
В. И. КРЫЛОВ, Н. Е. ВОЗЖЕВ, ПИ. МОЧАСТЫРНИИ

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

II

В. И. КРЫЛОВ, В. В. БОБКОВ,
П. И. МОНАСТЫРНЫЙ

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

ТОМ II

*Допущено Министерством
высшего и среднего специального образования СССР
в качестве учебного пособия
для студентов высших учебных заведений*



ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1977

Вычислительные методы, том II. В. И. Крылов, В. В. Бобков, П. И. Монастырный. Главная редакция физико-математической литературы изд-ва «Наука», М., 1977.

Книга является второй частью пособия, предназначенного для студентов высших технических учебных заведений, физических и механико-математических факультетов университетов. Она может служить справочником для всех лиц, которым приходится иметь дело с научными и техническими расчетами.

В книге содержится изложение методов решения обыкновенных дифференциальных уравнений, дифференциальных уравнений с частными производными и интегральных уравнений. Приведены также наиболее часто применяемые методы ускорения сходимости рядов и последовательностей. Кроме того, дано краткое изложение некоторых вопросов общей теории вычислительных методов на основе функционального анализа.

ОГЛАВЛЕНИЕ

Предисловие	7
Глава 6. Решение обыкновенных дифференциальных уравнений	
§ 1. Методы решения задачи Коши. Вводные замечания . . .	10
§ 2. Построение одношаговых методов способом разложения решения в ряд Тейлора	14
1. Пошаговый вариант метода рядов (14). 2. Видоизменение метода рядов, не требующее вычисления производных правой части уравнения (15).	
§ 3. Способ Рунге—Кутта построения одношаговых методов	21
§ 4. Построение вычислительных правил на основе принципа последовательного повышения порядка точности результата	32
1. Случай уравнений первого порядка (32). 2. Случай уравнений высших порядков (33).	
§ 5. Сходимость одношаговых методов	42
§ 6. Главный член погрешности. Правило Рунге	53
§ 7. Методы типа двусторонних	59
§ 8. Многоступенчатые методы	72
1. Экстраполяционные методы Адамса (74). 2. Интерполяционные методы Адамса (79). 3. Методы с забеганием вперед (85).	
§ 9. Решение линейных граничных задач	93
1. Постановка задачи. Понятие о многоточечных и граничных задачах (93). 2. Метод дифференциальной прогонки (95).	
§ 10. Решение нелинейных граничных задач	100
1. Метод редукции к задачам Коши (101). 2. Метод линеаризации (104).	
§ 11. Метод сеток для решения линейных граничных задач . .	107
1. Постановка задачи. Идея метода сеток (107). 2. Методы замены обыкновенных дифференциальных уравнений и граничных условий системой алгебраических уравнений (108). 3. О разрешимости систем разностных уравнений (111). 4. Метод ортогональной прогонки (113). 5. Оценка погрешности и сходимость метода сеток (117).	
§ 12. Метод Галеркина и метод моментов	120
§ 13. Метод наименьших квадратов и метод Рунца	129
Литература	140

Глава 7. Решение дифференциальных уравнений в частных производных

§ 1. Разностные схемы, основные понятия	141
1. Сходимость и аппроксимация разностных схем (142).	
2. Связь аппроксимации и устойчивости со сходимостью (145).	
§ 2. Разностные схемы для уравнений параболического типа . .	146
1. Решение задачи Коши (146). 2. Об устойчивости двух- слойных разностных схем (151). 3. О разностных схемах расщепления (155). 4. Решение смешанных граничных задач (159). 5. Необходимое спектральное условие устой- чивости Неймана (162). 6. Экономичные разностные схемы (165).	
§ 3. Разностные схемы для уравнений эллиптического типа . .	175
1. Построение разностных аппроксимаций для уравнений (176). 2. Аппроксимация граничных условий (180). 3. По- строение разностной схемы в случае задачи Дирихле для уравнения Пуассона (183). 4. Об устойчивости разностной схемы (20) (184). 5. Метод матричной прогонки (188). 6. Правило Рунге (191). 7. Метод итераций для разностной задачи Дирихле (192).	
§ 4. Разностные схемы для уравнений гиперболического типа . .	201
1. Решение задачи Коши (203). 2. Решение смешанной задачи (208). 3. Об устойчивости явной трехслойной раз- ностной схемы (210).	
§ 5. Метод характеристик	215
1. Характеристики дифференциального уравнения (215). 2. Построение сетки характеристик (221). 3. Нахождение характеристических полосок (224).	
§ 6. Метод прямых	228
1. Поперечные схемы метода прямых для уравнений пара- болического типа (229). 2. Продольные схемы метода пря- мых для параболических уравнений (234). 3. Метод пря- мых для уравнений гиперболического типа (238). 4. Метод прямых для эллиптических уравнений (241).	
§ 7. Метод интегральных соотношений	244
§ 8. Понятие о методе моментов и методе Галёркина	248
Литература	251

Глава 8. Численное решение интегральных уравнений

§ 1. Введение	252
§ 2. Уравнения вольтеррова вида второго рода	258
1. Вид вычислительного правила и теорема о сходимости вычислительного процесса (258). 2. Правила вычислений, основанные на формуле Эйлера—Маклорена (264).	
§ 3. Уравнения Фредгольма второго рода. Метод квадратур . .	266
1. Метод квадратур (267). 2. Сходимость метода квадратур (273). 3. Интерполяционный квадратурный метод (275).	
§ 4. Метод замены ядра уравнения на вырожденное ядро . . .	278
1. Некоторые сведения об уравнениях с вырожденным ядром (278). 2. Замечания о методах построения вырожденного	

ядра, близкого к ядру уравнения (280). Оценка близости между решениями уравнений в зависимости от близости самих уравнений (282).

Литература	285

Глава 9. Улучшение сходимости рядов и последовательностей

§ 1. Введение	286
1. О задаче улучшения сходимости (286). 2. Связь с проблемой суммирования расходящихся рядов и определения пределов расходящихся последовательностей (289).	
§ 2. Интерполяционные методы преобразования последовательности	297
1. Некоторые общие сведения (297). 2. Интерполирование при помощи многочленов от $1/n$. Случай постоянного числа узлов (301). 3. Некоторые типы сходимости (308). 4. Вспомогательная лемма (309). 5. Некоторые теоремы об ускорении сходимости (312). 6. Интерполяционные методы преобразования с возрастающим числом узлов (314). 7. Обращение преобразования (316). 8. Теорема об ускорении сходимости (320).	
§ 3. Улучшение сходимости последовательностей, для которых погрешность приближения к пределу близка к показательной функции или линейной комбинации таких функций	325
1. Простейшее правило преобразования и некоторые его свойства (325). 2. Об улучшении сходимости (327). 3. Применение δ^2 -преобразования к улучшению сходимости и аналитическому продолжению степенного ряда (329). 4. Связь с методом интерполирования предельного значения (331). 5. О пути уточнения δ^2 -преобразования (333). 6. Правило ускорения сходимости последовательности, для которой погрешность приближения к пределу близка к линейной комбинации показательных функций (335). 7. Теорема об ускорении сходимости (342). 8. Об улучшении сходимости и аналитическом продолжении степенного ряда мероморфной функции (349).	
§ 4. Улучшение сходимости ряда с помощью выделения медленно сходящейся части	351
1. Введение (351). 2. Числовые ряды. Случай убывания членов ряда по степенному закону (352). 3. Улучшение сходимости тригонометрического ряда (355).	
Литература	357

Глава 10. Применение функционального анализа к построению теории некоторых разделов вычислительной математики

§ 1. Необходимые сведения из функционального анализа	358
1. Метрические пространства. Сходимость и полнота (358).	
2. Линейные нормированные пространства. Линейные операторы (361). 3. Дифференцирование нелинейных операторов (365).	

§ 2. Метод итерации для решения операторных уравнений . .	371
1. Метод итерации для операторных уравнений (372). 2. Теоремы о сходимости и единственности (373). 3. Приложение к системам уравнений с числовыми неизвестными (376).	
§ 3. Метод Ньютона для операторных уравнений	380
1. Правило построения приближений к решению (380). 2. Теорема о сходимости последовательности приближений (381).	
§ 4. Задача упрощения уравнений	387
1. Введение (387). 2. О разрешимости приближенного уравнения (389). 3. Оценка погрешности приближения и условия сходимости (392).	
Литература	394
Указатель обозначений	395
Предметный указатель	396

ПРЕДИСЛОВИЕ

Главы 6 и 7, занимающие большую часть книги, посвящены изложению основных сведений из теории численного решения дифференциальных уравнений. Такие уравнения имеют очень широкую область применения и лежат в основании многих разделов теоретической и, особенно, прикладной науки. Вид дифференциального уравнения тесно связан с природой исследуемого явления, и уравнения могут различаться между собой по их порядкам, по числу переменных, по типам, зависящим от того, будет ли процесс установившимся или развивающимся, по дополнительным условиям — будут ли они начальными, граничными или смешанными, по особенностям, вид которых зависит от характерных свойств процесса и т. д. Этот перечень приведен, чтобы пояснить, насколько многообразными могут быть дифференциальные уравнения, встречающиеся в приложениях. Для задачи каждого вида должны быть развиты, как правило, свои методы решения, и поэтому проблема численного решения уравнений является весьма разветвленной.

Учебная книга обязана быть в какой-то мере универсальной и должна содержать изложение всех основных и наиболее часто применяемых методов. Добиться же этого на небольшом числе страниц можно, если только отказаться от полного изложения и ограничиться описанием лишь идей методов и начал их теории. Чтобы не увеличивать принятый объем книги, авторы вынуждены были опустить или изложить очень кратко многие важные вопросы. Например, в проблеме сходимости последовательности приближенных решений к точному авторы могли рассмотреть только простейшие случаи. Теория конечно-разностных методов решения уравнений с частными производными, ввиду очень большого значения их в приложениях математики, за последние два десятилетия разраба-

тивалась особенно интенсивно, и здесь были получены многие важные результаты, авторы же имели возможность включить в книгу лишь основные понятия и некоторые простейшие факты. В очень сжатом виде изложены также вопросы приближенного решения интегральных уравнений. Для уравнений, содержащих интегралы с переменной границей, дано описание общей вычислительной схемы и приведена только одна конкретная вычислительная схема, основанная на формуле Эйлера — Маклорена.

Несколько более подробно рассмотрены уравнения с интегралами, имеющими фиксированные границы интегрирования. Для них были избраны два метода: метод квадратур и метод замены ядра уравнения $K(x, s)$ на так называемое вырожденное ядро.

В книгу включены также вопросы ускорения сходимости рядов и последовательностей. Авторы избрали здесь три следующие задачи.

1. Методы ускорения сходимости последовательностей, основанные на алгебраическом интерполировании предельного значения.

2. Методы, предназначенные для ускорения сходимости последовательностей, стремящихся к пределу по закону, близкому к изменению показательных функций или их линейных комбинаций.

3. Ускорение сходимости рядов при помощи выделения медленно сходящейся части.

Наконец, авторы считали полезным хотя бы кратко описать, как может быть применена теория операторов к построению общей теории приближенных и, в частности, численных методов математики.

Параграфы 1 — 8 главы 6 и §§ 5 — 7 главы 7 написаны В. В. Бобковым, §§ 9 — 13 главы 6 и §§ 1 — 4, 8 главы 7 написаны П. И. Монастырным. Весь остальной текст книги принадлежит В. И. Крылову.

Авторы

ГЛАВА 6

РЕШЕНИЕ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Среди задач, с которыми приходится иметь дело в вычислительной практике, значительную часть составляют различные задачи для обыкновенных дифференциальных уравнений. Такие задачи возникают как непосредственно при математическом моделировании многих реальных явлений, так и в качестве промежуточных при решении ряда более сложных математических задач. При этом, как правило, точное решение рассматриваемой задачи не удастся выразить через элементарные функции. Доля задач, решаемых в явном виде, в случае обыкновенных дифференциальных уравнений ничтожно мала. Обычно приходится прибегать к помощи приближенных методов решения подобных задач. Такие методы в зависимости от того, ищется ли приближенное решение в аналитическом виде или в виде таблицы чисел, часто разделяют на *аналитические* и *численные*. Подобное разбиение методов на две группы основано на достаточно грубых признаках и носит общий характер. Конкретный вид метода существенно зависит прежде всего от типа решаемой дифференциальной задачи. В случае обыкновенных дифференциальных уравнений в зависимости от того, ставятся ли дополнительные условия в одной или в нескольких точках отрезка изменения независимой переменной, задачи обычно подразделяют на *одноточечные* (задачи с начальными условиями, или задачи Коши) и *многоточечные*. Среди многоточечных задач наиболее часто в прикладных вопросах встречаются так называемые *граничные задачи*, когда дополнительные условия ставятся на концах рассматриваемого отрезка.

Наиболее простой из указанных выше задач является, вообще говоря, задача Коши. Заметим, что нередко решение многоточечных и, в частности, граничных задач сводят к решению задач с начальными условиями. Задачи Коши очень важны для приложений и в книге им уделено много внимания.

Рассмотрение методов решения обыкновенных дифференциальных уравнений мы начнем со случая задачи Коши для уравнений первого порядка.

§ 1. Методы решения задачи Коши.

Вводные замечания

Для простоты изложения основных идей вычислительных методов решения задач с начальными условиями будем рассматривать, как правило, случай одного обыкновенного дифференциального уравнения первого порядка. Обычно эти идеи легко переносятся на системы уравнений первого порядка и сравнительно просто обобщаются на случай уравнений высших порядков.

Пусть на отрезке $x_0 \leq x \leq X$ требуется найти решение $y(x)$ дифференциального уравнения

$$y' = f(x, y), \quad (1)$$

удовлетворяющее при $x = x_0$ начальному условию

$$y(x_0) = y_0. \quad (2)$$

Условия существования и единственности решения поставленной задачи Коши будем считать выполненными. Будем предполагать также, что функция $f(x, y)$ в некоторой области изменения ее аргументов обладает необходимой по ходу изложения дополнительной гладкостью. Основной нашей целью на данном этапе будет построение вычислительных правил нахождения приближенного решения рассматриваемой задачи.

Из общего курса обыкновенных дифференциальных уравнений хорошо известен *метод последовательных приближений*, или *метод Пикара* приближенного решения задачи (1), (2). Этот метод позволяет получать в аналитическом виде последовательность приближений $y_m(x)$, $m = 1, 2, \dots$, к решению $y(x)$ рассматриваемой задачи

Коши по следующему итерационному правилу:

$$y_m(x) = y_0 + \int_{x_0}^x f[t, y_{m-1}(t)] dt, \quad x_0 \leq x \leq X,$$

$$y_0(t) \equiv y_0, \quad x_0 \leq t \leq X, \quad m = 1, 2, \dots$$

Метод Пикара, однако, редко используется в практике вычислений. Одним из его существенных недостатков, препятствующих широкому практическому применению метода, является необходимость выполнения операции интегрирования при осуществлении каждой итерации.

Несколько более широкое распространение в вычислительной практике получил другой аналитический метод, основанный на идее разложения в ряд решения рассматриваемой задачи Коши. Особенно часто для этих целей используют ряд Тейлора. В этом случае вычислительные правила строятся особенно просто. Приближенное решение $y_m(x)$ исходной задачи ищут в виде

$$y_m(x) = \sum_{i=0}^m \frac{(x-x_0)^i}{i!} y^{(i)}(x_0), \quad x_0 \leq x \leq X, \quad (3)$$

где

$$y^{(0)}(x_0) = y(x_0) = y_0, \quad y^{(1)}(x_0) = y'(x_0) = f(x_0, y_0),$$

а значения $y^{(i)}(x_0)$, $i = 2, 3, \dots, m$, находят по формулам, полученным последовательным дифференцированием уравнения (1):

$$y^{(2)}(x_0) = y''(x_0) = f_x(x_0, y_0) + f(x_0, y_0) f_y(x_0, y_0),$$

$$y^{(3)}(x_0) = y'''(x_0) = f_{x^2}(x_0, y_0) +$$

$$+ 2f(x_0, y_0) f_{xy}(x_0, y_0) + f^2(x_0, y_0) f_{y^2}(x_0, y_0) +$$

$$+ f_y(x_0, y_0) [f_x(x_0, y_0) + f(x_0, y_0) f_y(x_0, y_0)], \quad (4)$$

$$\dots$$

$$y^{(m)}(x_0) =$$

$$= F_m(f; f_x, f_y; f_{x^2}, f_{xy}, f_{y^2}; \dots; f_{x^{m-1}}, \dots, f_{y^{m-1}})_{x=x_0, y=y_0}$$

(конкретный вид многочлена F_m не приведен из-за громоздкой записи). Для значений x , близких к x_0 , метод рядов (3) при достаточно большом m дает обычно хорошее приближение к точному решению $y(x)$ задачи (1),

(2). Однако с увеличением расстояния $|x - x_0|$ погрешность приближенного равенства $y(x) \approx y_m(x)$, вообще говоря, возрастает по абсолютной величине и правило (3) становится вовсе неприемлемым, когда x выходит из области сходимости соответствующего (3) ряда Тейлора.

Предпочтительными в таких случаях будут, например, численные методы решения задачи Коши, позволяющие в некоторых попарно близких друг другу фиксированных точках (узлах)

$$x_0 < x_1 < x_2 < \dots < x_N = X \quad (5)$$

последовательно находить значения $y_n \approx y(x_n)$, $n = 1, 2, \dots, N$, приближенного решения. Сходимость методов подобного типа не так жестко связана с длиной отрезка $[x_0, X]$, и их чаще кладут в основу стандартных программ для ЭВМ. Таким методам ниже будет уделено основное внимание.

Большинство численных методов решения рассматриваемой задачи Коши можно привести к виду

$$y_{n+1} = F(y_{n-q}, y_{n-q+1}, \dots, y_n, y_{n+1}, \dots, y_{n+s}), \quad (6)$$

где F — некоторая известная функция указанных аргументов, определяемая способом построения метода и зависящая от вида уравнения (1) и избранной сетки (5). При $q=0$, $0 \leq s \leq 1$ такие вычислительные правила обычно называют *одношаговыми*, а при $q \geq 1$ или $s > 1$ — *многошаговыми*. Как одношаговые, так и многошаговые методы вида (6) называют *явными* в случае $s=0$ и *неявными* при $s=1$. В случае $s > 1$ многошаговые правила часто называют методами *с забеганием вперед*. С примерами правил указанных типов, особенностями их использования и основными способами построения мы познакомимся несколько позже, а сейчас сделаем лишь несколько замечаний общего характера.

Если правило (6) является одношаговым, то вычисления по этому правилу можно начинать со значения $n=0$ и проводить до значения $n=N-1$ включительно. В случае же многошаговых методов указанного вида дискретный аргумент n может пробегать значения лишь от $n=q$ по $n=N-s$, что, вообще говоря, влечет за собой нарушение однородности вычислительного процесса и требует

применения специальных вычислительных правил для нахождения первых q значений $y_1, y_2, \dots, y_q$ приближенного решения и последних его $s-1$ значений $y_{N-s+2}, y_{N-s+3}, \dots, y_N$. Одношаговые методы в этом отношении являются предпочтительными. Удобнее пользоваться ими и в том случае, когда шаг сетки $h_n = x_{n+1} - x_n$ не является постоянным для всех значений n (скажем, величина шага интегрирования выбирается ЭВМ автоматически по результатам вычислений). Вид одношаговых правил не связан с величиной шага на предыдущем этапе вычислительного процесса и поэтому эти правила легко допускают изменение шага численного интегрирования, в то время как многошаговые методы нужно специальным образом приспособлять для этих целей. Полезно также иметь в виду, что при построении многошаговых методов (в силу их вида) должна использоваться информация об исходной задаче Коши на отрезке $[x_{n-q}, x_{n+s}]$, в то время как в случае одношаговых правил подобную информацию нельзя привлекать за пределами отрезка $[x_n, x_{n+1}]$. Это предопределяет концентрирование такой информации на малом отрезке в случае одношаговых методов и рассредоточение ее на отрезке длиной в несколько шагов интегрирования в случае многошаговых правил. Последние соображения дают основания ожидать, что в областях резкого изменения поведения функций предпочтительнее будут методы, при построении которых информация о таких функциях (при сохранении ее количества) привлекалась бы более концентрированно. Эти же соображения поясняют и основной недостаток одношаговых методов (в сравнении с многошаговыми), связанный с их трудоемкостью. Так как в случае одношаговых правил информация о решаемой задаче используется лишь в пределах одного шага, то при переходе от шага к шагу ее, вообще говоря, нужно каждый раз заново получать, в то время как в случае многошаговых методов предоставляется возможность повторного использования части такой информации на нескольких соседних этапах вычислений, что позволяет уменьшить затраты вычислительного труда на каждый шаг численного интегрирования. Трудоемкость одношаговых методов долгое время была основным препятствием на пути их широкого практического применения. До появления вычисли-

тельных машин такие методы чаще использовались лишь для построения начала таблицы, а основной счет выполнялся обычно с помощью многшаговых методов. С ростом быстродействия ЭВМ область применения одношаговых правил быстро расширяется. С рассмотрения таких правил мы и начнем изложение основных численных методов решения задачи Коши.

§ 2. Построение одношаговых методов способом разложения решения в ряд Тейлора

1. Пошаговый вариант метода рядов. Будем считать, что процесс решения задачи (1.1), (1.2) доведен до точки x_n ($0 \leq n < N$), и известно (точно или приближенно) соответствующее значение $y(x_n)$ искомого решения. Построим вычислительное правило для нахождения значения решения в очередной узловой точке $x_{n+1} = x_n + h_n$ сетки (1.5). Поскольку при построении одношаговых методов используется информация о решаемой задаче лишь в пределах одного шага интегрирования, то можно без ущерба для понимания не писать индекс, означающий номер шага процесса.

Чтобы по известному значению $y(x)$ соответствующего решения дифференциального уравнения (1.1) в узловой точке $x \geq x_0$ найти значение этого решения в очередной точке $x+h$ сетки (1.5), можно, очевидно, воспользоваться вычислительным правилом типа (1.3), положив там $x_0 = x$ и взяв вместо текущей точки x отрезка $[x_0, X]$ узловую точку $x+h$. Это позволит записать приближенное равенство

$$y_m(x+h) \approx \sum_{i=0}^m \frac{h^i}{i!} y^{(i)}(x), \quad (1)$$

которое может быть положено в основу соответствующего одношагового метода, если для вычисления значений $y^{(i)}(x)$, $i = 2, 3, \dots, m$, использовать формулы типа (1.4). При условии, что данное решение уравнения (1.1) имеет на рассматриваемом отрезке непрерывную производную порядка $m+1$, погрешность приближенного равенства (1) будет, очевидно, величиной порядка h^{m+1} и при малых $h > 0$ и больших m построенный пошаговый вариант метода рядов будет давать, как правило,

достаточно хорошее приближение к искомому значению решения. Привлекательной чертой полученного вычислительного метода является то обстоятельство, что искомое значение приближенного решения (см. (1)) разложено по последовательным главным частям, что позволяет в процессе решения задачи без дополнительных вычислительных затрат по величине последних слагаемых суммы (1) составить представление о *локальной* погрешности приближенного решения (погрешности нахождения значения $y(x+h)$ в предположении, что значение $y(x)$ известно точно). Однако такой одношаговый метод интегрирования дифференциальных уравнений при $m > 1$ все же редко используется в практике вычислений, так как его применение требует на каждом шаге нахождения значений $\frac{m(m+1)}{2}$ различных функций $f, f_x, f_y, f_{x^2}, f_{xy}, \dots, f_{y^{m-1}}$ (см. (1.4)). При использовании ЭВМ это сопряжено с написанием большого числа блоков вычисления значений указанных функций, что осложняет связь пользователя с машиной и увеличивает, как правило, время решения задачи. Поэтому данный метод редко кладут в основу стандартных программ решения задач Коши, хотя в специальных частных случаях (например, когда приходится многократно решать задачи, отличающиеся лишь начальными данными или мало отличающиеся правыми частями уравнений) использование метода может быть оправданным. В общем случае стандартные программы чаще строятся с использованием вычислительных методов, не требующих нахождения значений производных от правых частей уравнений. Ниже будут рассмотрены основные способы построения таких методов.

2. Видоизменение метода рядов, не требующее вычисления производных правой части уравнения. Естественно поставить задачу о таком усовершенствовании приведенного выше одношагового метода, которое сохраняло бы основные его достоинства, но не было связано с нахождением значений производных правой части уравнения (1.1). Чтобы выполнить последнее условие, производные $y^{(i)}(x)$, $i = 2, 3, \dots, m$, входящие в правую часть равенства (1), можно заменить по формулам численного дифференцирования (см. § 6 гл. 1) их приближенными

выражениями через значения функции y' и учесть, что $y'(x) = f[x, y(x)]$. Требование одношаговости конструируемых правил накладывает при этом свои условия на такую замену. Ниже на конкретных примерах будет рассмотрен один из возможных подходов к решению поставленной задачи.

В случае $m=1$ приближенное равенство (1) не требует вычисления производных правой части уравнения (1.1) и позволяет с погрешностью порядка h^2 находить значение $y(x_n + h)$ решения этого уравнения по известному его значению $y(x_n)$. Соответствующее одношаговое правило можно записать в виде

$$y_{n+1} = y_n + h f_n^{[2]}, \quad (2)$$

используя широко применяемое ниже обозначение

$$f(x_n + \alpha h, y_{n+\alpha}^{[k]}) = f_{n+\alpha}^{[k]}, \quad (3)$$

где $y_{n+\alpha}^{[k]}$ — найденное с локальной ошибкой порядка h^k приближенное значение решения в точке $x_n + \alpha h$.

Правило (2) впервые было построено Эйлером и носит его имя. Иногда его называют также правилом *ломаных* или методом *касательных*, чем подчеркивают простой геометрический смысл формулы. Погрешность r_{n+1} этой формулы можно, очевидно, записать в виде

$$r_{n+1} = \frac{h^2}{2} y''(x_n + \theta h), \quad 0 < \theta < 1.$$

При $m=2$ равенство (1) требует вычисления производной $y''(x_n)$ и дает возможность находить значение $y(x_n + h)$ с локальной ошибкой порядка h^3 . Чтобы не понизить порядок погрешности приближенного равенства

$$y(x_n + h) \approx y(x_n) + h y'(x_n) + \frac{h^2}{2} y''(x_n), \quad (4)$$

значение производной $y''(x_n)$ необходимо найти по крайней мере не хуже, чем с ошибкой порядка h , для чего, очевидно, достаточно иметь два значения функции $y'(x) = f[x, y(x)]$ из отрезка $x_n \leq x \leq x_n + h$ ($0 < h < 1$).

В точке x_n значением $f_n^{[3]}$ этой функции мы уже располагаем, так как, по предположению, на предыдущем

шаге процесса было найдено $y_n^{[3]}$. Заметим, что в силу локального характера наших рассуждений значение $y_n^{[3]}$ можно рассматривать как точное значение $y(x_n)$ того решения уравнения (1.1), значение которого в точке $x_n + h$ интересует нас на данном этапе вычислений.

Найдем теперь еще значение функции y' в точке $x_n + h^i$ ($i \geq 1$). Для этого с учетом уравнения (1.1) достаточно указать правило вычисления $y(x_n + h^i)$.

Очевидно, $y(x_n + h^i) = y(x_n) + h^i y'(x_n) + O(h^{2i})$. Поэтому справедлива следующая расчетная формула

$$y_{n+h^{i-1}}^{[2i]} = y_n^{[3]} + h^i f_n^{[3]}. \quad (5)$$

Так как

$$y''(x_n) = \frac{y'(x_n + h^i) - y'(x_n)}{h^i} + O(h^i),$$

$$y'(x_n + h^i) = f[x_n + h^i, y(x_n + h^i)],$$

то на основании (4) можно записать

$$\begin{aligned} y_{n+1}^{[3]} &= y_n^{[3]} + h f_n^{[3]} + \frac{h^{2-i}}{2} \left(f_{n+h^{i-1}}^{[2i]} - f_n^{[3]} \right) = \\ &= y_n^{[3]} + h \left(1 - \frac{h^{1-i}}{2} \right) f_n^{[3]} + \frac{h^{2-i}}{2} f_{n+h^{i-1}}^{[2i]}. \end{aligned} \quad (6)$$

Формулы (5), (6) можно рассматривать как семейство (зависящих от параметра $i \geq 1$) одношаговых методов решения задачи Коши (1.1), (1.2) с локальной погрешностью порядка h^3 .

При $i=1$ эти формулы принимают вид

$$y_{n+1}^{[2]} = y_n^{[3]} + h f_n^{[3]}, \quad (7)$$

$$y_{n+1}^{[3]} = y_n^{[3]} + \frac{h}{2} \left(f_n^{[3]} + f_{n+1}^{[2]} \right). \quad (8)$$

Они имеют предсказывающе-исправляющий характер: формула (7) служит для получения грубого приближения искомой величины $y(x_n + h)$, а по формуле (8) произво-

*) Здесь и ниже величины x_n и h считаются безразмерными.

дится уточнение полученного значения. Сравнение y_{n+1} с y_n дает возможность высказать суждение о локальной погрешности результата.

Заметим, что иногда на основе формулы (8) бывает полезным сделать одну итерацию. Это может (при сохранении порядка погрешности) несколько повысить точность приближения величины $y(x_n + h)$. К существенному увеличению объема вычислений эта итерация не приведет, так как значение f_{n+1} , которое при этом необходимо будет вычислить, можно использовать в качестве f_n на следующем этапе вычислений. И наоборот, если вместо формул (7), (8) применить формулы

$$y_{n+1} = y_n + hf_n, \quad (9)$$

$$y_{n+1} = y_n + \frac{h}{2} (f_n + f_{n+1}), \quad (10)$$

то значение f_{n+1} можно будет использовать в качестве f_n на следующем этапе счета. Погрешность приближенного решения в случае метода (9), (10), вообще говоря, несколько возрастет (при сохранении порядка), но это позволит почти вдвое уменьшить общий объем вычислений в сравнении с методом (7), (8). Подобное специфичное использование информации о решении, полученной на предыдущем этапе вычислений, не является, вообще говоря, характерным для одношаговых методов. При этом, в частности, несколько нарушается однородность вычислительного процесса, так как первый его этап будет сопряжен с нахождением двух значений функции $f(x, y)$, а каждый последующий будет требовать лишь одного обращения к блоку вычисления значений этой функции.

В случае $i=2$ формулы (5), (6) имеют вид

$$y_{n+h} = y_n + h^2 f_n, \quad (11)$$

$$y_{n+1} = y_n + hf_n + \frac{1}{2} (f_{n+h} - f_n) = y_n + \left(h - \frac{1}{2}\right) f_n + \frac{1}{2} f_{n+h}. \quad (12)$$

Увеличение значения i на единицу позволило, вообще говоря, несколько улучшить структуру остаточного члена вычислительного правила. Если в случае правила (7), (8) погрешность складывалась из одного слагаемого вида $\frac{h^3}{6} y'''(x_n + \theta h)$, $0 < \theta < 1$, представляющего собой ошибку приближенного равенства (4), и еще двух слагаемых также порядка h^3 , порождаемых соответственно неточностью замены производной y'' и ошибкой формулы Эйлера (5), то в случае (11), (12) два последних слагаемых остаточного члена будут уже, очевидно, величинами порядка h^4 . При $i = 3$ соответствующие слагаемые остатка станут величинами порядка h^5 и т. д. Однако увеличение значения i , улучшая структурные свойства остаточного члена вычислительного правила, предъявляет повышенные требования к выполнению вычислений, так как в этом случае в силу ограниченности разрядной сетки ЭВМ возможна потеря точности результата за счет операций вычитания близких величин и деления на малые по абсолютному значению числа (см. (6)).

Рассмотрим далее случай $m = 3$. Соответствующее разложение решения по формуле Тейлора будет сейчас обрываться на члене, содержащем вторую производную функции y' . Чтобы построить приближенное выражение для этой производной через значения функции, нужно иметь по меньшей мере три таких значения. Возьмем в качестве них $y'(x_n)$, $y'(x_n + h^i)$ и $y'(x_n + 2h^i)$, где $i \geq 2$ при $0 < h \leq 1/2$. Чтобы воспользоваться простейшими симметричными приближенными выражениями для первой и второй производных функции y' через эти значения, будем исходить из разложения решения по формуле Тейлора не около точки x_n , а около точки $x_n + h^i$:

$$\begin{aligned} y(x_n + h) &= y[x_n + h^i + h(1 - h^{i-1})] = \\ &= y(x_n + h^i) + h(1 - h^{i-1}) y'(x_n + h^i) + \\ &+ \frac{h^2(1 - h^{i-1})^2}{2} y''(x_n + h^i) + \frac{h^3(1 - h^{i-1})^3}{6} y'''(x_n + h^i) + O(h^4). \end{aligned}$$

Опираясь на это представление, можно записать следующее семейство вычислительных правил (подобно

изложенному в случае $m=2$):

$$\begin{aligned}
 {}^{[2i]}y_{n+h^{i-1}} &= {}^{[4]}y_n + h^i {}^{[4]}f_n, \\
 {}^{[3i]}y_{n+h^{i-1}} &= {}^{[4]}y_n + \frac{h^i}{2} \left({}^{[4]}f_n + {}^{[2i]}f_{n+h^{i-1}} \right), \\
 {}^{[3i]}y_{n+2h^{i-1}} &= {}^{[3i]}y_{n+h^{i-1}} + \frac{h^i}{2} \left(3 {}^{[3i]}f_{n+h^{i-1}} - {}^{[4]}f_n \right), \\
 {}^{[4]}y_{n+1} &= {}^{[3i]}y_{n+h^{i-1}} + h(1-h^{i-1}) {}^{[3i]}f_{n+h^{i-1}} + \\
 &\quad + \frac{h^{2-i}(1-h^{i-1})^2}{4} \left({}^{[3i]}f_{n+2h^{i-1}} - {}^{[4]}f_n \right) + \\
 &\quad + \frac{h^{3-2i}(1-h^{i-1})^3}{6} \left({}^{[3i]}f_{n+2h^{i-1}} - 2 {}^{[3i]}f_{n+h^{i-1}} + {}^{[4]}f_n \right) = \\
 &= {}^{[3i]}y_{n+h^{i-1}} + \frac{h^{2-i}(1-h^{i-1})^2}{2} \left(\frac{h^{1-i}-1}{3} - \frac{1}{2} \right) {}^{[4]}f_n + \\
 &\quad + h(1-h^{i-1}) \left[1 - \frac{(h^{1-i}-1)^2}{3} \right] {}^{[3i]}f_{n+h^{i-1}} + \\
 &\quad + \frac{h^{2-i}(1-h^{i-1})^2}{2} \left(\frac{h^{1-i}-1}{3} + \frac{1}{2} \right) {}^{[3i]}f_{n+2h^{i-1}}.
 \end{aligned}$$

При $i=2$, например, имеем расчетное правило

$$\begin{aligned}
 {}^{[4]}y_{n+h} &= {}^{[4]}y_n + h^2 {}^{[4]}f_n, \\
 {}^{[6]}y_{n+h} &= {}^{[4]}y_n + \frac{h^2}{2} \left({}^{[4]}f_n + {}^{[4]}f_{n+h} \right), \\
 {}^{[6]}y_{n+2h} &= {}^{[6]}y_{n+h} + \frac{h^2}{2} \left(3 {}^{[6]}f_{n+h} - {}^{[4]}f_n \right), \\
 {}^{[4]}y_{n+1} &= {}^{[6]}y_{n+h} + \frac{(1-h)^2}{2} \left(\frac{1-h}{3h} - \frac{1}{2} \right) {}^{[4]}f_n + \\
 &\quad + (1-h) \left[h - \frac{(1-h)^2}{3h} \right] {}^{[6]}f_{n+h} + \frac{(1-h)^2}{2} \left(\frac{1-h}{3h} + \frac{1}{2} \right) {}^{[6]}f_{n+2h}.
 \end{aligned}$$

Полезно иметь в виду, что многие из предлагаемых выше одношаговых методов позволяют, кроме приближенного значения $y(x_n+h)$, попутно находить также приближенные значения решения и в некоторых других точках отрезка $[x_n, x_n+h]$.

Описанным способом можно строить вычислительные методы и более высокого порядка точности. Следует,

однако, заметить, что при больших значениях m построение таких методов связано с приближенной заменой производных высоких порядков по интерполяционным формулам численного дифференцирования, а эта процедура, как известно, сопряжена обычно с повышенными требованиями к точности выполнения вычислений. Ниже мы рассмотрим другие способы получения одношаговых правил, которые непосредственно не связаны с подобной аппроксимацией производных.

§ 3. Способ Рунге — Кутта построения одношаговых методов

Изложим основную идею этого способа на примере задачи (1.1), (1.2).

Интегрируя уравнение (1.1) в пределах от x до $x+h$ ($0 < h < 1$), получим равенство

$$y(x+h) = y(x) + \int_x^{x+h} f[t, y(t)] dt, \quad (1)$$

которое посредством последнего интеграла связывает значения решения рассматриваемого уравнения в двух точках, удаленных друг от друга на расстояние шага h . Указав эффективный метод приближенного вычисления интеграла в (1), мы получим тем самым одно из правил численного интегрирования уравнения (1.1). В силу требования одношаговости конструируемых вычислительных правил при нахождении значения указанного интеграла мы можем использовать информацию об интегрируемой функции $y'(t) = f[t, y(t)]$ лишь на отрезке $[x, x+h]$. По постановке задачи значение этой функции в точке x нам известно. Поэтому для вычисления интеграла можно применить, например, одноточечную формулу левых прямоугольников

$$\int_a^b \varphi(t) dt \approx (b-a) \varphi(a).$$

В результате мы получим, очевидно, известный метод Эйлера (2.2). Погрешность формулы (2.2), как мы уже отмечали, является величиной порядка h^2 .

По формулам типа (1.4) в точке x можно найти и значения производных $y^{(i)}$, $i = 2, 3, \dots, m$, интегрируемой

функции y' . Используя для наших целей квадратурную формулу с единственным кратным узлом на левом конце отрезка интегрирования, мы приходим к пошаговому варианту типа (2.1) метода рядов.

Если же интеграл в (1) приближенно заменить квадратурной суммой двухточечной формулы трапеций (5.3.2), получим приближенное равенство

$$y(x+h) \approx y(x) + \frac{h}{2} [f(x, y(x)) + f(x+h, y(x+h))],$$

которое позволяет записать так называемый метод *трапеций* численного интегрирования уравнения (1.1):

$$y_{n+1} = y_n + \frac{h}{2} [f(x_n, y_n) + f(x_n+h, y_{n+1})]. \quad (2)$$

Погрешность последней формулы (см. (5.3.3)) будет величиной порядка h^3 .

Равенство (2) дает нам пример *неявного* одношагового метода. Неизвестное значение y_{n+1} здесь приходится находить путем решения численного уравнения (2). Обычно для этих целей используют итерационные методы. Подробнее на неявных методах решения дифференциальных уравнений мы остановимся несколько позже, а сейчас отметим лишь, что на основе равенства (2) с использованием формулы Эйлера (2.2) можно очевидным образом построить известное уже нам явное вычислительное правило (2.7), (2.8) типа предиктор-корректор того же порядка точности.

Более сложный пример неявного метода можно получить, если и на правом конце отрезка $[x, x+h]$ через значение функции $y(x+h)$ при помощи формул типа (1.4) выразить значения производных $y^{(i)}(x+h)$, $i = 2, 3, \dots, m$, и применить для вычисления интеграла в (1) соответствующую квадратурную формулу с двумя кратными узлами x и $x+h$. Однако использование такого неявного метода потребует на каждом шаге итераций вычисления значений производных правой части уравнения, что, как мы отмечали ранее, является существенным недостатком вычислительного правила. Чтобы построить методы более высокого порядка точности, не связанные с вычислением производных функции $f[x, y(x)]$, нужно привлечь дополнительную информацию по значениям интегрируемой

функции. Поскольку эта функция, как функция одного аргумента x , в случае задачи Коши (1.1), (1.2), в отличие от ее частного случая — рассмотренной уже нами ранее задачи вычисления неопределенного интеграла (5.10.1), — вообще говоря, неизвестна внутри отрезка интегрирования $[x, x+h]$, то использовать непосредственно квадратурные формулы с числом узлов $N > 2$ не удастся. Поэтому в способе Рунге — Кутта применяют следующий специальный метод приближенного вычисления интеграла в (1).

Для удобства записи используем обозначение $\Delta y = y(x+h) - y(x)$ и равенству (1) придадим новый вид, произведя замену переменной интегрирования $t = x + \alpha h$:

$$\Delta y = h \int_0^1 f[x + \alpha h, y(x + \alpha h)] d\alpha. \quad (3)$$

Чтобы на основе (3) построить одношаговый метод численного интегрирования уравнения (1.1), введем три набора параметров:

$$\alpha_1, \alpha_2, \dots, \alpha_q; \quad (\alpha)$$

$$\begin{aligned} \beta_{10}, \\ \beta_{20}, \beta_{21}, \\ \dots \dots \dots \end{aligned} \quad (\beta)$$

$$\begin{aligned} \beta_{q0}, \beta_{q1}, \dots, \beta_{q, q-1}; \\ A_0, A_1, \dots, A_q, \end{aligned} \quad (A)$$

выбором которых распорядимся в дальнейшем. При помощи двух первых наборов составим величины

$$\begin{aligned} \varphi_0 &= hf(x, y), \\ \varphi_1 &= hf(x + \alpha_1 h, y + \beta_{10} \varphi_0), \\ \varphi_2 &= hf(x + \alpha_2 h, y + \beta_{20} \varphi_0 + \beta_{21} \varphi_1), \\ &\dots \dots \dots \\ \varphi_q &= hf(x + \alpha_q h, y + \beta_{q0} \varphi_0 + \beta_{q1} \varphi_1 + \dots + \beta_{q, q-1} \varphi_{q-1}), \end{aligned}$$

которые при заданных (α) и (β) могут быть вычислены последовательно,

Хотя

$$\varphi_i = hf(x + \alpha_i h, y + \beta_{i0}\varphi_0 + \dots + \beta_{i,i-1}\varphi_{i-1}), \\ i = 1, 2, \dots, q,$$

вообще говоря, не равны значениям

$$hf[x + \alpha_i h, y(x + \alpha_i h)],$$

однако при соответствующем выборе параметров (β) их можно трактовать как приближенные значения интегрируемой функции $f[x + \alpha h, y(x + \alpha h)]$, умноженные на h . Это дает основание надеяться при помощи параметров (A) составить такую линейную комбинацию величин φ_i , $i = 0, 1, \dots, q$, которая будет являться аналогом квадратурной суммы и позволит вычислить приближенное значение приращения Δy (см. (3)):

$$\Delta y \approx \sum_{i=0}^q A_i \varphi_i. \quad (4)$$

Тем самым параметрам (α) , (β) и (A) можно придать некоторый квадратурный смысл.

Рассмотрим теперь задачу выбора этих параметров. Введем величину

$$r_q(h) = \Delta y - \sum_{i=0}^q A_i \varphi_i, \quad (5)$$

представляющую собой погрешность приближенного равенства (4). В предположении, что правая часть уравнения (1.1) является достаточно гладкой функцией, запишем следующее разложение этой величины:]

$$r_q(h) = \sum_{j=0}^k \frac{h^j}{j!} r_q^{(j)}(0) + \frac{h^{k+1}}{(k+1)!} r_q^{(k+1)}(\theta h), \quad 0 < \theta < 1.$$

На основании этого разложения можно утверждать, что если параметры (α) , (β) и (A) подобрать так, чтобы выполнялись условия

$$r_q^{(j)}(0) = 0, \quad j = 0, 1, \dots, k, \quad (6)$$

то погрешность (5) приближенного равенства (4) будет величиной порядка не ниже h^{k+1} :

$$r_q(h) = \frac{h^{k+1}}{(k+1)!} r_q^{(k+1)}(\theta h). \quad (7)$$

Число k при этом обычно называют *порядком точности* соответствующего метода. Такому определению порядка точности метода можно дать следующее объяснение. В случае рассматриваемой задачи Коши (1.1), (1.2), как и в случае изученной ранее задачи неопределенного интегрирования (5.10.1), оказывается, что если погрешность типа (5) расчетной формулы данного метода является величиной порядка h^{k+1} в достаточно широкой окрестности решения, то *погрешность метода* (та часть погрешности приближенного решения, которая определяется лишь неточностью самой формулы) для случая конечного отрезка интегрирования будет величиной порядка h^k . Доказательство этого факта будет приведено несколько позже, а сейчас продолжим рассмотрение вопроса о построении вычислительных правил по способу Рунге—Кутта.

Для выполнения условий (6) при возможно большем значении k величины $r_q^{(j)}(0)$, $j=0, 1, \dots, k$, выражают через значения функции $f(x, y)$ и ее частных производных и требуют обращения в нуль возможно большего числа этих величин для любой достаточно гладкой функции f . Иными словами, выбор параметров (α) , (β) и (A) осуществляют на основании требования, чтобы разложение

$$\Delta y = y(x+h) - y(x) = \frac{h}{1!} y'(x) + \frac{h^2}{2!} y''(x) + \frac{h^3}{3!} y'''(x) + \dots \quad (8)$$

и разложение по степеням h линейной комбинации $\sum_{i=0}^q A_i \varphi_i$ совпадали до членов с возможно более высокими степенями h в случае любой правой части уравнения (1.1).

При произвольном q систему уравнений для определения параметров (α) , (β) и (A) записать очень трудно. Поэтому мы ограничимся здесь рассмотрением лишь нескольких конкретных примеров построения одношаговых правил указанным способом.

Метод первого порядка точности. Зададимся минимальным значением $q=0$, что равнозначно введению лишь одного параметра A_0 . Приближенное равенство (4) в этом случае принимает вид

$$\Delta y \approx A_0 \varphi_0 = h A_0 f(x, y)$$

и погрешность (5) может быть записана следующим образом:

$$r_0(h) = y(x+h) - y(x) - hA_0f(x, y).$$

Тогда

$$r'_0(h) = y'(x+h) - A_0f(x, y),$$

$$r''_0(h) = y''(x+h).$$

Так как $r''_0(h)$ не зависит от A_0 , то уже при $j=2$ условие (6) в случае произвольной функции f удовлетворено быть не может. Поэтому $k=1$ и система (6) принимает вид

$$(1 - A_0)f(x, y) = 0.$$

Отсюда находим, что $A_0 = 1$. Следовательно (см. (4)),

$$\Delta y \approx hf(x, y)$$

и (см. (7))

$$r_0(h) = \frac{h^2}{2} r''_0(\theta h) = \frac{h^2}{2} y''(x + \theta h), \quad 0 < \theta < 1.$$

В простейшем случае $q=0$, таким образом, способ Рунге — Кутта приводит к известному методу Эйлера (2.2). Квадратурный смысл этого вычислительного правила был уже выяснен.

Методы второго порядка точности. При $q=1$ имеем

$$\begin{aligned} \Delta y &\approx A_0\varphi_0 + A_1\varphi_1 = \\ &= hA_0f(x, y) + hA_1f[x + \alpha_1h, y + h\beta_{10}f(x, y)]. \end{aligned}$$

С целью выбора введенных параметров $\alpha_1, \beta_{10}, A_0, A_1$ разложим Δy и $A_0\varphi_0 + A_1\varphi_1$ по степеням h .

Разложение (8) для Δy с учетом (1.1) можно записать в виде

$$\begin{aligned} \Delta y = y(x+h) - y(x) &= hf + \frac{h^2}{2}(f_x + ff_y) + \\ &+ \frac{h^3}{6}[f_{xx} + 2ff_{xy} + f^2f_{yy} + f_y(f_x + ff_y)] + O(h^4). \end{aligned} \quad (9)$$

Используя формулу Тейлора, для линейной комбинации $A_0\varphi_0 + A_1\varphi_1$ можно дать следующее представление:

$$\begin{aligned} A_0\varphi_0 + A_1\varphi_1 &= h(A_0 + A_1)f + h^2A_1(\alpha_1f_x + \beta_{10}ff_y) + \\ &+ \frac{h^3}{2}A_1(\alpha_1^2f_{xx} + 2\alpha_1\beta_{10}ff_{xy} + \beta_{10}^2f^2f_{yy}) + O(h^4). \end{aligned} \quad (10)$$

Сравним в разложениях (9), (10) коэффициенты при hf , h^2f_x и h^2ff_y . Тем самым на выбор четырех параметров α_1 , β_{10} , A_0 , A_1 будут наложены три условия:

$$A_0 + A_1 = 1, \quad A_1\alpha_1 = 1/2, \quad A_1\beta_{10} = 1/2.$$

Непосредственно из (9), (10) следует, что в случае $q = 1$ для произвольных f нельзя добиться совпадения всех членов с множителем h^3 за счет выбора введенных параметров. Поэтому при $q = 1$ вычислительные правила типа Рунге — Кутта будут иметь лишь второй порядок точности. Параметры α_1 , β_{10} , A_0 могут быть выражены через коэффициент A_1 по формулам

$$\alpha_1 = \beta_{10} = \frac{1}{2A_1}, \quad A_0 = 1 - A_1.$$

В качестве A_1 может быть взято, вообще говоря, произвольное отличное от нуля число. Например, при $A_1 = 1/2$ будем иметь

$$\Delta y = \frac{1}{2} (\varphi_0 + \varphi_1) + O(h^3), \quad (11)$$

$$\varphi_0 = hf(x, y), \quad \varphi_1 = hf(x + h, y + \varphi_0).$$

Вычислительное правило, построенное на основе равенства (11), имеет следующий квадратурный смысл. Так как

$$\varphi_1 = hf(x + h, y + hy') \approx hf[x + h, y(x + h)],$$

то линейная комбинация $\frac{1}{2}(\varphi_0 + \varphi_1)$ является аналогом квадратурной суммы формулы трапеций (5.3.2) при вычислении интеграла в правой части равенства (1).

Выбрав $A_1 = 1$, получим еще одно из широко известных вычислительных правил типа Рунге — Кутта второго порядка точности:

$$\Delta y \approx \varphi_1, \quad (12)$$

$$\varphi_0 = hf(x, y), \quad \varphi_1 = hf\left(x + \frac{h}{2}, y + \frac{\varphi_0}{2}\right).$$

Формула (12) является, очевидно, аналогом односточной квадратурной формулы средних прямоугольников.

Локальная погрешность любого из методов типа Рунге — Кутта второго порядка точности, как следует из

разложений (9), (10), может быть представлена в виде

$$r_1(h) = \frac{h^3}{6} [f_{xx} (1 - 3\alpha_1^2 A_1) + 2ff_{xy} (1 - 3\alpha_1 \beta_{10} A_1) + \\ + f^2 f_{yy} (1 - 3\beta_{10}^2 A_1) + f_y (f_x + ff_y)] + O(h^4). \quad (13)$$

Иногда свободный параметр A_1 выбирают так, чтобы в этом представлении можно было обратить в нуль хотя бы часть членов. Например, если учесть, что $\alpha_1 = \beta_{10} = \frac{1}{2A_1}$, и положить $A_1 = 3/4$, то правая часть равенства (13) существенно упростится:

$$r_1(h) = \frac{h^3}{6} f_y (f_x + ff_y) + O(h^4).$$

При таком выборе A_1 будем иметь

$$\Delta y \approx \frac{1}{4} (\varphi_0 + 3\varphi_1), \quad (14)$$

$$\varphi_0 = hf(x, y), \quad \varphi_1 = hf\left(x + \frac{2}{3}h, y + \frac{2}{3}\varphi_0\right).$$

Методы третьего порядка точности. С повышением требований к точности вычислительных правил типа Рунге — Кутта очень быстро возрастает громоздкость необходимых построений, хотя общая схема таких построений и не претерпевает существенных изменений. Поэтому в случае $q=2$ мы не станем здесь воспроизводить все выкладки и выпишем лишь ту систему уравнений, которым должны удовлетворять параметры (α) , (β) и (A) в методах типа Рунге — Кутта третьего порядка точности:

$$A_0 + A_1 + A_2 = 1, \quad A_1\alpha_1 + A_2\alpha_2 = \frac{1}{2}, \quad A_1\alpha_1^3 + A_2\alpha_2^3 = 1/3,$$

$$A_2\alpha_1\beta_{21} = 1/6, \quad \beta_{20} + \beta_{21} = \alpha_2, \quad \beta_{10} = \alpha_1.$$

Одно из решений этой системы шести уравнений с восемью неизвестными приводит к следующим формулам:

$$\Delta y \approx \frac{1}{6} (\varphi_0 + 4\varphi_1 + \varphi_2), \\ \varphi_0 = hf(x, y), \quad \varphi_1 = hf\left(x + \frac{h}{2}, y + \frac{\varphi_0}{2}\right), \quad (15) \\ \varphi_2 = hf(x + h, y - \varphi_0 + 2\varphi_1).$$

Это вычислительное правило является, очевидно, аналогом трехточечной квадратурной формулы Симпсона (5.3.5).

Часто встречается в практике вычислений и такой метод типа Рунге — Кутта третьего порядка точности:

$$\begin{aligned}\Delta y &\approx \frac{1}{4} (\varphi_0 + 3\varphi_2), \\ \varphi_0 &= hf(x, y), \quad \varphi_1 = hf\left(x + \frac{h}{3}, y + \frac{\varphi_0}{3}\right), \\ \varphi_2 &= hf\left(x + \frac{2}{3}h, y + \frac{2}{3}\varphi_1\right).\end{aligned}$$

По квадратурному смыслу приведенное вычислительное правило сходно с методом (14) второго порядка точности и еще раз косвенно подчеркивает достоинства последнего.

Методы четвертого порядка точности. В случае $q=3$ система ограничений на выбор параметров (α) , (β) и (A) вычислительных правил типа Рунге — Кутта четвертого порядка точности может быть приведена к виду

$$\begin{aligned}A_0 + A_1 + A_2 + A_3 &= 1, \quad A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3 = \frac{1}{2}, \\ A_1\alpha_1^2 + A_2\alpha_2^2 + A_3\alpha_3^2 &= \frac{1}{3}, \quad A_1\alpha_1^3 + A_2\alpha_2^3 + A_3\alpha_3^3 = \frac{1}{4}, \\ A_2\beta_{21}\alpha_1 + A_3\beta_{31}\alpha_1 + A_3\beta_{32}\alpha_2 &= \frac{1}{6}, \\ A_2\beta_{21}\alpha_2\alpha_1 + A_3\beta_{31}\alpha_3\alpha_1 + A_3\beta_{32}\alpha_3\alpha_2 &= \frac{1}{8}, \\ A_2\beta_{21}\alpha_1^2 + A_3\beta_{31}\alpha_1^2 + A_3\beta_{32}\alpha_2^2 &= \frac{1}{12}, \\ A_3\beta_{32}\beta_{21}\alpha_1 &= \frac{1}{24}, \quad \beta_{30} + \beta_{31} + \beta_{32} = \alpha_3, \\ \beta_{20} + \beta_{21} &= \alpha_2, \quad \beta_{10} = \alpha_1.\end{aligned}$$

Одним из методов, удовлетворяющих этим требованиям, является следующий аналог четырехточечной квадратурной формулы «трех восьмых» (5.3.10):

$$\begin{aligned}\Delta y &\approx \frac{1}{8} (\varphi_0 + 3\varphi_1 + 3\varphi_2 + \varphi_3), \\ \varphi_0 &= hf(x, y), \quad \varphi_1 = hf\left(x + \frac{h}{3}, y + \frac{\varphi_0}{3}\right), \\ \varphi_2 &= hf\left(x + \frac{2}{3}h, y - \frac{\varphi_0}{3} + \varphi_1\right), \\ \varphi_3 &= hf(x + h, y + \varphi_0 - \varphi_1 + \varphi_2).\end{aligned}$$

Особенно широко известно другое вычислительное правило типа Рунге — Кутта четвертого порядка точности:

$$\begin{aligned}\Delta y &\approx \frac{1}{6} (\varphi_0 + 2\varphi_1 + 2\varphi_2 + \varphi_3), \\ \varphi_0 &= hf(x, y), \quad \varphi_1 = hf\left(x + \frac{h}{2}, y + \frac{\varphi_0}{2}\right), \\ \varphi_2 &= hf\left(x + \frac{h}{2}, y + \frac{\varphi_1}{2}\right), \quad \varphi_3 = hf(x + h, y + \varphi_2).\end{aligned}\quad (16)$$

Так как справедливы приближенные равенства

$$\begin{aligned}2\varphi_1 + 2\varphi_2 &\approx 4hf\left[x + \frac{h}{2}, y\left(x + \frac{h}{2}\right)\right], \\ \varphi_3 &\approx hf[x + h, y(x + h)],\end{aligned}$$

то формулу (16), как и (15), тоже можно считать аналогом квадратурного правила Симпсона (5.3.5), хотя и не столь естественным, как (15). Однако если сопоставить порядки остаточных членов квадратурного правила Симпсона (см. (5.3.7)) и формул (15), (16), то можно сделать вывод, что вычислительное правило (16) является более точным аналогом формулы Симпсона. Это одношаговое правило наиболее часто служило основой стандартных программ решения задач с начальными условиями для уравнений вида (1.1) и систем таких уравнений.

Чтобы построить вычислительные правила типа Рунге — Кутта пятого и более высоких порядков точности, придется брать значения $q > 3$. В рассмотренных выше случаях увеличение q на единицу, как мы видели, позволяло повышать на порядок точность правила. Однако при $q = 4$ поднять порядок точности вычислительных правил по отношению к случаю $q = 3$ не удастся, так как параметров (α) , (β) и (A) при этом значении q еще не хватает для того, чтобы их выбором можно было для любой функции $f(x, y)$ обеспечить совпадение разложе-

ний Δy и $\sum_{i=0}^4 A_i \varphi_i$ по степеням h до членов с множителем h^5 включительно. В этом нетрудно убедиться непосредственно. Необходимые условия можно выполнить лишь при $q \geq 5$.

В качестве примера правила типа Рунге—Кутта пятого порядка точности можно привести следующий одношаговый метод:

$$\begin{aligned}\Delta y &\approx \frac{1}{192} (23\varphi_0 + 125\varphi_2 - 81\varphi_4 + 125\varphi_5), \\ \varphi_0 &= hf(x, y), \quad \varphi_1 = hf\left(x + \frac{h}{3}, y + \frac{\varphi_0}{3}\right), \\ \varphi_2 &= hf\left(x + \frac{2}{5}h, y + \frac{4\varphi_0 + 6\varphi_1}{25}\right), \\ \varphi_3 &= hf\left(x + h, y + \frac{\varphi_0 - 12\varphi_1 + 15\varphi_2}{4}\right), \\ \varphi_4 &= hf\left(x + \frac{2}{3}h, y + \frac{6\varphi_0 + 90\varphi_1 - 50\varphi_2 + 8\varphi_3}{81}\right), \\ \varphi_5 &= hf\left(x + \frac{4}{5}h, y + \frac{6\varphi_0 + 36\varphi_1 + 10\varphi_2 + 8\varphi_3}{75}\right).\end{aligned}$$

Дальнейшее повышение порядка точности методов типа Рунге—Кутта также связано, как правило, с быстрым возрастанием их трудоемкости, которую можно характеризовать, скажем, числом обращений на одном шаге процесса к вычислению значений правой части уравнения. Однако нужно учесть, что методы более высокого порядка точности обычно допускают использование большего шага h , что может снизить общие вычислительные затраты.

С примерами правил типа Рунге—Кутта при значениях $q > 5$ можно познакомиться в рекомендованной к данной главе литературе.

Одношаговые методы, построенные способом Рунге—Кутта, в отличие от вычислительных правил, полученных на основе разложения решения в ряд по последовательным главным частям, непосредственно не представляют простой возможности судить о локальной точности найденного значения приближенного решения по результатам промежуточных вычислений. Обычно в практике вычислений для этих целей используют следующий простой прием. Исходя из $y_n \approx y(x_n)$, приближение к $y(x_n + h)$ вычисляют по-разному: сначала данное вычислительное правило типа Рунге—Кутта применяют с избранным шагом h , а затем заново используют его дважды с шагом $h/2$. Сравнение так найденных двух приближен-

ных значений величины $y(x_{n+1})$ позволяет, во-первых, вынести суждение о локальной точности полученного приближения и, во-вторых, высказать мнение об удаче выбора шага интегрирования. Однако такая организация вычислений требует больших дополнительных затрат вычислительного труда: объем работы на один узел сетки возрастает почти втрое. Рассмотрим сейчас способ построения правил численного интегрирования дифференциальных уравнений, который позволяет во многих случаях полнее использовать результаты промежуточных вычислений и, в частности, получать одношаговые методы предсказывающе-исправляющего характера, что дает возможность практически без дополнительных вычислений судить о локальной погрешности полученного значения приближенного решения и об удаче выбора шага численного интегрирования. Заметим, что здесь и всюду, где речь идет о построении вычислительных правил, мы рассматриваем обычно приближенное решение без учета погрешностей, обусловленных ошибками округлений и неточным заданием исходных данных, тем самым понятие погрешности приближенного решения используется в более узком смысле погрешности метода.

§ 4. Построение вычислительных правил на основе принципа последовательного повышения порядка точности результата *)

1. Случай уравнений первого порядка. Для простоты изложения основной идеи предполагаемого способа построения методов численного решения дифференциальных уравнений будем сначала иметь в виду задачу Коши (1.1), (1.2). Как и в случае способа Рунге — Кутты, при построении вычислительных правил будем исходить из равенства типа (3.1), которое, подобно (3.3), для наших целей удобнее записать в виде

$$y(x_n + h) = y(x_n) + h \int_0^1 z_n(\alpha) d\alpha, \quad (1)$$

*) В. В. Бобков, Об одном способе построения одношаговых правил приближенного решения дифференциальных уравнений, Изв. АН БССР, сер. физ.-матем., № 4, 1967.

где $z_n(\alpha) = f[x_n + \alpha h, y(x_n + \alpha h)]$. Заменим интеграл квадратурной суммой

$$\sum_{i=0}^q A_i z_n(\alpha_i).$$

Тогда вместо (1) получим

$$\begin{aligned} y(x_n + h) &\approx y(x_n) + h \sum_{i=0}^q A_i z_n(\alpha_i) = \\ &= y(x_n) + h \sum_{i=0}^q A_i f[x_n + \alpha_i h, y(x_n + \alpha_i h)]. \end{aligned} \quad (2)$$

Выбор параметров $A_i, \alpha_i, i=0, 1, \dots, q$, в этом приближенном равенстве будем осуществлять, например, на основании требования, чтобы квадратурная формула

$$\int_0^1 z_n(\alpha) d\alpha \approx \sum_{i=0}^q A_i z_n(\alpha_i) \quad (3)$$

была точной для всевозможных алгебраических многочленов до степени $k-1$ ($0 < k \leq 2q+2$) включительно. Это приводит к следующей системе k уравнений с $2q+2$ неизвестными $A_i, \alpha_i, i=0, 1, \dots, q$:

$$\sum_{i=0}^q A_i = 1, \quad \sum_{i=0}^q A_i \alpha_i^j = \frac{1}{j+1}; \quad j=1, 2, \dots, k-1. \quad (4)$$

Заметим, что последняя система может быть получена и из требования, чтобы разложения по степеням h обеих частей приближенного равенства

$$y(x_n + h) - y(x_n) \approx h \sum_{i=0}^q A_i y'(x_n + \alpha_i h)$$

совпадали до членов с h^k включительно. Такой способ получения системы (4) позволяет, в частности, записать следующее представление для погрешности

$$r_{n+1} = y(x_n + h) - y(x_n) - h \sum_{i=0}^q A_i f[x_n + \alpha_i h, y(x_n + \alpha_i h)]$$

приближенного равенства (2):

$$r_{n+1} = h^{k+1} y^{(k+1)}(x_n) \left[\frac{1}{(k+1)!} - \frac{1}{k!} \sum_{i=0}^q A_i \alpha_i^k \right] + O(h^{k+2}). \quad (5)$$

Так как весовая функция в случае интеграла $\int_0^1 z_n(\alpha) d\alpha$ равна единице, то квадратурная формула вида (3), имеющая наибольшую алгебраическую степень точности $2q+1$, может быть построена и притом единственным образом для любого $q \geq 0$. Поэтому при $k=2q+2$ система (4) имеет единственное решение, при этом $0 < A_i \leq 1$, $0 < \alpha_i < 1$, $i=0, 1, \dots, q$ (см. пп. 5.4.1, 5.4.2). Следовательно, при $1 \leq k \leq 2q+2$ у этой системы существует хотя бы одно решение, и приближенное равенство (2) может быть построено.

Если бы в (2) все значения $y(x_n + \alpha_i h)$, $i=0, 1, \dots, q$, были известны точно, то это приближенное равенство позволяло бы найти искомое значение $y(x_n + h)$ соответствующего решения уравнения (1.1) по известному значению $y(x_n)$ этого решения с ошибкой порядка h^{k+1} (см. (5)). Хотя точными значениями величин $y(x_n + \alpha_i h)$, $i=0, 1, \dots, q$, мы не располагаем, но, подобно (2), заменив там h на $\alpha_i h$, нетрудно указать правила их приближенного вычисления через значения $y(x_n + \alpha_i \beta_j h)$, для нахождения которых, в свою очередь, можно построить подобные же рекурсивные формулы. При этом следует иметь в виду, что наличие множителя h перед суммой в (2) позволяет находить значения $y(x_n + \alpha_i h)$ с локальной ошибкой порядка h^k , а значения $y(x_n + \alpha_i \beta_j h)$ — с ошибкой порядка h^{k-1} и т. д. Параметры соответствующих приближенных равенств должны удовлетворять системе уравнений типа (4), в которой с понижением требований к точности на порядок следует отбрасывать последнее уравнение. При этом часто бывает целесообразным уменьшать и число q , определяющее количество подлежащих выбору параметров. Следуя такой схеме построений, мы придем, наконец, к приближенным равенствам

$$y(x_n + \alpha_i \beta_j \dots \gamma_k h) \approx y(x_n) + \alpha_i \beta_j \dots \gamma_k h f(x_n, y_n),$$

на которых процесс замыкается. Погрешность таких равенств будет, очевидно, величиной порядка h^2 . Они

получаются из равенств типа (2) в случае, когда квадратурная формула (3) является простейшей формулой левых прямоугольников.

Проиллюстрируем предлагаемый способ построения вычислительных правил на нескольких простых примерах.

Методы первого порядка точности. Система (4) в этом случае вырождается в единственное требование

$$\sum_{i=0}^q A_i = 1. \quad (6)$$

Параметры α_i , $i=0, 1, \dots, q$, могут принимать при этом, вообще говоря, любые фиксированные значения. Однако для случая одношаговых методов выбор этих параметров должен быть подчинен ограничению

$$0 \leq \alpha_i \leq 1.$$

Взяв, например, $q=0$, найдем, что $A_0=1$. Полагая $\alpha_0=0$, получим исходную формулу известного метода Эйлера (2.2)

$$y(x_{n+1}) = y(x_n) + hf[x_n, y(x_n)] + O(h^2).$$

При $\alpha_0=1$ получим простейший неявный метод

$$y_{n+1}^{[2]} = y_n^{[2]} + hf_{n+1}^{[2]}.$$

При записи последнего вычислительного правила использованы обозначения типа (2.3).

Методы второго порядка точности. В этом случае требование (6) нужно дополнить условием

$$\sum_{i=0}^q A_i \alpha_i = \frac{1}{2}. \quad (7)$$

При $q=0$ система (6), (7) имеет единственное решение $A_0=1$, $\alpha_0=1/2$, что приводит к следующему вычислительному правилу:

$$\begin{aligned} y_{n+1/2}^{[2]} &= y_n^{[3]} + \frac{h}{2} f_n^{[3]}, \\ y_{n+1}^{[3]} &= y_n^{[3]} + hf_{n+1/2}^{[2]}. \end{aligned} \quad (8)$$

При $q=1$ система (6), (7) принимает вид

$$A_0 + A_1 = 1, \quad A_0 \alpha_0 + A_1 \alpha_1 = \frac{1}{2}.$$

Выбрав, скажем, $\alpha_0 = 0$, $\alpha_1 = 1$, найдем, что $A_0 = A_1 = 1/2$, и получим неявный метод *трапеций* (3.2)

$$y_{n+1}^{[3]} = y_n^{[3]} + \frac{h}{2} (f_n^{[3]} + f_{n+1}^{[3]}).$$

Используя формулу Эйлера (2.2), это вычислительное правило можно преобразовать в известное уже нам явное правило (2.7), (2.8).

Методы третьего порядка точности. Требования (6), (7) здесь необходимо пополнить уравнением

$$\sum_{i=0}^q A_i \alpha_i^2 = \frac{1}{3}.$$

Этим условиям удовлетворяют, например, вычислительные правила

$$y_{n+1/4}^{[2]} = y_n^{[4]} + \frac{h}{4} f_n^{[i]}, \quad (9)$$

$$y_{n+1/2}^{[3]} = y_n^{[4]} + \frac{h}{2} f_{n+1/4}^{[2]}, \quad (10)$$

$$y_{n+1}^{[3]} = y_n^{[4]} + h f_{n+1/2}^{[3]}, \quad (11)$$

$$y_{n+1}^{[4]} = y_n^{[4]} + \frac{h}{6} (f_n^{[i]} + 4f_{n+1/2}^{[3]} + f_{n+1}^{[3]}), \quad (12)$$

где i может принимать значения 3 или 4. При $i=4$ построенное правило на один узел сетки (1.5) требует четырехкратного обращения к блоку нахождения значений правой части уравнения (1.1). В случае же $i=3$ точность результата, вообще говоря, несколько понижается (при сохранении порядка), однако в основном счете число обращений к блоку вычисления значений функции $f(x, y)$ сокращается до трех на узел сетки.

Заметим, что построенные численные методы, как и вычислительное правило (2.7), (2.8), имеют предсказывающе-исправляющий характер. Приближенное значение величины $y(x_{n+1})$, найденное по формуле (11) с локаль-

ной погрешностью порядка h^3 , уточняется затем по формуле (12). Сравнение значений $y_{n+1}^{[3]}$ и $y_{n+1}^{[4]}$ дает практическую возможность по ходу вычислений без дополнительных затрат составить представление о локальной точности полученного приближения к $y(x_{n+1})$. Такое сравнение, в частности, может быть положено в основу правила автоматического выбора шага интегрирования.

На примере приведенных вычислительных правил легко видеть, что предлагаемый выше способ построения методов численного интегрирования дифференциальных уравнений, как и способ, изложенный в п. 2.2, удовлетворяет принципу модульности, когда сложные вычислительные алгоритмы komponуются на основе более простых типовых расчетных формул. Так, например, в случае последнего вычислительного правила формула (9) записана на основе метода Эйлера (2.2), для равенств (10), (11) базовой является формула средних прямоугольников типа (8), а равенство (12) непосредственно связано с формулой Симпсона (5.3.5).

Методы четвертого порядка точности. Система требований к построению таких методов получается из (4) при $k=4$. В качестве примера приведем здесь лишь одно семейство вычислительных правил четвертого порядка точности, имеющих предсказывающе-исправляющий характер:

$$\begin{aligned}
 y_{n+1/m}^{[2]} &= y_n^{[5]} + \frac{h}{m} f_n^{[i]}, \\
 y_{n+1/3}^{[3]} &= y_n^{[5]} + \frac{h}{3} \left(\frac{6-m}{6} f_n^{[i]} + \frac{m}{6} f_{n+1/m}^{[2]} \right), \\
 y_{n+1/2}^{[4]} &= y_n^{[5]} + \frac{h}{8} \left(f_n^{[i]} + 3f_{n+1/3}^{[3]} \right), \\
 y_{n+1}^{[4]} &= y_n^{[5]} + \frac{h}{2} \left(f_n^{[i]} - 3f_{n+1/3}^{[3]} + 4f_{n+1/2}^{[4]} \right), \\
 y_{n+1}^{[5]} &= y_n^{[5]} + \frac{h}{6} \left(f_n^{[i]} + 4f_{n+1/2}^{[4]} + f_{n+1}^{[4]} \right).
 \end{aligned} \tag{13}$$

Здесь $m > 0$, а i может принимать значения 4 или 5. При $i=4$, в частности, соответствующий метод сопряжен

в основном счете лишь с четырехкратным обращением к блоку вычисления значений правой части уравнения (1.1) на один узел сетки (1.5). Следует отметить также, что значение $m=6$ является наибольшим, при котором

в правой части равенства для $y_{n+1/3}$ нет отрицательных коэффициентов, а при $m=9/2$ обращается в нуль одно из слагаемых с h^3 в остаточном члене этой формулы.

Подобным же образом могут быть построены численные методы и более высокого порядка точности, при этом увеличения точности вычислительных правил можно достигать как за счет привлечения новых значений функции y' , так и за счет использования значений последующих производных.

Все сказанное выше применительно к задаче (1.1), (1.2) без существенных затруднений может быть перенесено на случай задачи Коши для системы обыкновенных дифференциальных уравнений первого порядка. Это, в частности, дает возможность использовать построенные методы при решении уравнений высших порядков, сводя предварительно последние к системам уравнений первого порядка. Вычислительные правила можно строить и специально для случая уравнений высших порядков, при этом часто удастся получить более приемлемые алгоритмы.

2. Случай уравнений высших порядков. Изложенные выше способы построения вычислительных методов решения задачи Коши (1.1), (1.2) легко могут быть обобщены на случай уравнений высших порядков. Проведем такое обобщение только для последнего способа. Для простоты рассмотрим задачу Коши для уравнения второго порядка:

$$y'' = f(x, y, y'), \quad (14)$$

$$y(x_0) = y_0, \quad y'(x_0) = y'_0. \quad (15)$$

Подобно (1), (2), здесь можно записать, что

$$y'(x_n + h) = y'(x_n) + h \int_0^1 y''(x_n + \alpha h) d\alpha \approx$$

$$\approx y'(x_n) + h \sum_{i=0}^{q_1} A_i f[x_n + \alpha_i h, y(x_n + \alpha_i h), y'(x_n + \alpha_i h)]. \quad (16)$$

И аналогично

$$\begin{aligned}
 y(x_n + h) &= y(x_n) + hy'(x_n) + \\
 &+ h^2 \int_0^1 (1 - \beta) y''(x_n + \beta h) d\beta \approx y(x_n) + hy'(x_n) + \\
 &+ h^2 \sum_{i=0}^{q_0} B_i f[x_n + \beta_i h, y(x_n + \beta_i h), y'(x_n + \beta_i h)]. \quad (17)
 \end{aligned}$$

Подберем параметры A_i , α_i , $i=0, 1, \dots, q_1$, и B_i , β_i , $i=0, 1, \dots, q_0$, этих приближенных равенств так, чтобы их погрешности были величинами порядка h^{k+1} ($1 < k \leq \leq \min \{2q_1 + 2, 2q_0 + 3\}$). Эти требования, подобно изложенному в предыдущем пункте, приводят к следующим системам, вообще говоря, нелинейных уравнений относительно искомых значений параметров:

$$\sum_{i=0}^{q_1} A_i = 1, \quad \sum_{i=0}^{q_1} A_i \alpha_i^j = \frac{1}{j+1}, \quad j=1, 2, \dots, k-1, \quad (18)$$

$$\sum_{i=0}^{q_0} B_i = \frac{1}{2}, \quad \sum_{i=0}^{q_0} B_i \beta_i^{j-1} = \frac{1}{j(j+1)}, \quad j=2, 3, \dots, k-1. \quad (19)$$

Разрешимость системы (18) была изучена ранее (см. (4)). Подобным же образом выясняется этот вопрос и для системы (19).

Подобрав, согласно указанным требованиям, подходящие значения параметров, мы построим приближенные формулы (16), (17), погрешности которых, подобно (5), могут быть записаны соответственно в виде

$$r_{n+1}^{(1)} = h^{k+1} y^{(k+2)}(x_n) \left[\frac{1}{(k+1)!} - \frac{1}{k!} \sum_{i=0}^{q_1} A_i \alpha_i^k \right] + O(h^{k+2}),$$

$$r_{n+1}^{(0)} = h^{k+1} y^{(k+1)}(x_n) \left[\frac{1}{(k+1)!} - \frac{1}{(k-1)!} \sum_{i=0}^{q_0} B_i \beta_i^{k-1} \right] + O(h^{k+2}).$$

Процесс дальнейшего построения вычислительных правил для задачи (14), (15) совершенно аналогичен случаю задачи (1.1), (1.2). При этом полезно отметить, что наличие множителя h^2 перед суммой в (17) позволяет находить значения $y(x_n + \beta_i h)$, $y'(x_n + \beta_i h)$, $i=0, 1, \dots, q_0$,

по отношению к требуемой точности значения $y(x_n + h)$ грубее не только на порядок, как это было в случае приближенного равенства (2), а даже на два порядка. Это позволяет, в частности, быстрее уменьшать число точек, привлекаемых при построении промежуточных формул, и упрощает вид вычислительных правил.

В качестве примера для задачи (14), (15) приведем здесь лишь одно семейство вычислительных правил четвертого порядка точности, тесно связанных с вычислительными правилами из семейства (13):

$$\begin{aligned}
 {}^{[2]}y'_{n+1/m} &= {}^{[5]}y'_n + \frac{h}{m} {}^{[5,i]}f_n, \\
 {}^{[3]}y_{n+1/m} &= y_n + \frac{h}{m} y'_n + \frac{h^2}{2m^2} {}^{[5,i]}f_n, \\
 {}^{[3]}y'_{n+1/3} &= {}^{[5]}y'_n + \frac{h}{3} \left(\frac{6-m}{6} {}^{[5,i]}f_n + \frac{m}{6} {}^{[3,2]}f_{n+1/m} \right), \\
 {}^{[4]}y_{n+1/3} &= y_n + \frac{h}{3} y'_n + \frac{h^2}{9} \left(\frac{9-m}{18} {}^{[5,i]}f_n + \frac{m}{18} {}^{[3,2]}f_{n+1/m} \right), \\
 {}^{[4]}y'_{n+1/2} &= {}^{[5]}y'_n + \frac{h}{8} \left({}^{[5,i]}f_n + 3 {}^{[4,3]}f_{n+1/3} \right), \\
 {}^{[4]}y_{n+1/2} &= y_n + \frac{h}{2} y'_n + \frac{h^2}{16} \left({}^{[5,i]}f_n + {}^{[4,3]}f_{n+1/3} \right), \\
 {}^{[4]}y'_{n+1} &= {}^{[5]}y'_n + \frac{h}{2} \left({}^{[5,i]}f_n - 3 {}^{[4,3]}f_{n+1/3} + 4 {}^{[4,4]}f_{n+1/2} \right), \\
 {}^{[5]}y_{n+1} &= y_n + h y'_n + \frac{h^2}{6} \left({}^{[5,i]}f_n + 2 {}^{[4,4]}f_{n+1/2} \right), \\
 {}^{[5]}y'_{n+1} &= {}^{[5]}y'_n + \frac{h}{6} \left({}^{[5,i]}f_n + 4 {}^{[4,4]}f_{n+1/2} + {}^{[5,4]}f_{n+1} \right).
 \end{aligned} \tag{20}$$

Как и в случае вычислительных правил (13), здесь $m > 0$, а i может принимать значения 4 или 5. Смысл обозначений также аналогичен используемым в случае формул (13):

$${}^{[i,j]}f_{n+\alpha} = f(x_n + \alpha h, y_{n+\alpha}, y'_{n+\alpha}),$$

где ${}^{[i]}y_{n+\alpha}$ и ${}^{[j]}y'_{n+\alpha}$ есть приближенные значения величин $y(x_n + \alpha h)$ и $y'(x_n + \alpha h)$, найденные с локальными ошибками порядка h^i и h^j соответственно.

К приведенному набору формул (20) можно присоединить формулу

$$y_{n+1}^{[4]} = y_n^{[5]} + h y_n'^{[5]} + \frac{h^2}{2} f_{n+1/3}^{[4, 3]}$$

или формулу

$$y_{n+1}^{[5]*} = y_n^{[5]} + h y_n'^{[5]} + \frac{h^2}{24} \left(2 f_n^{[5, i]} + 9 f_{n+1, 3}^{[4, 3]} + f_{n+1}^{[5, j]} \right),$$

где j , как и i , также может принимать значения 4 или 5. Это дает возможность (при $j=4$ — без дополнительных обращений к блоку вычисления значений правой части уравнения (14)) сравнить между собой не только приближенные значения $y_{n+1}^{[4]}$, $y_{n+1}^{[5]}$ величины $y'(x_n + h)$, но и приближенные значения $y_{n+1}^{[4]}$, $y_{n+1}^{[5]}$, $y_{n+1}^{[5]*}$ величины $y(x_n + h)$, что позволит высказать более достоверные суждения о локальной точности полученных результатов и удаче выбора шага h .

При необходимости, присоединив к (20) формулу

$$y_{n+\gamma}^{[5]} = y_n^{[5]} + h \gamma y_n'^{[5]} + \frac{h^2 \gamma^2}{6} \left\{ [3 - \gamma(3 - \gamma)] f_n^{[5, i]} + 2\gamma(2 - \gamma) f_{n+1/2}^{[4, 4]} + \gamma(\gamma - 1) f_{n+1}^{[5, j]} \right\}, \quad (21)$$

можно с локальной ошибкой порядка h^5 найти значение решения и в любой внутренней точке $x_n + \gamma h$ ($0 < \gamma < 1$) отрезка $[x_n, x_n + h]$ (в случае $j=4$ — без дополнительных вычислений значений функции f).

В заключение заметим, что если бы мы предварительно задачу (14), (15) свели к соответствующей задаче для системы двух уравнений первого порядка и стали решать ее численно методом типа (13), то мы использовали бы приближенные значения функции $y(x)$ в точках $x_n + h/m$ и $x_n + h/3$ соответственно с локальной ошибкой порядка h^2 и h^3 , в то время как в случае вычислительного правила (20), построенного специально для уравнений второго порядка, в указанных точках используются приближенные значения функции $y(x)$, точность которых на порядок выше. В этом направлении можно, очевидно, достичь и большего. Например, если к равенствам (20), предвари-

тельно заменив в их правых частях слагаемое y_n на $y_n^{[6]}$, присоединить формулу

$$y_{n+1}^{[6]} = y_n^{[6]} + h y_n^{[5]} + \frac{h^2}{72} \left(9 f_n^{[5, i]} + 25 f_{n+2/5}^{[5, 4]} + 2 f_{n+1}^{[5, 4]} \right),$$

где $y_{n+2/5}^{[5]}$ находится из (21) при $\gamma = 2/5$, $j = 4$, а $y_{n+2/5}^{[4]}$ вычисляется по правилу

$$y_{n+2/5}^{[4]} = y_n^{[5]} + \frac{2h}{375} \left(38 f_n^{[5, i]} + 44 f_{n+1/2}^{[4, 4]} - 7 f_{n+1}^{[5, 4]} \right),$$

то получим возможность ценой вычисления одного дополнительного значения функции f повысить на порядок локальную точность нахождения $y(x_n + h)$ (при сохранении порядка скорости сходимости метода). В случае уравнений высших порядков появляются еще более широкие возможности для построения подобных вариантов вычислительных правил.

§ 5. Сходимость одношаговых методов

Для простоты записи и бóльшей наглядности снова будем иметь в виду задачу (1.1), (1.2). Как и раньше, основные результаты этого параграфа могут быть перенесены на случай систем уравнений первого порядка и обобщены на уравнения высших порядков.

Будем считать, что уравнение (1.1) на отрезке $[x_0, X]$ имеет единственное решение не только при начальных данных (1.2), но и в случае любых начальных условий вида

$$y(\xi) = \eta,$$

где

$$x_0 \leq \xi \leq X,$$

$$y[\xi, X, y(X, x_0, y_0) - \varepsilon] \leq \eta \leq y[\xi, X, y(X, x_0, y_0) + \varepsilon],$$

$$\varepsilon > 0,$$

а через $y(x, \xi, \eta)$ обозначено значение в точке x решения $y(x)$ уравнения (1.1) при начальном условии $y(\xi) = \eta$. Величина ε при этом может быть истолкована, например, как верхняя граница допустимой абсолютной погрешности в нахождении интересующего нас значения $y(X, x_0, y_0)$.

Предположим также, что при любых значениях аргументов из указанной выше области их изменения решение $y(x, \xi, \eta)$ непрерывно дифференцируемо по начальному данному η . Для такой производной, как известно из теории обыкновенных дифференциальных уравнений, справедливо равенство

$$\frac{\partial y(x, \xi, \eta)}{\partial \eta} = \exp \left\{ \int_{\xi}^x f_y[t, y(t, \xi, \eta)] dt \right\}.$$

Это равенство может быть просто получено путем интегрирования однородного линейного дифференциального уравнения первого порядка

$$\frac{\partial}{\partial t} \frac{\partial y(t, \xi, \eta)}{\partial \eta} = f_y[t, y(t, \xi, \eta)] \frac{\partial y(t, \xi, \eta)}{\partial \eta}$$

относительно неизвестной функции $z(t) = \frac{\partial y(t, \xi, \eta)}{\partial \eta}$ с начальным значением $z(\xi) = \frac{\partial y(\xi, \xi, \eta)}{\partial \eta} = 1$. При условии непрерывной дифференцируемости функции $f(t, y)$ по y , достаточном для существования непрерывной производной $\frac{\partial y(t, \xi, \eta)}{\partial \eta}$, последнее уравнение может быть получено непосредственно дифференцированием по η очевидного равенства

$$\frac{\partial}{\partial t} y(t, \xi, \eta) = f[t, y(t, \xi, \eta)]$$

с последующей перестановкой порядка дифференцирования.

Пусть, скажем, константа $M > 0$ ограничивает сверху значения производной $\frac{\partial y(x, \xi, \eta)}{\partial \eta}$ для любых значений переменных x, ξ, η из указанной выше области их изменения при естественном дополнительном условии $\xi \leq x$. Будем предполагать также, что решение $y(x, \xi, \eta)$ имеет непрерывные производные по x достаточно высокого порядка, определяемого точностью избранного для исследования одношагового метода вида

$$y_{n+1} = F(y_n), \quad n = 0, 1, \dots, N-1, \quad (1)$$

где заданная функция F зависит, естественно, от правой части уравнения (1.1), шага h , который сейчас будем считать постоянным, и точки x_n отрезка $[x_0, X]$ ($x_i = x_0 + ih$,

$i = 0, 1, \dots, N$, $Nh = X - x_0$). Через y_i ($i = 1, 2, \dots, N$) в (1) обозначено приближенное значение величины $y(x_i, x_0, y_0)$, получаемое рассматриваемым методом.

В действительности вследствие наличия ошибок округления вычисления по формуле (1) выполняются, как правило, неточно. Фактически найденные по этой формуле значения $\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_N$ не будут точно удовлетворять (1), а будут связаны соотношениями вида

$$\tilde{y}_{n+1} = F(\tilde{y}_n) - \delta_{n+1}, \quad n = 0, 1, \dots, N-1, \quad (2)$$

где через $\tilde{y}_0$ для единообразия записи обозначена величина реально участвующего в вычислениях начального значения $y_0 = y(x_0)$.

Разность $\tilde{\varepsilon}_0 = y_0 - \tilde{y}_0$ будем называть *погрешностью начального условия*, а величину $-\delta_{n+1}$, определяемую формулой (2), — *погрешностью округления* на $(n+1)$ -м шаге вычислительного процесса.

Пользователя интересует непосредственно величина *погрешности приближенного решения* в точке x_n ($n = 1, 2, \dots, N$), задаваемая разностью

$$\tilde{\varepsilon}_n = y(x_n, x_0, y_0) - \tilde{y}_n = y(x_n, x_0, y_0) - y(x_n, x_n, \tilde{y}_n).$$

При условии отсутствия ошибок начальных данных и округлений эту погрешность называют обычно *погрешностью метода*. Последняя задается разностью

$$\varepsilon_n = y(x_n, x_0, y_0) - y_n = y(x_n, x_0, y_0) - y(x_n, x_n, y_n)$$

и для данной задачи зависит лишь от точности избранного метода. В действительности же погрешность приближенного решения зависит, очевидно, от погрешности начального условия, ошибок округлений и точности самой расчетной формулы. *Погрешность формулы* (1) на каждом шаге реального вычислительного процесса можно задать, скажем, посредством равенства

$$y(x_{n+1}, x_n, \tilde{y}_n) = F[y(x_n, x_n, \tilde{y}_n)] + r_{n+1} = F(\tilde{y}_n) + r_{n+1}. \quad (3)$$

Величина r_{n+1} этой погрешности характеризует, насколько хорошо точное решение дифференциального уравнения (1.1), на которое мы вышли на предыдущем n -м шаге рассматриваемого вычислительного процесса, удовлетворяет на данном этапе вычислений аппроксимирующему разностному уравнению (1). Она отражает точность замены

на отрезке $[x_n, x_n + h]$ дифференциального уравнения (1.1) разностным уравнением (1) и представляет собой локальную погрешность избранного метода или ошибку приближенного решения на одном шаге вычислений без учета погрешности округления на данном шаге. С учетом же погрешности округления локальная ошибка приближенного решения может быть задана формулой

$$y(x_{n+1}, x_n, \tilde{y}_n) - \tilde{y}_{n+1} = r_{n+1} + \delta_{n+1}. \quad (4)$$

Эту формулу можно получить путем почленного вычитания равенства (2) из (3).

Обычно, как мы видим на примерах вычислительных правил, рассмотренных в предыдущих параграфах, при построении одношаговых методов добиваются того, чтобы погрешность формулы была малой величиной некоторого порядка $k+1$ ($k > 0$) в сравнении с шагом h . Что касается погрешности округлений, то любая ее малость, как и малость погрешности начального условия, принципиально говоря, всегда может быть обеспечена. В реальных же условиях уменьшения абсолютного значения погрешности округлений обычно достигают повышением значности счета и уровня рациональности его организации. Практические возможности такого уменьшения ограничены.

Если бы мы располагали эффективными оценками погрешностей формулы и округлений, то на основании равенства (4) можно было бы дать оценку и локальной погрешности приближенного решения на каждом шаге вычислительного процесса, что позволило бы оценить, насколько сильно на данном этапе вычислений мы можем отклониться от того решения уравнения (1.1), которое получено на предыдущем этапе. Нас же больше интересует оценка итогового отклонения от искомого решения. Такое отклонение не равно, вообще говоря, сумме локальных ошибок, допущенных на каждом из выполненных этапов процесса. Оно существенно зависит и от характера поведения интегральных кривых уравнения (1.1). Предположив, например, что мы допустили ошибку лишь на первом этапе процесса, а все остальные шаги его не уводили нас с того решения, на которое вышли вначале, мы будем иметь принципиально разную качественную оценку итогового отклонения в сравнении с начальным в зависимости от того, сближаются или расходятся интегральные кривые

уравнения (1.1) при изменении x в направлении численного интегрирования. Оценка итогового отклонения и будет нашей ближайшей задачей.

Отклонение от искомого решения мы допускаем, вообще говоря, как на каждом этапе вычислений по формуле (1), так и в случае неточного задания начального условия. Организуя вычислительный процесс при оговоренных ранее условиях, мы должны прежде всего позаботиться о том, чтобы ни одно из этих отклонений не вывело нас за пределы области G_ε , ограниченной прямыми $x = x_0$, $x = X$ и кривыми $y = y[x, X, y(X, x_0, y_0) - \varepsilon]$, $y = y[x, X, y(X, x_0, y_0) + \varepsilon]$. Для данной задачи при избранном методе типа (1) это требование можно обеспечить лишь за счет выбора величины шага интегрирования и уменьшения погрешности начального условия и округлений. Сформулировать условия, достаточные для обеспечения указанного требования, можно на основании разных принципов. Например, в рассматриваемом случае постоянного шага h можно поставить условие, чтобы допустимое итоговое отклонение от точного решения было в каком-то смысле равномерно распределено по всем этапам вычислений. Можно потребовать, скажем, чтобы влияние каждого из $N + 1$ возможных отклонений на итоговую ошибку в интересующей нас точке X не превышало по абсолютному значению величины $\frac{\varepsilon}{N+1}$. Тогда абсолютная величина погрешности приближенного решения в точке X заведомо не превысит ε , что и будет свидетельством выполнения исходного требования.

Первое отклонение от искомого решения может быть допущено за счет погрешности начального условия. На основе сформулированного выше принципа для этого отклонения должно быть выдвинуто требование

$$|y(X, x_0, y_0) - y(X, x_0, \tilde{y}_0)| \leq \frac{\varepsilon}{N+1}.$$

Если воспользоваться теоремой Лагранжа о конечном приращении, то левую часть последнего неравенства можно представить в виде

$$|y(X, x_0, y_0) - y(X, x_0, \tilde{y}_0)| = \left| (y_0 - \tilde{y}_0) \frac{\partial y(X, x_0, \eta_0)}{\partial \eta} \right|,$$

где

$$\tilde{\eta}_0 = \tilde{y}_0 + \theta_0 (y_0 - \tilde{y}_0), \quad 0 < \theta_0 < 1.$$

В силу этого указанное требование будет заведомо выполняться, если

$$\varepsilon_0 M \leq \frac{\varepsilon}{N+1},$$

или

$$\varepsilon_0 \leq \frac{\varepsilon h}{M(X - x_0 + h)}, \quad (5)$$

где величина ε_0 удовлетворяет неравенству

$$|\tilde{\varepsilon}_0| \leq \varepsilon_0. \quad (6)$$

При заданном шаге h , ограничения на выбор которого мы сформулируем ниже, условие (5), принципиально говоря, всегда может быть выполнено за счет уменьшения абсолютной величины погрешности начального значения.

Дальнейшие отклонения от искомого решения связаны с локальными ошибками приближенного решения, которые задаются равенствами вида (4). В силу сформулированного принципа при этом должно быть обеспечено выполнение требований

$$|y(X, x_n, \tilde{y}_n) - y(X, x_{n+1}, \tilde{y}_{n+1})| \leq \frac{\varepsilon}{N+1}, \quad n=0, 1, \dots, N-1. \quad (7)$$

Но

$$\begin{aligned} y(X, x_n, \tilde{y}_n) - y(X, x_{n+1}, \tilde{y}_{n+1}) &= \\ &= y[X, x_{n+1}, y(x_{n+1}, x_n, \tilde{y}_n)] - y(X, x_{n+1}, \tilde{y}_{n+1}) = \\ &= [y(x_{n+1}, x_n, \tilde{y}_n) - \tilde{y}_{n+1}] \frac{\partial y(X, x_{n+1}, \tilde{\eta}_{n+1})}{\partial \eta}, \end{aligned}$$

где

$$\tilde{\eta}_{n+1} = \tilde{y}_{n+1} + \theta_{n+1} [y(x_{n+1}, x_n, \tilde{y}_n) - \tilde{y}_{n+1}], \quad 0 < \theta_{n+1} < 1.$$

Поэтому требования (7) с учетом (4) принимают вид

$$\left| (r_{n+1} + \delta_{n+1}) \frac{\partial y(X, x_{n+1}, \tilde{\eta}_{n+1})}{\partial \eta} \right| \leq \frac{\varepsilon}{N+1}, \quad n=0, 1, \dots, N-1,$$

и будут заведомо выполняться, если

$$(r + \delta) M \leq \frac{\varepsilon}{N+1}$$

или

$$\frac{r+\delta}{h} \leq \frac{\varepsilon}{M(X-x_0+h)}, \quad (8)$$

где r и δ удовлетворяют неравенствам

$$|r_i| \leq r, \quad |\delta_i| \leq \delta, \quad i = 1, 2, \dots, N. \quad (9)$$

Условие (8) накладывает ограничения и на погрешность формулы и на погрешность округлений. Так как обычно $r = O(h^{k+1})$, где $k > 0$, то выбором достаточно малого шага $h > 0$ всегда можно гарантировать необходимую малость величины r . При фиксированном же h соответствующая малость величины δ , принципиально говоря, также всегда может быть обеспечена.

Будем считать выполненными указанные выше требования и получим интересующую нас оценку погрешности приближенного решения, для чего дадим сначала выражение этой ошибки через погрешность начального условия, погрешности формулы и округлений.

Ошибка

$$\tilde{\varepsilon}_n = y(x_n, x_0, y_0) - \tilde{y}_n = y(x_n, x_0, y_0) - y(x_n, x_n, \tilde{y}_n)$$

очевидным образом может быть записана в форме

$$\begin{aligned} \tilde{\varepsilon}_n = & y(x_n, x_0, y_0) - y(x_n, x_0, \tilde{y}_0) + \\ & + \sum_{i=1}^n [y(x_n, x_{i-1}, \tilde{y}_{i-1}) - y(x_n, x_i, \tilde{y}_i)]. \end{aligned} \quad (10)$$

Так как

$$y(x_n, x_{i-1}, \tilde{y}_{i-1}) = y[x_n, x_i, y(x_i, x_{i-1}, \tilde{y}_{i-1})],$$

то равенству (10) можно придать следующий вид:

$$\begin{aligned} \tilde{\varepsilon}_n = & y(x_n, x_0, y_0) - y(x_n, x_0, \tilde{y}_0) + \\ & + \sum_{i=1}^n \{y[x_n, x_i, y(x_i, x_{i-1}, \tilde{y}_{i-1})] - y(x_n, x_i, \tilde{y}_i)\}, \end{aligned}$$

откуда на основании теоремы Лагранжа о конечном приращении получаем

$$\begin{aligned} \tilde{\varepsilon}_n = & (y_0 - \tilde{y}_0) \frac{\partial y(x_n, x_0, \tilde{\eta}_0)}{\partial \eta} + \\ & + \sum_{i=1}^n [y(x_i, x_{i-1}, \tilde{y}_{i-1}) - \tilde{y}_i] \frac{\partial y(x_n, x_i, \tilde{\eta}_i)}{\partial \eta}, \end{aligned}$$

где, как и выше,

$$\tilde{\eta}_0 = \tilde{y}_0 + \theta_0 (y_0 - \tilde{y}_0), \quad 0 < \theta_0 < 1,$$

$$\tilde{\eta}_i = \tilde{y}_i + \theta_i [y(x_i, x_{i-1}, \tilde{y}_{i-1}) - \tilde{y}_i], \quad 0 < \theta_i < 1.$$

Поэтому, учитывая (4) и определение погрешности начального условия, можно записать

$$\tilde{\varepsilon}_n = \tilde{\varepsilon}_0 \frac{\partial y(x_n, x_0, \tilde{\eta}_0)}{\partial \eta} + \sum_{i=1}^n (r_i + \delta_i) \frac{\partial y(x_n, x_i, \tilde{\eta}_i)}{\partial \eta}, \quad (11)$$

$$n = 1, 2, \dots, N.$$

Формула (11) дает искомое представление погрешности приближенного решения в точке x_n через погрешность начального условия, погрешности формулы и округлений на всех выполненных этапах вычислительного процесса. Из этой формулы хорошо видна также зависимость характера накопления ошибок при счете на большое число шагов от индивидуальных свойств решаемого дифференциального уравнения. Такая зависимость может быть наглядно пояснена, скажем, с помощью приводимого в начале этого параграфа равенства

$$\frac{\partial y(x, \xi, \eta)}{\partial \eta} = \exp \left\{ \int_{\xi}^x f_y[t, y(t, \xi, \eta)] dt \right\}.$$

В случае, когда при $\xi < x$, например,

$$\frac{\partial y(x, \xi, \eta)}{\partial \eta} > 1 \quad \left(\frac{\partial f(x, y)}{\partial y} > 0 \right)$$

и интегральные кривые уравнения (1.1) при изменении x в направлении численного интегрирования расходятся, влияние погрешности начального условия и каждой из локальных ошибок приближенного решения, допущенных на предыдущих шагах вычислительного процесса, возрастает. Если же, скажем,

$$\frac{\partial y(x, \xi, \eta)}{\partial \eta} < 1 \quad \left(\frac{\partial f(x, y)}{\partial y} < 0 \right)$$

и интегральные кривые уравнения (1.1) сближаются, то влияние каждой из ошибок, наоборот, ослабевает,

Заметим, что, опираясь на формулу (11), можно получить вместо (5), (8) другое достаточное условие того, что значения приближенного решения ни на одном шаге вычислительного процесса не выходят за пределы указанной ранее области G_ε . Для этого в силу (11) с учетом (6), (9) достаточно, чтобы было выполнено неравенство

$$[\varepsilon_0 + (r + \delta) N] M \leq \varepsilon,$$

или

$$\varepsilon_0 + \frac{r + \delta}{h} (X - x_0) \leq \frac{\varepsilon}{M}.$$

Это условие, в отличие от (5), (8), не основано на принципе равномерной малости местных ошибок, отнесенных к интересующей нас точке (роль X может играть и любая другая фиксированная точка отрезка $[x_0, X]$), а имеет суммарный характер. Как и условия (5), (8), последнее условие является, конечно, завышенным, хотя, вообще говоря, и менее жестким. Оно заведомо выполняется, если удовлетворены ограничения (5), (8). Действительно,

$$\begin{aligned} \varepsilon_0 + \frac{r + \delta}{h} (X - x_0) &\leq \frac{\varepsilon h}{M(X - x_0 + h)} + \\ &+ \frac{\varepsilon}{M(X - x_0 + h)} (X - x_0) = \frac{\varepsilon}{M}. \end{aligned}$$

На основании формулы (11) можно получать также различные оценки погрешности приближенного решения. Для целей выяснения условий и скорости сходимости рассматриваемого вычислительного процесса можно ограничиться лишь следующей достаточно грубой равномерной оценкой, которая непосредственно вытекает из (11) с учетом неравенств (6), (9):

$$|\tilde{\varepsilon}_n| \leq \left[\varepsilon_0 + \frac{r + \delta}{h} (X - x_0) \right] M, \quad n = 1, 2, \dots, N. \quad (12)$$

В силу этой оценки можно утверждать, что если при $h \rightarrow 0$

$$\varepsilon_0 \rightarrow 0, \quad \frac{\delta}{h} \rightarrow 0, \quad \frac{r}{h} \rightarrow 0, \quad (13)$$

то в любой точке отрезка $[x_0, X]$ приближенное решение задачи (1.1), (1.2), полученное одношаговым методом типа (1), будет сходиться к точному решению этой задачи.

В частности, когда $\tilde{\epsilon}_0 = 0$, $\delta_i = 0$, $i = 1, 2, \dots, N$, и речь идет лишь об оценке погрешности метода (1), неравенства (12) принимают вид

$$|\epsilon_n| \leq \frac{r}{h} (X - x_0) M, \quad n = 1, 2, \dots, N, \quad (14)$$

а условия (13) переходят в следующее условие сходимости метода:

$$\frac{r}{h} \rightarrow 0 \quad \text{при} \quad h \rightarrow 0. \quad (15)$$

Так как обычно при построении одношаговых методов добиваются выполнения требования

$$r = O(h^{k+1}), \quad k > 0,$$

то условие (15) в этом случае заведомо выполняется и оценка (14) гарантирует в любой точке отрезка $[x_0, X]$ сходимость одношагового метода (1) со скоростью порядка h^k . По этой причине соответствующий метод называют обычно *методом k -го порядка точности*.

Чтобы в условиях реального вычислительного процесса в любой точке отрезка $[x_0, X]$ приближенное решение, полученное данным одношаговым методом k -го порядка точности, сходилась к точному решению исходной задачи Коши со скоростью порядка h^k , достаточно потребовать (см. (12)), чтобы выполнялись условия

$$\epsilon_0 = O(h^k), \quad \delta = O(h^{k+1}).$$

Хотя эти требования, принципиально говоря, и выполнимы, однако в реальных условиях, когда значность вычислительных машин ограничена, с уменьшением шага $h > 0$ практически трудно бывает столь быстро повышать точность выполнения вычислений. Обычно для реальных ЭВМ погрешности $\tilde{\epsilon}_0$ и δ_i , $i = 1, 2, \dots, N$, при изменении h остаются ограниченными снизу по абсолютному значению величинами. При сохранении значности ЭВМ погрешность $\tilde{\epsilon}_0$, как правило, вообще не изменяется. Следовательно (см. (11)), с уменьшением шага практически не изменяется и *неустраняемая погрешность* — часть погрешности приближенного решения, зависящая только от ошибки начального условия. *Вычислительная же погрешность* (часть погрешности приближенного решения,

источником которой являются лишь ошибки округлений) в силу оценки (12) из-за наличия множителя δ/h при $h \rightarrow 0$ может возрасти со скоростью порядка h^{-1} . Что же касается *погрешности метода*, то она, как мы уже отмечали, при этом обычно убывает со скоростью порядка h^k . Поэтому главную часть полной погрешности приближенного решения в зависимости от величины шага h составляют, как правило, либо погрешность метода (при относительно больших значениях h), либо вычислительная погрешность (при очень малых значениях шага). В случае грубого задания начального условия может, конечно, оказаться доминирующей и неустраняемая погрешность, однако при сильном изменении шага в сторону его увеличения или уменьшения на первый план, особенно при счете на большое число шагов, выходят, как правило, погрешность метода или вычислительная погрешность, при этом результат вычислений может оказаться неприемлемым по точности как при больших, так и при очень малых значениях шага. Естественно поставить задачу о нахождении такого значения h , при котором величина правой части оценки погрешности приближенного решения в данной точке была бы наименьшей. Это значение шага приведет, грубо говоря, к равным долям погрешности метода и вычислительной погрешности. Доля неустраняемой погрешности при этом должна быть также не больше. Тогда вычислительный процесс будет сбалансирован. В практике вычислений обычно стараются добиться того, чтобы неустраняемая и вычислительная погрешности не влияли на результат в пределах погрешности метода.

Многие из только что приведенных наводящих рассуждений основаны на оценке (12), которая является явно завышенной. Ошибки округлений, например, могут иметь разные знаки и частично компенсировать друг друга, не обязательно должна сохранять знак на каждом шаге процесса и погрешность формулы; составляющие части полной погрешности по шагам также, вообще говоря, могут давать отклонения не в одну сторону от решения. В действительности же все это, однако, вносит лишь незначительные искажения в нарисованную асимптотическую картину. Практикой вычислений правильность сделанных выводов обычно подтверждается.

Изложенное выше позволяет утверждать, что при организации вычислительного процесса необходимо следить за взаимным согласованием метода вычислений, величины шага h , требуемой точности результата, точности задания входных данных и точности вычислений. Иногда для такого согласования прибегают к помощи модельных задач. Частичное согласование этих факторов можно проводить, например, и на основании полученной оценки (12). Правда, использование этой оценки затруднено тем, что практически пригодные значения входящих в нее констант для большинства задач получить очень трудно. Величина r , например, обычно оценивается через максимумы абсолютных значений достаточно высоких производных от решения исходного уравнения. В редких случаях удастся получить удовлетворительные оценки ее. По той же причине априорная оценка (12) не часто используется и для суждения о точности окончательного результата. Для этих целей чаще применяют более простые в употреблении апостериорные оценки, пользоваться которыми можно лишь после проведения вычислений. При рассмотрении основных способов построения одношаговых методов мы иногда касались вопроса апостериорной оценки погрешности, особенно в тех случаях, когда сам метод был хорошо приспособлен для этих целей. Сейчас мы еще раз затронем эту проблему.

§ 6. Главный член погрешности. Правило Рунге

Будем иметь в виду по-прежнему задачу Коши (1.1), (1.2), предполагая выполненными требования к ней, сформулированные в предыдущем параграфе. Сейчас нас будет интересовать преимущественно лишь погрешность метода, т. е. та часть погрешности приближенного решения, которая зависит лишь от неточностей формулы. Считая равными нулю погрешности начального условия и округлений, подобно (5.11), можно записать

$$\varepsilon_n = \sum_{i=1}^n r_i \frac{\partial y(x_n, x_i, \eta_i)}{\partial \eta_i}, \quad n = 1, 2, \dots, N, \quad (1)$$

где

$$\eta_i = y_i + \theta_i [y(x_i, x_{i-1}, y_{i-1}) - y_i], \quad 0 < \theta_i < 1.$$

Пусть погрешность формулы (см. (5.4))

$$r_i = y(x_i, x_{i-1}, y_{i-1}) - y_i, \quad i = 1, 2, \dots, n,$$

представима в виде

$$y(x_i, x_{i-1}, y_{i-1}) - y_i = h^{k+1} \varphi(x_{i-1}, y_{i-1}) + O(h^{k+2}), \quad (2)$$

где $k > 0$, при этом известная функция $\varphi(x, y)$ в области G_ε , введенной в предыдущем параграфе, ограничена по абсолютному значению вместе со своими частными производными первого порядка.

Для большинства вычислительных методов типа (5.1) при естественных требованиях к гладкости правой части уравнения (1.1) представление остаточного члена в виде (2) получить нетрудно. Первое слагаемое такого представления при малых значениях шага $h > 0$ дает главную часть погрешности формулы. Во многих случаях бывает достаточным вместо учета всей погрешности $r_i, i = 1, 2, \dots, N$, ограничиться рассмотрением лишь ее главной части.

Поставим перед собой задачу, опираясь на представление (2) для локальной погрешности метода, выделить главную часть также и из суммарной погрешности (1) рассматриваемого метода. Использование главной части погрешности метода вместо всей погрешности более удобно для целей практики вычислений и часто бывает достаточным.

В силу (2) и оценки (5.14)

$$y(x_i, x_0, y_0) - y_i = O(h^k).$$

Поэтому, если предполагать решение $y(x, \xi, \eta)$ уравнения (1.1) в описанной ранее области изменения переменных x, ξ, η дважды непрерывно дифференцируемым по начальному данному η , то, опираясь на теорему Лагранжа о конечном приращении, значение

$$\frac{\partial y(x_n, x_i, \eta_i)}{\partial \eta}$$

в (1) можно заменить на

$$\frac{\partial y[x_n, x_i, y(x_i, x_0, y_0)]}{\partial \eta} + O(h^k).$$

И аналогично $\varphi(x_{i-1}, y_{i-1})$ можно заменить на

$$\varphi[x_{i-1}, y(x_{i-1}, x_0, y_0)] + O(h^k) = \varphi[x_i, y(x_i, x_0, y_0)] + O(h),$$

что позволяет записать (2) иначе:

$$r_i = h^{k+1} \varphi[x_i, y(x_i, x_0, y_0)] + O(h^{k+2}).$$

Тем самым, принимая во внимание, кроме того, что $nh \leq X - x_0$, формуле (1) можно придать вид

$$\begin{aligned} \varepsilon_n = h^{k+1} \sum_{i=1}^n \varphi[x_i, y(x_i, x_0, y_0)] \times \\ \times \frac{\partial y[x_n, x_i, y(x_i, x_0, y_0)]}{\partial \eta} + O(h^{k+1}), \quad (3) \\ n = 1, 2, \dots, N. \end{aligned}$$

Если теперь на основании квадратурной формулы правых прямоугольников

$$\int_{x_0}^{x_n} \psi(t) dt = h \sum_{i=1}^n \psi(x_i) + O(h),$$

где $x_i = x_0 + ih$, $h > 0$, записать равенство

$$\begin{aligned} h \sum_{i=1}^n \varphi[x_i, y(x_i, x_0, y_0)] \frac{\partial y[x_n, x_i, y(x_i, x_0, y_0)]}{\partial \eta} = \\ = \int_{x_0}^{x_n} \varphi[t, y(t, x_0, y_0)] \frac{\partial y[x_n, t, y(t, x_0, y_0)]}{\partial \eta} dt + O(h), \end{aligned}$$

то (3) можно переписать в виде

$$\begin{aligned} \varepsilon_n = h^k \int_{x_0}^{x_n} \varphi[t, y(t, x_0, y_0)] \times \\ \times \frac{\partial y[x_n, t, y(t, x_0, y_0)]}{\partial \eta} dt + O(h^{k+1}), \quad (4) \\ n = 1, 2, \dots, N. \end{aligned}$$

Так как

$$\frac{\partial y(x, \xi, \eta)}{\partial \eta} = \exp \left\{ \int_{\xi}^x f_y[\tau, y(\tau, \xi, \eta)] d\tau \right\},$$

то в (4)

$$\frac{\partial y[x_n, t, y(t, x_0, y_0)]}{\partial \eta}$$

можно заменить на

$$\exp \left\{ \int_t^{x_n} f_y \{ \tau, y[\tau, t, y(t, x_0, y_0)] \} d\tau \right\}$$

или, что то же самое, на

$$\exp \left\{ \int_t^{x_n} f_y [\tau, y(\tau, x_0, y_0)] d\tau \right\}.$$

Тогда будем иметь, что

$$\varepsilon_n = h^k \rho(x_n) + O(h^{k+1}), \quad n = 1, 2, \dots, N, \quad (5)$$

где

$$\rho(x) = \int_{x_0}^x \Phi[t, y(t, x_0, y_0)] \exp \left\{ \int_t^x f_y[\tau, y(\tau, x_0, y_0)] d\tau \right\} dt.$$

Если $\rho(x_n) \neq 0$, то величина $h^k \rho(x_n)$ при достаточно малых значениях шага $h > 0$ будет давать неплохое представление о значении погрешности метода в точке x_n . Поэтому величину $h^k \rho(x_n)$ называют обычно *главным членом погрешности метода* или просто *главным членом погрешности*. Она не дает, конечно, полной характеристики погрешности метода и, тем более, погрешности приближенного решения, однако нередко посредством ее можно получить достаточно хорошее представление об истинной величине ошибки. Это будет в том случае, когда неустранимая погрешность невелика по абсолютному значению в сравнении с погрешностью метода, а шаг h выбран настолько малым, чтобы величина $h^k \rho(x_n)$ достаточно хорошо представляла собой погрешность метода, но все же и не очень малым, чтобы вычислительная погрешность оставалась еще небольшой по абсолютному значению по отношению к погрешности метода. Практика вычислений показывает, что для данного класса задач и фиксированного типа ЭВМ существует ограниченная сверху и снизу область значений шагов, в которой главный член погрешности избранного метода дает хорошее представление о величине полной погрешности приближенного решения, при этом, конечно, имеется в виду, что требования к точности результата согласованы с величиной неустранимой погрешности, которая

от шага h не зависит, либо заранее обеспечена необходимая малость этой погрешности. При значениях h из этой области члены высших порядков в асимптотическом разложении погрешности метода вместе с вычислительной погрешностью малы по абсолютному значению в сравнении с главным членом погрешности. Такую область изменения h назовем *реальной областью асимптотики*. При подходе к границе этой области в сторону увеличения значений шагов наблюдается обычно постепенное ухудшение асимптотики, поскольку при этом в правой части (5) возрастает роль членов, содержащих высшие степени h . Ухудшение же асимптотики на выходе из этой области в сторону уменьшения значений шагов связано с возрастанием относительной доли вычислительной погрешности. Нижняя граница реальной области асимптотики существенно зависит от разрядной сетки вычислительной машины. Для машин с более широкой разрядной сеткой эта граница, вообще говоря, меньше, так как в этом случае обычно меньше величина δ (см., например, оценку (5.12)) и отношение δ/h достигает критических значений при меньших h .

На основании формулы (5) легко указать простое апостериорное правило оценки главного члена погрешности, известное под названием *правила Рунге*, которое часто используется в вычислительной практике.

Предположим, что в некоторой фиксированной точке x отрезка $[x_0, X]$ при двух достаточно малых значениях шага $h = h_1$ и $h = h_2$ данным одношаговым методом k -го порядка точности найдены соответственно два приближенных значения y_{h_1} и y_{h_2} решения $y(x)$ задачи Коши (1.1), (1.2). Тогда в силу (5) можно ожидать, что будут малы по абсолютному значению погрешности приближенных равенств

$$y(x) - y_{h_1} \approx h_1^k \rho(x),$$

$$y(x) - y_{h_2} \approx h_2^k \rho(x).$$

Исключая из этих равенств неизвестное значение $y(x)$, находим

$$\rho(x) \approx \frac{y_{h_1} - y_{h_2}}{h_1^k - h_2^k}. \quad (6)$$

Это дает возможность после проведения вычислений до точки x с шагами $h = h_1$ и $h = h_2$ получить приближенные

значения каждой из величин $h_i^k \rho(x)$, $i=1, 2$, и составить представление в погрешностях значений y_{h_1} и y_{h_2} . Кроме того, при заданной границе ε допустимой ошибки на основании приближенного равенства

$$\varepsilon \approx |\rho(x)| h_\varepsilon^k$$

по результатам этих вычислений можно будет выбрать практически более приемлемое при данных требованиях к точности значение шага

$$\varepsilon \approx \sqrt[k]{\varepsilon \frac{|h_2^k - h_1^k|}{|y_{h_1} - y_{h_2}|}}. \quad (7)$$

В практике вычислений наиболее часто в качестве значений шагов h_1 и h_2 избирают h и $h/2$. В этом случае формулы (6), (7) принимают соответственно вид

$$\begin{aligned} \rho(x) &\approx \frac{2^k (y_{h/2} - y_h)}{h^k (2^k - 1)}, \\ h_\varepsilon &\approx \frac{h}{2} \sqrt[k]{\frac{(2^k - 1) \varepsilon}{|y_{h/2} - y_h|}}. \end{aligned} \quad (8)$$

Используя (8), в частности, можно записать, что

$$y(x) - y_{h/2} \approx \frac{y_{h/2} - y_h}{2^k - 1}. \quad (9)$$

Эта простая формула, как и все приближенные равенства, основанные на представлении (5) для погрешности метода, имеет асимптотический характер и при конкретных значениях шага h не играет роли гарантированной оценки.

Заметим, что, записав равенство (9) в виде

$$y(x) \approx y_{h/2} + \frac{y_{h/2} - y_h}{2^k - 1}, \quad (10)$$

можно истолковать его как правило увеличения точности приближенного решения. Правило (10) является частным случаем более общего метода Рунге-Кутты ускорения сходимости последовательности. Подробнее о методах улучшения сходимости речь пойдет в гл. 9.

Отметим в заключение, что правило Рунге может быть использовано не только при решении обыкновенных дифференциальных уравнений одношаговыми методами. Это

правило можно применять всякий раз, когда асимптотическое поведение погрешности избранного метода имеет указанный выше характер. Широко пользуются этим правилом, например, в практике вычисления интегралов, при численном решении различных задач для дифференциальных уравнений с частными производными и в других вычислительных задачах.

§ 7. Методы типа двусторонних

Во многих практических задачах, приводящих к решению обыкновенных дифференциальных уравнений с начальными условиями, на первом плане стоит проблема получения результата с гарантированной точностью. Объем вычислений при этом учитывается не как первостепенный фактор. Решение этой проблемы значительно упростилось бы, если бы были построены вычислительные методы, которые позволяют одновременно находить два приближенных решения данной задачи Коши, соответствующие значения которых в рассматриваемых точках расположены по разные стороны от истинного значения искомого решения. Задача о построении таких методов издавна привлекала внимание математиков. Еще в начале этого столетия С. А. Чаплыгиным, например, были разработаны основные принципы построения одного из двусторонних методов. Метод Чаплыгина принадлежит к группе аналитических методов и позволяет строить на каждом отрезке интегрирования две последовательности кривых, приближающих сверху и снизу искомую интегральную кривую данного уравнения. Правда, этот метод накладывает довольно сильные ограничения на правую часть дифференциального уравнения и, кроме того, он трудно реализуется при работе на ЭВМ. Достаточное полное изложение метода Чаплыгина дано, например, в [2]. Мы остановимся здесь на рассмотрении методов несколько иного типа, которые не являются строго двусторонними и обеспечивают приближение к искомому значению точного решения с разных сторон лишь в смысле главного члена погрешности. Однако, в отличие от метода Чаплыгина, эти методы принадлежат к группе численных методов, легко реализуются на ЭВМ и являются более универсальными.

Предположим, что в случае задачи (1.1), (1.2) построено семейство таких методов типа (5.1), для которых выражение (6.2) погрешности формулы имеет вид

$$r_i = y(x_i, x_{i-1}, y_{i-1}) - y_i = h^{k+1} \gamma \psi(x_{i-1}, y_{i-1}) + O(h^{k+2}), \quad (1)$$

где γ — свободный числовой параметр семейства, а $\psi(x, y)$ — известная функция указанных аргументов. Тогда, согласно (6.5), для погрешности ε_n каждого из этих методов будет справедливо представление

$$\varepsilon_n = y(x_n, x_0, y_0) - y_n = h^k \rho(x_n) + O(h^{k+1}), \quad n = 1, 2, \dots, N, \quad (2)$$

где

$$\rho(x) = \gamma \int_{x_0}^x \psi[t, y(t, x_0, y_0)] \exp \left\{ \int_t^x f_y[\tau, y(\tau, x_0, y_0)] d\tau \right\} dt.$$

Если теперь среди вычислительных правил данного семейства избрать два таких метода, для которых соответствующие значения γ и γ параметра γ удовлетворяют $\gamma \times \times \gamma < 0$ (скажем, $\gamma = -\gamma$), то в точке x_n ($n = 1, 2, \dots, N$) в силу (2) главные части погрешностей избранных методов будут иметь противоположные знаки. Это дает основание надеяться на получение по избранным вычислительным правилам приближений к значению $y(x_n, x_0, y_0)$ с избытком и с недостатком, если, конечно, неустранимая погрешность мала по абсолютному значению в сравнении с погрешностью метода, а шаг $h > 0$ взят настолько малым, что величина $h^k \rho(x_n)$ в (2) достаточно хорошо представляет собой всю погрешность метода; при этом вычислительная погрешность еще не настолько велика, чтобы исказить результат вычислений в пределах принятой точности.

Семейства вычислительных правил, для которых локальная погрешность представима в виде (1), можно получать по-разному. За основу может быть взят, например, способ Рунге — Кутты построения одношаговых методов. Однако более просто такие методы строятся способом, указанным в § 4. Поэтому мы остановимся здесь лишь на последнем подходе к построению подобных вычислительных правил.

В п. 4.1 было показано, что погрешность приближенного равенства (4.2) при выполнении условий (4.4) может быть представлена в виде (4.5)

$$r_{n+1} = h^{k+1} y^{(k+1)}(x_n) \left[\frac{1}{(k+1)!} - \frac{1}{k!} \sum_{i=0}^q A_i \alpha_i^k \right] + O(h^{k+2}).$$

В этом представлении, как и всюду в § 4, в силу локального характера построений значение в точке x_n соответствующего решения уравнения (1.1) считается известным точно. При этом для упрощения записи зависимость решения от начальных данных в обозначениях не указывается (необходимые значения производных решения по формулам типа (1.4) могут быть точно выражены через это значение).

Если выбор параметров $A_i, \alpha_i, i=0, 1, \dots, q$, в (4.2) дополнительно к условиям (4.4) ограничить требованием

$$\frac{1}{(k+1)!} - \frac{1}{k!} \sum_{i=0}^q A_i \alpha_i^k = \gamma, \quad (3)$$

где γ — произвольное вещественное число, то тем самым будет описано семейство приближенных равенств типа (4.2), погрешность которых представима в виде (1):

$$r_{n+1} = h^{k+1} \gamma y^{(k+1)}(x_n) + O(h^{k+2}). \quad (4)$$

Конкретный вид функции $\psi(x, y)$ при этом может быть получен последовательным дифференцированием правой части уравнения (1.1). Для того чтобы на основании этих приближенных равенств можно было построить соответствующее семейство вычислительных методов, достаточно указать такие правила эффективного вычисления значений $y(x_n + \alpha_i h)$, $i=0, 1, \dots, q$, чтобы вносимая ими в (4.2) погрешность не изменяла главного члена ошибки (4). Для этого, очевидно, достаточно обеспечить нахождение значений $y(x_n + \alpha_i h)$, $i=0, 1, \dots, q$, с локальной погрешностью порядка h^{k+1} . Соответствующие вычислительные правила легко могут быть построены способом, изложенным в § 4.

Выделив из построенного так семейства вычислительных правил два метода, отвечающие значениям γ и

$\gamma \begin{pmatrix} [-] & [+], & [-] \\ \gamma & \gamma \cdot \gamma < 0 \end{pmatrix}$ параметра γ , мы получим два набора расчетных формул, позволяющих при выполнении оговоренных выше требований находить приближения с избытком и с недостатком к значениям искомого решения задачи (1.1), (1.2) в фиксированных точках отрезка $[x_0, X]$. Наиболее существенным из упомянутых требований является условие, чтобы величина $h^k p(x_n)$ в (2), знаком которой мы можем управлять за счет выбора знака параметра γ , действительно являлась главной частью погрешности метода. Последнее существенно зависит от того, будет ли действительно главным первый член в правой части равенства (4), иными словами, будет ли при выполнении условий (4.4), (3) слагаемое

$$h^{k+1} y^{(k+1)}(x_n) \left[\frac{1}{(k+1)!} - \frac{1}{k!} \sum_{i=0}^q A_i \alpha_i^k \right] \quad (5)$$

играть определяющую роль в асимптотическом разложении

$$r_{n+1} = \sum_{j=0}^{\infty} h^{j+1} y^{(j+1)}(x_n) \left[\frac{1}{(j+1)!} - \frac{1}{j!} \sum_{i=0}^q A_i \alpha_i^j \right] \quad (6)$$

погрешности приближенного равенства (4.2). Чтобы повысить роль слагаемого (5) в сумме (6), можно, кроме выполнения условий (4.4), (3), выбором параметров A_i , α_i , $i=0, 1, \dots, q$ (при соответствующем значении q), обеспечить также обращение в нуль нескольких следующих за слагаемым (5) членов разложения (6). Система ограничений на выбор A_i , α_i , $i=0, 1, \dots, q$, в этом случае может быть записана в виде

$$\sum_{i=0}^q A_i = 1, \quad \sum_{i=0}^q A_i \alpha_i^j = \frac{1}{j+1} - j! \gamma \delta_k^j, \quad j = 1, 2, \dots, p, \quad (7)$$

где δ_k^j — символ Кронекера, а заданием p ($p \geq k$) определяется число $p-k$ слагаемых в (6), следующих за (5), которые будут обращены в нуль. При $p=k$ требования (7) эквивалентны условиям (4.4), (3). Если выбор параметров в (4.2) осуществляется на основе этих требований,

то, чтобы не нарушить оговоренной выше структуры локальной погрешности при вычислении значений $y(x_n + \alpha_i h)$, $i = 0, 1, \dots, q$, соответствующие расчетные формулы необходимо выбрать так, чтобы обеспечить нахождение этих значений с погрешностями по крайней мере порядка h^{p+1} . Построение таких формул может быть просто осуществлено способом, изложенным в § 4.

С увеличением p возрастает, естественно, и громоздкость вычислительных правил. Поэтому значение p желательно брать не сильно превышающим k , а дальнейшего повышения роли главных членов погрешности формулы и погрешности метода можно добиваться за счет уменьшения шага $h > 0$ численного интегрирования. При этом удачу выбора h (и p) можно контролировать, сопоставляя, например, величину главного члена погрешности метода с величиной погрешности полученного приближения к решению и следя за согласованием этих погрешностей с требованиями к точности результата.

Описанный выше вычислительный процесс предполагает получение в каждой узловой точке x_n двух приближенных значений $y_n^{[+]}$ и $y_n^{[-]}$ величины $y(x_n)$, отвечающих соответственно значениям $\gamma^{[+]}$ и $\gamma^{[-]}$ параметра γ . При выполнении условий (7) в силу (2) главные члены погрешностей метода в каждом из этих случаев являются величинами порядка h^k ; при этом их знаки будут противоположными. Очевидно (см. (2)), что приближенное значение y_n величины $y(x_n)$, вычисленное по формуле

$$y_n = \frac{\gamma^{[+]} y_n^{[-]} - \gamma^{[-]} y_n^{[+]}}{\gamma^{[+]} - \gamma^{[-]}}, \quad (8)$$

будет (без учета ошибок начального условия и округлений) характеризоваться погрешностью уже по крайней мере порядка h^{k+1} . Можно показать, что погрешностью порядка h^{k+1} обладает и следующий метод получения приближений к искомому решению, не связанный с предварительным нахождением в каждой узловой точке x_n

двух приближений $y_n^{[+]}$ и $y_n^{[-]}$. В этом методе значение y_n вычисляется не по формуле (8), а находится непосредственно по одному из двух вычислительных правил

построенного выше семейства, отвечающих значениям $y^{[+]}$ или $y^{[-]}$ параметра y . При этом избранные правила используются не одновременно, как ранее, а применяются поочередно, так что в каждой узловой точке вычисления производятся лишь по одному из них. При такой организации вычислительного процесса не ставится задача о получении двусторонних приближений, а построенный вычислительный метод будет характеризоваться лишь тем обстоятельством, что на каждом шаге процесса главная часть локальной погрешности, как правило, меняет знак на противоположный. Этим и объясняется утверждаемое выше благоприятное поведение суммарной погрешности метода.

Только что описанный вычислительный процесс поочередного использования каждой из двусторонних формул требует приблизительно вдвое меньшего объема вычислений на узел сетки в сравнении с рассмотренным выше двусторонним методом, который предполагает независимое построение верхних и нижних приближений. Если же отказаться от принципа независимого нахождения верхних и нижних приближений, то на основании изложенного в § 4 способа построения вычислительных методов можно сконструировать и более экономичный процесс двусторонних приближений. Надежность двусторонности при этом не будет сильно ослаблена.

В описанном ранее методе основные расчетные формулы двусторонних приближений имели вид

$$\begin{aligned} y_{n+1}^{[+]} &= y_n^{[+]} + h \sum_{i=0}^q A_i f \left(x_n + \alpha_i h, y_{n+\alpha_i}^{[+]} \right), \\ y_{n+1}^{[-]} &= y_n^{[-]} + h \sum_{i=0}^q A_i f \left(x_n + \alpha_i h, y_{n+\alpha_i}^{[-]} \right); \end{aligned} \quad (9)$$

при этом одношаговые вычислительные правила нахождения $y_{n+\alpha_i}^{[+]}$ и $y_{n+\alpha_i}^{[-]}$, $i=0, 1, \dots, q$, предполагали

использование в качестве исходной информации значений $y_n^{[+]}$ и $y_n^{[-]}$ соответственно. Здесь под A_i , α_i и A_i , α_i , $i=0, 1, \dots, q$, следует понимать значения параметров A_i , α_i , $i=0, 1, \dots, q$, удовлетворяющие системе (7) при

$\gamma^{[+]}$ и $\gamma^{[-]}$ соответственно. Смысл других обозначений в (9) также вполне аналогичен используемому ранее.

Сейчас будем считать, что на каждом шаге процесса величина y_n предварительно вычислена по формулам типа (8), а основные расчетные формулы двусторонних приближений имеют вид

$$\begin{aligned} y_{n+1}^{[+]} &= y_n^{[+]} + h \sum_{i=0}^q A_i^{[+]} f\left(x_n + \alpha_i^{[+]} h, y_{n+\alpha_i^{[+]}}^{[+]}\right), \\ y_{n+1}^{[-]} &= y_n^{[-]} + h \sum_{i=0}^q A_i^{[-]} f\left(x_n + \alpha_i^{[-]} h, y_{n+\alpha_i^{[-]}}^{[-]}\right), \end{aligned} \quad (10)$$

при этом для нахождения $y_{n+\alpha_i^{[+]}}^{[+]}$ и $y_{n+\alpha_i^{[-]}}^{[-]}$, $i=0, 1, \dots$

$\dots, q$, применяются одношаговые правила, использующие в качестве исходной информации лишь одно вычисленное ранее значение y_n . Если (при достаточно большом q) поло-

жить $\alpha_i^{[+]} = \alpha_i^{[-]}$ для всех (или почти всех) значений $i=0, 1, \dots, q$, то построенный метод двусторонних приближений будет почти вдвое экономичнее соответствующего двустороннего метода типа (9). И в случае такого вычислительного процесса, подобно (2), для верхних и нижних приближений можно получить выражения главных членов погрешности метода, которые также будут различаться между собой лишь множителями $\gamma^{[+]}$ или $\gamma^{[-]}$.

В последнем методе получения двусторонних приближений можно сравнительно просто проследить за переменной знака множителя $y^{(k+1)}(x_n)$ в представлении типа (4) главного члена локальной погрешности. О перемене знака этого множителя при оговоренных ранее предположениях можно с достаточно высокой достоверностью сделать заключение на основании сравнения знаков величин

$$h \sum_{i=0}^q A_i^{[+]} f\left(x_n + \alpha_i^{[+]} h, y_{n+\alpha_i^{[+]}}^{[+]}\right) - h \sum_{i=0}^q A_i^{[-]} f\left(x_n + \alpha_i^{[-]} h, y_{n+\alpha_i^{[-]}}^{[-]}\right) \quad (11)$$

и

$$y_n^{[+]} - y_n^{[-]} \quad (n > 0). \quad (12)$$

Если знак разности (11) окажется противоположным знаку разности (12), естественно формулы (10) заменить на формулы

$$y_{n+1}^{[+]} = y_n^{[+]} + h \sum_{i=0}^q A_{if}^{[-]} \left(x_n + \alpha_i h, y_{n+\alpha_i}^{[-]} \right),$$

$$y_{n+1}^{[-]} = y_n^{[-]} + h \sum_{i=0}^q A_{if}^{[+]} \left(x_n + \alpha_i h, y_{n+\alpha_i}^{[+]} \right).$$

Заметим, что разность (12) равна $\varepsilon_n^{[-]} - \varepsilon_n^{[+]}$, где

$$\varepsilon_n^{[+]} = y(x_n) - y_n^{[+]}, \quad \varepsilon_n^{[-]} = y(x_n) - y_n^{[-]}.$$

Так как в случае достаточно малого шага $h > 0$ ошибки

$\varepsilon_n^{[+]}$ и $\varepsilon_n^{[-]}$ должны иметь противоположные знаки, то абсолютное значение этой разности должно давать хорошее представление о сумме абсолютных значений главных членов погрешностей двусторонних приближений. И аналогично абсолютное значение разности (11) при малых h должно хорошо представлять сумму абсолютных значений локальных погрешностей соответствующих методов. Сопоставляя значения величин (11) и (12), вычисленных при соблюдении условий малости абсолютных значений ошибок начального условия и округлений, с соответствующими величинами главных членов погрешностей и требуемой точностью результата, можно, в частности, высказать суждение об удаче выбора шага h с большей достоверностью, чем в случае двусторонних приближений (9), когда мы не имели непосредственно по ходу вычислений информации типа (11) о локальной погрешности метода. При этом, конечно, приходится считаться с тем, что даже грубое вычисление главных членов погрешностей является очень трудоемкой задачей. Поэтому при выборе шага часто руководствуются и другими соображениями.

Чтобы не загромождать изложение, не станем все сказанное выше иллюстрировать примерами конкретных вычислительных правил, а приведем только некоторые из простейших примеров.

Рассмотрим сначала случай, когда $k = p = 1$. Тогда система (7) принимает вид

$$\sum_{i=0}^q A_i = 1, \quad \sum_{i=0}^q A_i \alpha_i = \frac{1}{2} - \gamma.$$

При $q = 0$ находим

$$A_0 = 1, \quad \alpha_0 = \frac{1}{2} - \gamma.$$

Требование одношаговости $0 \leq \alpha_0 \leq 1$ приводит к следующим ограничениям на выбор параметра γ :

$$-\frac{1}{2} \leq \gamma \leq \frac{1}{2}.$$

Выбрав два значения $\overset{[+]}{\gamma} = 1/2$, $\overset{[-]}{\gamma} = -1/2$ параметра γ , удовлетворяющие этим ограничениям, найдем соответственно

$$\overset{[+]}{\alpha_0} = 0, \quad \overset{[-]}{\alpha_0} = 1, \quad \overset{[+]}{A_0} = \overset{[-]}{A_0} = 1.$$

Метод двусторонних приближений вида (9) в этом случае может быть записан в форме

$$\begin{aligned} \overset{[2]}{y_{n+1}} &= \overset{[2-]}{y_n} + h \overset{[2-]}{f_n}, \\ \overset{[2-]}{y_{n+1}} &= \overset{[2-]}{y_n} + h \overset{[2]}{f_{n+1}}, \\ \overset{[2+]}{y_{n+1}} &= \overset{[2+]}{y_n} + h \overset{[2+]}{f_n}. \end{aligned} \quad (13)$$

Верхние индексы здесь, как и ранее, дают указание на порядок локальной погрешности метода, с которой было найдено данное значение приближенного решения, а также, где это существенно, определяют знак соответствующего значения параметра γ .

Погрешности основных формул двусторонних приближений (см. (4)) будут при этом иметь вид

$$\overset{[\pm]}{r_{n+1}} = \pm \frac{1}{2} h^2 y''(x_n) + O(h^3).$$

Метод (13) требует на каждый узел сетки трех обращений к блоку вычисления значений правой части уравнения.

Более экономичным, наверное, будет следующий процесс получения двусторонних приближений:

$$\begin{aligned}
 y_{n+1} &= y_n + h \overset{[2]}{f}_n, \\
 y_{n+1} &= y_n + h \overset{[2-]}{f}_{n+1}, \\
 y_{n+1} &= y_n + h \overset{[2+]}{f}_n, \\
 y_{n+1} &= \frac{1}{2} \left(\overset{[2+]}{y}_{n+1} + \overset{[2-]}{y}_{n+1} \right).
 \end{aligned} \tag{14}$$

Здесь i может принимать значения 2 или 3. Для случая $i=2$ формулы (14) в основном счете сопряжены с нахождением лишь одного значения функции $f(x, y)$ на каждый узел сетки, при этом будут вычислены четыре, вообще говоря, разных приближенных значения интересующей нас величины $y(x_{n+1})$, сопоставление которых также может дать полезную информацию. В этом случае, правда, проявляются некоторые особенности конструкции вычислительных правил (14), не характерные, вообще говоря, для одношаговых методов. На первом этапе, например, придется вычислять не одно, как всюду далее, а два значения функции $f(x, y)$. При этом каждый из последующих шагов вычислительного процесса будет связан со специфичным использованием информации разного уровня точности об искомом значении решения, полученной на предыдущем шаге. При $i=3$ однородность вычислений будет восстановлена; при этом окажется несколько улучшенной, вообще говоря, и точность результата, однако при таком значении i вычислительный процесс будет требовать двух обращений к блоку вычисления значений функции $f(x, y)$ на каждый узел сетки.

Положим теперь в системе (7) $k=1$, $p=2$. Тогда будем иметь

$$\sum_{i=0}^q A_i = 1, \quad \sum_{i=0}^q A_i \alpha_i = \frac{1}{2} - \gamma, \quad \sum_{i=0}^q A_i \alpha_i^2 = \frac{1}{3}.$$

Приближенные равенства типа (4.2), подчиненные таким ограничениям, будут характеризоваться погрешностью следующего вида (см. (6)):

$$r_{n+1} = h^2 \gamma y''(x_n) + O(h^4). \tag{15}$$

При $q=1$ и $\alpha_0=0$ полученные ограничения значительно упрощаются:

$$A_0 + A_1 = 1, \quad A_1 \alpha_1 = \frac{1}{2} - \gamma, \quad A_1 \alpha_1^2 = \frac{1}{3}.$$

Отсюда находим, что

$$\alpha_1 = \frac{2}{3(1-2\gamma)}, \quad A_1 = \frac{3}{4}(1-2\gamma)^2, \quad A_0 = 1 - \frac{3}{4}(1-2\gamma)^2.$$

Из условия $0 < \alpha_1 \leq 1$ получаем следующее ограничение на область изменения параметра γ :

$$\gamma \leq \frac{1}{6}.$$

В частности, при $\overset{[+]}{\gamma} = 1/6$ и $\overset{[-]}{\gamma} = -1/6$ соответственно находим

$$\begin{aligned} \overset{[+]}{\alpha_0} &= 0, & \overset{[+]}{A_0} &= \frac{2}{3}, & \overset{[+]}{\alpha_1} &= 1, & \overset{[+]}{A_1} &= \frac{1}{3}, \\ \overset{[-]}{\alpha_0} &= 0, & \overset{[-]}{A_0} &= -\frac{1}{3}, & \overset{[-]}{\alpha_1} &= \frac{1}{2}, & \overset{[-]}{A_1} &= \frac{4}{3}. \end{aligned}$$

По этому набору параметров можно построить следующее вычислительное правило:

$$\begin{aligned} \overset{[2]}{y_{n+1/4}} &= \overset{[4]}{y_n} + \frac{h}{4} \overset{[i]}{f_n}, \\ \overset{[3]}{y_{n+1/2}} &= \overset{[4]}{y_n} + \frac{h}{2} \overset{[2]}{f_{n+1/4}}, \\ \overset{[3]}{y_{n+1}} &= \overset{[4]}{y_n} + h \overset{[3]}{f_{n+1/2}}, \\ \overset{[2+]}{y_{n+1}} &= \overset{[2+]}{y_n} + \frac{h}{3} \left(\overset{[i]}{2f_n} + \overset{[3]}{f_{n+1}} \right), \\ \overset{[2-]}{y_{n+1}} &= \overset{[2-]}{y_n} + \frac{h}{3} \left(-\overset{[i]}{f_n} + 4 \overset{[3]}{f_{n+1/2}} \right), \\ \overset{[4]}{y_{n+1}} &= \frac{1}{2} \left(\overset{[2+]}{y_{n+1}} + \overset{[2-]}{y_{n+1}} \right). \end{aligned}$$

Здесь индекс i может принимать значения 3 или 4.

На основании этих расчетных формул нетрудно записать и двусторонний метод вида (9), в котором верхние и нижние приближения находятся уже независимо.

Погрешность подобных двусторонних формул характеризуется равенствами (см. (15))

$$\overset{[\pm]}{r_{n+1}} = \pm \frac{1}{6} h^2 y''(x_n) + O(h^4).$$

Построим пример двусторонних методов при $k=p=2$. В этом случае система (7) принимает вид

$$\sum_{i=0}^q A_i = 1, \quad \sum_{i=0}^q A_i \alpha_i = \frac{1}{2}, \quad \sum_{i=0}^q A_i \alpha_i^2 = \frac{1}{3} - 2\gamma.$$

Возьмем $q=1$ и положим $\alpha_0=0$. Тогда будем иметь

$$\alpha_0=0, \quad A_0 = \frac{1-24\gamma}{4(1-6\gamma)}, \quad \alpha_1 = \frac{2}{3}(1-6\gamma), \quad A_1 = \frac{3}{4(1-6\gamma)}.$$

Тем самым будет описано однопараметрическое семейство приближенных равенств типа (4.2), погрешность которых (см. (4)) представима в виде

$$r_{n+1} = h^3 \gamma y'''(x_n) + O(h^4).$$

Из условия $0 < \alpha_1 \leq 1$ получаем следующие ограничения на выбор параметра γ этого семейства:

$$-\frac{1}{12} \leq \gamma < \frac{1}{6}.$$

Положив, например, $\gamma^{[+]} = \frac{1}{24}$, $\gamma^{[-]} = -\frac{1}{12}$, найдем

$$\alpha_0^{[+]} = 0, \quad A_0^{[+]} = 0, \quad \alpha_1^{[+]} = \frac{1}{2}, \quad A_1^{[+]} = 1,$$

$$\alpha_0^{[-]} = 0, \quad A_0^{[-]} = \frac{1}{2}, \quad \alpha_1^{[-]} = 1, \quad A_1^{[-]} = \frac{1}{2},$$

что позволит записать следующий метод двусторонних приближений вида (10):

$$y_{n+1/4}^{[2]} = y_n^{[4]} + \frac{h}{4} f_n^{[i]},$$

$$y_{n+1,2}^{[3]} = y_n^{[4]} + \frac{h}{2} f_{n+1/4}^{[2]},$$

$$y_{n+1}^{[3]} = y_n^{[4]} + h f_{n+1/2}^{[3]},$$

$$y_{n+1}^{[3+]} = y_n^{[4]} + h f_{n+1/2}^{[3]},$$

$$y_{n+1}^{[3-]} = y_n^{[4]} + \frac{h}{2} (f_n^{[i]} + f_{n+1}^{[3]}),$$

$$y_{n+1}^{[4]} = \frac{1}{3} (y_{n+1}^{[3-]} + 2 y_{n+1}^{[3+]}).$$

Индекс i здесь может принимать значения 3 или 4. При $i=3$ процесс нахождения двусторонних приближений в каждом (кроме первого) узле сетки сопряжен с вычислением трех новых значений функции $f(x, y)$.

Очевидным образом может быть записан и соответствующий двусторонний метод вида (9). Он, правда, будет требовать на каждый узел сетки уже семи обращений к блоку нахождения значений правой части уравнения (1.1).

Погрешности основных двусторонних формул при избранных значениях γ имеют вид

$$r_{n+1}^{[+]} = \frac{1}{24} h^3 y'''(x_n) + O(h^4),$$

$$r_{n+1}^{[-]} = -\frac{1}{12} h^3 y'''(x_n) + O(h^4).$$

Приведем без вывода еще один набор двусторонних формул вида (10), отвечающий случаю $k=p=3$:

$$y_{n+1/9}^{[2]} = y_n^{[5]} + \frac{h}{9} f_n^{[i]},$$

$$y_{n+2/9}^{[3]} = y_n^{[5]} + \frac{2h}{9} f_{n+1/9}^{[2]},$$

$$y_{n+1/3}^{[4]} = y_n^{[5]} + \frac{h}{12} (f_n^{[i]} + 3 f_{n+2/9}^{[3]}),$$

$$y_{n+2/3}^{[4]} = y_n^{[5]} + \frac{h}{3} (f_n^{[i]} - 3 f_{n+2/9}^{[3]} + 4 f_{n+1/3}^{[4]}),$$

$$y_{n+1}^{[4]} = y_n^{[5]} + \frac{h}{4} (f_n^{[i]} + 3 f_{n+2/3}^{[4]}),$$

$$y_{n+1}^{[4+]} = y_n^{[4+]} + \frac{h}{4} (f_n^{[i]} + 3 f_{n+2/3}^{[4]}),$$

$$y_{n+1}^{[4-]} = y_n^{[4-]} + \frac{h}{4} (3 f_{n+1/3}^{[4]} + f_{n+1}^{[4]}),$$

$$y_{n+1}^{[5]} = \frac{1}{2} (y_{n+1}^{[4+]} + y_{n+1}^{[4-]}).$$

Здесь i может принимать значения 4 или 5; при этом в случае $i=4$ использование указанных формул связано лишь с пятикратным обращением к блоку вычисления

значений функции $f(x, y)$ на каждый узел сетки (кроме узла x_1).

Соответствующие формулы типа (9) теперь легко могут быть выписаны.

Погрешности подобных формул двусторонних приближений задаются равенствами

$$r_{n+1}^{[\pm]} = \pm \frac{1}{216} h^4 y^{(4)}(x_n) + O(h^5).$$

Заметим, что при использовании указанных выше расчетных формул, как и в случае приведенных ранее двусторонних методов вида (10), в каждом узле x_{n+1} сетки будут получены четыре приближенных значения величины $y(x_{n+1})$, сопоставление которых может дать полезную информацию, в частности, в вопросе выбора шага h или при оценке точности результата.

§ 8. Многошаговые методы

В предыдущих параграфах этой главы мы рассматривали, как правило, лишь одношаговые методы решения задач с начальными условиями. Такие методы, обладая рядом удобных для практики вычислений особенностей, страдают одним существенным недостатком. При построении этих методов предполагается привлекать информацию о решаемой задаче только на отрезке длиной в один шаг, поэтому подобная информация на каждом этапе процесса должна быть, вообще говоря, получена заново, что предопределяет большую трудоемкость соответствующих вычислительных правил. Отказавшись от требования одношаговости, можно вычислительные методы строить так, чтобы часть полученной информации могла быть использована повторно на нескольких соседних шагах вычислительного процесса. При этом иногда оказывается целесообразным привлекать также и информацию с забеганием вперед за ту точку, значение решения в которой в настоящий момент мы вычисляем. Подобные методы, использующие информацию о решаемой задаче на отрезке длиной более одного шага, и будут основным объектом нашего изучения в этом параграфе.

Для простоты и большей наглядности изложения будем рассматривать сейчас, как и ранее, задачу Коши (1.1),

(1.2), имея в виду, что все основные результаты, полученные в этом случае, могут быть перенесены без больших затруднений на задачи с начальными условиями для дифференциальных уравнений любого порядка и систем таких уравнений.

Мы не станем здесь подробно останавливаться на вычислительных правилах достаточно общего вида (1.6), а ограничимся лишь методами типа

$$y_{n+1} = \sum_{i=0}^p a_i y_{n-i} + h \sum_{j=-s}^q A_j f(x_{n-j}, y_{n-j}), \quad (1)$$

позволяющими искать приближенное значение y_{n+1} решения рассматриваемой задачи в точке x_{n+1} сетки (1.5) в виде линейной комбинации нескольких известных, вообще говоря, приближенных значений y_{n-i} решения в точках x_{n-i} этой сетки с коэффициентами a_i , $i=0, 1, \dots, p$, и нескольких приближенных значений $f(x_{n-j}, y_{n-j})$ производной $y'(x) = f[x, y(x)]$ искомого решения в точках x_{n-j} с коэффициентами hA_j , $j=-s, -s+1, \dots, q$. При этом среди указанных значений производной могут быть и неизвестные (при $s \geq 1$). Вычислительные правила (1) очевидным образом могут быть обобщены, например, на тот случай, когда при разыскании y_{n+1} привлекается информация и о значениях производных более высокого порядка от решения.

Если $s < 1$, то вычислительные методы вида (1) обычно называют *явными*, при $s = 1$ — *неявными* (если, конечно, $A_{-1} \neq 0$), а в случае $s > 1$ при естественном условии

$\sum_{j=-s}^{-2} A_j^2 > 0$ такие правила вычислений часто называют методами *с забеганием вперед*.

Коэффициенты вычислительных правил вида (1) в общем случае неравномерных сеток (1.5) должны, вообще говоря, зависеть от избранной сетки. Как мы уже отмечали в § 1, в случае неравномерных сеток использование многошаговых методов усложняется. Здесь сетку (1.5) мы будем предполагать равномерной:

$$x_i = x_0 + ih, \quad i = 0, 1, \dots, N, \quad Nh = X - x_0.$$

В § 10 гл. 5 при рассмотрении задачи неопределенного интегрирования (5.10.1), являющейся частным

случае задачи (1.1), (1.2), были изучены вычислительные правила вида (5.10.2), сходные с методами (1) численного решения задачи Коши. Там было выяснено, что поведение погрешности рассматриваемого метода существенно зависит от корней характеристического уравнения (5.10.11), определяемого коэффициентами при значениях приближенного решения. Подобная зависимость имеет место и в более общем случае задачи Коши (1.1), (1.2). Опираясь на результаты этих исследований, положим $p=0$ в формулах (1). Тем самым будут заведомо исключены из рассмотрения все вычислительные правила указанного вида, по упомянутому признаку мало пригодные для вычислений (при этом, правда, будет отброшено также и значительное число приемлемых вычислительных методов). Ограничив себя рассмотрением вычислительных правил вида

$$y_{n+1} = y_n + h \sum_{i=-s}^q A_i f(x_{n-i}, y_{n-i}), \quad (2)$$

мы тем не менее сохраним для изучения все наиболее часто встречающиеся в вычислительной практике многошаговые методы.

Среди вычислительных правил вида (2) особенно широко известны *экстраполяционные* (при $s=0$) и *интерполяционные* (при $s=1$, $A_{-1} \neq 0$) *методы Адамса*, к рассмотрению которых мы и приступаем.

1. Экстраполяционные методы Адамса. Построение многошаговых методов вида

$$y_{n+1} = y_n + h \sum_{i=0}^q A_i f(x_{n-i}, y_{n-i}) \quad (3)$$

можно осуществлять, например, способом, изложенным в § 4.

Как и в случае одношаговых методов, будем исходить при этом из равенства (4.1). На его основе построим приближенное равенство (4.2), где параметрам α_i , $i=0, 1, \dots, q$, заранее придадим нужные для (3) значения

$$\alpha_i = -i, \quad i=0, 1, \dots, q. \quad (4)$$

Соответствующие значения параметров A_i , $i=0, 1, \dots, q$,

могут быть найдены из следующей системы (см. (4.4), (4)) линейных алгебраических уравнений:

$$\sum_{i=0}^q A_i = 1, \quad \sum_{i=0}^q A_i (-i)^j = \frac{1}{j+1}, \quad j = 1, 2, \dots, q. \quad (5)$$

Так как определитель этой системы есть определитель Вандермонда, а все $\alpha_i = -i$, $i = 0, 1, \dots, q$, различны, то значения параметров A_i , $i = 0, 1, \dots, q$, удовлетворяющие ограничениям (5), могут быть выбраны для любого $q \geq 0$ и притом единственным образом. При заданном q тем самым будет построен соответствующий экстраполяционный метод Адамса (3). В силу (4.5) погрешность формулы (3) при этом может быть записана в виде

$$r_{n+1} = h^{q+2} y^{(q+2)}(x_n) \left[\frac{1}{(q+2)!} - \frac{1}{(q+1)!} \sum_{i=0}^q A_i (-i)^{q+1} \right] + O(h^{q+3}). \quad (6)$$

Нетрудно непосредственно проверить, что те же вычислительные правила экстраполяционного метода Адамса (3) можно получить и следующим способом.

Для приближенного вычисления интеграла в (4.1) предварительно подынтегральную функцию $y'(x_n + \alpha h)$ на отрезке интегрирования аппроксимируем алгебраическим интерполяционным многочленом степени q , построенным по ее значениям в точках $x_n - ih$, $i = 0, 1, \dots, q$, предшествующих отрезку интегрирования. Такое интерполирование за пределы таблицы значений интерполируемой функции обычно называют *экстраполированием*, с чем и связано название метода (3). Осуществив затем простую операцию интегрирования алгебраического интерполяционного многочлена, мы приходим к приближенному равенству (4.2) при $\alpha_i = -i$, $i = 0, 1, \dots, q$, которое является основой формулы (3). Эта формула может быть записана по-разному в зависимости от того, в каком виде предварительно был представлен соответствующий интерполяционный многочлен. Если его взять, скажем, в форме Лагранжа (см. (1.3.5)), то экстраполяционный метод Адамса окажется записанным непосредственно в виде (3). Для коэффициентов A_i , $i = 0, 1, \dots, q$, в этом случае можно получить явные выражения, не прибегая к решению

системы (5). Эти коэффициенты выразятся через интегралы по отрезку $[x_n, x_{n+1}]$ от соответствующих коэффициентов Лагранжа и после простых преобразований могут быть записаны в виде

$$A_i = \frac{(-1)^i}{i!(q-i)!} \int_0^1 \frac{\alpha(\alpha+1) \dots (\alpha+q)}{\alpha+i} d\alpha, \quad i=0, 1, \dots, q.$$

Если же соответствующий интерполяционный многочлен для функции $y'(x) = f[x, y(x)]$ представить в форме Ньютона для интерполирования в конце таблицы (см. (1.4.3)), то экстраполяционный метод Адамса может быть записан через конечные разности функции $f(x, y)$. Такая форма записи метода позволяет получить интересующее нас значение приближенного решения в виде разложения по последовательным главным частям, что может быть использовано, например, для практической оценки локальной погрешности. Правда, оперирование с таблицами конечных разностей вызывает очевидные затруднения при работе на ЭВМ. Нужно иметь в виду также, что составление таких таблиц обычно сопряжено с быстрым накоплением вычислительной погрешности, особенно при высоких порядках разностей. Это предъявляет повышенные требования к точности выполнения вычислений. Поэтому в настоящее время в практике работы на ЭВМ часто отдают предпочтение форме записи (3) экстраполяционного метода Адамса.

Замена уравнения (1.1) разностным уравнением (3) при начальных условиях (1.2) приводит, вообще говоря, к некорректной задаче, так как задание лишь одного начального данного y_0 при $q > 0$ не выделяет единственного решения этого разностного уравнения порядка $q+1$. Чтобы построенное вычислительное правило позволяло находить единственное решение, достаточно задать дополнительно к y_0 значения $y_1, y_2, \dots, y_q$. Эти *начальные значения* приближенного решения могут быть получены по-разному. Для их нахождения можно использовать любой из рассмотренных выше одношаговых методов. Часто, чтобы не так сильно нарушать однородность вычислительного процесса, для построения начала таблицы значений приближенного решения конструируют специальные вычислительные алгоритмы, стараясь по возможности более

тесно привязать их к методу основного счета. Не останавливаясь здесь подробно на таких алгоритмах, заметим лишь, что точность построения начала таблицы должна быть согласована с точностью нахождения последующих значений решения исходной задачи. Как и в случае одношаговых методов, для метода (3), в частности, также может быть получена оценка погрешности приближенного решения, сходная с оценкой (5.12), где роль ϵ_0 будет играть соответствующая граница неточностей начальных данных. Основные выводы, которые были сделаны на основании (5.12) в отношении одношаговых методов, остаются справедливыми и для многошаговых вычислительных правил указанного вида. Очевидно также, что в случае рассматриваемых многошаговых методов можно получить и многие другие результаты, сходные с приведенными ранее для одношаговых правил. Повторно на этом мы не будем здесь останавливаться, а ограничимся лишь конкретными примерами экстраполяционных методов Адамса.

Возьмем $q=0$. Тогда, очевидно, из (3) получаем известный метод Эйлера

$$y_{n+1} = y_n + hf_n, \quad (7)$$

где

$$f_n = f(x_n, y_n).$$

Формула (6) для локальной погрешности методов в этом случае также приводит к известному результату

$$r_{n+1} = \frac{1}{2} h^2 y''(x_n) + O(h^3).$$

При $q=1$ система (5) имеет вид

$$A_0 + A_1 = 1, \quad A_1 = -1/2,$$

откуда находим

$$A_0 = 3/2, \quad A_1 = -1/2.$$

Метод Адамса (3) при таком значении q принимает вид

$$y_{n+1} = y_n + \frac{h}{2} (3f_n - f_{n-1}). \quad (8)$$

Погрешность последней формулы в силу (6) задается равенством

$$r_{n+1} = \frac{5}{12} h^3 y'''(x_n) + O(h^4).$$

В отличие от метода Эйлера (7), который не требует предварительного построения начала таблицы значений приближенного решения, экстраполяционный метод Адамса (8) уже предполагает предварительно нахождение величины y_1 .

В случае $q=2$ имеем

$$A_0 + A_1 + A_2 = 1, \quad A_1 + 2A_2 = -1/2, \quad A_1 + 4A_2 = 1/3,$$

откуда находим

$$A_0 = 23/12, \quad A_1 = -4/3, \quad A_2 = 5/12,$$

что позволяет записать следующий метод:

$$y_{n+1} = y_n + \frac{h}{12} (23f_n - 16f_{n-1} + 5f_{n-2}). \quad (9)$$

Его локальная погрешность задается формулой (см. (6))

$$r_{n+1} = \frac{3}{8} h^4 y^{(4)}(x_n) + O(h^5).$$

Вычислительное правило (9), очевидно, может применяться, начиная со значения $n=2$.

При $q=3$, как легко проверить непосредственно, (3) будет иметь вид

$$y_{n+1} = y_n + \frac{h}{24} (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}). \quad (10)$$

Погрешность формулы (10) задается равенством

$$r_{n+1} = \frac{251}{720} h^5 y^{(5)}(x_n) + O(h^6).$$

В заключение без подробного вывода дадим другой вид экстраполяционных методов Адамса, запись которого связана с использованием конечных разностей функции $y'(x) = f[x, y(x)]$:

$$\begin{aligned} y_{n+1} = & y_n + \varphi_n + \frac{1}{2} \Delta \varphi_{n-1} + \frac{5}{12} \Delta^2 \varphi_{n-2} + \frac{3}{8} \Delta^3 \varphi_{n-3} + \\ & + \frac{251}{720} \Delta^4 \varphi_{n-4} + \frac{95}{288} \Delta^5 \varphi_{n-5} + \frac{19\,087}{60\,480} \Delta^6 \varphi_{n-6} + \frac{5\,275}{17\,280} \Delta^7 \varphi_{n-7} + \\ & + \frac{1\,070\,017}{3\,628\,800} \Delta^8 \varphi_{n-8} + \frac{1\,082\,753}{7\,257\,600} \Delta^9 \varphi_{n-9} + \dots + C_q \Delta^q \varphi_{n-q}. \end{aligned} \quad (11)$$

Здесь

$$\varphi_i = h f_i, \quad C_q = \frac{1}{q!} \int_0^1 \alpha (\alpha + 1) \dots (\alpha + q - 1) d\alpha,$$

а $\Delta^j \varphi_i$ — конечная разность порядка j , составленная для функции $\varphi(x)$ по ее значениям в точках $x_i, x_{i+1}, \dots, x_{i+j}$ (см. § 2 гл. 1).

Погрешность формулы (11), как следует из описанного ранее способа ее построения, может быть представлена в виде

$$r_{n+1} = \frac{h^{q+2}}{(q+1)!} y^{(q+2)}(\xi) \int_0^1 \alpha (\alpha + 1) \dots (\alpha + q) d\alpha,$$

$$x_{n-q} \leq \xi \leq x_{n+1},$$

или

$$r_{n+1} = h^{q+2} y^{(q+2)}(x_n) C_{q+1} + O(h^{q+3}). \quad (12)$$

Формулы (11), (12), как мы уже отмечали, эквивалентны формулам (3), (6) соответственно. Однако в реальных условиях из-за наличия ошибок округлений вычислительные правила (3), (11), отвечающие одним и тем же значениям q , на одной и той же ЭВМ могут дать разные результаты при решении одной и той же задачи, так как они существенно отличаются организацией вычислений.

2. Интерполяционные методы Адамса. При $s = 1$ формула (2) принимает вид

$$y_{n+1} = y_n + h \sum_{i=-1}^q A_i f(x_{n-i}, y_{n-i}). \quad (13)$$

Построение таких вычислительных методов может быть осуществлено совершенно аналогично случаю экстраполяционных формул (3).

Положив

$$\alpha_i = -i, \quad i = -1, 0, 1, \dots, q,$$

для нахождения соответствующих значений параметров A_i ,

$i = -1, 0, 1, \dots, q$, получим систему линейных алгебраических уравнений

$$\sum_{i=-1}^q A_i = 1, \quad \sum_{i=-1}^q A_i (-i)^j = \frac{1}{j+1}, \quad j = 1, 2, \dots, q+1,$$

которая однозначно разрешима при любом $q \geq -1$.

Для коэффициентов A_i , $i = -1, 0, 1, \dots, q$, вычислительных методов вида (13), подобно случаю экстраполяционных методов Адамса, можно дать следующие явные выражения:

$$A_i = \frac{(-1)^{i+1}}{(i+1)!(q-i)!} \int_0^1 \frac{(\alpha-1)\alpha(\alpha+1)\dots(\alpha+q)}{\alpha+i} d\alpha,$$

$$i = -1, 0, 1, \dots, q.$$

Погрешность формулы (13), аналогично (6), может быть записана в виде

$$r_{n+1} = h^{q+3} y^{(q+3)}(x_n) \left[\frac{1}{(q+3)!} - \frac{1}{(q+2)!} \sum_{i=-1}^q A_i (-i)^{q+2} \right] +$$

$$+ O(h^{q+4}). \quad (14)$$

Приведем несколько примеров интерполяционных методов Адамса (13).

При $q = -1$ будем иметь

$$y_{n+1} = y_n + hf_{n+1},$$

$$r_{n+1} = -\frac{1}{2} h^2 y''(x_n) + O(h^3). \quad (15)$$

В случае $q = 0$ находим

$$y_{n+1} = y_n + \frac{h}{2} (f_{n+1} + f_n),$$

$$r_{n+1} = -\frac{1}{12} h^3 y'''(x_n) + O(h^4). \quad (16)$$

При $q = 1$ получаем

$$y_{n+1} = y_n + \frac{h}{12} (5f_{n+1} + 8f_n - f_{n-1}),$$

$$r_{n+1} = -\frac{1}{24} h^4 y^{(4)}(x_n) + O(h^5). \quad (17)$$

Взяв $q=2$, найдем

$$y_{n+1} = y_n + \frac{h}{24} (9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}),$$

$$r_{n+1} = -\frac{19}{720} h^5 y^{(5)}(x_n) + O(h^6). \quad (18)$$

Перечень подобных правил нетрудно продолжить. Не станем здесь выписывать другие вычислительные правила вида (13), а приведем без предварительных выкладок лишь иную форму записи интерполяционных методов Адамса, использующую конечные разности функции $f(x, y)$:

$$y_{n+1} = y_n + \varphi_{n+1} - \frac{1}{2} \Delta \varphi_n - \frac{1}{12} \Delta^2 \varphi_{n-1} - \frac{1}{24} \Delta^3 \varphi_{n-2} -$$

$$- \frac{19}{720} \Delta^4 \varphi_{n-3} - \frac{3}{160} \Delta^5 \varphi_{n-4} - \frac{863}{60480} \Delta^6 \varphi_{n-5} - \frac{275}{24195} \Delta^7 \varphi_{n-6} -$$

$$- \frac{33953}{3628800} \Delta^8 \varphi_{n-7} - \frac{57281}{7257600} \Delta^9 \varphi_{n-8} - \dots + C_{q+1} \Delta^{q+1} \varphi_{n-q}. \quad (19)$$

Здесь

$$\varphi_i = hf_i, \quad C_{q+1} = \frac{1}{(q+1)!} \int_{-1}^0 \alpha(\alpha+1) \dots (\alpha+q) d\alpha.$$

Погрешность (14) при этом может быть записана в виде

$$r_{n+1} = h^{q+3} y^{(q+3)}(x_n) C_{q+2} + O(h^{q+4}). \quad (20)$$

Формулы (19), (20) могут быть получены совершенно аналогично случаю экстраполяционных методов Адамса. Для этого достаточно в очевидном равенстве

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} y'(x) dx = y(x_n) + h \int_{-1}^0 y'(x_{n+1} + \alpha h) d\alpha$$

заменить функцию y' по интерполяционной формуле Ньютона для интерполирования в конце таблицы при узлах интерполирования $x_{n+1}, x_n, \dots, x_{n-q}$ и выполнить операцию интегрирования. В этом случае для $q \geq 0$ точки отрезка $[x_n, x_{n+1}]$, на котором производится замена функции y' по интерполяционной формуле, принадлежат наименьшему отрезку $[x_{n-q}, x_{n+1}]$, содержащему все узлы интерполирования. Таким образом, здесь выполняется не экстраполирование за пределы используемой таблицы

значений приближаемой функции, как это было в предыдущем пункте, а производится интерполирование в собственном смысле слова, чем и объясняется название методов (13), (19). Подобное расположение узлов интерполирования является, очевидно, более благоприятным, чем в экстраполяционном случае (см. представление (1.3.11) для остаточного члена интерполирования). Этим, в частности, объясняется тот факт, что абсолютные значения коэффициентов при конечных разностях в (19) убывают с ростом порядка разности быстрее, чем в случае метода (11), что предопределяет и более быстрое убывание абсолютных значений слагаемых в разложении (19) в сравнении с (11). По той же причине константа при $h^k y^{(k)}(x_n)$ в главном члене погрешности формулы (19) интерполяционного метода Адамса для $k > 2$ по абсолютной величине меньше соответствующей константы в случае экстраполяционного метода того же порядка точности. Заметим, что указанные константы, кроме того, имеют противоположные знаки (см., например, вид остаточных членов формул (7) и (15), (8) и (16), (9) и (17), (10) и (18)). В общем случае это непосредственно следует из вида коэффициентов в разложениях (11), (19). Поэтому при достаточно малых значениях $h > 0$ локальные ошибки экстраполяционного и интерполяционного методов Адамса одного и того же порядка точности также должны иметь противоположные знаки. В силу этого соответствующая линейная комбинация так полученных двух приближений к $y(x_{n+1})$ может дать результат на порядок точнее, если, конечно, ошибками округлений в пределах требуемой точности можно пренебречь.

Как по экстраполяционным, так и в случае интерполяционных методов Адамса приведенного выше вида вычисления можно начинать со значения $n = q$. При этом предварительно должно быть заготовлено начало таблицы значений приближенного решения до значения y_q включительно. Для одинаковых значений q при этом порядок малости относительно h погрешности (20) в случае интерполяционного метода Адамса будет на единицу выше соответствующего порядка малости погрешности (12) для экстраполяционного метода Адамса. Это обстоятельство является очевидным следствием того, что интерполяционный метод (13) по отношению к экстраполяционному (3)

строится с использованием одного дополнительного приближенного значения $f(x_{n+1}, y_{n+1})$ производной y' . При $q=0$, например, когда начало таблицы приближенных значений решения предварительно строить нет необходимости, методы (3) и (13) переходят в одношаговые методы Эйлера (2.2) и трапеций (3.2) с локальными ошибками порядка h^2 и h^3 соответственно (см. (7) и (16)).

Заметим, что построенные интерполяционные методы Адамса не дают явных выражений для нахождения y_{n+1} , а представляют собой лишь уравнения относительно этой неизвестной. Это значительно осложняет процесс вычисления y_{n+1} и является существенным недостатком метода. Обычно значение y_{n+1} находят лишь приближенно, пользуясь, как правило, каким-либо методом последовательных приближений. Как в форме (13), так и в форме (19) интерполяционный метод Адамса уже приведен к виду, удобному для итераций. Кроме того, заданием достаточно малого $h > 0$ при удачном выборе начального приближения можно обеспечить сходимость соответствующего итерационного процесса. В самом деле, уравнение относительно y_{n+1} можно (см. (13), (19)) записать в виде

$$y_{n+1} = \Phi(y_{n+1}),$$

где

$$\Phi(y_{n+1}) = h \left(1 + \sum_{i=1}^{q+1} C_i \right) f(x_{n+1}, y_{n+1}) + \\ + F(h, x_n, y_n, y_{n-1}, \dots, y_{n-q}),$$

а через $F(h, x_n, y_n, y_{n-1}, \dots, y_{n-q})$ обозначена известная функция указанных аргументов. Если при $x \in [x_0, X]$ $f(x, y)$ в достаточно широкой окрестности решения имеет непрерывную производную $\frac{\partial f(x, y)}{\partial y}$, то при удачном выборе исходного приближения $y_{n+1}^{(0)}$ за счет уменьшения шага $h > 0$ можно заведомо обеспечить выполнение условий сходимости процесса итераций $y_{n+1}^{(i+1)} = \Phi(y_{n+1}^{(i)})$, $i = 0, 1, \dots$ (см. § 2 гл. 4, теорема 1).

Начальное приближение к y_{n+1} можно выбирать по-разному. В случае интерполяционного метода Адамса в форме (19) вместо выбора исходного приближения $y_{n+1}^{(0)}$ к y_{n+1} можно выбрать соответствующее приближение $\Delta^{q+1}\Phi_{n-q}^{(0)}$ к $\Delta^{q+1}\Phi_{n-q}$. Так как конечная разность $\Delta^{q+1}\Phi_{n-q}$

является малой величиной по крайней мере порядка h^{q+2} в сравнении с шагом h (см. (1.2.10)), то при малых h и больших q она станет настолько малой по абсолютному значению, что нетрудно будет подобрать начальное приближение к ней с небольшой абсолютной погрешностью. Особенно просто это сделать в том случае, когда последняя участвующая в вычислениях конечная разность в пределах принятой точности является практически постоянной. Тогда в качестве начального приближения $\Delta^{q+1}\varphi_{n-q}^{(0)}$ можно взять известную величину $\Delta^{q+1}\varphi_{n-q-1}$, после чего по простым формулам составления конечных разностей найти соответствующие приближения

$$\Delta^q\varphi_{n-q+1}^{(0)}, \quad \Delta^{q-1}\varphi_{n-q+2}^{(0)}, \quad \dots, \quad \Delta\varphi_n^{(0)}, \quad \varphi_{n+1}^{(0)}.$$

Теперь уже по правилу (см. (19))

$$\begin{aligned} y_{n+1}^{(i+1)} = & y_n + \varphi_{n+1}^{(i)} - \frac{1}{2} \Delta\varphi_n^{(i)} - \\ & - \frac{1}{12} \Delta^2\varphi_{n-1}^{(i)} - \dots + C_{q+1} \Delta^{q+1}\varphi_{n-q}^{(i)} \end{aligned} \quad (21)$$

при $i=0$ легко вычислить значение $y_{n+1}^{(1)}$. Затем, исходя из $y_{n+1}^{(1)}$, можно по формуле

$$\varphi_{n+1}^{(i)} = hf(x_{n+1}, y_{n+1}^{(i)})$$

при $i=1$ и по простым правилам составления конечных разностей последовательно найти

$$\varphi_{n+1}^{(1)}, \quad \Delta\varphi_n^{(1)}, \quad \Delta^2\varphi_{n-1}^{(1)}, \quad \dots, \quad \Delta^{q+1}\varphi_{n-q}^{(1)},$$

после чего из (21) при $i=1$ получить $y_{n+1}^{(2)}$ и т. д.

Если же последняя конечная разность не является практически постоянной, то в силу малости ее абсолютного значения начальное приближение к ней с небольшой абсолютной погрешностью можно выбрать, экстраполируя значение $\Delta^{q+1}\varphi_{n-q}$ по нескольким предшествующим значениям конечной разности того же порядка.

При использовании интерполяционного метода Адамса в форме (13) обычно в качестве начального приближения к y_{n+1} берут соответствующее значение, полученное экстра-

поляционным методом (3). При этом часто ограничиваются лишь одной итерацией. В этом случае вычислительный процесс приобретает предсказывающе-исправляющий характер: по формуле (3) находится приближенное значение величины $y(x_{n+1})$ с локальной погрешностью порядка h^{q+2} , которое затем уточняется на порядок на основании формулы типа (13). Такая организация вычислений при использовании ЭВМ применяется наиболее часто.

Можно подобный процесс организовать и так, чтобы формула (3) применялась при значении q на единицу большем, чем в формуле (13). Тогда предсказывающая и исправляющая формулы будут характеризоваться погрешностями одного порядка малости относительно шага h , правда, первая из них все же будет давать более грубый, вообще говоря, результат, так как расположение узлов интерполирования, положенное в ее основу, как мы уже отмечали ранее, является менее удачным для минимизации остатка интерполирования в точках отрезка интегрирования $[x_n, x_{n+1}]$, чем в случае интерполяционного метода Адамса (13) с таким же числом узлов. Привлекательным при такой организации вычислений является то обстоятельство, что погрешность, вносимая теперь за счет неточности формулы типа (3) в остаток (14) формулы (13), не изменяет вида его главного члена. Поэтому, если шаг $h > 0$ достаточно мал и ошибки округлений не влияют на результат в пределах принятой точности, то локальные погрешности приближенных значений величины $y(x_{n+1})$, полученные в рассматриваемом вычислительном процессе по формулам экстраполяционного и интерполяционного методов Адамса, будут иметь противоположные знаки, что позволяет составить такую линейную комбинацию этих приближенных значений, которая будет на порядок точнее приближать искомое значение $y(x_{n+1})$. Коэффициенты этой линейной комбинации очевидным образом определяются через известные множители при $h^{q+3}y^{(q+3)}(x_n)$ в главных членах погрешностей соответствующих формул экстраполяционного и интерполяционного методов Адамса.

3. Методы с забеганием вперед. Выше мы уже отмечали, что при построении вычислительных методов решения задачи (1.1), (1.2) существенно, как расположена

относительно отрезка, где производится аппроксимация дифференциального уравнения (1.1), используемая для этого информация. В этом отношении при одном и том же числе привлекаемых значений функции $f(x, y)$ интерполяционный метод Адамса, например, оказывается предпочтительнее экстраполяционного. Еще более благоприятные по точности результаты можно ожидать в том случае, когда используемая информация будет расположена симметрично (или почти симметрично) относительно рассматриваемого отрезка. Если при построении многошаговых методов решения задачи (1.1), (1.2) исходить, скажем, из равенства (3.1)

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f[x, y(x)] dx,$$

то в указанном смысле значения функции $y'(x) = f[x, y(x)]$ естественно привлекать парами в следующих точках:

$$[x_n, x_{n+1}], [x_{n-1}, x_{n+2}], \dots, [x_{n-q}, x_{n+q+1}]. \quad (22)$$

В этом случае получим вычислительное правило вида (2) при $s = q + 1$:

$$y_{n+1} = y_n + h \sum_{i=-q-1}^q A_i f(x_{n-i}, y_{n-i}). \quad (23)$$

Формулы указанного вида называют обычно *методами типа Коуэлла*.

Строя такие методы способом, описанным в п. 1, следует вместо (4) задать

$$\alpha_i = -i, \quad i = -q-1, -q, \dots, q.$$

Тогда для определения значений параметров A_i , $i = -q-1, -q, \dots, q$, вместо (5) получим следующую систему линейных алгебраических уравнений:

$$\sum_{i=-q-1}^q A_i = 1, \quad \sum_{i=-q-1}^q A_i (-i)^j = \frac{1}{j+1}, \\ j = 1, 2, \dots, 2q+1.$$

Можно указать и другой способ нахождения коэффициентов A_i , $i = -q-1, -q, \dots, q$, вычислительных правил типа Коуэлла. Совершенно аналогично случаю методов Адамса для этих коэффициентов можно получить следующие явные выражения:

$$A_i = \frac{(-1)^{i+q+1}}{(i+q+1)!(q-i)!} \int_0^1 \frac{(\alpha-q-1)(\alpha-q) \dots (\alpha+q)}{\alpha+i} d\alpha,$$

$$i = -q-1, -q, \dots, q.$$

Погрешность формулы (23) в силу (4.5) может быть записана в виде

$$r_{n+1} = h^{2q+3} y^{(2q+3)}(x_n) \times$$

$$\times \left[\frac{1}{(2q+3)!} - \frac{1}{(2q+2)!} \sum_{i=-q-1}^q A_i (-i)^{2q+2} \right] + O(h^{2q+4}). \quad (24)$$

При $q=1$, например, соответствующая формула типа Коуэлла будет иметь вид

$$y_{n+1} = y_n + \frac{h}{24} (-f_{n+2} + 13f_{n+1} + 13f_n - f_{n-1}). \quad (25)$$

Ее погрешность задается равенством (см. (24))

$$r_{n+1} = \frac{11}{720} h^5 y^{(5)}(x_n) + O(h^6).$$

Заметим, что коэффициент при $h^5 y^{(5)}(x_n)$ в главной части остатка формулы (25), как и следовало ожидать, оказался меньше по абсолютному значению соответствующих коэффициентов в остаточных членах формул (10) и (18) экстраполяционного и интерполяционного методов Адамса.

Подобно случаю методов Адамса, формулы типа Коуэлла также могут быть записаны через конечные разности функции $f(x, y)$. При узлах (22) для интерполирования функции $y'(x) = f[x, y(x)]$ на отрезке $[x_n, x_{n+1}]$ естественно использовать формулу Ньютона — Бесселя вида (1.4.6). После выполнения в (3.1) простой операции

интегрирования алгебраического многочлена вместо (23) получим тогда следующую формулу:

$$y_{n+1} = y_n + \frac{\varphi_{n+1} + \varphi_n}{2} - \frac{1}{12} \frac{\Delta^2 \varphi_n + \Delta^2 \varphi_{n-1}}{2} + \\ + \frac{11}{720} \frac{\Delta^4 \varphi_{n-1} + \Delta^4 \varphi_{n-2}}{2} - \frac{191}{60 \cdot 480} \frac{\Delta^6 \varphi_{n-2} + \Delta^6 \varphi_{n-3}}{2} + \dots \\ \dots + C_{2q} \frac{\Delta^{2q} \varphi_{n-q-1} + \Delta^{2q} \varphi_{n-q}}{2}. \quad (26)$$

Здесь

$$\varphi_i = hf_i,$$

$$C_{2q} = \frac{1}{(2q)!} \int_0^1 (\alpha - q)(\alpha - q + 1) \dots (\alpha - 1) \alpha (\alpha + 1) \dots \\ \dots (\alpha + q - 1) d\alpha.$$

Коэффициенты C_{2i} в формуле (26) разностного метода Коуэлла с ростом i убывают быстрее соответствующих коэффициентов и экстраполяционного, и интерполяционного методов Адамса (см. (11), (19)). Погрешность этой формулы может быть записана в виде

$$r_{n+1} = h^{2q+3} y^{(2q+3)}(x_n) C_{2q+2} + O(h^{2q+4}).$$

В случае методов типа Коуэлла разыскиваемое значение y_{n+1} по формулам вида (23) или (26) выражается не только через известные величины $y_{n-q}, y_{n-q+1}, \dots, y_n$, но и через неизвестные значения $y_{n+1}, y_{n+2}, \dots, y_{n+q+1}$. Тем самым при каждом значении n равенства (23) или (26) дают нам не расчетные формулы для нахождения y_{n+1} , а уравнения, связывающие неизвестные значения $y_{n+1}, y_{n+2}, \dots, y_{n+q+1}$. Эти равенства можно записать, вообще говоря, лишь при $n = q, q+1, \dots, N-q-1$. Поэтому в случае методов типа Коуэлла однородность вычислительного процесса может нарушаться не только при нахождении начальных значений $y_1, y_2, \dots, y_q$ таблицы значений приближенного решения, но и при вычислении последних значений $y_{N-q+1}, y_{N-q+2}, \dots, y_N$ этой таблицы. Кроме того, использование самих формул типа Коуэлла также требует специальной организации вычислений; при этом в зависимости от формы записи (23) или (26) этих формул можно поступать по-разному.

Поясним, как можно вести вычисления на основе, скажем, формулы (23). Для определенности и простоты

записи будем считать $q=1$. Тогда (23) принимает вид (25). Эта формула представляет собой уравнение с двумя неизвестными y_{n+1} , y_{n+2} . Вычисления в этом случае можно начинать со значения $n=1$. Один из возможных вариантов организации таких вычислений может быть следующим:

$$\begin{aligned} y_2 &= y_1 + \frac{h}{2} (3f_1^{[i]} - f_0^{[i]}), \\ y_2 &= y_1 + \frac{h}{12} (5f_2^{[3]} + 8f_1^{[i]} - f_0^{[i]}), \\ y_{n+2} &= y_{n+1} + \frac{h}{12} (23f_{n+1}^{[4]} - 16f_n^{[i]} + 5f_{n-1}^{[i]}), \\ y_{n+1} &= y_n + \frac{h}{24} (-f_{n+2}^{[4]} + 13f_{n+1}^{[4]} + 13f_n^{[i]} - f_{n-1}^{[i]}). \end{aligned} \quad (27)$$

Здесь i может принимать значения 4 или 5.

Для построения такого алгоритма использованы формулы (8), (17), (9) и (25); при этом в целях большей наглядности применены обозначения типа (2.3).

Явное вычислительное правило (27) можно заменить другим, если, например, вместо двух первых его формул использовать неявную формулу (см. (17))

$$y_2 = y_1 + \frac{h}{12} (5f_2^{[4]} + 8f_1^{[i]} - f_0^{[i]}).$$

При записи метода типа Коуэлла в форме (26) вычисления можно организовать и по-другому. Поясним это, например, для случая $q=1$. Тогда формула (26) принимает вид

$$y_{n+1} = y_n + \frac{\varphi_{n+1} + \varphi_n}{2} - \frac{1}{12} \frac{\Delta^2 \varphi_n + \Delta^2 \varphi_{n-1}}{2}. \quad (28)$$

Из этого уравнения с двумя неизвестными y_{n+1} , y_{n+2} интересующее нас значение y_{n+1} будем находить итерационно.

Предположим, что, кроме дополнительного начального значения y_1 , предварительно найдено также значение y_2 . Будем вести вычисления с помощью формулы (28), начиная с $n=2$. Тогда наряду с y_n , φ_n нам будут известны также значения $\Delta \varphi_{n-1}$ и $\Delta^2 \varphi_{n-2}$. По $\Delta^2 \varphi_{n-2}$ выберем начальные приближения $\Delta^2 \varphi_{n-1}^{(0)}$ и $\Delta^2 \varphi_n^{(0)}$ к $\Delta^2 \varphi_{n-1}$ и $\Delta^2 \varphi_n$.

Проще это сделать в том случае, когда при выбранном шаге h вторые конечные разности практически постоянны в пределах принятой точности вычислений. В противном случае приходится прибегать к экстраполированию в столбце вторых разностей. По $\Delta^2 \varphi_{n-1}^{(0)}$ посредством простых формул составления конечных разностей последовательно вычислим $\Delta \varphi_n^{(0)}$ и $\varphi_{n+1}^{(0)}$, что позволит на основе (28) найти значение $y_{n+1}^{(1)}$ по формуле

$$y_{n+1}^{(i+1)} = y_n + \frac{\varphi_{n+1}^{(i)} + \varphi_n}{2} - \frac{1}{12} \frac{\Delta^2 \varphi_n^{(i)} + \Delta^2 \varphi_{n-1}^{(i)}}{2} \quad (29)$$

при $i=0$. Затем, исходя из $y_{n+1}^{(1)}$, по правилу

$$\varphi_{n+1}^{(i)} = hf(x_{n+1}, y_{n+1}^{(i)})$$

при $i=1$ можно найти $\varphi_{n+1}^{(1)}$ и, следовательно, вычислить $\Delta \varphi_n^{(1)}$ и $\Delta^2 \varphi_{n-1}^{(1)}$. Пользуясь полученным значением $\Delta^2 \varphi_{n-1}^{(1)}$, можно найти уточненное приближение $\Delta^2 \varphi_n^{(1)}$ к $\Delta^2 \varphi_n$, после чего из (21) при $i=1$ получим $y_{n+1}^{(2)}$ и т. д.

При численном решении задачи Коши (1.1), (1.2) для сравнительной характеристики методов можно рассматривать различные показатели. О трудоемкости метода естественно судить по числу обращений к блоку вычисления значений правой части уравнения на каждый узел сетки. Однородность вычислений можно характеризовать, в частности, числом дополнительных начальных значений приближенного решения, которые предварительно находят каким-то другим методом. Полезную дополнительную информацию о методах одного порядка точности можно получить, сравнив длины отрезков оси x , на которых при одном шаге вычислительного процесса рассматривается решение исходного дифференциального уравнения; при этом важным является также и расположение этих отрезков относительно отрезка интегрирования в (3.1).

Построенные методы типа Коуэлла по этим показателям можно сравнить с другими численными методами соответствующего порядка точности. Например, вычислительное правило (27) при $i=4$ характеризуется длиной $3h$ указанного выше отрезка с симметричным расположением его относительно $[x_n, x_{n+1}]$, одним дополнительным начальным значением и двумя обращениями на каждый узел сетки к блоку вычисления значений функции $f(x, y)$. В случае

соответствующего по порядку точности экстраполяционного метода Адамса длина рассматриваемого отрезка возрастает до $4h$, сильно нарушается симметрия его расположения относительно отрезка $[x_n, x_{n+1}]$, число дополнительных начальных значений увеличивается до трех, но зато этот метод требует нахождения лишь одного значения функции $f(x, y)$ на узел сетки. С другой стороны, одношаговые методы с локальной ошибкой порядка h^5 сопряжены уже с четырьмя обращениями к блоку вычисления значений функции $f(x, y)$ на узел сетки, но не требуют нахождения дополнительных начальных значений и характеризуются минимальной длиной h упомянутого выше отрезка.

Подобное положение методов типа Коуэлла, а также крайние позиции в отношении рассматриваемых показателей одношаговых методов и явных многошаговых вычислительных правил изучаемого выше вида наводят на мысль о разыскании таких методов с забеганием вперед, которые в указанном смысле носили бы еще более компромиссный характер. Ниже способом, указанным в § 4, мы построим простейшие примеры подобных вычислительных правил.

Положим, например, $q=4$ и $k=5$ в системе (4.4). Это дает возможность строить вычислительные правила с локальной ошибкой порядка h^6 , свободно располагаясь выбором пяти параметров.

Выберем по своему усмотрению параметры α_i , $i=0, 1, \dots, 4$, имея намерение за счет удачного расположения точек $x_n + \alpha_i h$, $i=0, 1, \dots, 4$, построить на основе (4.2) такие вычислительные правила с локальной ошибкой порядка h^6 , которые, во-первых, требовали бы в сравнении с соответствующими по порядку точности одношаговыми методами меньшего числа обращений к блоку вычисления значений функции $f(x, y)$ и, во-вторых, характеризовались бы в сравнении с экстраполяционным методом Адамса указанного порядка точности меньшим числом дополнительных начальных значений, лучшим расположением и меньшей длиной отрезка, на котором при одном шаге вычислительного процесса рассматривается решение исходного дифференциального уравнения. С этой целью при выборе α_i , $i=0, 1, \dots, 4$, не будем строго придерживаться требования одношаговости $0 \leq \alpha_i \leq 1$ и

в то же время будем задавать числа α_i не обязательно целыми.

Возьмем, например,

$$\alpha_0 = -1/2, \quad \alpha_1 = 0, \quad \alpha_2 = 1/2, \quad \alpha_3 = 1, \quad \alpha_4 = 3/2,$$

что позволит из (4.4) найти соответствующие значения

$$A_0 = A_4 = -1/180, \quad A_1 = A_3 = 34/180, \quad A_2 = 114/180.$$

Используя методику построения вычислительных методов, предложенную в § 4, на основе приведенных значений параметров можно получить следующие расчетные формулы:

$$\begin{aligned} y_{n+1}^{[4]} &= y_n^{[6]} + \frac{h}{6} \left(f_{n-1/2}^{[5]} - 2f_n^{[i]} + 7f_{n+1/2}^{[5]} \right), \\ y_{n+1}^{[5]} &= y_n^{[6]} + \frac{h}{6} \left(f_n^{[i]} + 4f_{n+1/2}^{[5]} + f_{n+1}^{[4]} \right), \\ y_{n+3/2}^{[5]} &= y_n^{[6]} + \frac{3h}{16} \left(-f_{n-1/2}^{[5]} + 5f_n^{[i]} - 3f_{n+1/2}^{[5]} + 7f_{n+1}^{[5]} \right), \\ y_{n+1}^{[6]} &= y_n^{[6]} + \frac{h}{180} \times \\ &\times \left(-f_{n-1/2}^{[5]} + 34f_n^{[i]} + 144f_{n+1/2}^{[5]} + 34f_{n+1}^{[5]} - f_{n+3/2}^{[5]} \right). \end{aligned} \quad (30)$$

Здесь i может принимать значения 5 или 6.

При $i=5$ построенный вычислительный метод требует в основном счете лишь трех обращений к блоку нахождения значений функции f на узел сетки (одношаговые правила соответствующего порядка точности связаны с вычислением не менее шести таких значений). Этот метод, кроме того, характеризуется симметричным расположением относительно $\{x_n, x_{n+1}\}$ и длиной $2h$ отрезка, на котором при одном шаге вычислительного процесса рассматривается решение исходного уравнения (1.1). При этом он предполагает нахождение лишь двух дополнительных начальных значений, для получения которых можно в случае $n=0$ использовать вычислительные правила (4.13) при шагах $-h/2$ и $h/2$. Для экстраполяционного метода Адамса такого порядка точности длина упомянутого отрезка равна $5h$; при этом предполагается предварительное нахождение четырех дополнительных начальных значений. Метод (30) является удобным и

в том отношении, что без дополнительных вычислительных затрат он дает возможность на каждом шаге процесса сопоставить три приближенных значения $y_{n+1}^{[4]}$, $y_{n+1}^{[5]}$, $y_{n+1}^{[6]}$ искомой величины $y(x_{n+1})$.

При выбранном выше исходном наборе параметров α_i , $i=0, 1, \dots, 4$, можно построить и другие варианты вычислительных правил типа (30). Например, в последнем методе вместо двух первых формул для нахождения $y_{n+1}^{[5]}$ можно использовать вычислительное правило

$$y_{n+1}^{[5]} = y_n + \frac{h}{6} \left(-f_{n-1}^{[5]} + 4f_n^{[5]} - 5f_n^{[4]} + 8f_{n+1/2}^{[5]} \right),$$

что позволит сократить до двух число вычисляемых значений функции f на каждый узел сетки в основном счете, однако при этом возрастет до трех число дополнительных начальных значений, несколько нарушится симметрия расположения относительно $[x_n, x_{n+1}]$ и увеличится до $2,5h$ длина отрезка, на котором при одном шаге процесса рассматривается решение исходного дифференциального уравнения.

§ 9. Решение линейных граничных задач

Наряду с задачами Коши для обыкновенных дифференциальных уравнений рассматриваются также граничные задачи. В этих задачах дополнительные условия, присоединяемые к дифференциальным уравнениям, задаются в виде уравнений, содержащих комбинации значений решения и его производных, взятых в нескольких точках отрезка, на котором разыскивается решение. Примером подобных задач может служить задача о нахождении средней линии нагруженной строительной балки, лежащей на нескольких точках опоры, задача о провисании нагруженной цепи, закрепленной в двух точках, и др.

Рассмотрим некоторые методы решения граничных задач.

1. Постановка задачи. Понятие о многоточечных и граничных задачах. Пусть на отрезке $a \leq x \leq b$ задана система линейных обыкновенных дифференциальных уравнений

Если $m > 2$, то граничная задача (1), (2) называется *многоточечной* или *m -точечной* задачей.

Отметим, что граничные задачи для линейных обыкновенных дифференциальных уравнений некоторого порядка n путем введения новых неизвестных функций могут быть записаны в виде (1), (2).

2. Метод дифференциальной прогонки. Пусть рассматривается граничная задача (1), (2). Решение этой задачи можно было бы найти как решение задачи Коши для уравнения (1), когда нам станет известным значение $y(t)$ в некоторой точке отрезка $[a, b]$. Если бы такое значение удалось как-либо определить, то тогда для нахождения $y(x)$ можно было бы применять метод Рунге — Кутты или метод Адамса, описанные нами выше. Идея метода прогонки как раз и состоит в том, что, используя уравнение (1) и граничное условие (2), находят значение искомого решения задачи (1), (2) в некоторой точке отрезка $[a, b]$, затем на всем отрезке $[a, b]$ вычисляют $y(t)$ как решение задачи Коши.

Рассмотрим функциональное соотношение вида

$$u^T(x) y(x) = \gamma(x), \quad (5)$$

где

$$u^T(x) = \begin{bmatrix} u_{11}(x) & u_{12}(x) & \dots & u_{1n}(x) \\ u_{21}(x) & u_{22}(x) & \dots & u_{2n}(x) \\ \dots & \dots & \dots & \dots \\ u_{l1}(x) & u_{l2}(x) & \dots & u_{ln}(x) \end{bmatrix}, \quad \gamma(x) = \begin{bmatrix} \gamma_1(x) \\ \gamma_2(x) \\ \vdots \\ \gamma_l(x) \end{bmatrix}, \quad 1 \leq l < n,$$

функции $u_{ij}(x)$, $\gamma_j(x)$ дифференцируемы, $y(x)$ — решение уравнения (1), T — знак транспонирования. Пусть при $x_1 = a$ известно, что $u(a) = u_1$, $\gamma(a) = \gamma_1$, где u_1 и γ_1 — заданные матрица и вектор. Предположим еще, что $\det(u_1^T u_1) \neq 0$. Поставим задачу: определить $u(x)$ и $\gamma(x)$ таким образом, чтобы $y(x)$ удовлетворяло при $a \leq x \leq b$ уравнению (1). Продифференцировав (5) и воспользовавшись формулой (1), получим

$$(u^{T'} - u^T A(x)) y = \gamma' - u^T f(x). \quad (6)$$

Выберем теперь $u(x)$ и $\gamma(x)$ так, чтобы (6) выполнялось при любых $y(x)$. Для этого достаточно положить

$$u^{T'} - u^T A(x) = 0$$

или

$$u' - A^T(x) u = 0, \quad (7)$$

и

$$\gamma' = u^T(x) f(x). \quad (8)$$

Поскольку $u(a)$ и $\gamma(a)$ известны, то, используя (7), (8), можно определить $u(x)$ и $\gamma(x)$ на всем отрезке $[a, b]$. Если бы при этом матрица $u(x)$ обладала свойством $\det(u^T(x) u(x)) \neq 0$ при всех $a \leq x \leq b$, то с помощью соотношения (5) и граничных условий (2) можно было бы определить значение $y(x)$, например, в точке $x=b$. Однако условие $\det(u^T(x) u(x)) \neq 0$ часто не выполняется, если уравнение (7) решается приближенно.

Усовершенствуем уравнения (7), (8). Представим матрицу $u(x)$ в виде

$$u(x) = v(x) \cdot \sigma(x), \quad (9)$$

где

$$v(x) = \begin{bmatrix} v_{11}(x) & v_{21}(x) & \dots & v_{l1}(x) \\ v_{12}(x) & v_{22}(x) & \dots & v_{l2}(x) \\ \dots & \dots & \dots & \dots \\ v_{1n}(x) & v_{2n}(x) & \dots & v_{ln}(x) \end{bmatrix},$$

$$\sigma(x) = \begin{bmatrix} \sigma_{11}(x) & \sigma_{12}(x) & \dots & \sigma_{1l}(x) \\ \sigma_{21}(x) & \sigma_{22}(x) & \dots & \sigma_{2l}(x) \\ \dots & \dots & \dots & \dots \\ \sigma_{l1}(x) & \sigma_{l2}(x) & \dots & \sigma_{ll}(x) \end{bmatrix}.$$

Функции $v_{ij}(x)$, $\sigma_{ij}(x)$ дифференцируемы. Предположим также, что на отрезке $a \leq x \leq b$ матрица $\sigma(x)$ не вырождена, причем $\sigma(a) = E$, где E — единичная матрица.

Получим для матрицы $v(x)$ дифференциальное уравнение. В силу формулы (9) и условия на матрицу $\sigma(x)$ имеем

$$v(a) = u_1. \quad (10)$$

Используя формулы (7) и (9), получим

$$v' - A^T(x) v + vS(x) = 0, \quad (11)$$

где

$$S(x) = \sigma'(x) \cdot \sigma^{-1}(x).$$

Матрицу $S(x)$ выберем таким образом, чтобы выполнялось условие

$$(v^T(x) \cdot v(x))' = v^{T'}(x) v(x) + v^T(x) v'(x) \equiv 0. \quad (12)$$

Для этого нужно положить

$$S(x) = (v^T(x) v(x))^{-1} v^T(x) A(x) v(x).$$

Таким образом, искомым для $v(x)$ будет уравнение

$$v' - A^T(x) v + v(x) (v^T(x) v(x))^{-1} v^T(x) A(x) v(x) = 0. \quad (13)$$

Отметим, что матрицу $v(x)$ теперь можно найти по уравнению (13) и начальному условию (10). При этом в силу свойства (12) матрицы $v(x)$ будем иметь

$$v^T(x) \cdot v(x) \equiv v^T(a) \cdot v(a) = u_1^T u_1$$

и, значит, $\det(v^T(x) v(x)) \neq 0$ при всех $a \leq x \leq b$.

Получим уравнение для вектор-функции $\omega(x)$, определяемой в соответствии с формулами (5) и (9) по правилу

$$v^T(x) y(x) = \omega(x), \quad (14)$$

где $\omega(x) = (\omega_1(x), \dots, \omega_l(x))^T$. Продифференцировав (14) и заменив в полученном выражении $v'(x)$ и $y'(x)$ по формулам (1) и (13), придем к уравнению вида

$$\omega' + v^T(x) A^T(x) v(x) (v^T(x) v(x))^{-1} \omega = v^T(x) f(x). \quad (15)$$

Вектор-функцию $\omega(x)$ можно вычислить как решение задачи Коши для уравнения (15) с начальным условием в точке $x = a$:

$$\omega(a) = \gamma_1.$$

Построенная здесь схема вычислений позволяет с помощью соотношения (14) переносить любое из граничных условий (2) в произвольную точку отрезка $[a, b]$. Это может быть использовано для определения значения $y(x)$ в точке $x = b$ следующим образом.

1. Полагаем $l = k_1$ и находим матрицу $v^T(x)$ размерности $(k_1 \times n)$ по уравнению (13) и начальному условию $v(x_1) = v(a) = \varphi_1$; при этом $a \leq x \leq x_2$. Соответствующий вектор $\omega(x)$ определяем на отрезке $[x_1, x_2]$, решая уравнение (15) с начальным условием $\omega(x_1) = \omega(a) = \alpha_1$. В точке $x = x_2$ получим $v^T(x_2) y(x_2) = \omega(x_2)$. Объединим это условие с условием (2), заданным в точке x_2 , и запишем их в виде

$$\psi_2^T y(x_2) = \beta_2,$$

где $\psi_2^T = \begin{bmatrix} v^T(x_2) \\ \varphi_2 \end{bmatrix}$ — матрица размерности $(l \times n)$, $l = k_1 + k_2$, $\beta_2 = \begin{bmatrix} \omega(x_2) \\ \alpha_2 \end{bmatrix}$ — вектор-столбец размерности $k_1 + k_2$.

Предполагаем, что $\det(\psi_2^T \psi_2) \neq 0$.

2. Полагаем $l = k_1 + k_2$ и находим на отрезке $[x_2, x_3]$ матрицу $v^T(x)$ размерности $(l \times n)$ по уравнению (13) и начальному условию $v(x_2) = \psi_2$; вектор $\omega(x)$ на этом же отрезке определяем по уравнению (15) с начальным условием $\omega(x_2) = \beta_2$. В точке $x = x_3$ получим $v^T(x_3) \mathbf{y}(x_3) = \omega(x_3)$, кроме того, в этой точке, согласно граничному условию (2), имеем соотношение $\varphi_3^T \mathbf{y}(x_3) = \alpha_3$. Объединив их, получим

$$\psi_3^T \mathbf{y}(x_3) = \beta_3,$$

где $\psi_3^T = \begin{bmatrix} v^T(x_3) \\ \varphi_3 \end{bmatrix}$ — матрица размерности $(l \times n)$, $l = k_1 + k_2 + k_3$, $\beta_3 = \begin{bmatrix} \omega(x_3) \\ \alpha_3 \end{bmatrix}$ — вектор-столбец размерности $k_1 + k_2 + k_3$.

Полагаем, что $\det(\psi_3^T \psi_3) \neq 0$ и по аналогии продолжаем прогонку решения на отрезках $[x_3, x_4], \dots, [x_{m-2}, x_{m-1}]$. В точке $x = x_{m-1}$ имеем

$$\psi_{m-1}^T \mathbf{y}(x_{m-1}) = \beta_{m-1},$$

где ψ_{m-1}^T — известная матрица размерности $(l \times n)$, $l = \sum_{i=1}^{m-1} k_i$, β_{m-1} — известный вектор-столбец той же размер-

ности. При $l = \sum_{i=1}^{m-1} k_i$ решаем на отрезке $[x_{m-1}, x_m]$, $x_m = b$, задачу Коши для уравнения (13) с начальным условием $v(x_{m-1}) = \psi_{m-1}$ и для уравнения (15) с начальным условием $\omega(x_{m-1}) = \beta_{m-1}$. В точке $x = x_m$ получим

$$v^T(x_m) \mathbf{y}(x_m) = \omega(x_m). \quad (16)$$

В этой же точке в силу граничного условия (2) имеем еще соотношение

$$\varphi_m^T \mathbf{y}(x_m) = \alpha_m. \quad (17)$$

Формулы (16) и (17) представляют в совокупности систему n линейных алгебраических уравнений с n неиз-

вестными $y_1(x_m), y_2(x_m), \dots, y_n(x_m)$. Предположим, что ранг этой системы n и, решив ее, например, по методу Гаусса, найдем

$$y(x_m) = \beta_m, \quad (18)$$

где $\beta_m = (\beta_1^{(m)}, \beta_2^{(m)}, \dots, \beta_n^{(m)})^T$ — известный вектор.

Искомое решение граничной задачи (1), (2) вычисляется теперь как решение задачи Коши для уравнения (1) с начальным условием (18) в точке $x_m = b$. Эту задачу приходится решать в направлении от точки b к началу отрезка — точке a . Для этой цели, как мы уже отмечали, могут быть использованы какие-либо методы численного решения задач Коши, например, методы Эйлера, Адамса, Рунге — Кутты и др.

Теория изложенного здесь метода разработана советским математиком А. А. Абрамовым.

Укажем на преимущества этого метода перед такими известными методами решения линейных граничных задач, как метод дополнительных функций и метод сопряженных уравнений (см. [4], §§ 8—10 гл. 3). Указанные методы основаны на использовании структуры общего решения для целей переноса граничных условий. Однако часто фундаментальная система решений однородного уравнения имеет тенденцию к вырождению, когда часть векторов в ней становится почти линейно зависимой. Это приводит к тому, что граничное условие, обладающее свойством невырожденности, скажем, в точке x_s , теряет это свойство при своем переносе в точку x_{s+1} из-за быстрого накопления погрешностей. В методе Абрамова этого не происходит. Действительно, пусть мы в точке $x = x_s$ имеем граничное условие

$$\varphi_s^T y(x_s) = \alpha_s. \quad (19)$$

Свойство невырожденности этой линейной комбинации характеризуется величиной $\det(\varphi_s^T \varphi_s)$, по условию $\det(\varphi_s^T \varphi_s) \neq 0$. В методе Абрамова перенос этого граничного условия из x_s в x_{s+1} осуществляется с помощью функционального отношения (14). Значит, при переносе граничного условия (19) из x_s в x_{s+1} мы получим в точке $x = x_{s+1}$ новое граничное условие

$$v^T(x_{s+1}) y(x_{s+1}) = \omega(x_{s+1}). \quad (20)$$

Свойство невырожденности линейной комбинации (20) характеризуется величиной $\det(v^T(x_{s+1})v(x_{s+1}))$, но в силу свойства (12) $\det(v^T(x) \cdot v(x)) \equiv \text{const}$ при $x_s \leq x \leq x_{s+1}$, и потому

$$\det(v^T(x_{s+1})v(x_{s+1})) = \det(v^T(x_s)v(x_s)).$$

Значит, линейные комбинации (19) и (20) в смысле невырожденности обладают одинаковыми свойствами.

§ 10. Решение нелинейных граничных задач

Рассмотрим при $a \leq x \leq b$ систему нелинейных дифференциальных уравнений

$$y' = F(x, y) \quad (1)$$

с граничным условием

$$g(y(a), y(b)) = h, \quad (2)$$

где $y(x) = (y_1(x), \dots, y_n(x))^T$ — вектор-функция, подлежащая определению, $F = (F_1, \dots, F_n)^T$ — известная вектор-функция,

$$F_i = F_i(x, y), \quad g = (g_1, \dots, g_n)^T, \quad g_i = g_i(y(a), y(b)),$$

g — известная вектор-функция, $h = (h_1, \dots, h_n)^T$ — заданный числовой вектор, $y(a)$, $y(b)$ — значения искомого решения граничной задачи (1), (2) соответственно в точках $x=a$ и $x=b$.

Требуется найти вектор-функцию $y(x)$, которая при $a < x < b$ удовлетворяла бы (1), а при $x=a$ и $x=b$ — граничному условию (2).

В приложениях граничное условие (2) часто задается в упрощенном виде, а именно:

$$By(a) = d, \quad Cy(b) = e, \quad (3)$$

где

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{k1} & b_{k2} & \dots & b_{kn} \end{bmatrix}, \quad C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{p1} & c_{p2} & \dots & c_{pn} \end{bmatrix}, \quad k + p = n,$$

B, C и d, e — соответственно известные матрицы и векторы и $\text{rang } B = k$, $\text{rang } C = p$, $d = (d_1, \dots, d_k)^T$, $e = (e_1, \dots, e_p)^T$.

Предположим, что граничная задача (1), (2) имеет единственное на $[a, b]$ решение.

1. Метод редукции к задачам Коши. Будем считать уравнение (1) таким, что для него задача Коши с начальным условием $y(a) = r$, где r — некоторый вектор из n -мерного евклидова пространства E_n , имеет единственное решение.

Рассмотрим задачу Коши

$$\begin{aligned} z' &= F(x, z), & a \leq x \leq b, \\ z(a) &= z^{(0)}, & z^{(0)} \in E_n, \end{aligned} \quad (4)$$

где $z(x) = (z_1(x), \dots, z_n(x))^T$, $z^{(0)} = (z_1^{(0)}, \dots, z_n^{(0)})^T$.

Решение задачи (4) можно записать в виде

$$z(x) = Z(x, z^{(0)}), \quad (5)$$

где $Z = (Z_1, \dots, Z_n)^T$ — некоторая вектор-функция, $Z_i = Z_i(x, z^{(0)})$. Положим в (5) $x = b$ и полученное значение $z(b)$ подставим в (2) вместо $y(b)$, а вместо $y(a)$ подставим $z(a) = z^{(0)}$, получим

$$G(z^{(0)}) \equiv g(z^{(0)}, Z(b, z^{(0)})).$$

Если бы было $G(z^{(0)}) = h$, то это означало бы, что $z^{(0)} = y(a)$, где $y(a)$ — значение решения граничной задачи (1), (2) в точке $x = a$, и $z(x)$ — решение задачи Коши (4) тождественно совпадало бы с $y(x)$.

Если $z^0 \neq y(a)$, то и $G(z^0) \neq h$. Таким образом, вычисление искомого значения $y(a)$ можно выполнить, решая систему нелинейных численных уравнений

$$G(z^{(0)}) = h \quad (6)$$

или в развернутом виде:

$$G_1(z_1^{(0)}, z_2^{(0)}, \dots, z_n^{(0)}) = h_1,$$

$$G_2(z_1^{(0)}, z_2^{(0)}, \dots, z_n^{(0)}) = h_2,$$

$$\dots \dots \dots$$

$$G_n(z_1^{(0)}, z_2^{(0)}, \dots, z_n^{(0)}) = h_n.$$

Будем считать, что система уравнений (6) имеет единственное решение $z^{(0)} = y(a)$.

Отметим здесь, что вид функций G_i нам неизвестен, однако мы можем вычислить значение $G(z^{(k)})$ для любого

$z^{(k)} \in E_n$. Этого можно добиться путем решения задачи Коши

$$\begin{aligned} z' &= F(x, z), \quad a \leq x \leq b, \\ z(a) &= z^{(k)}, \quad z^{(k)} \in E_n. \end{aligned} \quad (7)$$

Пусть $z^{(k)}(x)$ — решение задачи (7), тогда

$$G(z^{(k)}) \equiv g(z^{(k)}, Z(b, z^{(k)})) \equiv g(z^{(k)}, z^{(k)}(b)).$$

Теперь можно сформулировать алгоритм метода редукции к задачам Коши. Он состоит в следующем:

1) решаем систему нелинейных численных уравнений (6) и находим ее решение $z^{(0)} = y(a)$;

2) искомое решение граничной задачи (1), (2) находим как решение задачи Коши:

$$\begin{aligned} z' &= F(x, z), \quad a \leq x \leq b, \\ z(a) &= y(a). \end{aligned}$$

Как мы уже отмечали, будет выполняться тождество $z(x) \equiv y(x)$.

Для решения упоминавшихся здесь задач Коши могут быть использованы многие методы, в том числе методы Эйлера, Адамса, Рунге — Кутты и др.

Остановимся на вопросе численного решения системы уравнений (6). Для решения этой системы целесообразно избрать какой-либо итерационный метод, обладающий быстрой сходимостью. Покажем, как для этого может быть применен метод Ньютона.

Пусть $z^{(k)}$ — k -е приближение к решению $y(a)$ системы уравнений (6), тогда в методе Ньютона $(k+1)$ -е приближение может быть вычислено по формуле

$$z^{(k+1)} = z^{(k)} - v^{(k)}, \quad (8)$$

где $v^{(k)} = (v_1^{(k)}, v_2^{(k)}, \dots, v_n^{(k)})^T$ — вектор-столбец, определяемый как решение системы линейных алгебраических уравнений,

$$G_z^{(k)} v^{(k)} = G(z^{(k)}) - h, \quad (9)$$

$$G_z^{(k)} = \begin{bmatrix} \frac{\partial G_1^{(k)}}{\partial z_1} & \frac{\partial G_1^{(k)}}{\partial z_2} & \dots & \frac{\partial G_1^{(k)}}{\partial z_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial G_n^{(k)}}{\partial z_1} & \frac{\partial G_n^{(k)}}{\partial z_2} & \dots & \frac{\partial G_n^{(k)}}{\partial z_n} \end{bmatrix}, \quad G_i^{(k)} = G_i(z^{(k)}),$$

$k=0, 1, 2, \dots; i=1, 2, \dots, n$. Мы предполагаем, что $\det G_z^{(k)} \neq 0$. Точно вычислить элементы матрицы $G_z^{(k)}$ мы не можем, ибо неизвестен вид функций $G_i(z^{(k)})$. Однако эти элементы могут быть вычислены приближенно, например, по формулам

$$\frac{\partial G_i^{(k)}}{\partial z_j} \approx \frac{G_i(z^{(k)} + l_k e_j) - G_i(z^{(k)})}{l_k}, \quad (10)$$

где $l_k > 0$ — достаточно малое число, выбираемое вычислением

$$e_j = (0, \dots, 0, \underset{(j)}{1}, 0, \dots, 0)^T, \quad i, j = 1, 2, \dots, n.$$

Обозначим через $\tilde{G}_z^{(k)}$ матрицу, элементы которой вычисляются по формуле (10). Реально мы можем вместо последовательности $z^{(k)}$ вычислить последовательность $\tilde{z}^{(k)}$, определяемую по правилу

$$\tilde{z}^{(k+1)} = \tilde{z}^{(k)} - \tilde{v}^{(k)}, \quad (11)$$

где $\tilde{v}^{(k)}$ — решение системы уравнений

$$\tilde{G}_z^{(k)} \tilde{v}^{(k)} = G(\tilde{z}^{(k)}) - h, \quad (12)$$

$\tilde{z}^{(0)} = z^{(0)}, k=0, 1, 2, \dots$

Чтобы полностью уяснить вычислительную схему метода редукции к задаче Коши, покажем, как по $\tilde{z}^{(0)} = z^{(0)}$ ($z^{(0)}$ — некоторый вектор из E_n) вычислить $\tilde{z}^{(1)}$.

1. Решаем задачу Коши (7) при $z(a) = \tilde{z}^{(0)}$, вычисляем значение решения этой задачи в точке $x=b$ (обозначим его через $\tilde{z}^{(0)}(b)$), вычисляем $G(\tilde{z}^{(0)}) = g(\tilde{z}^{(0)}, \tilde{z}^{(0)}(b))$.

2. Решаем n задач Коши (7) при начальных условиях $z(a) = \tilde{z}^{(0)} + l_0 e_j, j=1, 2, \dots, n$; значения решений этих задач в точке $x=b$ обозначим соответственно

$$\tilde{z}^{(0,1)}(b), \tilde{z}^{(0,2)}(b), \dots, \tilde{z}^{(0,n)}(b).$$

Вычисляем $G(\tilde{z}^{(0)} + l_0 e_j)$ по правилу

$$G(\tilde{z}^{(0)} + l_0 e_j) = g(\tilde{z}^{(0)}, \tilde{z}^{(0,j)}(b)), \quad j=1, 2, \dots, n.$$

По формуле (10) вычисляем элементы матрицы $\tilde{G}_z^{(0)}$.

3. При $k=0$ решаем систему уравнений (12) и находим $\tilde{v}^{(0)}$. Первое приближение к $y(a)$ вычисляем по формуле (11) при $k=0$. Имеем: $\tilde{z}^{(1)} = \tilde{z}^{(0)} - \tilde{v}^{(0)}$.

4. Аналогично по $\tilde{z}^{(1)}$ вычисляется $\tilde{z}^{(2)}$ и т. д.

Пусть при некотором s мы с удовлетворительной точностью получили значение $\tilde{z}^{(s)} \approx y(a)$. Тогда приближенное решение граничной задачи (1), (2) вычисляется как решение задачи Коши для уравнения (1) с начальным условием $y(a) = \tilde{z}^{(s)}$.

Сходимость метода редукции к задачам Коши обуславливается сходимостью метода Ньютона (11), (12) с приближенной матрицей Якоби $\tilde{G}_z^{(k)}$. Условия сходимости такого метода сходны с условиями сходимости основного метода Ньютона, определяемого формулами (8), (9).

Отметим еще, что если на каждом номере итераций величину l_k выбирать из неравенства

$$l_k \leq \max_{1 \leq i \leq n} |G_i(\tilde{z}^{(k)}) - h_i|, \quad k=0, 1, 2, \dots,$$

то метод Ньютона, определяемый формулами (11), (12), может иметь квадратичную сходимость.

Укажем на некоторые особенности в выборе величины l_k . Если мы хотим, чтобы формула (10) была возможно более точной, мы должны по смыслу производной величину l_k выбирать возможно меньшей. Но тогда в выражении (10) в числителе будет результат вычитания двух близких по абсолютной величине чисел, и если эти числа были большими, то произойдет сильная потеря значащих цифр. Это приведет к тому, что производная в формуле (10) будет вычислена с большой погрешностью. Если же l_k будет выбрано недостаточно малым, то тогда по смыслу определения производной формула (10) не может дать хорошего по точности результата. По этим причинам вопрос оптимального выбора величины l_k требует в каждом конкретном случае специального рассмотрения.

2. Метод линеаризации. Задачу (1), (2) можно линеаризовать, используя идею метода Ньютона. Пусть $y^{(0)}(x) = (y_1^{(0)}(x), \dots, y_n^{(0)}(x))^T$ — нулевое приближение к искомому решению граничной задачи (1), (2). В предпо-

ложении, что функции F_i и g_i дважды непрерывно дифференцируемы по своим аргументам, можно записать следующие соотношения:

$$F_i(x, y(x)) \approx \\ \approx F_i(x, y^{(0)}(x)) + \sum_{j=1}^n \frac{\partial F_i(x, y^{(0)}(x))}{\partial y_j} (y_j(x) - y_j^{(0)}(x)), \quad (13)$$

$$g_i(y(a), y(b)) \approx g_i(y^{(0)}(a), y^{(0)}(b)) + \\ + \sum_{j=1}^n \frac{\partial g_i(y^{(0)}(a), y^{(0)}(b))}{\partial y_j(a)} (y_j(a) - y_j^{(0)}(a)) + \\ + \sum_{j=1}^n \frac{\partial g_i(y^{(0)}(a), y^{(0)}(b))}{\partial y_j(b)} (y_j(b) - y_j^{(0)}(b)), \quad i = 1, 2, \dots, n. \quad (14)$$

Покажем теперь, как, используя формулы (13), (14), на основании задачи (1), (2) вычислить первое приближение $y^{(1)}(x)$ к решению $y(x)$ по нулевому приближению $y^{(0)}(x)$. Введем обозначения:

$$F_y(x, y^{(0)}(x)) = \left(\frac{\partial F_i(x, y^{(0)}(x))}{\partial y_j} \right)_1^n, \\ G_{y(a)}(y^{(0)}(a), y^{(0)}(b)) = \left(\frac{\partial g_i(y^{(0)}(a), y^{(0)}(b))}{\partial y_j(a)} \right)_1^n, \\ G_{y(b)}(y^{(0)}(a), y^{(0)}(b)) = \left(\frac{\partial g_i(y^{(0)}(a), y^{(0)}(b))}{\partial y_j(b)} \right)_1^n;$$

это известные матрицы размерности $(n \times n)$. Теперь для определения $y^{(1)}(x)$ можно записать линейную граничную задачу:

$$y^{(1)'}(x) = F_y(x, y^{(0)}(x)) (y^{(1)}(x) - y^{(0)}(x)) + F(x, y^{(0)}(x)), \\ G_{y(a)}(y^{(0)}(a), y^{(0)}(b)) (y^{(1)}(a) - y^{(0)}(a)) + \\ + G_{y(b)}(y^{(0)}(a), y^{(0)}(b)) (y^{(1)}(b) - y^{(0)}(b)) + \\ + g(y^{(0)}(a), y^{(0)}(b)) = h. \quad (15)$$

В некоторых случаях для решения задачи (15) может быть использован метод дополнительных функций в такой

форме. Вычисляем фундаментальную систему решений $z^{(1)}(x), z^{(2)}(x), \dots, z^{(n)}(x)$ по правилу

$$\frac{dz^{(k)}(x)}{dx} = F_y(x, y^{(0)}(x)) z^{(k)}(x), \quad k = 1, 2, \dots, n,$$

$$z^{(k)}(a) = e_k, \quad e_k = (0, \dots, 0, 1, 0, \dots, 0)^T, \quad a \leq x \leq b.$$

Находим на $[a, b]$ функцию $Y(x)$ как решение задачи Коши:

$$Y'(x) = F_y(x, y^{(n)}(x)) (Y(x) - y^{(0)}(x)) + F(x, y^{(0)}(x)), \\ Y(a) = 0.$$

Искомое решение задачи (15) определяем по формуле

$$y^{(1)}(x) = Y(x) + c_1 z^{(1)}(x) + \dots + c_n z^{(n)}(x). \quad (16)$$

В этой формуле параметры $c_1, c_2, \dots, c_n$ вычисляются из условия, что функция $y^{(1)}(x)$, определяемая по (16) при $x=a$ и $x=b$, удовлетворяет граничному условию в (15), т. е. из системы уравнений

$$[G_{y(a)}(y^{(0)}(a), y^{(0)}(b)) + G_{y(b)}(y^{(0)}(a), y^{(0)}(b)) Z(b)] C = d,$$

где $Z(b) = [z^{(1)}(b), z^{(2)}(b), \dots, z^{(n)}(b)]$ — матрица размерности $n \times n$,

$$C = (c_1, c_2, \dots, c_n)^T,$$

$$d = h + G_{y(a)}(y^{(0)}(a), y^{(0)}(b)) \cdot y^{(0)}(a) - \\ - G_{y(b)}(y^{(0)}(a), y^{(0)}(b)) (Y(b) - y^{(0)}(b)) - g(y^{(0)}(a), y^{(0)}(b)).$$

После вычисления параметров $c_1, c_2, \dots, c_n$ решение задачи (15) можно вычислять по формуле (16). Но можно поступать и по-другому. Используя (16), вычисляем значение $y^{(1)}(x)$ при $x=a$ или $x=b$, а затем вычисляем на всем отрезке $[a, b]$ функцию $y^{(1)}(x)$ как решение задачи Коши для дифференциального уравнения (15) с начальным условием в точке $x=a$ или в точке $x=b$.

Аналогично по $y^{(1)}(x)$ вычисляется $y^{(2)}(x)$ и т. д. Обычно вычисления ведутся до тех пор, пока не будет достигнут такой номер k итерации, что будет выполняться условие

$$\max_{1 \leq i \leq n} \max_{a \leq x \leq b} |y_i^{(k+1)}(x) - y_i^{(k)}(x)| \leq \varepsilon, \quad (17)$$

где ε — некоторое положительное число, характеризующее

точность вычисления $y(x)$. Если $y^{(k)}(x)$ — такая вектор-функция, что для нее выполняется условие (17), то ее принимают за приближенное решение граничной задачи (1), (2).

Метод линеаризации, при обычных для метода Ньютона условиях, в частности при достаточно хорошем начальном приближении $y^{(0)}(x)$, имеет квадратичную сходимость.

Отметим еще, что если среди фундаментальных решений дифференциального уравнения (15) есть быстро растущие решения, то от метода дополнительных функций как средства решения задачи (15) следует отказаться из-за его неустойчивости к ошибкам округлений. Здесь целесообразно будет воспользоваться методом С. К. Годунова, описанным в [1].

§ 11. Метод сеток для решения линейных граничных задач

Этот метод является одним из наиболее универсальных как для обыкновенных дифференциальных уравнений, так и для уравнений в частных производных. Мы проведем изложение метода сеток на примере простейших линейных граничных задач для дифференциальных уравнений второго порядка.

1. Постановка задачи. Идея метода сеток. Пусть при $a \leq x \leq b$ рассматривается граничная задача для дифференциального уравнения

$$L(y) \equiv y'' + p(x)y' + q(x)y = f(x) \quad (1)$$

с условиями

$$l_a(y) \equiv \alpha_0 y(a) + \alpha_1 y'(a) = A, \quad (2)$$

$$l_b(y) \equiv \beta_0 y(b) + \beta_1 y'(b) = B. \quad (3)$$

Будем считать, что граничная задача (1)—(3) имеет единственное решение, это решение непрерывно на $[a, b]$ и имеет непрерывные производные на этом отрезке до четвертого порядка включительно.

Метод сеток для решения граничной задачи (1)—(3), равно как и для многих других задач, состоит в следующем.

1. Область задания дифференциального уравнения (1) — отрезок $[a, b]$ — заменяется некоторой дискретной (сеточной) областью. Это означает, что на отрезке $[a, b]$ выбирается некоторая система точек. Совокупность этих точек называется *сеткой*. Если положение каждой точки определяется по правилу $x_k = a + kh$, $k = 0, 1, \dots, N$, $h = \frac{b-a}{N}$, N — целое число, то сетку называют *равномерной*. Точки x_k называют *узлами* сетки.

2. Граничная задача (1) — (3) на множестве узлов, принадлежащих сетке, заменяется некоторой сеточной задачей. Под термином сеточная задача мы будем понимать некоторые соотношения между приближенными значениями решения граничной задачи (1) — (3) в узлах сетки. В рассматриваемом случае это будет система линейных алгебраических уравнений.

3. Полученная сеточная задача решается по какому-либо численному методу и тем самым находятся приближенные значения решения граничной задачи в узлах сетки. Это и является конечной целью метода сеток.

Можно указать как на основные в методе сеток на следующие вопросы.

1. Как заменить область задания дифференциальных уравнений, а в случае дифференциальных уравнений в частных производных еще и границу области некоторой сеточной областью?

2. Как заменить дифференциальное уравнение и граничные условия некоторыми сеточными соотношениями?

3. Будет ли полученная сеточная задача однозначно разрешимой, будет ли она устойчивой, сходящейся?

Смысл последних двух понятий будет пояснен позже, в частности в § 1 гл. 7.

2. Методы замены обыкновенных дифференциальных уравнений и граничных условий системой алгебраических уравнений. Возвратимся к граничной задаче (1) — (3). Выберем равномерную сетку: $x_k = a + kh$, $k = 0, 1, \dots, N$, $h = \frac{b-a}{N}$. Дифференциальное уравнение (1) будем рассматривать только во внутренних узлах сетки, т. е. будем полагать, что $x = x_k$, $k = 1, 2, \dots, N - 1$. Граничные условия (2), (3) рассмотрим при $x_0 = a$ и $x_N = b$.

Положим в (1) $x = x_k$:

$$L(y)|_{x=x_k} \equiv y''(x_k) + p(x_k)y'(x_k) + q(x_k)y(x_k) = f(x_k), \quad (4)$$

$$k = 1, 2, \dots, N-1.$$

Выразим $y'(x_k)$, $y''(x_k)$ через значения функции $y(x)$ в узлах x_{k-1} , x_k , x_{k+1} , т. е. через значения $y(x_{k-1})$, $y(x_k)$, $y(x_{k+1})$. Для этой цели мы воспользуемся интерполированием и численным дифференцированием на основе интерполяционных полиномов (см. т. I, § 6 гл. 1).

Имеем

$$y'(x_k) = \frac{y(x_k) - y(x_{k-1}))}{h} + r_k^{(1)}(h), \quad (5)$$

$$r_k^{(1)}(h) = \frac{h}{2} y''(x_k^{(1)}), \quad x_{k-1} < x_k^{(1)} < x_k;$$

$$y'(x_k) = \frac{y(x_{k+1}) - y(x_k)}{h} + r_k^{(2)}(h), \quad (6)$$

$$r_k^{(2)}(h) = -\frac{h}{2} y''(x_k^{(2)}), \quad x_k < x_k^{(2)} < x_{k+1};$$

$$y'(x_k) = \frac{y(x_{k+1}) - y(x_{k-1}))}{2h} + r_k^{(3)}(h), \quad (7)$$

$$r_k^{(3)}(h) = -\frac{h^2}{6} y'''(x_k^{(3)}), \quad x_{k-1} < x_k^{(3)} < x_{k+1};$$

$$y''(x_k) = \frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1}))}{h^2} + r_k^{(4)}(h), \quad (8)$$

$$r_k^{(4)}(h) = -\frac{h^2}{12} y^{IV}(x_k^{(4)}), \quad x_{k-1} < x_k^{(4)} < x_{k+1}.$$

Подставив в формулу (4) выражения (7) и (8) для $y'(x_k)$ и $y''(x_k)$, получим

$$L(y)|_{x=x_k} \equiv L_h(y(x_k)) + R_k(h) = f(x_k), \quad (9)$$

где

$$L_h(y(x_k)) = \frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1}))}{h^2} + p(x_k) \frac{y(x_{k+1}) - y(x_{k-1}))}{2h} +$$

$$+ q(x_k) y(x_k),$$

$$R_k(h) = r_k^{(4)}(h) + p(x_k) \cdot r_k^{(3)}(h).$$

Выражение $L_h(y(x_k))$ называется *разностным оператором* второго порядка, а величина $R_k(h)$ — *погрешностью аппроксимации* дифференциального оператора $L(y)$ разностным

оператором $L_h(y(x_k))$ на решении. Если для $R_k(h)$ выполняется условие

$$|R_k(h)| \leq Mh^2, \quad k = 1, 2, \dots, N-1,$$

где $M = \text{const}$, не зависящая от h , то говорят, что разностный оператор L_h аппроксимирует на решении дифференциальный оператор L с погрешностью второго порядка относительно h .

Пусть h достаточно мало, тогда в формуле (9) величиной $R_k(h)$ можно пренебречь и мы получим

$$L_h(y_k) = f(x_k), \quad k = 1, 2, \dots, N-1, \quad (10)$$

где при выполнении некоторых условий можно предположить, что $y_i \approx y(x_i)$, $i = 0, 1, 2, \dots, N$. Вообще говоря, всегда $y_i \neq y(x_i)$, и только при $R_k(h) \equiv 0$ можно ожидать, что будет $y_i = y(x_i)$, $i = 0, 1, 2, \dots, N$. Равенство (10) мы будем называть *разностной схемой*, аппроксимирующей уравнение $L(y) = f(x)$.

Отметим еще, что (10) есть система линейных алгебраических уравнений, число таких уравнений $N-1$, а матрица этой системы — трехдиагональная. Неизвестными являются $y_0, y_1, \dots, y_N$. Число этих неизвестных в системе равно $N+1$.

Обратимся к граничным условиям (2), (3). Используя (6) при $k=0$, из (2) получим

$$l_a(y) \equiv l_a^{(h)}(y(x_0)) + R_0(h) = A, \quad (11)$$

где

$$l_a^{(h)}(y(x_0)) = \alpha y(x_0) + \alpha_1 \left(\frac{y(x_1) - y(x_0)}{h} \right), \\ R_0(h) = \alpha_1 r_0^{(2)}(h).$$

Аналогично, используя (5) при $k=N$, из (3) получим

$$l_b(y) \equiv l_b^{(h)}(y(x_N)) + R_N(h) = B, \quad (12)$$

где

$$l_b^{(h)}(y(x_N)) = \beta_0 y(x_N) + \beta_1 \left(\frac{y(x_N) - y(x_{N-1})}{h} \right), \\ R_N(h) = \beta_1 r_N^{(1)}(h).$$

При достаточно малом h величинами $R_0(h)$ и $R_N(h)$, имеющими первый порядок малости относительно h , в выра-

жениях (11) и (12) можно пренебречь. Тогда вместо (11) и (12) будем иметь

$$l_a^{(h)}(y_0) = A, \quad (13)$$

$$l_b^{(h)}(y_N) = B. \quad (14)$$

Операторы $l_a^{(h)}(y_0)$ и $l_b^{(h)}(y_N)$ аппроксимируют соответственно граничные операторы $l_a(y)$ и $l_b(y)$ с погрешностью $O(h)$. Формулы (10), (13), (14) образуют в совокупности систему $N+1$ линейных алгебраических уравнений с неизвестными $y_0, y_1, \dots, y_N$. В методе сеток эту систему решают по какому-либо численному алгоритму и после этого полагают $y(x_k) \approx y_k, k=0, 1, \dots, N$.

Обратим внимание на то, что граничные условия (2), (3) при необходимости можно аппроксимировать разностными условиями с погрешностью второго порядка малости относительно h . Для этого достаточно, например, воспользоваться вместо (5) и (6) следующими формулами:

$$y'(x_0) = \frac{-y(x_2) + 4y(x_1) - 3y(x_0)}{2h} + O(h^2),$$

$$y'(x_N) = \frac{3y(x_N) - 4y(x_{N-1}) + y(x_{N-2})}{2h} + O(h^2).$$

3. О разрешимости систем разностных уравнений. Как мы видели, в методе сеток приближенное решение граничной задачи (1)–(3) вычисляется как решение системы линейных алгебраических уравнений, образуемых формулами (10), (13), (14). Выясним условия, при которых такая система уравнений будет однозначно разрешимой.

С целью упрощения доказательств будем считать, что в граничных условиях (2), (3) $\alpha_0 = \beta_0 = 1, \alpha_1 = \beta_1 = 0$. В этом случае сеточная задача (10), (13), (14) может быть записана в таком виде:

$$L_h(y_k) \equiv \frac{1}{h^2} \Lambda_h(y_k) = f_k, \quad k=1, 2, \dots, N-1, \quad (15)$$

$$y_0 = A, \quad y_N = B,$$

где

$$\Lambda_h(y_k) = A_k y_{k+1} + 2B_k y_k + C_k y_{k-1},$$

$$A_k = 1 + \frac{h}{2} p_k, \quad 2B_k = -2 + h^2 q_k, \quad C_k = 1 - \frac{h}{2} p_k,$$

$$p_k = p(x_k), \quad q_k = q(x_k), \quad f_k = f(x_k).$$

Перепишем систему уравнений (15) в ином виде, нормировав уравнения для внутренних точек сетки. Получим

$$\Lambda_h(y_k) = h^2 f_k, \quad k = 1, 2, \dots, N-1, \quad (16)$$

$$y_0 = A, \quad y_N = B.$$

Чтобы доказать однозначную разрешимость системы уравнений (16), достаточно показать, что соответствующая однородная система уравнений

$$\Lambda_h(z_k) = 0, \quad k = 1, 2, \dots, N-1, \quad (17)$$

$$z_0 = 0, \quad z_N = 0$$

имеет только нулевое решение, т. е. $z_k \equiv 0$. Тогда определитель системы уравнений (17) отличен от нуля, и так как он совпадает с определителем системы уравнений (16), то последняя будет однозначно разрешимой в силу правила Крамера.

Доказательство разрешимости системы уравнений (16) будем вести при двух предположениях относительно шага сетки h и коэффициента $q(x)$ уравнения (1).

Будем считать, что выполняются условия:

$$1) \quad \frac{h}{2} \max_{a \leq x \leq b} |p(x)| < 1, \quad (18)$$

$$2) \quad q(x) \leq 0 \quad \text{при} \quad a \leq x \leq b. \quad (19)$$

Предположим, что заданы некоторые числа $Y_0, Y_1, \dots, Y_N$, среди которых есть неравные. Имеет место

Лемма 1. Пусть выполняются условия (18), (19) и

$$\Lambda_h(Y_k) \geq 0 \quad (\Lambda_h(Y_k) \leq 0) \quad (20)$$

при $k = 1, 2, \dots, N-1$. Среди чисел $Y_0, Y_1, \dots, Y_N$ наибольшее положительное значение (наименьшее отрицательное значение) принимает или Y_0 , или Y_N .

Отметим, что эту лемму часто именуют принципом максимума.

Доказательство. Будем доказывать только первое утверждение при условии, что имеет место неравенство

$$\Lambda_h(Y_k) \geq 0, \quad k = 1, 2, \dots, N-1.$$

Предположим противное и будем считать, что наибольшее положительное значение числа Y_k принимают при

некотором k ($1 \leq k \leq N-1$), и пусть $Y_s = M > 0$ есть одно из таких значений, для которого Y_{s+1} или Y_{s-1} строго меньше M . Тогда

$$\Lambda_h(Y_s) \equiv A_s Y_{s+1} + 2B_s Y_s + C_s Y_{s-1} < (A_s + C_s)M + 2B_s M,$$

ибо A_s и C_s строго положительны в силу условия (18). В силу условия (19)

$$\Lambda_h(Y_s) < (A_s + C_s + 2B_s)M = h^2 q_s M \leq 0.$$

Таким образом, имеем $\Lambda_h(Y_s) < 0$. Но это противоречит условию (20). Значит, наше предположение неверно и наибольшее положительное значение будет иметь или Y_0 , или Y_N . Аналогично доказывается утверждение леммы и в случае условия $\Lambda_h(Y_k) \leq 0$.

Теорема 1. Пусть выполняются условия (18), (19). Тогда система уравнений (17) имеет только нулевое решение.

Доказательство. В силу формулы (17) $\Lambda_h(z_k) = 0$, $k = 1, 2, \dots, N-1$. Поэтому можно считать, что выполняются оба условия (20) леммы 1, т. е. $\Lambda_h(z_k) \geq 0$ и $\Lambda_h(z_k) \leq 0$. В этом случае по лемме 1 наибольшее положительное значение и наименьшее отрицательное значение среди чисел $z_0, z_1, \dots, z_N$ будут иметь числа z_0 или z_N , но в уравнениях (17) $z_0 = z_N = 0$, значит, все числа $z_k \equiv 0$. Теорема доказана.

Из теоремы 1 сразу следует однозначная разрешимость системы уравнений (16).

4. Метод ортогональной прогонки. Мы уже отмечали, что системы алгебраических уравнений, возникающие в методе сеток, имеют специальный вид. Так, например, система уравнений (16) имеет трехдиагональную матрицу, т. е. матрицу, у которой отличными от нуля являются только элементы главной диагонали и элементы двух соседних диагоналей, все же другие элементы матрицы равны нулю. В настоящее время разработан ряд численных алгоритмов, приспособленных к нахождению решений подобных систем. Одним из таких методов является метод ортогональной прогонки. Мы изложим этот метод на примере системы уравнений, подобной системе (16), но более общей.

Пусть задана система линейных алгебраических уравнений

$$\begin{aligned} a_k y_{k+1} + 2b_k y_k + c_k y_{k-1} &= h^2 d_k, \quad k=1, 2, \dots, N-1, \\ \sigma_0 y_0 + \sigma_1 y_1 &= \gamma, \quad \tau_0 y_{N-1} + \tau_1 y_N = \delta, \end{aligned} \quad (21)$$

где $a_k, b_k, c_k, d_k, \sigma_0, \sigma_1, \gamma, \tau_0, \tau_1, \delta$ — известные числа и $\sigma_0^2 + \sigma_1^2 = 1, \tau_0^2 + \tau_1^2 = 1, a_k \cdot c_k \neq 0, k=1, 2, \dots, N-1$. Предположим, что система уравнений (21) имеет единственное решение. Легко видеть, что система уравнений (10), (13), (14) и (16) является частным случаем системы (21).

Дадим вывод алгоритма *метода ортогональной прогонки*. Пусть $y_k, k=0, 1, \dots, N$, — искомое решение системы (21). Введем в рассмотрение новые неизвестные величины u_k, v_k , определяемые соотношениями

$$u_k = y_k \sin \alpha_k + y_{k-1} \cos \alpha_k, \quad (22)$$

$$v_k = y_k \cos \alpha_k - y_{k-1} \sin \alpha_k \quad (23)$$

и системой уравнений (21). Очевидно, что

$$y_k = u_k \sin \alpha_k + v_k \cos \alpha_k, \quad (24)$$

$$y_{k-1} = u_k \cos \alpha_k - v_k \sin \alpha_k. \quad (25)$$

Используя формулы (21)—(25), запишем разностные уравнения для сеточных функций u_k и v_k :

$$u_{k+1} = \frac{1}{\rho_k} \cdot \frac{c_k}{a_k} u_k + h^2 \frac{d_k}{a_k} \cdot \sin \alpha_{k+1}, \quad (26)$$

$$\begin{aligned} v_k &= \frac{1}{\rho_k} v_{k+1} + \frac{1}{\rho_k} \left[\sin \alpha_{k+1} \cdot \sin \alpha_k + \right. \\ &\quad \left. + \cos \alpha_{k+1} \left(\frac{c_k}{a_k} \cos \alpha_k + \frac{2b_k}{a_k} \cdot \sin \alpha_k \right) \right] u_k - \frac{\cos \alpha_{k+1}}{\rho_k} \cdot h^2 d_k, \end{aligned} \quad (27)$$

где

$$\rho_k = \left[\cos^2 \alpha_k + \left(\frac{c_k}{a_k} \sin \alpha_k - \frac{2b_k}{a_k} \cos \alpha_k \right)^2 \right]^{1/2}.$$

Величины $\sin \alpha_{k+1}, \cos \alpha_{k+1}$ определяются из условия

$$\cos \alpha_k \cdot \cos \alpha_{k+1} + \left(\frac{c_k}{a_k} \sin \alpha_k - \frac{2b_k}{a_k} \cos \alpha_k \right) \sin \alpha_{k+1} = 0. \quad (28)$$

Это условие выводится из требования, чтобы правая часть уравнения (26) не содержала значений сеточной

функции v_k . Из (28) выводятся следующие рекуррентные формулы:

$$\begin{aligned}\sin \alpha_{k+1} &= -\frac{\cos \alpha_k}{\rho_k}, \\ \cos \alpha_{k+1} &= \frac{\frac{c_k}{a_k} \sin \alpha_k - \frac{2b_k}{a_k} \cos \alpha_k}{\rho_k}.\end{aligned}\quad (29)$$

Заметим, что $\rho_k \neq 0$ при всех значениях k , ибо в противном случае нулю должен был бы равняться коэффициент c_k/a_k . Но это невозможно из-за предположения $a_k \cdot c_k \neq 0$.

Чтобы можно было вести вычисления по формулам (26), (27) и (29), надо вычислить начальные значения $\sin \alpha_1$, $\cos \alpha_1$, u_1 и v_N . С этой целью подставим в условие $\sigma_0 y_0 + \sigma_1 y_1 = \gamma$ значение y_0 , определяемое по формуле (25) при $k=1$, и y_1 , определяемое по формуле (24) при $k=1$. Получим

$$(\sigma_0 \sin \alpha_1 + \sigma_1 \cos \alpha_1) u_1 + (\sigma_0 \cos \alpha_1 - \sigma_1 \sin \alpha_1) v_1 = \gamma. \quad (30)$$

Поскольку мы намерены определить u_1 и v_N , то в формуле (30) v_1 можно считать произвольным. Чтобы исключить влияние v_1 на u_1 , положим в формуле (30)

$$\sigma_0 \cos \alpha_1 - \sigma_1 \sin \alpha_1 = 0.$$

Можно считать, что

$$\begin{aligned}\sin \alpha_1 &= \sigma_0, \\ \cos \alpha_1 &= \sigma_1.\end{aligned}\quad (31)$$

Теперь из (30) получим

$$u_1 = \gamma, \quad (32)$$

ибо по предположению $\sigma_0^2 + \sigma_1^2 = 1$.

Таким образом, используя формулы (29), (31) и (26), (32), можно последовательно вычислить

$$\begin{aligned}\sin \alpha_1, \quad \sin \alpha_2, \quad \dots, \quad \sin \alpha_N, \\ \cos \alpha_1, \quad \cos \alpha_2, \quad \dots, \quad \cos \alpha_N, \\ u_1, \quad u_2, \quad \dots, \quad u_N.\end{aligned}$$

Положим в формуле (22) $k=N$:

$$(\sin \alpha_N) y_N + (\cos \alpha_N) y_{N-1} = u_N. \quad (33)$$

Кроме соотношения (33), мы еще имеем второе граничное условие

$$\tau_1 y_N + \tau_0 y_{N-1} = \delta. \quad (34)$$

Пусть

$$\Delta = \tau_0 \sin \alpha_N - \tau_1 \cos \alpha_N \neq 0;$$

тогда из системы уравнений (33), (34) получим

$$\begin{aligned} y_{N-1} &= \frac{1}{\Delta} (\delta \sin \alpha_N - u_N \tau_1), \\ y_N &= \frac{1}{\Delta} (u_N \tau_0 - \delta \cos \alpha_N). \end{aligned}$$

Подставим эти значения в формулу (23), положив в ней $k = N$. Получим

$$v_N = \frac{1}{\Delta} (-\delta + (\tau_0 \cos \alpha_N + \tau_1 \sin \alpha_N) u_N). \quad (35)$$

Значения $u_N, u_{N-1}, \dots, u_1$ могут быть вычислены теперь по формулам (27) и (35) при $k = N-1, N-2, \dots, 1$.

Все изложенное выше позволяет сформулировать алгоритм метода ортогональной прогонки.

1. Вначале, используя формулы (29), (31) и (26), (32), последовательно вычисляем $\sin \alpha_k, \cos \alpha_k, u_k, k = 1, 2, \dots, N$.

2. Затем проверяем, выполняется ли условие $\Delta \neq 0$, и если оно выполняется, то по формулам (27), (35) вычисляем значения $v_N, v_{N-1}, \dots, v_1$.

3. По формуле (24) при $k = N$ и по формуле (25) при $k = 1$ вычисляем соответственно y_N и y_0 . Другие значения $y_1, y_2, \dots, y_{N-1}$ искомого решения системы уравнений (21) можно вычислять по любой из формул (24), (25).

Непосредственной проверкой можно убедиться, что определяемые таким образом значения $y_0, y_1, \dots, y_N$ удовлетворяют системе уравнений (21).

Укажем на следующие свойства метода ортогональной прогонки, важные в вычислительном смысле. Все формулы, по которым ведутся вычисления, являются рекуррентными, и потому они просто программируются. Вычисление величин $\sin \alpha_k, \cos \alpha_k$ благоприятно в смысле медленного накопления погрешностей округления из-за свойства $|\sin \alpha_k| \leq 1, |\cos \alpha_k| \leq 1$. Благоприятным в смысле

накопления погрешностей округления является и процесс вычисления величин u_k и v_k , так как при $\Delta \neq 0$ имеем

$$u_k^2 + v_k^2 \equiv y_k^2 + y_{k-1}^2.$$

Значит, сеточные функции u_k и v_k по своей величине такие же, как и искомое решение системы уравнений (21).

5. Оценка погрешности и сходимость метода сеток. Эти вопросы мы будем изучать на примере упрощенной задачи (1) — (3). Будем считать, что граничные условия (2), (3) заданы в форме

$$y(a) = A, \quad y(b) = B.$$

Учитывая (9) и (15), запишем сеточную задачу для точных значений граничной задачи (1)—(3) в следующем виде:

$$\begin{aligned} \frac{1}{h^2} \Lambda_h(y(x_k)) + R_k(h) &= f(x_k), \quad k = 1, 2, \dots, N-1, \\ y(x_0) &= A, \quad y(x_N) = B. \end{aligned} \quad (36)$$

Отметим, что в случае, когда решение граничной задачи (1) — (3) непрерывно и имеет непрерывные на $[a, b]$ производные до четвертого порядка включительно, для погрешности аппроксимации $R_k(h)$ верна оценка

$$|R_k(h)| \leq Mh^2, \quad k = 1, 2, \dots, N-1, \quad (37)$$

где M — постоянная, не зависящая от h .

Реально из-за наличия погрешностей округления сеточные значения будут вычисляться как решение системы линейных алгебраических уравнений, сходной с системой (15):

$$\begin{aligned} \frac{1}{h^2} \Lambda_h(y_k) &= f_k - \delta_k, \quad k = 1, 2, \dots, N-1, \\ y_0 &= A, \quad y_N = B, \end{aligned} \quad (38)$$

где через δ_k обозначены погрешности округлений. Будем считать, что для δ_k указана некоторая оценка сверху:

$$|\delta_k| \leq \delta, \quad k = 1, 2, \dots, N-1. \quad (39)$$

Введем следующие обозначения:

$$\varepsilon_k = y(x_k) - y_k,$$

$$\varepsilon(h) = \max_{0 \leq k \leq N} |\varepsilon_k|.$$

Метод сеток, определяемый формулой (38), называется равномерно сходящимся, если при $h \rightarrow 0$ выполняется условие: $\varepsilon(h) \rightarrow 0$.

Дадим оценку для $|\varepsilon_k|$ и установим условия сходимости метода сеток.

Используя формулы (36) и (38), получим

$$\Lambda_h(\varepsilon_k) = h^2(-R_k(h) + \delta_k), \quad k = 1, 2, \dots, N-1, \quad (40)$$

$$\varepsilon_0 = 0, \quad \varepsilon_N = 0.$$

Метод получения оценки для $|\varepsilon_k|$ будет заключаться в построении некоторой функции $u(x)$, которая будет мажорантой для $|\varepsilon_k|$, т. е. будет выполняться условие

$$|\varepsilon_k| \leq u(x_k), \quad u(x) \geq 0, \quad k = 0, 1, \dots, N.$$

Для построения мажоранты нам понадобится следующая лемма.

Пусть $t_0, t_1, \dots, t_N$ и $s_0, s_1, \dots, s_N$ — некоторые числа, причем $s_k \geq 0$, $k = 0, 1, \dots, N$.

Лемма 2. Пусть выполняются неравенства (18), (19) и

- 1) $\Lambda_h(s_k) \leq -|\Lambda_h(t_k)|$, $k = 1, 2, \dots, N-1$,
- 2) $|t_0| \leq s_0$, $|t_N| \leq s_N$.

Тогда

$$|t_k| \leq s_k, \quad k = 1, 2, \dots, N-1.$$

Доказательство. Рассмотрим сумму $s_k + t_k$. В силу первого условия леммы имеем

$$\Lambda_h(s_k + t_k) = \Lambda_h(s_k) + \Lambda_h(t_k) \leq 0.$$

По лемме 1 среди чисел $s_k + t_k$ наименьшее отрицательное значение могут иметь или $s_0 + t_0$, или $s_N + t_N$. Но по второму условию леммы эти числа неотрицательны, значит, все $s_k + t_k \geq 0$. Аналогично можно показать, что также и $s_k - t_k \geq 0$. Последние два неравенства означают, что $-s_k \leq t_k \leq s_k$ или $|t_k| \leq s_k$. Лемма доказана.

Рассмотрим вспомогательную задачу

$$\begin{aligned} L(v) &\equiv v'' + p(x) v' + q(x) v \equiv -1, \\ v(a) &= 0, \quad v(b) = 0, \end{aligned} \quad (41)$$

где $q(x) < 0$. Ее решение обозначим через $v(x)$. Покажем, что $v(x) > 0$ для всех $a < x < b$. Действительно, если бы функция $v(x)$ принимала на $[a, b]$ отрицательные значения, то существовала бы точка $\tilde{x}$ такая, что $v(\tilde{x}) < 0$, $v'(\tilde{x}) = 0$ и $v''(\tilde{x}) \geq 0$. Но тогда

$$L(v)|_{x=\tilde{x}} = v''(\tilde{x}) + p(\tilde{x}) v'(\tilde{x}) + q(\tilde{x}) \cdot v(\tilde{x}) > 0,$$

что противоречит (41), ибо $L(v(x)) = -1$ при всех $a < x < b$.

Определим $u(x)$ по формуле $u(x) = \alpha \cdot v(x)$, где α — положительное число, правило выбора которого мы укажем позже.

Имеем

$$\Lambda_h(v(x_k)) = h^2(-1 - R_h(v, h)),$$

где через $R(v, h)$ мы обозначили погрешность аппроксимации дифференциального оператора $L(v)$ разностным $\Lambda_h(v(x_k))$. Очевидно, что

$$|R_h(v, h)| \leq M^* h^2, \quad (42)$$

где M^* — постоянная, не зависящая от h .

Запишем сеточную задачу для $u(x)$:

$$\begin{aligned} \Lambda_h(u(x_k)) &= -\alpha h^2(1 + R_h(v, h)), \quad k = 1, 2, \dots, N-1, \\ u(x_0) &= 0, \quad u(x_N) = 0. \end{aligned} \quad (43)$$

Сравним задачи (40) и (43). Если нам удастся определить α так, чтобы было

$$\Lambda_h(u(x_k)) \leq -|\Lambda_h(\varepsilon_k)| \quad (44)$$

при $k = 1, 2, \dots, N-1$, то тогда в силу леммы 2 будет выполняться неравенство $|\varepsilon_k| \leq u(x_k)$. Используя оценку (37), имеем

$$|\Lambda_h(\varepsilon_k)| \leq h^2(Mh^2 + \delta). \quad (45)$$

Пусть h настолько мало, что $1 - M^*h^2 > 0$. Очевидно, что

$$\Lambda_h(u(x_k)) \leq -\alpha h^2(1 - M^*h^2). \quad (46)$$

Из (45), (46) следует, что неравенство (44) будет выполняться, если α выбрано из условия

$$-\alpha h^2 (1 - M^* h^2) \leq -h^2 (M h^2 + \delta).$$

Положим

$$\alpha = \frac{M h^2 + \delta}{1 - M^* h^2}.$$

Тогда

$$|\varepsilon_k| \leq u(x_k) = \alpha v(x_k) = \frac{h^2 M + \delta}{1 - M^* h^2} \cdot v(x_k), \quad k = 0, 1, \dots, N. \quad (47)$$

Оценка (47) является искомой оценкой погрешности метода сеток, определяемого формулой (38).

Пусть $V \geq \max_{a \leq x \leq b} v(x)$, тогда на основании (47) можно записать, что

$$\varepsilon(h) \leq \frac{h^2 M + \delta}{1 - M^* h^2} \cdot V. \quad (48)$$

Из этого неравенства видно, что $\varepsilon(h) \rightarrow 0$ при $h \rightarrow 0$ со скоростью $O(h^2)$, если $\delta \leq \delta_0 \cdot h^2$. Таким образом, установлена оценка погрешности (47), равномерная сходимость и скорость сходимости метода сеток.

Теорема 2. Пусть решение граничной задачи (1)–(3) с условиями частного вида $y(a) = A$, $y(b) = B$, единственно и непрерывно дифференцируемо на $[a, b]$ до четвертого порядка включительно. Если выполняются условия (18), (19), то метод сеток для рассматриваемой задачи будет равномерно сходящимся со скоростью $O(h^2)$, если $\delta \leq \delta_0 h^2$.

Теорема 2 имеет место для граничной задачи (1)–(3) и при более общих граничных условиях, в частности, она верна в случае, когда

$$l_a(y) \equiv \alpha_0 y(a) + \alpha_1 y'(a) = A, \quad l_b(y) \equiv \beta_0 y(b) + \beta_1 y'(b) = B$$

и $\alpha_0 \alpha_1 \leq 0$, $\beta_0 \beta_1 \geq 0$.

§ 12. Метод Галеркина и метод моментов

Решение ряда задач физики и механики иногда более выгодно находить в аналитической форме. Этим свойством обладают методы Галеркина, моментов, наименьших квадратов и др. По универсальности и простоте реали-

зации упомянутые методы часто уступают методу сеток и некоторым другим подобным методам. Вместе с тем они имеют и определенное преимущество перед методом сеток, поскольку позволяют находить решение в аналитической форме.

Рассмотрим методы Галеркина и моментов на примере граничной задачи для нелинейного дифференциального уравнения второго порядка.

Пусть при $a \leq x \leq b$ задано дифференциальное уравнение

$$F(x, y, y', y'') = 0 \quad (1)$$

с граничными условиями

$$l_a(y) \equiv \alpha_0 y(a) + \alpha_1 y'(a) = A, \quad (2)$$

$$l_b(y) \equiv \beta_0 y(b) + \beta_1 y'(b) = B. \quad (3)$$

Будем считать, что граничная задача (1)—(3) имеет на $[a, b]$ единственное решение и это решение непрерывно дифференцируемо до второго порядка включительно.

В методе моментов применительно к граничной задаче (1)—(3) рассматриваются две системы функций:

1) система функций $\{\psi_k(x)\}$, подчиненная условиям:

а) $\psi_k(x) \in C[a, b]$, $k = 0, 1, 2, \dots$;

б) функции $\psi_k(x)$ образуют замкнутую на $[a, b]$ систему;

2) система функций $\{\varphi_k(x)\}$, удовлетворяющая условиям:

а) $\varphi_k(x) \in C_2[a, b]$, $k = 0, 1, 2, \dots$;

б) при любом конечном n функции $\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x)$ линейно независимы на $[a, b]$;

в) функция $\varphi_0(x)$ удовлетворяет граничным условиям (2), (3), а функции $\varphi_k(x)$ удовлетворяют требованиям

$$l_a(\varphi_k) = 0, \quad l_b(\varphi_k) = 0, \quad k = 1, 2, \dots;$$

г) функции $\varphi_k(x)$ образуют в классе функций $C_2[a, b]$, удовлетворяющих условиям (2), (3), полную систему.

Сделаем некоторые пояснения. Систему функций $\{\psi_k(x)\}$ мы будем называть замкнутой на множестве интегрируемых на $[a, b]$ функций, если не существует такой непрерывной функции, кроме тождественного нуля, которая была бы ортогональной на $[a, b]$ ко всем

функциям $\psi_k(x)$. Таким образом, если система функций $\{\psi_k(x)\}$ замкнута, то из условия

$$\int_a^b f(x) \cdot \psi_k(x) dx = 0, \quad k = 0, 1, 2, \dots,$$

где $f(x)$ — непрерывная на $[a, b]$ функция, следует, что $f(x) \equiv 0$. Свойство полноты, о котором говорится в условии г), понимается так. Обозначим через G класс функций $y(x)$, принадлежащих $C_2[a, b]$ и удовлетворяющих граничным условиям (2), (3). Говорят, что система функций $\{\psi_k(x)\}$ полна в классе G , если для любого $\varepsilon > 0$ и любой функции $y(x) \in G$ можно указать такое n и такие параметры $a_1, a_2, \dots, a_n$, что имеет место неравенство

$$|y^{(i)}(x) - y_n^{(i)}(x)| < \varepsilon, \quad i = 0, 1, 2, \quad a \leq x \leq b,$$

где $y_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x)$. Это означает, что для любой допустимой функции $y(x) \in G$ найдется такая функция $y_n(x)$, которая на $[a, b]$ будет сколь угодно точно приближать функцию $y(x)$ вместе с ее производными $y'(x)$ и $y''(x)$.

Перейдем к построению приближенного решения граничной задачи (1)–(3) *методом моментов*.

Будем считать функцию $F(x, y, z, t)$ непрерывной по всем аргументам в области

$$\Omega = \{a \leq x \leq b, \quad -\infty < y, z, t < +\infty\}.$$

Составим линейную комбинацию из $n+1$ первых функций $\varphi_k(x)$:

$$y_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x), \quad (4)$$

где a_k — некоторые численные параметры, выбор которых будет указан ниже. Отметим, что функция $y_n(x)$ удовлетворяет граничным условиям (2), (3) при любых значениях a_k . Действительно, в силу условия в) и линейности граничных условий имеем

$$l_a(y_n) = l_a(\varphi_0) + \sum_{k=1}^n a_k l_a(\varphi_k) = A,$$

ибо $l_a(\varphi_0) = A$, а $l_a(\varphi_k) = 0$ при всех $k = 1, 2, \dots$. Аналогично можно показать, что и $l_b(y_n) = B$ при любых $a_1, a_2, \dots, a_n$.

Мы будем выбирать параметры $a_1, a_2, \dots, a_n$ таким образом, чтобы функция $y_n(x)$ в указываемом ниже смысле возможно лучше удовлетворяла дифференциальному уравнению (1). Подставив в $F(x, y, y', y'')$ вместо y, y', y'' соответственно выражения $y_n(x), y'_n(x), y''_n(x)$, получим

$$F(x, y_n, y'_n, y''_n) \equiv F_n(x).$$

Образуем обобщенные моменты функции $F_n(x)$:

$$\mu_k = \int_a^b F_n(x) \cdot \psi_k(x) dx, \quad k = 0, 1, \dots$$

По своему смыслу момент μ_k является некоторой функцией параметров $a_1, a_2, \dots, a_n$. Выберем их так, чтобы первые n моментов обратились в нуль:

$$\mu_k = 0 \quad \text{при} \quad k = 0, 1, \dots, n-1. \quad (5)$$

Эти равенства образуют систему n уравнений, из которой должны быть найдены коэффициенты $a_1, a_2, \dots, a_n$. Пусть мы решили систему (5) и нашли значения $a_1^*, a_2^*, \dots, a_n^*$. В методе моментов в качестве n -го приближения к решению граничной задачи (1)—(3) принимают функцию

$$y_n^*(x) = \varphi_0(x) + \sum_{k=1}^n a_k^* \varphi_k(x).$$

Почему функцию $y_n^*(x)$ можно рассматривать как приближенное решение задачи (1)—(3)? Поясним это наглядными соображениями. Если бы оказалось, что равенства $\mu_k = 0$ выполняются при всяких $k = 0, 1, 2, \dots$, то ввиду замкнутости системы функций $\{\psi_k(x)\}$ отсюда следовало бы, что

$$F_n^*(x) \equiv F_n(x, y_n^*, y_n^{*'}, y_n^{*''}) \equiv 0$$

и функция $y_n^*(x)$ была бы точным решением граничной задачи (1)—(3). Но в нашем распоряжении имеется только конечное число параметров $a_1, a_2, \dots, a_n$ и выполнить бесконечное число условий ($\mu_k = 0, k = 0, 1, 2, \dots$)

при помощи их выбора можно только в исключительных случаях. Например, в тех случаях, когда решение граничной задачи (1)–(3) является линейной комбинацией вида (4). Можно надеяться, вообще говоря, на выполнение лишь n условий $\mu_k = 0$, $k = 0, 1, \dots, n-1$. При этом можно ожидать, что если n — достаточно большое число, то функция $F_n^*(x)$ будет близкой к нулю: $F_n^*(x) = \varepsilon_n(x)$, $\max_{a \leq x \leq b} |\varepsilon_n(x)| < \varepsilon$ и $\varepsilon > 0$ — достаточно малое число. В этом случае функция $y_n^*(x)$ будет удовлетворять дифференциальному уравнению

$$F(x, y_n^*, y_n^{\prime}, y_n^{\prime\prime}) = \varepsilon_n(x), \quad (6)$$

которое мало отличается от заданного уравнения (1).

Можно ожидать, что для широкого класса функций F из условия (6) будет следовать достаточная малость на $[a, b]$ разности $|y(x) - y_n^*(x)|$, где $y(x)$ — решение граничной задачи (1)–(3).

Изложим алгоритм метода моментов применительно к граничной задаче (1)–(3):

1) в начале вычислений выбираем системы функций $\{\psi_k(x)\}$ и $\{\varphi_k(x)\}$, подчиняя их выбор сформулированным выше условиям;

2) затем решаем систему нелинейных численных уравнений (5) и выписываем ее решение $a_1^*, a_2^*, \dots, a_n^*$;

3) значения параметров a_k^* подставляем в формулу (4) и полученную функцию $y_n^*(x)$ принимаем за n -е приближение к искомому решению $y(x)$.

Как мы уже отмечали, о близости $y_n^*(x)$ к $y(x)$ во многих случаях можно судить по величине $\max_{a \leq x \leq b} |\varepsilon_n(x)|$, а иногда по абсолютной величине нескольких последних параметров: $a_s^*, a_{s+1}^*, \dots, a_n^*$.

Если уравнение (1) линейно, то запись системы уравнений для определения параметров a_k упростится.

Рассмотрим, например, уравнение

$$F(x, y, y', y'') \equiv L(y) - f(x) = 0,$$

где

$$L(y) = \frac{d}{dx} (p(x) y') - q(x) \cdot y, \quad p(x) \geq p_0 > 0,$$

$$p(x) \in C_1[a, b], \quad q(x) \geq 0, \quad a \leq x \leq b,$$

с граничными условиями $y(a) = A$, $y(b) = B$. В этом случае мы получим систему линейных алгебраических

уравнений вида

$$\sum_{k=1}^n C_{ki} a_k - D_i = 0, \quad i = 0, 1, \dots, n-1, \quad (7)$$

где

$$C_{ki} = \int_a^b L(\varphi_k) \cdot \psi_i(x) dx,$$

$$D_i = \int_a^b [f(x) - L(\varphi_0)] \psi_i(x) dx.$$

В методе Галеркина обе системы функций $\{\psi_k(x)\}$ и $\{\varphi_k(x)\}$ совпадают, поэтому система уравнений (5) примет следующую форму:

$$\int_a^b F(x, y_n, y'_n, y''_n) \varphi_k(x) dx = 0, \quad k = 1, 2, \dots, n. \quad (8)$$

Здесь мы считаем, что свойством замкнутости обладает система функций $\{\varphi_k(x)\}$, $k = 1, 2, \dots$. Если исходное дифференциальное уравнение было линейным самосопряженным, то в методе Галеркина система уравнений (7) в силу (8) примет вид

$$\sum_{k=1}^n C_{ki} a_k - D_i = 0, \quad i = 1, 2, \dots, n, \quad (9)$$

где

$$C_{ki} = \int_a^b L(\varphi_k) \cdot \varphi_i(x) dx,$$

$$D_i = \int_a^b [f(x) - L(\varphi_0)] \varphi_i(x) dx.$$

Запись коэффициентов C_{ki} можно упростить следующим образом:

$$C_{ki} = \int_a^b \left(\frac{d}{dx} (p(x) \varphi'_k) - q(x) \varphi_k(x) \right) \varphi_i(x) dx =$$

$$= p(x) \varphi'_k(x) \varphi_i(x) \Big|_a^b - \int_a^b (p(x) \varphi'_k(x) \varphi'_i(x) +$$

$$+ q(x) \varphi_k(x) \varphi_i(x)) dx = - \int_a^b (p(x) \varphi'_k(x) \varphi'_i(x) +$$

$$+ q(x) \varphi_k(x) \varphi_i(x)) dx,$$

ибо

$$\varphi'_k(a) \varphi_i(a) = 0 \quad \text{и} \quad \varphi'_k(b) \varphi_i(b) = 0.$$

Отметим, что $C_{hi} = C_{ik}$ и матрица системы уравнений (9) симметричная. Аналогично для D_i получим

$$D_i = \int_a^b (p(x) \varphi'_0(x) \varphi'_i(x) + q(x) \varphi_0(x) \varphi_i(x) + f(x) \varphi_i(x)) dx.$$

Покажем, что система уравнений (9) однозначно разрешима. Для этого мы установим, что соответствующая однородная система имеет только нулевое решение. Предположим противное и будем считать, что система уравнений

$$\sum_{k=1}^n C_{ki} b_k = 0, \quad i = 1, 2, \dots, n, \quad (10)$$

имеет ненулевое решение, т. е. $b_k \not\equiv 0$. Обозначим через

$$z_n(x) = b_1 \varphi_1(x) + b_2 \varphi_2(x) + \dots + b_n \varphi_n(x).$$

Отметим, что $z_n(x) \not\equiv 0$, ибо функции $\varphi_k(x)$ линейно независимы и, по предположению, $b_k \not\equiv 0$. Используя выражение для C_{ki} из (10), получим

$$\int_a^b (p(x) z'_n(x) \varphi'_i(x) + q(x) z_n(x) \varphi_i(x)) dx = 0, \quad (11)$$

$$i = 1, 2, \dots, n.$$

Умножим обе части соотношения (11) на b_i и просуммируем по i от 1 до n :

$$\int_a^b [p(x) (z'_n(x))^2 + q(x) z_n^2(x)] dx = 0.$$

Поскольку $p(x) \geq p_0 > 0$, то отсюда следует, что $z'_n(x) \equiv 0$ и, значит, $z_n(x) \equiv \text{const}$, но $z_n(a) = z_n(b) = 0$, поэтому $z_n(x) \equiv 0$. Мы пришли к противоречию. Наше предположение о том, что $b_k \not\equiv 0$ неверно. Значит, $b_k \equiv 0$, и, следовательно, система (9) однозначно разрешима.

Мы говорили, что функция $\varphi_0(x)$ должна удовлетворять условиям $l_a(\varphi_0) = A$, $l_b(\varphi_0) = B$. Для этого можно $\varphi_0(x)$ определить как многочлен первой степени $\varphi_0(x) =$

$= \alpha + \beta x$ и коэффициенты α и β выбрать как решение системы уравнений

$$\alpha_0 (\alpha + \beta a) + \alpha_1 \beta = A,$$

$$\beta_0 (\alpha + \beta b) + \beta_1 \beta = B.$$

Другие функции $\varphi_k(x)$ можно, например, в случае, когда $\alpha_1 = \beta_1 = 0$, $\alpha_0 = \beta_0 = 1$, вычислять по правилу

$$\varphi_k(x) = (x-a)^k (x-b), \quad k = 1, 2, \dots, n,$$

или

$$\varphi_k(x) = (x-a)(x-b)^k, \quad k = 1, 2, \dots, n.$$

Чтобы иметь представление о скорости сходимости метода Галеркина хотя бы для некоторых частных задач, приведем известный результат для следующей граничной задачи:

$$\begin{aligned} y'' - q(x)y &= f(x), \quad 0 \leq x \leq 1, \\ y(0) &= y(1) = 0. \end{aligned} \quad (12)$$

Пусть $\varphi_k(x) = \sin k\pi x$. Обозначим через $y_n^*(x) = \sum_{k=1}^n a_k^* \varphi_k(x)$ приближенное решение, полученное по методу Галеркина. Если

$$0 \leq q(x) \leq q; \quad f(x), \quad q(x), \quad \frac{f(x)}{\sqrt{q(x)}} \in L_2[0, 1],$$

то имеет место оценка

$$\|y(x) - y_n^*(x)\| \leq \left[\frac{q}{4(n+1)^2} + \frac{\sqrt{2}}{(n+1/2)^{3/2}} \right] \frac{\sqrt{q}}{\sqrt{3}\pi^2} \cdot \left\| \frac{f(x)}{\sqrt{q(x)}} \right\| = O(n^{-3/2}),$$

где через $y(x)$ обозначено решение граничной задачи (12) и

$$\left\| \frac{f(x)}{\sqrt{q(x)}} \right\| = \left(\int_0^1 \frac{f^2(x)}{q(x)} dx \right)^{1/2}.$$

Выше было дано изложение методов моментов и Галеркина, применительно к задаче (1)–(3) некоторого конкретного вида. Приведем краткое изложение схем методов моментов и Галеркина в общем случае для операторов, заданных в гильбертовом пространстве H .

Пусть линейный оператор A задан на множестве $D(A)$, всюду плотном в H со значениями, принадлежащими тому

же пространству H . Предположим, что пространство H обладает счетной и полной системой линейно независимых элементов.

Рассмотрим уравнение

$$Ay = f, \quad (13)$$

где y — искомый элемент из $D(A)$, подлежащий определению, f — некоторый элемент из H .

Выберем последовательность элементов $\{\varphi_k\}$, $\varphi_k \in D(A)$. Предположим, что любой элемент y из $D(A)$ однозначно представим сходящимся рядом

$$y = \sum_{k=1}^{\infty} \alpha_k \varphi_k,$$

где α_k — некоторые коэффициенты, зависящие от y . Обозначим через y_n n -е приближение к решению уравнения (13) и будем разыскивать это приближение в виде

$$y_n = \sum_{k=1}^n a_k \varphi_k, \quad (14)$$

где a_k — параметры, которые мы определим из условия ортогональности выражения $Ay_n - f$ к элементам $\varphi_1, \varphi_2, \dots, \varphi_n$. Будем считать, что имеют место соотношения

$$(Ay_n - f, \varphi_k) = 0, \quad k = 1, 2, \dots, n. \quad (15)$$

Подставляя сюда вместо y_n его выражение (14), получим систему линейных алгебраических уравнений

$$\sum_{k=1}^n A_{ki} a_k - B_i = 0, \quad i = 1, 2, \dots, n, \quad (16)$$

где

$$A_{ki} = (A\varphi_k, \varphi_i), \quad B_i = (f, \varphi_i).$$

Пусть система уравнений (16) однозначно разрешима и $a_1^*, a_2^*, \dots, a_n^*$ — ее решение. Тогда n -е приближение к элементу y — решению уравнения (13) в методе Галеркина вычисляют по формуле $y_n^* = \sum_{k=1}^n a_k^* \varphi_k$. Если бы выражение $Ay_n^* - f$ оказалось ортогональным ко всем элементам φ_k ($k = 1, 2, \dots, n, \dots$), то это бы означало, что

$Ay_n^* - f = 0$ и y_n^* было бы точным решением уравнения (13). Если же условие ортогональности (15) выполнено для достаточно больших n , то можно надеяться, что и y_n^* близко к точному решению уравнения (13).

В методе моментов параметры a_k определяются по-иному. Требуют выполнения условий ортогональности элемента $Ay_n - f$ к элементам последовательности $\{\psi_k\}$

$$(Ay_n - f, \psi_k) = 0, \quad k = 1, 2, \dots, n. \quad (17)$$

При этом предполагается, что элементы $\psi_1, \psi_2, \dots, \psi_n$ линейно независимы при любом конечном n и последовательность $\{\psi_k\}$ полна в H . Формула (17) есть система линейных алгебраических уравнений вида

$$\sum_{k=1}^n C_{ki} a_k - D_i = 0, \quad i = 1, 2, \dots, n, \quad (18)$$

где

$$C_{ki} = (A\varphi_k, \psi_i), \quad D_i = (f, \psi_i).$$

Пусть $a_1^*, a_2^*, \dots, a_n^*$ — единственное решение системы уравнений (18). В методе моментов, как и в методе Галеркина, n -е приближение к решению уравнения (13) вычисляется по формуле

$$y_n^* = \sum_{k=1}^n a_k^* \varphi_k.$$

Легко видеть, что при $\psi_k \equiv \varphi_k$ метод моментов совпадает с методом Галеркина. Отметим еще, что оба эти метода применимы и к решению нелинейных уравнений.

§ 13. Метод наименьших квадратов и метод Ритца

Рассмотрим алгоритм метода наименьших квадратов применительно к операторным уравнениям.

Пусть $L_2[a, b]$ — пространство вещественных функций, интегрируемых с квадратом на $[a, b]$, и A — линейный оператор со значениями из $L_2[a, b]$. Пусть A определен на множестве $D(A)$, всюду плотном в $L_2[a, b]$.

Предположим, что требуется найти решение уравнения

$$Ay = f, \quad (1)$$

где $f \in H$, $y \in D(A)$. Будем считать, что уравнение (1) имеет единственное решение.

Поставим в соответствие уравнению (1) функционал

$$J(y) = \|Ay - f\|^2. \quad (2)$$

Здесь знаком $\| \cdot \|$ помечена норма в пространстве $L_2[a, b]$. Если в L_2 определено скалярное произведение и z — некоторый элемент из L_2 , то

$$\|z\| = \sqrt{(z, z)} = \left[\int_a^b z^2(x) dx \right]^{1/2}.$$

Задачу отыскания решения уравнения (1) заменим эквивалентной задачей об определении элемента, минимизирующего в $D(A)$ функционал (2). Очевидно, что

$$\min_{y \in D(A)} J(y) = J(y^*) = 0$$

и элемент y^* является решением уравнения $Ay = f$. Метод решения уравнения (1), основанный на минимизации функционала (2), получил название *метода наименьших квадратов*.

Функционал $J(y)$ можно минимизировать различными способами. Можно это выполнить так. Выбираем последовательность линейно независимых элементов $\{\varphi_k\}$, $\varphi_k \in D(A)$. Определяем элемент y_n — n -е приближение к решению уравнения (1) по формуле

$$y_n = \sum_{k=1}^n a_k \varphi_k, \quad (3)$$

где a_k — некоторые параметры. Они определяются из условия, чтобы функционал $J(y_n) = \|Ay_n - f\|^2$ принимал минимальное значение на подпространстве $D_n(A) \in D(A)$, натянутом на элементы $\varphi_1, \varphi_2, \dots, \varphi_n$. Это требование приводит к системе линейных алгебраических уравнений для определения $a_1, a_2, \dots, a_n$:

$$\frac{\partial J(y_n)}{\partial a_k} = 0, \quad k = 1, 2, \dots, n. \quad (4)$$

Если $a_1^*, a_2^*, \dots, a_n^*$ — решение системы уравнений (4), то n -е приближение к решению y дается формулой (3) при подстановке в нее вместо a_k величин a_k^* .

Дадим более подробное описание метода наименьших квадратов на примере линейной граничной задачи для дифференциального уравнения второго порядка

$$L(y) \equiv y'' + p(x)y' + q(x)y = f(x) \quad (5)$$

с граничными условиями

$$l_a(y) \equiv \alpha_0 y(a) + \alpha_1 y'(a) = A, \quad (6)$$

$$l_b(y) \equiv \beta_0 y(b) + \beta_1 y'(b) = B. \quad (7)$$

Будем считать, что задача (5)—(7) имеет единственное решение и это решение вместе со своими производными до второго порядка включительно непрерывно на $[a, b]$.

В методе наименьших квадратов применительно к граничной задаче (5)—(7) рассматривается система функций $\{\varphi_k(x)\}$, которая выбирается так же, как это мы делали в методе моментов (см. стр. 121).

Приближенное решение задачи (5)—(7) ищется в виде

$$y_n(x) = \varphi_0(x) + \sum_{k=1}^n \alpha_k \varphi_k(x), \quad (8)$$

где α_k — некоторые параметры. Функция $y_n(x)$ удовлетворяет граничным условиям (6), (7) при любых значениях параметров α_k . Выберем их так, чтобы $y_n(x)$ возможно лучше удовлетворяла уравнению (5). Подставим $y_n(x)$ в уравнение (5), получим

$$L(y_n) - f(x) = \varepsilon_n(x).$$

Почти всегда $\varepsilon_n(x) \neq 0$. Мы заинтересованы в том, чтобы погрешность $\varepsilon_n(x)$ по абсолютной величине была возможно меньшей. Введем в рассмотрение величину δ_n^2 , определив ее по правилу

$$\delta_n^2 = \int_a^b \varepsilon_n^2(x) dx.$$

Параметры $\alpha_1, \alpha_2, \dots, \alpha_n$ выберем так, чтобы величина δ_n^2 принимала наименьшее значение.

Заметим, что $\delta_n^2 = J(y_n)$, если $J(y)$ определять по правилу $J(y) = \|L(y) - f\|^2 = (L(y) - f, L(y) - f)$. Здесь скалярное произведение имеет следующий смысл:

$$(g, h) = \int_a^b g(x) h(x) dx$$

для любых функций $g(x)$ и $h(x)$, принадлежащих $L_2[a, b]$.

Имеем

$$L(y_n) = L(\varphi_0) + \sum_{k=1}^n a_k L(\varphi_k),$$

$$\delta_n^2 = \int_a^b \left[\sum_{k=1}^n a_k L(\varphi_k) - f_0 \right]^2 dx,$$

где $f_0 = f - L(\varphi_0)$.

Для определения искомых значений параметров a_k записываем систему линейных алгебраических уравнений

$$\frac{1}{2} \cdot \frac{\partial \delta_n^2}{\partial a_i} = \int_a^b \left[\sum_{k=1}^n a_k L(\varphi_k) - f_0 \right] L(\varphi_i) dx = 0,$$

$$i = 1, 2, \dots, n.$$

Эту запись можно упростить. Имеем

$$\sum_{k=1}^n A_{ik} a_k - B_i = 0, \quad i = 1, 2, \dots, n, \quad (9)$$

где

$$A_{ik} = A_{ki} = \int_a^b L(\varphi_k) \cdot L(\varphi_i) dx,$$

$$B_i = \int_a^b f_0(x) \cdot L(\varphi_i) dx.$$

Отметим, что матрица системы уравнений (9) симметрична и ее определитель есть определитель Грама для функций $L(\varphi_1), L(\varphi_2), \dots, L(\varphi_n)$. Предположим, что система уравнений (9) однозначно разрешима и $a_1^*, a_2^*, \dots, a_n^*$ — ее решение. Тогда n -е приближение к решению $y(x)$ задачи (5)—(7) вычисляется по формуле (8), если в нее вместо a_k подставлены значения a_k^* .

Выясним, когда можно надеяться на однозначную разрешимость системы уравнений (9). Ответ на этот вопрос зависит не только от свойств системы функций $\{\varphi_k(x)\}$, но также и от природы рассматриваемой граничной задачи (5)—(7), в частности от того, имеет ли однородная граничная задача

$$L(z) = 0, \quad l_a(z) = 0, \quad l_b(z) = 0 \quad (10)$$

только нулевое решение. Будем далее предполагать, что

$\alpha_1 = \beta_1 = 0$, $\alpha_0 = \beta_0 = 1$. Пусть $D = (A_{ik})_1^n$ — матрица системы уравнений (9), а $|D|$ — ее определитель.

Теорема 1. *Если граничная задача (10) имеет только нулевое решение $z(x) \equiv 0$, то $|D| \neq 0$ и система уравнений (9) имеет единственное решение.*

Доказательство. Рассмотрим однородную систему уравнений

$$\sum_{k=1}^n A_{ik} b_k = 0, \quad i = 1, 2, \dots, n.$$

Докажем, что эта система имеет только нулевое решение $b_k \equiv 0$. Умножим уравнение номера i на b_i и просуммируем полученные выражения по i от 1 до n . Имеем

$$\sum_{i=1}^n b_i \sum_{k=1}^n A_{ik} b_k = \int_a^b \sum_{i=1}^n \sum_{k=1}^n b_i L(\varphi_i) \cdot b_k L(\varphi_k) dx = 0. \quad (11)$$

Введем обозначение

$$z_n(x) = \sum_{k=1}^n b_k \varphi_k(x).$$

Теперь из соотношения (11) получим

$$\int_a^b L^2(z_n) dx = 0.$$

Отсюда следует, что $L(z_n) \equiv 0$. В силу выбора функций $\varphi_k(x)$ имеем: $z_n(a) = z_n(b) = 0$. Следовательно, $l_a(z_n) = 0$, $l_b(z_n) = 0$. Таким образом, функция $z_n(x)$ удовлетворяет граничной задаче (10) и по условию теоремы должно быть

$z_n(x) \equiv 0$ или $\sum_{k=1}^n b_k \varphi_k(x) \equiv 0$. Поскольку функции $\varphi_1(x)$,

$\varphi_2(x), \dots, \varphi_n(x)$ линейно независимы, то из условия $z_n(x) \equiv 0$ следует, что $b_k \equiv 0$. Значит, $|D| \neq 0$ и система уравнений (9) имеет единственное решение. Теорема доказана.

Изучим вопрос о сходимости метода наименьших квадратов на примере граничной задачи (5)—(7) в частном случае, когда

$$l_a(y) \equiv y(a) = 0, \quad l_b(y) \equiv y(b) = 0. \quad (12)$$

Пусть $y^*(x)$ — точное решение задачи (5), (12), а $y_n^*(x)$ n -е приближение к $y^*(x)$, получаемое по методу наименьших квадратов. Заметим, что в случае граничных условий (12)

$$\varphi_0(x) \equiv 0 \text{ и } y_n^*(x) = \sum_{k=1}^n a_k^* \varphi_k(x).$$

Теорема 2. *Последовательность функций $\{y_n^*(x)\}$, получаемых по методу наименьших квадратов, сходится в метрике L_2 к точному решению $y^*(x)$, если выполняются следующие условия:*

1) *граничная задача (5), (12) имеет единственное решение $y^*(x)$;*

2) *существует такая постоянная M , что для любой дважды непрерывно дифференцируемой функции $y(x)$, обращающейся в нуль на концах отрезка $[a, b]$, имеет место неравенство*

$$\|L(y)\| \geq \frac{1}{M} \left(\int_a^b y^2(x) dx \right)^{1/2}.$$

Доказательство. В силу условия 1) и теоремы 1 последовательность решений $\{y_n^*(x)\}$ может быть построена. Последовательность $\{\varphi_k(x)\}$ полна в классе $C_2[a, b]$, поэтому для любого $\varepsilon > 0$ можно указать такое N и такие параметры $a_1, a_2, \dots, a_N$, что

$$\left\| L(y^*) - \sum_{k=1}^N a_k L(\varphi_k) \right\| < \frac{\varepsilon}{M}.$$

Это неравенство можно переписать в виде

$$\|f - L(y_N)\| < \frac{\varepsilon}{M}, \quad (13)$$

так как

$$L(y^*) = f, \quad \sum_{k=1}^N a_k L(\varphi_k) = L\left(\sum_{k=1}^N a_k \varphi_k\right) = L(y_N).$$

Неравенство (13) тем более будет выполняться, если в нем функцию $y_N(x)$ заменить на функцию $y_N^*(x)$, ибо последняя строится из условия минимизации квадратичной формы $J(y_N) = \|L(y_N) - f\|^2$ на множестве параметров $a_1, a_2, \dots, a_N$. Значит,

$$\|L(y^*) - L(y_N^*)\| < \frac{\varepsilon}{M}.$$

Очевидно неравенство

$$\|L(y^*) - L(y_n^*)\| = \|L(y^* - y_n^*)\| < \frac{\varepsilon}{M}.$$

Отсюда, используя условие 2) теоремы 2, получим

$$\|y^* - y_n^*\| \leq M \|L(y^* - y_n^*)\| < \varepsilon. \quad (14)$$

Поскольку ε может быть любым сколь угодно малым числом, то из (14) следует, что

$$\lim_{n \rightarrow \infty} \|y^* - y_n^*\| = 0.$$

Теорема доказана.

Неравенство (14) можно переписать в виде

$$\|y^* - y_n^*\| \leq M \|f - L(y_n^*)\|.$$

По этому неравенству можно судить о том, с какой скоростью $y_n^*(x)$ сходится при $n \rightarrow \infty$ к точному решению $y^*(x)$.

Изложим теперь вычислительную схему метода Ритца на частном примере граничной задачи для уравнения

$$(p(x)y')' - q(x)y = f(x) \quad (15)$$

с условиями

$$y(a) = A, \quad y(b) = B, \quad (16)$$

где $p(x) \geq p_0 > 0$, $p(x) \in C_1[a, b]$, $q(x) \geq 0$, $q(x)$ и $f(x)$ принадлежат классу $C[a, b]$. Будем предполагать, что задача (15), (16) имеет единственное решение.

Рассмотрим функционал

$$J(y) = \int_a^b [p(x)y'^2 + q(x)y^2 + 2f(x)y] dx \quad (17)$$

и попытаемся найти его минимальное значение на классе допустимых функций, т. е. функций, удовлетворяющих двум условиям:

- 1) эти функции непрерывно дифференцируемы на $[a, b]$;
- 2) они удовлетворяют граничным условиям (16).

Теорема 3. Если функция $y^*(x)$ доставляет минимальное значение функционалу $J(y)$ среди всех допустимых функций, то она является решением граничной задачи (15), (16).

Доказательство. Так как функция $y^*(x)$ является допустимой, то она удовлетворяет граничному условию (16) и остается доказать, что она удовлетворяет уравнению (15). Для этой цели воспользуемся одной из теорем вариационного исчисления. Пусть рассматривается функционал вида $G(y) = \int_a^b F(x, y, y') dx$ и требуется в

упомянутом выше классе допустимых функций найти наименьшее значение $G(y)$. Доказывается, что при достаточной гладкости F минимизирующая функция должна удовлетворять дифференциальному уравнению Эйлера

$$\frac{d}{dx} F_{y'} - F_y = 0.$$

В вариационной задаче $F \equiv p(x) y'^2 + q(x) y^2 + 2f(x) y$. Для нее $F_{y'} = 2p(x) y'$, $F_y = 2q(x) y + 2f(x)$. Поэтому уравнение Эйлера для минимизирующей функции $y^*(x)$ будет иметь вид

$$2(p(x) y^{*\prime})' - 2q(x) y^* - 2f(x) = 0,$$

или

$$(p(x) y^{*\prime})' - q(x) y^* = f(x),$$

что и требовалось доказать.

Таким образом, задачу вычисления решения $y^*(x)$ граничной задачи (15), (16) можно заменить задачей отыскания минимума функционала (17) в классе непрерывно дифференцируемых на $[a, b]$ функций, принимающих заданные значения $y(a) = A$ и $y(b) = B$ на концах отрезка. Эта вторая задача в вычислительном смысле во многих случаях оказывается предпочтительнее первой.

В методе Ритца n -е приближение к минимизирующей функции $y^*(x)$ вычисляется в виде

$$y_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x), \quad (18)$$

где a_k — численные параметры, $\varphi_k(x)$ — некоторые известные функции. Выбор этих функций подчиняют условиям

- 1) $\varphi_k(x) \in C_1[a, b]$, $k = 0, 1, 2, \dots$;
- 2) $\varphi_0(a) = A$, $\varphi_0(b) = B$, $\varphi_k(a) = \varphi_k(b) = 0$, $k = 1, 2, \dots$;

3) при любом конечном n функции $\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x)$ линейно независимы;

4) образованное по функциям $\varphi_k(x)$ семейство функций $\{y_n(x)\}$ обладает свойством C_1 -полноты.

Отметим, что при любом выборе параметров a_k функция $y_n(x) \in C_1[a, b]$, $y_n(a) = A$, $y_n(b) = B$. Подставив $y_n(x)$ в функционал $J(y)$, получим

$$J(y_n) = A_0 + 2 \sum_{k=1}^n A_k a_k + \sum_{i=1}^n \sum_{k=1}^n A_{ik} a_i a_k,$$

где

$$A_0 = \int_a^b [p(x) \varphi_0'^2 + q(x) \varphi_0^2 + 2f(x) \varphi_0] dx,$$

$$A_k = \int_a^b [p(x) \varphi_0' \varphi_k' + q(x) \varphi_0 \varphi_k + f(x) \varphi_k] dx,$$

$$A_{ik} = A_{ki} = \int_a^b [p(x) \varphi_k' \varphi_i' + q(x) \varphi_k \varphi_i] dx,$$

$$i, k = 1, 2, \dots, n.$$

Параметры $a_1, a_2, \dots, a_n$ определяем из системы линейных алгебраических уравнений

$$\frac{1}{2} \frac{\partial J(y_n)}{\partial a_k} \equiv \sum_{i=1}^n A_{ik} a_i + A_k = 0, \quad k = 1, 2, \dots, n. \quad (19)$$

Система уравнений (19) имеет единственное решение, это можно показать подобно тому, как доказывалась однозначная разрешимость системы уравнений (9).

Пусть $a_1^*, a_2^*, \dots, a_n^*$ — решение системы уравнений (19), тогда вычисляемая в методе Ритца функция $y_n^*(x) = \varphi_0(x) + \sum_{k=1}^n a_k^* \varphi_k(x)$ будет n -м приближением к решению $y^*(x)$ граничной задачи (15), (16).

Введем следующее обозначение:

$$\lim_{n \rightarrow \infty} J(y_n^*) = m = J(y^*). \quad (20)$$

Последовательность функций $\{y_n^*\}$ называется *минимизи-*

рующей, если она удовлетворяет условию (20). Отметим, что из (20) не всегда следует, что

$$\lim_{n \rightarrow \infty} y_n^*(x) = y^*(x).$$

То есть из того, что последовательность является минимизирующей, не всегда следует ее сходимость. Сходимость имеет место только при выполнении определенных условий.

Теорема 4. Пусть выполняются условия:

$$1) \quad p(x) \geq p_0 > 0, \quad p(x) \in C_1[a, b],$$

$$2) \quad q(x) \geq 0 \text{ и } q(x), f(x) \in C[a, b],$$

3) последовательность функций $\{y_n^*(x)\}$ является минимизирующей. Тогда эта последовательность функций будет сходящейся равномерно на $[a, b]$ к $y^*(x)$ — решению граничной задачи (15) — (16).

Доказательство. Имеем

$$|y^*(x) - y_n^*(x)| = \left| \int_a^x [y^{*'}(t) - y_n^{*'}(t)] dt \right| \leq \int_a^b |y^{*'}(t) - y_n^{*'}(t)| dt. \quad (21)$$

Используя неравенство Буняковского, получим

$$\int_a^b |y^{*'}(t) - y_n^{*'}(t)| dt \leq \sqrt{b-a} \left[\int_a^b (y^{*'}(t) - y_n^{*'}(t))^2 dt \right]^{1/2}. \quad (22)$$

Для интеграла в правой части неравенства (22) очевидным образом можно продолжить следующую цепочку неравенств:

$$\begin{aligned} \sqrt{b-a} \left[\int_a^b (y^{*'}(t) - y_n^{*'}(t))^2 dt \right]^{1/2} &\leq \\ &\leq \left(\frac{b-a}{p_0} \right)^{1/2} \left[\int_a^b p(t) (y^{*'}(t) - y_n^{*'}(t))^2 dt \right]^{1/2} \leq \\ &\leq \left(\frac{b-a}{p_0} \right)^{1/2} \left[\int_a^b [p(t) (y^{*'}(t) - y_n^{*'}(t))^2 + \right. \\ &\quad \left. + q(x) (y^*(t) - y_n^*(t))^2] dt \right]^{1/2}. \end{aligned} \quad (23)$$

Установим связь между последним интегралом из (23) и разностью $J(y_n^*) - J(y^*)$. Имеем

$$\begin{aligned} J(y_n^*) - J(y^*) &= \\ &= \int_a^b [p(x) y_n^{*'}^2 + q(x) y_n^{*2} + 2f(x) y_n^* - p(x) y^{*'}^2 - q(x) y^{*2} - \\ &\quad - 2f(x) y^*] dx = \int_a^b [p(x) (y_n^{*'} - y^{*'})^2 + q(x) (y_n^* - y^*)^2] dx + \\ &\quad + 2 \int_a^b [p(x) y_n^{*'} y^{*'} + q(x) y_n^* y^* - p(x) y^{*'}^2 - q y^{*2} + \\ &\quad + f(y_n^* - y^*)] dx = \int_a^b [p(x) (y_n^{*'} - y^{*'})^2 + q(x) (y_n^* - y^*)^2] dx + \\ &\quad + 2 \int_a^b [p(x) y^{*'} (y_n^{*'} - y^{*'}) + q(x) y^* (y_n^* - y^*) + f(y_n^* - y^*)] dx. \end{aligned}$$

Здесь второй интеграл равен нулю в силу уравнения (15). Чтобы в этом убедиться, достаточно воспользоваться интегрированием по частям и учесть, что $y_n^*(a) - y^*(a) = y_n^*(b) - y^*(b) = 0$. Таким образом,

$$J(y_n^*) - J(y^*) = \int_a^b [p(x) (y_n^{*'} - y^{*'})^2 + q(x) (y_n^* - y^*)^2] dx. \quad (24)$$

Теперь на основании неравенств (21) — (23) и равенства (24) получим

$$|y^*(x) - y_n^*(x)| \leq \left(\frac{b-a}{\rho_0} \right)^{1/2} \cdot (J(y_n^*) - J(y^*))^{1/2}. \quad (25)$$

В оценке (25) правая часть от x не зависит. Так как по условию теоремы $J(y_n^*) - J(y^*) \rightarrow 0$ при $n \rightarrow \infty$, то из неравенства (25) следует, что минимизирующая последовательность функций $\{y_n^*(x)\}$ сходится равномерно относительно x к функции $y^*(x)$ на отрезке $[a, b]$. Теорема доказана.

ЛИТЕРАТУРА

1. Бахвалов Н. С., Численные методы, т. I, М., «Наука», 1973.
2. Березин И. С., Жидков Н. П., Методы вычислений, т. 2, М., «Наука», 1962.
3. Крылов В. И., Бобков В. В., Монастырный П. И., Вычислительные методы высшей математики, т. 2, Минск, «Вышэйшая школа», 1975.
4. Ланс Д. Н., Численные методы для быстродействующих вычислительных машин, М., «Наука», 1962.
5. Мысовских И. П., Лекции по методам вычислений, М., «Наука», 1962.

ГЛАВА 7

РЕШЕНИЕ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ

§ 1. Разностные схемы, основные понятия

Дифференциальные уравнения в частных производных имеют широкие приложения в математической физике, гидродинамике, акустике и других областях знаний. В большинстве своем такие уравнения в явном виде не решаются. Поэтому широкое распространение получили методы приближенного их решения, в частности метод сеток.

Мы уделим основное внимание вопросу решения линейных дифференциальных уравнений второго порядка.

Построение различных схем метода сеток в случае уравнений в частных производных зависит от типа уравнений и вида граничных условий, присоединяемых к уравнениям.

В связи с этим сделаем несколько замечаний о классификации линейных дифференциальных уравнений второго порядка.

Пусть D — некоторая область изменения независимых переменных x, y , ограниченная контуром Γ . Говорят, что в области D задано линейное дифференциальное уравнение второго порядка для функции $u(x, y)$, если для любой точки из области D имеет место соотношение

$$L(u) \equiv a(x, y) \frac{\partial^2 u}{\partial x^2} + 2b(x, y) \frac{\partial^2 u}{\partial x \partial y} + c(x, y) \frac{\partial^2 u}{\partial y^2} + \\ + 2d(x, y) \frac{\partial u}{\partial x} + 2e(x, y) \frac{\partial u}{\partial y} + g(x, y) u = f(x, y), \quad (1)$$

где $a(x, y), b(x, y), \dots$ — коэффициенты, $f(x, y)$ — свободный член уравнения. Эти функции известны и их обычно считают

определенными в замкнутой области $\bar{D} = D + \Gamma$. Обозначим: $\delta(x, y) = b^2(x, y) - a(x, y)c(x, y)$. Уравнение $L(u) = f$ называют эллиптическим, параболическим или гиперболическим в области D , если соответственно выполняются условия

$$\delta(x, y) < 0, \quad \delta(x, y) = 0, \quad \delta(x, y) > 0$$

для всех (x, y) из области D .

В зависимости от типа дифференциального уравнения по-разному ставятся граничные и начальные условия, связанные с этим уравнением. Конкретно вид граничных и начальных условий мы рассмотрим при изучении схем метода сеток для соответствующих уравнений.

Выше мы привели общее линейное дифференциальное уравнение второго порядка. Построение и исследование разностных схем непосредственно на примере этого уравнения связано с большими записями и вычислениями. Поэтому далее мы будем рассматривать некоторые частные случаи:

1) уравнение Пуассона (эллиптическое уравнение)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y);$$

2) уравнение теплопроводности (параболическое уравнение)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t);$$

3) волновое уравнение (гиперболическое уравнение)

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y).$$

Дадим некоторые основные понятия и определения, необходимые для изучения элементов теории разностных схем.

1. Сходимость и аппроксимация разностных схем.

Пусть u есть решение дифференциального уравнения

$$L(u) = f, \quad (2)$$

заданного в некоторой области D , ограниченной контуром Γ .

Рассмотрим некоторое множество $D_h = \{M_h\}$, состоящее из изолированных точек M_h , принадлежащих области

$\bar{D} = D + \Gamma$. Число точек в множестве D_h будем характеризовать величиной h ; чем меньше h , тем большим будет число точек в D_h . Множество D_h называется *сеткой*, а точки M_h из этого множества — *узлами сетки*. Функция, определенная в узлах сетки, называется *сеточной функцией*.

Обозначим через U пространство непрерывных в D функций $v(x, y)$. Через U_h обозначим пространство, образованное совокупностью сеточных функций $v_h(x, y)$, определенных на D_h . В методе сеток осуществляется замена пространства U на пространство U_h .

Пусть $u(x, y)$ — точное решение уравнения (2) и $u(x, y) \in U$. Поставим задачу: вычислить значения $u_h(x, y)$. Эти значения в совокупности образуют таблицу, в которой число значений равно числу точек в D_h . Точно поставленную задачу удастся решить лишь в немногих случаях. Как правило, можно вычислить некоторые сеточные значения $u^{(h)}$, относительно которых при выполнении дополнительных условий (о которых мы еще скажем) можно утверждать, что

$$u^{(h)} \approx u_h(x, y).$$

Величины $u^{(h)}$ называются приближенными сеточными значениями решения $u(x, y)$. Для их вычисления строят систему численных уравнений, которую мы будем записывать в виде

$$L_h(u^{(h)}) = f^{(h)}, \quad (3)$$

где L_h есть разностный оператор, соответствующий оператору L , $f^{(h)} \in F_h$. Если $f(x, y) \in F$, то F_h образуется по F аналогично тому, как U_h образовывалось по U .

Формулу (3) будем называть *разностной схемой*.

Пусть в линейных пространствах U_h и F_h введены соответственно нормы $\|\cdot\|_{U_h}$ и $\|\cdot\|_{F_h}$, которые являются сеточными аналогами норм $\|\cdot\|_U$ и $\|\cdot\|_F$ в исходных пространствах U и F .

Будем говорить, что разностная схема (3) является сходящейся, если при $h \rightarrow 0$ выполняется условие

$$\|u_h(x, y) - u^{(h)}\|_{U_h} \rightarrow 0.$$

Если выполняется условие

$$\|u_h(x, y) - u^{(h)}\|_{U_h} \leq Ch^s,$$

где C — постоянная, не зависящая от h , $s > 0$, то говорят, что сходимость имеет место со скоростью порядка s относительно h .

Говорят, что разностная схема (3) аппроксимирует задачу (2) на решении $u(x, y)$, если

$$L_h(u_h(x, y)) = f^{(h)} + \delta f^{(h)}$$

и

$$\|\delta f^{(h)}\|_{F_h} \rightarrow 0 \quad \text{при} \quad h \rightarrow 0.$$

Величина $\delta f^{(h)}$ называется погрешностью аппроксимации.

Если $\|\delta f^{(h)}\|_{F_h} \leq Mh^\sigma$, где M — постоянная, не зависящая от h , $\sigma > 0$, то говорят, что разностная схема (3) аппроксимирует задачу (2) на решении $u(x, y)$ с погрешностью порядка σ относительно h .

Отметим, что смысл понятий сходимости и аппроксимации разностной схемы в случае линейных обыкновенных дифференциальных уравнений второго порядка нами уже частично обсуждался в § 11 гл. 6.

Введем важное понятие устойчивости разностной схемы (3).

Пусть L_h — линейный разностный оператор.

Разностная схема (3) называется *устойчивой*, если существует такое $h_0 > 0$, что для всех $h < h_0$ и любых $f^{(h)} \in F_h$ выполняются условия:

- 1) разностная схема (3) имеет единственное решение;
- 2) $\|u^{(h)}\|_{U_h} \leq M \|f^{(h)}\|_{F_h}$, где M — постоянная, не зависящая от h и $f^{(h)}$.

Из этого определения видно, что устойчивость разностной схемы существенным образом связывается с понятием норм, вводимых в пространствах U_h и F_h . Возможны такие случаи, когда условие 2) будет выполняться для некоторых норм и не будет выполняться для других норм. В каждом отдельном случае важно выяснить, по какой причине для рассматриваемой разностной схемы условие 2) не выполняется. Если оно не выполняется из-за того, что мы неудачно ввели определение норм в пространствах U_h и F_h , то следует попробовать другие способы введения норм в U_h и F_h с тем, чтобы добиться выполнения условия 2). Но если 2) не выполняется ни при каком (!) разумном выборе норм, то это будет означать, что разностная схема неустойчива.

Отметим, что устойчивость является внутренним свойством разностной схемы, и поэтому на устойчивость можно влиять, изменяя некоторые параметры, входящие в разностную схему, или вообще видоизменяя всю разностную схему.

2. Связь аппроксимации и устойчивости со сходимостью.

Между введенными выше понятиями сходимости, аппроксимации и устойчивости существует связь. Она состоит в том, что из аппроксимации и устойчивости следует сходимость,

Теорема 1. Пусть разностная схема $L_h(u^{(h)}) = f^{(h)}$ аппроксимирует задачу $L(u) = f$ на решение $u(x, y)$ с порядком $s > 0$ относительно h и устойчива. Тогда эта схема будет сходящейся и порядок ее сходимости будет совпадать с порядком аппроксимации, т. е. будет справедлива оценка

$$\|u_h(x, y) - u^{(h)}\|_{U_h} \leq Kh^s, \quad (4)$$

где K — постоянная, не зависящая от h .

Доказательство. По определению аппроксимации имеем

$$\|\delta f^{(h)}\|_{F_h} \leq Ch^s, \text{ где } \delta f^{(h)} = L_h(u_h(x, y)) - f^{(h)}.$$

Обозначим: $\varepsilon_h(x, y) = u_h(x, y) - u^{(h)}$. Легко видеть, что в силу линейности L_h для $\varepsilon_h(x, y)$ имеет место формула

$$L_h(\varepsilon_h(x, y)) = \delta f^{(h)}.$$

Отсюда, используя определение устойчивости, получим

$$\|\varepsilon_h(x, y)\|_{U_h} \leq M \|\delta f^{(h)}\|_{F_h} \leq M(Ch^s) = Kh^s,$$

где $K = MC$. Таким образом, оценка (4) установлена и теорема доказана.

Теперь можно высказать некоторые общие замечания, показывающие, как следует поступать при изучении вопросов построения и исследования разностных схем.

1. Вначале указывается правило выбора сетки, т. е. указывается правило замены области D и контура Γ некоторой сеточной областью. Чаще всего сетка выбирается прямоугольной и равномерной.

2. Потом указывается и строится конкретно одна или несколько разностных схем. Проверяется условие аппрок-

симации разностных схем и устанавливается порядок аппроксимации.

3. Доказывается устойчивость построенных разностных схем. Это один из наиболее важных и сложных вопросов. Если разностные схемы обладают аппроксимацией и устойчивостью, то о сходимости разностных схем судят по доказанной выше теореме.

4. Рассматривается вопрос численного решения разностных схем. В случае линейных разностных схем это будет система линейных алгебраических уравнений. Уже в двумерном случае порядок таких систем может быть очень большим. Это делает задачу численного решения упомянутых систем во многих случаях весьма трудной. Поэтому для решения систем уравнений, возникающих в методе сеток, разработаны и разрабатываются специальные методы решения, учитывающие особенности таких задач.

Ниже при рассмотрении вопросов численного решения параболических, эллиптических, гиперболических уравнений мы раскроем конкретное содержание введенных понятий и определений. При этом мы всюду будем стремиться следовать при изучении вопросов построения и исследования разностных схем по пути, указанному в пп. 1 — 4.

§. 2. Разностные схемы для уравнений параболического типа

В этом параграфе мы рассмотрим вопросы построения и исследования разностных схем, возникающих при численном решении параболических уравнений по методу сеток.

1. Решение задачи Коши. Рассмотрим задачу Коши для уравнения теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \varphi(x, t), \quad -\infty < x < +\infty, \quad t > 0, \quad (1)$$

с условием на прямой $t = 0$

$$u(x, 0) = \psi(x), \quad -\infty < x < +\infty. \quad (2)$$

Требуется найти функцию $u(x, t)$, которая при $t > 0$ и $-\infty < x < +\infty$ удовлетворяла бы уравнению (1), а при $t = 0$ выполняла бы условие (2). Будем считать, что за-

дача (1), (2) имеет в верхней полуплоскости единственное решение $u(x, t)$, непрерывное вместе со своими производными $\frac{\partial^i u}{\partial t^i}$ $i=1, 2$, и $\frac{\partial^k u}{\partial x^k}$, $k=1, 2, 3, 4$. Запишем задачу (1), (2) в виде $L(u)=f$. Для этого достаточно положить

$$L(u) \equiv \begin{cases} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}, & -\infty < x < +\infty, \quad t > 0, \\ u(x, 0), & -\infty < x < +\infty, \end{cases}$$

$$f \equiv \begin{cases} \varphi(x, t), & -\infty < x < +\infty, \quad t > 0, \\ \psi(x), & -\infty < x < +\infty, \quad t = 0. \end{cases}$$

Будем далее считать, что t изменяется в пределах $0 \leq t \leq T < +\infty$.

В рассматриваемом случае $D = \{-\infty < x < +\infty, 0 < t \leq T\}$, Γ — объединение прямых $t=0$ и $t=T$.

Выберем прямоугольную сетку и заменим область $\bar{D} = D + \Gamma$ сеточной областью D_h . К области D_h отнесем совокупность точек (узлов) (x_m, t_n) , координаты которых определяются по правилу

$$x_m = mh, \quad m = 0, \pm 1, \pm 2, \dots, \quad h > 0, \quad t_n = n\tau, \\ n = 0, 1, \dots, N, \quad \tau > 0, \quad N\tau \leq T < (N+1)\tau.$$

Заменим задачу $L(u)=f$ разностной схемой вида $L_h(u^{(h)})=f^{(h)}$. Обозначим через $u(x_m, t_n)$ точное значение решения задачи $L(u)=f$ в узле (x_m, t_n) , через u_m^n — соответствующее приближенное сеточное значение. Имеем

$$L(u)|_{(x_m, t_n)} \equiv \begin{cases} \frac{\partial u}{\partial t} \Big|_{(x_m, t_n)} - \frac{\partial^2 u}{\partial x^2} \Big|_{(x_m, t_n)}, \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 1, 2, \dots, N, \\ u(x, 0)|_{(x_m, t_n)}, \end{cases}$$

$$f|_{(x_m, t_n)} \equiv \begin{cases} \varphi(x, t)|_{(x_m, t_n)}, \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 1, 2, \dots, N, \\ \psi(x)|_{(x_m, t_n)}. \end{cases}$$

Для замены выражений $\frac{\partial u}{\partial t} \Big|_{(x_m, t_n)}$ и $\frac{\partial^2 u}{\partial x^2} \Big|_{(x_m, t_n)}$ разност-

ными отношениями воспользуемся формулами численного дифференцирования. Имеем

$$\frac{\partial u}{\partial t} \Big|_{(x_m, t_n)} = \frac{u(x_m, t_{n+1}) - u(x_m, t_n)}{\tau} - \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x_m, t_n^{(1)})}, \quad (3)$$

$$t_n < t_n^{(1)} < t_{n+1},$$

$$\frac{\partial u}{\partial t} \Big|_{(x_m, t_n)} = \frac{u(x_m, t_n) - u(x_m, t_{n-1})}{\tau} + \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x_m, t_n^{(2)})}, \quad (4)$$

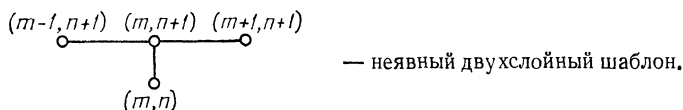
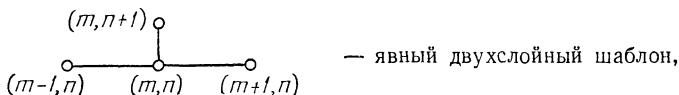
$$t_{n-1} < t_n^{(2)} < t_n,$$

$$\frac{\partial u}{\partial t} \Big|_{(x_m, t_n)} = \frac{u(x_m, t_{n+1}) - u(x_m, t_{n-1})}{2\tau} - \frac{\tau^2}{6} \frac{\partial^3 u}{\partial t^3} \Big|_{(x_m, t_n^{(3)})}, \quad (5)$$

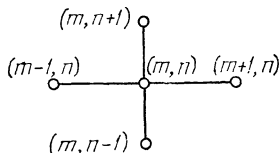
$$t_{n-1} < t_n^{(3)} < t_{n+1},$$

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} \Big|_{(x_m, t_n)} &= \\ &= \frac{u(x_{m+1}, t_n) - 2u(x_m, t_n) + u(x_{m-1}, t_n))}{h^2} - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(x_m^{(1)}, t_n)}, \quad (6) \\ &x_{m-1} < x_m^{(1)} < x_{m+1}. \end{aligned}$$

Назовем некоторую совокупность узлов, привлекаемых для замены задачи $L(u) = f$ в узле (x_m, t_n) разностной схемой $L_h(u^{(h)}) = f^{(h)}$, шаблоном. Приведем наиболее употребляемые для параболических уравнений шаблоны:



Явный трехслойный шаблон



некоторое время применялся в вычислительной практике.

Однако в настоящее время он не используется, так как оказалось, что соответствующая ему разностная схема неустойчива при любом выборе шага h — по пространству и шага τ — по времени.

Рассмотрим явный двухслойный шаблон. Имеем

$$L(u)|_{(x_m, t_n)} \equiv \begin{cases} \frac{u(x_m, t_{n+1}) - u(x_m, t_n)}{\tau} - \\ - \frac{u(x_{m+1}, t_n) - 2u(x_m, t_n) + u(x_{m-1}, t_n))}{h^2} + r_{mn}^{(1)}(h), \\ u(x_m, 0) + 0. \end{cases} \quad (7)$$

Здесь мы воспользовались формулами (3) и (6) и обозначили

$$r_{mn}^{(1)}(h) = -\frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x_m, t_n^{(1)})} - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(x_m^{(1)}, t_n)}.$$

Введем следующее обозначение:

$$f^{(h)} \equiv \begin{cases} \Phi(x_m, t_n), \\ \Psi(x_m). \end{cases} \quad (8)$$

Теперь на основании формул (7), (8) можно записать разностную схему для задачи $L(u) = f$:

$$L_h^{(1)}(u^{(h)}) = f^{(h)}, \quad (9)$$

где разностный оператор $L_h^{(1)}$ определяется по правилу

$$L_h^{(1)}(u^{(h)}) \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \\ u_m^0, \\ m = 0, \pm 1, \pm 2, \dots, \\ n = 0, 1, \dots, N-1. \end{cases}$$

Аналогично, если использовать неявный двухслойный шаблон, можно получить такую разностную схему:

$$L_h^{(2)}(u^{(h)}) = f^{(h)}, \quad (10)$$

где

$$L_h^{(2)}(u^{(h)}) \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2}, \\ u_m^0, \\ m = 0, \pm 1, \pm 2, \dots, \\ n = 0, 1, \dots, N-1, \\ f^{(h)} \equiv \begin{cases} \varphi(x_m, t_n), \\ \psi(x_m). \end{cases} \end{cases}$$

На основании формул (7) и (9) можно записать:

$$L_h^{(1)}(u_h(x, y)) = f^{(h)} + \delta^{(1)}f^{(h)},$$

где

$$\delta^{(1)}f^{(h)} \equiv \begin{cases} r_{mn}^{(1)}(h), \\ 0. \end{cases}$$

Аналогично, используя формулы (4), (6), (10), можно получить

$$L_h^{(2)}(u_h(x, y)) = f^{(h)} + \delta^{(2)}f^{(h)},$$

где

$$\delta^{(2)}f^{(h)} = \begin{cases} r_{mn}^{(2)}(h) = \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x_m, t_n^{(2)})} - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(x_m^{(1)}, t_{n+1})}, \\ 0. \end{cases}$$

Выясним порядок аппроксимации разностных схем (9) и (10). В качестве F_h возьмем линейное множество всех

пар ограниченных функций $g^{(h)} = \begin{cases} \alpha_m^n, \\ \beta_m. \end{cases}$ Норму в F_h определим по правилу

$$\|g^{(h)}\| = \max_{m, n} |\alpha_m^n| + \max_m |\beta_m|.$$

Если $\max_{m, n} |\alpha_m^n|$ или $\max_m |\beta_m|$ не достигается, то будем под нормой понимать величину $\|g^{(h)}\| = \sup_{m, n} |\alpha_m^n| + \sup_m |\beta_m|$.

Пусть $\tau = rh^s$, r, s — некоторые положительные числа. Предположим, что для $\frac{\partial^2 u}{\partial t^2}$ и $\frac{\partial^4 u}{\partial x^4}$ верны оценки:

$$\max_{(x, t) \in D} \left| \frac{\partial^2 u}{\partial t^2} \right| \leq M_2, \quad \max_{(x, t)} \left| \frac{\partial^4 u}{\partial x^4} \right| \leq M_4.$$

Тогда легко получить

$$\|\delta^{(1)}f^{(h)}\|_{F_h} = \max_{m, n} |r_{mn}^{(1)}(h)| \leq \left(\frac{r}{2} M_2 + \frac{h^{2-s}}{12} M_4\right) \cdot h^s, \quad (11)$$

$$\|\delta^{(2)}f^{(h)}\| = \max_{m, n} |r_{mn}^{(2)}(h)| \leq \left(\frac{r}{2} M_2 + \frac{h^{2-s}}{12} M_4\right) \cdot h^s. \quad (12)$$

Для параболических уравнений, как мы увидим далее, в случае схемы (9) можно взять $s=2$, в случае схемы (10) s можно взять равным единице.

Из формул (11), (12) следует, что разностные схемы (9), (10) аппроксимируют задачу $L(u)=f$ на $u(x, y)$ с погрешностью порядка s ($1 \leq s \leq 2$) относительно h .

Разностная схема (9) позволяет по значениям решения на нулевом слое, т. е. по значениям u_m^0 , $m=0, \pm 1, \pm 2, \dots$, вычислить значения на первом слое u_m^1 , $m=0, \pm 1, \pm 2, \dots$. Для этого достаточно в (9) положить $n=0$ и произвести вычисления, носящие рекурсионный характер. Потом по значениям u_m^1 можно аналогично при $n=1$ вычислить значения u_m^2 и т. д. В силу таких вычислительных свойств разностную схему (9) называют *явной*.

Разностная схема (10) упомянутыми выше свойствами не обладает. Действительно, если мы в (10) положим $n=0$, то в левой части полученной формулы будет линейная комбинация из значений $u_{m-1}^1, u_m^1, u_{m+1}^1, u_m^0$, в правой части будут значения известной функции $\varphi(x_m, 0)$ и $\psi(x_m)$. Для вычисления значений на первом слое $\dots, u_{-2}^1, u_{-1}^1, u_0^1, u_1^1, u_2^1, \dots$ в этом случае необходимо решать бесконечную систему линейных уравнений. По этой причине разностную схему (10) называют *неявной*.

2. Об устойчивости двухслойных разностных схем.

Определим норму в пространстве U_h по правилу

$$\|u^{(h)}\|_{U_h} = \max_{m, n} |u_m^n|.$$

Рассмотрим разностную схему (9). Выясним при каких значениях $r, \tau=rh^2$, возможна устойчивость этой схемы.

Для доказательства устойчивости надо показать, что разностная схема однозначно разрешима и при любых

$$g^{(h)} = \begin{cases} \alpha_m^n, \\ \beta_m, \end{cases} \quad g^{(h)} \in F_h \text{ имеет место оценка}$$

$$\|z^{(h)}\|_{U_h} \leq M \|g^{(h)}\|_{F_h},$$

где обозначено $M = \max(1, T)$; $M = 1$, если $T < 1$, и $M = T$, если $T \geq 1$.

На основании (16) можно записать

$$\max_{m, n} |z_m^n| \leq M \cdot \|g^{(h)}\|_{F_h} \quad \text{или} \quad \|z^{(h)}\|_{U_h} \leq M \cdot \|g^{(h)}\|_{F_h}.$$

Таким образом, разностная схема (9) при выполнении условия (14) на τ и h устойчива. Отметим, что условие (14) налагает жесткие ограничения на выбор τ , а именно, должно быть $\tau \leq \frac{1}{2} h^2$. Это приводит к тому, что если мы желаем сохранить устойчивость, то при вычислениях по схеме (9) шаг по времени τ приходится выбирать очень малым. Если бы мы рассматривали уравнение теплопроводности с переменным коэффициентом

$$\frac{\partial u}{\partial t} = a(x, t) \cdot \frac{\partial^2 u}{\partial x^2} + \varphi(x, t), \quad a(x, t) > 0,$$

то достаточное условие устойчивости, даваемое формулой (14), видоизменилось бы и, как это легко видеть, приняло бы форму

$$r = \frac{\tau}{h^2} \leq \frac{1}{2 \max_{(x, t) \in D} |a(x, t)|}. \quad (17)$$

Обратимся теперь к разностной схеме (10) и перепишем ее в виде

$$-r(u_{m+1}^{n+1} + u_{m-1}^{n+1}) + (1 + 2r) u_m^{n+1} = u_m^n + \tau \varphi(x_m, t_n), \quad (18)$$

$$u_m^0 = \psi(x_m),$$

$$m = 0, \pm 1, \pm 2, \dots, \quad n = 0, 1, \dots, N-1.$$

Посмотрим, какие надо проделать вычисления, чтобы, используя формулы (18), можно было вычислить, например, значения u_m^1 на первом временном слое по значениям u_m^0 ($m = 0, \pm 1, \pm 2, \dots$) на нулевом временном слое. Положив в формулах (18) $n = 0$, получим

$$-r(u_{m+1}^1 + u_{m-1}^1) + (1 + 2r) u_m^1 = u_m^0 + \tau \varphi(x_m, 0), \quad (19)$$

$$u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \dots$$

Формулы (19) представляют собой бесконечную систему линейных уравнений относительно неизвестных $\dots u_{-1}^1$,

$u_0^1, u_1^1, \dots$ Решение таких систем является сложной и трудоемкой задачей, поэтому разностные схемы (10) неудобны для задач Коши на бесконечных отрезках и применяются редко. Однако если отрезок оси x , на котором рассматривается задача Коши, конечен: $a \leq x \leq b$, $b - a \leq K$, а на прямых $x = a$ и $x = b$ дополнительно заданы некоторые ограничения на решение $u(x, t)$, то разностные схемы вида (10) оказываются весьма эффективными. В частности, как это будет показано в п. 4, такие схемы являются абсолютно устойчивыми, т. е. устойчивыми при любых значениях $r = \tau/h^2 > 0$.

Если, например, на отрезках прямых $x = a$ и $x = b$ заданы условия $u(a, t) = \gamma_0(t)$, $u(b, t) = \gamma_1(t)$, то вид системы (19) существенно изменится:

$$\begin{aligned} -r(u_{m+1}^1 + u_{m-1}^1) + (1 + 2r)u_m^1 &= \psi(x_m) + \tau\varphi(x_m, 0), \\ u_0^1 &= \gamma_0(t_1), \quad u_M^1 = \gamma_1(t_1), \\ m &= 1, 2, \dots, M-1, \quad h = \frac{b-a}{M}. \end{aligned} \quad (20)$$

Формулы (20) представляют собой систему $M+1$ линейных алгебраических уравнений относительно $u_0^1, u_1^1, \dots, u_M^1$. Эту систему можно решить по методу прогонки (см. § 11 гл. 6).

Проведенные рассуждения показывают, что реализация неявных разностных схем требует больших вычислительных затрат для вычисления решения на одном временном слое, но таких временных слоев может быть немного из-за того, что в этом случае отсутствуют ограничения на соотношение τ/h^2 . Если пользоваться явной разностной схемой, то вычисление решения на следующем слое осуществляется по рекурсионному правилу и связано с минимальными вычислительными затратами, однако из-за ограничения $\tau/h^2 \leq 1/2$ число временных слоев в случае явных схем может быть существенно большим по сравнению с числом временных слоев для неявных схем.

Выскажем здесь еще суждение о сходимости разностной схемы (9). Эта схема аппроксимирует задачу (1), (2) на решении $u(x, t)$ с погрешностью порядка $O(\tau + h^2)$ и устойчива при $r \leq 1/2$. Поэтому схема (9) будет сходящейся в силу теоремы 1 из § 1; при этом погрешность для приближенного решения будет величиной порядка $O(\tau + h^2)$.

3. О разностных схемах расщепления. В тех случаях, когда число пространственных переменных в задачах теплопроводности больше или равно двум, сильно возрастает объем вычислительной работы, которую необходимо выполнить при численной реализации разностных схем. Были предприняты попытки поиска таких схем, которые специально предназначены для решения многомерных задач и которые наряду со свойством устойчивости обладают свойством минимальности объема вычислений. Таковыми схемами являются, в частности, *схемы расщепления*. Рассмотрим правило построения схемы расщепления в случае двумерного уравнения теплопроводности.

Рассмотрим задачу Коши

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < t \leq T, \\ u(x, y, 0) &= \varphi(x, y), \quad -\infty < x, y < \infty, \end{aligned} \quad (21)$$

где $\varphi(x, y)$ — известная функция.

Обозначим $Au \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$. Выберем квадратную сетку по пространству с шагом $h > 0$ по направлению x и по направлению y . По времени выберем шаг τ из условия $N\tau \leq T < (N+1)\tau$. Таким образом, координаты некоторого узла (x_m, y_n, t_p) будут определяться по формулам

$$\begin{aligned} x_m &= mh, \quad y_n = nh, \quad t_p = p\tau, \\ m, n &= 0, \pm 1, \pm 2, \dots; \quad p = 0, 1, \dots, N. \end{aligned}$$

Предположим, что нам известно решение задачи (21) в момент времени $t = t_p$, т. е. известны значения функции $u(x, y, t_p)$.

В дальнейшем значения функции $u(x, y, t)$ при некотором $t = t_k$ будем обозначать через u^k .

Поставим задачу: выразить значение u^{p+1} через u^p и значение оператора A .

Имеем

$$\begin{aligned} u(x, y, t_{p+1}) &= u(x, y, t_p + \tau) = \\ &= u(x, y, t_p) + \frac{\tau}{1!} \frac{\partial u(x, y, t_p)}{\partial t} + O(\tau^2) = \\ &= u(x, y, t_p) + \frac{\tau}{1!} Au(x, y, t_p) + O(\tau^2) = \\ &= (E + \tau A) u(x, y, t_p) + O(\tau^2). \end{aligned}$$

Положим $A_1 \equiv \frac{\partial^2}{\partial x^2}$, $A_2 \equiv \frac{\partial^2}{\partial y^2}$, тогда $A = A_1 + A_2$.

Наряду с задачей (21) рассмотрим две вспомогательные задачи

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, \quad t_p < t \leq t_{p+1}, \quad (22)$$

$$v(x, y, t_p) = u(x, y, t_p), \quad -\infty < x, y < +\infty;$$

$$\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial y^2}, \quad t_p < t \leq t_{p+1}, \quad (23)$$

$$w(x, y, t_p) = v(x, y, t_{p+1}), \quad -\infty < x, y < +\infty.$$

Заметим, что задачи (22) и (23), в отличие от задачи (21), одномерные. Их можно решать последовательно, сначала задачу (22) (это позволит вычислить функцию $v(x, y, t_{p+1})$), затем задачу (23) (это позволит вычислить функцию $w(x, y, t_{p+1})$). Какова связь между значением $u(x, y, t_{p+1})$ и значением $w(x, y, t_{p+1})$?

Имеем

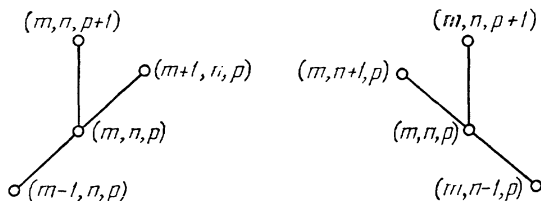
$$v(x, y, t_{p+1}) = (E + \tau A_1) u(x, y, t_p) + O(\tau^2),$$

$$\begin{aligned} w(x, y, t_{p+1}) &= (E + \tau A_2) w(x, y, t_p) + O(\tau^2) = \\ &= (E + \tau A_2)(x, y, t_{p+1}) + O(\tau^2) = \\ &= (E + \tau A_2)[(E + \tau A_1) u(x, y, t_p) + O(\tau^2)] = \\ &= (E + \tau A) u(x, y, t_p) + O(\tau^2) = u(x, y, t_{p+1}) + O(\tau^2). \end{aligned}$$

Таким образом, если мы последовательно решим задачи (22) и (23), то при $t = t_{p+1}$ получим значения функции $w(x, y, t_{p+1})$, которые отличаются от истинных значений $u(x, y, t_{p+1})$ — решения задачи (21) лишь на величину $O(\tau^2)$. Но при этом нам удалось заменить процесс решения двумерной задачи (21) процессом решения двух одномерных задач (22) и (23). В смысле объема вычислений второй процесс имеет преимущества перед первым и это делает изложенную здесь операцию расщепления выгодной.

Построим теперь две схемы расщепления для задачи (21), заменяя задачи (22) и (23) подходящими разностными аппроксимациями.

Возьмем два явных двухслойных шаблона видов



и, используя их соответственно для задач (22) и (23), запишем разностные схемы:

$$\frac{v_{mn}^{p+1} - v_{mn}^p}{\tau} = \Lambda_{xx}(v_{mn}^p), \quad (24)$$

$$v_{mn}^p = u_{mn}^p, \quad m, n = 0, \pm 1, \dots, \quad p = 0, 1, \dots, N-1,$$

$$\frac{w_{mn}^{p+1} - w_{mn}^p}{\tau} = \Lambda_{yy}(w_{mn}^p), \quad (25)$$

$$w_{mn}^p = v_{mn}^{p+1}, \quad m, n = 0, \pm 1, \dots, \quad p = 0, 1, \dots, N-1,$$

где приняты следующие обозначения:

$$\Lambda_{xx}(v_{mn}^p) = \frac{v_{m+1n}^p - 2v_{mn}^p + v_{m-1n}^p}{h^2},$$

$$\Lambda_{yy}(w_{mn}^p) = \frac{w_{mn+1}^p - 2w_{mn}^p + w_{mn-1}^p}{h^2}.$$

Формулы (24), (25) образуют в совокупности явную разностную схему расщепления для задачи (21). Численная реализация этой схемы осуществляется так. При $p=0$ значения u_{mn}^0 известны, поэтому в (24) при $p=0$ имеем $v_{mn}^0 = u_{mn}^0 = \varphi(x_m, y_n)$. Значения v_{mn}^{p+1} при $p=0$ вычисляем по явной формуле

$$v_{mn}^1 = \varphi(x_m, y_n) + \tau \Lambda_{xx}(v_{mn}^0).$$

Полагаем в (25) $p=0$ и вычисляем значение w_{mn}^1 по правилу

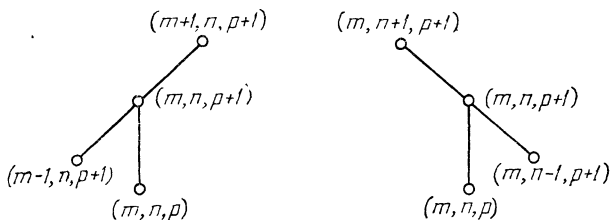
$$w_{mn}^1 = v_{mn}^1 + \tau \Lambda_{yy}(v_{mn}^1).$$

Таким образом, мы вычислим значения w_{mn}^1 ($m, n = 0, \pm 1, \dots$) на первом временном слое. Далее полагаем

$u_{mn}^1 \approx w_{mn}^1$, и это означает, что нами определены приближенные значения решения задачи (21) на первом временном слое. Аналогично по u_{mn}^1 вычисляем u_{mn}^2 и т. д., пока не вычислим значения u_{mn}^N на последнем временном слое.

Как мы видели, для устойчивости вычислений схемы (24) и (25) требуют жестких ограничений на шаг τ : $\tau \leq \leq \frac{1}{2} h^2$.

Рассмотрим теперь два неявных двухслойных шаблона видов



и по ним запишем для задач (22) и (23) разностные схемы

$$\frac{v_{mn}^{p+1} - v_{mn}^p}{\tau} = \Lambda_{xx}(v_{mn}^{p+1}), \quad (26)$$

$$v_{mn}^p = u_{mn}^p, \quad m, n = 0, \pm 1, \dots, \quad p = 0, 1, \dots, N-1,$$

$$\frac{w_{mn}^{p+1} - w_{mn}^p}{\tau} = \Lambda_{yy}(w_{mn}^{p+1}), \quad (27)$$

$$w_{mn}^p = v_{mn}^{p+1}, \quad m, n = 0, \pm 1, \dots, \quad p = 0, 1, \dots, N-1.$$

Формулы (26) и (27) образуют в совокупности неявную разностную схему расщепления для задачи (21). Численная реализация этой схемы расщепления осуществляется следующим образом. Полагаем в (26) $p=0$ и вычисляем значение v_{mn}^1 как решение системы линейных алгебраических уравнений

$$v_{mn}^1 - \tau \Lambda_{xx}(v_{mn}^1) = \varphi(x_m, y_n), \quad m, n = 0, \pm 1, \dots,$$

ибо $v_{mn}^0 = u_{mn}^0 = \varphi(x_m, y_n)$. Затем полагаем в (26) $p=0$ и находим w_{mn}^1 как решение системы линейных алгебраических уравнений

$$w_{mn}^1 - \tau \Lambda_{yy}(w_{mn}^1) = v_{mn}^1, \quad m, n = 0, \pm 1, \dots,$$

где v_{mn}^1 уже вычислено на первом этапе. Полагаем $u_{mn}^1 \approx w_{mn}^1$. Значит, мы вычислили приближенные значения u_{mn}^1 решения задачи (21) на первом временном слое. Далее по аналогии вычисляем $u_{mn}^2, u_{mn}^3, \dots, u_{mn}^N$. Решение упомянутых выше систем линейных алгебраических уравнений после их усечения *) можно осуществить по какому-либо алгоритму разностной прогонки.

Для устойчивости схем (26) и (27) никаких ограничений на шаги τ и h не требуется.

4. Решение смешанных граничных задач. Наряду с задачами Коши вида (21) для уравнений параболического типа рассматриваются также и смешанные граничные задачи, состоящие в том, что требуется определить функцию $u(x, t)$, которая в области $D = \{0 \leq x \leq 1, 0 < t \leq T\}$ удовлетворяет уравнению

$$\frac{\partial u}{\partial t} = \frac{\partial^2}{\partial x^2} + \varphi(x, t), \quad (28)$$

на начальной прямой при $t=0$ условию

$$u(x, 0) = \psi(x), \quad 0 \leq x \leq 1, \quad (29)$$

а на границе области D при $x=0$ и $x=1$ условиям

$$\left[\alpha_0(t) \cdot \frac{\partial u}{\partial x} + \beta_0(t) u \right]_{x=0} = \gamma_0(t), \quad (30)$$

$$\left[\alpha_1(t) \cdot \frac{\partial u}{\partial x} + \beta_1(t) u \right]_{x=1} = \gamma_1(t). \quad (31)$$

Здесь предполагаются известными функции $\varphi(x, t)$, $\psi(x)$ и $\alpha_i(t)$, $\beta_i(t)$, $\gamma_i(t)$, $i=0, 1$, причем $\alpha_i^2(t) + \beta_i^2(t) > 0$. Будем считать, что задача (28)—(31) имеет единственное решение и это решение непрерывно в D вместе со своими производными $\frac{\partial^k u}{\partial t^k}$, $k=1, 2$, $\frac{\partial^l u}{\partial x^l}$, $l=1, 2, 3, 4$.

Выберем прямоугольную равномерную сетку, определив в ней узлы по правилу

$$x_m = mh, \quad m=0, 1, \dots, M, \quad h=1/M > 0,$$

$$t_n = n\tau, \quad n=0, 1, \dots, N, \quad \tau > 0, \quad N\tau \leq T < (N+1)\tau.$$

*) С. Г. Михлин, Прямые методы в математической физике, М.—Л., 1950, § 51, стр. 303.

Способы замены уравнения (28) и начального условия (29) мы рассматривали выше (см. разностные схемы (9) и (10)). Укажем замену граничных условий (30) и (31) сеточными условиями. Используя формулы, аналогичные формулам (3) и (4), при $m=0$ и $m=M$:

$$\left. \frac{\partial u}{\partial x} \right|_{(x_0, t_n)} = \frac{u(x_1, t_n) - u(x_0, t_n)}{h} - \frac{h}{2} \left. \frac{\partial^2 u}{\partial x^2} \right|_{(x_0^{(1)}, t_n)}, \quad x_0 < x_0^{(1)} < x_1,$$

$$\left. \frac{\partial u}{\partial x} \right|_{(x_M, t_n)} = \frac{u(x_M, t_n) - u(x_{M-1}, t_n)}{h} + \frac{h}{2} \left. \frac{\partial^2 u}{\partial x^2} \right|_{(x_M^{(1)}, t_n)},$$

$$x_{M-1} < x_M^{(1)} < x_M,$$

получим на основании условий (30) и (31) при $t=t_n$ выражения

$$\begin{aligned} \alpha_{0n} \cdot \frac{u(x_1, t_n) - u(x_0, t_n)}{h} + \beta_{0n} u(x_0, t_n) = \\ = \gamma_{0n} + \frac{\alpha_{0n}}{2} h \left. \frac{\partial^2 u}{\partial x^2} \right|_{(x_0^{(1)}, t_n)}, \end{aligned} \quad (32)$$

$$\begin{aligned} \alpha_{1n} \cdot \frac{u(x_M, t_n) - u(x_{M-1}, t_n)}{h} + \beta_{1n} u(x_M, t_n) = \\ = \gamma_{1n} - \frac{\alpha_{1n}}{2} h \left. \frac{\partial^2 u}{\partial x^2} \right|_{(x_M^{(1)}, t_n)}, \end{aligned} \quad (33)$$

где $\alpha_{in} = \alpha_i(t_n)$, $\beta_{in} = \beta_i(t_n)$, $\gamma_{in} = \gamma_i(t_n)$, $i=0,1$.

Отбрасывая в выражениях (32) и (33) члены, имеющие первый порядок относительно h , построим сеточные граничные условия

$$\alpha_{0n} \frac{u_1^n - u_0^n}{h} + \beta_{0n} u_0^n = \gamma_{0n}, \quad (34)$$

$$\alpha_{1n} \frac{u_M^n - u_{M-1}^n}{h} + \beta_{1n} u_M^n = \gamma_{1n}. \quad (35)$$

Далее в целях упрощений будем полагать, что

$$\alpha_i(t) \equiv 0, \quad \beta_i(t) \equiv 1, \quad i=0,1.$$

Используя формулы (9), (10) и (34), (35), запишем две разностные схемы, аппроксимирующие задачу (28) — (31) с погрешностью порядка $O(\tau + h^2)$.

Имеем

$$L_h^{(1)}(u^{(h)}) = f^{(h)}, \quad (36)$$

где

$$L_h^{(1)}(u^{(h)}) \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \\ m = 1, 2, \dots, M-1, n = 0, 1, \dots, N-1, \\ u_m^0, m = 0, 1, \dots, M, \\ u_0^{n+1}, n = 0, 1, \dots, N-1, \\ u_M^{n+1}, n = 0, 1, \dots, N-1, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \varphi_m^n, m = 1, 2, \dots, M-1, n = 0, 1, \dots, N-1, \\ \psi_m, m = 0, 1, \dots, M, \\ \gamma_{0n+1}, n = 0, 1, \dots, N-1, \\ \gamma_{1n+1}, n = 0, 1, \dots, N-1, \end{cases}$$

$$L_h^{(2)}(u^{(h)}) = f^{(h)}, \quad (37)$$

где

$$L_h^{(2)}(u^{(h)}) \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2}, \\ m = 1, 2, \dots, M-1, n = 0, 1, \dots, N-1, \\ u_m^0, m = 0, 1, \dots, M, \\ u_0^{n+1}, n = 0, 1, \dots, N-1, \\ u_M^{n+1}, n = 0, 1, \dots, N-1. \end{cases}$$

В схеме (37) правая часть $f^{(h)}$ определяется так же, как и в схеме (36).

Разностная схема (36) явная, она, как и схема (9), устойчива при $\tau \leq \frac{1}{2} h^2$; разностная схема (37) неявная, она устойчива при любых отношениях τ и h . Как мы отмечали раньше, схема (10) тоже устойчива при любых τ и h .

Очевидно, что разностные схемы (36) и (37) аппроксимируют задачу (28)–(31) (при условии, что $\alpha_i(t) \equiv 0$, $\beta_i(t) \equiv 1$, $i = 0, 1$) с погрешностью $O(\tau + h^2)$. В силу теоремы 1 из п. 2 § 1 разностная схема (37) сходится при любых значениях $r = \frac{\tau}{h^2} > 0$, разностная же схема (36) будет сходящейся при $r = \frac{\tau}{h^2} \leq \frac{1}{2}$.

Для многих задач несложным путем могут быть указаны необходимые условия устойчивости. Такие условия позволяют сразу отбрасывать те схемы, для которых они не выполняются, и не проводить для этих схем доказательств устойчивости (поскольку в случае нарушения необходимого условия устойчивости быть не может).

5. Необходимое спектральное условие устойчивости Неймана. Вывод этого условия дадим на примере задачи Коши для уравнения теплопроводности

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + \varphi(x, t), \quad -\infty < x < +\infty, \quad 0 < t \leq T, \quad (38)$$

$$u(x, 0) = \psi(x), \quad a^2 = \text{const.}$$

По аналогии с (9) и (10) напомним две разностные схемы для задачи (38):

$$L_h^{(1)}(u^{(h)}) = f^{(h)}, \quad (39)$$

где

$$L_h^{(1)}(u^{(h)}) \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - a^2 \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 0, 1, \dots, N-1, \\ u_m^0, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \varphi(x_m, t_n), \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 0, 1, \dots, N-1, \\ \psi_m, \end{cases}$$

$$L_h^{(2)}(u^{(h)}) = f^{(h)}, \quad (40)$$

где

$$L_h^{(2)} \equiv \begin{cases} \frac{u_m^{n+1} - u_m^n}{\tau} - a^2 \frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2}, \\ m = 0, \pm 1, \pm 2, \dots, \\ n = 0, 1, 2, \dots, N-1, \\ u_m^0, \end{cases}$$

$f^{(h)}$ определяется так же, как и в формуле (39).

Очевидно, что разностные схемы (39) и (40) аппроксимируют задачу (38) с погрешностью $O(\tau + h^2)$.

Получим необходимое условие устойчивости разностной схемы (39).

Пусть эта схема устойчива, тогда по определению устойчивости существует такое $h_0 > 0$, что для всех $h < h_0$ и любых $f^{(h)} \in F_h$ выполняется условие

$$\|u^{(h)}\|_{U_h} \leq M \|f^{(h)}\|_{F_h}, \quad (41)$$

где M — постоянная, не зависящая от h и $f^{(h)}$. Перепишем неравенство (41), раскрывая содержание понятия норма. Получим

$$\max_{m, n} |u_m^n| \leq M \left(\max_{m, n} |\varphi_m^n| + \max_m |\psi_m| \right). \quad (42)$$

Отметим, что в (41) $f^{(h)}$ есть любой элемент из F_h и

$$f^{(h)} \equiv \begin{cases} \varphi_m^n, \\ \psi_m. \end{cases}$$

Так что, в частности, может быть $\varphi_m^n \equiv \varphi(x_m, t_n)$, $\psi_m = \psi(x_m)$ и, следовательно, $f^{(h)}$ будет совпадать с правой частью разностной схемы (39). Поскольку неравенство (42) должно иметь место для любых $f^{(h)} \in F_h$, положим

$$f^{(h)} \equiv \begin{cases} 0, \\ \psi_m. \end{cases}$$

Тогда (42) примет вид

$$\max_m |u_m^n| \leq M \max_m |u_m^0|, \quad n = 0, 1, \dots, N, \quad (43)$$

ибо здесь u_m^n есть решение разностной схемы со специальной правой частью

$$L_h^{(1)}(u^{(h)}) = f^{(h)} \equiv \begin{cases} 0, \\ \psi_m. \end{cases} \quad (44)$$

Говорят, что разностная схема (39) *устойчива по начальным данным*, если для нее имеет место оценка (43).

Оценка (43) должна иметь место при некоторых частных значениях u_m^0 . Положим

$$u_m^0 = \psi_m \equiv e^{i\alpha m}, \quad (45)$$

где α — численный параметр, $-\infty < \alpha < +\infty$.

Выпишем решение задачи (44) при условии (45). Сделать это можно, используя метод разделения переменных. Положим

$$u_m^n = \lambda^n e^{i\alpha m}. \quad (46)$$

Подберем здесь численный параметр λ так, чтобы выполнялось однородное разностное уравнение из схемы (44). Подставляя u_m^n , определяемое по формуле (46), в это однородное уравнение, получим

$$\frac{\lambda^{n+1}e^{i\alpha m} - \lambda^n e^{i\alpha m}}{\tau} - a^2 \frac{\lambda^n e^{i\alpha(m+1)} - 2\lambda^n e^{i\alpha m} + \lambda^n e^{i\alpha(m-1)}}{h^2} = 0.$$

Сокращая на $\lambda^n e^{i\alpha m}$ и преобразовывая оставшуюся часть, получим

$$\lambda = 1 - 4a^2 r \sin^2 \frac{\alpha}{2}, \quad r = \frac{\tau}{h^2}. \quad (47)$$

На основании (46) можно записать:

$$\max_m |u_m^n| \leq |\lambda|^n, \quad n = 0, 1, \dots, N, \quad (48)$$

так как $|e^{i\alpha m}| = 1$.

Теперь видно, что правая часть (48) может быть ограниченной при любых сколь угодно больших значениях n , например, в случае, когда

$$|\lambda| \leq 1 + C\tau, \quad (49)$$

где C — постоянная, не зависящая от τ . Действительно,

$$|\lambda|^n = (1 + C\tau)^n \leq (1 + C\tau)^N \leq (1 + C\tau)^{\frac{T}{\tau}} \leq e^{CT}.$$

Из формулы (47) следует, что неравенство (49) будет выполняться при

$$r \leq \frac{1}{2a^2}. \quad (50)$$

Условие (49) называется *необходимым спектральным условием устойчивости Неймана*. Для разностной схемы (39) эти условия будут выполняться только для значений r , удовлетворяющих неравенству (50), т. е. при $\tau \leq \frac{1}{2a^2} h^2$.

Если проделать аналогичные выкладки в случае разностной схемы (40), то для λ мы получим

$$\lambda = \frac{1}{1 + 4ra^2 \sin^2 \frac{\alpha}{2}}.$$

Из этого выражения видно, что для разностной схемы (40) необходимое спектральное условие Неймана будет выполняться при любых $r > 0$.

Отметим здесь, что при $r > \frac{1}{2a^2}$ разностная схема (39) перестает быть устойчивой. Действительно, в этом случае при $\alpha = \pi$ из (47) имеем

$$\lambda = 1 - 4a^2r.$$

И, значит, $|\lambda| > 1$, ибо $2a^2r > 1$. Еще получим

$$u_m^n = (1 - 4a^2r)^n \cdot e^{i\pi m} = (1 - 4a^2r)^n \cdot (-1)^m.$$

Поэтому

$$|u_m^n| \rightarrow +\infty \quad \text{при} \quad n \rightarrow +\infty$$

и устойчивости не может быть.

Если в уравнении (38) коэффициент a^2 зависит от x, t , т. е. $a^2 \equiv a^2(x, t)$, то в этом случае для исследования устойчивости разностных схем вида (39) и (40), в которых a^2 заменено на $a^2(x_m, t_n)$, может быть применен принцип замороженных коэффициентов (см. [3], § 26).

Отметим, что разностная схема (40) опять будет удовлетворять необходимому спектральному условию устойчивости Неймана при любых $r > 0$, а разностная схема (39) — при $r \leq \frac{1}{2 \max_{(x, t)} a^2(x, t)}$.

6. Экономичные разностные схемы. Построение таких разностных схем мы рассмотрим на примере смешанной граничной задачи для уравнения теплопроводности

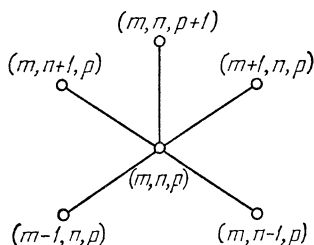
$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 \leq x, y \leq 1, \quad 0 < t \leq T, \\ u(x, y, 0) &= \psi(x, y), \\ u(x, y, t)|_{\Gamma} &= 0, \end{aligned} \quad (51)$$

где Γ — боковые грани прямоугольного параллелепипеда $D = \{0 \leq x, y \leq 1, \quad 0 < t \leq T\}$. Будем предполагать, что задача (51) имеет единственное решение $u(x, y, t)$ и это решение непрерывно в $\bar{D} = D + \Gamma$ вместе со своими производными $\frac{\partial^i u}{\partial t^i}$, $i = 1, 2$, $\frac{\partial^k u}{\partial x^k}$, $\frac{\partial^k u}{\partial y^k}$, $k = 3, 4$.

Выберем квадратную сетку по пространству и координаты узлов (x_m, y_n, t_p) определим по правилу

$$(x_m, y_n, t_p) = (mh, nh, p\tau), \quad m, n = 0, 1, \dots, M, \quad h = \frac{1}{M}, \\ \tau > 0, \quad p = 0, 1, \dots, N, \quad N\tau \leq T < (N+1)\tau.$$

Рассмотрим явный шаблон



и по аналогии, например, с (24), (25) запишем разностную схему для задачи (51)

$$L_h^{(1)}(u^{(h)}) = f^{(h)}, \quad (52)$$

где

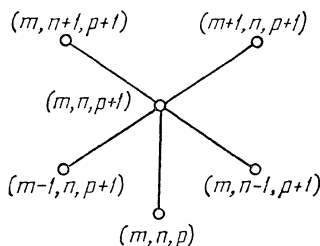
$$L_h^{(1)}(u^{(h)}) \equiv \begin{cases} \frac{u_{mn}^{p+1} - u_{mn}^p}{\tau} - \Lambda_{xx}(u_{mn}^p) - \Lambda_{yy}(u_{mn}^p), \\ m, n = 1, 2, \dots, M-1, p = 0, 1, \dots, N-1, \\ u_{mn}^0, \quad m, n = 0, 1, \dots, M, \\ u_{mn}^{p+1}, (mh, nh, (p+1)\tau) \in \Gamma_h. \end{cases}$$

Через Γ_h мы обозначили узлы сетки, принадлежащие Γ ,

$$f^{(h)} \equiv \begin{cases} 0, \\ \psi(x_m, y_n), \quad m, n = 0, 1, \dots, M, \\ 0. \end{cases}$$

Очевидно, что разностная схема (52) явная и что она аппроксимирует задачу (51) с погрешностью порядка $O(\tau + h^2)$. С помощью принципа максимума можно установить, что при $\tau \leq \frac{1}{4}h^2$ разностная схема (52) устойчива.

Возьмем «неявный» шаблон



и, используя формулы, подобные (26), (27), запишем разностную схему для задачи (51):

$$L_h^{(3)}(u^{(h)}) = f^{(h)}, \quad (53)$$

где

$$L_h(u^{(h)}) \equiv \begin{cases} \frac{u_{mn}^{p+1} - u_{mn}^p}{\tau} - \Lambda_{xx}(u_{mn}^{p+1}) - \Lambda_{yy}(u_{mn}^{p+1}), \\ m, n = 1, 2, \dots, M-1, n = 0, 1, \dots, N-1, \\ u_{mn}^0, \quad m, n = 0, 1, \dots, M, \\ u_{mn}^{p+1}, (mh, nh, (p+1)\tau) \in \Gamma_h, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} 0, \\ \psi(x_m, y_n), \quad m, n = 0, 1, \dots, M, \\ 0. \end{cases}$$

Разностная схема (53) неявная, она аппроксимирует задачу (51) с погрешностью порядка $O(\tau + h^2)$ и устойчива при любых значениях $r > 0$, $r = \tau/h^2$.

Обозначим через u^k совокупность значений сеточной функции на слое номера k , т. е. $u^k = \{u_{mn}^k, m, n = 0, 1, \dots, M\}$, $k = 0, 1, \dots, N$.

Пусть нам известны значения u^p . Для того чтобы вычислить значения u^{p+1} по явной разностной схеме (52), мы должны выполнить число арифметических операций, пропорциональное числу неизвестных u_{mn}^{p+1} . Этих неизвестных будет $(M-1)^2$, так как значения на Γ_h известны: $u_{mn}^{p+1} = 0$, если узел принадлежит Γ_h .

Введем понятие *экономичной разностной схемы*. Будем говорить, что двумерная разностная схема экономична,

если для вычисления значений u^{p+1} по значениям u^p эта схема требует выполнения числа арифметических операций, пропорционального числу неизвестных значений на слое $p+1$.

Очевидно, что разностная схема (52) экономична, но, как мы уже отмечали, вычисления по этой схеме будут устойчивыми только при выполнении условия $\tau \leq \frac{1}{4} h^2$, которое является весьма жестким ограничением на шаг τ .

Если мы будем использовать разностную схему (52), то для вычисления u^{p+1} по u^p нужно будет решать систему линейных алгебраических уравнений, в которой число уравнений есть $(M-1)^2$ и этих уравнений столько же, сколько неизвестных. Даже при условии, что эта система уравнений будет решаться по методу Гаусса, одному из наиболее быстродействующих методов, число операций, необходимых для вычисления u^{p+1} по u^p , будет приблизительно пропорционально $[(M-1)^2]^3$. И, следовательно, схему (53) отнести к экономичным схемам нельзя. Однако при этом следует отметить еще раз, что важным положительным свойством этой схемы является ее абсолютная устойчивость.

Поставим задачу построить разностную схему, которая аппроксимировала бы задачу (51) и обладала свойствами:

1) экономичности; 2) абсолютной устойчивости.

Искомую разностную схему можно записать, используя идею расщепления (см. п. 3) и формулы (26), (27).

Получим

$$L_h(u^{(h)}) = f^{(h)}, \quad (54)$$

где

$$L_h(u^{(h)}) \equiv \begin{cases} \frac{u_{mn}^{p+1} - u_{mn}^p}{\tau} - \Lambda_{yy}(u_{mn}^{p+1}), \\ m, n = 1, 2, \dots, M-1, p = 0, 1, \dots, N-1, \\ u_{mn}^{p+1}, (mh, nh, (p+1)\tau) \in \Gamma_h, \\ u_{mn}^p, m, n = 0, 1, \dots, M, p = 0, 1, \dots, N-1; \end{cases}$$

$$f^{(h)} \equiv \begin{cases} 0, m, n = 1, 2, \dots, M-1, p = 0, 1, \dots, N-1, \\ 0, (mh, nh, (p+1)\tau) \in \Gamma_h, \\ \tilde{u}_{mn}^{p+1}, m, n = 0, 1, \dots, M, p = 0, 1, \dots, N-1. \end{cases}$$

Сеточная функция $\tilde{u}_{mn}^{p+1}$, фигурирующая в определении $f^{(h)}$, вычисляется как решение вспомогательной разностной схемы

$$\frac{\tilde{u}_{mn}^{p+1} - u_{mn}^p}{\tau} = \Lambda_{xx}(\tilde{u}_{mn}^{p+1}),$$

$$m, n = 1, 2, \dots, M-1, \quad p = 0, 1, \dots, N-1, \quad (55)$$

$$u_{mn}^0 = \psi_{mn}, \quad m, n = 0, 1, \dots, M,$$

$$\tilde{u}_{mn}^{p+1} = 0, \quad (mh, nh, (p+1)\tau) \in \Gamma_h.$$

Доказано, что разностная схема (54) аппроксимирует задачу (51) с погрешностью порядка $O(\tau + h^2)$, абсолютно устойчива (об этом подробно написано в [3], § 32) и экономична. Здесь мы уделим большее внимание вопросу реализации разностной схемы (54) и установлению свойства ее экономичности. Проследим, как по значениям $u_{mn}^0 = \psi_{mn}$ вычислить значения u_{mn}^1 на первом слое и затем по значениям u_{mn}^1 — значения u_{mn}^2 и т. д.

Положив в (55) $p=0$, получим

$$\frac{\tilde{u}_{mn}^1 - u_{mn}^0}{\tau} = \frac{\tilde{u}_{m+1n}^1 - 2\tilde{u}_{mn}^1 + \tilde{u}_{m-1n}^1}{h^2}, \quad m = 1, 2, \dots, M-1, \quad (56)$$

$$\tilde{u}_{0n}^1 = 0, \quad \tilde{u}_{Mn}^1 = 0, \quad n = 1, 2, \dots, M-1.$$

Формулы (56) представляют собой при фиксированном n линейную граничную задачу для разностного уравнения второго порядка.

Опишем одну из возможных реализаций метода разностной прогонки применительно к задаче (56). Пусть n фиксировано ($1 \leq n \leq M-1$). Перепишем (56) в виде

$$r\tilde{u}_{m+1n}^1 - (1+2r)\tilde{u}_{mn}^1 + r\tilde{u}_{m-1n}^1 = \psi_{mn},$$

$$m = 1, 2, \dots, M-1, \quad (57)$$

$$\tilde{u}_{0n}^1 = 0, \quad \tilde{u}_{Mn}^1 = 0, \quad r = \tau/h^2 > 0.$$

Пусть $\tilde{u}_{kn}^1$ ($0 \leq k \leq M$) — искомое решение задачи (57). Предположим, что между двумя соседними значениями этого решения существует связь

$$\tilde{u}_{in}^1 = \alpha_{in}\tilde{u}_{i+1n}^1 + \beta_{in}, \quad i = 0, 1, \dots, M-1, \quad (58)$$

где α_{in} , β_{in} — некоторые числовые коэффициенты. При $i=0$ определим α_{0n} и β_{0n} таким образом, чтобы выпол-

нялось левое граничное условие $\tilde{u}_{0n}^1 = 0$. Для этого достаточно положить

$$\alpha_{0n} = 0, \quad (59)$$

$$\beta_{0n} = 0. \quad (60)$$

Возьмем $i = m - 1$. Значение $\tilde{u}_{in}^1$, определяемое по формуле (58), подставим в (57) и затем выполним элементарные преобразования. Получим

$$\tilde{u}_{mn}^1 = \frac{r}{(1+2r) - r\alpha_{m-1n}} \cdot \tilde{u}_{m+1n}^1 - \frac{\psi_{mn} - r\beta_{m-1n}}{(1+2r) - r\alpha_{m-1n}}. \quad (61)$$

Выберем α_{in} и β_{in} так, чтобы выражения (58) при $i = m$ и (61) совпадали. Это дает формулы

$$\alpha_{mn} = \frac{r}{(1+2r) - r\alpha_{m-1}}, \quad (62)$$

$$\beta_{mn} = \frac{r\beta_{m-1n} - \psi_{mn}}{(1+2r) - r\alpha_{m-1}}. \quad (63)$$

Теперь можно указать алгоритм вычисления значений $\tilde{u}_{1n}^1, \tilde{u}_{2n}^1, \dots, \tilde{u}_{M-1n}^1$ по методу разностной прогонки:

1) по формулам (62), (59) и (63), (60) при $m = 1, 2, \dots, M-1$ последовательно вычисляем $\alpha_{1n}, \alpha_{2n}, \dots, \alpha_{M-1n}$ и $\beta_{1n}, \beta_{2n}, \dots, \beta_{M-1n}$; этот процесс называют *прямым ходом метода прогонки*;

2) искомые значения $\tilde{u}_{1n}^1, \dots, \tilde{u}_{M-1n}^1$ вычисляем по формуле (58) при $i = M-1, M-2, \dots, 1, 0$. Отметим, что в силу (57) $\tilde{u}_{Mn}^1 = 0$. Процесс вычисления значений $\tilde{u}_{M-1n}^1, \tilde{u}_{M-2n}^1, \dots, \tilde{u}_{1n}^1$ называют *обратным ходом метода прогонки*.

Легко видеть, что $0 \leq \alpha_{mn} < 1$, и потому метод прогонки устойчив к ошибкам округлений.

Такие вычисления надо проделать для каждого n , т. е. надо вычислить таблицы значений

$$\alpha_{11}, \alpha_{21}, \dots, \alpha_{M-11}; \quad \beta_{11}, \beta_{21}, \dots, \beta_{M-11};$$

$$\tilde{u}_{M-11}^1, \tilde{u}_{M-21}^1, \dots, \tilde{u}_{11}^1,$$

$$\alpha_{12}, \alpha_{22}, \dots, \alpha_{M-12}; \quad \beta_{12}, \beta_{22}, \dots, \beta_{M-12};$$

$$\tilde{u}_{M-12}^1, \tilde{u}_{M-22}^1, \dots, \tilde{u}_{12}^1,$$

$$\dots \dots \dots$$

$$\alpha_{1M-1}, \alpha_{2M-1}, \dots, \alpha_{M-1M-1}; \quad \beta_{1M-1}, \beta_{2M-1}, \dots, \beta_{M-1M-1};$$

$$\tilde{u}_{M-1M-1}^1, \tilde{u}_{M-2M-1}^1, \dots, \tilde{u}_{1M-1}^1.$$

Для дальнейших вычислений потребуется только таблица из значений $\tilde{u}_{mn}^1$; ее и следует хранить в памяти ЭВМ.

Обратимся теперь к задаче (54). Положив $p=0$, получим

$$\frac{u_{mn}^1 - \tilde{u}_{mn}^1}{\tau} = \frac{u_{mn+1}^1 - 2u_{mn}^1 + u_{mn-1}^1}{h^2},$$

$$u_{m0}^1 = 0, \quad u_{mM}^1 = 0, \quad n = 1, 2, \dots, M-1, \quad (64)$$

$$m = 1, 2, \dots, M-1.$$

Пусть в (64) m фиксировано, тогда эту задачу можно записать в виде

$$ru_{mn+1}^1 - (1+2r)u_{mn}^1 + ru_{mn-1}^1 = \tilde{u}_{mn}^1,$$

$$n = 1, 2, \dots, M-1, \quad (65)$$

$$u_{m0}^1 = 0, \quad u_{mM}^1 = 0, \quad r = \tau/h^2 > 0.$$

Метод прогонки для задачи (65) полностью аналогичен методу прогонки для задачи (57). Укажем только на порядок вычислений.

1) По формулам

$$\gamma_{mn} = \frac{r}{(1+2r) - r\gamma_{mn-1}}, \quad \gamma_{m0} = 0,$$

$$\delta_{mn} = \frac{r\delta_{mn-1} - \tilde{u}_{mn}^1}{(1+2r) - r\gamma_{mn-1}}, \quad \delta_{m0} = 0,$$

вычисляем $\gamma_{m1}, \gamma_{m2}, \dots, \gamma_{mM-1}$ и $\delta_{m1}, \delta_{m2}, \dots, \delta_{mM-1}$.

2) Приближенные значения u_{mn}^1 на первом слое вычисляем по формуле:

$$u_{mi}^1 = \gamma_{mi}u_{mi+1}^1 + \delta_{mi}, \quad u_{mM}^1 = 0, \quad i = M-1, M-2, \dots, 1, 0.$$

Вычисления, указанные в 1) и 2), надо проделать для каждого m , $1 \leq m \leq M-1$. Таким образом, и здесь надо вычислить три таблицы.

$$\gamma_{11}, \gamma_{12}, \dots, \gamma_{1M-1}; \quad \delta_{11}, \delta_{12}, \dots, \delta_{1M-1};$$

$$u_{11}^1, u_{12}^1, \dots, u_{1M-1}^1,$$

$$\gamma_{21}, \gamma_{22}, \dots, \gamma_{2M-1}; \quad \delta_{21}, \delta_{22}, \dots, \delta_{2M-1};$$

$$u_{21}^1, u_{22}^1, \dots, u_{2M-1}^1,$$

$$\dots \dots \dots$$

$$\gamma_{M-11}, \gamma_{M-12}, \dots, \gamma_{M-1M-1}; \quad \delta_{M-11}, \delta_{M-12}, \dots, \delta_{M-1M-1};$$

$$u_{M-11}^1, u_{M-12}^1, \dots, u_{M-1M-1}^1.$$

Отметим, что как при вычислении таблиц для α_{mn} , β_{mn} , $\tilde{u}_{mn}^1$, так и для вычисления таблиц для γ_{mn} , δ_{mn} , u_{mn}^1 , необходимо было выполнить число арифметических операций, которое пропорционально $(M-1)^2$, т. е. пропорционально числу неизвестных u_{mn}^1 . Поэтому схема (54) экономична.

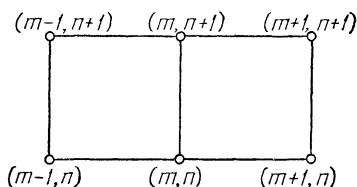
Далее вычисления аналогичны. А именно, при вычислении $\tilde{u}_{mn}^2$ поступаем так же, как и при вычислении $\tilde{u}_{mn}^1$, только $u_{mn}^0 = \psi_{mn}$ заменяем на u_{mn}^1 , при вычислении u_{mn}^2 правую часть $\tilde{u}_{mn}^1$ заменяем на $\tilde{u}_{mn}^2$ и т. д. Конечным итогом вычислений будет получение значений u_{mn}^p при $m, n = 1, 2, \dots, M-1, p = 1, 2, \dots, N$.

В заключение параграфа о разностных схемах для уравнений параболического типа мы приведем несколько разностных аппроксимаций для уравнения

$$L(u) \equiv \frac{\partial u}{\partial t} - a \frac{\partial^2 u}{\partial x^2} = 0,$$

где $a = \text{const} > 0$. Эти схемы часто использовались в вычислительной практике и их полезность и хорошие свойства были многократно подтверждены.

1. *Схема Дж. Кранка и Р. Николсона*. Она строится по шеститочечному двухслойному шаблону



и имеет вид

$$L_h(u^{(h)}) \equiv \frac{u_m^{n+1} - u_m^n}{\tau} - a \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n + u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{2h^2} = 0.$$

Эта схема неявная, она абсолютно устойчива и для погрешности аппроксимации $r_{mn}(h)$ верно представление

$$r_{mn}(h) = O(\tau^2 + h^2).$$

2. *Неявная двухслойная шеститочечная разностная аппроксимация.* Эта схема строится по тому же шаблону точек, что и схема Дж. Кранка и Р. Николсона, и имеет вид

$$L_h(u^{(h)}) \equiv \frac{u_m^{n+1} - u_m^n}{\tau} - a \frac{\gamma(u_{m+1}^n + 2u_m^n + u_{m-1}^n) + (1-\gamma)(u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1})}{h^2} = 0,$$

где $\gamma = \text{const}$, $0 \leq \gamma \leq 1$.

При $0 \leq \gamma < \frac{1}{2}$ схема устойчива, если

$$a \frac{\tau}{h^2} = \text{const} \leq \frac{1}{2(1-2\gamma)}.$$

При $\frac{1}{2} < \gamma \leq 1$ схема абсолютно устойчива.

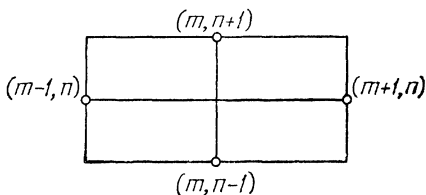
Для погрешности аппроксимации в обоих случаях верно представление $r_{mn}(h) = O(\tau + h^2)$. Если положить

$$\gamma = \frac{1}{2} - \frac{h^2}{12a\tau},$$

то схема будет устойчивой, но порядок ее аппроксимации по сравнению с двумя предыдущими случаями будет более высокий, а именно

$$r_{mn}(h) = O(\tau^2) = O(h^4).$$

3. *Схема Е. Дюфорта и С. Франкела.* Она строится по четырехточечному трехслойному шаблону вида



и имеет форму

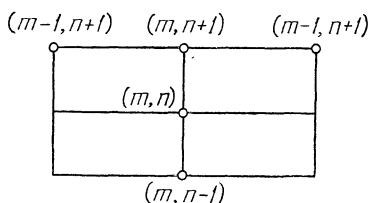
$$L_h(u^{(h)}) \equiv \frac{u_m^{n+1} - u_m^{n-1}}{2\tau} - a \frac{u_{m+1}^n - u_m^{n+1} - u_m^{n-1} + u_{m-1}^n}{h^2} = 0.$$

Эта схема явная, абсолютно устойчивая. Для погрешности аппроксимации верно представление

$$r_{mn}(h) = O(\tau + h^2) + O\left[\left(\frac{\tau}{h}\right)^2\right]$$

при условии, что $\frac{\tau}{h} \rightarrow 0$, когда $\tau, h \rightarrow 0$.

4. *Неявная пятиточечная трехслойная схема.* Здесь рассматривается шаблон вида



Имеем

$$L_h(u^{(h)}) \equiv \frac{3}{2} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{1}{2} \frac{u_m^n - u_m^{n-1}}{\tau} - a \frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2} = 0.$$

Схема абсолютно устойчива,

$$r_{mn}(h) = O(\tau^2 + h^2).$$

Отметим еще, что при изучении устойчивости разностных схем мы рассмотрели только простейшие приемы и методы исследования устойчивости так называемых модельных задач. Мы рассмотрели принцип максимума, метод разделения переменных (которым мы воспользовались при выводе необходимого спектрального условия устойчивости Неймана) и частично принцип замороженных коэффициентов.

Другие методы исследования устойчивости разностных схем, а также подробное рассмотрение современной теории и состояния этого вопроса даны в работах А. А. Самарского, в частности в его монографии [6].

§ 3. Разностные схемы для уравнений эллиптического типа

Как мы уже отмечали (см. § 1), уравнение

$$L(u) \equiv a(x, y) \frac{\partial^2 u}{\partial x^2} + 2b(x, y) \frac{\partial^2 u}{\partial x \partial y} + c(x, y) \frac{\partial^2 u}{\partial y^2} + \\ + 2d(x, y) \frac{\partial u}{\partial x} + 2e(x, y) \frac{\partial u}{\partial y} + g(x, y) \cdot u = f(x, y) \quad (1)$$

называется эллиптическим в области D , если

$$\delta(x, y) \equiv b^2(x, y) - a(x, y) \cdot c(x, y) < 0 \quad (2)$$

для всех (x, y) из D . К уравнениям эллиптического типа на границе Γ области D присоединяются граничные условия трех видов:

1) граничные условия первого рода

$$u|_{\Gamma} = \varphi(M), \quad (3)$$

где φ — известная функция, M — точка контура Γ ;

2) граничные условия второго рода

$$\frac{\partial u}{\partial n} \Big|_{\Gamma} = \varphi(M), \quad (4)$$

где φ — известная функция, M — точка контура Γ , $\frac{\partial u}{\partial n}$ — производная по внешней нормали;

3) граничные условия третьего рода

$$\left[\alpha(x, y) \frac{\partial u}{\partial n} + \beta(x, y) \cdot u \right]_{\Gamma} = \varphi(M), \quad (5)$$

где α , β , φ — известные функции, M — точка контура Γ , $\frac{\partial u}{\partial n}$ — производная по внешней нормали.

Требуется определить функцию $u(x, y)$, которая в области D удовлетворяла бы уравнению (1), а на границе Γ области D — одному из трех условий.

Задача (1), (3) называется задачей Дирихле, задача (1), (4) называется задачей Неймана и задача (1), (5) называется смешанной граничной задачей.

Рассмотрим элементы теории разностных схем для эллиптических уравнений. Область D будем считать ограниченной и ее границу Γ — простой (без кратных точек) кусочно-гладкой линией.

1. Построение разностных аппроксимаций для уравнений. Пусть в уравнении (1) смешанная производная $\frac{\partial^2 u}{\partial x \partial y}$ отсутствует, т. е. $b(x, y) \equiv 0$. Построим разностную аппроксимацию для такого уравнения в предположении, что оно имеет наиболее простой вид

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y). \quad (6)$$

Уравнение (6) называют уравнением Пуассона.

Выберем прямоугольную сетку, определив положение узлов (x_m, y_n) по правилу

$$\begin{aligned} x_m &= mh, \quad m = 0, \pm 1, \pm 2, \dots; \\ y_n &= nl, \quad n = 0, \pm 1, \pm 2, \dots; \quad h > 0, \quad l > 0. \end{aligned}$$

К сеточной области D_h отнесем все узлы, принадлежащие области $\bar{D} = D + \Gamma$ (рис. 1).

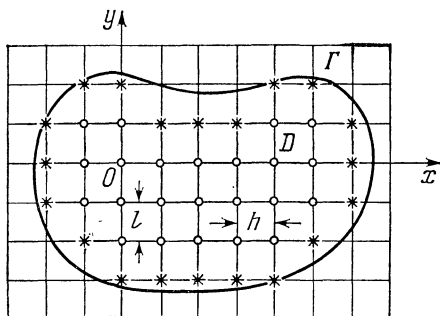
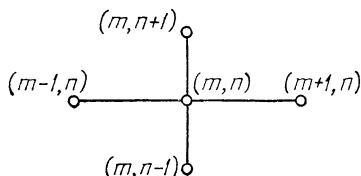


Рис. 1.

Возьмем пятиточечный шаблон



Пользуясь расположением точек в этом шаблоне, разобьем

узлы области на две категории: внутренние и граничные. Некоторый узел (m, n) будем считать внутренним, если он сам и четыре соседних точки шаблона принадлежат области D_h (такие узлы помечены на рис. 1 знаком $\circ$), и множество их мы обозначим через D_h° . Остальные узлы назовем граничными (они помечены на рис. 1 знаком $*$), и множество этих узлов обозначим через Γ_h : таким образом, $D_h = D_h^\circ + \Gamma_h$.

Из сказанного здесь следует, что разбиение узлов области D_h на внутренние и граничные зависит от того, какой шаблон выбран для аппроксимации дифференциального уравнения.

Пусть узел $(m, n) \in D_h^\circ$. Замену дифференциального уравнения (6) разностным будем осуществлять только во внутренних узлах.

Имеем

$$\frac{\partial^2 u}{\partial x^2} \Big|_{(x_m, y_n)} + \frac{\partial^2 u}{\partial y^2} \Big|_{(x_m, y_n)} = f(x_m, y_n). \quad (7)$$

Воспользовавшись формулой, подобной (6.11.8), и аналогичной формулой для $\frac{\partial^2 u}{\partial y^2}$, получим на основании (7)

$$\begin{aligned} & \frac{u(x_{m+1}, y_n) - 2u(x_m, y_n) + u(x_{m-1}, y_n))}{h^2} - \frac{h^2}{12} \cdot \frac{\partial^4 u}{\partial x^4} \Big|_{(x_m^{(1)}, y_n)} + \\ & + \frac{u(x_m, y_{n+1}) - 2u(x_m, y_n) + u(x_m, y_{n-1}))}{l^2} - \\ & - \frac{l^2}{12} \cdot \frac{\partial^4 u}{\partial y^4} \Big|_{(x_m, y_n^{(1)})} = f(x_m, y_n), \quad (m, n) \in D_h^\circ, \quad (8) \\ & x_{m-1} < x_m^{(1)} < x_{m+1}, \quad y_{n-1} < y_n^{(1)} < y_{n+1}. \end{aligned}$$

Пусть функции $\frac{\partial^4 u(x, y)}{\partial x^4}$ и $\frac{\partial^4 u(x, y)}{\partial y^4}$ ограничены по абсолютной величине в области $\bar{D}$; тогда в формуле (8) при достаточно малых h и l можно пренебречь членами, содержащими в качестве множителей h^2 и l^2 , и мы получим искомое разностное уравнение

$$L_h^{(1)}(u^{(h)}) = f^{(h)}, \quad (9)$$

где

$$L_h^{(1)}(u^{(h)}) \equiv \frac{u_{m+1n} - 2u_{mn} + u_{m-1n}}{h^2} + \frac{u_{mn+1} - 2u_{mn} + u_{mn-1}}{l^2},$$

$$(m, n) \in D_h^\circ,$$

$$f^{(h)} \equiv f(x_m, y_n), \quad (m, n) \in D_h^\circ.$$

Здесь через u_{mn} мы обозначаем приближенное сеточное значение решения уравнения (6), так что $u_{mn} \approx u(x_m, y_n)$ и $u(x, y)$ — решение уравнения (6). В отличие от уравнений параболического типа, в обозначении сеточного значения u_{mn} мы оба индекса ставим внизу; в силу их равноправности, они указывают на изменения положения узла на плоскости. В случае уравнения параболического типа аналогичное значение обозначалось через u_m^n и различной постановкой индексов подчеркивалась в сущности различная роль, которую играют в нестационарных задачах переменные по времени и пространству.

В силу определений и понятий, введенных в § 1, можно из формул (8), (9) получить

$$L_h^{(1)}(u_h(x, y)) = f^{(h)} + \delta f^{(h)}, \quad (10)$$

где

$$\delta f^{(h)} \equiv \frac{h^2}{12} \cdot \frac{\partial^4 u}{\partial x^4} \Big|_{(x_m^{(1)}, y_n)} + \frac{l^2}{12} \cdot \frac{\partial^4 u}{\partial y^4} \Big|_{(x_m, y_n^{(1)})}, \quad (m, n) \in D_h^\circ.$$

Следовательно, при сделанных выше предположениях относительно $\frac{\partial^4 u}{\partial x^4}$ и $\frac{\partial^4 u}{\partial y^4}$, имеет место оценка

$$\|\delta f^{(h)}\|_{F_h} \leq Mh^2, \quad (11)$$

где M — постоянная, не зависящая от h . Считаем, что $l = \alpha h$, α — некоторое положительное число.

Оценка (11) означает, что разностное уравнение (9) аппроксимирует дифференциальное уравнение (6) на решении $u(x, y)$ с погрешностью порядка $O(h^2)$.

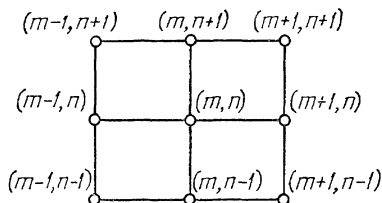
Рассмотрим теперь случай, когда в уравнение (1) входит смешанная производная $\frac{\partial^2 u}{\partial x \partial y}$ и это уравнение

имеет, например, такой вид:

$$a \frac{\partial^2 u}{\partial x^2} + 2b \frac{\partial^2 u}{\partial x \partial y} + c \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad (12)$$

где a, b, c — постоянные и $b^2 - ac < 0$.

Для замены дифференциального уравнения (12) разностным выберем шаблон



Используя этот шаблон, выделим в области D_h множество внутренних узлов D_h° и множество граничных узлов Γ_h . Будем считать узел (m, n) внутренним и пометить его знаком $\circ$ (рис. 2), если этот узел и восемь соседних узлов,

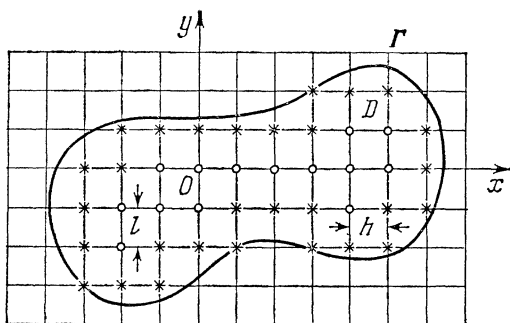


Рис. 2.

содержащихся в шаблоне, принадлежат D_h . Оставшиеся узлы области D_h отнесем к числу граничных, они помечены знаком $*$; множество таких узлов обозначим через Γ_h .

Запишем разностное выражение для замены $\frac{\partial^2 u}{\partial x \partial y}$. Чтобы не загромождать записи, остатки записывать не будем, ограничимся только указанием на порядок остатков.

Используя формулу, аналогичную формуле (5), получим

$$\begin{aligned} \frac{\partial^2 u}{\partial x \partial y} \Big|_{(x_m, y_n)} &= \frac{\partial}{\partial x} \left(\frac{\partial u}{\partial y} \right) \Big|_{(x_m, y_n)} = \\ &= \frac{\left(\frac{\partial u}{\partial y} \right)_{(x_{m+1}, y_n)} - \left(\frac{\partial u}{\partial y} \right)_{(x_{m-1}, y_n)}}{2h} + O(h^2) = \\ &= \frac{1}{2h} \left[\frac{u(x_{m+1}, y_{n+1}) - u(x_{m+1}, y_{n-1})}{2l} + \right. \\ &\quad \left. + O(l^2) - \frac{u(x_{m-1}, y_{n+1}) - u(x_{m-1}, y_{n-1})}{2l} - O(l^2) \right] + O(h^2) = \\ &= \frac{u(x_{m+1}, y_{n+1}) - u(x_{m+1}, y_{n-1}) - u(x_{m-1}, y_{n+1}) + u(x_{m-1}, y_{n-1})}{4hl} + \\ &\quad + O(h^2 + l^2), \end{aligned}$$

где предполагается, что $l/h = \text{const}$.

Теперь, подобно тому, как было построено разностное уравнение (9), можно построить для уравнения (12) разностное уравнение, обладающее аппроксимацией порядка $O(h^2 + l^2)$. Имеем

$$\begin{aligned} a \frac{u_{m+1n} - 2u_{mn} + u_{m-1n}}{h^2} + 2b \frac{u_{m+1n+1} - u_{m+1n-1} - u_{m-1n+1} + u_{m-1n-1}}{4hl} + \\ + c \frac{u_{mn+1} - 2u_{mn} + u_{mn-1}}{l^2} = f(x_m, y_n) \end{aligned}$$

для всех узлов (m, n) из области D_h .

2. Аппроксимация граничных условий. Рассмотрим, как можно заменить граничные условия вида (3) и (4) разностными условиями.

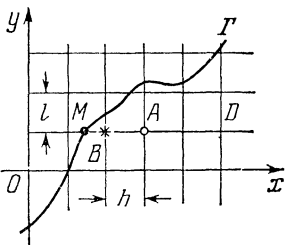


Рис. 3.

Отметим, что условия (3) и (4) заменяются разностными условиями на множестве граничных узлов Γ_h .

Пусть (m, n) — некоторый узел из Γ_h , обозначим его буквой B ; $(m+1, n)$ — внутренний узел, ближайший к B по направлению x , обозначим его буквой A . Буквой M обозначим точку контура Γ , ближайшую к B по направлению x (рис. 3 *).

*) Аналогично можно поступать в том случае, когда ближайшая к B точка Γ будет лежать в направлении оси y .

Координаты этих точек такие: $M(x_m - \delta, y_n)$, $0 < \delta < h$, $B(x_m, y_n)$, $A(x_{m+1}, y_n)$. По условию (3) имеем

$$u|_{\Gamma} = \varphi(M).$$

Значит, можно положить

$$u_{mn} = \varphi(M) \quad (13)$$

для узлов $(m, n) \in \Gamma_h$. Какова погрешность формулы (13)? Имеем

$$\varphi(M) = u(x_m - \delta, y_n) = u(x_m, y_n) - \frac{\delta}{1!} \frac{\partial u}{\partial x} \Big|_{(x_m^{(1)}, y_n)},$$

$x_m - \delta < x_m^{(1)} < x_m$. Отсюда получим

$$u(x_m, y_n) - \varphi(M) = u(x_m, y_n) - u_{mn} = - \frac{\delta}{1!} \frac{\partial u}{\partial x} \Big|_{(x_m^{(1)}, y_n)}.$$

Из этой формулы следует, что погрешность формулы (13) будет иметь первый порядок относительно h в предположении, что $\delta = \alpha h$, α — некоторое положительное число. Если точки M и B совпадают, то формула (13) будет точной.

Точность вычисления u_{mn} при $(m, n) \in \Gamma_h$ можно повысить, если воспользоваться еще значением $u(x, y)$ в точке A . Имеем

$$u(M) = u(x_m - \delta, y_n) = u(B) - \frac{\delta}{1!} \frac{\partial^2 u}{\partial x} \Big|_{(B)} + \frac{\delta^2}{2!} \frac{\partial^2 u}{\partial x^2} \Big|_{(\tilde{B})}, \quad (14)$$

где $\tilde{B}$ — некоторая средняя точка между M и B ,

$$u(A) = u(x_m + h, y_n) = u(B) + \frac{h}{1!} \frac{\partial u}{\partial x} \Big|_{(B)} + \frac{h^2}{2!} \frac{\partial^2 u}{\partial x^2} \Big|_{(\tilde{A})}, \quad (15)$$

где $\tilde{A}$ — некоторая средняя точка между B и A . Исключив из (14) $\frac{\partial u}{\partial x} \Big|_{(B)}$ с помощью формулы (15), получим

$$u(B) = \frac{h\varphi(M) + \delta u(A)}{h + \delta} + O(h^2).$$

Отбросив здесь величину $O(h^2)$, получим разностное граничное условие, аппроксимирующее граничное условие (3) в узле $(m, n) \in \Gamma_h$ с погрешностью порядка $O(h^2)$:

$$u_{mn} = \frac{h\varphi(M) + \delta u_{m+1n}}{h + \delta}. \quad (16)$$

Формулы вида (13) или (16) могут быть записаны для любого узла (m, n) , принадлежащего Γ_h .

Теперь обратимся к задаче замены граничного условия (4) разностным. Рассмотрим рис. 4.



Рис. 4.

Поясним обозначения на этом рисунке: B — граничный узел с координатами (x_m, y_n) , M — ближайшая к B точка контура Γ , A — внутренний узел с координатами (x_{m-1}, y_n) , C — граничный узел с координатами (x_m, y_{n-1}) , n — внешняя нормаль к Γ в точке M . Обозначим угол между n и Ox через α , между n и Oy — через β . Легко видеть, что $\beta = \frac{3}{2}\pi + \alpha$. Поставим задачу. Заменить граничное условие $\left. \frac{\partial u}{\partial n} \right|_{\Gamma} = \varphi(M)$ разностным в граничном узле B . По определению имеем

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x} \cos \alpha + \frac{\partial u}{\partial y} \cos \beta = \frac{\partial u}{\partial x} \cos \alpha + \frac{\partial u}{\partial y} \sin \alpha.$$

Далее мы предполагаем, что в точке B направление нормали n сохраняется таким же, как и в M . Поскольку расстояние между B и M есть величина порядка $O(h)$, то это предположение связано с внесением погрешности того же порядка $O(h)$. Значит,

$$\left. \frac{\partial u}{\partial n} \right|_{(M)} \approx \left. \frac{\partial u}{\partial n} \right|_{(B)}.$$

Поэтому окончательно получим

$$\frac{u(x_m, y_n) - u(x_{m-1}, y_n)}{h} \cdot \cos \alpha + \frac{u(x_m, y_n) - u(x_m, y_{n-1})}{l} \cdot \sin \alpha + O(h + l) = \varphi(M)$$

или, используя приближенные сеточные значения,

$$\frac{u_{mn} - u_{m-1n}}{h} \cos \alpha + \frac{u_{mn} - u_{mnl-1}}{l} \sin \alpha = \varphi(M). \quad (17)$$

Формула (17) представляет собой искомое разностное условие, аппроксимирующее в узле $(m, n) \in \Gamma_h$ граничное

условие (4) с погрешностью порядка $O(h+l)$. Выражения вида (17) должны быть записаны для всех граничных узлов $(m, n) \in \Gamma_h$, после чего будут получены разностные граничные условия, аппроксимирующие граничное условие (4). Как видим, процедура замены граничных условий разностными может оказаться громоздкой и сложной, особенно если контур Γ имеет непростую форму.

Замена граничных условий вида (5) может быть осуществлена с помощью формул вида (13), (16), (17).

Рассмотрим вопросы сходимости и устойчивости разностных схем для уравнений эллиптического типа на примере задачи Дирихле для уравнения Пуассона.

3. Построение разностной схемы в случае задачи Дирихле для уравнения Пуассона. Пусть в области $D = \{0 < x < a, 0 < y < b\}$ задано уравнение Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad (18)$$

а на границе Γ области D — условие Дирихле

$$u|_{\Gamma} = \varphi(M), \quad (19)$$

где φ — известная функция, M — точка контура Γ . Будем считать, что задача (18), (19) имеет единственное решение $u(x, y)$ в области $\bar{D} = D + \Gamma$ и это решение имеет непрерывные в D производные $\frac{\partial^4 u}{\partial x^4}$ и $\frac{\partial^4 u}{\partial y^4}$.

Возьмем прямоугольную сетку, положив

$$\begin{aligned} x_m &= mh, \quad m=0, 1, \dots, M, \quad h=a/M; \\ y_n &= nl, \quad n=0, 1, \dots, N, \quad l=b/N. \end{aligned}$$

Для аппроксимации уравнения (18) выберем пятиточечный шаблон (см. п. 1) и к множеству внутренних узлов D_h° отнесем все узлы, лежащие в D , а к множеству граничных узлов Γ_h отнесем узлы, лежащие на Γ .

Используя формулы вида (9) и (13), запишем разностную схему, аппроксимирующую задачу (18), (19) с погрешностью порядка $O(h^2 + l^2)$:

$$L_h(u^{(h)}) = f^{(h)}, \quad (20)$$

где

$$L_h(u^{(h)}) \equiv \begin{cases} \frac{u_{m+1n} - 2u_{mn} + u_{m-1n}}{h^2} + \frac{u_{mn+1} - 2u_{mn} + u_{mn-1}}{l^2}, \\ m = 1, 2, \dots, M-1, n = 1, 2, \dots, N-1, \\ u_{mn}, (mh, nl) \in \Gamma_h, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} f(x_m, y_n), m = 1, 2, \dots, M-1, n = 1, 2, \dots, N-1, \\ \varphi(x_m, y_n), (mh, nl) \in \Gamma_h. \end{cases}$$

Разностная схема (20) представляет собой систему линейных алгебраических уравнений, число уравнений этой системы равно $(M-1) \times (N-1)$, столько же и неизвестных, это — значения u_{mn} при $m = 1, 2, \dots, M-1$, $n = 1, 2, \dots, N-1$.

4. Об устойчивости разностной схемы (20). Доказательство устойчивости разностной схемы (20) сводится к выявлению двух свойств.

1. Разностная схема $L_h(z^{(h)}) = g^{(h)}$, где $g^{(h)} \equiv \begin{cases} \alpha_{mn}, \\ \beta_{mn} \end{cases}$, $g^{(h)}$ — произвольный элемент из F_h , однозначно разрешима.
2. Имеет место оценка

$$\|z^{(h)}\|_{U_h} \leq C \cdot \|g^{(h)}\|_{F_h},$$

где C — постоянная, не зависящая от h и $g^{(h)}$. Напомним, что, как и прежде, нормы определяются по правилу

$$\|z^{(h)}\|_{U_h} = \max_{m, n} |z_{mn}|,$$

$$\|g^{(h)}\|_{F_h} = \max_{m, n} |\alpha_{mn}| + \max_{m, n} |\beta_{mn}|.$$

Введем следующее обозначение:

$$\Lambda_h(u^{(h)}) \equiv \Lambda_{xx}(u^{(h)}) + \Lambda_{yy}(u^{(h)}).$$

Лемма 1. Пусть $v^{(h)} = \{v_{mn}\}$, $v_{mn} \not\equiv \text{const}$, — некоторая сеточная функция, определенная на $D_h = D_h^\circ + \Gamma_h$. Если выполняется условие

$$\Lambda_h(v^{(h)})'_{(x_m, y_n)} \geq 0, \quad (21)$$

где $(x_m, y_n) \in D_h^\circ$, то $v^{(h)}$ достигает своего наибольшего значения на $D^{(h)}$ в граничных точках, т. е. на Γ_h .

Доказательство. Предположим противное. Пусть $v^{(h)}$ имеет свое наибольшее значение v_{ij} в области D_h° . Мы можем считать, что хотя бы одно из значений v_{i+1j} , v_{i-1j} , v_{ij+1} , v_{ij-1} строго меньше v_{ij} . Тогда получим

$$\Lambda_h(v^{(h)})|_{(x_i, y_j)} = \Lambda_{xx}(v^{(h)})|_{(x_i, y_j)} + \Lambda_{yy}(v^{(h)})|_{(x_i, y_j)} < 0, \quad (22)$$

ибо

$$\begin{aligned} \Lambda_{xx}(v^{(h)})|_{(x_i, y_j)} &= \frac{v_{i+1j} - 2v_{ij} + v_{i-1j}}{h^2} \\ &= \frac{(v_{i+1j} - v_{ij}) + (v_{i-1j} - v_{ij})}{h^2} \leq 0, \end{aligned}$$

$$\begin{aligned} \Lambda_{yy}(v^{(h)})|_{(x_i, y_j)} &= \frac{v_{ij+1} - 2v_{ij} + v_{ij-1}}{l^2} = \\ &= \frac{(v_{ij+1} - v_{ij}) + (v_{ij-1} - v_{ij})}{l^2} \leq 0, \end{aligned}$$

при этом хотя бы одна из этих величин строго отрицательна в силу предположения о значении v_{ij} . Но (22) противоречит условию (21) леммы, и потому наше предположение неверно. Лемма доказана.

Лемма 2. Пусть $v^{(h)} = \{v_{mn}\}$, $v_{mn} \not\equiv \text{const}$, — некоторая сеточная функция, определенная на $D_h = D_h^\circ + \Gamma_h$. Если выполняется условие

$$\Lambda_h(v^{(h)})|_{(x_m, y_n)} \leq 0, \quad (23)$$

где $(x_m, y_n) \in D_h^\circ$, то $v^{(h)}$ достигает своего наименьшего значения на D_h в граничных точках, т. е. на Γ_h .

Доказательство аналогично предыдущему.

Теорема (принцип максимума). Каждое решение разностного уравнения

$$\Lambda_h(v^{(h)})|_{(x_m, y_n)} = 0, \quad (x_m, y_n) \in D_h^\circ,$$

принимает свое наибольшее и наименьшее значение в некоторых точках границы Γ_h .

Доказательство следует из лемм 1, 2.

Применим эту теорему к доказательству однозначной разрешимости разностной схемы

$$L_h(z^{(h)}) = g^{(h)} \quad (24)$$

при любых $g^{(h)} \in F_h$. С этой целью рассмотрим однородную разностную схему $L_h(z^{(h)}) = 0$ и покажем, что она

имеет только нулевое решение $z^{(h)} \equiv 0$. Отсюда будет следовать однозначная разрешимость схемы (24).

Подробно разностную схему $L_h(z^{(h)}) = 0$ можно записать в виде

$$L(z^{(h)}) \equiv \begin{cases} \Lambda_h(z^{(h)})|_{(x_m, y_n)} = 0, & (x_m, y_n) \in D_h^*, \\ z^{(h)}|_{(x_m, y_n)} = 0, & (x_m, y_n) \in \Gamma_h. \end{cases}$$

Если бы $z^{(h)} \not\equiv 0$, то тогда свое наибольшее значение и свое наименьшее значение величина $z^{(h)}$ в силу принципа максимума могла бы иметь в точках Γ_h ; но на Γ_h $z^{(h)} \equiv 0$, значит; и всюду в $D_h = D_h^* + \Gamma_h$ будет $z^{(h)} \equiv 0$. Поэтому разностная схема (24) однозначно разрешима.

Теперь покажем, что разностная схема (24) обладает свойством

$$\max_{m, n} |z_{mn}| \leq C \left(\max_{(m, n) \in D_h^*} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \right), \quad (25)$$

где C — постоянная, не зависящая от h и $g^{(h)} \equiv \begin{cases} \alpha_{mn}, \\ \beta_{mn}. \end{cases}$ Если мы это установим, то тем самым будет доказана устойчивость разностной схемы (20).

Оценку (25) мы получим путем построения мажоранты для функции $|z^{(h)}|$ по правилу С. Гершгорина.

Обратим внимание на следующий факт. Если

$$P(x, y) = \sum_{0 \leq i+j \leq 2} a_{ij} x^i y^j$$

есть некоторый многочлен второй степени от двух аргументов x и y , то

$$\Lambda_h(P^{(h)})|_{(x_m, y_n)} \equiv \left[\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} \right]_{(x_m, y_n)} = 2(a_{20} + a_{02}). \quad (26)$$

Действительно,

$$\Lambda_h(P^{(h)}) \equiv \Lambda_{xx}(P^{(h)}) + \Lambda_{yy}(P^{(h)}).$$

Кроме того, имеем

$$\Lambda_{xx}(P^{(h)})|_{(x_m, y_n)} \equiv \frac{\partial^2 P}{\partial x^2} \Big|_{(x_m, y_n)} = 2a_{20},$$

$$\Lambda_{yy}(P^{(h)})|_{(x_m, y_n)} \equiv \frac{\partial^2 P}{\partial y^2} \Big|_{(x_m, y_n)} = 2a_{02},$$

ибо $\frac{\partial^{k+l} P}{\partial x^k \partial y^l} \equiv 0$ при $k+l \geq 3$.

Мажоранту $Z(x, y)$ определим по правилу

$$Z(x, y) = \frac{1}{4} [(a^2 + b^2) - (x^2 + y^2)] \max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}|.$$

Поясним геометрический смысл функции $S(x, y) \equiv (a^2 + b^2) - (x^2 + y^2)$. Рассмотрим рис. 5, на котором изображена область D с контуром Γ .

Длина диагонали OA равна $\sqrt{a^2 + b^2}$, поэтому уравнение кривой $S(x, y) = 0$ есть уравнение окружности с центром в начале координат и радиусом, равным $OA = \sqrt{a^2 + b^2}$. Таким образом, если $(x, y) \in \bar{D} = D + \Gamma$, то $S(x, y) \geq$

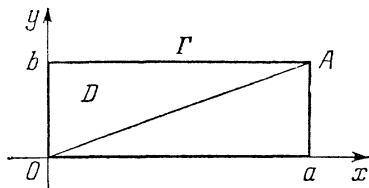


Рис. 5.

≥ 0 , причем $S(x, y) = 0$ только для одной точки из $\bar{D}$, когда $x = a$, $y = b$. Отметим, что эта точка не принадлежит D_h , и потому

$$S(x, y)|_{(x_m, y_n) \in D_h} > 0.$$

Покажем, что

$$|z^{(h)}| \leq Z^{(h)}. \quad (27)$$

Рассмотрим сумму $z^{(h)} + Z^{(h)}$. На множестве Γ_h имеем

$$[z^{(h)} + Z^{(h)}]_{(x_m, y_n) \in \Gamma_h} = z^{(h)}|_{\Gamma_h} + Z^{(h)}|_{\Gamma_h} = \beta_{mn}|_{(x_m, y_n) \in \Gamma_h} + \frac{1}{4} S(x, y)|_{(x_m, y_n) \in \Gamma_h} \max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \geq 0. \quad (28)$$

Кроме того,

$$\begin{aligned} \Lambda_h(z^{(h)} + Z^{(h)})|_{D_h^\circ} &= \Lambda_h(z^{(h)})|_{D_h^\circ} + \Lambda_h(Z^{(h)})|_{D_h^\circ} = \\ &= \alpha_{mn} + (- \max_{(m, n) \in D_h^\circ} |\alpha_{mn}|) \leq 0. \end{aligned} \quad (29)$$

Из (29) в силу леммы 2 следует, что функция $z^{(h)} + Z^{(h)}$ принимает свое наименьшее значение в точках Γ_h , но по (28) на Γ_h выполнено неравенство $z^{(h)} + Z^{(h)} \geq 0$. Значит, всюду в D_h

$$z^{(h)} + Z^{(h)} \geq 0. \quad (30)$$

Если рассмотреть разность $z^{(h)} - Z^{(h)}$ и воспользоваться леммой 1, то всюду в D_h получим неравенство

$$z^{(h)} - Z^{(h)} \leq 0. \quad (31)$$

Из (30) и (31) следует, что

$$|z^{(h)}| \leq Z^{(h)} \quad (32)$$

для всех $(x_m, y_n) \in D_h$. Из оценки (32) получим, что

$$\begin{aligned} |z_{mn}| &\leq \frac{1}{4} S(x_m, y_n) \max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \leq \\ &\leq \frac{1}{4} (a^2 + b^2) \max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \leq \\ &\leq \max \left\{ \frac{1}{4} (a^2 + b^2), 1 \right\} \cdot \left[\max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \right] = \\ &= C \cdot \left[\max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \right], \end{aligned}$$

где $C = \max \left\{ \frac{1}{4} (a^2 + b^2), 1 \right\}$.

Таким образом, имеем

$$\max_{m, n} |z_{mn}| \leq C \left[\max_{(m, n) \in D_h^\circ} |\alpha_{mn}| + \max_{(m, n) \in \Gamma_h} |\beta_{mn}| \right]$$

и, значит, оценка (25) установлена. Тем самым доказана устойчивость разностной схемы (20).

Разностная схема (20) аппроксимирует задачу (18), (19) с погрешностью порядка $O(h^2)$ (мы предполагаем, что $h/l = \text{const}$) и устойчива. В силу теоремы 1 из § 1 эта схема будет сходящейся со скоростью $O(h^2)$.

Для решения систем линейных алгебраических уравнений, определяемых формулой (20), разработан специальный метод — метод матричной прогонки.

5. Метод матричной прогонки. Перепишем разностную схему (20) в ином виде:

$$\begin{aligned} u_{m+1n} - 2u_{mn} + u_{m-1n} + \\ + \alpha (u_{mn+1} - 2u_{mn} + u_{mn-1}) = h^2 f(x_m, y_n) \quad (33) \\ m = 1, 2, \dots, M-1, n = 1, 2, \dots, N-1, \\ u_{0n} = \varphi(0, y_n), u_{Mn} = \varphi(a, y_n), n = 1, 2, \dots, N-1, \\ u_{m0} = \varphi(x_m, 0), u_{mN} = \varphi(x_m, b), \\ m = 1, 2, \dots, M-1, h/l = \alpha > 0. \end{aligned}$$

Пусть $M \gg N$. Введем обозначение

$$u_m = \begin{bmatrix} u_{m1} \\ u_{m2} \\ \vdots \\ u_{mN-1} \end{bmatrix}, \quad m = 0, 1, \dots, M. \quad (34)$$

Положим в формулах (33) $n = 1, 2, \dots, N-1$ и, используя (34), запишем систему уравнений (33) в векторной форме:

$$u_{m+1} + Au_m + u_{m-1} = f_m, \quad m = 1, 2, \dots, M-1, \quad (35)$$

$$u_0 = \varphi_0, \quad u_M = \varphi_a,$$

где A — трехдиагональная матрица порядка $N-1$ вида

$$A = \begin{bmatrix} -(2+2\alpha) & \alpha & 0 & \dots & 0 & 0 & 0 \\ \alpha & -(2+2\alpha) & \alpha & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha & -(2+2\alpha) & \alpha \\ 0 & 0 & 0 & \dots & 0 & \alpha & -(2+2\alpha) \end{bmatrix},$$

$$f_m = \begin{bmatrix} h^2 f(x_m, y_1) - \alpha \varphi(x_m, 0) \\ h^2 f(x_m, y_2) \\ \dots \\ h^2 f(x_m, y_{N-2}) \\ h^2 f(x_m, y_{N-1}) - \alpha \varphi(x_m, b) \end{bmatrix},$$

$$\varphi_0 = \begin{bmatrix} \varphi(0, y_1) \\ \varphi(0, y_2) \\ \vdots \\ \varphi(0, y_{N-1}) \end{bmatrix}, \quad \varphi_a = \begin{bmatrix} \varphi(a, y_1) \\ \varphi(a, y_2) \\ \vdots \\ \varphi(a, y_{N-1}) \end{bmatrix}.$$

Задача (35) аналогична задаче (57) из § 2 и ее отличие состоит лишь в том, что она имеет векторную форму. Формулы метода матричной прогонки подобны формулам (58) — (60), (62), (63), поэтому мы не будем приводить здесь их вывод и сразу изложим алгоритм метода прогонки в векторном виде.

1. По формуле

$$R_{m+1} = -(A + R_m)^{-1}, \quad m = 1, 2, \dots, M-1,$$

полагая $R_1 = 0$, вычисляем матрицы $R_m = (R_{ij}^{(m)})_{i=1}^{N-1}$, $m = 1, 2, \dots, M$, затем по формуле

$$s_{m+1} = R_{m+1}(s_m - f_m), \quad m = 1, 2, \dots, M-1,$$

полагая $s_1 = \varphi_0$, вычисляем векторы $s_m = (s_1^{(m)}, s_2^{(m)}, \dots, s_{N-1}^{(m)})^T$, T — знак транспонирования, $m = 1, 2, \dots, M$.

2. По формуле

$$u_{m-1} = R_m u_m + s_m, \quad m = M, M-1, \dots, 1,$$

полагая $u_M = \varphi_a$, вычисляем последовательно искомые значения решения задачи (35): $u_M, u_{M-1}, \dots, u_1, u_0$.

Устойчивость описанного здесь алгоритма метода матричной прогонки зависит от свойств матриц R_m . Если окажется, что $\|R_m\|_{III} \leq 1$ для всех m , то устойчивость к ошибкам округлений метода матричной прогонки будет очевидной. Пусть λ_i ($i = 1, 2, \dots, N-1$) — собственные значения матрицы A ; известно (см., например, [4]), что

$$\lambda_i \leq -2, \quad i = 1, 2, \dots, N-1. \quad (36)$$

Мы рассматриваем евклидову норму матриц, понимая под нормой матрицы $\|A\|_{III}$ корень квадратный из наибольшего собственного значения матрицы A^*A , где $*$ помечен переход к комплексно сопряженной и транспонированной матрице. В нашем случае A — вещественная матрица, поэтому $A^*A = A'A$. Пусть σ — ненулевой $(N-1)$ -мерный вектор-столбец. В силу свойств евклидовой нормы и неравенства (36) имеем

$$\|A\sigma\|_{III} \geq \min_i |\lambda_i| \cdot \|\sigma\|_{III} \geq 2\|\sigma\|_{III}, \quad (37)$$

ибо $|\lambda_i| \geq 2$.

Предположим, что при некотором k $\|R_k\|_{III} \leq 1$, и покажем, что $\|R_{k+1}\|_{III} \leq 1$. Тогда это будет означать, что все $\|R_m\|_{III} \leq 1$, $m = 1, 2, \dots, M$, так как $\|R_1\|_{III} = 0$. Наряду с вектором σ рассмотрим вектор τ , определяемый по формуле

$$\tau = -(A + R_k)\sigma = R_{k+1}\sigma.$$

Имеем

$$\begin{aligned} \|\tau\|_{III} &= \|A\sigma + R_k\sigma\|_{III} \geq \|A\sigma\|_{III} - \|R_k\sigma\|_{III} \geq \\ &\geq 2\|\sigma\|_{III} - \|\sigma\|_{III} = \|\sigma\|_{III}. \end{aligned}$$

Таким образом,

$$\|\tau\|_{III} \geq \|\sigma\|_{III}, \quad \|\tau\|_{III} = \|R_{k+1}\sigma\|_{III}.$$

Значит, $\|R_{k+1}\|_{III} \leq 1$, что и требовалось установить.

Мы описали метод матричной прогонки для случая, когда направление прогонки выбирается совпадающим

с направлением оси Ox . Такое направление особенно выгодно брать при условии $M \gg N$, так как основной объем вычислений в методе прогонки падает на определение матриц R_{m+1} по формуле $R_{m+1} = -(A + R_m)^{-1}$; этот объем будет тем меньше, чем меньше порядок матриц $A + R_m$. При условии, что $M \gg N$, выгоднее обратить M матриц порядка $N - 1$, чем N матриц порядка $M - 1$. Если задача (33) такова, что $N \gg M$, то алгоритм метода прогонки следует видоизменить, взяв за направление прогонки ось Oy .

6. Правило Рунге. Укажем здесь один практический прием, позволяющий на основе вычислений судить о том, с какой точностью получены приближенные сеточные значения решения.

Пусть $u(x, y)$ — точное решение некоторой граничной задачи, а $u_h(x, y)$ — приближенное решение этой задачи, полученное по методу сеток с шагами h и l , $l/h = \text{const}$. В методе сеток часто известен порядок относительно h погрешности

$$\varepsilon_h(x, y) = u(x, y) - u_h(x, y). \quad (38)$$

Предположим, что для $\varepsilon_h(x, y)$ имеет место представление $\varepsilon_h(x, y) = K(x, y) \cdot h^p$, верное с точностью до величин порядка $O(h^{p+1})$, где $K(x, y)$ — некоторая положительная, ограниченная в области задания граничной задачи функция, p — положительное число. При шаге, в два раза большем h , получим

$$\varepsilon_{2h}(x, y) = u(x, y) - u_{2h}(x, y), \quad (39)$$

$$\varepsilon_{2h}(x, y) = K(x, y) \cdot (2h)^p = 2^p \varepsilon_h(x, y). \quad (40)$$

Из формул (38)–(40) можно вывести правило для определения $\varepsilon_h(x, y)$ через сеточные значения $u_h(x, y)$ и $u_{2h}(x, y)$

$$\varepsilon_h(x, y) = \frac{u_h(x, y) - u_{2h}(x, y)}{2^p - 1}. \quad (41)$$

Формула (41) носит приближенный характер, ибо приближенной была формула $\varepsilon_h(x, y) = K(x, y) h^p$, на основе которой получено выражение (41). Достоинство такого

выражения для $\varepsilon_h(x, y)$ состоит в том, что эту величину можно реально вычислить. Можно ожидать, что значение

$$\tilde{u}_h(x, y) = u_h(x, y) + \frac{u_h(x, y) - u_{2h}(x, y)}{2^p - 1}$$

будет более точным, чем $u_h(x, y)$. В этом и состоит правило Рунге для уточнения сеточных значений $u_h(x, y)$. В реальных вычислениях часто поступают следующим образом. Находят решение при шаге h , потом вычисляют решение при шаге $2h$ и сравнивают значения $u_h(x, y)$ и $u_{2h}(x, y)$ в одинаковых узлах. Если эти значения совпадают на заданном числе знаков, то решение $u_h(x, y)$ принимают за искоемое. Если такого совпадения нет, то шаг h делят пополам и вычисляют решение $u_{h/2}(x, y)$. Далее контроль точности в вычислении значений $u_{h/2}(x, y)$ осуществляют аналогично предыдущему.

7. Метод итераций для разностной задачи Дирихле. Здесь мы рассмотрим итерационные методы, приспособленные для решения разностных схем вида (20). Основное внимание будет уделено наиболее простым вариантам *метода установления* [3, 6].

Как мы видели, разностная схема (20), которую мы перепишем в виде

$$\begin{aligned} \Lambda_{xx}(u_{mn}) + \Lambda_{yy}(u_{mn}) &= f(x_m, y_n), \\ m &= 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1, \\ u_{mn} &= \varphi(x_m, y_n), \quad (mh, nh) \in \Gamma_h, \end{aligned} \quad (42)$$

где

$$\begin{aligned} \Lambda_{xx}(u_{mn}) &= \frac{1}{h^2} (u_{m+1n} - 2u_{mn} + u_{m-1n}), \\ \Lambda_{yy}(u_{mn}) &= \frac{1}{j^2} (u_{mn+1} - 2u_{mn} + u_{mn-1}), \end{aligned}$$

аппроксимирует задачу Дирихле для уравнения Пуассона

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= f(x, y), \quad (x, y) \in D = \{0 < x < a, 0 < y < b\}, \\ u|_{\Gamma} &= \varphi(Q), \quad Q - \text{точка контура } \Gamma. \end{aligned} \quad (43)$$

Сопоставим задачу (43) с родственной ей нестационарной задачей о распространении тепла

$$\begin{aligned} \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} - f(x, y), \quad (x, y) \in D, \quad t > 0, \\ v|_{\Gamma} &= \varphi(Q), \quad Q - \text{точка контура } \Gamma, \\ v(x, y, 0) &= \psi(x, y), \quad (x, y) \in D, \end{aligned} \quad (44)$$

где $\psi(x, y)$ — произвольная функция, выбираемая так, чтобы она возможно мало отличалась от решения задачи (43).

В задаче (44) источник тепла $f(x, y)$ и температура на границе $\varphi(Q)$ не зависят от времени t . Естественно ожидать поэтому, что при $t \rightarrow \infty$ будет выполняться соотношение $\frac{\partial v}{\partial t} \rightarrow 0$, откуда

$$\lim_{t \rightarrow \infty} v(t, x, y) = u(x, y),$$

поскольку в этом случае задачи (43) и (44) совпадают. Можно предположить, что для достаточно больших значений t , например для $t > T$, будет с необходимой точностью верно приближенное равенство

$$v(t, x, y) \approx u(x, y).$$

На этой закономерности основана идея метода решения стационарных задач, состоящая в замене их подходящими нестационарными задачами. Этот метод и ряд его модификаций принято называть *методом установления*.

Запишем для задачи (44) простейшую двухслойную разностную схему (см., например, (2.52)):

$$\begin{aligned} \frac{v_{mn}^{p+1} - v_{mn}^p}{\tau} &= \Lambda_{xx}(v_{mn}^p) + \Lambda_{yy}(v_{mn}^p) - f(x_m, y_n), \\ m &= 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1, \\ p &= 0, 1, 2, \dots, \\ v_{mn}^{p+1} &= \varphi(x_m, y_n), \quad (mh, nl) \in \Gamma_h, \\ v_{mn}^0 &= \psi(x_m, y_n), \quad m = 1, 2, \dots, M-1, \\ n &= 1, 2, \dots, N-1, \end{aligned} \quad (45)$$

и двухслойную разностную схему переменных направлений (см., например, разностную схему (2.54))

$$\begin{aligned}\frac{w_{mn} - v_{mn}^p}{\tau} &= \frac{1}{2} [\Lambda_{xx}(w_{mn}) + \Lambda_{yy}(v_{mn}^p) - f(x_m, y_n)], \\ \frac{v_{mn}^{p+1} - w_{mn}}{\tau} &= \frac{1}{2} [\Lambda_{xx}(v_{mn}^p) + \Lambda_{yy}(w_{mn}) - f(x_m, y_n)], \\ m &= 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1, \\ p &= 0, 1, 2, \dots, \\ v_{mn}^{p+1} &= w_{mn} = \varphi(x_m, y_n), \quad (mh, nl) \in \Gamma_h, \\ v_{mn}^0 &= \psi(x_m, y_n), \quad m = 1, 2, \dots, M-1, \\ n &= 1, 2, \dots, N-1.\end{aligned}\tag{46}$$

Доопределим функцию $\psi(x, y)$ так, чтобы при $(mh, nl) \in \Gamma_h$ выполнялось условие

$$\psi(x_m, y_n) = \varphi(x_m, y_n).$$

Обратим внимание на преимущества, которые имеют разностные схемы (45) и (46) перед разностной схемой (42). Последняя представляет собой систему линейных алгебраических уравнений с числом неизвестных $(M-1)(N-1)$. При больших M и N решение таких систем на ЭВМ представляет собой довольно трудную задачу. В то же время вычисления по формуле (45) весьма просты и носят несложный рекурсионный характер. Вычисления по формуле (46) сложнее, чем вычисления по формуле (45), но и они намного проще вычислений, которые необходимо проделать в задаче (42). Это станет очевидным, если мы при решении задачи (46) воспользуемся методом разностной прогонки по направлению переменной x для вычисления значений $\{w_{mn}\}$ и методом разностной прогонки по направлению переменной y для вычисления значений $\{v_{mn}^p\}$. Схему (45) будем называть простейшим явным методом установления, а схему (46) — методом переменных направлений.

Остановимся на выявлении условий сходимости явного метода установления (45).

Пусть u_{mn} — решение разностной схемы (42), а v_{mn}^p — решение разностной схемы (45) при некотором фиксиро-

ванном p . Положим $\varepsilon_{mn}^p = v_{mn}^p - u_{mn}$ и определим норму по правилу

$$\|\varepsilon^p\| = \left(\sum_{m,n} |\varepsilon_{mn}^p|^2 \right)^{1/2}.$$

Отметим, что по условию $\varepsilon_{mn}^p|_{\Gamma_h} = 0$ при всех $p = 0, 1, 2, \dots$. Выясним, когда и с какой скоростью $\|\varepsilon^p\| \rightarrow 0$ при $p \rightarrow \infty$.

Если в разностной схеме (42) уравнение записать в эквивалентном виде

$$0 = \frac{u_{mn} - u_{mn}}{\tau} = \Lambda_{xx}(u_{mn}) + \Lambda_{yy}(u_{mn}) - f(x_m, y_n),$$

то на основании этой формулы и формул (42) и (45) можно получить разностную схему для погрешности ε_{mn}^p :

$$\begin{aligned} \frac{\varepsilon_{mn}^{p+1} - \varepsilon_{mn}^p}{\tau} &= \Lambda_{xx}(\varepsilon_{mn}^p) + \Lambda_{yy}(\varepsilon_{mn}^p), \\ m &= 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1, \\ p &= 0, 1, 2, \dots, \\ \varepsilon_{mn}^{p+1} &= 0, \quad (mh, nl) \in \Gamma_h, \\ \varepsilon_{mn}^0 &= \psi(x_m, y_n) - u_{mn}, \\ m &= 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1. \end{aligned} \quad (47)$$

В целях упрощения анализа задачи (47) будем считать, что $a=b=1$, $h=l$ и, следовательно, $M=N$.

Решение разностной схемы (47) можно выписать в явном виде, если воспользоваться собственными значениями λ_{kl} и собственными функциями $\psi_{mn}^{(k,l)}$ разностного оператора.

Известно ([3], § 27), что

$$\begin{aligned} \lambda_{kl} &= 1 - \frac{4\tau}{h^2} \left(\sin^2 \frac{k\pi}{2M} + \sin^2 \frac{l\pi}{2M} \right), \quad k = 1, \dots, M, \quad l = 1, \dots, M, \\ \psi_{mn}^{(kl)} &= 2 \sin \frac{k\pi m}{M} \sin \frac{l\pi n}{M} \quad m = 0, 1, \dots, M, \quad n = 0, 1, \dots, M, \end{aligned}$$

а функции $\psi_{mn}^{(kl)}$ образуют ортонормированный базис. Решение схемы (47) на слое номера p обозначим через ε^p и там, где в этом не будет особой необходимости, далее

не будем указывать зависимость этой величины от m, n , аналогично будем писать $\psi^{(kl)}$ вместо $\psi_{mn}^{(kl)}$. Имеем

$$\varepsilon^p = \sum_{k, l=1}^{M-1} c_{kl} \lambda_{kl}^p \psi^{(k, l)}, \quad (48)$$

где c_{kl} — некоторые постоянные.

Функции $\psi^{(k, l)}$ выбраны таким образом, чтобы выполнялось условие $\psi_{mn}^{(k, l)}|_{\Gamma_h} = 0$. Поэтому будет $\varepsilon_{mn}^p|_{\Gamma_h} = 0$ и, значит, решение ε^p , определяемое по формуле (48), удовлетворяет разностному уравнению и граничному условию, фигурирующим в (47). Чтобы выполнялось еще и начальное условие $\varepsilon_{mn}^0 = \psi(x_m, y_n) - u_{mn}$, т. е.

$$\sum_{kl}^{M-1} c_{kl} \lambda_{kl}^0 \psi_{mn}^{(kl)} = \psi(x_m, y_n) - u_{mn},$$

нужно в качестве постоянных c_{kl} взять разностные коэффициенты Фурье для функции $\psi(x_m, y_n) - u_{mn}$. Таким образом, получим

$$c_{kl} = h^2 \sum_{m, n=0}^M (\psi(x_m, y_n) - u_{mn}) 2 \sin \frac{k\pi m}{M} \cdot \sin \frac{l\pi n}{M}. \quad (49)$$

Формулу (48) можно записать в виде

$$\varepsilon^p = \sum_{k, l=1}^{M-1} \mu_{kl}^p \psi^{(k, l)},$$

где $\mu_{kl}^p = c_{kl} \cdot \lambda_{kl}^p$. Из этой записи видно, что числа μ_{kl}^p являются коэффициентами Фурье при разложении погрешности ε^p по функциям $\psi^{(k, l)}$. Поэтому

$$(\varepsilon^p, \varepsilon^p) = \sum_{k, l=1}^{M-1} |c_{kl} \lambda_{kl}^p|^2$$

и, значит,

$$\begin{aligned} \|\varepsilon^p\| &= \left(\sum_{k, l=1}^{M-1} |c_{kl} \lambda_{kl}^p|^2 \right)^{1/2}, \\ \|\varepsilon^0\| &= \left(\sum_{k, l=1}^{M-1} |c_{kl}|^2 \right)^{1/2}. \end{aligned} \quad (50)$$

Обозначим $\max_{k, l} |\lambda_{kl}|$ через λ ; тогда $|\lambda_{kl}| \leq \lambda$.

На основании (50) получим

$$\|\varepsilon^p\| \leq \lambda^p \|\varepsilon^0\|. \quad (51)$$

Если $\lambda < 1$, то из формулы (51) следует, что $\|\varepsilon^p\| \rightarrow 0$ при $p \rightarrow \infty$; это стремление $\|\varepsilon^p\|$ к нулю тем быстрее, чем меньше λ . По определению λ — функция от τ , т. е. $\lambda = \lambda(\tau)$, так как

$$\max_{k, l} |\lambda_{kl}| = \max_{k, l} \left| 1 - \frac{4\tau}{h^2} \left(\sin^2 \frac{k\pi}{2M} + \sin^2 \frac{l\pi}{2M} \right) \right| = \lambda(\tau).$$

Выберем τ таким образом, чтобы $\lambda(\tau)$ имело наименьшее значение. Обозначим

$$\rho(k, l) = \sin^2 \frac{k\pi}{2M} + \sin^2 \frac{l\pi}{2M}, \quad 1 \leq k, l \leq M-1.$$

Очевидно, что

$$2 \sin^2 \frac{\pi}{2M} \leq \rho(k, l) \leq 2 \cos^2 \frac{\pi}{2M}. \quad (52)$$

Рассмотрим семейство функций

$$\sigma_{kl}(\tau) = 1 - \frac{4}{h^2} \rho(k, l) \tau$$

при $1 \leq k, l \leq M-1$ и $\tau > 0$. Все графики функций этого семейства расположены в области P , заключенной

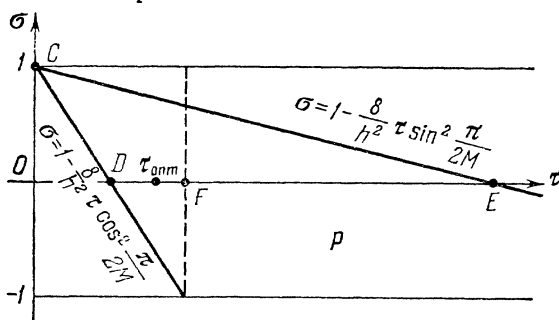


Рис. 6.

между двумя прямыми, проходящими через точки C , D и C , E (рис. 6). Здесь $C(0, 1)$, $D(\tau^*, 0)$,

$$\tau^* = \frac{h^2}{8 \cos^2 \frac{\pi}{2M}}, \quad E(\tau^{**}, 0),$$

$$\tau^{**} = \frac{h^2}{8 \sin^2 \frac{\pi}{2M}}, \quad F(\tau_0, 0), \quad \tau_0 = \frac{h^2}{4 \cos^2 \frac{\pi}{2M}}.$$

Из рис. 6 видно, что подходящие значения τ расположены в интервале $0 < \tau < \tau_0$. Если $\tau \geq \tau_0$, то возможны значения $|\sigma(\tau)| \geq 1$. Оптимальное значение τ должно находиться в интервале $\tau^* < \tau < \tau_0$, так как на этом интервале функция $\sigma(\tau) = 1 - \frac{8}{h^2} \tau \sin^2 \frac{\pi}{2M}$ убывает и $\sigma(\tau) < 1$, а функция $|\sigma(\tau)| = \left| 1 - \frac{8}{h^2} \tau \cos^2 \frac{\pi}{2M} \right|$ возрастает и $|\sigma(\tau)| < 1$. Поэтому оптимальное значение $\tau_{\text{опт}}$ может быть определено из условия равенства

$$1 - \frac{8}{h^2} \tau \sin^2 \frac{\pi}{2M} = \left| 1 - \frac{8}{h^2} \tau \cos^2 \frac{\pi}{2M} \right|$$

на отрезке $\tau^* < \tau < \tau_0$. Простые вычисления дают

$$\tau_{\text{опт}} = h^2/4.$$

Из формулы $\lambda_{kl} = 1 - \frac{4}{h^2} \rho(k, l) \tau$ при $\tau = \tau_{\text{опт}}$ получим

$$\lambda_{\text{опт}} = \max_{k, l} |\lambda_{kl}| \Big|_{\tau=\tau_{\text{опт}}} = 1 - 2 \sin^2 \frac{\pi}{2M}.$$

Если теперь мы захотим, чтобы $\|\varepsilon^p\|$ была меньше некоторого наперед заданного числа $\delta > 0$ ($\delta < \|\varepsilon^0\|$), то мы должны число p определить из условия

$$\|\varepsilon^p\| \leq \left(1 - 2 \sin^2 \frac{\pi}{2M} \right)^p \cdot \|\varepsilon^0\| < \delta.$$

Отсюда получим

$$p \geq \frac{\ln \delta - \ln \|\varepsilon^0\|}{\ln \left(1 - 2 \sin^2 \frac{\pi}{2M} \right)}$$

или, в частности,

$$p \approx \frac{2M^2}{\pi^2} (\ln \|\varepsilon^0\| - \ln \delta) = O(M^2).$$

Исследуем теперь сходимость метода переменных направлений, определяемого формулами (46). Как и ранее, в целях упрощения будем считать, что $a = b = 1$, $h = l$ и $M = N$.

Положим $\varepsilon_{mn}^p = v_{mn}^p - u_{mn}$, $\delta_{mn} = \omega_{mn} - u_{mn}$. Подобно формулам (47), на основании схем (42) и (46) можно

записать задачу для определения погрешностей ε_{mn}^p и δ_{mn} :

$$\begin{aligned}\frac{\delta_{mn} - \varepsilon_{mn}^p}{\tau} &= \Lambda_{xx}(\delta_{mn}) + \Lambda_{yy}(\varepsilon_{mn}^p), \\ \frac{\varepsilon_{mn}^{p+1} - \delta_{mn}}{\tau} &= \Lambda_{xx}(\delta_{mn}) + \Lambda_{yy}(\varepsilon_{mn}^{p+1}),\end{aligned}\quad (53)$$

$$\delta_{mn}|_{\Gamma_h} = \varepsilon_{mn}^p|_{\Gamma_h} = 0,$$

$$\varepsilon_{mn}^0 = \psi(x_m, y_n) - u_{mn},$$

$$m, n = 1, 2, \dots, M-1, \quad p = 0, 1, 2, \dots$$

Для погрешности ε_{mn}^p в этом случае имеет место прежнее представление, даваемое формулой (48), с той лишь разницей, что теперь собственные значения λ_{kl} вычисляются по другому правилу:

$$\lambda_{kl} = \frac{\left(1 - 4\tau M^2 \sin^2 \frac{\pi k}{2M}\right) \left(1 - 4\tau M^2 \sin^2 \frac{\pi l}{2M}\right)}{\left(1 + 4\tau M^2 \sin^2 \frac{\pi k}{2M}\right) \left(1 + 4\tau M^2 \sin^2 \frac{\pi l}{2M}\right)}. \quad (54)$$

Обозначим $q = \max_{k, l} |\lambda_{kl}|$ и по аналогии с (51) получим

$$\|\varepsilon^p\| \leq q^p \|\varepsilon^0\|. \quad (55)$$

В силу формулы (54) должно быть $|\lambda_{kl}| < 1$. Значит, $q < 1$. Поэтому из (55) следует, что $\lim_{p \rightarrow \infty} \|\varepsilon^p\| = 0$. Сходимость метода переменных направлений установлена. Как видно, скорость этой сходимости зависит от величины q и особенно от того, насколько q меньше 1. В связи с этим вновь встает задача выбора τ таким образом, чтобы $\max_{k, l} |\lambda_{kl}|$ был возможно меньшим. Обозначим

$$\rho_k = \frac{1 - 4\tau M^2 \sin^2 \frac{\pi k}{2M}}{1 + 4\tau M^2 \sin^2 \frac{\pi k}{2M}}, \quad k = 1, 2, \dots, M-1.$$

Тогда $\lambda_{kl} = \rho_k \rho_l$. Поскольку

$$\sin^2 \frac{\pi}{2M} \leq \sin^2 \frac{\pi k}{2M} \leq \cos^2 \frac{\pi}{2M},$$

то

$$\rho_{M-1} \leq \rho_k \leq \rho_1.$$

Отметим еще, что

$$\max_{k, l} |\lambda_{kl}| = \max_k |\rho_k| \cdot \max_l |\rho_l| = \left(\max_k |\rho_k| \right)^2 = \rho_1^2.$$

Здесь $\rho_1 = \rho_1(\tau) = \max_k |\rho_k|$. Выберем τ таким образом,

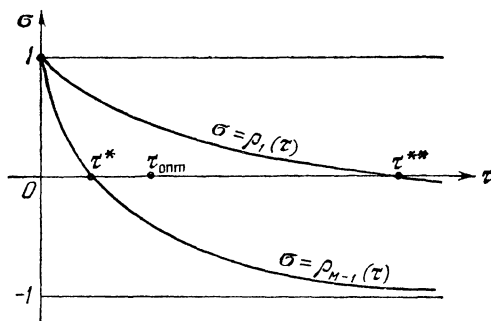


Рис. 7.

чтобы $\rho_1(\tau)$ было наименьшим. Рассмотрим графики функций (см. рис. 7)

$$\rho_k(\tau) = \frac{1 - 4\tau M^2 \sin^2 \frac{\pi k}{2M}}{1 + 4\tau M^2 \sin^2 \frac{\pi k}{2M}}.$$

Здесь

$$\tau^* = \frac{1}{4M^2 \cos^2 \frac{\pi}{2M}}, \quad \tau^{**} = \frac{1}{4M^2 \sin^2 \frac{\pi}{2M}}.$$

Из этого рисунка видно, что $\tau^* < \tau_{\text{опт}} < \tau^{**}$ и при $\tau_0 = \tau_{\text{опт}}$ должно выполняться соотношение

$$\rho_1(\tau_0) = -\rho_{M-1}(\tau_0).$$

Значит, для определения τ_0 получается уравнение

$$\frac{1 - 4\tau_0 M^2 \sin^2 \frac{\pi}{2M}}{1 + 4\tau_0 M^2 \sin^2 \frac{\pi}{2M}} = - \frac{1 - 4\tau_0 M^2 \cos^2 \frac{\pi}{2M}}{1 + 4\tau_0 M^2 \cos^2 \frac{\pi}{2M}},$$

откуда

$$\tau_0 \approx \frac{1}{\sqrt{2} \pi M}.$$

При $\tau = \tau_0$ имеем

$$\max_{k, l} |\lambda_{kl}| = 1 - \frac{\sqrt{2}\pi}{M} + O\left(\frac{1}{M^2}\right).$$

Если теперь мы захотим, используя метод переменных направлений (46), добиться того, чтобы норма $\|\varepsilon^p\|$ не превосходила наперед заданного δ ($\delta < \|\varepsilon^0\|$), то необходимое для этого число итераций p определяется из условия

$$\|\varepsilon^p\| = \left(1 - \frac{\sqrt{2}\pi}{M}\right)^p \|\varepsilon^0\| \leq \delta.$$

Отсюда получим оценку для p

$$p \geq \frac{\ln \|\varepsilon^0\| - \ln \delta}{-\ln\left(1 - \frac{\sqrt{2}\pi}{M}\right)} \approx \frac{M}{\sqrt{2}\pi} (\ln \|\varepsilon^0\| - \ln \delta) = O(M).$$

Отметим, что в методе установления при применении явной схемы для достижения такой точности требовалось выполнение $O(M^2)$ итераций. Поэтому, если M велико, то метод с применением переменных направлений будет выгоднее метода установления с явной схемой.

Искомое решение задачи (43) с помощью разностной схемы (42) может быть вычислено с точностью $O(h^2) = O\left(\frac{1}{M^2}\right)$. Поэтому и число δ следует задавать согласованным с этой точностью. Во всяком случае брать его по порядку меньшим величины $O\left(\frac{1}{M^2}\right)$ бессмысленно.

Если шаг τ периодически менять, чтобы удерживать вблизи нулевой точки значения λ_{kl} , то можно добиться ускорения сходимости метода установления. На этой идее основано несколько модификаций метода установления; со многими из них можно ознакомиться по книгам [3, 6].

§ 4. Разностные схемы для уравнений гиперболического типа

Будем рассматривать линейные дифференциальные уравнения 2-го порядка гиперболического типа. Такие уравнения часто встречаются в физических задачах, связанных с переносом энергии и с процессами колебаний. Метод сеток для уравнений гиперболического типа по своей структуре имеет много общего с методом сеток для уравнений эллиптического и параболического типов.

Мы уделим основное внимание типичному и наиболее простому по виду уравнению с постоянными коэффициентами гиперболического типа — волновому уравнению

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad (1)$$

где $f(x, y)$ — известная функция, которую мы будем считать дважды непрерывно дифференцируемой в некоторой нижеуказываемой области D .

В связи с уравнением (1) будем рассматривать следующие задачи.

Задача Коши. В области $D = \{y > 0, -\infty < x < +\infty\}$ найти дважды непрерывно дифференцируемую функцию $u(x, y)$, которая в этой области удовлетворяла бы уравнению (1), а на прямой $y = 0$ — начальным условиям

$$u(x, 0) = \varphi(x), \quad \left. \frac{\partial u}{\partial y} \right|_{y=0} = \psi(x), \quad (2)$$

где φ и ψ — заданные функции.

Смешанная граничная задача. В области $D = \{y > 0, \alpha < x < \beta\}$ найти дважды непрерывно дифференцируемую функцию $u(x, y)$, которая в этой области удовлетворяла бы уравнению (1), а на границе Γ области D при $y = 0$ — начальным условиям (2) и при $x = \alpha$, $x = \beta$ — одному из трех граничных условий:

а) (условиям первого рода)

$$u(\alpha, y) = \mu_1(y), \quad u(\beta, y) = \mu_2(y); \quad (3)$$

б) (условиям второго рода)

$$\left. \frac{\partial u}{\partial x} \right|_{x=\alpha} = \sigma_1(y), \quad \left. \frac{\partial u}{\partial x} \right|_{x=\beta} = \sigma_2(y); \quad (4)$$

в) (условиям третьего рода)

$$\begin{aligned} \left[\tau_1(y) \frac{\partial u}{\partial x} + \tau_2(y) u \right]_{x=\alpha} &= \omega_1(y), \\ \left[\delta_1(y) \frac{\partial u}{\partial x} + \delta_2(y) u \right]_{x=\beta} &= \omega_2(y); \end{aligned} \quad (5)$$

где $\mu_i(y)$, $\sigma_i(y)$, $\tau_i(y)$, $\delta_i(y)$, $\omega_i(y)$, $i = 1, 2$, — известные функции. Коэффициенты $\tau_i(y)$ и $\delta_i(y)$ подчинены естественным условиям

$$|\tau_1(y)| + |\tau_2(y)| > 0, \quad |\delta_1(y)| + |\delta_2(y)| > 0.$$

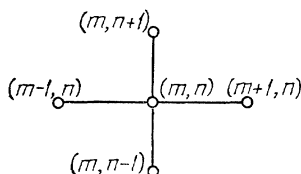
Отметим, что условия (3) и (4) являются частным случаем условий (5).

1. Решение задачи Коши. Рассмотрим метод сеток решения задачи Коши для дифференциального уравнения (1) с начальными условиями (2).

Выберем прямоугольную сетку, положив

$$\begin{aligned} x_m &= mh, \quad m = 0, \pm 1, \pm 2, \dots, \\ y_n &= nl, \quad n = 0, 1, 2, \dots, \quad h > 0, \quad l > 0. \end{aligned}$$

Рассмотрим трехслойный пятиточечный шаблон



и в соответствии с этим шаблоном определим множество D_h° внутренних узлов и множество Γ_h граничных узлов. К множеству D_h° отнесем узлы $(x_m, y_n) \in D$, к множеству Γ_h отнесем узлы, лежащие на прямой $y=0$. Вся сеточная область $D_h = D_h^\circ + \Gamma_h$ будет состоять из узлов $(x_m, y_n) \in \bar{D} = D + \Gamma$.

Используя взятый шаблон и формулу (2.6), а также аналогичную формулу для замены $\frac{\partial^2 u}{\partial y^2}$, получим разностную схему

$$L_h(u^{(h)}) = f^{(h)}, \quad (6)$$

где

$$L_h(u^{(h)}) \equiv \begin{cases} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} - \frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{l^2}, \\ u_m^0, \\ \frac{u_m^1 - u_m^0}{l}, \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 1, 2, \dots, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} f(x_m, y_n), \\ \varphi(x_m), \\ \psi(x_m), \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 1, 2, \dots, \end{cases}$$

Эта схема аппроксимирует уравнение (1) с погрешностью порядка $O(h^2 + l^2)$, а начальные условия — с погрешностью порядка $O(l)$. Действительно, на основании формул (2.3) и (2.6) и по определению аппроксимации получим

$$L_h(u_h(x, y)) = f^{(h)} + \delta f^{(h)},$$

где через $u_h(x, y)$ обозначены значения точного решения задачи (1), (2) на сетке D_h , а для $\delta f^{(h)}$ имеет место представление:

$$\delta f^{(h)} \equiv \begin{cases} \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4} \Big|_{(x_m^{(1)}, y_n)} - \frac{l^2}{12} \frac{\partial^4 u}{\partial y^4} \Big|_{(x_m, y_n^{(1)})}, \\ x_{m-1} < x_m^{(1)} < x_{m+1}, \quad y_{n-1} < y_n^{(1)} < y_{n+1}, \\ 0, \\ \frac{l}{2} \frac{\partial^2 u}{\partial y^2} \Big|_{(x_m, y_0^{(1)})}, \quad y_0 < y_0^{(1)} < y_1, \\ m = 0, \pm 1, \pm 2, \dots, \quad n = 1, 2, \dots \end{cases}$$

Порядок аппроксимации начальных условий можно повысить. Для этой цели заменим $\frac{\partial u}{\partial y} \Big|_{(x_m, 0)}$ по формуле

$$\frac{\partial u}{\partial y} \Big|_{(x_m, 0)} = \frac{u(x_m, y_1) - u(x_m, y_{-1})}{2l} - \frac{l^2}{6} \cdot \frac{\partial^3 u}{\partial y^3} \Big|_{(x_m, y_0^{(2)})}, \quad (7)$$

где $y_{-1} = -l$, $-l < y_0^{(2)} < l$. При этом мы предполагаем, что решение задачи (1), (2) продолжимо в область $G = \{-l \leq y \leq 0, -\infty < x < +\infty\}$. Из формулы (7) получим сеточное условие, аппроксимирующее начальное условие $\frac{\partial u}{\partial y} \Big|_{y=0} = \psi(x)$ с погрешностью порядка $O(l^2)$:

$$\frac{u_m^1 - u_m^{-1}}{2l} = \psi(x_m), \quad m = 0, \pm 1, \pm 2, \dots \quad (8)$$

Значение u_m^{-1} (8) можно исключить с помощью разностного уравнения из (8) при $n = 0$. Имеем

$$\frac{u_{m+1}^0 - 2u_m^0 + u_{m-1}^0}{h^2} - \frac{u_m^1 - 2u_m^0 + u_m^{-1}}{l^2} = f(x_m, y_0),$$

и потому (8) можно преобразовать к виду

$$\frac{u_m^1 - u_m^0}{l} = \psi(x_m) + \frac{1}{2} l \Delta_{xx}(u_m^0) - \frac{1}{2} l f(x_m, y_0), \quad (9)$$

где

$$\Lambda_{xx}(u_m^0) \equiv \frac{u_{m+1}^0 - 2u_m^0 + u_{m-1}^0}{h^2}.$$

Таким образом, вместо разностной схемы (6) можно записать другую разностную схему, аппроксимирующую задачу (1), (2) с погрешностью порядка $O(h^2 + l^2)$:

$$L_h(u^{(h)}) = f^{(h)}, \quad (10)$$

где оператор L_h вычисляется так же, как и в формуле (6), а

$$f^{(h)} \equiv \begin{cases} f(x_m, y_n), \\ \varphi(x_m), \\ \psi(x_m) + \frac{1}{2} l \Lambda_{xx}(u_m^0) - \frac{1}{2} l f(x_m, 0), \\ m=0, \pm 1, \pm 2, \dots, n=1, 2, \dots \end{cases}$$

Остановимся на вопросе численной реализации схемы (10). Покажем, как по значениям u_m^0 и u_m^1 вычислить значения u_m^2 .

В силу (6) и (10) имеем

$$u_m^0 = \varphi(x_m),$$

$$u_m^1 = \varphi(x_m) + l \left(\psi(x_m) + \frac{1}{2} l \Lambda_{xx}(\varphi(x_m)) - \frac{1}{2} l f(x_m, 0) \right), \quad (11)$$

$$m=0, \pm 1, \pm 2, \dots$$

Разностное уравнение в схеме (10) перепишем в виде

$$u_m^{n+1} = 2u_m^n + l^2 \Lambda_{xx}(u_m^n) - u_m^{n-1} - l^2 f(x_m, y_n), \quad (12)$$

$$m=0, \pm 1, \pm 2, \dots, n=1, 2, \dots$$

При $n=1$ по формуле (12) вычислим значения u_m^2 , $m=0, \pm 1, \pm 2, \dots$; значения u_m^0 и u_m^1 известны в силу (11). Затем по (12) при $n=2$ вычисляем значения u_m^3 через уже известные значения u_m^1 , u_m^2 и т. д.

В случае параболических уравнений мы видели, что численная реализация явных разностных схем связана с жестким ограничением на шаг τ по времени. Можно ожидать, что и численная реализация явной разностной схемы (10) возможна лишь при определенных ограничениях на h и l . Покажем, каким может быть отношение $l/h = \gamma > 0$ в случае гиперболического уравнения.

Рассмотрим однородное уравнение, соответствующее уравнению (1),

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0. \quad (13)$$

Присоединим к этому уравнению начальные условия (2). Разностную схему (6) применительно к задаче (13), (2) можно записать в виде

$$u_m^{n+1} = \gamma^2 u_{m+1}^n + (2 - 2\gamma^2) u_m^n + \gamma^2 u_{m-1}^n - u_m^{n-1}, \quad (14)$$

$$u_m^0 = \varphi(x_m), \quad u_m^1 = \varphi(x_m) + l\psi(x_m),$$

$$m = 0, \pm 1, \pm 2, \dots, n = 1, 2, \dots$$

Предположим, что в узле S с координатами (x_i, y_i) , где i — любое целое число, $j \geq 2$, нужно вычислить значение u_i^j , пользуясь формулами (14). Положим в (14) $n = j - 1$, тогда мы видим, что u_i^j определяется значениями u_{i+1}^{j-1} , u_i^{j-1} , u_{i-1}^{j-1} и u_i^{j-2} . Если $j - 2 > 1$, то значения u_{i+1}^{j-1} , u_i^{j-1} , u_{i-1}^{j-1} , u_i^{j-2} в свою очередь по формуле (14) могут быть выражены через значения на низших слоях

$$u_{i+2}^{j-2}, u_{i+1}^{j-2}, u_i^{j-2}, u_{i-1}^{j-2}, u_{i-2}^{j-2}, u_{i+1}^{j-3}, u_i^{j-3}, u_{i-1}^{j-3}, u_i^{j-4}.$$

В конечном счете значение u_i^j может быть выражено через значения u_m^0 ($m = i + s$, $s = 0, \pm 1, \dots, \pm j - 2$) и u_m^1 ($m = i + s$, $s = 0, \pm 1, \dots, \pm j - 1$). Все эти значения

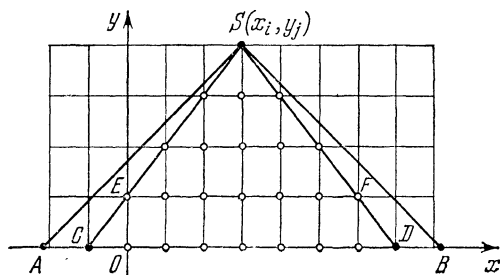


Рис. 8.

находятся внутри треугольника (рис. 8), вершина которого лежит в точке S , одна сторона лежит на прямой Ox , две другие стороны проходят через точку S и образуют с осью Ox углы SDC и SCD , равные соответственно $\arctg(-\gamma)$

и $\operatorname{arctg} \gamma$, $\gamma = l/h = \operatorname{const} > 0$. Треугольник SCD называют *треугольником определенности разностной схемы* (14).

Таким образом, значение u_i^j в точке S определяется разностным уравнением (14) и начальными значениями u_m^0 и u_m^1 , лежащими на отрезках CD и EF .

Пусть теперь мы хотим вычислить в этой же точке S точное значение решения $u(x, y)$ задачи (13), (2). Из теории уравнений математической физики известно, что значение решения $u(x, y)$ в точке S определяется уравнением (13) и начальными условиями, содержащимися на отрезке, высекаемом характеристиками, проходящими через точку S , на прямой $y=0$, т. е. начальными условиями на отрезке AB . Характеристики уравнения (13) перпендикулярны друг другу и образуют с осью Ox углы, равные $\pi/4$ и $3\pi/4$. Это — углы SAB и SBx . Треугольник SAB называют *треугольником определенности дифференциального уравнения* (13).

На рис. 8 изображен случай, когда $\angle SAB < \angle SCD$ и $\operatorname{tg}(\angle SCD) = \gamma = l/h > 1$. Такой выбор соотношения шагов ($l/h = \gamma = \operatorname{const} > 1$) оказывается неудачным по следующей причине. В этом случае треугольник определенности разностного уравнения содержится внутри треугольника определенности дифференциального уравнения, поэтому начальных условий, задаваемых на отрезке CD , будет недостаточно для определения решения в точке S . Если мы, например, изменим как-то начальные условия на отрезках AC и DB , то это может изменить решение задачи (13), (2) во всей области D , в том числе и в точке S . Однако сеточное значение u_i^j в точке S от таких изменений не будет зависеть и остается прежним. Значит, при $l/h = \gamma = \operatorname{const} > 1$ нельзя надеяться на сходимость решения разностной схемы (14) к решению задачи Коши для уравнения (13). Поскольку разностная схема (14) аппроксимирует задачу (13), (2), то она не может быть устойчивой, ибо в противном случае из аппроксимации и устойчивости следовала бы сходимость. Мы показали, что при $\gamma > 1$ сходимости быть не может, значит, не может быть и устойчивости. Можно надеяться на устойчивость при условии, что $\triangle ASB$ содержится внутри $\triangle CSD$. В нашем случае это будет при $l/h = \gamma \leq 1$. Итак, в разностной схеме (14) для устойчивости необходимо наложить условие $\gamma \leq 1$.

Если уравнение гиперболического типа имеет более общий вид, чем (13), то для него, как правило, треугольник определенности будет криволинейным, криволинейным может быть также и треугольник определенности соответствующего разностного уравнения. Однако упомянутая закономерность (она впервые была высказана Курантом, Фридрихсом и Леви), что треугольник определенности дифференциального уравнения содержится в треугольнике определенности разностного уравнения, должна сохраниться, если мы хотим надеяться на устойчивость.

Достаточные условия устойчивости будут рассмотрены ниже.

2. Решение смешанной задачи. Пусть уравнение

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad (15)$$

задано в области

$$D = \{0 < y \leq Y < +\infty, \quad \alpha < x < \beta\}.$$

Будем считать, что к уравнению (15) присоединены начальные условия

$$u(x, 0) = \varphi(x), \quad \frac{\partial u}{\partial y} \Big|_{y=0} = \psi(x), \quad \alpha \leq x \leq \beta, \quad (16)$$

и граничные условия третьего рода

$$\begin{aligned} \left[\tau_1(y) \frac{\partial u}{\partial x} + \tau_2(y) u \right]_{x=\alpha} &= \omega_1(y), \\ \left[\delta_1(y) \frac{\partial u}{\partial x} + \delta_2(y) u \right]_{x=\beta} &= \omega_2(y). \end{aligned} \quad (17)$$

Предположим, что смешанная граничная задача (15) — (17) имеет единственное решение в области $\bar{D} = D + \Gamma$, Γ — контур области D , и это решение вместе с производными до 4-го порядка по x и по y включительно непрерывно в $\bar{D}$.

Выберем прямоугольную равномерную сетку, положив $x_m = mh$, $m = 0, 1, \dots, M$, $h = \frac{\beta - \alpha}{M}$, $y_n = nl$,

$$n = 0, 1, \dots, N, \quad Nl \leq Y < (N+1)l, \quad l > 0.$$

Для замены уравнения (15) разностным воспользуемся, как и в случае задачи Коши, пятиточечным явным трехслойным шаблоном. По этому шаблону сеточную область D_h разобьем на два множества: множество D_h^i внутрен-

них узлов, $D_h^n = \{(x_m, y_n), m = 1, 2, \dots, N-1, n = 1, 2, \dots, N-1\}$, и множество Γ_h граничных узлов, $\Gamma_h = \{(x_m, y_n), m = 0, 1, \dots, M, n = 0; m = 0, n = 0, 1, \dots, N; m = M, n = 0, 1, \dots, N\}$.

На множестве D_h^n уравнение (15) и начальные условия (16) аппроксимируются разностной схемой вида (10) (далее будет приведена ее подробная запись). Для замены граничных условий на прямых $x = \alpha$ и $x = \beta$ воспользуемся формулами типа формул (23) и (24). Получим

$$\tau_{1n} \frac{u_1^n - u_0^n}{h} + \tau_{2n} u_0^n = \omega_{1n}, \quad (18)$$

$$\delta_{1n} \frac{u_M^n - u_{M-1}^n}{h} + \delta_{2n} u_M^n = \omega_{2n}, \quad n = 0, 1, \dots, N,$$

где $\tau_{1n} = \tau_1(y_n)$, $\tau_{2n} = \tau_2(y_n)$ и т. д.

Легко видеть, что разностные граничные условия (18) аппроксимируют граничные условия (17) с погрешностью порядка $O(h)$.

Теперь на основании формул (10) и (18) запишем разностную схему, аппроксимирующую граничную задачу (15) — (17) с погрешностью $O(h + l^2)$:

$$L_h(u^{(h)}) = f^{(h)}, \quad (19)$$

где

$$L_h(u^{(h)}) \equiv \begin{cases} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} - \frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{l^2}, \\ m = 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1, \\ u_m^n, \quad m = 0, 1, \dots, M, \\ \frac{u_m^1 - u_m^0}{l}, \quad m = 0, 1, \dots, M, \\ \tau_{1n} \frac{u_1^n - u_0^n}{h} + \tau_{2n} u_0^n, \quad n = 0, 1, \dots, N, \\ \delta_{1n} \frac{u_M^n - u_{M-1}^n}{h} + \delta_{2n} u_M^n, \quad n = 0, 1, \dots, N, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} f(x_m, y_n), \quad m = 1, 2, \dots, M-1, \quad n = 1, 2, \dots, N-1, \\ \varphi(x_m), \quad m = 0, 1, \dots, M, \\ \psi(x_m) + \frac{1}{2} l \Lambda_{xx}(u_m^n) - \frac{1}{2} f(x_m, 0), \quad m = 0, 1, \dots, M, \\ \omega_{1n}, \quad n = 0, 1, \dots, N, \\ \omega_{2n}, \quad n = 0, 1, \dots, N. \end{cases}$$

Отметим, что если $\tau_1(y) = \delta_1(y) \equiv 0$ и $\tau_2(y) = \delta_2(y) \equiv 1$, то разностная схема (19) аппроксимирует граничную задачу (15) — (17) с погрешностью $O(h^2 + l^2)$.

Численная реализация разностной схемы (19) осуществляется почти так же, как и численная реализация схемы (10). Вначале, используя формулы вида (11), вычисляем значения на нулевом слое u_m^0 , потом значения на первом слое u_m^1 ; в обоих случаях m изменяется в пределах $0 \leq m \leq M$. Далее по уравнению (12) при $n=1$ вычисляем значения $u_1^2, u_2^2, \dots, u_{M-1}^2$. Для вычисления на втором слое значений u_0^3 и u_M^3 воспользуемся разностными граничными условиями (18) при $n=2$. Получим

$$u_0^3 = \frac{\tau_{12}}{\tau_{12} - h\tau_{22}} u_1^2 - \frac{h\omega_{12}}{\tau_{12} - h\tau_{22}},$$

$$u_M^3 = \frac{\delta_{12}}{\delta_{12} + h\delta_{22}} u_{M-1}^2 + \frac{h\omega_{22}}{\delta_{12} + h\delta_{22}},$$

в предположении, что $\tau_{12} - h\tau_{22} \neq 0$ и $\delta_{12} + h\delta_{22} \neq 0$. Таким образом, мы указали, как по значениям u_m^0 и u_m^1 , $m=0, 1, \dots, M$, вычислить значения u_m^2 . Аналогично по значениям u_m^1 , u_m^2 , $m=0, 1, \dots, M$, вычисляются значения u_m^3 и т. д.

3. Об устойчивости явной трехслойной разностной схемы. Исследование устойчивости приведем на примере смешанной граничной задачи (15) — (17) при условии, что $\alpha=0$, $\beta=1$, $\tau_1(y) = \delta_1(y) \equiv 0$, $\tau_2(y) = \delta_2(y) \equiv 1$. В этом случае для разностной схемы (19) правая часть будет прежней, а оператор L_h примет вид

$$L_h(u^{(h)}) \equiv \begin{cases} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} - \frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{l^2}, \\ m=1, 2, \dots, M-1, n=1, 2, \dots, N-1, \\ u_m^0, \quad m=0, 1, \dots, M, \\ \frac{u_m^1 - u_m^0}{l}, \quad m=0, 1, \dots, M, \\ u_0^n, \quad n=0, 1, \dots, N, \\ u_M^n, \quad n=0, 1, \dots, N. \end{cases}$$

Будем исследовать устойчивость разностной схемы $L_h(u^{(h)}) = f^{(h)}$ по начальным условиям. Для этого мы должны установить два свойства:

1) разностная схема $L_h(z^{(h)}) = g^{(h)}$, где

$$g^{(h)} \equiv \begin{cases} 0, \\ \alpha_m^0, & m=0, 1, \dots, M, \\ \beta_m^0 + \frac{1}{2} \Lambda_{xx}(z_m^0), & m=0, 1, \dots, M, \\ 0, \\ 0, \end{cases}$$

однозначно разрешима;

2) верна оценка

$$\|z^{(h)}\|_{U_h} \leq C \cdot \|g^{(h)}\|_{F_h},$$

где C — постоянная, не зависящая от h и $g^{(h)}$. Правило выбора норм в пространствах U_h и F_h будет оговорено позже. При доказательстве устойчивости разностной схемы $L_h(u^{(h)}) = f^{(h)}$ мы будем следовать доказательству, приведенному в [6].

Прежде всего отметим, что рассматриваемая разностная схема явная, и поэтому свойство 1) очевидно.

Для установления свойства 2) разностную схему $L_h(z^{(h)}) = g^{(h)}$ запишем в развернутом и преобразованном виде:

$$\Lambda_{yy}(z_m^n) = \Lambda_{xx}(z_m^n), \quad m=1, 2, \dots, M-1, \\ n=1, 2, \dots, N-1, \quad (20)$$

$$z_m^0 = \alpha_m^0, \quad z_m^1 = \alpha_m^0 + l\gamma_m^0, \quad m=0, 1, \dots, M, \quad (21)$$

где

$$\gamma_m^0 = \beta_m^0 + \frac{1}{2} \Lambda_{xx}(\alpha_m^0), \\ z_0^n = 0, \quad z_M^n = 0, \quad n=0, 1, \dots, N. \quad (22)$$

Решение задачи (20) — (22) будем искать методом разделения переменных. Найдем частные решения уравнения (20) в виде

$$z_m^n(k) = X_m(k) \cdot Y^n(k) \neq 0.$$

После подстановки $z_m^n(k)$ в уравнение (20) и некоторых преобразований получим

$$\frac{\Lambda_{xx}(X_m(k))}{X_m(k)} = \frac{\Lambda_{yy}(Y^n(k))}{Y^n(k)} = -\lambda_k, \quad (23)$$

где λ_k — некоторое число. Из (23) и граничных условий (22) получаем для $X_m(k)$ задачу на собственные значения

$$\Lambda_{xx}(X_m(k)) + \lambda_k X_m(k) = 0, \\ X_0(k) = X_M(k) = 0, \quad X_m(k) \not\equiv 0.$$

Она имеет решения

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{\pi kh}{2}, \quad X_m(k) = \sqrt{2} \sin \pi kmh.$$

Для $Y^n(k)$ получим разностное уравнение второго порядка

$$\Lambda_{yy}(Y^n(k)) + \lambda_k Y^n(k) = 0$$

или

$$Y^{n+1}(k) - (2 - l^2 \lambda_k) Y^n(k) + Y^{n-1}(k) = 0. \quad (24)$$

Для него характеристическое уравнение имеет вид

$$q^2 - (2 - 2\alpha_k)q + 1 = 0,$$

где $\alpha_k = \frac{1}{2} l^2 \lambda_k$. Корни этого уравнения $q_{1,2} = 1 - \alpha_k \pm \pm \sqrt{\alpha_k^2 - 2\alpha_k}$. Если $0 < \alpha_k < 2$, то корни $q_{1,2} = 1 - \alpha_k \pm \pm i \sqrt{\alpha_k(2 - \alpha_k)}$ комплексные и $|q_{1,2}| = 1$. Введем новую переменную σ_k , полагая

$$\cos \sigma_k = 1 - \alpha_k, \quad \sin \sigma_k = \sqrt{\alpha_k(2 - \alpha_k)}.$$

Тогда

$$q_1(k) = e^{i\sigma_k}, \quad q_2(k) = e^{-i\sigma_k}.$$

Общее решение уравнения (24) имеет вид

$$Y^n(k) = C_k (q_1(k))^n + D_k (q_2(k))^n = A_k \cos n\sigma_k + B_k \sin n\sigma_k,$$

где A_k и B_k — произвольные постоянные.

Теперь решение задачи (20) — (24) может быть записано в форме (см., например, п. 3.7)

$$z_m^n = \sum_{k=1}^{M-1} (A_k \cos n\sigma_k + B_k \sin n\sigma_k) \cdot X_m(k). \quad (25)$$

Пусть a_k и b_k — коэффициенты разложений

$$\alpha_m^0 = \sum_{k=1}^{M-1} a_k X_m(k), \quad \beta_m^0 + \frac{1}{2} \Lambda_{xx}(\alpha_m^0) = \sum_{k=1}^{M-1} b_k X_m(k). \quad (26)$$

Потребуем, чтобы z_m^n , определяемое по формулам (25),

удовлетворяло начальным условиям (21). Это возможно при следующих значениях A_k и B_k :

$$A_k = \alpha_k^0, \quad B_k = \frac{1 - \cos \sigma_k}{\sin \sigma_k} \alpha_k^0 + \frac{l}{\sin \sigma_k} \left(\beta_k^0 + \frac{1}{2} \Lambda_{xx}(\alpha_k^0) \right). \quad (27)$$

Подставим эти значения A_k и B_k в (25) и, выполнив преобразования, получим

$$z_m^n = \sum_{k=1}^{M-1} \left(\frac{\cos \left(n - \frac{1}{2} \right) \sigma_k}{\cos \frac{1}{2} \sigma_k} a_k + \frac{l \cdot \sin n \sigma_k}{\sin \sigma_k} b_k \right) X_m(k). \quad (28)$$

Оценим величину $\rho(z^n) = \left[\sum_{m=1}^{M-1} (z_m^n)^2 \right]^{1/2}$. Имеем

$$\cos \sigma_k = 1 - \alpha_k, \quad \sin \sigma_k = \sqrt{\alpha_k(2 - \alpha_k)}, \quad \alpha_k = \frac{1}{2} l^2 \lambda_k.$$

Наложим на шаги h и l ограничение

$$\frac{l^2}{h^2} \leq \frac{1}{1 + \varepsilon}, \quad (29)$$

где $\varepsilon > 0$ — любое число. Тогда

$$\alpha_k \leq \frac{2}{1 + \varepsilon}, \quad k = 1, 2, \dots, M-1. \quad (30)$$

Значит,

$$\cos \frac{1}{2} \sigma_k = \sqrt{1 - \frac{1}{2} \alpha_k} \geq \sqrt{\frac{\varepsilon}{1 + \varepsilon}}. \quad (31)$$

Оценим величину $\frac{\sin \sigma_k}{l}$ снизу. Имеем

$$\frac{\sin \sigma_k}{l} = \frac{2 \sin \frac{1}{2} \sigma_k}{l} \cos \frac{1}{2} \sigma_k \geq \frac{2 \sin \frac{1}{2} \sigma_k}{l} \sqrt{\frac{\varepsilon}{1 + \varepsilon}}.$$

Но так как

$$\frac{2 \sin \frac{1}{2} \sigma_k}{l} = \frac{\sqrt{2(1 - \cos \sigma_k)}}{l} = \frac{\sqrt{2\alpha_k}}{l} = \sqrt{\lambda_k},$$

то

$$\frac{\sin \sigma_k}{l} \geq \sqrt{\lambda_k \frac{\varepsilon}{1 + \varepsilon}}. \quad (32)$$

Из (28) следует неравенство

$$\rho(z^n) \leq \rho(r^n) + \rho(s^n), \quad (33)$$

где

$$r_m^n = \sum_{k=1}^{M-1} \frac{\cos\left(n - \frac{1}{2}\right) \sigma_k}{\cos \frac{1}{2} \sigma_k} a_k X_m(k),$$

$$s_m^n = \sum_{k=1}^{M-1} \frac{l \sin n \sigma_k}{\sin \sigma_k} b_k X_m(k).$$

Воспользовавшись формулами (26) и оценками (31), (32), получим

$$\begin{aligned} \rho(r^n) &\leq \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left[\sum_{m=1}^{M-1} \left(\sum_{k=1}^{M-1} a_k X_m(k) \right)^2 \right]^{1/2} = \\ &= \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left[\sum_{m=1}^{M-1} (\alpha_m^0)^2 \right]^{1/2}, \\ \rho(s^n) &\leq \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left[\sum_{m=1}^{M-1} \frac{b_m^2}{\lambda_m} \right]^{1/2}. \end{aligned}$$

Таким образом, оценка (33) окончательно может быть записана в виде

$$\rho(z^n) \leq \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left\{ \left[\sum_{m=1}^{M-1} (\alpha_m^0)^2 \right]^{1/2} + \left[\sum_{m=1}^{M-1} \frac{b_m^2}{\lambda_m} \right]^{1/2} \right\}, \quad (34)$$

где $n = 2, 3, \dots, N$.

В оценке (34) правая часть от n не зависит и может быть истолкована как некоторая норма в пространстве F_h . Это означает, что для схемы $L_h(u^{(h)}) = f^{(h)}$ выполняется и свойство 2).

Следовательно, достаточным для устойчивости по начальным данным разностной схемы $L_h(u^{(h)}) = f^{(h)}$ является условие (29), налагающее на шаги h и l ограничение $\frac{l^2}{h^2} \leq \frac{1}{1+\varepsilon}$, где $\varepsilon > 0$.

Отметим, что для уравнений гиперболического типа рассматриваются также и неявные разностные схемы; об этом подробно сказано в [6].

§ 5. Метод характеристик

Для каждой точки P из области D определения решения исходной дифференциальной задачи существует, очевидно, такая наибольшая область $G(P) \in D$, что значение искомого решения в точке P зависит от соответствующей информации об этом решении во всех точках области $G(P)$. При построении сходящихся сеточных схем необходимо позаботиться о том, чтобы в сколь угодно малой окрестности любой точки области $G(P)$ при достаточно малых значениях шагов сетки имелась хотя бы одна точка сеточной области, от информации о разностном решении в которой зависит значение этого решения в точке P . В случае уравнений гиперболического типа подобная область может быть описана с использованием характеристик исходного уравнения. Последние часто имеют достаточно сложный вид и, кроме того, могут зависеть от искомого решения. Поэтому при использовании простейших разностных схем, рассматриваемых ранее, не всегда оказывается учтенным высказанное выше требование. Более подходящим в таких случаях может быть численное интегрирование вдоль характеристик, которое и лежит в основе одноименного метода. При этом обычно используется сетка характеристик или аппроксимирующая ее сетка. В тех случаях, когда характеристики зависят от искомого решения, построение такой сетки приходится проводить одновременно с нахождением решения. Подобная организация вычислений позволяет достаточно точно определять область влияния исходных данных и, следовательно, строго учитывать распространение возмущений, что придает методу характеристик четкий физический смысл. Ниже мы изложим основную идею этого метода на примере задачи Коши для квазилинейного уравнения второго порядка гиперболического типа.

1. Характеристики дифференциального уравнения. Рассмотрим относительно неизвестной функции $u = u(x, y)$ дифференциальное уравнение второго порядка

$$ar + 2bs + ct + f = 0, \quad (1)$$

где

$$r = u_{xx}, \quad s = u_{xy}, \quad t = u_{yy},$$

а функции

$$a = a(x, y, u, p, q), \quad b = b(x, y, u, p, q), \\ c = c(x, y, u, p, q), \quad f = f(x, y, u, p, q)$$

являются известными функциями переменных $x, y, u, p = u_x, q = u_y$.

Такое уравнение обычно называют *квазилинейным*. В частном случае, когда

$$a = a(x, y), \quad b = b(x, y), \quad c = c(x, y), \quad f = f(x, y, u, p, q),$$

уравнение (1) будем называть *почти линейным*. *Линейным* же это уравнение будем называть в случае, когда неизвестная функция u и ее частные производные входят в (1) линейно.

Если линейность уравнения не оговаривается особо, будем иметь в виду более общий квазилинейный случай. При этом функции a, b, c, f будем считать достаточно гладкими в некоторой области изменения их аргументов.

Присоединим к уравнению (1), скажем, начальные условия (условия Коши)

$$u|_l = \varphi(Q), \quad p|_l = \psi_1(Q), \quad q|_l = \psi_2(Q), \quad (2)$$

где l есть некоторая регулярная кривая в плоскости переменных x, y , а φ, ψ_1, ψ_2 — известные достаточно гладкие функции точки Q этой кривой. Функции φ, ψ_1, ψ_2 при этом не могут задаваться независимо. Вдоль линии l они должны быть подчинены условию согласования

$$du = p dx + q dy. \quad (3)$$

Поставленную дифференциальную задачу более наглядно можно пояснить геометрически.

Уравнение (1) задает в трехмерном пространстве переменных x, y, u множество поверхностей, которые обычно называют интегральными поверхностями данного уравнения. Дополнительные же условия (2), (3) геометрически означают следующее. Равенство $u|_l = \varphi(Q)$ задает в указанном пространстве над линией l некоторую линию l_0 . Задание же на l производных $p = u_x$ и $q = u_y$, согласованных с u посредством условия (3), определяет в каждой точке линии l_0 касательную плоскость. В сово-

купности это равносильно заданию бесконечно узкой полоски интегральной поверхности. Будем называть эту полоску *полоской первого порядка* l_1 . Поставленная задача Коши, таким образом, состоит в том, чтобы полоску первого порядка l_1 , заданную условиями (2), (3), с помощью уравнения (1) расширить до интегральной поверхности $u = u(x, y)$.

Возникает вопрос, всегда ли можно осуществить такое расширение и единственно ли оно. Выясним сначала, можно ли с помощью уравнения (1) расширить полоску l_1 хотя бы до полоски второго порядка l_2 , т. е. можно ли по условиям (2), (3) с учетом (1) найти на линии l значения вторых производных r, s, t (если подобное расширение понимать в буквальном смысле слова, то после нахождения на линии l значений r, s, t его можно приближенно осуществить, например, с помощью формулы Тейлора).

Поскольку задание некоторой функции на линии предполагает выполнение условий согласования типа (3), то применительно к функциям p, q на линии l будем иметь

$$\begin{aligned} dp &= r dx + s dy, \\ dq &= s dx + t dy. \end{aligned}$$

С учетом (1) это позволяет для неизвестных значений r, s, t на линии l указать следующую систему трех линейных алгебраических уравнений

$$\begin{aligned} ar + 2bs + ct &= -f, \\ r dx + s dy &= dp, \\ s dx + t dy &= dq. \end{aligned} \tag{4}$$

Если для данной полоски l_1 определитель

$$\Delta = \begin{vmatrix} a & 2b & c \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} = a dy^2 - 2b dx dy + c dx^2 \tag{5}$$

системы (4) отличен от нуля, то r, s, t на l могут быть найдены единственным образом. Тем самым полоска l_1 может быть однозначно расширена до полоски l_2 . Подобным же образом нетрудно убедиться, что в случае $\Delta \neq 0$ однозначно определяются и полоски более высоких

порядков, что дает возможность однозначно осуществить требуемое расширение.

Если же определитель (5) равен нулю на l_1 , то система уравнений (4) либо не имеет решения, либо имеет бесчисленное множество решений. Условие

$$\Delta = a \, dy^2 - 2b \, dx \, dy + c \, dx^2 = 0 \quad (6)$$

может выполняться и не для всей полосы l_1 . В дальнейшем для нас особый интерес будет представлять именно та полоска, для которой условие (6) выполняется всюду, при этом система (4) остается совместной. Такую полоску l_1 станем называть *характеристической полоской*. Соответствующую линию l_0 будем называть *характеристической кривой (характеристикой) в пространстве переменных x, y, u* , а ее проекцию l — *проекцией характеристической кривой (характеристики) на плоскости переменных x, y* .

В случае характеристических полосок значения вторых производных решения на l определяются с помощью исходного дифференциального уравнения лишь с точностью до слагаемого, являющегося произвольным решением однородной системы линейных алгебраических уравнений, соответствующей системе (4). Такие полоски допускают неоднозначное расширение до полосок второго порядка, и по ним возможно ветвление интегральных поверхностей.

Так как в случае характеристических полосок выполняется условие $\Delta = 0$, но система (4) при этом разрешима, то уравнения этой системы линейно зависимы. Выяснив эту зависимость, мы найдем для элементов таких полосок дополнительно к (3), (6) новые соотношения.

Требование совместности системы (4) означает, что ранг матрицы коэффициентов этой системы должен совпадать с рангом ее расширенной матрицы. Так как на линии l определитель матрицы коэффициентов системы (4) равен нулю, а по крайней мере один из ее миноров второго порядка

$$\begin{vmatrix} dx & dy \\ 0 & dx \end{vmatrix}, \quad \begin{vmatrix} dy & 0 \\ dx & dy \end{vmatrix}$$

отличен от нуля, то ранг этой матрицы, а следовательно

и ранг расширенной матрицы системы (4), равен двум. Отсюда вытекает равенство нулю всех миноров третьего порядка у последней матрицы, что дает возможность записать:

$$-\begin{vmatrix} a & 2b & -f \\ dx & dy & dp \\ 0 & dx & dq \end{vmatrix} = a(dx dp - dy dq) + 2b dx dq + f dx^2 = 0,$$

$$-\begin{vmatrix} a & c & -f \\ dx & 0 & dp \\ 0 & dy & dq \end{vmatrix} = a dy dp + c dx dq + f dx dy = 0,$$

$$-\begin{vmatrix} 2b & c & -f \\ dy & 0 & dp \\ dx & dy & dq \end{vmatrix} = 2b dy dp + c(dy dq - dx dp) + f dy^2 = 0.$$

Получив из (6) выражение

$$dy = \frac{b \pm \sqrt{b^2 - ac}}{a} dx \quad (7)$$

и используя его в каждом из трех полученных выше равенств, во всех случаях после простых преобразований приходим к соотношению

$$(a dp + f dx)(b \pm \sqrt{b^2 - ac}) + ac dq = 0. \quad (8)$$

Итак, в случае характеристических полосок дополнительно к условиям согласования (3) и требованию (6), из которого следуют соотношения (7), должны выполняться также равенства (8). В совокупности все эти соотношения

$$\begin{aligned} a dy - (b \pm \sqrt{b^2 - ac}) dx &= 0, \\ (a dp + f dx)(b \pm \sqrt{b^2 - ac}) + ac dq &= 0, \\ du &= p dx + q dy \end{aligned} \quad (9)$$

называют обычно *уравнениями характеристических полосок* или *уравнениями характеристик*. Два уравнения, получающиеся из первого равенства (9) при выборе знака у корня, называют *уравнениями направления характеристик*, а остальные — *дифференциальными соотношениями вдоль характеристик*.

Уравнения направления характеристик, записанные в виде

$$\frac{dy}{dx} = \frac{b \pm \sqrt{b^2 - ac}}{a}, \quad (10)$$

можно рассматривать как обыкновенные дифференциальные уравнения первого порядка относительно неизвестной функции $y = y(x)$. Правые части уравнений (10) содержат дискриминант $ac - b^2$ исходного уравнения (1). Если в некоторой точке данной интегральной поверхности выполнено условие $ac - b^2 < 0$, то через эту точку будут проходить две разные характеристики с разными проекциями на плоскость переменных x, y . Подобные точки обычно называют *гиперболическими* точками интегральной поверхности. В случае, если условие $ac - b^2 < 0$ выполняется для всех точек интегральной поверхности, эту поверхность называют гиперболической. Если же все интегральные поверхности данного дифференциального уравнения вида (1) в некоторой области изменения переменных x, y удовлетворяет условию гиперболичности, то в этой области рассматриваемое уравнение относят к гиперболическому типу. Подобным же образом через условия $ac - b^2 = 0$ и $ac - b^2 > 0$ вводят соответственно *параболические* и *эллиптические* уравнения. Для параболических уравнений обе характеристики сливаются в одну, а в эллиптическом случае вещественных характеристик нет.

В дальнейшем мы будем рассматривать лишь гиперболические уравнения. Для таких уравнений можно говорить о двух семействах характеристических полосок. Вдоль каждой из двух характеристик, выходящих из данной точки интегральной поверхности, в силу (8) выполняются разные дифференциальные соотношения. Если ввести обозначения

$$\lambda^{(1)} = \frac{b - \sqrt{b^2 - ac}}{a}, \quad \lambda^{(2)} = \frac{b + \sqrt{b^2 - ac}}{a},$$

то после умножения (8) на $b \pm \sqrt{b^2 - ac}$ уравнения характеристических полосок можно привести к виду

$$\begin{aligned} dy - \lambda^{(1)} dx &= 0, \\ a(dp + \lambda^{(2)} dq) + f dx &= 0, \\ du &= p dx + q dy \end{aligned} \quad (11)$$

для характеристических полосок первого семейства или к виду

$$\begin{aligned} du - \lambda^{(2)} dx &= 0, \\ a(dp + \lambda^{(1)} dq) + f dx &= 0, \\ du &= p dx + q dy \end{aligned} \quad (12)$$

для характеристических полосок второго семейства.

Уравнения (11), (12) позволяют для каждого решения исходного уравнения (1) находить как характеристики обоих семейств, так и само решение вдоль этих характеристик вместе с его первыми производными. При этом интересующие нас величины не могут, вообще говоря, находиться независимо, так как в случае квазилинейного уравнения (1) величины $\lambda^{(i)}$, $i = 1, 2$, являются функциями переменных x, y, u, p, q . В частных случаях линейного или почти линейного уравнений вида (1), когда коэффициенты a, b, c и $\lambda^{(1)}, \lambda^{(2)}$, а значит, и проекции характеристик на плоскость переменных x, y , не зависят от значений u, p, q , указанные проекции могут быть найдены как решения обыкновенных дифференциальных уравнений

$$\frac{dy}{dx} = \lambda^{(1)}(x, y), \quad (13)$$

$$\frac{dy}{dx} = \lambda^{(2)}(x, y) \quad (14)$$

соответственно для характеристик первого и второго семейств. В таких случаях уравнения (11), (12) можно решать в два этапа: сначала с помощью уравнений направления характеристик (13), (14) строится сетка характеристик, а затем из дифференциальных соотношений на построенных характеристиках находятся величины u, p, q .

Рассмотрение численных методов решения уравнений характеристических полосок мы и начнем со случая, когда возможно подобное упрощение процесса решения.

2. Построение сетки характеристик. Будем считать, что коэффициенты a, b, c в уравнении (1) не зависят от u, p, q ; при этом в некоторой достаточно широкой области изменения переменных x, y выполняется условие гиперболичности $ac - b^2 < 0$. Условия Коши (2), (3) будем предполагать заданными на некоторой линии l , не

принимавшей ни в одной точке характеристических направлений этого уравнения.

Через крайние точки M_0 , M_N этой линии проведем характеристики обоих семейств исходного гиперболического уравнения (1). Тем самым будет построена область, состоящая из двух криволинейных треугольников, в которой существует единственное решение поставленной задачи Коши. Будем решать эту задачу численно лишь в одном из таких треугольников M_0PM_N (рис. 9). Для

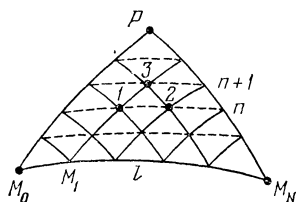


Рис. 9.

точки P этот треугольник будет играть роль области $G(P)$, о которой шла речь в начале этого параграфа. Аппроксимируем эту область сеточной областью следующим образом. Выберем на линии l , несущей начальные данные, в дополнение к точкам M_0 , M_N еще $N - 1$ (не обязательно равноотстоя-

щих) точек $M_1, M_2, \dots, M_{N-1}$ и проведем через каждую из этих точек обе характеристики исходного уравнения. Точки пересечения этих характеристик, принадлежащие треугольнику M_0PM_N , примем за узлы аппроксимирующей сеточной области. Будем находить эти точки последовательно по слоям, принимая за нулевой слой точки $M_0, M_1, \dots, M_N$ линии l .

Допустим, что все узлы n -го слоя ($n \geq 0$) сетки характеристик построены. Тогда любая точка 3 с координатами x_3, y_3 слоя номера $n + 1$ может быть найдена как точка пересечения характеристик первого и второго семейств исходного уравнения, проведенных соответственно через точки $1(x_1, y_1)$ и $2(x_2, y_2)$ n -го слоя сетки. Такие характеристики могут быть построены как решения задач Коши для уравнений (13) и (14) при начальных данных $y(x_i) = y_i$ для $i = 1$ и $i = 2$. Интегрируя последние уравнения соответственно в пределах от x_1 до x_3 и от x_2 до x_3 , получим

$$y_3 = y_1 + \int_{x_1}^{x_3} \lambda^{(1)}(x, y) dx,$$

$$y_3 = y_2 + \int_{x_2}^{x_3} \lambda^{(2)}(x, y) dx.$$
(15)

Так как на рассматриваемых участках характеристик зависимость y от x , вообще говоря, неизвестна, то координаты x_3 , y_3 точки пересечения избранных характеристик обычно приходится находить из (15) приближенно. Рассмотрим простейшие вычислительные правила, применяемые для этих целей.

Аналог метода Эйлера. Если интегралы в правых частях равенств (15) заменить приближенно соответствующими квадратурными суммами простейшей формулы левых прямоугольников, то можно записать

$$y_3 \approx y_1 + \lambda_1^{(1)} (x_3 - x_1), \quad y_3 \approx y_2 + \lambda_2^{(2)} (x_3 - x_2), \quad (16)$$

где

$$\lambda_j^{(i)} = \lambda^{(i)}(x_j, y_j).$$

Погрешность каждого из приближенных равенств (16) будет, очевидно, величиной порядка h^2 (см. выражение для остаточного члена формулы Эйлера (6.2.2)), где

$$h = \max \{ |x_3 - x_1|, |x_3 - x_2| \}.$$

Равенства (16) дают нам систему двух приближенных линейных алгебраических уравнений относительно неизвестных значений x_3 , y_3 .

Аналог метода трапеций. Чтобы получить более точные, чем (16), равенства, связывающие координаты точек 1, 2, 3, заменим интегралы в (15) по двухточечной квадратурной формуле трапеций (см. (5.3.2)). Тогда можно записать

$$\begin{aligned} y_3 &\approx y_1 + \frac{\lambda_1^{(1)} + \lambda_3^{(1)}}{2} (x_3 - x_1), \\ y_3 &\approx y_2 + \frac{\lambda_2^{(2)} + \lambda_3^{(2)}}{2} (x_3 - x_2). \end{aligned} \quad (17)$$

Погрешность каждого из таких приближенных равенств будет уже величиной порядка h^3 (см. выражение (5.3.3) для остаточного члена формулы трапеций).

Система (17) является аналогом метода трапеций (6.3.2) численного интегрирования обыкновенных дифференциальных уравнений вида (6.1.1). Она, вообще говоря, нелинейна. Неизвестные значения x_3 , y_3 из этой системы, как правило, находят итерациями. Начальные прибли-

жения x_{30}, y_{30} к этим значениям можно получить из линейной системы типа (16)

$$y_{30} = y_1 + \lambda_1^{(1)} (x_{30} - x_1), \quad y_{30} = y_2 + \lambda_2^{(2)} (x_{30} - x_2).$$

Уточнение же исходных значений можно осуществлять, например, посредством последовательного решения линейных систем

$$y_{3i+1} = y_1 + \frac{\lambda_1^{(1)} + \lambda^{(1)}(x_{3i}, y_{3i})}{2} (x_{3i+1} - x_1),$$

$$y_{3i+1} = y_2 + \frac{\lambda_2^{(2)} + \lambda^{(2)}(x_{3i}, y_{3i})}{2} (x_{3i+1} - x_2), \quad i = 0, 1, 2, \dots,$$

которые очевидным образом строятся на основе формул (17). Обычно при этом ограничиваются одной-двумя итерациями.

Еще более точные вычислительные правила построения сетки характеристик можно получить, если сконструировать соответствующие аналоги более точных методов (как одношаговых, так и многошаговых) численного интегрирования уравнения вида (6.1.1). Достаточно удобные расчетные формулы можно построить, например, по аналогии с одношаговыми методами, рассматриваемыми в § 4 гл. 6. Для каждого уровня точности использование таких формул сводится к последовательному решению фиксированного числа систем двух линейных алгебраических уравнений с двумя неизвестными. Аналоги многошаговых методов менее удобны для этих целей, так как они предполагают привлечение на каждой характеристике не одного, а нескольких найденных ранее узлов сетки, которые расположены, вообще говоря, неравномерно. Все это предопределяет громоздкость и неоднородность соответствующих вычислительных правил.

3. Нахождение характеристических полосок. Сначала будем иметь в виду случай, когда сетка характеристик предварительно построена. Значения величин u, p, q в узлах этой сетки будем искать последовательно по слоям (см. рис. 9). В простейшем случае эта задача сводится к нахождению величин $u_3 = u(x_3, y_3)$, $p_3 = p(x_3, y_3)$, $q_3 = q(x_3, y_3)$ по известным значениям $u_i = u(x_i, y_i)$, $p_i = p(x_i, y_i)$, $q_i = q(x_i, y_i)$, $i = 1, 2$. Для нахождения этих величин используем из уравнений ха-

рактических полосок (11), (12) лишь дифференциальные соотношения

$$\begin{aligned} a(dp + \lambda^{(2)} dq) + f dx &= 0, \\ du &= p dx + q dy, \end{aligned} \quad (18)$$

$$\begin{aligned} a(dp + \lambda^{(1)} dq) + f dx &= 0, \\ du &= p dx + q du \end{aligned} \quad (19)$$

вдоль характеристик первого и второго семейств соответственно. Подобно изложенному в предыдущем пункте, на основании (18), (19) можно записать точные равенства

$$\begin{aligned} \int_1^3 [a(dp + \lambda^{(2)} dq) + f dx] &= 0, \\ \int_2^3 [a(dp + \lambda^{(1)} dq) + f dx] &= 0, \end{aligned} \quad (20)$$

$$u_3 = \frac{u_1 + u_2}{2} + \frac{1}{2} \left[\int_1^3 (p dx + q dy) + \int_2^3 (p dx + q dy) \right].$$

Прибегая, как и ранее, к различным способам приближенной замены интегралов в этих равенствах, получим различные вычислительные методы нахождения величин u_3 , p_3 , q_3 .

Аналог метода Эйлера. Использование простейших квадратурных формул левых прямоугольников позволяет на основании (20) записать приближенные равенства

$$\begin{aligned} a_1(p_3 - p_1) + a_1\lambda_1^{(2)}(q_3 - q_1) + f_1(x_3 - x_1) &\approx 0, \\ a_2(p_3 - p_2) + a_2\lambda_2^{(1)}(q_3 - q_2) + f_2(x_3 - x_2) &\approx 0, \\ u_3 \approx \frac{u_1 + u_2}{2} + \frac{1}{2} [p_1(x_3 - x_1) + q_1(y_3 - y_1) + \\ &+ p_2(x_3 - x_2) + q_2(y_3 - y_2)], \end{aligned} \quad (21)$$

где $a_i = a(x_i, y_i)$, $f_i = f(x_i, y_i, u_i, p_i, q_i)$.

Тем самым относительно неизвестных значений u_3 , p_3 , q_3 получим приближенную систему трех линейных алгебраических уравнений, которая может быть решена последовательно: сначала из первых двух ее уравнений находят приближенные значения величин p_3 , q_3 , после чего из последнего в системе (21) уравнения можно непосредственно вычислить приближенное значение функции u в точке $З$.

Аналог метода трапеций. Замена интегралов в (20) квадратурными суммами двухточечной формулы трапеций позволяет указать для приближенного нахождения величин u_3 , p_3 , q_3 следующую систему численных уравнений:

$$\begin{aligned} \frac{a_1 + a_3}{2} (p_3 - p_1) + \frac{a_1 \lambda_1^{(2)} + a_3 \lambda_3^{(2)}}{2} (q_3 - q_1) + \\ + \frac{f_1 + f_3}{2} (x_3 - x_1) \approx 0, \\ \frac{a_2 + a_3}{2} (p_3 - p_2) + \frac{a_2 \lambda_2^{(1)} + a_3 \lambda_3^{(1)}}{2} (q_3 - q_2) + \\ + \frac{f_2 + f_3}{2} (x_3 - x_2) \approx 0, \\ u_3 \approx \frac{u_1 + u_2}{2} + \frac{1}{2} \left[\frac{p_1 + p_3}{2} (x_3 - x_1) + \frac{q_1 + q_3}{2} (y_3 - y_1) + \right. \\ \left. + \frac{p_2 + p_3}{2} (x_3 - x_2) + \frac{q_2 + q_3}{2} (y_3 - y_2) \right]. \end{aligned} \quad (22)$$

В отличие от (21), первые два уравнения системы (22) уже, вообще говоря, нелинейны; при этом неизвестные u_3 , p_3 , q_3 входят в каждое уравнение системы.

Систему (22) можно решать итерационно. Исходные значения u_{30} , p_{30} , q_{30} при этом обычно находят из линейной системы типа (21), а процесс их уточнения проводят также на основании линейных систем вида

$$\begin{aligned} \frac{a_1 + a_3}{2} (p_{3i+1} - p_1) + \frac{a_1 \lambda_1^{(2)} + a_3 \lambda_3^{(2)}}{2} (q_{3i+1} - q_1) + \\ + \frac{f_1 + f(x_3, y_3, u_{3i}, p_{3i}, q_{3i})}{2} (x_3 - x_1) = 0, \\ \frac{a_2 + a_3}{2} (p_{3i+1} - p_2) + \frac{a_2 \lambda_2^{(1)} + a_3 \lambda_3^{(1)}}{2} (q_{3i+1} - q_2) + \\ + \frac{f_2 + f(x_3, y_3, u_{3i}, p_{3i}, q_{3i})}{2} (x_3 - x_2) = 0, \\ u_{3i+1} = \frac{u_1 + u_2}{2} + \frac{1}{2} \left[\frac{p_1 + p_{3i+1}}{2} (x_3 - x_1) + \frac{q_1 + q_{3i+1}}{2} (y_3 - y_1) + \right. \\ \left. + \frac{p_2 + p_{3i+1}}{2} (x_3 - x_2) + \frac{q_2 + q_{3i+1}}{2} (y_3 - y_2) \right], \quad i = 0, 1, 2, \dots \end{aligned}$$

При каждом фиксированном значении i процесс решения таких систем, как и в случае аналога метода Эйлера, распадается на два более простых этапа.

Подобно изложенному в § 4 гл. 6, могут быть построены и более точные вычислительные методы нахождения характеристических полосок. Не останавливаясь

здесь на рассмотрении таких методов, сделаем лишь несколько замечаний относительно нахождения характеристических полосок для более общего случая квазилинейного уравнения (1), когда характеристики зависят от решения поставленной задачи Коши. В этом случае сетка характеристик должна строиться одновременно с нахождением значений функций u , p , q в ее узлах. Равенства типа (15) сейчас не могут рассматриваться независимо от (20). В квазилинейном случае, очевидно, необходимо одновременно рассматривать все пять равенств:

$$\begin{aligned} y_3 - y_1 &= \int_1^3 \lambda^{(1)}(x, y, u, p, q) dx, \\ y_3 - y_2 &= \int_2^3 \lambda^{(2)}(x, y, u, p, q) dx, \\ \int_1^3 \{a(x, y, u, p, q) [dp + \lambda^{(2)}(x, y, u, p, q) dq] + \\ &\quad + f(x, y, u, p, q) dx\} = 0, \\ \int_2^3 \{a(x, y, u, p, q) [dp + \lambda^{(1)}(x, y, u, p, q) dq] + \\ &\quad + f(x, y, u, p, q) dx\} = 0, \\ u_3 &= \frac{u_1 + u_2}{2} + \frac{1}{2} \left[\int_1^3 (p dx + q dy) + \int_2^3 (p dx + q dy) \right]. \end{aligned}$$

На основании этих равенств, подобно изложенному выше, можно построить, например, следующий аналог метода Эйлера:

$$\begin{aligned} y_3 - y_1 &\approx (x_3 - x_1) \lambda^{(1)}(x_1, y_1, u_1, p_1, q_1), \\ y_3 - y_2 &\approx (x_3 - x_2) \lambda^{(2)}(x_2, y_2, u_2, p_2, q_2), \\ a(x_1, y_1, u_1, p_1, q_1) \times \\ &\quad \times [(p_3 - p_1) + \lambda^{(2)}(x_1, y_1, u_1, p_1, q_1)(q_3 - q_1)] + \\ &\quad + f(x_1, y_1, u_1, p_1, q_1)(x_3 - x_1) \approx 0, \\ a(x_2, y_2, u_2, p_2, q_2) \times \\ &\quad \times [(p_3 - p_2) + \lambda^{(1)}(x_2, y_2, u_2, p_2, q_2)(q_3 - q_2)] + \\ &\quad + f(x_2, y_2, u_2, p_2, q_2)(x_3 - x_2) \approx 0, \\ u_3 &\approx \frac{u_1 + u_2}{2} + \\ &\quad + \frac{1}{2} [p_1(x_3 - x_1) + q_1(y_3 - y_1) + p_2(x_3 - x_2) + q_2(y_3 - y_2)]. \end{aligned}$$

Решение этой линейной системы может быть расчленено на три простых этапа: приближенные координаты точки Z находятся из системы двух ее первых уравнений, значения p_z, q_z вычисляются приближенно с помощью двух следующих уравнений, после чего из последнего уравнения непосредственно находится приближенное значение функции u в точке Z . При этом, очевидно, координаты узлов следующего слоя сетки характеристик могут быть приближенно вычислены лишь в том случае, если известны не только координаты узлов предыдущего слоя, но и значения функций u, p, q в этих узлах.

Аналог метода трапеций в квазилинейном случае приведет, как и ранее, к нелинейной системе пяти численных уравнений, однако каждое уравнение системы будет содержать уже все пять неизвестных величин x_z, y_z, u_z, p_z, q_z .

В процессе итераций решение этой нелинейной системы, подобно изложенному выше, можно свести к последовательному решению линейных систем; при этом решение каждой из них может быть расчленено на три более простых этапа.

При некоторых дополнительных ограничениях с использованием, скажем, способа построения вычислительных правил, основанных на принципе последовательного повышения порядка точности результата (см. § 4 гл. 6), могут быть построены и более точные методы решения уравнений характеристик.

§ 6. Метод прямых

В методе прямых, в отличие от метода сеток, операцию дифференцирования аппроксимируют не по всем, а только по некоторым избранным переменным, тем самым дифференциальные уравнения приближенно заменяют дифференциально-разностными с меньшим числом непрерывных независимых переменных. Этот метод можно рассматривать как предельный случай метода сеток, когда шаги разбиения по части независимых переменных стремятся к нулю. Метод прямых, таким образом, позволяет понизить размерность дифференциальной задачи. Подобное понижение размерности бывает особенно оправданным в тех случаях, когда роли независимых переменных ока-

зываются неравнозначными. Такие задачи нередко встречаются в приложениях.

С основными вариантами метода мы познакомимся ниже на примере линейных дифференциальных уравнений в частных производных второго порядка, хотя метод применим и в случае нелинейных уравнений любых порядков и систем таких уравнений.

1. Поперечные схемы метода прямых для уравнений параболического типа. Пусть в полуполосе $0 \leq x \leq 1$, $t \geq 0$ поставлена следующая граничная задача:

$$u_t = L(u) + f(x, t), \quad (1)$$

$$u(x, 0) = \varphi(x), \quad (2)$$

$$u(0, t) = \psi_0(t), \quad u(1, t) = \psi_1(t). \quad (3)$$

Здесь

$$L(u) = a(x, t) u_{xx} + b(x, t) u_x + c(x, t) u.$$

Существование и единственность достаточно гладкого решения рассматриваемой задачи предполагаются. Будем искать его приближенно методом прямых в прямоугольнике $\Pi \{0 \leq x \leq 1, 0 \leq t \leq T < \infty\}$. В зависимости от того, будем ли мы при этом аппроксимировать производные по x или по t , получим продольный или поперечный вариант метода. В соответствии с историческим развитием вопроса рассмотрим сначала поперечный его вариант.

Обозначим через Π_τ решеточную область $0 < x < 1$, $t = t_n = n\tau$, $n = 1, 2, \dots, N$ ($N\tau = T$), а под ее границей Γ_τ будем понимать множество, состоящее из отрезка $0 \leq x \leq 1$ прямой $t = 0$ и точек $(0, t_n)$, $(1, t_n)$, $n = 1, 2, \dots, N$.

Рассмотрим задачу (1) — (3) на $\Pi_\tau + \Gamma_\tau$. Производную $u_t(x, t)$ в (1) при $t = t_n$, $n = 1, 2, \dots, N$, заменим приближенно левосторонним разностным отношением

$$u_t(x, t_n) \approx \frac{u(x, t_n) - u(x, t_{n-1})}{\tau}. \quad (4)$$

Тогда для решеточной функции, задаваемой системой функций $u_n = u_n(x)$, $n = 0, 1, \dots, N$, приближенно представляющих функцию $u(x, t)$ на отрезках $0 \leq x \leq 1$ прямых $t = t_n$, $n = 0, 1, \dots, N$, в $\Pi_\tau + \Gamma_\tau$ можно поставить

следующую дифференциально-разностную граничную задачу:

$$\frac{u_n(x) - u_{n-1}(x)}{\tau} = L_n(u_n) + f_n(x), \quad (5)$$

$$u_0(x) = \varphi(x), \quad (6)$$

$$u_n(0) = \psi_0(t_n), \quad u_n(1) = \psi_1(t_n). \quad (7)$$

Здесь

$$L_n(u_n) = a_n(x) u_n''(x) + b_n(x) u_n'(x) + c_n(x) u_n(x),$$

$$a_n(x) = a(x, t_n), \quad b_n(x) = b(x, t_n),$$

$$c_n(x) = c(x, t_n), \quad f_n(x) = f(x, t_n).$$

Задача (5)–(7) распадается на N граничных задач для обыкновенных дифференциальных уравнений второго порядка. Действительно, так как функция $u_0(x)$ известна из начального условия (6), то неизвестные функции $u_n(x) \approx u(x, t_n)$, $n = 1, 2, \dots, N$, можно находить последовательно, решая при каждом значении $n = 1, 2, \dots, N$ одно обыкновенное дифференциальное уравнение второго порядка вида (5) с соответствующими граничными условиями (7).

Построенная вычислительная схема метода прямых достаточно четко поясняет геометрический смысл названия метода. В силу этого в многомерном случае рассматриваемый метод часто называют *методом плоскостей* или *методом гиперплоскостей*.

На примере вычислительной схемы (5)–(7) можно проследить связь между методом прямых и методом сеток. Если дополнительно к прямым $t = t_n = n\tau$, $n = 0, 1, \dots, N$, в прямоугольнике Π провести прямые $x = x_m = mh$, $m = 0, 1, \dots, M$ ($Mh = 1$) и во всех внутренних точках (x_m, t_n) , $m = 1, 2, \dots, M-1$, $n = 1, 2, \dots, N$, полученной сеточной области, кроме замены типа (4), произвести также приближенную замену производных u_{xx} , u_x их простейшими симметричными выражениями через значения функции u , то, с учетом дополнительных условий (2), (3), получим широко распространенную в практике вычислений двухслойную трехстолбцовую неявную схему метода сеток. В пределе при $h \rightarrow 0$ эта схема, очевидно, переходит в схему (5)–(7) метода прямых. В § 2 для случая уравнения теплопроводности было показано, что

подобная неявная схема метода сеток может применяться при любом отношении шагов τ и h . Поэтому предельный переход при $h \rightarrow 0$ в ней оправдан, и можно надеяться, что схема (5) — (7) метода прямых будет вполне приемлемой для вычислений.

Параллельно с этой схемой можно построить и другую поперечную схему метода, заменив производную u_t не левосторонним, а правосторонним разностным отношением. Однако следует ожидать, что такая схема метода прямых будет уже мало пригодной для вычислений. Действительно, в § 2 было выяснено, что явная схема метода сеток для уравнения теплопроводности, построенная с использованием правосторонней замены производной u_t , может применяться для счета на большое число шагов по t при условии типа $\tau \leq O(h^2)$. Последняя же вычислительная схема метода прямых получается из подобной явной схемы метода сеток предельным переходом, когда $h \rightarrow 0$. При этом неизбежно будет нарушено упомянутое выше соотношение между h и τ .

По построению очевидно, что погрешность аппроксимации вычислительной схемы метода прямых будет, вообще говоря, меньше соответствующей погрешности связанной с ней разностной схемы метода сеток. В схеме (5) — (7), например, эта погрешность возникает лишь за счет приближенной замены (4) и имеет вид

$$r_n(x) = \frac{\tau}{2} u_{tt}(x, t_n - \theta_n \tau), \quad 0 < \theta_n < 1, \quad n = 1, 2, \dots, N. \quad (8)$$

При построении соответствующей вычислительной схемы метода сеток дополнительно к (8) будут внесены ошибки и за счет приближенной замены производных u_{xx} , u_x . Правда, в последнем случае аппроксимирующая задача будет, как правило, более простой с точки зрения ее реализации. Однако в ряде частных случаев метод прямых может оказаться предпочтительнее и в этом отношении. Например, поперечная схема (5) — (7) особенно удобна в том случае, когда коэффициенты уравнения (1) не зависят от x . Соответствующая аппроксимирующая задача приводит тогда к последовательному решению граничных задач для обыкновенных дифференциальных уравнений второго порядка с постоянными коэффициентами.

Если в прямоугольнике Π производная u_{tt} ограничена по абсолютному значению, то погрешность (8) будет величиной порядка τ . Для задач с более гладким решением можно ставить вопрос о построении более точных вычислительных схем. Повысить порядок погрешности аппроксимации можно следующим простым приемом.

Рассмотрим уравнение (1) при $t=t_n$ и $t=t_{n-1}$. Производную u_t в первом случае заменим левосторонним разностным отношением, а во втором — правосторонним. Тогда можно записать

$$\begin{aligned}\frac{u(x, t_n) - u(x, t_{n-1})}{\tau} &= L(u)|_{t=t_n} + f_n(x) + \frac{\tau}{2} u_{tt}(x, t_n - \theta_n \tau), \\ \frac{u(x, t_n) - u(x, t_{n-1})}{\tau} &= L(u)|_{t=t_{n-1}} + f_{n-1}(x) - \frac{\tau}{2} u_{tt}(x, t_n - \theta'_n \tau), \\ 0 &< \theta'_n < 1,\end{aligned}$$

откуда

$$\begin{aligned}\frac{u(x, t_n) - u(x, t_{n-1})}{\tau} &= \\ &= \frac{L(u)|_{t=t_n} + L(u)|_{t=t_{n-1}}}{2} + \frac{f_n(x) + f_{n-1}(x)}{2} + \rho_n(x),\end{aligned}$$

где

$$\begin{aligned}\rho_n(x) &= \frac{\tau}{4} [u_{tt}(x, t_n - \theta_n \tau) - u_{tt}(x, t_n - \theta'_n \tau)] = \\ &= \frac{\tau^2}{4} (\theta'_n - \theta) u_{ttt}(x, t_n - \tilde{\theta}_n \tau), \quad 0 < \tilde{\theta}_n < 1.\end{aligned}\quad (9)$$

Соответствующая схема метода прямых принимает вид

$$\begin{aligned}\frac{u_n(x) - u_{n-1}(x)}{\tau} &= \frac{L_n(u_n) + L_{n-1}(u_{n-1})}{2} + \frac{f_n(x) + f_{n-1}(x)}{2}, \\ 0 &< x < 1,\end{aligned}\quad (10)$$

$$u_0(x) = \varphi(x), \quad 0 \leq x \leq 1, \quad (11)$$

$$u_n(0) = \psi_0(t_n), \quad u_n(1) = \psi_1(t_n), \quad n = 1, 2, \dots, N. \quad (12)$$

Поперечная схема (10) — (12) точнее схемы (5) — (7) в том смысле, что порядок малости в сравнении с τ ее погрешности аппроксимации на единицу выше, чем в случае схемы (5) — (7) (см. остаточные члены (8) и (9)).

Каждое дифференциальное уравнение в рассмотренных выше поперечных схемах метода прямых связывает приближенные значения неизвестной функции $u(x, t)$ на двух соседних прямых. Такие двухслойные схемы метода особенно часто используются в практике вычислений. Дальнейшего повышения порядка точности поперечных схем можно добиваться, например, на пути построения многослойных схем метода. Заменяя в уравнении (1) производную u_t на прямой $t=t_n$ ($n=q, q+1, \dots, N$) приближенным выражением через значения функции u при $t=t_n, t=t_{n-1}, \dots, t=t_{n-q}$, нетрудно по аналогии с рассмотренным выше построить пример $(q+1)$ -слойной поперечной схемы метода прямых. Если приближенные значения функции u на начальных прямых $t=t_1, t=t_2, \dots, t=t_{q-1}$ были как-то найдены предварительно, то построенная $(q+1)$ -слойная схема дает возможность находить приближенные значения функции u на очередных прямых $t=t_q, t=t_{q+1}, \dots, t=t_N$ последовательно как решения соответствующих обыкновенных уравнений второго порядка с граничными условиями вида (7).

Через значения функции u при $t=t_n, t=t_{n-1}, \dots, t=t_{n-q}$ можно приближенно выразить производную u_t не только при $t=t_n$, но и на любой из прямых $t=t_{n-1}, t=t_{n-2}, \dots, t=t_{n-q}$. Соответственно этому получим различные варианты $(q+1)$ -слойных поперечных схем метода. Необходимо отметить, что полученные так схемы не всегда пригодны для вычислений, так как они могут оказаться неустойчивыми. Комбинируя их, как и в случае двухслойных схем, можно построить $(q+1)$ -слойные схемы повышенной точности аппроксимации. В этом случае вместо (10) будем иметь более общие уравнения

$$\frac{1}{\tau} \sum_{i=0}^q A_i u_{n-i}(x) = \sum_{i=0}^q B_i [L_{n-i}(u_{n-i}) + f_{n-i}(x)], \quad (13)$$

$$n = q, q+1, \dots, N.$$

Аппроксимацию уравнения (1) можно и непосредственно искать в виде (13) с неопределенными коэффициентами A_i и $B_i, i=0, 1, \dots, q$. Выбор этих параметров можно подчинить, например, требованиям максимального порядка аппроксимации и устойчивости вычислений.

2. Продольные схемы метода прямых для параболических уравнений. Будем иметь в виду, как и ранее, первую граничную задачу (1) — (3). Однако сейчас при построении вычислительных схем метода прямых будем аппроксимировать операцию дифференцирования не по времени t , а по пространственной переменной x . Для этого, подобно случаю поперечных схем, предварительно исходную область приближенно заменим решеточной областью, проводя прямые $x = x_n = nh$, $n = 0, 1, \dots, N$ ($Nh = 1$). На каждой из внутренних прямых $x = x_n$, $n = 1, 2, \dots, N - 1$, производные u_{xx} и u_x аппроксимируем через значения функции u на нескольких соседних прямых. Если для этих целей использовать, скажем, простейшие симметричные выражения, то, подобно случаю поперечных схем метода, можно записать:

$$u'_n(t) = a_n(t) \frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{h^2} + b_n(t) \frac{u_{n+1}(t) - u_{n-1}(t)}{2h} + c_n(t) u_n(t) + f_n(t), \quad t > 0, \quad (14)$$

$$u_0(t) = \psi_0(t), \quad u_N(t) = \psi_1(t), \quad t \geq 0, \quad (15)$$

$$u_n(0) = \varphi(x_n), \quad n = 1, 2, \dots, N - 1. \quad (16)$$

Здесь

$$u_n(t) \approx u(x_n, t),$$

$$a_n(t) = a(x_n, t),$$

$$b_n(t) = b(x_n, t),$$

$$c_n(t) = c(x_n, t),$$

$$f_n(t) = f(x_n, t).$$

Тем самым нахождение приближенных значений $u_n(t)$ решения $u(x, t)$ на прямых $x = x_n$, $n = 1, 2, \dots, N - 1$, сводится к решению задачи Коши для системы $N - 1$ линейных обыкновенных дифференциальных уравнений первого порядка.

Погрешность аппроксимации задачи (1) — (3) задачей (14) — (16) предопределяется неточностью замены производных u_{xx} , u_x соответствующими простейшими симметричными выражениями через значения функции u и будет, очевидно, величиной порядка h^2 , если решение $u(x, t)$ обладает достаточной гладкостью (скажем, функция u

имеет в рассматриваемой области ограниченную по абсолютному значению четвертую производную по x). В тех случаях, когда решение будет более гладким, можно ставить вопрос о построении вычислительных схем, характеризующихся более высоким порядком погрешности аппроксимации. Подобно случаю поперечных схем, повышать точность аппроксимации можно по-разному. Остановимся здесь лишь на некоторых из простейших подходов к решению этой проблемы.

Прежде всего обратим внимание на то, что в случае аппроксимирующей задачи (14) — (16) при каждом из значений $n = 1, 2, \dots, N - 1$ соответствующее обыкновенное дифференциальное уравнение первого порядка связывает три функции $u_{n-1}(t)$, $u_n(t)$, $u_{n+1}(t)$, но при этом каждое отдельно взятое уравнение содержит производную лишь одной из них. Можно записать несколько схем типа (14) — (16), для которых соответствующие обыкновенные дифференциальные уравнения, как и выше, будут связывать три соседних по номеру функции, но при этом схемы могут отличаться друг от друга, в частности, тем, что будут содержать производные разных функций (из указанных выше). Линейным комбинированием таких схем можно построить новые вычислительные схемы, для которых каждое обыкновенное дифференциальное уравнение первого порядка будет связывать не только три соседних по номеру функции, но и производные каждой из них. Удачным выбором соответствующей линейной комбинации, подобно случаю поперечных схем, можно получить схему с более высоким порядком погрешности аппроксимации.

При построении подобных вычислительных схем исходную задачу в рассматриваемой области удобнее записать в виде

$$\rho(x, t) u_t = u_{xx} + c(x, t) u + f(x, t), \quad (17)$$

$$u(x, 0) = \varphi(x), \quad (18)$$

$$u(0, t) = \psi_0(t), \quad u(1, t) = \psi_1(t). \quad (19)$$

Такая форма записи может быть легко получена из прежней, если в уравнении (1) заменой неизвестной функции исключить слагаемое с первой производной по x .

Рассмотрев уравнение (17) на трех соседних прямых $x = x_{n+1}$, $x = x_n$, $x = x_{n-1}$, можно составить линейную комбинацию соответствующих равенств с произвольными коэффициентами A , B , C . При этом, в частности, будет образована следующая комбинация значений второй производной:

$$Au_{xx}(x_{n+1}, t) + Bu_{xx}(x_n, t) + Cu_{xx}(x_{n-1}, t). \quad (20)$$

Представив эту комбинацию по возможности более точно через значения функции u , скажем, в форме

$$\frac{\alpha u(x_{n+1}, t) + \beta u(x_n, t) + \gamma u(x_{n-1}, t)}{h^2}, \quad (21)$$

очевидным образом можем построить интересующую нас аппроксимацию задачи (17) — (19) (заметим, что уравнение (17) не содержит производной u_x).

Коэффициенты A , B , C и α , β , γ подберем так, чтобы в разложениях по степеням h выражений (20) и (21) совпадали слагаемые до членов с возможно более высокими степенями h . Указанные разложения легко могут быть выписаны, если предварительно с помощью формулы Тейлора составить разложения около точки (x_n, t) значений $u(x_n \pm h, t)$ и $u_{xx}(x_n \pm h, t)$ функций u и u_{xx} . Простыми вычислениями можно проверить, что выбором параметров A , B , C и α , β , γ удастся добиться совпадения таких разложений, вообще говоря, лишь до членов с h^3 включительно; при этом указанные параметры должны быть подчинены ограничениям

$$\gamma = \alpha, \quad \beta = -2\alpha, \quad A = C = \frac{1}{12}\alpha, \quad B = \frac{5}{6}\alpha.$$

Тем самым непосредственно может быть проверена справедливость равенства

$$\begin{aligned} \frac{u_{xx}(x_{n+1}, t) + 10u_{xx}(x_n, t) + u_{xx}(x_{n-1}, t)}{12} = \\ = \frac{u(x_{n+1}, t) - 2u(x_n, t) + u(x_{n-1}, t)}{h^2} + r_n(t), \end{aligned} \quad (22)$$

где

$$r_n(t) = \frac{h^4}{240} \frac{\partial^6 u(\xi_n, t)}{\partial x^6}, \quad x_{n-1} < \xi_n < x_{n+1}. \quad (23)$$

Теперь уже, используя (22), нетрудно записать следующую вычислительную схему продольного варианта метода прямых:

$$\begin{aligned} & \frac{\rho_{n+1}(t) u'_{n+1}(t) + 10\rho_n(t) u'_n(t) + \rho_{n-1}(t) u'_{n-1}(t)}{12} = \\ & = \frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{h^2} + \\ & + \frac{c_{n+1}(t) u_{n+1}(t) + 10c_n(t) u_n(t) + c_{n-1}(t) u_{n-1}(t)}{12} + \\ & + \frac{f_{n+1}(t) + 10f_n(t) + f_{n-1}(t)}{12}, \quad t > 0, \end{aligned} \quad (24)$$

$$u_0(t) = \psi_0(t), \quad u_N(t) = \psi_1(t), \quad t \geq 0, \quad (25)$$

$$u_n(0) = \varphi(x_n), \quad n = 1, 2, \dots, N-1. \quad (26)$$

Смысл используемых в этой записи обозначений вполне аналогичен применяемым ранее.

Если решение исходной задачи имеет ограниченную по абсолютному значению производную $\frac{\partial^6 u(x, t)}{\partial x^6}$, то в силу (23) порядок погрешности аппроксимации в случае схемы (24) — (26) будет на две единицы выше, чем в случае схемы типа (14) — (16).

Дальнейшего повышения точности аппроксимации при построении продольных схем метода прямых можно добиваться за счет привлечения информации более чем на трех соседних прямых. При этом, как и в случае поперечных схем, аппроксимирующие уравнения можно искать непосредственно в виде равенств, аналогичных (13), с произвольными коэффициентами, подчиняя выбор последних соответствующим требованиям. Следует, однако, заметить, что в случае продольных многослойных схем при аппроксимации дифференциального уравнения на прямых, близких к граничным, обычно возникают трудности, связанные с нарушением однородности процесса приближения. В некоторых частных случаях, правда, эти трудности удастся существенно ослабить за счет использования граничных данных.

3. Метод прямых для уравнений гиперболического типа.

Пусть в полуполосе $0 \leq x \leq 1$, $t \geq 0$ поставлена следующая граничная задача:

$$u_{tt} + g(x, t) u_t = L(u) + f(x, t), \quad (27)$$

$$u(x, 0) = \varphi(x), \quad u_t(x, 0) = \Phi(x), \quad (28)$$

$$u(0, t) = \psi_0(t), \quad u(1, t) = \psi_1(t). \quad (29)$$

Здесь $L(u)$ имеет такой же вид, как и в (1). Будем считать, что $a(x, t) \geq a > 0$. Существование и единственность достаточно гладкого решения задачи (27) — (29) предполагаются. Будем искать это решение методом прямых в прямоугольнике $\Pi \{0 \leq x \leq 1, 0 \leq t \leq T < \infty\}$.

Как и в случае параболических уравнений, здесь также можно говорить о продольном и поперечном вариантах метода. Приведем сначала пример поперечной схемы.

Прямоугольник Π аппроксимируем решеточной областью $\Pi_\tau + \Gamma_\tau$; при этом, в отличие от параболического случая (см. п. 1), к границе Γ_τ рассматриваемой области отнесем также и интервал $0 < x < 1$ прямой $t = t_1 = \tau$. На каждой прямой $t = t_{n+1}$ для $n = 1, 2, \dots, N-1$ производные u_t и u_{tt} заменим следующими левосторонними разностными отношениями первого и второго порядков соответственно:

$$u_t(x, t_{n+1}) \approx \frac{u(x, t_{n+1}) - u(x, t_n)}{\tau}, \quad (30)$$

$$u_{tt}(x, t_{n+1}) \approx \frac{u(x, t_{n+1}) - 2u(x, t_n) + u(x, t_{n-1}))}{\tau^2}. \quad (31)$$

Кроме того, на прямой $t = 0$ для аппроксимации производной u_t используем правостороннюю замену

$$u_t(x, 0) \approx \frac{u(x, \tau) - u(x, 0)}{\tau}. \quad (32)$$

Это позволит для решеточной функции, задаваемой системой функций $u_n(x) \approx u(x, t_n)$, $n = 0, 1, \dots, N$, поставить

в области $\Pi_t + \Gamma_\tau$ следующую дифференциально-разностную граничную задачу:

$$\frac{u_{n+1}(x) - 2u_n(x) + u_{n-1}(x)}{\tau^2} + g_{n+1}(x) \frac{u_{n+1}(x) - u_n(x)}{\tau} = L_{n+1}(u_{n+1}) + f_{n+1}(x), \quad (33)$$

$$u_0(x) = \varphi(x), \quad \frac{u_1(x) - u_0(x)}{\tau} = \Phi(x), \quad (34)$$

$$u_{n+1}(0) = \psi_0(t_{n+1}) \quad u_{n+1}(1) = \psi_1(t_{n+1}). \quad (35)$$

Здесь использованы обозначения, совершенно аналогичные применяемым в параболическом случае.

Процесс решения аппроксимирующей задачи (33) — (35) очевидным образом может быть расчленен на следующие этапы. Сначала по заданной функции $u_0(x) = \varphi(x)$ из (34) непосредственно получаем

$$u_1(x) = \varphi(x) + \tau \Phi(x).$$

После этого последовательно находим функции $u_2(x)$, $u_3(x)$, ..., $u_N(x)$; при этом нахождение каждой из них сводится к решению одного обыкновенного дифференциального уравнения второго порядка вида (33) с граничными условиями первого рода (35).

Так как ошибки приближенных равенств (30) — (32) являются малыми величинами, вообще говоря, лишь порядка τ , то построенная вычислительная схема характеризуется только первым порядком погрешности аппроксимации. Если при аппроксимации дифференциального уравнения (27) для замены производной u_t вместо (30) использовать более точное равенство *)

$$u_t(x, t_{n+1}) \approx \frac{3u(x, t_{n+1}) - 4u(x, t_n) + u(x, t_{n-1}))}{2\tau}, \quad (36)$$

то при малых значениях $\tau > 0$, как правило, удается уменьшить часть погрешности аппроксимации исходного уравнения, зависящую от неточности замены первой производной по времени. При некоторых дополнительных предположениях более точно можно заменить первую

*) Разлагая значения $u(x, t_{n+1})$, $u(x, t_n)$ функции $u(x, t)$ по формуле Тейлора около точки (x, t_{n+1}) , нетрудно показать, что погрешность приближенного равенства (36) будет величиной порядка τ^2 .

производную и при аппроксимации начального условия. Предположив, например, что решение исходной задачи продолжимо хотя бы на один шаг τ вниз под прямую $t=0$, можно вместо (32) использовать более точное симметричное представление первой производной:

$$u_t(x, 0) \approx \frac{u(x, \tau) - u(x, -\tau)}{2\tau}. \quad (37)$$

Введенную при этом новую неизвестную функцию $u_{-1}(x)$ можно исключить, рассмотрев уравнение (33) дополнительно и при $n=0$.

Использование приближенных равенств вида (36), (37) позволит, как правило, несколько уменьшить погрешность замены исходной задачи, однако порядок этой погрешности, вообще говоря, не возрастет, так как при аппроксимации уравнения (27) была использована грубая левосторонняя замена второй производной. Если же для замены производной u_{tt} использовать более точное симметричное представление через значения функции u на трех соседних прямых, то соответствующая вычислительная схема, следует ожидать, будет мало пригодной для вычислений. К такому выводу можно прийти с помощью рассуждений, подобных проведенным в п. 1 и основанных на связи метода прямых с методом сеток. Заметим, что последнее еще в большей мере относится, очевидно, к вычислительной схеме, основанной на правосторонней замене второй производной.

Чтобы построить поперечную схему метода прямых, характеризующуюся более высоким порядком погрешности аппроксимации в сравнении с (33) — (35), можно, подобно параболическому случаю, составить, скажем, полусумму (33) и сходного уравнения, основанного на правосторонних заменах производных u_t , u_{tt} . Аппроксимацию начального условия при этом естественно проводить с помощью приближенного равенства (37). Погрешность аппроксимации такой вычислительной схемы будет, очевидно, величиной порядка τ^2 .

Если воспользоваться формулой типа (22), то можно построить и более точную трехслойную аппроксимацию исходного уравнения, в котором предварительно с помощью соответствующей замены искомой функции исключен член с первой производной по времени. Подобно параболическому случаю, дальнейшего повышения порядка

точности замены исходного уравнения можно ожидать на пути построения многослойных аппроксимаций, сходных с (13).

Процесс построения поперечных схем метода прямых для гиперболических уравнений, как мы видели, во многом схож со случаем уравнений параболического типа. Подобные аналогии еще более сильно проявляются в случае продольных схем метода. Если, например, в рассматриваемой области вместо прямых, параллельных оси x , провести прямые

$$x = x_n = nh, \quad n = 0, 1, \dots, N \quad (Nh = 1),$$

то, поступая аналогично изложенному в п. 2, для задачи (27) — (29) можно построить следующую продольную схему метода прямых:

$$u_n''(t) + g_n(t) u_n'(t) = a_n(t) \frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{h^2} + \\ + b_n(t) \frac{u_{n+1}(t) - u_{n-1}(t)}{2h} + c_n(t) u_n(t) + f_n(t), \quad t > 0, \quad (38)$$

$$u_0(t) = \psi_0(t), \quad u_N(t) = \psi_1(t), \quad t \geq 0, \quad (39)$$

$$u_n(0) = \varphi(x_n), \quad u_n'(0) = \Phi(x_n), \quad n = 1, 2, \dots, N-1. \quad (40)$$

При записи этой схемы использованы обозначения, вполне аналогичные принятым ранее.

Аппроксимирующая задача (38) — (40) может быть истолкована как задача Коши для системы $N-1$ линейных обыкновенных дифференциальных уравнений второго порядка относительно неизвестных функций $u_n(t)$, $n = 1, 2, \dots, N-1$, приближенно представляющих искомую функцию $u(x, t)$ на прямых $x = x_n$, $n = 1, 2, \dots, N-1$.

По способу построения схема (38) — (40) совершенно аналогична схеме (14) — (16). Подобное можно наблюдать и при построении продольных схем более высокого порядка аппроксимации.

4. Метод прямых для эллиптических уравнений. Пусть в прямоугольнике $\Pi \{0 \leq x \leq X, 0 \leq y \leq Y\}$ существует и единственно решение задачи Дирихле

$$u_{yy} + L(u) = f(x, y), \quad (41)$$

$$\begin{aligned} u(x, 0) &= \varphi(x), \quad u(x, Y) = \Phi(x), \\ u(0, y) &= \psi(y), \quad u(X, y) = \Psi(y), \end{aligned} \quad (42)$$

где

$$L(u) = a(x, y) u_{xx} + b(x, y) u_x + c(x, y) u, \quad a(x, y) \geq a > 0.$$

Будем предполагать решение достаточно гладким и станем искать его приближенно методом прямых.

Для задачи (41), (42), в отличие от рассмотренных ранее эволюционных задач, исчезает принципиальное различие между поперечным и продольным вариантами метода: в обоих случаях мы получаем в качестве аппроксимирующей граничную задачу для системы обыкновенных дифференциальных уравнений второго порядка.

В качестве примера построим здесь применительно к рассматриваемой задаче Дирихле одну из простейших вычислительных схем метода.

Прямоугольник Π аппроксимируем решеточной областью, проведя, скажем, прямые $y = y_n = nh$, $n = 0, 1, \dots, N$ ($Nh = Y$). Производную u_{yy} при $y = y_n$, $n = 1, 2, \dots, N - 1$, выразим приближенно через значения функции u , опираясь на известную формулу

$$u_{yy}(x, y_n) = \frac{u(x, y_{n+1}) - 2u(x, y_n) + u(x, y_{n-1}))}{h^2} - \frac{h^2}{12} \frac{\partial^4 u(x, y_n + \theta h)}{\partial y^4}, \quad -1 < \theta < 1. \quad (43)$$

Тогда, используя обозначения, совершенно аналогичные применяемым ранее, можно записать

$$\frac{u_{n+1}(x) - 2u_n(x) + u_{n-1}(x)}{h^2} + L_n(u_n) = f_n(x), \quad 0 < x < X, \quad (44)$$

$$u_0(x) = \varphi(x), \quad u_N(x) = \Phi(x), \quad 0 \leq x \leq X, \quad (45)$$

$$u_n(0) = \psi(y_n), \quad u_n(X) = \Psi(y_n), \quad n = 1, 2, \dots, N - 1. \quad (46)$$

В силу (43) дифференциально-разностная задача (44) — (46) аппроксимирует исходную задачу Дирихле с погрешностью порядка h^2 . Она сводится к решению системы обыкновенных дифференциальных уравнений второго порядка с граничными условиями первого рода.

Опираясь, например, на формулу типа (22), можно построить и более точную трехслойную аппроксимацию дифференциального уравнения (41). Соответствующая вычислительная схема метода прямых в случае достаточно

гладкого решения исходной задачи (см. (23)) будет характеризоваться погрешностью аппроксимации порядка h^4 .

Как и в случае эволюционных задач, дальнейшего повышения порядка погрешности аппроксимации можно ожидать, например, на пути построения многослойных схем метода. Способы построения таких схем и основные трудности, которые могут возникнуть при этом, очень близки рассмотренному ранее случаю продольных схем для уравнений параболического и гиперболического типов.

Вопрос построения вычислительных схем метода прямых обсуждался выше на примере ориентированных по осям прямоугольных областей. В случае областей более сложной формы применение этого метода может быть сопряжено с определенными осложнениями. В самом деле, проекции g_n , $n = 0, 1, \dots, N$, на ось координат внутренних сечений области избранными прямыми совпадут при любом числе прямых только для соответствующим образом ориентированного прямоугольника. В общем же случае n -е уравнение рассматриваемой вычислительной схемы метода, связывающее приближенные значения искомого решения на соседних прямых номеров от $n - p$ до $n + q$, без дополнительных предположений определено лишь на пересечении

$$G_n = \bigcap_{-p \leq i \leq q} g_{n+i}.$$

Если же считать такие уравнения выполняющимися и вне G_n , то даже для очень простых областей аппроксимирующая задача при некоторых N может оказаться неразрешимой. Чтобы преодолеть подобные затруднения, иногда предпочитают продолжать непрерывно внутри области функцию, задающую граничные условия, а уравнения аппроксимирующей системы рассматривают лишь на соответствующих пересечениях G_n , при этом вычислительный алгоритм приходится строить специальным образом (подробнее об этом см., например, в [2], гл. 10, § 8, п. 2).

Часто удается избежать подобных затруднений, разбивая рассматриваемую область на полосы не прямыми, а кривыми линиями, избранными с учетом формы области. Один из методов понижения размерности задач, допускающих подобное разбиение, мы сейчас и рассмотрим.

§ 7. Метод интегральных соотношений

Пусть в области D с границами $x=a$, $x=b$, $y=0$, $y=\delta=\delta(x)>0$ рассматривается следующая система m дифференциальных уравнений в частных производных дивергентного вида:

$$\frac{\partial}{\partial x} P_i(x, y, u_1, u_2, \dots, u_m) + \frac{\partial}{\partial y} Q_i(x, y, u_1, u_2, \dots, u_m) = F_i(x, y, u_1, u_2, \dots, u_m), \quad i=1, 2, \dots, m. \quad (1)$$

Здесь $u_1, u_2, \dots, u_m$ — искомые функции независимых переменных x, y , а P_i, Q_i, F_i — известные достаточно гладкие функции переменных $x, y, u_1, u_2, \dots, u_m$. Будем считать также, что имеется в общей сложности m дополнительных условий на прямых $x=a$, $x=b$ и m условий на границах $y=0$, $y=\delta(x)$.

Предполагая существование и единственность достаточно гладкого решения поставленной задачи, зададимся целью свести приближенно решение этой задачи для дифференциальных уравнений в частных производных к решению некоторой аппроксимирующей задачи для обыкновенных дифференциальных уравнений. Для этого от системы дифференциальных уравнений (1) перейдем к системе интегральных соотношений следующего вида:

$$\int_0^{\delta(x)} \frac{\partial P_i}{\partial x} f_i(y) dy + \int_0^{\delta(x)} \frac{\partial Q_i}{\partial y} f_i(y) dy = \int_0^{\delta(x)} F_i f_i(y) dy, \quad (2)$$

$$i=1, 2, \dots, m.$$

Такой переход заключается в интегрировании по y от 0 до $\delta(x)$ каждого уравнения исходной системы, умноженного предварительно на некоторую известную функцию $f_i(y)$, $i=1, 2, \dots, m$. Функции $f_i(y)$, $i=1, 2, \dots, m$, при этом играют роль сглаживающих функций. Для различных значений i эти функции могут и совпадать между собой.

Соотношения (2) легко могут быть приведены к виду

$$\begin{aligned} \frac{d}{dx} \int_0^{\delta(x)} P_i f_i(y) dy - \delta'(x) P_{i(\delta)} f_i(\delta) + Q_{i(\delta)} f_i(\delta) - \\ - Q_{i(0)} f_i(0) - \int_0^{\delta(x)} Q_i f'_i(y) dy = \int_0^{\delta(x)} F_i f_i(y) dy, \\ i=1, 2, \dots, m, \quad (3) \end{aligned}$$

где для сокращения записей введены обозначения типа

$$f_i[\delta(x)] = f_i(\delta),$$

$$S_{i(\alpha)} = S_i[x, y, u_1(x, y), u_2(x, y), \dots, u_m(x, y)]_{i=\alpha}(x).$$

Отметим, что за счет умножения исходных уравнений на сглаживающие функции и последующего интегрирования часто удается построить интегральные соотношения вида (3), более благоприятные для приближенной замены, чем дифференциальные уравнения (1). При аппроксимации соотношений (3) можно поступить, например, следующим образом.

Область D разобьем на N , вообще говоря, криволинейных полос линиями

$$y = y_n(x) = \frac{n}{N} \delta(x), \quad n = 0, 1, \dots, N, \quad (4)$$

и каждую из функций $P_i, Q_i, F_i, i = 1, 2, \dots, m$, представим тогда интерполяционно через ее значения на границах всех полос. Например, для функции P_i получим приближенное равенство

$$P_i(x, y, u_1, u_2, \dots, u_m) \approx \sum_{n=0}^N \omega_n(y) P_{i(n)}, \quad (5)$$

где через $\omega_n(y)$, $n = 0, 1, \dots, N$, обозначены интерполяционные коэффициенты, конкретный вид которых зависит от выбора базисных функций, положенных в основу интерполирования (в случае алгебраического интерполирования в качестве $\omega_n(y)$, $n = 0, 1, \dots, N$, могут быть взяты интерполяционные множители Лагранжа (см. п. 1.3.1)).

Будем искать сейчас лишь значения $u_{i(n)}(x) = u_i[x, y_n(x)]$, $i = 1, 2, \dots, m$; $n = 0, 1, \dots, N$, функций $u_i(x, y)$, $i = 1, 2, \dots, m$, на линиях (4). С целью согласования числа интегральных соотношений с числом подлежащих разысканию функций $u_{i(n)}(x)$, запишем равенства вида (3) для каждой из функций $f_{ij}(y)$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, N$, где $f_{ij}(y)$ выбираются заранее с учетом имеющейся информации об исходной задаче; при этом для каждого значения $i = 1, 2, \dots, m$ функции $f_{i1}, f_{i2}, \dots, f_{iN}$ должны быть линейно независимыми (для

различных значений i избираемые функции могут совпадать). Тогда будем иметь

$$\begin{aligned} \frac{d}{dx} \int_0^{\delta(x)} P_{if_{ij}}(y) dy - \delta'(x) P_{i(\delta)} f_{ij}(\delta) + Q_{i(\delta)} f_{ij}(\delta) - \\ - Q_{i(0)} f_{ij}(0) - \int_0^{\delta(x)} Q_{if'_{ij}}(y) dy = \int_0^{\delta(x)} F_{if_{ij}}(y) dy, \\ i = 1, 2, \dots, m; \quad j = 1, 2, \dots, N. \end{aligned}$$

Заменив в этих соотношениях функции P_i , Q_i , F_i , $i = 1, 2, \dots, m$, их приближенными выражениями типа (5), получим для $m(N+1)$ приближенных значений $u_{in}(x)$, $i = 1, 2, \dots, m$; $n = 0, 1, \dots, N$, функций $u_{i(n)}(x)$ получим систему mN обыкновенных дифференциальных уравнений первого порядка

$$\begin{aligned} \frac{d}{dx} \sum_{n=0}^N \alpha_{ijn}(x) P_{in} - \delta'(x) P_{iN} f_{ij}(\delta) + Q_{iN} f_{ij}(\delta) - \\ - Q_{i0} f_{ij}(0) - \sum_{n=0}^N \beta_{ijn} Q_{in} = \sum_{n=0}^N \alpha_{ijn}(x) F_{in}, \\ i = 1, 2, \dots, m; \quad j = 1, 2, \dots, N, \end{aligned} \quad (6)$$

где

$$\begin{aligned} \alpha_{ijn}(x) &= \int_0^{\delta(x)} \omega_n(y) f_{ii}(y) dy, \\ \beta_{ijn}(x) &= \int_0^{\delta(x)} \omega_n(y) f'_{ij}(y) dy, \end{aligned}$$

$$P_{in} = P_i[x, y_n(x), u_{1n}(x), u_{2n}(x), \dots, u_{mn}(x)]$$

(аналогично определяются Q_{in} и F_{in}). Используя условия на границе области D , часто удается поставить для системы (6) граничную задачу, которая позволяет единственным образом найти интересующие нас функции $u_{in}(x)$, $i = 1, 2, \dots, m$; $n = 0, 1, \dots, N$.

Описанный выше метод понижения размерности исходной дифференциальной задачи обычно называют *обобщенным методом интегральных соотношений*, в отличие от *простого метода интегральных соотношений*, в котором сглаживающие функции не вводятся, но, с целью согла-

сования числа интегральных соотношений с числом неизвестных функций, интегрирование выполняется не поперек всей области D , а лишь от нижней до верхней границы каждой полосы, на которые исходная область разбивается линиями (4). В этом случае вместо соотношений (3) будем иметь следующую систему интегральных соотношений:

$$\frac{d}{dx} \int_{y_{n-1}(x)}^{y_n(x)} P_i dy - y'_n(x) P_{i(n)} + y'_{n-1}(x) P_{i(n-1)} + \\ + Q_{i(n)} - Q_{i(n-1)} = \int_{y_{n-1}(x)}^{y_n(x)} F_i dy, \quad (7)$$

$$i = 1, 2, \dots, m; \quad n = 1, 2, \dots, N.$$

Как и в случае обобщенного метода интегральных соотношений, функции P_i , F_i , $i = 1, 2, \dots, m$, стоящие под знаком интеграла в равенствах (7), заменяются приближенными, например, интерполяционными выражениями через их значения на границах полос; при этом может быть использована как сквозная интерполяция вида (5), так и кусочная интерполяция, когда приближение осуществляется по значениям рассматриваемой функции лишь на части из линий разбиения (4), что позволяет обычно упростить вид аппроксимирующей задачи (удобные и достаточно точные двухслойные схемы метода можно получить, пользуясь линейным интерполированием по значениям функции на границах полосы, поперек которой выполняется интегрирование).

Простой метод интегральных соотношений можно трактовать как частный случай обобщенного. Он, например, получается из последнего, если в качестве функций $f_{ij}(y)$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, N$, взять «ступенчатые» функции

$$f_{ij}(y) = \begin{cases} 0, & y < y_{j-1}, \\ 1, & y_{j-1} \leq y \leq y_j, \\ 0, & y_j < y. \end{cases}$$

В свою очередь простой метод интегральных соотношений можно рассматривать, очевидно, как обобщение метода прямых.

§ 8. Понятие о методе моментов и методе Галёркина

Применительно к обыкновенным дифференциальным уравнениям эти методы были рассмотрены нами в § 12 гл. 6. Здесь мы изложим вычислительные схемы метода моментов и метода Галёркина на примере линейного уравнения

$$L(u) \equiv a(x, y) \frac{\partial^2 u}{\partial x^2} + 2b(x, y) \frac{\partial^2 u}{\partial x \partial y} + c(x, y) \frac{\partial^2 u}{\partial y^2} + \\ + 2d(x, y) \frac{\partial u}{\partial x} + 2e(x, y) \frac{\partial u}{\partial y} + g(x, y) u = f(x, y). \quad (1)$$

Будем предполагать, что коэффициенты этого уравнения и свободный член $f(x, y)$ есть функции, непрерывные в области D , ограниченной контуром Γ . Предположим также, что это уравнение имеет в области $\bar{D} = D + \Gamma$ эллиптический вид, т. е. всюду в этой области выполняется неравенство

$$b^2(x, y) - a(x, y) \cdot c(x, y) < 0.$$

Присоединим к уравнению (1) условие на границе Γ в виде

$$u|_{\Gamma} = \varphi(Q), \quad (2)$$

где φ — известная функция, Q — точка на контуре Γ .

Поставим задачу: найти функцию $u(x, y)$, которая в области D удовлетворяла бы уравнению (1), а на контуре Γ — граничному условию (2).

Для построения вычислительной схемы метода моментов рассмотрим две системы функций

$$\{\varphi_k(x, y)\}, \quad \{\psi_k(x, y)\}, \quad k = 0, 1, 2, \dots$$

Выбор функций $\varphi_k(x, y)$ и $\psi_k(x, y)$ связывается с видом уравнения (1) и видом граничного условия (2). В рассматриваемом случае на функции $\varphi_k(x, y)$, $\psi_k(x, y)$ накладываются следующие условия:

1) функции $\psi_k(x, y)$, $k = 0, 1, 2, \dots$, непрерывны в области D ;

2) функции $\psi_k(x, y)$, $k=0, 1, 2, \dots$, образуют на классе непрерывных в области D функций замкнутую систему в том смысле, что из выполнения условия

$$\iint_D F(x, y) \cdot \psi_k(x, y) dx dy = 0, \quad k=0, 1, 2, \dots,$$

где $F(x, y)$ — некоторая непрерывная в D функция, следует, что $F(x, y) \equiv 0$;

3) функции $\varphi_k(x, y)$, $k=0, 1, 2, \dots$, дважды непрерывно дифференцируемы в D ;

4) при любом конечном n функции $\varphi_1(x, y), \dots, \varphi_n(x, y)$ линейно независимы в D ;

5) функция $\varphi_0(x, y)$ удовлетворяет граничному условию (2), т. е. $\varphi_0|_{\Gamma} = \varphi(Q)$, а $\varphi_k|_{\Gamma} = 0$, $k=1, 2, \dots$;

6) для любой дважды непрерывно дифференцируемой в D функции $u(x, y)$, удовлетворяющей условию (2), и любого $\varepsilon > 0$, найдутся такие n и $a_1, a_2, \dots, a_n$, что выполняются условия:

$$\begin{aligned} \left| \frac{\partial^i u(x, y)}{\partial x^i} - \frac{\partial^i u_n(x, y)}{\partial x^i} \right| &< \varepsilon, \\ \left| \frac{\partial^i u(x, y)}{\partial y^i} - \frac{\partial^i u_n(x, y)}{\partial y^i} \right| &< \varepsilon, \quad i=0, 1, 2, \\ \left| \frac{\partial^2 u(x, y)}{\partial x \partial y} - \frac{\partial^2 u_n(x, y)}{\partial x \partial y} \right| &< \varepsilon, \end{aligned}$$

где $u_n(x, y) = \varphi_0(x, y) + \sum_{k=1}^n a_k \varphi_k(x, y)$, $a_1, a_2, \dots, a_n$ — некоторые числовые параметры.

В методе моментов функция $u_n(x, y)$ принимается за n -е приближение к решению граничной задачи (1,2), если параметры a_k , $k=1, 2, \dots, n$, определяются по следующему правилу. Подставим в $L(u)$ вместо u линейную комбинацию u_n . Точное выполнение уравнения $L(u_n) - f = 0$ равносильно условию

$$\iint_D [L(u_n) - f] \psi_k(x, y) dx dy = 0 \quad \text{при} \quad k=0, 1, 2, \dots$$

Отметим, что, имея в своем распоряжении n произвольных числовых параметров a_k ($k=1, 2, \dots, n$), нельзя

рассчитывать за счет их выбора удовлетворить бесконечному числу условий.

Определим a_k из условий ортогональности $L(u_n) - f$ к первым n функциям $\psi_k(x, y)$:

$$\iint_D [L(u_n) - f] \psi_k(x, y) dx dy = 0, \quad k = 0, 1, \dots, n-1. \quad (3)$$

Формулы (3) представляют собой систему линейных алгебраических уравнений, которую мы запишем в виде

$$\sum_{i=1}^n A_{ki} a_i + A_k = 0, \quad k = 0, 1, \dots, n-1, \quad (4)$$

где

$$A_{ki} = \iint_D L(\varphi_i) \psi_k dx dy,$$

$$A_k = \iint_D [L(\varphi_0) - f] \psi_k dx dy.$$

Пусть $a_1^*, a_2^*, \dots, a_n^*$ есть решение системы уравнений (4), тогда n -е приближение к искомому решению граничной задачи (1), (2) в методе моментов вычисляется по формуле

$$u_n^*(x, y) = \varphi_0(x, y) + \sum_{k=1}^n a_k^* \varphi_k(x, y).$$

Метод Галёркина является частным случаем метода моментов. В нем система функций $\{\varphi_k(x, y)\}$ выбирается так же, как в методе моментов, а функции $\psi_k(x, y)$ определяются по правилу

$$\psi_k(x, y) = \varphi_{k+1}(x, y), \quad k = 0, 1, 2, \dots$$

Отметим, что если оператор L имеет вид

$$L(u) \equiv -\frac{\partial}{\partial x} \left(p(x, y) \frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial y} \left(p(x, y) \frac{\partial u}{\partial y} \right) + q(x, y) u \quad (5)$$

и $p(x, y) \geq p_0 > 0$, $q(x, y) \geq 0$, $p(x, y)$ — непрерывно дифференцируемая функция, $q(x, y)$ — непрерывная функция в области D , то метод Галёркина для задачи (1), (2) с оператором (5) обладает сходимостью в среднем *).

*) Говорят, что последовательность $\{u_n(x, y)\}$ сходится в среднем к $u(x, y)$ в области D , если выполнено условие

$$\lim_{n \rightarrow \infty} \iint_D (u_n(x, y) - u(x, y))^2 dx dy = 0.$$

В монографии [5] читатель может найти достаточно полное изложение теории метода Галёркина и вопросов, связанных с его численной реализацией.

ЛИТЕРАТУРА

1. Бахвалов Н. С., Численные методы, т. 1, М., «Наука», 1973.
2. Березин И. С., Жидков Н. П., Методы вычислений, т. 2, М., «Наука», 1962.
3. Годунов С. К., Рябенский В. С., Разностные схемы, М., «Наука», 1973.
4. Крылов В. И., Бобков В. В., Монастырный П. И., Вычислительные методы высшей математики, т. 2, Минск, «Вышэйшая школа», 1975.
5. Михлин С. Г., Численная реализация вариационных методов, М., «Наука», 1966.
6. Самарский А. А., Введение в теорию разностных схем, М., «Наука», 1971.

Г Л А В А 8

ЧИСЛЕННОЕ РЕШЕНИЕ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

§ 1. Введение

Идеи методов численного решения интегральных уравнений мы постараемся выяснить на примере простого случая, когда неизвестная функция зависит от одного аргумента и интегралы, в которых она содержится, являются однократными и неособенными. Такие уравнения не исчерпывают всех видов интегральных уравнений, встречающихся в приложениях, но с ними приходится сталкиваться достаточно часто; кроме того, методы их решения являются типичными для других случаев и знание этих методов полезно для решения более сложных уравнений.

В виде, достаточно общем для нашей цели, эти уравнения можно записать в форме

$$F(x, \varphi(x), z) = 0, \quad (1)$$

где $\varphi(x)$ есть искомая функция, а буквой z обозначен один из следующих интегралов:

$$z = \int_a^b f(x, s, \varphi(s)) ds \text{ или } z = \int_a^x f(x, s, \varphi(s)) ds. \quad (2)$$

С формальной точки зрения второй из интегралов является частным случаем первого, когда функция f такова, что $f(x, s, \varphi) \equiv 0$ при $s > x$ и любых значениях φ . Но это свойство функции f имеет настолько глубокие следствия, что делает теории уравнений с интегралами первого или второго видов (2) принципиально различными.

Уравнения, содержащие интегралы первого из видов (2), называются уравнениями Фредгольма, а уравнения с интегралами второго вида — уравнениями Вольтерра.

Остановим внимание на отличительных особенностях каждого из видов уравнений и укажем на то, какое влияние они оказывают на методы приближенного их решения.

В уравнениях вольтеррова вида значения функции φ в точке x определяются ее значениями в предшествующих точках $t < x$, значения же в последующих точках $t > x$ не влияют на $\varphi(x)$. Для приближенного решения уравнения на отрезке $a \leq x \leq b$, где разыскивается φ , берут сетку точек $a \leq x_0 < x_1 < \dots < x_N \leq b$ и находят приближенно значения $\varphi(x_k) = \varphi_k$ только в точках сетки. Для этого интегральное уравнение заменяют близкой к нему в каком-то смысле системой $N+1$ численных уравнений между величинами φ_k . Эта система строится так, чтобы по известным значениям $\varphi_0, \dots, \varphi_{k-1}$ могло быть построено уравнение для нахождения φ_k

$$\Phi_k(\varphi_0, \varphi_1, \dots, \varphi_{k-1}, \varphi_k) = 0. \quad (3)$$

Часто такое уравнение строят в форме

$$\varphi_k = g_k(\varphi_0, \dots, \varphi_{k-1}, \varphi_k), \quad (4)$$

позволяющей удобно применять метод итерации. В частных случаях уравнение (4) может иметь форму рекурсионного правила, позволяющего вычислять φ_k по предшествующим значениям φ без решения уравнения

$$\varphi_0 = a_0, \quad \varphi_k = g_k(\varphi_0, \varphi_1, \dots, \varphi_{k-1}) \quad (k = 1, 2, \dots, N).$$

Для уравнений Фредгольма система для нахождения φ_k должна иметь иную форму, чем (3). Интеграл z первого из видов (2) зависит от значений $\varphi(x)$ на всем отрезке $a \leq x \leq b$. Поэтому система должна быть такой, чтобы в каждое ее уравнение входили значения φ_k во всех точках x_k сетки и система должна решаться «в целом», а не последовательно.

Ниже мы ограничимся рассмотрением линейных уравнений. В фредгольмовом случае линейные уравнения имеют вид

$$A(x) \varphi(x) + \int_a^b K(x, s) \varphi(s) ds = F(x), \quad (5)$$

а для случая Вольтерра —

$$A(x)\varphi(x) + \int_a^x K(x, s)\varphi(s)ds = F(x). \quad (6)$$

Когда коэффициент $A(x)$ тождественно равен нулю, уравнения (5) и (6) называют уравнениями первого рода. Они имеют вид

для случая Фредгольма

$$\int_a^b K(x, s)\varphi(s)ds = F(x), \quad (*)$$

для случая Вольтерра

$$\int_a^x K(x, s)\varphi(s)ds = F(x). \quad (**)$$

Если же коэффициент $A(x)$ отличен от нуля всюду на отрезке $[a, b]$, уравнения (5) и (6) называются уравнениями второго рода. Они, очевидно, могут быть приведены к следующим каноническим формам:

в случае уравнения Фредгольма

$$\varphi(x) + \int_a^b K(x, s)\varphi(s)ds = f(x), \quad a \leq x \leq b, \quad (7)$$

в случае уравнения Вольтерра

$$\varphi(x) + \int_a^x K(x, s)\varphi(s)ds = f(x), \quad a \leq x \leq b. \quad (8)$$

В дальнейшем будут рассматриваться только эти уравнения. Функции $f(x)$ и $K(x, s)$ предполагаются непрерывными в областях соответственно $a \leq x \leq b$, $a \leq s \leq x \leq b$ для уравнений Вольтерра и $a \leq x \leq b$, $a \leq s, x \leq b$ для уравнений Фредгольма.

Прежде чем перейти к изложению методов численного решения таких уравнений, мы объясним причины, побудившие нас отказаться от рассмотрения уравнений первого рода. Это даст нам возможность на их примере

ознакомиться с классом задач, которые получили название некорректных. Содержание приведенного термина требует пояснения. Рассмотрим какую-либо задачу. Все величины, определяющие решение, называются входными данными задачи. Например, для уравнения (7) входными данными являются свободный член $f(x)$, ядро $K(x, s)$ и отрезок интегрирования $[a, b]$. Для решения уравнений с частными производными входными данными будут: само уравнение или, точнее говоря, параметры, определяющие уравнение (его коэффициенты, свободные члены и т. д.), затем присоединяемые к уравнению начальные или граничные условия, или и те и другие.

Очень часто входные данные зависят от измерений и, стало быть, известны только приближенно. Поэтому особый интерес имеют задачи, для которых выполняются следующие два условия: 1) решение существует при любых входных данных, когда они изменяются в некоторых допустимых границах; 2) решение задачи мало изменяется при малом изменении входных данных, т. е. является устойчивым относительно изменений этих данных. Такие задачи называются *корректными* или *корректно поставленными*.

Но среди прикладных задач встречаются такие, изучение которых имеет большое значение, но которые не обладают каким-либо из указанных свойств.

Задачи, которые не имеют решения или решения которых неустойчивы относительно входных данных, называются *некорректными* (некорректно поставленными).

Уравнение первого рода вида Фредгольма (*) является примером такой задачи. По существу, равенство (*) дает интегральное преобразование функции $\varphi(s)$ в $F(x)$. В нем φ является оригиналом и F — изображением.

Для простоты будем считать ядро $K(x, s)$ непрерывной функцией в квадрате $a \leq s, x \leq b$ и оригинал $\varphi(s)$ — принадлежащим множеству I_0 абсолютно интегрируемых на $[a, b]$ функций. При этих предположениях интеграл, стоящий слева в равенстве (*), будет непрерывной на $[a, b]$ функцией x . Преобразование будет отображать множество I_0 во множество непрерывных на $[a, b]$ функций C_0 .

Может оказаться, что множество изображений составит только часть C_0 . Тогда, если в качестве $F(x)$ мы возьмем непрерывную функцию, не входящую в эту часть,

уравнение (*) не будет иметь решения $\varphi(s)$ в множестве I_0 и задача (*) будет неразрешимой. Например, пусть ядро $K(x, s)$ непрерывно и имеет непрерывную частную производную по аргументу x в квадрате $a \leq s, x \leq b$. Интеграл в (*) будет непрерывно дифференцируемой функцией x на $[a, b]$ и преобразование (*) отображает I_0 в множество непрерывно дифференцируемых на $[a, b]$ функций. Поэтому, если в уравнении (*) $F(x)$ есть непрерывная, но не дифференцируемая на $[a, b]$ функция, уравнение не разрешимо в I_0 .

Приведенные рассуждения показывают, что уравнение (*) разрешимо не при всяких входных данных и, стало быть, не обладает первым из указанных свойств корректных задач.

Проверим также, что оно не обладает и свойством устойчивости решения. Пусть $\varphi(s)$ удовлетворяет уравнению (*). Изменим значительно φ на очень коротком частичном отрезке $[\alpha, \beta]$ и вновь полученную функцию обозначим φ_1 . После изменения (*) перейдет в новое равенство

$$\int_a^b K(x, s) \varphi_1(s) ds = F_1(x). \quad (***)$$

Если длина $\beta - \alpha$ частичного отрезка взята достаточно малой, то новая функция $F_1(x)$ будет при всех x мало отличаться от $F(x)$. Мы могли бы добиться близости F_1 к F , изменяя значительно φ не на одном частичном отрезке, а на многих отрезках, сумма длин которых будет достаточно малой, и также получили бы два интегральных уравнения (*) и (***) с одним и тем же ядром K и с близкими свободными членами F и F_1 , но с решениями φ и φ_1 , значительно различающимися между собой. Это показывает неустойчивость решения φ относительно изменений свободного члена F .

Несколько иначе обстоит дело с уравнением первого рода вольтеррова вида (**). Допустим, что ядро $K(x, s)$ и свободный член $F(x)$ имеют непрерывные производные по x в областях $a \leq s \leq x \leq b$ и $a \leq x \leq b$ соответственно.

Отметим предварительно, что при $x = a$ левая часть уравнения (**) обращается в нуль, и поэтому необходимым условием существования решения будет $F(a) = 0$.

Вычислим производную от обеих частей уравнения (**):

$$K(x, x) \varphi(x) + \int_a^x K_x(x, s) \varphi(s) ds = F'(x). \quad (****)$$

Здесь возможны три следующих случая.

1) Когда

$$K(x, x) \neq 0 \quad (a \leq x \leq b),$$

это уравнение путем деления на $K(x, x)$ приводится к уравнению второго рода (8), задача решения которого является корректной. Методы решения изложены в § 2, но можно применить аналогичные методы непосредственно к (**).

2) Пусть

$$K(x, x) \equiv 0, \quad a \leq x \leq b.$$

Тогда получится однородное уравнение, аналогичное (**):

$$\int_a^x K_x(x, s) \varphi(s) ds = F'(x).$$

К нему можно применить предыдущие рассуждения: проверить, будет ли выполняться необходимое условие разрешимости $F'(a) = 0$, вычислить производные от обеих частей уравнения и т. д.

Особенным здесь будет случай, когда правая или левая части уравнения не имеют производной или производная от интеграла не может быть вычислена по обычному правилу. Такой случай требует дополнительного изучения.

3) Искключительным является случай, когда функция $K(x, x)$ не есть тождественный нуль, но обращается в нуль на некотором множестве точек на $[a, b]$. Тогда интегральное уравнение (****) имеет особые точки и должно быть исследовано, так как общие методы решения здесь могут оказаться неприменимыми.

В предыдущих главах книги рассматривались методы решения только корректных и неособенных задач. Такие методы для решения некорректных задач или задач с особенностями могут оказаться неприменимыми. Для решения задач такого типа создаются свои методы, которые в книге не излагаются. Все это и заставило авторов отказаться от рассмотрения уравнений первого рода.

§ 2. Уравнения вольтеррова вида второго рода

1. Вид вычислительного правила и теорема о сходимости вычислительного процесса. Допустим, что для решения уравнения (8) на конечном отрезке $a \leq x \leq b$ взята сетка равноотстоящих точек с шагом h : $x_k = a + kh$ [$k = 0, 1, \dots, N$; $a + Nh \leq b < a + (N+1)h$].

Положим в интегральном уравнении $x = x_k$ и рассмотрим систему равенств

$$\begin{aligned} \varphi(x_0) &= f(a), \\ \varphi(x_n) + \int_a^{x_n} K(x_n, s) \varphi(s) ds &= f(x_n) \quad (n = 1, \dots, N). \end{aligned} \quad (1)$$

Пусть для вычисления интегрального члена последнего равенства взята квадратурная формула с узлами в точках $x_0, x_1, \dots, x_n$

$$\int_a^{x_n} K(x_n, s) \varphi(s) ds = h \sum_{i=0}^n A_{ni} K_{ni} \varphi(x_i) + r_n. \quad (2)$$

Здесь $K_{ni} = K(x_n, x_i)$ и r_n — остаток квадратурного правила. Множитель h перед суммой выделен для того, чтобы коэффициенты A_{ni} были безразмерными величинами. Внося это представление интеграла в равенства (1), приведем их к виду

$$\begin{aligned} \varphi(x_0) &= f_0, \\ \varphi(x_n) + h \sum_{i=0}^n A_{ni} K_{ni} \varphi(x_i) + r_n &= f_n, \quad f_n = f(x_n) \\ &\quad (n = 1, \dots, N). \end{aligned} \quad (3)$$

Отбрасывая остаток r_n , получим уравнения для приближенных значений функции φ в узлах x_n . Эти значения обозначим через φ_n . При вычислениях уравнения для φ_n будут выполняться неточно. Если погрешность, с которой они выполняются, обозначить через δ_n , то точное правило вычисления φ_n может быть записано так:

$$\varphi_n + h \sum_{i=0}^n A_{ni} K_{ni} \varphi_i = f_n + \delta_n \quad (n = 1, 2, \dots, N). \quad (4)$$

Рассмотрим погрешность приближенных значений $\varepsilon_n = \varphi_n - \varphi(x_n)$ ($n=0, 1, \dots$); при этом $\varphi(x_0) = f(x_0)$, а φ_0 есть приближенное значение, полученное при вычислении $f(x_0)$. Уравнение для погрешности при $n \geq 1$ получится, если из (4) вычесть почленно уравнение (3)

$$\varepsilon_n + h \sum_{i=0}^n A_{ni} K_{ni} \varepsilon_i = r_n + \delta_n \quad (n=1, 2, \dots, N). \quad (5)$$

Для получения оценки ε_n , гарантирующей равномерную сходимость приближенного решения φ_n к точному решению $\varphi(x)$, предположим, что при всех значениях $n=1, \dots, N$ выполняются неравенства

$$|r_n| \leq r = r(h), \quad |\delta_n| \leq \delta = \delta(h).$$

Введем числа A, K , определив их неравенствами

$$|A_{ni}| \leq A < \infty \quad (i \leq n, n=1, 2, \dots), \quad |K(x, s)| \leq K.$$

В качестве A и K могут быть взяты, например,

$$A = \max_{in} |A_{ni}| \quad \text{и} \quad K = \max_{xs} |K(x, s)|.$$

Выделим в сумме, стоящей слева в (3), слагаемое с ε_n и перепишем равенство в виде

$$(1 + hA_{nn}K_{nn}) \varepsilon_n = r_n + \delta_n - h \sum_{i=1}^{n-1} A_{ni} K_{ni} \varepsilon_i \quad (n \geq 1). \quad (6)$$

Предположим, что при всех достаточно малых h и $n=1, 2, \dots, N$ выполняется неравенство

$$h |A_{nn}K_{nn}| \leq q < 1 \quad (n=1, \dots, N). \quad (7)$$

Тогда из (6) получается рекурсионное неравенство для погрешностей ε_n

$$|\varepsilon_n| \leq \frac{1}{1-q} \left(r + \delta + hAK \sum_{i=1}^{n-1} |\varepsilon_i| \right). \quad (8)$$

Введем величины E_n ($n=0, 1, \dots$), определив их начальным условием $E_0 = |\varepsilon_0|$ и рекурсионным соотношением

$$E_n = \frac{1}{1-q} \left(r + \delta + hAK \sum_{i=0}^{n-1} E_i \right), \quad i=1, 2, \dots \quad (9)$$

Легко проверяется, что $|\varepsilon_n| \leq E_n$. Для этого воспользуемся индукцией. При $n=0$ неравенство, очевидно, верно, так как $|\varepsilon_0| = E_0$. Пусть $|\varepsilon_n| \leq E_i$ при $i=0, 1, \dots, n-1$. Тогда из (8) и (9) следует

$$|\varepsilon_n| \leq \frac{1}{1-q} \left(r + \delta + hAK \sum_{i=0}^{n-1} E_i \right) = E_n.$$

Для упрощения последующих записей введем обозначения:

$$\frac{1}{1-q} (r + \delta) = \alpha, \quad \frac{1}{1-q} hAK = \beta.$$

Соотношение (9) примет вид

$$E_n = \alpha + \beta \sum_{i=0}^{n-1} E_i. \quad (10)$$

Убедимся в том, что величина

$$E_n = (\alpha + \beta E_0) (1 + \beta)^{n-1} \quad (n=1, 2, \dots) \quad (11)$$

является решением этого уравнения. Для этого подставим в правую часть (10) значение E_i ($i \geq 1$). Получим

$$\begin{aligned} \alpha + \beta \sum_{i=0}^{n-1} E_i &= \alpha + \beta E_0 + \beta (\alpha + \beta E_0) \sum_{i=1}^{n-1} (1 + \beta)^{i-1} = \\ &= (\alpha + \beta E_0) \left\{ 1 + \beta \cdot \frac{1}{\beta} [(1 + \beta)^{n-1} - 1] \right\} = \\ &= (\alpha + \beta E_0) (1 + \beta)^{n-1} = E_n. \end{aligned}$$

Этот результат показывает, что величина (11) действительно удовлетворяет уравнению (10).

Изложенное выше позволяет утверждать, что при всяких значениях $n=1, 2, \dots$ для погрешности ε_n верна оценка

$$\begin{aligned} |\varepsilon_n| &\leq (\alpha + \beta |\varepsilon_0|) (1 + \beta)^{n-1} = \\ &= \frac{1}{1-q} \left(r + \delta + hAK |\varepsilon_0| \right) \left(1 + \frac{1}{1-q} hAK \right)^{n-1}. \end{aligned}$$

Получим равномерную относительно n оценку погрешности. Воспользуемся очевидными неравенствами $n \leq \leq N \leq \frac{b-a}{h}$ и заменим показатель степени $n-1$ большей

величиной $\frac{1}{h}(b-a)$. Воспользуемся также тем, что при $t \rightarrow +0$ величина $(1+t)^{1/t}$ возрастает и стремится к e . Все указанное позволит записать систему неравенств

$$\begin{aligned} \left(1 + \frac{1}{1-q} hAK\right)^{n-1} &\leq \left(1 + \frac{1}{1-q} hAK\right)^{\frac{b-a}{h}} = \\ &= \left[\left(1 + \frac{1}{1-q} hAK\right)^{\frac{1-q}{AK} \cdot \frac{1}{h}}\right]^{\frac{b-a}{1-q} AK} < e^{\frac{b-a}{1-q} AK}, \end{aligned}$$

что дает следующую равномерную относительно n оценку ε_n :

$$|\varepsilon_n| \leq \frac{1}{1-q} e^{\frac{b-a}{1-q} AK} (r + \delta + AK |\varepsilon_0| h). \quad (12)$$

Полученная оценка погрешности позволяет высказать заключение о сходимости приближенного решения φ_n ($n=0, 1, \dots, N$) к точному решению $\varphi(x)$ ($a \leq x \leq b$). Предположим, что число узлов $N+1$ сетки неограниченно возрастает; при этом для определенности допустим, что шаг сетки всегда берется равным $\frac{1}{N}(b-a) = h$. Если окажется, что для избранного квадратурного правила (9) верхняя граница $r(h)$ погрешности может быть взята стремящейся к нулю и если точность вычислений выбрать так, чтобы граница δ погрешностей δ_n вычислений стремилась к нулю, то правая часть неравенства (12) будет стремиться к нулю. Поэтому можно утверждать, что приближенное решение φ_n ($n=0, 1, \dots, N$) будет равномерно сходиться к точному решению $\varphi(x)$. Последнее означает, что для любого $\varepsilon > 0$ существует такое значение h_0 , что при всяких $h < h_0$ будут выполняться неравенства

$$|\varphi_n - \varphi(x_n)| < \varepsilon \quad (n=0, 1, \dots, N). \quad (13)$$

Заметим, что в неравенстве (12) справа в скобках стоят слагаемые, зависящие от различных факторов. Первое из них r определяется свойствами интегрального уравнения и избранным для вычислений квадратурным

правилom; второе — δ учитывает погрешности, которые совершаются при нахождении приближенных значений φ_n , и, наконец, третье связано с погрешностью ε_0 вычисления начального значения $\varphi(x_0)$ решения. Величиной двух последних членов можно при вычислениях распорядиться, увеличивая или уменьшая точность вычислений. Этой возможностью обычно пользуются так, чтобы соответствующие всем членам доли погрешности были сравнимы по величине или по меньшей мере имели одинаковый порядок малости.

Сформулируем теперь теорему о сходимости, вытекающую из проделанных исследований.

Будем считать отрезок $[a, b]$ произвольным отрезком конечной длины, ядро $K(x, s)$ непрерывной функцией в области $a \leq s \leq x \leq b$ (следовательно, ограниченной) и свободный член $f(x)$ непрерывным на отрезке $a \leq x \leq b$. При этих условиях уравнение (1.8) имеет и притом единственное решение $\varphi(x)$.

Теорема 1. Пусть для вычислительного процесса (4) выполняются условия:

1) погрешность r_n квадратурной формулы (2) стремится к нулю равномерно относительно n , когда $h \rightarrow 0$;

2) существует число A такое, что при всех значениях i, n, h выполняется неравенство

$$|A_{ni}| \leq A < \infty;$$

3) погрешность δ_n вычислительного правила стремится к нулю равномерно относительно n при $h \rightarrow 0$.

Тогда при всяких достаточно малых h приближенное решение φ_n может быть построено по правилу (4) и для него будет выполняться неравенство (13) при любом заданном положительном ε , если h будет достаточно малым.

Закончим настоящий пункт указанием на необходимость составления начала расчетной таблицы при применении правила (4). Коэффициенты A_{ni} обычно выбирают так, чтобы локальная погрешность r_n правила имела во всех узлах один и тот же порядок малости относительно h , т. е. чтобы погрешность r_n при всяких n имела представление вида

$$r_n = h^k C_n(h), \quad C_n(h) \rightarrow C_n \neq 0 \quad \text{при} \quad h \rightarrow 0.$$

Предположим теперь, что нам необходимо вычислить φ_1 . Для этого рассмотрим квадратурную формулу (2) при $n = 1$:

$$\int_a^{a+h} K(x_1, s) \varphi(s) ds = h [A_{10} K_{10} \varphi(a) + A_{11} K_{11} \varphi(a+h)] + r_1,$$

$$x_0 = a, \quad x_1 = a + h.$$

Она содержит два численных параметра A_{10} и A_{11} , которые мы можем выбрать так, чтобы формула имела возможно более высокую степень точности. Максимум, что мы можем получить в этом отношении — это сделать формулу точной для многочленов первой степени. Тогда формула станет известным правилом трапеций, для которого $A_{10} \equiv A_{11} = \frac{1}{2}$:

$$\int_a^{a+h} K(x_1, s) \varphi(s) ds = \frac{h}{2} [K_{10} \varphi(a) + K_{11} \varphi(a+h)] + r_1.$$

Остаточный член r_1 этого правила приведен в равенстве (5.3.3), которое для рассматриваемого интеграла имеет вид

$$r_1 = -\frac{1}{12} h^3 \left[\frac{\partial^2}{\partial s^2} K \varphi \right]_{s=\eta}, \quad a \leq \eta \leq a + h.$$

Отсюда видно, что r_1 имеет относительно h третий порядок малости. Если окажется, что принятый в вычислениях порядок k малости r_n больше трех, то формулу (4) нельзя применять к нахождению φ_1 . Тогда эта величина должна быть найдена с необходимой точностью предварительно, до применения вычислительного правила (4).

Аналогично, чтобы воспользоваться правилом (4) для нахождения φ_2 , нужно положить в (2) $n = 2$ и рассмотреть квадратурную формулу

$$\int_a^{a+2h} K(x_2, s) \varphi(s) ds =$$

$$= h [A_{20} K_{20} \varphi(a) + A_{21} K_{21} \varphi(a+h) + A_{22} K_{22} \varphi(a+2h)] + r_2.$$

Наивысшая степень алгебраической точности, которую можно достичь при помощи выбора коэффициентов A_{20} ,

A_{21} , A_{22} , равна 3 и достигается она в правиле парабол, для которого $A_{20} = A_{22} = \frac{1}{3}$, $A_{21} = \frac{4}{3}$:

$$\begin{aligned} \int_a^{a+2h} K(x_2, s) \varphi(s) ds = \\ = \frac{h}{3} [K_{20}\varphi(a) + 4K_{21}\varphi(a+h) + K_{22}\varphi(a+2h)] + r_2. \end{aligned}$$

Форма остаточного члена правила парабол указана в равенстве (5.3.7):

$$r_2 = -\frac{1}{90} h^5 \left[\frac{\partial^4}{\partial s^4} K\varphi \right]_{s=\eta}, \quad a \leq \eta \leq a+2h.$$

Остаток r_2 имеет пятый порядок малости относительно h , и если окажется, что $k > 5$, то правило вычислений (4) нецелесообразно применять для вычисления φ_2 , и это значение, так же как и φ_1 , необходимо найти заранее, до применения (4) и т. д.

Значения φ_n , которые должны быть найдены предварительно, составят начало расчетной таблицы. Оно зависит от порядка k малости r_n . Если принять во внимание степени точности интерполяционных квадратурных правил, о которых говорилось в § 3 гл. 5, то можно сказать, что в начало таблицы должны войти значения φ_n по меньшей мере с индексами n , для которых выполняются неравенства

$$\begin{aligned} n < k-3, \quad \text{когда } n - \text{четное число,} \\ n < k-2, \quad \text{когда } n - \text{нечетное число.} \end{aligned}$$

Значения φ_n , входящие в начало таблицы, можно находить, например, решая интегральное уравнение в окрестности точки a при помощи степенного ряда, если функция $f(x)$ и ядро $K(x, s)$ являются аналитическими функциями, или применяя для вычисления значений решения формулу трапеций с уменьшенным шагом h .

2. Правила вычислений, основанные на формуле Эйлера — Маклорена. В качестве примера построения вычислительного правила вида (4) рассмотрим задачу использования для этой цели разностного варианта

формулы Эйлера — Маклорена *)

$$\begin{aligned} \int_a^{a+nh} y(x) dx = & h \left[\frac{1}{2} y_0 + y_1 + \dots + y_{n-1} + \frac{1}{2} y_n \right] - \\ & - \frac{h}{12} (\Delta y_{n-1} - \Delta y_0) - \frac{h}{24} (\Delta^2 y_{n-2} + \Delta^2 y_0) - \\ & - \frac{19h}{720} (\Delta^3 y_{n-3} - \Delta^3 y_0) - \frac{3h}{160} (\Delta^4 y_{n-4} + \Delta^4 y_0) - \dots \\ & \dots - h C_{2m} [\Delta^{2m} y_{n-2m} + \Delta^{2m} y_0] + r_{n, 2m}, \quad (14) \end{aligned}$$

где $r_{n, 2m} = O(h^{2m+1})$, если функция y имеет на $[a, a+nh]$ непрерывную производную порядка $2m+1$ и $2m < n$.

Если, как и выше, рассмотреть сетку равноотстоящих точек $x_n = a + nh$, $h = \frac{1}{N}(b-a)$, положить в уравнении (1.8) $x = x_n$, применить к вычислению интеграла правило Грегори (14) и, наконец, отбросить остаточный член $r_{n, 2m}$, то для нахождения приближенных значений φ_n функции φ получим систему уравнений

$$\begin{aligned} \varphi_n + h \left[\frac{1}{2} K_{n0} \varphi_0 + K_{n1} \varphi_1 + \dots + K_{nn-1} \varphi_{n-1} + \frac{1}{2} K_{nn} \varphi_n \right] - \\ - \frac{h}{12} [\Delta_s (K_{nn-1} \varphi_{n-1}) - \Delta_s (K_{n0} \varphi_0)] - \\ - \frac{h}{24} [\Delta_s^2 (K_{nn-2} \varphi_{n-2}) + \Delta_s^2 (K_{n0} \varphi_0)] - \dots \\ \dots - h C_{2m} [\Delta_s^{2m} (K_{nn-2m} \varphi_{n-2m}) + \Delta_s^{2m} (K_{n0} \varphi_0)] = f_n. \quad (15) \end{aligned}$$

Знаком Δ_s обозначен оператор конечной разности по аргументу s :

$$\begin{aligned} \Delta_s (K_{nl} \varphi_l) &= K_{nl+1} \varphi_{l+1} - K_{nl} \varphi_l, \\ \Delta_s^2 (K_{nl} \varphi_l) &= K_{nl+2} \varphi_{l+2} - 2K_{nl+1} \varphi_{l+1} + K_{nl} \varphi_l, \dots \end{aligned}$$

Применять к вычислениям правило (15) можно при $n \geq 2m$, при этом необходимо составить начало расчетной таблицы, которое содержало бы предварительно найденные значения $\varphi_0, \dots, \varphi_{2m-1}$. Кроме того, конечная разность высшего порядка $2m$, применяемая в вычислениях, должна при принятой точности вычислений сохра-

*) Том I, (5.8.5).

нять постоянное значение или, что равносильно, разности порядка, большего $2m$, должны обращаться в нуль при этой точности.

Остановимся более подробно на простейшем и типичном случае $m=1$, когда в вычислениях применяются конечные разности по s до второго порядка:

$$\begin{aligned} \varphi_n + h \left[\frac{1}{2} K_{n0} \varphi_0 + K_{n1} \varphi_1 + \dots + K_{nn-1} \varphi_{n-1} + \frac{1}{2} K_{nn} \varphi_n \right] - \\ - \frac{h}{12} [\Delta_s (K_{nn-1} \varphi_{n-1}) - \Delta_s (K_{n0} \varphi_0)] - \\ - \frac{h}{24} [\Delta_s^2 (K_{nn-2} \varphi_{n-2}) + \Delta_s^2 (K_{n0} \varphi_0)] = \\ = \varphi_n + h T_n + \frac{h}{12} A_n + \frac{h}{24} B_n + \frac{3}{8} h K_{nn} \varphi_n = f_n, \\ T_n = \sum_{i=0}^{n-1} K_{ni} \varphi_i - \frac{1}{2} K_{n0} \varphi_0, \quad A_n = K_{nn-1} \varphi_{n-1} + \Delta_s (K_{n0} \varphi_0), \\ B_n = A_n + \Delta_s (K_{nn-2} \varphi_{n-2}) - \Delta_s (K_{n1} \varphi_1). \end{aligned}$$

Отсюда получаются следующие расчетные формулы для φ_n :

$$\begin{aligned} \varphi_2 = \left(1 - \frac{1}{3} h K_{22} \right)^{-1} \left(f_2 - \frac{4}{3} h K_{21} \varphi_1 - \frac{1}{3} h K_{20} \varphi_0 \right), \\ \varphi_n = \left(1 - \frac{3}{8} h K_{nn} \right)^{-1} \left(f_n - h T_n - \frac{h}{12} A_n - \frac{h}{24} B_n \right), \\ n = 3, 4, 5, \dots \end{aligned} \quad (16)$$

Для применения их к вычислениям шаг h должен быть взят настолько малым, чтобы для принятой точности вторые разности $\Delta_s^2 (K_{nl} \varphi_l)$ ($l=0, 1, \dots, n-2$) сохраняли постоянные значения. Чтобы начать вычисления, необходимо предварительно найти значения $\varphi_0 = f_0 = f(x_0)$ и φ_1 .

§ 3. Уравнения Фредгольма второго рода. Метод квадратур

Задача численного решения уравнений Фредгольма является более сложной, чем для уравнений Вольтерра, и имеет значительно более развитую теорию. В книге будут рассмотрены только два метода: метод квадратур и метод замены ядра уравнения на вырожденное ядро.

1. Метод квадратур. Пусть дано интегральное уравнение второго рода вида Фредгольма

$$\varphi(x) = f(x) + \lambda \int_a^b K(x, s) \varphi(s) ds. \quad (1)$$

На отрезке интегрирования $[a, b]$ возьмем произвольно n точек $a \leq x_1 < x_2 < \dots < x_n \leq b$ и для нахождения приближенных значений функции φ в них построим систему уравнений. Такую систему можно получить, если к вычислению интеграла, входящего в уравнение, применить какое-либо правило приближенного интегрирования. Интеграл зависит от параметра x , поэтому коэффициенты правила и его остаточный член будут, вообще говоря, некоторыми функциями от x :

$$\int_a^b K(x, s) \varphi(s) ds = \sum_{k=1}^n C_k(x) \varphi(x_k) + \rho(x). \quad (2)$$

Подставив это выражение интеграла в уравнение (1) и положив там последовательно

$$x = x_1, x_2, \dots, x_n,$$

получим для точных значений $\varphi(x)$ систему уравнений

$$\varphi(x_i) = f_i + \lambda \sum_{k=1}^n C_k(x_i) \varphi(x_k) + \lambda \rho(x_i). \quad (3)$$

Остаток $\rho(x)$ в квадратурном правиле (2) по сравнению с квадратурной суммой имеет обычно малую величину. Поэтому, если в равенствах (3) пренебречь малыми величинами $\lambda \rho(x_i)$, можно рассчитывать на то, что получится алгебраическая система уравнений, решение которой φ_i ($i = 1, \dots, n$) будет близким к значениям $\varphi(x_i)$ точного решения интегрального уравнения

$$\varphi_i - \lambda \sum_{k=1}^n C_k(x_i) \varphi_k = f_i \quad (i = 1, 2, \dots, n). \quad (4)$$

Определитель системы

$$\Delta(\lambda) = D(E - \lambda C), \quad (5)$$

где

$$C = \begin{bmatrix} C_1(x_1) & \dots & C_n(x_1) \\ \dots & \dots & \dots \\ C_1(x_n) & \dots & C_n(x_n) \end{bmatrix}$$

есть многочлен от λ степени n .

Предположим, что λ не есть собственное значение уравнения (1) и уравнение, следовательно, имеет только одно решение. Тогда можно ожидать, что определитель $\Delta(\lambda)$ системы будет отличен от нуля и система (4) также имеет единственное решение. Но это надлежит проверять в каждой частной задаче.

Будем считать, что $\Delta(\lambda) \neq 0$. При численном решении системы каждое уравнение удовлетворяется с некоторым округлением чисел, и если погрешности округлений обозначить через $-\delta_i$, то для φ_i будут выполняться равенства

$$\varphi_i - \lambda \sum_{k=1}^n C_i(x_k) \varphi_k = f_i - \delta_i \quad (6)$$

и, следовательно,

$$\varphi_i = \frac{1}{\Delta(\lambda)} \sum_{k=1}^n \Delta_{ki} [f_k - \delta_k], \quad (7)$$

где Δ_{ki} есть алгебраическое дополнение элемента определителя $\Delta(\lambda)$ с индексами k, i .

Теперь рассмотрим погрешность приближенного решения $\varepsilon_i = \varphi(x_i) - \varphi_i$ и получим ее оценку. Точные значения $\varphi(x_i)$ решения удовлетворяют уравнениям (3), которые отличаются от уравнений (4) для φ_i только заменой $-\delta_i$ на $\lambda \rho(x_i)$, и поэтому для $\varphi(x_i)$ верны выражения вида (7) с такой же заменой:

$$\varphi(x_i) = \frac{1}{\Delta(\lambda)} \sum_{k=1}^n \Delta_{ki} [f_k + \lambda \rho(x_k)]. \quad (8)$$

Если теперь отсюда вычесть почленно (7), получится представление для ε_i :

$$\varepsilon_i = \frac{1}{\Delta(\lambda)} \sum_{k=1}^n \Delta_{ki} [\lambda \rho(x_k) + \delta_k]. \quad (9)$$

Оно дает возможность получить оценку ε_i .

Пусть для погрешности $\rho(x)$ квадратурного правила (8) верно неравенство $|\rho(x)| \leq \rho = \rho(n)$. В качестве $\rho(n)$ может быть взята величина $\rho(n) = \max_x |\rho(x)|$. Предположим также, что для погрешности δ_k , вызванной округлениями, выполняется неравенство $|\delta_k| \leq \delta = \delta(n)$ ($k = 1, \dots, n$). Величина $\delta(n)$ определяется принятым вычислительным процессом.

Из представления (9) вытекает следующая оценка для ϵ_i :

$$|\epsilon_i| \leq B[\lambda|\rho + \delta|], \quad (10)$$

где

$$B \geq \frac{1}{|\Delta(\lambda)|} \sum_{k=1}^n |\Delta_{ki}|.$$

Отметим попутно, что значения $\Delta(\lambda)$ и Δ_{ki} находятся в процессе вычислений и поэтому B является вычислимой величиной. В оценке (10) остается еще найти величину ρ . С этой целью возвратимся к остаточному члену $\rho(x)$ квадратурного правила (2) при произвольном значении x :

$$\rho(x) = \int_a^b K(x, s) \varphi(s) ds - \sum_{k=1}^n C_k(x) \varphi(x_k).$$

Внесем в него вместо $\varphi(x)$ выражение, стоящее в правой части уравнения (1):

$$\begin{aligned} \rho(x) &= \int_a^b K(x, s) \left[f(s) + \lambda \int_a^b K(s, u) \varphi(u) du \right] ds - \\ &- \sum_{k=1}^n C_k(x) \left[f(x_k) + \lambda \int_a^b K(x_k, u) \varphi(u) du \right] = \\ &= \int_a^b K(x, s) f(s) ds - \sum_{k=1}^n C_k(x) f(x_k) + \\ &+ \lambda \left[\int_a^b \int_a^b K(x, s) K(s, u) \varphi(u) du ds - \right. \\ &- \left. \sum_{k=1}^n C_k(x) \int_a^b K(x_k, u) \varphi(u) du \right] = \\ &= \rho_f(x) + \lambda \int_a^b \rho_K(x, u) \varphi(u) du, \quad (11) \end{aligned}$$

где

$$\rho_f(x) = \int_a^b K(x, s) f(s) ds - \sum_{k=1}^n C_k(x) f(x_k),$$

$$\rho_K(x, u) = \int_a^b K(x, s) K(s, u) ds - \sum_{k=1}^n C_k(x) K(x_k, u).$$

Очевидно, $\rho_f(x)$ есть остаток приближенного интегрирования по правилу (2) свободного члена f заданного уравнения, а $\rho_K(x, u)$ — остаток интегрирования по тому же правилу функции $K(s, u)$. Обе эти функции известны и поэтому две следующие величины: $\rho_f = \max_x |\rho_f(x)|$ и

$\rho_K = \max_x \int_a^b |\rho_K(x, u)| du$ мы вправе считать также известными.

Из представления (11) получается следующая оценка для ρ :

$$|\rho(x_i)| \leq \rho \leq \max_x |\rho(x)| \leq \rho_f + |\lambda| \rho_K H, \quad (12)$$

$$H = \max_x |\varphi(x)|.$$

Полученное неравенство совместно с соотношением (10) приводит к выписанной ниже оценке погрешности ε_i :

$$|\varepsilon_i| \leq B[|\lambda|(\rho_f + |\lambda| \rho_K H) + \delta] \quad (i = 1, \dots, n). \quad (13)$$

Эта оценка во многих случаях позволяет сделать заключение о сходимости вычислительного процесса и о порядке малости погрешности ε_i . Для этой цели ниже мы воспользуемся ею, а сейчас отметим недостаток, не позволяющий без оговорок применять ее в численных расчетах: она содержит число H , являющееся наибольшим абсолютным значением неизвестной функции *). Этот недостаток можно устранить и оценить H через вычислимые величины (см. (16)).

*) Для нестрогих ориентировочных расчетов число H можно часто заменить на $\max_i |\varphi_i|$, основываясь на том, что если сетка точек x_i достаточно густо заполняет отрезок $[a, b]$ и на сетке значения $\varphi(x_i)$ и φ_i ($i = 1, \dots, n$) близки между собой, то $\max_x |\varphi(x)|$ и $\max_i |\varphi_i|$ должны мало отличаться друг от друга.

Для этого потребуется определить приближенное решение не только в точках x_i , но и на всем отрезке $[a, b]$. Такую интерполяцию приближенного решения с сетки на отрезок можно выполнить многими способами, например, при помощи алгебраического интерполирования или какого-нибудь другого.

Мы воспользуемся «естественным» в рассматриваемой задаче интерполированием через квадратурное правило. Возвратимся к уравнению (1) и вычислим в нем интеграл при помощи квадратурного правила (2):

$$\varphi(x) = f(x) + \lambda \sum_{k=1}^n C_k(x) \varphi(x_k) + \lambda \rho(x). \quad (14)$$

Заменим в правой части точные значения $\varphi(x_k)$ функции φ в узлах x_k на вычисленные приближенные значения φ_k и отбросим «малый» остаток $\lambda \rho(x)$:

$$\tilde{\varphi}(x) = f(x) + \lambda \sum_{k=1}^n C_k(x) \varphi_k. \quad (15)$$

Полученную так функцию $\tilde{\varphi}(x)$ примем за приближенное решение на всем отрезке $[a, b]$. Ввиду уравнений (4) ясно, что $\tilde{\varphi}(x)$ в узлах x_i принимает значения φ_i соответственно. Поэтому $\tilde{\varphi}(x)$ выполняет интерполирование в обычном смысле сеточной функции φ_i ($i = 1, \dots, n$) на весь отрезок $[a, b]$. В основу интерполирования здесь положены функции $C_k(x)$. Так как они определяются ядром $K(x, s)$ интегрального уравнения, то (15) есть интерполирование φ_i через ядро $K(x, s)$, что в нашей задаче является естественным.

Получаемая погрешность имеет значение

$$\varepsilon(x) = \varphi(x) - \tilde{\varphi}(x) = \lambda \left[\sum_{k=1}^n C_k(x) \varepsilon_k + \rho(x) \right].$$

Отсюда, если ввести обозначение $\max_x \sum_{k=1}^n |C_k(x)| = C$, то для $\varepsilon(x)$ с помощью оценок (12) и (13) получится неравенство

$$|\varepsilon(x)| \leq |\lambda| \{B[|\lambda|(\rho_f + |\lambda| \rho_K H) + \delta]C + (\rho_f + |\lambda| \rho_K H)\}.$$

Введем, наконец, величину $\tilde{H} = \max_x |\tilde{\varphi}(x)|$. Она может быть вычислена, так как $\tilde{\varphi}(x)$ выражается только через величины, находящиеся в процессе вычислений.

Из равенства $\varphi(x) = \tilde{\varphi}(x) + \varepsilon(x)$ следует $H \leq \tilde{H} + \max_x |\varepsilon(x)|$.

Если воспользоваться полученной выше оценкой для $|\varepsilon(x)|$, то будем иметь неравенство

$$H \leq \tilde{H} + |\lambda| \{BC [|\lambda|(\rho_f + |\lambda|\rho_K H) + \delta] + (\rho_f + |\lambda|\rho_K H)\}.$$

Отсюда при условии, что $|\lambda|^2 \rho_K (1 + |\lambda| BC) < 1$ вытекает неравенство для H :

$$H \leq \frac{\tilde{H} + |\lambda| [\rho_f (1 + |\lambda| BC) + \delta BC]}{1 - |\lambda|^2 \rho_K (1 + |\lambda| BC)}. \quad (16)$$

Оно дополняет оценку (13) погрешности ε_i для тех случаев, когда необходимо знать границу для численного значения величины H .

В предыдущем изложении узлы x_k считались произвольными. Сделаем сейчас несколько замечаний об их выборе и соответствующих им квадратурных правилах. Когда необходимо составить таблицу значений решения φ , то наибольший интерес представляет случай равноотстоящих значений аргумента $s_k = a + hk$ ($k = 0, 1, \dots, n$), $h = \frac{1}{n}(b - a)$. Тогда для вычисления интеграла могут быть применены, например, правила трапеций, парабол, трех восьмых и т. д.

В случае трапеций формула (2) будет иметь вид

$$\begin{aligned} \int_a^b K(x, s) \varphi(s) ds &= h \left[\frac{1}{2} K(x, s_0) \varphi(s_0) + K(x, s_1) \varphi(s_1) + \dots \right. \\ &\quad \left. \dots + K(x, s_{n-1}) \varphi(s_{n-1}) + \frac{1}{2} K(x, b) \varphi(b) \right] - \\ &\quad - \frac{(b-a)^3}{12n^2} \left[\frac{\partial^2}{\partial s^2} (K\varphi) \right]_{s=\xi}, \\ a &\leq \xi \leq b, \end{aligned}$$

В случае парабол

$$\int_a^b K(x, s) \varphi(s) ds = h \left\{ \frac{1}{3} \left[K(x, s_0) \varphi(s_0) + K(x, s_n) \varphi(s_n) \right] + \right. \\ \left. + \frac{4}{3} \sum_m K(x, s_{2m+1}) \varphi(s_{2m+1}) + \frac{2}{3} \sum_m K(x, s_{2m}) \varphi(s_{2m}) \right\} - \\ - \frac{1}{90} \frac{(b-a)^5}{n^4} \left[\frac{\partial^5}{\partial s^5} (K\varphi) \right]_{s=\xi}, \quad a \leq \xi \leq b$$

и т. д. Правила приближенного интегрирования этого вида имеют невысокую степень точности и для получения малой погрешности часто бывает необходимо брать малое значение h и в соответствии с этим большое число узлов x_k . Тогда система вида (4) будет иметь большое число уравнений и ее решение потребует большое количество вычислений. Ввиду этого часто предпочитают применять квадратурные правила, имеющие высокую степень точности, например правило Гаусса или его аналоги.

2. Сходимость метода квадратур. Вычислительный процесс метода квадратур определяется двумя таблицами: бесконечной таблицей узлов квадратурных правил

$$X = \begin{bmatrix} x_1^1 & & & \\ x_1^2, & x_2^2 & & \\ \dots & \dots & \dots & \\ x_1^n, & x_2^n, & \dots, & x_n^n \\ \dots & \dots & \dots & \dots \end{bmatrix} \quad (17)$$

и таблицей коэффициентов $C_k^i(x)$ этих правил

$$A = \begin{bmatrix} C_1^1(x) & & & \\ C_1^2(x) & C_2^2(x) & & \\ \dots & \dots & \dots & \\ C_1^n(x) & C_2^n(x) & \dots & C_n^n(x) \\ \dots & \dots & \dots & \dots \end{bmatrix} \quad (18)$$

Таблицы составлены так, что в строке номера n помещены узлы и коэффициенты квадратурного правила вида (2), которое применяется для приближенного вычисления интеграла на шаге номера n процесса:

$$\int_a^b K(x, s) \varphi(s) ds = \sum_{k=1}^n C_k^n(x) \varphi(x_k^n) + \rho_\varphi^n(x). \quad (19)$$

В вычислительном процессе все величины, зависящие от C_k^n и x_k^n , будут некоторыми функциями от n , например остаточный член $\rho^n(x)$ квадратурного правила будет зависеть как от x , так и от n . В оценке (13) от номера n шага будут зависеть как сама погрешность ϵ_i , так и величины B , ρ_f , ρ_K . От n будет зависеть и погрешность δ , так как, если мы хотим сделать процесс сходящимся к точному решению, то с ростом n должны неограниченно увеличивать точность решения системы вида (4). Указанная зависимость от n отмечена при помощи приписки величине верхнего индекса n .

Для процесса оценка вида (13) запишется в следующей форме:

$$|\epsilon_i^n| \leq B^n [|\lambda|(\rho_f^n + |\lambda| \rho_K^n H) + \delta^n] \quad (i = 1, \dots, n). \quad (20)$$

Такая оценка погрешности позволяет считать доказанной приводимую ниже теорему о сходимости вычислительного процесса метода квадратур.

Теорема 1. *Предположим, что λ не есть собственное значение ядра $K(x, s)$ и уравнение (1) имеет, следовательно, единственное решение.*

Пусть для вычислительного процесса метода квадратур, определяемого равенством (19), выполняются условия:

1) погрешности $\rho_{f(s)}^n(x)$ и $\rho_{K(x,s)}^n(x)$ квадратурного правила (19) для функций $\varphi(s) = f(s)$ и $\varphi(s) = K(s, u)$ таковы, что при неограниченном росте n будет соответственно

$$B^n \rho_{f(s)}^n(x) \rightarrow 0 \quad (21)$$

равномерно относительно x и

$$B^n \rho_{K(s,u)}^n(x) \rightarrow 0 \quad (22)$$

равномерно относительно x и u ;

2) погрешности — δ_i^n решения уравнений систем вида (4) уменьшаются настолько быстро при увеличении n , что имеет место соотношение

$$B^n \delta^n \rightarrow 0, \quad \delta^n = \max_i |\delta_i^n|. \quad (23)$$

Тогда для всякого положительного ϵ существует $N(\epsilon)$ такое, что при $n > N(\epsilon)$ будут выполняться неравенства

$$|\varphi(x_i^n) - \varphi_i^n| < \epsilon \quad (i = 1, 2, \dots, n). \quad (24)$$

Приведенная теорема имеет ту особенность, что ее условия (21) и (22) должны проверяться в процессе вычислений, а условие (23) должно учитываться при определении точности решения системы уравнений для нахождения φ_i^n ($i = 1, \dots, n$).

3. Интерполяционный квадратурный метод. Этот метод является частным случаем общего квадратурного метода и основан на интерполировании неизвестной функции φ . Для определенности изложения будем иметь в виду алгебраическое интерполирование.

Выберем на отрезке $[a, b]$ n различных точек $a \leq x_1 < x_2 < \dots < x_n \leq b$ и выполним интерполирование функции φ по ее значениям в этих точках. Воспользуемся лагранжевым представлением интерполяционного многочлена

$$\begin{aligned} \varphi(x) &= \sum_{k=1}^n \frac{\omega(x)}{(x-x_k)\omega'(x_k)} \varphi(x_k) + r(x) = \\ &= \sum_{k=1}^n l_k(x) \varphi(x_k) + r(x). \end{aligned} \quad (25)$$

Внесем это представление φ в интегральный член уравнения (1):

$$\begin{aligned} \varphi(x) &= f(x) + \lambda \sum_{k=1}^n L_k(x) \varphi(x_k) + \lambda \rho(x), \\ L_k(x) &= \int_a^b K(x, s) l_k(s) ds, \quad \rho(x) = \int_a^b K(x, s) r(s) ds. \end{aligned} \quad (26)$$

Чтобы получить систему уравнений для $\varphi(x_n)$, отбросим в (26) остаточный член $\lambda \rho(x)$ и в полученном приближенном равенстве положим x равным $x_1, x_2, \dots, x_n$:

$$\varphi_i = f_i + \lambda \sum_{k=1}^n L_k(x_i) \varphi_k \quad (i = 1, \dots, n). \quad (27)$$

Эта система является частным случаем более общей системы (4), когда $C_k(x_i) = L_k(x_i)$.

После решения системы (27) мы будем знать приближенные значения φ_i решения φ в точках x_i ($i = 1, \dots, n$).

Чтобы найти приближенные значения φ всюду на $[a, b]$, можно выполнить алгебраическое интерполирование φ по значениям φ_i , как было сделано выше для получения уравнений (27). Но можно ожидать, что лучшая точность будет достигнута, если воспользоваться «естественным» интерполированием φ через ядро $K(x, s)$. Оно получится, если в равенстве (26) отбросить остаточный член $\lambda r(x)$:

$$\tilde{\varphi}(x) = f(x) + \lambda \sum_{k=1}^n L_k(x) \varphi(x_k). \quad (28)$$

Остановимся еще на вопросе выбора узлов x_k в интерполяционном методе. Если точное решение $\varphi(x)$ является многочленом от x степени не выше $n-1$, то интерполирование $\varphi(x)$ будет точным и остаточный член $r(x)$ в (25) — тождественно равным нулю. В соответствии с этим остаточный член $\lambda r(x)$ в представлении (26) также будет тождественным нулем и системе (27) будут удовлетворять точные значения решения $\varphi_i = \varphi(x_i)$. Эти значения могут быть найдены из системы и по ним построено точное решение $\varphi(x)$.

Это верно при всяких узлах x_k ($k=1, \dots, n$), лишь бы они были различными. Поэтому говорят, что интерполяционный метод при любых узлах x_k имеет степень точности не меньше $n-1$. Естественно попытаться выяснить, можно ли путем выбора узлов сделать интерполяционный метод точным для любых многочленов $\varphi(s)$ степени n или, может быть, для всяких многочленов некоторой степени $m > n$.

Остановимся на требовании точности метода для многочленов степени n . Это требование эквивалентно тому, чтобы выполнялись точно равенства

$$\int_a^b K(x_i, s) \varphi(s) ds = \sum_{k=1}^n L_k(x_i) \varphi(x_k) \quad (i=1, \dots, n) \quad (29)$$

для любых многочленов $\varphi(s)$ степени n .

Теорема 2. Для того чтобы равенства (29) были точными для всякого многочлена $\varphi(s)$ степени n , необходимо и достаточно выполнение условий:

1) коэффициенты $L_k(x_i)$ должны быть значениями при $x=x_i$ функции $L_k(x)$, указанной в равенствах (26);

2) многочлен $\omega(s) = (s - x_1) \dots (s - x_n)$ должен удовлетворять следующим условиям ортогональности:

$$\int_a^b K(x_i, s) \omega(s) ds = 0, \quad i = 1, 2, \dots, n. \quad (30)$$

Сделаем замечание по поводу этой теоремы. Равенства (30) дают систему n уравнений, из которой должны быть найдены узлы x_i ($i = 1, \dots, n$). Уравнения определяются ядром $K(x, s)$, и поэтому для каждого ядра узлы x_i будут, вообще говоря, своими. Последнее очень затрудняет практическое использование указанного способа повышения степени точности интерполяционного метода.

Доказательство. Проверим необходимость условий теоремы. Предположим, что равенство (29) верно для всякого многочлена $\varphi(s)$ степени n . Рассмотрим коэффициент Лагранжа $l_m(s)$, отвечающий узлу x_m :

$$l_m(s) = \frac{\omega(s)}{(s - x_m) \omega'(x_m)}.$$

Это есть многочлен степени $n - 1$, поэтому для него равенство (29) должно быть точным. С другой стороны, $l_m(x_i) = 0$ ($i \neq k$) и $l_m(x_m) = 1$ и в сумме соотношения (29) сохранится только слагаемое, отвечающее $k = m$. Тогда равенство (29) примет вид

$$\int_a^b K(x_i, s) l_m(s) ds = L_m(x_i),$$

что доказывает необходимость первого условия.

Положим теперь $\varphi(s) = \omega(s)$. Многочлен $\omega(s)$ имеет степень n , и для него (29) выполняется точно. Но так как $\omega(x_k) = 0$ ($k = 1, \dots, n$), то равенство (29) совпадает с (30). Этим доказывается необходимость второго условия теоремы.

Проверим достаточность условий. Пусть $\varphi(x)$ есть произвольный многочлен степени n . Разделим его на $\omega(s)$. В частном получится постоянная величина, которую обозначим через α , а в остатке — некоторый многочлен $Q(s)$ степени не больше $n - 1$:

$$\varphi(s) = \alpha \omega(s) + Q(s).$$

Ввиду соотношения $\omega(x_k) = 0$, очевидно, будет $Q(x_k) = \varphi(x_k)$ и

$$\int_a^b K(x_i, s) \varphi(s) ds = \alpha \int_a^b K(x_i, s) \omega(s) ds + \int_a^b K(x_i, s) Q(s) ds.$$

По второму условию теоремы первый из интегралов, стоящих справа, равен нулю. Так как $Q(s)$ имеет степень $\leq n-1$, то для него интерполяционное правило интегрирования

$$\int_a^b K(x, s) \varphi(s) ds = \sum_{k=1}^n L_k(x) \varphi(x_k) + \rho(x),$$

$$L_k(x) = \int_a^b K(x, s) l_k(s) ds$$

выполняется при всяких x точно (без остатка $\rho(x)$); в частности, при $x = x_i$ имеем

$$\int_a^b K(x_i, s) Q(s) ds = \sum_{k=1}^n L_k(x_i) Q(x_k) = \sum_{k=1}^n L_k(x_i) \varphi(x_k).$$

Поэтому

$$\int_a^b K(x_i, s) \varphi(s) ds = \sum_{k=1}^n L_k(x_i) \varphi(x_k)$$

и равенства (29) действительно точны для всякого многочлена $\varphi(s)$ степени n .

§ 4. Метод замены ядра уравнения на вырожденное ядро

1. Некоторые сведения об уравнениях с вырожденным ядром. Ядро уравнения $K(x, s)$ называют вырожденным, если оно представимо в виде

$$K(x, s) = \sum_{i=1}^n \alpha_i(x) \beta_i(s). \quad (1)$$

Функции α_i ($i = 1, \dots, n$) можно, очевидно, считать линейно независимыми. Аналогичное можно сказать о функциях β_i ($i = 1, \dots, n$).

Когда в интегральном уравнении (1) ядро $K(x, s)$ имеет форму (1), то его решение, если оно существует, представимо в следующем виде:

$$\varphi(x) = f(x) + \sum_{j=1}^n A_j \alpha_j(x), \quad (2)$$

где коэффициенты A_j есть некоторые неизвестные постоянные

$$A_j = \lambda \int_a^b \beta_j(s) \varphi(s) ds.$$

Для нахождения этих коэффициентов может быть просто построена линейная система уравнений. Введем следующие обозначения:

$$\beta_{ij} = \int_a^b \beta_i(s) \alpha_j(s) ds, \quad f_i = \int_a^b f(s) \beta_i(s) ds.$$

Подставим в уравнение (3.1) вместо φ представление (2):

$$\begin{aligned} f(x) + \sum_{i=1}^n A_i \alpha_i(x) &= \\ &= f(x) + \lambda \int_a^b \sum_{i=1}^n \alpha_i(x) \beta_i(s) \left[f(s) + \sum_{j=1}^n A_j \alpha_j(s) \right] ds. \end{aligned}$$

Ввиду линейной независимости функций $\alpha_i(x)$ коэффициенты при них в левой и правой частях равенства должны быть одинаковыми и после сравнения их мы получим систему уравнений для A_i :

$$A_i = \lambda f_i + \lambda \sum_{j=1}^n \beta_{ij} A_j \quad (i = 1, \dots, n). \quad (3)$$

Если ввести векторы $A = (A_1, \dots, A_n)$, $f = (f_1, \dots, f_n)$ и матрицу

$$B = \begin{bmatrix} \beta_{11} & \dots & \beta_{1n} \\ \vdots & \dots & \vdots \\ \beta_{n1} & \dots & \beta_{nn} \end{bmatrix},$$

то систему (3) можно записать в виде

$$A = \lambda f + \lambda B A. \quad (4)$$

Будем считать, что определитель этой системы отличен от нуля:

$$\Delta\lambda = D(E - \lambda B) = \begin{vmatrix} 1 - \lambda\beta_{11} & -\lambda\beta_{12} & \dots & -\lambda\beta_{1n} \\ -\lambda\beta_{21} & 1 - \lambda\beta_{22} & \dots & -\lambda\beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda\beta_{n1} & -\lambda\beta_{n2} & \dots & 1 - \lambda\beta_{nn} \end{vmatrix} \neq 0.$$

Решением системы (4) является следующий вектор:

$$A = (E - \lambda B)^{-1} f. \quad (5)$$

Чтобы дать явное выражение для его составляющих, обозначим $\Delta_{ij}(\lambda)$ алгебраическое дополнение элемента определителя $\Delta(\lambda)$, стоящего в строке номера i и столбце номера j . Тогда составляющая A_i вектора A может быть записана в виде

$$A_i = \lambda \frac{1}{\Delta(\lambda)} \sum_{j=1}^n \Delta_{ji} f_j.$$

Внося эти значения коэффициентов A_i в равенство (2), получим представление решения через известные величины

$$\begin{aligned} \varphi(x) &= f(x) + \lambda \frac{1}{\Delta(\lambda)} \sum_{i=1}^n \sum_{j=1}^n \Delta_{ji} \alpha_i(x) \int_a^b f(s) \beta_j(s) ds = \\ &= f(x) + \lambda \int_a^b \frac{\Delta(x, s; \lambda)}{\Delta(\lambda)} f(s) ds. \quad (6) \end{aligned}$$

Резольвента интегрального уравнения здесь имеет простое явное выражение

$$\Gamma(x, s; \lambda) = \frac{1}{\Delta(\lambda)} \Delta(x, s; \lambda) = \frac{1}{\Delta(\lambda)} \sum_{i=1}^n \sum_{j=1}^n \Delta_{ji} \alpha_i(x) \beta_j(s).$$

Построение решения $\varphi(x)$ приводится в случае вырожденного ядра $K(x, s)$ к вычислению величин f_i , β_{ij} и к решению линейной системы (3).

2. Замечания о методах построения вырожденного ядра, близкого к ядру уравнения. Метод приближенного решения, который рассматривается в настоящем параграфе, основан на простом соображении: если в заданном

интегральном уравнении ядро заменить на близкое к нему вырожденное ядро, то решение нового вспомогательного уравнения, как следует ожидать, будет близким к решению заданного уравнения. Построение же решения вспомогательного уравнения будет намного проще ввиду вырожденности ядра. В следующем пункте будет доказана теорема, подтверждающая такое ожидание: решение вспомогательного уравнения может быть сделано сколь угодно близким к решению заданного уравнения, если, во-первых, число λ не есть собственное значение заданного уравнения и, во-вторых, вырожденное ядро вспомогательного уравнения взято достаточно близким к заданному ядру. Сейчас мы укажем некоторые приемы построения вырожденных ядер.

Применение степенного ряда. Этот метод, наверно, применим в случае, когда ядро $K(x, s)$ есть аналитическая функция от x или s , регулярная в достаточно широкой окрестности отрезка $[a, b]$. Пусть $K(x, s)$ есть аналитическая функция от s в области $|s - c| < R$, где c есть середина отрезка $[a, b]$ и $R > \frac{1}{2}(b - a)$. В этом случае $K(x, s)$ можно разложить в степенной ряд по степеням $s - c$, сходящийся в круге $|s - c| < R$; коэффициенты этого ряда будут зависеть от x :

$$K(x, s) = C_0(x) + C_1(x)(s - c) + \dots + C_n(x)(s - c)^n + \dots,$$

$$C_n(x) = \frac{1}{n!} \frac{\partial^n}{\partial s^n} K(x, c).$$

В качестве вырожденного ядра $\tilde{K}(x, s)$ здесь может быть взят отрезок степенного ряда

$$\tilde{K}(x, s) = C_0(x) + C_1(x)(s - c) + \dots + C_n(x)(s - c)^n.$$

Аналогично можно поступить и в тех случаях, когда ядро есть аналитическая функция от x или от обоих аргументов.

Использование ортогональных разложений. Пояснение этого способа сделаем на примере ряда Фурье. Для упрощения записи будем считать, что ядро $K(x, s)$ задано в квадрате $0 \leq x, s \leq 1$. Предположим, что $K(x, s)$ может быть разложено по переменной s

в равномерно сходящийся относительно x и s ряд Фурье, коэффициенты которого будут зависеть от x :

$$K(x, s) = a_0(x) + \sum_{k=1}^{\infty} [a_k(x) \cos 2\pi kx + b_k(x) \sin 2\pi kx],$$

$$a_0(x) = \int_0^1 K(x, s) ds, \quad a_k = 2 \int_0^1 K(x, s) \cos 2\pi ks ds,$$

$$b_k = 2 \int_0^1 K(x, s) \sin 2\pi ks ds.$$

В качестве вырожденного ядра $\tilde{K}(x, s)$ может быть взят отрезок этого ряда. Очевидно, подобным же образом можно использовать ряд Фурье по аргументу x или кратный ряд Фурье по обоим аргументам x и s .

Тригонометрические ряды были взяты лишь для определенности записи и, очевидно, можно воспользоваться любыми другими ортогональными разложениями.

Интерполяционные методы. Выберем на отрезке $[a, b]$ n узлов $s_1, s_2, \dots, s_n$ и проинтерполируем по аргументу s функцию $K(x, s)$ по ее значениям во взятых точках. При этом будет получено приводимое ниже приближенное представление $K(x, s)$, которое может быть принято за вырожденное ядро:

$$\tilde{K}(x, s) = \sum_{k=1}^n \frac{\omega(x)}{(s - s_k) \omega'(s_k)} K(x, s_k),$$

$$\omega(s) = (s - s_1) \dots (s - s_n).$$

3. Оценка близости между решениями уравнений в зависимости от близости самих уравнений. Пусть задано интегральное уравнение

$$\varphi(x) = f(x) + \lambda \int_a^b K(x, s) \varphi(s) ds \quad (7)$$

и одновременно с ним рассматривается измененное уравнение

$$\tilde{\varphi}(x) = \tilde{f}(x) + \lambda \int_a^b \tilde{K}(x, s) \tilde{\varphi}(s) ds. \quad (8)$$

В рассматриваемом сейчас методе роль измененного уравнения играет уравнение с вырожденным ядром.

Нас будут интересовать два следующих основных вопроса:

1) если известна разрешимость измененного уравнения, то при каких условиях близости можно быть уверенным в разрешимости заданного уравнения и как может быть оценена разность между решениями;

2) что можно сказать о сходных вопросах для измененного уравнения по известной разрешимости заданного уравнения.

Так как уравнения (7) и (8) в этих вопросах играют сходные роли, то достаточно получить ответы на первый из вопросов.

Теорема 1. *Предположим, что для уравнений (7) и (8) выполняются условия:*

1) для ядер $K(x, s)$ и $\tilde{K}(x, s)$ при всяких x из $[a, b]$ выполняется неравенство

$$\int_a^b |K(x, s) - \tilde{K}(x, s)| ds \leq h; \quad (9)$$

2) при всяких x и s из отрезка $[a, b]$ для резольвенты $\tilde{\Gamma}(x, s; \lambda)$ уравнения (8) верна оценка

$$|\tilde{\Gamma}(x, s; \lambda)| \leq B; \quad (10)$$

3) для свободных членов f и $\tilde{f}$ уравнений верно неравенство

$$|f(x) - \tilde{f}(x)| \leq \eta, \quad x \in [a, b]; \quad (11)$$

4) числа λ , h , η , B удовлетворяют условию

$$1 - |\lambda| h (1 + |\lambda| B) > 0. \quad (12)$$

Тогда можно утверждать, что уравнение (7) имеет единственное решение $\varphi(x)$ и для решений уравнений (7) и (8) верна оценка

$$|\varepsilon(x)| = |\varphi(x) - \tilde{\varphi}(x)| \leq (\eta + |\lambda| h M) (1 + |\lambda| B), \quad (13)$$

где $M = \max_x |\varphi(x)|$, при этом для M выполняется неравенство

$$M \leq \frac{\tilde{M} + \eta(1 + |\lambda| B)}{1 - |\lambda| h (1 + |\lambda| B)}, \quad \tilde{M} = \max_x |\tilde{\varphi}(x)|. \quad (14)$$

Доказательство. По условию (10) резольвента уравнения (8) ограничена, поэтому λ не является собственным значением уравнения и уравнение имеет единственное решение. Предположим, что уравнение (7) имеет решение $\varphi(x)$. Ниже увидим, что в условиях теоремы это предположение будет оправдано. Рассмотрим разность $\varepsilon(x) = \varphi(x) - \tilde{\varphi}(x)$ между решениями и построим для нее уравнение. С этой целью из (7) вычтем почленно (8)

$$\begin{aligned}\varepsilon(x) &= \varphi(x) - \tilde{\varphi}(x) = \\ &= f(x) - \tilde{f}(x) + \lambda \int_a^b K(x, s) \varphi(s) ds - \lambda \int_a^b \tilde{K}(x, s) \tilde{\varphi}(s) ds = \\ &= f(x) - \tilde{f}(x) + \lambda \int_a^b [K(x, s) - \tilde{K}(x, s)] \varphi(s) ds + \\ &\quad + \lambda \int_a^b \tilde{K}(x, s) [\varphi(s) - \tilde{\varphi}(s)] ds = \\ &= f^*(x) + \lambda \int_a^b \tilde{K}(x, s) [\varphi(s) - \tilde{\varphi}(s)] ds.\end{aligned}$$

Полученное равенство является интегральным уравнением для $\varepsilon(x) = \varphi(x) - \tilde{\varphi}(x)$ со свободным членом $f^*(x)$ и ядром $\tilde{K}(x, s)$. Из него с помощью резольвенты $\tilde{\Gamma}$ получается явное выражение $\varphi - \tilde{\varphi}$:

$$\varphi(x) - \tilde{\varphi}(x) = f^*(x) + \lambda \int_a^b \tilde{\Gamma}(x, s; \lambda) f^*(s) ds. \quad (15)$$

Ввиду неравенства

$$\begin{aligned}|f^*(x)| &\leq |f(x) - \tilde{f}(x)| + \\ &\quad + |\lambda| \cdot \left| \int_a^b [K(x, s) - \tilde{K}(x, s)] \varphi(s) ds \right| \leq \eta + |\lambda| hM\end{aligned}$$

из (14) вытекает следующая оценка:

$$\begin{aligned}|\varphi(x) - \tilde{\varphi}(x)| &\leq \eta + |\lambda| hM + |\lambda| B(\eta + |\lambda| hM) = \\ &= (\eta + |\lambda| hM)(1 + |\lambda| B).\end{aligned} \quad (16)$$

В правую часть неравенства входит неизвестная величина M . Если K и $\tilde{K}$, так же как f и $\tilde{f}$, «близки» между собой, то, как правило, φ и $\tilde{\varphi}$ будут также мало отли-

чатся и при нестрогих оценках M в (16) часто бывает можно заменить на $\tilde{M}$ и получить нестрогую оценку $\varphi - \tilde{\varphi}$, имеющую только ориентировочное значение и применимую не во всех случаях. Однако в условиях теоремы для M можно получить строгую оценку. В самом деле,

$$|\varphi(x)| = |\tilde{\varphi} + (\varphi - \tilde{\varphi})| \leq \tilde{M} + (\eta + |\lambda|hM)(1 + |\lambda|B),$$

и так как это неравенство верно при всяких x , то

$$M \leq \tilde{M} + (\eta + |\lambda|hM)(1 + |\lambda|B).$$

Отсюда и на основании условия (11) следует

$$M \leq \frac{\tilde{M} + \eta(1 + |\lambda|B)}{1 - |\lambda|h(1 + |\lambda|B)}.$$

Этот результат показывает, что все решения уравнения (7) в условиях теоремы ограничены одним числом. Но тогда λ не может быть собственным значением ядра $K(x, s)$ и уравнение (7) имеет решение и только одно.

ЛИТЕРАТУРА

1. Березин И. С., Жидков Н. П., Методы вычислений, т. 2, Физматгиз, М., 1959, гл. 10, § 10.
2. Канторович Л. В., Крылов В. И., Приближенные методы высшего анализа, Физматгиз, М.—Л., 1962, гл. 2.
3. Крылов В. И., Бобков В. В., Монастырный П. И., Вычислительные методы высшей математики, т. 2, Минск, «Вышэйшая школа», 1975, гл. 8.

*) Если λ есть собственное значение ядра $K(x, s)$, то к решению неоднородного уравнения можно прибавить решение однородного уравнения, умноженное на любую постоянную, и решения неоднородного уравнения поэтому не могут быть ограничены в совокупности.

ГЛАВА 9

УЛУЧШЕНИЕ СХОДИМОСТИ РЯДОВ И ПОСЛЕДОВАТЕЛЬНОСТЕЙ

§ 1. Введение

1. О задаче улучшения сходимости. Ряды и последовательности очень часто и успешно применяются для решения большого числа практически важных задач и исследования многих вопросов теории. Но нередко оказывается, что получающиеся ряды или последовательности сходятся настолько медленно, что хотя они в принципе и решают стоящую перед исследователем задачу, но ими невозможно реально пользоваться, так как для нахождения суммы ряда с нужной точностью необходимо было бы взять чрезмерно много членов ряда или найти очень много членов последовательности, чтобы последний из найденных членов с хорошей точностью давал предел последовательности.

В этих случаях естественно поставить задачу об ускорении сходимости, т. е. о таком преобразовании заданного ряда в другой ряд, который имел бы такую же сумму, но сходил бы быстрее. Аналогично для последовательности: ее нужно преобразовать в новую последовательность, имеющую тот же предел, но сходящуюся к нему более быстро, чем заданная.

Ускорение сходимости есть только одна из задач, которые рассматриваются в проблеме улучшения сходимости. Иногда полезно рассмотреть иные задачи, например задачу преобразования произвольно сходящейся последовательности в монотонную или колеблющуюся около предела последовательность и т. п.

Ниже мы будем заниматься преимущественно задачей ускорения сходимости.

С целью сокращения изложения напомним, что сходимости ряда и последовательности равносильны между собой: сходимость ряда $u_1 + u_2 + \dots + u_n + \dots$ к сумме s , по определению, есть сходимость последовательности частичных сумм $s_1, s_2, \dots, s_n, \dots$ ($s_n = u_1 + \dots + u_n$) к пределу s ; наоборот, сходимость последовательности s_n ($n = 1, 2, \dots$) к s равносильна сходимости ряда $s_1 + (s_2 - s_1) + \dots + (s_n - s_{n-1}) + \dots$ к сумме s . Поэтому ясно, что каждый результат, установленный для последовательности, может быть с необходимыми изменениями преобразован в равносильный результат для ряда, и наоборот. Это позволит доказывать утверждения только для одного из объектов — ряда или последовательности, — избирая его, как будет удобнее или нагляднее.

Пусть дана сходящаяся последовательность s_n ($n = 1, 2, \dots; s_n \rightarrow s$). Рассмотрим погрешность приближения $r_n = s - s_n$. Ускорить сходимость s_n к s означает, что из r_n нужно выделить главную часть, т. е. представить r_n в виде суммы двух слагаемых $r_n = \omega_n^1 + r_n^1$, из которых r_n^1 стремится к нулю быстрее, чем ω_n^1 [$r_n^1 = o(\omega_n^1)$], и затем главную часть ω_n^1 прибавить к s_n :

$$s_n^1 = s_n + \omega_n^1, \quad s = s_n^1 + r_n^1.$$

Новая последовательность s_n^1 будет, очевидно, более быстро сходиться к s , чем s_n .

Если одного шага преобразования недостаточно для получения нужного ускорения сходимости, мы должны такие же преобразования выполнить с s_n^1 и r_n^1 : из r_n^1 выделить главную часть, представив r_n^1 в форме суммы $r_n^1 = \omega_n^2 + r_n^2$, где $r_n^2 = o(\omega_n^2)$ ($n \rightarrow \infty$), прибавить ω_n^2 к s_n^1 и т. д.

В общем случае для всякого k нужно разложить r_n на $k+1$ слагаемых так, чтобы выполнялись условия

$$r_n = \omega_n^1 + \omega_n^2 + \dots + \omega_n^k + r_n^k, \quad \omega_n^{s+1} = o(\omega_n^s) \quad (1)$$

$$(s = 1, \dots, k-1), \quad r_n^k = o(\omega_n^k).$$

Чтобы выполнить такое разложение, необходимо иметь некоторые достаточно полные сведения о законе, по которому r_n стремится к нулю при возрастании n .

Приведем из предыдущих глав примеры выделения из погрешности главной части и разложений вида (1).

I. В § 2 гл. 4 говорилось о том, что метод одношаговой итерации $x_{n+1} = \varphi(x_n)$ ($n = 0, 1, \dots$) для нахождения решения x^* уравнения $x = \varphi(x)$, наверное, сходится к x^* при выполнении двух условий: 1) функция φ имеет непрерывную производную $\varphi'(x)$ в окрестности x^* и притом $|\varphi'(x^*)| < 1$; 2) начальное приближение x_0 взято достаточно близким к решению x^* . Когда приближение x_n близко к x^* , то для его погрешности $\varepsilon_n = x_n - x^*$ с нужной точностью выполняется приближенное равенство $\varepsilon_{n+1} \approx \varphi'(x^*) \varepsilon_n$. Можно считать, что это равенство выполняется, начиная с $n = 0$. Из него следует явное выражение для ε_n :

$$\varepsilon_n \approx [\varphi'(x^*)]^n \varepsilon_0 = \varepsilon_0 q^n, \quad (2)$$

показывающее, что с ростом n погрешность ε_n убывает приблизительно по закону геометрической прогрессии. Приближенное же значение (2) для ε_n дает главную часть погрешности.

Эти факты были использованы в п. 4.3.3 для ускорения сходимости метода простой итерации и построения правила преобразования (4.3.9) такого, что для новой последовательности погрешность убывает по квадратичному закону вида (4.3.11).

II. Напомним, что погрешность квадратурного правила трапеций, если функция f имеет достаточно высокий порядок дифференцируемости, представима в виде (см. п. 5.8.2, (5.8.4))

$$\begin{aligned} R_n &= \int_a^b f(x) dx - T_n = \\ &= - \sum_{s=1}^k \frac{h^{2s}}{(2s)!} B_{2s} [f^{(2s-1)}(b) - f^{(2s-1)}(a)] + \rho_{2s+2}(f), \end{aligned} \quad (3)$$

где

$$\begin{aligned} T_n &= h \left[\frac{1}{2} f(a) + f(a+h) + \dots + \frac{1}{2} f(b) \right], \\ h &= \frac{1}{n} (b-a), \quad \rho_{2s+2} = O(h^{2s+2}). \end{aligned}$$

Когда n неограниченно возрастает, это равенство дает представление погрешности R_n в форме суммы бесконечно малых величин относительно $1/n$ возрастающих порядков

и имеет вид (1). Главным членом в R_n является величина $-\frac{h^2}{12}[f'(b) - f'(a)]$ и т. д.

В общем виде проблема улучшения сходимости последовательности имеет следующий смысл: по заданной сходящейся последовательности s_k ($k = 1, 2, \dots; s_k \rightarrow s$) нужно построить новую последовательность

$$\sigma_n = f_n(s_1, s_2, \dots) \quad (n = 1, 2, \dots), \quad (4)$$

удовлетворяющую условиям:

1) новая последовательность сходится к тому же пределу s , что и s_n ;

2) сходимость σ_n к s «лучше» в каком-то смысле сходимости s_n к s .

Критерий сравнения сходимостей должен быть при этом точно указан.

В преобразовании (4) можно воспользоваться выбором функций f_n , и если мы хотим добиться «улучшения» сходимости, например ее ускорения, то выбор f_n должен быть как-то согласован с законом сходимости $s_n \rightarrow s$. Обычно, чтобы избежать чрезмерной специализации правил для такого согласования, избирают некоторое множество S сходящихся последовательностей s_n , близких между собой по характеру стремления к пределу. В качестве таких множеств берут в первую очередь множества последовательностей с особенно часто встречающимися в приложениях видами сходимости. Примерами таких множеств могут служить последовательности s_n , сходящиеся по закону, близкому к тому, как стремятся к пределу рациональные функции, или же показательные функции $s + Ae^{-\alpha n}$ и их линейные комбинации и т. д. Преобразование (4) выбирают так, чтобы оно ускоряло сходимость не одной фиксированной последовательности s_n , а всех последовательностей взятого множества S . Некоторые способы такого выбора f_n будут рассмотрены ниже.

2. Связь с проблемой суммирования расходящихся рядов и определения пределов расходящихся последовательностей. Задача улучшения сходимости последовательностей и рядов в некоторых сторонах своих связана с другой классической задачей, имеющей почти столь же длительную историю. Ее содержание указано в назва-

нии пункта. Поясним смысл этой задачи двумя примерами.

1. Первый из них возьмем из области гармонического анализа. Пусть f есть 2π — периодическая функция, абсолютно интегрируемая *) на $(-\pi, \pi]$. Для нее могут быть вычислены коэффициенты Фурье

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx, \quad (5)$$

$$n = 0, 1, \dots$$

Рассмотрим обратную задачу, которую часто называют *гармоническим синтезом*: по известным коэффициентам Фурье a_n и b_n функции f нужно найти саму функцию f .

Напомним некоторые факты, связанные с такой задачей. Предположим, что мы изменили функцию f на множестве точек, имеющем меру нуль **), и получили новую функцию f_1 , которая пусть также является абсолютно интегрируемой. При таком изменении коэффициенты a_n и b_n , очевидно, сохраняют свои значения, так что функции f и f_1 будут иметь одинаковые коэффициенты Фурье.

Верно и обратное: если две абсолютно интегрируемые функции f и f_1 , имеют одинаковые коэффициенты Фурье, то они могут отличаться между собой только на множестве меры нуль.

Поэтому коэффициенты Фурье определяют функцию f лишь с точностью до ее значений на множестве нулевой меры.

Пусть даны числа a_n и b_n и известно, что они являются коэффициентами Фурье некоторой функции f , но ее значения нам неизвестны и мы должны найти их по a_n и b_n . Для этого обычно составляют ряд Фурье, отвечающий f ,

$$f(x) \sim \frac{1}{2} a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx). \quad (6)$$

*) Имеется в виду абсолютная интегрируемость по Риману. Для функций, интегрируемых по Лебегу, некоторые оговорки, которые делаются в дальнейшем изложении, будут излишними и могут быть опущены.

**) Говорят, что множество точек имеет меру нуль, если оно может быть покрыто счетным множеством интервалов, сумма длин которых может быть сделана меньше любого положительного числа ε .

Известен следующий простой результат: если ряд (6) равномерно на $[-\pi, \pi]$ сходится, то сумма его $s(x)$ будет иметь a_n и b_n своими коэффициентами Фурье и будет, следовательно, одной из разыскиваемых функций f .

Но может оказаться, что этот ряд будет расходиться на некотором множестве точек отрезка $[-\pi, \pi]$ или даже всюду на этом отрезке *). В этих случаях мы сталкиваемся со следующим: ряд (6) содержит достаточно полную информацию о функции f , определяющую ее всюду на $[-\pi, \pi]$, кроме множества точек меры нуль, но обычное понимание суммы ряда здесь не приводит к цели и не дает возможности найти функцию f . Чтобы получить правило для вычисления f по a_n и b_n , нужно каким-то целесообразным образом изменить понятие суммы ряда. Такие правила указываются в теории рядов Фурье. Простейшим из них является метод средних арифметических: нужно составить частичные суммы ряда (6)

$$s_n(x) = \frac{1}{2} a_0 + \sum_{k=1}^n (a_k \cos kx + b_k \sin kx)$$

и образовать из них средние арифметические величин

$$\sigma_n(x) = \frac{1}{n} [s_1(x) + \dots + s_n(x)].$$

Доказывается, что если f есть непрерывная на $[-\pi, \pi]$ функция **), то $\lim_{n \rightarrow \infty} \sigma_n(x) = f(x)$; при этом сходимость будет равномерной на $[-\pi, \pi]$.

*) Существуют, например, непрерывные 2π -периодические функции, для которых соответствующий им ряд Фурье расходится на всюду плотном на $[-\pi, \pi]$ множестве точек и такие функции f , абсолютно интегрируемые по Лебегу, для которых этот ряд расходится всюду на $[-\pi, \pi]$. (Примеры можно найти в книге: Н. К. Бари, Тригонометрические ряды, М., Гостехиздат, 1961, гл. V, §§ 17, 20).

**) В общем случае, когда f есть любая абсолютно интегрируемая функция, она может быть найдена по ряду (6) при помощи следующего предельного перехода:

$$\lim_{r \rightarrow 1-0} \left[\frac{1}{2} a_0 + \sum_{n=1}^{\infty} r^n (a_n \cos nx + b_n \sin nx) \right] = f(x)$$

всюду, кроме множества меры нуль.

2. В качестве второго примера рассмотрим задачу аналитического продолжения степенного ряда. Пусть аналитическая функция $f(z)$ комплексной переменной z задана степенным рядом

$$f(z) = \sum_{v=0}^{\infty} c_v z^v = c_0 + c_1 z + \dots, \quad (7)$$

сходящимся в круге $|z| < R$. Ряд представляет f и дает возможность вычислить ее значения только внутри круга и, может быть, в некоторых точках граничной окружности, там, где он сходится. Но область существования функции f может быть значительно шире, чем круг $|z| < R$.

Степенной ряд определяет функцию f во всей области ее существования, но если взять любую точку z в этой области, лежащую вне круга $|z| < R$, то вычислить там значение $f(z)$ при обычном понимании суммы ряда (7) мы не сможем, так как в этой точке ряд расходится. Значение $f(z)$ естественно считать обобщенной суммой расходящегося ряда (7). Вычислить же его через коэффициенты ряда c_v и координату z можно только в том случае, если отказаться от обычного понятия суммы ряда (7), подвергнуть его предварительно целесообразному преобразованию и лишь потом воспользоваться классическим понятием предела. Некоторые правила преобразования степенного ряда, применяемые для аналитического продолжения его суммы за круг сходимости, указываются в курсах теории функций комплексной переменной и книгах по теории расходящихся рядов.

Пусть дана произвольная последовательность s_n ($n = 1, 2, \dots$). Применим к ней преобразование (4), которое кратко запишем в виде

$$\sigma = f(s). \quad (8)$$

Если окажется, что новая последовательность σ_n сходится и $\lim \sigma_n = \sigma$, то ее предел σ называют f -пределом или обобщенным f -пределом заданной последовательности и пишут

$$\sigma = f - \lim_n s_n.$$

Величина σ зависит, очевидно, не только от последовательности s_n , но и от выбора преобразования, т. е. от

выбора функций f_n . Такой выбор не должен, очевидно, противоречить обычному пределу последовательности: если заданная последовательность s_n сходится к пределу s и если к s_n применимо преобразование (4) и обобщенный предел σ существует, то должно быть $\sigma = s$. Это требование оставляет еще достаточно большой произвол в выборе f_n .

Кроме того, в каждой конкретной задаче к преобразованию предъявляется свое требование, зависящее от цели задачи: если строится преобразование для обобщенного суммирования ряда Фурье и если обобщенный предел существует, то он обязан быть одной из функций f , для которой числа a_n и b_n будут коэффициентами Фурье; когда строится преобразование для обобщенного суммирования степенного ряда в точке z , лежащей вне круга сходимости $|z| < R$, то обобщенный предел должен давать одно из аналитических продолжений f в точке z (или избранную ветвь продолжения) и т. д.

Для более полной характеристики преобразований, которые будут рассматриваться в последующем изложении, полезно ознакомиться с некоторыми понятиями, относящимися к обобщенным пределам.

Рассмотрим некоторое множество S сходящихся последовательностей s_n . Говорят, что преобразование (4) сохраняет сходимость на S , если преобразование применимо ко всякой последовательности $s_n \in S$ и новая последовательность σ_n сходится: $\sigma_n \rightarrow \sigma$, при этом не обязательно предел σ_n совпадает с пределом s_n .

Если S есть множество всех сходящихся последовательностей и из $s_n \rightarrow s$ следует $\sigma_n \rightarrow \sigma$, то говорят просто, что преобразование сохраняет сходимость.

Когда из $s_n \rightarrow s$ ($s_n \in S$) следует $\sigma_n \rightarrow s$, то говорят, что преобразование сохраняет предел на S , или просто сохраняет предел, если s_n есть любая сходящаяся последовательность. В последнем случае, когда из $s_n \rightarrow s$ всегда следует $\sigma_n \rightarrow s$, преобразование (4) часто называют *регулярным*.

Признаки сохранения сходимости или предела при преобразовании изучены с достаточной полнотой лишь в случае линейных преобразований

$$\sigma_n = \sum_{m=1}^{\infty} a_{nm} s_m \quad (n = 1, 2, \dots). \quad (9)$$

Теорема 1 (о сохранении сходимости). *Для того чтобы преобразование (9) было применимо ко всякой сходящейся последовательности s_n ($s_n \rightarrow s$) и сохраняло сходимость, необходимо и достаточно выполнение условий:*

1) *существует число M такое, что при всяких $n = 1, 2, \dots$ выполняется неравенство*

$$A_n = \sum_{m=1}^{\infty} |a_{nm}| \leq M < \infty; \quad (10)$$

2) *при $n \rightarrow \infty$*

$$B_n = \sum_{m=1}^{\infty} a_{nm} \rightarrow \alpha; \quad (11)$$

3) *для каждого фиксированного значения m при $n \rightarrow \infty$ выполнено соотношение*

$$a_{nm} \rightarrow \alpha_m \quad (m = 1, 2, \dots). \quad (12)$$

При выполнении этих условий предельное значение σ_n дается равенством

$$\sigma = \lim \sigma_n = \left(\alpha - \sum_{m=1}^{\infty} \alpha_m \right) s + \sum_{n=1}^{\infty} \alpha_n s_n. \quad (13)$$

Доказательство теоремы мы опускаем и сделаем лишь некоторые пояснения ее условий. Первое из них не имеет простого и наглядного истолкования. Но можно показать, что если оно не выполняется и числа A_n не будут ограничены в совокупности, то существует такая сходящаяся последовательность s_n , для которой соответствующая последовательность σ_n будет неограниченной и, следовательно, расходящейся.

Если мы возьмем стационарную последовательность $s_1 = s_2 = \dots = 1$, ей будет отвечать последовательность $\sigma_n = B$ и условие 2) означает, что она должна сходиться к некоторому пределу α .

Наконец, условие 3) имеет следующий смысл. Возьмем последовательность s_n вида $0, \dots, 0, \overset{m}{1}, 0, 0, \dots$, состоящую из нулей, и только на месте номера m стоит единица. Она, очевидно, сходится к нулю. Соответствующая ей последовательность σ_n есть $\sigma_n = a_{nm}$ (m фиксировано, $n = 1, 2, \dots$). Условие 3) является записью требования сходимости такой последовательности σ_n .

Сделаем еще два простых замечания. Рассмотрим бесконечную матрицу линейного преобразования

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & \dots \\ a_{21} & a_{22} & \dots & a_{2m} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Числа B_n суть суммы элементов этой матрицы по строкам. Условие 2) требует, чтобы суммы B_n имели конечный предел при $n \rightarrow \infty$. Элементы a_{nm} , где m фиксировано, стоят в столбце номера m матрицы. Условие 3) требует, чтобы последовательности элементов, стоящих в каждом столбце матрицы, имели конечный предел.

Отметим также, что условия 2) и 3) могут быть объединены в одно более простое и наглядное. Последовательности вида $s_1, s_2, \dots, s_m, s, s, s, \dots$, у которых все значения, начиная с некоторого места, являются одинаковыми, называются *почти — или квазистационарными*. При помощи таких последовательностей условия 2) и 3) могут быть заменены одним: всякая почти стационарная последовательность $s_1, \dots, s_m, s, s, \dots$ преобразуется в сходящуюся последовательность σ_n .

Теорема о сохранении предела является простым следствием высказанной выше теоремы 1.

Теорема 2. *Для того чтобы преобразование (9) сохраняло предел, т. е. чтобы из $s_n \rightarrow s$ следовало $\sigma_n \rightarrow s$, необходимо и достаточно выполнение условий:*

1) *существует число M такое, что при $n = 1, 2, \dots$ выполняется неравенство*

$$A_n = \sum_{m=1}^{\infty} |a_{nm}| \leq M < \infty; \quad (14)$$

2) *при $n \rightarrow \infty$*

$$B_n = \sum_{m=1}^{\infty} a_{nm} \rightarrow 1; \quad (15)$$

3) *для каждого фиксированного m и $n \rightarrow \infty$ выполняется соотношение*

$$a_{nm} \rightarrow 0 \quad (m = 1, 2, \dots). \quad (16)$$

Первые условия теорем 1 и 2 совпадают. Второе условие теоремы 2 означает сохранение предела для всякой стационарной последовательности $s_n \equiv s$.

Коэффициент a_{nm} стоит множителем при s_m и соответствующий член в сумме (9) есть $a_{nm}s_m$. Поэтому условие (16) означает уничтожение влияния на предел σ_n ($n \rightarrow \infty$) любого значения s_m с фиксированным номером m . Полезно обратить внимание на то, что в теореме 1, в которой не требуется сохранение предела, на предельное значение новой последовательности $\sigma = \lim_n \sigma_n$ может оказывать влияние любое значение s_n заданной последовательности, как это видно в равенстве (13). В теореме 2 о сохранении предела влияние отдельных значений s , естественно, должно быть исключено.

Две рассмотренные задачи — об улучшении сходимости и о суммировании расходящихся рядов или пределах расходящихся последовательностей — связаны между собой. Во-первых, для решения их применяется один и тот же аппарат преобразований (4), и хотя эта связь является формальной, но она иногда позволяет дополнять исследования одной из задач результатами из другой. Во-вторых, задачи имеют некоторую внутреннюю связь, объяснить которую можно наиболее просто на частном случае. Причиной расходимости ряда $\sum_n a_n$ часто является недоста-

точно быстрое убывание модулей членов a_n . Для введения «обобщенной суммы» заданный ряд преобразуют в другой $\sum b_n$ [$b_n = \varphi(a_1, a_2, \dots)$], и если он сходится, то его сумму принимают за обобщенную сумму данного ряда. Так как $\sum_n b_n$ сходится, его члены b_n будут убывать, вообще го-

воря, более быстро, чем a_n , и преобразование $\sum_n a_n \rightarrow \sum_n b_n$ ускоряет убывание модулей членов ряда.

Поэтому можно ожидать, что каждое преобразование последовательности или ряда, позволяющее ввести понятие обобщенных сумм или пределов, должно, по крайней мере в некоторых случаях, улучшать сходимость. Можно также ожидать, что преобразования, построенные для улучшения сходимости, иногда будут применимы к определению обобщенных пределов и сумм. Примеры этого можно найти ниже.

§ 2. Интерполяционные методы преобразования последовательности

1. Некоторые общие сведения. Задача интерполирования предельного значения последовательности в общей форме будет рассматриваться только для пояснения содержания некоторых понятий, полезных в дальнейшем изложении. Для этой цели достаточно ограничиться случаем интерполирования при помощи линейных комбинаций.

Пусть дана сходящаяся последовательность s_n ($n = 1, 2, \dots$; $s_n \rightarrow s$). Ее можно рассматривать как функцию целочисленного аргумента n , область определения которой есть множество натуральных чисел плюс бесконечно далекая точка, где функция имеет некоторое значение $s = s_\infty$, подлежащее разысканию.

Допустим, что выбрана система функций $\omega_0(n)$, $\omega_1(n)$, $\dots$ целочисленного аргумента n ($n = 1, 2, \dots$), которая будет лежать в основе интерполирования: из функций $\omega_i(n)$ будут составляться линейные комбинации, служащие для интерполирования s_n ,

$$P_k(n) = a_0\omega_0(n) + a_1\omega_1(n) + \dots + a_k\omega_k(n). \quad (1)$$

Выбор $\omega_i(n)$ естественно подчинить следующим условиям:

- 1) никакая из функций $\omega_i(n)$ не есть линейная комбинация конечного числа других функций $\omega_j(n)$ ($j \neq i$);
- 2) каждая функция $\omega_i(n)$ ($i = 1, 2, \dots$) имеет конечное предельное значение при $n \rightarrow \infty$

$$\omega_i(n) \rightarrow \omega_i; \quad (2)$$

- 3) при всех k и для любых $k+1$ различных значений $n_0, n_1, \dots, n_k$ аргумента n выполняется неравенство *)

$$D_k = D_k(n_0, n_1, \dots, n_k) = \begin{vmatrix} \omega_0(n_0) & \omega_1(n_0) & \dots & \omega_k(n_0) \\ \omega_0(n_1) & \omega_1(n_1) & \dots & \omega_k(n_1) \\ \dots & \dots & \dots & \dots \\ \omega_0(n_k) & \omega_1(n_k) & \dots & \omega_k(n_k) \end{vmatrix} \neq 0. \quad (3)$$

Ниже будет ясно, что третье условие равносильно требованию возможности и единственности интерполиро-

*) Вместо этого часто говорят, что функции $\omega_i(n)$ ($n = 0, 1, \dots$) образуют систему Чебышева.

Отсюда сразу получается

$$P_k(n) = \frac{-1}{D_k} \begin{vmatrix} 0 & \omega_0(n) & \dots & \omega_k(n) \\ s_{n0} & & & \\ \vdots & & D_k & \\ s_{nk} & & & \end{vmatrix}. \quad (7)$$

Предельный переход при $n \rightarrow \infty$ дает следующее выражение для σ :

$$\lim_n P_k(n) = \sigma = \frac{-1}{D_k} \begin{vmatrix} 0 & \omega_0 & \dots & \omega_k \\ s_{n0} & & & \\ \vdots & & D_k & \\ s_{nk} & & & \end{vmatrix}. \quad (8)$$

Необходимо отметить, что погрешность $s - \sigma$ приближенного равенства $s \approx \sigma$ зависит от выбора функций $\omega_i(n)$ и узлов n_j ($j = 0, 1, \dots, n$). Обратим внимание на одну сторону этой зависимости: чтобы можно было сделать погрешность $s - \sigma$ сколь угодно малой, необходимо функции $\omega_i(n)$ взять такими, чтобы s_n можно было приблизить при помощи $P_k(n)$ сколь угодно точно, выбирая надлежащим образом k и коэффициенты a_j ($j = 0, 1, \dots, k$). Это приводит к следующему понятию полноты.

Пусть рассматривается некоторое множество S сходящихся последовательностей s_n .

Множество линейных комбинаций $P_k(n)$ называется *полным* в S , если для каждого $\varepsilon > 0$ и всякой последовательности $s_n \in S$ существуют такое k и такие коэффициенты a_j ($j = 0, 1, \dots, k$), что при всех значениях $n = 1, 2, \dots$ выполняется неравенство

$$|s_n - P_k(n)| < \varepsilon.$$

В качестве четвертого условия, которому подчиняют выбор $\omega_i(n)$, берут обычно следующее:

4) функции $\omega_i(n)$ таковы, что множество линейных комбинаций $P_k(n)$ является полным в множестве S всех сходящихся последовательностей s_n .

Равенство (8) дает интерполяционное правило преобразования последовательности s_n в число σ . Чтобы получить преобразование последовательности s_n в последова-

тельность σ_p ($p = 1, 2, \dots$), нужно указать бесконечную матрицу узлов

$$N = \begin{pmatrix} n_0^1 & n_1^1 & \dots & n_{k_1}^1 \\ n_0^2 & n_1^2 & \dots & n_{k_2}^2 \\ \dots & \dots & \dots & \dots \\ n_0^p & n_1^p & \dots & n_{k_p}^p \\ \dots & \dots & \dots & \dots \end{pmatrix}, \quad (9)$$

которая записана так, что в ее строке номера p стоят узлы, используемые для получения значения σ_p :

$$\lim_n P_{k_p}(n) = \sigma_p = \frac{-1}{D_{k_p}(n_0^p \dots n_{k_p}^p)} \begin{vmatrix} 0 & \omega_0 & \dots & \omega_n \\ s_{n_0^p} & \dots & \dots & \dots \\ \vdots & & & \\ s_{n_{k_p}^p} & & & \dots \end{vmatrix} \quad (10)$$

($p = 1, 2, \dots$).

Применяя полученное правило преобразования (6) при $p = 1, 2, \dots$, построим новую последовательность $\sigma_1, \sigma_2, \dots, \sigma_p, \dots$.

Естественно поставить вопрос ее сходимости. Оказывается, что сделанных предположений для этого, вообще говоря, недостаточно. Дело в том, что свойство полноты гарантирует существование для всякой сходящейся последовательности s_n такой последовательности $P_k(n)$, которая сходится к s_n равномерно относительно $n = 1, 2, \dots$. Для такой последовательности $P_k(\infty) \rightarrow s$ ($k \rightarrow \infty$ по некоторым значениям). При этом на выбор линейных комбинаций $P_k(n)$ не налагается ограничений: величины k и a_j могут быть любыми. В нашей проблеме построение $P_k(n)$ определяется равенством (7) и все значения $\omega_i(n_j)$ находятся по таблице узлов (9). Если функции $\omega_i(n)$ и таблица (9) заданы, то комбинации $P_k(n)$ ($k = 1, 2, \dots$) вполне определены, и мы не имеем права ничего выбирать, а тогда сходимости $\sigma_p \rightarrow s$ может не быть.

Если функции $\omega_i(n)$ таковы, что соответствующая им система $P_k(n)$ не обладает свойством полноты в рассматриваемом семействе S , то, по-видимому, следует считать выбор функций $\omega_i(n)$ неудачным, так как невозможно рассчитывать на сходимость $\sigma_p \rightarrow s$ для всех s_n из семейства S .

Если же система $P_k(n)$ обладает полнотой в S , то это еще не может гарантировать сходимость $\sigma_p \rightarrow s$ при любых $s_n \subset S$. Проблема сходимости должна быть исследована дополнительно. В ней мы имеем дело со следующими объектами: множеством S последовательностей, функциями $\omega_i(n)$ ($i=0, 1, \dots$), лежащими в основе интерполирования, и таблицей N узлов интерполирования (9). Необходимо выяснить, как должны быть связаны между собой эти объекты, чтобы можно было ответить на следующие основные вопросы.

1. Для какого множества S последовательностей s_n новая последовательность σ_n будет сходящейся?

2. В каком случае из $s_n \rightarrow s$ и $\sigma_n \rightarrow \sigma$ следует $\sigma = s$, т. е. когда преобразование (10) будет сохранять предел?

3. Для каких сходящихся последовательностей s_n преобразование (10) ускоряет (или улучшает) сходимость?

4. Как может быть оценена погрешность $\sigma_n - s$ новой последовательности?

2. Интерполирование при помощи многочленов от $1/n$. Случай постоянного числа узлов. В качестве функций $\omega_i(n)$ возьмем отрицательные степени n :

$$\omega_0(n) = 1, \quad \omega_1(n) = \frac{1}{n}, \quad \dots, \quad \omega_i(n) = \frac{1}{n^i}, \quad \dots$$

Их линейными комбинациями будут многочлены от $1/n$:

$$a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \dots + \frac{a_k}{n^k} = P_k\left(\frac{1}{n}\right).$$

Полнота всех $P_k(1/n)$ на множестве сходящихся последовательностей s_n является следствием теоремы Вейерштрасса о возможности равномерного приближения непрерывной функции алгебраическим многочленом.

Пусть $s_n \rightarrow s$ ($n \rightarrow \infty$). Введем новую переменную x , положив $x = n^{-1}$, $n = x^{-1}$. Рассмотрим функцию $v(x) = s_n$. Она определена в точках $x = 1, \frac{1}{2}, \frac{1}{3}, \dots$. Доопределим ее всюду на отрезке $[0, 1]$, интерполируя ее линейно на отрезках $\left[\frac{1}{k}, \frac{1}{k+1}\right]$ ($k = 1, 2, \dots$) и положив $v(0) = s$. Так построенная функция $v(x)$ будет непрерывной всюду на $[0, 1]$. Для любого $\varepsilon > 0$ существует многочлен $P_k(x) = a_0 + a_1x + \dots + a_kx^k$ некоторой степени k , такой, что при

всяких $x \in [0, 1]$ будет верно неравенство $|v(x) - P_k(x)| < \varepsilon$. В точках же $x = 1/n$ будет выполняться неравенство

$$\left| s_n - P_k\left(\frac{1}{n}\right) \right| < \varepsilon \quad (n = 1, 2, \dots),$$

доказывающее полноту $P_k(1/n)$.

Для интерполирования на шаге номера p и построения σ_p возьмем $k+1$ узлов $n = p, p+1, \dots, p+k$. Матрица (9) узлов интерполирования для всех шагов при $p = 1, 2, \dots$ имеет вид

$$N = \begin{vmatrix} 1 & 2 & \dots & k \\ 2 & 3 & \dots & k+1 \\ \dots & \dots & \dots & \dots \\ p & p+1 & \dots & p+k \\ \dots & \dots & \dots & \dots \end{vmatrix}.$$

Возможность и единственность интерполирования посредством $P_k(1/n)$ любой последовательности s_n по значениям ее в этих узлах вытекает из аналогичных результатов для интерполирования обычными алгебраическими многочленами, полученных в первой части книги (см. п. 1.1.1).

Выражение интерполирующего многочлена $P_k(1/n)$ может быть построено при помощи формулы Лагранжа и будет следующим:

$$P_k\left(\frac{1}{n}\right) = \sum_{j=0}^k \frac{\omega\left(\frac{1}{n}\right)}{\left(\frac{1}{n} - \frac{1}{p+j}\right) \omega'\left(\frac{1}{p+j}\right)} s_{p+j}, \quad (11)$$

где

$$\begin{aligned} \omega\left(\frac{1}{n}\right) &= \left(\frac{1}{n} - \frac{1}{p}\right) \left(\frac{1}{n} - \frac{1}{p+1}\right) \dots \left(\frac{1}{n} - \frac{1}{p+k}\right), \\ \omega'\left(\frac{1}{p+j}\right) &= \left(\frac{1}{p+j} - \frac{1}{p}\right) \dots \\ &\dots \left(\frac{1}{p+j} - \frac{1}{p+j-1}\right) \left(\frac{1}{p+j} - \frac{1}{p+j+1}\right) \dots \left(\frac{1}{p+j} - \frac{1}{p+k}\right) = \\ &= \frac{(-1)^j j! (k-j)!}{(p+j)^k p(p+1) \dots (p+j-1)(p+j+1) \dots (p+k)}. \end{aligned}$$

Предел $P_k(1/n)$ при $n \rightarrow \infty$ принимается за σ_p :

$$\sigma_p = \sum_{j=0}^k \frac{(-1)^{k-j} (p+j)^k}{j! (k-j)!} s_{p+j} = L(s_p) \quad (p = 1, 2, \dots). \quad (12)$$

Это равенство является записью линейного преобразования последовательности s_n в последовательность σ_p . Оно применимо ко всякой последовательности s_n независимо от сходимости или расходимости ее и является частным случаем общего линейного преобразования (1.9), когда

$$a_{pm} = \begin{cases} 0, & m < p, \\ \frac{(-1)^{k+p-m}}{(m-p)!(k+p-m)!}, & p \leq m \leq p+k, \\ 0, & m > p+k. \end{cases}$$

Формула (12) показывает, что при возрастании p коэффициенты преобразования растут как p^k , поэтому для них не выполняется первое условие теорем 1 и 2 и преобразование может, следовательно, переводить сходящуюся последовательность s_n в расходящуюся последовательность σ_p .

Обратим еще внимание на некоторые свойства коэффициентов преобразования (12). Функции $s_n \equiv 1$, $s_n = \frac{1}{n}$, ..., $s_n = \frac{1}{n^k}$ интерполируются многочленом $P_k(1/n)$ точно и преобразование (12) должно при всяких p точно давать их предельные значения, что приводит к следующим соотношениям:

$$\begin{aligned} L(1) &= \sum_{j=0}^k (-1)^{k-j} \frac{(p+j)^k}{j!(k-j)!} = 1, \\ L\left(\frac{1}{p^m}\right) &= \sum_{j=0}^k (-1)^{k-j} \frac{(p+j)^k}{j!(k-j)!} \cdot \frac{1}{(p+j)^m} = 0 \\ &(p=1, 2, \dots; m=1, 2, \dots, k). \end{aligned} \quad (13)$$

Чтобы ответить на некоторые из вопросов, поставленных в конце предыдущего пункта применительно к преобразованию (12), обратим это преобразование. Для этого перепишем его подробнее:

$$L(s_p) = \frac{(p+k)^k}{k!} s_{p+k} - \frac{(p+k-1)^k}{1!(k-1)!} s_{p+k-1} + \dots + (-1)^k \frac{p^k}{k!} s_p = \sigma_p. \quad (14)$$

Примем прежде всего во внимание то, что это преобра-

зование относительно s_p является линейным конечноразностным уравнением порядка k со свободным членом σ_p . Когда σ_p тождественно равно нулю и $L(s_p) = 0$, оно называется однородным. Существенно заметить следующее: из второго равенства (13) следует, что функции $s_p = \frac{1}{p}$, $s_p = \frac{1}{p^2}$, ..., $s_p = \frac{1}{p^k}$ являются решениями однородного уравнения $L(s_p) = 0$.

Полагая в (14) $p = 1$, получим равенство

$$L(s_1) = \frac{(k+1)^k}{k!} s_{k+1} - \frac{k^k}{1!(k-1)!} s_k + \dots + (-1)^k \frac{1}{k!} s_1 = \sigma_1,$$

из которого можно найти s_{k+1} , выразив его линейно через $s_1, \dots, s_k$ и σ_1 . При $k = 2$ формула (14) даст

$$L(s_2) = \frac{(k+2)^k}{k!} s_{k+2} - \frac{(k+1)^k}{1!(k-1)!} s_{k+1} + \dots + (-1)^k \frac{2^k}{k!} s_2 = \sigma_2,$$

откуда найдем s_{k+2} как линейную функцию от $s_2, \dots, s_{k+1}$ и σ_2 , и так как раньше s_{k+1} было выражено через $s_1, \dots, s_k, \sigma_1$, то окончательно найдем s_{k+2} как линейную функцию от $s_1, \dots, s_k, \sigma_1, \sigma_2$ и т. д.

Продолжая такие вычисления, мы представим s_p в виде линейной комбинации от $s_1, \dots, s_k, \sigma_1, \dots, \sigma_{p-k}$, коэффициенты которой зависят от p :

$$s_p = \sum_{j=1}^k \Gamma_p^j s_j + \sum_{i=1}^{p-k} G_p^i \sigma_i \quad (p = k+1, k+2, \dots). \quad (15)$$

Величины $s_1, \dots, s_k$, называемые начальными значениями решения, остаются произвольными и могут иметь любые значения. Когда они заданы, все дальнейшие значения s_p ($p > k$) решения уравнения определяются единственным образом.

Коэффициент Γ_p^j учитывает влияние, которое оказывает на s_p начальное значение s_j . Аналогично коэффициент G_p^i учитывает влияние значения σ_i свободного члена. Ввиду такой роли Γ_p^j и G_p^i их называют *функциями влияния* или *функциями Грина* соответствующих начальных значений и значений правой части.

Для нахождения Γ_p^i заметим прежде всего, что когда все σ_i ($i = 1, 2, \dots$) равны нулю и уравнение (14) однородно $L(s_p) = 0$, в представлении (15) для s_p останется

только сумма с s_j . Если, кроме того, все начальные значения, кроме s_j , равны нулю, а $s_i = 1$, то $s_p = \Gamma_p^j$. Поэтому можно сказать, что Γ_p^j есть решение однородного уравнения $L(s_p) = 0$ с начальными значениями $\Gamma_p^i = 0$ ($p = 1, \dots, k, p \neq j$), $\Gamma_j^j = 1$.

Γ_p^j имеет следующее выражение:

$$\Gamma_p^j = \frac{(-1)^{j-1}}{(j-1)!(k-j)!} \cdot \frac{j}{p} \left(\frac{1}{p} - 1\right) \left(\frac{1}{p} - \frac{1}{2}\right) \dots \\ \dots \left(\frac{1}{p} - \frac{1}{j-1}\right) \left(\frac{1}{p} - \frac{1}{j+1}\right) \dots \left(\frac{1}{p} - \frac{1}{k}\right). \quad (16)$$

Проверим это. Правая часть равенства (16) есть многочлен от $1/p$ степени k без свободного члена и является линейной комбинацией функций $s_p = \frac{1}{p}$, $s_p = \frac{1}{p^2}$, ..., $s_p = \frac{1}{p^k}$. Но каждая из них есть решение однородного уравнения $L(s_p) = 0$, поэтому и Γ_p^j также есть решение этого уравнения. То, что правая часть имеет нулевые начальные значения при $p = 1, \dots, k, p \neq j$, является очевидным, численный же множитель $(-1)^{j-1}[(j-1)!(k-j)!]^{-1}$ в правой части выбран так, чтобы она была равна единице при $p = j$.

Рассмотрим теперь коэффициент G_p^m . Он совпадает с решением s_p , если считать, что все s_j ($j = 1, \dots, k$) равны нулю и $\sigma_i = 0$ ($i \neq m$), $\sigma_m = 1$. Таким образом, G_p^j есть решение неоднородного уравнения

$$L(s_p) = \delta_p^m \quad (17)$$

с нулевыми начальными значениями.

Когда $p < m$, указанное уравнение будет однородным, так как $\delta_p^m = 0$, и ввиду нулевых начальных значений его решение будет тождественным нулем для $p < m + k$, так что

$$G_p^m = 0, \quad p = 1, 2, \dots, m + k - 1. \quad (18)$$

При $p = m$ уравнение (17) дает равенства

$$L(G_m^m) = \frac{(m+k)^k}{k!} G_{m+k}^m = 1, \quad G_{m+k}^m = \frac{k!}{(m+k)^k}.$$

Для $p > m$, так как $\delta_p^m = 0$, уравнение (17) вновь становится однородным. Можно, следовательно, сказать,

что G_p^m для $p > m$ есть решение однородного уравнения с следующими начальными значениями:

$$G_{m+1}^m = \dots = G_{m+k-1}^m = 0, \quad G_{m+k}^m = \frac{k!}{(m+k)^k}.$$

Покажем, что при $p \geq m+k$ функция G_p^m представима в форме

$$G_p^m = \frac{k}{p} \left(1 - \frac{m+1}{p}\right) \left(1 - \frac{m+2}{p}\right) \dots \left(1 - \frac{m+k-1}{p}\right) \quad (p \geq m+k). \quad (19)$$

Действительно, правая часть есть многочлен от $1/p$ степени k без свободного члена, она является линейной комбинацией решений $s_p = \frac{1}{p}$, $s_p = \frac{1}{p^2}$, ..., $s_p = \frac{1}{p^k}$ однородного уравнения $L(s_p) = 0$ и является поэтому решением однородного уравнения. Начальные же значения правой части при $p = m+1, m+2, \dots, m+k$ совпадают с начальными значениями G_p^m .

Таким образом, в равенстве (15), обращающем рассматриваемое преобразование (12), найдены все коэффициенты. Правило обращения можно считать известным. Оно позволяет дать ответ на два первых вопроса, сформулированных в конце предыдущего пункта. Если в этом равенстве считать величины s_j ($j = 1, \dots, k$) и σ_i ($i = 1, 2, \dots$) произвольными численными параметрами, то можно сказать, что множество всех последовательностей s_n , которые при преобразовании (12) переходят в сходящиеся последовательности σ_i , имеет параметрическое представление (15), где параметры s_j могут иметь произвольные значения, параметры же σ_i должны образовать сходящуюся последовательность, которая может быть любой.

Но если дана последовательность s_n , то проверить, что она может быть представлена в форме (15) со сходящейся последовательностью σ_i ($i = 1, 2, \dots$) можно, вероятно, в очень редких случаях. Поэтому возможность указанного представления s_n нельзя считать эффективным признаком того, что s_n при преобразовании (12) перейдет в сходящуюся последовательность.

Более простой и интересный ответ представление (15) позволяет дать на второй вопрос.

Теорема 1. Если последовательность σ_n сходится и $\lim \sigma_n = \sigma$, то всякая последовательность s_n , переходящая при преобразовании (12) в σ_n , также будет сходящейся и при этом $\lim s_n = \sigma$.

Теорема может быть сформулирована более кратко: если преобразование (12) сохраняет сходимость, то оно сохраняет и предел.

Доказательство. Рассмотрим отдельно суммы, входящие в (15):

$$s_p = s'_p + s''_p, \quad s'_p = \sum_{j=1}^k \Gamma_p^j s_j, \quad s''_p = \sum_{i=1}^{p-k} G_p^i \sigma_i.$$

Формула (16) позволяет сказать, что $\Gamma_p^j \rightarrow 0$ при $p \rightarrow \infty$, и поэтому при всяких s_j сумма $s'_p \rightarrow 0$ ($p \rightarrow \infty$).

Рассмотрим частный случай стационарной последовательности $s_n \equiv 1$. Для нее правило обращения (15), ввиду того, что при всех значениях p $L(1) = \sigma_p$, имеет вид

$$1 = \sum_{j=1}^k \Gamma_p^j + \sum_{i=1}^{p-k} G_p^i.$$

При $p \rightarrow \infty$ первая из сумм правой части стремится к нулю и, следовательно,

$$\sum_{i=1}^{p-k} G_p^i \rightarrow 1, \quad p \rightarrow \infty. \quad (20)$$

Рассмотрим теперь вторую сумму

$$s''_p = \sum_{i=1}^{p-k} G_p^i \sigma_i. \quad (21)$$

Равенство можно рассматривать как линейное преобразование сходящейся последовательности σ_i ($\lim \sigma_i = \sigma$) в последовательность s''_p . Это есть частный случай линей-

ного преобразования $s''_p = \sum_{i=1}^{\infty} a_{pi} \sigma_i$. Покажем, что преобразование (21) сохраняет предел. Применим к нему теорему 1 п. 1.2 о сохранении предела и убедимся в том, что для (21) условия теоремы выполняются. В самом деле, заметим прежде всего, что из (18) и (19) следует неотрицательность G_p^i при всех p .

Так как

$$A_p = \sum_{i=1}^{p-k} |G_p^i| = \sum_{i=1}^{p-k} G_p^i \rightarrow 1 \quad (p \rightarrow \infty),$$

то множество чисел A_p ограничено: $A_p \leq M < \infty$ и первое условие теоремы выполнено.

Ввиду того же соотношения (20)

$$B_p = \sum_{i=1}^{p-k} G_p^i \rightarrow 1 \quad (p \rightarrow \infty)$$

и второе условие теоремы также выполняется.

Наконец, из представления (19) следует, что если m фиксировано и $p \rightarrow \infty$, то $G_p^m \rightarrow 0$. Поэтому выполнено третье условие теоремы. Все сказанное убеждает в том, что преобразование (21) действительно сохраняет предел, и, стало быть, $\lim_p s_p'' = \sigma$.

Из равенства $s_p = s_p' + s_p''$ вытекает, что $s_p \rightarrow 0 + \sigma = \sigma$ при $p \rightarrow \infty$, что доказывает теорему.

3. Некоторые типы сходимости. Выше мы обращали внимание на то, что методы улучшения сходимости должны быть тесно связаны с законами стремления последовательностей к их пределам. В этом пункте будут указаны некоторые виды сходимостей, когда правила преобразований, основанные на алгебраическом интерполировании, могут ускорить сходимость.

Сходимость почти алгебраического типа. Говорят, что последовательность s_n имеет почти алгебраический тип сходимости к пределу степени r , если она представима в форме

$$s_n = c_0 + \frac{c_1}{n} + \frac{c_2}{n^2} + \dots + \frac{c_r + \gamma_r(n)}{n^r}, \quad \gamma_r(n) \rightarrow 0 \quad (n \rightarrow \infty), \quad (22)$$

где $c_0, \dots, c_r$ — постоянные числа, причем свободный член c_0 здесь есть предельное значение s_n .

Асимптотический степенной тип сходимости. Переменную s_n называют обладающей асимптотическим степенным типом сходимости, если существует

степенной ряд $a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \dots$ такой, что при любом $r = 1, 2, \dots$ верно соотношение

$$\left[s_n - \left(a_0 + \frac{a_1}{n} + \dots + \frac{a_r}{n^r} \right) \right] n^r \rightarrow 0 \quad (n \rightarrow \infty). \quad (23)$$

Это равносильно представлению s_n в виде (22) при любом r . Поэтому принадлежность s_n асимптотическому степенному типу сходимости равносильна принадлежности s_n типу почти алгебраической сходимости сколь угодно высокой степени.

Аналитический тип сходимости. Последовательность s_n называют обладающей аналитическим типом сходимости, если существует число N такое, что при всяком $n > N$ переменная s_n является суммой степенного ряда вида

$$s_n = c_0 + \frac{c_1}{n} + \dots + \frac{c_r}{n^r} + \dots \quad (24)$$

Введем комплексную переменную z и рассмотрим аналитическую функцию $F(z)$, определенную степенным рядом

$$F(z) = c_0 + \frac{c_1}{z} + \dots + \frac{c_r}{z^r} + \dots$$

Ряд сходится по меньшей мере в области $|z| > N$.

В целых точках $z = n$ функция $F(z)$ принимает такие же значения, как и s_n .

Внутри области регулярности аналитическая функция изменяется очень плавно и принадлежность s_n аналитическому типу сходимости означает, что s_n приближается к предельному значению особенно «гладко».

Указанные три типа сходимости часто встречаются в прикладных задачах и выработка правил ускорения сходимости для них имеет не только теоретическое, но и практическое значение.

4. Вспомогательная лемма. Выше мы указывали на значения, которые принимает оператор $L(s_p)$ на нескольких первых отрицательных степенях p [см. (13)]:

$$L(1) = 1, \quad L(p^{-m}) = 0, \quad m = 1, 2, \dots, k.$$

Нам потребуются значения этого оператора при $m > k$.

Лемма 1. Если $m > k$, то

$$L(p^{-m}) = \frac{1}{p(p+1)\dots(p+k)} a_{m-k-1}, \quad (25)$$

$$a_v = \sum_{\alpha_0 + \alpha_1 + \dots + \alpha_k = v} \frac{1}{p^{\alpha_0} (p+1)^{\alpha_1} \dots (p+k)^{\alpha_k}}, \quad v \geq 0, \alpha_i \geq 0.$$

Доказательство. Введем новую переменную $t = \frac{1}{n}$ и выполним интерполирование функции $s_n = \frac{1}{p^n} = t^m$ по ее значениям в узлах $t = \frac{1}{p}, \frac{1}{p+1}, \dots, \frac{1}{p+k}$:

$$P_k\left(\frac{1}{n}\right) = \Pi_k(t) = \sum_{j=0}^k \frac{\omega(t)}{\left(t - \frac{1}{p+j}\right) \omega'\left(\frac{1}{p+j}\right)} \cdot \frac{1}{(p+j)^m},$$

$$\omega(t) = \left(t - \frac{1}{p}\right) \dots \left(t - \frac{1}{p+k}\right).$$

Возьмем представление остатка интерполирования $t^m - \Pi_k(t)$ в форме контурного интеграла (см. п. 1.3.2, (1.3.18))

$$t^m - \Pi_k(t) = \frac{\omega(t)}{2\pi i} \int_l \frac{u^m}{(u-t)\omega(u)} du.$$

Здесь l есть произвольная замкнутая линия, охватывающая отрезок $\left[0, \frac{1}{p}\right]$. Напомним, что значение $L\left(\frac{1}{p^n}\right) = \sigma_p$ получается из $P_k\left(\frac{1}{n}\right) = \Pi_k(t)$ в пределе при $n \rightarrow \infty$ или $t \rightarrow 0$. Перейдем в представлении остатка к пределу при $t \rightarrow 0$:

$$0 - \Pi_k(0) = \sigma_p = \frac{\omega(0)}{2\pi i} \int_l \frac{u^m}{u\omega(u)} du$$

или ввиду того, что $\omega(0) = \frac{(-1)^{k+1}}{p(p+1)\dots(p+k)}$,

$$L(p^{-m}) = \sigma_p = \frac{(-1)^k}{p(p+1)\dots(p+k)} \cdot \frac{1}{2\pi i} \int_l \frac{u^m}{u\omega(u)} du. \quad (26)$$

Интеграл может быть легко вычислен просто. Будем считать контур интегрирования l линией, охватывающей

$\left[0, \frac{1}{p}\right]$, например окружностью с центром в начале координат и достаточно большого радиуса. Тогда $|u|$ на l будет большой величиной и будут возможны проводимые ниже разложения в степенные ряды по степеням $1/u$, при этом ряды будут сходиться равномерно относительно u на l и допустимым будет почленное интегрирование ряда, которое выполняется ниже. Подынтегральная функция в соотношении (26) представляется в виде

$$\begin{aligned} \frac{u^m}{u\omega(u)} &= u^{m-1} \cdot \frac{1}{u^{k+1} \left(1 - \frac{1}{up}\right) \left(1 - \frac{1}{u(p+1)}\right) \dots \left(1 - \frac{1}{u(p+k)}\right)} = \\ &= u^{m-k-2} \sum_{\alpha_0=0}^{\infty} \frac{1}{u^{\alpha_0} p^{\alpha_0}} \sum_{\alpha_1=0}^{\infty} \frac{1}{u^{\alpha_1} (p+1)^{\alpha_1}} \dots \sum_{\alpha_k=0}^{\infty} \frac{1}{u^{\alpha_k} (p+k)^{\alpha_k}} = \\ &= u^{m-k-2} \sum_{v=0}^{\infty} \frac{1}{u^v} \sum_{\alpha_0+\alpha_1+\dots+\alpha_k=v} \frac{1}{p^{\alpha_0} (p+1)^{\alpha_1} \dots (p+k)^{\alpha_k}}. \end{aligned}$$

При вычислении интеграла $\int_l \frac{u^m}{u\omega(u)} du$ заменим интегрируемую функцию полученным для нее степенным рядом и выполним почленное интегрирование. Если принять во внимание равенства

$$\int_l u^s ds = \begin{cases} 0, & s \neq -1, \\ 2\pi i, & s = -1, \end{cases}$$

то будет ясно, что

$$\begin{aligned} \int_l \frac{u^m}{u\omega(u)} du &= \\ &= \sum_{v=0}^{\infty} \sum_{\alpha_0+\dots+\alpha_k=v} \frac{1}{p^{\alpha_0} (p+1)^{\alpha_1} \dots (p+k)^{\alpha_k}} \int_l u^{m-k-2-v} du = \\ &= 2\pi i \sum_{\alpha_0+\dots+\alpha_k=m-k-1} \frac{1}{p^{\alpha_0} (p+1)^{\alpha_1} \dots (p+k)^{\alpha_k}}. \end{aligned}$$

Для завершения доказательства леммы осталось внести полученное значение интеграла в равенство (26).

Необходимо сделать одно замечание, полезное для дальнейшего изложения. Когда p неограниченно возра-

стает, величина a_v при $v \geq 1$ стремится к нулю, при этом порядок малости a_v будет следующим: $a_v = O(p^{-v})$, и этот порядок является точным, так как в сумме, которой определяется a_v , все слагаемые положительны.

Из равенства (25) следует, что при возрастании p , в силу соотношений $\frac{1}{p(p+1)\dots(p+k)} = O(p^{-k-1})$ и $a_{m-k-1} = O(p^{-m+k+1})$, значение оператора $L(p^{-m})$ будет малой величиной следующего точно указанного порядка:

$$L\left(\frac{1}{p^m}\right) = O\left(\frac{1}{p^m}\right). \quad (27)$$

5. Некоторые теоремы об ускорении сходимости. Нашей целью будет выяснение условий, при которых преобразование (12) будет ускорять сходимость последовательностей, принадлежащих трем типам, определенным в п. 3.

Пусть s_n принадлежит почти алгебраическому типу сходимости степени r и имеет следующее представление:

$$s_n = s + \frac{c_q}{n^q} + \frac{c_{q+1}}{n^{q+1}} + \dots + \frac{c_r}{n^r} + \frac{\gamma_r(n)}{n^r}, \quad (28)$$

где

$$\gamma_r(n) \rightarrow O(n \rightarrow \infty), \quad c_q \neq 0.$$

Скорость сходимости $s_n \rightarrow s$ определяется, очевидно, членом $c_q n^{-q}$; q называют *порядком сходимости*.

Применим к s_n преобразование (12)

$$\begin{aligned} \sigma_n = L(s_n) = s + c_q L\left(\frac{1}{n^q}\right) + c_{q+1} L\left(\frac{1}{n^{q+1}}\right) + \\ + \dots + c_r L\left(\frac{1}{n^r}\right) + L(n^{-r} \gamma_r(n)). \end{aligned} \quad (29)$$

Оценим последнее слагаемое. Если считать $\gamma_r(n)$ произвольной величиной, стремящейся к нулю при $n \rightarrow \infty$, то будет верна следующая неулучшаемая оценка:

$$\begin{aligned} |L(n^{-r} \gamma_r(n))| &= \left| \sum_{j=0}^k (-1)^{k-j} \frac{(n+j)^k}{j! (k-j)!} \cdot \frac{\gamma_r(n+j)}{(n+j)^r} \right| \leq \\ &\leq \max_j |\gamma_r(n+j)| \sum_{j=0}^k \frac{1}{j! (k-j)!} \cdot \frac{1}{(n+j)^{r-k}} = \\ &= \max_j |\gamma_r(n+j)| O\left(\frac{1}{n^{r-k}}\right). \end{aligned}$$

Так как $\max_i |\gamma_r(n+j)| \rightarrow 0$ ($n \rightarrow \infty$), то $L(n^{-r}\gamma_r(n)) = o\left(\frac{1}{n^{r-k}}\right)$.

Рассмотрим сначала случай $k < q$. Тогда, в соответствии с леммой, $L\left(\frac{1}{n^q}\right)$ отлично от нуля и будет иметь значение, указанное в (25) при $m = q$. Остальные члены в правой части равенства (29), имеющие вид $L\left(\frac{1}{n^i}\right)$ ($i > q$), обладают более высоким порядком малости.

Главным членом разности $\sigma_n - s$ может быть либо $L(n^{-q})$, имеющий порядок $O(n^{-q})$, либо последний член, порядок которого $o(n^{k-r})$. Поэтому в рассматриваемом случае $k < q$ ускорения сходимости при преобразовании, вообще говоря, не произойдет и не исключено даже замедление сходимости, когда $L(n^{-r}\gamma_r(n))$ стремится к нулю менее быстро, чем n^{-q} .

Указанные результаты можно было предвидеть заранее, так как если степень точности интерполирования k взять меньше порядка сходимости, то при такой несогласованности между k и q невозможно ожидать ускорения сходимости.

Положим теперь $k \geq q$. При такой степени точности интерполирования $L(n^{-q}) = \dots = L(n^{-k}) = 0$, несколько первых членов в представлении (29) для $L(s_n)$ исчезнут и $L(s_n)$ примет вид

$$L(s_n) = s + c_{k+1}L\left(\frac{1}{n^{k+1}}\right) + \dots + c_rL\left(\frac{1}{n^r}\right) + L(n^{-r}\gamma_r(n)).$$

Остающиеся справа члены вида $L(n^{-i})$ ($i \geq k+1 > q$) будут иметь порядок малости более высокий, чем $L(n^{-q})$. Вопрос об ускорении сходимости решится последним членом $L(n^{-r}\gamma_r(n)) = o\left(\frac{1}{n^{r-k}}\right)$. Чтобы он стремился к нулю быстрее $L(n^{-q})$, достаточно выполнить неравенство $r - k \geq q$ или $k \leq r - q$.

Изложенное позволяет считать доказанной следующую теорему.

Теорема 2. Для того чтобы преобразование (12) увеличивало скорость сходимости последовательности вида (28), достаточно степень точности k преобразования (12) взять так, чтобы выполнялись неравенства

$$q \leq k \leq r - q. \quad (30)$$

Более просто формулируется правило ускорения сходимости для последовательностей, принадлежащих асимптотическому и аналитическому типу сходимости. Для них показатель степени r может быть взят как угодно большим, и поэтому вторая часть $k \leq r - q$ условия (30) всегда выполняется. Поэтому верна

Теорема 3. Для того чтобы преобразование (12) увеличивало скорость сходимости последовательностей асимптотического степенного и аналитического типа сходимости достаточно степень точности интерполяционного преобразования k взять не меньше, чем порядок сходимости q последовательности s_n :

$$k \geq q.$$

6. Интерполяционные методы преобразования с возрастающим числом узлов. По поводу преобразования (12), построенного в п. 2, можно сделать замечание, побуждающее изменить постановку задачи интерполирования предельного значения s и построения σ_p , а именно: на шаге номера p интерполирование s_n выполнялось по значениям s_n в узлах $p, p+1, \dots, p+k$, что же касается предшествующих узлов $1, 2, \dots, p-1$, то значения s_n в них не использовались. Поэтому естественно построить интерполяционные методы, которые используют все предшествующие значения s_n : при интерполировании предельного значения s и построении σ_n на шаге номера n пользуются узлами $1, 2, \dots, n$.

Соответствующая такому процессу преобразований таблица N узлов будет треугольной:

$$N = \begin{vmatrix} 1 & & & & \\ 1 & 2 & & & \\ 1 & 2 & 3 & & \\ \dots & \dots & \dots & \dots & \\ 1 & 2 & 3 & \dots & n \\ \dots & \dots & \dots & \dots & \dots \end{vmatrix}.$$

Правило построения приближения σ_n к предельному значению s здесь может быть получено из (12), если положить там $p=1$, $k=n-1$:

$$\sigma_n = \sum_{j=0}^{n-1} \frac{(-1)^{n-1-j} (1+j)^{n-1}}{j! (n-1-j)!} s_{n+j}.$$

После изменения индекса $j+1=j'$ и отбрасывания штриха у нового индекса преобразование запишется в следующей более удобной форме:

$$\sigma_n = \sum_{j=1}^n \frac{(-1)^{n-j} j^n}{j! (n-j)!} s_j = L^*(s_n). \quad (31)$$

Отметим некоторые свойства этого преобразования.

1. Степень точности преобразования. Напомним, что для получения σ_n должно быть выполнено интерполирование переменной s_n по ее значениям в n точках $1, 2, \dots, n$ с помощью многочлена от $\frac{1}{p}$ степени $n-1$ $P_{n-1}\left(\frac{1}{p}\right)$. Такое интерполирование будет точным для случаев $s_p \equiv 1$, $s_p = \frac{1}{p}$, $\dots$, $s_p = \frac{1}{p^{n-1}}$. Поэтому преобразование (31) должно при всяких n давать точно предельные значения этих последовательностей и, стало быть, должны точно выполняться равенства

$$L^*(1) = \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} = 1, \quad (32)$$

$$L^*\left(\frac{1}{n^m}\right) = \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \cdot \frac{1}{j^m} = 0 \quad (m = 1, 2, \dots, n-1). \quad (33)$$

2. Преобразование (31) применимо ко всякой переменной s_n , сходящейся или расходящейся. Представляет интерес (по крайней мере теоретический) выяснение вопроса, может ли преобразование (31) переводить расходящиеся последовательности в сходящиеся. Несколько позже будет доказано, что это не так и что из предположения о сходимости σ_n вытекает сходимость s_n .

Можно проверить, что преобразование (31) не всегда сохраняет сходимость и может сходящуюся последовательность s_n переводить в расходящуюся последовательность σ_n . Для этого достаточно проверить, что для коэффициентов преобразования не выполняется первое условие теоремы 1 из п. 1.1. Действительно, рассмотрим коэффициент, соответствующий значению $j=n$: $a_{nn} = \frac{n^n}{n!}$. Так как $n! = n^n e^n \sqrt{2\pi n} (1 + \varepsilon_n)$ ($\varepsilon_n \rightarrow 0$), то $a_{nn} = e^n (2\pi n)^{-1/2} (1 + \delta_n)$

($\delta_n \rightarrow 0$) и при $n \rightarrow \infty$ этот коэффициент будет быстро и неограниченно возрастать. Поэтому для преобразования (31) величины $A_n = \sum_{j=1}^{\infty} |a_{nj}|$ не могут быть ограниченными в совокупности.

3. В сумме (31) при больших n будут присутствовать большие по абсолютной величине коэффициенты, при этом соседние будут иметь противоположные знаки. Ввиду этого преобразование (31) при больших n очень часто связано с большой потерей значащих цифр, а это затрудняет применение преобразования.

7. Обращение преобразования. Полагая $p = 1, 2, \dots$, запишем преобразование (31) подробно

$$\begin{aligned} s_1 &= \sigma_1, \\ -c_2^1 s_1 + 2^2 c_3^2 s_2 &= 2! \sigma_2, \\ c_3^1 s_1 - 2^3 c_3^2 s_2 + 3^3 c_3^3 s_3 &= 3! \sigma_3, \\ &\dots \end{aligned}$$

Задача обращения преобразования есть задача решения бесконечной системы уравнений с треугольной матрицей. Она, очевидно, имеет единственное решение. Первое уравнение дает значение s_1 ; пользуясь им, из второго уравнения находим s_2 , выражая его через σ_1 и σ_2 . Из третьего уравнения, используя уже полученные s_1 и s_2 , находим s_3 , выразив его через $\sigma_1, \sigma_2, \sigma_3$ и т. д. Все эти вычисления могут быть проделаны и будет найдено явное выражение s_n с любым номером n через $\sigma_1, \dots, \sigma_n$. Оно оказывается очень простым:

$$s_n = \sum_{l=1}^n l \frac{n(n-1) \dots (n-l+1)}{n^{l+1}} \sigma_l \quad (n = 1, 2, \dots). \quad (34)$$

Убедиться в его правильности можно путем подстановки приведенных выражений s_n в уравнения (31)

$$\begin{aligned} L^*(s_n) &= \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} s_j = \\ &= \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \sum_{l=1}^j l \frac{j(j-1) \dots (j-l+1)}{j^{l+1}} \sigma_l. \end{aligned}$$

Если изменить порядок суммирования по индексам j и l , то получим

$$L^*(s_n) = \sum_{l=1}^n l \sigma_l \sum_{j=l}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \cdot \frac{j(j-1) \dots (j-l+1)}{j^{l+1}}.$$

Можно заменить нижнюю границу изменения индекса j и суммировать в границах $1 \leq j \leq n$; при этом сумма не изменится, так как все добавленные слагаемые равны нулю, например при $j=1$ обратится в нуль скобка $(j-1)$ и так далее до значения $j=l-1$, когда в нуль обращается скобка $(j-l+1)$.

Внутренняя сумма при этом станет результатом преобразования (12) переменной

$$s_n = \frac{n(n-1) \dots (n-l+1)}{n^l}. \quad (35)$$

В самом деле,

$$\begin{aligned} L^* \left(\frac{n(n-1) \dots (n-l+1)}{n^l} \right) &= \\ &= \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \cdot \frac{j(j-1) \dots (j-l+1)}{j^{l+1}}. \end{aligned} \quad (36)$$

Если в $j^{l-1}[j(j-1) \dots (j-l+1)]$ раскрыть скобки и расположить слагаемые по степеням j^{-1} , то получим многочлен от j^{-1} вида $\frac{1}{j} + \frac{a_2}{j^2} + \dots + (-1)^{l-1} (l-1)! \frac{1}{j^l}$. Поэтому правая часть (36) является суммой нескольких выражений вида (33), для которых $m=1, 2, \dots, l$, и при $l \leq n-1$ все такие суммы, как указано в (33), равны нулю. Поэтому в выражении для $L^*(s_n)$ при суммировании по l может быть отличным от нуля только слагаемое для $l=n$:

$$L^*(s_n) = n \sigma_n \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \cdot \frac{j(j-1) \dots (j-n+1)}{j^{n+1}}.$$

Так как

$$\begin{aligned} j^{-n-1}[j(j-1) \dots (j-n+1)] &= \\ &= \frac{1}{j} + \frac{a'_2}{j^2} + \dots + (-1)^{n-1} (n-1)! \frac{1}{j^n}, \end{aligned}$$

то среди сумм по индексу j ввиду (33) отличной от нуля может быть лишь сумма с последним членом $(-1)^{n-1} \times \times (n-1)! \frac{1}{j^n}$:

$$L^*(s_n) = n\sigma_n (-1)^{n-1} (n-1)! \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \cdot \frac{1}{j^n}. \quad (37)$$

Последняя сумма есть значение оператора L^* для последовательности $s_t = \frac{1}{t^n}$ и легко вычисляется*). Проинтерполируем s_t по значениям в узлах $t=1, 2, \dots, n$ с помощью многочлена $P_{n-1}\left(\frac{1}{t}\right)$ степени $n-1$ от $\frac{1}{t}$ и рассмотрим остаток интерполирования $r\left(\frac{1}{t}\right) = \frac{1}{t^n} - P_{n-1}\left(\frac{1}{t}\right)$. Это есть многочлен от $\frac{1}{t}$ степени n со старшим коэффициентом, равным единице. Корни $r\left(\frac{1}{t}\right)$ известны: это есть $\frac{1}{t} = \frac{1}{1}, \frac{1}{t} = \frac{1}{2}, \dots, \frac{1}{t} = \frac{1}{n}$, поэтому

$$r\left(\frac{1}{t}\right) = \frac{1}{t^n} - P_{n-1}\left(\frac{1}{t}\right) = \left(\frac{1}{t} - \frac{1}{1}\right)\left(\frac{1}{t} - \frac{1}{2}\right) \dots \left(\frac{1}{t} - \frac{1}{n}\right).$$

При неограниченном увеличении t $P_{n-1}\left(\frac{1}{t}\right) \rightarrow L^*\left(\frac{1}{t^n}\right)$.

Правая часть стремится к $(-1)^n \frac{1}{n!}$. Поэтому в пределе получится равенство

$$\begin{aligned} \sum_{j=1}^n (-1)^{n-j} \frac{j^n}{j! (n-j)!} \cdot \frac{1}{j^n} &= L^*\left(\frac{1}{j^n}\right) = \\ &= \lim_t P_{n-1}\left(\frac{1}{t}\right) = -(-1)^n \frac{1}{n!}. \end{aligned}$$

Из (37) найдем выражение для $L^*(s_n)$:

$$L^*(s_n) = n\sigma_n (-1)^{n-1} (n-1)! \left[-(-1)^n \frac{1}{n!} \right] = \sigma_n$$

*) Индекс n последовательности заменен на t , чтобы отличить его от обозначения n точки, в которой вычисляется значение оператора L^* .

и значения s_n , указанные в равенстве (34), действительно удовлетворяют уравнениям (31) и обращают, следовательно, это преобразование.

Сделаем несколько замечаний об обратном преобразовании (34). Оно является частным случаем общего линейного преобразования $s_n = \sum_{l=1}^{\infty} a_{nl} \sigma_l$, когда его коэффициенты имеют значения

$$a_{nl} = \begin{cases} l \frac{n(n-1) \dots (n-l+1)}{n^{l+1}}, & l \leq n, \\ 0, & l > n. \end{cases} \quad (38)$$

Отметим некоторые свойства коэффициентов:

1. $a_{nl} \geq 0$ при любых n, l .
2. При фиксированном значении l и $n \rightarrow \infty$ коэффициент a_{nl} является, как видно из верхнего равенства (38), величиной порядка $O\left(\frac{1}{n}\right)$.

3. Можно просто проверить, что сумма коэффициентов в преобразовании (34) равна единице при любых n . Действительно, если $s_n \equiv 1$, то, так как постоянная величина интерполируется многочленом от $1/n$ всегда точно, из (31) следует, что $\sigma_n \equiv 1$. Ввиду же однозначной обратимости этого преобразования для (34) можно утверждать, что из $\sigma_n \equiv 1$ вытекает $s_n \equiv 1$ и, следовательно,

$$B_n = \sum a_{nl} = \sum_{l=1}^n l \frac{n \dots (n-l+1)}{n^{l+1}} = 1 \quad (n = 1, 2, \dots).$$

Покажем, теперь, что преобразование (34) сохраняет предел. Для этого достаточно проверить выполнение условий теоремы 2 § 1 для (34).

Так как коэффициенты преобразования неотрицательны, то сумма их модулей A_n совпадает с суммой B_n ; последняя по третьему из указанных свойств равна единице:

$$A_n = \sum |a_{nl}| = \sum a_{nl} = B_n = 1.$$

Первое из условий теоремы 2 действительно выполнено. Ввиду $B_n = 1$ ($n = 1, 2, \dots$) второе условие также выполняется.

Второе из отмеченных свойств говорит о том, что для преобразования (34) выполняется и третье условие теоремы 2 § 1.

Этот результат позволяет для преобразования (31) сформулировать следующую теорему.

Теорема 4. *Если последовательность σ_n , полученная в результате преобразования (31), сходится и $\lim \sigma_n = \sigma$, то исходная последовательность s_n также будет сходящейся и при этом $\lim s_n = \lim \sigma_n = \sigma$.*

Скажем еще несколько слов о природе обратного преобразования (34). Так как его коэффициенты положительны и их сумма равна единице, то s_n есть не что иное, как среднее взвешенное значение*), составленное из $\sigma_1, \dots, \sigma_n$ с весами $m_l(n) = l \frac{n(n-1) \dots (n-l+1)}{n^{l+1}}$.

Выше мы обращали внимание на то, что преобразование (31) может переводить сходящиеся последовательности s_n в расходящиеся последовательности σ_n . Представляет интерес (главным образом теоретический) выяснить, какое множество составляют сходящиеся последовательности s_n , переходящие при преобразовании (31) в сходящиеся последовательности σ_n . Обратное преобразование и теорема 4 позволяют высказать приводимый ниже результат.

Теорема 5. *Если ко всем сходящимся последовательностям σ_n применить преобразования средних взвешенных (34), то получится множество последовательностей s_n , которые при преобразовании (31) переходят в сходящиеся последовательности.*

Последовательности s_n этого множества будут сходящимися, и для них преобразование (31) сохраняет не только сходимость, но и предел.

8. Теорема об ускорении сходимости. Укажем теперь класс последовательностей s_n , применительно к которым преобразование (31) дает ускорение сходимости. Как будет видно ниже, это ускорение будет сильным, но при

*) Среднее взвешенное величин $x_1, \dots, x_n$ с весами $m_1, \dots, m_n$ есть $\xi = \frac{m_1 x_1 + \dots + m_n x_n}{m_1 + \dots + m_n}$. Если x_i есть координаты масс m_i , то ξ есть координата центра инерции системы масс.

преобразовании будет, как правило, происходить значительная потеря точности.

Пусть s_n имеет аналитический тип сходимости и при больших n представляется сходящимся степенным рядом

$$s_n = c_0 + \frac{c_1}{n} + \frac{c_2}{n^2} + \dots + \frac{c_r}{n^r} + \dots, \quad n \geq N. \quad (39)$$

Здесь c_0 есть предельное значение s_n при $n \rightarrow \infty$: $c_0 = s$. Параллельно с s_n будем рассматривать аналитическую функцию комплексной переменной z , определив ее рядом

$$F(z) = c_0 + \frac{c_1}{z} + \frac{c_2}{z^2} + \dots$$

Она регулярна по меньшей мере в круге $|z| > N$ и значения ее в целых точках n совпадают с s_n :

$$F(n) = s_n, \quad n \geq N. \quad (40)$$

Напомним, что для построения σ_n мы интерполировали функцию s_t ($t = 1, 2, \dots$) по значениям $s_1, \dots, s_n$ посредством многочлена $P_{n-1}\left(\frac{1}{t}\right)$ степени $n-1$ от $1/t$ и полагали $\sigma_n = \lim_{t \rightarrow \infty} P_{n-1}\left(\frac{1}{t}\right)$, что привело нас к праву (31).

Нам нужно оценить погрешность $\sigma_n - s$. Чтобы воспользоваться аппаратом контурных интегралов в привычной форме, заменим переменную t , положив $t = 1/u$, $u = 1/t$, и введем функцию $\varphi(u) = F\left(\frac{1}{u}\right) = F(z) = c_0 + c_1 u + c_2 u^2 + \dots$. Она регулярна в круге $|u| < 1/N$ и связана с последовательностью s_n равенствами

$$\varphi\left(\frac{1}{n}\right) = s_n, \quad n \geq N.$$

Интерполирование s_n многочленом $P_{n-1}\left(\frac{1}{t}\right)$ по значениям s_t в узлах $1, 2, \dots, n$ совпадает с интерполированием $\varphi(u)$ по значениям в узлах $1, \frac{1}{2}, \dots, \frac{1}{n}$ с помощью многочлена $P_{n-1}(u)$. В частности, очевидно, $\sigma_n = \lim_{u \rightarrow 0} P_{n-1}(u) = P_{n-1}(0)$.

Рассмотрим точки $1, \frac{1}{2}, \frac{1}{3}, \dots$, они сгущаются к началу координат $u=0$, и лишь конечное число их может лежать вне круга $|u| < 1/N$. Пусть это будут точки $1, \frac{1}{2}, \dots, \frac{1}{m}$, так что все следующие точки $\frac{1}{m+1}, \frac{1}{m+2}, \dots$ принадлежат открытому кругу $|u| < 1/N$.

Возьмем произвольную достаточно гладкую замкнутую линию l , лежащую в круге $|u| < 1/N$ и содержащую отрезок $\left[0, \frac{1}{m+1}\right]$ внутри себя.

Составим интеграл *)

$$\frac{1}{2\pi i} \int_l \frac{\varphi(u)}{u\omega(u)} du, \quad (41)$$

$$\omega(u) = (u-1)\left(u-\frac{1}{2}\right) \dots \left(u-\frac{1}{n}\right)$$

и вычислим его, считая $n > m$. Он равен сумме вычетов интегрируемой функции в особых точках, лежащих внутри контура l . Числитель $\varphi(u)$ не имеет особенностей ни на самом контуре, ни внутри его. Поэтому особыми точками интегрируемой функции будут нули знаменателя $u\omega(u)$, находящиеся внутри l . Это будут точки $u=0$, $u = \frac{1}{n}, \dots, u = \frac{1}{m+1}$. Все они будут простыми полюсами функции $\frac{\varphi}{u\omega}$, и вычеты в них, найденные с помощью известных простых правил, имеют следующие значения:

$$\text{Res} \left[\frac{\varphi(u)}{u\omega(u)} \right]_{u=0} = \frac{\varphi(0)}{\omega(0)}.$$

Далее, так как $\omega\left(\frac{1}{j}\right) = 0$,

$$\text{Res} \left[\frac{\varphi(u)}{u\omega(u)} \right]_{u=1/j} = \frac{\varphi\left(\frac{1}{j}\right)}{\frac{1}{j} \omega'\left(\frac{1}{j}\right)}.$$

*) Он связан с интегральным представлением остатка интерполирования (см. п. 1.3.18), но эта связь сейчас несущественна.

Следовательно,

$$\begin{aligned} \frac{1}{2\pi i} \int_l \frac{\varphi(u)}{u\omega(u)} du &= \frac{\varphi(0)}{\omega(0)} + \sum_{j=m+1}^n \frac{\varphi\left(\frac{1}{j}\right)}{\frac{1}{j} \omega'\left(\frac{1}{j}\right)} = \\ &= \frac{1}{\omega(0)} \left[\varphi(0) + \sum_{j=m+1}^n \frac{\omega(0)}{\frac{1}{j} \omega'\left(\frac{1}{j}\right)} s_j \right] = \\ &= \frac{1}{\omega(0)} \left[\varphi(0) + \sum_{j=1}^n \frac{\omega(0)}{\frac{1}{j} \omega'\left(\frac{1}{j}\right)} s_j - \sum_{j=1}^m \frac{\omega(0)}{\frac{1}{j} \omega'\left(\frac{1}{j}\right)} s_j \right]. \end{aligned}$$

Приведенные ниже вычисления не требуют пояснений:

$$\varphi(0) = c_0 = s, \quad \omega(0) = (-1)^n \frac{1}{n!},$$

$$\begin{aligned} \omega'\left(\frac{1}{j}\right) &= \\ &= \left(\frac{1}{j} - 1\right) \left(\frac{1}{j} - \frac{1}{2}\right) \dots \left(\frac{1}{j} - \frac{1}{j-1}\right) \left(\frac{1}{j} - \frac{1}{j+1}\right) \dots \left(\frac{1}{j} - \frac{1}{n}\right) = \\ &= (-1)^{j-1} \frac{j! (n-j)!}{j^{n-1} n!}. \end{aligned}$$

Формула Лагранжа для интерполяционного многочлена $P_{n-1}(u)$ имеет вид

$$P_{n-1}(u) = \sum_{j=1}^n \frac{\omega(u)}{\left(u - \frac{1}{j}\right) \omega'\left(\frac{1}{j}\right)} s_j.$$

Она показывает, что первая из сумм в выражении, полученном для интеграла, есть значение $P_{n-1}(0) = \sigma_n$, взятое со знаком минус. Таким образом для интеграла будем иметь соотношение

$$\frac{1}{2\pi i} \int_l \frac{\varphi(u)}{u\omega(u)} du = (-1)^n n! \left[s - \sigma_n + \sum_{j=1}^m (-1)^{n-j} \frac{j^n}{j! (n-j)!} s_j \right];$$

отсюда

$$s - \sigma_n = - \sum_{j=1}^m (-1)^{n-j} \frac{j^n}{j! (n-j)!} s_j + \frac{(-1)^n}{n! 2\pi i} \int_l \frac{\varphi(u)}{u\omega(u)} du.$$

Оценим сначала первое слагаемое в правой части. Очевидна оценка

$$\frac{j^n}{j!(n-j)!} = \frac{j^n(n-j+1) \dots n}{j!n!} \leq \frac{j^n n^j}{j!n!} \leq \frac{m^n n^m}{n!}.$$

Если для $j = 1, 2, \dots, m$ верно неравенство $|s_j| \leq M$, то

$$\left| \sum_{j=1}^m (-1)^{n-j} \frac{j^n}{j!(n-j)!} s_j \right| \leq M m \frac{m^n n^m}{n!} = A \frac{m^n n^m}{n!}.$$

Для нашей цели достаточной является грубая оценка интеграла. Так как ни одна из точек $\frac{1}{j}$ ($j = 1, 2, \dots$) не лежит на линии l и эти точки сгущаются к $u=0$, то существует такое положительное число δ , что при всяком j будет $\left| u - \frac{1}{j} \right| \geq \delta$ и $|u| \geq \delta$. Поэтому

$$|\omega(u)| = \left| (u-1) \left(u - \frac{1}{2}\right) \dots \left(u - \frac{1}{n}\right) \right| \geq \delta^n;$$

ввиду ограниченности $v(u)$ на l существует число B такое, что

$$\left| \frac{1}{2\pi i} \int_l \frac{v(u)}{u\omega(u)} du \right| \leq \frac{B}{\delta^n}.$$

Таким образом, для разности $s - \sigma_n$ верна оценка

$$|s - \sigma_n| \leq A \frac{m^n n^m}{n!} + \frac{C}{n! \delta^n}. \quad (42)$$

Из проведенных рассуждений вытекает

Теорема 6. Если последовательность s_n имеет аналитический тип сходимости и $s_n \rightarrow s$, то существуют положительные числа A , C и δ такие, что для последовательности σ_n , получающейся из s_n при преобразовании (31), верна оценка (42).

Заметим теперь, что если s_n не приводится при больших n к постоянной величине, то в ряде (39) среди коэффициентов c_r ($r \geq 1$) существуют отличные от нуля. Пусть первый из них есть c_q и разложение (39) имеет вид

$$s_n = c_0 + \frac{c_q}{n^q} + \frac{c_{q+1}}{n^{q+1}} + \dots, \quad c_q \neq 0.$$

В этом случае, очевидно, для погрешности приближения $s - s_n = c_0 - s_n$ верно точное соотношение

$$|s - s_n| = O(n^{-q}).$$

Правая же часть оценки (42) ввиду быстрого возрастания $n!$ ($n \rightarrow \infty$) убывает быстрее, чем любая отрицательная степень n . Поэтому сходимость $s_n \rightarrow s$ будет много быстрее, чем $s_n \rightarrow s$.

§ 3. Улучшение сходимости последовательностей, для которых погрешность приближения к пределу близка к показательной функции или линейной комбинации таких функций

1. Простейшее правило преобразования и некоторые его свойства. Сначала рассмотрим самый простой случай, когда последовательность s_n мало отличается от показательной функции и представима в виде

$$s_n = s + Aq^n (1 + \varepsilon_n), \text{ где } \varepsilon_n \rightarrow 0 \text{ при } n \rightarrow \infty. \quad (1)$$

С такими последовательностями мы встречались в первой части книги, например при рассмотрении задачи ускорения сходимости при решении системы линейных уравнений или нахождения собственных значений и векторов методом итерации, при решении нелинейных уравнений тем же методом и в других вопросах.

Правила ускорения сходимости этих последовательностей указывались в весьма краткой форме, часто без достаточного выяснения наглядной стороны вопроса и без изучения свойств этих правил. Сейчас мы остановимся более подробно на правиле улучшения сходимости последовательностей вида (1). Начнем с рассмотрения показательной функции

$$s_n = s + Aq^n, \quad Aq \neq 0, \quad q \neq 1, \quad n = 1, 2, \dots \quad (2)$$

Величину s назовем постоянной частью показательной функции и Aq^n — переменной ее частью. Если $|q| < 1$, то, очевидно, $s_n \rightarrow s$ при $n \rightarrow \infty$. Когда же $|q| \geq 1$ и $q \neq 1$, последовательность s_n будет расходиться.

Независимо от значения q и, следовательно, независимо от сходимости или расходимости s_n , найдем выражение s через значения s_n . Для этого потребуется исклю-

чить из равенств (2) A и q и получить соотношение только между s и значениями s_n . Возьмем равенства (2), отвечающие трем смежным значениям $n-1$, n , $n+1$ аргумента n :

$$s_{n-1} - s = Aq^{n-1}, \quad s_n - s = Aq^n, \quad s_{n+1} - s = Aq^{n+1}.$$

Если второе из них разделить почленно на первое и третье — на второе, получатся два выражения для q , а приравнивание их даст уравнение для s

$$\frac{s_n - s}{s_{n-1} - s} = \frac{s_{n+1} - s}{s_n - s}.$$

Из него находим

$$s = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}}.$$

Это равенство позволяет при всяком n найти точно постоянную часть показательной функции (2), а, стало быть, и ее предел s , если он существует, по трем значениям s_n смежных номеров. Оно было положено в основу правила преобразования последовательностей: по любой заданной последовательности s_n ($n=1, 2, \dots$) составляют новую последовательность

$$\sigma_n = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}} \quad (n=2, 3, \dots). \quad (3)$$

Это преобразование часто называют *преобразованием Эйткена* или *δ^2 -преобразованием*.

Говорят, что преобразование (3) применимо к последовательности s_n , если для всех достаточно больших значений n знаменатель $\Delta^2 s_{n-1} = s_{n+1} - 2s_n + s_{n-1}$ отличен от нуля.

Условия, при которых преобразование (3) сохраняет предел, указываются в следующей простой теореме, дающей эти условия в форме, доступной для применения.

Теорема 1. Пусть последовательность s_n сходится к пределу s и к ней применимо преобразование (3). Для того чтобы при преобразовании сохранялся предел, т. е. чтобы $\sigma_n \rightarrow s$, необходимо и достаточно выполнение условия

$$\frac{(s_{n+1} - s_n)(s_n - s_{n-1})}{s_{n+1} - 2s_n + s_{n-1}} = \frac{\Delta s_n \Delta s_{n-1}}{\Delta^2 s_{n-1}} \rightarrow 0 \quad (n \rightarrow \infty). \quad (4)$$

Доказательство. Легко проверяется, что преобразование (3) может быть записано в виде

$$\delta_n = s_n - \frac{(s_{n+1} - s_n)(s_n - s_{n-1})}{s_{n+1} - 2s_n + s_{n-1}}. \quad (5)$$

Если $s_n \rightarrow s$, то для сходимости $\sigma_n \rightarrow s$ необходимо и достаточно выполнения условия (4).

Укажем еще, какую форму будет иметь δ^2 -преобразование для рядов. Пусть задан ряд

$$a_1 + a_2 + \dots + a_n + \dots \quad (6)$$

Возьмем последовательность его частичных сумм $s_n = a_1 + a_2 + \dots + a_n$ ($n = 1, 2, \dots$) и применим к ней преобразование (3). Если принять во внимание, что $s_{n+1} = s_n + a_{n+1}$ и $s_{n-1} = s_n - a_n$, для σ_n получим

$$\sigma_n = s_n - \frac{a_n a_{n+1}}{a_{n+1} - a_n} = a_1 + a_2 + \dots + a_n + \frac{a_n a_{n+1}}{a_n - a_{n+1}}. \quad (7)$$

Преобразование возможно, если $a_n - a_{n+1} \neq 0$, начиная с некоторого места в ряде.

Как показывают равенства (5) и (7), преобразование (3) сводится к тому, что к s_n добавляются дополнительные члены, имеющие простое строение: в случае последовательности такой член указан в (5), в случае же ряда — в (7), где он имеет тот же смысл, но более простую запись.

2. Об улучшении сходимости. Правила (3) и (5) построены так, что если переменная s_n имеет форму (2) и сходится к s , то эти преобразования при всяком n дают предельное значение s . Аналогично правило (7) при всяком n дает сумму ряда (6), если он есть сходящаяся геометрическая прогрессия $A + Aq + Aq^2 + \dots$. Простота дополнительных членов в (6) и (7) заставляет ожидать, что эти правила позволят ускорить сходимость последовательности и ряда, если они достаточно близки к (2) или к прогрессии соответственно. Поясним это более подробно на случае ряда.

Предположим, что ряд (6) сходится и имеет сумму s . Если ввести остаток ряда $r_n = a_{n+1} + a_{n+2} + \dots$ и заменить s_n на $s - r_n$, правило преобразования (7) можно записать в форме

$$\sigma_n = s - r_n + \frac{a_n a_{n+1}}{a_n - a_{n+1}}.$$

Так как $r_n \rightarrow 0$ ($n \rightarrow \infty$), условием сходимости σ_n к s будет следующее: $\frac{a_n a_{n+1}}{a_n - a_{n+1}} \rightarrow 0$ ($n \rightarrow \infty$). Если окажется, что величина $\rho_n = r_n - \frac{a_n a_{n+1}}{a_n - a_{n+1}}$ стремится к нулю при $n \rightarrow \infty$ быстрее, чем r_n , то новая последовательность σ_n будет сходиться к s быстрее, чем s_n . Это означает, что дробь $\frac{a_n a_{n+1}}{a_n - a_{n+1}}$ должна быть главной частью остатка r_n . Но r_n есть бесконечный ряд, указанный выше, а бесконечным рядом для дроби будет, вообще говоря, геометрическая прогрессия

$$\frac{a_n a_{n+1}}{a_n - a_{n+1}} = \frac{a_{n+1}}{1 - q_n} = a_{n+1} (1 + q_n + q_n^2 + \dots), \quad q_n = \frac{a_{n+1}}{a_n},$$

и чем больше n , тем меньшим должно быть различие между этими двумя рядами.

Получим теперь количественные признаки улучшения сходимости. Рассмотрим последовательность s_n , представляемую в форме (1). Преобразование (3) для нее даст следующий результат

$$\sigma_n = s + A q^{n+1} \frac{\varepsilon_{n+1} - 2\varepsilon_n + \varepsilon_{n-1} + \varepsilon_{n+1}\varepsilon_{n-1} - \varepsilon_n^2}{(1-q)^2 + q^2\varepsilon_{n+1} - 2q\varepsilon_n + \varepsilon_{n-1}}. \quad (8)$$

Так как $\varepsilon_n \rightarrow 0$ ($n \rightarrow \infty$), то знаменатель дроби будет стремиться к величине $(1-q)^2$, отличной от нуля ввиду $q \neq 1$, и при всех достаточно больших n будет отличным от нуля.

1. Когда $|q| < 1$, переменная s_n стремится к s , при этом разность $s_n - s$ будет величиной порядка $O(q^n)$. Новая последовательность σ_n , как вытекает из (8), также сходится к s и разность $\sigma_n - s$ имеет порядок малости $O(q^n \varepsilon_n)$. Это верно для любого закона стремления ε_n к нулю. Преобразование (3) в этом случае даст ускорение сходимости; при этом ускорение будет тем сильнее, чем быстрее ε_n стремится к нулю.

2. При $|q| = 1$ и $q \neq 1$ переменная s_n , оставаясь ограниченной, не имеет предела, тогда как σ_n будет ввиду $\varepsilon_n \rightarrow 0$ сходиться к s и сходимость здесь будет не медленнее, чем при стремлении ε_n к нулю.

3. Если $|q| > 1$, $A \neq 0$, то модуль s_n будет неограниченно возрастать, когда $n \rightarrow \infty$, поведение же σ_n зависит от свойств ε_n , в частности, если для ε_n выполняется

условие $q^n \varepsilon_n \rightarrow 0$ ($n \rightarrow \infty$), то σ_n сходится к s , причем погрешность приближения имеет порядок $\sigma_n - s = O(q^n \varepsilon_n)$.

Приведенные рассуждения позволяют сформулировать следующую теорему.

Теорема 2. Преобразование (3) применимо к последовательности (1) при всяких достаточно больших n и верны следующие утверждения:

1) При $|q| < 1$ полученная после преобразования последовательность σ_n сходится к s и преобразование ускоряет сходимость, так как $s_n - s = O(q^n)$, тогда как $\sigma_n - s = O(q^n \varepsilon_n)$;

2) Когда $|q| = 1$, $q \neq 1$, последовательность s_n не имеет предела, последовательность же σ_n сходится к s и $\sigma_n - s = O(\varepsilon_n)$;

3) Пусть $|q| > 1$ и переменная s_n , следовательно, неограниченно возрастает по модулю при $n \rightarrow \infty$. Если выполняется условие

$$q^n \varepsilon_n \rightarrow 0 \quad (n \rightarrow \infty), \quad (9)$$

то новая последовательность σ_n будет сходиться к s и при этом

$$\sigma_n - s = O(q^n \varepsilon_n).$$

Теорема показывает, что для последовательностей вида (1) преобразование (3) позволяет не только ускорять сходимость, когда s_n сходится, но и ввести обобщенный предел для s_n в случае расходимости s_n , если выполняется условие (9).

3. Применение δ^2 -преобразования к улучшению сходимости и аналитическому продолжению степенного ряда. Пусть функция $f(z)$ задана степенным рядом, сходящимся в круге радиуса R :

$$f(z) = c_0 + c_1 z + c_2 z^2 + \dots + c_n z^n + \dots, \quad |z| < R, \quad (10)$$

и пусть известно также, что на окружности круга сходимости $f(z)$ имеет единственную особую точку, являющуюся полюсом первого порядка.

В этом случае $f(z)$ представима в следующем виде:

$$f(z) = \frac{\alpha a}{a-z} + \varphi(z), \quad a = Re^{i\varphi}, \quad (11)$$

где функция $\varphi(z)$ регулярна в круге радиуса ρ большего, чем R , и может быть там разложена в степенной ряд

$$\varphi(z) = b_0 + b_1 z + b_2 z^2 + \dots, \quad |z| < \rho. \quad (12)$$

Ряд (10) для f есть сумма ряда (12) и геометрической прогрессии $\frac{\alpha a}{a-z} = \frac{\alpha}{1-\frac{z}{a}} = \sum_{v=0}^{\infty} \alpha a^{-v} z^v$:

$$f(z) = \sum_{v=0}^{\infty} c_v z^v = \sum_{v=0}^{\infty} (\alpha a^{-v} + b_v) z^v. \quad (13)$$

Рассмотрим конечную сумму последнего ряда, считая $z \neq a$ и $|z| < \rho$

$$\begin{aligned} s_n &= \sum_{v=0}^{n-1} (\alpha a^{-v} + b_v) z^v = \\ &= \frac{\alpha a}{a-z} - \frac{\alpha a}{a-z} \left(\frac{z}{a}\right)^n + \sum_{v=0}^{\infty} b_v z^v - \sum_{v=n}^{\infty} b_v z^v = \\ &= \frac{\alpha a}{a-z} + \varphi(z) - \frac{\alpha a}{a-z} \left(\frac{z}{a}\right)^n \left[1 + \frac{a-z}{\alpha a} \left(\frac{a}{z}\right)^n \sum_{v=n}^{\infty} b_v z^v\right]. \end{aligned}$$

Она имеет форму (1) при $s = \frac{\alpha a}{a-z} + \varphi(z) = f(z)$,

$$A = -\frac{\alpha a}{a-z}, \quad q = \frac{z}{a}, \quad \varepsilon_n = \frac{a-z}{\alpha a} \left(\frac{a}{z}\right)^n \sum_{v=n}^{\infty} b_v z^v.$$

Оценим ε_n . Для этой цели воспользуемся известной теоремой о коэффициентах степенного ряда: так как ряд (11) сходится в круге $|z| < \rho$, то для всякого $\delta > 0$ существует такое число M , что верно неравенство $|b_v| \leq M(\rho - \delta)^{-v}$. Будем считать δ настолько малой величиной, что $|z| < \rho - \delta$. Для ε_n получим следующее неравенство:

$$\begin{aligned} |\varepsilon_n| &\leq \left| \frac{a-z}{\alpha a} \right| \left| \frac{a}{z} \right|^n \cdot \sum_{v=n}^{\infty} M(\rho - \delta)^{-v} |z|^v = \\ &= M \left| \frac{a-z}{\alpha a} \right| \frac{\rho - \delta}{\rho - \delta - |z|} \cdot \frac{|a|^n}{(\rho - \delta)^n}. \end{aligned}$$

Применим теперь к последовательности частичных сумм s_n δ^2 -преобразование. В результате получим новую последовательность σ_n вида (8) с указанными несколькими строками выше значениями s , A , q и ε_n . Произведение $q^n \varepsilon_n$ в рассматриваемом вопросе имеет следующую оценку:

$$|q^n \varepsilon_n| = \left| \frac{z}{a} \right|^n |\varepsilon_n| \leq M \left| \frac{a-z}{\alpha a} \right| \frac{\rho - \delta}{\rho - \delta - |z|} \cdot \frac{|z|^n}{(\rho - \delta)^n}. \quad (14)$$

Отсюда ввиду того, что $|z| < \rho - \delta$, сразу же следует соотношение $|q^n \varepsilon_n| \rightarrow 0$, а из теоремы 2 и из выражения (8) для σ_n вытекает, что дополнительный к s член в правой части (8) стремится к нулю не медленнее, чем $|z|^n (\rho - \delta)^{-n}$. Поэтому σ_n при всяком значении z из круга $|z| < \rho$ и $z \neq a$ сходится к $s = f(z)$. При этом, если проследить оценки, можно без труда увидеть, что сходимость $\sigma_n \rightarrow f(z)$ будет равномерной относительно z в любой замкнутой области, лежащей в круге $|z| < \rho$ и не содержащей точку $z = a$.

Когда точка z принадлежит кругу $|z| < R$, ряд (10) сходится. Главной частью его остатка $r_n = \sum_{v=n}^{\infty} (\alpha a^{-v} z^v + b_v z^v)$ будет более медленно сходящаяся часть $\sum_{v=n}^{\infty} \alpha a^{-v} z^v = \frac{\alpha a}{a-z} \left(\frac{z}{a} \right)^n$. Вторая группа слагаемых имеет больший радиус сходимости и, следовательно, члены ее убывают быстрее с ростом n , чем в главной части. При $n \rightarrow \infty$ остаток стремится к нулю и имеет порядок малости $O\left[\left(\frac{|z|}{R}\right)^n\right]$. Разность $\sigma_n - f(z)$ имеет, как видно из оценки (14), порядок малости $O\left[\left(\frac{|z|}{\rho - \delta}\right)^n\right]$ более высокий чем r_n , и применение преобразования (3) к s_n дает значительное ускорение сходимости. Когда же точка z принадлежит кольцу $[R \leq |z| < \rho, z \neq a]$, то ряд (10) расходится и последовательность σ_n сходится к аналитическому продолжению степенного ряда (10).

4. Связь с методом интерполирования предельного значения. Пусть дана произвольная последовательность s_n ($n = 1, 2, \dots$). Рассмотрим показательную функцию $f(n) = B + Aq^n$, содержащую три параметра B , A , q .

Выберем их так, чтобы в трех узлах $n = p-1, p, p+1$ функция f принимала такие же значения, как и s_n : $f(p-1) = s_{p-1}$, $f(p) = s_p$, $f(p+1) = s_{p+1}$.

Для нахождения значений параметров будем иметь следующую систему уравнений:

$$B + Aq^{p-1} = s_{p-1}, \quad B + Aq^p = s_p, \quad B + Aq^{p+1} = s_{p+1}.$$

При ее решении за неизвестные проще принять $B, Aq^p = u, q$. Система примет вид: $B + \frac{1}{q}u = s_{p-1}$, $B + u = s_p$, $B + qu = s_{p+1}$. Она решается просто. Значения параметров, найденные из нее, зависят от значения p (это отмечено индексом (p) , поставленным около них):

$$B^{(p)} = \frac{s_{p+1}s_{p-1} - s_p^2}{s_{p+1} - 2s_p + s_{p-1}}, \quad u^{(p)} = (Aq^p)^{(p)} = \frac{(s_{p+1} - s_p)(s_p - s_{p-1})}{s_{p+1} - 2s_p + s_{p-1}},$$

$$q^{(p)} = \frac{s_{p+1} - s_p}{s_p - s_{p-1}}.$$

Подстановка этих выражений в $f(n)$ приводит к следующему выражению для интерполирующей функции:

$$f_p(n) = \frac{1}{s_{p+1} - 2s_p + s_{p-1}} \times$$

$$\times \left[s_{p+1}s_{p-1} - s_p^2 + (s_{p+1} - s_p)(s_p - s_{p-1}) \left(\frac{s_{p+1} - s_p}{s_p - s_{p-1}} \right)^{n-p} \right]. \quad (15)$$

Построение f_p возможно при выполнении условий

$$s_{p+1} - 2s_p + s_{p-1} \neq 0, \quad s_p - s_{p-1} \neq 0. \quad (16)$$

Будем считать, что неравенства соблюдаются для всех достаточно больших p .

Устремим теперь n к ∞ . Предельное значение у $f_p(n)$ существует лишь при условии

$$\left| \frac{s_{p+1} - s_p}{s_p - s_{p-1}} \right| < 1. \quad (17)$$

Считая условие (17) выполненным, примем $\lim_n f_p(n)$ за элемент новой последовательности и обозначим его через σ_p :

$$\sigma_p = \lim_n f_p(n) = B^{(p)} = \frac{s_{p+1}s_{p-1} - s_p^2}{s_{p+1} - 2s_p + s_{p-1}}.$$

Это правило преобразования такое же, как (3), но получение его потребовало больших условий, чем для (3).

5. О пути уточнения δ^2 -преобразования. В рассмотренном выше множестве последовательностей (1) погрешность приближения s_n к пределу s изменялась по закону, близкому к геометрической прогрессии или показательному закону

$$s_n - s = Aq^n (1 + \varepsilon_n).$$

При построении преобразования (3), предназначенного для улучшения сходимости s_n к s , считалась известной структура только главной части Aq^n погрешности; при этом параметры ее A и q предполагались любыми.

В некоторых случаях бывает известной структура не только главных частей, но и нескольких следующих по порядку малости частей погрешности. Тогда естественно изменить преобразование (3) так, чтобы учесть не только главные, но и следующие известные части.

Законы изменения погрешностей, строение их и, в частности, разложение их налагаемые с последовательно возрастающими скоростями убывания, зависят от вида задачи и метода ее решения. Последние могут быть многообразными и уточнять методы ускорения их сходимости можно в многих отношениях. На примере мы поясним причины, лежащие в основе обобщения метода (3), которое будет излагаться ниже. Этот пример связан с переходом от метода итерации для одного численного уравнения к применению его для решения систем уравнений и с переходом от одной последовательности к системе последовательностей.

Предварительно рассмотрим простую задачу о решении методом итерации линейной системы. Запишем систему в виде

$$x = b + Ax,$$

где x и b есть k -мерные векторы и A есть квадратная $k \times k$ -матрица. После выбора начального приближения x^0 к решению, все дальнейшие приближения находятся по правилу

$$x^n = b + Ax^{n-1}, \quad n = 1, 2, \dots \quad (18)$$

Пусть x^* есть точное решение системы

$$x^* = b + Ax^*. \quad (19)$$

Рассмотрим погрешность приближения номера n : $\delta^n = \mathbf{x}^n - \mathbf{x}^*$. Для нее верно соотношение, получающееся при почленном вычитании двух предшествующих равенств:

$$\delta^n = A\delta^{n-1}.$$

Если это соотношение применить n раз, то для δ^n получится выражение через погрешность δ^0 начального значения

$$\delta^n = A^n \delta^0.$$

Предположим, что матрица A может быть подобным преобразованием приведена к диагональному виду

$$A = S [\lambda_1, \lambda_2, \dots, \lambda_k] S^{-1}.$$

Тогда *)

$$A^n = S [\lambda_1^n, \lambda_2^n, \dots, \lambda_k^n] S^{-1}$$

и

$$\delta^n = S [\lambda_1^n, \lambda_2^n, \dots, \lambda_k^n] S^{-1} \delta^0.$$

Поэтому для приближения $\mathbf{x}^n$ к решению $\mathbf{x}^*$ верно представление

$$\mathbf{x}^n = \mathbf{x}^* + \delta^n = \mathbf{x}^* + S [\lambda_1^n, \lambda_2^n, \dots, \lambda_k^n] S^{-1} \delta^0.$$

Отсюда вытекает, что каждая составляющая x_i^n ($i = 1, \dots, k$) вектора $\mathbf{x}^n$ является функцией следующего вида:

$$x_i^n = x_i^* + a_{i1}\lambda_1^n + \dots + a_{ik}\lambda_k^n \quad (i = 1, 2, \dots, k). \quad (20)$$

Параметры x_i^* , a_{i1} , $\dots$, a_{ik} , λ_1 , $\dots$, λ_k таких представлений нам неизвестны, так как они либо зависят от $\mathbf{x}^*$, либо требуют вычислений, превосходящих по сложности решение линейной системы, и в последующих рассуждениях их нужно считать произвольными.

Для простоты будем считать, что собственные значения λ_i матрицы A различных по модулю и занумерованы в порядке убывания модулей

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_k|.$$

*) Здесь и ниже запись λ_m^n означает степень числа λ_m .

В правой части равенства (20) главным членом при больших n будет сумма $x_i^* + a_{i1}\lambda_1^n$, если только $a_{i1} \neq 0$, и равенство может быть записано так, как указано в (1):

$$x_i^n = x_i^* + a_{i1}\lambda_1^n(1 + \varepsilon_i^n), \quad \varepsilon_i^n = \frac{a_{i2}}{a_{i1}}\left(\frac{\lambda_2}{\lambda_1}\right)^n + \dots \quad (21)$$

Преобразование (3) учитывало только постоянный член x_i^* переменной x_i^n и первый член $a_{i1}\lambda_1^n$ переменной части, все остальные считались произвольными. В рассматриваемой задаче нам известно строение не только первого члена погрешности, но и строение всех остальных ее членов. Этим можно воспользоваться при создании правила улучшения сходимости, применяя для этого все члены с известной структурой или часть их. К каким уточнениям это приводит, будет выяснено в следующем пункте.

Выше был приведен очень простой пример, показывающий возможность и желательность уточнения δ^2 -преобразования. Но существуют другие задачи, в которых главный и дополнительные члены погрешности имеют сходное строение и где построенные ниже правила ускорения сходимости будут иметь применения.

6. Правило ускорения сходимости последовательности, для которой погрешность приближения к пределу близка к линейной комбинации показательных функций. Сначала определим само преобразование. Пусть s_n ($n = 1, 2, \dots$) есть произвольная последовательность. Введем два связанных с ней определителя

$$D_{kn}(s_n) = \begin{vmatrix} 1 & 1 & \dots & 1 \\ \Delta s_{n-k} & \Delta s_{n-k+1} & \dots & \Delta s_n \\ \Delta s_{n-k+1} & \Delta s_{n-k+2} & \dots & \Delta s_{n+1} \\ \dots & \dots & \dots & \dots \\ \Delta s_{n-1} & \Delta s_n & \dots & \Delta s_{n+k-1} \end{vmatrix}, \quad \Delta s_n = s_{n+1} - s_n,$$

$$D_{kn}^*(s) = \begin{vmatrix} s_{n-k} & s_{n-k+1} & \dots & s_n \\ \Delta s_{n-k} & \Delta s_{n-k+1} & \dots & \Delta s_n \\ \Delta s_{n-k+1} & \Delta s_{n-k+2} & \dots & \Delta s_{n+1} \\ \dots & \dots & \dots & \dots \\ \Delta s_{n-1} & \Delta s_n & \dots & \Delta s_{n+k-1} \end{vmatrix}.$$

С их помощью построим преобразование последовательности s_n в новую последовательность σ_n , положив

$$\sigma_n = \sigma_{kn} = \frac{D_{kn}^*(s_n)}{D_{kn}(s_n)}, \quad n = k+1, k+2, \dots \quad (22)$$

Ниже это правило будем называть S_k -преобразованием или *преобразованием Шенкса* порядка k . Условием его применимости является неравенство $D_{kn} \neq 0$. Будем говорить, что преобразование (22) применимо к последовательности s_n , если для всяких достаточно больших n $D_{kn} \neq 0$.

Для $k=1$ правило (22) совпадает с δ^2 -преобразованием:

$$\sigma_{1n} = \frac{\begin{vmatrix} s_{n-1} & s_n \\ \Delta s_{n-1} & \Delta s_n \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ \Delta s_{n-1} & \Delta s_n \end{vmatrix}} = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}} = \delta^2 s_n.$$

При $k=2$ преобразование (22) имеет вид:

$$\begin{aligned} \sigma_{2n} &= \frac{\begin{vmatrix} s_{n-2} & s_{n-1} & s_n \\ \Delta s_{n-2} & \Delta s_{n-1} & \Delta s_n \\ \Delta s_{n-1} & \Delta s_n & \Delta s_{n+1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ \Delta s_{n-2} & \Delta s_{n-1} & \Delta s_n \\ \Delta s_{n-1} & \Delta s_n & \Delta s_{n+1} \end{vmatrix}} = \\ &= \frac{s_{n-2} [\Delta s_{n-1} \Delta s_{n+1} - (\Delta s_n)^2] + s_{n-1} [\Delta s_{n-1} \Delta s_n - \Delta s_{n-2} \Delta s_{n+1}] +}{\Delta^2 s_{n-2} \Delta^2 s_n - (\Delta^2 s_{n-1})^2} + \\ &\quad + \frac{s_n [\Delta s_{n-2} \Delta s_n - (\Delta s_{n-1})^2]}{\Delta^2 s_{n-2} \Delta^2 s_n - (\Delta^2 s_{n-1})^2}. \quad (23) \end{aligned}$$

Оно является простейшим уточнением δ^2 -преобразования.

Выясним теперь, для какого множества сходящихся последовательностей можно наверное при применении преобразования (22) ожидать ускорения сходимости. Для этого предварительно рассмотрим следующую комбинацию показательных функций:

$$s_n = s + c_1 q_1^n + c_2 q_2^n + \dots + c_k q_k^n \quad (n = 1, 2, \dots). \quad (24)$$

Основания q_i ($i = 1, \dots, k$) показательных функций предполагаются различными, отличными от нуля и единицы. Коэффициенты c_i можно считать отличными от нуля.

Если все q_i будут по модулю меньше единицы, величина s_n при $n \rightarrow \infty$ будет стремиться к своей постоянной

появятся вторые разности, в первой же строке всюду, кроме первого места, получатся нули:

$$D_{kn} = \begin{vmatrix} 1 & 0 & \dots & 0 \\ \Delta s_{n-k} & \Delta^2 s_{n-k} & \dots & \Delta^2 s_{n-1} \\ \dots & \dots & \dots & \dots \\ \Delta s_{n-1} & \Delta^2 s_{n-1} & \dots & \Delta^2 s_{n+k-2} \end{vmatrix} = \begin{vmatrix} \Delta^2 s_{n-k} & \dots & \Delta^2 s_{n-1} \\ \dots & \dots & \dots \\ \Delta^2 s_{n-1} & \dots & \Delta^2 s_{n+k-2} \end{vmatrix}.$$

Так как последовательность s_n имеет форму (24), то для нее справедливо равенство

$$\Delta^2 s_n = s_{n+2} - 2s_{n+1} + s_n = \sum_{i=1}^k c_i q_i^n (q_i - 1)^2 = \sum_{i=1}^k c'_i q_i^n, \\ c'_i = c_i (q_i - 1)^2.$$

Воспользуемся правилом умножения определителей «строка на строку»:

$$D_{kn} = \begin{vmatrix} \sum_{i=1}^k c'_i q_i^{n-k} & \dots & \sum_{i=1}^k c'_i q_i^{n-1} \\ \dots & \dots & \dots \\ \sum_{i=1}^k c'_i q_i^{n-1} & \dots & \sum_{i=1}^k c'_i q_i^{n+k-2} \end{vmatrix} = \\ = \begin{vmatrix} c'_1 q_1^{n-k} & \dots & c'_k q_k^{n-k} \\ \dots & \dots & \dots \\ c'_1 q_1^{n-1} & \dots & c'_k q_k^{n-1} \end{vmatrix} \cdot \begin{vmatrix} 1 & \dots & 1 \\ q_1 & \dots & q_k \\ \dots & \dots & \dots \\ q_1^{k-1} & \dots & q_k^{k-1} \end{vmatrix}.$$

Если из столбцов первого из перемножаемых определителей вынести множители соответственно $c'_1 q_1^{n-k} = c_1 (q_1 - 1)^2 q_1^{n-k} \dots c'_k q_k^{n-k} = c_k (q_k - 1)^2 q_k^{n-k}$, то получится определитель Вандермонда $W_k(q_1, \dots, q_k)$ такой же, как второй множитель и

$$D_{kn} = c_1 \dots c_k (q_1 - 1)^2 \dots (q_k - 1)^2 q_1^{n-k} \dots q_k^{n-k} W_k^2(q_1, \dots, q_k). \quad (31)$$

По предположению, $c_i \neq 0$, $q_i \neq 0$, $q_i \neq 1$, $q_i \neq q_j$ ($i \neq j$) и, следовательно, $D_{kn} \neq 0$.

Поэтому из (26) для s получим

$$s = \frac{D_{kn}^*(s_n)}{D_{kn}(s_n)}. \quad (32)$$

Сравнение найденного значения s с преобразованием (22)

позволяет сказать, что если последовательность s_n имеет вид (24), то при всяком значении $n = k + 1, k + 2, \dots$ преобразование дает постоянную часть s переменной (24), каковы бы ни были c_i, q_i , удовлетворяющие указанным выше условиям.

Напомним, что при доказательстве этого результата мы предполагали основания q_i показательных функций в (24) различными между собой. Если же среди q_i будут одинаковые, то определитель Вандермонда $W_k(q_1, \dots, q_k)$ в (31), а вместе с ним и D_{kn} обратятся в нуль, и преобразование (22) к такой последовательности s_n будет неприменимым. В этом случае для получения результата, аналогичного указанному выше, необходимо изменить вид последовательности (24). Сейчас мы сделаем краткое замечание о том, какой должна быть она взята в случае совпадения некоторых q_i .

Рассмотрим последовательность вида

$$s_n = s + c_1 z_n^{(1)} + c_2 z_n^{(2)} + \dots + c_k z_n^{(k)},$$

где функции $z_n^{(i)}$ оставим на время произвольными линейно независимыми. Выясним, каким условиям они должны удовлетворять, чтобы в линейных комбинациях вида

$$L(s_{m-k}) = a_k s_{m-k} + a_{k-1} s_{m-k+1} + \dots + a_1 s_{m-1} + s_m, \\ m = n, \dots, n+k,$$

стоящих слева в равенствах (26) и аналогичных ему, получающихся при замене n на $n+1, \dots, n+k$, коэффициенты при параметрах c_i обращались в нуль и эти параметры в $L(s_{m-k})$ не входили. Так как

$$L(s_{m-k}) = s(a_k + a_{k-1} + \dots + 1) + \\ + \sum_{i=1}^k c_i (a_k z_{m-k}^{(i)} + a_{k-1} z_{m-k+1}^{(i)} + \dots + z_m^{(i)}) = \\ = s(a_k + a_{k-1} + \dots + 1) + \sum_{i=1}^k c_i L(z_{m-k}^{(i)}),$$

то для исключения c_i достаточно выбрать функции $z_m^{(i)}$ так, чтобы $L(z_{m-k}^{(i)}) = 0$ ($i = 1, \dots, k; m = k+1, k+2, \dots$). Последнее говорит о том, что $z_m^{(i)}$ должны быть решениями

однородного линейного разностного уравнения порядка k с постоянными коэффициентами

$$a_k s_{m-k} + a_{k-1} s_{m-k+1} + \dots + a_1 s_{m-1} + s_m = 0.$$

Его решения, как известно *), определяются характеристическим уравнением, имеющим в нашем случае вид

$$a_k + a_{k-1}q + a_{k-2}q^2 + \dots + a_1q^{k-1} + q^k = 0.$$

Пусть оно имеет l различных корней $q_1, \dots, q_l$ кратностей соответственно $k_1, \dots, k_l$. Для кратностей, очевидно, выполняется соотношение $k_1 + \dots + k_l = k$. Корню q_i отвечает k_i независимых частных решений

$$q_i^n, nq_i^n, \dots, n^{k_i-1}q_i^n \quad (i = 1, \dots, l).$$

Общее число таких решений равно $k_1 + \dots + k_l = k$, и они образуют полную систему линейно независимых решений уравнения. Примем их за $z_n^{(i)}$. Соответствующая им последовательность s_n будет иметь вид

$$s_n = s + \sum_{i=1}^l \left(c_1^{(i)} + c_2^{(i)} n + \dots + c_{k_i}^{(i)} n^{k_i-1} \right) q_i^n, \quad (33)$$

$$k_1 + \dots + k_l = k; \quad q_i \neq 1, 0; \quad c_{k_i}^{(i)} \neq 0.$$

Если коэффициенты $a_1, a_2, \dots, a_k$ определить равенством

$$(q - q_1)^{k_1} \dots (q - q_l)^{k_l} = q^k + a_1 q^{k-1} + \dots + a^k,$$

то в линейных комбинациях $L(s_{m-k})$ будут отсутствовать члены с коэффициентами $c_j^{(i)}$ ($j = 1, \dots, k_i$; $i = 1, \dots, l$) и будут верны соотношения

$$\begin{aligned} L(s_{m-k}) &= a_k s_{m-k} + a_{k-1} s_{m-k+1} + \dots + a_1 s_{m-1} + s_m = \\ &= s(a_k + \dots + a_1 + 1), \quad m = n, n+1, \dots, n+k, \end{aligned}$$

из которых сразу же следуют системы (28) и (29). Из них в свою очередь путем исключения коэффициентов $a_1, \dots, a_k$, зависящих от q_i ($i = 1, \dots, l$), получится равенство (30) для нахождения s через значения $s_{n-k}, \dots, s_{n+k}$.

Выше для переменной s_n , имеющей вид (24), было найдено простое явное выражение (31) для D_{kn} , позволившее убедиться в том, что $D_{kn} \neq 0$. Когда s_n имеет вид (33), явное выражение для D_{kn} является значительно

*) С таким уравнением мы встречались в п. 5.10.1 (см. уравнение (5.10.9)).

более сложным, и его получение потребовало бы громоздких вычислений. Убедиться в выполнении неравенства можно было бы сравнительно просто, если воспользоваться некоторыми результатами теории конечноразностных уравнений, но изложение их также потребовало бы некоторого места и отвлекло бы от основной темы. Примем без доказательства, что для последовательности (33) будет $D_{kn} \neq 0$.

Тогда из (30) для s следует выражение (32). Это приводит к формулируемой ниже теореме, частным случаем которой является утверждение, высказанное раньше для последовательности (24).

Теорема 3. *К последовательности s_n вида (33) преобразование (22) применимо при любом $n = k + 1, k + 2, \dots$ и для каждого n дает постоянную часть s последовательности s_n .*

Отметим, что если дана последовательность s_n , имеющая вид (33), то для определения преобразования (22), обладающего указанным в теореме 3 свойством, необходимо знать только число $k = k_1 + \dots + k_l$, и не нужно знания q_i, k_i и других параметров формулы (33).

7. Теорема об ускорении сходимости. Если последовательности (24) или (33) сходятся, то s для них является предельным значением и преобразование (22) при всех значениях n дает это предельное значение, так как тогда по теореме 3 $\sigma_n = s$ ($n = k + 1, k + 2, \dots$). Это есть крайний случай ускорения сходимости при преобразовании, когда предельное значение получается сразу.

Если последовательность s_n не совпадает с (33), а в каком-то смысле близка к такому виду, то преобразование (22) даст величину σ_n , не совпадающую с s , а мало отличающуюся от s , и можно ожидать, что новая последовательность σ_n будет сходиться быстрее к пределу s , чем к нему сходится s_n . Нам надлежит выяснить, какого рода «близость» s_n к (33) для этого достаточна. Начнем с выяснения достаточных условий сходимости σ_n и ограничимся простым случаем, когда последовательность s_n близка к (22). Рассмотрим последовательность вида

$$s_n = s + c_1 q_1^n + \dots + c_k q_k^n + \varepsilon_n = s + z_n + \varepsilon_n, \quad (34)$$

$$c_i \neq 0, \quad q_i \neq 0; 1 \quad (i = 1, \dots, k), \quad c_i \neq c_j \quad (i \neq j).$$

Будем предполагать, что величина ε_n является «малой» сравнительно с показательными функциями и для нее выполняются условия

$$\frac{\varepsilon_n}{q_i^n} \rightarrow 0, \quad n \rightarrow \infty \quad (i = 1, \dots, k). \quad (35)$$

Выделим в определителях $D_{k,n}(s_n)$ и $D_{k,n}^*(s_n)$ главные части. Возьмем $D_{k,n}$ и преобразуем его, вычитая из последнего столбца предпоследний и т. д. и первый столбец из второго. Это позволит привести D_{kn} к виду, содержащему только вторые разности

$$D_{kn} = \begin{vmatrix} 1 & 0 & \dots & 0 \\ \Delta s_{n-k} & \Delta^2 s_{n-k} & \dots & \Delta^2 s_{n-1} \\ \dots & \dots & \dots & \dots \\ \Delta s_{n-1} & \Delta^2 s_{n-1} & \dots & \Delta^2 s_{n+k-2} \end{vmatrix} = \begin{vmatrix} \Delta^2 s_{n-k} & \dots & \Delta^2 s_{n-1} \\ \dots & \dots & \dots \\ \Delta^2 s_{n-1} & \dots & \Delta^2 s_{n+k-2} \end{vmatrix}.$$

Ввиду

$$\begin{aligned} \Delta^2 s_m &= \Delta^2 z_m + \Delta^2 \varepsilon_m = \sum_{j=1}^k c_j q_j^m (q_j^2 - 2q_j + 1) + \Delta^2 \varepsilon_m = \\ &= \sum_{j=1}^k c'_j q_j^m + \eta_m, \quad c'_j = c_j (q_j - 1)^2, \quad \eta_m = \Delta^2 \varepsilon_m \end{aligned}$$

справедлива формула

$$D_{kn} = \begin{vmatrix} \sum_{j=1}^k c'_j q_j^{n-k} + \eta_{n-k} & \dots & \sum_{j=1}^k c'_j q_j^{n-1} + \eta_{n-1} \\ \dots & \dots & \dots \\ \sum_{j=1}^k c'_j q_j^{n-1} + \eta_{n-1} & \dots & \sum_{j=1}^k c'_j q_j^{n+k-2} + \eta_{n+k-2} \end{vmatrix}.$$

Каждый столбец в D_{kn} можно рассматривать как сумму k столбцов с элементами $(c'_j q_j^{n-i}, \dots, c'_j q_j^{n-i+k})$, которые будем называть столбцами первого вида, и одного столбца второго вида с элементами $(\eta_{n-i}, \dots, \eta_{n-i+k-1})$. Поэтому D_{kn} можно представить как сумму $(k+1)^k$ определителей того же порядка k , в каждом из которых столбцами будут отдельные слагаемые столбцы этих сумм. Приведем запись суммы к простейшей форме. Возьмем один из слагаемых определителей. В нем будет несколько столбцов первого вида. Их всегда можно переставить на

первые места; при этом может измениться знак определителя, и если такие перестановки выполнить во всех слагаемых определителях, получим для D_{kn} разложение следующего вида:

$$D_{kn} = \sum \pm \begin{vmatrix} c'_a q_a^{n-\alpha} & \dots & c'_d q_d^{n-\gamma} \eta_{n-l} & \dots & \eta_{n-m} \\ c'_a q_a^{n-\alpha+k-1} & \dots & c'_d q_d^{n-\gamma+k-1} \eta_{n-l+k-1} & \dots & \eta_{n-m+k-1} \end{vmatrix}.$$

Из столбцов первого вида можно вынести общие множители $c'_a q_a^{n-\alpha}, \dots, c'_d q_d^{n-\gamma}$. Отделим в них множители $q_a^n, \dots, q_d^n$, зависящие от n , а остальные объединим в постоянный множитель c' :

$$D_{kn} = \sum \pm c' q_a^n \dots q_d^n \begin{vmatrix} 1 & \dots & 1 & \eta_{n-l} & \dots & \eta_{n-m} \\ q_a & \dots & q_d & \eta_{n-l+1} & \dots & \eta_{n-m+1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ q_a^{k-1} & \dots & q_d^{k-1} & \eta_{n-l+k-1} & \dots & \eta_{n-m+k-1} \end{vmatrix}.$$

Если среди индексов $a, \dots, d$ есть одинаковые, то соответствующий определитель равен нулю, так как в нем будут два одинаковых столбца, и при суммировании следует принимать во внимание лишь слагаемые с разными индексами $a, \dots, d$.

Разделим D_{kn} на произведение $q_1^n \dots q_k^n$. При почленном делении суммы множитель $q_a^n \dots q_d^n$ сократится. Оставшееся после сокращения произведение обозначим через $q_e^n \dots q_h^n$. Число оставшихся множителей будет равно числу столбцов второго вида в соответствующем слагаемом определителе. Внесем множители $q_e^{-n}, \dots, q_h^{-n}$ под знак определителя и присоединим их по одному к элементам каждого столбца второго вида:

$$\begin{aligned} D_{kn} q_1^{-n} \dots q_k^{-n} &= \\ &= \sum \pm c' \begin{vmatrix} 1 & \dots & 1 & q_e^{-n} \eta_{n-l} & \dots & q_h^{-n} \eta_{n-m} \\ q_a & \dots & q_d & q_e^{-n} \eta_{n-l+1} & \dots & q_h^{-n} \eta_{n-m+1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ q_a^{k-1} & \dots & q_d^{k-1} & q_e^{-n} \eta_{n-l+k-1} & \dots & q_h^{-n} \eta_{n-m+k-1} \end{vmatrix}. \end{aligned} \quad (36)$$

Выясним, какое значение имеет сумма, стоящая справа при больших n . Для определенности будем считать, что основания q_i перенумерованы в порядке убывания модулей:

$$|q_1| \geq |q_2| \geq \dots \geq |q_k|.$$

Для величин ε_n выполняются условия (35), и так как $\eta_n = \varepsilon_{n+2} - 2\varepsilon_{n+1} + \varepsilon_n$, то для η_n будут выполнены аналогичные условия:

$$q_i^{-n} \eta_n \rightarrow 0 \quad (n \rightarrow \infty).$$

Каждый определитель, стоящий под знаком суммы и имеющий хотя бы один столбец второго вида, будет стремиться к нулю при $n \rightarrow \infty$ и является величиной порядка $o(q_k^{-n} \varepsilon_n)$. Число таких определителей не зависит от n , и их сумма с коэффициентами c' будет малой величиной такого же порядка.

Рассмотрим теперь сумму членов в (36), в которых определители не содержат столбцов второго вида. Это будет разложение определителя, получающегося из $D_{kn} q_1^{-n} \dots q_k^{-n}$, если в D_{kn} заменить ε_n тождественным нулем. Такой определитель без множителя $q_1^{-n} \dots q_k^{-n}$ рассмотрен в предыдущем пункте перед равенством (31) и его значение указано в (31). Все изложенное позволяет сказать, что равенство (36) при больших n имеет форму

$$D_{kn} q_1^{-n} \dots q_k^{-n} = c_1 \dots c_k (q_1 - 1)^2 \dots (q_k - 1)^2 q_1^{-k} \dots \\ \dots q_k^{-k} W^2(q_1, \dots, q_k) [1 + o(q_k^{-n} \varepsilon_n)]$$

и, следовательно, при больших n

$$D_{kn} = c_1 \dots c_k (q_1 - 1)^2 \dots (q_k - 1)^2 q_1^{n-k} \dots \\ \dots q_k^{n-k} W_k^2(q_1, \dots, q_k) [1 + o(q_k^{-n} \varepsilon_n)]. \quad (37)$$

Сходные вычисления проделаем для определителя D_{kn}^* :

$$D_{kn}^*(s_n) = \begin{vmatrix} s + z_{n-k} + \varepsilon_{n-k} & \dots & s + z_n + \varepsilon_n \\ \Delta s_{n-k} & \dots & \Delta s_n \\ \dots & \dots & \dots \\ \Delta s_{n-1} & \dots & \Delta s_{n+k-1} \end{vmatrix} = \\ = s D_{kn} + \begin{vmatrix} z_{n-k} + \varepsilon_{n-k} & \dots & z_n + \varepsilon_n \\ \Delta(z_{n-k} + \varepsilon_{n-k}) & \dots & \Delta(z_n + \varepsilon_n) \\ \dots & \dots & \dots \\ \Delta(z_{n-1} + \varepsilon_{n-1}) & \dots & \Delta(z_{n+k-1} + \varepsilon_{n+k-1}) \end{vmatrix} = \\ = s D_{kn} + D_{kn}^*(z_n + \varepsilon_n).$$

Последний из определителей $-D_{kn}^*(z_n + \varepsilon_n)$ — приведем к более удобной для преобразований форме. Первую строку прибавим ко второй, после чего на месте второй строки получим $z_{n-k+1} + \varepsilon_{n-k+1} \dots z_{n+1} + \varepsilon_{n+1}$. Новую вторую строку прибавим к третьей и т. д. После этого получим

$$D_{kn}^*(z_n + \varepsilon_n) = \begin{vmatrix} z_{n-k} + \varepsilon_{n-k} & \dots & z_n + \varepsilon_n \\ \dots & \dots & \dots \\ z_n + \varepsilon_n & \dots & z_{n+k} + \varepsilon_{n+k} \end{vmatrix} =$$

$$= \begin{vmatrix} \sum_{j=1}^k c_j q_j^{n-k} + \varepsilon_{n-k} & \dots & \sum_{j=1}^k c_j q_j^n + \varepsilon_n \\ \dots & \dots & \dots \\ \sum_{j=1}^k c_j q_j^n + \varepsilon_n & \dots & \sum_{j=1}^k c_j q_j^{n+k} + \varepsilon_{n+k} \end{vmatrix}.$$

Каждый столбец последнего определителя есть сумма $k+1$ столбцов, и подобно тому, как это сделано выше для D_{kn} , $D_{kn}^*(z_n + \varepsilon_n)$ можно представить в форме суммы $(k+1)^{k+1}$ определителей порядка $k+1$ следующего вида:

$$D_{kn}^*(z_n + \varepsilon_n) =$$

$$= \sum \pm \begin{vmatrix} c_a q_a^{n-\alpha} & \dots & c_d q_d^{n-\gamma} & \dots & \varepsilon_{n-l} & \dots & \varepsilon_{n-m} & \varepsilon_{n-p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ c_a q_a^{n-\alpha+k} & \dots & c_d q_d^{n-\gamma+k} & \dots & \varepsilon_{n-l+k} & \dots & \varepsilon_{n-m+k} & \varepsilon_{n-p} \end{vmatrix}.$$

Если из столбцов определителей, стоящих под знаком суммы, вынести множители $c_a q_a^{n-\alpha} \dots, c_d q_d^{n-\gamma}$, отделить в них части $q_a^n, \dots, q_d^n$, зависящие от n , а остальные множители объединить в один, не зависящий от n , и обозначить его через C , то получим следующее представление:

$$D_{kn}^*(z_n + \varepsilon_n) =$$

$$= \sum \pm C q_a^n \dots q_d^n \begin{vmatrix} 1 & \dots & 1 & \varepsilon_{n-l} & \dots & \varepsilon_{n-m} \varepsilon_{n-p} \\ q_a & \dots & q_d & \varepsilon_{n-l+1} & \dots & \varepsilon_{n-m+1} \varepsilon_{n-p+1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ q_a^k & \dots & q_d^k & \varepsilon_{n-l+k} & \dots & \varepsilon_{n-m+k} \varepsilon_{n-p+k} \end{vmatrix}.$$

Рассмотрим сначала определители, в которых нет столбцов со значениями ε_n . В них будет $k+1$ столбцов с величинами $q_a \dots q_d$, и так как среди этих величин может быть только k различных, то все такие определители равны нулю ввиду того, что в них обязательно будут одинако-

вые столбцы. Отличными от нуля могут быть только определители, имеющие столбцы со значениями ε_n . Кроме того, если в таких определителях некоторые из чисел $q_a, \dots, q_d$ окажутся равными, определитель обратится в нуль, и мы вправе считать $q_a, \dots, q_d$ различными.

Обе части последнего равенства разделим на произведение $q_1^n \dots q_k^n$. При почленном делении суммы множители $q_a^n, \dots, q_d^n$ сократятся нацело, при этом оставшихся степеней q , которые назовем $q_l^n, \dots, q_h^n$, будет на единицу меньше числа столбцов со значениями величины ε . Внесем множители $q_l^{-n}, \dots, q_h^{-n}$ под знак определителя и присоединим по одному к столбцам с ε , кроме последнего, который останется свободным:

$$q_1^{-n} \dots q_k^{-n} D_{kn}^* (z_n + \varepsilon_n) = \\ = \sum \pm C \begin{vmatrix} 1 & \dots & 1 & q_l^{-n} \varepsilon_{n-l} & \dots & q_h^{-n} \varepsilon_{n-m} & \varepsilon_{n-p} \\ q_a & \dots & q_d & q_l^{-n} \varepsilon_{n-l+1} & \dots & q_h^{-n} \varepsilon_{n-m+1} & \varepsilon_{n-p+1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ q_a^k & \dots & q_d^k & q_l^{-n} \varepsilon_{n-l+k} & \dots & q_h^{-n} \varepsilon_{n-m+k} & \varepsilon_{n-p+k} \end{vmatrix}.$$

Среди определителей, стоящих под знаком суммы, наиболее низкую по порядку малости оценку будут иметь те, где значения ε стоят только в одном столбце (последнем в нашей записи), и такие определители должны быть оценены величиной $O(\varepsilon_n)$. Если в определителе будет несколько столбцов со значениями ε , то порядок малости их оценок будет не ниже чем $O(\varepsilon_n) \left(\frac{\varepsilon_n}{q_k^n} \right)$. Поэтому правая

часть последнего равенства при больших n будет величиной $O(\varepsilon_n)$ и для $D_{kn}(z_n + \varepsilon_n)$ получится соотношение

$$D_{kn}^* (z_n + \varepsilon_n) = q_1^n \dots q_k^n O(\varepsilon_n). \quad (38)$$

Возвратимся к преобразованию (22)

$$\sigma_n = \frac{D_{kn}^* (s_n)}{D_{kn}} = \frac{D_{kn}^* (s + z_n + \varepsilon_n)}{D_{kn}} = s + \frac{D_{kn}^* (z_n + \varepsilon_n)}{D_{kn}}.$$

На основании равенств (37) и (38) можно сказать, что

$$\sigma_n = s + O(\varepsilon_n), \quad (39)$$

и верна

Теорема 4. Если последовательность имеет форму (34) и для нее выполняется условие (35), то при всяких достаточно больших n к ней применимо преобразование (22); при этом новая последовательность σ_n , полученная из s_n после преобразования, сходится к значению s , если $\varepsilon_n \rightarrow 0$, когда $n \rightarrow \infty$, причем $\sigma_n - s = O(\varepsilon_n)$.

Допустим вновь, что последовательность s_n имеет вид (34) и для нее выполняются условия (35). Пусть, кроме того, s_n сходится. Последнее может, очевидно, быть лишь в том случае, когда все основания q_i по модулю меньше единицы: $1 > |q_1| \geq |q_2| \geq \dots \geq |q_k| > 0$. Пределом последовательности является число s . Скорость стремления к нулю погрешности $s_n - s = c_1 q_1^n + \dots + c_k q_k^n + \varepsilon_n$ определяется наиболее медленно убывающей показательной функцией, так что $s_n - s = O(q_1^n)$. Применим к s_n теорему 4. Условие $\varepsilon_n \rightarrow 0$ ($n \rightarrow \infty$) в рассматриваемом случае выполнено, так как из $\frac{\varepsilon_n}{q_k^n} \rightarrow 0$ следует, что $\varepsilon_n = o(q_k^n)$.

Соотношение (39) здесь имеет вид

$$\sigma_n - s = O(\varepsilon_n) = o(q_k^n). \quad (40)$$

Преобразование s_n в σ_n может дать сильное увеличение скорости сходимости.

Изложенное позволяет сформулировать приводимую ниже теорему.

Теорема 5. Пусть для последовательности s_n вида (34) выполнены условия (35) и $1 > |q_1| \geq |q_2| \geq \dots \geq |q_k| > 0$. Тогда верны утверждения:

1) s_n сходится к s и скорость сходимости характеризуется равенством

$$s_n - s = O(q_1^n);$$

2) к s_n при всех достаточно больших n применимо преобразование (22), новая последовательность σ_n также сходится к s и ускорение сходимости, получающееся при преобразовании, характеризуется равенством (40).

Аналоги теорем 4 и 5 могут быть доказаны для последовательностей вида (33), но они требуют более громоздких вычислений и по этой причине опущены.

8. Об улучшении сходимости и аналитическом продолжении степенного ряда мероморфной функции. Пусть аналитическая функция $f(z)$ комплексной переменной z задана степенным рядом

$$f(z) = c_0 + c_1 z + \dots + c_n z^n + \dots, \quad (41)$$

сходящемся в круге $|z| < R$. Пусть, кроме того, известно, что $f(z)$ в круге большего радиуса $|z| < \rho$ ($\rho > R$) имеет конечное число особых точек, которые являются полюсами функции f . Рассмотрим первоначально простейший случай, когда все полюсы имеют первый порядок. Тогда $f(z)$ представима в форме

$$f(z) = \sum_{j=1}^k \frac{\alpha_j a_j}{a_j - z} + \varphi(z), \quad (42)$$

$$a_i \neq a_j \quad (i \neq j), \quad R \leq |a_j| < \rho \quad (j = 1, \dots, k),$$

где $\varphi(z)$ регулярна в круге $|z| < \rho$ и может быть, следовательно, разложена в степенной ряд

$$\varphi(z) = \sum_{v=0}^{\infty} b_v z^v, \quad |z| < \rho. \quad (43)$$

Ряд (42) является суммой k степенных рядов для элементарных функций $\alpha_j a_j (a_j - z)^{-1}$ и ряда (43)

$$\sum_{v=0}^{\infty} c_v z^v = \sum_{j=1}^k \alpha_j \sum_{v=0}^{\infty} \left(\frac{z}{a_j}\right)^v + \sum_{v=0}^{\infty} b_v z^v.$$

В этом представлении функции f нам известны лишь коэффициенты c_v , число k и то, что ряд (42) сходится в круге $|z| < \rho$. Считая $|z| < \rho$, рассмотрим частичную сумму заданного ряда

$$\begin{aligned} s_n &= \sum_{v=0}^{n-1} c_v z^v = \sum_{j=1}^k \alpha_j \sum_{v=0}^{n-1} \left(\frac{z}{a_j}\right)^v + \sum_{v=0}^{n-1} b_v z^v = \\ &= \sum_{j=1}^k \alpha_j \left[\frac{a_j}{a_j - z} - \frac{a_j}{a_j - z} \left(\frac{z}{a_j}\right)^n \right] + \sum_{v=0}^{\infty} b_v z^v - \sum_{v=n}^{\infty} b_v z^v = \\ &= \sum_{j=1}^k \frac{\alpha_j a_j}{a_j - z} + \varphi(z) - \sum_{j=1}^k \frac{\alpha_j a_j}{a_j - z} \left(\frac{z}{a_j}\right)^n - \sum_{v=n}^{\infty} b_v z^v. \end{aligned}$$

Величина s_n имеет форму (34) при

$$s = \sum_{j=1}^k \frac{\alpha_j a_j}{a_j - z} + \varphi(z), \quad c_j = -\frac{\alpha_j a_j}{a_j - z},$$

$$q_j = \frac{z}{a_j}, \quad \varepsilon_n = -\sum_{v=n}^{\infty} b_v z^v.$$

Проверим выполнение условия (35). Так как ряд (42) сходится в круге $|z| < \rho$, то для всякого $\delta > 0$ существует такое число M , что выполняется неравенство $|b_v| \leq M |\rho - \delta|^{-v}$. Будем считать δ настолько малым, что

$$|z| < \rho - \delta, \quad |a_j| < \rho - \delta,$$

$$|\varepsilon_n| \leq \sum_{v=n}^{\infty} M \frac{|z|^v}{(\rho - \delta)^v} = M \frac{\rho - \delta}{\rho - \delta - |z|} \cdot \left| \frac{z}{\rho - \delta} \right|^n.$$

Поэтому

$$\left| \frac{\varepsilon_n}{q_j} \right|^n \leq M \frac{\rho - \delta}{\rho - \delta - |z|} \cdot \left| \frac{z}{\rho - \delta} \right|^n \cdot \left| \frac{a_j}{z} \right|^n = M \frac{\rho - \delta}{\rho - \delta - |z|} \left| \frac{a_j}{\rho - \delta} \right|^n.$$

Отсюда ввиду $|a_j| < \rho - \delta$ следует соотношение (35). Наконец, последнее условие теоремы 4 $\varepsilon_n \rightarrow 0$ ($n \rightarrow \infty$), очевидно, выполняется, так как ряд (43) в точке z сходится.

Теорема 4 применима к последовательности частичных сумм s_n ряда (41), и новая последовательность σ_n , полученная из s_n преобразованием (22), сходится к

$$\sum_{j=1}^k \frac{\alpha_j a_j}{a_j - z} + \varphi(z)$$

при всяких z из круга $|z| < \rho$, отличных от a_j ($j = 1, \dots, k$), и предел σ_n дает аналитическое продолжение суммы степенного ряда (41) в этом круге.

Если z лежит в круге $|z| < R$, то последовательность σ_n по теореме 5 будет сходиться к s , причем скорость сходимости характеризуется соотношением $\sigma_n - s = O(\varepsilon_n) =$

$= O\left(\sum_{v=n}^{\infty} b_v z^v\right)$, и, так как ряд (43) сходится в круге $|z| < \rho$ ($\rho > R$), будет больше, чем у заданного ряда (41).

Случаем простых полюсов функции $f(z)$ мы ограничились для облегчения записи. Без доказательства отметим, что справедлива более общая теорема.

Теорема 6. Пусть функция $f(z)$, заданная степенным рядом (41), имеет внутри круга $|z| < \rho$ ($\rho > R$) конечное число особых точек $a_1, \dots, a_l$, являющихся полюсами, и известно, что сумма порядков этих полюсов равна k . Тогда преобразование (22) применимо к последовательности частичных сумм s_n степенного ряда (41). Новая последовательность σ_n будет сходиться к $f(z)$ всюду в круге $|z| < \rho$, кроме точек a_i ($i = 1, \dots, l$); при этом сходимость σ_n к $f(z)$ в круге $|z| < R$ будет более быстрой, чем сходимость s_n к $f(z)$.

§ 4. Улучшение сходимости ряда с помощью выделения медленно сходящейся части

1. Введение. Пусть рассматривается сходящийся численный или функциональный ряд

$$u_1 + u_2 + \dots + u_n + \dots \quad (1)$$

Если сходимость является недостаточно быстрой, можно попытаться ускорить ее с помощью выделения из ряда медленно сходящейся части. Достигают этого обычно путем разложения заданного ряда на сумму двух сходящихся рядов

$$\sum_{n=1}^{\infty} u_n = \sum_{n=1}^{\infty} v_n + \sum_{n=1}^{\infty} w_n. \quad (2)$$

Эти ряды выбирают, чтобы первый из них имел известную сумму или легко находимую, второй же ряд должен быть более быстро сходящимся, чем (1).

Члены рядов связаны между собой очевидным соотношением: $u_n = v_n + w_n$. Так как третий ряд с членами w_n должен сходиться быстрее ряда (1), то его члены w_n должны стремиться к нулю быстрее u_n и величина v_n при неограниченном росте n должна быть главной частью u_n . Разложение (2) есть в первую очередь операция выделения из членов u_n главных частей при условии $n \rightarrow \infty$. Такое выделение может быть сделано многими способами и разные значения v_n могут отличаться между собой лишь малыми величинами более высокого порядка малости, чем они сами. Произволом в выборе v_n пользуются обычно для того, чтобы сумма ряда $\sum v_n$

Второй из рядов сходится тем быстрее, чем больше k . Что же касается первого ряда, то его сумма выражается точно через значения дзета-функции Римана

$$\sum_{n=1}^{\infty} v_n = \sum_{n=1}^{\infty} \left(\frac{a_0}{n^\mu} + \frac{a_1}{n^{\mu+1}} + \dots + \frac{a_k}{n^{\mu+k}} \right) = \\ = a_0 \zeta(\mu) + a_1 \zeta(\mu+1) + \dots + a_k \zeta(\mu+k).$$

Численные значения этого выражения могут быть во многих случаях найдены с помощью известных таблиц функций *) $\zeta(x)$.

В случае, когда μ есть целое число, большее единицы, вычисления упрощаются и могут быть выполнены без обращения к функции $\zeta(x)$.

Заменим разложение по отрицательным степеням n другим разложением — по обратным величинам факториальных многочленов.

Напомним, что *факториальным многочленом степени p* называют следующую функцию:

$$\Pi_p(x) = x(x+1)\dots(x+p-1). \quad (5)$$

Присоединяя множитель $n^{-\mu}$ к $F(n)$, можно считать, что рассматривается ряд

$$\sum_{n=1}^{\infty} F(n), \quad (6)$$

где функция $F(n)$ имеет представление

$$F(n) = \frac{a_2}{n^2} + \frac{a_3}{n^3} + \dots + \frac{a_k}{n^k} + \frac{\gamma_k(n)}{n^k}.$$

Такое разложение может быть заменено близким к нему разложением

$$F(n) = \frac{b_2}{n(n+1)} + \frac{b_3}{n(n+1)(n+2)} + \dots \\ \dots + \frac{b_k}{n(n+1)\dots(n+k-1)} + \delta_k(n), \quad (7) \\ \delta_k(n) \rightarrow 0 \quad (n \rightarrow \infty).$$

*) См., например, А. В. Лебедев, Р. М. Федорова, Справочник по математическим таблицам, М., Изд-во АН СССР, 1956, гл. XI; Н. М. Брунова, Справочник... Добавление № 1, М., Изд-во АН СССР, 1959.

Его параметры b_i и $\delta_k(n)$ определяются равенствами

$$b_2 = \lim_n n(n+1)F(n),$$

$$b_3 = \lim_n n(n+1)(n+2) \left[F(n) - \frac{b_2}{n(n+1)} \right],$$

.....

$$b_k = \lim_n n(n+1) \dots$$

$$\begin{aligned} & \dots (n+k-1) \left[F(n) - \left(\frac{b_2}{n(n+1)} + \dots + \frac{b_{k-1}}{n(n+1) \dots (n+k-2)} \right) \right], \\ & \frac{\delta_k(n)}{n(n+1) \dots (n+k-1)} = \\ & = F(n) - \left(\frac{b_2}{n(n+1)} + \dots + \frac{b_k}{n(n+1) \dots (n+k-1)} \right). \end{aligned}$$

Равенство (7) позволяет разделить ряд (6) на две части

$$\begin{aligned} \sum_{n=1}^{\infty} u_n &= \sum_{n=1}^{\infty} F(n) = \\ &= \sum_{n=1}^{\infty} \left(\frac{b_2}{n(n+1)} + \frac{b_3}{n(n+1)(n+2)} + \dots \right. \\ & \quad \left. \dots + \frac{b_k}{n(n+1) \dots (n+k-1)} \right) + \sum_{n=1}^{\infty} \frac{\delta_k(n)}{n(n+1) \dots (n+k-1)}. \end{aligned}$$

Если число k достаточно велико, то второй из рядов правой части будет быстро сходящимся. Сумма первого ряда вычисляется точно. Если обратить внимание на то, что

$$\begin{aligned} & \frac{1}{n(n+1) \dots (n+m-1)} = \\ & = \left[\frac{1}{n(n+1) \dots (n+m-2)} - \frac{1}{(n+1)(n+2) \dots (n+m-1)} \right] \frac{1}{m-1} \end{aligned}$$

и, следовательно,

$$\begin{aligned} & \sum_{n=1}^{\infty} \frac{1}{n(n+1) \dots (n+m-1)} = \\ & = \frac{1}{m-1} \sum_{n=1}^{\infty} \left[\frac{1}{n(n+1) \dots (n+m-2)} - \frac{1}{(n+1) \dots (n+m-1)} \right] = \\ & = \frac{1}{(m-1)(m-1)!}, \end{aligned}$$

то можно сказать, что сумма первого ряда равна

$$\frac{b_2}{1 \cdot 1!} + \frac{b_3}{2 \cdot 2!} + \dots + \frac{b_k}{(k-1)(k-1)!}. \quad (8)$$

3. Улучшение сходимости тригонометрического ряда.

Как пример задачи ускорения сходимости функциональных рядов будет рассмотрена задача для сходящегося тригонометрического ряда

$$S(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \quad (-\infty < x < \infty). \quad (9)$$

Ввиду сходимости ряда его коэффициенты a_n и b_n стремятся к нулю при $n \rightarrow \infty$. Предположим, что при больших значениях n коэффициенты имеют следующие представления:

$$\begin{aligned} a_n &= \frac{\alpha_1}{n} + \frac{\alpha_2}{n^2} + \dots + \frac{\alpha_k}{n^k} + \frac{1}{n^k} \varepsilon_k(n), \quad \varepsilon_k(n) \rightarrow 0, \\ b_n &= \frac{\beta_1}{n} + \frac{\beta_2}{n^2} + \dots + \frac{\beta_k}{n^k} + \frac{1}{n^k} \delta_k(n), \quad \delta_k(n) \rightarrow 0. \end{aligned} \quad (10)$$

Если эти выражения для a_n и b_n внести в ряд (9), то его можно разложить на сумму нескольких слагаемых рядов

$$\begin{aligned} S(x) &= a_0 + \left[\alpha_1 \sum_{n=1}^{\infty} \frac{\cos nx}{n} + \beta_1 \sum_{n=1}^{\infty} \frac{\sin nx}{n} \right] + \\ &+ \left[\alpha_2 \sum_{n=1}^{\infty} \frac{\cos nx}{n^2} + \beta_2 \sum_{n=1}^{\infty} \frac{\sin nx}{n^2} \right] + \dots \\ &\dots + \left[\alpha_k \sum_{n=1}^{\infty} \frac{\cos nx}{n^k} + \beta_k \sum_{n=1}^{\infty} \frac{\sin nx}{n^k} \right] + \\ &+ \sum_{n=1}^{\infty} \frac{1}{n^k} [\varepsilon_k(n) \cos nx + \delta_k(n) \sin nx]. \end{aligned}$$

Последний из них будет сходиться тем быстрее, чем больше число k . Что же касается всех других рядов

правой части равенства, то суммы их известны в теории рядов Фурье и значения сумм приводятся ниже *).

$$\begin{aligned}
 & \text{I} \left\{ \begin{aligned} \sum_{n=1}^{\infty} \frac{\sin nx}{n} &= \frac{\pi - x}{2}, & 0 < x < 2\pi, \\ \sum_{n=1}^{\infty} \frac{\cos nx}{n} &= -\ln\left(2 \sin \frac{x}{2}\right) = \psi(x), & 0 < x < 2\pi, \end{aligned} \right. \\
 & \text{II} \left\{ \begin{aligned} \sum_{n=1}^{\infty} \frac{\cos nx}{n^2} &= \frac{(x - \pi)^2}{4} - \frac{\pi^2}{12}, & 0 \leq x \leq 2\pi, \\ \sum_{n=1}^{\infty} \frac{\sin nx}{n^2} &= -\int_0^x \ln\left(2 \sin \frac{t}{2}\right) dt = \int_0^x \psi(t) dt, & 0 \leq x \leq 2\pi, \end{aligned} \right. \\
 & \text{III} \left\{ \begin{aligned} \sum_{n=1}^{\infty} \frac{\sin nx}{n^3} &= \frac{x^3 - 3\pi x^2 + 2\pi^2 x}{12}, & 0 \leq x \leq 2\pi, \\ \sum_{n=1}^{\infty} \frac{\cos nx}{n^3} &= -\int_0^x (x - t) \psi(t) dt + \sum_{n=1}^{\infty} \frac{1}{n^3}, & 0 \leq x \leq 2\pi, \\ \sum_{n=1}^{\infty} \frac{1}{n^3} &= \zeta(3) = 1,202\,056\,903 \dots, \end{aligned} \right. \\
 & \text{IV} \left\{ \begin{aligned} \sum_{n=1}^{\infty} \frac{\cos nx}{n^4} &= \frac{x^4 - 4\pi x^3 + 4\pi^2 x^2}{48} + \frac{\pi^4}{90}, & 0 \leq x \leq 2\pi, \\ \sum_{n=1}^{\infty} \frac{\sin nx}{n^4} &= -\frac{1}{2} \int_0^x (x - t)^2 \psi(t) dt + x \sum_{n=1}^{\infty} \frac{1}{n^3}, & 0 \leq x \leq 2\pi, \end{aligned} \right. \\
 & \dots \dots \dots
 \end{aligned}$$

Выше был рассмотрен основной случай, когда в тригонометрическом ряду могут содержаться гармоники

*) Первую пару равенств можно найти, например, в книге: Г. М. Фихтенгольц, Курс дифференциального и интегрального исчисления, т. III, М., «Наука», 1970, гл. XIX, § 2. Каждая следующая пара равенств может быть получена из предыдущей при помощи почленного интегрирования по переменному отрезку $[0, x]$. Последовательность равенств этим путем может быть продолжена сколь угодно далеко.

любых кратных частот. В частных случаях, когда, например, тригонометрический ряд содержит гармоники только нечетных кратностей, вычисления могут, очевидно, быть упрощены путем построения равенств, аналогичных приведенным выше, но содержащим только нечетные значения n .

ЛИТЕРАТУРА

1. Крылов В. И., Бобков В. В., Монастырный П. И., Вычислительные методы высшей математики, т. 2, Минск, «Вышэйшая школа», 1975, гл. 9.
2. Харди Г., Расходящиеся ряды, М., ИЛ, 1951.

ГЛАВА 10

ПРИМЕНЕНИЕ ФУНКЦИОНАЛЬНОГО АНАЛИЗА К ПОСТРОЕНИЮ ТЕОРИИ НЕКОТОРЫХ РАЗДЕЛОВ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

В вычислительных методах математики к середине текущего столетия было накоплено обширное количество знаний и построен большой арсенал методов численного решения научных и технических задач. После появления быстродействующих вычислительных машин количество задач, поддающихся численному решению, увеличилось во много раз. Было создано много новых методов, учитывающих свойства новых задач и свойства машин. Вычислительной математике грозила опасность стать собранием большого числа частных приемов вычислений, недостаточно объединенных по идеям и не всегда теоретически полно обоснованных.

Потребность в упорядочении и идейном объединении методов вычислений стала настоятельно ощущаться. Такую возможность объединить и получать из общего источника много частных, родственных по своему характеру результатов, дал функциональный анализ, в частности общая теория операторов. К настоящему времени в этом направлении сделано очень многое. Из большого числа полученных здесь результатов авторы избрали только три вопроса, считая, что их достаточно для того, чтобы сделать понятной полезность абстрактных разделов математики в задаче объединения и упрощения разрозненных знаний.

§ 1. Необходимые сведения из функционального анализа

1. Метрические пространства. Сходимость и полнота. Множество X произвольных элементов x называется метрическим пространством, если каждой паре элементов x, y из X соответствует число $\rho(x, y)$, которое называют

расстоянием между x и y и которое удовлетворяет следующим условиям, называемым *аксиомами метрики*:

1) $\rho(x, y) \geq 0$ и $\rho(x, y) = 0$ в том и только в том случае, когда $x = y$;

2) $\rho(x, y) = \rho(y, x)$;

3) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

Первое из условий называется *аксиомой различения*, второе — *аксиомой симметрии* и третье — *аксиомой треугольника*.

Говорят, что последовательность элементов x_n ($n = 1, 2, \dots$) сходится к элементу x^* , если $\rho(x_n, x^*) \rightarrow 0$ ($n \rightarrow \infty$). Это записывается либо в форме $x_n \rightarrow x^*$, либо в форме $\lim x_n = x^*$.

1. Если $x_n \rightarrow x^*$ и $y_n \rightarrow y^*$, то $\rho(x_n, y_n) \rightarrow \rho(x^*, y^*)$.

Действительно, двукратное применение аксиомы треугольника позволяет написать:

$$\begin{aligned} \rho(x_n, y_n) &\leq \rho(x_n, x^*) + \rho(x^*, y_n) \leq \\ &\leq \rho(x_n, x^*) + \rho(x^*, y^*) + \rho(y^*, y_n). \end{aligned}$$

Поэтому

$$\rho(x_n, y_n) - \rho(x^*, y^*) \leq \rho(x_n, x^*) + \rho(y^*, y_n).$$

Если поменять местами пары элементов x_n, y_n и x^*, y^* , то получится неравенство с другим знаком левой части и, стало быть,

$$|\rho(x^*, y^*) - \rho(x_n, y_n)| \leq \rho(x_n, x^*) + \rho(y^*, y_n).$$

Поэтому из того, что $\rho(x_n, x^*) \rightarrow 0$ и $\rho(y_n, y^*) \rightarrow 0$, следует $\rho(x_n, y_n) \rightarrow \rho(x^*, y^*)$.

Таким образом, расстояние $\rho(x, y)$ есть непрерывная функция своих аргументов x и y .

2. Если $x_n \rightarrow x^*$ и $x_n \rightarrow y^*$, то $x^* = y^*$, т. е. сходящаяся последовательность не может иметь два различных предела.

В самом деле, $0 = \rho(x_n, x_n) \rightarrow \rho(x^*, y^*)$. Следовательно, $\rho(x^*, y^*) = 0$ и $x^* = y^*$.

Нам потребуется еще понятие полноты пространства. Всякая сходящаяся последовательность x_n выполняет условие Больцано — Коши, так как, если $x_n \rightarrow x$, то для

всякого $\varepsilon > 0$ существует номер n_0 такой, что при $n \geq n_0$ будет $\rho(x_n, x) < \frac{1}{2} \varepsilon$. Возьмем $n > n_0$ и $m > n_0$. Тогда

$$\rho(x_n, x_m) \leq \rho(x_n, x) + \rho(x, x_m) \leq \frac{1}{2} \varepsilon + \frac{1}{2} \varepsilon = \varepsilon.$$

Обратное утверждение, вообще говоря, не всегда верно, так как можно указать такое метрическое пространство, для которого из выполнения признака Больцано — Коши для последовательности x_n не следует существование элемента x , к которому сходится последовательность x_n .

Метрическое пространство называется *полным*, если для любой последовательности x_n , выполняющей условие Больцано — Коши, в X будет существовать предельный элемент.

Рассмотрим числовое пространство R_n , элементами которого являются упорядоченные совокупности n действительных (комплексных) чисел $x = (x_1, x_2, \dots, x_n)$. Определить в нем расстояние между элементами можно многими способами. Приведем три наиболее употребительных метрики.

1. *Кубическая* или m -метрика. Здесь

$$\rho_I(x, y) = \rho_m(x, y) = \max_i |x_i - y_i|.$$

2. *Октаэдрическая* или s -метрика:

$$\rho_{II}(x, y) = \rho_s(x, y) = \sum_{i=1}^n |x_i - y_i|.$$

3. *Сферическая* или l -метрика

$$\rho_{III}(x, y) = \rho_l(x, y) = \left[\sum_{i=1}^n |x_i - y_i|^2 \right]^{1/2}.$$

Легко проверить, что для каждого из этих расстояний выполняются аксиомы метрики. Во всех указанных метриках сходимость последовательности элементов $x^{(m)} \rightarrow x^*$ ($m \rightarrow \infty$) равносильна сходимостям по координатам и в каждой из метрик R_n является полным метрическим пространством.

2. Линейные нормированные пространства. Линейные операторы. Множество X называется *линейным* *), если для каждого двух его элементов определена сумма $x + y$, являющаяся элементом из X , и для любого элемента x и числа λ определено произведение λx , также являющееся элементом X , и обе эти операции подчинены условиям:

- 1) коммутативность сложения $x + y = y + x$;
- 2) ассоциативность сложения $(x + y) + z = x + (y + z)$;
- 3) существует элемент 0 , который называют нулевым, такой, что при всяком $x \in X$ будет $x + 0 = x$;
- 4) для всякого x существует такой элемент $-x$, называемый противоположным, что $x + (-x) = 0$;
- 5) умножение — ассоциативно: $\lambda(\mu x) = (\lambda\mu)x$;
- 6) верны два распределительных закона:

$$(\lambda + \mu)x = \lambda x + \mu x, \quad \lambda(x + y) = \lambda x + \lambda y;$$

$$7) 1 \cdot x = x;$$

$$8) 0 \cdot x = 0;$$

$$9) \text{ если } \lambda x = 0 \text{ и } x \neq 0, \text{ то } \lambda = 0.$$

Линейное множество X называется *нормированным*, если каждому $x \in X$ соответствует число $\|x\|$, называемое *нормой* x , для которого выполняются условия:

$$1) \|x\| \geq 0, \text{ и } \|x\| = 0, \text{ только когда } x = 0;$$

$$2) \|x + y\| \leq \|x\| + \|y\|;$$

$$3) \|\lambda x\| = |\lambda| \cdot \|x\|.$$

Линейное нормированное пространство всегда может быть метризовано, и для этого достаточно за расстояние между элементами x и y взять норму разности $x - y$:

$$\rho(x, y) = \|x - y\|.$$

Линейное нормированное и полное пространство часто называют *пространством типа Банаха* или, сокращенно, *В-пространством*. Ниже приведены примеры таких пространств.

1. Пространство $C[a, b]$. Рассмотрим множество функций $x(t)$, непрерывных на конечном отрезке $[a, b]$. Под сложением элементов этого множества понимают обычное сложение функции и под умножением элемента на число —

*) Линейные множества часто называют *векторными* пространствами.

обычное умножение функции на число. Под нормой функции понимают максимум ее абсолютного значения

$$\|x(t)\| = \max_t |x(t)|.$$

Сходимость последовательности элементов есть равномерная на $[a, b]$ сходимость последовательности непрерывных функций.

Пространство $C[a, b]$ является полным.

2. Пространство $L_p[a, b]$. Возьмем множество функций $x(t)$, измеримых и суммируемых со степенью p ($p \geq 1$) на конечном отрезке $[a, b]$. Норма функции определяется следующим интегралом:

$$\|x(t)\|^p = \int_a^b |x(t)|^p dt.$$

Две функции, отличающиеся между собой на множестве меры нуль, считаются совпадающими. Пространство $L_p[a, b]$ является полным.

Пусть X и Y есть два линейных нормированных пространства и A есть оператор, определенный на X и со значениями в пространстве Y : $y = A(x)$. Оператор A называется *непрерывным* на элементе x , если из сходимости $x_n \rightarrow x$ в X следует сходимость $A(x_n) \rightarrow A(x)$ в Y . Говорят, что оператор A непрерывен в X , если A непрерывен *) на каждом элементе $x \in X$.

A называют *аддитивным оператором*, если для всяких двух элементов x_1 и x_2 из X верно равенство $A(x_1 + x_2) = Ax_1 + Ax_2$.

Укажем некоторые свойства аддитивных операторов:

1. $A(0) = 0$. Действительно, $A(0) = A(0 + 0) = A(0) + A(0)$ и $A(0) = 0$.

2. $A(-x) = -A(x)$. В самом деле, $0 = A(0) = A[x + (-x)] = A(x) + A(-x)$ и $A(-x) = -A(x)$.

3. Пусть A есть аддитивный и непрерывный оператор, тогда он *однородный*, т. е. для любого действительного числа λ верно равенство

$$A(\lambda x) = \lambda A(x).$$

*) Приведенное определение непрерывности оператора сохраняет свой смысл, если X и Y есть два любых пространства, в которых определено понятие предела.

Ввиду аддитивности A при всяком целом значении p будет $A(px) = pA(x)$.

Далее, при целых q $A(x) = A\left(q \frac{1}{q} x\right) = qA\left(\frac{1}{q} x\right)$ и $A\left(\frac{1}{q} x\right) = \frac{1}{q} A(x)$. Следовательно, при всяких целых p и q $A\left(\frac{p}{q} x\right) = \frac{p}{q} A(x)$.

Возьмем произвольное действительное число λ и рассмотрим последовательность рациональных чисел r_n , сходящуюся к λ . Ввиду непрерывности оператора A

$$A(\lambda x) = \lim A(r_n x) = \lim r_n A(x) = \lambda A(x).$$

Оператор A называется *ограниченным*, если существует число M такое, что при всяких x из X верно неравенство

$$\|A(x)\| \leq M \|x\|. \quad (1)$$

Аддитивный оператор A называют *линейным*, если он непрерывен.

Теорема 1. *Для того чтобы аддитивный оператор был линейным, необходимо и достаточно, чтобы он был ограниченным.*

Доказательство. Необходимость. Предположим, что оператор A неограничен. Тогда существует последовательность элементов x_n такая, что будет

$$\|A(x_n)\| \geq n \|x_n\|.$$

Возьмем элемент

$$x'_n = \frac{1}{n} \frac{x_n}{\|x_n\|}.$$

При $n \rightarrow \infty$ $x'_n \rightarrow 0$. Но

$$\|A(x'_n)\| = \frac{1}{n} \frac{1}{\|x_n\|} \|A(x_n)\| \geq 1$$

и оказывается, что при $x'_n \rightarrow 0$ значение $A(x'_n)$ не стремится к нулевому элементу и A не является непрерывным на нулевом элементе. Сделанное выше предположение о неограниченности A противоречит его непрерывности.

Достаточность. Пусть $\|x - x_n\| \rightarrow 0$.

$$\|A(x) - A(x_n)\| = \|A(x - x_n)\| \leq M \|x - x_n\| \rightarrow 0 \quad (n \rightarrow \infty),$$

и поэтому

$$A(x_n) \rightarrow A(x).$$

Нормой оператора A называют наименьшее число M , для которого при всяких x выполняется неравенство (1). Норму обозначают $\|A\|$.

Теорема 2. Норма линейного оператора A равна верхней грани значений $\|A(x)\|$ в единичном шаре $\|x\| \leq 1$:

$$\|A\| = \sup_{\|x\| \leq 1} \|A(x)\|. \quad (2)$$

В самом деле, при $\|x\| \leq 1$

$$\|A(x)\| \leq \|A\| \|x\| \leq \|A\|. \quad (3)$$

С другой стороны, по определению нормы оператора для всякого $\varepsilon > 0$ существует элемент x' такой, что $\|A(x')\| > (\|A\| - \varepsilon) \|x'\|$. Положим $x = \frac{x'}{\|x'\|}$.

$$\|A(x)\| = \frac{1}{\|x'\|} \|A(x')\| > \frac{1}{\|x'\|} (\|A\| - \varepsilon) \|x'\| = \|A\| - \varepsilon.$$

Так как $\|x\| = 1$, отсюда получится

$$\sup_{\|x\| \leq 1} \|A(x)\| > \|A\| - \varepsilon.$$

Ввиду произвольности ε это неравенство совместно с (3) доказывает справедливость (2).

Оператор V , определенный в пространстве Y и имеющий значения в X , называется *обратным* для оператора A и обозначается A^{-1} , если он удовлетворяет двум следующим условиям:

- 1) при всяких $x \in X$ $VA(x) = x$,
 - 2) при всех $y \in Y$ $AV(y) = y$.
- (4)

Отметим, что если оператор A имеет обратный $V = A^{-1}$, то A осуществляет взаимно однозначное отображение X на Y . Действительно, по второму условию каждый элемент y есть результат преобразования некоторого $x \in X$, например в качестве такого элемента может быть взят $x = A^{-1}(y)$. Остается проверить, если $x_1 \neq x_2$, то $A(x_1) \neq A(x_2)$. Действительно, если $A(x_1) = A(x_2)$, то по первому условию будет $x_1 = VA(x_1) = VA(x_2) = x_2$.

В качестве примера рассмотрим линейное преобразование ν -мерного векторного пространства X_ν с элементами $x(x_1, \dots, x_\nu)$ в μ -мерное векторное пространство Y_μ с элементами $y(y_1, \dots, y_\mu)$:

$$y = A(x).$$

При линейном преобразовании $X_\nu \rightarrow Y_\mu$ составляющие y_i вектора y будут линейными функциями составляющих x_j вектора x и само преобразование определится системой равенств

$$y_i = \sum_{j=1}^{\nu} a_{ij} x_j, \quad i = 1, 2, \dots, \mu.$$

Это есть обычное линейное преобразование с прямоугольной матрицей переменных x_j ($j = 1, \dots, \nu$) в переменные y_i ($i = 1, \dots, \mu$).

3. Дифференцирование нелинейных операторов. Нам потребуются некоторые факты теории линейных и билинейных операторов.

Пусть X и Y — линейные нормированные пространства. Возьмем множество всех линейных операторов, переводящих X в Y . Обозначим его $[X \rightarrow Y]$. Это множество можно при надлежащем определении операций сложения элементов, умножения их на число и выборе нормы сделать линейным нормированным пространством. Если A — линейный оператор, принадлежащий $[X \rightarrow Y]$ и λ — произвольное число, то под произведением $\lambda A = \tilde{A}$ понимают оператор, определенный равенством $\tilde{A}(x) = \lambda A(x)$.

Пусть A_1 и A_2 два любых элемента из $[X \rightarrow Y]$. Сумму $A = A_1 + A_2$ определяют равенством

$$A(x) = A_1(x) + A_2(x).$$

Так как A_1 и A_2 — аддитивные операторы, то A также аддитивный оператор. Далее

$$\begin{aligned} \|A(x)\| &= \|A_1(x) + A_2(x)\| \leq \\ &\leq \|A_1(x)\| + \|A_2(x)\| \leq (\|A_1\| + \|A_2\|) \|x\|. \end{aligned}$$

Поэтому A — ограниченный и, следовательно, линейный оператор со следующей оценкой нормы:

$$\|A\| \leq \|A_1\| + \|A_2\|.$$

За нулевой элемент в $[X \rightarrow Y]$ примем нулевой оператор A_0 :

$$A_0(x) = 0 \quad \text{при всех } x \in X.$$

Все аксиомы линейного множества, как можно легко проверить, при выбранном определении умножения на число и сложения для $[X \rightarrow Y]$ будут выполняться.

За норму A , как элемента из $[X \rightarrow Y]$, принимают норму оператора $A(x)$:

$$\|A\| = \sup_{\|x\| \leq 1} \|A(x)\|.$$

Введем понятие билинейного оператора. Пусть каждой паре элементов $x, x' \in X$ ставится в соответствие некоторый элемент $y \in Y$:

$$y = B(x', x).$$

Оператор $B(x', x)$ называется *билинейным*, когда выполняются следующие условия:

1) при всяких элементах $x, x', \tilde{x}, \tilde{x}' \in X$ и числах $\alpha, \tilde{\alpha}$ верны равенства

$$\begin{aligned} B(\alpha x' + \tilde{\alpha} \tilde{x}', x) &= \alpha B(x', x) + \tilde{\alpha} B(\tilde{x}', x), \\ B(x', \alpha x + \tilde{\alpha} \tilde{x}) &= \alpha B(x', x) + \tilde{\alpha} B(x', \tilde{x}); \end{aligned} \quad (5)$$

2) существует число M такое, что при любых $x', x \in X$ верно неравенство

$$\|B(x', x)\| \leq M \|x'\| \cdot \|x\|. \quad (6)$$

Нормой оператора $B(x', x)$ называют наименьшее значение M , при котором выполняется (6).

Пример. Пусть X_v есть v -мерное векторное пространство с элементами $x(x_1, x_2, \dots, x_v)$ и Y_μ есть μ -мерное векторное пространство, элементы которого — $y(y_1, y_2, \dots, y_\mu)$. Предположим, что $y = B(x', x)$ есть билинейный оператор из X_v в Y_μ . Каждая составляющая y_i элемента y является, очевидно, билинейной формой от составляющих элементов $x'(x'_1, \dots, x'_v)$ и $x(x_1, \dots, x_v)$

$$y_i = \sum_{j,k=1}^v a_{jk}^i x'_j x_k \quad (i = 1, \dots, \mu). \quad (7)$$

$y = B(x', x)$ — μ -мерный вектор, все составляющие которого есть билинейные формы вида (7).

Норма оператора $B(x', x)$ зависит от выбора норм в пространствах X_ν и Y_μ . Пусть в них приняты кубические (или m -) нормы

$$\|x\|_m = \max_i |x_i|, \quad \|y\|_m = \max_i |y_i|.$$

Тогда

$$\begin{aligned} \|y\|_m = \|B(x', x)\| &= \max_i \sum_{j,k} |a_{jk}^i x'_j x_k| \leq \\ &\leq \max_i \sum_{j,k} |a_{jk}^i| \|x'\|_m \|x\|_m \end{aligned}$$

и, следовательно,

$$\|B\|_m \leq \max_i \sum_{jk} |a_{jk}^i|.$$

При определении второй производной от оператора мы встретимся с отображением пространства X на множество операторов $[X \rightarrow Y]$. Обозначим $[X \rightarrow [X \rightarrow Y]]$ множество таких отображений. Существует тесная связь между ними и билинейными операторами.

Пусть U есть оператор, отображающий X в $[X \rightarrow Y]$. Возьмем элемент x' из X и рассмотрим $U(x') = A_{x'}$. Это есть оператор из $[X \rightarrow Y]$, поэтому $y = A_{x'} x = U(x') x = B(x', x)$ является элементом из Y . Оператор $B(x', x)$ удовлетворяет второму из условий (5), так как оператор $A_{x'} x$ линеен относительно x и удовлетворяет первому условию (5) ввиду того, что $A_{x'} = U(x')$ является линейным относительно x' .

Неравенство (6) также выполняется при некотором M , так как

$$\|B(x', x)\| \leq \|A_{x'}\| \|x\| = \|Ux'\| \|x\| \leq \|U\| \|x'\| \|x\|,$$

и можно положить $M = \|U\|$. Таким образом, $B(x', x)$ действительно есть билинейный оператор и

$$\|B\| \leq \|U\|. \quad (8)$$

Указанным выше способом каждому оператору U из $[X \rightarrow [X \rightarrow Y]]$ ставится в соответствие билинейный оператор $B(x', x)$. Убедимся теперь в том, что при указанном законе соответствия каждый билинейный опера-

тор $B(x', x)$ есть отображение некоторого оператора U из $[X \rightarrow [X \rightarrow Y]]$.

Фиксируем элемент x' . Оператор $B(x', \cdot) = U(x') = = A_{x'}$ принадлежит множеству $[X \rightarrow Y]$, и ввиду неравенств

$$\|U(x')\| = \sup_{\|x\| \leq 1} \|U(x')(x)\| = \sup_{\|x\| \leq 1} \|B(x', x)\| \leq \|B\| \|x'\|,$$

оператор U ограничен, а следовательно, линеен и принадлежит множеству $[X \rightarrow [X \rightarrow Y]]$. Из неравенств вытекает также, что $\|U\| \leq \|B\|$. Сравнение последнего результата с неравенством (8) позволяет утверждать, что

$$\|B\| = \|U\|. \quad (9)$$

Определим производную первого порядка от оператора. Как и раньше, будем считать X и Y линейными нормированными пространствами и предположим, что дан оператор

$$y = f(x), \quad (10)$$

преобразующий X в Y . Оператор f называют *дифференцируемым на элементе x* , если существует линейный оператор H , принадлежащий $[X \rightarrow Y]$, такой, что верно неравенство

$$\|f(x + \Delta x) - f(x) - H(\Delta x)\| \leq \|\Delta x\| \varepsilon(\|\Delta x\|), \quad (11)$$

где $\varepsilon(\delta) \rightarrow 0$ ($\delta \rightarrow 0$). Линейный оператор H называют *производной от оператора f на элементе x* :

$$H = f'(x). \quad (12)$$

Заметим, что $H(\Delta x)$ имеет смысл дифференциала оператора *) и является аналогом известного в курсах анализа дифференциала функции.

Для примера рассмотрим преобразование ν -мерного векторного пространства X_ν в μ -мерное векторное пространство Y_μ . Такое преобразование $y = f(x)$ равносильно заданию μ функций

$$y_i = f_i(x_1, x_2, \dots, x_\nu), \quad i = 1, 2, \dots, \mu. \quad (13)$$

*) $f(x + \Delta x) - f(x)$ есть изменение оператора при переходе от элемента x к элементу $x + \Delta x$ и $H(\Delta x)$ есть главная линейная часть этого изменения.

Для вычисления производной от оператора f нужно из его изменения $f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x})$ выделить главную часть, линейно зависящую от $\Delta\mathbf{x} = (\Delta x_1, \dots, \Delta x_v)$:

$$f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) = H(\Delta\mathbf{x}) + \varepsilon(\mathbf{x}, \Delta\mathbf{x}),$$

где

$$\frac{\|\varepsilon(\mathbf{x}, \Delta\mathbf{x})\|}{\|\Delta\mathbf{x}\|} \rightarrow 0 \quad (\|\Delta\mathbf{x}\| \rightarrow 0).$$

Возьмем одну из составляющих изменения оператора

$$\begin{aligned} f_i(\mathbf{x} + \Delta\mathbf{x}) - f_i(\mathbf{x}) &= \\ &= f_i(x_1 + \Delta x_1, \dots, x_v + \Delta x_v) - f_i(x_1, \dots, x_v) = \Delta f_i. \end{aligned}$$

Главной частью приращения функции f_i , линейной относительно $\Delta x_1, \dots, \Delta x_v$, является полный дифференциал функции f_i , так что

$$\{H(\Delta\mathbf{x})\}_i = \sum_{j=1}^v \frac{\partial f_i(\mathbf{x})}{\partial x_j} \Delta x_j. \quad (14)$$

Это говорит о том, что производная f' есть линейное преобразование $X_v \rightarrow Y_\mu$, матрица которого есть $v \times \mu$ матрица Якоби с элементами, вычисленными для точки $\mathbf{x}$:

$$f'(\mathbf{x}) = \left[\frac{\partial f_i}{\partial x_j} \right] \quad \begin{pmatrix} i = 1, 2, \dots, \mu; \\ j = 1, 2, \dots, v \end{pmatrix}.$$

Определим теперь вторую производную от оператора f . Первая производная $H = f'(\mathbf{x})$ есть линейный оператор, принадлежащий $[X \rightarrow Y]$. Он зависит от элемента $\mathbf{x}$, на котором вычисляется f' , как от своего аргумента, и каждому элементу $\mathbf{x}$ отвечает свой линейный оператор. Может оказаться, что оператор f' имеет производную на элементе $\mathbf{x}$. Производную от f' называют второй производной от оператора f и обозначают f'' . Она принадлежит множеству $[X \rightarrow [X \rightarrow Y]]$.

Выше было показано, что рассмотрение множества операторов $[X \rightarrow [X \rightarrow Y]]$ равносильно рассмотрению множества билинейных операторов $B(\mathbf{x}', \mathbf{x})$ при выборе определенного, указанного выше закона соответствия между элементами этих множеств. Такой равносильностью пользуются в вычислениях, так как билинейные формы часто бывают удобнее.

В качестве примера вычислим вторую производную от рассмотренного выше оператора, преобразующего векторное пространство X_ν , имеющего размерность ν , в векторное пространство Y_μ размерности μ . Выберем произвольный элемент x_0 из X_ν и вычислим f'' на x_0 .

По определению производной (11) нужно найти такой линейный оператор U из множества $[X \rightarrow [X \rightarrow Y]]$, для которого выполняется неравенство

$$\|f'(x_0 + \Delta x)(x) - f'(x_0)(x) - U(\Delta x)\| \leq \|\Delta x\| \varepsilon(\|\Delta x\|). \quad (15).$$

Полезно преобразовать это условие к другой форме, более удобной для вычислений. Зафиксируем какой-либо элемент $x' \in X_\nu$ и положим $\Delta x = tx'$. Разделив (15) на t и перейдя к пределу при $t \rightarrow 0$, получим равенство

$$\lim_{t \rightarrow 0} \frac{1}{t} [f'(x_0 + tx')(x) - f'(x_0)(x)] = U(x')(x) = A_{x'}(x). \quad (16)$$

Или, так как $A_{x'}(x) = U(x')(x) = B(x', x)$,

$$\lim_{t \rightarrow 0} \frac{1}{t} [f'(x_0 + tx')(x) - f'(x_0)(x)] = B(x', x). \quad (17)$$

Значение $B(x', x)$ принадлежит пространству Y_μ и является вектором $y(y_1, \dots, y_\mu)$. Его составляющие y_i есть билинейные формы от $(x'_1, \dots, x'_\nu)$ и $(x_1, \dots, x_\nu)$ и их значения записываются в виде (7)

$$\{B(x', x)\}_i = \sum_{jk=1}^{\nu} a_{jk}^i x'_j x_k. \quad (18)$$

Векторное равенство (17) равносильно ν численным равенствам

$$\lim_{t \rightarrow 0} \frac{1}{t} \left[\sum_{k=1}^{\nu} \left(\frac{\partial}{\partial x_k} f_i(x_0 + tx') - \frac{\partial}{\partial x_k} f_i(x_0) \right) x_k \right] = \{B(x', x)\}_i.$$

Отсюда получается

$$\sum_{jk=1}^{\nu} \frac{\partial^2}{\partial x_j \partial x_k} f_i(x_0) x'_j x_k = \{B(x', x)\}_i$$

и, следовательно,

$$a_{jk}^i = \frac{\partial^2}{\partial x_j \partial x_k} f_i(x_0).$$

Для § 3 нам потребуются теоремы об изменении оператора и о главной части изменения. Мы ограничились тем, что сформулировали эти теоремы, но для краткости опустили их доказательства, которые можно найти в статье [2] (с. 166—168) или в [1] гл. 9 (с. 252—253).

Теорема 2. Пусть f есть дифференцируемый оператор. Для его изменения верна оценка

$$\|f(x + \Delta x) - f(x)\| \leq \sup_{\theta} \|f'(x + \theta \Delta x)\| \|\Delta x\| \quad (0 \leq \theta \leq 1). \quad (19)$$

Теорема 3. Если f есть дважды дифференцируемый оператор, то для разницы между изменением функционала f и главной линейной частью его верна оценка

$$\begin{aligned} \|f(x + \Delta x) - f(x) - f'(x) \Delta x\| &\leq \\ &\leq \frac{1}{2} \max_{\theta} \|f''(x + \theta \Delta x)\| \cdot \|\Delta x\|^2. \end{aligned} \quad (20)$$

§ 2. Метод итерации для решения операторных уравнений

Метод итерации применялся в гл. 2 первого тома для решения систем линейных уравнений и в гл. 4 для решения любых уравнений или их систем, и в обоих случаях речь шла о нахождении значений неизвестных чисел. Но совершенно очевидно, что идея метода итерации может быть применена для решения уравнений с неизвестными любой природы. Им пользуются, в частности, для решения многообразных задач, где неизвестными являются функции одного или многих аргументов; при этом в уравнения могут входить не только сами функции, но и некоторые операторы от них, такие, как операторы дифференцирования, интегрирования, смещения аргумента и т. д.

Было бы целесообразно не рассматривать отдельно применение метода итерации к большому числу различных видов уравнений, а построить общую теорию метода итерации, не зависящую от вида уравнений, и из нее

как частные случаи получать необходимые результаты для отдельных конкретных задач.

Простейший результат общей теории метода итерации — принцип сжатых изображений — будет изложен в настоящем параграфе.

1. Метод итерации для операторных уравнений. Предположим, что рассматривается множество X элементов x произвольной природы: это могут быть числа или их системы, функции или их системы и т. д. Пусть на этом множестве задан оператор $y = \varphi(x)$, значения которого принадлежат тому же множеству. Этот оператор ставит в соответствие каждому элементу x из X некоторый элемент из того же множества. Действие такого оператора может быть истолковано как преобразование множества X в себя. Рассмотрим уравнение

$$x = \varphi(x). \quad (1)$$

Решить его — значит найти такие элементы X , которые при преобразовании X оператором φ переходят в себя, иначе говоря, остаются неподвижными при таком преобразовании.

Будем считать, что такие неподвижные элементы существуют, и рассмотрим итерационный метод построения последовательных приближений для их нахождения. Пусть каким-то способом найдено исходное приближение x^0 к неподвижному элементу. Все остальные приближения находятся по следующему закону:

$$x^{n+1} = \varphi(x^n) \quad (n = 0, 1, \dots). \quad (2)$$

Заметим, что при принятом выше предположении о том, что значения $\varphi(x)$ принадлежат области определения X оператора φ , приближения x^n могут быть построены для всех значений n . В некоторых задачах такое предположение не будет выполняться, и тогда может возникнуть проблема выяснения условий о возможности построения бесконечной последовательности x^n .

Пусть последовательность x^n построена. Для нее мы должны выяснить следующие вопросы: 1) будет ли она сходиться в каком-то принятом нами смысле? 2) если x^* есть элемент, к которому сходится x^n , то какова скорость сходимости?

2. Теоремы о сходимости и единственности. Рассмотрим случай, когда X есть полное метрическое пространство. Обозначим его метрику через $\rho(x, y)$.

Теорема 1. *Предположим, что выполняются условия:*

1) *оператор $\varphi(x)$ определен в замкнутом шаре S радиуса δ с центром в x^0 :*

$$\rho(x, x^0) \leq \delta; \quad (3)$$

2) *для всяких элементов x и y из шара S верно соотношение*

$$\rho[\varphi(x), \varphi(y)] \leq q\rho(x, y), \quad 0 \leq q < 1, \quad (4)$$

где q — постоянная величина;

3) *для начального элемента x^0 выполнено неравенство*

$$\rho[\varphi(x^0), x^0] \leq t; \quad (5)$$

4) *для чисел δ, q, t , соблюдается условие*

$$\frac{t}{1-q} \leq \delta. \quad (6)$$

Тогда верны следующие утверждения:

1) *итерационные приближения x^n , вычисляемые по правилу (2), могут быть найдены при всяком значении n , и все они принадлежат шару S ;*

2) *последовательность x^n сходится к некоторому элементу x^* из S :*

$$\lim_n x^n = x^* \quad (x^* \in S);$$

3) *предельный элемент x^* есть решение заданного уравнения (1);*

4) *для приближения x^n верна следующая оценка близости к решению x^* :*

$$\rho[x^n, x^*] \leq \frac{t}{1-q} q^n. \quad (7)$$

Перед доказательством сделаем пояснительные замечания относительно условий теоремы. Начнем со второго условия и запишем его в виде

$$\frac{\rho[\varphi(x), \varphi(y)]}{\rho(x, y)} \leq q.$$

Здесь знаменатель есть расстояние между элементами x и y до преобразования, числитель $\rho[\varphi(x), \varphi(y)]$ — рассто-

ание между теми же элементами после преобразования шара S при помощи оператора φ . Отношение имеет смысл коэффициента изменения расстояния между элементами при преобразовании. Он не больше числа q , и так как $q < 1$, то при преобразовании происходит уменьшение всех расстояний между элементами шара.

Если исходное приближение x^0 случайно окажется решением уравнения (1), то $\varphi(x^0) = x^0$ и будет $\rho[\varphi(x^0), x^0] = 0$. Если же x^0 не будет совпадать с решением, но будет близким к нему элементом, то расстояние $\rho[\varphi(x^0), x^0]$ будет малой величиной и число m может быть взято малым. Когда x^0 лежит на большом расстоянии от решения уравнения, то трудно ожидать, что m будет иметь малое значение. Ввиду этого число m в условии (5) тесно связано с близостью x^0 к точному решению уравнения.

Условия теоремы означают следующее. Когда в некотором шаре S достаточно большого радиуса δ с центром в x^0 при преобразовании оператором $y = \varphi(x)$ происходит достаточно сильное сжатие (q не близко к единице) и x^0 взято достаточно близко к решению (m есть достаточно малая величина), то уравнение имеет решение в шаре S и к нему сходятся итерационные последовательные приближения x^n . Численная же характеристика всех этих условий дается в неравенстве (6).

Теорему 1 ввиду указанного наглядного содержания часто называют *принципом сжатых отображений*.

Доказательство аналогично доказательству сходной теоремы в гл. 4. т. I.

Убедимся сначала в том, что приближения x^n могут быть построены при всяких n , принадлежат шару S и выполняется неравенство

$$\rho(x^n, x^{n+1}) \leq mq^n. \quad (8)$$

Для $n=0$ это очевидно ввиду того, что x^0 есть центр шара, и значение $x^1 = \varphi(x^0)$ может быть найдено, так как оператор φ на элементе x^0 определен и, кроме того,

$$\rho(x^0, x^1) = \rho(x^0, \varphi(x^0)) \leq m \leq \frac{m}{1-q} \leq \delta,$$

поэтому x^1 принадлежит S .

Предположим теперь, что $x^0, x^1, \dots, x^n$ построены, принадлежат S и для них выполнены неравенства

$$\rho(x^k, x^{k+1}) \leq mq^k \quad (k=0, 1, \dots, n-1). \quad (9)$$

Приближение $x^{n+1} = \varphi(x^n)$ может быть вычислено, так как $x^n \in S$. По второму условию теоремы и по индуктивному предположению

$$\begin{aligned} \rho(x^n, x^{n+1}) &= \rho[\varphi(x^{n-1}), \varphi(x^n)] \leq q\rho(x^{n-1}, x^n) \leq \\ &\leq q \cdot mq^{n-1} = mq^n \end{aligned}$$

и неравенство (9) для $k=n$ действительно выполнено. Осталось проверить, что x^{n+1} принадлежит шару S . Если применить несколько раз аксиому треугольника, получим

$$\begin{aligned} \rho(x^0, x^{n+1}) &\leq \rho(x^0, x^n) + \rho(x^n, x^{n+1}) \leq \dots \\ &\dots \leq \rho(x^0, x^1) + \rho(x^1, x^2) + \dots + \rho(x^n, x^{n+1}) \leq \\ &\leq m + mq + \dots + mq^n < \frac{m}{1-q} \leq \delta. \end{aligned}$$

Отсюда видно, что $\rho(x^0, x^{n+1}) \leq \delta$ и $x^{n+1} \in S$.

Проверим, что для последовательности x^n выполняется признак Больцано — Коши

$$\begin{aligned} \rho(x^n, x^{n+p}) &\leq \rho(x^n, x^{n+1}) + \rho(x^{n+1}, x^{n+2}) + \dots \\ &\dots + \rho(x^{n+p-1}, x^{n+p}) \leq mq^n + mq^{n+1} + \dots + mq^{n+p-1} = \\ &= \frac{m}{1-q} (q^n - q^{n+p}) < \frac{m}{1-q} q^n. \end{aligned}$$

Так как $0 \leq q < 1$, из неравенства $\rho(x^n, x^{n+p}) < \frac{m}{1-q} q^n$ вытекает, что для любого $\varepsilon > 0$ при всяких достаточно больших n и всех p будет $\rho(x^n, x^{n+p}) \leq \varepsilon$.

Из признака Больцано — Коши ввиду полноты пространства следует существование элемента x^* , к которому сходятся x^n :

$$x^n \rightarrow x^* \quad (n \rightarrow \infty).$$

Переходя в неравенстве $\rho(x^0, x^n) \leq \delta$ к пределу при $n \rightarrow \infty$, получим $\rho(x^0, x^*) \leq \delta$, что доказывает принадлежность x^* шару S .

Наконец, переходя в неравенстве $\rho[\varphi(x^n), \varphi(x^*)] \leq q\rho(x^n, x^*)$ к пределу при $n \rightarrow \infty$, ввиду $\rho(x^n, x^*) \rightarrow 0$ получим, что $\rho[\varphi(x^n), \varphi(x^*)] \rightarrow 0$, т. е. $\varphi(x^n) \rightarrow$

обозначать совокупность значений функций $(\varphi_1, \varphi_2, \dots, \varphi_m)$. Система запишется в краткой векторной форме

$$\mathbf{x} = \varphi(\mathbf{x}).$$

Пусть как-то выбрано начальное приближение к решению $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_m^0)$. Все следующие приближения будут находиться по правилу итераций

$$\mathbf{x}^{n+1} = \varphi(\mathbf{x}^n).$$

Для выяснения условий, достаточных для сходимости последовательности приближений $\mathbf{x}^n$ ($n = 0, 1, \dots$), применим теорему 1. Для этого мы должны в R_n ввести метрику. Сделать это можно многими способами, и мы остановимся на трех классических метриках.

1. Случай кубической метрики $\rho(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$. В шаре $\rho(\mathbf{x}, \mathbf{x}^0) \leq \delta$ возьмем две произвольные точки $\mathbf{x} = (x_1, x_2, \dots, x_m)$ и $\mathbf{y} = (y_1, y_2, \dots, y_m)$. Оценим изменение функции φ_i :

$$\begin{aligned} |\varphi_i(\mathbf{x}) - \varphi_i(\mathbf{y})| &= |\varphi_i(x_1, x_2, \dots, x_m) - \varphi_i(y_1, y_2, \dots, y_m)| = \\ &= \left| \sum_{j=1}^m \left(\frac{\partial \varphi_i}{\partial x_j} \right)^* (x_j - y_j) \right| \leq \left(\sum_{j=1}^m \left| \frac{\partial \varphi_i}{\partial x_j} \right| \right)^* \max_j |x_j - y_j| = \\ &= \left(\sum_{j=1}^m \left| \frac{\partial \varphi_i}{\partial x_j} \right| \right)^* \rho(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Звездочкой у скобок отмечено то, что значение функции, стоящей в скобках, должно быть взято в некоторой точке отрезка прямой, соединяющего $\mathbf{x}$ и $\mathbf{y}$. Чтобы найти оценку изменения, не зависящую от i и положения точек $\mathbf{x}, \mathbf{y}$, заменим последнюю из скобок ее максимальным значением по $\mathbf{x}$ и i . Тогда получим

$$\begin{aligned} |\varphi_i(\mathbf{x}) - \varphi_i(\mathbf{y})| &\leq \max_i |\varphi_i(\mathbf{x}) - \varphi_i(\mathbf{y})| = \rho[\varphi(\mathbf{x}), \varphi(\mathbf{y})] \leq \\ &\leq \max_i \max_{\mathbf{x}} \sum_{j=1}^m \left| \frac{\partial \varphi_i}{\partial x_j} \right| \cdot \rho(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Отсюда видно, что в качестве числа q в неравенстве (4) может быть взята величина

$$q = \max_i \max_x \sum_{j=1}^m \left| \frac{\partial \varphi_i}{\partial x_j} \right|.$$

Все это позволяет сформулировать теорему о сходимости итерационной последовательности в случае кубической метрики.

Теорема 3. Пусть:

1) функции $\varphi_i(x_1, \dots, x_m)$ определены и непрерывно дифференцируемы в области

$$|x_i - x_i^0| \leq \delta \quad (i = 1, \dots, m); \quad (11_1)$$

2) в этой области

$$\max_x \sum_{j=1}^m \left| \frac{\partial \varphi_i}{\partial x_j} \right| \leq q < 1 \quad (i = 1, \dots, m); \quad (12_1)$$

3) начальные приближения неизвестных удовлетворяют неравенствам

$$|x_i^0 - \varphi_i(x_1^0, \dots, x_m^0)| \leq M \quad (i = 1, \dots, m); \quad (13_1)$$

4) для чисел δ , q и M выполняется условие

$$\frac{M}{1-q} \leq \delta. \quad (14_1)$$

Тогда:

1) система (10) в области (11₁) имеет решение $(x_1^*, x_2^*, \dots, x_m^*)$ и к нему сходится итерационная последовательность приближений $(x_1^n, \dots, x_m^n)$;

2) скорость сходимости может быть охарактеризована неравенствами

$$|x_i^n - x_i^*| \leq \frac{M}{1-q} q^n \quad (i = 1, \dots, m). \quad (15_1)$$

2. Случай октаэдрической метрики $\rho(x, y) = \sum_{j=1}^m |x_j - y_j|$. Аналогично теореме 3 для октаэдрической метрики в R_n может быть доказана

Теорема 4. Пусть:

1) функции $\varphi_i(x_1, \dots, x_m)$ ($i = 1, \dots, m$) определены и непрерывны в области

$$\sum_{j=1}^m |x_j - x_j^0| \leq \delta; \quad (11_2)$$

2) удовлетворяют в этой области условию

$$\sum_{i=1}^m \max_j \max_x \left| \frac{\partial \varphi_i}{\partial x_j} \right| \leq q < 1; \quad (12_2)$$

3) для начальных приближений неизвестных $x_1^0, \dots, x_m^0$ выполняется неравенство

$$\sum_{i=1}^m |x_i^0 - \varphi_i(x_1^0, \dots, x_m^0)| \leq M; \quad (13_2)$$

4) для чисел δ, q, M соблюдается условие

$$\frac{M}{1-q} \leq \delta. \quad (14_2)$$

Тогда:

1) система (10) в области (11₂) имеет решение $(x_1^*, \dots, x_m^*)$ и к нему сходится итерационная последовательность приближений $(x_1^n, \dots, x_m^n)$;

2) скорость сходимости характеризуется неравенством

$$\sum_{i=1}^m |x_i^* - x_i^n| \leq \frac{M}{1-q} q^n. \quad (15_2)$$

3. Случай шаровой метрики $\rho(x, y) = \left(\sum_{i=1}^m |x_i - y_i|^2 \right)^{1/2}$. Аналогично теоремам 3 и 4 может быть доказана следующая теорема о сходимости итерационного процесса.

Теорема 5. Пусть:

1) функции $\varphi_i(x_1, \dots, x_m)$ определены и непрерывны в области

$$\sum_{i=1}^m |x_i - x_i^0|^2 \leq \delta^2; \quad (11_3)$$

2) удовлетворяют в этой области условию

$$\sum_{i=1}^m \max_x \sum_{j=1}^m \left| \frac{\partial \varphi_i}{\partial x_j} \right|^2 \leq q < 1; \quad (12_3)$$

3) для начальных приближений $(x_1^0, \dots, x_m^0)$ верно неравенство

$$\sum_{i=1}^m |x_i^0 - \varphi_i(x_1^0, \dots, x_m^0)|^2 \leq M^2; \quad (13_3)$$

4) для чисел δ , q и M соблюдается условие

$$\frac{M}{1-q} \leq \delta. \quad (14_3)$$

Тогда:

1) система (10) в области (11₃) имеет решение $(x_1^*, \dots, x_m^*)$ и к нему сходится итерационная последовательность приближений $(x_1^n, \dots, x_m^n)$;

2) скорость сходимости можно характеризовать неравенством

$$\left\{ \sum_{i=1}^m |x_i^* - x_i^n|^2 \right\}^{1/2} \leq \frac{M}{1-q} q^n. \quad (15_3)$$

§ 3. Метод Ньютона для операторных уравнений

В гл. 4 т. I рассматривался метод Ньютона для решения уравнений с численными неизвестными и для систем таких уравнений. Но ясно, что основная мысль этого метода — приведение решения нелинейного уравнения к решению последовательности линейных уравнений — может применяться к уравнениям других видов: к многочисленным нелинейным задачам для дифференциальных уравнений обыкновенных и с частными производными, для нелинейных интегральных уравнений и т. д. В настоящем параграфе будут рассмотрены начала общей теории метода Ньютона для операторных уравнений достаточно общего вида.

1. Правило построения приближений к решению. Предположим, что X и Y являются полными, линейными, нормированными пространствами.

Пусть в некоторой области D пространства X задан нелинейный оператор $y = f(x)$ со значениями в Y .

Будем рассматривать уравнение

$$f(x) = 0. \quad (1)$$

В правой части его стоит нулевой элемент пространства Y . Допустим, что каким-либо способом найдено исходное приближение x_0 к решению. Для выделения из уравнения главной линейной части положим $f(x) = f(x_0) + [f(x) - f(x_0)]$ и рассмотрим изменение оператора $f(x) - f(x_0)$. Так как x_0 берется обычно близким к точному решению x и норма их разности $\|x - x_0\|$ будет малой величиной, изменение оператора $f(x) - f(x_0)$ будет отличаться, как показывает неравенство (1.17) от $f'(x_0) \times (x - x_0)$ «весьма мало», и поэтому заданное уравнение (1) будет «близким» к следующему приближенному линейному уравнению:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) = 0. \quad (2)$$

Можно ожидать, что его решение, которое мы обозначим через x_1 , ближе к точному решению, чем x_0 . Когда для $f'(x_0)$ существует обратный оператор $[f'(x_0)]^{-1}$, x_1 получим в форме

$$x_1 = x_0 - [f'(x_0)]^{-1} f(x_0).$$

Для x_1 можно повторить проделанные рассуждения и по x_1 построить второе улучшенное приближение x_2 и так далее. Правило Ньютона для нахождения следующего приближения по предыдущему имеет вид

$$x_{n+1} = x_n - [f'(x_n)]^{-1} f(x_n). \quad (3)$$

Построение последовательности приближений x_n ($n = 0, 1, \dots$) возможно при выполнении двух условий: 1) все приближения x_n принадлежат области D ; 2) при всяком значении n существует обратный оператор $[f'(x_n)]^{-1}$.

2. Теорема о сходимости последовательности приближений. Докажем теорему о разрешимости уравнения, о возможности построения последовательности приближений x_n и о сходимости ее к решению. Известно несколько теорем такого содержания, мы приведем одну из них,

аналогичную теореме о сходимости процесса Ньютона для численных уравнений, доказанную в гл. 4.

Теорема 1. *Предположим, что выполняются условия:*

1) оператор $f(x)$ определен в замкнутом шаре

$$\|x - x_0\| \leq \delta, \quad (4)$$

дважды дифференцируем там, при этом вторая производная от него ограничена по норме числом K :

$$\|f''(x)\| \leq K, \quad \|x - x_0\| \leq \delta; \quad (5)$$

2) оператор $f'(x_0)$ имеет обратный $\Gamma_0 = [f'(x_0)]^{-1}$, для нормы которого выполнена оценка

$$\|\Gamma_0\| \leq B; \quad (6)$$

3) для начального приближения x_0 к решению верно неравенство

$$\|\Gamma_0 f(x_0)\| \leq \eta; \quad (7)$$

4) величины K, B, η удовлетворяют условию

$$h = KB\eta \leq \frac{1}{2}; \quad (8)$$

5) для числа δ верно неравенство

$$\frac{1 - \sqrt{1 - 2h}}{h} \eta \leq \delta. \quad (9)$$

Тогда:

1) в шаре (4) уравнение $f(x) = 0$ имеет решение;

2) в вычислительном процессе Ньютона (3) приближение x_n может быть построено при любом значении n , все x_n принадлежат шару (4) и последовательность x_n сходится к решению x^* уравнения;

3) для приближения x_n верно следующее неравенство, характеризующее скорость сходимости:

$$\|x^* - x_n\| \leq t^* - t_n, \quad (10)$$

где t^* есть меньший корень уравнения

$$P(t) = \frac{1}{2} K t^2 - \frac{1}{B} t + \frac{\eta}{B} = 0$$

и t_n есть приближение к нему, построенное по методу Ньютона при начальном значении $t_0 = 0$.

Доказательство. Предварительно необходимо напомнить *) некоторые факты, касающиеся уравнения $P(t) = 0$ и последовательности t_n . Если для параметров K, B, η уравнения выполнено условие (8), то корни уравнения действительны и положительны, при этом меньший из них есть $t^* = h^{-1} [1 - \sqrt{1 - 2h}]$. Ньютонова последовательность приближений t_n к t^* , начатая с $t_0 = 0$, является монотонно возрастающей и скорость приближения t_n к t^* достаточно наглядно характеризуется неравенством

$$0 < t^* - t_n \leq \frac{1}{2^{n-1}} (2h)^{2^n - 1} \cdot \eta.$$

В ходе доказательства нам потребуется простая лемма о существовании обратного оператора.

Лемма 1. *Предположим, что в полном нормированном линейном пространстве X задан линейный оператор H , преобразующий X в себя. Если $\|H\| \leq q < 1$, то для оператора $I - H$ существует обратный ему $(I - H)^{-1}$ и при этом $\|(I - H)^{-1}\| \leq \frac{1}{1 - q}$.*

Доказательство. Рассмотрим в пространстве X ряд $Ax = x + Hx + H^2x + \dots$ и покажем, что он сходится. Применим к нему признак Больцано — Коши. Возьмем отрезок ряда и оценим его норму

$$\begin{aligned} \|H^n x + H^{n+1} x + \dots + H^{n+p} x\| &\leq \\ &\leq \|H^n \cdot x\| + \dots + \|H^{n+p} x\| \leq (q^n + \dots + q^{n+p}) \|x\| = \\ &= \frac{1 - q^{p+1}}{1 - q} q^n \|x\| < \frac{\|x\|}{1 - q} q^n. \end{aligned}$$

Для всякого $\varepsilon > 0$ последняя часть системы неравенств будет меньше ε при всех достаточно больших n и любых p , поэтому для рассматриваемого ряда признак Больцано — Коши выполняется. Ввиду же полноты пространства X ряд будет сходящимся и сумма его Ax будет некоторым элементом пространства X . Кроме того, $\|Ax\| \leq \frac{1}{1 - q} \|x\|$ и, стало быть, $\|A\| \leq \frac{1}{1 - q}$.

*) См. п. 4.5.1, т. I

Приведенные ниже равенства являются очевидными:

$$(I - H)Ax = (x + Hx + H^2x + \dots) - (Hx + H^2x + \dots) = x, \\ A(I - H)x = (x - Hx) + (Hx - H^2x) + \dots = x.$$

Они позволяют утверждать, что $A = (I - H)^{-1}$ и, значит, $\|(I - H)^{-1}\| = \|A\| \leq \frac{1}{1-q}$.

Докажем, что при принятых в теореме условиях приближения x_n могут быть построены для любых n , все x_n принадлежат шару (4) и для x_n выполняются неравенства

$$\|x_{n+1} - x_n\| \leq t_{n+1} - t_n \quad (n=0, 1, \dots). \quad (11)$$

Для этого прибегнем к индукции. Начальное приближение x_0 принадлежит шару (4), и по второму условию существует обратный оператор $\Gamma_0 = [f'(x_0)]^{-1}$. Поэтому может быть построено первое приближение $x_1 = x_0 - \Gamma_0 f(x_0)$. Далее ввиду (17) $\|x_1 - x_0\| = \|\Gamma_0 f(x_0)\| \leq \eta$. Из соотношения $t_1 - t_0 = -\frac{P(t_0)}{P'(t_0)} = -\frac{P(0)}{P'(0)} = \eta$ вытекает $\|x_1 - x_0\| \leq t_1 - t_0$, и неравенство (11) при $n=0$ выполняется.

Предположим, что приближения $x_0, x_1, \dots, x_n$ построены, все принадлежат шару (4) и для них выполняются неравенства

$$\|x_{k+1} - x_k\| \leq t_{k+1} - t_k \quad (k=0, 1, \dots, n-1). \quad (12)$$

Приближение x_{n+1} должно быть найдено по правилу (3). Мы убедимся в возможности его построения, если докажем существование оператора $[f'(x_n)]^{-1}$. Для этой цели воспользуемся леммой и предположением о существовании обратного оператора $\Gamma_0 = [f'(x_0)]^{-1}$.

Рассмотрим оператор $H = \Gamma_0 [f'(x_n) - f'(x_0)]$. Он отображает X на себя. Оценим его норму. Разность $f'(x_n) - f'(x_0)$ есть изменение оператора f' при переходе от элемента x_0 к x_n . Здесь может быть применено неравенство (1.19), где f должно быть заменено на f' , x на x_0 и Δx на $x_n - x_0$:

$$\|f'(x_n) - f'(x_0)\| \leq \\ \leq \sup_{0 \leq \theta \leq 1} \|f''[x_0 + \theta(x_n - x_0)]\| \cdot \|x_n - x_0\| \leq K \|x_n - x_0\|.$$

Воспользовавшись неравенствами (12), получим для $\|x_n - x_0\|$:

$$\begin{aligned}\|x_n - x_0\| &= \|(x_n - x_{n-1}) + (x_{n-1} - x_{n-2}) + \dots + (x_1 - x_0)\| \leq \\ &\leq (t_n - t_{n-1}) + (t_{n-1} - t_{n-2}) + \dots + t_1 - t_0 = t_n - t_0 = t_n;\end{aligned}$$

отсюда

$$\|f'(x_n) - f'(x_0)\| \leq Kt_n.$$

Для $\|H\|$ из цепочки неравенств

$$\begin{aligned}\|H\| &\leq \|\Gamma_0\| \cdot \|f'(x_n) - f'(x_0)\| \leq BKt_n < BKt^* = \\ &= BK\eta \frac{1 - \sqrt{1 - 2h}}{h} = 1 - \sqrt{1 - 2h} \leq 1\end{aligned}$$

получается оценка

$$\|H\| \leq BKt_n < 1.$$

По лемме оператор $I + H$ имеет обратный

$$(I + H)^{-1} = \{I + \Gamma_0[f'(x_n) - f'(x_0)]\}^{-1}$$

и верна оценка его нормы

$$\|(I + H)^{-1}\| \leq \frac{1}{1 - BKt_n}.$$

Положим $\Gamma_n = (I + H)^{-1} \Gamma_0$ и проверим, что это есть нужный нам оператор $[f'(x_n)]^{-1}$:

$$\begin{aligned}\Gamma_n &= (I + H)^{-1} \Gamma_0 = [I + \Gamma_0(f'(x_n) - f'(x_0))]^{-1} \Gamma_0 = \\ &= \{\Gamma_0^{-1} [I + \Gamma_0(f'(x_n) - f'(x_0))]\}^{-1} = \\ &= \{f'(x_0) + (f'(x_n) - f'(x_0))\}^{-1} = \{f'(x_n)\}^{-1}.\end{aligned}$$

Этим доказано существование оператора $\{f'(x_n)\}^{-1} = \Gamma_n$. Получим оценку его нормы

$$\|\{f'(x_n)\}^{-1}\| = \|\Gamma_n\| \leq \frac{1}{1 - BKt_n} B = \left(\frac{1}{B} - Kt_n\right)^{-1} = -\frac{1}{P'(t_n)}.$$

Найдем еще оценку нормы $f(x_n)$. Приближение x_n находится по правилу Ньютона (3)

$$x_n = x_{n-1} - [f'(x_{n-1})]^{-1} f(x_{n-1})$$

или

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0.$$

Поэтому для $f(x_n)$ верно представление

$$f(x_n) = [f(x_n) - f(x_{n-1})] - f'(x_{n-1})(x_n - x_{n-1}).$$

К полученной разности между изменением оператора f и главной линейной частью изменения могут быть применены теорема 3 из § 1 и неравенство (1.20). Это приводит к следующему результату:

$$\begin{aligned} \|f(x_n)\| &\leq \frac{1}{2} \max_{0 \leq \theta \leq 1} \|f''[x_{n-1} + \theta(x_n - x_{n-1})]\| \cdot \|x_n - x_{n-1}\|^2 \leq \\ &\leq \frac{1}{2} K(t_n - t_{n-1})^2. \end{aligned}$$

Заметим, что если бы рассматривалось не операторное уравнение $f(x) = 0$, а численное уравнение $P(t) = 0$ и мы хотели бы оценить $P(t_n)$, то на этом пути для $P(t_n)$ нашли бы точное значение

$$\begin{aligned} P(t_n) &= P(t_n) - [P(t_{n-1}) + P'(t_{n-1})(t_n - t_{n-1})] = \\ &= \frac{1}{2} P''(t_{n-1})(t_n - t_{n-1})^2 = \frac{1}{2} K(t_n - t_{n-1})^2. \end{aligned}$$

Для $f(x_n)$ будем иметь оценку

$$\|f(x_n)\| \leq P(t_n).$$

Все полученное дает возможность оценить $\|x_{n+1} - x_n\|$:

$$\|x_{n+1} - x_n\| = \| [f'(x_n)]^{-1} \cdot f(x_n) \| \leq -\frac{P(t_n)}{P'(t_n)} = t_{n+1} - t_n.$$

Неравенство (11) для x_{n+1} действительно верно. Принадлежность x_{n+1} шару (4) проверяется просто:

$$\begin{aligned} \|x_{n+1} - x_0\| &\leq \|x_{n+1} - x_n\| + \|x_n - x_{n-1}\| + \dots + \|x_1 - x_0\| \leq \\ &\leq (t_{n+1} - t_n) + (t_n - t_{n-1}) + \dots + (t_1 - t_0) = t_{n+1} - t_0 = \\ &= t_{n+1} < t^* = \frac{1 - \sqrt{1 - 2h}}{h} \eta \leq \delta. \end{aligned}$$

Для доказательства сходимости x_n ($n \rightarrow \infty$) применим признак Больцано — Коши

$$\begin{aligned} \|x_{n+p} - x_n\| &\leq \\ &\leq \|x_{n+p} - x_{n+p-1}\| + \|x_{n+p-1} - x_{n+p-2}\| + \dots + \|x_{n+1} - x_n\| \leq \\ &\leq (t_{n+p} - t_{n+p-1}) + (t_{n+p-1} - t_{n+p-2}) + \dots + (t_{n+1} - t_n) = t_{n+p} - t_n, \\ &\|x_{n+p} - x_n\| \leq t_{n+p} - t_n. \end{aligned} \quad (13)$$

Так как последовательность t_n сходится, для нее признак Больцано — Коши выполняется. Из последнего

неравенства следует, что признак выполняется и для последовательности x_n . Ввиду полноты пространства X последовательность x_n является также сходящейся: $\lim x_n = x^*$. Все x_n принадлежат шару $\|x - x_0\| \leq \delta$, поэтому x^* также принадлежит ему.

Утверждение о скорости сходимости получится из (13), если там фиксировать n и увеличивать неограниченно p . В пределе получится неравенство (10).

Утверждение о том, что предельный элемент x^* есть решение уравнения $f(x) = 0$, можно доказать, воспользовавшись правилом Ньютона (3). Запишем его в виде

$$f(x_{n+1}) + f'(x_n)(x_{n+1} - x_n) = 0.$$

Будем считать n неограниченно возрастающим. Так как $x_n \rightarrow x^*$ и $x_{n+1} \rightarrow x^*$, то $\|x_{n+1} - x_n\| \rightarrow 0$. $f(x)$ и $f'(x)$ предполагались дифференцируемыми всюду в шаре, в частности на элементе x^* . Поэтому они будут непрерывными на x^* и в пределе последнее равенство даст $f(x^*) = 0$, что доказывает утверждение.

Замечание. Если воспользоваться оценкой $t^* - t_n$, указанной в начале доказательства теоремы, можно неравенство (10) для оценки скорости сходимости заменить другим, более грубым, но более наглядным:

$$\|x^* - x_n\| \leq \frac{1}{2^{n-1}} (2h)^{2^n - 1} \cdot \eta. \quad (14)$$

§ 4. Задача упрощения уравнений

Умение заменять сложную задачу на «близкую» к ней более простую позволяет в многих случаях очень сильно расширить круг разрешимых проблем. Возможность такой замены особенно важна при использовании вычислительных машин, так как тогда всякая задача должна быть приведена только к операциям с числами, в частности уравнение любого вида должно быть приведено к численному уравнению или к конечной системе таких уравнений.

1. Введение. Предположим, что X и Y есть два любых множества, элементы которых будем обозначать соответственно x и y . Пусть на X задан оператор $y =$

$= f(x)$, значения которого принадлежат Y . Возьмем на Y произвольный элемент y^* и рассмотрим уравнение

$$f(x) = y^*. \quad (1)$$

Ниже его мы будем называть точным уравнением.

Допустим теперь, что с целью упрощения уравнения множества X и Y заменены на X_n и Y_n , при этом преобразования $X \rightarrow X_n$ и $Y \rightarrow Y_n$ выполняются операторами φ_n и ψ_n , которые часто называют *операторами сноса*.

Наконец, предположим, что точное уравнение (1) заменяется «ближким» в каком-то смысле к нему приближенным уравнением

$$f_n(x_n) = y_n^* = \psi_n y^*. \quad (2)$$

Здесь $x_n \in X_n$, $y_n^* \in Y_n$ и оператор f_n выполняет преобразования X_n в Y_n .

После того, как найдено решение x_n^* приближенного уравнения (2), необходимо по нему построить элемент $x^{(n)}$ во множестве X , который будет принят за приближение к решению x^* точного уравнения. Пусть необходимое преобразование X_n в X выполняется оператором $\bar{\varphi}_n$. Принято называть x_n^* *каркасом* или *скелетом приближения* $x^{(n)}$, а $\bar{\varphi}_n$ — *оператором восполнения*.

Схема действий операторов и множеств указана на рис. 10.

Нас будут интересовать следующие основные вопросы. Известно, что точное уравнение (1) имеет на некоторой части множества X единственное решение x^* . При каких условиях приближенное уравнение (2) будет иметь решение x_n^* ? В каких случаях последовательность приближений $x^{(n)}$ будет сходиться к точному решению x^* уравнения (1), и если сходимость имеет место, то как может быть оценена погрешность $x^* - x^{(n)}$ приближений?

Результаты исследований здесь могут быть получены только при некоторых предположениях о множествах X , Y , X_n , Y_n и об операторах f , f_n , φ_n , ψ_n , $\bar{\varphi}_n$.

Ниже всюду будем считать множества X , Y , X_n , Y_n линейными, нормированными и полными пространствами, а операторы f , f_n , φ_n , ψ_n , $\bar{\varphi}_n$ — линейными.

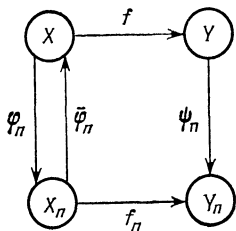


Рис. 10.

2. О разрешимости приближенного уравнения. Предварительно докажем следующую лемму о разрешимости уравнения.

Лемма 1. Пусть X и Y есть пространства типа Банаха и A — линейный оператор из X в Y .

Если для каждого $y \in Y$ существует элемент $x \in X$ такой, что

$$\|Ax - y\| \leq q\|y\|, \quad \|x\| \leq N\|y\|, \quad (3)$$

где q и N есть некоторые числа, не зависящие от y , причем $q < 1$, то уравнение

$$Ax = y \quad (4)$$

имеет при всяком $y \in Y$ решение $x \in X$ и выполняется неравенство

$$\|x\| \leq \frac{N}{1-q} \|y\|. \quad (5)$$

Перед доказательством заметим, что лемме может быть придано следующее истолкование. Элемент x можно рассматривать как приближенное решение уравнения (4). В лемме утверждается, что если при каждом $y \in Y$ уравнение (4) имеет приближенное решение, для которого выполняются неравенства (3), то уравнение (4) при любом $y \in Y$ имеет точное решение и для него выполняется неравенство (5).

Доказательство. Пусть y — элемент из Y . Положим $y_1 = y$ и построим последовательность элементов y_n по следующему правилу. По предположению, для y_1 существует элемент x_1 такой, что

$$\|Ax_1 - y_1\| \leq q\|y_1\|, \quad \|x_1\| \leq N\|y_1\|.$$

Положим $y_2 = y_1 - Ax_1$. В силу предположения, сделанного в лемме, для y_2 существует элемент x_2 , для которого будут верны неравенства

$$\|Ax_2 - y_2\| \leq q\|y_2\| \leq q^2\|y_1\|, \quad \|x_2\| \leq N\|y_2\| \leq Nq\|y_1\|$$

и т. д. В результате получим две последовательности элементов x_k, y_k таких, что

$$y_{k+1} = y_k - Ax_k, \\ \|y_k\| \leq q^{k-1}\|y_1\|, \quad \|x_k\| \leq Nq^{k-1}\|y_1\|.$$

Если просуммировать почленно равенства $y_{k+1} = y_k - Ax_k$ ($k = 1, 2, \dots, n$), получим следующее представление y_{n+1} :

$$y_{n+1} = y_1 - A(x_1 + \dots + x_n). \quad (6)$$

Ряд $x_1 + \dots + x_n + \dots$ сходится ввиду оценки $\|x_k\| \leq Nq^{k-1}\|y_1\|$ и полноты пространства X . Обозначим его сумму через x . Левая часть равенства y_{n+1} будет стремиться к нулевому элементу, так как $\|y_{n+1}\| \leq q^n\|y_1\|$. В пределе из (6) получится

$$0 = y - Ax.$$

Поэтому x есть точное решение уравнения (4). Кроме того, очевидно,

$$\|x\| \leq \sum_{k=1}^{\infty} \|x_k\| \leq N \sum_{k=1}^{\infty} q^{k-1} \|y_1\| = \frac{N}{1-q} \|y_1\|.$$

Лемма доказана.

Будем, как указано выше, считать, что точное уравнение

$$Ax = y \quad (7)$$

и приближенное

$$A_n x_n = y_n \quad (8)$$

являются линейными. Оба оператора сноса φ_n , ψ_n и оператор восполнения $\bar{\varphi}_n$ также предположим линейными, пространства же X , Y , X_n , Y_n — линейными, нормированными и полными.

Допустим, что оператор A точного уравнения имеет обратный A^{-1} и $x = A^{-1}y$. Надеяться на разрешимость приближенного уравнения (8) можно только в том случае, если оно окажется близким в каком-то смысле (еще подлежащем уточнению) к точному уравнению. Близость зависит от операторов φ_n , ψ_n сноса и уравнения (8). За локальную погрешность, порождаемую этими операторами при обработке элемента x , может быть принята величина

$$\varepsilon_n(x) = \|A_n \varphi_n x - \psi_n Ax\|_{Y_n}. \quad (9)$$

Схема ее получения указана на рис. 11.

Нам потребуется также предположение о характере сноса $Y \rightarrow Y_n$ оператором ψ_n . Будем считать, что для каждого $y_n \in Y_n$ существует элемент y из Y такой, что выполняются неравенства

$$\begin{aligned} \|\psi_n y - y_n\| &\leq \alpha_n \|y_n\|_{Y_n}, \\ \|y\|_Y &\leq \beta_n \|y_n\|_{Y_n}, \end{aligned} \quad (10)$$

где α_n и β_n есть некоторые числа, не зависящие от элемента y_n .

Теорема 1. Если для точного и приближенного уравнений (7) и (8) выполняются условия:

- 1) для оператора A существует обратный A^{-1} ;
- 2) существует число γ_n , не зависящее от элемента x , такое, что для погрешности $\varepsilon_n(x)$ выполняется неравенство

$$\varepsilon_n(x) \leq \gamma_n \|x\|_X; \quad (11)$$

- 3) для оператора ψ_n , выполняющего снос $Y \rightarrow Y_n$, верно предположение (10);

- 4) для чисел α_n , β_n , γ_n и $\|A^{-1}\|$ справедливо неравенство

$$q = \alpha_n + \beta_n \gamma_n \|A^{-1}\| < 1; \quad (12)$$

тогда приближенное уравнение (8) имеет решение x_n при всяком элементе y_n из Y_n и для нормы x_n верна оценка

$$\|x_n\|_{X_n} \leq \frac{N}{1-q} \|y_n\|_{Y_n}, \quad N = \beta_n \|\varphi_n A^{-1}\|. \quad (13)$$

Доказательство. Покажем, что для уравнения (8) выполняются условия леммы 1. Возьмем любой элемент y_n , принадлежащий Y_n . По третьему предположению теоремы существует элемент $y \in Y$, для которого выполнены неравенства (10). По y построим элемент $x = A^{-1}y$, затем положим $x_n = \varphi_n x = \varphi_n A^{-1}y$. Проверим для пары элементов x_n , y_n первое неравенство условия (3) леммы 1:

$$\begin{aligned} \|A_n x_n - y_n\| &= \|(A_n \varphi_n x - \psi_n A x) + (\psi_n y - y_n)\|_{Y_n} \leq \\ &\leq \|A_n \varphi_n x - \psi_n A x\|_{Y_n} + \|\psi_n y - y_n\|_{Y_n} \leq \varepsilon_n(x) + \alpha_n \|y_n\|_{Y_n}. \end{aligned}$$

Ввиду того, что $\varepsilon_n(x) \leq \gamma_n \|x\|_X$ и

$$\|x\|_X = \|A^{-1}y\|_X \leq \|A^{-1}\| \cdot \|y\|_Y \leq \|A^{-1}\| \beta_n \|y_n\|_{Y_n},$$

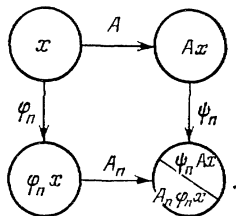


Рис. 11.

для $\|A_n \mathbf{x}_n - \mathbf{y}_n\|_{Y_n}$ получим оценку

$$\|A_n \mathbf{x}_n - \mathbf{y}_n\|_{Y_n} \leq (\beta_n \gamma_n \|A^{-1}\| + \alpha_n) \|\mathbf{y}_n\|_{Y_n} = q \|\mathbf{y}_n\|_{Y_n}.$$

Так как $q < 1$ по условию (12), первое из неравенств (3) действительно выполняется.

Второе неравенство (3) проверяется так же просто. Ввиду $\mathbf{x}_n = \varphi_n \mathbf{x} = \varphi_n A^{-1} \mathbf{y}$ получим

$$\|\mathbf{x}_n\|_{X_n} \leq \|\varphi_n A^{-1}\| \cdot \|\mathbf{y}\|_Y \leq \beta_n \|\varphi_n A^{-1}\| \cdot \|\mathbf{y}_n\|_{Y_n}.$$

Отсюда следует, что второе неравенство (3) выполняется при

$$N = \beta_n \|\varphi_n A^{-1}\|.$$

Утверждения теоремы 1 получаются, если к уравнению (8) применить лемму 1.

3. Оценка погрешности приближения и условия сходимости. Предположим, что точное уравнение

$$A\mathbf{x} = \mathbf{y}^* \quad (14)$$

имеет решение $\mathbf{x}^*$, что же касается приближенного уравнения

$$A_n \mathbf{x}_n = \mathbf{y}_n^* = \varphi_n \mathbf{y}^*, \quad (15)$$

то будем считать, что оператор A_n при всех $n = 1, 2, \dots$ имеет обратный A_n^{-1} . Решением приближенного уравнения является

$$\mathbf{x}_n^* = A_n^{-1} \mathbf{y}_n^*.$$

По нему на шаге номера n с помощью оператора выполнения $\bar{\varphi}_n$ строится приближение номера n к точному решению

$$\mathbf{x}^{(n)} = \bar{\varphi}_n \mathbf{x}_n^* = \bar{\varphi}_n A_n^{-1} \mathbf{y}_n^*.$$

Нас интересует погрешность приближения $\mathbf{x}^* - \mathbf{x}^{(n)}$. Ниже мы оценим ее норму $\rho_n = \|\mathbf{x}^* - \mathbf{x}^{(n)}\|_X$ и получим условия, при которых имеет место сходимость $\mathbf{x}^{(n)} \rightarrow \mathbf{x}^*$ в норме пространства X :

$$\begin{aligned} \rho_n &= \|\mathbf{x}^* - \mathbf{x}^{(n)}\|_X = \|\mathbf{x}^* - \bar{\varphi}_n \mathbf{x}_n^*\|_X = \\ &= \|(\mathbf{x}^* - \bar{\varphi}_n \varphi_n \mathbf{x}^*) + (\bar{\varphi}_n \varphi_n \mathbf{x}^* - \bar{\varphi}_n \mathbf{x}_n^*)\|_X \leq \\ &\leq \|\mathbf{x}^* - \bar{\varphi}_n \varphi_n \mathbf{x}^*\|_X + \|\bar{\varphi}_n\| \cdot \|\varphi_n \mathbf{x}^* - \mathbf{x}_n^*\|_X. \end{aligned} \quad (16)$$

Первая из величин, стоящая в последней части системы соотношений, имеет наглядный смысл: разность $\mathbf{x}^* - \bar{\varphi}_n \varphi_n \mathbf{x}^*$ есть, очевидно, несогласованность операций сноса φ_n и восполнения $\bar{\varphi}_n$ на решении. Норма разности есть мера этой несогласованности. Она зависит от точного решения $\mathbf{x}^*$, нам неизвестного, и должна быть оценена в каждой частной задаче на основании дополнительных сведений. Нередко бывает известен порядок малости этой величины.

Во второй член входит разность $\varphi_n \mathbf{x}^* - \mathbf{x}_n^*$, которая характеризует несогласованность на точном решении $\mathbf{x}^*$ между операцией сноса φ_n и приближенным уравнением (15). Она может быть оценена через величину $\varepsilon_n(\mathbf{x})$, введенную раньше в (9). Ввиду $A_n \mathbf{x}_i^* = \psi_n \mathbf{y}^* = \psi_n A \mathbf{x}^*$ будем иметь

$$\begin{aligned} \|\varphi_n \mathbf{x}^* - \mathbf{x}_n^*\|_X &= \|A_n^{-1} (A_n \varphi_n \mathbf{x}^* - A_n \mathbf{x}_n^*)\|_{X_n} = \\ &= \|A_n^{-1} (A_n \varphi_n \mathbf{x}^* - \psi_n A \mathbf{x}^*)\|_{X_n} \leq \\ &\leq \|A_n^{-1}\| \cdot \|A_n \varphi_n \mathbf{x}^* - \psi_n A \mathbf{x}^*\|_{Y_n} = \|A_n^{-1}\| \varepsilon_n(\mathbf{x}^*). \end{aligned}$$

В силу (16) это приводит к следующей оценке нормы погрешности:

$$\|\mathbf{x}^* - \mathbf{x}^{(n)}\|_X \leq \|\mathbf{x}^* - \bar{\varphi}_n \varphi_n \mathbf{x}^*\|_X + \|\bar{\varphi}_n\| \cdot \|A_n^{-1}\| \cdot \varepsilon_n(\mathbf{x}^*). \quad (17)$$

Величина $\varepsilon_n(\mathbf{x}^*)$ также зависит от неизвестного точного решения $\mathbf{x}^*$ и должна быть оценена для каждого конкретного типа задач. Порядок ее малости часто может быть указан.

Оценка (17) позволяет сформулировать теорему о сходимости.

Теорема 2. Если для приближенного уравнения (15) и операторов сноса и восполнения φ_n , ψ_n , $\bar{\varphi}_n$ выполняются условия:

1) мера несогласованности на точном решении $\mathbf{x}^*$ операторов сноса φ_n и восполнения $\bar{\varphi}_n$ стремится к нулю при $n \rightarrow \infty$

$$\|\mathbf{x}^* - \bar{\varphi}_n \varphi_n \mathbf{x}^*\|_X \rightarrow 0;$$

2) для погрешности $\varepsilon_n(\mathbf{x}^*)$ верно соотношение

$$\|\bar{\varphi}_n\| \cdot \|A_n^{-1}\| \cdot \varepsilon_n(\mathbf{x}^*) \rightarrow 0, \quad n \rightarrow \infty;$$

тогда имеет место сходимость в норме пространства X приближенного решения $x^{(n)}$ к точному решению x^* :

$$\|x^* - x^{(n)}\| \rightarrow 0, \quad n \rightarrow \infty.$$

Значение оценки (17) и теоремы 2 состоит преимущественно в том, что они выясняют строение погрешности и указывают направление исследований, которые должны быть проведены в конкретных задачах.

ЛИТЕРАТУРА

1. Гавурин М. К., Лекции по методам вычислений, М., «Наука», 1971.
2. Канторович Л. В., Функциональный анализ и прикладная математика, М., УМН 3, вып. 6(28), 1948.
3. Коллатц Л., Функциональный анализ и вычислительная математика, М., ИЛ, 1971.

УКАЗАТЕЛЬ ОБОЗНАЧЕНИЙ

$\approx$ 14	$C[a, b]$ 121, 361	$\gg$ 189
$[k]$	$C_2[a, b]$ 121	$\ A\ _{III}$ 190
$y_{n+\alpha}$ 16	$L_2[a, b]$ 129	$\sim$ 290
$[k]$	D_h 143	$Res(f(u))$ 322
$f_{n+\alpha}$ 16	U_h 143	$\rho(x, y)$ 359
$[i, j]$	F_h 143	$\rho_I(x, y)$ 360
$f_{n+\alpha}$ 40	$\ \cdot\ _{U_h}$ 143	$\rho_m(x, y)$ 360
$[\pm]$	$\ \cdot\ _{F_h}$ 143	$\rho_{II}(x, y)$ 360
y_n 63	$u^{(i)}$ 143	$\rho_s(x, y)$ 360
$[\pm]$	$f^{(h)}$ 143	$\rho_{III}(x, y)$ 360
$y_{n+\frac{[\pm]}{\alpha_i}}$ 64	$\left. \frac{\partial u}{\partial x} \right (x_m, t_n)$ 147	$\rho_l(x, y)$ 360
$[k\pm]$	Λ_{xx} 157	$L_p[a, b]$ 362
y_n 67	Λ_{yy} 157	$[X \rightarrow Y]$ 365
$[k\pm]$	$L_h(u^{(h)})$ 160	$\ A\ $ 366
f_n 67	$\Lambda_h(u^{(h)})$ 184	$\ \cdot\ _m$ 367
v^T 95		$[X \rightarrow [X \rightarrow Y]]$ 367
L_h 109		
$R_k(h)$ 109		
Λ_h 111		

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Аддитивный оператор 362
Аксиома различия 359
— симметрии 359
— треугольника 359
Аксиомы линейного пространства 361
— метрики 359
— нормы 361
Аналитический метод 9
— тип сходимости 309
Асимптотический степенной тип сходимости 308
Аппроксимация граничных условий 180
— оператора 142
— разностной схемы 142, 176

Гармонический синтез 290
Главный член погрешности метода 56

Двухточечная граничная задача 94

Задача граничная линейная 107
— — многоточечная 95
— — нелинейная 100
— — смешанная 159, 202
— корректно поставленная 255
— Коши 146
— многоточечная 95
— некорректно поставленная 255

Интегральные уравнения Вольтерра 253
— — — второго рода 254
— — — первого рода 254
— — Фредгольма 253
— — — второго рода 254
— — — первого рода 254
Интерполяционное преобразование с возрастающим числом узлов 314
— — — постоянным числом узлов 301
Интерполяционный квадратурный метод 275

Каркас приближенного решения 388
Квазистационарная последовательность 295
Корректно поставленная задача 255
Кубическая метрика 360

Метод аналитический 9
— Галеркина 120, 248
— гиперплоскостей 230
— дифференциальной прогонки 95
— замены ядра интегрального уравнения на вырожденное 278
— интегральных соотношений обобщенный 246
— касательных 16
— k -го порядка точности 51
— линеаризации 104

- Метод ломаных 16
 — матричной прогонки 188
 — многошаговый 12, 72
 — моментов 122, 248
 — наименьших квадратов 129
 — неявный 12, 73
 — Ньютона для операторных уравнений 380
 — одношаговый 12
 — ортогональной прогонки 113
 — Пикара 10
 — плоскостей 230
 — последовательных приближений 10
 — прямых 228
 — — для гиперболических уравнений 238
 — — — параболических уравнений 229
 — — — эллиптических уравнений 244
 — Рунге 136
 — рядов 11
 — сеток для решения линейной граничной задачи 107
 — трапеций 22, 36
 — установления 192
 — характеристик 215
 — Чаплыгина 59
 — явный 12, 73
 Методы Адамса интерполяционные 74, 79
 — — экстраполяционные 74
 — с забеганием вперед 12, 73, 85
 — типа двусторонних 59
 — — Коуэлла 86
 Метрика 359
 — кубическая 360, 377
 — октаэдрическая 360, 378
 — сферическая 360
 — шаровая 379
 Метрическое пространство 358
 Минимизирующая последовательность функций 137
 Многоточечная задача 95

 Некорректно поставленные задачи 255
 Необходимое спектральное условие устойчивости Неймана 164

 Несогласованность операторов снос и восполнения 393
 Неявная двухслойная шеститочечная разностная аппроксимация 173
 — пятиточечная трехслойная схема 174
 — разностная схема 151
 Неявный метод установления 194
 Норма 361
 — оператора 364
 — — билинейного 366
 Нормированное линейное пространство 361

 Обобщенный метод интегральных соотношений 246
 Обратный метод прогонки 170
 — оператор 364
 Октаэдрическая метрика 360
 Оператор аддитивный 362
 — билинейный 366
 — восполнения 388
 — дифференцируемый на элементе пространства 368
 — линейный 363
 — непрерывный 362
 — обратный 364
 — ограниченный 362
 — однородный 362
 — разностный 109
 — сноса 388
 Оператора норма 364
 Оценка апостериорная 53
 — априорная 53
 — погрешности метода 51
 — — приближенного решения 51

 Погрешность аппроксимации 144
 — — дифференциального оператора 109
 — вычислительная 51
 — локальная 15
 — метода 25, 52
 — начального условия 44

- Погрешность неустраняемая
51
— округлений 44
— приближенного решения 44
— формулы 44
Полное метрическое пространство 360
Полоска первого порядка 217
Поперечная схема метода прямых 229
Порядок точности метода 25
Правило ломаных 16
— Ньютона для операторных уравнений 380
— Рунге 57, 191
— Рунге — Кутты 31
Преобразование последовательности регулярное 289, 293
— —, сохраняющее предел 293
— — — сходимость 293
— Ценкса 336
— Эйткена 326
Принцип максимума 112
— сжатых отображений 374
Продольная схема метода прямых 234
Производная от оператора 368
Простой метод интегральных соотношений 246
Пространство линейное 361
— линейных операторов 365
— полное 360
— типа Банаха 361
Прямой метод прогонки 170
- Разностная схема 110, 143
— — неявная 151
— — расщепления 155
— — устойчивая 144
— — экономичная 167
— — явная 151
Разностный оператор 109
Реальная область асимптотики 57
Регулярное преобразование 293
- Сетка 108, 143
— равномерная 108
Сеточная функция 143
- Система Чебышева 297
Скелет приближения 388
Смешанная граничная задача 202
Сферическая метрика 360
Схема Дюфорта и Франкела 173
— Кранка и Николсона 172
— разностная 110, 143
Схемы расщепления 155
Сходимость аналитического типа 309
— асимптотического степенного типа 308
— метода Рунге 138
— — сеток 117
— одношаговых методов 42
— почти алгебраического типа 308
— разностных схем 144
- Теорема о сохранении предела 295
— — — сходимости 294
Треугольник определенности дифференциального уравнения 207
— — разностной схемы 207
- Узел внутренний 177
— граничный 177
— сетки 108, 142
Уравнения направления характеристик 219
— характеристик 219
— характеристических полосок 219
Устойчивость разностных схем 144
— схемы по начальным данным 163
- Факториальный многочлен 353
Функция влияния значения свободного члена 304
— — начального значения 304
— Грина 304

- Характеристика 218
Характеристическая кривая 218
— полоска 218
- Шаблон 148
- Экономичные разностные схемы 167
Экстраполирование 75
- Экстерполяционные методы
Адамса 74
Экстерполяция 75
- Явная разностная схема 151
Явные вычислительные методы 73
Явный метод установления 194
Ядро 255
— вырожденное 278
В-пространство 361
 δ^2 -преобразование 326

*Владимир Иванович Крылов
Владимир Васильевич Бобков
Петр Ильич Монастырский*

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ
Т о м II

М., 1977 г., 400 стр. с илл.

Редакторы *Е. Ю. Ходан, Е. В. Шикин.*

Техн. редактор *С. Я. Шкляр.*

Корректор *Е. В. Сидоркина.*

Сдано в набор 01.11. 1976 г. Подписано к печати 19.04. 1977 г. Бумага 84×108¹/₃₂ тип. № 3. Физ. печ. л. 12,5. Условн. печ. л. 21. Уч.-изд. л. 20,74. Тираж 37 000 экз. Цена книги 99 коп. Заказ № 947.

Издательство «Наука»
Главная редакция
физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

Ордена Трудового Красного Знамени Ленинградское производственно-техническое объединение «Печатный Двор» имени А. М. Горького Союзполиграфпрома при Государственном комитете Совета Министров СССР по делам издательств, полиграфии и книжной торговли. 197136, Ленинград, П-136, Гатчинская ул., 26.