества \mathfrak{M}_n . Из определения оператора перехода S_j в схеме (2) и свойств (6), (7) получим

$$\|S_j\| = \max_k |\lambda_k(S_j)| \leq \max_{\substack{\delta_1 \leq x \leq \Delta_1 \\ \delta_2 \leq y \leq \Delta_2}} \left| \frac{\omega_j^{(2)} - x}{\omega_j^{(1)} + x} \frac{\omega_j^{(1)} - y}{\omega_j^{(2)} + y} \right|$$

или в силу замены (15)

$$\|S_j\| \leq \max_{\eta \leq u \leq 1} \left| \frac{x_j - u}{x_j + u} \right|^2.$$

Так как все x_j принадлежат интервалу $(\eta, 1)$, то отсюда следует, что $\|S_j\| < 1$ для любого j . Поэтому итерационный метод (2), (3) будет устойчив к ошибкам округления при любом порядке выбора x_j из множества \mathfrak{M}_n , например, $x_j = \mu_j$, $j = 1, 2, \dots, n$.

В заключение этого пункта покажем, что для построенного набора параметров $\omega_j^{(1)}$ и $\omega_j^{(2)}$ операторы $\omega_j^{(\alpha)} E + A_\alpha$, $\alpha = 1, 2$, для любого j положительно определены в H . Действительно, из (27) получим

$$\frac{\partial \omega_j^{(1)}}{\partial x_j} = \frac{r-st}{(1+tx_j)^2} > 0, \quad \frac{\partial \omega_j^{(2)}}{\partial x_j} = \frac{r-st}{(1-tx_j)^2} > 0.$$

Так как знаменатели в (27) не обращаются в нуль и $\eta < x_j < 1$, то отсюда и из (18) найдем

$$\delta_2 = \frac{r\eta+s}{1+t\eta} \leq \omega_j^{(1)} \leq \frac{r+s}{1+t} = \Delta_2, \quad \delta_1 = \frac{r\eta-s}{1-t\eta} \leq \omega_j^{(2)} \leq \frac{r-s}{1-t} = \Delta_1. \quad (31)$$

Следовательно, в силу предположения (7) из (31) получим

$$\omega_j^{(1)} E + A_1 \geq (\delta_1 + \delta_2) E, \quad \omega_j^{(2)} E + A_2 \geq (\delta_1 + \delta_2) E,$$

и так как $\delta_1 + \delta_2 > 0$ по предположению (8), то утверждение доказано.

§ 2. Примеры применения метода

1. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Рассмотрение примеров применения метода переменных направлений начнем с решения разностной задачи Дирихле для уравнения Пуассона в прямоугольнике.

Пусть на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$, введенной в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, требуется найти решение задачи

$$\begin{aligned}\Lambda y &= (\Lambda_1 + \Lambda_2)y = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \\ \Lambda_\alpha y &= y_{x_\alpha x_\alpha}, \quad \alpha = 1, 2.\end{aligned}\quad (1)$$

Обозначим через H пространство сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2.$$

Операторы A , A_1 и A_2 определим на H следующим образом:

$Ay = -\Lambda \dot{y}$, $A_\alpha y = -\Lambda_\alpha \dot{y}$, $\alpha = 1, 2$, где $y \in H$, $\dot{y} \in \dot{H}$, $y(x) = \dot{y}(x)$ для $x \in \omega$, а \dot{H} — множество сеточных функций, заданных на $\bar{\omega}$ и обращающихся в нуль на γ .

Разностная задача (1) может быть тогда записана в виде операторного уравнения $Au = f$, где $A = A_1 + A_2$.

Как мы знаем (см. п. 5 § 1 гл. V), операторы A_α самосопряжены в H и имеют границы δ_α и Δ_α :

$$\delta_\alpha = \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta_\alpha = \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \alpha = 1, 2,$$

которые совпадают с минимальным и максимальным собственными значениями разностных операторов Λ_α . Осталось проверить выполнение условия перестановочности операторов A_1 и A_2 . Используя определение операторов A_α и разностных операторов Λ_α , получим

$$A_1 A_2 y = \dot{y}_{\bar{x}_1 \bar{x}_1 \bar{x}_2 \bar{x}_2} = \dot{y}_{\bar{x}_2 \bar{x}_2 \bar{x}_1 \bar{x}_1} = A_2 A_1 y,$$

что и требовалось доказать.

Итак, условия, требуемые для применения метода переменных направлений в коммутативном случае, для рассматриваемого примера выполнены.

Используя определение операторов A_1 и A_2 , алгоритм метода переменных направлений для рассматриваемого примера можно записать в следующем виде:

$$\omega_{k+1}^{(1)} y_{k+1/2} - \Lambda_1 y_{k+1/2} = \omega_{k+1}^{(1)} y_k + \Lambda_2 y_k + \varphi, \quad h_1 \leq x_1 \leq l_1 - h_1, \quad (2)$$

$$y_{k+1/2}(x) = g(x), \quad x_1 = 0, \quad l_1, \quad h_2 \leq x_2 \leq l_2 - h_2,$$

$$\omega_{k+1}^{(2)} y_{k+1} - \Lambda_2 y_{k+1} = \omega_{k+1}^{(2)} y_{k+1/2} + \Lambda_1 y_{k+1/2} + \varphi, \quad h_2 \leq x_2 \leq l_2 - h_2, \quad (3)$$

$$y_{k+1}(x) = g(x), \quad x_2 = 0, \quad l_2, \quad h_1 \leq x_1 \leq l_1 - h_1,$$

причем $y_k(x) = g(x)$ при $x \in \gamma$ для любого $k \geq 0$. Таким образом, алгоритм метода состоит в последовательном решении для каж-

дого фиксированного x_2 трехточечных краевых задач (2) по направлению x_1 для определения $y_{k+1/2}$ на ω и решении для каждого x_1 краевых задач (3) по направлению x_2 для определения нового итерационного приближения y_{k+1} на ω . Чередование направлений, по которым решаются краевые задачи (2), (3), дало название метода — метод переменных направлений.

Для нахождения решения задач (2), (3) можно воспользоваться методом прогонки. Запишем уравнения (2), (3) в виде трехточечной системы и проверим выполнение достаточных условий устойчивости метода прогонки. Уравнения будут иметь вид

$$-y_{k+1/2}(i+1, j) + (2 + h_1^2 \omega_{k+1}^{(1)}) y_{k+1/2}(i, j) - y_{k+1/2}(i-1, j) = \\ = \varphi_1(i, j), \quad 1 \leq i \leq N_1 - 1, \quad (4)$$

$$y_{k+1/2}(0, j) = g(0, j), \quad y_{k+1/2}(N_1, j) = g(N_1, j), \\ 1 \leq j \leq N_2 - 1,$$

где

$$\varphi_1(i, j) = \frac{h_2^2}{h_1^2} [y_k(i, j+1) - (2 + h_2^2 \omega_{k+1}^{(1)}) y_k(i, j) + \\ + y_k(i, j-1) + h_2^2 \varphi(i, j)]; \\ -y_{k+1}(i, j+1) + (2 + h_2^2 \omega_{k+1}^{(2)}) y_{k+1}(i, j) - y_{k+1}(i, j-1) = \varphi_2(i, j), \\ 1 \leq j \leq N_2 - 1, \quad (5)$$

$$y_{k+1}(i, 0) = g(i, 0), \quad y_{k+1}(i, N_2) = g(i, N_2), \quad 1 \leq i \leq N_1 - 1,$$

где

$$\varphi_2(i, j) = \frac{h_2^2}{h_1^2} [y_{k+1/2}(i+1, j) - \\ - (2 + h_1^2 \omega_{k+1}^{(2)}) y_{k+1/2}(i, j) + y_{k+1/2}(i-1, j) + h_1^2 \varphi(i, j)].$$

Так как для данного примера $\delta_1 > 0$, $\delta_2 > 0$, то в силу неравенств (31) п. 4 § 1 параметры $\omega_k^{(1)}$ и $\omega_k^{(2)}$ положительны. Поэтому в трехточечных уравнениях (4) и (5) коэффициенты при $y_{k+1/2}(i, j)$ и $y_{k+1}(i, j)$ преобладают над остальными коэффициентами. Следовательно, метод прогонки, примененный к задачам (4), (5), будет устойчив по отношению к ошибкам округления.

Подсчитаем число арифметических действий, которое нужно затратить на реализацию одного итерационного шага в методе (2), (3) для рассматриваемого примера. Достаточно подсчитать число действий для задачи (4), для задачи (5) подсчет проводится аналогично.

Формулы метода прогонки для задачи (4) имеют вид (j фиксировано):

$$y_{k+1/2}(i, j) = \alpha_i y_{k+1/2}(i+1, j) + \beta_i, \quad 1 \leq i \leq N_1 - 1, \\ y_{k+1/2}(N_1, j) = g(N_1, j),$$

$$\alpha_{i+1} = 1/(C - \alpha_i), \quad i = 1, 2, \dots, N_1 - 1, \quad \alpha_1 = 0, \quad C = 2 + h_1^2 \omega_{k+1}^{(1)}, \\ \beta_{i+1} = \alpha_{i+1} (\varphi_1(i, j) + \beta_i), \\ i = 1, 2, \dots, N_1 - 1, \quad \beta_1 = g(0, j).$$

Заметим, что прогоночные коэффициенты α_i не зависят от j и поэтому могут быть вычислены один раз с затратой $2(N_1 - 1)$ арифметических операций. Далее, на вычисление $\varphi_1(i, j)$ на сетке ω потребуется $6(N_1 - 1)(N_2 - 1)$ арифметических операций. Прогоночные коэффициенты β_i и решение $y_{k+1/2}$ нужно считать заново при каждом j . Для этого потребуется $4(N_1 - 1)(N_2 - 1)$ действий. Всего для нахождения $y_{k+1/2}$ на сетке ω при заданном y_k потребуется $Q_1 = 10(N_1 - 1)(N_2 - 1) + 2(N_1 - 1)$ арифметических действий. Для нахождения y_{k+1} из (15) по вычисленному $y_{k+1/2}$ потребуется $Q_2 = 10(N_1 - 1)(N_2 - 1) + 2(N_2 - 1)$ действий. Итак, для рассматриваемого примера реализация одного итерационного шага в методе переменных направлений осуществляется за

$$Q = 20(N_1 - 1)(N_2 - 1) + 2(N_1 - 1) + 2(N_2 - 1) \quad (6)$$

арифметических действий.

Оценим теперь число итераций n , достаточное для получения решения с заданной точностью ε . В частном случае, когда область \bar{G} есть квадрат со стороной l ($l_1 = l_2 = l$) и сетка ω квадратная с $N_1 = N_2 = N$ ($h_1 = h_2 = l/N$), будем иметь

$$\delta_1 = \delta_2 = \delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \Delta_1 = \Delta_2 = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l}.$$

Из (21) и (28) § 1 получим следующую оценку для числа итераций:

$$n \geq n_0(\varepsilon) = 0,1 \ln \frac{4}{\eta} \ln \frac{4}{\varepsilon}, \quad \eta = \delta/\Delta = \operatorname{tg}^2 \frac{\pi h}{2l}$$

или для малых h

$$n_0(\varepsilon) = 0,2 \ln(4N/\pi) \ln(4/\varepsilon), \quad (7)$$

т. е. число итераций пропорционально логарифму от числа неизвестных N по одному направлению.

Из (6), (7) получим следующую оценку для числа арифметических действий $Q(\varepsilon)$, затрачиваемых на нахождение решения разностной задачи (1) методом переменных направлений с точностью ε :

$$Q(\varepsilon) = nQ = 4N^2 \ln(4N/\pi) \ln(4/\varepsilon). \quad (8)$$

Чтобы сравнить этот метод с прямым методом полной редукции (см. § 3 гл. III), перейдем в (8) от натуральных логарифмов к логарифмам по основанию 2.

Получим

$$Q(\varepsilon) \approx 2,12 N^2 \log_2 (4N/\pi) \log_2 (4/\varepsilon).$$

Так как погрешность аппроксимации разностной схемы (1) есть $O(h^2)$, то ε целесообразно выбирать равным $O(h^2)$.

Если взять $\varepsilon = 4/N^2$, то получим

$$Q(\varepsilon) = 4,24 N^2 \log_2 N \log_2 (4N/\pi).$$

При $N = 64$ получим $\varepsilon \approx 10^{-3}$ и

$$Q(\varepsilon) \approx 27,6 N^2 \log_2 N.$$

Сравнение с оценкой числа действий для метода полной редукции показывает, что для указанной сетки метод переменных направлений требует примерно в 5,5 раза больше арифметических действий, чем метод полной редукции. С увеличением N и уменьшением ε это различие увеличивается.

Для рассматриваемого частного случая приведем число итераций n в зависимости от числа узлов N по одному направлению для $\varepsilon = 10^{-4}$.

Для сравнения приведем и число итераций для других рассмотренных выше методов.

Таблица 11

N	Метод простой итерации	Явный чебышевский метод	Метод верхней релаксации	Попеременно-треугольный метод	Метод переменных направлений
32	1909	101	65	16	8
64	7642	202	128	23	10
128	30577	404	257	32	11

Из таблицы следует, что наименьшее число итераций требуется для метода переменных направлений. По числу итераций ему уступает попеременно-треугольный метод с чебышевскими параметрами, который был рассмотрен в главе X.

Замечание. Если для рассматриваемой задачи (1) рассмотреть метод переменных направлений с постоянными параметрами, т. е. $\omega_j^{(1)} \equiv \omega^{(1)}$, $\omega_j^{(2)} \equiv \omega^{(2)}$, $\tau_j = \omega^{(1)} + \omega^{(2)}$, то из формулы (25) § 1 получим в силу равенства $\operatorname{dn} \left(\frac{1}{2} K'(k), k' \right) = \sqrt{k}$, что параметр $\kappa_j \equiv \sqrt{\eta_j}$. В частном случае, когда $\delta_1 = \delta_2 = \delta$, $\Delta_1 =$

$=\Delta_2=\Delta$, ранее в п. 3 § 1 было получено следующее соотношение между параметрами $\omega_j^{(1)}$, $\omega_j^{(2)}$ и κ_j : $\omega_j^{(1)}=\omega_j^{(2)}=\Delta\kappa_j$. Так как при этом $\eta=\delta/\Delta$, то отсюда находим $\omega^{(1)}=\omega^{(2)}=\sqrt{\delta\Delta}$.

2. Третья краевая задача для эллиптического уравнения с разделяющимися переменными. Пусть в прямоугольнике $\bar{G}=\{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha=1, 2\}$ требуется найти решение следующей краевой задачи:

$$\begin{aligned} Lu &= \sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x_\alpha) \frac{\partial u}{\partial x_\alpha} \right) - qu = -f(x), \quad x \in G, \\ k_\alpha(x_\alpha) \frac{\partial u}{\partial x_\alpha} &= \kappa_{-\alpha} u - g_{-\alpha}(x), \quad x_\alpha = 0, \\ -k_\alpha(x_\alpha) \frac{\partial u}{\partial x_\alpha} &= \kappa_{+\alpha} u - g_{+\alpha}(x), \quad x_\alpha = l_\alpha, \quad \alpha = 1, 2. \end{aligned} \quad (9)$$

Будем предполагать, что выполнены следующие условия:

$$0 \leqslant c_{1, \alpha} \leqslant k_\alpha(x_\alpha) \leqslant c_{2, \alpha}, \quad \kappa_{\pm\alpha} = \text{const} \geqslant 0, \quad \alpha = 1, 2, \quad (10)$$

$$q = \text{const} \geqslant 0, \quad \sum_{\alpha=1}^2 \kappa_{\pm\alpha}^2 + q^2 \neq 0.$$

Краевая задача Неймана ($\kappa_{\pm\alpha}=0$) для случая $q=0$ будет рассмотрена отдельно в главе XII. Условия (10) обеспечивают существование и единственность решения задачи (9).

На прямоугольной сетке $\omega=\{x_{ij}=(ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha=l_\alpha/N_\alpha, \alpha=1, 2\}$ задаче (9) соответствует разностная краевая задача

$$\Lambda y = (\Lambda_1 + \Lambda_2) y = -\varphi(x), \quad x \in \bar{\omega}, \quad (11)$$

где разностные операторы Λ_1 и Λ_2 и правая часть φ определяются следующим образом:

$$\Lambda_\alpha y = \begin{cases} \frac{2}{h_\alpha} a_\alpha(h_\alpha) y_{x_\alpha} - \left(0,5q + \frac{2}{h_\alpha} \kappa_{-\alpha} \right) y, & x_\alpha = 0, \\ (a_\alpha(x_\alpha) y_{x_\alpha})_{x_\alpha} - 0,5qy, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ -\frac{2}{h_\alpha} a_\alpha(l_\alpha) y_{x_\alpha} - \left(0,5q + \frac{2}{h_\alpha} \kappa_{+\alpha} \right) y, & x_\alpha = l_\alpha \end{cases}$$

для $0 \leqslant x_\beta \leqslant l_\beta$, $\beta=3-\alpha$, $\alpha=1, 2$ и $\varphi=f+\varphi_1+\varphi_2$,

$$\varphi_\alpha(x) = \begin{cases} \frac{2}{h_\alpha} g_{-\alpha}(x), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} g_{+\alpha}(x), & x_\alpha = l_\alpha. \end{cases}$$

Обозначим через H пространство сеточных функций, заданных на $\bar{\omega}$, скалярное произведение в котором определено формулой

$$(u, v) = \sum_{x \in \bar{\omega}} u(x)v(x)\tilde{h}_1(x_1)\tilde{h}_2(x_2),$$

$$\tilde{h}_\alpha(x_\alpha) = \begin{cases} 0,5h_\alpha, & x_\alpha = 0, \\ h_\alpha, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha. \end{cases}$$

Операторы A , A_1 и A_2 определим на H соотношениями $Ay = -\Lambda y$, $A_\alpha y = -\Lambda_\alpha y$, $\alpha = 1, 2$. В § 2 гл. V было показано, что так определенные операторы A_1 и A_2 самосопряжены и перестановочные. Кроме того, в силу условий (10) оператор A положительно определен в H (т. е. $\delta_1 + \delta_2 > 0$). Осталось найти границы операторов A_1 и A_2 , т. е. постоянные δ_α и Δ_α в неравенствах $\delta_\alpha E \leq A_\alpha \leq \Delta_\alpha E$, $\alpha = 1, 2$.

Найдем сначала δ_α . Из определения операторов A_α и разностных формул Грина получим

$$\begin{aligned} (A_\alpha y, y) &= - \sum_{x_\beta=0}^{l_\beta} \sum_{x_\alpha=h_\alpha}^{l_\alpha-h_\alpha} [(a_\alpha y_{\bar{x}_\alpha})_{x_\alpha} - 0,5qy] y h_1 \tilde{h}_2 - \\ &\quad - \sum_{x_\beta=0}^{l_\beta} [a_\alpha(h_\alpha) y_{x_\alpha} - (\kappa_{-\alpha} + \frac{h_1}{4} q)y] y \Big|_{x_\alpha=0} \tilde{h}_2 + \\ &\quad + \sum_{x_\beta=0}^{l_\beta} [a_\alpha(l_\alpha) y_{\bar{x}_\alpha} + (\kappa_{+\alpha} + \frac{h_1}{4} q)y] y \Big|_{x_\alpha=l_\alpha} \tilde{h}_2 = \\ &= \sum_{x_\beta=0}^{l_\beta} \sum_{x_\alpha=h_\alpha}^{l_\alpha} a_\alpha y_{\bar{x}_\alpha}^2 h_1 \tilde{h}_2 + \\ &\quad + \sum_{x_\beta=0}^{l_\beta} (\kappa_{-\alpha} y^2 \Big|_{x_\alpha=0} + \kappa_{+\alpha} y^2 \Big|_{x_\alpha=l_\alpha}) \tilde{h}_2 + 0,5q(y^2, 1). \end{aligned}$$

Отсюда найдем, что если $q = \kappa_{-\alpha} = \kappa_{+\alpha} = 0$, то $\delta_\alpha = 0$. Если хотя бы одна из величин q , $\kappa_{-\alpha}$ или $\kappa_{+\alpha}$ отлична от нуля, то δ_α можно найти следующим образом. В силу леммы 16 § 2 гл. V будем иметь

$$(y^2, 1)_{\bar{\omega}_\alpha} \leq \max_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha) (A_\alpha y, y)_{\bar{\omega}_\alpha}, \quad (12)$$

где $v^\alpha(x_\alpha)$ есть решение трехточечной краевой задачи

$$\begin{aligned} (a_\alpha(x_\alpha) v_{\bar{x}_\alpha}^\alpha)_{x_\alpha} - 0,5qv &= -1, \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} a_\alpha(h_\alpha) v_{x_\alpha}^\alpha - \left(0,5q + \frac{2}{h_\alpha} \kappa_{-\alpha}\right) v^\alpha &= -1, \quad x_\alpha = 0, \\ -\frac{2}{h_\alpha} a_\alpha(l_\alpha) v_{\bar{x}_\alpha}^\alpha - \left(0,5q + \frac{2}{h_\alpha} \kappa_{+\alpha}\right) v^\alpha &= -1, \quad x_\alpha = l_\alpha, \end{aligned} \quad (13)$$

а скалярное произведение определяется следующим образом:

$$(u, v)_{\bar{\omega}_\alpha} = \sum_{x_\alpha=0}^{l_\alpha} u(x_\alpha) v(x_\alpha) h_\alpha(x_\alpha).$$

Умножая (12) на $h_\beta(x_\beta)$ и суммируя по x_β от 0 до l_β , получим

$$(y^2, 1) \leq \max_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha) (A_\alpha y, y)$$

и, следовательно,

$$\delta_\alpha = \frac{1}{\max_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha)}, \quad \alpha = 1, 2.$$

Найдем теперь Δ_α . Оператору A_α соответствует трехдиагональная матрица a_α . Обозначим через \mathcal{D} диагональную часть матрицы A_α , т. е. $\mathcal{D}y = d_\alpha(x_\alpha)y$,

$$d_\alpha(x_\alpha) = \begin{cases} 0,5q + \frac{2}{h_\alpha} \kappa_{-\alpha} + \frac{2}{h_\alpha^2} a_\alpha(h_\alpha), & x_\alpha = 0, \\ 0,5q + \frac{1}{h_\alpha^2} (a_\alpha(x_\alpha) + a_\alpha(x_\alpha + h_\alpha)), & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ 0,5q + \frac{2}{h_\alpha} \kappa_{+\alpha} + \frac{2}{h_\alpha^2} a_\alpha(l_\alpha), & x_\alpha = l_\alpha. \end{cases}$$

Рассмотрим задачу на собственные значения

$$A_\alpha y - \lambda \mathcal{D}y = 0, \quad x \in \bar{\omega}. \quad (14)$$

Легко показать, что если λ есть собственное значение задачи (14), то $2 - \lambda$ — тоже собственное значение. Следовательно,

$$\lambda_{\min} \mathcal{D} \leq A_\alpha \leq (2 - \lambda_{\min}) \mathcal{D}$$

или

$$(A_\alpha y, y) \leq (2 - \lambda_{\min}) (\mathcal{D}y, y) \leq (2 - \lambda_{\min}) \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha(x_\alpha) (y, y).$$

Поэтому в качестве Δ_α можно взять

$$\Delta_\alpha = (2 - \lambda_{\min}) \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha(x_\alpha).$$

Осталось найти λ_{\min} . Если $q = \kappa_{-\alpha} = \kappa_{+\alpha} = 0$, то оператор A_α вырожденный и $\lambda_{\min} = 0$. Иначе, в силу замечания 2 леммы 14 § 2 гл. V, будем иметь

$$(d_\alpha y, y)_{\bar{\omega}_\alpha} \leq \max_{x_\alpha \in \bar{\omega}_\alpha} w^\alpha(x_\alpha) (A_\alpha y, y)_{\bar{\omega}_\alpha}, \quad (15)$$

где $w^\alpha(x_\alpha)$ есть решение следующей краевой задачи:

$$\begin{aligned} \left(a_\alpha w_{x_\alpha}^\alpha \right)_{x_\alpha} - 0,5q w^\alpha &= -d_\alpha(x_\alpha), \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} a_\alpha(h_\alpha) w_{x_\alpha}^\alpha - \left(0,5q + \frac{2}{h_\alpha} \kappa_{-\alpha} \right) w^\alpha &= -d_\alpha(0), \quad x_\alpha = 0, \\ -\frac{2}{h_\alpha} a_\alpha(l_\alpha) w_{x_\alpha}^\alpha - \left(0,5q + \frac{2}{h_\alpha} \kappa_{+\alpha} \right) w^\alpha &= -d_\alpha(l_\alpha), \quad x_\alpha = l_\alpha. \end{aligned} \quad (16)$$

Умножая (15) на $\tilde{h}_\beta(x_\beta)$ и суммируя по x_β от 0 до l_β , получим

$$(\mathcal{D}y, y) \leq \max_{\bar{x}_\alpha \in \bar{\omega}_\alpha} w^\alpha(x_\alpha) (A_\alpha y, y)$$

и, следовательно,

$$\lambda_{\min} \geq \frac{1}{\max_{x_\alpha \in \bar{\omega}_\alpha} w^\alpha(x_\alpha)}.$$

Итак, если $q = \kappa_{-\alpha} = \kappa_{+\alpha} = 0$, то

$$\delta_\alpha = 0, \quad \Delta_\alpha = 2 \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha(x_\alpha),$$

иначе

$$\begin{aligned} \delta_\alpha &= \frac{1}{\max_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha)}, \\ \Delta_\alpha &= \left(2 - \frac{1}{\max_{x_\alpha \in \bar{\omega}_\alpha} w^\alpha(x_\alpha)} \right) \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha(x_\alpha), \end{aligned}$$

где $v^\alpha(x_\alpha)$ и $w^\alpha(x_\alpha)$ — решения задач (13) и (16). Вся необходимая для применения метода переменных направлений априорная информация найдена. Используя формулы теоремы 1, найдем итерационные параметры метода и оценим требуемое число итераций.

Приведем теперь формулы алгоритма метода переменных направлений для рассматриваемого примера. Учитывая определение операторов A_1 , A_2 и правой части φ , получим

$$\begin{aligned} \omega_{k+1}^{(1)} y_{k+1/2} - \Lambda_1 y_{k+1/2} &= \omega_{k+1}^{(2)} y_k + \Lambda_2 y_k + \varphi, \\ 0 \leq x_1 \leq l_1, \quad 0 \leq x_2 \leq l_2, \\ \omega_{k+1}^{(2)} y_{k+1} - \Lambda_2 y_{k+1} &= \omega_{k+1}^{(2)} y_{k+1/2} + \Lambda_1 y_{k+1/2} + \varphi, \\ 0 \leq x_2 \leq l_2, \quad 0 \leq x_1 \leq l_1. \end{aligned}$$

Здесь, в отличие от задачи Дирихле, трехточечные краевые задачи должны решаться и по границе сетки $\bar{\omega}$, а начальное приближение y_0 есть произвольная сеточная функция, заданная на всей сетке ω .

Используя условия (10), можно показать, что для рассматриваемого примера, как и для случая задачи Дирихле, для числа итераций n справедлива следующая асимптотическая по h оценка:

$$n \geq n_0(\varepsilon) = O(\ln |h| \ln \varepsilon), \quad |h|^2 = h_1^2 + h_2^2.$$

Отметим, что все проведенные здесь рассмотрения сохраняют силу и в случае, когда $\bar{\omega}$ — произвольная неравномерная прямоугольная сетка в области \bar{G} . Нужно лишь заменить введенные здесь операторы Λ_α операторами на неравномерной сетке.

Подчеркнем, что предположения постоянства $q, \kappa_{\pm\alpha}$ и зависимости коэффициентов a_α только от x_α существенны. Если не выполнено хотя бы одно из этих предположений, то условие коммутативности операторов A_1 и A_2 не будет выполнено.

В заключение отметим, что метод переменных направлений можно применять для решения разностных аналогов уравнения (9) и при других краевых условиях. В частности, на каждой стороне прямоугольника \bar{G} может быть задано одно из краевых условий первого, второго или третьего рода с постоянными $\kappa_{\pm\alpha}$.

3. Разностная задача Дирихле повышенного порядка точности. Рассмотрим еще один пример применения метода переменных направлений. Пусть на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$, введенной в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, требуется найти решение разностной задачи Дирихле повышенного порядка точности для уравнения Пуассона

$$\begin{aligned} \Lambda y &= (\Lambda_1 + \Lambda_2 + (\kappa_1 + \kappa_2) \Lambda_1 \Lambda_2) y = -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \quad (17)$$

где $\Lambda_\alpha y = y_{x_\alpha x_\alpha}$, $\kappa_\alpha = h_\alpha^2/12$, $\alpha = 1, 2$.

Здесь

$$\varphi = \tilde{f} + \kappa_1 \Lambda_1 \tilde{f} + \kappa_2 \Lambda_2 \tilde{f},$$

где $\tilde{f}(x)$ — правая часть исходного дифференциального уравнения

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -\tilde{f}(x), \quad x \in G, \quad u(x) = g(x), \quad x \in \Gamma.$$

Разностная схема (17) при указанном выборе $\varphi(x)$ имеет точность $O(|h|^4)$, $|h|^2 = h_1^2 + h_2^2$, а на квадратной сетке ($h_1 = h_2 = h$) при соответствующем выборе $\varphi(x)$

$$\varphi = \tilde{f} + \frac{h^2}{12} (\Lambda_1 + \Lambda_2) \tilde{f} + \frac{h^4}{360} (\Lambda_1^2 + 4\Lambda_1 \Lambda_2 + \Lambda_2^2) \tilde{f}$$

имеет точность $O(h^6)$.

Вводя операторы $A_\alpha y = -\Lambda_\alpha \dot{y}$, где $y \in H$, $\dot{y} \in \dot{H}$ и H — пространство сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2,$$

а \dot{H} — множество сеточных функций, обращающихся в нуль на γ , запишем (17) в операторном виде

$$Au = f, \quad (18)$$

где $A = A_1 + A_2 - (\kappa_1 + \kappa_2)A_1A_2$.

Как было неоднократно показано, операторы A_1 и A_2 обладают следующими свойствами: A_1 и A_2 самосопряжены в H и перестановочны

$$A_\alpha = A_\alpha^*, \quad \alpha = 1, 2, \quad A_1A_2 = A_2A_1, \quad (19)$$

оператор A_α имеет границы δ_α и Δ_α , где

$$\delta_\alpha = \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta_\alpha = \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha},$$

$$\delta_\alpha E \leq A_\alpha \leq \Delta_\alpha E, \quad \delta_\alpha > 0, \quad \alpha = 1, 2. \quad (20)$$

Имеет место

Лемма 1. *Если выполнены условия (19), (20) и $\kappa_\alpha \Delta_\alpha < 1$, то операторы*

$$\bar{A}_\alpha = (E - \kappa_\alpha A_\alpha)^{-1} A_\alpha, \quad \alpha = 1, 2, \quad (21)$$

самосопряжены в H , перестановочны и имеют границы $\bar{\delta}_\alpha$ и $\bar{\Delta}_\alpha$, т. е.

$$\bar{\delta}_\alpha E \leq \bar{A}_\alpha \leq \bar{\Delta}_\alpha E, \quad \bar{\delta}_\alpha > 0, \quad \alpha = 1, 2,$$

где $\bar{\delta}_\alpha$ и $\bar{\Delta}_\alpha$ определяются формулами

$$\bar{\delta}_\alpha = \frac{\delta_\alpha}{1 - \kappa_\alpha \delta_\alpha}, \quad \bar{\Delta}_\alpha = \frac{\Delta_\alpha}{1 - \kappa_\alpha \Delta_\alpha}. \quad (22)$$

Действительно, существование оператора \bar{A}_α следует из положительной определенности оператора $E - \kappa_\alpha A_\alpha$, если выполнено условие $\kappa_\alpha \Delta_\alpha < 1$. Далее, представляя \bar{A}_α в виде $\bar{A}_\alpha = (A_\alpha^{-1} - \kappa_\alpha E)^{-1}$ и учитывая самосопряженность операторов A_α , A_α^{-1} и $A_\alpha^{-1} - \kappa_\alpha E$, получим

$$\left(\frac{1}{\Delta_\alpha} - \kappa_\alpha \right) E \leq (A_\alpha^{-1} - \kappa_\alpha E) \leq \left(\frac{1}{\delta_\alpha} - \kappa_\alpha \right) E.$$

Отсюда следует утверждение леммы. Перестановочность операторов \bar{A}_1 и \bar{A}_2 следует из перестановочности A_1 и A_2 . Лемма доказана.

Для рассматриваемого примера условия леммы 1 выполнены, так как $\kappa_\alpha \Delta_\alpha < 1/3$.

Преобразуем теперь уравнение (18). Для этого запишем (18) в виде

$$A_1(E - \kappa_2 A_2) u + A_2(E - \kappa_1 A_1) u = f. \quad (23)$$

Применяя к (23) оператор $(E - \kappa_1 A_1)^{-1}(E - \kappa_2 A_2)^{-1}$ и учитывая перестановочность всех операторов, получим из (23) эквивалентное (18) уравнение

$$\bar{A}u = (\bar{A}_1 + \bar{A}_2)u = \bar{f}, \quad (24)$$

где \bar{A}_1 и \bar{A}_2 определены в (21), а $\bar{f} = (E - \kappa_1 A_1)^{-1}(E - \kappa_2 A_2)^{-1}f$. Итак, решение уравнения (18) сведено к решению уравнения (24) с самосопряженными перестановочными операторами \bar{A}_1 и \bar{A}_2 , границы которых заданы в (22).

Для нахождения приближенного решения уравнения (24) воспользуемся методом переменных направлений

$$\bar{B}_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + \bar{A}y_k = \bar{f}, \quad k = 0, 1, \dots, y_0 \in H, \quad (25)$$

где

$$\bar{B}_k = (\omega_k^{(1)} E + \bar{A}_1)(\omega_k^{(2)} E + \bar{A}_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}.$$

Итерационные параметры $\omega_k^{(1)}$ и $\omega_k^{(2)}$ находятся по формулам теоремы 1, в которых δ_α и Δ_α заменены на $\bar{\delta}_\alpha$ и $\bar{\Delta}_\alpha$. Все необходимые условия применимости метода переменных направлений здесь выполнены.

Осталось рассмотреть алгоритм, реализующий итерационный метод (25). Перепишем (25) в виде

$$(\omega_{k+1}^{(1)} E + \bar{A}_1)(\omega_{k+1}^{(2)} E + \bar{A}_2)y_{k+1} = (\omega_{k+1}^{(2)} E - \bar{A}_1)(\omega_{k+1}^{(1)} E - \bar{A}_2)y_k + \tau_{k+1}\bar{f}. \quad (26)$$

В п. 4 § 1 было показано, что итерационные параметры $\omega_k^{(1)}$ и $\omega_k^{(2)}$ для любого k удовлетворяют неравенствам $\bar{\delta}_2 \leq \omega_k^{(1)} \leq \bar{\Delta}_2$, $\bar{\delta}_1 \leq \omega_k^{(2)} \leq \bar{\Delta}_1$.

Так как для рассматриваемого примера $\bar{\delta}_\alpha > 0$, то, деля левую и правую части (26) на $\omega_{k+1}^{(1)}\omega_{k+1}^{(2)}$ и обозначая $\tau_{k+1}^{(1)} = 1/\omega_{k+1}^{(1)}$, $\tau_{k+1}^{(2)} = 1/\omega_{k+1}^{(2)}$, получим

$$(E + \tau_{k+1}^{(1)}\bar{A}_1)(E + \tau_{k+1}^{(2)}\bar{A}_2)y_{k+1} = \\ = (E - \tau_{k+1}^{(2)}\bar{A}_1)(E - \tau_{k+1}^{(1)}\bar{A}_2) + (\tau_{k+1}^{(1)} + \tau_{k+1}^{(2)})\bar{f}.$$

Применим к обеим частям этого равенства оператор

$$(E - \kappa_1 A_1)(E - \kappa_2 A_2)$$

и учтем, что все операторы перестановочны и

$$(E - \kappa_\alpha A_\alpha)(E + \tau_{k+1}^{(\alpha)}\bar{A}_\alpha) = E + (\tau_{k+1}^{(\alpha)} - \kappa_\alpha)A_\alpha, \\ (E - \kappa_\alpha A_\alpha)(E - \tau_{k+1}^{(\beta)}\bar{A}_\alpha) = E - (\tau_{k+1}^{(\beta)} + \kappa_\alpha)A_\alpha, \\ \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

В результате получим

$$(E + (\tau_{k+1}^{(1)} - \kappa_1) A_1)(E + (\tau_{k+1}^{(2)} - \kappa_2) A_2) y_{k+1} = \\ = (E - (\tau_{k+1}^{(2)} + \kappa_1) A_1)(E - (\tau_{k+1}^{(1)} + \kappa_2) A_2) y_k + (\tau_{k+1}^{(1)} + \tau_{k+1}^{(2)}) f. \quad (27)$$

Итерационная схема (27) эквивалентна следующей схеме:

$$(E + (\tau_{k+1}^{(1)} - \kappa_1) A_1) y_{k+1/2} = (E - (\tau_{k+1}^{(1)} + \kappa_2) A_2) y_k + (\tau_{k+1}^{(1)} - \kappa_1) f, \quad (28)$$

$$(E + (\tau_{k+1}^{(2)} - \kappa_2) A_2) y_{k+1} = (E - (\tau_{k+1}^{(2)} + \kappa_1) A_1) y_{k+1/2} + (\tau_{k+1}^{(2)} + \kappa_1) f. \quad (29)$$

Эквивалентность (27) и (28), (29) доказывается следующим образом. Умножая (28) на $\tau_{k+1}^{(2)} + \kappa_1$, (29) на $-(\tau_{k+1}^{(1)} - \kappa_1)$ и складывая результаты, получим

$$(\tau_{k+1}^{(1)} + \tau_{k+1}^{(2)}) y_{k+1/2} = (\tau_{k+1}^{(1)} - \kappa_1) (E + (\tau_{k+1}^{(2)} - \kappa_2) A_2) y_{k+1} + \\ + (\tau_{k+1}^{(2)} + \kappa_1) (E - (\tau_{k+1}^{(1)} + \kappa_2) A_1) y_k. \quad (30)$$

Подставляя (30) в (28), получим после несложных преобразований (27). Обратный ход рассуждений очевиден.

Учитывая определение операторов A_1 и A_2 , схему (28) и (29) можно записать в виде обычной разностной схемы

$$(E - (\tau_{k+1}^{(1)} - \kappa_1) \Lambda_1) y_{k+1/2} = (E + (\tau_{k+1}^{(1)} + \kappa_2) \Lambda_2) y_k + (\tau_{k+1}^{(1)} - \kappa_1) \varphi \quad (31)$$

для $h_1 \leqslant x_1 \leqslant l_1 - h_1$,

$$y_{k+1/2} = g(x) + (\kappa_1 + \kappa_2) \Lambda_2 g(x), \quad x_1 = 0, \quad l_1.$$

Краевую задачу (31) нужно последовательно решать для $h_2 \leqslant x_2 \leqslant l_2 - h_2$. В результате будет найдено $y_{k+1/2}$ для $0 \leqslant x_1 \leqslant l_1$, $h_2 \leqslant x_2 \leqslant l_2 - h_2$. Далее,

$$(E - (\tau_{k+1}^{(2)} - \kappa_2) \Lambda_2) y_{k+1} = (E + (\tau_{k+1}^{(2)} + \kappa_1) \Lambda_1) y_{k+1/2} + (\tau_{k+1}^{(2)} + \kappa_1) \varphi \quad (32)$$

для $h_2 \leqslant x_2 \leqslant l_2 - h_2$,

$$y_{k+1} = g(x), \quad x_2 = 0, \quad l_2.$$

Краевую задачу (32) нужно последовательно решать для $h_1 \leqslant x_1 \leqslant l_1 - h_1$. В результате будет найдено y_{k+1} .

Если сравнить числа итераций метода переменных направлений для разностной схемы второго порядка точности, рассмотренной в п. 1 § 2, и для схемы повышенного порядка точности, то в последнем случае число итераций будет несколько большим. В частном случае $l_1 = l_2 = l$, $N_1 = N_2 = N$ для $N = 10$ это увеличение происходит на 1%, а для $N = 100$ на 4%. Объем вычислений на каждую итерацию для обеих схем практически одинаков, а различие в числе итераций незначительно. Так как схема повышенного порядка точности позволяет пользоваться более грубой сеткой для достижения заданной точности решения дифференциальной задачи, то ее применение особенно выгодно в тех случаях, когда решение дифференциальной задачи обладает достаточной гладкостью.

Напомним, что в § 3 гл. III для решения задачи (17) мы рассмотрели прямой метод — метод редукции. Как для схемы второго порядка точности, так и для рассматриваемой здесь схемы, прямой метод будет требовать меньшего числа арифметических действий, чем метод переменных направлений с оптимальными параметрами.

§ 3. Метод переменных направлений в общем случае

1. Случай неперестановочных операторов. Пусть требуется найти решение линейного операторного уравнения

$$Au = f \quad (1)$$

с невырожденным оператором A , который представим в виде суммы двух самосопряженных неперестановочных операторов A_1 и A_2 с границами δ_1 , Δ_1 и δ_2 , Δ_2 :

$$\begin{aligned} A_\alpha &= A_\alpha^*, \quad \delta_\alpha E \leq A_\alpha \leq \Delta_\alpha E, \quad \alpha = 1, 2, \quad \delta_1 + \delta_2 > 0, \\ A &= A_1 + A_2. \end{aligned} \quad (2)$$

Для приближенного решения уравнения (1) рассмотрим двухслойную схему метода переменных направлений с двумя итерационными параметрами $\omega^{(1)}$ и $\omega^{(2)}$:

$$\begin{aligned} B \frac{y_{k+1} - y_k}{\tau} + Ay_k &= f, \quad k = 0, 1, \dots, y_0 \in H, \\ B &= (\omega^{(1)}E + A_1)(\omega^{(2)}E + A_2), \quad \tau = \omega^{(1)} + \omega^{(2)}. \end{aligned} \quad (3)$$

Здесь итерационные параметры и оператор B не зависят от номера итерации k .

Как и в случае перестановочных операторов A_1 и A_2 , итерационное приближение y_{k+1} для схемы (3) может быть найдено по следующему алгоритму:

$$\begin{aligned} (\omega^{(1)}E + A_1)y_{k+1/2} &= (\omega^{(1)}E - A_2)y_k + f, \\ (\omega^{(2)}E + A_2)y_{k+1} &= (\omega^{(2)}E - A_1)y_{k+1/2} + f, \quad k = 0, 1, \dots \end{aligned}$$

Исследуем сходимость итерационной схемы (3) и найдем оптимальные значения параметров $\omega^{(1)}$ и $\omega^{(2)}$. Предполагая, что оператор $\omega^{(2)}E + A_2$ не вырожден, изучим сходимость (3) в энергетическом пространстве H_D , где $D = (\omega^{(2)}E + A_2)^2$. В силу (2) оператор D самосопряжен в H , а из указанного выше предположения следует, что D положительно определен в H .

Для погрешности $z_k = y_k - u$ получим из (3) однородное уравнение

$$\begin{aligned} z_{k+1} &= Sz_k, \quad k = 0, 1, \dots, \quad z_0 = y_0 - u, \\ S &= (\omega^{(2)}E + A_2)^{-1}(\omega^{(1)}E + A_1)^{-1}(\omega^{(2)}E - A_1)(\omega^{(1)}E - A_2). \end{aligned} \quad (4)$$

Перейдем в (4) к задаче для эквивалентной погрешности

$$x_k = (\omega^{(2)}E + A_2)z_k. \quad (5)$$

Получим

$$\begin{aligned}x_{k+1} &= \bar{S}x_k, \quad k = 0, 1, \dots, \quad \bar{S} = \bar{S}_1 \bar{S}_2, \\ \bar{S}_1 &= (\omega^{(1)}E + A_1)^{-1}(\omega^{(2)}E - A_1), \\ \bar{S}_2 &= (\omega^{(2)}E + A_2)^{-1}(\omega^{(1)}E - A_2).\end{aligned}\tag{6}$$

Так как в силу сделанной замены (5) имеет место равенство $\|x_k\| = \|z_k\|_D$, то достаточно исследовать поведение нормы эквивалентной погрешности x_k в пространстве H . Из (6) найдем

$$\|x_{k+1}\| \leq \|\bar{S}\| \|x_k\| \leq \|\bar{S}_1\| \|\bar{S}_2\| \|x_k\|, \quad k = 0, 1, \dots$$

и, следовательно,

$$\|z_n\|_D = \|x_n\| \leq \|\bar{S}\|^n \|x_0\| \leq (\|\bar{S}_1\| \|\bar{S}_2\|)^n \|z_0\|_D.\tag{7}$$

Оценим норму операторов \bar{S}_1 и \bar{S}_2 . Предположим, что операторы $\omega^{(\alpha)}E + A_\alpha$, $\alpha = 1, 2$, неотрицательны. Тогда из п. 4 § 1 гл. V в силу (2) получим

$$\|\bar{S}_1\| \leq \max_{\delta_1 \leq x \leq \Delta_1} \left| \frac{\omega^{(2)} - x}{\omega^{(1)} + x} \right|, \quad \|\bar{S}_2\| \leq \max_{\delta_2 \leq y \leq \Delta_2} \left| \frac{\omega^{(1)} - y}{\omega^{(2)} + y} \right|$$

и, следовательно,

$$\|\bar{S}_1\| \|\bar{S}_2\| \leq \max_{\substack{\delta_1 \leq x \leq \Delta_1 \\ \delta_2 \leq y \leq \Delta_2}} |R_1(x, y)|, \quad R_1(x, y) = \frac{\omega^{(2)} - x}{\omega^{(1)} + x} \frac{\omega^{(1)} - y}{\omega^{(2)} + y}.$$

Учитывая оценку (7), поставим задачу выбрать параметры $\omega^{(1)}$ и $\omega^{(2)}$ из условия

$$\min_{\omega^{(1)}, \omega^{(2)}} \max_{\substack{\delta_1 \leq x \leq \Delta_1 \\ \delta_2 \leq y \leq \Delta_2}} |R_1(x, y)|.$$

Эта задача является частным случаем задачи, решенной в § 1 настоящей главы. При помощи дробно-линейного преобразования переменных x и y (см. (15), (21)–(24) в § 1) поставленная задача сводится к задаче нахождения параметра κ^* из условия

$$\max_{\eta \leq u \leq 1} \left| \frac{\kappa^* - u}{\kappa^* + u} \right| = \min_{\kappa} \max_{\eta \leq u \leq 1} \left| \frac{\kappa - u}{\kappa + u} \right| = \rho.\tag{8}$$

При этом параметры $\omega^{(1)}$ и $\omega^{(2)}$ выражаются через κ^* по формулам

$$\omega^{(1)} = \frac{r\kappa^* + s}{1 + t\kappa^*}, \quad \omega^{(2)} = \frac{r\kappa^* - s}{1 - t\kappa^*},$$

а для погрешности z_n имеет место оценка

$$\|z_n\|_D \leq \rho^{2n} \|z_0\|_D.$$

Кроме того, в п. 4 § 1 было показано, что при оптимальном выборе κ^* операторы $\omega^{(\alpha)}E + A_\alpha$ положительно определены, если $\delta_1 + \delta_2 > 0$. Следовательно, в силу (2) наши предположения

о неотрицательности операторов $\omega^{(\alpha)}E + A_\alpha$, $\alpha = 1, 2$, будут заведомо выполнены.

Приведем решение задачи (8) независимо от результатов п. 4 § 1. Рассмотрим функцию $\varphi(u) = (\kappa - u)/(\kappa + u)$ на отрезке $0 < \eta \leq u \leq 1$ при $\kappa > 0$. Эта функция монотонно убывает по u , следовательно,

$$\max_{\eta \leq u \leq 1} |\varphi(u)| = \max \left(\left| \frac{\eta - \kappa}{\eta + \kappa} \right|, \left| \frac{1 - \kappa}{1 + \kappa} \right| \right).$$

Отсюда легко получить, что минимум по κ этого выражения достигается для κ^* , которое определяется из равенства

$$\frac{\kappa^* - \eta}{\kappa^* + \eta} = \frac{1 - \kappa^*}{1 + \kappa^*}.$$

Отсюда найдем

$$\kappa^* = V\eta, \quad \min_{\kappa} \max_{\eta \leq u \leq 1} \left| \frac{\kappa - u}{\kappa + u} \right| = \rho = \frac{1 - V\eta}{1 + V\eta}.$$

Итак, доказана

Теорема 2. Пусть выполнены условия (2), а параметры $\omega^{(1)}$ и $\omega^{(2)}$ выбраны по формулам

$$\omega^{(1)} = \frac{rV\eta + s}{1 + tV\eta}, \quad \omega^{(2)} = \frac{rV\eta - s}{1 - tV\eta},$$

где r, s, t и η определены в (21)–(24) § 1. Метод переменных направлений (3) сходится в H_D , и для погрешности z_n имеет место оценка

$$\|z_n\|_D \leq \rho^{2n} \|z_0\|_D, \quad \rho = \frac{1 - V\eta}{1 + V\eta},$$

где $D = (\omega^{(2)}E + A_2)^2$. Для числа итераций n справедлива оценка

$$n = n_0(\varepsilon) = \ln \varepsilon / (2 \ln \rho) \approx \ln \frac{1}{\varepsilon} / (4V\eta).$$

2. Разностная задача Дирихле для эллиптического уравнения с переменными коэффициентами. Рассмотрим пример применения метода переменных направлений в некоммутативном случае. Пусть на прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$, введенной в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, требуется найти решение следующей разностной задачи:

$$\Lambda y = \sum_{\alpha=1}^2 (a_\alpha(x) y_{x_\alpha})_{x_\alpha} - q(x) y = -\varphi(x), \quad x \in \omega, \quad (9)$$

$$y(x) = g(x), \quad x \in \gamma,$$

где коэффициенты разностной схемы удовлетворяют условиям

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad 0 \leq d_1 \leq q(x) \leq d_2. \quad (10)$$

В пространстве H сеточных функций, заданных на ω , со скалярным произведением $(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2$, операторы A_1 и A_2 определим следующим образом: $A_\alpha y = -\Delta_\alpha \dot{y} = -(a_\alpha \dot{y})_{x_\alpha} + 0,5q\ddot{y}$, $\alpha = 1, 2$, $y \in H$, $\dot{y} \in \dot{H}$, где, как обычно, $\dot{y}(x) = 0$ на γ .

Введенные операторы A_α являются самосопряженными в H , и если $a_\alpha(x)$ зависит только от переменной x_α и $q(x)$ есть постоянная, то операторы A_1 и A_2 будут перестановочны. В общем же случае перестановочность не будет иметь места, и для решения уравнения (1), соответствующего разностной задаче (9), можно применить рассмотренный в п. 1 § 3 метод переменных направлений (3).

Алгоритм метода имеет простой вид

$$\begin{aligned}\omega^{(1)}y_{k+1/2} - \Lambda_1 y_{k+1/2} &= \omega^{(1)}y_k + \Lambda_2 y_k + \varphi, \quad h_1 \leq x_1 \leq l_1 - h_1, \\ y_{k+1/2}(x) &= g(x), \quad x_1 = 0, \quad l_1, \quad h_2 \leq x_2 \leq l_2 - h_2, \\ \omega^{(2)}y_{k+1} - \Lambda_2 y_{k+1} &= \omega^{(2)}y_{k+1/2} + \Lambda_1 y_{k+1/2} + \varphi, \quad h_2 \leq x_2 \leq l_2 - h_2, \\ y_{k+1}(x) &= g(x), \quad x_2 = 0, \quad l_2, \quad h_1 \leq x_1 \leq l_1 - h_1.\end{aligned}$$

Осталось найти границы δ_α и Δ_α операторов A_α , $\alpha = 1, 2$. Так как условия (10) выполнены, то из леммы 14 § 2 гл. V получим

$$(y^2, 1)_{\omega_\alpha} \leq \kappa_\alpha(x_\beta)(A_\alpha y, y)_{\omega_\alpha}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2, \quad (11)$$

где $\kappa_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x)$, а $v^\alpha(x)$ есть решение следующей трехточечной краевой задачи:

$$\begin{aligned}\left(a_\alpha v_{x_\alpha}^\alpha\right)_{x_\alpha} - 0,5qv^\alpha &= -1, \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ v^\alpha(x) &= 0, \quad x_\alpha = 0, \quad l_\alpha, \quad h_\beta \leq x_\beta \leq l_\beta - h_\beta.\end{aligned}$$

Скалярное произведение по ω_α определяется следующим образом:

$$(u, v)_{\omega_\alpha} = \sum_{x_\alpha=h_\alpha}^{l_\alpha-h_\alpha} u(x)v(x)h_\alpha = \sum_{x_\alpha \in \omega_\alpha} u(x)v(x)h_\alpha.$$

Умножая (11) на h_β и суммируя по x_β , получим

$$\left(\frac{1}{\kappa_\alpha} y^2, 1\right) \leq (A_\alpha y, y), \quad \alpha = 1, 2.$$

Следовательно, в качестве δ_α можно взять

$$\delta_\alpha = \min_{h_\beta \leq x_\beta \leq l_\beta - h_\beta} \frac{1}{\kappa_\alpha(x_\beta)}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Найдем теперь Δ_α . Будем поступать по аналогии с п. 2 § 2. Обозначим через \mathcal{D} диагональную часть матрицы A_α , соответствующей оператору A_α :

$$\mathcal{D}y = d_\alpha(x) y,$$

$$d_\alpha(x) = \begin{cases} 0,5q(x) + \frac{1}{h_\alpha^2}(a_\alpha(x_\alpha, x_\beta) + a_\alpha(x_\alpha + h_\alpha, x_\beta)), \\ h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ h_\beta \leq x_\beta \leq l_\beta - h_\beta. \end{cases}$$

Тогда имеет место неравенство

$$(A_\alpha y, y) \leq (2 - \lambda_{\min})(\mathcal{D}y, y) \leq (2 - \lambda_{\min}) \max_{x \in \omega} d_\alpha(x) (y, y),$$

где λ_{\min} — постоянная из операторного неравенства $\lambda_{\min}\mathcal{D} \leq A_\alpha$. Найдем λ_{\min} . Из леммы 14 § 2 гл. V получим

$$(d_\alpha y^*, 1)_{\omega_\alpha} \leq \rho_\alpha(x_\beta) (A_\alpha y, y)_{\omega_\alpha}, \quad (12)$$

где $\rho_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} w^\alpha(x)$, а $w^\alpha(x)$ есть решение следующей трехточечной краевой задачи:

$$\left(a_\alpha w_{x_\alpha}^\alpha \right)_{x_\alpha} - 0,5qw^\alpha = -d_\alpha(x), \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha,$$

$$w^\alpha(x) = 0, \quad x_\alpha = 0, \quad l_\alpha, \quad h_\beta \leq x_\beta \leq l_\beta - h_\beta.$$

Умножая (12) на h_β и суммируя по ω_β , получим

$$\left(\frac{d_\alpha}{\rho_\alpha} y^*, 1 \right) \leq (A_\alpha y, y), \quad \alpha = 1, 2.$$

Следовательно, в качестве λ_{\min} можно взять

$$\lambda_{\min} = \min_{h_\beta \leq x_\beta \leq l_\beta - h_\beta} \frac{1}{\rho_\alpha(x_\beta)},$$

поэтому Δ_α есть

$$\Delta_\alpha = \left(2 - \frac{1}{\max_{x_\beta} \rho_\alpha(x_\beta)} \right) \max_{x \in \omega} d_\alpha(x), \quad \alpha = 1, 2.$$

Итак, априорная информация, требуемая для применения метода переменных направлений, найдена. Используя условия (10), можно показать, что величина η , определяющая скорость сходимости метода, для рассматриваемого примера есть $O(|h|^2)$, где $|h|^2 = h_1^2 + h_2^2$. Поэтому в силу теоремы 2 для числа итераций будет справедлива оценка

$$n = O \left(\frac{1}{|h|} \ln \frac{1}{\varepsilon} \right).$$

Рассмотрим модельную задачу. Пусть разностная схема (9) задана на квадратной сетке в единичном квадрате ($N_1 = N_2 = N$,

$l_1 = l_2 = 1$). Коэффициенты $a_1(x)$, $a_2(x)$ и $q(x)$ выберем следующим образом:

$$\begin{aligned} a_1(x) &= 1 + c[(x_1 - 0,5)^2 + (x_2 - 0,5)^2], \\ a_2(x) &= 1 + c[0,5 - (x_1 - 0,5)^2 - (x_2 - 0,5)^2], \\ q(x) &\equiv 0, \quad c > 0. \end{aligned}$$

В этом случае в неравенствах (10) $c_1 = 1$, $c_2 = 1 + 0,5c$, $d_1 = d_2 = 0$, меняя параметр c , будем получать коэффициенты разностной схемы (9) с различными экстремальными характеристиками.

Приведем число итераций для рассмотренного метода переменных направлений в зависимости от отношения c_2/c_1 и от числа узлов N по одному направлению для $\epsilon = 10^{-4}$.

Т а б л и ц а 12

c_2/c_1	$N=32$	$N=64$	$N=128$
2	65	132	264
8	90	187	380
32	110	233	482
128	122	264	556
512	128	282	603

Сравним этот метод с методом верхней релаксации (см. § 2 гл. IX), попаременно-треугольным методом (см. § 2 гл. X) и неявным чебышевским методом (см. п. 3 § 2 гл. VI). По числу итераций рассмотренный метод переменных направлений уступает методу верхней релаксации и попаременно-треугольному методу, но превосходит неявный чебышевский метод в 1,5—2 раза. Однако по объему вычислительной работы метод переменных направлений будет уступать и неявному чебышевскому методу.

ГЛАВА XII

МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЙ С НЕЗНАКОПРЕДЕЛЕННЫМИ И ВЫРОЖДЕННЫМИ ОПЕРАТОРАМИ

В главе изучаются прямые и итерационные методы решения уравнений с невырожденным и незнакоопределенным оператором, с комплексным оператором, а также с вырожденным оператором. В § 1 для уравнения с незнакоопределенным оператором рассмотрены метод с чебышевскими параметрами и метод вариационного типа. В § 2 для уравнения с комплексным оператором специального вида построены методы простой итерации и переменных направлений с комплексными итерационными параметрами. В § 3 изучены общие итерационные методы решения уравнений с вырожденным оператором, когда оператор на верхнем слое невырожден. Параграф 4 посвящен построению специальных прямых и итерационных методов для уравнений с вырожденным оператором.

§ 1. Уравнения с действительным незнакоопределенным оператором

1. Итерационная схема. Задача выбора итерационных параметров. Пусть в гильбертовом пространстве H дано уравнение

$$Au = f \quad (1)$$

с линейным невырожденным оператором A . Для решения уравнения (1) рассмотрим неявную двухслойную итерационную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad (2)$$

с невырожденным оператором B и произвольным $y_0 \in H$.

Итерационные схемы вида (2) изучались в главах VI, VIII, где были предложены некоторые способы выбора итерационных параметров τ_k в зависимости от свойств операторов A , B и D . Напомним, что D есть самосопряженный положительно определенный оператор, который порождает энергетическое пространство H_D . Было показано, что для сходимости в H_D рассмотренных итерационных методов требуется положительная определенность оператора

$$C = D^{-1/2} (DB^{-1}A) D^{-1/2}. \quad (3)$$

Для конкретных операторов D это требование приводит к следующим условиям на операторы A и B :

1) оператор A должен быть положительно определен в H , если $D = A$, B или $A^*B^{-1}A$;

2) оператор B^*A должен быть положительно определен в H , если $D = A^*A$ или B^*B .

Существуют задачи, для которых эти требования не выполнены, т. е. либо оператор A не является знакоопределенным, либо трудно найти такой оператор B , чтобы B^*A был положительно определенным оператором. В качестве примера таких задач можно привести задачу Дирихле для уравнения Гельмгольца в прямоугольнике

$$\begin{aligned} y_{\bar{x}_1 \bar{x}_1} + y_{\bar{x}_2 \bar{x}_2} + m^2 y &= 0, & x \in \omega, \\ y(x) &= g(x), & x \in \gamma, \end{aligned}$$

где $m^2 > 0$.

Данный параграф посвящен построению неявных двухслойных итерационных методов для случая, когда оператор C является невырожденным незнакоопределенным в H оператором. Здесь мы будем рассматривать только действительные операторы C , комплексный случай изучается в § 2.

Переходим к построению итерационных методов. В уравнении

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A) z_k, \quad k = 0, 1, \dots, 2n-1,$$

для погрешности $z_k = y_k - u$ итерационной схемы (2) сделаем замену $z_k = D^{-1/2} x_k$ и перейдем к уравнению для эквивалентной погрешности x_k :

$$x_{k+1} = (E - \tau_{k+1} C) x_k, \quad k = 0, 1, \dots, 2n-1, \quad (4)$$

где оператор C определен в (3). Так как оператор C незнакоопределен, то очевидно, что норма оператора $E - \tau_{k+1} C$ будет больше либо равна единице для любого τ_{k+1} .

Рассмотрим теперь уравнение, связывающее погрешности на четных итерациях. Из (4) получим

$$x_{2k+2} = (E - \tau_{2k+2} C) (E - \tau_{2k+1} C) x_{2k}, \quad k = 0, 1, \dots, n-1. \quad (5)$$

Если обозначить

$$\omega_{k+1} = -\tau_{2k+2} \tau_{2k+1}, \quad k = 0, 1, \dots, n-1, \quad (6)$$

и потребовать, чтобы итерационные параметры τ_{2k+2} и τ_{2k+1} для любого k удовлетворяли соотношению

$$1/\tau_{2k+2} + 1/\tau_{2k+1} = 2\alpha, \quad k = 0, 1, \dots, n-1, \quad (7)$$

где α — неопределенная пока постоянная, то (5) можно записать в виде

$$x_{2k+2} = (E - \omega_{k+1} \bar{C}) x_{2k}, \quad k = 0, 1, \dots, \bar{C} = C^2 - 2\alpha C. \quad (8)$$

Если ω_{k+1} и α найдены, то параметры τ_{2k+2} и τ_{2k+1} в силу (6) и (7) определяются по формулам

$$\begin{aligned}\tau_{2k+1} &= -\alpha\omega_{k+1} - \sqrt{\alpha^2\omega_{k+1}^2 + \omega_{k+1}}, \\ \tau_{2k+2} &= -\alpha\omega_{k+1} + \sqrt{\alpha^2\omega_{k+1}^2 + \omega_{k+1}}, \\ k &= 0, 1, \dots, n-1.\end{aligned}\quad (9)$$

Из (8) получим

$$\begin{aligned}x_{2n} &= \prod_{j=1}^n (E - \omega_j \bar{C}) x_0, \\ \|x_{2n}\| &\leq \left\| \prod_{j=1}^n (E - \omega_j \bar{C}) \right\| \|x_0\|.\end{aligned}\quad (10)$$

Так как оператор \bar{C} зависит от α , то требование положительной определенности оператора \bar{C} будет одним из условий, которым подчинен выбор параметра α . Кроме того, из (10) следует, что параметры ω_j , $1 \leq j \leq n$, и параметр α нужно выбрать из условия минимума нормы разрешающего оператора $\prod_{j=1}^n (E - \omega_j \bar{C})$.

Эта задача о наилучшем выборе итерационных параметров ω_j и α , а следовательно, и параметров τ_k для схемы (2) будет решена ниже. Сначала установим связь предлагаемого способа построения итерационного метода со способом, основанным на трансформации Гаусса, для случая самосопряженного оператора C .

Заметим, что замена $u = D^{-1/2}x$, $f = BD^{-1/2}\varphi$ позволяет записать исходное уравнение (1) в следующем виде:

$$Cx = \varphi, \quad (11)$$

где оператор C определен в (3). Используя (11), получим

$$\bar{C}x = C^2x - 2\alpha Cx = (C - 2\alpha E)\varphi = \bar{\varphi}. \quad (12)$$

Далее, если обозначить $v_k = D^{1/2}y_k$, где y_k — итерационное приближение в схеме (2), то легко найдем

$$x_k = D^{1/2}z_k = D^{1/2}y_k - D^{1/2}u = v_k - x.$$

Подставляя x_k в (8) и учитывая (12), получим итерационную схему

$$\frac{v_{2k+2} - v_{2k}}{\omega_{k+1}} + \bar{C}v_{2k} = \bar{\varphi}, \quad k = 0, 1, \dots \quad (13)$$

Итак, схема (13) есть явная двухслойная схема для преобразованного уравнения (12).

Пусть $C = C^*$. Напомним, что в этом случае первая трансформация Гаусса состоит в переходе от уравнения (11) к уравнению $\bar{C}x = C^2x = C\varphi = \bar{\varphi}$. Так как C — невырожденный оператор,

то оператор C^2 будет положительно определенным в H . Поэтому указанное преобразование приводит нас к случаю знакоопределенного оператора. Для решения уравнения с таким оператором можно использовать двухслойную схему вида (13), заменив \bar{C} на $\bar{\bar{C}}$ и $\bar{\varphi}$ на $\bar{\bar{\varphi}}$. Очевидно, что такой метод есть частный случай (при $\alpha = 0$) рассматриваемого нами метода.

2. Преобразование оператора в самосопряженном случае. Будем предполагать, что оператор \bar{C} самосопряжен в H . Тогда оператор $\bar{C} = C^2 - \alpha C$ также самосопряжен в H . Нашей ближайшей целью будет выбрать параметр α так, чтобы оператор \bar{C} был положительно определен, и найти границы $\gamma_1 = \gamma_1(\alpha)$ и $\gamma_2 = \gamma_2(\alpha)$ этого оператора, т. е. величины из неравенств

$$\gamma_1 E \leq \bar{C} \leq \gamma_2 E, \quad \gamma_1 > 0. \quad (14)$$

Если указанное значение для α существует, то в силу оценки

$$\left\| \prod_{j=1}^n (E - \omega_j \bar{C}) \right\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} \left| \prod_{j=1}^n (1 - \omega_j t) \right|$$

задача нахождения параметров ω_j , $j = 1, 2, \dots, n$, сводится к построению полинома $P_n(t)$ степени n , нормированного условием $P_n(0) = 1$ и наименее уклоняющегося от нуля на отрезке $[\gamma_1, \gamma_2]$ положительной полуоси. Эта задача была изучена нами ранее в главе VI при построении чебышевского метода. Решение имеет вид

$$\omega_k = \frac{\omega_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ -\cos \frac{2i-1}{2n} \pi, \quad i = 1, 2, \dots, n \right\},$$

где $k = 1, 2, \dots, n$,

$$\omega_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

При этом в силу (10) для погрешности x_{2n} будет справедлива оценка

$$\|x_{2n}\| \leq q_n \|x_0\|, \quad q_n = 2\rho_1^n / (1 + \rho_1^{2n}).$$

Отсюда следует, что выбор параметра α должен быть подчинен условию максимума отношения γ_1/γ_2 .

Найдем оптимальное значение для параметра α . Пусть собственные значения μ оператора C заключены на отрезках $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$. Так как оператор C незнакоопределен и невырожден, то

$$\dot{\gamma}_1 \leq \dot{\gamma}_2 < 0 < \dot{\gamma}_3 \leq \dot{\gamma}_4. \quad (15)$$

Найдем собственные значения λ оператора $\bar{C} = C^2 - 2\alpha C$. Легко видеть, что собственные значения операторов \bar{C} и C связаны соотношением

$$\lambda = \mu^2 - 2\alpha\mu, \quad \mu \in \Omega, \quad (16)$$

где Ω состоит из двух отрезков $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$.

Найдем сначала ограничения на α , которые обеспечивают положительность собственных значений λ , т. е. положительную определенность оператора \bar{C} . Анализируя неравенство $\mu^2 - 2\alpha\mu > 0$, находим, что оно имеет место для μ , меняющегося вне интервала $[0, 2\alpha]$. Поэтому это неравенство будет выполнено для $\mu \in \Omega$, если α удовлетворяет условию

$$\dot{\gamma}_2 < 2\alpha < \dot{\gamma}_3. \quad (17)$$

Будем считать, что (17) выполнено. Из (16) получим, что преобразование $\lambda = \lambda(\mu) = \mu^2 - 2\alpha\mu$ отображает отрезок $[\dot{\gamma}_1, \dot{\gamma}_2]$ на отрезок $[\lambda_2, \lambda_1]$, а отрезок $[\dot{\gamma}_3, \dot{\gamma}_4]$ на отрезок $[\lambda_3, \lambda_4]$, где $\lambda_i = \lambda(\dot{\gamma}_i)$, $1 \leq i \leq 4$. Таким образом, все собственные значения оператора \bar{C} положительны и расположены на отрезках $[\lambda_2, \lambda_1] \cup [\lambda_3, \lambda_4]$. Поэтому в неравенствах (14) следует положить

$$\gamma_1 = \min(\lambda_2, \lambda_3), \quad \gamma_2 = \max(\lambda_1, \lambda_4). \quad (18)$$

Выберем теперь $2\alpha \in (\dot{\gamma}_2, \dot{\gamma}_3)$ из условия максимума отношения γ_1/γ_2 . Из (18) получим

$$\begin{aligned} \gamma_1 &= \begin{cases} \lambda_2 = \dot{\gamma}_2(\dot{\gamma}_2 - 2\alpha), & \dot{\gamma}_2 < 2\alpha \leq \dot{\gamma}_2 + \dot{\gamma}_3, \\ \lambda_3 = \dot{\gamma}_3(\dot{\gamma}_3 - 2\alpha), & \dot{\gamma}_2 + \dot{\gamma}_3 \leq 2\alpha < \dot{\gamma}_3, \end{cases} \\ \gamma_2 &= \begin{cases} \lambda_4 = \dot{\gamma}_4(\dot{\gamma}_4 - 2\alpha), & 2\alpha \leq \dot{\gamma}_1 + \dot{\gamma}_4, \\ \lambda_1 = \dot{\gamma}_1(\dot{\gamma}_1 - 2\alpha), & \dot{\gamma}_1 + \dot{\gamma}_4 \leq 2\alpha. \end{cases} \end{aligned}$$

Введем следующие обозначения: $\Delta_1 = \dot{\gamma}_2 - \dot{\gamma}_1$, $\Delta_2 = \dot{\gamma}_4 - \dot{\gamma}_3$, и рассмотрим два случая.

1) Пусть сначала $\Delta_1 \leq \Delta_2$, т. е. $\dot{\gamma}_2 + \dot{\gamma}_3 \leq \dot{\gamma}_1 + \dot{\gamma}_4$. В этом случае для $\xi = \gamma_1/\gamma_2$ получим следующее выражение:

$$\xi = \xi(\alpha) = \begin{cases} \frac{\dot{\gamma}_2(\dot{\gamma}_2 - 2\alpha)}{\dot{\gamma}_4(\dot{\gamma}_4 - 2\alpha)}, & \dot{\gamma}_2 < 2\alpha \leq \dot{\gamma}_2 + \dot{\gamma}_3, \text{ возрастает по } \alpha, \\ \frac{\dot{\gamma}_3(\dot{\gamma}_3 - 2\alpha)}{\dot{\gamma}_4(\dot{\gamma}_4 - 2\alpha)}, & \dot{\gamma}_2 + \dot{\gamma}_3 \leq 2\alpha \leq \dot{\gamma}_1 + \dot{\gamma}_4, \text{ убывает по } \alpha, \\ \frac{\dot{\gamma}_3(\dot{\gamma}_3 - 2\alpha)}{\dot{\gamma}_1(\dot{\gamma}_1 - 2\alpha)}, & \dot{\gamma}_1 + \dot{\gamma}_4 \leq 2\alpha, \text{ убывает по } \alpha. \end{cases}$$

Следовательно, в этом случае оптимальное значение α есть

$$\alpha = \alpha_0 = (\dot{\gamma}_2 + \dot{\gamma}_3)/2, \quad (19)$$

причем условие (17) выполнено. При $\alpha = \alpha_0$ имеем

$$\gamma_1 = \lambda_2 = \lambda_3 = -\dot{\gamma}_2 \dot{\gamma}_3, \quad (20)$$

$$\gamma_2 = \lambda_4 = \dot{\gamma}_4 (\Delta_2 - \Delta_1) - \dot{\gamma}_1 \dot{\gamma}_4 \geq \lambda_1. \quad (21)$$

2) Пусть теперь $\Delta_1 \geq \Delta_2$, т. е. $\dot{\gamma}_2 + \dot{\gamma}_3 \geq \dot{\gamma}_1 + \dot{\gamma}_4$. В этом случае будем иметь

$$\xi = \xi(\alpha) = \begin{cases} \frac{\dot{\gamma}_2 (\dot{\gamma}_2 - 2\alpha)}{\dot{\gamma}_4 (\dot{\gamma}_4 - 2\alpha)}, & 2\alpha \leq \dot{\gamma}_1 + \dot{\gamma}_4, \text{ возрастает по } \alpha, \\ \frac{\dot{\gamma}_2 (\dot{\gamma}_2 - 2\alpha)}{\dot{\gamma}_1 (\dot{\gamma}_1 - 2\alpha)}, & \dot{\gamma}_1 + \dot{\gamma}_4 \leq 2\alpha \leq \dot{\gamma}_2 + \dot{\gamma}_3, \text{ возрастает по } \alpha, \\ \frac{\dot{\gamma}_3 (\dot{\gamma}_3 - 2\alpha)}{\dot{\gamma}_1 (\dot{\gamma}_1 - 2\alpha)}, & \dot{\gamma}_2 + \dot{\gamma}_3 \leq 2\alpha < \dot{\gamma}_3, \text{ убывает по } \alpha. \end{cases}$$

Следовательно, и в этом случае оптимальное значение параметра α определяется формулой (19), значение γ_1 дано в (20), а

$$\gamma_2 = \lambda_1 = \dot{\gamma}_1 (\Delta_2 - \Delta_1) - \dot{\gamma}_1 \dot{\gamma}_4 \geq \lambda_4. \quad (22)$$

Итак, доказана

Лемма 1. Пусть собственные значения оператора C заключены на отрезках $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$, $\dot{\gamma}_2 < 0 < \dot{\gamma}_3$. Тогда для оператора $\bar{C} = C^2 - \alpha C$ при $\alpha = \alpha_0 = (\gamma_2 + \gamma_3)/2$ справедливы неравенства

$$\gamma_1 E \leq \bar{C} \leq \gamma_2 E, \quad \gamma_1 > 0,$$

где

$$\gamma_1 = -\dot{\gamma}_2 \dot{\gamma}_3, \quad \gamma_2 = \max [\dot{\gamma}_4 (\Delta_2 - \Delta_1), \dot{\gamma}_1 (\Delta_2 - \Delta_1)] - \dot{\gamma}_1 \dot{\gamma}_4.$$

Для указанного значения α отношение γ_1/γ_2 максимально.

Утверждения леммы следуют из (19)–(22). Отметим, что $\alpha_0 = 0$ лишь в случае, когда $\gamma_2 = -\gamma_3$.

3. Итерационный метод с чебышевскими параметрами. Выше мы рассмотрели двухслойную итерационную схему (2), параметры τ_k , $k = 1, 2, \dots, 2n$, которой выражаются через ω_k , $1 \leq k \leq n$ и α по формулам (9). При этом параметры ω_k являются итерационными параметрами чебышевского метода и определяются соответствующими формулами, а необходимая для этого априорная информация и оптимальное значение параметра α даны в лемме 1.

Заметим, что мы предполагали принадлежность собственных значений μ самосопряженного оператора C отрезкам $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$. Из определения (3) оператора C следует, что собственные значения оператора C одновременно являются и собственными значениями следующей задачи:

$$Au - \mu Bu = 0. \quad (23)$$

Чтобы убедиться в этом, достаточно умножить это уравнение слева на оператор $D^{1/2}B^{-1}$ и сделать замену, полагая $u = D^{-1/2}v$. Заметим, что оператор C будет самосопряжен в H , если самосопряжен оператор $DB^{-1}A$.

Сформулируем полученные результаты в виде теоремы.

Теорема 1. Пусть оператор $DB^{-1}A$ самосопряжен в H и собственные значения задачи (23) принадлежат отрезкам $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$, $\dot{\gamma}_1 \leq \dot{\gamma}_2 < 0 < \dot{\gamma}_3 \leq \dot{\gamma}_4$. Для итерационного процесса (2) с параметрами

$$\tau_{2k-1} = -\alpha_0 \omega_k - \sqrt{\alpha_0^2 \omega_k^2 + \omega_k}, \quad \tau_{2k} = -\alpha_0 \omega_k + \sqrt{\alpha_0^2 \omega_k^2 + \omega_k}, \\ k = 1, 2, \dots, n,$$

справедлива оценка

$$\|z_{2n}\|_D \leq q_n \|z_0\|_D,$$

где

$$\omega_k = \frac{\omega_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \quad 1 \leq k \leq n, \\ \omega_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \\ \alpha_0 = 0.5(\dot{\gamma}_2 + \dot{\gamma}_3), \quad \gamma_1 = -\dot{\gamma}_2 \dot{\gamma}_3, \\ \gamma_2 = \max [\dot{\gamma}_4(\Delta_2 - \Delta_1), \dot{\gamma}_1(\Delta_2 - \Delta_1)] - \dot{\gamma}_1 \dot{\gamma}_4, \\ \Delta_1 = \dot{\gamma}_2 - \dot{\gamma}_1, \quad \Delta_2 = \dot{\gamma}_4 - \dot{\gamma}_3.$$

Итерационный метод (2) с указанными параметрами τ_k будем называть чебышевским методом.

Рассмотрим некоторые частные случаи. Пусть $\Delta_1 = \Delta_2$, т. е. длины отрезков $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$ одинаковы. В этом случае имеем

$$\gamma_1 = -\dot{\gamma}_2 \dot{\gamma}_3, \quad \gamma_2 = -\dot{\gamma}_1 \dot{\gamma}_4, \quad \xi = \frac{\dot{\gamma}_2 \dot{\gamma}_3}{\dot{\gamma}_1 \dot{\gamma}_4}.$$

Покажем, что в рассматриваемом случае построенный набор параметров τ_k наилучший. Это утверждение нужно доказывать, поскольку при построении параметров τ_k для схемы (2) мы наложили n условий (7), следовательно, выбор параметров был подчинен дополнительным ограничениям.

Из (5) и (8) найдем, что

$$x_{2n} = Q_{2n}(C)x_0 = P_n(\bar{C})x_0,$$

где

$$Q_{2n}(C) = \prod_{j=1}^{2n} (E - \tau_j C) = P_n(\bar{C}) = \prod_{j=1}^n (E - \omega_j \bar{C}). \quad (24)$$

Рассмотрим соответствующие алгебраические полиномы $Q_{2n}(\mu)$ и $P_n(\lambda)$ ($\lambda = \mu^2 - 2\alpha\mu$). Если параметры ω_j выбраны указанным

в теореме 1 образом, то полином $P_n(\lambda)$ следующим образом выражается через полином Чебышева 1-го рода $T_n(\lambda)$ (см. гл. VI, § 2, п. 1):

$$P_n(\lambda) = q_n T_n\left(\frac{1-\omega_0 \lambda}{\rho_0}\right), \quad P_n(0) = 1,$$

$$\max_{\gamma_1 \leq \lambda \leq \gamma_2} |P_n(\lambda)| = q_n.$$

Заметим, что в точках $\gamma_1 = \lambda_0 < \lambda_1 < \dots < \lambda_n = \gamma_2$, где

$$\lambda_k = \frac{1 - \rho_0 \cos \frac{k\pi}{n}}{\omega_0}, \quad k = 0, 1, \dots, n,$$

полином $P_n(\lambda)$ достигает экстремальных на $[\gamma_1, \gamma_2]$ значений:

$$P_n(\lambda_k) = (-1)^k q_n, \quad k = 0, 1, \dots, n. \quad (25)$$

Так как в силу (24) имеет место равенство $Q_{2n}(\mu) = P_n(\lambda)$, где λ и μ связаны соотношением $\lambda = \mu^2 - 2\alpha\mu$, то из (25) найдем

$$Q_{2n}(\mu_k^-) = Q_{2n}(\mu_k^+) = (-1)^k q_n, \quad k = 0, 1, \dots, n, \quad (26)$$

где μ_k^- и μ_k^+ — корни квадратного уравнения

$$\mu_k^2 - 2\alpha\mu_k - \lambda_k = 0, \quad k = 0, 1, \dots, n. \quad (27)$$

Далее, для рассматриваемого случая преобразование $\lambda = \lambda(\mu) = \mu^2 - 2\alpha\mu$ отображает каждый из отрезков $[\dot{\gamma}_1, \dot{\gamma}_2]$ и $[\dot{\gamma}_3, \dot{\gamma}_4]$ на один отрезок $[\gamma_1, \gamma_2]$. При этом точкам $\mu = \dot{\gamma}_2$, $\mu = \dot{\gamma}_3$ соответствует $\lambda = \gamma_1$, а $\mu = \dot{\gamma}_1$ и $\mu = \dot{\gamma}_4$ соответствует $\lambda = \gamma_2$. Поэтому корни уравнения (27) расположены следующим образом:

$$\dot{\gamma}_1 = \mu_n^- < \mu_{n-1}^- < \dots < \mu_0^- = \dot{\gamma}_2, \quad \dot{\gamma}_3 = \mu_0^+ < \mu_1^+ < \dots < \mu_n^+ = \dot{\gamma}_4.$$

Предположим теперь, что построенный в теореме 1 набор параметров τ_k не наилучший. Это означает, что существует другой полином степени не выше $2n$ вида

$$\bar{Q}_{2n}(\mu) = \prod_{j=1}^{2n} (1 - \bar{\tau}_j \mu),$$

для которого

$$\max_{\mu \in \Omega} |\bar{Q}_{2n}(\mu)| < q_n, \quad \Omega = [\dot{\gamma}_1, \dot{\gamma}_2] \cup [\dot{\gamma}_3, \dot{\gamma}_4].$$

Рассмотрим разность $R_{2n}(\mu) = Q_{2n}(\mu) - \bar{Q}_{2n}(\mu)$, которая есть полином степени не выше $2n$. Доказательство существования $2n+2$ корней у полинома $R_{2n}(\mu)$ приведет нас к выводу о неверности сделанного выше предположения.

Для доказательства рассмотрим значения $R_{2n}(\mu)$ в точках μ_k , $0 \leq k \leq n$. Так как, по предположению, $-q_n < \bar{Q}_{2n}(\mu) < q_n$, $\mu \in \Omega$, то

$$R_{2n}(\mu_k) = Q_{2n}(\mu_k) - \bar{Q}_{2n}(\mu_k) = (-1)^k q_n - \bar{Q}_{2n}(\mu_k)$$

и $R_{2n}(\mu_k) < 0$, если k — нечетное, $R_{2n}(\mu_k) > 0$, если k — четное. Следовательно, при переходе от μ_k к μ_{k+1} , $0 \leq k \leq n-1$, полином $R_{2n}(\mu)$ меняет знак. Поэтому на отрезке $[\dot{\gamma}_1, \dot{\gamma}_2]$ существует n корней этого полинома. Аналогично, рассматривая значения $R_{2n}(\mu)$ в точках μ_k^+ , $0 \leq k \leq n$, докажем существование n корней и на отрезке $[\dot{\gamma}_3, \dot{\gamma}_4]$. Далее, так как

$$\begin{aligned} R_{2n}(\dot{\gamma}_2) &= R_{2n}(\mu_0^-) > 0, & R_{2n}(\dot{\gamma}_3) &= R_{2n}(\mu_0^+) > 0, \\ R_{2n}(0) &= 0, \end{aligned}$$

то в интервале $(\dot{\gamma}_2, \dot{\gamma}_3)$ есть либо два различных (один из которых нуль) корня полинома $R_{2n}(\mu)$, либо нуль есть кратный корень. Следовательно, на отрезке $[\dot{\gamma}_1, \dot{\gamma}_4]$ полином $R_{2n}(\mu)$ имеет $2n+2$ корня, что невозможно.

Итак, для случая $\Delta_1 = \Delta_2$ построенный в теореме 1 набор параметров τ_k наилучший.

Пусть теперь $\Delta_1 \leq \Delta_2$. В этом случае имеем $\dot{\gamma}_1 = -\dot{\gamma}_2 \dot{\gamma}_3$, $\dot{\gamma}_2 = \dot{\gamma}_4 (\Delta_2 - \Delta_1) - \dot{\gamma}_1 \dot{\gamma}_4$. Так как $\dot{\gamma}_2 = \dot{\gamma}_4 (\Delta_2 - \Delta_1) - \dot{\gamma}_1 \dot{\gamma}_4 = \dot{\gamma}_4 (\dot{\gamma}_4 - \dot{\gamma}_3 - \dot{\gamma}_2)$, то $\dot{\gamma}_1$ и $\dot{\gamma}_2$ не зависят от $\dot{\gamma}_1$. Следовательно, для любого $\dot{\gamma}_1$ из интервала $\dot{\gamma}_2 + \dot{\gamma}_3 - \dot{\gamma}_4 \leq \dot{\gamma}_1 \leq \dot{\gamma}_2$ имеем один и тот же набор параметров τ_k , и итерационный метод (2) сходится с одинаковой скоростью для любого $\dot{\gamma}_1$ из указанного интервала.

В заключение заметим, что построенный в теореме 1 набор параметров τ_k будет наилучшим и для случая, когда $n = 1$, а Δ_1 и Δ_2 не обязательно равны. Это есть случай циклического метода простой итерации, для которого в схеме (2) $\tau_{2k-1} = \tau_1$, $\tau_{2k} = \tau_2$, $k = 1, 2, \dots$, а τ_1 и τ_2 находятся по формулам теоремы 1 для $n = 1$ ($\omega_1 = \omega_0$)

$$\tau_1 = -\alpha_0 \omega_0 - \sqrt{\alpha_0^2 \omega_0^2 + \omega_0}, \quad \tau_2 = -\alpha_0 \omega_0 + \sqrt{\alpha_0^2 \omega_0^2 + \omega_0},$$

где $\omega_0 = 2/(\dot{\gamma}_1 + \dot{\gamma}_2)$. Так как в этом случае имеем

$$\begin{aligned} x_{2n} &= \prod_{j=1}^n (E - \omega_0 \bar{C}) x_0 = (E - \omega_0 \bar{C})^n x_0, \\ \|E - \omega_0 \bar{C}\| &\leq \rho_0, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\dot{\gamma}_1}{\dot{\gamma}_2}, \end{aligned}$$

то для погрешности z_{2n} итерационной схемы (2) будем иметь оценку

$$\|z_{2n}\|_D \leq \rho_0^n \|z_0\|_D.$$

Так как в силу (6) и (7) два параметра τ_1 и τ_2 заменяются на параметры ω_1 и α , а последние выбираются оптимальным образом ($\omega_1 = \omega_0$, $\alpha = \alpha_0$), то, действительно, параметры τ_1 и τ_2 для метода простой итерации выбраны наилучшими.

4. Итерационные методы вариационного типа. Выше мы рассмотрели итерационные методы для случая самосопряженного оператора $DB^{-1}A$, когда не все собственные значения задачи (23) одного знака. При этом сходимость итерационного метода (2) обеспечивалась построением специального набора итерационных параметров. Рассмотрим теперь итерационные методы вида (2), сходимость которых при обычном выборе итерационных параметров обеспечивается структурой оператора B . С таким способом построения итерационных схем мы имели дело в методе симметризации уравнения (см. гл. VI, § 4, п. 4) и при изучении методов минимальных погрешностей и сопряженных погрешностей в главе VIII.

Пусть оператор B имеет вид

$$B = (A^*)^{-1}\tilde{B}, \quad (28)$$

где \tilde{B} — произвольный самосопряженный и положительно определенный оператор. В качестве оператора D возьмем \tilde{B} . Тогда $DB^{-1}A = A^*A$, $C = \tilde{B}^{-1/2}A^*A\tilde{B}^{-1/2}$. Если оператор B не вырожден и незакоопределен, то оператор C все равно положительно определен. Кроме того, оператор C самосопряжен в H . Поэтому, если заданы γ_1 и γ_2 в неравенствах $\gamma_1 E \leq C \leq \gamma_2 E$, $\gamma_1 > 0$ или в эквивалентных им неравенствах

$$\gamma_1 \tilde{B} \leq A^*A \leq \gamma_2 \tilde{B}, \quad \gamma_1 > 0, \quad (29)$$

то параметры τ_k в (2) можно выбрать по формулам чебышевского двухслойного метода (см. гл. VI, § 2, п. 1)

$$\begin{aligned} \tau_k &= \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_k^* = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \\ &\quad k = 1, 2, \dots, n, \\ \tau_0 &= \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \end{aligned} \quad (30)$$

Итак, имеет место

Теорема 2. Пусть A невырожденный оператор. Для итерационного метода (2), (28) с параметрами (30), где γ_1 и γ_2 заданы в (29), имеет место оценка

$$\|z_n\|_{\tilde{B}} \leq q_n \|z_0\|_{\tilde{B}}, \quad q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Если постоянные γ_1 и γ_2 в (29) либо неизвестны, либо могут быть оценены слишком грубо, можно воспользоваться рассмотренными в главе VIII итерационными методами вариационного типа.

Если для схемы (2), (28) параметры τ_k выбирать по формулам

$$\tau_{k+1} = \frac{(r_k, r_k)}{(A\omega_k, r_k)}, \quad k = 0, 1, \dots,$$

где $r_k = Ay_k - f$ — невязка, а ω_k — поправка, определяемая из уравнения $\tilde{B}\omega_k = A^*r_k$, то получим метод минимальных погрешностей (см. гл. VIII, § 2, п. 4). Как известно, для погрешности z_n этого метода имеет место оценка $\|z_n\|_{\tilde{B}} \leq \rho_0^n \|z_0\|_{\tilde{B}}$, где ρ_0 определено в (30).

Если рассмотреть трехслойную итерационную схему

$$\begin{aligned} By_{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \tau_{k+1}\alpha_{k+1}f, \quad k \geq 1, \\ By_1 &= (B - \tau_1A)y_0 + \tau_1f, \quad y_0 \in H, \end{aligned}$$

где оператор B определен в (28), и выбрать итерационные параметры α_{k+1} и τ_{k+1} по формулам

$$\tau_{k+1} = \frac{(r_k, r_k)}{(A\omega_k, r_k)}, \quad k = 0, 1, \dots,$$

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \frac{1}{\alpha_k}\right)^{-1}, \quad k = 1, 2, \dots, \quad \alpha_1 = 1,$$

то получим метод сопряженных погрешностей (см. гл. VIII, § 4, п. 1). Для погрешности этого метода верна оценка

$$\|z_n\|_{\tilde{B}} \leq q_n \|z_0\|_{\tilde{B}}.$$

5. Примеры. Рассмотрим применение построенных выше методов к нахождению решения разностной задачи Дирихле для уравнения Гельмгольца в прямоугольнике

$$\begin{aligned} y_{x_1 x_1}^- + y_{x_2 x_2}^- + m^2 y &= -f(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma, \end{aligned} \tag{31}$$

где $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$, а γ — граница сетки ω .

Сведем задачу (31) к операторному уравнению (1). В данном случае H — пространство сеточных функций, заданных на ω со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2, \quad u \in H, v \in H.$$

Определим оператор R следующим образом: $Ry = -\Lambda \dot{y}$, $y \in H$, $\dot{y} \in \dot{H}$ и $y(x) = \underline{y}(x)$, $x \in \omega$, где \dot{H} — множество сеточных функций, заданных на ω и обращающихся в нуль на γ , а Λ есть разностный оператор Лапласа $\Lambda y = y_{x_1 x_1}^- + y_{x_2 x_2}^-$. Тогда оператор A определяется равенством $A = R - m^2 E$. Так как оператор R самосопряжен в H и имеет собственные значения

$$\lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad \lambda_{k_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi h_\alpha}{2l_\alpha}, \quad 1 \leq k_\alpha \leq N_\alpha - 1,$$

то оператор A также самосопряжен в H , и его собственные значения μ_k выражаются через λ_k по формуле

$$\mu_k = \lambda_k - m^2, \quad k = (k_1, k_2), \quad 1 \leq k_\alpha \leq N_\alpha - 1, \quad \alpha = 1, 2. \quad (32)$$

Пусть m^2 не совпадает ни с одним λ_k . Обозначим через λ_{m_1} и λ_{m_2} ближайшие к m^2 собственные значения λ_k соответственно снизу и сверху, т. е.

$$\lambda_{m_1} < m^2 < \lambda_{m_2}. \quad (33)$$

В этом случае оператор A не вырожден и незнакоопределен.

Для решения уравнения (1) с указанным оператором A рассмотрим явную итерационную схему (2) ($B = E$). Если положить $D = E$, то оператор $DB^{-1}A$ совпадает с A и является самосопряженным в H . Выбор итерационных параметров в этом случае можно осуществить, используя теорему 1. Из (23) получим, что необходимая априорная информация задается границами отрезков $[\dot{\gamma}_1, \dot{\gamma}_2]$, $[\dot{\gamma}_3, \dot{\gamma}_4]$ положительной и отрицательной полуосей, на которых расположены собственные значения оператора A .

Из (32) и (33) найдем $\dot{\gamma}_1 = \delta - m^2$, $\dot{\gamma}_2 = \lambda_{m_1} - m^2$, $\dot{\gamma}_3 = \lambda_{m_2} - m^2$, $\dot{\gamma}_4 = \Delta - m^2$, где

$$\delta = \min_k \lambda_k = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \max_k \lambda_k = \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}.$$

Найдем теперь γ_1 , γ_2 и величину $\sqrt{\xi}$, которая определяет число итераций для рассматриваемого метода, так что $n \geq n_0(\varepsilon) = \ln(2/\varepsilon)/(2\sqrt{\xi})$. Из формул теоремы 1 найдем

$$\begin{aligned} \gamma_1 &= (m^2 - \lambda_{m_1})(\lambda_{m_2} - m^2), \\ \gamma_2 &= \begin{cases} (\Delta - m^2)(\Delta + m^2 - \lambda_{m_1} - \lambda_{m_2}), & \lambda_{m_1} + \lambda_{m_2} \leq (\Delta + \delta), \\ (m^2 - \delta)(\lambda_{m_1} + \lambda_{m_2} - m^2 - \delta), & \lambda_{m_1} + \lambda_{m_2} \geq (\Delta + \delta). \end{cases} \end{aligned}$$

Отношение $\xi = \gamma_1/\gamma_2$ зависит от m^2 . Чтобы получить представление о качестве рассматриваемого итерационного метода, найдем значение m^2 из интервала $(\lambda_{m_1}, \lambda_{m_2})$, при котором ξ максимальна. Получим

$$m^2 = 0,5(\lambda_{m_1} + \lambda_{m_2}),$$

при этом

$$\begin{aligned} \gamma_1 &= \left(\frac{\lambda_{m_2} - \lambda_{m_1}}{2} \right)^2, \\ \gamma_2 &= \begin{cases} (\Delta - m^2)^2, & 2m^2 \leq \Delta + \delta, \\ (m^2 - \delta)^2, & 2m^2 \geq \Delta + \delta. \end{cases} \end{aligned}$$

Если m^2 мало, т. е. λ_{m_1} и λ_{m_2} близки к δ , то $\gamma_1 = O(1)$, а $\gamma_2 = (\Delta - m^2)^2 = O\left(\frac{1}{|h|^4}\right)$. В этом случае наилучшее $\xi = O(|h|^4)$. Если λ_{m_1} и λ_{m_2} близки к Δ , то снова получим $\xi = O(|h|^4)$. Лишь для случая, когда λ_{m_1} и λ_{m_2} близки к $0,5(\Delta + \delta)$, получим

$$\gamma_1 = O\left(\frac{1}{|h|^2}\right) \text{ и } \gamma_2 = O\left(\frac{1}{|h|^4}\right),$$

так что

$$\xi = O(|h|^2).$$

Заметим, что разностная задача (31) может быть решена одним из прямых методов, рассмотренных нами в главах III, IV: либо методом полной редукции, либо методом разделения переменных. Возникающие при этом трехточечные краевые задачи следует решать, в отличие от случая $m=0$, методом немонотонной прогонки.

§ 2. Уравнения с комплексным оператором

1. Метод простой итерации. Пусть в комплексном гильбертовом пространстве H дано уравнение

$$Au + qu = f, \quad (1)$$

где A — эрмитов оператор, а $q = q_1 + iq_2$ — комплексное число. Для приближенного решения уравнения (1) рассмотрим явную двухслойную схему

$$\frac{y_{k+1} - y_k}{\tau} + (A + qE)y_k = f, \quad 'k = 0, 1, \dots, \quad y_0 \in H, \quad (2)$$

где $\tau = \tau_1 + i\tau_2$ — комплексный итерационный параметр.

Будем предполагать, что $q_1 \neq 0$, а γ_1 и γ_2 — постоянные в неравенствах

$$\gamma_1 E \leq A \leq \gamma_2 E. \quad (3)$$

Исследуем сходимость итерационной схемы (2) в энергетическом пространстве H ($D = E$) и найдем оптимальное значение для итерационного параметра τ . Используя (1) и (2), запишем уравнение для погрешности $x_k = y_k - u$ в виде:

$$x_{k+1} = Sx_k, \quad k = 0, 1, \dots, \quad S = E - \tau C, \quad (4)$$

где

$$C = A + qE.$$

Из (4) найдем

$$x_n = S^n x_0, \quad \|x_n\| \leq \|S^n\| \|x_0\|. \quad (5)$$

Изучим оператор перехода от итерации к итерации. Так как оператор A эрмитов, то

$$C^* = A + \bar{q}E, \quad C^*C = CC^*,$$

т. е. оператор C является нормальным оператором. Поэтому нормальным является и оператор S . Известно (см. гл. V, § 1, п. 2), что для нормального оператора S справедливы следующие соотношения:

$$\|S^n\| = \|S\|^n, \quad \|S\| = \sup_{x \neq 0} \frac{|(Sx, x)|}{(x, x)}.$$

Поэтому из (5) следует, что задача выбора итерационного параметра τ сводится к нахождению его из условия минимума нормы оператора S .

Решим эту задачу. Из (3) будет следовать, что

$$z = \frac{(Cx, x)}{(x, x)} \in \Omega,$$

$$\Omega = \{z = z_1 + a(z_2 - z_1), \quad 0 \leq a \leq 1, \quad z_1 = \gamma_1 + q, \quad z_2 = \gamma_2 + q\},$$

где Ω — отрезок в комплексной плоскости, соединяющий точки z_1 и z_2 . Поэтому

$$\|S\| = \sup_{x \neq 0} \frac{|(Sx, x)|}{(x, x)} = \sup_{z \in \Omega} |1 - \tau z|$$

и параметр τ ищется, исходя из условия $\min_{\tau} \max_{z \in \Omega} |1 - \tau z|$.

Исследуем функцию $\varphi(z) = |1 - \tau z|$. Так как линии уровня $|1 - \tau z| = \rho_0$ есть концентрические окружности с центром в точке $1/\tau$ и радиуса $R = \rho_0/|\tau|$, то для оптимального значения параметра $\tau = \tau_0$ точки z_1 и z_2 должны лежать на одной линии уровня. Следовательно, должны выполняться равенства

$$|1 - \tau_0 z_1| = \rho_0, \quad |1 - \tau_0 z_2| = \rho_0,$$

причем $|1 - \tau_0 z| \leq \rho_0$ для $z \in \Omega$.

Запишем эти равенства в эквивалентном виде

$$\left| \frac{1 - \tau_0 z_2}{1 - \tau_0 z_1} \right| = 1, \quad \rho_0 = \frac{|z_2 - z_1|}{|z_1|} \left| \frac{z_2}{z_1} - \frac{1 - \tau_0 z_2}{1 - \tau_0 z_1} \right|.$$

Так как в силу первого равенства при изменении τ_0 комплексное число

$$z = \frac{1 - \tau_0 z_2}{1 - \tau_0 z_1}$$

пробегает единичную окружность в комплексной плоскости с центром в начале координат, то ρ_0 будет минимально, если выполняется равенство

$$\frac{1-\tau_0 z_2}{1-\tau_0 z_1} = -\frac{z_2}{z_1} \frac{|z_1|}{|z_2|}.$$

Это условие дает следующее значение τ_0 :

$$\tau_0 = \frac{|z_2|/z_2 + |z_1|/z_1}{|z_1| + |z_2|}. \quad (6)$$

При этом значении $\tau = \tau_0$ для нормы оператора S верна оценка

$$\|S\| = \rho_0 = \frac{|z_2 - z_1|}{|z_1| + |z_2|}, \quad (7)$$

используя которую, получим для погрешности x_n итерационной схемы (2) оценку

$$\|x_n\|_B \leq \rho_0^n \|x_0\|_B. \quad (8)$$

Найдем теперь условия, при выполнении которых $\rho_0 < 1$. Так как справедливо неравенство

$$|z_2 - z_1| = |z_1| \left| \frac{z_2}{|z_1|} - \frac{z_1}{|z_1|} \right| \leq |z_1| \left(1 + \frac{|z_2|}{|z_1|} \right) = |z_1| + |z_2|,$$

причем здесь достигается равенство лишь при выполнении условия

$$\frac{z_1}{|z_1|} = -\frac{z_2}{|z_1|} \frac{|z_1|}{|z_2|} = -\frac{z_2}{|z_2|}, \quad (9)$$

то $\rho_0 < 1$, если (9) не имеет места.

В рассматриваемом нами случае $z_1 = \gamma_1 + q$ и $z_2 = \gamma_2 + q$. Из (9) легко находим, что в двух случаях $\rho_0 < 1$: либо $q_2 \neq 0$ и γ_1 и γ_2 любые, либо $q_2 = 0$, но γ_1 и γ_2 подчинены условию $(\gamma_1 + q_1)(\gamma_2 + q_1) > 0$. Будем считать далее эти условия выполнеными. Тогда итерационный процесс (2) будет сходящимся.

Теорема 3. Пусть A — эрмитов оператор и выполнены неравенства (3). Итерационный процесс (2) с параметром

$$\tau = \tau_0 = \frac{1}{|\gamma_1 + q| + |\gamma_2 + q|} \left(\frac{|\gamma_1 + q|}{\gamma_1 + q} + \frac{|\gamma_2 + q|}{\gamma_2 + q} \right)$$

сходится в H , и для погрешности имеет место оценка (8), где

$$\rho_0 = \frac{|\gamma_2 - \gamma_1|}{|\gamma_1 + q| + |\gamma_2 + q|}.$$

Замечание. Выше была решена задача о нахождении оптимального параметра τ из условия $\min_{\tau} \max_{z \in \Omega} |1 - \tau z|$, где Ω — отрезок комплексной плоскости, соединяющий две точки z_1 и z_2 . Легко найти решение этой задачи и в случае, когда Ω есть

круг с центром в точке z_0 радиуса $r_0 < |z_0|$, т. е. не включающий в себя начало координат. Решение поставленной задачи имеет вид

$$\tau_0 = \frac{1}{z_0}, \quad \sup_{z \in \Omega} |1 - \tau_0 z| = \rho_0 = \frac{r_0}{|z_0|} < 1.$$

Рассмотрим теперь использование построенного метода для нахождения решения следующей разностной задачи:

$$\begin{aligned} Au - qu &= -\varphi(x), \quad x \in \omega, \\ u(x) &= g(x), \quad x \in \gamma, \quad q = q_1 + iq_2, \\ \Lambda &= \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha u = (a_\alpha u_{x_\alpha})_{x_\alpha}, \quad \alpha = 1, 2, \end{aligned} \quad (10)$$

где $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ — сетка в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, а коэффициенты $a_\alpha(x)$ вещественны и удовлетворяют условиям

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad x \in \bar{\omega}. \quad (11)$$

В рассматриваемом случае H — пространство комплекснозначных сеточных функций, заданных на ω , со скалярным произведением

$$(u, v) = \sum_{x \in \omega} u(x) \bar{v}(x) h_1 h_2.$$

Задача (10) записывается в виде уравнения (1), где оператор A определяется обычным образом: $Ay = -\Lambda y$, где $\dot{y} \in \dot{H}$, $y(x) = \dot{y}(x)$ для $x \in \omega$, $\dot{y}(x) = 0$, $x \in \gamma$.

Для решения построенного уравнения (1) рассмотрим явную итерационную схему (2).

Используя разностные формулы Грина для комплекснозначных функций, а также неравенства (11), убедимся в том, что оператор A является эрмитовым в H , а в неравенствах (3)

$$\gamma_1 = c_1 \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha},$$

$$\gamma_2 = c_2 \sum_{\alpha=1}^2 \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}.$$

Если выбрать итерационный параметр τ согласно теореме 3, то для погрешности $x_n = y_n - u$ будет иметь место оценка (8), где ρ_0 определено в теореме 3.

В частном случае, когда $l_1 = l_2 = l$, $N_1 = N_2 = N$ и $q_1 = O(1)$, $q_2 = O(1)$, получим $\rho_0 = 1 - O(N^{-2})$. Следовательно, для достижения заданной точности ε потребуется выполнить $n_0(\varepsilon) = O\left(N^2 \ln \frac{1}{\varepsilon}\right)$ итераций.

2. Метод переменных направлений. Рассмотрим снова уравнение (1) и предположим, что оператор A можно представить в виде суммы двух эрмитовых перестановочных операторов A_1 и A_2 :

$$A = A_1 + A_2, \quad A_1 A_2 = A_2 A_1, \quad A_\alpha = A_\alpha^*, \quad \alpha = 1, 2. \quad (12)$$

Пусть δ и Δ —границы операторов A_1 и A_2 , т. е.

$$\delta E \leq A_\alpha \leq \Delta E, \quad \alpha = 1, 2. \quad (13)$$

Для решения уравнения (1) рассмотрим неявную двухслойную итерационную схему (2), в которой оператор B задан следующим образом:

$$B = (\omega E + A_1 + q_0 E) (\omega E + A_2 + q_0 E), \quad q_0 = 0,5q, \quad (14)$$

а параметры τ и ω связаны соотношением $\tau = 2\omega$. Аналогичную итерационную схему мы получили в главе XI при построении метода переменных направлений. Заметим, что для нахождения y_{k+1} в схеме (2), (14) можно воспользоваться следующим алгоритмом:

$$\begin{aligned} (\omega E + C_1) y_{k+1/2} &= (\omega E - C_2) y_k + f, \\ (\omega E + C_2) y_{k+1} &= (\omega - C_1) y_{k+1/2} + f, \quad k = 0, 1, \dots, \end{aligned}$$

где для сокращения обозначений $C_\alpha = A_\alpha + q_0 E$, $\alpha = 1, 2$.

Переходим к исследованию сходимости схемы (2), (14) в норме H . Воспользовавшись перестановочностью операторов A_1 и A_2 , получим уравнение для погрешности z_k

$$z_{k+1} = S_1 S_2 z_k, \quad k = 0, 1, \dots, \quad (15)$$

$$S_\alpha = (\omega E + C_\alpha)^{-1} (\omega E - C_\alpha), \quad \alpha = 1, 2, \quad (16)$$

причем операторы S_1 и S_2 перестановочны. Из (15) найдем

$$z_n = S_1^n S_2^n z_0, \quad \|z_n\| \leq \|S_1^n\| \|S_2^n\| \|z_0\|. \quad (17)$$

Оценим норму оператора S_α^n , $\alpha = 1, 2$. Поскольку C_α —нормальный оператор ($C_\alpha^* C_\alpha = C_\alpha C_\alpha^*$, $\alpha = 1, 2$), то нормальным будет и оператор S_α . Поэтому $\|S_\alpha^n\| = \|S_\alpha\|^n$ и достаточно оценить норму самого оператора S_α .

Так как норма нормального оператора равна его спектральному радиусу (см. гл. V, § 1, п. 2), то из (16) получим

$$\|S_\alpha\| = \max_{\lambda_\alpha} \left| \frac{\omega - \lambda_\alpha}{\omega + \lambda_\alpha} \right|, \quad (18)$$

где λ_α —собственные значения оператора C_α . В силу сделанных

предположений (12) и (13) относительно операторов A_α получим, что $\lambda_\alpha \in \Omega = \{z = z_1 + a(z_2 - z_1), 0 \leq a \leq 1, z_1 = \delta + q_0, z_2 = \Delta + q_0\}$ для $\alpha = 1, 2$. Следовательно, из (18) получим, что

$$\|S_\alpha\| \leq \max_{z \in \Omega} \left| \frac{\omega - z}{\omega + z} \right|, \quad \alpha = 1, 2. \quad (19)$$

Поставим теперь задачу выбрать параметр ω из условия минимума правой части неравенства (19).

Рассмотрим дробно-линейное отображение

$$w = (\omega - z)/(\omega + z), \quad \omega \neq 0, \quad (20)$$

устанавливающее соотношение между точками z -плоскости и точками w -плоскости. Из свойств преобразования (20) следует, что окружностям $|w| = \rho_0$ в w -плоскости при $\rho \neq 1$ соответствуют окружности в z -плоскости, а единичной окружности соответствует в z -плоскости прямая, проходящая через начало координат. Точки указанной прямой имеют аргумент, отличающийся от аргумента ω на $\pm \pi/2$.

Найдем в z -плоскости центр и радиус окружности, соответствующей окружности $|w| = \rho_0 \neq 1$ в w -плоскости. Для этого выразим из (20) z через w :

$$z = \omega(1-w)/(1+w),$$

и, используя это соотношение, вычислим

$$\left| \frac{1+|w|^2}{1-|w|^2} \omega - z \right| = \left| \frac{1+|w|^2}{1-|w|^2} \omega - \frac{1-w}{1+w} \omega \right| = \frac{2|\omega|}{|1-|w|^2|} \cdot \frac{|w+|w|^2|}{|1+w|}.$$

Так как

$$|w+|w|^2| = |w+w\bar{w}| = |w||-1+\bar{w}| = |w||1+w|,$$

то окончательно получим

$$\left| \frac{1+|w|^2}{1-|w|^2} \omega - z \right| = \frac{2|\omega||w|}{|1-|w|^2|}.$$

Отсюда следует, что окружностям $|w| = \rho_0 < 1$ соответствуют окружности в z -плоскости с центром в точке z_0 радиуса R , где

$$z_0 = \frac{1+\rho_0^2}{1-\rho_0^2} \omega, \quad R = \frac{2\rho_0|\omega|}{1-\rho_0^2}. \quad (21)$$

Заметим, кроме того, что в силу взаимной однозначности отображения (20), равенства

$$\left| \frac{\omega-z}{\omega+z} \right| = \rho_0 < 1, \quad |z_0 - z| = R \quad (22)$$

эквивалентны.

Вернемся к поставленной задаче. Рассмотрим функцию

$$\varphi(z) = |w| = \left| \frac{\omega - z}{\omega + z} \right|.$$

Из сказанного выше следует, что линии уровня $\varphi(z) = \rho_0$ при $\rho_0 < 1$ есть окружности с центром в точке z_0 радиуса R , где z_0 и R определены в (21). Для разных ρ_0 эти окружности не пересекаются, причем окружность, соответствующая меньшему значению ρ_0 , лежит внутри окружности, соответствующей большему значению ρ_0 . Отсюда получим, что для оптимального значения $\omega = \omega_0$ точки z_1 и z_2 должны лежать на одной линии уровня:

$$\left| \frac{\omega_0 - z_1}{\omega_0 + z_1} \right| = \rho_0 < 1, \quad \left| \frac{\omega_0 - z_2}{\omega_0 + z_2} \right| = \rho_0 < 1, \quad (23)$$

при этом будет выполняться равенство

$$\max_{z \in \Omega} \left| \frac{\omega_0 - z}{\omega_0 + z} \right| = \rho_0.$$

Параметр ω_0 должен быть выбран из условия минимума ρ_0 .

Найдем оптимальное ω_0 и вычислим ρ_0 . Из (23) в силу (22) получим

$$|z_0 - z_1| = R_0, \quad |z_0 - z_2| = R_0,$$

$$z_0 = \frac{1 + \rho_0^2}{1 - \rho_0^2} \omega_0, \quad R_0 = \frac{2\rho_0 |\omega_0|}{1 - \rho_0^2}$$

или

$$\left| \frac{z_0 - z_2}{z_0 - z_1} \right| = 1, \quad \frac{2\rho_0}{1 + \rho_0^2} = \frac{R_0}{|z_0|} = \frac{|z_2 - z_1|}{|z_1| \left| \frac{z_2 - z_0 - z_2}{z_1 - z_0 - z_1} \right|}. \quad (24)$$

Заметим, что ρ_0 минимально, когда минимально $\frac{2\rho_0}{1 + \rho_0^2}$, а это имеет место, если потребовать выполнения равенства

$$\frac{z_0 - z_2}{z_0 - z_1} = - \frac{z_2}{z_1} \frac{|z_1|}{|z_2|}. \quad (25)$$

Подставляя это выражение в (24), получим

$$\frac{2\rho_0}{1 + \rho_0^2} = \frac{|z_2 - z_1|}{|z_1| + |z_2|}.$$

Отсюда легко найдем

$$\gamma_1 = \frac{(1 - \rho_0)^2}{1 + \rho_0^2} = \frac{|z_1| + |z_2| - |z_2 - z_1|}{|z_1| + |z_2|},$$

$$\gamma_2 = \frac{(1 + \rho_0)^2}{1 + \rho_0^2} = \frac{|z_1| + |z_2| + |z_2 - z_1|}{|z_1| + |z_2|}, \quad \xi = \frac{\gamma_1}{\gamma_2} = \left(\frac{1 - \rho_0}{1 + \rho_0} \right)^2. \quad (26)$$

Следовательно,

$$\rho_0 = \frac{1 - V\bar{\xi}}{1 + V\bar{\xi}}, \quad \frac{1 - \rho_0^2}{1 + \rho_0^2} = \sqrt{\gamma_1 \gamma_2}$$

и, кроме того,

$$z_0 = \frac{1 + \rho_0^2}{1 - \rho_0^2} \omega_0 = \frac{\omega_0}{V\sqrt{\gamma_1 \gamma_2}}.$$

Подставляя это выражение в (25), найдем оптимальное значение параметра ω_0 :

$$\omega_0 = \frac{|z_1| + |z_2|}{|z_1|/z_1 + |z_2|/z_2} \sqrt{\gamma_1 \gamma_2}. \quad (27)$$

Итак, для оптимального $\omega = \omega_0$ получена оценка нормы оператора S_α : $\|S_\alpha\| \leq \rho_0$, $\alpha = 1, 2$. Подставляя ее в (17), найдем оценку для погрешности z_n :

$$\|z_n\| \leq \rho_0^{2n} \|z_0\|, \quad \rho_0 = \frac{1 - V\bar{\xi}}{1 + V\bar{\xi}}, \quad \bar{\xi} = \frac{\gamma_1}{\gamma_2}. \quad (28)$$

Рассуждая, как и в методе простой итерации, найдем, что неравенство $\gamma_1 > 0$, а вместе с ним и неравенство $\rho_0 < 1$ будут иметь место в двух случаях: либо при $q_2 \neq 0$, либо при $q_2 = 0$, но $(\delta + 0,5q_1)(\Delta + 0,5q_1) > 0$.

Итак, доказана следующая

Теорема 4. Пусть выполнены условия (12), заданы δ и Δ из неравенств (13) и либо $q_2 \neq 0$, либо $q_2 = 0$ и $(\delta + 0,5q_1)(\Delta + 0,5q_1) > 0$. Для метода переменных направлений (2), (14), в котором итерационный параметр $\omega = \omega_0$ выбран по формуле (27), а $\tau = 2\omega_0$, справедлива оценка (28), где γ_1 и γ_2 определены в (26), а $z_1 = \delta + 0,5q$ и $z_2 = \Delta + 0,5q$.

Замечание 1. Решение задачи $\min_{\omega} \max_{z \in \Omega} \left| \frac{\omega - z}{\omega + z} \right|$, где Ω — круг с центром в точке z_0 радиуса $r_0 < |z_0|$ имеет вид

$$\omega_0 = z_0 V\sqrt{\gamma_1 \gamma_2}, \quad \rho_0 = \max_{z \in \Omega} \left| \frac{\omega - z}{\omega + z} \right| = \frac{1 - V\bar{\xi}}{1 + V\bar{\xi}}, \quad \bar{\xi} = \frac{\gamma_1}{\gamma_2},$$

где $\gamma_1 = 1 - r_0/|z_0|$, $\gamma_2 = 1 + r_0/|z_0|$.

Замечание 2. Если вместо неравенств (13) заданы неравенства $\delta_\alpha E \leq A_\alpha \leq \Delta_\alpha E$, $\alpha = 1, 2$, то в теореме 4 следует положить $\delta = \min(\delta_1, \delta_2)$, $\Delta = \max(\Delta_1, \Delta_2)$.

§ 3. Общие итерационные методы для уравнений с вырожденным оператором

1. Итерационные схемы в случае невырожденного оператора B .

Пусть в конечномерном гильбертовом пространстве $H = H_N$ дано уравнение

$$Au = f \quad (1)$$

с линейным вырожденным оператором A . Последнее означает, что равенство $Au = 0$ имеет место для некоторого $u \neq 0$. Напомним (см. гл. V, § 2, п. 2) сведения, относящиеся к проблеме решения уравнения (1).

Пусть $\ker A$ — ядро оператора A , т. е. множество элементов $u \in H$, для которых $Au = 0$. Через $\text{im } A$ — образ оператора A — обозначим множество элементов вида $y = Au$, где $u \in H$. Известно, что имеют место следующие ортогональные разложения пространства H в прямые суммы двух подпространств:

$$H = \ker A \oplus \text{im } A^*, \quad H = \ker A^* \oplus \text{im } A. \quad (2)$$

Это означает, что любой элемент $u \in H$ можно представить в виде $u = \bar{u} + \tilde{u}$, где $\bar{u} \in \text{im } A^*$ и $\tilde{u} \in \ker A$, причем $(\bar{u}, \tilde{u}) = 0$. Аналогично $u = \bar{u} + \tilde{u}$, где $\bar{u} \in \text{im } A$ и $\tilde{u} \in \ker A^*$, $(\bar{u}, \tilde{u}) = 0$.

Пусть в уравнении (1) $f = \bar{f} + \tilde{f}$, где $\bar{f} \in \text{im } A$, $\tilde{f} \in \ker A^*$. Обобщенным решением (1) называется элемент $u \in H$, для которого $Au = \bar{f}$; оно доставляет минимум функционалу $\|Au - f\|$. Обобщенное решение не единственно и определяется с точностью до элемента из $\ker A$. Нормальное решение — обобщенное решение, имеющее минимальную норму. Нормальное решение единственно и принадлежит $\text{im } A^*$.

Нашей задачей является построение методов, позволяющих приближенно находить нормальное решение уравнения (1). При этом будем требовать, чтобы приближенное решение, так же как и точное нормальное решение, принадлежало подпространству $\text{im } A^*$.

Для решения поставленной задачи будем использовать неявную двухслойную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H. \quad (3)$$

Сначала изучим случай невырожденного в H оператора B . Общие требования к итерационному процессу таковы:

а) итерации проводятся по схеме (3), приближение $y_n \in \text{im } A^*$, в то время как промежуточные приближения y_k могут принадлежать H ;

б) конкретная структура подпространств $\ker A$, $\ker A^*$, $\text{im } A$ и $\text{im } A^*$ в процессе итераций не используется.

Найдем условия на оператор B , начальное приближение y_0 и параметры τ_k , $k = 1, 2, \dots, n$, которые обеспечивают выполнение сформулированных выше требований.

Условия 1. Пусть оператор B такой, что

$$Bu \in \ker A^*, \quad \text{если } u \in \ker A, \quad (4)$$

$$Bu \in \text{im } A, \quad \text{если } u \in \text{im } A^*. \quad (5)$$

Имеет место

Лемма 2. Если для операторов A и B справедливы равенства
 $A^*B = CA$, $BA^* = AD$, (6)

где C и D —некоторые операторы, то условия (4) и (5) выполнены.

Действительно, пусть выполнены равенства (6). Если $u \in \ker A$, то $Au = 0$ и, следовательно, $A^*Bu = CAu = 0$. Поэтому $Bu \in \ker A^*$, и (4) выполнено. Пусть теперь $u \in \text{im } A^*$, т. е. $u = A^*v$, где $v \in H$. Тогда $Bu = BA^*v = ADv \in \text{im } A$. Следовательно, условие (5) выполнено. Лемма доказана.

Следствие. Для случая $A = A^*$ условия леммы 2 будут выполнены, если операторы A и B перестановочны: $AB = BA$.

Приведем еще ряд утверждений, вытекающих из (4) и (5).

Лемма 3. Пусть выполнены условия (4) и (5). Тогда

$$B^{-1}u \in \ker A, \quad \text{если } u \in \ker A^*, \quad (7)$$

$$B^{-1}u \in \text{im } A^*, \quad \text{если } u \in \text{im } A, \quad (8)$$

и оператор AB^{-1} не вырожден на $\text{im } A$.

Действительно, пусть $u \in \ker A^*$ и $u \neq 0$. Обозначим $v = B^{-1}u$ и предположим, что $v \in \text{im } A^*$. Тогда в силу (5) $u = Bu \in \text{im } A$. Но так как $u \neq 0$, а пространства $\text{im } A$ и $\ker A^*$ ортогональны, то сделанное предположение неверно. Следовательно, $v = -B^{-1}u \in \ker A$, и (7) доказано. Аналогично доказывается (8).

Докажем теперь невырожденность AB^{-1} на подпространстве $\text{im } A$. Действительно, пусть $u \in \text{im } A$. Тогда в силу (8) $B^{-1}u \in \ker A^*$ и, следовательно, $B^{-1}u \perp \ker A$. Отсюда получим, что $AB^{-1}u \neq 0$, и поэтому $(AB^{-1}u, AB^{-1}u) > 0$. Лемма доказана.

Вернемся теперь к схеме (3) и посмотрим, что дает условие 1. В соответствии с разложением H в виде (2) представим f и y_k для любого k в виде

$$\begin{aligned} f &= \bar{f} + \tilde{f}, & \bar{f} &\in \text{im } A, & \tilde{f} &\in \ker A^*, \\ y_k &= \bar{y}_k + \tilde{y}_k, & \bar{y}_k &\in \text{im } A^*, & \tilde{y}_k &\in \ker A. \end{aligned} \quad (9)$$

Используя (9), запишем схему (3) следующим образом:

$$B \frac{\bar{y}_{k+1} - \bar{y}_k}{\tau_{k+1}} + B \frac{\tilde{y}_{k+1} - \tilde{y}_k}{\tau_{k+1}} + A \bar{y}_k = \bar{f} + \tilde{f}, \quad k = 0, 1, \dots \quad (10)$$

Из (4) и (5) получим, что первое слагаемое левой части (10) принадлежит $\text{im } A$, а второе — $\ker A^*$. Поэтому из (10) найдем уравнение

$$B \frac{\bar{y}_{k+1} - \bar{y}_k}{\tau_{k+1}} + A \bar{y}_k = \tilde{f}, \quad k = 0, 1, \dots, \quad \bar{y}_0 \in \text{im } A^* \quad (11)$$

для компоненты $\bar{y}_k \in \text{im } A^*$ и уравнение

$$B \frac{\tilde{y}_{k+1} - \tilde{y}_k}{\tau_{k+1}} = \tilde{f}, \quad k = 0, 1, \dots, \quad \tilde{y}_0 \in \ker A \quad (12)$$

для компоненты $\tilde{y}_k \in \ker A$.

Найдем условия, при выполнении которых $y_n \in \text{im } A^*$. Из (9) следует, что если $\tilde{y}_n = 0$, то $y_n = \tilde{y}_n \in \text{im } A^*$. Найдем из (12) явное выражение для \tilde{y}_n и приравняем его нулю. Тогда сформулированное требование а) будет выполнено.

Из (12) получим

$$\tilde{y}_{k+1} = \tilde{y}_k + \tau_{k+1} B^{-1} \tilde{f} = \dots = \tilde{y}_0 + \sum_{j=1}^{k+1} \tau_j B^{-1} \tilde{f}.$$

Отсюда следуют

Условия 2. Пусть $y_0 = A^* \varphi$, где $\varphi \in H$, а параметры τ_k , $k = 1, 2, \dots, n$, удовлетворяют требованию

$$\sum_{j=1}^n \tau_j = 0, \quad (13)$$

если $f \in H$. Если $f \perp \ker A^*$, то ограничение на параметры τ_k не накладывается.

Поясним выбор начального приближения y_0 . Так как для любого $\varphi \in H$ имеем, что $y_0 = A^* \varphi \in \text{im } A^*$, то в разложении (9) $\tilde{y}_0 = 0$ и $y_0 = \tilde{y}_0$. В частности, выбирая $\varphi = 0$, получим начальное приближение $y_0 = 0$.

Итак, если выполнены условия 2, то $y_n = \tilde{y}_n$. Поэтому итерационный процесс (3) будет сходиться и давать приближенное нормальное решение уравнения (1), если сходится итерационный процесс (11), т. е. если последовательность \tilde{y}_k сходится к нормальному решению \bar{u} .

Замечание 1. Условия 2 позволяют выделить из итерационного приближения y_n его проекцию на $\text{im } A^*$, т. е. найти \tilde{y}_n без использования самих подпространств $\ker A$, $\text{im } A$, $\ker A^*$ или $\text{im } A^*$. Далее, если известно, что $\sum_{j=1}^n \tau_j \|B^{-1} \tilde{f}\|$ мала, т. е. мала $\|\tilde{y}_n\|$, то в силу равенства $\|y_n - \bar{u}\| = \|\tilde{y}_n - \bar{u}\| + \|\tilde{y}_n\|$ в качестве приближенного решения можно взять y_n и отказаться от ограничения (13). В этом случае $y_n \in \text{im } A^*$.

Замечание 2. При условии, что все элементы подпространства $\ker A^*$ известны, можно ограничиться рассмотрением случая $f \perp \ker A^*$, вычитая при необходимости из f его проекцию на $\ker A^*$. Если рассмотреть нестационарный итерационный процесс

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \dots, \quad y_0 = A^* \varphi,$$

и потребовать выполнения условий 1, где B заменено на B_k , $k = 1, 2, \dots$, то все $y_k \in \text{im } A^*$ и никаких дополнительных ограничений на τ_k накладывать не нужно.

2. Итерационный метод минимальных невязок. Рассмотрим теперь задачу о выборе итерационных параметров τ_k для схемы (3). Будем предполагать, что оператор B удовлетворяет условиям 1, а ограничения на выбор начального приближения y_0 и на параметры τ_k задаются условиями 2.

Выше было показано, что параметры τ_k следует выбирать из условия сходимости итерационного процесса (11) к нормальному решению \bar{u} уравнения (1).

Изучим итерационную схему (11). Сначала заметим, что оператор $D = A^*A$ является положительно определенным на $\text{im } A^*$. Действительно, пусть $u \in \text{im } A^*$ и $u \neq 0$. Так как $u \perp \ker A$, то $Au \neq 0$ и, следовательно, $(Du, u) = \|Au\|^2 > 0$. Оператор D порождает энергетическое пространство H_D , состоящее из элементов $\text{im } A^*$, скалярное произведение в котором определяется обычным образом: $(u, v)_D = (Du, v)$, $u \in \text{im } A^*$, $v \in \text{im } A^*$.

Поставим теперь задачу выбрать параметры τ_{k+1} в схеме (11) из условия минимума $\|\bar{z}_{k+1}\|_D$, где \bar{z}_{k+1} — погрешность: $\bar{z}_{k+1} = \bar{y}_{k+1} - \bar{u}$, $A\bar{u} = \bar{f}$ и \bar{u} — нормальное решение уравнения (1).

Для погрешности $\bar{z}_k \in \text{im } A^*$ из (11) получим следующее уравнение:

$$\bar{z}_{k+1} = (E - \tau_{k+1} B^{-1} A) \bar{z}_k. \quad (14)$$

Отсюда найдем

$$\|\bar{z}_{k+1}\|_D^2 = \|\bar{z}_k\|_D^2 - 2\tau_{k+1} (AB^{-1}A\bar{z}_k, A\bar{z}_k) + \tau_{k+1}^2 \|AB^{-1}A\bar{z}_k\|^2.$$

Заметим, что в силу леммы 3 $\|AB^{-1}A\bar{z}_k\| > 0$, $(A\bar{z}_k \in \text{im } A)$. Поэтому минимум $\|\bar{z}_{k+1}\|_D^2$ достигается при

$$\tau_{k+1} = \frac{(AB^{-1}A\bar{z}_k, A\bar{z}_k)}{(AB^{-1}A\bar{z}_k, AB^{-1}A\bar{z}_k)} \quad (15)$$

и равен

$$\|\bar{z}_{k+1}\|_D^2 = \rho_{k+1}^2 \|\bar{z}_k\|_D^2, \quad \rho_{k+1}^2 = 1 - \frac{(AB^{-1}A\bar{z}_k, A\bar{z}_k)^2}{\|AB^{-1}A\bar{z}_k\|^2 \|A\bar{z}_k\|^2}. \quad (16)$$

Формула (15) еще непригодна для вычислений, так как содержит неизвестные величины. Преобразуем ее. Используя разложение (9), получим

$$A\bar{z}_k = A\bar{y}_k - \bar{f} = Ay_k - \bar{f} = r_k + \tilde{f}, \quad (17)$$

где $r_k = Ay_k - \bar{f}$ — невязка. Так как $\tilde{f} \in \ker A^*$, то в силу леммы 3 $B^{-1}\tilde{f} \in \ker A$ и, следовательно, $AB^{-1}A\bar{z}_k = AB^{-1}r_k$. Подставляя это выражение, а также (17) в (15) и учитывая равенство $A^*\tilde{f} = 0$, получим

$$\tau_{k+1} = \frac{(AB^{-1}r_k, r_k)}{(AB^{-1}r_k, AB^{-1}r_k)} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}, \quad (18)$$

где поправка w_k находится из уравнения $Bw_k = r_k$.

Отметим, что (18) совпадает с формулой для итерационного параметра τ_{k+1} метода минимальных невязок, рассмотренного в главе VIII для уравнения с невырожденным оператором A .

Получим теперь оценку скорости сходимости для построенного метода. Умножим (14) слева на A , вычислим норму левой и правой частей и, учитывая, что $\|Az_k\| = \|\bar{z}_k\|_D$, получим следующую оценку:

$$\|\bar{z}_{k+1}\|_D \leq \|E - \tau_{k+1}AB^{-1}\|_{\text{im } A} \|\bar{z}_k\|_D \quad (19)$$

для любого τ_{k+1} . Из (16) и (19) получим для любого τ_{k+1}

$$\rho_{k+1} \leq \|E - \tau_{k+1}AB^{-1}\|_{\text{im } A}. \quad (20)$$

Если обозначить

$$\rho_0 = \min_{\tau} \|E - \tau AB^{-1}\|_{\text{im } A},$$

то из (16) и (20) будет следовать оценка для погрешности

$$\|\bar{z}_{k+1}\|_D \leq \rho_0 \|\bar{z}_k\|_D. \quad (21)$$

Здесь обозначение $\|S\|_{\text{im } A}$ используется для обозначения нормы оператора S в подпространстве $\text{im } A$.

Если $\rho_0 < 1$, то итерационный метод (11), (18) будет сходиться в H_D и из (21) получим, что

$$\|\bar{z}_k\|_D \leq \rho_0^k \|\bar{z}_0\|_D, \quad k = 0, 1, \dots \quad (22)$$

Осталось только подчинить выбор параметров τ_k условию (13), если $\tilde{f} \neq 0$. Для этого поступим следующим образом. Выполним по схеме (3) $(n-1)$ итерацию, выбирая $y_0 = A^* \varphi$, где $\varphi \in H$, используя для параметров τ_{k+1} , $k = 0, 1, \dots, n-2$ формулу (18). Выполним еще одну итерацию, выбирая

$$\tau_n = - \sum_{j=1}^{n-1} \tau_j.$$

Тогда условие (13) будет выполнено и, следовательно, $y_n = \bar{y}_n$. Оценим теперь норму погрешности $z_n = y_n - \bar{u}$ в H_D . Так как $y_n = \bar{y}_n$, то из (11) получим

$$y_n = \bar{y}_{n-1} - \tau_n B^{-1}(A\bar{y}_{n-1} - \tilde{f}) = \bar{y}_{n-1} - \tau_n B^{-1}A\bar{z}_{n-1}.$$

Отсюда

$$z_n = \bar{z}_{n-1} - \tau_n B^{-1}A\bar{z}_{n-1}$$

и после умножения на A будем иметь

$$Az_n = (E - \tau_n AB^{-1}) A\bar{z}_{n-1}.$$

Вычисляя норму, получаем оценку

$$\|z_n\|_D \leq \|E - \tau_n AB^{-1}\|_{\text{im } A} \|\bar{z}_{n-1}\|_D.$$

Подставляя сюда (22) и учитывая, что в силу выбора y_0 имеем равенство $\bar{y}_0 = y_0$, найдем

$$\|y_n - \bar{u}\|_D \leq \|E - \tau_n A B^{-1}\|_{\text{im } A} \rho_0^{n-1} \|y_0 - \bar{u}\|_D. \quad (23)$$

Рассмотрим частные случаи.

1) Пусть $B = E$, а оператор A самосопряжен в H . Пусть γ_1 и γ_2 — постоянные в неравенствах

$$\gamma_1(x, x) \leq (Ax, x) \leq \gamma_2(x, x), \quad \gamma_1 > 0, \quad Ax \neq 0. \quad (24)$$

В этом случае условия 1 выполнены.

Найдем ρ_0 и оценим норму оператора в (23). Так как оператор A самосопряжен в H , то, используя (24), получим

$$\|E - \tau A\|_{\text{im } A} = \sup_{Au \neq 0} \left| 1 - \tau \frac{(Au, u)}{(u, u)} \right| \leq \max_{\gamma_1 \leq t \leq \gamma_2} |1 - \tau t|.$$

С вычислением указанного максимума и выбором τ из условия его минимума мы встречались в главе VI при изучении метода простой итерации. Там было найдено, что

$$\min_{\tau} \max_{\gamma_1 \leq t \leq \gamma_2} |1 - \tau t| = \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Итак, ρ_0 найдено. Далее, при $B = E$ формула (15) для параметра τ_{k+1} записывается в виде

$$\tau_{k+1} = \frac{(Ax, x)}{(Ax, Ax)}, \quad x = A\bar{z}_k \in \text{im } A.$$

Так как $A = A^*$ и $\gamma_1 > 0$, то неравенства (24) эквивалентны следующим неравенствам (см. гл. V, § 1, п. 3):

$$\gamma_1(Ax, x) \leq (Ax, Ax) \leq \gamma_2(Ax, x), \quad Ax \neq 0.$$

Поэтому параметры τ_k для $k \leq n-1$ удовлетворяют неравенствам $1/\gamma_2 \leq \tau_k \leq 1/\gamma_1$. Отсюда найдем оценку

$$0 < -\tau_n = \sum_{j=1}^{n-1} \tau_j \leq \frac{n-1}{\gamma_1}. \quad (25)$$

Оценим норму оператора в (23). Учитывая (24) и (25), получим

$$\begin{aligned} \|E - \tau_n A\|_{\text{im } A} &\leq \max_{\gamma_1 \leq t \leq \gamma_2} |1 - \tau_n t| = \\ &= 1 - \tau_n \gamma_2 \leq 1 + (n-1) \frac{\gamma_2}{\gamma_1} = 1 + (n-1) \frac{1 + \rho_0}{1 - \rho_0}. \end{aligned}$$

Подставим эту оценку в (23) и найдем

$$\|y_n - \bar{u}\|_D \leq \rho_0^{n-1} \left[1 + (n-1) \frac{1 + \rho_0}{1 - \rho_0} \right] \|y_0 - \bar{u}\|_D. \quad (26)$$

2) Пусть $B = B^*$, $A = A^*$ и $AB = BA$. Пусть γ_1 и γ_2 —постоянны в неравенствах

$$\gamma_1(Bx, x) \leq (Ax, x) \leq \gamma_2(Bx, x), \quad \gamma_1 > 0, \quad Ax \neq 0. \quad (27)$$

В этом случае условия 1 выполнены, оператор AB^{-1} самосопряжен в H и можно показать, что для погрешности метода (3), (18) будет верна оценка (26).

3) Пусть операторы B^*A и AB^* самосопряжены в H , а γ_1 и γ_2 —постоянны в (27). В этом случае в силу леммы 2 условия 1 выполнены. Кроме того, оператор AB^{-1} будет самосопряжен в H . Можно показать, что и в этом случае оценка (26) имеет место.

3. Метод с чебышевскими параметрами. Рассмотрим теперь итерационные методы (3), параметры τ_k для которых выбираются с использованием априорной информации об операторах A и B .

Сначала приведем некоторые вспомогательные утверждения, необходимые нам для дальнейшего изложения.

Лемма 4. Пусть выполнены условия

$$A = A^* \geq 0, \quad B = B^* > 0, \quad AB = BA \quad (28)$$

и заданы постоянные γ_1 и γ_2 в неравенствах

$$\gamma_1(Bx, x) \leq (Ax, x) \leq \gamma_2(Bx, x), \quad \gamma_1 > 0, \quad Ax \neq 0. \quad (29)$$

Обозначим через D один из операторов A , B или $AB^{-1}A$ и определим на подпространстве $\text{im } A$ оператор C

$$C = D^{-1/2} (DB^{-1}A) D^{-1/2}.$$

Оператор C самосопряжен в $\text{im } A$ и удовлетворяет неравенствам

$$0 < \gamma_1(x, x) \leq (Cx, x) \leq \gamma_2(x, x), \quad x \in \text{im } A. \quad (30)$$

Действительно, из (28) и следствия к лемме 2 вытекает выполнение условий 1. Далее, оператор D самосопряжен в H и положительно определен на $\text{im } A$. Для примера докажем положительную определенность оператора $D = AB^{-1}A$. Пусть $u \in \text{im } A$ и $u \neq 0$. Так как $(Du, u) = (B^{-1}Au, Au)$, а оператор B^{-1} положительно определен в силу ограниченности и положительной определенности оператора B , то $(Du, u) \geq 0$, причем равенство нулю возможно лишь при выполнении условия $Au = 0$. Но это противоречит сделанным предположениям.

Оператор D отображает $\text{im } A$ на $\text{im } A$, поэтому существует $D^{-1/2}$, который также отображает это подпространство на себя. Следовательно, на $\text{im } A$ можно определить указанный в лемме оператор C . Переход от (29) к (30) доказывается так же, как это было сделано в гл. VI, § 2, п. 3. Лемма доказана.

Лемма 5. Пусть выполнены условия

$$B^*A = A^*B, \quad AB^* = BA^* \quad (31)$$

и заданы γ_1 и γ_2 в (29). Обозначим $C_1 = AB^{-1}$ и $C_2 = B^{-1}A$. Операторы C_1 и C_2 самосопряжены в H и удовлетворяют неравенствам

$$\gamma_1(x, x) \leq (C_1 x, x) \leq \gamma_2(x, x), \quad \gamma_1 > 0, \quad x \in \text{im } A, \quad (32)$$

$$\gamma_1(x, x) \leq (C_2 x, x) \leq \gamma_2(x, x), \quad \gamma_1 > 0, \quad x \in \text{im } A^*. \quad (33)$$

Самосопряженность операторов C_1 и C_2 непосредственно следует из (31). Докажем для примера (32). Рассмотрим задачу на собственные значения

$$AB^{-1}v - \lambda v = 0, \quad v \in H. \quad (34)$$

Так как оператор AB^{-1} самосопряжен в H , то существует ортонормированная система собственных функций задачи (34) $\{v_1, v_2, \dots, v_p, v_{p+1}, \dots, v_N\}$. Пусть v_1, \dots, v_p — функции, соответствующие собственному значению $\lambda = 0$, а v_{p+1}, \dots, v_N соответствуют ненулевым λ . Легко видеть, что $v_i \notin \ker A^*$, $1 \leq i \leq p$, $v_i \in \text{im } A$, $p+1 \leq i \leq N$, и в силу разложения H на подпространства (2), функции v_{p+1}, \dots, v_N образуют в $\text{im } A$ базис. Тогда для $x \in \text{im } A$ имеем

$$x = \sum_{k=p+1}^N a_k v_k, \quad C_1 x = \sum_{k=p+1}^N \lambda_k a_k v_k,$$

и в силу ортогональности собственных функций

$$(x, x) = \sum_{k=p+1}^N a_k^2, \quad (C_1 x, x) = \sum_{k=p+1}^N \lambda_k a_k^2.$$

Отсюда получим неравенства

$$\min_{p+1 \leq k \leq N} \lambda_k (x, x) \leq (C_1 x, x) \leq \max_{p+1 \leq k \leq N} \lambda_k (x, x).$$

Осталось найти минимальное и максимальное собственные значения, соответствующие собственным функциям задачи (34), принадлежащим $\text{im } A$. Запишем (34) в виде

$$A u_k - \lambda_k B u_k = 0, \quad p+1 \leq k \leq N, \quad (35)$$

где $u_k = B^{-1} v_k \in \text{im } A^*$ и, следовательно, $A u_k \neq 0$. Умножая (35) скалярно на u_k и используя (29), получим, что

$$\min_{p+1 \leq k \leq N} \lambda_k = \gamma_1, \quad \max_{p+1 \leq k \leq N} \lambda_k = \gamma_2.$$

Неравенства (32) доказаны. Справедливость (33) устанавливается аналогично. Лемма доказана.

Обратимся теперь к задаче выбора итерационных параметров для схемы (3). С учетом условий 2 запишем ее в следующем виде:

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad y_0 \in A^* \varphi, \quad \sum_{j=1}^n \tau_j = 0. \quad (36)$$

Если условия 1 будут выполнены, то параметры τ_k нужно выбрать из условия сходимости схемы (11) при указанном выше ограничении на сумму τ_j .

Рассмотрим уравнение (14) для погрешности схемы (11). Если выполнены условия леммы 4, то, полагая в (14) $\bar{z}_k = D^{-1/2}x_k$, где D — один из операторов леммы 4, получим следующее уравнение для эквивалентной погрешности:

$$x_{k+1} = (E - \tau_{k+1}C)x_k, \quad k = 0, 1, \dots, x_k \in \text{im } A. \quad (37)$$

Оператор C также определен в лемме 4.

Если выполнены условия леммы 5, то обозначая $B\bar{z}_k = x_k$ или $A\bar{z}_k = x_k$, получим уравнение

$$x_{k+1} = (E - \tau_{k+1}C_1)x_k, \quad k = 0, 1, \dots, x_k \in \text{im } A. \quad (38)$$

В этом случае $\|x_k\| = \|\bar{z}_k\|_D$, где $D = B^*B$ или A^*A . Если обозначить $\bar{z}_k = x_k$, то получим уравнение

$$x_{k+1} = (E - \tau_{k+1}C_2)x_k, \quad k = 0, 1, \dots, x_k \in \text{im } A^*, \quad (39)$$

и в этом случае $\|x_k\| = \|\bar{z}_k\|_D$, где $D = E$. Операторы C_1 и C_2 определены в лемме 5.

Итак, во всех рассматриваемых случаях мы получили уравнение вида

$$x_{k+1} = (E - \tau_{k+1}C)x_k, \quad k = 0, 1, \dots, x_k \in H_1 \quad (40)$$

в подпространстве H_1 , причем в силу лемм 4 и 5 оператор C самосопряжен в H_1 , действует в H_1 и удовлетворяет неравенствам

$$\gamma_1(x, x) \leq (Cx, x) \leq \gamma_2(x, x), \quad \gamma_1 > 0, x \in H_1, \quad (41)$$

где γ_1 и γ_2 взяты из неравенств (29).

Из (40) найдем

$$x_n = \prod_{j=1}^n (E - \tau_j C) x_0, \quad (42)$$

$$\|x_n\| \leq \|P_n(C)\| \|x_0\|, \quad P_n(C) = \prod_{j=1}^n (E - \tau_j C).$$

Учитывая самосопряженность C и неравенства (41), получим

$$\|P_n(C)\| \leq \max_{\gamma_1 \leq t \leq \gamma_2} |P_n(t)|.$$

Легко видеть, что

$$\sum_{j=1}^n \tau_j = -P'_n(0),$$

поэтому полином $P_n(t)$ нормирован двумя условиями

$$P_n(0) = 1, \quad P'_n(0) = 0. \quad (43)$$

Следовательно, мы приходим к задаче построения полинома степени n , удовлетворяющего условиям (43) и наименее уклоняющегося от нуля на отрезке $0 < \gamma_1 \leq t \leq \gamma_2$. Построение такого полинома полностью решает проблему выбора итерационных параметров τ_k для схемы (3).

Точное решение этой задачи нам неизвестно, и мы приведем иное решение проблемы. Как и в рассмотренном выше методе минимальных невязок, оставим произвол в выборе параметров $\tau_1, \tau_2, \dots, \tau_{n-1}$, а условие $\sum_{j=1}^n \tau_j = 0$ удовлетворим за счет выбора τ_n по формуле

$$\tau_n = - \sum_{j=1}^{n-1} \tau_j.$$

Из (42) получим следующую оценку:

$$\|x_n\| \leq \|P_{n-1}(C)\| \|E - \tau_n C\| \|x_0\|, \quad P_{n-1}(C) = \prod_{j=1}^{n-1} (E - \tau_j C). \quad (44)$$

Выберем теперь параметры $\tau_1, \tau_2, \dots, \tau_{n-1}$ из условия минимума нормы операторного полинома $P_{n-1}(C)$. Так как на $P_{n-1}(C)$ никаких дополнительных ограничений не налагается, то решение поставленной задачи имеет вид (см. гл. VI, § 2):

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_{n-1} = \left\{ \cos \frac{(2k-1)\pi}{2(n-1)}, \quad 1 \leq k \leq n-1 \right\}, \quad (45)$$

$k = 1, 2, \dots, n-1$, где приняты обозначения

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

При этом

$$P_{n-1}(t) = q_{n-1} T_{n-1} \left(\frac{1 - \tau_0 t}{\rho_0} \right), \quad \|P_{n-1}(C)\| \leq q_{n-1}, \quad (46)$$

где $T_{n-1}(x)$ — полином Чебышева 1-го рода степени $n-1$,

$$q_k = 2\rho_1^k / (1 + \rho_1^{2k}), \quad \rho_1 = (1 - \sqrt{\xi}) / (1 + \sqrt{\xi}).$$

Осталось найти явное выражение для τ_n . Из (46) найдем

$$\tau_n = - \sum_{j=1}^{n-1} \tau_j = P''_{n-1}(0) = - \frac{(n-1)\tau_0}{\rho_0} q_{n-1} U_{n-2} \left(\frac{1}{\rho_0} \right), \quad (47)$$

где $U_{n-2}(x)$ — полином Чебышева 2-го рода степени $n-2$. Здесь было использовано соотношение $T'_m(x) = mU_{m-1}(x)$. Вычислим $U_{n-2}(1/\rho_0)$. Так как $\rho_0 < 1$, то из явного вида для $U_{n-2}(x)$ (см. гл. I, § 4, п. 2):

$$U_{n-2}(x) = \frac{(x + \sqrt{x^2 - 1})^{n-1} - (x - \sqrt{x^2 - 1})^{-(n-1)}}{2 \sqrt{x^2 - 1}}, \quad |x| \geq 1,$$

получим в результате несложных выкладок

$$U_{n-2} \left(\frac{1}{\rho_0} \right) = \frac{1 - \rho_1^{2(n-1)}}{2\rho_1^{n-1}} \cdot \frac{\rho_0}{\sqrt{1 - \rho_0^2}}.$$

Подставим это выражение в (47) и найдем

$$\tau_n = - \frac{(n-1) \tau_0}{\sqrt{1 - \rho_0^2}} \frac{1 - \rho_1^{2(n-1)}}{1 + \rho_1^{2(n-1)}}. \quad (48)$$

Учитывая самосопряженность C и неравенства (41), формулу (48) и равенство $\tau_0 \gamma_2 = 1 + \rho_0$, получим

$$\begin{aligned} \|E - \tau_n C\| &\leq \max_{\gamma_1 \leq t \leq \gamma_2} |1 - \tau_n t| = 1 - \tau_n \gamma_2 = \\ &= 1 + (n-1) \sqrt{\frac{1 + \rho_0}{1 - \rho_0}} \frac{1 - \rho_1^{2(n-1)}}{1 + \rho_1^{2(n-1)}} \leq 1 + (n-1) \sqrt{\frac{1 + \rho_0}{1 - \rho_0}}. \end{aligned} \quad (49)$$

Подставляя (49) и (46) в (44), получим следующую оценку для нормы эквивалентной погрешности x_n :

$$\|x_n\| \leq \left(1 + (n-1) \sqrt{\frac{1 + \rho_0}{1 - \rho_0}} \right) q_{n-1} \|x_0\|$$

при условии, что параметры $\tau_1, \tau_2, \dots, \tau_n$ выбраны по формулам (45) и (48).

Теорема 5. Пусть итерационные параметры $\tau_k, k=1, \dots, n$, для схемы (3) выбраны по формулам (45) и (48) и $y_0 = A^* \varphi$. Тогда для погрешности верна оценка

$$\|y_n - \bar{u}\|_D \leq \left(1 + (n-1) \sqrt{\frac{1 + \rho_0}{1 - \rho_0}} \right) q_{n-1} \|y_0 - \bar{u}\|_P,$$

где \bar{u} — нормальное решение уравнения (1), а D определяется следующим образом: $D = A$, B или $AB^{-1}A$, если выполнены условия леммы 4; $D = B^*B$, A^*A или E , если выполнены условия леммы 5. Априорной информацией для метода с чебышевскими параметрами являются постоянные γ_1 и γ_2 из неравенств (29).

§ 4. Специальные методы

1. Разностная задача Неймана для уравнения Пуассона в прямоугольнике. На примере указанной задачи преиллюстрируем применение итерационной схемы с переменным оператором B_k к решению уравнения с вырожденным оператором A .

Пусть в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ требуется найти решение уравнения Пуассона

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -\varphi(x), \quad x \in G, \quad (1)$$

удовлетворяющее следующим краевым условиям:

$$\begin{aligned} \frac{\partial u}{\partial x_\alpha} &= -g_{-\alpha}(x_\beta), & x_\alpha &= 0, \quad \beta = 3 - \alpha, \\ -\frac{\partial u}{\partial x_\alpha} &= -g_{+\alpha}(x_\beta), & x_\alpha &= l_\alpha, \quad \alpha = 1, 2. \end{aligned} \quad (2)$$

На прямоугольной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ задаче (1), (2) соответствует следующая разностная задача:

$$\begin{aligned} \Lambda y &= -f(x), \quad x \in \bar{\omega}, \\ \Lambda &= \Lambda_1 + \Lambda_2, \quad f(x) = \varphi(x) + \frac{2}{h_1} \varphi_1(x) + \frac{2}{h_2} \varphi_2(x), \end{aligned} \quad (3)$$

где

$$\begin{aligned} \Lambda_\alpha y &= \begin{cases} \frac{2}{h_\alpha} y_{x_\alpha}, & x_\alpha = 0, \\ y_{\bar{x}_\alpha x_\alpha}, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ -\frac{2}{h_\alpha} y_{\bar{x}_\alpha}, & x_\alpha = l_\alpha, \end{cases} \\ \varphi_\alpha(x) &= \begin{cases} g_{-\alpha}(x_\beta), & x_\alpha = 0, \\ 0, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ g_{+\alpha}(x_\beta), & x_\alpha = l_\alpha. \end{cases} \end{aligned} \quad (4)$$

Пространство H состоит из сеточных функций, заданных на сетке $\bar{\omega}$, со скалярным произведением $(u, v) = \sum_{x \in \bar{\omega}} u(x)v(x)\bar{h}_1(x_1)\bar{h}_2(x_2)$,

где $\bar{h}_\alpha(x_\alpha)$ — средний шаг,

$$\bar{h}_\alpha(x_\alpha) = \begin{cases} h_\alpha, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ 0,5h_\alpha, & x_\alpha = 0, l_\alpha, \alpha = 1, 2. \end{cases}$$

Оператор A определим как сумму операторов A_1 и A_2 , где $A_\alpha = -\Lambda_\alpha, \alpha = 1, 2$. Тогда задачу (3) можно записать в виде операторного уравнения

$$Au = f \quad (5)$$

с указанным оператором A .

Отметим следующие свойства операторов A_1 и A_2 . Операторы A_1 и A_2 самосопряжены в H и перестановочны, т. е.

$$A_\alpha = A_\alpha^*, \quad \alpha = 1, 2, \quad A_1 A_2 = A_2 A_1.$$

Эти свойства позволяют, используя метод разделения переменных, решить задачу на собственные значения для оператора A : $Au = \lambda u$. Действуя по аналогии со случаем задачи Дирихле, под-

робно рассмотренным в п. 1 § 2 гл. IV, получим решение задачи в виде

$$\lambda_{k_1 k_2} = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad \lambda_{k_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi}{2N_\alpha}, \quad k_\alpha = 0, 1, \dots, N_\alpha,$$

$$\mu_{k_1 k_2}(i, j) = \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leq k_\alpha \leq N_\alpha, \quad \alpha = 1, 2,$$

$$\mu_{k_\alpha}^{(\alpha)}(m) = \begin{cases} \sqrt{\frac{2}{l_\alpha}} \cos \frac{k_\alpha \pi m}{N_\alpha}, & 1 \leq k_\alpha \leq N_\alpha - 1, \\ \sqrt{\frac{1}{l_\alpha}} \cos \frac{k_\alpha \pi m}{N_\alpha}, & k_\alpha = 0, N_\alpha, \quad \alpha = 1, 2. \end{cases}$$

При этом имеем

$$A_\alpha \mu_{k_1 k_2} = \lambda_{k_\alpha}^{(\alpha)} \mu_{k_1 k_2}, \quad A \mu_{k_1 k_2} = \lambda_{k_1 k_2} \mu_{k_1 k_2}.$$

Отсюда следует, что оператор A имеет простое собственное значение, равное нулю, которому соответствует собственная функция $\mu_{00}(i, j) \equiv 1/\sqrt{l_1 l_2}$. Эта функция образует базис в подпространстве $\ker A$. Функции $\mu_{k_1 k_2}(i, j)$ при $0 \leq k_\alpha \leq N_\alpha$ и $k_1 + k_2 \neq 0$ образуют базис в подпространстве $\text{im } A$.

Для решения уравнения (5) рассмотрим итерационную схему метода переменных направлений

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (6)$$

$$B_k = (\omega_k^{(1)} E + A_1)(\omega_k^{(2)} E + A_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}.$$

Чтобы не накладывать на $\{\tau_k\}$ и операторы $\{B_k\}$ дополнительных ограничений, связанных с выделением компоненты $\bar{y}_n \in \text{im } A$, потребуем, чтобы правая часть f была ортогональна $\ker A$. Если заданное f не удовлетворяет этому условию, то заменим его в (6) на $f_1 = f - (f, \mu_{00}) \mu_{00}$.

Заметим, что операторы B_k и A для любого k перестановочны. Поэтому в силу следствия к лемме 2 условия 1 будут выполнены (в них B нужно заменить на оператор B_k). Кроме того в силу леммы 3 оператор B_k^{-1} отображает $\text{im } A$ на $\text{im } A$.

Воспользуемся установленными фактами для изучения сходимости схемы (6). Так как $f \in \text{im } A$, то, предполагая, что $y_k \in \text{im } A$, получим из (6), что

$$y_{k+1} = y_k - \tau_{k+1} B_{k+1}^{-1} (A y_k - f) \in \text{im } A.$$

Поэтому, если выбрать $y_0 = 0$, то $y_0 \in \text{im } A$, следовательно, для любого $k \geq 0$ итерационные приближения $y_k \in \text{im } A$. Следовательно, схема (6) может рассматриваться только на подпространстве $\text{im } A$.

Исследуем сходимость схемы (6) в $\text{im } A$ по норме пространства H_D , в качестве D можно взять один из операторов E , A или A^2 . Каждый из этих операторов будет положительно опре-

делен в ім A . Способ изучения сходимости схемы (6) точно такой, какой был использован в главе XI при построении метода переменных направлений в невырожденном случае. Поэтому ограничимся лишь формулировкой задачи о наилучшем выборе параметров, опуская все необходимые для этого выкладки.

Наилучшие параметры $\omega_j^{(1)}$ и $\omega_j^{(2)}$ для схемы (6) должны быть выбраны из условия

$$\min_{\{\omega_j\}} \max_{x, y \in \Omega} \left| \prod_{j=1}^n \frac{\omega_j^{(2)} - x}{\omega_j^{(1)} + x} \frac{\omega_j^{(1)} - y}{\omega_j^{(2)} + y} \right| = \rho_n,$$

где $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$, $\Omega_1 = \{\lambda_1^{(1)} \leq x \leq \lambda_{N_1}^{(1)}, \lambda_1^{(2)} \leq y \leq \lambda_{N_2}^{(2)}\}$,

$$\Omega_2 = \{\lambda_1^{(1)} \leq x \leq \lambda_{N_1}^{(1)}, y = 0\} \quad \text{и} \quad \Omega_3 = \{x = 0, \lambda_1^{(2)} \leq y \leq \lambda_{N_2}^{(2)}\}.$$

При этом для погрешности $z_n = y_n - \bar{u}$, где \bar{u} — нормальное решение уравнения (5), будет верна оценка

$$\|z_n\|_D \leq \rho_n \|z_0\|_D.$$

Отметим, что для рассматриваемого примера условие ортогональности \bar{u} к $\ker A$ записывается в виде $(u, 1) = 0$. Любое другое решение уравнения (5) отличается от нормального решения \bar{u} на функцию, равную постоянной на сетке $\bar{\omega}$. Поэтому одно из возможных решений задачи (3) можно выделить, фиксируя значение этого решения в одном узле сетки $\bar{\omega}$.

Сформулированная выше задача для параметров отличается от рассмотренной нами в § 1 гл. XI, но может быть к ней сведена путем некоторых упрощений и ценою уменьшения возможной скорости сходимости итерационного метода. Обозначим

$$\delta = \min_{\alpha} \lambda_1^{(\alpha)}, \quad \Delta = \max_{\alpha} \lambda_{N_{\alpha}}^{(\alpha)}, \quad \eta = \frac{\delta}{\Delta},$$

$$\kappa_j = \frac{1}{\Delta} \omega_j^{(1)} = \frac{1}{\Delta} \omega_j^{(2)}, \quad j = 1, 2, \dots, n.$$

Используя эти обозначения и структуру области Ω , задачу выбора параметров можно сформулировать так: выбрать κ_j , $1 \leq j \leq n$, из условий

$$\min_{\{\kappa_j\}} \max_{\eta \leq u \leq 1} |r_n(u, \kappa)| = \bar{\rho}_n, \quad r_n(u, \kappa) = \prod_{j=1}^n \frac{\kappa_j - u}{\kappa_j + u}.$$

При этом, очевидно, $\bar{\rho}_n > \rho_n$.

Именно такая задача и была рассмотрена в § 1 гл. XI. Напомним, что там были получены формулы для κ_j и числа итераций $n = n_0(\varepsilon)$, которые гарантировали выполнение неравенства $\bar{\rho}_n^2 \leq \varepsilon$. Так как здесь нам нужно обеспечить оценку $\bar{\rho}_n \leq \varepsilon$, то в формулы для κ_j и $n_0(\varepsilon)$ гл. XI нужно вместо ε подставить ε^2 .

Тогда для погрешности метода (6) будет выполняться оценка $\|z_n\|_D \leq \varepsilon \|z_0\|_D$. Приведем вид оценки для числа итераций: $n \geq n_0(\varepsilon)$, $n_0(\varepsilon) = \frac{1}{\pi^2} \ln \frac{4}{\eta} \ln \frac{4}{\varepsilon^2}$. Для примера, если $l_1 = l_2 = l$ и $h_1 = h_2 = h$, то

$$\delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \Delta = \frac{4}{h^2}, \quad \eta = \sin^2 \frac{\pi h}{2}, \quad n_0(\varepsilon) = O(\ln h \ln \varepsilon).$$

Следовательно, для задачи Неймана метод переменных направлений, имея такую же по порядку оценку числа итераций, как и для случая задачи Дирихле, требует фактически в два раза больше итераций.

Заметим, что так как итерационные параметры κ_j удовлетворяют оценке (см. § 1, гл. XI) $\eta < \kappa_j < 1$, то параметры $\omega_j^{(1)}$ и $\omega_j^{(2)}$ принадлежат интервалу (δ, Δ) . Поэтому операторы $\omega_k^{(\alpha)} E + A_\alpha$ положительно определены в H , и для их обращения можно использовать алгоритм обычной трехточечной прогонки.

2. Прямой метод для задачи Неймана. Рассмотрим теперь прямой метод — комбинацию метода разделения переменных и метода редукции — для решения разностной задачи (3). Напомним, что такой метод был построен в п. 2 § 3 гл. IV для следующей краевой задачи: в области G задано уравнение (1), на сторонах $x_2 = 0$ и $x_2 = l_2$ заданы краевые условия (2), а на сторонах $x_1 = 0$ и $x_1 = l_1$ вместо условий второго рода (2) были заданы краевые условия третьего рода

$$\begin{aligned} \frac{\partial u}{\partial x_1} &= \kappa_{-1} u - g_{-1}(x_1), \quad x_1 = 0, \\ -\frac{\partial u}{\partial x_2} &= \kappa_{+1} u - g_{+1}(x_1), \quad x_1 = l_1, \end{aligned}$$

причем κ_{-1} и κ_{+1} неотрицательные константы, одновременно не равные нулю. Соответствующая разностная задача отличается от задачи (3) лишь определением оператора Λ_1 . Там мы имели дело с оператором Λ_1 :

$$\Lambda_1 y = \begin{cases} \frac{2}{h_1} (y_{x_1} - \kappa_{-1} y), & x_1 = 0, \\ y_{\bar{x}_1 x_1}, & h_1 \leq x_1 \leq l_1 - h_1, \\ \frac{2}{h_1} (-y_{\bar{x}_1} - \kappa_{+1} y), & x_1 = l_1. \end{cases}$$

Требование необращения одновременно в нуль κ_{-1} и κ_{+1} гарантировало разрешимость разностной задачи и единственность решения. В алгоритме же метода это требование использовалось лишь при решении трехточечных краевых задач для коэффициентов Фурье искомого решения. Поэтому для решения задачи (3) формально можно воспользоваться алгоритмом, приведенным в п. 2 § 3 гл. IV, полагая в нем $\kappa_{-1} = \kappa_{+1} = 0$, и

отдельно обсудить вопрос о решении возникающих трехточечных краевых задач.

Вернемся к задаче (3). Будем считать, что $f \perp \ker A$, т. е. есть $\langle f, 1 \rangle = 0$. Тогда задача разрешима, нормальное решение \bar{u} ортогонально $\ker A$, а одно из возможных решений можно выделить, фиксируя его значение в одном узле сетки $\bar{\omega}$. В рассматриваемом алгоритме выделение одного из возможных решений удобно осуществить, фиксируя не само решение в узле, а один из коэффициентов Фурье. Пусть $y(i, j)$ — решение задачи (3). Тогда нормальное решение \bar{u} можно найти по формуле

$$\bar{u} = y - (y, \mu_{00}) \mu_{00}, \quad \mu_{00}(i, j) = 1/\sqrt{l_1 l_2}. \quad (7)$$

Приведем теперь алгоритм прямого метода решения задачи Неймана (3) для уравнения Пуассона в прямоугольнике.

1) Для $0 \leq i \leq N_1$ вычисляются значения функции

$$\begin{aligned} \varphi(i, j) = \\ = \begin{cases} 2[f(i, 0) + f(i, 1)] - h_2^2 \Lambda_1 f(i, 0), & j = 0, \\ f(i, 2j-1) + f(i, 2j+1) + 2f(i, 2j) - h_2^2 \Lambda_1 f(i, 2j), & 1 \leq j \leq M_2 - 1, \\ 2[f(i, N_2) + f(i, N_2-1)] - h_2^2 \Lambda_1 f(i, N_2), & j = M_2, \\ \kappa_{-1} = \kappa_{+1} = 0. \end{cases} \end{aligned}$$

2) По алгоритму быстрого преобразования Фурье вычисляются коэффициенты Фурье функции $\varphi(i, j)$:

$$z_{k_2}(i) = \sum_{j=0}^{M_2} \rho_j \varphi(i, j) \cos \frac{k_2 \pi j}{M_2}, \quad 0 \leq k_2 \leq M_2, \quad 0 \leq i \leq N_1.$$

3) Решаются трехточечные краевые задачи

$$\begin{aligned} 4 \sin^2 \frac{k_2 \pi}{2N_2} w_{k_2}(i) - h_2^2 \Lambda_1 w_{k_2}(i) = h_2^2 z_{k_2}(i), & \quad 0 \leq i \leq N_1, \\ 4 \cos^2 \frac{k_2 \pi}{2N_2} y_{k_2}(i) - h_2^2 \Lambda_1 y_{k_2}(i) = w_{k_2}(i), & \quad 0 \leq i \leq N_1 \end{aligned} \quad (8)$$

для $0 \leq k_2 \leq M_2$, в результате находятся коэффициенты Фурье $y_{k_2}(i)$ функции $y(i, j)$.

4) По алгоритму быстрого преобразования Фурье находится решение задачи на четных строках сетки $\bar{\omega}$

$$y(i, 2j) = \sum_{k_2=0}^{M_2} \rho_{k_2} y_{k_2}(i) \cos \frac{k_2 \pi j}{M_2}, \quad 0 \leq j \leq M_2, \quad 0 \leq i \leq N_1,$$

и решаются трехточечные краевые задачи

$$2y(i, 2j-1) - h_2^2 \Lambda_1 y(i, 2j-1) = \\ = h_2^2 f(i, 2j-1) + u(i, 2j-2) + u(i, 2j), \\ 0 \leq i \leq N_1, \quad 1 \leq j \leq M_2,$$

для нахождения решения на нечетных строках.

Здесь использованы обозначения

$$M_2 = 0.5N_2, \quad \rho_j = \begin{cases} 1, & 1 \leq j \leq M_2 - 1, \\ 0.5, & j = 0, M_2, \end{cases}$$

оператор Λ_1 определен в (4), и предполагается, что N_2 есть степень 2. Число действий описанного метода будет равно $O(N^2 \log_2 N)$ для $N_1 = N_2 = N$.

Выделение одного решения из совокупности решений задачи (3) в приведенном алгоритме осуществляется следующим образом. Из всех трехточечных краевых задач, которые требуется решить, лишь одна задача (8) при $k_2 = 0$ имеет неединственное решение. Выделение здесь одного из решений обеспечит решение поставленной задачи. Разностная задача (8) при $k_2 = 0$ имеет вид

$$\Lambda_1 w_0(i) = -z_0(i), \quad 0 \leq i \leq N_1,$$

или

$$(w_0)_{\bar{x}_1, x_1} = -z_0(i), \quad 1 \leq i \leq N_1 - 1, \\ \frac{2}{h_1} (w_0)_{x_1} = -z_0(0), \quad i = 0, \\ -\frac{2}{h_1} (w_0)_{\bar{x}_1} = -z_0(N_1), \quad i = N_1. \quad (9)$$

Несложно показать, используя ортогональность $f(i, j)$ к $\mu_{00}(i, j)$, что сеточная функция $z_0(i)$ ортогональна функции $\mu_0(i) = 1/\sqrt{I_1}$ в смысле скалярного произведения

$$(u, v)_1 = \sum_{x_1=0}^{l_1} u(x_1) \bar{v}(x_1) \bar{h}_1(x_1).$$

А так как $\mu_0(i)$ является базисом в подпространстве $\ker \Lambda_1$, то задача (9) имеет решения. Выделим одно из решений, фиксируя значение $w_0(i)$ при каком-либо i , $0 \leq i \leq N_1$. Положим, например, $w_0(N_1) = 0$ и исключим из (9) краевое условие при $i = N_1$. Полученная в результате такой замены разностная задача легко решается методом прогонки.

После того как одно из решений $y(i, j)$ задачи (3) будет найдено по описанному выше алгоритму, нормальное решение \bar{u} , если в нем есть необходимость, определяется по формуле (7).

В заключение отметим, что аналогичная процедура выделения одного из возможных решений может быть использована и в методе полной редукции, когда он используется для решения разностной задачи Неймана.

3. Итерационные схемы с вырожденным оператором B . Наличие прямых методов обращения оператора Лапласа в прямоугольнике в случае краевых условий Неймана позволяет использовать такие операторы в качестве оператора B в неявных итерационных схемах решения вырожденных уравнений. Так как в этом случае оператор B вырожден, то необходимо заново изучить проблему выбора итерационных параметров.

Рассмотрим итерационные методы решения уравнения (5) при следующих предположениях: 1) оператор A самосопряжен и вырожден; 2) известно ядро оператора A , т. е. задан базис в $\ker A$; 3) правая часть f уравнения (5) принадлежит $\text{im } A$, т. е. $f = \bar{f} \in \text{im } A$. Этому условию легко удовлетворить, так как известен базис в $\ker A$. При этом нормальное решение \bar{u} уравнения (5) является классическим, оно удовлетворяет соотношению

$$A\bar{u} = f. \quad (10)$$

Отметим, что в силу самосопряженности оператора A имеет место следующее ортогональное разложение пространства H :

$$H = \ker A \bigoplus \text{im } A. \quad (11)$$

Для решения уравнения (5) рассмотрим неявную двухслойную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (12)$$

с вырожденным оператором B . Ставится задача найти при помощи (12) приближение к одному из решений уравнения (5).

Сформулируем теперь дополнительные предположения относительно операторов A и B . Пусть B —самосопряженный в H оператор и $\ker B = \ker A$. Кроме того, пусть для любого $x \in \text{im } A$ выполняются неравенства

$$\begin{aligned} \gamma_1(Bx, x) &\leq (Ax, x) \leq \gamma_2(Bx, x), \quad \gamma_1 > 0, \quad Ax \neq 0, \\ (Bx, x) &> 0. \end{aligned} \quad (13)$$

Заметим, что из условий $B = B^*$, $\ker B = \ker A$ и (11) следует совпадение $\text{im } A$ и $\text{im } B$.

Изучим схему (12). В соответствии с (11) представим y_k в виде суммы

$$y_k = \bar{y}_k + \tilde{y}_k, \quad \bar{y}_k \in \text{im } A, \quad \tilde{y}_k \in \ker A.$$

Из (12) получим следующее уравнение для y_{k+1} :

$$By_{k+1} = \varphi_k, \quad (14)$$

где $\varphi_k = By_k - \tau_{k+1}(Ay_k - f)$.

Так как $f \in \text{im } A$ и $\text{im } B = \text{im } A$, то $\varphi_k \in \text{im } A$ при любом y_k . Следовательно, $\varphi_k \perp \ker B$, и уравнение (14) имеет совокупность решений в обычном смысле, а нормальное его решение \bar{y}_{k+1} удовлетворяет уравнению

$$B\bar{y}_{k+1} = \varphi_k. \quad (15)$$

Заметим, что в силу равенства $B\bar{y}_k = A\bar{y}_k = 0$ имеем

$$\varphi_k = B\bar{y}_k - \tau_{k+1}(A\bar{y}_k - f). \quad (16)$$

Поэтому компонента \bar{y}_k итерационного приближения $y_k, \bar{y}_k \in \ker A$, не оказывает никакого влияния на \bar{y}_{k+1} . Отсюда следует вывод—при решении уравнения (14) достаточно найти какое-либо его решение и лишь после окончания процесса итераций вычислить проекцию y_n на $\text{im } A$, т. е. найти \bar{y}_n .

Рассмотрим теперь вопрос о выборе итерационного параметра τ_k . В силу сказанного выше его следует выбрать так, чтобы последовательность \bar{y}_k стремилась к нормальному решению \bar{u} уравнения (5). Из (10), (15) и (16) получим следующую задачу для погрешности $z_k = y_k - \bar{u}$:

$$B\bar{z}_{k+1} = (B - \tau_{k+1}A)\bar{z}_k, \quad k = 0, 1, \dots, \quad (17)$$

где $\bar{z}_k \in \text{im } A$ для любого $k \geq 0$.

Так как в подпространстве $\text{im } A$ операторы A и B в силу (13) положительно определены, то схему (17) можно обычным образом исследовать на сходимость по норме энергетического пространства H_D , где $D = A$, B или $AB^{-1}A$. Так как в этом случае оператор $DB^{-1}A$ самосопряжен, то параметры τ_k можно выбрать по формулам чебышевского итерационного метода (см. § 2 гл. VI)

$$\begin{aligned} \tau_k &= \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ \cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \quad 1 \leq k \leq n, \\ \tau_0 &= \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad (18) \\ n &\geq n_0(\varepsilon) = \ln(0.5\varepsilon)/\ln \rho_1, \end{aligned}$$

используя γ_1 и γ_2 из неравенств (13). Тогда для погрешности \bar{z}_n после n итераций будет верна оценка

$$\|\bar{z}_n\|_D \leq \varepsilon \|\bar{z}_0\|_D.$$

Смысл рассмотрения итерационных методов с вырожденным оператором B заключается в следующем. Если оператор B таков,

что нахождение решения уравнения (14) осуществляется значительно проще, чем исходного уравнения (5), а отношение ξ не слишком мало, то такой способ приближенного решения уравнения (5) может оказаться целесообразным.

Приведем пример одной разностной задачи, на которой проиллюстрируем предложенный метод. Пусть на прямоугольной сетке

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\},$$

введенной в прямоугольнике \bar{G} , требуется найти решение задачи Неймана для эллиптического уравнения с переменными коэффициентами

$$\begin{aligned} \Lambda y &= -f(x), \quad x \in \bar{\omega}, \\ \Lambda = \Lambda_1 + \Lambda_2, \quad f(x) &= \varphi(x) + \frac{2}{h_1} \varphi_1(x) + \frac{2}{h_2} \varphi_2(x), \end{aligned} \quad (19)$$

где

$$\begin{aligned} \Lambda_\alpha y &= \begin{cases} \frac{2}{h_\alpha} a_\alpha^{+1} y_{x_\alpha}, & x_\alpha = 0, \\ (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ -\frac{2}{h_\alpha} a_\alpha y_{\bar{x}_\alpha}, & x_\alpha = l_\alpha, \end{cases} \\ \varphi_\alpha(x_\alpha) &= \begin{cases} g_{-\alpha}(x_\beta), & x_\alpha = 0, \\ 0, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ g_{+\alpha}(x_\beta), & x_\alpha = l_\alpha, \end{cases} \end{aligned}$$

$$a_1^{+1}(x) = a_1(x_1 + h_1, x_2), \quad a_2^{+1}(x) = a_2(x_1, x_2 + h_2).$$

Предполагается, что коэффициенты $a_1(x)$ и $a_2(x)$ удовлетворяют условиям

$$0 < c_1 \leq a_\alpha(x) \leq c_2, \quad \alpha = 1, 2, x \in \bar{\omega}. \quad (20)$$

Схема (19) есть разностный аналог задачи Неймана для эллиптического уравнения

$$\begin{aligned} \frac{\partial}{\partial x_1} \left(k_1(x) \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k_2(x) \frac{\partial u}{\partial x_2} \right) &= -\varphi(x), \quad x \in G, \\ k_\alpha \frac{\partial u}{\partial x_\alpha} &= -g_{-\alpha}(x_\beta), \quad x_\alpha = 0, \quad \beta = 3 - \alpha, \\ -k_\alpha \frac{\partial u}{\partial x_\alpha} &= -g_{+\alpha}(x_\beta), \quad x_\alpha = l_\alpha, \quad \alpha = 1, 2. \end{aligned}$$

Пространство H определено в п. 1. Вводя оператор $A = -\Lambda$, запишем разностную задачу (19) в виде уравнения (5). Легко проверить, что $A = A^*$, а разностные формулы Грина дают

$$(Ay, y) = \sum_{\alpha=1}^2 \left(a_\alpha y_{\bar{x}_\alpha}^2, 1 \right)_\alpha, \quad (21)$$

где использованы следующие обозначения:

$$(u, v)_\alpha = \sum_{x_\beta=0}^{l_\beta} \sum_{x_\alpha=h_\alpha}^{l_\alpha} u(x) v(x) \tilde{h}_\beta(x_\beta) h_\alpha, \quad \beta=3-\alpha, \alpha=1, 2.$$

Нетрудно показать, что оператор \bar{A} вырожден и для любых коэффициентов $a_\alpha(x)$, удовлетворяющих (20), ядро оператора \bar{A} составляют сеточные функции, являющиеся постоянными на $\bar{\omega}$. Поэтому в качестве базиса в $\ker A$ может быть взята известная нам функция $\mu_{00}(i, j) = 1/\sqrt{l_1 l_2}$.

Определим теперь оператор $B = -\dot{\Lambda}$, где $\dot{\Lambda} = \dot{\Lambda}_1 + \dot{\Lambda}_2$,

$$\dot{\Lambda}_\alpha y = \begin{cases} \frac{2}{h_\alpha} y_{x_\alpha}, & x_\alpha = 0, \\ y_{\bar{x}_\alpha x_\alpha}, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ -\frac{2}{h_\alpha} y_{\bar{x}_\alpha}, & x_\alpha = l_\alpha, \quad \alpha = 1, 2. \end{cases}$$

Оператор B самосопряжен в H , а в п. 1 было отмечено, что базис в $\ker B$ образует именно функция $\mu_{00}(i, j)$. Следовательно, $\ker A$ известно и $\ker A = \ker B$. Если к тому же взять проекцию f на $\text{im } A$ и заменить ею в случае необходимости правую часть в схеме (19), то все требования, предъявляемые к схеме (12) и уравнению (5), будут выполнены.

Для применения итерационного метода (12), (18) осталось указать γ_1 и γ_2 в неравенствах (13). Так как

$$(By, y) = \sum_{\alpha=1}^2 (y_{\bar{x}_\alpha}^2, 1)_\alpha, \quad (22)$$

а подпространства $\text{im } A$ и $\text{im } B$ совпадают и состоят из сеточных функций, не являющихся постоянными на сетке $\bar{\omega}$, то из (20) — (22) получим, что $\gamma_1 = c_1$, $\gamma_2 = c_2$. Необходимая априорная информация найдена.

Из оценки (18) для числа итераций видно, что оно не зависит от числа неизвестных в задаче, а определяется лишь отношением c_1/c_2 . Далее, в силу выбора оператора B уравнение (14) для y_{k+1} есть разностная задача Неймана для уравнения Пуассона в прямоугольнике. Решение ее можно найти прямым методом, изложенным в п. 2 с затратой $O(N^2 \log_2 N)$ арифметических действий. Тогда общее число действий для предлагаемого метода, которое следует затратить для получения решения (19) с точностью ϵ , будет равно $Q(\epsilon) = O(N^2 \log_2 N |\ln \epsilon|)$.

ГЛАВА XIII

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

В главе изучаются итерационные методы решения нелинейных разностных схем. В § 1 излагается общая теория итерационных методов для абстрактного нелинейного операторного уравнения в гильбертовом пространстве при различных предположениях относительно оператора. В § 2 рассматривается применение общей теории к решению разностных аналогов краевых задач для квазилинейных эллиптических уравнений второго порядка.

§ 1. Итерационные методы. Общая теория

1. Метод простой итерации для уравнений с монотонным оператором. В предыдущих главах были изучены итерационные методы решения линейного операторного уравнения первого рода

$$Au = f, \quad (1)$$

заданного в гильбертовом пространстве H . Большинство построенных методов были линейными и сходились со скоростью геометрической прогрессии.

Перейдем теперь к изучению методов решения уравнения (1) в случае, когда A — произвольный нелинейный оператор, действующий в H . Эта глава посвящена конструированию итерационных методов для решения нелинейных уравнений (1). Построение таких методов основано, как правило, на использовании в неявных итерационных схемах линейного оператора B , близкого в некотором смысле к нелинейному оператору A . Ниже при различных предположениях относительно операторов A , B и D будут доказаны общие теоремы сходимости в H_D решения неявной двухслойной итерационной схемы

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H. \quad (2)$$

Изучение итерационной схемы (2) начнем со случая монотонного оператора A . Напомним, что оператор A , заданный в вещественном гильбертовом пространстве, называется *монотонным*, если

$$(Au - Av, u - v) \geqslant 0, \quad u, v \in H,$$

и сильно монотонным, если существует такое число $\delta > 0$, что для любых $u, v \in H$

$$(Au - Av, u - v) \geq \delta \|u - v\|^2. \quad (3)$$

Из теоремы 11 главы V следует существование и единственность в шаре $\|u\| \leq \frac{1}{\delta} \|A0 - f\|$ решения уравнения (1) с сильно монотонным оператором, являющимся непрерывным в конечномерном пространстве H .

Будем предполагать, что B — линейный ограниченный и положительно определенный в H оператор, а D — самосопряженный положительно определенный в H оператор. Пусть, кроме того, заданы постоянные γ_1 и γ_2 в неравенствах

$$(DB^{-1}(Au - Av), B^{-1}(Au - Av)) \leq \gamma_2 (DB^{-1}(Au - Av), u - v), \quad (4)$$

$$(DB^{-1}(Au - Av), u - v) \geq \gamma_1 (D(u - v), u - v), \quad (5)$$

причем $\gamma_1 > 0$.

Лемма 1. Пусть выполнены условия (4), (5). Тогда уравнение (1) однозначно разрешимо при любой правой части.

В самом деле, запишем уравнение (1) в эквивалентном виде

$$u = Su, \quad (6)$$

где нелинейный оператор S определяется следующим образом:

$$Su = u - \tau B^{-1} Au + \tau B^{-1} f, \quad \tau > 0.$$

Покажем, что в H_D оператор S при $\tau < 2/\gamma_2$ является равномерно сжимающим, т. е. для любых $u, v \in H$ справедлива оценка

$$\|Su - Sv\|_D \leq p(\tau) \|u - v\|_D, \quad p(\tau) < 1, \quad (7)$$

причем $p(\tau)$ не зависит от u и v . Тогда утверждение леммы будет следовать из теоремы 8 главы V о сжатых отображениях.

Имеем

$$\begin{aligned} \|Su - Sv\|_D^2 &= (D(Su - Sv), (Su - Sv)) = \|u - v\|_D^2 - \\ &- 2\tau(DB^{-1}(Au - Av), u - v) + \tau^2(DB^{-1}(Au - Av), B^{-1}(Au - Av)). \end{aligned}$$

Из (4), (5) найдем при $\tau < 2/\gamma_2$

$$\begin{aligned} \|Su - Sv\|_D^2 &\leq \|u - v\|_D^2 - \tau(2 - \tau\gamma_2)(DB^{-1}(Au - Av), u - v) \leq \\ &\leq p^2(\tau) \|u - v\|_D^2, \end{aligned}$$

где

$$p^2(\tau) = 1 - \tau(2 - \tau\gamma_2)\gamma_1. \quad (8)$$

Так как $\tau < 2/\gamma_2$, то $p(\tau) < 1$. Лемма доказана.

Исследуем теперь сходимость итерационной схемы (2) в предположении, что условия (4), (5) выполнены.

Из (2) найдем

$$y_{k+1} = y_k - \tau B^{-1} A y_k + \tau B^{-1} f = S y_k, \quad (9)$$

где нелинейный оператор S определен выше. Так как решение u уравнения (1) удовлетворяет соотношению (6), то из (6)–(9) получим

$$\begin{aligned} y_{k+1} - u &= Sy_k - Su, \quad k = 0, 1, \dots, \\ \|y_{k+1} - u\|_D &\leq \|Sy_k - Su\|_D \leq \rho^2(\tau) \|y_k - u\|_D^2, \end{aligned}$$

где $\rho^2(\tau)$ определено в (8). Нетрудно видеть, что наилучшая оценка скорости сходимости достигается, когда $\rho(\tau)$ минимально, т. е. при $\tau = \tau_0 = 1/\gamma_2$. При этом $\rho_0 = \rho(\tau_0) = \sqrt{1-\xi}$, $\xi = \gamma_1/\gamma_2$. Итак, доказана

Теорема 1. Пусть выполнены условия (4), (5). Итерационный метод (2) с $\tau = \tau_0 = 1/\gamma_2$ сходится в H_D , и для погрешности имеет место оценка

$$\|y_n - u\|_D \leq \rho_0^n \|y_0 - u\|_D, \quad \rho_0 = \sqrt{1-\xi}, \quad \xi = \gamma_1/\gamma_2,$$

где u – решение уравнения (1). Для числа итераций верна оценка

$$n \geq n_0(\varepsilon) = \ln \varepsilon / \ln \rho_0.$$

Заметим, что для линейного оператора A условия (4), (5) можно записать в виде

$$(DB^{-1}Ay, B^{-1}Ay) \leq \gamma_2(DB^{-1}Ay, y), \quad (DB^{-1}Ay, y) \geq \gamma_1(Dy, y).$$

Следовательно, в этом случае они совпадают с условиями, налагаемыми на операторы A , B и D , когда оператор $DB^{-1}A$ является несамосопряженным в H . При этом построенный здесь метод переходит в первый вариант метода простой итерации для несамосопряженного случая (см. п. 2 § 4 гл. VI).

Отметим, что вместо (4) можно потребовать выполнения условия

$$\|B^{-1}(Au - Av)\|_D \leq \bar{\gamma}_2 \|u - v\|_D, \quad (10)$$

которое при $D = B = E$ является условием Липшица для оператора A . Из (10) и (5) следует неравенство (4) с $\gamma_2 = \bar{\gamma}_2^2/\gamma_1$.

Если оператор B самосопряжен и положительно определен в H , то в качестве оператора D можно взять B . Тогда условия (4), (5) будут иметь вид

$$\begin{aligned} (B^{-1}(Au - Av), Au - Av) &\leq \gamma_2 ((Au - Av), u - v), \\ (Au - Av, u - v) &\geq \gamma_1 (B(u - v), u - v), \quad \gamma_1 > 0. \end{aligned}$$

Если B несамосопряжен и невырожден, то при $D = B^*B$ условия (4), (5) имеют вид

$$\begin{aligned} (Au - Av, Au - Av) &\leq \gamma_2 (Au - Av, B(u - v)), \\ (Au - Av, B(u - v)) &\geq \gamma_1 (B(u - v), B(u - v)), \quad \gamma_1 > 0. \end{aligned}$$

При $D = B = E$ условие (5) означает, что оператор A должен быть сильно монотонным в H .

2. Итерационные методы для случая дифференцируемого оператора. Улучшение оценки скорости сходимости метода простой итерации для уравнения (1) может быть достигнуто за счет более сильных ограничений на оператор A . Именно, будем считать, что оператор A имеет производную Гато. Напомним, что линейный оператор $A'(u)$ называется *производной Гато* оператора A в точке $u \in H$, если для любого $v \in H$ справедливо соотношение

$$\lim_{t \rightarrow 0} \left\| \frac{A(u+tv) - A(u)}{t} - A'(u)v \right\| = 0.$$

Если оператор A имеет производную Гато в каждой точке пространства H , то справедливо неравенство Лагранжа

$$\|Au - Av\| \leq \sup_{0 < t \leq 1} \|A'(u + t(v-u))\| \|u - v\|, \quad u, v \in H,$$

и для любых u, v и $w \in H$ существует $t \in [0, 1]$ такое, что

$$(Au - Av, w) = (A'(u + t(v-u))z, w), \quad z = u - v. \quad (11)$$

Вернемся к исследованию сходимости итерационной схемы (2). Имеет место

Теорема 2. Пусть оператор A имеет в шаре $\Omega(r) = \{v : \|u - v\|_D \leq r\}$ производную Гато $A'(v)$, которая при любом $v \in \Omega(r)$ удовлетворяет неравенствам

$$\begin{aligned} (DB^{-1}A'(v)y, B^{-1}A'(v)y) &\leq \gamma_2(DB^{-1}A'(v)y, y), \\ (DB^{-1}A'(v)y, y) &\geq \gamma_1(Dy, y), \quad \gamma_1 > 0 \end{aligned} \quad (12)$$

для любого $y \in H$. Итерационный метод (2) с $\tau = 1/\gamma_2$ и $y_0 \in \Omega(r)$ сходится в H_D , и для погрешности верна оценка

$$\|y_n - u\|_D \leq \rho^n \|y_0 - u\|_D, \quad (13)$$

где u — решение уравнения (1), а $\rho = \sqrt{1-\xi}$, $\xi = \gamma_1/\gamma_2$. Если оператор $DB^{-1}A'(v)$ самосопряжен в H при $v \in \Omega(r)$ и выполнены неравенства

$$\gamma_1(Dy, y) \leq (DB^{-1}A'(v)y, y) \leq \gamma_2(Dy, y), \quad \gamma_1 > 0 \quad (14)$$

для любого $v \in \Omega(r)$ и $y \in H$, то при $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$ для итерационного процесса (2) верна оценка (13) с $\rho = \rho_0 = (1-\xi)/(1+\xi)$.

Действительно, из уравнения для погрешности

$$y_{k+1} - u = Sy_k - Su, \quad Sv = v - \tau B^{-1}Av + \tau B^{-1}f$$

и неравенства Лагранжа получим

$$\|y_{k+1} - u\|_D = \|Sy_k - Su\|_D \leq \sup_{0 < t \leq 1} \|S'(v_k)\|_D \|y_k - u\|_D, \quad (15)$$

где $v_k = y_k + t(u - y_k) \in \Omega(r)$, если $y_k \in \Omega(r)$. Так как $S'(v_k) = E - \tau B^{-1}A'(v_k)$, то задача сводится к оценке в H_D нормы ли-

нейного оператора $E - \tau B^{-1} A'(v_k)$. Из определения нормы оператора имеем

$$\begin{aligned}\|S'(v_k)\|_D^2 &= \sup_{y \neq 0} \frac{(S'(v_k)y, S'(v_k)y)_D}{(y, y)_D} = \\ &= \sup_{y \neq 0} \frac{(DS'(v_k)y, S'(v_k)y)}{(Dy, y)} = \sup_{z \neq 0} \frac{((E - \tau C(v_k))z, (E - \tau C(v_k))z)}{(z, z)} = \\ &= \|E - \tau C(v_k)\|^2,\end{aligned}$$

где $C(v_k) = D^{-1/2} (DB^{-1} A'(v_k)) D^{-1/2}$ и была сделана замена $y = D^{-1/2} z$.

Подставляя найденное соотношение в (15), получим

$$\|y_{k+1} - u\|_D \leq \sup_{0 \leq t \leq 1} \|E - \tau C(v_k)\| \|y_k - u\|_D.$$

Из (12) найдем, что оператор $C(v_k)$ при любом $v_k \in \Omega(r)$ удовлетворяет неравенствам

$$\begin{aligned}(C(v_k)y, C(v_k)y) &\leq \gamma_2 (C(v_k)y, y), \\ (C(v_k)y, y) &\geq \gamma_1 (y, y).\end{aligned}$$

Напомним, что требуемая оценка для нормы линейного оператора $E - \tau C(v_k)$ при указанных предположениях была получена в п. 2 § 4 гл. VI. Именно, при $\tau = 1/\gamma_3$ имеем $\|E - \tau C(v_k)\| \leq \rho$, где $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$. Первое утверждение теоремы доказано. Аналогично доказывается и второе. В этом случае оператор $C(v_k)$ самосопряжен в H , и оценка для нормы оператора $E - \tau C(v_k)$ была ранее получена в п. 2 § 3 гл. VI. Теорема 2 доказана.

В главе VI, помимо использованной здесь оценки для нормы оператора $E - \tau C(v_k)$ в несамосопряженном случае, была получена другая оценка в предположении, что заданы три числа γ_1 , γ_2 и γ_3 в неравенствах

$$\bar{\gamma}_1 E \leq C(v_k) \leq \bar{\gamma}_2 E, \quad \|C(v_k)\| \leq \bar{\gamma}_3, \quad \bar{\gamma}_1 > 0,$$

где $C_1 = 0,5(C - C^*)$ — кососимметрическая часть оператора C . В этом случае при $\tau = \tau_0(1 - \kappa\rho)$ верна оценка $\|E - \tau C(v_k)\| \leq \bar{\rho}$, где

$$\kappa = \frac{\bar{\gamma}_3}{\sqrt{\bar{\gamma}_1 \bar{\gamma}_2 + \bar{\gamma}_3}}, \quad \tau_0 = \frac{2}{\bar{\gamma}_1 + \bar{\gamma}_2}, \quad \bar{\rho} = \frac{1 - \bar{\xi}}{1 + \bar{\xi}}, \quad \bar{\xi} = \frac{1 - \kappa}{1 + \kappa} \frac{\bar{\gamma}_1}{\bar{\gamma}_2}. \quad (16)$$

Теорема 3. Пусть оператор A имеет в шаре $\Omega(r)$ производную Гата $A'(v)$, которая при любом $v \in \Omega(r)$ удовлетворяет неравенствам

$$\begin{aligned}\bar{\gamma}_1 (Dy, y) &\leq (DB^{-1} A'(v)y, y) \leq \bar{\gamma}_2 (Dy, y), \quad \gamma_1 > 0, \\ \|0,5(DB^{-1} A'(v) - A'^*(v)(B^*)^{-1} D)y\|_{D^{-1}}^2 &\leq \bar{\gamma}_3^2 (Dy, y).\end{aligned} \quad (17)$$

Тогда при $\tau = \tau_0(1 - \kappa\rho)$ и $y_0 \in \Omega(r)$ итерационный метод (2) сходится в H_D , и для погрешности верна оценка (13), где $\rho = \bar{\rho}$ определено в (16).

Покажем теперь, что если оператор $A'(w)$ при $w \in \Omega(r)$ удовлетворяет условиям (17), то для любых $u, v \in \Omega(r)$ выполнены неравенства (4), (5) с постоянными $\gamma_1 = \bar{\gamma}_1$, $\bar{\gamma}_2 = (\bar{\gamma}_2 + \bar{\gamma}_3)^2 / \bar{\gamma}_1$. Тогда из леммы 1 будет следовать однозначная разрешимость уравнения (1).

В силу (11) имеем для $u, v \in \Omega(r)$ и $t \in [0, 1]$

$$(DB^{-1}Au - DB^{-1}Av, u - v) = (Ry, y), \quad R = DB^{-1}A'(w),$$

где $y = u - v$, $w = u + t(v - u) \in \Omega(r)$. Из (17) получим $(Ry, y) \geq \bar{\gamma}_1(Dy, y)$, т. е. выполнено неравенство (5) с $\gamma_1 = \bar{\gamma}_1$.

Далее имеем $(DB^{-1}Au - DB^{-1}Av, z) = (Ry, z)$. Представим оператор R в виде суммы $R = R_0 + R_1$, где $R_0 = 0,5(R + R^*)$ — симметрическая, а $R_1 = 0,5(R - R^*) = 0,5(DB^{-1}A'(w) - A''(w)(B^{*-1}D))$ — кососимметрическая части оператора R .

В силу неравенства Коши—Буняковского и условия (17) получим

$$\begin{aligned} (R_1y, z) &= (D^{-1/2}R_1y, D^{1/2}z) \leq (D^{-1}R_1y, R_1y)^{1/2} (Dz, z)^{1/2} = \\ &= \|R_1y\|_{D^{-1}} (Dz, z)^{1/2} \leq \bar{\gamma}_3 (Dy, y)^{1/2} (Dz, z)^{1/2}. \end{aligned}$$

Из обобщенного неравенства Коши—Буняковского найдем

$$\begin{aligned} (R_0y, z) &\leq (R_0y, y)^{1/2} (R_0z, z)^{1/2} = \\ &= (Ry, y)^{1/2} (Rz, z)^{1/2} \leq \bar{\gamma}_2 (Dy, y)^{1/2} (Dz, z)^{1/2}. \end{aligned}$$

Итак, мы получим неравенство

$$(Ry, z) \leq (\bar{\gamma}_2 + \bar{\gamma}_3) (Dy, y)^{1/2} (Dz, z)^{1/2}.$$

Полагая $z = B^{-1}(Au - Av)$ и используя (5), будем иметь

$$\begin{aligned} (DB^{-1}(Au - Av), B^{-1}(Au - Av)) &\leq \\ &\leq \frac{(\bar{\gamma}_2 + \bar{\gamma}_3)^2}{\bar{\gamma}_1} (DB^{-1}(Au - Av), u - v). \end{aligned}$$

Утверждение доказано.

3. Метод Ньютона—Канторовича. В теоремах 2 и 3 мы предполагали, что производная Гато $A'(v)$ существует и удовлетворяет соответствующим неравенствам для $v \in \Omega(r) = \{v : \|u - v\|_D \leq r\}$, где u — решение уравнения (1).

Из доказательства теорем следует, что достаточно потребовать на каждой итерации $k = 0, 1, \dots$ выполнения этих неравенств лишь для $v \in \Omega(r_k)$, где $r_k = \|u - y_k\|_D$.

В этом случае γ_1 и $\bar{\gamma}_2$ (а также и $\bar{\gamma}_1$, $\bar{\gamma}_2$ и $\bar{\gamma}_3$) могут зависеть от номера итераций k . Если выбирать итерационный параметр τ по формулам теорем 2 и 3, то получим нестационарный итерационный процесс (2) с $\tau = \tau_{k+1}$.

Более того, можно рассмотреть итерационный процесс

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H, \quad (18)$$

оператор $B = B_{k+1}$, в котором также зависит от номера итераций. Как выбрать операторы B_k ? Если оператор A линеен, то $A'(v) = A$ для любого $v \in H$. Тогда из теорем 2 и 3 следует, что при $B = A'(v) = A$ скорость сходимости итерационного метода (2) максимальна. Именно, при любом начальном приближении y_0 получим, что $y_1 = u$.

Выберем теперь оператор B_{k+1} в случае нелинейного оператора A следующим образом: $B_{k+1} = A'(y_k)$. Получим итерационную схему

$$A'(y_k) \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, y_0 \in H. \quad (19)$$

В соответствии с принятой терминологией итерационный процесс (19) является нелинейным. При $\tau_k \equiv 1$ он называется *методом Ньютона—Канторовича*. Для оценки скорости сходимости итерационного процесса (19) можно воспользоваться теоремами 2 и 3, в которых B следует заменить на $A'(y_k)$. В частности, для случая $D = E$ при $\tau_{k+1} = 1/\gamma_2$ имеет место оценка

$$\|y_{k+1} - u\| \leq \rho \|y_k - u\|, \quad \rho = \sqrt{1 - \gamma_1/\gamma_2} < 1, \quad (20)$$

где γ_1 и γ_2 взяты из неравенств (12) теоремы 2

$$\begin{aligned} \|A'(y_k)^{-1} A'(v)y\|^2 &\leq \gamma_2 ((A'(y_k)^{-1} A'(v)y, y)), \\ ((A'(y_k)^{-1} A'(v)y, y)) &\geq \gamma_1 (y, y), \quad \gamma_1 > 0 \end{aligned}$$

при $y \in H$, $v \in \Omega(r_k)$ и $r_k = \|u - y_k\|$. Из (20) следует, что $r_{k+1} = \|y_{k+1} - u\| \leq \rho r_k < r_k$ и, следовательно, $r_k \rightarrow 0$ при $k \rightarrow \infty$. Поэтому, если производная $A'(v)$ как функция от v со значениями в пространстве линейных операторов непрерывна в окрестности решения, то при $k \rightarrow \infty$ имеем, что $\gamma_1 \rightarrow 1$ и $\gamma_2 \rightarrow 1$. Это приведет к ускорению сходимости итерационного метода (19) при увеличении номера итерации k .

Приведенные рассуждения показывают, что методы вида (19) имеют при некоторых дополнительных предположениях о гладкости оператора $A'(v)$ более высокую скорость сходимости, чем скорость сходимости геометрической прогрессии.

Рассмотрим метод Ньютона—Канторовича (19) с $\tau_k \equiv 1$. Исследуем сходимость этого метода при следующих предположениях: 1) выполнены неравенства

$$\|A'(v) - A'(w)\| \leq \alpha \|v - w\|, \quad \alpha \geq 0, \quad (21)$$

$$\|A'(v)y\| \geq \frac{1}{\beta} \|y\|, \quad y \in H, \quad \beta > 0 \quad (22)$$

для $v, w \in \Omega(r)$; 2) начальное приближение y_0 принадлежит шару $\Omega(\bar{r})$, где $\bar{r} = \min(r, 1/(\alpha\beta))$.

Теорема 4. Если выполнены предположения 1) и 2), то для погрешности итерационного метода (19) с $\tau_k = 1$ верна оценка

$$\|y_n - u\| \leq \frac{1}{\alpha\beta} (\alpha\beta \|y_0 - u\|)^{2^n}. \quad (23)$$

Действительно, из (19) получим следующее соотношение:

$$\begin{aligned} A'(y_k)(y_{k+1} - u) &= A'(y_k)(y_k - u) - (Ay_k - Au) = Ty_k - Tu, \\ Tu &= A'(y_k)u - Au, \end{aligned}$$

где u — решение уравнения (1). Отсюда в силу неравенства Лагранжа для нелинейного оператора T получим

$$\|A'(y_k)(y_{k+1} - u)\| = \|Ty_k - Tu\| \leq \sup_{0 \leq t \leq 1} \|T'(v_k)\| \|y_k - u\|,$$

где $v_k = y_k + t(u - y_k)$. Из определения оператора T будем иметь

$$T'(v_k) = A'(y_k) - A'(v_k).$$

Предположим, что $y_k \in \Omega(\bar{r})$. Так как $\bar{r} \leq r$, то $y_k \in \Omega(r)$, и, следовательно, $v_k \in \Omega(r)$. Из неравенства (21) будем иметь

$$\begin{aligned} \|T'(v_k)\| &= \|A'(y_k) - A'(v_k)\| \leq \alpha \|y_k - v_k\| = \alpha t \|y_k - u\|, \\ \sup_{0 \leq t \leq 1} \|T'(v_k)\| &\leq \alpha \|y_k - u\|. \end{aligned}$$

Таким образом, найдена оценка

$$\|A'(y_k)(y_{k+1} - u)\| \leq \alpha \|y_k - u\|^2.$$

Используя неравенство (22), отсюда получим

$$\|y_{k+1} - u\| \leq \alpha\beta \|y_k - u\|^2. \quad (24)$$

Заметим, что так как $\|y_k - u\| \leq \bar{r}$ и $\alpha\beta\bar{r} \leq 1$, то

$$\|y_{k+1} - u\| \leq \alpha\beta\bar{r} \|y_k - u\| \leq \|y_k - u\| \leq \bar{r}.$$

Следовательно, из условия $y_k \in \Omega(\bar{r})$ вытекает, что $y_{k+1} \in \Omega(\bar{r})$.

Так как $y_0 \in \Omega(\bar{r})$, то по индукции находим, что $y_k \in \Omega(\bar{r})$ для любого $k \geq 0$. Поэтому оценка (24) верна для любого $k \geq 0$.

Решим неравенство (24). Умножим его на $\alpha\beta$ и обозначим $q_k = \alpha\beta \|y_k - u\|$. Для q_k получим неравенство $q_{k+1} \leq q_k^2$, $k = 0, 1, \dots$

По индукции легко доказать, что его решение имеет вид $q_n \leq q_0^{2^n}$, $n \geq 0$. Следовательно, имеем оценку,

$$\alpha\beta \|y_n - u\| \leq (\alpha\beta \|y_0 - u\|)^{2^n}.$$

Отсюда и следует утверждение теоремы.

Замечание 1. Если начальное приближение y_0 выбрано так, что $\bar{r} \leq \rho/(\alpha\beta)$, $\rho < 1$, то из (23) следует оценка

$$\|y_n - u\| \leq \rho^{2^n-1} \|y_0 - u\|$$

и оценка

$$n \geq n_0(\varepsilon) = \log_2 (\ln \varepsilon / \ln \rho + 1)$$

для числа итераций.

Замечание 2. Если вместо условия (21) выполнено неравенство

$$\|A'(v) - A'(w)\| \leq \alpha \|v - w\|^p, \quad p \in (0, 1],$$

то для погрешности верна оценка

$$\|y_n - u\| \leq \frac{1}{\sqrt[p]{\alpha\beta}} (\sqrt[p]{\alpha\beta} \|y_0 - u\|)^{(p+1)^n},$$

$$\left(\bar{r} = \min \left(r, \frac{1}{\sqrt[p]{\alpha\beta}} \right) \right).$$

При доказательстве теоремы 4 мы получили оценку для погрешности (23). Эта оценка бесполезна с точки зрения практического ее применения, но она важна для теории метода, поскольку показывает, как осуществляется сходимость вблизи решения u .

Теорема 4 позволяет находить области несуществования решения. Действительно, теорема утверждает, что y_k сходится к u , если $\|y_0 - u\| \leq \bar{r}$. Поэтому, если итерации не сходятся, то в шаре $\|y_0 - v\| \leq \bar{r}$ с центром в точке y_0 решений уравнения (1) нет.

Отметим, что если оператор A имеет в шаре $\Omega(r)$ вторую производную Гато, то в неравенстве (21)

$$\alpha = \sup_{0 \leq t \leq 1} \|A''(v + t(w - v))\|.$$

При реализации итерационной схемы (19) для каждого k нужно решать линейное операторное уравнение

$$A'(y_k)v = F(y_k), \quad (25)$$

где

$$F(y_k) = A'(y_k)y_k - \tau_{k+1}(Ay_k - f). \quad (26)$$

Если v — точное решение уравнения (25), то в (19) $y_{k+1} = v$.

Оператор $A'(y_k)$ необходимо вычислять на каждой итерации, и это может потребовать больших вычислительных затрат. Рассмотрим пример. Пусть оператор A соответствует системе нелинейных уравнений

$$\varphi_i(u) = 0, \quad i = 1, 2, \dots, m, \quad u = (u_1, u_2, \dots, u_m).$$

Производная Гато $A'(y)$ в точке $y = (y_1, y_2, \dots, y_m)$ есть квадратная матрица с элементами $a_{ij}(y)$, где

$$a_{ij}(y) = \left. \frac{\partial \varphi_i(u)}{\partial u_j} \right|_{u=y}, \quad i, j = 1, 2, \dots, m.$$

Следовательно, на каждой итерации нужно вычислять m^2 элементов матрицы $A'(y)$, тогда как число неизвестных в задаче равно m .

Чтобы избежать вычисления производной $A'(y_k)$ на каждой итерации, используют следующую модификацию схемы (19):

$$A'(y_{km}) \frac{y_{km+i+1} - y_{km+i}}{\tau_{km+i+1}} + A y_{km+i} = f,$$

$$i = 0, 1, \dots, m-1, \quad k = 0, 1, \dots$$

Здесь производная A' вычисляется через каждые m итераций и используется для нахождения промежуточных приближений $y_{km+1}, y_{km+2}, \dots, y_{(k+1)m}$. При $m=1$ получим итерационную схему (19).

4. Двухступенчатые итерационные методы. Итерационную схему (19) целесообразно использовать в случае, когда оператор $A'(y_k)$ легко обратим. При этом точное решение v уравнения (25) принимается за новое итерационное приближение y_{k+1} , которое удовлетворяет схеме (19). Таким образом, мы получаем итерационную схему, оператор B_{k+1} в которой задан в явном виде: $B_{k+1} = A'(y_k)$.

Если уравнение (25) решается приближенно, например при помощи вспомогательного (внутреннего) итерационного метода и в качестве y_{k+1} берется n -е итерационное приближение v_n , то B_{k+1} удовлетворяет общей схеме (18) с некоторым $B_{k+1} \neq A'(y_k)$. В этом случае явный вид оператора B_{k+1} не используется, а знание его структуры необходимо лишь для исследования сходимости итерационной схемы (18). Построенные таким способом итерационные методы иногда называют двухступенчатыми, подразумевая под этим специальный алгоритм обращения оператора B_{k+1} .

Опишем более подробно общую схему построения двухступенчатых методов. Пусть для решения линейного уравнения (25) используется какой-либо неявный двухслойный итерационный метод

$$\bar{B}_{n+1} \frac{v_{n+1} - v_n}{\omega_{n+1}} + A'(y_k) v_n = F(y_k), \quad n = 0, 1, \dots, m-1, \quad (27)$$

где $F(y_k)$ определено в (26), $\{\omega_n\}$ — набор итерационных параметров, \bar{B}_{n+1} — операторы в H , которые могут зависеть от y_k , а $v_0 = y_k$.

Выразим v_n через y_k . Сначала найдем уравнение для погрешности $z_n = v_n - v$, где v — решение уравнения (25). Из (25) и (27) найдем

$$z_{n+1} = S_{n+1} z_n, \quad n = 0, 1, \dots, \quad S_n = E - \omega_n \bar{B}_n^{-1} A'(y_k)$$

и, следовательно,

$$\begin{aligned} z_m &= v_m - v = T_m z_0 = T_m (v_0 - v), \quad T_m = S_m S_{m-1} \dots S_1, \\ v_m &= (E - T_m) v + T_m y_k. \end{aligned} \quad (28)$$

Из (25), (26) получим

$$v = [A'(y_k)]^{-1} F(y_k) = y_k - \tau_{k+1} [A'(y_k)]^{-1} (Ay_k - f).$$

Подставляя найденное v в (28), будем иметь

$$y_{k+1} = v_m = y_k - \tau_{k+1} (E - T_m) [A'(y_k)]^{-1} (Ay_k - f).$$

Отсюда следует, что y_{k+1} удовлетворяет итерационной схеме (18), если обозначить

$$B_{k+1} = A'(y_k) (E - T_m)^{-1}. \quad (29)$$

Итак, реализация одного шага двухступенчатого метода состоит в вычислении $F(y_k)$ по формуле (26) и выполнении m итераций по схеме (27) с начальным приближением $v_0 = y_k$. Полученное приближение v_m берется в качестве y_{k+1} .

Рассмотрим итерационную схему (18), (29). Для оценки скорости сходимости можно использовать теоремы 2 и 3, в которых B заменено на B_{k+1} , а τ на τ_{k+1} . Недостатком такого выбора параметра τ является то, что нужно достаточно точно оценить γ_1 , γ_2 и γ_3 .

Отметим, что при построении двухступенчатого метода можно было бы исходить не из уравнения (25), а из «близкого» к нему уравнения

$$Rv = F(y_k),$$

где линейный оператор R в некотором смысле эквивалентен оператору $A'(y_k)$. В этом случае в итерационной схеме (18) имеем

$$B_{k+1} \equiv B = R(E - T_m)^{-1}.$$

Исследуем этот случай более подробно. Пусть выполнены условия

$$R = R^* > 0, \quad T_m^* R = RT_m, \quad (30)$$

$$\|T_m\|_R \leq q < 1. \quad (31)$$

Лемма 2. Пусть выполнены условия (30), (31). Тогда оператор $B = R(E - T_m)^{-1}$ самосопряжен и положительно определен в H , и справедливы неравенства

$$(1 - q) B \leq R \leq (1 + q) B. \quad (32)$$

Рассмотрим оператор $B^{-1} = (E - T_m) R^{-1}$. Из (30) найдем $(E - T_m^*) R = R(E - T_m)$ или $R^{-1}(E - T_m^*) = (E - T_m) R^{-1}$. Следовательно, оператор B^{-1} самосопряжен в H .

Так как в силу (30) оператор T_m самосопряжен в H_R , то

$$\|T_m\|_R = \sup_{x \neq 0} \frac{|(T_m x, x)_R|}{(x, x)_R} = \sup_{x \neq 0} \frac{|(RT_m x, x)|}{(Rx, x)} \leq q < 1.$$

Следовательно, для любого $x \in H$ имеем неравенство

$$|(RT_m x, x)| \leq q(Rx, x).$$

Полагая здесь $x = R^{-1}y$, получим

$$|(T_m R^{-1}y, y)| \leq q(R^{-1}y, y),$$

поэтому для $y \in H$ найдем

$$(1-q)(R^{-1}y, y) \leq ((E - T_m)R^{-1}y, y) \leq (1+q)(R^{-1}y, y).$$

Итак, получена оценка

$$(1-q)R^{-1} \leq B^{-1} \leq (1+q)R^{-1}. \quad (33)$$

Так как R^{-1} и B^{-1} —самосопряженные в H операторы и $q < 1$, то из леммы 9 § 1 гл. V следует, что неравенства (33) и (32) эквивалентны. Лемма доказана.

Лемма 3. Пусть оператор A имеет в шаре $\Omega(r)$ производную Гато $A'(v)$, которая при любом $v \in \Omega(r)$ удовлетворяет неравенствам

$$c_1(Ry, y) \leq (A'(v)y, y) \leq c_2(Ry, y), \quad c_1 > 0, \quad (34)$$

$$\|0.5[A'(v) - (A'(v))^*]y\|_{R^{-1}}^2 \leq c_3^2(Ry, y), \quad c_3 \geq 0. \quad (35)$$

и пусть выполнены условия (30), (31). Тогда выполнены неравенства (17) теоремы 3, где

$$\begin{aligned} \bar{\gamma}_1 &= c_1(1-q), \quad \bar{\gamma}_2 = c_2(1+q), \quad \bar{\gamma}_3 = c_3(1+q)^2, \\ D &= B = R(E - T_m). \end{aligned}$$

Действительно, в силу леммы 2 оператор D самосопряжен и положительно определен в H . Кроме того, неравенства (17) при $D = B$ имеют вид

$$\bar{\gamma}_1(By, y) \leq (A'(v)y, y) \leq \bar{\gamma}_2(By, y), \quad (36)$$

$$\|0.5[A'(v) - (A'(v))^*]y\|_{B^{-1}}^2 \leq \bar{\gamma}_3^2(By, y). \quad (37)$$

Неравенства (36) с указанными в лемме 3 $\bar{\gamma}_1$ и $\bar{\gamma}_2$ следуют из (32) и (34), а (37) вытекает из (32), (33) и (35), так как

$$\begin{aligned} \|z\|_{B^{-1}}^2 &= (B^{-1}z, z) \leq (1+q)(R^{-1}z, z) = (1+q)\|z\|_{R^{-1}}^2, \\ (Rz, z) &\leq (1+q)(Bz, z). \end{aligned}$$

Используя лемму 3, можно доказать аналог теоремы 3 для двухступенчатого метода.

Теорема 5. Пусть выполнены условия леммы 3, и двухступенчатый метод построен на основе уравнения $Rv = F(y_k)$ с использованием разрешающего оператора T_m . Если в итерационной схеме (18) с $B_{k+1} \equiv B = R(E - T_m)^{-1}$, описывающей этот двухступенчатый метод, выбрать $\tau_k \equiv \tau_0(1 - \kappa\rho)$ и $y_0 \in \Omega(r)$, то для погрешности будет верна оценка

$$\|y_n - u\|_B \leq \bar{\rho}^n \|y_0 - u\|_B,$$

где u — решение уравнения (1), $\bar{\varrho}$, x и τ_0 определены в (16) с $\bar{\gamma}_1$, $\bar{\gamma}_2$ и $\bar{\gamma}_3$, указанными в лемме 3.

5. Другие итерационные методы. В этом пункте мы приведем краткое описание некоторых итерационных методов, также используемых для решения уравнения (1) с нелинейным оператором A .

Пусть $\Phi(u)$ — дифференцируемый по Гато функционал, заданный в H . Оператор A , действующий в H , называется *потенциальным*, если существует дифференцируемый функционал $\Phi(u)$ такой, что $Au = \text{grad } \Phi(u)$ для всех u . Здесь градиент функционала $\Phi(u)$ определяется равенством $\frac{d}{dt} \Phi(u + tv)|_{t=0} = (\text{grad } \Phi(u), v)$. Примером потенциального оператора может служить ограниченный линейный самосопряженный оператор A , действующий в гильбертовом пространстве H . Он порождается функционалом $\Phi(u) = 0,5(Au, u)$.

Пусть оператор A непрерывно дифференцируем в H . Оператор A является потенциальным тогда и только тогда, когда производная Гато $A'(v)$ есть самосопряженный в H оператор.

Если оператор A потенциален, то формула

$$\Phi(u) = \int_0^1 (A(u_0 + t(u - u_0)), u - u_0) dt,$$

где u_0 — произвольный, но фиксированный элемент H , дает способ построения функционала $\Phi(u)$ по оператору A .

Если оператор A порожден градиентом строго выпуклого функционала, то производная $A'(v)$ является положительно определенным в H оператором для любого $v \in H$. В этом случае для приближенного решения уравнения (25) можно использовать итерационные методы вариационного типа, например в (27) итерационные параметры ω_{n+1} выбирать по формулам методов скользящего спуска, минимальных невязок и т. д.

Рассмотрим для примера двухступенчатый метод (18), (29), для которого $\tau_{k+1} = 1$, а в схеме (27) $m = 1$ и $\bar{B}_1 = E$. Тогда $B_{k+1} = E/\omega_1$. Если для вспомогательного итерационного процесса (27) параметр ω_1 выбрать по формулам метода минимальных невязок (или минимальных поправок), то получим (см. п. 2, 3 § 2 гл. VIII)

$$\omega_1 = \frac{(A'(y_k) r_k, r_k)}{\|A'(y_k) r_k\|^2}, \quad r_k = Ay_k - f. \quad (38)$$

В этом случае двухступенчатый итерационный метод описывается формулой

$$\frac{y_{k+1} - y_k}{\omega_1} + Ay_k = f, \quad k = 0, 1, \dots, \quad (39)$$

где ω_1 определено в (38).

В ситуации, когда оператор A не является потенциальным, параметр ω_1 можно выбрать по формулам метода минимальных погрешностей, полагая в (27) $\bar{B}_1 = [(A'(y_k))^*]^{-1}$ и

$$\omega_1 = \frac{(r_k, r_k)}{\|(A'(y_k))^* r_k\|^2}, \quad r_k = Ay_k - f. \quad (40)$$

В этом случае двухступенчатый метод имеет вид

$$\frac{y_{k+1} - y_k}{\omega_1} + (A'(y_k))^* Ay_k = (A'(y_k))^* f, \quad k = 0, 1, \dots, \quad (41)$$

где ω_1 определено в (40).

Легко видеть, что в методе (38), (39) параметр ω_1 выбирается из условия минимума $\|A'(y_k)(y_{k+1} - y_k) + Ay_k - f\|$, а в методе (40), (41) — из условия минимума нормы $\|y_{k+1} - y_k + [A'(y_k)]^{-1}(Ay_k - f)\|$.

Задача решения уравнения $Au = f$ в случае потенциального оператора иногда может быть заменена задачей минимизации функционала, порождающего этот оператор. Заметим, что всегда имеется простой способ преобразовать задачу решения уравнения (1) в задачу минимизации, даже если оператор A не является потенциальным.

Действительно, пусть $\Phi(u)$ — функционал, заданный в H и имеющий единственную точку минимума $u = 0$. В качестве примера такого функционала можно привести $\Phi(u) = (Du, u)$, где D — самосопряженный положительно определенный в H оператор. Далее, для заданного уравнения (1) рассмотрим функционал

$$F(u) = \Phi(Au - f), \quad u \in H.$$

Если уравнение (1) имеет решение u , то, очевидно, оно доставляет минимум функционалу $F(u)$.

Опишем метод минимизации функционала (метод спуска). Пусть уравнение (1) порождено градиентом строго выпуклого функционала $\Phi(u)$. Пусть минимизирующая последовательность строится согласно итерационной схеме (19), т. е. по формуле

$$y_{k+1} = y_k - \tau_{k+1} [A'(y_k)]^{-1} (Ay_k - f), \quad k = 0, 1, \dots \quad (42)$$

Обозначим

$$w_k = [A'(y_k)]^{-1} \operatorname{grad} \Phi(y_k), \quad (43)$$

где в силу сделанных предположений $\operatorname{grad} \Phi(y_k) = Ay_k - f$. Запишем (42) в виде

$$y_{k+1} = y_k - \tau_{k+1} w_k.$$

Отметим, что оператор $A'(y_k)$ положительно определен и самосопряжен в H . Далее, из определения производной Гато функционала имеем

$$\lim_{\tau_{k+1} \rightarrow 0} \left[\frac{\Phi(y_k - \tau_{k+1} w_k) - \Phi(y_k)}{\tau_{k+1}} \right] + (\operatorname{grad} \Phi(y_k), w_k) = 0.$$

Так как $A'(y_k)w_k = \text{grad } \Phi(y_k)$, то

$$(\text{grad } \Phi(y_k), w_k) = (A'(y_k)w_k, w_k) > 0.$$

Следовательно, существует такое $\tau_{k+1} > 0$, что $\Phi(y_{k+1})$ будет строго меньше $\Phi(y_k)$.

Если минимизирующая последовательность $\{y_k\}$ строится по явной схеме (18) ($B_k \equiv E$), т. е. по формулам

$$y_{k+1} = y_k - \tau_{k+1}(Ay_k - f),$$

то переход от y_k к y_{k+1} осуществляется по направлению градиента функционала $\Phi(u)$ в точке y_k . Такие методы принято называть *методами градиентного спуска*. Существуют некоторые алгоритмы выбора итерационных параметров τ_k , но на этих вопросах мы не будем здесь останавливаться.

Приведем в заключение обобщение явного метода сопряженных градиентов, который используется для минимизации функционала при указанных выше предположениях. Формулы алгоритма Флетчера—Ривса имеют вид:

$$\begin{aligned} y_{k+1} &= y_k - a_{k+1}w_k, & k = 0, 1, \dots, \\ w_k &= \text{grad } \Phi(y_k) + b_k w_{k-1}, & k = 1, 2, \dots, \\ w_0 &= \text{grad } \Phi(y_0), \end{aligned}$$

где

$$b_k = \frac{\|\text{grad } \Phi(y_k)\|^2}{\|\text{grad } \Phi(y_{k-1})\|^2}, \quad k = 1, 2, \dots,$$

а параметр a_{k+1} выбирается из условия минимума $\Phi(y_k - a_{k+1}w_k)$. Эта задача об отыскании минимума функции одной переменной решается одним из методов численного анализа.

§ 2. Методы решения нелинейных разностных схем

1. Разностная схема для одномерного эллиптического квазилинейного уравнения. Изложенную в § 1 общую теорию итерационных методов будем применять для нахождения приближенного решения нелинейных эллиптических разностных схем. Начнем с простейших примеров.

Рассмотрим третью краевую задачу для одномерного квазилинейного уравнения в дивергентном виде

$$\begin{aligned} Lu &= \frac{d}{dx} k_1 \left(x, u, \frac{du}{dx} \right) - k_0 \left(x, u, \frac{du}{dx} \right) = -\varphi(x), \quad 0 \leq x \leq l, \\ k_1 \left(x, u, \frac{du}{dx} \right) &= \kappa_0(u) - \mu_0, \quad x = 0, \\ -k_1 \left(x, u, \frac{du}{dx} \right) &= \kappa_1(u) - \mu_1, \quad x = l. \end{aligned} \tag{1}$$

Будем предполагать, что функции $k_1(x, p_0, p_1)$, $k_0(x, p_0, p_1)$, $\kappa_0(p_0)$ и $\kappa_1(p_0)$ непрерывны по p_0 и p_1 и выполнены условия сильной эллиптичности

$$\sum_{\alpha=0}^1 [k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)](p_\alpha - q_\alpha) \geq c_1 \sum_{\alpha=0}^1 (p_\alpha - q_\alpha)^2, \quad (2)$$

$$[\kappa_\alpha(p_0) - \kappa_\alpha(q_0)](p_0 - q_0) \geq 0, \quad \alpha = 0, 1, \quad (3)$$

где $c_1 > 0$ — положительная постоянная, $0 \leq x \leq l$, $|p_0|, |q_0|, |p_1|, |q_1| < \infty$.

На равномерной сетке $\bar{\omega} = \{x_i = ih, i = 0, 1, \dots, N, hN = l\}$ задаче (1) поставим в соответствие разностную схему

$$\Lambda y_i = -f_i, \quad 0 \leq i \leq N, \quad (4)$$

где

$$f_i = \begin{cases} \varphi(0) + \frac{2}{h}\mu_0, & i = 0, \\ \varphi(x_i), & 1 \leq i \leq N-1, \\ \varphi(l) + \frac{2}{h}\mu_1, & i = N. \end{cases}$$

Разностный оператор Λ определяется формулами:

$$\begin{aligned} \Lambda y_i = & \frac{1}{2} \{ [k_1(x, y, y_x)]_{\bar{x}} + [k_1(x, y, y_{\bar{x}})]_x - \\ & - k_0(x, y, y_x) - k_0(x, y, y_{\bar{x}}) \}_t, \quad 1 \leq i \leq N-1, \end{aligned}$$

$$\begin{aligned} \Lambda y_0 = & \frac{1}{h} [k_1(0, y_0, y_{x,0}) + k_1(h, y_1, y_{\bar{x},1})] - \\ & - k_0(0, y_0, y_{x,0}) - \frac{2}{h} \kappa_0(y_0), \quad i = 0, \end{aligned}$$

$$\begin{aligned} \Lambda y_N = & -\frac{1}{h} [k_1(l-h, y_{N-1}, y_{x,N-1}) + k_1(l, y_N, y_{\bar{x},N})] - \\ & - k_0(l, y_N, y_{\bar{x},N}) - \frac{2}{h} \kappa_1(y_N), \quad i = N. \end{aligned}$$

Если в пространстве $H = H(\bar{\omega})$ определить нелинейный оператор A соотношением $A = -\Lambda$, то разностная схема (4) запишется в виде операторного уравнения $Au = f$.

Исследуем свойства нелинейного оператора A , действующего из H в H . Напомним, что скалярное произведение в $H(\bar{\omega})$ определяется по формуле

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0.5h(u_0 v_0 + u_N v_N),$$

а через $(u, v)_{\omega+}$ и $(u, v)_{\omega-}$ обозначаются суммы

$$(u, v)_{\omega+} = \sum_{i=1}^N u_i v_i h, \quad (u, v)_{\omega-} = \sum_{i=0}^{N-1} u_i v_i h,$$

так что

$$(u, v) = \frac{1}{2} [(u, v)_{\omega+} + (u, v)_{\omega-}].$$

Покажем, что при выполнении условий (2), (3) оператор A является сильно монотонным в $H(\bar{\omega})$, т. е. выполнено неравенство

$$(Au - Av, u - v) \geq c_1 \|u - v\|^2, \quad c_1 > 0, \quad (5)$$

где c_1 определено в (2).

Обозначим $p_0 = \bar{p}_0 = u_i$, $q_0 = \bar{q}_0 = v_i$, $p_1 = u_{x,i}$, $\bar{p}_1 = u_{\bar{x},i}$, $q_1 = v_{x,i}$, $\bar{q}_1 = v_{\bar{x},i}$. Используя определение оператора A , формулы суммирования по частям (см. (7), (9) § 2 гл. V) и условия (2), (3), получим

$$\begin{aligned} (Au - Av, u - v) &= (\Lambda v - \Lambda u, u - v) = \\ &= \frac{1}{2} \sum_{i=1}^N h \left\{ \sum_{\alpha=0}^1 [k_\alpha(x, \bar{p}_0, \bar{p}_1) - k_\alpha(x, \bar{q}_0, \bar{q}_1)] (\bar{p}_\alpha - \bar{q}_\alpha) \right\}_i + \\ &\quad + \frac{1}{2} \sum_{i=0}^{N-1} h \left\{ \sum_{\alpha=0}^1 [k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)] (p_\alpha - q_\alpha) \right\}_i + \\ &+ [\kappa_1(\bar{p}_0) - \kappa_1(\bar{q}_0)] (\bar{p}_0 - \bar{q}_0) \Big|_{i=N} + [\kappa_0(p_0) - \kappa_0(q_0)] (p_0 - q_0) \Big|_{i=0} \geq \\ &\geq \frac{c_1}{2} \sum_{i=1}^N h \sum_{\alpha=0}^1 (\bar{p}_\alpha - \bar{q}_\alpha)_i^2 + \frac{c_1}{2} \sum_{i=0}^{N-1} h \sum_{\alpha=0}^1 (p_\alpha - q_\alpha)_i^2. \end{aligned}$$

Учитывая равенство $u_{x,i} = u_{\bar{x},i+1}$, запишем полученную оценку в виде

$$\begin{aligned} (Au - Av, u - v) &\geq \frac{c_1}{2} [(u - v, u - v)_{\omega+} + (u - v, u - v)_{\omega-} + \\ &\quad + \sum_{i=1}^N h (u - v)_{x,i}^2 + \sum_{i=0}^{N-1} h (u - v)_{\bar{x},i}^2] = \\ &= c_1 [\|u - v\|^2 + ((u - v)_{\bar{x}}^2, 1)_{\omega+}] \geq c_1 \|u - v\|^2. \end{aligned}$$

Из замечания 2 к лемме 12 главы V следует, что эта оценка не может быть улучшена.

Итак, установлена сильная монотонность оператора A . В силу непрерывности функций $k_\alpha(x, p_0, p_1)$ и $\kappa_\alpha(p_0)$, оператор A непрерывен в H . Поэтому из теоремы 11 главы V следует существование и единственность в шаре $\|u\| \leq \frac{1}{c_1} \|A0 - f\|$ решения уравнения $Au = f$ и, следовательно, разностной задачи (4).

Если $k_\alpha(x, p_0, p_1)$ и $\kappa_\alpha(p_0)$, $\alpha = 0, 1$ — непрерывно дифференцируемые функции своих аргументов, то вместо (2), (3) можно

использовать другие достаточные условия, обеспечивающие сильную монотонность оператора A .

Будем предполагать, что выполняются условия

$$c_1 \sum_{\alpha=0}^1 \xi_\alpha^2 \leqslant \sum_{\alpha, \beta=0}^1 a_{\alpha\beta}(x, p_0, p_1) \xi_\alpha \xi_\beta \leqslant c_2 \sum_{\alpha=0}^1 \xi_\alpha^2, \quad c_1 > 0, \quad (6)$$

$$0 \leqslant \sigma_\alpha(p_0) \leqslant c_3, \quad \alpha = 1, 2, \quad (7)$$

где $\xi = (\xi_0, \xi_1)$ — произвольный вектор и

$$a_{\alpha\beta}(x, p_0, p_1) = \frac{\partial k_\alpha(x, p_0, p_1)}{\partial p_\beta}, \quad \sigma_\alpha(p_0) = \frac{\partial \kappa_\alpha(p_0)}{\partial p_0}, \quad \alpha, \beta = 0, 1.$$

Покажем, что из условий (6), (7) следуют (2), (3). Действительно, имеют место равенства

$$\begin{aligned} k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1) &= \\ &= \int_0^1 \frac{d}{dt} k_\alpha(x, tp_0 + (1-t)q_0, tp_1 + (1-t)q_1) dt = \\ &= (p_0 - q_0) \int_0^1 \frac{\partial k_\alpha(x, s_0, s_1)}{\partial s_0} dt + (p_1 - q_1) \int_0^1 \frac{\partial k_\alpha(x, s_0, s_1)}{\partial s_1} dt = \\ &= \sum_{\beta=0}^1 (p_\beta - q_\beta) \int_0^1 a_{\alpha\beta}(x, s_0, s_1) dt, \quad \alpha = 0, 1, \end{aligned}$$

где $s_0 = tp_0 + (1-t)q_0$, $s_1 = tp_1 + (1-t)q_1$. Умножая это равенство на $p_\alpha - q_\alpha$ и суммируя его по α от 0 до 1, получим с учетом (6)

$$\begin{aligned} \sum_{\alpha=0}^1 [k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)] (p_\alpha - q_\alpha) &= \\ &= \int_0^1 \sum_{\alpha, \beta=0}^1 a_{\alpha\beta}(x, s_0, s_1) (p_\alpha - q_\alpha) (p_\beta - q_\beta) dt \geqslant \\ &\geqslant c_1 \int_0^1 \sum_{\alpha=0}^1 (p_\alpha - q_\alpha)^2 dt = c_1 \sum_{\alpha=0}^1 (p_\alpha - q_\alpha)^2. \end{aligned}$$

Итак, неравенство (2) получено. Аналогично из (7) получим неравенство (3)

$$[\kappa_\alpha(p_0) - \kappa_\alpha(q_0)] (p_0 - q_0) = \int_0^1 \frac{\partial \kappa_\alpha(s_0)}{\partial s_0} dt (p_0 - q_0)^2 \geqslant 0.$$

Таким образом, условия (6), (7) гарантируют существование и единственность решения разностной задачи (4).

Найдем теперь производную Гато оператора A , предполагая, что функции $k_\alpha(x, p_0, p_1)$ и $\kappa_\alpha(p_0)$, $\alpha = 0, 1$, имеют ограниченные производные по p_0 и p_1 нужного порядка.

Из определения производной Гато нелинейного оператора будем иметь

$$\begin{aligned} A'(u)y_i = & -\frac{1}{2}\{[a_{11}(x, u, u_x)y_x]_{\bar{x}, i} + [a_{11}(x, u, u_{\bar{x}})y_{\bar{x}}]_{x, i} + \\ & + [a_{10}(x, u, u_x)y]_{\bar{x}, i} + [a_{10}(x, u, u_{\bar{x}})y]_{x, i}\} + \\ & + \frac{1}{2}\{a_{01}(x, u, u_x)y_{x, i} + a_{01}(x, u, u_{\bar{x}})y_{\bar{x}, i} + \\ & + [a_{00}(x, u, u_x) + a_{00}(x, u, u_{\bar{x}})]y_i\}, \quad 1 \leq i \leq N-1. \end{aligned}$$

При $i=0$ получим

$$\begin{aligned} A'(u)y_0 = & -\frac{1}{h}[a_{11}(0, u_0, u_{x, 0}) + a_{11}(h, u_1, u_{\bar{x}, 1}) - \\ & - ha_{01}(0, u_0, u_{x, 0}) + ha_{10}(h, u_1, u_{\bar{x}, 1})]y_{x, 0} + \frac{2}{h}[\sigma_0(u_0) - \\ & - \frac{1}{2}a_{10}(0, u_0, u_{x, 0}) - \frac{1}{2}a_{10}(h, u_1, u_{\bar{x}, 1}) + \frac{h}{2}a_{00}(0, u_0, u_{x, 0})]y_0, \end{aligned}$$

а при $i=N$ будем иметь

$$\begin{aligned} A'(u)y_N = & \frac{1}{h}[a_{11}(l-h, u_{N-1}, u_{x, N-1}) + a_{11}(l, u_N, u_{\bar{x}, N}) + \\ & + ha_{01}(l, u_N, u_{\bar{x}, N}) - ha_{10}(l-h, u_{N-1}, u_{x, N-1})]y_{\bar{x}, N} + \frac{2}{h}[\sigma_1(u_N) + \\ & + \frac{1}{2}a_{10}(l, u_N, u_{\bar{x}, N}) + \frac{1}{2}a_{10}(l-h, u_{N-1}, u_{x, N-1}) + \\ & + \frac{h}{2}a_{00}(l, u_N, u_{\bar{x}, N})]y_N. \end{aligned}$$

Отметим, что при вычислении $A'(u)y_0$ и $A'(u)y_N$ были использованы соотношения

$$y_1 = y_0 + hy_{x, 0}, \quad y_{N-1} = y_N - hy_{\bar{x}, N}. \quad (8)$$

Исследуем свойства производной Гато $A'(u)$ оператора A .

Лемма 4. Если выполнены условия

$$\frac{\partial k_1(x, p_0, p_1)}{\partial p_0} = \frac{\partial k_0(x, p_0, p_1)}{\partial p_1}, \quad (9)$$

то $A'(u)$ — самосопряженный в H оператор. При выполнении условий (6), (7) он положительно определен в H .

В самом деле, используя формулы суммирования по частям, а также соотношения (8), получим

$$\begin{aligned}
 (A'(u)y, z) = & \frac{1}{2} \sum_{i=0}^{N-1} h [a_{11}(x, u, u_x) y_x z_x + a_{10}(x, u, u_x) y z_x + \\
 & + a_{01}(x, u, u_x) y_x z + a_{00}(x, u, u_x) y z]_i + \\
 & + \frac{1}{2} \sum_{i=1}^N h [a_{11}(x, u, u_{\bar{x}}) y_{\bar{x}} z_{\bar{x}} + a_{10}(x, u, u_{\bar{x}}) y z_{\bar{x}} + \\
 & + a_{01}(x, u, u_{\bar{x}}) y_{\bar{x}} z + a_{00}(x, u, u_{\bar{x}}) y z]_i + \sigma_0(u_0) y_0 z_0 + \sigma_1(u_N) y_N z_N. \tag{10}
 \end{aligned}$$

Сравнивая это выражение с выражением для $(y, A'(u)z)$, получим, что при условии $a_{10}(x, p_0, p_1) = a_{01}(x, p_0, p_1)$, которое является другой формой записи для (9), оператор $A'(u)$ для любого $u \in H$ самосопряжен в H .

Пусть теперь выполнены условия (6), (7). Полагая в (10) $z_i \equiv y_i$, получим

$$\begin{aligned}
 (A'(u)y, y) \geq & \frac{c_1}{2} \left[\sum_{i=0}^{N-1} h(y_i^2 + y_{x,i}^2) + \sum_{i=1}^N h(y_i^2 + y_{\bar{x},i}^2) \right] = \\
 & = c_1 [(y, y) + (y_{\bar{x}}^2, 1)_{\omega+}] \geq c_1 (y, y), \tag{11}
 \end{aligned}$$

т. е. оператор $A'(u)$ положительно определен в H . Лемма доказана.

Заметим, что в силу теоремы 2 главы V из положительной определенности производной Гато непрерывного оператора A следует, что он является сильно монотонным. Таким образом, при выполнении условий (6), (7) оператор A сильно монотонный.

Полагая в (10) $z_i \equiv y_i$, получим в силу условий (6), (7) оценку сверху

$$\begin{aligned}
 (A'(u)y, y) \leq & \frac{c_2}{2} \left[\sum_{i=0}^{N-1} h(y_i^2 + y_{x,i}^2) + \sum_{i=1}^N h(y_i^2 + y_{\bar{x},i}^2) \right] + \\
 & + c_3(y_0^2 + y_N^2) = c_2 [(y, y) + (y_{\bar{x}}^2, 1)_{\omega+}] + c_3(y_0^2 + y_N^2).
 \end{aligned}$$

Из неравенства (36) леммы 15 главы V при $\varepsilon = 1$ найдем, что

$$y_0^2 + y_N^2 \leq c_4 [(y, y) + (y_{\bar{x}}^2, 1)_{\omega+}], \quad c_4 = \frac{8+l^2}{l \sqrt{16+l^2}}. \tag{12}$$

Следовательно, имеем

$$(A'(u)y, y) \leq \gamma_2 [(y, y) + (y_{\bar{x}}^2, 1)_{\omega+}], \quad \gamma_2 = c_2 + c_3 c_4. \tag{13}$$

Определим в пространстве $H = H(\bar{\omega})$ линейный оператор R , отображающий H на \dot{H} , по формулам

$$Ry_i = \begin{cases} -\frac{2}{h} y_{x, 0} + y_0, & i = 0, \\ -y_{\bar{x}, i} + y_i, & 1 \leq i \leq N-1, \\ \frac{2}{h} y_{\bar{x}, N} + y_N, & i = N. \end{cases}$$

Из первой разностной формулы Грина найдем

$$(Ry, y) = (y, y) + (y_{\bar{x}}^2, 1)_{\omega+}. \quad (14)$$

Тогда из (11), (13), (14) легко следует, что при выполнении условий (6), (7) для производной Гато $A'(u)$ оператора A справедливы неравенства

$$\gamma_1(Ry, y) \leq (A'y, y) \leq \gamma_2(Ry, y), \quad (15)$$

где $\gamma_1 = c_1 > 0$, $\gamma_2 = c_2 + c_3 c_4$, т. е. операторы R и A' энергетически эквивалентны с постоянными, не зависящими от шага сетки h .

Напомним, что выше было получено неравенство

$$(Au - Av, u - v) \geq c_1 [\|u - v\|^2 + ((u - v)_{\bar{x}}^2, 1)_{\omega+}],$$

если выполнены условия (2), (3). Отсюда и из (14) следует, что при выполнении условий (2), (3) имеет место оценка

$$(Au - Av, u - v) \geq \gamma_1 (R(u - v), u - v), \quad \gamma_1 = c_1 > 0. \quad (16)$$

Покажем теперь, что для любых $u, v \in H$ верна оценка

$$(R^{-1}(Au - Av), Au - Av) \leq \gamma_2 (Au - Av, u - v), \quad (17)$$

где $\gamma_2 = c_2(1 + c_4)$, если выполнены условия

$$\sum_{\alpha=0}^1 [k_{\alpha}(x, p_0, p_1) - k_{\alpha}(x, q_0, q_1)]^2 \leq c_2 \sum_{\alpha=0}^1 [k_{\alpha}(x, p_0, p_1) - k_{\alpha}(x, q_0, q_1)] (p_{\alpha} - q_{\alpha}), \quad (18)$$

$$[\kappa_{\alpha}(p_0) - \kappa_{\alpha}(q_0)]^2 \leq c_2 [\kappa_{\alpha}(p_0) - \kappa_{\alpha}(q_0)] (p_0 - q_0).$$

Действительно, для доказательства (17) достаточно получить для любых $u, v, z \in H$ оценку

$$(Au - Av, z)^2 \leq \gamma_2 (Au - Av, u - v) (Rz, z). \quad (19)$$

Тогда, полагая здесь $z = R^{-1}(Au - Av)$, будем иметь (17).

Обозначим:

$$\begin{aligned} p_0 &= \bar{p}_0 = u_i, & q_0 &= \bar{q}_0 = v_i, & s_0 &= \bar{s}_0 = z_i, \\ p_1 &= u_{x, i}, & \bar{p}_1 &= u_{\bar{x}, i}, & q_1 &= v_{x, i}, & \bar{q}_1 &= v_{\bar{x}, i}, \\ s_1 &= z_{x, i}, & \bar{s}_1 &= z_{\bar{x}, i}. \end{aligned}$$

Используя определение оператора A и формулы суммирования по частям, получим

$$\begin{aligned}
 (Au - Av, z)^2 &= (\Lambda v - \Lambda u, z)^2 = \\
 &= \left\{ \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)], s_\alpha)_{\omega^-} + \right. \\
 &\quad + \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, \bar{p}_0, \bar{p}_1) - k_\alpha(x, \bar{q}_0, \bar{q}_1)], \bar{s}_\alpha)_{\omega^+} + \\
 &\quad \left. + [\varkappa_1(\bar{p}_0) - \varkappa_1(\bar{q}_0)] \bar{s}_0|_{i=N} + [\varkappa_0(p_0) - \varkappa_0(q_0)] s_0|_{i=0} \right\}^2.
 \end{aligned}$$

Используя неравенство Коши—Буняковского, последовательно найдем

$$\begin{aligned}
 (Au - Av, z)^2 &\leqslant \\
 &\leqslant \left\{ \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)]^2, 1)_{\omega^-}^{1/2} (s_\alpha^2, 1)_{\omega^-}^{1/2} + \right. \\
 &\quad + \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, \bar{p}_0, \bar{p}_1) - k_\alpha(x, \bar{q}_0, \bar{q}_1)]^2, 1)_{\omega^+}^{1/2} (\bar{s}_\alpha^2, 1)_{\omega^+}^{1/2} + \\
 &\quad \left. + [\varkappa_1(\bar{p}_0) - \varkappa_1(\bar{q}_0)] \bar{s}_0|_{i=N} + [\varkappa_0(p_0) - \varkappa_0(q_0)] s_0|_{i=0} \right\}^2 \leqslant \\
 &\leqslant \left\{ \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)]^2, 1)_{\omega^-} + \right. \\
 &\quad + \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, \bar{p}_0, \bar{p}_1) - k_\alpha(x, \bar{q}_0, \bar{q}_1)]^2, 1)_{\omega^+} + \\
 &\quad \left. + [\varkappa_1(\bar{p}_0) - \varkappa_1(\bar{q}_0)]^2_{i=N} + [\varkappa_0(p_0) - \varkappa_0(q_0)]^2_{i=0} \right\} \times \\
 &\quad \times \left\{ \frac{1}{2} \sum_{\alpha=0}^1 [(s_\alpha^2, 1)_{\omega^-} + (\bar{s}_\alpha^2, 1)_{\omega^+}] + \bar{s}_0^2|_{i=N} + s_0^2|_{i=0} \right\}.
 \end{aligned}$$

Учитывая, что верно равенство

$$\begin{aligned}
 (Au - Av, u - v) &= \\
 &= \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)](p_\alpha - q_\alpha), 1)_{\omega^-} + \\
 &\quad + \frac{1}{2} \sum_{\alpha=0}^1 ([k_\alpha(x, \bar{p}_0, \bar{p}_1) - k_\alpha(x, \bar{q}_0, \bar{q}_1)](\bar{p}_\alpha - \bar{q}_\alpha), 1)_{\omega^+} + \\
 &\quad + [\varkappa_1(\bar{p}_0) - \varkappa_1(\bar{q}_0)](\bar{p}_0 - \bar{q}_0)|_{i=N} + [\varkappa_0(p_0) - \varkappa_0(q_0)](p_0 - q_0)|_{i=0},
 \end{aligned}$$

а также что в силу (12), (14) и введенных обозначений

$$\begin{aligned} \frac{1}{2} \sum_{\alpha=0}^1 [(s_\alpha^2, 1)_\omega - + (\bar{s}_\alpha^2, 1)_\omega +] + \bar{s}_0^2|_{t=N} + s_0^2|_{t=0} = \\ = \frac{1}{2} \left[(z^2, 1)_\omega + + (z^2, 1)_\omega - + \left(z_x^2, 1 \right)_\omega + + (z_x^2, 1)_\omega - \right] + z_N^2 + z_0^2 = \\ = (z^2, 1) + \left(z_x^2, 1 \right)_\omega + + z_N^2 + z_0^2 \leqslant (1 + c_4)(Rz, z), \end{aligned}$$

получим оценку (19), если выполнены условия (18). Утверждение доказано.

2. Метод простой итерации. Рассмотрим теперь итерационные методы решения построенной нелинейной разностной схемы (4). Предположим сначала, что выполнены условия (2), (3) и (18).

Для решения уравнения (4) воспользуемся неявным методом простой итерации

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (20)$$

где $A = -\Lambda$, $B = R$ и оператор R определен выше. Из (20) следует, что для нахождения y_{k+1} при заданном y_k требуется решить линейное уравнение

$$By_{k+1} = \varphi, \quad \varphi = By_k - \tau(Ay_k - f)$$

или в развернутом виде

$$\begin{aligned} -y_{k+1}(i-1) + cy_{k+1}(i) - y_{k+1}(i+1) &= h^2 \varphi(i), \quad 1 \leq i \leq N-1, \\ cy_{k+1}(0) - 2y_{k+1}(1) &= h^2 \varphi(0), \quad i=0, \\ -2y_{k+1}(N-1) + cy_{k+1}(N) &= h^2 \varphi(N), \quad i=N, \end{aligned}$$

где $c = 2 + h^2$. Так как $c > 2$, то разностная краевая задача может быть решена методом монотонной прогонки с затратой $O(N)$ арифметических операций.

Осталось указать значение итерационного параметра τ и дать оценку для числа требуемых итераций. Так как выполнены условия (2), (3) и (18), то имеют место оценки (16) и (17), которые можно записать в виде

$$\begin{aligned} (Au - Av, u - v) &\geq \gamma_1(B(u - v), u - v), \quad \gamma_1 = c_1 > 0, \\ (B^{-1}(Au - Av), Au - Av) &\leq \gamma_2(Au - Av, u - v), \quad \gamma_2 = c_2(1 + c_4), \end{aligned} \quad (21)$$

где c_1 задано в (2), c_2 — в (18) и c_4 — в (12).

Так как оператор B является самосопряженным и положительно определенным, то сходимость метода (20) исследуем в энергетическом пространстве H_D , где $D = B$. Для указанного выбора оператора D неравенства (21) совпадают с неравенствами (4), (5). Поэтому для выбора итерационного параметра τ можно воспользоваться теоремой 1. Получим, что при $\tau = 1/\gamma_2 =$

$= 1/(c_2(1+c_4))$ итерационный метод (20) сходится в H_D , и для погрешности верна оценка $\|y_n - u\|_B \leq \rho^n \|y_0 - u\|_B$, $\rho = \sqrt{1-\xi}$, $\xi = \gamma_1/\gamma_2$ при любом начальном приближении y_0 .

Итак, если выполнены условия (2), (3), (18), то итерационный метод простой итерации (20) с указанным значением параметра τ позволяет получить решение нелинейной разностной схемы (4) с точностью ε за $n \geq n_0(\varepsilon)$ итераций, где

$$n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho} = \frac{2 \ln \varepsilon}{\ln \left(1 - \frac{c_1}{c_2(1+c_4)} \right)}.$$

Так как постоянные c_1 , c_2 и c_4 не зависят от шага сетки h , то число итераций $n_0(\varepsilon)$ зависит лишь от ε и не меняется с измельчением сетки.

Рассмотрим теперь итерационный метод (20) в предположении, что выполнены (6), (7) для производных $a_{\alpha\beta} = \partial k_\alpha / \partial p_\beta$ и $\sigma_\alpha = \partial \kappa_\alpha / \partial p_0$, а также условия симметрии (9). Тогда для производной Гато оператора A будут справедливы неравенства (15), которые в силу выбора $B = R$ можно записать в виде

$$\gamma_1(By, y) \leq (A'(v)y, y) \leq \gamma_2(By, y), \quad v, y \in H, \quad (22)$$

где $\gamma_1 = c_1$, $\gamma_2 = c_2 + c_3 c_4$, c_1 , c_2 и c_3 определены в (6), (7), а c_4 — в (12).

Пусть $D = B$. Тогда оператор $DB^{-1}A'(v) = A'(v)$ в силу леммы 4 будет самосопряжен в H , и, следовательно, выполнены условия теоремы 2, а неравенства (22) совпадают с неравенствами (14). Поэтому параметр τ в схеме (20) следует взять равным $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$. При этом для погрешности $y_n - u$ и для числа итераций будут верны оценки

$$\|y_n - u\|_B \leq \rho_0^n \|y_0 - u\|_B, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2} = \frac{c_1}{c_2 + c_3 c_4},$$

$$n \geq n_0(\varepsilon) = \ln \varepsilon / \ln \rho_0.$$

Здесь, так же как и для предыдущего метода, число итераций не зависит от шага сетки h . Для выбранного оператора B в силу первой разностной формулы Грина будем иметь следующие представления для нормы $\|z\|_B$:

$$\|z\|_B^2 = (z, z) + \left(z_x^2, 1 \right)_{\omega+}.$$

Мы рассмотрели методы решения нелинейной разностной схемы, аппроксимирующей квазилинейное одномерное уравнение на равномерной сетке. Не представляет труда перенести эти рассмотрения на случай произвольной неравномерной сетки, а также на разностные схемы, аппроксимирующие основные граничные задачи для квазилинейного эллиптического уравнения второго порядка в прямоугольнике.

3. Итерационные методы для разностных квазилинейных эллиптических уравнений в прямоугольнике. В прямоугольнике $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ с границей Γ требуется найти решение уравнения

$$\sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) - k_0 \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) = -\varphi(x), \quad x \in G, \quad (23)$$

удовлетворяющее краевым условиям третьего рода

$$\begin{aligned} k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) &= \kappa_{-\alpha}(x, u) - g_{-\alpha}(x), \quad x_\alpha = 0, \\ -k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) &= \kappa_{+\alpha}(x, u) - g_{+\alpha}(x), \quad x_\alpha = l_\alpha, \quad \alpha = 1, 2. \end{aligned} \quad (24)$$

Предположим, как и в одномерном случае, что выполнены следующие условия. Функции $k_\alpha(x, p)$ и $\kappa_{\pm\alpha}(p_0)$ непрерывны по $p = (p_0, p_1, p_2)$ и p_0 и, кроме того,

$$\begin{aligned} \sum_{\alpha=0}^2 [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha) &\geq c_1 \sum_{\alpha=0}^2 (p_\alpha - q_\alpha)^2, \quad c_1 > 0, \\ \sum_{\alpha=0}^2 [k_\alpha(x, p) - k_\alpha(x, q)]^2 &\leq c_2 \sum_{\alpha=0}^2 [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha), \\ [\kappa_{\pm\alpha}(p_0) - \kappa_{\pm\alpha}(q_0)]^2 &\leq c_2 [\kappa_{\pm\alpha}(p_0) - \kappa_{\pm\alpha}(q_0)](p_0 - q_0), \quad \alpha = 1, 2, \end{aligned}$$

где $c_1 > 0$ и $c_2 > 0$, $x \in \bar{G}$ и $|p|, |q| < \infty$.

Введем в области \bar{G} прямоугольную равномерную сетку

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}.$$

Простейшая разностная схема, соответствующая задаче (23), (24), имеет вид

$$\begin{aligned} \Lambda y &= -f, \quad x \in \bar{\omega}, \\ \Lambda = \Lambda_1 + \Lambda_2, \quad f &= \varphi + 2\varphi_1/h_1 + 2\varphi_2/h_2, \end{aligned} \quad (25)$$

где

$$\varphi_\alpha(x) = \begin{cases} g_{-\alpha}(x), & x_\alpha = 0, \\ 0, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ g_{+\alpha}(x), & x_\alpha = l_\alpha, 0 \leq x_{3-\alpha} \leq l_{3-\alpha}, \end{cases}$$

а операторы Λ_α , $\alpha = 1, 2$, определены формулами:

1) для $h_\beta \leq x_\beta \leq l_\beta - h_\beta$ имеем

$$\begin{aligned}\Lambda_\alpha y &= \frac{1}{2} \left\{ \left[k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) \right]_{x_\alpha} + \left[k_\alpha(x, y, y_{x_1}, y_{x_2}) \right]_{\bar{x}_\alpha} \right\} - \\ &\quad - \frac{1}{4} \left[k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_0(x, y, y_{x_1}, y_{x_2}) \right], \quad h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha; \\ \Lambda_\alpha y &= \frac{1}{h_\alpha} \left[k_\alpha^{+1\alpha}(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_\alpha(x, y, y_{x_1}, y_{x_2}) \right] - \\ &\quad - \frac{1}{2} k_0(x, y, y_{x_1}, y_{x_2}) - \frac{2}{h_\alpha} \kappa_{-\alpha}(x, y), \quad x_\alpha = 0; \\ \Lambda_\alpha y &= -\frac{1}{h_\alpha} [k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_\alpha^{-1\alpha}(x, y, y_{x_1}, y_{x_2})] - \\ &\quad - \frac{1}{2} k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - \frac{2}{h_\alpha} \kappa_{+\alpha}(x, y), \quad x_\alpha = l_\alpha;\end{aligned}$$

2) для $x_\beta = 0$ имеем

$$\Lambda_\alpha y = [k_\alpha(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_\alpha} - \frac{1}{2} k_0(x, y, y_{x_1}, y_{x_2})$$

при $h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha$;

$$\Lambda_\alpha y = \frac{2}{h_\alpha} k_\alpha(x, y, y_{x_1}, y_{x_2}) - k_0(x, y, y_{x_1}, y_{x_2}) - \frac{2}{h_\alpha} \kappa_{-\alpha}(x, y)$$

при $x_\alpha = 0$;

$$\Lambda_\alpha y = -\frac{2}{h_\alpha} k_\alpha^{-1\alpha}(x, y, y_{x_1}, y_{x_2}) - \frac{2}{h_\alpha} \kappa_{+\alpha}(x, y), \quad x_\alpha = l_\alpha;$$

3) для $x_\beta = l_\beta$ имеем

$$\Lambda_\alpha y = [k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_\alpha} - \frac{1}{2} k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})$$

при $h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha$;

$$\Lambda_\alpha y = \frac{2}{h_\alpha} k_\alpha^{+1\alpha}(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - \frac{2}{h_\alpha} \kappa_{-\alpha}(x, y), \quad x_\alpha = 0;$$

$$\Lambda_\alpha y = -\frac{2}{h_\alpha} k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - \frac{2}{h_\alpha} \kappa_{+\alpha}(x, y)$$

при $x_\alpha = l_\alpha$.

Здесь $\beta = 3 - \alpha$, $\alpha = 1, 2$ и использованы обозначения

$$\begin{aligned}k_1^{+11}(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})|_{x_{ij}} &= \\ &= k_1(x_{i+1, j}, y(i+1, j), y_{\bar{x}_1}(i+1, j), y_{\bar{x}_2}(i+1, j)),\end{aligned}$$

а также аналогичные обозначения для k_1^{-11} и $k_2^{\pm 12}$.

В пространстве H сеточных функций, заданных на $\bar{\omega}$, определим скалярное произведение

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \bar{h}_1(i) \bar{h}_2(j) u(i, j) v(i, j),$$

$$\bar{h}_\alpha(k) = \begin{cases} h_\alpha, & 1 \leq k \leq N_\alpha - 1, \\ 0,5h_\alpha, & k = 0, N_\alpha \end{cases}$$

и операторы $A_\alpha = -\Lambda_\alpha$, $\alpha = 1, 2$, $A = A_1 + A_2$, $R = R_1 + R_2$, где

$$R_\alpha y = \begin{cases} -\frac{2}{h_\alpha} y_{x_\alpha} + \frac{1}{2} y, & x_\alpha = 0, \\ -y_{\bar{x}_\alpha x_\alpha} + \frac{1}{2} y, & h_\alpha \leq x_\alpha \leq l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} y_{\bar{x}_\alpha} + \frac{1}{2} y, & x_\alpha = l_\alpha, \alpha = 1, 2, \end{cases}$$

и $0 \leq x_\beta \leq l_\beta$. Тогда разностная схема (25) запишется в виде операторного уравнения

$$Au = f \quad (26)$$

с нелинейным оператором A .

Используя сделанные выше предположения относительно коэффициентов $k_\alpha(x, p)$ и $\kappa_{\pm\alpha}(p_0)$, как и в одномерном случае, получим, что имеют место неравенства (16) и (17), где c_4 — постоянная из неравенства

$$\begin{aligned} \sum_{j=0}^{N_2} \bar{h}_2(j) [y^2(0, j) + y^2(N_1, j)] + \sum_{i=0}^{N_1} \bar{h}_1(i) [y^2(i, 0) + y^2(i, N_2)] &\leq \\ &\leq c_4 \left[\sum_{i=0}^{N_1} \sum_{j=0}^{N_2} y^2(i, j) \bar{h}_1(i) \bar{h}_2(j) + \sum_{j=0}^{N_2} \sum_{i=0}^{N_1} h_1 \bar{h}_2(j) y_{\bar{x}_1}^2(i, j) + \right. \\ &\quad \left. + \sum_{i=0}^{N_1} \sum_{j=1}^{N_2} \bar{h}_1(i) h_2 y_{x_2}^2(i, j) \right]. \quad (27) \end{aligned}$$

Покажем, что

$$c_4 = V\bar{2}(16 + l^2)/(lV\sqrt{32 + l^2}), \quad l = \min(l_1, l_2). \quad (28)$$

Действительно, из неравенства (36) леммы 15 главы V при $\varepsilon = V\bar{2}$ получим

$$\begin{aligned} y^2(0, j) + y^2(N_1, j) &\leq \\ &\leq \frac{(16 + l_1^2)}{l_1 V\sqrt{32 + l_1^2}} \left[\sum_{i=1}^{N_1} h_1 y_{\bar{x}_1}^2(i, j) + \frac{1}{2} \sum_{i=0}^{N_1} \bar{h}_1(i) y^2(i, j) \right]. \end{aligned}$$

Отметим, что если здесь заменить l_1 на l , то неравенство лишь усилится. Умножим теперь левую и правую части полученного

неравенства на $\tilde{h}_2(j)$ и просуммируем по j от 0 до N_2 . Будем иметь

$$\begin{aligned} \sum_{j=0}^{N_2} \tilde{h}_2(j) [y^2(0, j) + y^2(N_1, j)] &\leq \\ &\leq c_4 \left[\sum_{j=0}^{N_2} \sum_{i=1}^{N_1} h_1 \tilde{h}_2(j) y_{x_1}^2(i, j) + \frac{1}{2} \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} y^2(i, j) \tilde{h}_1(i) \tilde{h}_2(j) \right], \quad (29) \end{aligned}$$

где c_4 определено в (28). Аналогично найдем

$$\begin{aligned} \sum_{i=0}^{N_1} \tilde{h}_1(i) [y^2(i, 0) + y^2(i, N_2)] &\leq \\ &\leq c_4 \left[\sum_{i=0}^{N_1} \sum_{j=1}^{N_2} \tilde{h}_1(i) h_2 y_{x_2}^2(i, j) + \frac{1}{2} \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} y^2(i, j) \tilde{h}_1(i) \tilde{h}_2(j) \right]. \quad (30) \end{aligned}$$

Складывая (29), (30), получим неравенство (27).

Для решения уравнения (26) можно воспользоваться неявным методом простой итерации (20), где $B=R$ и $\tau=1/\gamma_2=1/(c_2(1+c_4))$. Тогда в силу теоремы 1 итерационный метод (20) будет сходиться в H_B и для погрешности будет верна оценка

$$\|y_n - u\|_B \leq \rho^n \|y_0 - u\|_B, \quad \rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2 = c_1/(c_2(1+c_4)).$$

Следовательно, число итераций $n_0(\varepsilon)$, которое следует выполнить для достижения относительной точности ε , не будет зависеть от числа узлов сетки $\bar{\omega}$.

Для нахождения y_{k+1} имеем задачу

$$Ry_{k+1} = \varphi, \quad \varphi = Ry_k - \tau(Ay_k - f).$$

Так как оператор R соответствует второй краевой задаче для разностного уравнения с постоянными коэффициентами, то указанная задача может быть решена прямыми методами, описанными в главах III и IV, с затратой $O(N^2 \log_2 N)$ арифметических действий ($N_1 = N_2 = N = 2^n$). Если функции $k_\alpha(x, p)$ и $x_{\pm\alpha}(x, p_0)$ дифференцируемы, то оператор A имеет производную Гато, которая является самосопряженным в H оператором, если выполнены условия

$$a_{\alpha\beta}(x, p) = a_{\beta\alpha}(x, p), \quad \alpha, \beta = 0, 1, 2, \quad (31)$$

где $a_{\alpha\beta}(x, p) = \frac{\partial k_\alpha(x, p)}{\partial p_\beta}$. Можно показать, что если кроме (31) выполнены условия

$$\begin{aligned} c_1 \sum_{\alpha=0}^2 \xi_\alpha^2 &\leq \sum_{\alpha, \beta=0}^2 a_{\alpha\beta}(x, p) \xi_\alpha \xi_\beta \leq c_2 \sum_{\alpha=0}^2 \xi_\alpha^2, \quad c_1 > 0, \\ 0 &\leq \frac{\partial x_{\pm\alpha}(x, p_0)}{\partial p_0} \leq c_3, \quad \alpha = 1, 2, \end{aligned}$$

то верны неравенства (15), где $\gamma_1 = c_1$, $\gamma_2 = c_2 + c_3 c_4$, а c_4 определено в (28). Тогда в итерационном методе (20) с $B = R$ параметр τ можно выбрать равным $\tau_0 = 2/(\gamma_1 + \gamma_2)$. В силу теоремы 4 для погрешности будет верна оценка

$$\|y_n - u\|_B \leq \rho_0^n \|y_0 - u\|_B, \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Пусть теперь требуется найти решение первой краевой задачи в прямоугольнике \bar{G}

$$\sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) - k_0 \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) = -\varphi(x), \\ x \in G, \quad (32) \\ u(x) = 0, \quad x \in \Gamma.$$

Предположим, что функции $k_\alpha(x, p)$ непрерывны по $p = (p_0, p_1, p_2)$ и выполнены условия

$$\begin{aligned} \sum_{\alpha=1}^2 [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha) &\geq c_1 \sum_{\alpha=1}^2 (p_\alpha - q_\alpha)^2, \quad c_1 > 0, \\ [k_0(x, p) - k_0(x, q)](p_0 - q_0) &\geq 0, \\ \sum_{\alpha=0}^2 [k_\alpha(x, p) - k_\alpha(x, q)]^2 &\leq c_2 \sum_{\alpha=0}^2 [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha), \end{aligned} \quad (33)$$

где $c_1 > 0$, $c_2 > 0$ для $x \in \bar{G}$ и $|p|, |q| < \infty$.

Задача (32) на прямоугольной равномерной сетке $\bar{\omega} = \omega \cup \gamma$, введенной ранее, поставим в соответствие разностную схему

$$\Lambda y = -f, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma, \quad (34)$$

где $f = \varphi$, а разностный оператор Λ определен следующим образом:

$$\begin{aligned} \Lambda y = \Lambda^- y &= \frac{1}{2} \{ [k_1(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_1} + [k_1(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_1} + \\ &+ [k_2(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_2} + [k_2(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_2} - \\ &- k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - k_0(x, y, y_{x_1}, y_{x_2}) \}. \end{aligned}$$

Приведем еще две возможные аппроксимации:

$$\begin{aligned} \Lambda y = \Lambda^+ y &= \frac{1}{2} \{ [k_1(x, y, y_{\bar{x}_1}, y_{x_2})]_{x_1} + [k_1(x, y, y_{x_1}, y_{\bar{x}_2})]_{\bar{x}_1} + \\ &+ [k_2(x, y, y_{x_1}, y_{\bar{x}_2})]_{x_2} + [k_2(x, y, y_{\bar{x}_1}, y_{x_2})]_{\bar{x}_2} - \\ &- k_0(x, y, y_{\bar{x}_1}, y_{x_2}) - k_0(x, y, y_{x_1}, y_{\bar{x}_2}) \} \end{aligned}$$

и $\Lambda = \frac{1}{2} (\Lambda^- + \Lambda^+)$.

Для данного примера H — пространства сеточных функций, заданных на ω , скалярное произведение в котором определено формулой

$$(u, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2+1} h_1 h_2 u(i, j) v(i, j).$$

Если в уравнения схемы (34) подставить $y|_{\gamma} = 0$, то получим разностную схему $\bar{\Delta}y = -f$. Определяя оператор $A = -\bar{\Delta}$, запишем полученную схему в виде операторного уравнения (26) в пространстве H .

Используя условия (33), получим для всех трех аппроксимаций, что соответствующий оператор A удовлетворяет неравенствам (16), (17):

$$(Au - Av, u - v) \geq \gamma_1 (R(u - v), u - v), \quad \gamma_1 = c_1 > 0,$$

$$(R^{-1}(Au - Av), Au - Av) \leq \gamma_2 (Au - Av, u - v), \quad \gamma_2 = c_2(1 + c_4),$$

где

$$c_4 = \frac{1}{\delta}, \quad \delta = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2} \geq \frac{8}{l_1^2} + \frac{8}{l_2^2},$$

а оператор R соответствует разностному оператору Лапласа $Ry = -\mathcal{R}y$, $y(x) = \dot{y}(x)$ для $x \in \omega$ и $\dot{y}(x) = 0$ для $x \in \gamma$, $\mathcal{R}u = u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2}$.

Для решения уравнения (26) воспользуемся методом простой итерации (20) с $B = R$ и $\tau = 1/\gamma_2$. В силу теоремы 1 будем иметь оценку

$$\|y_n - u\|_B \leq \rho^n \|y_0 - u\|_B, \quad \rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1 / \gamma_2.$$

Как и раньше, для решения уравнения $Ry_{k+1} = Ry_k - \tau(Ay_k - f)$ можно использовать прямые методы полной редукции или разделения переменных, предложенные в главах III и IV.

4. Итерационные методы для слабонелинейных уравнений. В прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ рассмотрим слабонелинейное эллиптическое уравнение второго порядка

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} - k_0 \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) = 0, \quad x \in G \quad (35)$$

с краевыми условиями первого рода

$$u(x) = 0, \quad x \in \Gamma. \quad (36)$$

Слабонелинейность уравнения (35) означает, что функция $k_0(x, p_0, p_1, p_2)$ определена при $x \in \bar{G}$ и $|p_0|, |p_1|, |p_2| < \infty$ и непрерывна по x при фиксированных p_0, p_1, p_2 , а также существуют

производные от функции $k_0(x, p_0, p_1, p_2)$ по p_0 , p_1 и p_2 , которые удовлетворяют условиям

$$c_2 \geq \frac{\partial k_0}{\partial p_0} \geq 0, \quad \left| \frac{\partial k_0}{\partial p_\alpha} \right| \leq M, \quad \alpha = 1, 2. \quad (37)$$

На прямоугольной равномерной сетке $\bar{\omega} = \omega \cup \gamma$, введенной ранее, разностная схема, соответствующая задаче (35), (36), имеет вид

$$\begin{aligned} \Lambda y &= 0, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma, \\ \Lambda y &= \mathcal{R}y - \frac{1}{2} [k_0(x, y, y_{x_1}, y_{x_2}) + k_0(x, y, y_{x_1}, y_{x_2})], \end{aligned} \quad (38)$$

где $\mathcal{R}y = y_{x_1 x_1} + y_{x_2 x_2}$ — разностный оператор Лапласа.

Определим теперь разностный оператор $\Lambda'(v)$, зависящий от v :

$$\begin{aligned} \Lambda'(v)y &= \mathcal{R}y - \frac{1}{2} [a_{01}(x, v, v_{x_1}, v_{x_2})y_{x_1} + \\ &\quad + a_{01}(x, v, v_{x_1}, v_{x_2})y_{x_1} + a_{02}(x, v, v_{x_1}, v_{x_2})y_{x_2} + \\ &\quad + a_{02}(x, v, v_{x_1}, v_{x_2})y_{x_2} + (a_{00}(x, v, v_{x_1}, v_{x_2}) + a_{00}(x, v, v_{x_1}, v_{x_2}))y], \end{aligned}$$

где

$$a_{0\alpha}(x, p_0, p_1, p_2) = \frac{\partial k_0(x, p_0, p_1, p_2)}{\partial p_\alpha}, \quad \alpha = 0, 1, 2.$$

В пространстве H сеточных функций, заданных на ω , определим операторы:

$$Ay = -\Lambda \dot{y}, \quad Ry = -\mathcal{R} \dot{y}, \quad A'(v)y = -\Lambda'(\dot{v}) \dot{y},$$

где

$$y(x) = \dot{y}(x), \quad v(x) = \dot{v}(x) \quad \text{для } x \in \omega$$

и

$$\dot{y}(x) = 0, \quad \dot{v}(x) = 0 \quad \text{для } x \in \gamma.$$

Оператор $A'(v)$ — производная Гато оператора A . Используя эти обозначения, разностную схему запишем в виде операторного уравнения (26).

Если $k_0(x, p_0, p_1, p_2)$ не зависит от p_1 и p_2 , т. е.

$$k_0(x, p_0, p_1, p_2) = k_0(x, p_0),$$

то

$$a_{01}(x, p) = a_{02}(x, p) = 0.$$

В этом случае оператор $A'(v)$ самосопряжен в H .

Используя оценку снизу для разностного оператора $(-\mathcal{R})$

$$(-\mathcal{R} \dot{y}, \dot{y}) = -(\dot{y}_{x_1 x_1} + \dot{y}_{x_2 x_2}, \dot{y}) \geq \delta(\dot{y}, \dot{y}),$$

где

$$\delta = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2} \geq \frac{8}{l_1^2} + \frac{8}{l_2^2},$$

условия (37) при $M=0$ и равенства

$$-(\Lambda'(\dot{v})\dot{y}, \dot{y}) = -(\mathcal{R}\dot{y}, \dot{y}) + (a_{00}(x, \dot{v})\dot{y}, \dot{y}),$$

получим

$$\gamma_1(Ry, y) \leq (A'(v)y, y) \leq \gamma_2(Ry, y),$$

где

$$\gamma_1 = 1, \quad \gamma_2 = 1 + c_2/\delta.$$

Следовательно, если для рассматриваемого «самосопряженного» случая использовать итерационный метод (20) с $B=D=R$ и $\tau=\tau_0=2/(\gamma_1+\gamma_2)$, то в силу теоремы 2 для погрешности будет верна оценка

$$\|y_n - u\|_B \leq \rho_0^n \|y_0 - u\|_B, \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Оператор R в схеме (20) можно обращать одним из прямых методов.

Г Л А В А XIV

**ПРИМЕРЫ РЕШЕНИЯ СЕТОЧНЫХ
ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ**

В § 1 рассмотрены некоторые способы построения неявных итерационных схем, в частности, основанных на выделении регуляризатора. Методом решения систем эллиптических уравнений посвящен § 2. Здесь рассмотрено применение общей теории для решения некоторых задач теории упругости.

§ 1. Способы построения неявных итерационных схем

1. Принцип регуляризации в общей теории итерационных методов. В главах VI—VIII, XII, XIII была изложена общая теория итерационных методов, используемых для решения операторного уравнения

$$Au = f. \quad (1)$$

В общей теории итерационных методов мы не используем конкретную структуру операторов итерационной схемы—теория использует минимум информации общего функционального характера относительно операторов. Это позволяет указать (при фиксированных операторах схемы) общие принципы конструирования оптимальных итерационных методов. Например, если операторы A и B двухслойной итерационной схемы

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H \quad (2)$$

удовлетворяют условиям

$$B = B^* > 0, \quad A = A^* > 0, \quad (3)$$

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \quad (4)$$

то набор чебышевских итерационных параметров τ_k :

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \quad 1 \leq k \leq n,$$

где

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}$$

является наилучшим.

Какими требованиями следует руководствоваться при выборе оператора B ? В § 3 главы V отмечалось, что выбор B должен быть подчинен двум требованиям: 1) обеспечению наиболее быстрой сходимости метода; 2) экономичности обращения этого оператора.

Для приведенного выше примера первое требование выполнено, если энергия оператора B будет близка к энергии оператора A , т. е. в неравенствах (4) близки γ_1 и γ_2 . Чтобы удовлетворить второму требованию, нужно из класса операторов B , близких по энергии к оператору A , выбрать наиболее легко обратимый.

Как конструировать легко обратимые операторы? Очевидно, что если B^1, B^2, \dots, B^p — легко обратимые операторы, то оператор $B = B^1 B^2 \dots B^p$, являющийся их произведением, также легко обратим.

Отметим, что, в отличие от множителей, сам оператор B может иметь сложную структуру. Например, пусть $B^\alpha = E + \omega R_\alpha$, $\alpha = 1, 2$, где R_α — оператор, соответствующий разностному оператору $(-\mathcal{R}_\alpha)$: $\mathcal{R}_\alpha y = y_{\tilde{x}_\alpha x_\alpha}$, $\alpha = 1, 2$. Оператору B^α соответствует трехточечный разностный оператор, который обращается методом прогонки с затратой числа арифметических действий, пропорционального числу неизвестных в задаче. Оператор $B = B^1 B^2$ имеет девятиточечный шаблон и ему соответствует разностный оператор \mathcal{B} :

$$\mathcal{B}y = y - \omega \sum_{\alpha=1}^2 y_{\tilde{x}_\alpha x_\alpha} + \omega^2 y_{\tilde{x}_1 x_1 \tilde{x}_2 x_2}.$$

Усложнение структуры оператора B позволяет увеличить отношение $\xi = \gamma_1/\gamma_2$, что приводит к увеличению скорости сходимости итерационного метода.

При построении оператора B можно исходить из некоторого оператора $R = R^* > 0$ (регуляризатора), энергетически эквивалентного A и B :

$$c_1 R \leq A \leq c_2 R, \quad c_2 \geq c_1 > 0, \quad (5)$$

$$\dot{\gamma}_1 B \leq R \leq \dot{\gamma}_2 B, \quad \dot{\gamma}_2 \geq \dot{\gamma}_1 > 0. \quad (6)$$

Тогда справедливы неравенства (4) с постоянными $\gamma_1 = c_1 \dot{\gamma}_1$, $\gamma_2 = c_2 \dot{\gamma}_2$, причем

$$\xi = \gamma_1/\gamma_2 = (c_1/c_2) \dot{\xi}, \quad \dot{\xi} = \dot{\gamma}_1/\dot{\gamma}_2.$$

В чем состоит смысл введения регуляризатора R ? Для сеточных эллиптических краевых задач оператор R обычно выбирается так, чтобы постоянные c_1 и c_2 в неравенствах (5) не зависели от параметров сетки (от числа узлов сетки). Например,

если оператор A соответствует разностному оператору с переменными коэффициентами

$$\Lambda y = (a_1 y_{\bar{x}_1})_{x_1} + (a_2 y_{\bar{x}_2})_{x_2}, \quad 0 < c_1 \leq a_\alpha \leq c_2,$$

заданному на равномерной сетке $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leq i \leq N_1, 0 \leq j \leq N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$, введенной в прямоугольнике $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$, так что $Ay = -\Lambda \dot{y}$, где $y(x) = \dot{y}(x)$ для $x \in \omega$ и $\dot{y}(x) = 0$ для $x \in \gamma$, то в качестве R можно взять оператор, соответствующий разностному оператору Лапласа $\mathcal{R}y = (\mathcal{R}_1 + \mathcal{R}_2)y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$, $Ry = -\mathcal{R}\dot{y}$, где операторы \mathcal{R}_α определены выше.

Пользуясь разностными формулами Грина, легко показать (см. п. 8 § 2 гл. V), что операторы A и B самосопряжены в H и выполнены неравенства (5). Здесь H — пространство сеточных функций, заданных на ω , скалярное произведение в котором определяется формулой $(u, v) = \sum_{x \in \omega} u(x)v(x)h_1h_2$.

Пусть теперь оператор A соответствует разностному эллиптическому оператору, содержащему смешанные производные

$$\Lambda y = \sum_{\alpha, \beta=1}^2 0,5 \left[(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha} \right],$$

и выполнены условия сильной эллиптичности:

$$c_1 \sum_{\alpha=1}^2 \xi_\alpha^2 \leq \sum_{\alpha, \beta=1}^2 k_{\alpha\beta}(x) \xi_\alpha \xi_\beta \leq c_2 \sum_{\alpha=1}^2 \xi_\alpha^2, \quad c_1 > 0.$$

Возьмем в качестве регуляризатора определенный выше оператор R . В п. 8 § 2 гл. V было показано, что для рассматриваемых операторов A и R выполнены неравенства (5).

Приведем еще один пример. Пусть оператор A соответствует разностному оператору Лапласа повышенного порядка точности

$$\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 x_1 \bar{x}_2 x_2}.$$

Покажем, что если в качестве оператора R выбрать указанный выше оператор, то верны неравенства (5) с постоянными $c_1 = 2/3$, $c_2 = 1$.

Действительно, используя первую разностную формулу Грина и равенство $y_{\bar{x}_1 x_1 \bar{x}_2 x_2} = y_{\bar{x}_1 \bar{x}_2 x_1 x_2}$, справедливое для сеточных функций, заданных на прямоугольной сетке $\bar{\omega}$, получим

$$\begin{aligned} -(\Lambda \dot{y}, \dot{y}) &= \left(\dot{y}_{\bar{x}_1}^2, 1 \right)_1 + \left(\dot{y}_{\bar{x}_2}^2, 1 \right)_2 - \frac{h_1^2 + h_2^2}{12} \left(\dot{y}_{\bar{x}_1 \bar{x}_1}^2, 1 \right)_{12}, \\ -(\mathcal{R}\dot{y}, \dot{y}) &= \left(\dot{y}_{\bar{x}_1}^2, 1 \right)_1 + \left(\dot{y}_{\bar{x}_2}^2, 1 \right)_2. \end{aligned} \quad (7)$$

Здесь обозначено

$$(u, v)_1 = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} u(i, j) v(i, j) h_1 h_2,$$

$$(u, v)_2 = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2} u(i, j) v(i, j) h_1 h_2,$$

$$(u, v)_{12} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u(i, j) v(i, j) h_1 h_2.$$

Из (7) следует оценка $A \leq R$, т. е. в (5) $c_2 = 1$. Далее, учитывая, что $\dot{y}_{\bar{x}_2}(x) = 0$ при $x_1 = 0$, l_1 и $\dot{y}_{\bar{x}_1} = 0$ при $x_2 = 0$, l_2 , из леммы 12 главы V получим оценку

$$\left(\dot{y}_{\bar{x}_1 \bar{x}_2}^2, 1 \right)_{12} \leq \frac{4}{h_2^2} \left(\dot{y}_{\bar{x}_1}^2, 1 \right)_1 \quad (8)$$

и аналогичным образом

$$\left(\dot{y}_{\bar{x}_1 \bar{x}_2}^2, 1 \right)_{12} = \left(\dot{y}_{\bar{x}_2 \bar{x}_1}^2, 1 \right)_{12} \leq \frac{4}{h_1^2} \left(\dot{y}_{\bar{x}_2}^2, 1 \right)_2. \quad (9)$$

Умножая (8) на $h_2^2/12$, а (9)—на $h_1^2/12$ и складывая полученные неравенства, будем иметь

$$\frac{h_1^2 + h_2^2}{12} \left(\dot{y}_{\bar{x}_1 \bar{x}_2}^2, 1 \right)_{12} \leq \frac{1}{3} \left[\left(\dot{y}_{\bar{x}_1}^2, 1 \right)_1 + \left(\dot{y}_{\bar{x}_2}^2, 1 \right)_2 \right].$$

Отсюда и из (7) следует оценка $A \geq 2R/3$. Утверждение доказано.

Рассмотренные примеры показывают, что для различных операторов A в качестве регуляризатора может быть выбран один и тот же оператор R . Поэтому задача построения оператора B для неявной итерационной схемы упрощается. Оператор B строится из условия близости по энергии к регуляризатору R . Класс регуляризаторов существенно более узкий, чем класс, содержащий операторы A . Если оператор B уже выбран и, следовательно, постоянные γ_1 и γ_2 в неравенствах (6) найдены, то для каждого конкретного оператора A останется найти лишь постоянные c_1 и c_2 в неравенствах (5).

Основная трудность при использовании регуляризатора переносится на получение оценок для γ_1 и γ_2 . Наиболее часто оператор B имеет факторизованный вид, причем множители зависят от некоторых итерационных параметров. Тем самым задается семейство операторов B определенной структуры, характеризуемое указанными параметрами. Эти параметры следует выбирать из условия максимума ξ . Некоторые примеры, в которых изучаются факторизованные операторы, будут рассмотрены в следующем пункте. Здесь же отметим, что в качестве оператора B иногда можно взять регуляризатор R ($\gamma_1 = \gamma_2 = 1$).

2. Итерационные схемы с факторизованным оператором. В п. 1 принцип регуляризации был проиллюстрирован на примере самосопряженного оператора A . В этом случае неравенства (4) являются следствием неравенств (5) и (6).

В главе VI было показано, что если оператор A несамосопряжен в H , а энергетическое пространство H_D порождается самосопряженным положительно определенным оператором D , где D есть либо B , либо $A^*B^{-1}A$, то неравенства (4) следует заменить неравенствами

$$\gamma_1(Bx, x) \leq (Ax, x), \quad (B^{-1}Ax, Ax) \leq \gamma_2(Ax, x), \quad \gamma_1 > 0 \quad (10)$$

или неравенствами

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad (B^{-1}A_1x, A_1x) \leq \gamma_3^2(Bx, x), \quad \gamma_1 > 0, \quad (11)$$

где $A_1 = 0,5(A - A^*)$ — несамосопряженная часть оператора A . Пусть оператор $B = B^* > 0$ построен, исходя из регуляризатора R , и выполнены неравенства (6). Тогда, если оператор R удовлетворяет условиям

$$c_1(Rx, x) \leq (Ax, x), \quad (R^{-1}Ax, Ax) \leq c_2(Ax, x), \quad c_1 > 0, \quad (10')$$

то имеют место неравенства (10) с постоянными $\gamma_1 = c_1 \dot{\gamma}_1$, $\gamma_2 = c_2 \dot{\gamma}_2$.

Действительно, из леммы 9 главы V и неравенства (6) следует, что верны неравенства $\dot{\gamma}_1 R^{-1} \leq B^{-1} \leq \dot{\gamma}_2 \cdot R^{-1}$. Отсюда получим

$$(B^{-1}Ax, Ax) \leq \dot{\gamma}_2 (R^{-1}Ax, Ax) \leq c_2 \dot{\gamma}_2 (Ax, x).$$

Аналогично доказывается, что если оператор R удовлетворяет условиям

$$c_1 R \leq A \leq c_2 R, \quad (R^{-1}A_1x, A_1x) \leq c_3^2(Rx, x), \quad c_1 > 0, \quad (11')$$

то верны неравенства (11) с постоянными $\gamma_1 = c_1 \dot{\gamma}_1$, $\gamma_2 = c_2 \dot{\gamma}_2$, $\gamma_3 = c_3 \dot{\gamma}_2$.

Таким образом, и в случае несамосопряженного оператора A необходимо уметь получать оценки для $\dot{\gamma}_1$ и $\dot{\gamma}_2$, входящих в неравенства (6).

Займемся теперь получением неравенств (6) для самосопряженных операторов R и B . Рассмотрим два случая:

1) Оператор R представлен в виде суммы $R = R_1 + R_2$ сопряженных друг другу операторов R_1 и R_2 :

$$R_2 = R_1^*, \quad (12)$$

так что $(R_1x, x) = (R_2x, x) = 0,5(Rx, x)$, $x \in H$, а оператор B имеет вид

$$B = (E + \omega R_1)(E + \omega R_2), \quad (13)$$

где $\omega > 0$ — параметр.

2) Оператор R представим в виде суммы $R=R_1+R_2+\dots+R_p$, $p \geq 2$, самосопряженных попарно перестановочных операторов R_α , $\alpha=1, 2, \dots, p$, так что

$$R_\alpha = R_\alpha^*, \quad R_\alpha R_\beta = R_\beta R_\alpha, \quad \alpha, \beta = 1, 2, \dots, p, \quad (14)$$

а оператор B факторизован и имеет вид

$$B = \prod_{\alpha=1}^p (E + \omega R_\alpha), \quad (15)$$

где $\omega > 0$ — параметр.

В каждом из случаев оператор B является самосопряженным в H . Особо подчеркнем универсальность выбора оператора B в виде (13), где операторы R_1 и R_2 удовлетворяют условию (12).

Наша задача состоит в получении оценок для $\dot{\gamma}_1$ и $\dot{\gamma}_2$, содержащихся в (6), и выборе итерационного параметра ω из условия максимума отношения $\xi = \dot{\gamma}_1 / \dot{\gamma}_2$.

Каждый случай исследуем отдельно. Первый случай был подробно изучен в главе X, посвященной попеременно-треугольному методу. Поэтому ограничимся здесь лишь формулировкой результатов.

Теорема 1. Пусть выполнены условия (12) и в неравенствах

$$R \geq \delta E, \quad (R_2 x, R_2 x) \leq \frac{\Delta}{4} (Rx, x), \quad \delta > 0 \quad (16)$$

заданы постоянные δ и Δ . Тогда при оптимальном значении параметра $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$ оператор B , определенный равенством (13), удовлетворяет неравенствам (6) с постоянными

$$\dot{\gamma}_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \dot{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}.$$

Отметим, что можно рассмотреть более общий, нежели (13), вид оператора B , а именно:

$$B = (\mathcal{D} + \omega R_1) \mathcal{D}^{-1} (\mathcal{D} + \omega R_2),$$

где $\mathcal{D} = \mathcal{D}^* > 0$. Из леммы 1 главы X следует, что теорема 1 остается справедливой, необходимо лишь заменить (16) следующими неравенствами:

$$R \geq \delta \mathcal{D}, \quad (\mathcal{D}^{-1} R_2 x, R_2 x) \leq \frac{\Delta}{4} (Rx, x).$$

Здесь оператор \mathcal{D} играет роль дополнительного итерационного параметра.

Оператор B легко обратим, например, в случае, когда оператору R_1 соответствует нижняя треугольная матрица, R_2 — верхняя треугольная матрица, \mathcal{D} — диагональная матрица. Если оператор R соответствует разностному эллиптическому оператору, то указанные треугольные матрицы в каждой строке будут иметь конечное, не зависящее от числа узлов сетки число нену-

левых элементов. Поэтому обращение каждого множителя, входящего в оператор B , может быть осуществлено с затратой числа действий, пропорционального числу неизвестных в задаче.

Рассмотрим теперь второй случай.

Теорема 2. Пусть оператор B имеет вид (15), выполнены условия (14) и заданы границы операторов R_α :

$$\delta_\alpha E \leq R_\alpha \leq \Delta_\alpha E, \quad \delta_\alpha > 0, \quad \alpha = 1, 2, \dots, p.$$

Тогда при оптимальном значении параметра ω

$$\omega = \omega_0 = \frac{1}{\Delta} \frac{1-\eta^{1/p}}{\eta^{1/p}-\eta}$$

оператор B удовлетворяет неравенствам (6) с постоянными

$$\dot{\gamma}_1 = \frac{p\Delta}{(1+\omega_0\Delta)^p}, \quad \dot{\gamma}_2 = \dot{\gamma}_1 \frac{p-k(1-\eta)}{p\eta^{k/p}}, \quad \eta = \frac{\delta}{\Delta},$$

где

$$\delta = \min_\alpha \delta_\alpha, \quad \Delta = \max_\alpha \Delta_\alpha, \quad k = \left[\frac{p}{1-\eta} - \frac{\eta^{1/p}}{1-\eta^{1/p}} \right],$$

[a]—целая часть числа a .

Ввиду громоздкости доказательства мы его не приводим. Отметим лишь, что в силу условий (14) оператор B перестановочен с операторами R_α , $\alpha = 1, 2, \dots, p$, и поэтому

$$\dot{\gamma}_1 = \min_{\delta \leq x_\alpha \leq \Delta} \frac{x_1 + x_2 + \dots + x_p}{\prod_{\alpha=1}^p (1+\omega x_\alpha)}, \quad \dot{\gamma}_2 = \max_{\delta \leq x_\alpha \leq \Delta} \frac{x_1 + x_2 + \dots + x_p}{\prod_{\alpha=1}^p (1+\omega x_\alpha)}.$$

Отметим частные случаи теоремы 2. Если $p=2$, то

$$k=1, \quad \omega_0 = \frac{1}{V\delta\Delta}, \quad \dot{\gamma}_1 = \frac{2\delta}{(1+V\eta)^2}, \quad \dot{\gamma}_2 = \frac{\delta}{V\eta} \frac{1+\eta}{(1+V\eta)^2}.$$

Если $p=3$, то

$$k=2, \quad \omega_0 = \frac{1}{\sqrt[3]{\delta\Delta} (\sqrt[3]{\delta} + \sqrt[3]{\Delta})}, \quad \dot{\gamma}_1 = 3\delta \left(\frac{1-\eta^{2/3}}{1-\eta} \right)^3,$$

$$\dot{\gamma}_2 = \frac{\delta(1+2\eta)}{\eta^{2/3}} \left(\frac{1-\eta^{2/3}}{1-\eta} \right)^3.$$

Для случая $p=2$ можно получить лучшие оценки для $\dot{\gamma}_1$ и $\dot{\gamma}_2$, вводя в оператор

$$B = (E + \omega_1 R_1)(E + \omega_2 R_2) \tag{17}$$

два параметра ω_1 и ω_2 , которые учитывают, что границы операторов R_1 и R_2 различны. Имеет место

Теорема 3. Пусть оператор B имеет вид (17), выполнены условия

$$R_\alpha = R_\alpha^*, \quad \alpha = 1, 2, \quad R_1 R_2 = R_2 R_1,$$

и заданы границы операторов R_1 и R_2 :

$$\delta_\alpha E \leq R_\alpha \leq \Delta_\alpha E, \quad \alpha = 1, 2, \quad \delta_1 + \delta_2 > 0.$$

Тогда при оптимальных значениях параметров ω_1 и ω_2

$$\omega_1 = \frac{1+t\sqrt{\eta}}{r\sqrt{\eta}+s}, \quad \omega_2 = \frac{1-t\sqrt{\eta}}{r\sqrt{\eta}-s}$$

выполнены неравенства (6) с постоянными

$$\dot{\gamma}_1 = \frac{4\sqrt{\eta}}{(\omega_1+\omega_2)(1+\sqrt{\eta})^2}, \quad \dot{\gamma}_2 = \frac{2(1+\eta)}{(\omega_1+\omega_2)(1+\sqrt{\eta})^2},$$

где

$$r = \frac{\Delta_2 + \Delta_1 b}{1+b}, \quad s = \frac{\Delta_2 - \Delta_1 b}{1+b}, \quad t = \frac{1-b}{1+b}, \quad \eta = \frac{1-a}{1+a},$$

$$a = \sqrt{\frac{(\Delta_1 - \delta_1)(\Delta_2 - \delta_2)}{(\Delta_1 + \delta_2)(\Delta_2 + \delta_1)}}, \quad b = \frac{\Delta_2 + \delta_1}{\Delta_1 - \delta_1} a.$$

Для доказательства теоремы выполним замену, полагая

$$R_1 = (r\bar{R}_1 - sE)(E - t\bar{R}_1)^{-1}, \quad R_2 = (r\bar{R}_2 + sE)(E + t\bar{R}_2)^{-1}$$

с указанными значениями для r , s , t . Можно показать, что определяемые таким образом операторы

$$\bar{R}_1 = (R_1 + sE)(rE + tR_1)^{-1}, \quad \bar{R}_2 = (R_2 - sE)(rE - tR_2)^{-1}$$

удовлетворяют условиям $\bar{R}_\alpha = \bar{R}_\alpha^*$, $\alpha = 1, 2$, $\bar{R}_1 \bar{R}_2 = \bar{R}_2 \bar{R}_1$ и имеют одинаковые границы $\eta E \leq \bar{R}_\alpha \leq E$, $\eta > 0$, $\alpha = 1, 2$. Далее, так как операторы $E - t\bar{R}_1$ и $E + t\bar{R}_2$ самосопряжены и положительно определены, то существуют перестановочные операторы $(E - t\bar{R}_1)^{1/2}$ и $(E + t\bar{R}_2)^{1/2}$. Положим

$$x = (E - t\bar{R}_1)^{1/2}(E + t\bar{R}_2)^{1/2}y.$$

Получим

$$(Bx, x) = (1 - \omega_1 s)(1 + \omega_2 s)(\bar{B}y, y), \quad (18)$$

$$(Rx, x) = (r - st)(\bar{R}y, y), \quad (19)$$

где $\bar{B} = (E + \bar{\omega}\bar{R}_1)(E + \bar{\omega}\bar{R}_2)$, $\bar{R} = \bar{R}_1 + \bar{R}_2$,

$$\bar{\omega} = \frac{\omega_1 r - t}{1 - \omega_1 s} = \frac{\omega_2 r + t}{1 + \omega_2 s}. \quad (20)$$

Из (20) найдем

$$2\bar{\omega} = \frac{\omega_1 r - t}{1 - \omega_1 s} + \frac{\omega_2 r + t}{1 + \omega_2 s} = \frac{(r - st)(\omega_1 + \omega_2)}{(1 - \omega_1 s)(1 + \omega_2 s)}.$$

Отсюда и из (18), (19) получим

$$\frac{(Rx, x)}{(Bx, x)} = \frac{2\bar{\omega}}{\omega_1 + \omega_2} \frac{(\bar{R}y, y)}{(\bar{B}y, y)}. \quad (21)$$

Используя теорему 2, получим, что при

$$\bar{\omega} = \omega_0 = 1/\sqrt{\eta} \quad (22)$$

имеют место неравенства

$$\dot{\gamma}_1(\bar{R}y, y) \leq (\bar{B}y, y) \leq \dot{\gamma}_2(\bar{R}y, y), \quad (23)$$

где

$$\dot{\gamma}_1 = \frac{2\eta}{(1+\sqrt{\eta})^2}, \quad \dot{\gamma}_2 = \frac{\sqrt{\eta}(1+\eta)}{(1+\sqrt{\eta})^2}.$$

Следовательно, из (20) и (22) получим оптимальные значения для параметров ω_1 и ω_2 :

$$\omega_1 = \frac{1+t\sqrt{\eta}}{r\sqrt{\eta}+s}, \quad \omega_2 = \frac{1-t\sqrt{\eta}}{r\sqrt{\eta}-s},$$

а из (21) и (23) следуют неравенства (6) с постоянными $\dot{\gamma}_1$ и $\dot{\gamma}_2$, указанными в формулировке теоремы 3. Теорема 3 доказана.

3. Способ неявного обращения оператора B (двуухступенчатый метод). В п. 2 мы изучили способ построения неявных итерационных схем, характеризующийся тем, что оператор B задается конструктивно в виде произведения легко обратимых операторов. Рассмотрим еще один способ, в котором итерационное приближение y_{k+1} находится в результате вспомогательной процедуры, которую можно трактовать как неявное обращение некоторого оператора B .

Напомним, что общая идея такого способа рассмотрена в п. 4 § 3 гл. V. В п. 4 § 1 гл. XIII этот способ использовался при построении итерационного метода решения уравнения с нелинейным оператором A . Там же были сформулированы условия, позволяющие получить оценки для $\dot{\gamma}_1$ и $\dot{\gamma}_2$, входящих в неравенства (6).

Изложим полученные результаты. Пусть итерационное приближение y_{k+1} находится по формуле схемы с поправкой: $y_{k+1} = y_k - \tau_{k+1}w^p$, а поправка w^p есть приближенное решение вспомогательного уравнения

$$Rw = r_k, \quad r_k = Ay_k - f. \quad (24)$$

Здесь R — регуляризатор, удовлетворяющий неравенствам (5) для случая самосопряженного оператора A и удовлетворяющий неравенствам (10') или (11') для несамосопряженного A .

Пусть уравнение (24) решается при помощи какой-либо двухслойной итерационной схемы, так что погрешность $z^m = w^m - w$ удовлетворяет уравнению

$$z^{m+1} = S_{m+1}z^m, \quad m = 0, 1, \dots, p-1, \quad z^0 = w^0 - w,$$

где S_{m+1} — оператор перехода от m -й к $(m+1)$ -й итерации.

Выбирая $w^0 = 0$, из равенств

$$z^p = w^p - w = T_p(w^0 - w), \quad T_p = \prod_{m=1}^p S_m,$$

$$w = R^{-1}r_k,$$

будем иметь

$$w^p = B^{-1}r_k, \quad \text{где } B = R(E - T_p)^{-1}.$$

Подставляя найденное для w^p выражение в (23), получим неявную итерационную схему (2) с указанным оператором B .

Теорема 4. Пусть выполнены условия

$$R = R^* > 0, \quad T_p^* R = R T_p, \quad \|T_p\|_R \leq q < 1.$$

Тогда оператор $B = R(E - T_p)^{-1}$ самосопряжен и положительно определен в H и верны неравенства (6) с постоянными $\gamma_1 = 1 - q$, $\gamma_2 = 1 + q$.

Для доказательства см. лемму 2 главы XIII.

Замечание. Если операторы R и T_p самосопряжены и перестановочны и $\|T_p\| \leq q < 1$, то справедливы утверждения теоремы 4.

Описанным выше способом мы построили неявную двухслойную итерационную схему. Если же исходить из формул

$$y_{k+1} = \alpha_{k+1} y_k + (1 - \alpha_{k+1}) y_{k-1} - \tau_{k+1} \alpha_{k+1} w_k^p, \quad k = 1, 2, \dots,$$

$$y_1 = y_0 - \tau_1 w_0^p,$$

а поправку w_k^p для любого $k = 0, 1, \dots$ находить как приближенное решение уравнения (24), то мы получим неявную трехслойную итерационную схему

$$By_{k+1} = \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \alpha_{k+1}\tau_{k+1}f, \quad (25)$$

$$By_1 = (B - \tau_1 A)y_0 + \tau_1 f.$$

В заключение отметим, что итерационные параметры τ_k для схемы (2) и τ_k , α_k для схемы (25) выбираются согласно общей теории итерационных методов. Здесь возникает задача выбора оптимального числа итераций p для вспомогательного итерационного процесса. Поясним ситуацию. Пусть для простоты вспомогательный процесс является стационарным ($S_m \equiv S$), операторы R и S самосопряжены и перестановочны и выполнено условие $\|S\| \leq p$. Тогда $q = p^p$, т. е.

$$p = \ln q / \ln p. \quad (26)$$

Операторы A и B удовлетворяют неравенствам (4) с постоянными

$$\gamma_1 = c_1(1 - q), \quad \gamma_2 = c_2(1 + q).$$

Если итерационные параметры τ_k в схеме (2) выбраны по формулам чебышевского метода, то для числа итераций верна оценка

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \ln(0,5\varepsilon)/\ln\rho_1,$$

где $\rho_1 = \rho_1(q) = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}$, $\xi = \frac{\gamma_1}{\gamma_2} = \frac{c_1}{c_2} \frac{1-q}{1+q}$. Тогда общее число итераций $k = pn$ оценивается величиной

$$k \geq k_0(\varepsilon), \quad k_0(\varepsilon) = \frac{\ln 0,5\varepsilon}{\ln \rho} \frac{\ln q}{\ln \rho_1(q)}.$$

Отсюда следует, что величина q , определяющая, согласно (26), число внутренних итераций, должна быть выбрана из условия минимума функции $\varphi(q) = \ln q/\ln \rho_1(q)$. Эта задача может быть решена численно.

§ 2. Системы эллиптических уравнений

1. Задача Дирихле для системы эллиптических уравнений в p -мерном параллелепипеде. Пусть $\mathbf{u} = (u^1(x), u^2(x), \dots, u^{m_0}(x))$ и $\mathbf{f} = (f^1(x), f^2(x), \dots, f^{m_0}(x))$ — векторы размерности m_0 , $x = (x_1, x_2, \dots, x_p)$ — точка p -мерного пространства, $k = (k_{\alpha\beta})$ — клеточная матрица размера $p \times p$, так что клетка $k_{\alpha\beta} = (k_{\alpha\beta}^{sm}(x))$ является матрицей размера $m_0 \times m_0$:

$$k = \begin{vmatrix} k_{11} & k_{12} & \dots & k_{1p} \\ k_{21} & k_{22} & \dots & k_{2p} \\ \dots & \dots & \dots & \dots \\ k_{p1} & k_{p2} & \dots & k_{pp} \end{vmatrix}, \quad k_{\alpha\beta} = \begin{vmatrix} k_{\alpha\beta}^{11} & k_{\alpha\beta}^{12} & \dots & k_{\alpha\beta}^{1m_0} \\ k_{\alpha\beta}^{21} & k_{\alpha\beta}^{22} & \dots & k_{\alpha\beta}^{2m_0} \\ \dots & \dots & \dots & \dots \\ k_{\alpha\beta}^{m_0 1} & k_{\alpha\beta}^{m_0 2} & \dots & k_{\alpha\beta}^{m_0 m_0} \end{vmatrix}.$$

В p -мерном параллелепипеде $\bar{G} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2, \dots, p\}$ с границей Γ рассмотрим задачу Дирихле для системы эллиптических уравнений:

$$\begin{aligned} L\mathbf{u} &= \sum_{\alpha, \beta=1}^p \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta} \frac{\partial \mathbf{u}}{\partial x_\beta} \right) = -\mathbf{f}(x), \quad x \in G, \\ \mathbf{u}(x) &= \mathbf{g}(x), \quad x \in \Gamma. \end{aligned} \tag{1}$$

Если перейти от векторной к скалярной записи, то задача (1) запишется в виде системы

$$\begin{aligned} (L\mathbf{u})^s &= -f^s(x), \quad x \in G, \\ u^s(x) &= g^s(x), \quad x \in \Gamma, \quad s = 1, 2, \dots, m_0, \end{aligned}$$

где

$$(L\mathbf{u})^s = \sum_{\alpha, \beta=1}^p \sum_{m=1}^{m_0} \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}^{sm}(x) \frac{\partial u^m}{\partial x_\beta} \right). \tag{2}$$

Будем предполагать, что выполнено условие сильной эллиптичности

$$c_1 \sum_{\alpha=1}^p |\xi_\alpha|^2 \leq \sum_{\alpha, \beta=1}^p (k_{\alpha\beta} \xi_\alpha, \xi_\beta) \leq c_2 \sum_{\alpha=1}^p |\xi_\alpha|^2, \quad (3)$$

где $c_1 > 0$, $c_2 > 0$ — постоянные, не зависящие от x , $\xi_\alpha = (\xi_\alpha^1, \xi_\alpha^2, \dots, \xi_\alpha^{m_0})$, $\alpha = 1, 2, \dots, p$, — произвольные векторы,

$$|\xi_\alpha|^2 = \sum_{s=1}^{m_0} (\xi_\alpha^s)^2, \quad (k_{\alpha\beta} \xi_\alpha, \xi_\beta) = \sum_{s, m=1}^{m_0} k_{\alpha\beta}^{sm} \xi_\alpha^s \xi_\beta^m.$$

Отметим, что левое неравенство (3) означает положительную определенность матрицы k .

Построим разностную схему, аппроксимирующую задачу (1). Для этого в области \bar{G} введем прямоугольную равномерную сетку

$$\bar{\omega} = \{x_i = (i_1 h_1, \dots, i_p h_p) \in \bar{G}, 0 \leq i_\alpha \leq N_\alpha, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2, \dots, p\}$$

с границей γ , так что $\bar{\omega} = \omega \cup \gamma$. На сетке $\bar{\omega}$ будем рассматривать векторные сеточные функции, компонентами которых являются сеточные функции p дискретных переменных, например $\mathbf{y} = (y^1, y^2, \dots, y^{m_0})$, причем $y^s = y^s(x_i) = y^s(i_1, i_2, \dots, i_p)$.

Разностная задача Дирихле для системы (1) на сетке $\bar{\omega}$ в векторной записи имеет вид

$$\begin{aligned} \Lambda^- \mathbf{y} &= \sum_{\alpha, \beta=1}^p 0.5 [(k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}] = -\varphi(x), \quad x \in \omega, \\ \mathbf{y}(x) &= \mathbf{g}(x), \quad x \in \gamma. \end{aligned}$$

Переходя к скалярной записи, получим систему

$$\begin{aligned} (\Lambda^- \mathbf{y})^s &= -\varphi^s(x), \quad x \in \omega, \\ y^s(x) &= g^s(x), \quad x \in \gamma, \quad s = 1, 2, \dots, m_0, \end{aligned} \quad (4)$$

где

$$(\Lambda^- \mathbf{y})^s = \sum_{\alpha, \beta=1}^p \sum_{m=1}^{m_0} 0.5 [(k_{\alpha\beta}^{sm} y_{\bar{x}_\beta}^m)_{\bar{x}_\alpha} + (k_{\alpha\beta}^{sm} y_{x_\beta}^m)_{\bar{x}_\alpha}].$$

Оператор Λ^- , как и в случае скалярного эллиптического уравнения, допускает другую запись, именно:

$$\begin{aligned} \Lambda^- \mathbf{y} &= \sum_{\alpha=1}^p 0.5 [(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{\bar{x}_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \\ &\quad + \sum_{\alpha \neq \beta}^{1 \div p} 0.5 [(k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}]. \end{aligned}$$

Отметим, что для аппроксимации дифференциального оператора L можно также использовать и другие, отличные от Λ^- разностные операторы, например

$$\begin{aligned}\Lambda^+ \mathbf{y} = \sum_{\alpha=1}^p 0,5 [k_{\alpha\alpha} \mathbf{y}_{\bar{x}_\alpha}]_{x_\alpha} + (k_{\alpha\alpha} \mathbf{y}_{x_\alpha})_{\bar{x}_\alpha}] + \\ + \sum_{\alpha \neq \beta}^{1+p} 0,5 [(k_{\alpha\beta} \mathbf{y}_{x_\beta})_{x_\alpha} + (k_{\alpha\beta} \mathbf{y}_{\bar{x}_\beta})_{\bar{x}_\alpha}]\end{aligned}$$

или

$$\begin{aligned}\Lambda^0 \mathbf{y} = 0,5 (\Lambda^- + \Lambda^+) \mathbf{y} = \\ = \sum_{\alpha=1}^p 0,5 [(k_{\alpha\alpha} \mathbf{y}_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} \mathbf{y}_{x_\alpha})_{\bar{x}_\alpha}] + \sum_{\alpha \neq \beta}^{1+p} (k_{\alpha\beta} \mathbf{y}_{x_\beta})_{\bar{x}_\alpha}.\end{aligned}$$

Введем пространство H векторных сеточных функций, заданных на ω , и определим в нем скалярное произведение

$$\begin{aligned}(\mathbf{u}, \mathbf{v}) = \sum_{s=1}^{m_0} (\mathbf{u}^s, \mathbf{v}^s), \quad (\mathbf{u}^s, \mathbf{v}^s) = \sum_{x \in \omega} u^s(x) v^s(x) h_1 h_2 \dots h_p, \\ \mathbf{u} = (u^1, u^2, \dots, u^{m_0}), \quad \mathbf{v} = (v^1, v^2, \dots, v^{m_0}), \quad \mathbf{u}, \mathbf{v} \in H.\end{aligned}$$

Определим разностный оператор Лапласа:

$$\mathcal{R} \mathbf{y} = \sum_{\alpha=1}^p \mathbf{y}_{\bar{x}_\alpha} |_{x_\alpha}, \quad (\mathcal{R} \mathbf{y})^s = \sum_{\alpha=1}^p y_{\bar{x}_\alpha}^s |_{x_\alpha}.$$

В пространстве H обычным образом определим операторы A и R :

$$A \mathbf{y} = -\Lambda^- \dot{\mathbf{y}}, \quad R \mathbf{y} = -\mathcal{R} \dot{\mathbf{y}}, \quad \mathbf{y} \in H,$$

где $\dot{\mathbf{y}}(x) = \mathbf{y}(x)$ для $x \in \omega$ и $\dot{\mathbf{y}}(x) = 0$, если $x \in \gamma$.

Используя введенные обозначения и подправляя очевидным образом правую часть уравнения (4) в приграничных узлах, запишем разностную задачу (4) в виде операторного уравнения

$$A \mathbf{u} = \mathbf{f}, \tag{5}$$

заданного в гильбертовом пространстве H .

Пользуясь разностной формулой Грина для скалярных сеточных функций, условиями (3) и предполагая, что выполнены условия симметрии

$$k_{\alpha\beta}^{sm} = k_{\beta\alpha}^{ns}, \quad \alpha, \beta = 1, 2, \dots, p, \quad s, m = 1, 2, \dots, m_0, \tag{6}$$

получим, что операторы R и A самосопряжены в H и энергетически эквивалентны с постоянными c_1 и c_2 , т. е. имеют место операторные неравенства

$$c_1 R \leq A \leq c_2 R, \quad c_1 > 0. \tag{7}$$

Для нахождения приближенного решения уравнения (5) воспользуемся неявным двухслойным итерационным методом с чебышевскими параметрами

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (8)$$

где

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \\ k = 1, 2, \dots, n,$$

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \\ n \geq n_0(\epsilon) = \ln(0.5\epsilon)/\ln \rho_1,$$

а γ_1 и γ_2 — постоянные энергетической эквивалентности самоспряженных операторов A и B :

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0, \quad A = A^*, \quad B = B^*. \quad (9)$$

Если в качестве оператора B выбрать определенный выше оператор R , то из (7) получим, что в неравенствах (9) $\gamma_1 = c_1$ и $\gamma_2 = c_2$. Следовательно, число итераций метода (8) не зависит в рассматриваемом случае от числа узлов сетки: $n = O(\ln(2/\epsilon))$.

Из определения операторов A и B следует, что для нахождения y_{k+1} по известному предыдущему приближению y_k необходимо решить следующую разностную задачу:

$$\mathcal{R}y_{k+1} = -F_k, \quad x \in \omega, \quad F_k = \tau_{k+1}(\Lambda^{-}y_k + \Phi) - \mathcal{R}y_k, \\ y_{k+1} = g, \quad x \in \gamma.$$

В скалярном виде эта задача записывается в виде системы

$$\sum_{\alpha=1}^p (y_{k+1}^s)_{x_\alpha x_\alpha} = -F_k^s(x), \quad x \in \omega, \\ y_{k+1}^s(x) = g^s(x), \quad x \in \gamma, \quad s = 1, 2, \dots, m_0. \quad (10)$$

Так как каждое уравнение системы (10) может быть решено независимо от других уравнений, то нахождение приближения y_{k+1} сводится к решению m_0 разностных задач Дирихле в p -мерном параллелепипеде на равномерной прямоугольной сетке ω .

Если для решения p -мерной разностной задачи Дирихле для уравнения Пуассона использовать метод разделения переменных с алгоритмом быстрого дискретного преобразования Фурье, то можно показать, что потребуется $q \approx 4pN^p \log_2 N$ ($N_1 = N_2 = \dots = N_p = N = 2^n$) арифметических действий. Следовательно, для решения системы (10) потребуется $Q_{m_0} = m_0 q$ действий, а всего для нахождения решения разностной задачи (4) с точностью ϵ

необходимо затратить $Q = nQ_{m_0} = nm_0q = O\left(m_0 p N^p \ln \frac{2}{\varepsilon} \log_2 N\right)$ арифметических операций.

Рассмотрим теперь попеременно-треугольный итерационный метод. Итерационная схема имеет вид (8), где B есть факторизованный оператор $B = (E + \omega R_1)(E + \omega R_2)$, $R_1 = R_2^*$, $R_1 + R_2 = R$. Операторы R_1 и R_2 определяются при помощи разностных операторов \mathcal{R}_1 и \mathcal{R}_2 следующим образом: $R_\alpha \mathbf{y} = -\mathcal{R}_\alpha \dot{\mathbf{y}}$, $\alpha = 1, 2$, $\mathbf{y}(x) = \dot{\mathbf{y}}(x)$ для $x \in \omega$ и $\dot{\mathbf{y}}(x) = 0$ для $x \in \gamma$, где

$$\mathcal{R}_1 \mathbf{y} = -\sum_{\alpha=1}^p \frac{1}{h_\alpha} \mathbf{y}_{x_\alpha}^-, \quad \mathcal{R}_2 \mathbf{y} = \sum_{\alpha=1}^p \frac{1}{h_\alpha} \mathbf{y}_{x_\alpha}^+.$$

Так же, как и в скалярном случае, доказывается, что выполнены неравенства $R \geqslant \delta E$, $R_1 R_2 \leqslant \frac{\Delta}{4} R$, где

$$\delta = \sum_{\alpha=1}^p \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \sum_{\alpha=1}^p \frac{4}{h_\alpha^2}.$$

Из общей теории попеременно-треугольного метода (см. § 1 главы X) следует, что при оптимальном значении параметра $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$ имеют место операторные неравенства

$$\dot{\gamma}_1 B \leqslant R \leqslant \dot{\gamma}_2 B, \quad \dot{\gamma}_1 > 0, \quad (11)$$

где $\dot{\gamma}_1 = \frac{\delta}{2(1 + \sqrt{\eta})}$, $\dot{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}$, $\eta = \frac{\delta}{\Delta}$.

Сравнивая (7), (9) и (11), находим, что операторы A и B удовлетворяют неравенствам (9) с $\gamma_1 = c_1 \dot{\gamma}_1$ и $\gamma_2 = c_2 \dot{\gamma}_2$.

Используя для схемы (8) чебышевский набор параметров τ_k , получим, что построенный попеременно-треугольный итерацион-

ный метод требует $n = O\left(\frac{1}{\sqrt{|h|}} \sqrt{\frac{c_2}{c_1}} \ln \frac{2}{\varepsilon}\right)$ итераций, где

$|h|^2 = h_1^2 + h_2^2 + \dots + h_p^2$. Так как переход от y_k к y_{k+1} осуществляется по явным формулам с затратой $O(m_0 N_1 N_2 \dots N_p)$ арифметических действий, то общее число действий, требуемое для нахождения решения задачи (4) с точностью ε , оценивается величиной

$$Q = O\left(m_0 N^{p+0.5} \sqrt{\frac{c_2}{c_1}} \ln \frac{2}{\varepsilon}\right),$$

если $l_1 = l_2 = \dots = l_p$, $N_1 = N_2 = \dots = N_p = N$.

В заключение отметим, что рассмотренные выше итерационные методы сходятся в энергетическом пространстве H_D , где в качестве оператора D можно взять один из операторов A , B или $AB^{-1}A$.

2. Система уравнений теории упругости. Рассмотрим систему уравнений стационарной теории упругости (уравнений Ламэ)

$$L\mathbf{u} = \mu \Delta \mathbf{u} + (\lambda + \mu) \operatorname{grad} \operatorname{div} \mathbf{u} = -\mathbf{f}(x), \quad (12)$$

где $\mathbf{u} = (u^1, u^2, \dots, u^p)$, $\mathbf{f} = (f^1, f^2, \dots, f^p)$, $x = (x_1, x_2, \dots, x_p)$, $\lambda > 0$ и $\mu > 0$ — постоянные Ламэ.

Напишем уравнение (12) в виде системы

$$(Lu)^s = \mu \sum_{\alpha=1}^p \frac{\partial^2 u^s}{\partial x_\alpha^2} + (\lambda + \mu) \sum_{\beta=1}^p \frac{\partial^2 u^\beta}{\partial x_\beta \partial x_s} = -f^s, \quad s = 1, 2, \dots, p. \quad (13)$$

При $p=2$ систему (13) можно записать в виде

$$\begin{aligned} (\lambda + 2\mu) \frac{\partial^2 u^1}{\partial x_1^2} + \mu \frac{\partial^2 u^1}{\partial x_2^2} + (\lambda + \mu) \frac{\partial^2 u^2}{\partial x_1 \partial x_2} &= -f^1(x_1, x_2), \\ (\lambda + \mu) \frac{\partial^2 u^1}{\partial x_1 \partial x_2} + \mu \frac{\partial^2 u^2}{\partial x_1^2} + (\lambda + 2\mu) \frac{\partial^2 u^2}{\partial x_2^2} &= -f^2(x_1, x_2). \end{aligned}$$

Эта система описывает равновесие однородного изотропного упругого твердого тела в случае плоской деформации. Неизвестные функции $u^1(x_1, x_2)$ и $u^2(x_1, x_2)$ имеют смысл перемещений точек по направлениям осей Ox_1 и Ox_2 соответственно.

Для системы (12) может быть поставлена задача отыскания вектора $\mathbf{u}(x)$, удовлетворяющего уравнению (12) в области G и принимающего на границе Γ заданные значения

$$\mathbf{u}(x) = \mathbf{g}(x), \quad x \in \Gamma. \quad (14)$$

Сравнивая (13) с (2), находим, что система (12), (14) может быть записана в виде (1), где $m_0 = p$,

$$k_{\alpha\beta}^{sm} = \mu \delta_{\alpha\beta} \delta_{sm} + (\lambda + \mu) [\theta \delta_{as} \delta_{\beta m} + (1 - \theta) \delta_{am} \delta_{\beta s}], \quad (15)$$

а θ — произвольная постоянная, $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$ Действительно, подставляя (15) в (2), будем иметь

$$\begin{aligned} (Lu)^s &= \sum_{\alpha, \beta=1}^p \sum_{m=1}^p \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta}^{sm} \frac{\partial u^m}{\partial x_\beta} \right) = \mu \sum_{\alpha, \beta=1}^p \sum_{m=1}^p \delta_{\alpha\beta} \delta_{sm} \frac{\partial^2 u^m}{\partial x_\alpha \partial x_\beta} + \\ &+ (\lambda + \mu) \left[\theta \sum_{\alpha, \beta=1}^p \sum_{m=1}^p \delta_{as} \delta_{\beta m} \frac{\partial^2 u^m}{\partial x_\alpha \partial x_\beta} + (1 - \theta) \sum_{\alpha, \beta=1}^p \sum_{m=1}^p \delta_{am} \delta_{\beta s} \frac{\partial^2 u^m}{\partial x_\alpha \partial x_\beta} \right] = \\ &= \mu \sum_{\alpha=1}^p \frac{\partial^2 u^s}{\partial x_\alpha^2} + (\lambda + \mu) \left[\theta \sum_{\beta=1}^p \frac{\partial^2 u^\beta}{\partial x_s \partial x_\beta} + (1 - \theta) \sum_{\alpha=1}^p \frac{\partial^2 u^\alpha}{\partial x_\alpha \partial x_s} \right] = \\ &= \mu \sum_{\alpha=1}^p \frac{\partial^2 u^s}{\partial x_\alpha^2} + (\lambda + \mu) \sum_{\beta=1}^p \frac{\partial^2 u^\beta}{\partial x_s \partial x_\beta}. \end{aligned}$$

Утверждение доказано.

Найдем теперь постоянные c_1 и c_2 в неравенствах (3). Покажем, что $c_1 = \mu$. Имеем

$$\begin{aligned} \sum_{s=1}^p \sum_{\alpha, \beta=1}^p k_{\alpha \beta}^{s \bar{s}} \xi_{\alpha}^s \xi_{\beta}^{\bar{s}} &= \mu \sum_{\alpha, s=1}^p (\xi_{\alpha}^s)^2 + \\ &+ (\lambda + \mu) \left[\theta \sum_{\alpha, s=1}^p \xi_{\alpha}^s \xi_{\bar{s}}^s + (1 - \theta) \sum_{\alpha, s=1}^p \xi_{\alpha}^s \xi_{s \bar{s}}^{\alpha} \right] = \\ &= \mu \sum_{\alpha, s=1}^p (\xi_{\alpha}^s)^2 + (\lambda + \mu) \left[\theta \left(\sum_{\alpha=1}^p \xi_{\alpha}^{\alpha} \right)^2 + (1 - \theta) \sum_{\alpha, s=1}^p \xi_{\alpha}^s \xi_{s \bar{s}}^{\alpha} \right]. \end{aligned} \quad (16)$$

Полагая здесь $\theta = 1$, найдем

$$\sum_{\alpha, \beta=1}^p (k_{\alpha \beta} \xi_{\alpha}, \xi_{\beta}) = \mu \sum_{\alpha=1}^p |\xi_{\alpha}|^2 + (\lambda + \mu) \left(\sum_{\alpha=1}^p \xi_{\alpha}^{\alpha} \right)^2 \geq \mu \sum_{\alpha=1}^p |\xi_{\alpha}|^2.$$

Нетрудно показать также, что $c_2 = \lambda + 2\mu$. Полагая в (16) $\theta = 0$ и используя неравенство Коши—Буняковского, получим

$$\begin{aligned} \sum_{\alpha, \beta=1}^p (k_{\alpha \beta} \xi_{\alpha}, \xi_{\beta}) &= \mu \sum_{\alpha=1}^p |\xi_{\alpha}|^2 + (\lambda + \mu) \sum_{\alpha, s=1}^p \xi_{\alpha}^s \xi_{s \bar{s}}^{\alpha} \leqslant \\ &\leqslant \mu \sum_{\alpha=1}^p |\xi_{\alpha}|^2 + \frac{(\lambda + \mu)}{2} \left[\sum_{\alpha, s=1}^p (\xi_{\alpha}^s)^2 + \sum_{\alpha, s=1}^p (\xi_{s \bar{s}}^{\alpha})^2 \right] = \\ &= \mu \sum_{\alpha=1}^p |\xi_{\alpha}|^2 + (\lambda + \mu) \sum_{\alpha, s=1}^p (\xi_{\alpha}^s)^2 = (\lambda + 2\mu) \sum_{\alpha=1}^p |\xi_{\alpha}|^2. \end{aligned}$$

Построим теперь разностную схему, аппроксимирующую задачу (12), (14). Подставляя (15) в разностную схему (4), будем иметь

$$(\Lambda - y)^s = \mu \sum_{\alpha=1}^p y_{x_{\alpha} x_{\alpha}}^s + 0,5 (\lambda + \mu) \sum_{\beta=1}^p \left(y_{x_{\beta} x_s}^{\beta} + y_{x_{\beta} x_{\bar{s}}}^{\beta} \right) = -\varphi^s, \quad x \in \omega, \quad (17)$$

$$y^s(x) = g^s(x), \quad x \in \gamma, \quad s = 1, 2, \dots, p,$$

где $\bar{\omega} = \omega \cup \gamma$ — сетка, введенная в п. 1.

Остается определить операторы A и R , как это было сделано в п. 1. Условие симметрии (6) выполнено, поэтому, пользуясь первой разностной формулой Грина, находим, что операторы A и R самосопряжены в H и имеют место неравенства $c_1 R \leqslant A \leqslant c_2 R$, где $c_1 = \mu$, $c_2 = \lambda + 2\mu$.

Здесь дальнейшие рассуждения совпадают с теми, которые проводились в п. 1. Так, итерационный метод (8) с $B = R$ и чебышевскими параметрами τ_k характеризуется следующей оценкой для числа итераций:

$$n \geq n_0(\varepsilon) = \frac{\ln 0,5\varepsilon}{\ln \rho_1}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{c_1}{c_2} = \frac{\mu}{\lambda + 2\mu},$$

а попеременно-треугольный метод, построенный на основе регуляризатора R , характеризуется этой же оценкой, где

$$\xi = \frac{\mu}{\lambda + 2\mu} \frac{2\sqrt{\eta}}{1 + \sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta},$$

$$\delta = \sum_{\alpha=1}^p \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \sum_{\alpha=1}^p \frac{4}{h_\alpha^2}.$$

Таким образом, для попеременно-треугольного метода число

итераций пропорционально $\sqrt{\frac{\lambda + 2\mu}{\mu}} = \sqrt{2 + \frac{\lambda}{\mu}}$:

$$n_0(\varepsilon) = \sqrt{2 + \frac{\lambda}{\mu}} n_0^*(\varepsilon),$$

где $n_0^*(\varepsilon)$ — число итераций для решения p -мерного разностного уравнения Пуассона попеременно-треугольным методом.

ГЛАВА XV

МЕТОДЫ РЕШЕНИЯ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ В КРИВОЛИНЕЙНЫХ ОРТОГОНАЛЬНЫХ КООРДИНАТАХ

В этой главе рассматриваются примеры решения разностных задач, аппроксимирующих краевые задачи для эллиптических уравнений в криволинейных системах координат. Для задач в цилиндрической и полярной системах координат выясняются условия применимости прямых и итерационных методов, в частности метода переменных направлений.

В § 1 приведена постановка краевых задач для дифференциальных уравнений. Параграф 2 посвящен изложению прямых и итерационных методов решения разностных задач в (r, z) — геометрии, а также задач на поверхности цилиндра. В § 3 рассмотрены методы решения разностных задач в круге, кольце и кольцевом секторе.

§ 1. Постановка краевых задач для дифференциальных уравнений

1. Эллиптические уравнения в цилиндрической системе координат. Пусть задано уравнение Пуассона

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} = -f(x), \quad x = (x_1, x_2, x_3). \quad (1)$$

Если для этого уравнения ставится задача отыскания решения в конечном круговом цилиндре или в кольцевой трубе, то его естественно рассматривать в цилиндрических координатах. В этой системе координат уравнение Пуассона (1) имеет вид

$$L_{r\varphi z}u = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \varphi^2} + \frac{\partial^2 u}{\partial z^2} = -f(r, \varphi, z), \quad (2)$$

где $r = \sqrt{x_1^2 + x_2^2}$, $\operatorname{tg} \varphi = x_2/x_1$, $z = x_3$.

Уравнение (1) описывает, например, стационарное распределение температуры $u = u(x_1, x_2, x_3)$ в однородной среде. Если среда неоднородна, но изотропна, то вместо (1) следует рассмотреть уравнение

$$Lu = \operatorname{div}(k \operatorname{grad} u) = \sum_{\alpha=1}^3 \frac{\partial}{\partial x_\alpha} \left(k(x) \frac{\partial u}{\partial x_\alpha} \right) = -f(x), \quad (3)$$

которому в (r, φ, z) -системе соответствует уравнение

$$L_{r\varphi z}u = \frac{1}{r} \frac{\partial}{\partial r} \left(rk \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial}{\partial \varphi} \left(k \frac{\partial u}{\partial \varphi} \right) + \frac{\partial}{\partial z} \left(k \frac{\partial u}{\partial z} \right) = -f. \quad (4)$$

Если среда анизотропна, т. е. коэффициент теплопроводности зависит не только от точки, но и от направления, то вместо (3) будем иметь уравнение со смешанными производными

$$Lu = \sum_{\alpha, \beta=1}^3 \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta} \frac{\partial u}{\partial x_\beta} \right) = -f(x). \quad (5)$$

Уравнению (5) в цилиндрической системе координат соответствует уравнение

$$\begin{aligned} L_{r\varphi z}u = & \frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \frac{\bar{k}_{12}}{r} \frac{\partial u}{\partial \varphi} + \bar{k}_{13} \frac{\partial u}{\partial z} \right) \right] + \\ & + \frac{1}{r} \frac{\partial}{\partial \varphi} \left(\bar{k}_{21} \frac{\partial u}{\partial r} + \frac{\bar{k}_{22}}{r} \frac{\partial u}{\partial \varphi} + \bar{k}_{23} \frac{\partial u}{\partial z} \right) + \\ & + \frac{\partial}{\partial z} \left(\bar{k}_{31} \frac{\partial u}{\partial r} + \frac{\bar{k}_{32}}{r} \frac{\partial u}{\partial \varphi} + \bar{k}_{33} \frac{\partial u}{\partial z} \right) = -f(r, \varphi, z), \end{aligned} \quad (6)$$

где коэффициенты $\bar{k}_{\alpha\beta}$ выражаются через $k_{\alpha\beta}$ по формулам:

$$\begin{aligned} \bar{k}_{11} &= k_{11} \cos^2 \varphi + (k_{12} + k_{21}) \sin \varphi \cos \varphi + k_{22} \sin^2 \varphi, \\ \bar{k}_{12} &= k_{12} \cos^2 \varphi + (k_{22} - k_{11}) \sin \varphi \cos \varphi - k_{21} \sin^2 \varphi, \\ \bar{k}_{21} &= k_{21} \cos^2 \varphi + (k_{22} - k_{11}) \sin \varphi \cos \varphi - k_{12} \sin^2 \varphi, \\ \bar{k}_{22} &= k_{11} \sin^2 \varphi - (k_{12} + k_{21}) \sin \varphi \cos \varphi + k_{22} \cos^2 \varphi, \\ \bar{k}_{13} &= k_{13} \cos \varphi + k_{23} \sin \varphi, \quad \bar{k}_{23} = k_{23} \cos \varphi - k_{13} \sin \varphi, \\ \bar{k}_{31} &= k_{31} \cos \varphi + k_{32} \sin \varphi, \quad \bar{k}_{32} = k_{32} \cos \varphi - k_{31} \sin \varphi, \\ \bar{k}_{33} &= k_{33}. \end{aligned}$$

Уравнение (6) называется *уравнением со смешанными производными в цилиндрической системе координат*. Если $\bar{k}_{\alpha\beta} = 0$ для $\alpha \neq \beta$, то (6) принимает вид

$$L_{r\varphi z}u = \frac{1}{r} \frac{\partial}{\partial r} \left(r \bar{k}_1 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial}{\partial \varphi} \left(\bar{k}_2 \frac{\partial u}{\partial \varphi} \right) + \frac{\partial}{\partial z} \left(\bar{k}_3 \frac{\partial u}{\partial z} \right) = -f, \quad (7)$$

где $\bar{k}_\alpha = \bar{k}_{\alpha\alpha}$, $\alpha = 1, 2, 3$, и называется *уравнением без смешанных производных*.

Отметим, что если $k_{\alpha\beta} = k_{\beta\alpha}$, то $\bar{k}_{\alpha\beta} = \bar{k}_{\beta\alpha}$ и наоборот. Приведенные выше уравнения (2) и (4) являются частными случаями уравнения (7), соответствующими $\bar{k}_\alpha \equiv 1$ и $\bar{k}_\alpha \equiv k$.

Уравнение (5) будет сильно эллиптическим, если существует постоянная $c_1 > 0$ такая, что для любых ξ_1 , ξ_2 и ξ_3 справедливо неравенство

$$\sum_{\alpha, \beta=1}^3 k_{\alpha\beta}(x) \xi_\alpha \xi_\beta \geq c_1 \sum_{\alpha=1}^3 \xi_\alpha^2. \quad (8)$$

Если в (8) сделать замену, полагая

$$\xi_1 = \bar{\xi}_1 \cos \varphi - \bar{\xi}_2 \sin \varphi, \quad \xi_2 = \bar{\xi}_1 \sin \varphi + \bar{\xi}_2 \cos \varphi, \quad \xi_3 = \bar{\xi}_3,$$

то неравенство (8) примет вид

$$\sum_{\alpha, \beta=1}^3 \bar{k}_{\alpha \beta} \bar{\xi}_{\alpha} \bar{\xi}_{\beta} \geq c_1 \sum_{\alpha=1}^3 \bar{\xi}_{\alpha}^2. \quad (9)$$

На практике наиболее часто встречаются два случая.

А) В осесимметрическом случае коэффициенты и правая часть уравнения, а также само решение не зависят от угла φ . При этом уравнение (6) упрощается:

$$L_{rz} u = \frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \bar{k}_{13} \frac{\partial u}{\partial z} \right) \right] + \frac{\partial}{\partial z} \left(\bar{k}_{31} \frac{\partial u}{\partial r} + \bar{k}_{33} \frac{\partial u}{\partial z} \right) = -f(r, z), \quad (10)$$

а в отсутствие смешанных производных уравнение, соответствующее (7), имеет вид

$$L_{rz} u = \frac{1}{r} \frac{\partial}{\partial r} \left(r \bar{k}_1 \frac{\partial u}{\partial r} \right) + \frac{\partial}{\partial z} \left(\bar{k}_3 \frac{\partial u}{\partial z} \right) = -f(r, z). \quad (11)$$

Б) В плоском случае коэффициенты, правая часть и решение уравнения (6) не зависят от z , и, следовательно, уравнение (6) принимает вид

$$L_{r\varphi} u = \frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \frac{\bar{k}_{12}}{r} \frac{\partial u}{\partial \varphi} \right) \right] + \frac{1}{r} \frac{\partial}{\partial \varphi} \left(\bar{k}_{21} \frac{\partial u}{\partial r} + \frac{\bar{k}_{22}}{r} \frac{\partial u}{\partial \varphi} \right) = -f(r, \varphi). \quad (12)$$

Если смешанные производные отсутствуют, то уравнение имеет вид

$$L_{r\varphi} u = \frac{1}{r} \frac{\partial}{\partial r} \left(r \bar{k}_1 \frac{\partial u}{\partial r} \right) + \frac{1}{r} \frac{\partial}{\partial \varphi} \left(\bar{k}_2 \frac{\partial u}{\partial \varphi} \right) = -f(r, \varphi). \quad (13)$$

В плоском случае говорят, что (12) и (13) есть эллиптические уравнения в полярной системе координат.

Отметим, что при $\bar{k}_{\alpha} = 1$, $\alpha = 1, 2, 3$ формулы (11) и (13) описывают уравнение Пуассона в (r, z) - и (r, φ) -системах координат.

Иногда требуется решить уравнение Пуассона или более общее эллиптическое уравнение на поверхности цилиндра радиуса R . В этом случае

$$L_{\varphi z} u = \frac{1}{R} \frac{\partial}{\partial \varphi} \left(\frac{\bar{k}_{22}}{R} \frac{\partial u}{\partial \varphi} + \bar{k}_{33} \frac{\partial u}{\partial z} \right) + \frac{\partial}{\partial z} \left(\frac{\bar{k}_{32}}{R} \frac{\partial u}{\partial \varphi} + \bar{k}_{33} \frac{\partial u}{\partial z} \right) = -f(\varphi, z), \quad (14)$$

а уравнение (7) без смешанных производных принимает вид

$$L_{\varphi z} u = \frac{1}{R^2} \frac{\partial}{\partial \varphi} \left(\bar{k}_2 \frac{\partial u}{\partial \varphi} \right) + \frac{\partial}{\partial z} \left(\bar{k}_3 \frac{\partial u}{\partial z} \right) = -f(\varphi, z). \quad (14')$$

Отметим, что замена $\varphi' = R\varphi$ позволяет свести эти уравнения к обычным эллиптическим уравнениям с переменными коэффициентами.

2. Краевые задачи для уравнений в цилиндрической системе координат. Рассмотрим сначала осесимметрический случай. Так как решение не зависит от угла φ , то в цилиндрических координатах (r, z) область, где ищется решение, есть прямоугольник $\bar{G} = \{l_1 \leq r \leq L_1, l_3 \leq z \leq L_3, l_1 \geq 0\}$. Если исходная область представляет собой кольцевой (полый) цилиндр, то $l_1 > 0$.

Поставим краевые задачи для уравнения (10) в прямоугольнике \bar{G} . В области G задается уравнение (10), на сторонах $r = L_1$, $z = l_3$ и $z = L_3$ задается одно из краевых условий первого, второго или третьего рода. Например, краевые условия третьего рода имеют вид

$$\begin{aligned} -\bar{k}_{11} \frac{\partial u}{\partial r} - \bar{k}_{13} \frac{\partial u}{\partial z} &= \kappa_1^+ u - g_1^+(z), \quad r = L_1, \\ \bar{k}_{31} \frac{\partial u}{\partial r} + \bar{k}_{33} \frac{\partial u}{\partial z} &= \kappa_3^- u - g_3^-(r), \quad z = l_3, \\ -\bar{k}_{31} \frac{\partial u}{\partial r} - \bar{k}_{33} \frac{\partial u}{\partial z} &= \kappa_3^+ u - g_3^+(r), \quad z = L_3. \end{aligned} \quad (15)$$

При $l_1 = 0$ уравнение (10) имеет особенность на оси $r = 0$. В этом случае обычно интересуются ограниченным решением. Если $l_1 > 0$, то на стороне $r = l_1$ может быть задано одно из краевых условий первого, второго или третьего рода. Например, краевое условие третьего рода имеет вид

$$\bar{k}_{11} \frac{\partial u}{\partial r} + \bar{k}_{13} \frac{\partial u}{\partial z} = \kappa_1^- u - g_1^-(z), \quad r = l_1 > 0. \quad (16)$$

Если $l_1 = 0$, то ограниченное решение выделяется условием

$$\lim_{r \rightarrow 0} r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \bar{k}_{13} \frac{\partial u}{\partial z} \right) = 0. \quad (17)$$

В условиях (15), (16) $\kappa_1^\pm(z)$ и $\kappa_3^\pm(r)$ неотрицательные функции. Если на границе прямоугольника \bar{G} заданы краевые условия второго рода ($\kappa_\alpha^\pm \equiv 0$), то задача (10), (15), (16) разрешима лишь при выполнении условия

$$\begin{aligned} \int_{l_1}^{L_1} \int_{l_3}^{L_3} r f(r, z) dr dz + \int_{l_3}^{L_3} [L_1 g_1^+(z) + l_1 g_1^-(z)] dz + \\ + \int_{l_1}^{L_1} r [g_3^+(r) + g_3^-(r)] dr = 0. \end{aligned} \quad (18)$$

В этом случае решение не единственно и определяется с точностью до постоянной, т. е. $u(r, z) = u_0(r, z) + \text{const}$, где $u_0(r, z)$ — какое-либо решение.

Рассмотрим теперь уравнение (14) на поверхности цилиндра. В координатах (φ, z) область, где ищется решение, есть прямоугольник $\bar{G} = \{l_2 \leq \varphi \leq L_2, l_3 \leq z \leq L_3, L_2 - l_2 \leq 2\pi\}$.

На сторонах $z = l_3$ и $z = L_3$ могут быть заданы краевые условия первого, второго или третьего рода, например

$$\begin{aligned} \frac{1}{R} \bar{k}_{32} \frac{\partial u}{\partial \varphi} + \bar{k}_{33} \frac{\partial u}{\partial z} &= \kappa_3^- u - g_3^-(\varphi), \quad z = l_3, \\ -\frac{1}{R} \bar{k}_{32} \frac{\partial u}{\partial \varphi} - \bar{k}_{33} \frac{\partial u}{\partial z} &= \kappa_3^+ u - g_3^+(\varphi), \quad z = L_3. \end{aligned} \quad (19)$$

Такого рода краевые условия могут быть заданы на сторонах $\varphi = l_2$ и $\varphi = L_2$, если поверхность не замкнута ($L_2 - l_2 < 2\pi$). Например, краевые условия третьего рода имеют вид

$$\begin{aligned} \frac{1}{R^2} \bar{k}_{22} \frac{\partial u}{\partial \varphi} + \frac{1}{R} \bar{k}_{23} \frac{\partial u}{\partial z} &= \kappa_2^- u - g_2^-(z), \quad \varphi = l_2, \\ -\frac{1}{R^2} \bar{k}_{22} \frac{\partial u}{\partial \varphi} - \frac{1}{R} \bar{k}_{23} \frac{\partial u}{\partial z} &= \kappa_2^+ u - g_2^+(z), \quad \varphi = L_2. \end{aligned} \quad (20)$$

Здесь $\kappa_2^\pm(z) \geq 0$ и $\kappa_3^\pm(\varphi) \geq 0$.

Условие разрешимости задачи (14), (19), (20) при $\kappa_\alpha^\pm \equiv 0$ имеет вид

$$\int_{l_1}^{L_2} \int_{l_3}^{L_3} f(\varphi, z) d\varphi dz + \int_{l_3}^{L_3} [g_2^+(z) + g_2^-(z)] dz + \int_{l_1}^{L_2} [g_3^+(\varphi) + g_3^-(\varphi)] d\varphi = 0.$$

Если поверхность замкнута ($L_2 - l_2 = 2\pi$), то стороны $\varphi = l_2$ и $\varphi = L_2$ отождествляются и ставится задача отыскания периодического с периодом 2π решения уравнения (14), удовлетворяющего на сторонах $z = l_3$ и $z = L_3$ одному из перечисленных выше условий. Если при этом в условиях (19) $\kappa_3^\pm \equiv 0$, то условие разрешимости (с точностью до постоянной) указанной задачи имеет вид

$$\int_{l_1}^{L_2} \int_{l_3}^{L_3} f(\varphi, z) d\varphi dz + \int_{l_3}^{L_3} [g_3^+(\varphi) + g_3^-(\varphi)] d\varphi = 0.$$

Сформулируем теперь постановки краевых задач для уравнения (12), заданного в полярной системе координат, в случае, когда рассматриваемая область в декартовых переменных есть круг, кольцо или кольцевой сектор. Указанным областям в (r, φ) -координатах соответствует прямоугольник $\bar{G} = \{l_1 \leq r \leq L_1, l_2 \leq \varphi \leq L_2, l_1 \geq 0, L_2 - l_2 \leq 2\pi\}$.

Пусть сначала исходная область есть круг. Уравнение (12) задается в G ; при $r = L_1$ ставится одно из краевых условий первого, второго или третьего рода. Например, краевое условие третьего рода имеет вид

$$-\bar{k}_{11} \frac{\partial u}{\partial r} - \frac{\bar{k}_{12}}{r} \frac{\partial u}{\partial \varphi} = \kappa_1^+ u - g_1^+(\varphi), \quad r = L_1. \quad (21)$$

Для корректности задачи (12), (21) необходимо наложить дополнительное условие в центре круга. Обычно ищут ограниченное при $r=0$ решение. Это решение удовлетворяет условию

$$\lim_{r \rightarrow 0} r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \frac{\bar{k}_{12}}{r} \frac{\partial u}{\partial \varphi} \right) = 0. \quad (22)$$

Ввиду того, что в полярной системе координат точка $r=0$ плоскости (x_1, x_2) имеет произвольную координату φ , то все точки стороны прямоугольника \bar{G} при $r=0$ отождествляются. При этом $u(0, \varphi) = u_0 = \text{const}$ для $l_1 \leq \varphi \leq L_2$ в силу непрерывности решения.

Далее, стороны $\varphi = l_2$ и $\varphi = L_2$ отождествляются и ставится задача отыскания периодического с периодом 2π решения уравнения (12), удовлетворяющего перечисленным выше условиям.

В случае, когда при $r=L_1$ задано краевое условие второго рода (21) с $\kappa_1^+(\varphi) \equiv 0$, то решение задачи существует, если выполнено условие

$$\int_0^{2\pi} \int_0^{L_1} r f(r, \varphi) dr d\varphi + L_1 \int_0^{2\pi} g_1^+(\varphi) d\varphi = 0. \quad (23)$$

Решение при этом не единственно и определено с точностью до постоянной.

Пусть теперь исходная область есть кольцо, т. е. $l_1 > 0$. В этом случае ищется периодическое с периодом 2π решение уравнения (12), удовлетворяющее на сторонах $r=l_1$ и $r=L_1$ одному из краевых условий первого, второго или третьего рода. Приведем вид краевого условия третьего рода на внутренней стороне кольца

$$\bar{k}_{11} \frac{\partial u}{\partial r} + \frac{\bar{k}_{12}}{r} \frac{\partial u}{\partial \varphi} = \kappa_1^- u - g_1^-(\varphi), \quad r = l_1, \quad (24)$$

где $\kappa_1^-(\varphi) \geq 0$.

Если заданы краевые условия второго рода (21), (24) с $\kappa_1^\pm(\varphi) \equiv 0$, то решение поставленной задачи существует, если выполнено условие

$$\int_0^{2\pi} \int_{l_1}^{L_1} r f(r, \varphi) dr d\varphi + \int_0^{2\pi} [L_1 g_1^+(\varphi) + l_1 g_1^-(\varphi)] d\varphi = 0. \quad (25)$$

В этом случае решение определено с точностью до постоянной.

Если область есть кольцевой сектор ($l_1 > 0$, $L_2 - l_2 < 2\pi$), то ставится задача о нахождении решения уравнения (12), удовлетворяющего на сторонах прямоугольника \bar{G} одному из условий первого, второго или третьего рода, в частности условиям

(21), (24) при $r = L_1$ и $r = l_1$ и краевым условиям третьего рода

$$\begin{aligned} \bar{k}_{21} \frac{\partial u}{\partial r} + \frac{\bar{k}_{22}}{r} \frac{\partial u}{\partial \varphi} &= \kappa_2^- u - g_2^-(r), \quad \varphi = l_2, \\ -\bar{k}_{21} \frac{\partial u}{\partial r} - \frac{\bar{k}_{22}}{r} \frac{\partial u}{\partial \varphi} &= \kappa_2^+ u - g_2^+(r), \quad \varphi = L_2, \end{aligned} \quad (26)$$

при $\varphi = l_2$ и $\varphi = L_2$, $\kappa_2^\pm(r) \geq 0$.

Если заданы краевые условия второго рода (21), (24), (26) с $\kappa_1^\pm(\varphi) \equiv 0$, $\kappa_2^\pm(r) \equiv 0$, то решение задачи существует, если выполнено условие

$$\int_{l_2}^{L_2} \int_{l_1}^{L_1} r f(r, \varphi) dr d\varphi + \int_{l_2}^{L_2} (L_1 g_1^+ + l_1 g_1^-) d\varphi + \int_{l_1}^{L_1} (g_2^- + g_2^+) dr = 0. \quad (27)$$

При этом решение не единственное и определяется с точностью до постоянной.

§ 2. Решение разностных задач в цилиндрической системе координат

1. Разностные схемы без смешанных производных в осесимметрическом случае. Рассмотрим краевые задачи для эллиптических уравнений, не содержащих смешанные производные, в цилиндрической системе координат в случае осевой симметрии.

В прямоугольнике $\bar{G} = \{l_1 \leq r \leq L_1, l_3 \leq z \leq L_3, l_1 \geq 0\}$ требуется найти решение уравнения

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r k_1 \frac{\partial u}{\partial r} \right) + \frac{\partial}{\partial z} \left(k_3 \frac{\partial u}{\partial z} \right) - qu = -f(r, z), \quad (r, z) \in G, \quad (1)$$

удовлетворяющее на границе прямоугольника \bar{G} следующим краевым условиям:

1) на стороне $r = l_1$, $l_3 \leq z \leq L_3$,

$$u(r, z) = g_1^-(z), \quad \text{если } l_1 > 0, \quad (2)$$

или

$$k_1 \frac{\partial u}{\partial r} = \kappa_1^- u - g_1^-(z), \quad \text{если } l_1 > 0, \quad (3)$$

$$\lim_{r \rightarrow 0} r k_1 \frac{\partial u}{\partial r} = 0, \quad \text{если } l_1 = 0;$$

2) на стороне $r = L_1$, $l_3 \leq z \leq L_3$,

$$u(r, z) = g_1^+(z) \quad (4)$$

или

$$-k_1 \frac{\partial u}{\partial r} = \kappa_1^+ u - g_1^+(z); \quad (5)$$

3) на стороне $z = l_s$, $l_1 \leq r \leq L_1$,

$$u(r, z) = g_s^-(r) \quad (6)$$

или

$$k_s \frac{\partial u}{\partial z} = \kappa_s^- u - g_s^-(r); \quad (7)$$

4) на стороне $z = L_s$, $l_1 \leq r \leq L_1$,

$$u(r, z) = g_s^+(r) \quad (8)$$

или

$$-k_s \frac{\partial u}{\partial z} = \kappa_s^+ u - g_s^+(r). \quad (9)$$

Предполагается, что коэффициенты удовлетворяют условиям

$$k_1(r, z) \geq c_1 > 0, \quad k_s(r, z) \geq c_s > 0, \quad q(r, z) \geq 0,$$

$$\kappa_1^\pm(z) \geq 0, \quad \kappa_s^\pm(r) \geq 0.$$

В случае, когда $q \equiv 0$ и $\kappa_\alpha^\pm \equiv 0$ в краевых условиях (3), (5), (7), (9) или $l_1 = 0$ и заданы краевые условия второго рода (5), (7), (9), требуем выполнения условия разрешимости (см. (18) § 1).

Будем рассматривать любые комбинации краевых условий (2)–(9). Построим разностные схемы, соответствующие указанным краевым задачам.

Введем в области \bar{G} произвольную неравномерную прямоугольную сетку

$$\bar{\omega} = \{(r_i, z_k) \in \bar{G}, \quad r_i = r_{i-1} + h_1(i), \quad 1 \leq i \leq N_1, \quad r_0 = l_1, \quad r_{N_1} = L_1, \\ z_k = z_{k-1} + h_s(k), \quad 1 \leq k \leq N_s, \quad z_0 = l_s, \quad z_{N_s} = L_s\},$$

определим средние шаги

$$\bar{h}_\alpha(m) = \begin{cases} 0,5h_\alpha(1), & m = 0, \\ 0,5[h_\alpha(m) + h_\alpha(m+1)], & 1 \leq m \leq N_\alpha - 1, \\ 0,5h_\alpha(N_\alpha), & m = N_\alpha, \quad \alpha = 1, 3, \end{cases}$$

и сеточную функцию одного переменного

$$\rho(i) = r_i, \quad 1 \leq i \leq N_1, \quad \rho(0) = \begin{cases} \frac{1}{4}h_1(1), & l_1 = 0, \\ l_1, & l_1 > 0. \end{cases}$$

В простейшем случае непрерывных коэффициентов k_1 , k_s , q и f коэффициенты разностной схемы будем определять по формулам

$$a_1(i, k) = \bar{r}_i k_1(\bar{r}_i, \bar{z}_k), \quad a_s(i, k) = k_s(r_i, \bar{z}_k), \\ d(i, k) = q(r_i, z_k), \quad \varphi(i, k) = f(r_i, z_k),$$

где $\bar{r}_i = r_i - 0,5h_1(i)$, $\bar{z}_k = z_k - 0,5h_s(k)$.

Используя введенные обозначения, аппроксимируем (1) разностным уравнением

$$\frac{1}{\rho} (a_1 y_r)_r + (a_3 y_z)_z - dy = -\varphi, \quad 1 \leq i \leq N_1 - 1, \quad 1 \leq k \leq N_3 - 1. \quad (10)$$

Краевые условия первого рода (2), (4), (6), (8) аппроксимируем точно:

$$y(0, k) = g_1^-(z_k), \quad 0 \leq k \leq N_3, \quad (11)$$

$$y(N_1, k) = g_1^+(z_k), \quad 0 \leq k \leq N_3, \quad (12)$$

$$y(i, 0) = g_3^-(r_i), \quad 0 \leq i \leq N_1, \quad (13)$$

$$y(i, N_3) = g_3^+(r_i), \quad 0 \leq i \leq N_1. \quad (14)$$

Разностный аналог краевых условий (3) имеет вид

$$\frac{a_1^{+1}}{\rho h_1} y_r + (a_3 y_z)_z - \left(d + \frac{\kappa_1^-}{h_1} \right) y = -\varphi - \frac{g_1^-}{h_1}, \quad i = 0, \quad (15)$$

где $1 \leq k \leq N_3 - 1$ и $\kappa_1^- = g_1^- = 0$, если $i_1 = 0$. Краевые условия (5), (7), (9) аппроксимируются следующим образом:

$$-\frac{a_1}{\rho h_1} y_r + (a_3 y_z)_z - \left(d + \frac{\kappa_1^+}{h_1} \right) y = -\varphi - \frac{g_1^+}{h_1}, \quad i = N_1, \quad (16)$$

где $1 \leq k \leq N_3 - 1$,

$$\frac{1}{\rho} (a_1 y_r)_r + \frac{a_3^{+1}}{h_3} y_z - \left(d + \frac{\kappa_3^-}{h_3} \right) y = -\varphi - \frac{g_3^-}{h_3}, \quad k = 0, \quad (17)$$

$$\frac{1}{\rho} (a_1 y_r)_r - \frac{a_3}{h_3} y_z - \left(d + \frac{\kappa_3^+}{h_3} \right) y = -\varphi - \frac{g_3^+}{h_3}, \quad k = N_3, \quad (18)$$

где $1 \leq i \leq N_1 - 1$. Здесь использованы обозначения $a_1^{+1} = a_1(i+1, k)$, $a_3^{+1} = a_3(i, k+1)$.

Если на пересекающихся сторонах прямоугольника \bar{G} заданы краевые условия третьего рода, то в угловых узлах сетки $\bar{\omega}$ ставятся краевые условия

$$\frac{a_1^{+1}}{\rho h_1} y_r + \frac{a_3^{+1}}{h_3} y_z - \left(d + \frac{\kappa_1^-}{h_1} + \frac{\kappa_3^-}{h_3} \right) y = -\varphi - \frac{g_1^-}{h_1} - \frac{g_3^-}{h_3}, \quad i = k = 0, \quad (19)$$

$$-\frac{a_1}{\rho h_1} y_r + \frac{a_3^{+1}}{h_3} y_z - \left(d + \frac{\kappa_1^+}{h_1} + \frac{\kappa_3^+}{h_3} \right) y = -\varphi - \frac{g_1^+}{h_1} - \frac{g_3^+}{h_3}, \quad i = N_1, k = 0, \quad (20)$$

$$\frac{a_1^{+1}}{\rho h_1} y_r - \frac{a_3}{h_3} y_z - \left(d + \frac{\kappa_1^-}{h_1} + \frac{\kappa_3^+}{h_3} \right) y = -\varphi - \frac{g_1^-}{h_1} - \frac{g_3^+}{h_3}, \\ i = 0, k = N_3, \quad (21)$$

$$-\frac{a_1}{\rho h_1} y_r - \frac{a_3}{h_3} y_z - \left(d + \frac{\kappa_1^+}{h_1} + \frac{\kappa_3^+}{h_3} \right) y = -\varphi - \frac{g_1^+}{h_1} - \frac{g_3^+}{h_3}, \\ i = N_1, k = N_3. \quad (22)$$

Как и раньше, если $l_1 = 0$, то в (19) и (21) следует положить $\mathbf{x}_1^- = g_1^- = 0$.

Заметим, что разностная задача (10), (15)–(22) с краевыми условиями третьего рода на каждой стороне прямоугольника \bar{G} может быть записана в компактном виде

$$\begin{aligned} \Lambda y &= -f, \quad 0 \leq i \leq N_1, \quad 0 \leq k \leq N_3, \\ \Lambda &= \Lambda_1 + \Lambda_3, \quad f = \varphi + \varphi_1/\hbar_1 + \varphi_3/\hbar_3, \end{aligned} \quad (23)$$

где

$$\varphi_1(i, k) = \begin{cases} g_1^-, & i = 0, \\ 0, & 1 \leq i \leq N_1 - 1, \\ g_1^+, & i = N_1, \end{cases} \quad \varphi_3(i, k) = \begin{cases} g_3^-, & k = 0, \\ 0, & 1 \leq k \leq N_3 - 1, \\ g_3^+, & k = N_3, \end{cases} \quad (24)$$

а разностные операторы Λ_1 и Λ_3 задаются формулами

$$\Lambda_1 y = \begin{cases} \frac{a_1^{+1}}{\rho \hbar_1} y_r - \left(d_1 + \frac{\kappa_1^-}{\hbar_1} \right) y, & i = 0, \\ \frac{1}{\rho} (a_1 y_r)_r - d_1 y, & 1 \leq i \leq N_1 - 1, \\ -\frac{a_1}{\rho \hbar_1} y_r - \left(d_1 + \frac{\kappa_1^+}{\hbar_1} \right) y, & i = N_1, \quad 0 \leq k \leq N_3, \end{cases} \quad (25)$$

$$\Lambda_3 y = \begin{cases} \frac{a_3^{+1}}{\hbar_3} y_z - \left(d_3 + \frac{\kappa_3^-}{\hbar_3} \right) y, & k = 0, \\ (a_3 y_z)_z - d_3 y, & 1 \leq k \leq N_3 - 1, \\ -\frac{a_3}{\hbar_3} y_z - \left(d_3 + \frac{\kappa_3^+}{\hbar_3} \right) y, & k = N_3, \quad 0 \leq i \leq N_1. \end{cases} \quad (26)$$

Здесь $d_1 + d_3 = d$, $d_1 \geq 0$ и $d_3 \geq 0$.

Найдем условия разрешимости разностной схемы (23) в случае $d \equiv 0$ и $\kappa_\alpha^\pm \equiv 0$, $\alpha = 1, 2$.

В пространстве H сеточных функций, заданных на $\bar{\omega}$, определим скалярное произведение по формуле

$$(u, v) = \sum_{i=0}^{N_1} \sum_{k=0}^{N_3} u(i, k) v(i, k) \rho(i) \hbar_1(i) \hbar_3(k). \quad (27)$$

Определим операторы A_1 и A_3 , действующие в H , полагая $A_\alpha = -\Lambda_\alpha$, $\alpha = 1, 3$. Тогда разностную схему (23) можно записать в виде операторного уравнения

$$Au = f, \quad A = A_1 + A_3. \quad (28)$$

Используя первую разностную формулу Грина, найдем для случая $d \equiv 0$ и $\kappa_\alpha^\pm \equiv 0$, что

$$(Au, v) = \sum_{i=1}^{N_1} \sum_{k=0}^{N_3} h_1(i) \bar{h}_3(k) (a_1 u_r v_r)_{ik} + \\ + \sum_{i=0}^{N_1} \sum_{k=1}^{N_3} \bar{h}_1(i) h_3(k) \rho(i) (a_3 u_z v_z)_{ik} = (u, Av).$$

Следовательно, оператор A самосопряжен в H и неотрицателен, причем $(Au, u) = 0$ лишь в случае, когда $u(i, k) \equiv \text{const}$ или $u(i, k) \equiv 0$. Отсюда в силу неравенства Коши—Буняковского

$$(Au, u)^2 \leq (Au, Au)(u, u)$$

следует, что $Au = 0$ для $u \neq 0$, если u есть константа на $\bar{\omega}$. Таким образом, ядро оператора \bar{A} состоит из сеточных функций, равных постоянным на сетке $\bar{\omega}$. Поэтому задача (28) разрешима, если выполнено условие $(f, 1) = 0$ или, в силу определения f , — условие

$$\sum_{i=0}^{N_1} \sum_{k=0}^{N_3} \rho \Phi \bar{h}_1 \bar{h}_2 + \sum_{k=0}^{N_3} \bar{h}_3 (\rho g^-_1 + \rho g^+_1) + \sum_{i=0}^{N_1} \bar{h}_1 \rho [g^-_3 + g^+_3] = 0. \quad (29)$$

Условие (29) есть разностный аналог условия (18) разрешимости дифференциальной задачи, соответствующей разностной задаче (23).

Если условие (29) выполнено, то решение задачи (23) в случае $d \equiv 0$ и $\kappa_\alpha^\pm \equiv 0$ существует, но не единственно, два любых решения отличаются на постоянную. Поэтому одно из возможных решений можно выделить, фиксируя значение $u(i, k)$ в каком-либо узле сетки $\bar{\omega}$.

2. Прямые методы. Рассмотрим случай, для которого разностные задачи (10)–(22) могут быть решены одним из прямых методов, изложенных в главах III и IV.

Пусть коэффициенты k_1 , k_3 и q уравнения (1) не зависят от z , т. е. $k = k_1(r)$, $k_3 = k_3(r)$, $q = q(r)$, в краевых условиях третьего рода (3), (5) коэффициенты κ_1^+ и κ_1^- постоянны, а в условиях (7), (9) $\kappa_3^- = \kappa_3^+ \equiv 0$.

Допускаются любые комбинации краевых условий (2)–(9). Предполагается, что сетка $\bar{\omega}$ равномерна по z , т. е. $h_3(k) \equiv h_3$, и может быть неравномерной по r . При указанных предположениях разностные задачи (10)–(22) могут быть решены либо методом полной редукции, либо комбинированным методом неполной редукции и разделения переменных.

Проиллюстрируем возможность применения прямых методов на примере, в котором на сторонах $r = l_1$ и $r = L_1$ заданы краевые условия третьего (второго) рода (3), (5), а при $z = l_3$ и

$z = L_3$ — второго рода. Другие комбинации краевых условий рассматриваются аналогично.

Разностная схема, соответствующая поставленной задаче, имеет вид (23). В силу сделанных выше предположений коэффициенты разностной схемы определяются по формулам (ср. п. 1) $a_1 = a_1(i) = \bar{r}_i k_1(\bar{r}_i)$, $a_3 = a_3(i) = k_3(r_i)$, $d = d(i) = q(r_i)$, так что $a_3^{+1} = a_3$. В определении (25) разностного оператора Λ_1 выберем $d_1 = d$, а в формулах (26), задающих оператор Λ_3 , положим $\kappa_3^- = \kappa_3^+ = 0$, $d_3 = 0$. Так как сетка $\bar{\omega}$ равномерна по z , то в (26) разностное выражение $(a_3 y_z)_{\bar{z}}$ следует заменить на $a_3 y_{zz}$.

Сведем теперь разностную задачу (23) к системе трехточечных векторных уравнений. Для этого введем вектор неизвестных

$$\mathbf{Y}_k = (y(0, k), y(1, k), \dots, y(N_1, k)), \quad 0 \leq k \leq N_3,$$

содержащий значения искомой сеточной функции на k -й строке сетки $\bar{\omega}$, и вектор правых частей

$$\mathbf{F}_k = (\theta_0 f(0, k), \theta_1 f(1, k), \dots, \theta_{N_1} f(N_1, k)), \quad 0 \leq k \leq N_3,$$

где $\theta_i = h_3^2/a_3(i)$, $0 \leq i \leq N_1$. Определим квадратную матрицу C , полагая

$$CY_k = ((2E - \theta_0 \Lambda_1) y(0, k), \dots, (2E - \theta_{N_1} \Lambda_1) y(N_1, k)).$$

Используя эти обозначения, разностную задачу (23) запишем в векторном виде

$$\begin{aligned} CY_0 - 2Y_1 &= \mathbf{F}_0, & k = 0, \\ -Y_{k-1} + CY_k - Y_{k+1} &= \mathbf{F}_k, & 1 \leq k \leq N_3 - 1, \\ 2Y_{N_3-1} + CY_{N_3} &= \mathbf{F}_{N_3}, & k = N_3. \end{aligned} \quad (30)$$

Для того чтобы убедиться в этом, достаточно умножить каждое уравнение схемы (23) на $(-\theta_i)$ и перейти к векторной записи.

Напомним, что метод полной редукции для системы (30) был построен в п. 1 § 4 гл. III. Комбинированный метод неполной редукции и разделения переменных был рассмотрен в п. 2 § 3 гл. IV. Здесь отличие от рассмотренных в главах III и IV примеров заключается в ином определении оператора Λ_1 . Но так как разностный оператор Λ_1 по-прежнему трехточечный, то это отличие не влияет ни на конструкцию этих методов, ни на характер зависимости числа арифметических операций от числа узлов сетки $\bar{\omega}$. Если $N_3 = 2^n$, то число арифметических операций для указанных методов оценивается величиной $O(N_1 N_3 \log_2 N_3)$.

В заключение отметим, что применение комбинированного метода с выделением одного из решений в вырожденном случае ($d \equiv 0$, $\kappa_1^- = \kappa_2^+ \equiv 0$) подробно описано в п. 2 § 4 гл. XII для декартовой системы координат.

3. Метод переменных направлений. Рассмотрим теперь частный случай задачи (1)–(9), для которого $k_1 = k_1(r)$, $k_3 = k_3(z)$, $q = \text{const}$, $\kappa_\alpha^\pm = \text{const}$, $\alpha = 1, 3$, а на сторонах прямоугольника \bar{G}

задана любая комбинация краевых условий (2)–(9). В этом случае переменные в задаче (1)–(9) разделяются.

Предполагается, что сетка ω —произвольная, неравномерная по каждому направлению. При сделанных предположениях разностные задачи (10)–(22) могут быть решены методом переменных направлений с оптимальным набором итерационных параметров, который приведен в главе XI для случая декартовой системы координат.

Проиллюстрируем применение этого метода на примере, в котором на сторонах прямоугольника \bar{G} заданы краевые условия третьего рода (3), (5), (7), (9). Разностная схема, соответствующая задаче (1), (3), (5), (7), (9), имеет вид (23), где операторы A_1 и A_3 определены в (25), (26), а коэффициенты a_1 , a_3 , d_1 и d_3 задаются формулами $a_1(i) = \bar{r}_i k_1(\bar{r}_i)$, $a_3(k) = k_3(\bar{z}_k)$, $d_1 = d_3 = 0,5d$, $d = q$.

В п. 1 было показано, что разностная задача (23) может быть записана в виде операторного уравнения (28)

$$Au = f, \quad A = A_1 + A_3$$

в гильбертовом пространстве H сеточных функций, заданных на $\bar{\omega}$. Укажем основные свойства операторов A_1 и A_3 :

- 1) операторы A_1 и A_3 перестановочны, $A_1 A_3 = A_3 A_1$;
- 2) A_1 и A_3 —самосопряженные операторы, $(A_\alpha u, v) = (u, A_\alpha v)$;
- 3) операторы A_1 и A_3 —неотрицательные ограниченные операторы, т. е. для любого $u \in H$ выполнены неравенства

$$\delta_\alpha(u, u) \leq (A_\alpha u, u) \leq \Delta_\alpha(u, u), \quad \delta_\alpha \geq 0, \quad \Delta_\alpha > 0, \quad \alpha = 1, 3. \quad (31)$$

Действительно, перестановочность операторов A_1 и A_3 следует из структуры операторов A_1 и A_3 и предположения относительно коэффициентов k_1 , k_3 , q и κ_α^\pm .

Далее, используя определение (27) скалярного произведения в H и разностные формулы Грина, получим для A_1 и любых $u, v \in H$ равенство

$$(A_1 u, v) = \sum_{i=1}^{N_1} \sum_{k=0}^{N_3} h_1(i) \bar{h}_3(k) (a_1 u_{\bar{r}} v_{\bar{r}})_{ik} + d_1(u, v) + \\ + \sum_{k=0}^{N_3} \bar{h}_3(k) [\kappa_1^- \rho u v|_{i=0} + \kappa_1^+ \rho u v|_{i=N_1}] \quad (32)$$

и аналогичное равенство для A_3

$$(A_3 u, v) = \sum_{k=1}^{N_3} \sum_{i=0}^{N_1} \rho(i) \bar{h}_1(i) h_3(k) (a_3 u_{\bar{z}} v_{\bar{z}})_{ik} + \\ + d_3(u, v) + \sum_{i=0}^{N_1} \rho(i) \bar{h}_1(i) [\kappa_3^- u v|_{k=0} + \kappa_3^+ u v|_{k=N_3}]. \quad (33)$$

Меняя местами u и v , убеждаемся в самосопряженности операторов A_1 и A_3 .

Если положить здесь $u=v$ и учесть условия $k_1 \geq c_1 > 0$, $k_3 \geq c_1 > 0$, $q \geq 0$, $\kappa_\alpha^\pm \geq 0$, $\alpha=1, 3$, то найдем, что операторы A_1 и A_3 неотрицательны, т. е. $(A_\alpha u, u) \geq 0$. Если выполнено условие

$$d_\alpha^2 + (\kappa_\alpha^-)^2 + (\kappa_\alpha^+)^2 \neq 0, \quad \alpha = 1, 3, \quad (34)$$

то соответствующее δ_α положительно. Пусть (34) выполнено.

Дадим оценку для δ_α снизу.

Из леммы 16 главы V для фиксированного i , $0 \leq i \leq N_1$, получим оценку

$$\begin{aligned} \delta_3 \sum_{k=0}^{N_3} h_3(k) u^2(i, k) &\leq \sum_{k=1}^{N_3} h_3(k) a_3(k) u_z^2(i, k) + \\ &+ d_3 \sum_{k=0}^{N_3} h_3(k) u^2(i, k) + \kappa_3^- u^2(i, 0) + \kappa_3^+ u^2(i, N_3), \end{aligned} \quad (35)$$

где $1/\delta_3 = \max_{0 \leq k \leq N_3} v(k)$, $v(k)$ есть решение краевой задачи

$$\begin{aligned} (a_3 v_z)_z - d_3 v &= -1, \quad 1 \leq k \leq N_3 - 1, \\ \frac{a_3^{+1}}{h_3} v_z - \left(d_3 + \frac{\kappa_3^-}{h_3} \right) v &= -1, \quad k = 0, \\ -\frac{a_3}{h_3} v_z - \left(d_3 + \frac{\kappa_3^+}{h_3} \right) v &= -1, \quad k = N_3. \end{aligned} \quad (36)$$

Так как выполнено условие (34), то решение задачи (36) существует и единственno. Умножая теперь (35) на $\rho(i) \bar{h}_1(i)$ и суммируя по i от 0 до N_1 , получим неравенство $\delta_3(u, u) \leq (A_3 u, u)$. Решая численно задачу (36), определим δ_3 . Итак, постоянная δ_3 найдена. Аналогично оценивается постоянная δ_1 : $1/\delta_1 = \max_{0 \leq i \leq N} \bar{v}(i)$, где $\bar{v}(i)$ — решение краевой задачи

$$\begin{aligned} \frac{1}{\rho} (a_1 \bar{v}_r)_r - d_1 \bar{v} &= -1, \quad 1 \leq i \leq N_1 - 1, \\ \frac{a_1^{+1}}{\rho \bar{h}_1} \bar{v}_r - \left(d_1 + \frac{\kappa_1^-}{\bar{h}_1} \right) \bar{v} &= -1, \quad i = 0, \\ -\frac{a_1}{\rho \bar{h}_1} \bar{v}_r - \left(d_1 + \frac{\kappa_1^+}{\bar{h}_1} \right) \bar{v} &= -1, \quad i = N_1. \end{aligned} \quad (37)$$

Получим теперь оценки для Δ_1 и Δ_3 . Из (33) при $u=v$ найдем

$$\begin{aligned} (A_3 u, u) &= \sum_{i=0}^{N_1} \rho(i) \bar{h}_1(i) \left[\sum_{k=1}^{N_3} h_3(k) a_3(k) u_z^2(i, k) + \right. \\ &\quad \left. + d_3 \sum_{k=0}^{N_3} h_3(k) u^2(i, k) + \kappa_3^- u^2(i, 0) + \kappa_3^+ u^2(i, N_3) \right]. \end{aligned}$$

Оценим выражение, стоящее в квадратных скобках. Из леммы 16 главы V получим

$$d_3 \sum_{k=0}^{N_3} u^2(i, k) \tilde{h}_3(k) + \kappa_3^- u^2(i, 0) + \kappa_3^+ u^2(i, N_3) \leq m_1 \left[\sum_{k=1}^{N_3} a_3(k) u_z^2(i, k) h_3(k) + \sum_{k=0}^{N_3} \tilde{h}_3(k) u^2(i, k) \right], \quad (38)$$

где $m_1 = \max_{0 \leq k \leq N_3} w(k)$, а $w(k)$ — решение краевой задачи

$$\begin{aligned} (a_3 w_z)_z - w &= -d_3, \quad 1 \leq k \leq N_3 - 1, \\ \frac{a_3^{+1}}{\tilde{h}_3} w_z - w &= -\left(d_3 + \frac{\kappa_3^-}{\tilde{h}_3}\right), \quad k = 0, \\ -\frac{a_3}{\tilde{h}_3} w_z - w &= -\left(d_3 + \frac{\kappa_3^+}{\tilde{h}_3}\right), \quad k = N_3. \end{aligned} \quad (39)$$

Используя лемму 17 главы V, будем иметь

$$\sum_{k=1}^{N_3} a_3(k) u_z^2(i, k) h_3(k) \leq m_2 \sum_{k=0}^{N_3} \tilde{h}_3(k) u^2(i, k), \quad (40)$$

где

$$m_2 = \max \left(\frac{a_3(N_3)}{\tilde{h}_3^2(N_3)}, \frac{a_3(1)}{\tilde{h}_3^2(0)}, \max_{1 \leq k \leq N_3 - 1} \frac{2}{\tilde{h}_3(k)} \left[\frac{a_3(k)}{\tilde{h}_3(k)} + \frac{a_3(k+1)}{\tilde{h}_3(k+1)} \right] \right).$$

Из (38) и (40) следует оценка

$$\begin{aligned} \sum_{k=1}^{N_3} h_3(k) a_3(k) u_z^2(i, k) + d_3 \sum_{k=0}^{N_3} \tilde{h}_3(k) u^2(i, k) + \kappa_3^- u^2(i, 0) + \\ + \kappa_3^+ u^2(i, N_3) \leq \Delta_3 \sum_{k=0}^{N_3} \tilde{h}_3(k) u^2(i, k), \quad \Delta_3 = m_1 + m_2(1 + m_1). \end{aligned}$$

Умножая полученное неравенство на $\rho(i) \tilde{h}_1(i)$ и суммируя его по i от 0 до N_1 , будем иметь оценку $(A_3 u, u) \leq \Delta_3(u, u)$.

Аналогичным образом находится Δ_1 : $\Delta_1 = \bar{m}_1 + \bar{m}_2(1 + \bar{m}_1)$, где $\bar{m}_1 = \max_{0 \leq i \leq N_1} \bar{w}(i)$, $\bar{w}(i)$ — решение краевой задачи

$$\begin{aligned} \frac{1}{\rho} (a_1 \bar{w}_z)_z - \bar{w} &= -d_1, \quad 1 \leq i \leq N_1 - 1, \\ \frac{a_1^{+1}}{\rho \tilde{h}_1} \bar{w}_z - \bar{w} &= -\left(d_1 + \frac{\kappa_1^-}{\tilde{h}_1}\right), \quad i = 0, \\ -\frac{a_1}{\rho \tilde{h}_1} \bar{w}_z - \bar{w} &= -\left(d_1 + \frac{\kappa_1^+}{\tilde{h}_1}\right), \quad i = N_1. \end{aligned} \quad (41)$$

причем

$$\begin{aligned} \bar{m}_1 = \max \left(\frac{a_1(N_1)}{\rho(N_1) \tilde{h}_1^2(N_1)}, \frac{a_1(1)}{\rho(0) \tilde{h}_1^2(0)}, \right. \\ \left. \max_{1 \leq i \leq N_1 - 1} \frac{2}{\rho(i) \tilde{h}_1(i)} \left[\frac{a_1(i)}{\tilde{h}_1(i)} + \frac{a_1(i+1)}{\tilde{h}_1(i+1)} \right] \right). \end{aligned}$$

Решая численно задачу (41), определим \bar{m}_1 и, следовательно, Δ_1 . Таким образом, постоянные δ_α и Δ_α , $\alpha = 1, 3$, фигурирующие в неравенствах (31), найдены.

Напомним, что итерационная схема метода переменных направлений для операторного уравнения (28) имеет вид (см. гл. XI)

$$\begin{aligned} B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k &= f, \quad k = 0, 1, \dots, \quad y_0 \in H, \\ B_k = (\omega_k^{(1)} E + A_1) (\omega_k^{(3)} E + A_3), \quad \tau_k &= \omega_k^{(1)} + \omega_k^{(3)}. \end{aligned} \quad (42)$$

В п. 4 § 1 гл. XI для итерационной схемы (42), операторы A_1 и A_3 , которой удовлетворяют перечисленным выше свойствам 1)–3), был построен оптимальный набор параметров $\omega_k^{(1)}$ и $\omega_k^{(2)}$, $k = 1, 2, \dots, n$. При использовании этого набора параметров относительная точность $\varepsilon > 0$ ($\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_D$, $D = A$, E) достигается, если выполнить $n \geq n_0(\varepsilon)$ итераций, где

$$n_0(\varepsilon) = \frac{1}{\pi^2} \ln \frac{4}{\eta} \ln \frac{4}{\varepsilon}, \quad \eta = \frac{1-a}{1+a}, \quad a = \sqrt{\frac{(\Delta_1 - \delta_1)(\Delta_3 - \delta_3)}{(\Delta_1 + \delta_3)(\Delta_3 + \delta_1)}}.$$

Набор оптимальных параметров $\omega_k^{(1)}$ и $\omega_k^{(2)}$ для случая второй краевой задачи ($d = 0$, $\kappa_\alpha^\pm = 0$) был построен в п. 1 § 4 гл. XII.

4. Решение уравнений, заданных на поверхности цилиндра.

Рассмотрим теперь метод решения разностных аналогов краевых задач для эллиптического уравнения без смешанных производных, заданного на поверхности цилиндра радиуса R . Ограничимся рассмотрением замкнутой по φ поверхности цилиндра, так как методы решения задач в случае незамкнутой поверхности ничем не отличаются от методов решения плоских задач в декартовых переменных.

Итак, в области $\bar{G} = \{l_2 \leq \varphi \leq L_2, l_3 \leq z \leq L_3, L_2 - l_2 = 2\pi\}$ ищется решение уравнения

$$\frac{1}{R^2} \frac{\partial}{\partial \varphi} \left(k_2 \frac{\partial u}{\partial \varphi} \right) + \frac{\partial}{\partial z} \left(k_3 \frac{\partial u}{\partial z} \right) - qu = -f(\varphi, z), \quad (\varphi, z) \in G, \quad (43)$$

периодическое по φ с периодом 2π , удовлетворяющее на сторонах $z = l_3$ и $z = L_3$ либо краевым условиям первого рода $u(\varphi, z) = g_3^-(\varphi)$ при $z = l_3$, $u(\varphi, z) = g_3^+(\varphi)$ при $z = L_3$, либо второго или третьего рода

$$\begin{aligned} k_3 \frac{\partial u}{\partial z} &= \kappa_3^- u - g_3^-(\varphi), \quad z = l_3, \\ -k_3 \frac{\partial u}{\partial z} &= \kappa_3^+ u - g_3^+(\varphi), \quad z = L_3, \end{aligned} \quad (44)$$

либо любой их комбинации. Предполагается, что коэффициенты удовлетворяют условиям

$$k_2(\varphi, z) \geq c_1 > 0, \quad k_3(\varphi, z) \geq c_1 > 0, \quad q(\varphi, z) \geq 0, \quad \kappa_3^\pm(\varphi) \geq 0.$$

В области \bar{G} введем произвольную неравномерную сетку
 $\bar{\omega} = \{(\varphi_j, z_k) \in \bar{G}, \varphi_j = \varphi_{j-1} + h_2(j), 1 \leq j \leq N_2, \varphi_0 = l_2,$
 $\varphi_{N_2} = L_2, z_k = z_{k-1} + h_3(k), 1 \leq k \leq N_3, z_0 = l_3, z_{N_3} = L_3\}$
и определим средний шаг

$$\bar{h}_2(j) = \begin{cases} 0,5 [h_2(1) + h_2(N_2)], & j=0, \\ 0,5 [h_2(j) + h_2(j+1)], & 1 \leq j \leq N_2-1. \end{cases} \quad (45)$$

Средний шаг $\bar{h}_3(k)$ определен выше.

Уравнение (43) с учетом условия периодичности аппроксимируем следующим образом:

$$(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} + (a_3 y_z)_{\hat{z}} - dy = -\psi, \quad 0 \leq j \leq N_2-1, \quad 1 \leq k \leq N_3-1, \quad (46)$$

где использованы соотношения $y(j, k) = y(N_2 + j, k), j = 0, -1, a_2(0, k) = a_2(N_2, k), h_2(0) = h_2(N_2)$, являющиеся следствием условия периодичности. В случае гладких коэффициентов k_2, k_3, q и f коэффициенты в уравнении (46) можно выбрать, например, так:

$$a_2(j, k) = \frac{1}{R^2} k_2 (\varphi_j - 0,5 \bar{h}_2(j), z_k), \quad d(j, k) = q(\varphi_j, z_k), \\ a_3(j, k) = k_3 (\varphi_j, z_k - 0,5 \bar{h}_3(k)), \quad \psi(j, k) = f(\varphi_j, z_k).$$

Краевые условия первого рода аппроксимируются точно

$$y(j, 0) = g_3^-(\varphi_j), \quad k=0, \quad y(j, N_3) = g_3^+(\varphi_j), \quad k=N_3 \quad (47)$$

для $0 \leq j \leq N_2-1$, а разностный аналог краевых условий (44) третьего рода имеет вид для $0 \leq j \leq N_2-1$:

$$(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} + \frac{a_3^{+1}}{\bar{h}_3} y_z - \left(d + \frac{\kappa_3^-}{\bar{h}_3} \right) y = -\psi - \frac{g_3^-}{\bar{h}_3}, \quad k=0, \\ (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - \frac{a_3^-}{\bar{h}_3} y_z - \left(d + \frac{\kappa_3^+}{\bar{h}_3} \right) y = -\psi - \frac{g_3^+}{\bar{h}_3}, \quad k=N_3. \quad (48)$$

В задаче (46), (47) неизвестными являются значения $y(j, k)$ для $0 \leq j \leq N_2-1, 1 \leq k \leq N_3-1$, а в задаче (46), (48) — для тех же значений j и $0 \leq k \leq N_3$.

Найдем условие разрешимости разностной задачи (46), (48) в случае, когда $d \equiv 0, \kappa_3^\pm \equiv 0$. Сначала запишем схему (46), (48) в виде

$$\Lambda y = -f, \quad 0 \leq j \leq N_2-1, \quad 0 \leq k \leq N_3, \\ \Lambda = \Lambda_2 + \Lambda_3, \quad f = \psi + \psi_3 / \bar{h}_3, \quad (49)$$

где разностный оператор Λ_3 определен в (26) с $d_3 = d$, оператор Λ_2 задается формулой $\Lambda_2 y = (a_2 y_{\bar{\varphi}})_{\hat{\varphi}}, 0 \leq j \leq N_2-1$,

$$\psi_3(j, k) = \begin{cases} g_3^-(\varphi_j), & k=0, \\ 0, & 1 \leq k \leq N_3-1, \\ g_3^+(\varphi_j), & k=N_3. \end{cases}$$

Пусть теперь $d \equiv 0$ и $\kappa_3^\pm \equiv 0$. Обозначим через H пространство сеточных функций, заданных на $\bar{\omega}^* = \{(\varphi_j, z_k) \in \bar{\omega}, 0 \leq j \leq N_2 - 1, 0 \leq k \leq N_3\}$, скалярное произведение в котором определим формулой

$$(u, v) = \sum_{j=0}^{N_2-1} \sum_{k=0}^{N_3} u(j, k) v(j, k) \tilde{h}_2(j) \tilde{h}_3(k).$$

Определим операторы A_2 и A_3 , действующие в H , равенствами: $A_3 = -\Lambda_3$, $A_2 y = -\Lambda_2 \bar{y}$, где $y(j, k) = \bar{y}(j, k)$ для $0 \leq j \leq N_2 - 1$, $0 \leq k \leq N_3$ и y удовлетворяет условию периодичности $y(j, k) = \bar{y}(N_2 + j, k)$, $j = 0, -1$.

Используя введенные обозначения, запишем разностную схему (49) в виде операторного уравнения

$$Au = f, \quad A = A_2 + A_3. \quad (50)$$

Учитывая условия периодичности, при помощи разностной формулы Грина получим

$$\begin{aligned} (Au, v) = -(\Lambda u, \bar{v}) &= \sum_{k=0}^{N_3} \sum_{j=0}^{N_2-1} \tilde{h}_3(k) h_2(j) (a_2 \bar{u}_\varphi \bar{v}_\varphi)_{jk} + \\ &+ \sum_{k=1}^{N_3} \sum_{j=0}^{N_2-1} \tilde{h}_2(j) h_3(k) (a_3 \bar{u}_z \bar{v}_z)_{jk} = (u, Av). \end{aligned}$$

Следовательно, оператор A самосопряжен в H . Кроме того, рассматривая значения (Au, u) , найдем, что ядро оператора A состоит из сеточных функций, принимающих на сетке ω^* постоянные значения. Поэтому решение разностной задачи (49) существует, если выполнено условие $(f, 1) = 0$. Подставляя сюда f из (49), получим

$$\sum_{j=0}^{N_2-1} \sum_{k=0}^{N_3} \tilde{h}_2(j) \tilde{h}_3(k) \psi(j, k) + \sum_{j=0}^{N_2-1} \tilde{h}_2(j) [g_3^-(\varphi_j) + g_3^+(\varphi_j)] = 0.$$

При выполнении этого условия решение разностной задачи (46), (48) при $d \equiv 0$ и $\kappa_3^\pm = 0$ существует и два любых ее решения отличаются на постоянную.

Рассмотрим случаи, когда решение разностных задач (46)–(48) может быть найдено прямыми методами, изложенными в главах III и IV.

Первый случай. Коэффициенты k_2 , k_3 и q уравнения (43) зависят только от φ , $\kappa_3^\pm = \text{const}$ и сетка ω равномерна по z . Разностная задача (46), (48) может быть записана в виде системы трехточечных векторных уравнений

$$\begin{aligned} (C + 2\alpha E) Y_0 - 2Y_1 &= F_0, & k = 0, \\ -Y_{k-1} + CY_k - Y_{k+1} &= F_k, & 1 \leq k \leq N_3 - 1, \\ -2Y_{N_3-1} + (C + 2\beta E) Y_{N_3} &= F_{N_3}, & k = N_3, \end{aligned} \quad (51)$$

где $N_3 = 2^n$, $n > 0$ — целое число,

$$\begin{aligned} \mathbf{Y}_k &= (y(0, k), y(1, k), \dots, y(N_2 - 1, k)), \\ \mathbf{F}_k &= (\theta_0 f(0, k), \theta_1 f(1, k), \dots, \theta_{N_2 - 1} f(N_2 - 1, k)), \\ CY_k &= ((2E - \theta_0 \Lambda_2) y(0, k), \dots, (2E - \theta_{N_2 - 1} \Lambda_2) y(N_2 - 1, k)) \end{aligned}$$

для $0 \leq k \leq N_3$. Оператор Λ_2 определен выше, $f(j, k)$ задано в (49) и $\theta_j = h_3^2/a_3(j)$, $\alpha = h_3 x_3^\pm$, $\beta = h_3 x_3^\mp$.

Напомним, что в п. 3 § 4 гл. III для решения задачи (51) при условии $\alpha^2 + \beta^2 \neq 0$ был построен метод полной редукции. Если $\alpha = \beta = 0$, но $d \neq 0$, то алгоритм метода изложен в п. 1 § 4 гл. III. Для последнего случая в п. 2 § 3 гл. IV был построен комбинированный метод неполной редукции и разделения переменных.

Второй случай. Коэффициенты k_2, k_3 и q зависят только от z , $x_3^\pm = \text{const}$ и сетка $\bar{\omega}$ равномерна по φ . Разностная задача (46), (48) записывается в виде системы трехточечных векторных уравнений

$$\begin{aligned} -Y_{N_2 - 1} + CY_0 - Y_1 &= \mathbf{F}_0, & j = 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= \mathbf{F}_j, & 1 \leq j \leq N_2 - 2, \\ -Y_{N_2 - 2} + CY_{N_2 - 1} - Y_0 &= \mathbf{F}_{N_2 - 1}, & j = N_2 - 1. \end{aligned} \quad (52)$$

Здесь $N_2 = 2^n$, $n > 0$ — целое число,

$$\begin{aligned} Y_j &= (y(j, 0), y(j, 1), \dots, y(j, N_3)), \\ \mathbf{F}_j &= (\theta_0 f(j, 0), \theta_1 f(j, 1), \dots, \theta_{N_3} f(j, N_3)), \\ CY_j &= ((2E - \theta_0 \Lambda_3) y(j, 0), \dots, (2E - \theta_{N_3} \Lambda_3) y(j, N_3)), \end{aligned}$$

где $0 \leq j \leq N_2 - 1$. Разностный оператор Λ_3 определен в (26) с $d_3 = d$ и $\theta_k = h_3^2/a_3(k)$, $0 \leq k \leq N_3$. Задача (52) может быть решена методом полной редукции, построенным в п. 2 § 4 гл. III или комбинированным методом, использующим алгоритм быстрого дискретного преобразования Фурье действительной периодической функции. Этот алгоритм построен в п. 4 § 1 гл. IV.

В каждом из рассмотренных случаев прямые методы реализуются с затратой $O(N_2 N_3 n)$ действий.

В заключение отметим, что если коэффициенты удовлетворяют условиям $k_2 = k_2(\varphi)$, $k_3 = k_3(z)$, $g = \text{const}$, $x_3^\pm = \text{const}$, а сетка неравномерна по каждому направлению, то для нахождения решения задачи (46), (48) можно использовать метод переменных направлений с оптимальным набором параметров:

$$\begin{aligned} B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k &= f, \quad k = 0, 1, \dots, y_0 \in H, \\ B_k &= (\omega_k^{(2)} E + A_2)(\omega_k^{(3)} E + A_3), \quad \tau_k = \omega_k^{(2)} + \omega_k^{(3)}. \end{aligned}$$

Здесь оператор $A_3 = -\Lambda_3$, $A_2 y = -\Lambda_2 \bar{y}$, где разностный оператор Λ_3 определен в (26) с $d_3 = 0,5d$, а $\Lambda_2 y = (a_2 y_\varphi)_\varphi - 0,5 dy$. Постоянные δ_α и Δ_α , являющиеся границами оператора A_α ,

оцениваются следующим образом: δ_3 и Δ_3 найдены в п. 3 § 2, постоянная δ_2 находится точно: $\delta_2 = 0,5d$, а в качестве Δ_2 можно взять

$$\Delta_2 = \max_{0 \leq i \leq N_2 - 1} \left[\frac{2}{h_2(i)} \left(\frac{a_2(j)}{h_2(j)} + \frac{a_2(j+1)}{h_2(j+1)} \right) + \frac{d}{2} \right].$$

§ 3. Решение разностных задач в полярной системе координат

1. Разностные схемы для уравнений в круге и кольце. Рассмотрим методы решения разностных схем для эллиптических уравнений без смешанных производных в полярной системе координат. Сначала изучим случай, когда область, где ищется решение, есть круг или кольцо в декартовой системе координат. В полярной системе координат указанным областям соответствует прямоугольник $\bar{G} = \{l_1 \leq r \leq L_1, l_2 \leq \varphi \leq L_2, l_1 \geq 0, L_2 - l_2 = 2\pi\}$. Требуется найти решение уравнения

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r k_1 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial}{\partial \varphi} \left(k_2 \frac{\partial u}{\partial \varphi} \right) - q u = -f, \quad (r, \varphi) \in G, \quad (1)$$

являющееся периодическим по φ с периодом 2π и удовлетворяющее на границе прямоугольника \bar{G} условиям:

1) при $r = L_1, l_2 \leq \varphi \leq L_2$ либо краевым условиям первого рода

$$u(r, \varphi) = g_1^+(\varphi), \quad (2)$$

либо второго или третьего рода

$$-k_1 \frac{\partial u}{\partial r} = \kappa_1^+ u - g_1^+(\varphi); \quad (3)$$

2) при $r = l_1 > 0, l_2 \leq \varphi \leq L_2$ либо краевым условиям первого рода

$$u(r, \varphi) = g_1^-(\varphi), \quad (4)$$

либо второго или третьего рода

$$k_1 \frac{\partial u}{\partial r} = \kappa_1^- u - g_1^-(\varphi); \quad (5)$$

при $r = l_1 = 0$ ставится условие

$$\lim_{r \rightarrow 0} r k_1 \frac{\partial u}{\partial r} = 0, \quad (6)$$

выделяющее ограниченное решение.

Предполагается, что коэффициенты удовлетворяют условиям $k_1(r, \varphi) \geq c_1 > 0, k_2(r, \varphi) \geq c_1 > 0, q(r, \varphi) \geq 0, \kappa_1^\pm(\varphi) \geq 0$.

Будем рассматривать любые комбинации краевых условий (2)–(5). Построим разностные схемы, соответствующие указанным краевым задачам.

В области \bar{G} введем произвольную неравномерную прямоугольную сетку

$$\bar{\omega} = \{(r_i, \varphi_j) \in \bar{G}, r_i = r_{i-1} + h_1(i), 1 \leq i \leq N_1, r_0 = l_1, \\ r_{N_1} = L_1, \varphi_j = \varphi_{j-1} + h_2(j), 1 \leq j \leq N_2, \varphi_0 = l_2, \varphi_{N_2} = L_2\}.$$

Средний шаг $\bar{h}_1(i)$ определен в п. 1 § 2, а шаг $\bar{h}_2(j)$ — в п. 4 § 2 формулой (45). Определим сеточную функцию $\rho(i)$:

$$\rho(i) = \begin{cases} l_1 + \frac{1}{4} h_1(1), & i = 0, \\ r_i + \frac{1}{4} [h_1(i+1) - h_1(i)], & 1 \leq i \leq N_1 - 1, \\ L_1 - \frac{1}{4} h_1(N_1), & i = N_1. \end{cases} \quad (7)$$

В простейшем случае непрерывных коэффициентов k_1, k_2, q и f коэффициенты разностной схемы будем определять по формулам

$$a_1(i, j) = \bar{r}_i k_1(\bar{r}_i, \varphi_j), \quad a_2(i, j) = k_2(r_i, \bar{\varphi}_j), \\ d(i, j) = q(r_i, \varphi_j), \quad \psi(i, j) = f(r_i, \varphi_j),$$

где $\bar{r}_i = r_i - 0,5h_1(i)$, $\bar{\varphi}_j = \varphi_j - 0,5h_2(j)$.

Используя введенные обозначения, аппроксимируем (1) разностным уравнением

$$\Lambda y = \frac{1}{\rho} (a_1 y_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2} (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - dy = -\psi, \\ 1 \leq i \leq N_1 - 1, 0 \leq j \leq N_2 - 1. \quad (8)$$

Здесь для компактности записи использованы соотношения

$$y(i, j) = y(i, N_2 + j), \quad j = 0, -1, \quad a_2(i, 0) = a_2(i, N_2), \\ h_2(0) = h_2(N_2), \quad (9)$$

являющиеся следствием условия периодичности.

Краевые условия (2), (4) аппроксимируются точно

$$y(N_1, j) = g_1^+(j), \quad y(0, j) = g_1^-(j), \quad 0 \leq j \leq N_2 - 1. \quad (10)$$

Разностный аналог краевых условий третьего рода (3), (5) имеет вид (для $0 \leq j \leq N_2 - 1$)

$$\Lambda y = -\frac{a_1}{\rho \bar{h}_1} y_{\bar{r}} + \frac{1}{\rho^2} (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - \left(d + \frac{r g_1^+}{\rho \bar{h}_1} \right) y = -\psi - \frac{r g_1^+}{\rho \bar{h}_1}, \quad i = N_1, \quad (11)$$

$$\Lambda y = \frac{a_1^{+1}}{\rho \bar{h}_1} y_{\bar{r}} + \frac{1}{\rho^2} (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - \left(d + \frac{r g_1^-}{\rho \bar{h}_1} \right) y = -\psi - \frac{r g_1^-}{\rho \bar{h}_1}, \quad i = 0. \quad (12)$$

Здесь использованы соотношения (9).

Осталось построить разностное краевое условие на стороне $r = l_1$ для случая, когда $l_1 = 0$. Так как все узлы, лежащие на

стороне $r=0$, отождествляются, то

$$y(0, j) = y_0, \quad 0 \leq j \leq N_2 - 1. \quad (13)$$

Далее, так как начало координат является внутренней точкой круга, то, записывая уравнение (1) в декартовой системе координат и аппроксимируя его на радиально-кольцевой сетке при условии (6), получим

$$\begin{aligned} \Delta y &= \frac{1}{2\pi\rho h_i} \sum_{j=0}^{N_2-1} a_i^{+1} y_r h_2 - dy = -\Psi, \quad i=0, \\ d(0, j) &= d_0, \quad \Psi(0, j) = \psi_0, \quad 0 \leq j \leq N_2 - 1. \end{aligned} \quad (14)$$

Здесь y_0 , d_0 и ψ_0 — значения соответствующих сеточных функций в центре круга.

Итак, в случае круга имеем нелокальное краевое условие (13), (14) на стороне $r=0$ прямоугольника \bar{G} . Разностные схемы построены.

Для разностной аппроксимации уравнения (1) в окрестности $r=0$ часто используется другая сетка по r , в которой точка $r=0$ не содержится:

$$\begin{aligned} \bar{\omega} &= \{(r_i, \varphi_j) \in \bar{G}, r_i = (i + 0,5) h_1, 0 \leq i \leq N_1, r_{N_1} = l_1, \\ &\quad \varphi_j = \varphi_{j-1} + h_2(j), 1 \leq j \leq N_2, \varphi_0 = l_2, \varphi_{N_2} = L_2\} \end{aligned}$$

(для простоты предполагаем, что сетка по r равномерна).

Тогда $a_1(i, j) = \bar{r}_i k_1(\bar{r}_i, \varphi_j)$, $a_2(i, j) = k_2(r_i, \varphi_j)$ и т. д., где $\bar{r}_i = ih_1$. Уравнения (8) остаются без изменений, а при $i=0$ пишется следующее разностное уравнение:

$$\Delta y = \frac{\bar{r}_1}{r_0 h_1} a_1(1, j) y_r(1, j) + \frac{1}{r_0} (a_2 y_\varphi)_\varphi - dy = -\Psi$$

(здесь $r_0 = 0,5h_1$, $\bar{r}_1 = h_1$), которое является аналогом краевого условия третьего рода.

Условие при $r=0$ отсутствует; определить значение y при $r=0$ из написанных разностных уравнений нельзя.

2. Разрешимость разностных краевых задач. В п. 1 были построены разностные схемы, аппроксимирующие задачи (1)–(6). Для круга схема задана формулами (8), (10), (11), (14), для кольца — формулами (8), (10), (12). Исследуем вопрос о разрешимости указанных схем.

Обозначим через $\bar{\omega}^*$ часть сетки $\bar{\omega}$: $\bar{\omega}^* = \{(r_i, \varphi_j) \in \bar{\omega}, 0 \leq i \leq N_1, 0 \leq j \leq N_2 - 1\}$. Пространство H состоит из сеточных функций, заданных на $\bar{\omega}^*$ и удовлетворяющих дополнительному условию $y(0, j) = \text{const}$, $0 \leq j \leq N_2 - 1$, если $l_1 = 0$. Скалярное произведение в H определим формулой

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} u(i, j) v(i, j) \rho(i) h_1(i) h_2(j).$$

Можно показать, что если функция $\rho(i)$ определена формулами (7), то верно равенство

$$(1, 1) = 0,5(L_1^2 - l_1^2)(L_2 - l_2) = \pi(L_1^2 - l_1^2), \quad (15)$$

т. е. квадрат нормы функции, тождественно равной единице на $\bar{\omega}^*$, равен площади круга ($l_1 = 0$) или кольца ($l_1 > 0$). Кроме того, если рассматриваемая область есть круг, то, используя постоянство по j при $i=0$ сеточных функций из H и равенство

$$\sum_{j=0}^{N_2-1} \bar{h}_2(j) = L_2 - l_2 = 2\pi,$$

можно получить следующее выражение для введенного выше скалярного произведения:

$$(u, v) = \rho(0) \bar{h}_1(0) 2\pi u_0 v_0 + \sum_{i=1}^{N_1} \sum_{j=0}^{N_2-1} u(i, j) v(i, j) \rho(i) \bar{h}_1(i) \bar{h}_2(j), \quad (16)$$

где $u_0 = u(0, j)$, $v_0 = v(0, j)$.

Исследуем разрешимость разностных задач (8), (11), (13), (14) при $l_1 = 0$ и (8), (11), (12) при $l_1 > 0$, если $d \equiv 0$, $\kappa_1^+ = \kappa_1^- \equiv 0$. Запишем указанные выше разностные задачи в виде операторного уравнения

$$Au = f, \quad (17)$$

где оператор A определим следующим образом: $Ay = -\Lambda \bar{y}$, $y(i, j) = \bar{y}(i, j)$ для $0 \leq i \leq N_1$, $0 \leq j \leq N_2 - 1$ и \bar{y} удовлетворяет условиям периодичности (9), кроме того, $y(0, j) = \bar{y}(0, j) = \text{const}$.

Рассмотрим сначала оператор A , соответствующий разностному оператору Λ задачи (8), (11), (13), (14). Учитывая, что первая разностная формула Грина для функций, удовлетворяющих условию периодичности (9), принимает вид

$$\sum_{j=0}^{N_2-1} (a_2 u_{-\bar{\varphi}})_{\bar{\varphi}} v \bar{h}_2 = - \sum_{i=0}^{N_1} a_2 u_{-\bar{\varphi}} v_{-\bar{\varphi}} \bar{h}_2,$$

будем иметь с учетом (16)

$$\begin{aligned} (Au, v) &= -(\Lambda \bar{u}, \bar{v}) = \\ &= \sum_{i=0}^{N_1} \bar{h}_1 \left(\sum_{j=1}^{N_2} h_1 a_1 \bar{u}_{\bar{i}} \bar{v}_{\bar{j}} + \sum_{i=0}^{N_1} \rho \bar{h}_1 d \bar{u} \bar{v} + r \kappa_1^+ \bar{u} \bar{v} |_{i=N_1} \right) + \\ &\quad + \sum_{i=1}^{N_1} \frac{\bar{h}_1}{\rho} \sum_{j=0}^{N_2-1} h_2 a_2 \bar{u}_{\bar{\varphi}} \bar{v}_{\bar{\varphi}} = -(\bar{u}, \Lambda \bar{v}) = (u, Av). \end{aligned}$$

Следовательно, оператор A самосопряжен в H .

Для оператора A , соответствующего разностному оператору Λ задачи (8), (11), (12), получим аналогичное равенство $(Au, v) =$

$$\sum_{i=0}^{N_2-1} \hbar_2 \left(\sum_{i=1}^{N_1} h_1 a_1 \bar{u}_r \bar{v}_r + \sum_{i=0}^{N_1} \rho \hbar_1 d \bar{u} \bar{v} + r \kappa_1 \bar{u} \bar{v}|_{i=0} + r \kappa_1^+ \bar{u} \bar{v}|_{i=N_1} \right) + \\ + \sum_{i=0}^{N_1} \frac{\hbar_1}{\rho} \sum_{j=0}^{N_2-1} h_2 a_2 \bar{u}_{\bar{\varphi}} \bar{v}_{\bar{\varphi}} = (u, Av),$$

из которого следует самосопряженность оператора A .

Если $d \equiv 0$, $\kappa_1^\pm \equiv 0$, то из самосопряженности оператора A , неравенства Коши—Буняковского $(Au, u) \leq \|Au\| \|u\|$ следует, что ядро оператора A состоит из сеточных функций, равных постоянной на сетке $\bar{\omega}^*$. Поэтому условие существования решения операторного уравнения (17) имеет вид $(f, 1) = 0$. Для задачи (8), (11), (13), (14) ему соответствует условие

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \psi(i, j) \rho(i) \hbar_1(i) \hbar_2(j) + L_1 \sum_{j=0}^{N_2-1} \hbar_2(j) g_1^+(j) = 0, \quad (18)$$

являющееся разностным аналогом условия (23) § 1. Для задачи (8), (11), (12) условие разрешимости имеет вид

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \psi(i, j) \rho(i) \hbar_1(i) \hbar_2(j) + \sum_{j=0}^{N_2-1} \hbar_2(j) [L_1 g_1^+(j) + l_1 g_1^-(j)] = 0$$

и является аналогом условия (25) § 1, обеспечивающего разрешимость соответствующей дифференциальной задачи для кольца.

Если указанные условия выполнены, то решения рассмотренных задач существуют и любые два решения отличаются на постоянную. Нормальное решение этих задач удовлетворяет условию $(y, 1) = 0$.

Пусть \bar{y} — одно из решений, которое можно найти, например, фиксируя искомое решение в одном узле сетки. Тогда, учитывая равенство (15), получим, что функция

$$\bar{y} = y - \frac{(y, 1)}{\pi(L_1^2 - l_1^2)} = y - \frac{(y, 1)}{(1, 1)}$$

является нормальным решением.

Замечание. Если определить сеточную функцию $\rho(i)$ формулами

$$\rho(i) = r_i, \quad 1 \leq i \leq N_1, \quad \rho(0) = \begin{cases} h_1(0)/4, & l_1 = 0, \\ l_1, & l_1 > 0, \end{cases}$$

то изменится лишь равенство (15) для случая, когда $l_1 = 0$. В этом случае будем иметь

$$(1, 1) = \pi L_1^2 + \frac{h_1^2(1)}{4} \pi = \pi \left(L_1^2 + \frac{h_1^2(1)}{4} \right).$$

3. Принцип суперпозиции для задачи в круге. Решение разностных задач в круге осложнено наличием нелокального краевого условия (14), задаваемого при $i = 0$. Заметим, что если задача вырождена, а условие разрешимости (18) выполнено, то одно из решений удобно выделять, фиксируя его значение в центре круга, т. е. задавая $y(0, j) = y_0$, $0 \leq j \leq N_2 - 1$. В этом случае условие (14) не используется, и полученная задача с заданным y_0 аналогична задаче, поставленной для кольца с краевым условием первого рода на внутренней окружности. Пусть теперь разностная задача (8), (11), (13), (14) не вырождена. Покажем, что ее решение можно найти, решая две вспомогательные задачи с локальными краевыми условиями первого рода при $i = 0$, $0 \leq j \leq N_2 - 1$.

Будем искать решение задачи (8), (11), (13), (14) в виде

$$y(i, j) = v(i, j) + y_0 w(i, j), \quad 0 \leq i \leq N_1, \quad 0 \leq j \leq N_2 - 1, \quad (19)$$

где y_0 — значение искомого решения в центре круга, а $v(i, j)$ и $w(i, j)$ удовлетворяют условиям периодичности

$$v(i, j) = v(i, N_2 + j), \quad w(i, j) = w(i, N_2 + j), \quad j = 0, -1$$

и являются решениями следующих краевых задач:

$$\left. \begin{aligned} \frac{1}{\rho} (a_1 v_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2} (a_2 v_{\bar{\varphi}})_{\hat{\varphi}} - dv &= -\psi, \quad 1 \leq i \leq N_1 - 1, \\ &0 \leq j \leq N_2 - 1, \end{aligned} \right\} \quad (20)$$

$$v(0, j) = 0, \quad t = 0,$$

$$\left. \begin{aligned} -\frac{a_1}{\rho h_1} v_{\bar{r}} + \frac{1}{\rho^2} (a_2 v_{\bar{\varphi}})_{\hat{\varphi}} - \left(d + \frac{r \kappa_1^+}{\rho h_1} \right) v &= -\psi - \frac{r g_1^+}{\rho h_1}, \quad i = N_1, \\ \Lambda w = \frac{1}{\rho} (a_1 w_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2} (a_2 w_{\bar{\varphi}})_{\hat{\varphi}} - dw &= 0, \quad 1 \leq i \leq N_1 - 1, \\ &0 \leq j \leq N_2 - 1, \\ w(0, j) &= 1, \quad t = 0, \\ -\frac{a_1}{\rho h_1} w_{\bar{r}} + \frac{1}{\rho^2} (a_2 w_{\bar{\varphi}})_{\hat{\varphi}} - \left(d + \frac{r \kappa_1^+}{\rho h_1} \right) w &= 0, \quad i = N_1. \end{aligned} \right\} \quad (21)$$

Очевидно, что функция y , определяемая согласно (19), удовлетворяет уравнению (8) и условиям (11), (13). Осталось определить y_0 . Подставляя (19) в неиспользованное еще условие (14)

и учитывая краевые условия для v и w , получим

$$y_0 = \frac{\left[2\pi\rho\hbar_1\psi_0 + \sum_{j=0}^{N_2-1} a_1^{+1}v_r\hbar_2(j) \right]_{i=0}}{\left[2\pi\rho\hbar_1d - \sum_{j=0}^{N_2-1} a_1^{+1}w_r\hbar_2(j) \right]_{i=0}}. \quad (22)$$

Покажем, что знаменатель в (22) отличен от нуля. Для этого умножим уравнение (21) скалярно на w . Используя краевые условия для w , соотношения периодичности и разностные формулы Грина, получим

$$0 = \sum_{i=1}^{N_1-1} \sum_{j=0}^{N_2-1} (\Lambda w) w \rho \hbar_1 \hbar_2 = - \sum_{j=0}^{N_2-1} \hbar_2 (a_1^{+1} w_r|_{i=0} + L_1 \kappa_1^+ w^2|_{i=N_1}) - \\ - \sum_{i=1}^{N_1} \sum_{j=0}^{N_2-1} a_1 w_r^2 \hbar_1 \hbar_2 - \sum_{i=1}^{N_1} \sum_{j=0}^{N_2-1} \hbar_1 \left[\frac{\hbar_2}{\rho} a_2 w_\varphi^2 + \hbar_2 \rho d w^2 \right].$$

Так как функция w отлична от постоянной, $d \geq 0$, $a_\alpha \geq c_1 > 0$, $\alpha = 1, 2$, и $\kappa_1^+ \geq 0$, причем $d^2 + (\kappa_1^+)^2 \neq 0$, то отсюда получим, что

$$\sum_{j=0}^{N_2-1} a_1^{+1} w_r \hbar_2|_{i=0} < 0$$

и, следовательно, знаменатель в формуле (22) отличен от нуля.

Итак, решение исходной задачи (8), (11), (13), (14) сведено к решению двух задач (20) и (21) с локальными краевыми условиями и нахождению y_0 по формуле (22). Искомое решение y находится по формуле (19).

Отметим, что если на стороне $r = L_1$ задано краевое условие первого рода $y(N_1, j) = g_1^+(\varphi_j)$, то для функций v и w вместо условий третьего рода в (20) и (21) следует положить $v(N_1, j) = g_1^+(\varphi_j)$ и $w(N_1, j) = 0$ для $0 \leq j \leq N_2 - 1$. Формула (22) для y_0 сохраняется. Если коэффициенты k_1, k_2, q и κ_1^+ не зависят от φ , то не зависит от φ и решение w задачи (21). В этом случае для функции w мы имеем одномерную задачу

$$\begin{aligned} \frac{1}{\rho} (a_1 w_r)_r - dw = 0, \quad 1 \leq i \leq N_1 - 1, \\ w(0, j) = 1, \quad i = 0, \\ -\frac{a_1}{\rho \hbar_1} w_r - \left(d + \frac{r \kappa_1^+}{\rho \hbar_1} \right) w = 0, \quad i = N_1, \end{aligned}$$

которая решается методом прогонки.

4. Прямые методы решения уравнений в круге и кольце. Из сказанного выше следует, что достаточно ограничиться рассмотрением методов решения разностных задач (8), (10)–(12).

Сначала изучим случай, для которого указанные разностные задачи могут быть решены одним из прямых методов, изложенных в главах III и IV.

Пусть коэффициенты k_1 , k_2 и q уравнения (1) не зависят от φ : $k_1 = k_1(r)$, $k_2 = k_2(r)$, $q = q(r)$. Такая ситуация имеет место для уравнения Пуассона в полярной системе координат. Пусть, кроме того, в краевых условиях третьего рода (11), (12) κ_1^- и κ_1^+ — постоянные. Предполагается, что сетка $\bar{\omega}$ равномерна по φ , т. е. $h_2(j) \equiv h_2$, и может быть неравномерной по r . При указанных предположениях разностное уравнение (8) с любой комбинацией краевых условий (10)–(12) может быть решено либо методом полной редукции, либо комбинированным методом неполной редукции и разделения переменных.

Проиллюстрируем возможность применения прямых методов на примере, в котором на сторонах $r = l_1$ и $r = L_1$ заданы краевые условия третьего (второго) рода (11), (12). Другие комбинации краевых условий рассматриваются аналогично.

В силу сделанных предположений коэффициенты разностной схемы определяются по формулам

$$a_1(i) = \bar{r}_i k_1(\bar{r}_i), \quad a_2(i) = k_2(r_i), \quad d(i) = q(r_i),$$

и так как сетка $\bar{\omega}$ равномерна по φ , то разностный оператор $(a_2 y_{\varphi})_{\hat{\varphi}}$ заменяется на $a_2 y_{\varphi}$.

Сведем разностную задачу (8), (11), (12) к системе трехточечных векторных уравнений

$$\begin{aligned} -Y_{N_2-1} + CY_0 - Y_1 &= F_0, \quad j = 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leq j \leq N_2 - 2, \\ -Y_{N_2-2} + CY_{N_2-1} - Y_0 &= F_{N_2-1}, \quad j = N_2 - 1. \end{aligned} \quad (23)$$

Здесь для $0 \leq j \leq N_2 - 1$ использованы обозначения:

$$\begin{aligned} Y_j &= (y(0, j), y(1, j), \dots, y(N_1, j)), \\ F_j &= (\theta_0 f(0, j), \theta_1 f(1, j), \dots, \theta_{N_1} f(N_1, j)), \\ CY_j &= ((2E - \theta_0 \Lambda_1) y(0, j), \dots, (2E - \theta_{N_1} \Lambda_1) y(N_1, j)), \end{aligned}$$

где

$$f(i, j) = \begin{cases} \psi(0, j) + \frac{l_1 g_1^-(\varphi_j)}{\rho(0) \bar{h}_1(0)}, & i = 0, \\ \psi(i, j), & 1 \leq i \leq N_1 - 1, \\ \psi(N_1, j) + \frac{L_1 g_1^+(\varphi_j)}{\rho(N_1) \bar{h}_1(N_1)}, & i = N_1, \end{cases} \quad (24)$$

разностный оператор Λ_1 действует следующим образом:

$$\Lambda_1 y = \begin{cases} \frac{a_1^{+1}}{\rho h_1} y_r - \left(d + \frac{r \kappa_1^-}{\rho h_1} \right) y, & i = 0, \\ \frac{1}{\rho} (a_1 y_r)_r - dy, & 1 \leq i \leq N_1 - 1, \\ -\frac{a_1}{\rho h_1} y_r - \left(d + \frac{r \kappa_1^+}{\rho h_1} \right) y, & i = N_1, \end{cases} \quad (25)$$

и, наконец, $\theta_i = \rho^2(i) h_2^2 / a_2(i)$, $0 \leq i \leq N_1$.

Система (23) получается из (8), (11) и (12) умножением каждого уравнения на соответствующее θ_i и переходом к векторной записи.

Напомним, что алгоритм метода полной редукции для системы (23) описан в п. 2 § 4 гл. III. В комбинированном методе используется алгоритм быстрого дискретного преобразования Фурье, который приведен в п. 4 § 1 гл. IV. Эти методы характеризуются оценкой $O(N_1 N_2 \log_2 N_2)$ арифметических действий при $N_2 = 2^n$.

5. Метод переменных направлений. Пусть теперь коэффициенты в уравнении (1) и краевых условиях (3), (5) удовлетворяют условиям $k_1 = k_1(r)$, $k_2 = k_2(\varphi)$, $q = \text{const}$, $\kappa_1^\pm = \text{const}$, т. е. для задачи (1), (3), (5) применим метод разделения переменных. Предполагается, что сетка ω неравномерна по каждому направлению. Рассмотрим разностное уравнение (8) с любой комбинацией краевых условий (10)–(12). При сделанных предположениях переменные в разностной схеме разделяются, и ее приближенное решение может быть найдено при помощи метода переменных направлений с оптимальным набором итерационных параметров.

Для примера рассмотрим задачу (8), (11), (12) с краевыми условиями третьего рода при $r = l_1$ и $r = L_1$. Запишем эту задачу в виде

$$\begin{aligned} \bar{\Lambda}y &= -\bar{f}, \quad 0 \leq i \leq N_1, \quad 0 \leq j \leq N_2 - 1, \\ \bar{\Lambda} &= \bar{\Lambda}_1 + \bar{\Lambda}_2, \quad \bar{f} = \rho^2 f, \end{aligned} \quad (26)$$

где $\bar{\Lambda}_1 = \rho^2 \Lambda_1$, оператор Λ_1 определен в (25), оператор $\bar{\Lambda}_2$ задается равенством $\bar{\Lambda}_2 y = (a_2 y_\varphi^-)_{\hat{\varphi}}$, причем выполнены соотношения (9), а правая часть f определена в (24). Уравнение (26) получено из (8), (11) и (12) умножением на ρ^2 .

В силу сделанных предположений коэффициенты разностной схемы (26) выбираются по формулам $a_1(i) = r_i k_1(r_i)$, $a_2(j) = k_2(\varphi_j)$, $d = q = \text{const}$.

первого рода. Постоянные δ_α и Δ_α оцениваются так же, как в рассмотренном ранее случае, только в (30) и (32) краевые условия третьего рода следует заменить на условия $v(0)=0$, $v(N_1)=0$ и $w(0)=0$, $w(N_1)=0$.

В заключение отметим, что при $d=0$, $\kappa_i^\pm=0$ задача (8), (11), (12) вырождена, и если выполнено условие разрешимости

$$\sum_{l=0}^{N_1} \sum_{j=0}^{N_2-1} \psi \varrho \hbar_1 \hbar_2 + \sum_{j=0}^{N_2-1} \hbar_2 [L_1 g_1^+ + l_1 g_1^-] = 0,$$

то задача имеет неединственное решение. Для этого случая набор параметров $\omega_k^{(1)}$ и $\omega_k^{(2)}$ для метода переменных направлений (29) построен в п. 1 § 4 гл. XII.

6. Решение разностных задач в кольцевом секторе. Рассмотрим методы решения разностных краевых задач для эллиптического уравнения без смешанных производных, заданного в кольцевом секторе.

В области $\bar{G} = \{l_1 \leq r \leq L_1, l_2 \leq \varphi \leq L_2, l_1 > 0, L_2 - l_2 < 2\pi\}$ требуется найти решение уравнений (1), удовлетворяющее на сторонах $r = l_1$ и $r = L_1$ одному из краевых условий (2)–(5), а на сторонах $\varphi = l_2$ и $\varphi = L_2$ одному из условий

$$u(r, \varphi) = g_2^-(r), \quad \varphi = l_2 \quad (34)$$

или

$$\frac{k_2}{r} \frac{\partial u}{\partial \varphi} = \kappa_2^- u - g_2^+(r), \quad \varphi = l_2, \quad (35)$$

$$u(r, \varphi) = g_2^+(r), \quad \varphi = L_2, \quad (36)$$

или

$$-\frac{k_2}{r} \frac{\partial u}{\partial \varphi} = \kappa_2^+ u - g_2^+(r), \quad \varphi = L_2. \quad (37)$$

Предполагается, что коэффициенты удовлетворяют условиям $k_1(r, \varphi) \geq c_1 > 0$, $k_2(r, \varphi) \geq c_1 > 0$, $q(r, \varphi) \geq 0$, $\kappa_1^\pm(\varphi) \geq 0$, $\kappa_2^\pm(r) \geq 0$.

В области \bar{G} вводится произвольная неравномерная прямоугольная сетка $\bar{\omega}$ (см. п. 1 § 3):

$$\begin{aligned} \bar{\omega} = \{(r_i, \varphi_j) \in \bar{G}, \quad r_i = r_{i-1} + h_1(i), \quad 1 \leq i \leq N_1, \quad r_0 = l_1, \\ r_{N_1} = L_1, \quad \varphi_j = \varphi_{j-1} + h_2(j), \quad 1 \leq j \leq N_2, \quad \varphi_0 = l_2, \quad \varphi_{N_2} = L_2\} \end{aligned}$$

и определяются средние шаги $\bar{h}_1(i)$ и $\bar{h}_2(j)$:

$$\bar{h}_\alpha(m) = \begin{cases} 0,5h_\alpha(1), & m=0, \\ 0,5[h_\alpha(m) + h_\alpha(m+1)], & 1 \leq m \leq N_\alpha-1, \\ 0,5h_\alpha(N_\alpha), & m=N_\alpha, \quad \alpha=1, 2. \end{cases}$$

Уравнение (1) аппроксимируется разностным уравнением

$$\frac{1}{\rho} (a_1 y_r)_r + \frac{1}{\rho^2} (a_2 y_\varphi)_\varphi - dy = -\psi, \quad (38)$$

$$1 \leq i \leq N_1 - 1, \quad 1 \leq j \leq N_2 - 1.$$

Краевые условия первого рода (2), (4), (34), (36) аппроксимируются точно:

$$y(N_1, j) = g_1^+(r_j), \quad y(0, j) = g_1^-(r_j), \quad 0 \leq j \leq N_2, \quad (39)$$

$$y(i, N_2) = g_2^+(r_i), \quad y(i, 0) = g_2^-(r_i), \quad 0 \leq i \leq N_1. \quad (40)$$

Условия третьего рода (3) и (5), заданные при $r = L_1$ и $r = l_1$, заменяются для $1 \leq j \leq N_2 - 1$ условиями (11) и (12).

Разностный аналог краевых условий (35) и (37) имеет вид

$$\frac{1}{\rho} (a_1 y_r)_r + \frac{a_2^{+1}}{\rho^2 h_2} y_\varphi - \left(d + \frac{\kappa_2^-}{\rho h_2} \right) y = -\psi - \frac{g_2^-}{\rho h_2}, \quad j = 0, \quad (41)$$

$$\frac{1}{\rho} (a_1 y_r)_r - \frac{a_2^-}{\rho^2 h_2} y_\varphi - \left(d + \frac{\kappa_2^+}{\rho h_2} \right) y = -\psi - \frac{g_2^+}{\rho h_2}, \quad j = N_2. \quad (42)$$

Если на пересекающихся сторонах прямоугольника заданы краевые условия третьего рода, то в угловых узлах сетки ставятся следующие краевые условия:

$$\frac{a_1^{+1}}{\rho h_1} y_r + \frac{a_2^{+1}}{\rho^2 h_2} y_\varphi - \left(d + \frac{r \kappa_1^-}{\rho h_1} + \frac{\kappa_2^-}{\rho h_2} \right) y = -\psi - \frac{r g_1^-}{\rho h_1} - \frac{g_2^-}{\rho h_2}, \quad (43)$$

если $i = j = 0$;

$$-\frac{a_1^-}{\rho h_1} y_r + \frac{a_2^{+1}}{\rho^2 h_2} y_\varphi - \left(d + \frac{r \kappa_1^+}{\rho h_1} + \frac{\kappa_2^-}{\rho h_2} \right) y = -\psi - \frac{r g_1^+}{\rho h_1} - \frac{g_2^-}{\rho h_2}, \quad (44)$$

если $i = N_1, j = 0$;

$$\frac{a_1^{+1}}{\rho h_1} y_r - \frac{a_2^-}{\rho^2 h_2} y_\varphi - \left(d + \frac{r \kappa_1^-}{\rho h_1} + \frac{\kappa_2^+}{\rho h_2} \right) y = -\psi - \frac{r g_1^-}{\rho h_1} - \frac{g_2^+}{\rho h_2}, \quad (45)$$

если $i = 0, j = N_2$; и, наконец,

$$-\frac{a_1^-}{\rho h_1} y_r - \frac{a_2^-}{\rho^2 h_2} y_\varphi - \left(d + \frac{r \kappa_1^+}{\rho h_1} + \frac{\kappa_2^+}{\rho h_2} \right) y = -\psi - \frac{r g_1^+}{\rho h_1} - \frac{g_2^+}{\rho h_2}, \quad (46)$$

если $i = N_1, j = N_2$.

Если рассматривается разностная задача (38), (11), (12), (41)–(46) с $d \equiv 0$ и $\kappa_\alpha^\pm \equiv 0$, $\alpha = 1, 2$, то решение существует, если выполнено условие

$$\sum_{j=0}^{N_2} \sum_{i=0}^{N_1} \rho h_1 h_2 \psi + \sum_{j=0}^{N_2} h_2 (L_1 g_1^+ + l_1 g_1^-) + \sum_{i=0}^{N_1} h_1 (g_2^- + g_2^+) = 0,$$

являющееся разностным аналогом условия (27) § 1 разрешимости соответствующей задачи для дифференциального уравнения.

В пространстве H сеточных функций, заданных на $\bar{\omega}^*$, определим скалярное произведение

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \frac{h_1(i) h_2(j)}{\rho(i)} u(i, j) v(i, j). \quad (27)$$

Операторы A_1 и A_2 , действующие в H , определим обычным образом: $A_\alpha y = -\Lambda_\alpha \bar{y}$, где $y(i, j) = \bar{y}(i, j)$ для $0 \leq i \leq N_1$, $0 \leq j \leq N_2 - 1$ и \bar{y} удовлетворяет соотношениям периодичности (9). Тогда схема (26) может быть записана в виде операторного уравнения

$$Au = \bar{f}, \quad A = A_1 + A_2 \quad (28)$$

в пространстве H .

Для решения уравнения (28) используем метод переменных направлений, итерационная схема которого имеет вид

$$\begin{aligned} B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k &= \bar{f}, \quad k = 0, 1, \dots, \quad y_0 \in H, \\ B_k &= (\omega_k^{(1)} E + A_1)(\omega_k^{(2)} E + A_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}. \end{aligned} \quad (29)$$

Самосопряженность операторов A_1 и A_2 в пространстве H устанавливается при помощи разностной формулы Грина, а перестановочность их проверяется непосредственно.

Найдем теперь границы операторов A_1 и A_2 , т. е. постоянные δ_α и Δ_α , $\alpha = 1, 2$, в неравенствах

$$\delta_\alpha(u, u) \leq (A_\alpha u, u) \leq \Delta_\alpha(u, u).$$

Найдем сначала δ_2 и Δ_2 . Так как для функции $\bar{u}(i, j)$, удовлетворяющей условию периодичности (9), имеем

$$(A_2 u, u) = -(\Lambda_2 \bar{u}, \bar{u}) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \frac{h_1(i) h_2(j)}{\rho(i)} \left(a_2 u \frac{\bar{u}}{\varphi} \right)_{ij},$$

то

$$\delta_2 = 0, \quad \Delta_2 = \max_{0 \leq j \leq N_2 - 1} \left[\frac{a_2(j+1)}{h_2(j+1)} + \frac{a_2(j)}{h_2(j)} \right] \frac{2}{h_2(j)}.$$

Здесь были учтены соотношения (9) для a_2 и h_2 .

Далее, используя аналог леммы 16 главы V, найдем, что δ_1 можно оценить следующим образом: $1/\delta_1 = \max_{0 \leq i \leq N_1} v(i)$, где $v(i)$ — решение краевой задачи

$$\begin{aligned} \rho(a_1 v_r) \hat{r} - d \rho^2 v &= -1, \quad 1 \leq i \leq N_1 - 1, \\ \frac{\rho a_1^{+1}}{h_1} v_r - \left(d + \frac{r \kappa_1^-}{\rho h_1} \right) \rho^2 v &= -1, \quad i = 0, \\ -\frac{\rho a_1^-}{h_1} v_r - \left(d + \frac{r \kappa_1^+}{\rho h_1} \right) \rho^2 v &= -1, \quad i = N_1. \end{aligned} \quad (30)$$

Задача (30) решается методом прогонки.

Получим теперь оценку для Δ_1 . Используя первую разностную формулу Грина и определение (27) для скалярного произведения, найдем

$$(A_1 u, u) = -(\bar{A}_1 \bar{u}, \bar{u}) = \\ = \sum_{j=0}^{N_2-1} \bar{h}_2(j) \left[\sum_{i=1}^{N_1} h_1(i) a_1(i) \bar{u}_r^2(i, j) + d \sum_{i=0}^{N_1} \bar{h}_1(i) \rho(i) \bar{u}^2(i, j) + \right. \\ \left. + l_1 \kappa_1^- \bar{u}^2(0, j) + L_1 \kappa_1^+ \bar{u}^2(N_1, j) \right].$$

Оценим выражение, стоящее в квадратных скобках. Из аналого леммы 16 главы V получим оценку

$$d \sum_{i=0}^{N_1} \bar{h}_1(i) \rho(i) \bar{u}^2(i, j) + l_1 \kappa_1^- \bar{u}^2(0, j) + L_1 \kappa_1^+ \bar{u}^2(N_1, j) \leq \\ \leq m_1 \left[\sum_{i=1}^{N_1} (a_1 \bar{u}_r^2)_{ij} h_1(i) + \sum_{i=0}^{N_1} \frac{\bar{h}_1(i)}{\rho(i)} \bar{u}^2(i, j) \right], \quad (31)$$

где $m_1 = \max_{0 \leq i \leq N_1} w(i)$, а $w(i)$ есть решение задачи

$$\begin{aligned} \rho(a_1 w_r)_{\hat{r}} - w &= -d\rho^2, \quad 1 \leq i \leq N_1 - 1, \\ \frac{\rho a_i^{+1}}{\bar{h}_1} w_r - w &= -\left(d + \frac{r \kappa_1^-}{\rho \bar{h}_1}\right) \rho^2, \quad i = 0, \\ -\frac{\rho a_i}{\bar{h}_1} w_r - w &= -\left(d + \frac{r \kappa_1^+}{\rho \bar{h}_1}\right) \rho^2, \quad i = N_1. \end{aligned} \quad (32)$$

Далее, из аналого леммы 17 главы V получим оценку

$$\sum_{i=1}^{N_1} a_1(i) \bar{u}_r^2(i, j) h_1(i) \leq m_2 \sum_{i=0}^{N_1} \frac{\bar{h}_1(i)}{\rho(i)} \bar{u}^2(i, j), \quad (33)$$

где

$$m_2 = \max \left(\frac{a_1(N_1) \rho(N_1)}{\bar{h}_1^2(N_1)}, \frac{a_1(1) \rho(0)}{\bar{h}_1^2(0)}, \max_{1 \leq i \leq N_1 - 1} \frac{2\rho(i)}{\bar{h}_1(i)} \left[\frac{a_1(i)}{h_1(i)} + \frac{a_1(i+1)}{h_1(i+1)} \right] \right).$$

Из (31) и (33) следует оценка

$$\sum_{i=1}^{N_1} h_1 a_1 \bar{u}_r^2 + d \sum_{i=0}^{N_1} \bar{h}_1 \rho \bar{u}^2 + l_1 \kappa_1^- \bar{u}^2|_{i=0} + L_1 \kappa_1^+ \bar{u}^2|_{i=N_1} \leq \\ \leq \Delta_1 \sum_{i=0}^{N_1} \frac{\bar{h}_1}{\rho} \bar{u}^2, \quad \Delta_1 = m_1 + m_2 (1 + m_1).$$

Умножая это неравенство на $\bar{h}_2(j)$ и суммируя по j от 0 до $N_2 - 1$, получим $(A_1 u, u) \leq \Delta_1(u, u)$.

Итак, постоянные δ_α и Δ_α , $\alpha = 1, 2$, найдены. Напомним, что формулы для итерационных параметров $\omega_k^{(1)}$ и $\omega_k^{(2)}$ были получены в п. 4 § 1 гл. XI.

Аналогичным образом строится метод переменных направлений для разностной задачи (8), (10) с краевыми условиями

При этом любые два решения указанной задачи отличаются на постоянную.

Приведенное утверждение доказывается почти так же, как это было сделано в п. 2 § 3, для случая круга и кольца. Здесь скалярное произведение в пространстве H сеточных функций, заданных на ω , определяется формулой

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} u(i, j) v(i, j) \rho(i) \tilde{h}_1(i) \tilde{h}_2(j). \quad (47)$$

Отметим, что коэффициенты a_1, a_2, q и функция $\rho(i)$ в данном пункте определяются, как и в п. 1 § 3.

Сделаем замечание относительно методов решения построенных разностных задач. Если коэффициенты k_1, k_2, q зависят только от r, φ — постоянные, а $x_2^\pm = 0$, если заданы краевые условия (3), (5), (35), (37), и сетка ω равномерна по φ , то соответствующие разностные задачи могут быть решены прямыми методами, построенными в главах III и IV.

Если выполнены условия $k_1 = k_1(r), k_2 = k_2(\varphi), q = \text{const}, x_\alpha^\pm = \text{const}$ и сетка ω неравномерна по каждому из направлений, то для решения разностных задач можно использовать метод переменных направлений с оптимальным набором параметров. В этом случае, так же как было сделано в предыдущем пункте, разностные уравнения следует предварительно умножить на $\rho^\alpha(i)$.

7. Общий случай переменных коэффициентов. Рассмотрим теперь случай, когда переменные не разделяются и решение разностной краевой задачи находится итерационным методом.

Пусть, например, требуется найти решение задачи Дирихле для уравнения (1) на сетке $\bar{\omega}$ в предположениях, что сетка $\bar{\omega}$ равномерна по $\varphi (h_2(j) = h_2)$, $q = 0$, а коэффициенты k_1 и k_2 удовлетворяют условиям

$$0 < c_1 \leq k_\alpha(r, \varphi) \leq c_2, \quad \alpha = 1, 2. \quad (48)$$

При этих предположениях разностная задача записывается в виде

$$\begin{aligned} \Lambda y &= \frac{1}{\rho} (a_1 y_r)_r + \frac{1}{\rho^2} (a_2 y_\varphi)_\varphi = -\psi, \quad (r, \varphi) \in \omega, \\ y(r, \varphi) &= g(r, \varphi), \quad (r, \varphi) \in \gamma, \end{aligned} \quad (49)$$

где

$$\begin{aligned} a_1(i, j) &= \bar{r}_i k_1(\bar{r}_i, \bar{\varphi}_j), \quad a_2(i, j) = k_2(r_i, \bar{\varphi}_j), \\ \bar{r}_i &= r_i - 0.5h_1(i), \quad \bar{\varphi}_j = \varphi_j - 0.5h_2. \end{aligned} \quad (50)$$

В пространстве H сеточных функций, заданных на ω , определим скалярное произведение

$$(u, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} u(i, j) v(i, j) \rho(i) \tilde{h}_1(i) \tilde{h}_2,$$

и операторы A и R , действующие в H , $Ay = -\Lambda \dot{y}$, $Ry = -\gamma \dot{y}$,
где $y(r, \varphi) = \dot{y}(r, \varphi)$ для $(r, \varphi) \in \omega$ и $\dot{y}(r, \varphi) = 0$ для $(r, \varphi) \in \gamma$.
Здесь разностный оператор \mathcal{R} определяется соотношением

$$\mathcal{R}y = \frac{1}{\rho} (\bar{r}y_{\bar{r}})_{\bar{r}} + \frac{1}{\rho^2} y_{\bar{\varphi}\varphi}, \quad (r, \varphi) \in \omega.$$

Используя разностные формулы Грина, можно проверить, что операторы A и R самосопряжены в H и, кроме того, для любого $y \in H$ имеют место равенства

$$(Ay, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} a_i \dot{y}_r^2 h_1 h_2 + \sum_{j=1}^{N_2} \sum_{i=1}^{N_1-1} \frac{a_2}{\rho} \dot{y}_{\varphi}^2 \bar{h}_1 h_2,$$

$$(Ry, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} \bar{r} \dot{y}_r^2 h_1 h_2 + \sum_{j=1}^{N_2} \sum_{i=1}^{N_1-1} \frac{1}{\rho} \dot{y}_{\varphi}^2 \bar{h}_1 h_2.$$

Отсюда и из (48), (50) следует, что операторы A и R энергетически эквивалентны с постоянными $\gamma_1 = c_1$ и $\gamma_2 = c_2$:

$$\gamma_1 (Ry, y) \leq (Ay, y) \leq \gamma_2 (Ry, y), \quad \gamma_1 > 0. \quad (51)$$

Разностная задача (49) может быть записана в виде операторного уравнения

$$Au = f$$

с определенным выше оператором A . Для ее решения используем неявную итерационную схему

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H, \quad (52)$$

где $B = R$.

Из общей теории итерационных методов, изложенной в главе VI, следует, что если параметры τ_{k+1} в схеме (52) выбрать по формулам чебышевского метода

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k},$$

$$\mu_k \in \mathfrak{M}_n^* = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad 1 \leq i \leq n \right\}, \quad k = 1, 2, \dots, n,$$

то для погрешности $z_n = y_n - u$ будет верна оценка

$$\|y_n - u\|_D \leq \varepsilon \|y_0 - u\|_D,$$

где $D = A$ или $D = B$, $D = AB^{-1}A$, а число итераций удовлетворяет оценке

$$n \geq n_0(\varepsilon) = \ln(0.5\varepsilon) / \ln \rho_1.$$

Здесь

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Так как γ_1 и γ_2 не зависят от шагов сетки $\bar{\omega}$, то число итераций пропорционально $|\ln 0,5\epsilon|$ и не меняется при изменении сетки.

Для нахождения y_{k+1} получим разностную задачу

$$\mathcal{R}y_{k+1} = -F, \quad (r, \varphi) \in \omega, \quad y_{k+1} = g, \quad (r, \varphi) \in \omega$$

с известной правой частью $F = -\mathcal{R}y_k + \tau_{k+1}(\Lambda y_k + \psi)$. Отметим, что эта задача удовлетворяет всем условиям, позволяющим находить ее решение одним из прямых методов, например методом полной редукции с затратой $O(N_1 N_2 \log_2 N_2)$ арифметических действий при $N_2 = 2^n$. Таким образом, общее число действий, которое необходимо затратить для нахождения решения рассмотренной разностной задачи с точностью ϵ , оценивается величиной $O(N_1 N_2 \log_2 N_2 \ln(2/\epsilon))$.

Аналогичным образом могут быть построены, при соответствующих предположениях, итерационные методы решения поставленных в предыдущих параграфах разностных краевых задач в цилиндрической и полярной системах координат.

ДОПОЛНЕНИЕ

Построение полинома, наименее уклоняющегося от нуля

1. В § 2 гл. VI при рассмотрении двухслойных итерационных схем была сформулирована задача: построить полином степени n , принимающий в нуле значение 1, максимум модуля которого на отрезке $[\gamma_1, \gamma_2]$ минимален.

Решим эту задачу. Нам будет удобно проводить все исследования не на отрезке $[\gamma_1, \gamma_2]$, а на отрезке $[-1, 1]$. Для этого сделаем линейную замену переменной, переводящую отрезок $\gamma_1 \leq t \leq \gamma_2$ в отрезок $-1 \leq x \leq 1$, а точку γ_1 в точку 1. Эта замена имеет вид

$$t = \frac{1 - \rho_0 x}{\tau_0}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

При такой замене точке $t=0$ соответствует точка $x=1/\rho_0 > 1$.

Таким образом, сформулированная выше задача эквивалентна задаче: среди всех полиномов степени n , принимающих в точке $x=1/\rho_0 > 1$ значение 1, найти наименее уклоняющийся от нуля на отрезке $[-1, 1]$.

Это классическая чебышевская задача теории аппроксимации функций, решение которой хорошо известно, но нам полезно будет это решение найти заново. Для этого нам понадобится

Теорема 1. *Каковы бы ни были непрерывные на $[-1, 1]$ функции $g(x) > 0$ и $f(x)$, существует единственный полином $P_n(x)$ степени не выше n такой, что*

$$q_n = \max_{-1 \leq x \leq 1} g(x) |f(x) - P_n(x)| = \min_{\left\{ \begin{array}{c} R_k(x) \\ k \leq n \end{array} \right\}} \max_{-1 \leq x \leq 1} g(x) |f(x) - R_k(x)|.$$

Этот полином вполне характеризуется следующим свойством: число последовательных точек отрезка $[-1, 1]$, в которых функция $g(x)(f(x) - P_n(x))$ принимает с чередующимися знаками значение q_n , не меньше $n+2$.

Преобразуем поставленную задачу к задаче, фигурирующей в теореме 1. Учитывая, что искомый полином принимает значение 1 в точке $x=1/\rho_0$, представим его в виде

$$P_n(x) = 1 - \left(\frac{1}{\rho_0} - x \right) R_{n-1}(x) = \frac{1 - \rho_0 x}{\rho_0} \left[\frac{\rho_0}{1 - \rho_0 x} - R_{n-1}(x) \right],$$

где $R_{n-1}(x)$ — полином степени не выше $n-1$.

Отсюда следует, что наша задача сводится к задаче отыскания полинома $R_{n-1}(x)$ степени не выше $n-1$, дающего наилучшее равномерное приближение с весом $g(x) = (1 - \rho_0 x)/\rho_0 > 0$ функции $f(x) = \rho_0/(1 - \rho_0 x)$ на отрезке $[-1, 1]$.

Именно эта задача и фигурирует в теореме 1.

Поэтому на основании теоремы 1 существует по меньшей мере $n+1$ точек x_1, x_2, \dots, x_{n+1} отрезка $[-1, 1]$, в которых искомый полином $P_n(x)$ принимает с чередующимися знаками значение q_n .

Покажем сначала, что таких точек должно быть ровно $n+1$. Действительно, для того чтобы непрерывная функция более чем в $n+1$ последовательных точках отрезка $[-1, 1]$ могла принимать отличные от нуля значения q_n с чередующимися знаками, она должна обратиться в нуль не меньше чем в n точках.

Так как полином $P_n(x)$ отличен от тождественно нулевого, то на отрезке $[-1, 1]$ он может обратиться в нуль не более чем в n точках. Тем самым, искомый многочлен $P_n(x)$ на $[-1, 1]$ значение q_n с чередующимися знаками принимает ровно $n+1$ раз.

Охарактеризуем эти точки. Если полином $P_n(x)$ во внутренней точке отрезка $[-1, 1]$ принимает максимальное значение, то производная $P'_n(x)$ в этой точке обращается в нуль. Но степень $P'_n(x)$ равна $n-1$ и, следовательно, производная искомого полинома может обратиться в нуль лишь в $n-1$ точках. Поэтому искомый полином имеет $n-1$ внутренних экстремальных точек на $[-1, 1]$ и следовательно, два краевых экстремума, т. е.

$$|P_n(-1)| = |P_n(1)| = q_n.$$

Итак, мы имеем

$$P_n(\omega_j) = 0, \quad j = 1, 2, \dots, n, \quad |P_n(x_j)| = q_n, \quad j = 1, 2, \dots, n+1,$$

где ω_j — корни полинома, а x_j — экстремальные точки

$$-1 = x_{n+1} < \omega_n < x_n < \dots < \omega_2 < x_2 < \omega_1 < x_1 = 1.$$

Кроме того, так как $P_n(1/\rho_0) = 1$ и все корни полинома $P_n(x)$ лежат на отрезке $[-1, 1]$, то $P_n(1) = q_n$ и, следовательно, справедливы равенства

$$P_n(x_j) = (-1)^{j-1} q_n, \quad j = 1, 2, \dots, n+1. \quad (1)$$

Имеет место

Лемма 1. Полином $P_n(x)$, который среди всех многочленов n -й степени, принимающих значение 1 при $x = 1/\rho_0$, наименее уклоняется от нуля на отрезке $[-1, 1]$, удовлетворяет дифференциальному уравнению

$$(1-x^2)(P')^2 = n^2(q_n^2 - P^2). \quad (2)$$

Действительно, по доказанному точки x_2, x_3, \dots, x_n есть простые нули полинома $P'_n(x)$. Очевидно, что эти точки являются двукратными нулями полинома $q_n^2 - P_n^2(x)$, а по доказанному точки $x_{n+1} = -1$ и $x_1 = 1$ являются простыми нулями этого полинома. Поэтому полиномы $(1-x^2)(P'_n(x))^2$ и $q_n^2 - P_n^2(x)$ степени 2 имеют одни и те же нули. Следовательно, они пропорциональны, т. е.

$$(1-x^2)(P')^2 = c(q_n^2 - P_n^2(x)).$$

Приравнивая коэффициенты при старших степенях x у обоих полиномов, находим $c = n^2$. Лемма доказана.

2. Переходим к построению полинома $P_n(x)$, используя уравнение (2). Это уравнение, помимо неизвестной функции $P_n(x)$, содержит еще неизвестный параметр q_n . Мы не будем отдельно фиксировать дополнительные условия, которые однозначно определяют решение уравнения (2), а будем пользоваться всей известной информацией относительно $P_n(x)$.

Рассмотрим сначала уравнение (2) на отрезке $[-1, 1]$. В этом случае $|P_n(x)| \leq q_n$, и, следовательно, из левой и правой частей уравнения (2) можно извлечь корень

$$\pm \frac{dP}{\sqrt{q_n^2 - P^2}} = n \frac{dx}{\sqrt{1-x^2}}, \quad 0 \leq x \leq 1. \quad (3)$$

Исследуем левую часть (3). Если $P_n(x_{j+1}) = q_n$, то при изменении x от x_{j+1}

до x_j функция $P_n(x)$ убывает от q_n до $-q_n$. При этом дифференциал dP отрицателен, и поэтому в левой части уравнения (3) следует взять знак минус. Аналогично находим, что если $P_n(x_{j+1}) = -q_n$, то следует взять знак плюс. Учитывая (1), получим, что на отрезке $[x_{j+1}, x_j]$ уравнение (3) должно быть записано в виде

$$(-1)^{j-1} \frac{dP}{\sqrt{q_n^2 - P^2}} = n \frac{dx}{\sqrt{1-x^2}}, \quad x \in [x_{j+1}, x_j], \quad j=1, 2, \dots, n. \quad (4)$$

Получим теперь выражение для $P_n(x)$ на отрезке $[-1, 1]$. Пусть x — любая точка отрезка $[-1, 1]$, и для определенности пусть x принадлежит, например, отрезку $[x_{k+1}, x_k]$.

Проинтегрируем правую часть уравнения (4) по x от x до 1. Получим

$$n \int_x^1 \frac{dx}{\sqrt{1-x^2}} = n \arcsin x \Big|_x^1 = n \arccos x.$$

Проинтегрируем левую часть уравнения (4). Когда x меняется от x_{j+1} до x_j , функция $P(x)$ меняется от $P(x_{j+1}) = (-1)^{j-1} q_n$ до $P(x_j) = (-1)^{j-1} q_n$. Поэтому

$$(-1)^{j-1} \int_{P(x_{j+1})}^{P(x_j)} \frac{dP}{\sqrt{q_n^2 - P^2}} = \int_{-q_n}^{q_n} \frac{dP}{\sqrt{q_n^2 - P^2}} = \arcsin \frac{P}{q_n} \Big|_{-q_n}^{q_n} = \pi.$$

Далее, при интегрировании левой части (4) от $P(x)$ до $P(x_k)$ получим

$$(-1)^{k-1} \int_{P(x)}^{P(x_k)} \frac{dP}{\sqrt{q_n^2 - P^2}} = \int_{(-1)^{k-1} P(x)}^{q_n} \frac{dP}{\sqrt{q_n^2 - P^2}} = \arccos (-1)^{k-1} \frac{P(x)}{q_n}.$$

Так как

$$\int_x^1 \frac{dx}{\sqrt{1-x^2}} = \int_x^{x_k} \frac{dx}{\sqrt{1-x^2}} + \sum_{l=1}^{k-1} \int_{x_{j+1}}^{x_j} \frac{dx}{\sqrt{1-x^2}},$$

то окончательно получим

$$n \arccos x = (k-1) \pi + \arccos (-1)^{k-1} \frac{P(x)}{q_n}. \quad (5)$$

Отсюда найдем

$$P_n(x) = q_n \cos(n \arccos x), \quad |x| \leq 1. \quad (6)$$

Полагая в (5) $x = \omega_k \in [x_{k+1}, x_k]$, найдем корни полинома $P_n(x)$

$$\omega_k = \cos \frac{(2k-1)\pi}{2n}, \quad k=1, 2, \dots, n.$$

Формула (6) определяет полином $P_n(x)$ для $x \in [-1, 1]$. Найдем вид полинома $P_n(x)$ для $|x| \geq 1$ и определим q_n . Для этого заметим, что

$$\omega_{n-k+1} = \cos \left(\pi - \frac{2k-1}{2n} \pi \right) = -\omega_k, \quad k=1, 2, \dots, n.$$

Поэтому $P_n(-x) = (-1)^n P_n(x)$ и, следовательно, достаточно определить $P_n(x)$ для $x \geq 1$.

Исследуем уравнение (2) при $x \geq 1$. В этом случае его следует переписать следующим образом:

$$(x^2 - 1)(P')^2 = n^2 (P^2 - q_n^2), \quad x \geq 1.$$

Так как $x \geq 1$, то $P(x) \geq q_n$ и функция возрастает. Поэтому, извлекая корень, получим

$$\frac{dP}{\sqrt{P^2 - q_n^2}} = n \frac{dx}{\sqrt{x^2 - 1}}.$$

При интегрировании правой части этого уравнения от 1 до x левая часть будет интегрироваться от q_n до $P_n(x)$. Поэтому

$$\begin{aligned} \int_{q_n}^{P_n(x)} \frac{dP}{\sqrt{P^2 - q_n^2}} &= \ln \left(\frac{P_n(x)}{q_n} + \sqrt{\frac{P_n^2(x)}{q_n^2} - 1} \right) = \operatorname{arccch} \frac{P_n(x)}{q_n} = \\ &= n \int_1^x \frac{dx}{\sqrt{x^2 - 1}} = n \ln(x + \sqrt{x^2 - 1}) = n \operatorname{arccch} x. \end{aligned} \quad (7)$$

Отсюда получим

$$P_n(x) = q_n \operatorname{ch}(n \operatorname{arccch} x), \quad x \geq 1.$$

Так как $P_n(x) = (-1)^n P_n(-x)$, то для $x \leq 1$ найдем

$$P_n(x) = (-1)^n q_n \operatorname{ch}(n \operatorname{arccch}(-x)) = q_n \operatorname{ch}(n \operatorname{arccch} x), \quad x \leq -1.$$

Таким образом, для $|x| \geq 1$ получим следующее выражение для полинома $P_n(x)$:

$$P_n(x) = q_n \operatorname{ch}(n \operatorname{arccch} x), \quad |x| \geq 1. \quad (8)$$

Найдем теперь q_n . Полагая в (8) $x = 1/\rho_0$, и учитывая, что $P_n(1/\rho_0) = 1$, получим

$$q_n = 1/\operatorname{ch}(n \operatorname{arccch}(1/\rho_0)).$$

С другой стороны, полагая в (7) $x = 1/\rho_0$, найдем

$$\ln \frac{1 + \sqrt{1 - q_n^2}}{q_n} = n \ln \frac{1 + \sqrt{1 - \rho_0^2}}{\rho_0} = n \ln \frac{1}{\rho_1},$$

где

$$\rho_1 = \frac{\rho_0}{1 + \sqrt{1 - \rho_0^2}} = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad \rho_0 = \frac{2\rho_1}{1 + \rho_1^2}.$$

Следовательно,

$$q_n = \frac{1}{\operatorname{ch}\left(n \operatorname{arccch} \frac{1}{\rho_0}\right)} = \frac{2\rho_1^n}{1 + \rho_1^{2n}} < 1. \quad (9)$$

Объединяя (6) и (8), получим

$$P_n(x) = q_n T_n(x) = T_n(x)/T_n(1/\rho_0), \quad (10)$$

где

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leq 1 \\ \operatorname{ch}(n \operatorname{arccosh} x), & |x| \geq 1. \end{cases}$$

Полином $T_n(x)$ называется полиномом Чебышева первого рода степени n .
Итак, поставленная задача полностью решена. Ее решение дается формулами (9), (10). Возвращаясь к переменной t , получим искомый полином

$$Q_n(t) = P_n\left(\frac{1 - \tau_0 t}{\rho_0}\right) = q_n T_n\left(\frac{1 - \tau_0 t}{\rho_0}\right),$$

который наименее уклоняется от нуля на отрезке $[\gamma_1, \gamma_2]$.

ЛИТЕРАТУРА

1. Вазов В., Форсайт Дж. Разностные методы решения дифференциальных уравнений в частных производных.— М.: ИЛ., 1963.
2. Гельфond А. О. Исчисление конечных разностей.— М.: Наука, 1967.
3. Каrчевский М. М., Ляшко А. Д. Разностные схемы для нелинейных задач математической физики.— Казань: Ротапринт, изд. Каз. гос. ун., 1976.
4. Красносельский М. А., Вайникко Г. М. и др. Приближенное решение операторных уравнений.— М.: Наука, 1969.
5. Марчук Г. И., Методы вычислительной математики.— Новосибирск: Наука, 1973.
6. Оганесян Л. А., Ривкинд В. Я., Руховец Л. А. Вариационно-разностные методы решения эллиптических уравнений, ч. 1 и 2.— В сб.: Дифференциальные уравнения и их применение, вып. 5, Вильнюс, Пяргале, 1973, вып. 8, Вильнюс, Пяргале, 1974.
7. Орtega Дж., Рейнboldt В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными.— М.: Мир, 1975.
8. Самарский А. А. Введение в теорию разностных схем.— М.: Наука, 1971 (имеется библиография до 1971 г.).
9. Самарский А. А. Теория разностных схем.— М.: Наука, 1977.
10. Самарский А. А., Гулин А. В. Устойчивость разностных схем.— М.: Наука, 1973.
11. Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений.— М.: Наука, 1976.
12. Самарский А. А., Карамзин Ю. Н., Разностные уравнения,— М.: Знание, 1978.
13. Фадеев Д. К., Фадеева В. Н. Вычислительные методы линейной алгебры.— М.: Физматгиз, 1963.
14. Young D. M. Iterative solution of large linear systems:— N.— Y., L.: Acad. Press., 1971.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Алгоритм дискретного преобразования
Фурье 164
Асимптотическое свойство 338
- Задача на собственные значения 63
- Итерационные методы вариационного типа 332
— — двухступенчатые 509, 540
— — с факторизованным оператором 536
— — треугольные 389
Итерационный метод верхней релаксации 375
— — градиентного спуска 514
— — Зейделя 369
— — минимальных невязок 342, 482
— — погрешностей 344, 469
— — поправок 343
— — Ньютона — Канторовича 506
— — переменных направлений 432, 453, 475
— — полпеременно-треугольный 396, 411
— — простой итерации 285
— — скорейшего спуска 341
— — сопряженных градиентов 355
— — направлений 353
— — невязок 355
— — погрешностей 356, 469
— — поправок 356
— — стационарный трехслойный 322
— — чебышевский (Ричардсона) 269, 465
- Канонический вид итерационной схемы двухслойной 260
— — — трехслойной стандартного типа 261, 315
- Метод вариации постоянных 42
— прогонки 76
— — матричной 107
— — немонотонной 94
— — ортогональной 112
— —, потоковый вариант 84
— — циклической 87
— разделения переменных 190
— редукция 136
— установления 259
- Невязка 275
Нормальное решение 227, 479
- Обобщенное решение 227, 479
Оператор монотонный 221, 501
— непрерывный 216
— нормальный 219
— перехода 265, 268
— положительно определенный 221
— потенциальный 512
— разрешающий 265, 268
— самосопряженный 219
— сильно монотонный 221, 501
— сопряженный 218
Операторы коммутативные 217
— энергетически эквивалентные 221
- Полином Чебышева I рода 57
— II рода 58
Полуутирационный метод Чебышева 318
Поправка 265, 275
Принцип сжатых отображений 228
— регуляризации 532
Производная Гато 216, 503
- Разностная схема 24
Разностные производные 27
— тождества 233
— формулы Грина 234
Разностный оператор 26
Регуляризатор 533
- Сетка 24
Сеточная функция 26
— — векторная 26
Сеточное уравнение 30
Собственное значение оператора 225, 226
Собственный элемент оператора 225, 227
Спектральный радиус 218, 376
- Упорядоченный чебышевский набор параметров 270, 280
Ускорение сходимости 360
Устойчивость вычислительная 275
— по априорным данным 324
- Функция Грина разностного оператора 239
- Числовой радиус оператора 220, 293
- Ядро оператора 217

*Александр Андреевич Самарский,
Евгений Сергеевич Николаев*

**МЕТОДЫ РЕШЕНИЯ
СЕТОЧНЫХ УРАВНЕНИЙ**

М., 1978 г., 592 стр. с илл.

Редактор Т. Н. Галишникова.
Техн. редактор С. Я. Шкляр.
Корректор Н. Д. Дорохова.

ИБ № 2049

Сдано в набор 24.03.78. Подписано к печати 16.08.78.
Бумага 60×90^{1/16}, тип. № 1. Литературная гарнитура.
Высокая печать. Условн. печ. л. 37. Уч.-изд. л. 36,37.
Тираж 18 000 экз. Заказ № 2566. Цена книги 1 р. 40 к.

Издательство «Наука»
Главная редакция физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

Ордена Октябрьской Революции
и ордена Трудового Красного Знамени
Первая Образцовая типография имени А. А. Жданова
Союзполиграфпрома при Государственном комитете
Совета Министров СССР по делам издательств,
полиграфии и книжной торговли.
Москва, М-54, Валовая, 28