И. И. Ляшко
В. Л. Макаров
А. А. Скоробогатько

МЕТОДЫ ВЫЧИСЛЕНИЙ

И. И. Ляшко
В. Л. Макаров
А. А. Скоробогатько

МЕТОДЫ ВЫЧИСЛЕНИЙ

(Численный анализ. Методы решения задач математической физики)

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

КИЕВ
ГОЛОВНОЕ ИЗДАТЕЛЬСТВО
ИЗДАТЕЛЬСКОГО ОБЪЕДИНЕНИЯ
«ВИЩА ШКОЛА»
1977

530.1
Л99

УДК 518.12(075.8)

Методы вычислений. (Численный анализ. Методы решения задач математической физики). Ляшко И. И., Макаров В. Л., Скоробогатько А. А. Киев, Издательское объединение «Вища школа», 1977, 408 с.

В учебном пособии освещены численные методы математики, применяемые для решения различных задач с помощью современных вычислительных машин. Рассматриваются общие вопросы численного анализа, численные методы решения задач алгебры, проекционные и разностные методы решения задач математической физики.

Предназначено для студентов вузов, обучающихся по специальности «Прикладная математика», а также может быть использовано аспирантами и инженерами, работающими в области прикладной математики.

Табл. 7. Ил. 13. Список лит.: 98 назв.

Редакция литературы по кибернетике, электронике и энергетике

Зав. редакцией *А. В. Дьячков*

Л $\frac{20204-283}{М\,211\,(04)-77}$ БЗ—1—7—77

© Издательское объединение «Вища школа», 1977.

ПРЕДИСЛОВИЕ

Настоящее учебное пособие написано в соответствии с программой курса «Методы вычислений», читаемого авторами на протяжении нескольких последних лет в Киевском государственном университете на факультете кибернетики. Пособие содержит также ряд вопросов, которые не входят в программу и могут быть использованы в специальных курсах.

Пособие состоит из двух частей и приложения. Первая часть посвящена аппроксимации линейных операторов. В ней приводятся наиболее распространенные постановки задач аппроксимации линейных операторов, доказательства ряда теорем о наилучших приближениях в нормированных пространствах. В изложении теории интерполирования вместо традиционной диаграммы Фрезера приведен чисто аналитический способ получения интерполяционных формул. В главе, посвященной среднеквадратическим и равномерным приближениям, используется «геометрический» язык гильбертовых пространств, позволяющий более тесно связать равномерные приближения с общей теорией приближения функций в нормированных пространствах. При выводе и исследовании квадратурных формул, в том числе и формулы Эйлера, используется подход, основанный только на теории интерполирования, что позволило логически объединить все приводимые в книге квадратурные формулы.

Вторая часть учебного пособия посвящена приближенным методам решения операторных уравнений. Основное внимание при этом уделяется изложению приближенных методов решения задач математической физики. Центральное место здесь отводится различным способам построения и исследования устойчивости разностных схем. В основу изложения положены работы А. А. Самарского и его учеников.

Большое внимание уделено примерам построения конкретных разностных схем. Частично затронуты вопросы построения разностных схем для бесконечных областей, основанные на использовании регуляризирующих операторов.

Авторы сочли целесообразным остановиться более подробно на одном из прямых методов решения конечно-разностных уравнений, разработанном в Киевском университете и получившем название метода суммарных представлений. Как и в любом прямом методе, его

применение связано с построением явных формул, по которым находится решение разностной задачи. Вид разностного уравнения играет существенную роль при построении таких формул. Идея метода суммарных представлений излагается на примере краевых задач, связанных с дифференциальными уравнениями эллиптического типа в канонических областях.

В этой же части рассматриваются различные итерационные методы решения сеточных уравнений. Поскольку в большинстве итерационных методов конкретная структура полученной системы не используется, то теорию исследования сходимости этих методов можно строить с единой точки зрения. Идеи функционального анализа позволяют перенести основные результаты на операторные уравнения более общего вида и существенно упростить общий подход при доказательстве их сходимости. Основное внимание при этом уделено одношаговым и двухшаговым методам решения линейных операторных уравнений.

Среди итерационных методов решения нелинейных уравнений излагаются основные идеи метода последовательных приближений, метода Ньютона, продолжения решения по параметру и метода спуска. Определенное внимание уделено построению одношаговых и многошаговых методов решения задачи Коши для обыкновенных дифференциальных уравнений и проекционным методам решения операторных уравнений.

Как справочный материал в приложении приводятся некоторые сведения из функционального анализа, теории специальных функций, дискретного анализа, необходимые при изучении вопросов, изложенных в настоящем пособии.

Авторы выражают благодарность доктору физ.-мат. наук проф. А. Н. Костовскому, доктору техн. наук проф. Я. М. Григоренко за полезные советы и замечания, способствовавшие улучшению рукописи.

Отзывы и пожелания просим направлять по адресу: 252054, Киев, 54, ул. Гоголевская, 7, Головное издательство издательского объединения «Вища школа».

Часть I

АППРОКСИМАЦИЯ ЛИНЕЙНЫХ ОПЕРАТОРОВ

Глава 1

ОБЩИЕ ВОПРОСЫ АППРОКСИМАЦИИ ЛИНЕЙНЫХ ОПЕРАТОРОВ

§ 1. ПОСТАНОВКА ЗАДАЧ АППРОКСИМАЦИИ ЛИНЕЙНЫХ ОПЕРАТОРОВ

Многие задачи вычислительной математики можно интерпретировать следующим образом. Пусть $\mathbf{B}$ — банахово пространство, $\mathbf{Y}$ — линейное нормированное пространство над полем вещественных чисел R_1 и пусть заданы линейные операторы $F : \mathbf{B} \to \mathbf{Y}$ и $F_n : \mathbf{B} \to \mathbf{Y}$, причем такие, что $\forall f \in \mathbf{B}$ можно достаточно просто находить $F_n(f)$, $n = 0, 1, \ldots$.

Общая задача аппроксимации оператора F состоит в нахождении такой последовательности операторов F_n, что $\forall \varepsilon > 0$ существует такой номер n, для которого выполняется неравенство

$$\| F(f) - F_n(f) \| \leqslant \varepsilon, \qquad \forall f \in \mathbf{B}.$$

Оператором F_n и заменяется приближенно оператор F.

Аналогично ставится аппроксимационная задача, когда оператор F не заменяется, а приближение достигается за счет замены f элементом более «простого» множества M, являющегося подмножеством $\mathbf{B}$, т. е. $M \subset \mathbf{B}$.

Частными случаями задачи аппроксимации линейных операторов являются: интерполирование; приближение функций; численное интегрирование; численное дифференцирование.

Конкретизируем общую задачу аппроксимации линейных операторов применительно к каждому из перечисленных выше случаев.

ЗАДАЧА ТЕОРИИ ИНТЕРПОЛИРОВАНИЯ

Пусть $\mathbf{B} = C(S)$ — система непрерывных действительных функций, заданных на компакте S, с обычным определением сложения функций и умножения их на действительные числа, с чебышевской нормой $\|f\| = \max\limits_{x \in S} |f(x)|$. Пусть $x_i \in S$, $i = 0, 1, 2, \ldots, n$ — различные точки (узлы), в которых известны значения функции $f(x)$. Положим $F(f) = f(x)$ и введем следующие определения:

Определение 1. Рассмотрим конечную или счетную совокупность линейно независимых достаточно простых функций $\{\varphi_i(x)\} \in \mathbf{B}$.

Возьмем первые $(n+1)$ элементов из $\{\varphi_i(x)\}$. Всевозможные линейные комбинации

$$\varphi(x) = \sum_{i=0}^{n} a_i \varphi_i(x) \tag{1}$$

с действительными коэффициентами a_i назовем *обобщенными многочленами* по системе $\varphi_i(x)$, $i = 0, 1, \ldots, n$.

Обобщенные многочлены вида (1) образуют линейное подмножество $M_n \subset \mathbf{B}$. Поставим задачу приближения функционала $F(f) = f(x)$ в фиксированной точке $x \neq x_i$ обобщенными многочленами следующим образом: найти обобщенный многочлен $\varphi(x) \in M_n$, такой, что

$$f(x_i) = \varphi(x_i), \qquad i = 0, 1, \ldots, n. \tag{2}$$

Если такой обобщенный многочлен можно построить, то полагаем $F_n(f) = \varphi(x)$, $f(x) \approx \varphi(x)$, а $\varphi(x)$ называем *обобщенным интерполяционным многочленом* для функции $f(x)$.

Для того чтобы поставленная задача решалась однозначно, необходимо наложить на систему функций $\{\varphi_i(x)\}$ дополнительные ограничения.

Определение 2. Назовем систему функций $\varphi_i(x)$ $(i = 0, 1, \ldots, n)$ *системой Чебышева* на компакте S, если любой обобщенный многочлен по этой системе, у которого хотя бы один коэффициент отличен от нуля, имеет на S не более n нулей. Подпространство M_n в этом случае называют подпространством, удовлетворяющим *условию Хаара*. Очевидно, требование линейной независимости системы $\{\varphi_i(x)\}$ является необходимым для того, чтобы эта система функций была системой Чебышева.

Теорема 1. *Для того чтобы $\forall f(x) \in C(S)$ и любого набора $n+1$ различных точек $x_i \in S$ $(i = 0, 1, \ldots, n)$ существовал обобщенный интерполяционный многочлен $\varphi(x)$, необходимо и достаточно, чтобы система функций $\{\varphi_i(x)\}$ являлась системой Чебышева на S. При этом обобщенный интерполяционный многочлен будет единственным.*

Доказательство. Для справедливости первого утверждения теоремы необходимо и достаточно, чтобы система линейных алгебраических уравнений

$$\sum_{p=0}^{n} a_p \varphi_p(x_i) = f(x_i), \qquad i = 0, 1, \ldots, n \tag{3}$$

имела решения относительно a_p $(p = 0, 1, \ldots, n)$ при любом выборе попарно различных узлов $x_i \in S$ $(i = 0, 1, \ldots, n)$ и любом выборе чисел $f(x_i)$, $i = 0, 1, \ldots, n$. Это возможно тогда и только тогда, когда определитель системы

$$\Phi = \left| \varphi_p(x_i) \right|_{i=\overline{0,n}}^{p=\overline{0,n}} \neq 0$$

при любом выборе попарно различных точек $x_i \in S$. Покажем, что необходимым и достаточным условием для этого является условие, что система функций $\{\varphi_p(x)\}$ — система Чебышева на S. Действитель-

но, если определитель $\Phi = 0$, то существуют постоянные b_p ($p = 0, 1, \ldots, n$) такие, что

$$\sum_{p=0}^{n} b_p \varphi_p(x_i) = 0, \qquad i = 0, 1, \ldots, n$$

(ввиду линейной зависимости столбцов Φ), т. е. обобщенный многочлен $\varphi(x) = \sum_{p=0}^{n} b_p \varphi_p(x)$ имеет на S $n+1$ корень, что противоречит определению системы Чебышева. Обратно, если $\{\varphi_p(x)\}$ не образует систему Чебышева, то существует нетривиальный обобщенный многочлен $\varphi(x) = \sum_{p=0}^{n} C_p \varphi_p(x)$ такой, что он обращается в нуль не менее чем в $n+1$ различной точке, принадлежащей $[a, b]$. Отсюда следует линейная зависимость столбцов определителя Φ, когда последовательностью точек x_i является $n+1$ различный нуль обобщенного многочлена

$$\varphi(x) = \sum_{p=0}^{n} C_p \varphi_p(x).$$

Поскольку определитель системы линейных алгебраических уравнений (3) отличен от нуля, то эта система имеет единственное решение. Теорема доказана полностью.

Если построить обобщенный интерполяционный многочлен вида (1) для функции $f(x)$ по системе Чебышева, то его можно представить в следующем виде:

$$\varphi(x) = \sum_{i=0}^{n} f(x_i)\, \Phi_i(x), \tag{4}$$

где $\Phi_i(x)$ — *фундаментальные обобщенные многочлены,* построенные по чебышевской системе функций $\varphi_i(x)$, со свойствами:

$$\Phi_i(x_j) = \delta_{ij}, \qquad i, j = 0, 1, \ldots, n, \tag{5}$$

где δ_{ij} — символ Кронекера.

Запись явного выражения обобщенного многочлена $\Phi_j(x)$ через функции $\varphi_i(x)$, очевидно, не представляет труда и мы оставляем ее в качестве упражнения для читателя.

Величину

$$R(x) = f(x) - \varphi(x) \tag{6}$$

будем называть *остаточным членом интерполяционной формулы.*

ЗАДАЧА ПРИБЛИЖЕНИЯ ФУНКЦИЙ

На практике часто бывает необходимо многократно вычислять значение некоторой функции $f(x)$, например значения элементарных функций e^x, $\ln x$, $\sin x$, $\cos x$ и других (особенно это касается работы на ЭВМ). Запоминать и хранить таблицы таких функций, которые должны быть достаточно большими, а затем тратить время на поиск нужного значения в таблице нецелесообразно. Поэтому часто для

нахождения значения функции $f(x)$ с точностью ε ее заменяют другой, легко вычисляемой функцией $\varphi(x)$ (например, многочленом), значения которой на всем рассматриваемом отрезке $[a, b]$ изменения x отличаются от значения $f(x)$ не больше чем на ε, и затем вычисляют $\varphi(x)$ в нужной точке.

Приведем общую постановку описанной выше ситуации.

Пусть $\mathbf{R}$ — линейное нормированное пространство и $f \in \mathbf{R}$ — элемент, который требуется приблизить. Возьмем в $\mathbf{R}$ $n+1$ линейно-независимых элементов φ_i $(i = 0, 1, \ldots, n)$ и образуем $(n+1)$-мерное линейное подпространство M_n всевозможных линейных комбинаций

$$\Phi = \sum_{i=0}^{n} C_i \varphi_i \tag{7}$$

с действительными коэффициентами C_i, $i = 0, 1, \ldots, n$.

Рассмотрим числовое множество

$$\Delta(f, \Phi) = \|f - \Phi\|, \tag{8}$$

где f — фиксированный, а Φ — произвольный элементы из $\mathbf{R}$ и M_n соответственно. Это числовое множество ограничено снизу и, следовательно, существует такое число $\Delta(f)$, что

$$\Delta(f) = \inf_{\Phi \in M_n} \Delta(f, \Phi). \tag{9}$$

Таким образом, приходим к задаче: найти элемент $\Phi_0 \in M_n$, для которого

$$\Delta(f) = \|f - \Phi_0\|. \tag{10}$$

Определение 3. Элемент $\Phi_0 \in M_n$, для которого выполняется равенство (10), называется *элементом наилучшего приближения* для f в M_n или проекцией f на M_n.

Теорема 2. *$\forall f \in \mathbf{R}$ в M_n существует элемент наилучшего приближения. Причем множество всех элементов наилучшего приближения выпукло.*

Доказательство. Функция

$$\varphi(C_0, C_1, \ldots, C_n) = \Delta(f, \Phi) = \|f - \Phi\| = \left\| f - \sum_{i=0}^{n} C_i \varphi_i \right\|$$

в силу непрерывности нормы является непрерывной функцией своих аргументов C_i, $i = 0, 1, \ldots, n$. Для $\|\Phi\| > 2\|f\|$ имеем

$$\|f - \Phi\| > \|f\| \geqslant \Delta(f),$$

поэтому целесообразно рассматривать лишь элементы замкнутого шара $\|\Phi\| \leqslant 2\|f\|$. На этом шаре непрерывная функция φ согласно теореме Вейерштрасса достигает своего минимального значения. Следовательно, существует, по крайней мере, один элемент Φ_0, для которого выполнено (10). Если

$$\Phi_0 = \sum_{i=0}^{n} C_i \varphi_i, \quad \hat{\Phi}_0 = \sum_{i=0}^{n} \hat{C}_i \varphi_i$$

— два элемента наилучшего приближения, то $\|f-\Phi_0\|=\|f-\hat{\Phi}_0\|=\Delta(f)$. В случае если $\Delta(f)=0$, имеем $f=\Phi_0=\hat{\Phi}_0$.

Рассмотрим случай $\Delta(f)>0$. Пусть m — точка отрезка, соединяющего Φ_0 с $\hat{\Phi}_0$:

$$m=a\Phi_0+b\hat{\Phi}_0, \quad a,\ b\geqslant 0, \quad a+b=1.$$

Тогда

$$\Delta(f)\leqslant\|f-m\|=\|a(f-\Phi_0)+b(f-\hat{\Phi}_0)\|\leqslant a\|f-\Phi_0\|+$$

$$+b\|f-\hat{\Phi}_0\|=\Delta(f),$$

следовательно, $\|f-m\|=\Delta(f)$, т. е. m является элементом наилучшего приближения и множество всех элементов наилучшего приближения — выпукло.

В описанном выше способе приближения $F(f)=f$ и $F_n(f)=\Phi_0$.

Часто на практике в качестве $\mathbf{R}$ рассматривают множество $C(S)$ непрерывных на компакте S функций, а в качестве M_n — некоторое множество обобщенных многочленов вида (7).

Лемма 1. *Пусть S_1 — подмножество точек S, в которых*

$$|f(x)-\Phi_0(x)|=\Delta(f,\Phi_0).$$

Для того чтобы обобщенный многочлен $\Phi_0(x)$ являлся многочленом наилучшего приближения к $f(x)\in C(S)$, необходимо и достаточно, чтобы $\forall\Phi(x)$ вида (7) $\exists\bar{x}\in S_1$ такой, что

$$\Phi(\bar{x})[f(\bar{x})-\Phi_0(\bar{x})]\leqslant 0.$$

Доказательство. Для доказательства *достаточности* возьмем произвольный обобщенный многочлен $\Phi_1(x)\neq\Phi_0(x)$ вида (7). Тогда по условию леммы найдется, по крайней мере, одна точка $\bar{x}\in S_1$, для которой будет иметь место неравенство

$$[\Phi_1(\bar{x})-\Phi_0(\bar{x})][f(\bar{x})-\Phi_0(\bar{x})]\leqslant 0.$$

Отсюда следует, что

$$|f(\bar{x})-\Phi_1(\bar{x})|=|f(\bar{x})-\Phi_0(\bar{x})+\Phi_0(\bar{x})-\Phi_1(\bar{x})|=$$

$$=\Delta(f,\Phi_0)+|\Phi_0(\bar{x})-\Phi_1(\bar{x})|\geqslant\Delta(f,\Phi_0), \qquad \forall\,\Phi_1(x)\in M_n,$$

а значит $\Phi_0(x)$ — многочлен наилучшего приближения.

Доказательство *необходимости* будем проводить от противного. Пусть $\exists\Phi_1(x)\in M_n$ такой, что

$$\Phi_1(x)[f(x)-\Phi_0(x)]>0, \qquad \forall\,x\in S_1.$$

Поскольку S_1 является подмножеством компакта S, то любая бесконечная последовательность точек из S_1 — сходящаяся и предельная точка ее в силу определения S_1 принадлежит S_1. Следовательно, S_1 — замкнуто и существует

$$\min_{x\in S_1}\Phi_1(x)[f(x)-\Phi_0(x)]=m>0.$$

Разобьем S на два подмножества S^+ и $S^- = S \setminus S^+$ такие, что

$$\Phi_1(x)[f(x) - \Phi_0(x)] > \frac{m}{2}, \qquad \forall x \in S^+. \tag{11}$$

Очевидно $S^+ \supset S_1$ и является открытым множеством в то время, как $\overline{S}^- = S^-$.

Введем обозначения

$$M = \max_{x \in S} |\Phi_1(x)|, \quad \mu = \Delta(f, \Phi_0) - \max_{x \in S^-} |f(x) - \Phi_0(x)|. \tag{12}$$

Легко видеть, что $\mu > 0$, ибо $S^- \not\supset S_1$. Рассмотрим обобщенный многочлен

$$\Phi_2(x) = \Phi_0(x) + \lambda\Phi_1(x), \tag{13}$$

где λ — пока произвольный положительный параметр, и оценим величину $\Delta(f, \Phi_2)$. Будем иметь

$$|f(x) - \Phi_2(x)| = |f(x) - \Phi_0(x) - \lambda\Phi_1(x)| \leqslant |f(x) - \Phi_0(x)| + \\ + \lambda|\Phi_1(x)| \leqslant \Delta(f, \Phi_0) - \mu + \lambda M, \qquad \forall x \in S^- \tag{14}$$

(здесь мы воспользовались соотношениями (12));

$$|f(x) - \Phi_2(x)|^2 = [f(x) - \Phi_0(x)]^2 + \lambda^2\Phi_1^2(x) - 2\lambda\Phi_1(x)[f(x) - \\ - \Phi_0(x)] \leqslant \Delta^2(f, \Phi_0) + \lambda^2 M^2 - \lambda m, \qquad \forall x \in S^+ \tag{15}$$

(здесь мы воспользовались соотношениями (11) и (12)).

Положим теперь

$$\lambda = \min\left\{\frac{\mu}{2M}, \frac{m}{2M^2}\right\}.$$

При таком выборе λ из неравенств (14) и (15) заключаем, что

$$|f(x) - \Phi_2(x)| < \Delta(f, \Phi_0), \qquad \forall x \in S$$

и, следовательно, $\Phi_0(x)$ не является многочленом наилучшего приближения. Полученное противоречие завершает доказательство леммы.

Лемма 2. *Пусть $f(x) \in C(S)$, $\Phi(x)$ — обобщенный многочлен построенный по системе Чебышева $\{\varphi_i\}$, вида (7). Если уравнение*

$$|f(x) - \Phi(x)| = \Delta(f, \Phi) \tag{16}$$

имеет в S меньше $n + 1$ различных корней, то $\Phi(x)$ не является многочленом наилучшего приближения для $f(x)$.

Доказательство. Пусть $x_i \in S$ $(i = 0, 1, \ldots, m, m < n)$ — различные корни уравнения (16). Возьмем любые отличающиеся друг от друга точки $x_{m+k} \in S$ $(k = 1, 2, \ldots, n - m)$, которые не совпадают с корнями уравнения (16). Построим обобщенный многочлен $\Phi_1(x)$ таким образом, чтобы он удовлетворял условиям

$$\Phi_1(x_i) = f(x_i) - \Phi(x_i), \qquad i = 0, 1, \ldots, n.$$

Такой обобщенный многочлен $\Phi_1(x)$ существует, так как $\{\varphi_i(x)\}$ образует систему Чебышева на компакте S. Тогда будем иметь

$$\Phi_1(x_i)[f(x_i) - \Phi(x_i)] = [f(x_i) - \Phi(x_i)]^2 > 0, \qquad \forall i = 0, 1, \ldots, m$$

и в силу леммы 1 $\Phi(x)$ не является многочленом наилучшего приближения. Противоречие доказывает лемму.

Ответ на вопрос, каковы должны быть условия, обеспечивающие единственность обобщенного многочлена наилучшего приближения, дает *теорема Хаара:*

Теорема 3. *Для того чтобы* $\forall f(x) \in C(S)$ *существовал единственный обобщенный многочлен наилучшего приближения, необходимо и достаточно, чтобы система* $\{\varphi_i(x)\}$ *была системой Чебышева на* S.

Доказательство. *Необходимость.* Будем проводить доказательство от противного. Предположим, что в S имеются $n+1$ различные точки x_i $(i = 0, 1, \ldots, n)$ такие, что

$$\left|\varphi_i(x_j)\right|_{i=\overline{0,n}}^{j=\overline{0,n}} = 0. \tag{17}$$

Отсюда следует, что

$$\sum_{i=0}^{n} C_i \varphi_k(x_i) = 0, \qquad \forall k = 0, 1, \ldots, n,$$

где $C_i \in R_1$ и $\sum\limits_{i=0}^{n} C_i^2 > 0$, и что

$$\sum_{i=0}^{n} C_i \Phi(x_i) = 0, \qquad \forall \Phi(x) \in M_n. \tag{18}$$

Кроме того, из сделанного предположения (17) следует существование такого нетривиального обобщенного многочлена $\Phi_0(x)$, что $\Phi_0(x_i) = 0$, $i = 0, 1, \ldots, n$. Найдем число λ, удовлетворяющее условию

$$\max_{x \in S} |\lambda \Phi_0(x)| \leqslant 1,$$

и любую функцию $g(x) \in S$, для которой

$$|g(x)| \leqslant 1, \quad \forall x \in S, \quad g(x_i) = \operatorname{sign} C_i, \text{ если } C_i \neq 0, \tag{19}$$
$$i = 0, 1, \ldots, n.$$

Функция

$$f(x) = g(x)\,[1 - |\lambda \Phi_0(x)|],$$

очевидно, также будет обладать свойствами (19). Докажем, что для $f(x)$ имеется бесконечно много обобщенных многочленов наилучшего приближения. Действительно,

$$\max_{x \in S} |f(x) - \Phi(x)| = \Delta(f, \Phi) \geqslant 1, \qquad \forall \Phi \in M_n,$$

ибо в противном случае на основании равенств

$$f(x_i) = \operatorname{sign} C_i, \quad C_i \neq 0, \qquad i = 0, 1, \ldots, n$$

должно быть справедливо соотношение

$$\operatorname{sign} \Phi(x_i) = \operatorname{sign} f(x_i) = \operatorname{sign} C_i, \qquad \forall C_i \neq 0,$$

что невозможно из-за (18).

С другой стороны, для любого ε такого, что $|\varepsilon| \leqslant 1$, будем иметь

$$|f(x) - \varepsilon\lambda\Phi_0(x)| \leqslant |f(x)| + |\varepsilon\lambda\Phi_0(x)| = |g(x)|[1 - |\lambda\Phi_0(x)|] +$$
$$+ |\varepsilon\lambda\Phi_0(x)| \leqslant 1 - |\lambda\Phi_0(x)| + |\varepsilon\lambda\Phi_0(x)| = 1 - (1 - |\varepsilon|)|\lambda\Phi_0(x)| \leqslant 1,$$

так что $\varepsilon\lambda\Phi_0(x)$ $\forall\varepsilon$ ($|\varepsilon| \leqslant 1$) — многочлен наилучшего приближения для $f(x)$. Необходимость доказана.

Достаточность. Пусть в противоположность утверждению функция $f(x)$ имеет два различных обобщенных многочлена наилучшего приближения $\Phi_0(x)$, $\overline{\Phi_0(x)}$.

Так как

$$\left|\frac{1}{2}(\Phi_0(x) + \overline{\Phi_0(x)} - f(x))\right| \leqslant \frac{1}{2}|\Phi_0(x) - f(x)| + \frac{1}{2}|\overline{\Phi_0(x)} - f(x)|,$$

то $\Phi_1(x) = \frac{1}{2}[\Phi_0(x) + \overline{\Phi_0(x)}]$ также будет многочленом наилучшего приближения.

В силу леммы 2, уравнение

$$|f(x) - \Phi_1(x)| = \Delta(f)$$

имеет, по крайней мере, $n + 1$ различных нулей $x_i \in S$, $i = 0, 1, \ldots$ $\ldots, n$. Но для того чтобы имело место соотношение

$$|f(x_i) - \Phi_1(x_i)| = \Delta(f), \qquad \forall i = 0, 1, \ldots, n,$$

необходимо выполнение равенства

$$f(x_i) - \Phi_0(x_i) = f(x_i) - \Phi_0(x_i) = \pm\Delta(f), \qquad \forall i = 0, 1, \ldots, n,$$

из которого вытекает, что нетривиальный обобщенный многочлен $\Phi_0(x) - \overline{\Phi_0(x)}$ имеет $n + 1$ различный нуль, что невозможно. Достаточность доказана.

Пусть $\mathbf{B} = \mathbf{H}$ — гильбертово пространство, $M_n \subset \mathbf{H}$ и имеет тот же смысл, что и раньше, тогда будет справедлива такая теорема:

Теорема 4. *$\forall f \in \mathbf{H}$ в M_n существует элемент наилучшего приближения и притом единственный.*

Доказательство. Существование элемента наилучшего приближения Φ следует из теоремы 2. Докажем его единственность. Для этого сначала покажем, что будет иметь место соотношение

$$(f - \Phi_0, \Phi) = 0, \qquad \forall \Phi \in M_n.$$

Допустим противное, т. е. предположим, что $\exists \Phi_1 \in M_n$, для которого $(f - \Phi_0, \Phi_1) = \alpha \neq 0$, причем можно считать, не уменьшая общности, что $\|\Phi_1\| = 1$. Рассмотрим элемент $\Phi_2 = \Phi_0 + \alpha\Phi_1$ и оценим норму:

$$\|f - \Phi_2\|^2 = (f - \Phi_2, f - \Phi_2) = (f - \Phi_0, f - \Phi_0) - \alpha(\Phi_1, f - \Phi_0) -$$
$$- \bar{\alpha}(f - \Phi_0, \Phi_1) + \alpha\bar{\alpha}(\Phi_1, \Phi_1) = \|f - \Phi_0\|^2 - |\alpha|^2.$$

Отсюда следует, что

$$\|f - \Phi_2\| < \|f - \Phi_0\|,$$

что невозможно, так как Φ_0 — элемент наилучшего приближения.

Предположим теперь, что существует второй элемент наилучшего приближения $\hat{\Phi}_0 \neq \Phi_0$. Тогда в силу предыдущего

$$(f - \Phi_0, \Phi) = (f - \hat{\Phi}_0, \Phi) = 0, \qquad \forall \Phi \in M_n$$

и, в частности,

$$(f - \Phi_0, \Phi_0 - \hat{\Phi}_0) = (f - \hat{\Phi}_0, \Phi_0 - \hat{\Phi}_0) = 0.$$

Но

$$\|\Phi_0 - \hat{\Phi}_0\|^2 = (\Phi_0 - \hat{\Phi}_0, \Phi_0 - \hat{\Phi}_0) = ((\Phi_0 - f) + (f - \hat{\Phi}_0), \Phi_0 - \hat{\Phi}_0) =$$

$$= (\Phi_0 - f, \Phi_0 - \hat{\Phi}_0) + (f - \hat{\Phi}_0, \Phi_0 - \hat{\Phi}_0) = 0,$$

а это означает, что $\Phi_0 = \hat{\Phi}_0$. Приходим к противоречию. Этим завершается доказательство теоремы.

ЗАДАЧА ЧИСЛЕННОГО ИНТЕГРИРОВАНИЯ

Пусть **B** — пространство непрерывных и интегрируемых на отрезке $[a, b]$ с весом $\rho(x) > 0$ функций, а оператор $F : \mathbf{B} \to R_1$ является следующим функционалом:

$$F(f) = \int_a^b f(x)\,\rho(x)\,dx. \tag{20}$$

Задача численного интегрирования состоит в построении приближенных формул вида

$$F(f) = \int_a^b f(x)\,\rho(x)\,dx \approx F_n(f) = \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}), \tag{21}$$

т. е. в нахождении функционалов $F_n(f)$.

Здесь $x_k^{(n)}$ $(k = 1, 2, \ldots, n)$ называют *узлами* или *абсциссами квадратурной формулы*, а числа $C_k^{(n)}$ — *коэффициентами* или *весами квадратурной формулы*. Требуется, чтобы узлы $x_k^{(n)}$ и коэффициенты $C_k^{(n)}$ не зависели от выбора функции $f(x)$ из рассматриваемого класса функций.

Величину

$$R_n(f) = \int_a^b f(x)\,\rho(x)\,dx - \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) \tag{22}$$

будем называть *остаточным членом квадратурной формулы.*

Возможны различные подходы к построению квадратурных формул вида (21).

а) Квадратурные формулы с наилучшей оценкой на классе функций. Пусть

$$R_n = \sup_{f \in \mathbf{B}} R_n(f), \tag{23}$$

требуется определить узлы и веса квадратурной формулы (21) (с частичными ограничениями на них или без ограничений), чтобы величина

R_n была наименьшей. Такие квадратурные формулы называют формулами с наилучшей оценкой на классе функций **B**.

б) **Квадратурные формулы наилучшей степени точности.** Квадратурные формулы вида (21), точные для обобщенных многочленов максимально высокой степени, построенных по функциям $\varphi_i(x)$ ($i = 0, 1, \ldots$), образующим систему Чебышева на $[a, b]$, будем называть квадратурными формулами наилучшей степени точности относительно системы $\{\varphi_i(x)\}$. Максимальный порядок m обобщенного многочлена, для которого квадратурная формула точна, называют *степенью точности квадратурной формулы* относительно системы $\{\varphi_i(x)\}$. При этом либо веса, либо узлы квадратурной формулы (21) могут заранее фиксироваться.

в) **Интерполяционные квадратурные формулы.** Представим функцию $f(x) \in \mathbf{B}$ в виде

$$f(x) = \varphi(x) + R(x), \tag{24}$$

где $\varphi(x)$ — обобщенный интерполяционный многочлен (4), а $R(x)$ — остаточный член интерполяционной формулы. Тогда

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

$$= \sum_{k=0}^{n} C_k^{(n)} f(x_k^{(n)}) + R(f), \tag{25}$$

где

$$C_k^{(n)} = \int_a^b \rho(x)\Phi_k(x)\,dx, \qquad k = 0, 1, \ldots, n;$$

$$R(f) = \int_a^b \rho(x) R(x)\,dx. \tag{26}$$

Числа $C_k^{(n)}$ не зависят от функции $f(x)$ и их можно раз и навсегда вычислить. Квадратурные формулы, веса которых находятся по формулам (26), называют *квадратурными формулами интерполяционного типа.*

Отметим, что интерполяционная квадратурная формула с $n + 1$ узлом имеет степень точности не меньше n относительно системы функций $\varphi_i(x)$ ($i = 0, 1, \ldots, n$), ибо для любой линейной комбинации

$$\psi_n(x) = \sum_{i=0}^{n} a_i \varphi_i(x)$$

будет иметь место соотношение $R(\psi_n) = 0$.

ЗАДАЧА ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ

Пусть **B** — система действительных функций, определенных и дифференцируемых на отрезке $[a, b]$. Для любого фиксированного $x \in [a, b]$ положим $F(f) = \dfrac{d^k f(x)}{dx^k}$ и будем искать приближение к $F(f)$

с помощью замены функции f более простой функцией, например интерполяционным многочленом. Задача численного дифференцирования состоит в отыскании формул вида

$$F(f) = \frac{d^k f}{dx^k} \approx \sum_{i=0}^{n} C_i f(x_i). \tag{27}$$

Величину

$$R(x) = \frac{d^k f}{dx^k} - \sum_{i=0}^{n} C_i f(x_i) \tag{28}$$

называют *погрешностью формулы численного дифференцирования.* Отыскание коэффициентов C_i и узлов x_i формулы численного дифференцирования (27) так же, как и в случае численного интегрирования, может проводиться с помощью тех же подходов.

Пусть необходимо приближенно найти $f''(x_i)$ по формуле (27) с условиями: $a = 0$, $b = 1$, $x_i = \frac{i}{n+1}$, и формула (27) должна иметь наивысшую точность относительно тригонометрической системы Чебышева $\sin \pi x$, $\sin 2\pi x, \ldots$. Можно проверить, что i-я компонента вектора

$$f^{(2)} = P\Lambda P f, \text{ где } P = \sqrt{\frac{2}{n+1}} \left[\sin \frac{ij\pi}{n+1}\right]_{i=\overline{1,n}}^{j=\overline{1,n}}, \Lambda = [-k^2\pi^2]_{k=\overline{1,n}},$$

будет давать искомую формулу.

Действительно, для произвольной линейной комбинации

$$f_n(x) = \sum_{k=1}^{n} a_n \sin k\pi x$$

ввиду соотношения ортогональности

$$\sum_{i=1}^{n} \sin k\pi x_i \sin s\pi x_i = \sum_{i=1}^{n} \sin \frac{ik\pi}{n+1} \sin \frac{si\pi}{n+1} = \delta_{k,s} \sqrt{\frac{n+1}{2}}$$

будем иметь

$$f_n^{(2)} = P\Lambda P f_n = f_n'',$$

где

$$f_n'' = \left(-\sum_{k=1}^{n} a_n k^2\pi^2 \sin k\pi x_i\right)_{i=1}^{n},$$

т. е. след на сетку функции $f_n''(x)$ совпадает с вектором $P\Lambda P f_n = f_n^{(2)}$.

Задача численного дифференцирования по сравнению с задачами теории интерполирования и численного интегрирования имеет существенно отличительную черту. Задача дифференцирования является некорректной в $C[a, b]$, т. е. сколь угодно близкие функции могут иметь сколь угодно удаленные производные в смысле расстояния пространства $C[a, b]$. Этим обстоятельством объясняется малая точность формул численного дифференцирования, особенно если учесть, что значения функций, участвующие в формулах, как правило, несут погрешность.

§ 2. ЕДИНЫЙ СПОСОБ ПОСТРОЕНИЯ ФОРМУЛ ИНТЕРПОЛЯЦИОННОГО ТИПА ДЛЯ ПРИБЛИЖЕНИЯ ЛИНЕЙНЫХ ФУНКЦИОНАЛОВ

Рассмотрим сначала случай, когда **B** — пространство непрерывных действительных функций, заданных на отрезке $[a, b]$, а узлы x_i ($i = 0, 1, \ldots, n$) различны и заранее фиксированы (фиксированные узлы).

Для приближения функционала F будем искать формулу интерполяционного типа

$$F(f) \approx F_n(f) = \sum_{i=0}^{n} C_i^{(n)} f(x_i), \qquad f \in \mathbf{B}, \tag{1}$$

приближающую линейный функционал F. Поскольку (1) является формулой интерполяционного типа, она должна быть точной для обобщенного интерполяционного многочлена $\varphi(x)$ (4), § 1, построенного по чебышевской системе функций $\varphi_i(x)$ ($i = 0, 1, 2, \ldots, n$) на $[a, b]$. Отсюда следует, что веса $C_i^{(n)}$ должны определяться формулами

$$C_i^{(n)} = F(\Phi_i), \qquad i = 0, 1, \ldots, n, \tag{2}$$

где $\Phi_i(x)$ ($i = 0, 1, \ldots, n$) — обобщенные фундаментальные многочлены (см. § 1). Здесь при выводе (1) применен так называемый *метод аналитической замены.*

Согласно другому способу получения формул вида (1), требуют, чтобы формула (1) была точной для любой функции из чебышевской системы $\varphi_i(x)$ ($i = 0, 1, \ldots, n$), т. е. веса $C_i^{(n)}$ находятся из системы уравнений

$$F(\varphi_j) = \sum_{i=0}^{n} C_i^{(n)} \varphi_j(x_i), \qquad j = 0, 1, \ldots, n. \tag{3}$$

Но так как обобщенный интерполяционный многочлен n-й степени для любой функции $\varphi_j(x)$ совпадает с самой этой функцией, то легко видеть, что описанные выше два способа получения формулы (1) эквивалентны.

В некоторых случаях, когда матрица

$$T(\varphi_0, \varphi_1, \ldots, \varphi_n) = [\varphi_j(x)]_{i=\overline{0,n}}^{j=\overline{0,n}} \tag{4}$$

системы (3) легко обратима и часто используется, она может быть раз и навсегда обращена. Тогда, если вектор $\mu^{(n)} = (\mu_i)_{i=0}^{n}$ обобщенных моментов функционала F известен, где

$$\mu_j = F(\varphi_j), \qquad j = 0, 1, \ldots, n, \tag{5}$$

то вектор весов $C^{(n)} = (C_i^{(n)})_{i=0}^{n}$ находится из формулы

$$C^{(n)} = T^{-1}(\varphi_0, \varphi_1, \ldots, \varphi_n)\, \mu^{(n)}. \tag{6}$$

Приведем несколько примеров.

Пример 1. Пусть $\varphi_j(x) = P_j(x)$, где $P_j(x)$ — ортогональные многочлены, соответствующие распределению $d\alpha(x)$, а $x_{i+1,\, n+1}^{P}$ ($i = 0, 1, \ldots, n$) — нули многочлена $P_{n+1}(x)$. Тогда справедлива следующая лемма:

Лемма 1. Если $\varphi_j(x) = P_j(x)$, $x_i = x^P_{i+1,n+1}$ $(i, j = 0, 1, \ldots, n)$, *то имеет место соотношение*

$$T(P_0, \ldots, P_n)\,\Lambda_{n+1}T^*(P_0, \ldots, P_n) = D,$$

где $\Lambda_{n+1} = [\lambda_{jn+1}]_{j=1}^{n+1}$, $D\,[d_j^2]_{j=0}^n$ — *диагональные матрицы;* λ_{jn+1} — *числа Кристоффеля в форме механических квадратур Гаусса-Якоби,* $d_j = \left[\int\limits_a^b P_j^2(x)\,d\alpha(x)\right]^{\frac{1}{2}}$ — *норма полинома* $P_j(x)$.

Доказательство. Обозначим $A = T(P_0, \ldots, P_n)\,\Lambda_{n+1}T^*(P_0, \ldots, P_n)$. Тогда элементы $a_{i+1,j+1}$ матрицы A будут выражаться формулой

$$a_{i+1,j+1} = \sum_{k=0}^{n} \lambda_{k+1,n+1} P_j(x_k)\,P_i(x_k), \qquad i, j = 0, 1, \ldots, n.$$

Так как $P_i(x)\,P_j(x)$ — многочлен степени не выше $2n$, то исходя из формулы механической квадратуры (см. § 2, гл. 4) имеем

$$a_{i+1,j+1} = \sum_{k=0}^{n} \lambda_{k+1,n+1} P_j(x_k)\,P_i(x_k) = \int\limits_a^b P_j(x)\,P_i(x)\,d\alpha(x) = \delta_{ij} d_i^2,$$

что и требовалось доказать.

Исходя из этой леммы нетрудно получить

$$T^{-1}(P_0, \ldots, P_n) = \Lambda_{n+1}T^*(P_0, \ldots, P_n)\,D^{-1}. \tag{7}$$

Заметим, что из леммы следует ортогональность строк матрицы $T(P_0, \ldots, P_n)$ с весами $\lambda_{j,n+1}$. У таких матриц столбцы также ортогональны по отношению к надлежащим весам, что следует из приводимой ниже леммы.

Лемма 2. Пусть y_{rs} $(r, s = 0, \ldots, m-1)$ *ортогональны в том смысле, что*

$$\sum_{r=0}^{m-1} a_r y_{rs} \bar{y}_{rt} = \delta_{st}\rho_s \qquad (s, t = 0, \ldots, m-1), \tag{8}$$

где a_r, ρ_s — *вещественны и положительны. Тогда*

$$\sum_{r=0}^{m-1} \frac{y_{sr}\bar{y}_{tr}}{\rho_r} = \frac{\delta_{st}}{a_s} \qquad (s, t = 0, \ldots, m-1). \tag{8'}$$

Доказательство. Из соотношений ортогональности (8) заключаем, что m векторов $\boldsymbol{y}_s = (y_{0s}, \ldots, y_{m-1,s})$ $(s = 0, 1, \ldots, m-1)$ линейно-независимы. Таким образом, произвольный вектор $(u_0, \ldots, u_{m-1})$ можно выразить в виде

$$u_n = \sum_{p=0}^{m-1} \frac{v_p y_{np}}{\rho_p} \qquad (n = 0, 1, \ldots, m-1),$$

где ρ_p — нормирующие множители; v_p — коэффициенты Фурье. Умножая последнее равенство на $a_n\bar{y}_{nr}$ и суммируя по n, получим согласно (8)

$$\sum_{n=0}^{m-1} a_n u_n \bar{y}_{nr} = \sum_{n=0}^{m-1} a_n \bar{y}_{nr} \sum_{p=0}^{m-1} \frac{v_p y_{np}}{\rho_p} = \sum_{p=0}^{m-1} \frac{v_p}{\rho_p} \sum_{n=0}^{m-1} a_n \bar{y}_{nr} y_{np} = v_r.$$

Подставляя это в выражение для u_n, имеем

$$u_n = \sum_{p=0}^{m-1} y_{np}\rho_p^{-1} \sum_{q=0}^{m-1} a_q u_q \bar{y}_{qp} = \sum_{q=0}^{m-1} u_q a_q \sum_{p=0}^{m-1} \frac{y_{np}\bar{y}_{qp}}{\rho_p}.$$

Здесь u_n произвольны и поэтому (8′) можно получить путем сравнения коэффициентов при u_q.

Пример 2. Пусть $\varphi_j(x) = x^j$, $j = 0, 1, \ldots, n$, тогда матрица (4) является матрицей Вандермонда, которая легко обратима, и для нее имеет место формула

$$T^{-1}(1, x, \ldots, x^n) = \left[\frac{a_{m,k}}{\omega_n'(x_m)}\right]_{m=\overline{0,n}}^{k=\overline{0,n}}, \tag{9}$$

где $a_{m,k}$ — коэффициенты многочлена

$$\frac{\omega_n(x)}{x - x_m} = \prod_{\substack{i=0\\ i\neq m}}^{n} (x - x_i) = \sum_{i=0}^{n} a_{i,m} x^i.$$

Для случая, когда узлы равноотстоящие и $x_i = -\frac{n}{2} + i$, где $i = 0, 1, \ldots, n$, имеются таблицы.

Переходя к случаю свободных узлов, заметим, что формула (1) имеет уже $2n + 2$ параметра. Для их определения можно воспользоваться другим способом, потребовав, чтобы формула (1) была точной для любой из функций $\varphi_j(x)$ ($j = 0, 1, \ldots, 2n + 1$), образующих чебышевскую систему. Тогда получим следующую нелинейную систему уравнений

$$\mu^{(2n+1)} = T(\varphi_0, \varphi_1, \ldots, \varphi_{2n+1}) C^{(2n+1)}, \tag{10}$$

или в скалярном виде

$$\mu_k = \sum_{i=0}^{n} \varphi_k(x_i) C_i^{(2n+1)}, \qquad k = 0, 1, \ldots, 2n + 1. \tag{10′}$$

Существует хорошо известный метод для решения системы линейных уравнений такого вида. Сначала определим обобщенный многочлен

$$\Phi(x) = \sum_{i=0}^{n+1} a_i \varphi_i(x), \qquad a_{n+1} = 1, \tag{11}$$

который обладает свойствами

$$\Phi(x_j) = 0, \qquad j = 0, 1, \ldots, n,$$

что возможно, ибо функции $\varphi_i(x)$ образуют систему Чебышева на отрезке $[a, b]$.

Произведем с первыми $n + 1$ уравнениями системы (10′) следующее преобразование. Умножив первое уравнение (10′) на a_0, следующее на a_1 и так далее, последнее на a_n и просуммировав их, получим

$$\sum_{i=0}^{n} a_i \mu_i = 0 = \sum_{i=0}^{n} \Phi(x_i) C_i^{(2n+1)}.$$

Чтобы получить следующее уравнение, относительно системы Чебышева $\varphi_i(x)$ предположим

$$\varphi_{i+j}(x) = \varphi_i(x) \varphi_j(x). \tag{12}$$

Тогда, повторяя описанное выше преобразование со следующими $n + 1$ уравнениями системы (10′), начиная со второго, будем иметь:

$$\sum_{i=0}^{n} a_i \mu_{i+1} = \sum_{i=0}^{n} \varphi_1(x_i) \Phi(x_i) C_i^{(2n+1)} = 0.$$

Продолжая аналогичные преобразования, окончательно получим систему

$$\sum_{i=0}^{n} a_i \mu_{i+k} = 0, \qquad k = 0,\ 1,\ 2,\ \ldots,\ n, \quad (a_n = 1). \tag{13}$$

Определитель системы (13) называется «персимметричным». Если предположить, что система (13) имеет единственное решение, то после ее решения фактически получаем явный вид обобщенного многочлена (11). После определения его корней $x_i^{(n)}$ ($i = 0, 1, \ldots, n$) задача сводится к предыдущей, т. е. к случаю фиксированных узлов.

Наконец рассмотрим смешанный случай, когда один или несколько узлов в формуле (1) фиксированы, а остальные — свободны. Рассмотрим случай, когда узел $x_0 = a$, а остальные узлы — пока произвольные: $x_k \in (a,\ b)$, $k = 1, 2, \ldots, n$. Тогда вместе с весовыми коэффициентами $C_i^{(n+1)}$ ($i = 0, 1, \ldots, n$) в нашем распоряжении находится $2n + 1$ параметр. Определяющее уравнение (10′) запишем в виде

$$\mu_k = C_0^{(2n+1)} \varphi_k(a) + \sum_{i=0}^{n} C_i^{(2n+1)} \varphi_k(x_i), \qquad k = 0,\ 1,\ \ldots,\ 2n. \tag{14}$$

Частично исключим a, умножая каждое из уравнений (14) на $\varphi_1(a)$ и вычитая его из следующего за ним уравнения:

$$\bar{\mu}_k = \mu_{k+1} - \varphi_1(a)\,\mu_k = \sum_{i=1}^{n} C_i^{(2n+1)} [\varphi_1(x_i) - \varphi_1(a)]\, \varphi_k(x_i),$$

$$k = 0,\ 1,\ \ldots,\ 2n-1.$$

Здесь использовано соотношение (12). Теперь образуем обобщенный многочлен

$$\Phi(x) = \sum_{i=0}^{n} a_i \varphi_i(x), \quad \Phi(x_k) = 0, \qquad k = 1,\ 2,\ \ldots,\ n, \quad a_n = 1,$$

используя лишь неизвестные узловые точки, и повторим процесс исключения, примененный выше:

$$\sum_{k=0}^{n} a_k \bar{\mu}_{k+j} = \sum_{i=1}^{n} C_i^{(2n+1)} [\varphi_i(x_i) - \varphi_1(a)]\, \varphi_j(x_i)\, \Phi(x_i) = 0,$$

$$j = 0,\ 1,\ \ldots,\ n-1.$$

После этого можно найти a_k и неизвестные x_k как корни обобщенного многочлена $\Phi(x)$.

§ 3. СИСТЕМЫ ЧЕБЫШЕВА И ИХ СВОЙСТВА

В предыдущих параграфах было показано, что основой при построении аппроксимаций линейных операторов является система функций Чебышева.

Естественно, возникают следующие вопросы: для многих ли компактов S существуют подпространства $M \subset C(S)$, удовлетворяющие

условию Хаара, и всегда ли существует для данного компакта S подпространство $M \subset C(S)$, удовлетворяющее условию Хаара?

Приведем две теоремы, уточняющие взаимосвязь между S и возможностью найти подпространство $M \subset C(S)$, удовлетворяющее условию Хаара.

Предварительно введем следующее понятие: систему трех непрерывных отображений $f_i : [0, 1] \to S$ $(i = 1, 2, 3)$ таких, что $f_i(t)$ отличны от постоянных, $f_i(0) = a \in S$ и $f_i(t) \neq f_j(t')$ для $i \neq j$ и $\forall t, t' > 0$, назовем триподом.

Теорема 1. *Если компакт S содержит некоторый трипод, то в $C(S)$ нет подпространства размерности, большей или равной двум, которое удовлетворяло бы условию Хаара.*

Доказательство. Определитель $\Phi(x_1, \ldots, x_n) = \det[\varphi_i(x_j)]_{i=\overline{1,n}}^{j=\overline{1,n}}$ непрерывен по всем переменным $x_1, x_2, \ldots, x_n$. Зафиксируем $x_3, x_4, \ldots, x_n$ и рассмотрим функцию

$$\varphi(x_1, x_2) = \Phi(x_1, x_2, \ldots, x_n).$$

Пусть $x_1 = f_1(t_1), x_2 = f_2(t_2)$. Опишем непрерывные пути, которые последовательно проходят переменные x_1, x_2:

1) $x_1 : f_1(t_1) \to f_1(0) = a \to f_3(t_3)$,
2) $x_2 : f_2(t_2) \to f_2(0) = a \to f_1(t_1)$,
3) $x_1 : f_3(t_3) \to f_3(0) = a \to f_2(t_2)$.

В результате таких действий в определителе Φ меняются местами столбцы и значит он меняет знак. Следовательно, найдется такое промежуточное положение $\bar{x}_1, \bar{x}_2$, когда $\varphi(\bar{x}_1, \bar{x}_2) = 0$, что приводит к противоречию.

Следующую теорему приведем без доказательства.

Теорема 2. *Пусть S — компактное множество в $\mathbf{R}_m$. Подпространство $M \subset S$ размерности $n \geqslant 2$, удовлетворяющее условию Хаара, существует тогда и только тогда, когда S гомеоморфен замкнутой части окружности.*

Пусть S является конечным отрезком числовой прямой. Рассмотрим достаточные условия, при которых система функций $\varphi_i(x)$ $(i = 0, 1, \ldots, n)$ будет системой Чебышева на отрезке $[a, b]$.

Наложим на функции $\{\varphi_i(x)\}_{i=0}^{n}$ следующие ограничения:

1) $\varphi_i(x) \in C^{(n+1)}[a, b]$;
2) все вронскианы

$$W[\varphi_0, \varphi_1, \ldots, \varphi_k] = \left| \varphi_p^{(i)}(x) \right|_{p=\overline{0,k}}^{i=\overline{0,k}} \neq 0,$$

$$(k = 0, 1, \ldots, n) \ \forall x \in [a, b].$$

Докажем следующее обобщение теоремы Ролля.

Теорема 3. *Пусть $f(x) \in C^{(n+1)}[a, b]$ и имеет на этом промежутке $n + 2$ корня. Тогда $\exists \xi \in [a, b]$ такая, что выражение*

$$L_{n+1}(f) = \frac{W[\varphi_0, \varphi_1, \ldots, \varphi_n, f]}{W[\varphi_0, \varphi_1, \ldots, \varphi_n]} \tag{1}$$

обращается в нуль в точке ξ.

Доказательство. Для дифференциального уравнения

$$L_{k+1}[\varphi] = \frac{W[\varphi_0, \ldots, \varphi_k, \varphi]}{W[\varphi_0, \ldots, \varphi_k]} = 0, \qquad k = 0, 1, \ldots, L_0[\varphi] = \varphi$$

функции $\varphi_i(x)$ $(i = 0, 1, \ldots, k)$ образуют фундаментальную систему, причем коэффициент при старшей производной в этом уравнении равен 1. Рассмотрим другое дифференциальное выражение

$$M_{k+1}[\varphi] = \frac{d}{dx} L_k[\varphi] - b_k(x) L_k[\varphi] = 0,$$

где

$$b_k(x) = \frac{\frac{d}{dx} L_k[\varphi_k]}{L_k[\varphi_k]}, \qquad k = 1, 2, \ldots, b_0(x) = \frac{\varphi_0'(x)}{\varphi_0(x)},$$

для которого по построению функции $\varphi_i(x)$ $(i = 0, 1, \ldots, k)$ также образуют фундаментальную систему ($M_{k+1}[\varphi_k] = 0$ за счет выбора коэффициента $b_k(x)$) и коэффициент при старшей производной в нем также равен 1. Но тогда

$$L_{k+1}[\varphi] = M_{k+1}[\varphi] = \frac{d}{dx} L_k[\varphi] - b_k(x) L_k[\varphi].$$

Рассмотрим теперь функцию

$$\psi_{k+1}(x) = L_k[f(x)] \exp\left[-\int_0^x b_k(x)\,dx\right], \qquad k = 1, 2, \ldots, n,$$

$$\psi_1(x) = f(x) \frac{\varphi_0(0)}{\varphi_0(x)}.$$

Она удовлетворяет условию

$$\frac{d}{dx} \psi_{k+1}(x) = \exp\left[-\int_0^x b_k(x)\,dx\right] \left\{\frac{d}{dx} L_k[f] - b_k(x) L_k[f]\right\} =$$

$$= \exp\left[-\int_0^x b_k(x)\,dx\right] L_{k+1}[f]. \tag{2}$$

Функция $\psi_1(x)$ обращается $n + 2$ раза в нуль на $[a, b]$, тогда функция $\psi_1'(x)$ обращается, по крайней мере, $n + 1$ раз в нуль на $[a, b]$ и согласно формуле (2) столько же раз в нуль обратится на $[a, b]$ функция $L_1[f]$, а следовательно, и $\psi_2(x)$ и т. д. Наконец, показываем, что найдется, по крайней мере, одна точка $\xi \in [a, b]$ такая, что $L_{n+1}[f(\xi)] = 0$, что и требовалось доказать.

Теорема 4. *Если* $\varphi_i(x) \in C^{n+1}[a, b]$ $\forall i = 0, 1, \ldots, n$ *и вронскианы* $W[\varphi_0, \varphi_1, \ldots, \varphi_k] \neq 0$ *на* $[a, b]$ *при всех* $k = 0, 1, \ldots, n$, *то функции* $\varphi_0(x), \varphi_1(x), \ldots, \varphi_n(x)$ *образуют систему Чебышева.*

Доказательство. Предположим, что это не так. Тогда найдется такая линейная комбинация

$$f(x) = \sum_{p=0}^{n} C_p \varphi_p(x), \quad \left(C_p \in R_1, \quad \sum_{p=0}^{n} C_p^2 \neq 0\right),$$

которая обращается в нуль, по крайней мере, в $n+1$ различных точках отрезка $[a, b]$. Тогда по теореме 3 $\exists\, \xi \in [a, b]$ такая, что $L_n[f(\xi)] = 0$. Но

$$L_n[f] = \frac{W[\varphi_0, \ldots, \varphi_{n-1}, f]}{W[\varphi_0, \ldots, \varphi_{n-1}]} = C_n \frac{W[\varphi_0, \ldots, \varphi_n]}{W[\varphi_0, \ldots, \varphi_{n-1}]}.$$

Так как вронскианы $W[\varphi_0, \ldots, \varphi_k]$ $(k = 0, 1, \ldots, n)$ не обращаются в нуль ни в одной точке $x \in [a, b]$, то должно быть $C_n = 0$. Таким образом, найдется $n+1$ различных точек отрезка $[a, b]$, в которых функция

$$f(x) = \sum_{p=0}^{n-1} C_p \varphi_p(x)$$

обращается в нуль. Тогда, снова применяя теорему 3, найдем, что $\exists\, \xi \in [a, b]$ такая, что $L_{n-1}[f(\xi)] = 0$. Проводя те же рассуждения, что и раньше, найдем, что $C_{n-1} = 0$. Продолжая этот процесс, приходим к выводу, что все коэффициенты $C_i = 0$ $(i = 0, 1, \ldots, n)$ вопреки нашему предположению.

Теорема 5. *Пусть $p_0(x) > 0$, $r_0(x) > 0$ и $q_0(x)$ — вещественные непрерывные функции, определенные на ограниченном интервале $[a, b]$. Пусть $\lambda_0 < \lambda_1 < \lambda_2 < \ldots$ и уравнение*

$$(p_0(x)u')' + [q_0(x) + \lambda_n r_0(x)]u = 0, \qquad n = 0, 1, \ldots \tag{3}$$

имеет (вещественное) решение $u_n(x)$ при $x \in [a, b]$, имеющее не более n нулей и такое, что существует $\lim\limits_{x \to a, b} [u_n(x)/u_0(x)]$. Тогда система функций $u_i(x)$ $(i = 0, 1, \ldots, n)$ образует систему Чебышева на $[a, b]$.

Рассмотрим определитель

$$\Phi(x) = \Phi(x, x_1, \ldots, x_n) = \begin{vmatrix} \varphi_0(x) & \varphi_0(x_1) & \ldots & \varphi_0(x_n) \\ \varphi_1(x) & \varphi_1(x_1) & \ldots & \varphi_1(x_n) \\ \cdot & \cdot & \cdot & \cdot \\ \varphi_n(x) & \varphi_n(x_1) & \ldots & \varphi_n(x_n) \end{vmatrix},$$

где $x_i \in [a, b]$ $(i = 1, 2, \ldots, n)$ — попарно различные точки.

Лемма 1. *Пусть $a < x_1 < \ldots < x_n < b$. Тогда функция $\Phi(x)$ сохраняет знак на каждом из интервалов (x_i, x_{i+1}), $i = 0, 1, \ldots, n$, $x_0 = a$, $x_{n+1} = b$, при этом знаки $\Phi(x)$ в последовательных интервалах чередуются.*

Доказательство. Так как $\Phi(x)$ есть обобщенный многочлен по системе Чебышева $\{\varphi_k(x)\}$, обращающийся в нуль в точках x_i $(i = 1, 2, \ldots, n)$, то ни в какой другой точке отрезка $[a, b]$ он не может обращаться в нуль. Для доказательства того, что на двух соседних интервалах $\Phi(x)$ имеет противоположные знаки, заметим, что знак $\Phi(x)$ не изменится, если мы будем как угодно непрерывно перемещать точки $x, x_1, \ldots, x_n$ на отрезке $[a, b]$, не меняя их взаимного расположения.

Рассмотрим два последовательных интервала (x_k, x_{k+1}) и (x_{k+1}, x_{k+2}) и пусть $\xi \in (x_k, x_{k+1})$, $\eta \in (x_{k+1}, x_{k+2})$. В силу нашего замечания, знак $\Phi(\xi)$ будет совпадать со знаком определителя Φ, в котором ξ смещена в положение x_{k+1}, а x_{k+1} — в положение η, так как

при этом взаимные расположения точек $\xi, x_1, \ldots, x_n$ не изменяются, т. е.

$$\operatorname{sign}\Phi(\xi) = \operatorname{sign}\Phi(\xi, x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) =$$
$$= \operatorname{sign}\Phi(x_{k+1}, x_1, \ldots, x_k, \eta, x_{k+2}, \ldots, x_n).$$

Но определители $\Phi(x_{k+1}, x_1, \ldots, x_k, \eta, x_{k+2}, \ldots, x_n)$ и $\Phi(\eta, x_1, \ldots, x_k, x_{k+1}, \ldots, x_n)$ отличаются только знаками. Таким образом, $\operatorname{sing}\Phi(\xi) = -\operatorname{sing}\Phi(\eta)$, что и требовалось доказать.

Определение 1. Семейство функций $\{\varphi_i(x)\}_{i=0}^{n}$, образующих систему Чебышева на $[a, b]$, называется *периодической системой* Чебышева на $[a, b]$, если

$$\varphi_i(a) = \varphi_i(b), \qquad i = 0, 1, \ldots, n$$

и, кроме того, a и b считаются за один нуль.

Доказанная лемма 1 позволяет установить, каким должен быть порядок n периодической системы Чебышева на отрезке $[a, b]$.

Лемма 2. *Порядок n периодической системы Чебышева на отрезке $[a, b]$ — четное число.*

Доказательство. Из леммы 1 следует, что при возрастании x от a до b определитель $\Phi(x)$ изменяет свой знак n раз и $\Phi(a) = \Phi(b)$. Это возможно лишь при n четном, что и требовалось доказать.

Приведем примеры систем Чебышева, основываясь на теореме 5.

Пример 1. Пусть $p_0(x) = r_0(x) \equiv 1$, $q_0(x) \equiv 0$, $\lambda_n = n^2$, $a = 0$, $b = 2\pi$, тогда согласно лемме 2 система тригонометрических функций

$$1, \sin x, \cos x, \ldots, \sin nx, \cos nx \qquad (n = 0, 1, \ldots)$$

образует периодическую систему Чебышева на интервале $[0, 2\pi]$.

Пример 2. Положим $p_0(x) = x^2$, $q_0(x) \equiv 0$, $r_0(x) \equiv 1$, $\lambda_n = -n(n+1)$, тогда система функций x^i $(i = 0, 1, \ldots)$ образует систему Чебышева на любом отрезке $[a, b] \in (-\infty, \infty)$. Заметим, что отсюда следует существование и единственность алгебраического интерполяционного многочлена.

Пример 3. Пусть $p_0(x) = r_0(x) = x$, $q_0(x) \equiv 0$, $\lambda_n = \sqrt{\mu_n}$, $n = 1, 2, \ldots$, μ_n — нули функции Бесселя первого рода нулевого порядка $J_0(x)$, тогда система функций $J_0(\sqrt{\mu_k}x)$ $(k = 1, 2, \ldots)$ образует систему Чебышева на любом отрезке $[a, b] \in (0, 1)$.

Глава 2

ИНТЕРПОЛИРОВАНИЕ

§ 1. ИНТЕРПОЛИРОВАНИЕ АЛГЕБРАИЧЕСКИМИ МНОГОЧЛЕНАМИ

Пусть в постановке общей задачи интерполирования (см. § 1, гл. 1) компакт S совпадает с отрезком вещественной оси $[a, b]$ и в точках $x_i \in [a, b]$ $(i = 0, 1, \ldots, n)$ известны значения функции $f(x) \in C[a, b]$ и ее производных до $(\alpha_i - 1)$-го порядка включительно

$$f^{(j)}(x_i) = y_i^{(j)}, \qquad j = 0, 1, \ldots, \alpha_i - 1; \quad i = 0, 1, \ldots, n. \qquad (1)$$

Взяв в качестве базисной системы функций $\{\varphi_k(x)\}_{k=0}^{m}$ последовательность многочленов $\varphi_k(x) \in \pi_k$, поставим задачу отыскания

такого многочлена

$$\Phi(x) = \sum_{k=0}^{m} C_k \varphi_k(x), \tag{2}$$

который удовлетворял бы условиям

$$\Phi^{(j)}(x_i) = y_i^{(j)}, \qquad j = 0, 1, \ldots, \alpha_i - 1; \quad i = 0, 1, \ldots, n, \tag{3}$$

где

$$m = \sum_{i=0}^{n} \alpha_i - 1. \tag{4}$$

При выполнении (4) количество параметров C_k в (2) совпадает с количеством условий (3). Если многочлен $\Phi(x)$ с перечисленными выше свойствами существует, то тем самым определен линейный оператор $H_m : C[a, b] \to \pi_m$, т. е. $H_m(f) = \Phi(x)$.

Очевидно, для существования и единственности интерполяционного многочлена $\Phi(x)$, кроме выполнения (4), требуется, чтобы

$$\begin{vmatrix} \varphi_0(x) & \varphi_1(x_0) & \ldots & \varphi_m(x_0) \\ \varphi_0'(x_0) & \varphi_1'(x_0) & \ldots & \varphi_m'(x_0) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_0^{(\alpha_0-1)}(x_0) & \varphi_1^{(\alpha_0-1)}(x_0) & \ldots & \varphi_m^{(\alpha_0-1)}(x_0) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_m(x_n) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_0^{(\alpha_n-1)}(x_n) & \varphi_1^{(\alpha_n-1)}(x_n) & \ldots & \varphi_m^{(\alpha_n-1)}(x_n) \end{vmatrix} \neq 0. \tag{4'}$$

Рассмотрим далее эту задачу для случая, когда $\varphi_i(x) = x^i$, и найдем общий вид многочлена вида (2). Такой многочлен $H_m(x)$ будем называть *интерполяционным многочленом Эрмита.*

Построим многочлены $H_{ij}(x)$ степени не выше m, удовлетворяющие следующим условиям:

$$H_{ij}^{(p)}(x_k) = \delta_{ik}\delta_{jp}, \tag{5}$$

$$p = 0, 1, \ldots, \alpha_k - 1, \qquad k = 0, 1, \ldots, n,$$

тогда

$$H_m(x) = \sum_{i=0}^{n} \sum_{j=0}^{\alpha_i - 1} y_i^{(j)} H_{ij}(x). \tag{6}$$

Задача, таким образом, сводится к нахождению $H_{ij}(x)$. Так как $H_{ij}(x)$ имеет в точках $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ нули соответственно кратности $\alpha_0, \alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n$, а в точке x_i нуль кратности j, то

$$H_{ij}(x) = (x - x_0)^{\alpha_0}(x - x_1)^{\alpha_1} \ldots (x - x_{i-1})^{\alpha_{i-1}}(x - x_i)^j \times$$
$$\times (x - x_{i+1})^{\alpha_{i+1}} \ldots (x - x_n)^{\alpha_n} \overline{H}_{ij}(x), \tag{7}$$

где $\overline{H}_{ij}(x)$ — многочлен степени $\alpha_i - j - 1$, не обращающийся в нуль при $x = x_i$. Представим его в виде

$$\overline{H}_{ij}(x) = A_{ij}^{(0)} + A_{ij}^{(1)}(x - x_i) + \cdots + A_{ij}^{(\alpha_i - j - 1)}(x - x_i)^{\alpha_i - j - 1}. \tag{8}$$

Если обозначить

$$\Omega(x) = (x - x_0)^{\alpha_0} (x - x_1)^{\alpha_1} \ldots (x - x_n)^{\alpha_n}, \tag{9}$$

то

$$A_{ij}^{(0)} + A_{ij}^{(1)} (x - x_i) + \cdots + A_{ij}^{(\alpha_i - j - 1)} (x - x_i)^{\alpha_i - j - 1} = \frac{(x - x_i)^{\alpha_i - j}}{\Omega(x)} H_{ij}(x). \tag{10}$$

Подставляя сюда $x = x_i$, получим

$$A_{ij}^{(0)} = \lim_{x \to x_i} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right]. \tag{11}$$

В формуле (11) первое отношение непрерывно при $x = x_i$. Следовательно,

$$\lim_{x \to x_i} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right] = \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]_{x = x_i}.$$

Предел второго отношения найдем по правилу Лопиталя:

$$\lim_{x \to x_i} \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right] = \lim_{x \to x_i} \frac{H_{ij}^{(j)}(x)}{j!} = \frac{1}{j!}.$$

Таким образом,

$$A_{ij}^{(0)} = \frac{1}{j!} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]_{x = x_i}. \tag{12}$$

Для коэффициентов $A_{ij}^{(k)}$ аналогично получаем выражение:

$$A_{ij}^{(k)} = \frac{1}{k!} \lim_{x \to x_i} \frac{d^k}{dx^k} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right]. \tag{13}$$

Применим правило Лейбница для дифференцирования произведения:

$$\frac{d^k}{dx^k} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right] = \sum_{p=0}^{k} C_k^p \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right]^{(k-p)}.$$

Производные

$$\left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)}$$

непрерывны при $x = x_i$, поэтому

$$\lim_{x \to x_i} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} = \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)}_{x = x_i}.$$

Для определения

$$\lim_{x \to x_i} \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right]^{(k-p)}$$

воспользуемся тем же приемом, что и для нахождения $A_{ij}^{(k)}$. Многочлен $H_{ij}(x)$ имеет степень не выше m и делится на $(x - x_i)^j$. Следовательно, его можно записать в виде

$$H_{ij}(x) = B_{ij}^{(0)} (x - x_i)^j + B_{ij}^{(1)} (x - x_i)^{j+1} + \cdots + B_{ij}^{(m-j)} (x - x_i)^m,$$

или

$$\frac{H_{ij}(x)}{(x-x_i)^j} = B_{ij}^{(0)} + B_{ij}^{(1)}(x-x_i) + \cdots + B_{ij}^{(m-j)}(x-x_i)^{m-j}.$$

Отсюда

$$\lim_{x\to x_i}\left[\frac{H_{ij}(x)}{(x-x_i)}\right]^{(k-p)} = (k-p)!\ B_{ij}^{(k-p)}.$$

Но так как $B_{ij}^{(k-p)}$ являются коэффициентами разложения по степеням $(x-x_i)$, то

$$B_{ij}^{(k-p)} = \frac{H_{ij}^{(j+k-p)}(x_i)}{(j+k-p)!}.$$

В нашем случае

$$j+k-p \leqslant j+k \leqslant j+\alpha_i-j-1 = \alpha_i-1.$$

Таким образом, $B_{ij}^{(k-p)}$ отличны от нуля только при $p=k$ и в этом случае

$$B_{ij}^{(0)} = \frac{1}{j!}.$$

Итак,

$$A_{ij}^{(k)} = \frac{1}{k!}\lim_{x\to x_i}\frac{d^k}{dx^k}\left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)}\frac{H_{ij}(x)}{(x-x_i)^j}\right] = \frac{1}{k!j!}\left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)}\right]^{(k)}_{x=x_i} \tag{14}$$

и

$$H_{ij}(x) = \frac{1}{j!}\frac{\Omega(x)}{(x-x_i)^{\alpha_i-j}}\sum_{k=0}^{\alpha_i-j-1}\frac{1}{k!}\left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)}\right]^{(k)}_{x=x_i}(x-x_i)^k. \tag{15}$$

Учитывая формулу (6), имеем:

$$H_m(x) = \sum_{i=0}^{n}\sum_{j=0}^{\alpha_i-1}\sum_{k=0}^{\alpha_i-j-1} y_i^{(j)}\frac{1}{k!}\frac{1}{j!}\left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)}\right]^{(k)}_{x=x_i}\frac{\Omega(x)}{(x-x_i)^{\alpha_i-j-k}}. \tag{16}$$

Лемма 1. *Интерполяционный многочлен* (16) *удовлетворяет условиям* (5) *и является единственным.*

Доказательство. То, что многочлен $H_m(x)$ удовлетворяет условиям (5), следует из построения. Доказательство единственности проводим от противного. Пусть имеется два многочлена $\overline{H}_m(x)$ и $\overline{\overline{H}}_m(x)$, удовлетворяющих условиям (5). Тогда их разность

$$\overline{H}_m(x) - \overline{\overline{H}}_m(x)$$

представляла бы многочлен степени не выше m, имеющий на отрезке $[a, b]$ $m+1$ корень (с учетом их кратности). Пришли к противоречию. Этим завершается доказательство леммы.

Положим в формуле (16) $\alpha_0 = \alpha_1 = \ldots = \alpha_n = 1$. Тогда $\Omega(x) = \omega_n(x) = (x-x_0)(x-x_1)\ldots(x-x_n)$; $k=0$,

$$\left[\frac{x-x_i}{\Omega(x)}\right]^{(k)}_{x=x_i} = \left[\frac{x-x_i}{\omega_n(x)}\right]_{x=x_i} = \frac{1}{\omega_n'(x_i)}$$

и формула (16) принимает вид

$$L_n(x) = \sum_{i=0}^{n} y_i \frac{\omega_n(x)}{(x - x_i)\,\omega_n'(x_i)} = \sum_{i=0}^{n} y_i Q_{ni}(x). \qquad (17)$$

Многочлен $L_n(x)$ называется *интерполяционным многочленом Лагранжа.* Нахождение этих многочленов связано с большой вычислительной работой. Также велика вычислительная работа при получении значения $L_n(x)$ для какого-то фиксированного значения x. Если найден полином Лагранжа, построенный по узлам $x_0, x_1, \ldots, x_n$, то это мало чем помогает при построении многочлена Лагранжа для узлов $x_0, x_1, \ldots, x_n, x_{n+1}$ и поэтому необходимо усовершенствовать формулы Лагранжа с целью упрощения вычислительного процесса.

Запишем вид интерполяционного многочлена Лагранжа для случая равноотстоящих узлов.

Пусть

$$x_1 - x_0 = x_2 - x_1 = \cdots = x_n - x_{n-1} = h, \quad y_i = f(x_i)$$

и $\frac{x - x_0}{h} = t$. Тогда

$$\frac{\omega_n(x)}{(x - x_i)\,\omega_n'(x_i)} = \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)} =$$

$$= \frac{th\,(th - h)\ldots[th - (i-1)h]\,[th - (i+1)h]\ldots(th - nh)}{ih\,(i-1)\ldots h\,(-h)\ldots[-(n-i)h]} =$$

$$= \frac{t(t-1)\ldots(t-n)}{t-i} \cdot \frac{(-1)^{n-i}}{i!\,(n-i)!} =$$

$$= (-1)^{n-i} C_n^i \frac{1}{t-i} \cdot \frac{t(t-1)\ldots(t-n)}{n!}.$$

Подставляя это выражение в (17), будем иметь

$$L_n(x) = L_n(x_0 + th) = (-1)^n \frac{t(t-1)\ldots(t-n)}{n!} \sum_{i=0}^{n} (-1)^i \frac{C_n^i y_i}{t-i}. \qquad (18)$$

В последнем выражении коэффициенты, стоящие перед y_i,

$$(-1)^n C_n^i \frac{t(t-1)\ldots(t-n)}{(t-i)\,n!}$$

не зависят ни от функции $f(x)$, ни от шага h. Поэтому их можно протабулировать и использовать для любой функции $f(x)$ и шага h. Такие таблицы составлены и называются *таблицами коэффициентов Лагранжа.*

В случае, если требуется найти не общее выражение $L_n(x)$, а лишь его значения при некоторых x, то удобно пользоваться так называемой *интерполяционной схемой Эйткена.* По этой схеме значение интерполяционного многочлена для какого-то значения x находится путем последовательного применения единообразного процесса.

Введем следующее выражение

$$L_{01}(x) = \frac{\begin{vmatrix} y_0 & x_0 - x \\ y_1 & x_1 - x \end{vmatrix}}{x_1 - x_0}.$$

Ввиду того что

$$L_{01}(x_i) = y_i, \qquad i = 1,\ 2$$

и, кроме того, $L_{01}(x)$ является многочленом первой степени относительно x, то $L_{01}(x)$ решает задачу многочленного интерполирования по двум узлам. Точно также можно образовать $L_{12}(x)$, $L_{23}(x)$ и т. д.

Рассмотрим далее выражение

$$L_{012}(x) = \frac{\begin{vmatrix} L_{01}(x) & x_0 - x \\ L_{12}(x) & x_2 - x \end{vmatrix}}{x_2 - x_0}.$$

Легко проверить, что будут справедливы равенства

$$L_{012}(x_i) = y_i, \qquad i = 0,\ 1,\ 2.$$

Следовательно, $L_{012}(x)$ совпадает с интерполяционным многочленом, принимающим в точках x_0, x_1, x_2 соответственно значения y_0, y_1, y_2. Вообще,

$$L_{0123\ldots n}(x) = \frac{1}{x_n - x_0} \begin{vmatrix} L_{012\ldots(n-1)}(x) & x_0 - x \\ L_{123\ldots n}(x) & x_n - x \end{vmatrix} \tag{19}$$

будет интерполяционным многочленом, принимающим в точках x_0, x_1, ..., x_n соответственно значения y_0, y_1, ..., y_n, причем порядок точек и их нумерация при этом не имеют значения. Каждый многочлен $L_{012\ldots k}(x)$ получается из $L_{012\ldots(k-1)}(x)$ и $L_{123\ldots k}(x)$ так же, как и $L_{01}(x)$ получается из y_0 и y_1. Вычислительная схема для получения интерполяционного многочлена будет выглядеть следующим образом:

x_i	y_i	$x_i - x$	$L_{i-1,\,i}$	$L_{i-2,\,i-1,\,i}$	$L_{i-3,\,i-2,\,i-1,\,i}$	$L_{i-4,\,i-3,\,i-2,\,i-1,\,i}$
x_0	y_0	$x_0 - x$				
x_1	y_1	$x_1 - x$	$L_{01}(x)$			
x_2	y_2	$x_2 - x$	$L_{12}(x)$	$L_{012}(x)$		
x_3	y_3	$x_3 - x$	$L_{23}(x)$	$L_{123}(x)$	$L_{0123}(x)$	
x_4	y_4	$x_4 - x$	$L_{34}(x)$	$L_{234}(x)$	$L_{1234}(x)$	$L_{01234}(x)$
x_5	y_5	$x_5 - x$	$L_{45}(x)$	$L_{345}(x)$	$L_{2345}(x)$	$L_{12345}(x)$

Применяя такую схему, мы можем постепенно подключать все новые и новые значения x_i до тех пор, пока станет ясно, что точность уже не возрастает. Как видим, интерполяционный процесс Эйткена характерен своим единообразием и поэтому легко реализуется на ЭВМ.

Отметим, что в силу леммы 1 интерполяционный многочлен, построенный по схеме Эйткена, совпадает с интерполяционным многочленом Лагранжа и отличается от последнего лишь формой записи.

Пусть $x_{r-q}, x_{r-q+1} \ldots, x_r, \ldots, x_{r+p}$ — узлы интерполирования, которые мы разместим в произвольном порядке $x_{s_1}, x_{s_2}, \ldots, x_{s_{p+q+1}}$. Чтобы зафиксировать выбранный порядок следования интерполяционных узлов, обозначим через s перестановку чисел $r-q, r-q+1, \ldots$ $\ldots, r+p$, при которой каждому числу $r-q+i-1$ поставлено во взаимно-однозначное соответствие число s_i, при всех $i = 1, 2, \ldots, p+$ $+q+1$, т. е. $s: (r-q+i-1) \to s_i, i = 1, 2, \ldots, p+q+1$.

Пусть $L_k^s(x)$ — интерполяционный многочлен Лагранжа, построенный для функции $f(x)$ по узлам $x_{s_1}, x_{s_2}, \ldots, x_{s_{k+1}}$. Тогда

$$L_{p+q}^s(x) = L_0^s(x) + [L_1^s(x) - L_0^s(x)] + [L_2^s(x) - L_1^s(x)] + \\ + \cdots + [L_{p+q}^s(x) - L_{p+q-1}^s(x)]. \tag{20}$$

Рассмотрим разность $L_k^s(x) - L_{k-1}^s(x)$. По построению — это многочлен k-й степени, обращающийся в нуль в узлах $x_{s_1}, x_{s_2}, \ldots, x_{s_k}$, следовательно, он имеет вид:

$$L_k^s(x) - L_{k-1}^s(x) = A \prod_{i=1}^{k} (x - x_{s_i}), \tag{21}$$

где A — некоторая постоянная. Чтобы определить A, положим $x =$ $= x_{s_{k+1}}$, тогда получим

$$L_k^s(x_{s_{k+1}}) - L_{k-1}^s(x_{s_{k+1}}) = f(x_{s_{k+1}}) - L_{k-1}^s(x_{s_{k+1}}) = A \prod_{i=1}^{k} (x_{s_{k+1}} - x_{s_i}).$$

Отсюда находим

$$A = \frac{f(x_{s_{k+1}})}{\prod\limits_{i=1}^{k} (x_{s_{k+1}} - x_{s_i})} - \frac{\sum\limits_{j=1}^{k} f(x_{s_j}) \dfrac{\omega_{k-1}^s(x_{s_{k+1}})}{(x_{s_{k+1}} - x_{s_j})[\omega_{k-1}^s(x)]_{x=x_{s_j}}}}{\prod\limits_{i=1}^{k} (x_{s_{k+1}} - x_{s_i})} =$$

$$= \sum_{j=1}^{k+1} \frac{f(x_{s_j})}{\prod\limits_{\substack{i=1 \\ i \neq j}}^{k+1} (x_{s_j} - x_{s_i})} = f(x_{s_1}; x_{s_2}; \ldots; x_{s_{k+1}}), \tag{22}$$

где $\omega_{k-1}^s(x) = \prod\limits_{i=1}^{k} (x - x_{s_i})$. Здесь мы воспользовались свойством симметрии разделенных разностей (приложение, § 3).

С учетом (21), (22), формула (20) примет вид

$$L_{p+q}^s(x) = f(x_{s_1}) + (x - x_{s_1}) f(x_{s_1}; x_{s_2}) + \cdots + \\ + \prod_{i=1}^{p+q} (x - x_{s_i}) f(x_{s_1}; \ldots; x_{s_{p+q+1}}). \tag{23}$$

Ввиду того что на практике таблица разделенных разностей имеет вполне определенную структуру, не всякий интерполяционный многочлен вида (23) может быть построен по такой таблице.

Определение 1. Назовем перестановку s *допустимой* для интерполяционного многочлена $L^s_{p+q}(x)$, если при всех $i = 1, 2, \ldots, p+q+1$ последовательность $s_1, s_2, \ldots, s_i$ может быть переупорядочена в последовательность $t_1, t_2, \ldots, t_i$, для которой $t_k - t_{k-1} = 1$ $(k = 2, 3, \ldots, i)$.

Лемма 2. *Пусть s и $\bar{s}$ — две допустимые перестановки, тогда:*

1) *интерполяционные многочлены $L^s_{p+q}(x)$ и $L^{\bar{s}}_{p+q}(x)$ вида* (23) *могут быть построены по таблице разделенных разностей* (приложение, § 3);

2) $L^s_{p+q}(x) \equiv L^{\bar{s}}_{p+q}(x) \equiv L_{p+q}(x)$;

3) $\alpha L^s_{p+q}(x) + \beta L^{\bar{s}}_{p+q}(x) \equiv L_{p+q}(x)$, $\forall\, \alpha, \beta$ *таких, что* $\alpha + \beta = 1$, *где $L_{p+q}(x)$ — интерполяционный многочлен Лагранжа, построенный по узлам $x_{r-q}, x_{r-q+1}, \ldots, x_{r+p}$.*

Доказательство очевидно.

Пользуясь леммой 2, получим из формулы (23) наиболее распространенные интерполяционные многочлены:

а) при $r = q = 0$, $p = n$, $s : (i) \to (i)$, $i = 0, 1, \ldots, n$

$$L_n(x) = \sum_{k=0}^{n} \prod_{i=0}^{k-1} (x - x_i) f(x_0; x_1; \ldots; x_k), \quad \left(\prod_{i=0}^{-1} = 1\right) \tag{24}$$

— *формула Ньютона для интерполирования вперед;*

б) при $r = q = 0$, $p = n$, $s : (i) \to (n - i)$, $i = 0, 1, \ldots, n$

$$L_n(x) = \sum_{k=0}^{n} \prod_{i=0}^{k-1} (x - x_{n-i}) f(x_n; x_{n-1}; \ldots; x_{n-k}) \tag{25}$$

— *формула Ньютона для интерполирования назад;*

в) при $r = 0$, $q = n$, $p = n + 1$, $s : (-n, -n+1, \ldots, n+1) \to$ $\to (0, 1, -1, 2, \ldots, -n, n+1)$

$$L_{2n+1}(x) = f(x_0) + (x - x_0) f(x_0; x_1) + \cdots + \\ + (x - x_0)(x - x_1) \ldots (x - x_{-n}) f(x_0; x_1; \ldots; x_{n+1}). \tag{26}$$

— *формула Гаусса для интерполирования вперед;*

г) при $r = 0$, $q = n$, $p = n + 1$, $s : (-n, -n+1, \ldots, n+1) \to$ $\to (1, 0, 2, -1, \ldots, n+1; -n)$

$$L_{2n+1}(x) = f(x_1) + (x - x_1) f(x_1; x_0) + \cdots + \\ + (x - x_1)(x - x_0) \ldots (x - x_{n+1}) f(x_1; x_0; \ldots; x_{-n}) \tag{27}$$

— *формула Гаусса для интерполирования назад;*

д) $r = 0$, $q = n$, $p = n + 1$.

Среднее арифметическое формул Гаусса (26) и (27) носит название *интерполяционной формулы Бесселя:*

$$L_{2n+1}(x) = \frac{1}{2}[f(x_0) + f(x_1)] + \left(x - \frac{x_0 + x_1}{2}\right) f(x_0; x_1) + \cdots + \\ + (x - x_0)(x - x_1)(x - x_{-1}) \ldots (x - x_n)\left(x - \frac{x_{n+1} + x_{-n}}{2}\right) \times \\ \times f(x_0; x_1; x_{-1}; \ldots; x_{n+1}); \tag{28}$$

е) $r = 0$, $q = p = n$.

Среднее арифметическое двух формул Гаусса, получающихся из (23) с помощью подстановок

$$s:(-n,\ -n+1,\ \ldots,\ n)\to(0,\ 1,\ -1,\ 2,\ \ldots,\ n,\ -n);$$

$$\bar{s}:(-n,\ -n+1,\ \ldots,\ n)\to(0,\ -1,\ 1,\ -2,\ \ldots,\ -n,\ n),$$

носит название *интерполяционной формулы Стирлинга* и имеет вид:

$$L_{2n}(x)=f(x_0)+(x-x_0)\frac{1}{2}[f(x_0;\ x_1)+f(x_0;\ x_{-1})]+$$

$$+(x-x_0)\left(x-\frac{x_{-1}+x_1}{2}\right)f(x_0;\ x_1;\ x_{-1})+\cdots+(x-x_0)(x-x_1)\times$$

$$\times(x-x_{-1})\ldots(x-x_{n-1})(x-x_{-n+1})\left(x-\frac{x_{-n}+x_n}{2}\right)\times$$

$$\times f(x_0;\ x_1;\ x_{-1};\ \ldots;\ x_{-n}). \qquad (29)$$

Формула записи интерполяционного многочлена в виде (23) более удобна для вычислений, чем формула Лагранжа. Добавление одного или нескольких узлов не приводит к повторению всей проделанной работы заново, как это было при вычислениях по формуле Лагранжа.

Если узлы интерполирования равноотстоящие, то формулы (24) — (28) несколько упрощаются. Так, чтобы получить формулу Ньютона для интерполирования вперед в случае равных промежутков $x_i = x_0 + ih$ $(i = 0, 1, \ldots, n)$, заменим в формуле (24) разделенные разности их выражениями через конечные разности (приложение, § 3), в результате чего получим

$$L_n(x)=L_n(x_0+th)=f_0+t\Delta f_0+\cdots+\frac{t(t-1)\ldots(t-n+1)}{n!}\Delta^n f_0, \qquad (24')$$

где $t=\dfrac{x-x_0}{h}$.

Аналогично, если исходить из формулы (25), то приходим к интерполяционному многочлену вида

$$L_n(x)=L_n(x_n+th)=f_n+t\Delta f_{n-1}+\cdots+\frac{t(t+1)\ldots(t+n-1)}{n!}\Delta^n f_0, \qquad (25')$$

где $t=\dfrac{x-x_n}{h}$.

Формулы Ньютона для интерполирования вперед и назад (24) и (25) применяются в том случае, когда точка x лежит соответственно вблизи начала и вблизи конца таблицы.

Полагая в формулах Гаусса (26) и (27) $x = x_0 + th$, получим

$$L_{2n+1}(x)=L_{2n+1}(x_0+th)=f_0+t\Delta f_0+\frac{t(t-1)}{2!}\Delta^2 f_{-1}+$$

$$+\cdots+\frac{t(t^2-1)\ldots(t^2-n^2)}{(2n+1)!}\Delta^{2n+1}f_{-n}; \qquad (26')$$

$$L_{2n+1}(x) = L_{2n+1}(x_0 + th) = f_1 + \frac{t-1}{1!}\Delta f_0 + \frac{(t-1)\,t}{2!}\Delta^2 f_0 +$$

$$+ \cdots + \frac{t(t^2-1)\ldots(t^2-(n-1)^2)(t-n)(t-n-1)}{(2n+1)!}\Delta^{2n+1} f_{-n}. \quad (27')$$

Формулу (26′) применяют, когда точка x лежит на интервале $\left(x_0, x_0 + \frac{h}{2}\right]$, а (27′) — когда точка x лежит на $\left[x_0 - \frac{h}{2}, x_0\right)$.

Формулу Бесселя для равных промежутков находим как полусумму формул Гаусса (26′) и (27′) и она будет иметь вид

$$L_{2n+1}(x) = L_{2n+1}(x_0 + th) = \frac{f_0 + f_1}{2} + \left(t - \frac{1}{2}\right)\Delta f_0 +$$

$$+ \frac{t(t-1)}{2!} \cdot \frac{\Delta^2 f_{-1} + \Delta^2 f_0}{2} + \cdots +$$

$$+ \frac{t(t^2-1)\ldots(t^2-(n-1)^2)(t-n)\left(t-\frac{1}{2}\right)}{(2n+1)!}\Delta^{2n+1} f_{-n}. \quad (28')$$

Применять формулу (28′) целесообразно особенно при интерполировании на средину отрезка, т. е. когда $t = \frac{1}{2}$. В этом случае коэффициенты при разностях нечетного порядка обращаются в нуль и формула значительно упрощается.

Приведем, наконец, формулу Стирлинга для равных промежутков. Она получается из (29) заменой $x = x + th$ и имеет вид

$$L_{2n}(x) = L_{2n}(x_0 + th) = f_0 + \frac{t}{2}[\Delta f_0 + \Delta f_{-1}] + \frac{t^2}{2!}\Delta^2 f_{-1} +$$

$$+ \cdots + \frac{t^2(t^2-1)\ldots(t^2-(n-1)^2)}{(2n)!}\Delta^2 f_{-n}. \quad (29')$$

Формула Стирлинга используется, когда x такое, что $|t| \leqslant \frac{1}{4}$.

Отметим, что формула Лагранжа и все ее видоизменения будут пригодны и для функций комплексного переменного.

§ 2. ИНТЕРПОЛИРОВАНИЕ ПЕРИОДИЧЕСКИХ ФУНКЦИЙ

В случае, если интерполируемая функция $f(x)$ обладает свойством $f(a) = f(b)$, естественно, чтобы базисные функции $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_n(x)$ удовлетворяли такому же условию, т. е. $\varphi_i(a) = \varphi_i(b)$. Такие функции можно рассматривать как периодические с периодом $\tau = b - a$. Таким образом, приходим к необходимости использования для построения обобщенного многочлена периодической системы Чебышева на отрезке $[a, b]$, о которой уже говорили ранее (см. § 3, гл. 1). Не ограничивая общности, положим $a = 0$, $b = 2\pi$. Тогда согласно примеру 1, § 3, гл. 1 функции

$$1, \sin x, \cos x, \ldots, \sin nx, \cos nx \quad (1)$$

образуют периодическую систему Чебышева на отрезке $[0, 2\pi]$. Отсюда следует, что любые два тригонометрические многочлена вида

$$T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx), \tag{2}$$

совпадающие в $2n+1$ попарно различных точках из промежутка $[0, 2\pi]$, тождественно равны между собой.

В силу этого замечания и того, что система (1) является системой функций Чебышева, справедливо такое утверждение: для любой периодической функции $f(x)$ с периодом 2π при любом наборе из $2n+1$ попарно различных узлов $x_0, x_1, \ldots, x_{2n} \in [0, 2\pi)$ существует единственный тригонометрический многочлен $T_n(x)$, являющийся интерполяционным многочленом для $f(x)$ по данной системе узлов интерполирования, т. е. удовлетворяющий условиям

$$T_n(x_j) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx_j + b_k \sin kx_j) = f(x_j) \tag{3}$$

$$(j = 0, 1, 2, \ldots, 2n).$$

В дальнейшем будем рассматривать случай, когда $f(x)$ — 2π — периодическая функция, заданная на $[0, 2\pi]$, так как любой отрезок $[a, b]$ линейной заменой переменного можно привести к $[0, 2\pi]$.

Покажем, что тригонометрическим интерполяционным многочленом для функции $f(x)$ по системе узлов $x_0, x_1, \ldots, x_{2n}$ ($x_i \in [0, 2\pi)$, $x_i \neq x_j$ при $i \neq j$) будет функция

$$T_n(x) = \sum_{i=0}^{2n} f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^{2n} \frac{\sin \frac{x - x_j}{2}}{\sin \frac{x_i - x_j}{2}} = \sum_{i=0}^{2n} f(x_i)\, \Phi_i(x). \tag{4}$$

Действительно, то, что функция $T_n(x)$ удовлетворяет условиям (3), следует из самого вида формулы (4) ($\Phi_i(x)$ являются обобщенными фундаментальными многочленами), а справедливость представления $T_n(x)$ в виде (2) легко устанавливается с помощью элементарных тригонометрических преобразований, которые рекомендуются читателю в качестве упражнения.

Если функция $f(x)$ четная на отрезке $[-\pi, \pi]$, то по значениям функции в $n+1$ точке $x_0, x_1, \ldots, x_n$ ($x_i \in [0, \pi)$, $x_i \neq x_j$ при $i \neq j$) можно построить четный тригонометрический многочлен. Нетрудно проверить, что таким многочленом будет

$$T_n(x) = \sum_{i=0}^{n} f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{\cos x - \cos x_j}{\cos x_i - \cos x_j}. \tag{5}$$

Если же функция $f(x)$ нечетная на $[-\pi, \pi]$, то по ее значениям в узлах $x_1, x_2, \ldots, x_n$ ($x_i \in (0, \pi)$, $x_i \neq x_j$ при $i \neq j$) можно построить следующий нечетный интерполяционный многочлен:

$$T_n(x) = \sum_{i=1}^{n} f(x_i) \frac{\sin x}{\sin x_i} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{\cos x - \cos x_j}{\cos x_i - \cos x_j}. \tag{6}$$

Практическое построение интерполяционных тригонометрических многочленов весьма громоздко. Эта задача несколько упрощается в случае равноотстоящих узлов.

Так, при $x_i = x_0 + ih$, где $i = 0, 1, 2, \ldots, 2n$, $h = \frac{2\pi}{2n+1}$, $0 < x_0 \leqslant \frac{2\pi}{2n+1}$, формула (4) приобретает вид

$$T_n(x) = \sum_{i=0}^{2n} f(x_i) \frac{1}{2n+1} \frac{\sin\left((2n+1)\frac{x-x_i}{2}\right)}{\sin\left(\frac{x-x_i}{2}\right)}. \tag{7}$$

§ 3. АНАЛИЗ ПОГРЕШНОСТИ ИНТЕРПОЛЯЦИОННЫХ ФОРМУЛ

Пусть функция $f(x)$ определена в n узлах интерполяции $x_\nu \in [a, b]$ ($\nu = 1, 2, \ldots, n$), а $T(x)$ — *ее интерполирующая функция*, т. е.

$$f^{(\mu)}(x_i) = T^{(\mu)}(x_i), \qquad \mu = 0, 1, \ldots, m_i - 1, \quad i = 1, 2, \ldots, k, \tag{1}$$

причем $\sum_{p=1}^{k} m_p = n$.

Теорема 1. *Пусть $f(x) \in C^{(n)}[a, b]$, $T(x)$ — ее интерполирующая функция с кратными узлами интерполяции, причем $T(x) \in C^{(n)}[a, b]$.*

Если $\exists\, g(x) \in C^{(n)}[a, b]$ и такая, что

$$g^{(\mu)}(x_i) = 0, \qquad \mu = 0, 1, \ldots, m_i - 1; \quad i = 1, 2, \ldots, k, \tag{2}$$

$$g^{(n)}(x) \neq 0, \quad \forall x \in [a, b],$$

то $\forall x \in [a, b]$ и $x \neq x_i$ общий вид остаточного члена интерполяционной формулы определяется следующим соотношением:

$$R(x) = f(x) - T(x) = \frac{f^{(n)}(\xi) - T^{(n)}(\xi)}{g^{(n)}(\xi)} g(x). \tag{3}$$

$$\xi \in (a, b).$$

Доказательство. Рассмотрим функцию

$$F(t) = f(t) - T(t) - \lambda g(t). \tag{4}$$

Эта функция $F(t) \in C^{(n)}[a, b]$ и $F^{(\mu)}(x_i) = 0$; $\mu = 0, 1, \ldots, m_i - 1$, $i = 1, 2, \ldots, k$. Выберем λ из условия $F(x) = 0$, т. е.

$$\lambda = \frac{f(x) - T(x)}{g(x)}, \tag{5}$$

тогда функция $F(t)$ при $t \in [a, b]$ обращается в нуль в $(n+1)$-й точке с учетом кратности, и на основании обобщенной теоремы Ролля $\exists\, \xi \in (a, b)$, для которой $F^{(n)}(\xi) = 0$. Учитывая формулы (4) и (5), из последнего соотношения получаем утверждение теоремы.

Рассмотрим частные случаи формулы (3). 1. Пусть все интерполяционные узлы простые (не кратные) и $g(x) = \prod_{i=1}^{n}(x - x_i)$, тогда, если

$T(x)$ — интерполяционный многочлен $(n-1)$-й степени, то формула (3) дает остаточный член n-точечной интерполяционной формулы Лагранжа:

$$R(x) = f(x) - T(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{n} (x - x_i). \qquad (6)$$

2. Пусть $T(x) = H_{n-1}(x)$—интерполяционный многочлен $(n-1)$-й степени с кратными интерполяционными узлами и $g(x) = \prod_{i=1}^{k} (x - x_i)^{m_i}$, $\sum_{i=1}^{k} m_i = n$, тогда $H_{n-1}(x)$ является интерполяционным многочленом Эрмита и формула (3) принимает вид

$$R(x) = f(x) - T(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{k} (x - x_i)^{m_i}. \qquad (7)$$

На основании этой формулы можно доказать равномерную сходимость $H_m(x)$ к $f(x)$ на $[a, b]$, если $f(x)$ — целая функция. Однако для интерполяционных многочленов Эрмита Фейер в 1930 г. доказал более сильное утверждение:

Если $x_i^{(n)}$ $(i = 0, 1, 2, \ldots, n)$ являются корнями многочлена Чебышева $T_{n+1}(x)$ и многочлен Эрмита $H_{2n+1}(x)$ удовлетворяет условиям

$$H_{2n+1}(x_i^{(n)}) = f(x_i^{(n)}); \quad H'_{2n+1}(x_i^{(n)}) = y'_{in}, \qquad i = 0, 1, 2, \ldots, n,$$

где y'_{in} — произвольные числа, удовлетворяющие условию

$$\lim_{n\to\infty} \max_i \frac{|y'_{in}| \ln n}{n} = 0,$$

а $f(x)$ — произвольная функция, непрерывная на отрезке $[-1, 1]$, то

$$\lim_{n\to\infty} H_{2n+1}(x) = f(x),$$

причем сходимость равномерная на $[-1, 1]$.

3. Если в предыдущем случае $x_1 = x_2 = \ldots = x_n$, то $T(x)$ — многочлен Тейлора, а формула (7) — остаточный член формулы Тейлора в форме Коши.

Если все вычисления проводятся точно, то интерполяционный многочлен совпадает с заданной функцией $f(x)$ в узлах интерполяции $x_0, x_1, \ldots, x_n$. Однако он будет отличен от нее в остальных точках (за счет погрешности метода). Исключение представляет тот случай, когда сама функция $f(x)$ является многочленом степени не выше n. Тогда $f(x)$ и $L_n(x)$ тождественно совпадают.

Погрешности возникают также от того, что y_i могут оказаться приближенными (неустранимая погрешность) и в процессе вычислений будет возникать погрешность за счет округлений (погрешность округлений).

Если интерполируемая функция $f(x)$ имеет на $[a, b]$ непрерывные производные до n-го порядка и $f^{(n)}(x)$ дифференцируема на $[a, b]$,

то, как видно из (6), погрешность метода для многочлена Лагранжа имеет вид

$$R(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)(x-x_1)\ \ldots\ (x-x_n), \tag{8}$$

где $\xi \in (a, b)$. Полагая здесь $M_{n+1} = \sup\limits_{x\in[a,b]} |f^{(n+1)}(x)|$, получим

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|(x-x_0)(x-x_1)\ \ldots\ (x-x_n)|. \tag{9}$$

Как видно из (3), отклонение $f(x)$ от $L_n(x)$ определяется величинами $f^{(n+1)}(\xi)$ и $\omega_n(x)$. Если о первой величине можно только сказать, в каких пределах она изменяется, то вторую можно изменять посредством надлежащего выбора точек x_i.

Решим такую задачу: выбрать узлы x_i так, чтобы величина $\sup\limits_{x\in[a,b]} |\omega_n(x)|$ была наименьшей. Для этого придется воспользоваться некоторыми свойствами многочленов Чебышева первого рода (приложение, § 2):

$$T_n(x) = \cos[n \arccos x], \qquad |x| \leqslant 1.$$

Нули многочленов Чебышева первого рода определяются формулой

$$x_m = \cos\frac{2m+1}{2n}\pi; \qquad m = 0,\ 1,\ \ldots,\ n-1.$$

Заметим, что $\max\limits_{x\in[-1,1]} |T_n(x)|$ на отрезке $[-1,1]$ равен 1 и достигается в $n+1$ точках $x_m = \cos\frac{m\pi}{n}$ $(m = 0, 1, \ldots, n)$. Если теперь в качестве отрезка интерполирования взять $[-1,1]$, а в качестве узлов — корни многочлена Чебышева x_m, то

$$\omega_n(x) = \frac{1}{2^n}T_{n+1}(x) \text{ и } \sup_{x\in[-1,1]} |\omega_n(x)| = \frac{1}{2^n}.$$

Теорема 2. *Среди всех многочленов n-й степени со старшим коэффициентом 1 многочлен*

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

наименее отклоняется от нуля.

Доказательство. Покажем, что какой бы многочлен $P_n(x)$ степени n со старшим коэффициентом 1 мы ни взяли, $\sup\limits_{x\in[-1,1]} |P_n(x)| \geqslant \frac{1}{2^{n-1}}$. Действительно, если бы это было не так, то разность $\frac{1}{2^{n-1}}T_n(x) - P_n(x)$ представляла бы собой многочлен степени $n-1$, принимающий в $n+1$ точках $x_m = \cos\frac{m\pi}{n}$ $(m = 0, 1, \ldots, n)$ попеременно положительные и отрицательные значения. Поэтому он должен иметь хотя бы n нулей, что невозможно. Противоречие доказывает теорему.

Из доказанной теоремы следует: если отрезок интерполирования есть $[-1,1]$, то $\sup\limits_{x\in[-1,1]} |\omega_n(x)|$ принимает наименьшее значение при условии, что в качестве узлов интерполирования взяты корни многочлена Чебышева, и оценка (9) в этом случае примет вид

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{2^n (n+1)!}. \tag{10}$$

Если интерполирование производится на произвольном отрезке $[a, b]$, то заменой

$$x = \frac{1}{2}[(b-a)z + (b+a)], \quad z = \frac{1}{b-a}[2x - b - a]$$

он переводится в $[-1,1]$. При этом корни многочлена $T_{n+1}(x)$ перейдут в

$$x_m = \frac{1}{2}\left[(b-a)\cos\frac{2m+1}{2n+2}\pi + (b+a)\right], \qquad m = 0, 1, \ldots, n$$

и оценка в этом случае имеет вид

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} \cdot \frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{11}$$

Полученные результаты дают наилучшую оценку в целом по всему отрезку $[a, b]$. Мы воспользовались тем свойством многочленов $\overline{T}_n(x) = \frac{1}{2^{n-1}} T_n(x)$, что для них $\sup\limits_{x\in[-1,1]} = |\overline{T}_n(x)|$ имеет наименьшее значение среди всех многочленов степени n с коэффициентом 1 при старшей степени. Благодаря этому свойству многочлены $\overline{T}_n(x)$ получили название *многочленов, наименее уклоняющихся от нуля.*

Рассмотрим теперь другую задачу: при фиксированных узлах интерполирования изучить, для каких промежутков изменения остаточный член будет принимать бóльшие значения и для каких меньшие. Для решения такой задачи нужно изучить поведение функции $\omega_n(x)$ при фиксированных $x_0, x_1, \ldots, x_n$. Многочлен $\omega_n(x)$ обращается в нуль в точках $x_0, x_1, \ldots, x_n$, меняет знак, переходя через эти точки, и где-то в промежутках между ними принимает попеременно то максимальное, то минимальное значение.

Абсолютные значения этих экстремумов будут равны друг другу только в том случае, если $x_0, x_1, \ldots, x_n$ являются корнями многочлена

$$\cos\left[(n+1)\arccos\frac{2x-b-a}{b-a}\right].$$

При интерполировании в других случаях вблизи бóльших по абсолютной величине экстремумов можно ожидать бóльшую погрешность.

Ограничимся случаем равноотстоящих узлов, т. е. будем предполагать, что

$$x_1 - x_0 = x_2 - x_1 = \cdots = x_n - x_{n-1} = h.$$

Если обозначить $t = \frac{x - x_0}{h}$, то будем иметь:

$$\omega_n(x) = \omega_n(x_0 + th) = h^{n+1} t(t-1)(t-2)\ldots(t-n).$$

Изучим поведение функции

$$\varphi(t) = t(t-1)(t-2)\ \ldots\ (t-n)$$

при различных значениях t. Заметим, что эта функция будет четной или нечетной относительно точки $\left(\frac{n}{2}, 0\right)$ в зависимости от четности n. Действительно,

$$\varphi(t) = (-1)^{n+1}\varphi(n-t).$$

Далее заметим, что

$$\varphi(t+1) = (t+1)\,t\,(t-1)\ \ldots\ (t+1-n) = \frac{t+1}{t-n}\varphi(t).$$

Значит, если разбить отрезок $[0, n]$ на части $[0, 1], [1, 2], \ldots, [(n-1), n]$, то значение функции на отрезке $[i, i+1]$ будет получаться из соответствующего значения функции на предыдущем отрезке путем умножения его на $\frac{t+1}{t-n}$. Этот множитель всегда отрицателен при $t \in (0, n)$, поэтому знак у функции $\varphi(t)$ будет чередоваться при переходе от интервала к интервалу. Абсолютная величина этого множителя будет меньше 1 на $\left[0, \frac{n-1}{2}\right]$. Таким образом, экстремальные значения $\varphi(t)$ будут убывать по абсолютной величине до середины отрезка $[0, n]$ и затем в силу симметрии снова возрастать. Вне пределов отрезка $[0, n]$ функция $\varphi(t)$ быстро возрастает по абсолютной величине.

Отсюда делаем следующие выводы:

1. Оценка остаточного члена формулы Лагранжа будет особенно велика для значений x, лежащих вне отрезка $[x_0, x_n]$, т. е. если проводить экстраполирование, то следует ожидать большой погрешности.

2. При интерполировании для значений x, лежащих неблизко к узлам интерполирования, точность будет больше для средних отрезков и меньше для крайних.

Анализ неустранимой погрешности формулы Лагранжа показывает, что при изменении t на отрезке $[0, n]$ она сравнительно невелика. При увеличении n неустранимая погрешность незначительно возрастает. Минимальные погрешности получаются в средних отрезках $[i, i+1]$ при изменении t от 0 до n. При экстраполяции опять получаются значительные погрешности.

Оценки ошибок округления здесь не приводятся, так как они во многом определяются программой вычислений.

Остаточный член интерполяционной формулы Ньютона точно такой же, как и формулы Лагранжа, но его можно записать и в другой форме. Для этого, воспользовавшись свойством симметрии разделенных разностей (приложение, § 3), запишем

$$f(x;\ x_0;\ x_1;\ \ldots;\ x_n) = \frac{f(x)}{(x-x_0)(x-x_1)\ \ldots\ (x-x_n)} + \\ + \frac{f(x_0)}{(x_0-x)(x_0-x_1)\ \ldots\ (x_0-x_n)} + \ \cdots\ + \\ + \frac{f(x_n)}{(x_n-x)(x_n-x_0)\ \ldots\ (x_n-x_{n-1})}. \qquad (12)$$

Отсюда

$$f(x) = f(x_0)\frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)} + \cdots +$$

$$+ f(x_n)\frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})} +$$

$$+ (x-x_0)(x-x_1)\ldots(x-x_n) f(x;\ x_0;\ x_1;\ \ldots;\ x_n) =$$

$$= L_n(x) + (x-x_0)(x-x_1)\ldots(x-x_n) f(x;\ x_0;\ x_1;\ \ldots;\ x_n). \quad (13)$$

Значит

$$R_n(x) = f(x) - L_n(x) = (x-x_0)(x-x_1)\ldots(x-x_n) f(x;\ x_0;\ \ldots;\ x_n). \quad (14)$$

В частности, если $f(x)$ имеет производную порядка $n+1$, то получим

$$f(x;\ x_0;\ \ldots;\ x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad (15)$$

где ξ — некоторая точка, принадлежащая наименьшему промежутку, содержащему все точки $x_0, x_1, \ldots, x_n, x$.

Разделенная разность, входящая в (14), может быть найдена только в том случае, когда нам известно $f(x)$. Но тогда нет смысла использовать формулу Ньютона. Однако в некоторых случаях остаточный член в форме (14) можно использовать для фактической оценки погрешности формулами Ньютона.

Пусть нам известно, что разделенные разности порядка $n+1$ и $n+2$ сохраняют постоянные знаки на рассматриваемом отрезке. Запишем равенства

$$f(x) = \sum_{i=0}^{n}(x-x_0)\ldots(x-x_{i-1}) f(x_0;\ x_1;\ \ldots;\ x_i) +$$

$$+ (x-x_0)\ldots(x-x_n) f(x;\ x_0;\ \ldots;\ x_n),$$

$$f(x) = \sum_{i=0}^{n+1}(x-x_0)\ldots(x-x_{i-1}) f(x_0;\ x_1;\ \ldots;\ x_i) +$$

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

из которых с учетом (14) получаем

$$R_n(x) = (x-x_0)(x-x_1)\ldots(x-x_n) f(x_0;\ x_1;\ \ldots;\ x_{n+1}) +$$

$$+ R_{n+1}(x). \quad (16)$$

Для заданного x всегда можно подобрать x_{n+1} так, что $R_n(x)$ и $R_{n+1}(x)$ будут иметь различные знаки. Действительно, если $f(x;\ x_0;\ \ldots;\ x_n)$ и $f(x;\ x_0;\ \ldots;\ x_{n+1})$ имеют одинаковые знаки, то берем $x_{n+1} > x$; если они имеют разные знаки, то берем $x_{n+1} < x$. Тогда

$$\operatorname{sign}[R_n(x)] = \operatorname{sign}[(x-x_0)(x-x_1)\ldots(x-x_n) f(x_0;\ x_1;\ \ldots;\ x_{n+1})]$$

и

$$|R_n(x)| < |(x-x_0)(x-x_1)\ldots(x-x_n) f(x_0;\ x_1;\ \ldots;\ x_{n+1})|. \quad (17)$$

В том случае, если $f(x_{n+1})$ известно, можем фактически оценить $R_n(x)$.

Рассмотрим еще случай, когда на отрезке $[a, b]$, где берутся x и узлы интерполирования, функция $f(x)$ имеет производную $f^{(n+2)}(x)$, сохраняющую знак. Покажем, что в этом случае $f(x; x_0; \ldots; x_n)$ — монотонная функция от x на $[a, b]$. Для этого рассмотрим выражение

$$z = \frac{f(\bar{x}; x_0; \ldots; x_n) - f(\bar{\bar{x}}; x_0; \ldots; x_n)}{\bar{x} - \bar{\bar{x}}},$$

где $\bar{x}$ и $\bar{\bar{x}}$ — некоторые точки отрезка $[a, b]$. В силу симметрии разделенных разностей относительно своих аргументов будем иметь:

$$z = \frac{f(\bar{x}; x_0; \ldots; x_n) - f(x_0; x_1, \ldots; x_n; \bar{\bar{x}})}{\bar{x} - \bar{\bar{x}}},$$

а это есть разделенная разность $f(\bar{x}, \bar{\bar{x}}; x_0; \ldots; x_n)$ порядка $n + 2$ от функции $f(x)$. На основании равенства (15) получаем

$$z = f(\bar{x}; \bar{\bar{x}}; x_0; \ldots; x_n) = \frac{f^{(n+2)}(\xi)}{(n+2)!},$$

откуда следует, что z сохраняет знак на $[a, b]$. Из доказанной монотонности разделенной разности $f(x; x_0; \ldots; x_n)$ вытекают неравенства:

$$f(a; x_0; x_1; \ldots; x_n) \leqslant f(x; x_0; \ldots; x_n) \leqslant f(b; x_0; \ldots; x_n),$$
$$\text{если } f^{(n+2)}(x) > 0, \quad \forall x \in [a, b]; \tag{18}$$
$$f(b; x_0; \ldots; x_n) \leqslant f(x; x_0; \ldots; x_n) \leqslant f(a; x_0; \ldots; x_n),$$
$$\text{если } f^{(n+2)}(x) < 0, \quad \forall x \in [a, b].$$

В этих случаях остаточный член можно оценить, если известны $f(a)$ и $f(b)$.

Известно, что для многочленов разделенные разности, начиная с некоторого порядка, обращаются в нуль. Для функций, не являющихся многочленами, это утверждение не имеет места. Можно показать, что разделенные разности от целых функций стремятся к нулю. Но на практике стремления разделенных разностей к нулю не будет вследствие того, что сами исходные данные обычно бывают приближенными, а в процессе вычисления разделенных разностей еще производится округление. Поэтому часто бывает так, что сначала разделенные разности убывают с повышением порядка, а затем «ведут себя неправильно» и снова растут.

Узлы интерполирования, лежащие ближе всего к интерполируемому значению x, окажут большее влияние на интерполяционный многочлен, лежащие дальше,— меньшее. Целесообразно поэтому за x_0 и x_1 взять ближайшие к x узлы интерполирования и по ним произвести линейную интерполяцию. Затем постепенно привлекать соседние узлы так, чтобы они возможно симметричнее располагались относительно x. Полученные при этом поправки обычно бывают незначительны.

Приведем остаточные члены интерполяционных формул Ньютона для интерполирования вперед и назад.

Для первой формулы имеем:

$$R_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!} =$$

$$= \frac{h^{n+1} f^{(n+1)}(\xi)}{(n+1)!} t(t-1) \ldots (t-n); \tag{19}$$

для второй —

$$R_n(x) = (x - x_n)(x - x_{n-1}) \ldots (x - x_0) \frac{f^{(n+1)}(\xi)}{(n+1)!} =$$

$$= \frac{h^{n+1} f^{(n+1)}(\xi)}{(n+1)!} t(t+1) \ldots (t+n). \tag{20}$$

В некоторых случаях, особенно когда значения f_t получены из эксперимента, бывает очень трудно оценить величину производной $f^{(n+1)}(\xi)$. Дадим простой, хотя и грубый способ такой оценки. Как известно,

$$f(x_0;\ x_0 + h;\ \ldots;\ x_0 + (n+1)h) = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{21}$$

С другой стороны,

$$f(x_0;\ x_0 + h;\ \ldots;\ x_0 + (n+1)h) = \frac{\Delta^{n+1} f_0}{h^{n+1}(n+1)!}. \tag{22}$$

Считая, что на рассматриваемом отрезке производная $f^{(n+1)}(x)$, а значит, и разности $\Delta^{n+1} f_0$ изменяются не сильно, можно заменить производную, входящую в остаточный член, разностью и получим

$$R_n(x) \approx \frac{t(t-1) \ldots (t-n)}{(n+1)!} \Delta^{n+1} f_0. \tag{23}$$

Аналогичную процедуру можно проделать и для второй формулы. Но следует еще раз подчеркнуть, что полученные оценки грубы и применять их можно только в случае крайней необходимости. Если не выполнено условие того, что производная изменяется незначительно, то можно получить абсолютно неправильный результат.

В заключение рассмотрим остаточный член интерполяционного многочлена для функций комплексного переменного.

Пусть C — простая замкнутая кривая, $f(z)$ — аналитическая на C и внутри C функция. Пусть узлы интерполирования $z_0, z_1, \ldots, z_n$ лежат внутри C. Рассмотрим интеграл

$$P(z) = \frac{1}{2\pi i} \int_C \frac{\omega(\xi) - \omega(\xi)}{\omega(\xi)(\xi - z)} f(\xi)\, d\xi, \tag{24}$$

где $\omega(z) = (z - z_0)(z - z_1) \ldots (z - z_n)$. Подынтегральная функция аналитична на C и внутри C, за исключением точек $z_0, z_1, \ldots, z_n$, поэтому интеграл будет равен сумме вычетов относительно каждой из этих точек.

Но

$$\lim_{\xi \to z_k} \frac{\omega(\xi) - \omega(z)}{\omega(\xi)(\xi - z)} f(\xi)(\xi - z_k) = f(\xi_k) \frac{\omega(z)}{\omega'(z_k)(z - z_k)},$$

откуда

$$P(z) = \sum_{k=0}^{n} f(z_k) \frac{\omega(z)}{\omega'(z_k)(z - z_k)} = L_n(z) \tag{25}$$

и $P(z)$ имеет точно такой же вид, как многочлен Лагранжа. Представим $P(z)$ в виде разности

$$P(z) = \frac{1}{2\pi i} \int_C \frac{f(\xi)}{\xi - z} d\xi - \frac{\omega(z)}{2\pi i} \int_C \frac{f(\xi)}{\omega(\xi)(\xi - z)} d\xi.$$

В силу интегральной формулы Коши первый интеграл равен $f(z)$, следовательно,

$$f(z) = P(z) + \frac{\omega(z)}{2\pi i} \int_C \frac{f(\xi)}{\omega(\xi)(\xi - z)} d\xi. \tag{26}$$

Отсюда видно, что остаточный член интерполяционной формулы Лагранжа в нашем случае имеет вид

$$R(z) = \frac{\omega(z)}{2\pi i} \int_C \frac{f(\xi)}{\omega(\xi)(\xi - z)} d\xi. \tag{27}$$

§ 4. СХОДИМОСТЬ ИНТЕРПОЛЯЦИОННЫХ ФОРМУЛ

При использовании интерполяционных формул на практике не всегда удается произвести оценку остаточного члена, ибо входящие в него высшие производные, как правило, не известны. Поэтому наличие сходимости интерполяционного процесса гарантирует, что при достаточно большом количестве узлов интерполяции можно достаточно хорошо приблизиться к интерполируемой функции.

Пусть $f(x) \in C[a, b]$. Для каждого натурального n выберем на отрезке $[a, b]$ $n+1$ различных точек $x_i^{(n)}$ $(i = 0, 1, \ldots, n;\ n = 0, 1, \ldots)$ и построим по этим наборам точек последовательность интерполяционных многочленов: по n-му набору $x_i^{(n)}$ $(i = 0, 1, \ldots, n)$ строим интерполяционный многочлен $L_n(x)$.

Определение 1. Интерполяционный процесс называется *сходящимся*, если

$$\lim_{n\to\infty} L_n(x) = f(x), \qquad x \in [a, b], \tag{1}$$

и *равномерно сходящимся*, если сходимость последнего выражения равномерная.

Будем рассматривать $L_n(x)$ как оператор $L_n : C[a, b] \to \pi_n$, преобразующий $f(x) \in C[a, b]$ в элемент того же пространства — многочлен n-й степени. Оператор L_n — непрерывен. Фундаментальное значение в теории интерполирования имеет норма этого оператора, которая в силу (17), § 1, будет равна:

$$\|L_n\| = \max_{a \leqslant x \leqslant b} \sum_{j=0}^{n} |Q_{nj}(x)|, \qquad n = 0, 1, 2, \ldots \tag{2}$$

и называется *константой Лебега*.

По построению, $\forall f(x) \in \pi_k$ будем иметь

$$L_n(f) = f, \qquad \forall n \geqslant k.$$

Поэтому, чтобы применить следствие из теоремы Банаха — Штейнгауза к доказательству сходимости интерполяционного процесса (приложение, § 1), достаточно изучить поведение $\| L_n \|$ как функций от n.

Теорема 1. *Каковы бы ни были узлы $x_i^{(n)} \in [a, b]$ $(i = 0, 1, \ldots, n)$, всегда будет иметь место неравенство*

$$\| L_n \| \geqslant \frac{2}{\pi^2} \ln n + b(n), \tag{3}$$

где $b(n)$ — ограниченная функция.

Из теоремы 1 и следствия из теоремы Банаха — Штейнгауза следует такая теорема:

Теорема 2. *$\forall x_i^{(n)} \in [a, b]$ $(i = 0, 1, \ldots, n;\ n = 0, 1, \ldots)$, $\exists f(x) \in C[a, b]$, для которой интерполяционный процесс является расходящимся.*

Хотя, как следует из теоремы 2, не существует узлов $x_i^{(n)}$, которые обеспечивали бы сходимость интерполяционного процесса $\forall f(x) \in C[a, b]$, однако существуют такие системы узлов, которые лучше других в смысле величины $\| L_n \|$.

Теорема 3. *Пусть L_n^E — оператор интерполирования по равноотстоящим узлам на отрезке $[-1, 1]$, L_n^T — оператор интерполирования по узлам, совпадающим с нулями многочлена Чебышева первого рода $T_{n+1}(x)$. Тогда*

$$\frac{4}{\pi^2} \ln(n+1) + b(n) \leqslant \| L_n^T \| < \frac{4}{\pi} \ln(n+1) + 8, \tag{4}$$

$$\sqrt{e} \leqslant \overline{\lim_{n \to \infty}} (\| L_n^E \|)^{\frac{1}{n+1}} \leqslant 2, \tag{5}$$

где $b(n)$ имеет тот же смысл, что и в теореме 1.

Приведенная выше теорема 3 показывает, что $\| L_n^E \| \to \infty$ значительно быстрее, чем $\| L_n^T \|$.

Имеет место следующая теорема:

Теорема 4. *$\forall f(x) \in C[a, b]$ $\exists x_i^{(n)} \in [a, b]$ $(i = 0, 1, \ldots, n;\ n = 0, 1, \ldots)$ такие, что $L_n(x) \to f(x)$ в $C[a, b]$ и справедливо равенство*

$$\| L_n(f) - f \|_C = \Delta_n(f) = \inf_{\varphi \in \pi_n} \| \varphi - f \|_C. \tag{6}$$

Доказательство. Пусть $P_n(x)$ — наилучшее приближение к $f(x)$ в π_n. Тогда $\| P_n(x) - f(x) \|_C = \Delta_n(f)$. Так как множество π_n удовлетворяет условию Хаара, то по леммам 1 и 2, § 1, гл. 1 $\exists y_i^{(n)} \in [a, b]$ $(i = 0, 1, \ldots, n+1)$ такие, что будут справедливы равенства:

$$| P_n(y_i^{(n)}) - f(y_i^{(n)}) | = \| P_n - f \|_C, \qquad i = 0, 1, \ldots, n+1;$$

$$P_n(y_i^{(n)}) - f(y_i^{(n)}) = -[P_n(y_{i+1}^{(n)}) - f(y_{i+1}^{(n)})],$$

$$i = 0, 1, \ldots, n.$$

Поэтому функция $P_n(x) - f(x) \in C[a, b]$ и $\exists\, z_i^{(n)} \in [a, b]$ $(i = 0, 1, \ldots, n)$ такие, что

$$y_i^{(n)} \leqslant z_i^{(n)} \leqslant y_{i+1}^{(n)}, \quad P_n(z_i^{(n)}) - f(z_i^{(n)}) = 0, \qquad i = 0, 1, \ldots, n.$$

Выбрав в качестве искомых интерполяционных узлов эти точки $z_i^{(n)}$, получаем $L_n(x) = P_n(x)$, что и требовалось доказать.

Утверждение, аналогичное теореме 2, имеет место и для тригонометрических интерполяционных формул. Существуют модификации интерполяционных формул, в которых за счет повышения степени многочлена добиваются сходимости $\forall f \in C[a, b]$. Простейший пример такого рода — интерполяционный многочлен Джексона. Пусть

$$K_m(\theta) = \frac{2}{n}\left(\frac{\sin\frac{n}{2}\theta}{2\sin\frac{\theta}{2}}\right)^2 \text{— ядро Фейера}, \quad m = 2n - 1, \quad \theta_k = \beta + \frac{2\pi k}{n},$$

$k = 0, 1, \ldots, n - 1$ — узлы интерполяции. По $f \in C[0, 2\pi]$ построим многочлен Джексона:

$$t_m(f;\ \theta) = \frac{2}{n}\sum_{k=0}^{n-1} f(\theta_k)\, K_m(\theta - \theta_k). \tag{7}$$

На основании равенства

$$K_m(\theta) = \frac{2}{n}\left(\frac{\sin\frac{n}{2}\theta}{2\sin\frac{\theta}{2}}\right)^2 =$$

$$= \frac{1}{n}\left[\frac{n}{2} + (n-1)\cos\theta + \cdots + \cos(n-1)\theta\right], \tag{8}$$

тригонометрический многочлен $t_m(f;\ \theta)$ будет порядка $n - 1$. Можно проверить, что

$$t_m(f;\ \theta_k) = f(\theta_k), \quad t_m'(f;\ \theta_k) = 0, \qquad k = 0, 1, \ldots, n-1. \tag{9}$$

Следовательно, речь идет об эрмитовой интерполяции, но полагаем $t_m'(f;\ \theta_k) = 0$, а не $f'(\theta_k)$, так как $f'(x)$ может просто не существовать. Полагая в (7) $f \equiv 1$, получим

$$\|t_m\| = \frac{2}{n}\sum_{k=0}^{n-1} K_m(\theta - \theta_k) = 1. \tag{10}$$

На основании теоремы о сходимости последовательности линейных операторов (приложение, § 1), получаем, что $\forall f(x) \in C[0, 2\pi]$

$$\lim_{m\to\infty} t_m(f;\ \theta) = f(x). \tag{11}$$

Приведем без доказательства одну теорему о сходимости интерполяционных процессов, относящуюся к целым функциям. Напомним, что функция называется целой, если ее можно представить в виде всюду сходящегося степенного ряда

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k. \tag{12}$$

Теорема 5. *Пусть $f(x)$ — произвольная целая функция. Тогда последовательность построенных для нее интерполяционных многочленов $L_n(x)$ по узлам $x_k^{(n)} \in [a, b]$ ($k = 0, 1, \ldots, n$; $n = 0, 1, \ldots$) равномерно сходится на $[a, b]$ к $f(x)$.*

На примере функции $f(x) \in A_r(M)$, где $A_r(M)$ — множество функций, допускающих аналитическое продолжение внутрь эллипса $Э_r$ с фокусами в концах отрезка $[-1,1]$ и суммой полуосей, равной r, причем $|f(x)| \leqslant M$, $\forall x \in Э_r$, рассмотрим вопрос о сходимости интерполяционных многочленов вне отрезка $[-1,1]$, т. е. сходимость экстраполяционного процесса. Исследуем интерполяционные многочлены с узлами в нулях многочленов Чебышева $T_{n+1}(x)$ либо $U_{n+1}(x)$ (приложение, § 2).

Лемма 1. *Если $f(x) \in A_r(M)$, $L_n(x)$ — интерполяционный многочлен Лагранжа с узлами в нулях $T_{n+1}(x)$ либо $U_{n+1}(x)$, то*

$$|f(x) - L_n(x)| \leqslant CM\left(\frac{\rho}{r}\right)^n, \tag{13}$$

где ρ — полусумма осей эллипса, софокусного с $Э_r$ и проходящего через точку x; C — константа, зависящая только от r и ρ.

Доказательство. Поскольку для интерполяционного многочлена Лагранжа остаточный член можно записать через контурный интеграл (27), § 3:

$$|f(x) - L_n(x)| = \frac{1}{2\pi}\left|\int\limits_{Э_r} \frac{\omega(x)}{\omega(t)} \cdot \frac{f(t)\,dt}{x-t}\right|, \tag{14}$$

то отсюда следует оценка

$$|f(x) - L_n(x)| \leqslant \frac{M}{2\pi}\int\limits_{Э_r} \frac{|\omega(x)|}{|\omega(t)|} \cdot \frac{|dt|}{|x-t|}. \tag{14'}$$

Пусть $\omega(x) = T_{n+1}(x)$; если $t \in Э_r$, то

$$\frac{1}{2}(r^n - r^{-n}) \leqslant |T_{n+1}(t)| \leqslant \frac{1}{2}(r^n + r^{-n}).$$

Тогда из (14′) будет следовать

$$|f(x) - L_n(x)| \leqslant \frac{M(\rho^n + \rho^{-n})}{r^n - r^{-n}} \frac{1}{2\pi}\int \frac{|dt|}{|x-t|}.$$

Отсюда получаем (13) с константой $C = \tilde{C}\log\frac{\sqrt{r\rho}}{r-\rho}\max\left(1, \frac{1}{(r-1)\,n}\right)$, где $\tilde{C}$ — абсолютная постоянная. Если интерполировать с узлами в нулях многочлена $U_{n+1}(x)$, то результат будет тот же и лишь изменится константа C.

Из предыдущей леммы следует, что экстраполяция возможна в случае, когда $f(x)$ аналитична на $[-1,1]$. Осталось выяснить, каковы оптимальные узлы при экстраполяции. Понятно, что аналитичность $f(x)$ не только достаточна для возможности экстраполяции, но в известном смысле и необходима. Весьма реалистична следующая постановка

вопроса: определить на промежутке $[-1, 1]$ узлы так, чтобы ошибки, сделанные при вычислении функции в них, приводили к минимальной ошибке величины $P_n^*(\xi)$, где ξ, $(|\xi| > 1)$ — точка, в которой производится экстраполяция функции.

Полагая, что при вычислении значений функции $f(x)$ допускаются ошибки, имеющие случайный характер, приходим к задаче о

$$\min_{x_0, \ldots, x_n} \sum_{j=0}^{n} |Q_{nj}(\xi)|. \tag{15}$$

Лемма 2. (С. Н. Бернштейна). *В задаче* (15) *минимум достигается, когда узлы совпадают с нулями многочлена Чебышева* $U_{n+1}(x)$.

Доказательство. Примем для определенности $\xi > 1$. Легко видеть, что

$$Q(\xi) = \sum_{i=0}^{n} |Q_{nj}(\xi)| = \sum_{j=0}^{n} \varepsilon_j Q_{nj}(\xi),$$

где $\varepsilon_j = \pm 1$ и, следовательно, $Q(\xi)$ совпадает со значением некоторого многочлена степени не выше n, принимающего в узлах значения, равные по модулю единице. Многочлен Чебышева первого рода $T_n(x)$ принимает максимальные по модулю значения, равные единице, в нулях $U_{n+1}(x)$. Поэтому

$$T_n(\xi) = \sum_{j=0}^{n} T_n(x_j) Q_{nj}(\xi) \leqslant \sum_{j=0}^{n} |Q_{nj}(\xi)| = Q(\xi)$$

и знак равенства достигается только в том случае, когда узлы совпадают с нулями $U_{n+1}(x)$, что и требовалось доказать.

§ 5. НЕКОТОРЫЕ ВОПРОСЫ ПРИМЕНЕНИЯ ИНТЕРПОЛЯЦИОННЫХ ФОРМУЛ

Интерполяционные формулы очень широко применяются как в теоретических исследованиях, так и на практике. Остановимся лишь на некоторых аспектах применения теории интерполирования.

1. Обратное интерполирование

В практике вычислений нередко возникает такая задача: по заданному значению функции $f(x)$ определить соответствующее значение аргумента x, если в узлах x_i $(i = 0, 1, \ldots, n)$ известны величины $f(x_i) = y_i$ $(i = 0, 1, \ldots, n)$. Такая задача может быть решена методом обратного интерполирования.

Если заданная в точках x_i функция $f(x)$ монотонна, то искомое значение x — единственно. Примем переменную y за независимую, а x будем считать функцией от y. Тогда, написав по заданным узлам (y_i, x_i) $(i = 0, 1, \ldots, n)$ интерполяционный многочлен Лагранжа

$$x = \sum_{i=0}^{n} x_i \frac{(y - y_0)(y - y_1) \ldots (y - y_{i-1})(y - y_{i+1}) \ldots (y - y_n)}{(y_i - y_0)(y_i - y_1) \ldots (y_i - y_{i-1})(y_i - y_{i+1}) \ldots (y_i - y_n)}, \tag{1}$$

по заданному y определим x. Остаточный член в этом случае получается из остаточного члена формулы Лагранжа, если в последнем поменять местами x и y и заменить производные от прямой функции производными от обратной функции.

Если же заданная функция не монотонна, то изложенный выше прием неприменим. В этом случае, не меняя ролями функцию и аргумент, записываем какую-либо интерполяционную формулу. Затем, считая функцию известной, решаем полученное уравнение относительно аргумента. Однако, если число узлов велико, то придется решать уравнение высокой степени. Остановимся на итерационном способе решения такого уравнения, когда значения аргумента равноотстоящие.

Итак, для функции $y = f(x)$ запишем, например, интерполяционную формулу Ньютона интерполирования вперед:

$$y = y_0 + \frac{\Delta y_0}{1!} q + \frac{\Delta^2 y_0}{2!} q(q-1) + \ldots + \frac{\Delta^n y_0}{n!} q(q-1)\ldots(q-n+1), \tag{2}$$

где $q = \frac{x - x_0}{h}$.

Рассматривая последнее выражение как уравнение относительно q, преобразуем его к виду

$$q = \varphi(q) = \frac{y - y_0}{\Delta y_0} - \frac{\Delta^2 y_0}{2!\Delta y_0} q(q-1) - \ldots - \frac{\Delta^n y_0}{n!\Delta y_0} q(q-1)\ldots(q-n+1). \tag{2'}$$

Будем решать уравнение (2′) итерационно. Примем за начальное приближение для q величину

$$q_0 = \frac{y - y_0}{\Delta y_0}$$

и затем строим итерационный процесс по формуле

$$q_m = \varphi(q_{m-1}), \quad m = 1, 2, \ldots .$$

Во многих случаях при достаточно малых $h = x_{i+1} - x_i$ описанный процесс сходится к искомому корню. Условием сходимости является выполнение неравенства $|\varphi'(q)| \leqslant \alpha < 1$. На практике итерации продолжаются до тех пор, пока два последовательных значения q_m и q_{m+1} не совпадут с заданной точностью, и тогда полагают $q \approx q_{m+1}$. Заметим, что указанный метод итераций требует монотонности функций $f(x)$ только на интервале (x_0, x_1), где $y_0 < y < y_1$ или $y_0 > y > y_1$. Узлы же применяемой интерполяционной формулы могут лежать и вне интервала монотонности функции $f(x)$.

Оценим ошибку второго метода обратного интерполирования, т. е. когда функция $f(x)$ не монотонна.

Пусть $L_n(x)$ — интерполяционный многочлен Лагранжа для $f(x)$, построенный по узлам $x_0, x_1, \ldots, x_n$. Тогда

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)\ldots(x - x_n). \tag{3}$$

Предположим, что нужно найти такое $\bar{x}$, при котором $f(\bar{x}) = \bar{y}$ ($\bar{y}$ задано). Будем решать уравнение $L_n(x) = \bar{y}$. Получим некоторое значение $\bar{\bar{x}}$. Подставляя $\bar{\bar{x}}$ в (3), получим

$$f(\bar{\bar{x}}) - L_n(\bar{\bar{x}}) = f(\bar{\bar{x}}) - \bar{y} = f(\bar{\bar{x}}) - f(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(\bar{\bar{x}}).$$

Применяя теорему Лагранжа о конечных приращениях, будем иметь

$$(\bar{\bar{x}} - \bar{x}) f'(\eta) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(\bar{\bar{x}}),$$

где η находится между $\bar{x}$ и $\bar{\bar{x}}$. Если $[a, b]$ — интервал, содержащий $\bar{x}$ и $\bar{\bar{x}}$, и $\min\limits_{x\in[a,b]} |f'(x)| = m_1 \neq 0$, то из последнего равенства следует, что

$$|\bar{\bar{x}} - \bar{x}| \leqslant \frac{M_{n+1}}{m_1(n+1)!}|\omega_n(\bar{\bar{x}})|$$

(при этом предполагается, что уравнение $L_n(x) = \bar{y}$ решено точно).

2. Численное дифференцирование

Задача численного дифференцирования возникает в том случае, когда функция $f(x)$, для которой нужно найти производную, задана таблично либо имеет очень сложное аналитическое выражение. В таких случаях поступаем следующим образом.

Запишем функцию $f(x)$ в виде

$$f(x) = \varphi(x) + R(x),$$

где $\varphi(x)$ — интерполирующая функция, $R(x)$ — остаточный член интерполяционной формулы. Продифференцируем это тождество k раз (предполагаем, что $f(x)$ и $\varphi(x)$ имеют производные k-го порядка):

$$f^{(k)}(x) = \varphi^{(k)}(x) + R^{(k)}(x)$$

и за приближенное значение $f^{(k)}(x)$ примем $\varphi^{(k)}(x)$. Погрешность при этом есть $R^{(k)}(x)$. При замене функции $f(x)$ интерполирующей функцией $\varphi(x)$ предполагается, что остаточный член мал, однако отсюда отнюдь не следует, что мало и $R^{(k)}(x)$. И на самом деле, при таком способе вычисления $f^{(k)}(x)$ получается сравнительно большая погрешность, особенно при вычислении производных высших порядков. Это можно увидеть на следующем примере.

Рассмотрим две функции $f(x)$, $\tilde{f}(x) \in C[a, b]$, которые связаны соотношением

$$\tilde{f}(x) = f(x) + \frac{1}{n}\sin[n^2(x-a)].$$

Расстояние между ними в пространстве $C[a, b]$ есть

$$\|f - \tilde{f}\|_C = \max_{x\in[a,b]}\left|\frac{1}{n}\sin[n^2(x-a)]\right| \leqslant \frac{1}{n}$$

и его можно сделать как угодно малым за счет выбора n. Однако чебышевское расстояние между производными

$$\|f' - \tilde{f}'\|_C = \max_{x\in[a,b]} |n \cos [n^2 (x - a)]| = n$$

может принимать сколь угодно большие значения.

Этот пример показывает, что задача дифференцирования некорректна в $C[a, b]$. Этим и объясняется малая точность формул численного дифференцирования. Однако при вычислении первых и вторых производных по указанному выше методу во многих случаях можно получить приемлемый результат.

Будем исходить из интерполяционной формулы Ньютона интерполирования вперед для неравных промежутков

$$\begin{aligned} f(x) = f(x_0) + (x - x_0) f(x_0;\ x_1) + (x - x_0)(x - x_1) f(x_0;\ x_1;\ x_2) + \cdots \\ \cdots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f(x_0;\ x_1;\ \ldots;\ x_n) + \\ + (x - x_0)(x - x_1) \ldots (x - x_n) f(x;\ x_0;\ x_1;\ \ldots;\ x_n) = \\ = L_n(x) + R(x). \end{aligned} \tag{4}$$

Продифференцируем обе части (4) k раз, в результате чего получим

$$f^{(k)}(x) = L_n^{(k)}(x) + R^{(k)}(x), \tag{5}$$

где

$$R^{(k)}(x) = \sum_{i=0}^{k} C_k^i f^{(i)}(x;\ x_0;\ \ldots;\ x_n)\, \omega_n^{(k-i)}(x). \tag{6}$$

На основании определения производной, будет иметь место формула

$$\begin{aligned} f^{(i)}(x;\ x_0;\ x_1;\ \ldots;\ x_n) = i! \lim_{\delta\to 0} f(x;\ x+\delta;\ \ldots;\ x + \\ + i\delta;\ x_0;\ x_1;\ \ldots;\ x_n) = \\ = i! \lim_{\delta\to 0} \frac{f^{(n+i+1)}(\xi_i(x, \delta))}{(n+i+1)!} = i! \frac{f^{(n+i+1)}(\xi_i(x))}{(n+i+1)!}, \end{aligned} \tag{7}$$

где $\xi_i(x)$ — некоторая промежуточная точка. При выводе формулы (7) мы воспользовались соотношением (3), § 3, приложение.

После подстановки (7) в (6) получим:

$$R^{(k)}(x) = \sum_{i=0}^{k} \frac{k!}{(k-i)!\,(n+i+1)!} f^{(n+i+1)}(\xi_i)\, \omega_n^{(k-i)}(x), \tag{8}$$

где ξ_i — некоторые точки, заключенные в интервале между наибольшим и наименьшим из чисел $x, x_0, x_1, \ldots, x_n$.

Из (8) следует оценка погрешности формул численного дифференцирования

$$|f^{(k)}(x) - L_n^{(k)}(x)| \leqslant M_{n+k+1} \sum_{i=0}^{k} \frac{k!}{(k-i)!\,(n+i+1)!} |\omega_n^{(k-i)}(x)|, \tag{8'}$$

где $M_{n+k+1} = \max\limits_{0\leqslant i\leqslant n+k+1} \max\limits_{x\in[a,b]} |f^{(i)}(x)|$.

Рассмотрим случай равноотстоящих узлов $x_i - x_{i-1} = h$, $i = 1, 2, \ldots$ $\ldots, n$. Нетрудно видеть, что выражение $\omega_n^{(k-i)}(x)$ относительно шага h есть величина порядка $O(h^{n-k+i+1})$. Следовательно, из (6) будем иметь

$$R^{(k)}(x) = f(x;\ x_0;\ x_1;\ \ldots;\ x_n)\,\omega_n^{(k)}(x) + O(h^{n-k+2}) = O(h^{n-k+1}). \tag{9}$$

Если точка x, в которой мы ищем k-ю производную от функции $f(x)$, такова, что $\omega_n^{(k)}(x) = 0$, то порядок точности формулы численного дифференцирования повышается на единицу.

Рассмотрим наиболее часто применяемые формулы.

Пусть нам требуется вычислить значение k-й производной от функции $f(x)$ в точке $\bar{x} = x_0 + h\left\{\frac{k}{2}\right\}$, где символ $\{\cdot\}$ обозначает дробную часть числа. Построим для $f(x)$ интерполяционный многочлен $L_n(x)$ с узлами $x_{-l+2\left\{\frac{k}{2}\right\}}, \ldots, x_0, x_1, \ldots, x_l$, расположенными симметрично относительно точки $\bar{x}$, где $n = 2l - 2\left\{\frac{k}{2}\right\}$. В силу построения многочлен $\omega_n(x)$ относительно точки $\bar{x}$ будет четным или нечетным в зависимости от того, четно или нечетно k. Следовательно, в любом случае будем иметь

$$\omega_n^{(k)}(\bar{x}) = 0 \tag{10}$$

и неравенство (8′) с учетом (9) и (10) примет вид

$$|f^{(k)}(\bar{x}) - L_n^{(k)}(\bar{x})| = 0(h^{n-k+2}). \tag{11}$$

Положим, например, $k = 2$ и воспользуемся формулой Стирлинга (29), § 1, в результате получим формулу численного дифференцирования

$$f''(\bar{x}) = f''(x_0) \approx h^{-2}\sum_{i=1}^{l} 2(-1)^i \frac{[(i-1)!]^2}{(2i)!}\Delta^{2i} f_{-i}, \tag{12}$$

остаточный член которой согласно (8) и свойствам функции $\omega_n(x)$ имеет вид

$$f''(x_0) - L_{2l+1}(x_0) = \frac{2(-1)^l (l!)^2}{(2l+2)!} f^{(2l+2)}(\xi)\, h^{2l}. \tag{13}$$

При $l = 1, 2$ из (12) получаем наиболее часто применяемые формулы

$$\begin{gathered} f''(x_0) \approx h^{-2}\Delta^2 f_{-1} = \frac{f_1 - 2f_0 + f_{-1}}{h^2}; \\ f''(x_0) \approx h^{-2}\left(\Delta^2 f_{-1} - \frac{1}{12}\Delta^4 f_{-2}\right). \end{gathered} \tag{14}$$

Если нужно найти производную в начальном узле таблицы, то можно воспользоваться формулой Ньютона для интерполирования вперед (24′), § 1, дифференцируя которую будем иметь:

$$f'(x_0) \approx h^{-1}\sum_{i=0}^{n-1} \frac{(-1)^i}{(i+1)}\Delta^{i+1} f_0 = L_n'(x_0). \tag{15}$$

Формулы вида (15) носят название односторонних формул численного дифференцирования.

Для остаточного члена формулы (15) справедливо соотношение

$$f'(x_0) - L_n'(x_0) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_n'(x_0) = \frac{(-1)^n}{n+1} f^{(n+1)}(\xi)\, h^n. \qquad (16)$$

В заключение рассмотрим влияние некорректности дифференцирования при вычислениях приближенных значений производных с помощью формул численного дифференцирования.

Пусть табличные значения функции $f(x)$ известны не точно, а с некоторой погрешностью. Причем об этой погрешности известно только, что абсолютная величина ее не превосходит ε.

Пусть необходимо определить $f'(x_0)$. Воспользовавшись формулой (15), при $n = 1$ будем иметь

$$f'(x_0) \approx \frac{f_1 - f_0}{h}, \qquad (17)$$

где погрешность (17) согласно (16) оценивается с помощью неравенства

$$\left| f'(x_0) - \frac{f_1 - f_0}{h} \right| \leqslant \frac{M_2 h}{2}, \quad M_2 = \max_{x \in [a,b]} |f''(x)|. \qquad (18)$$

Но вместо правой части (17) вычисляем фактически другую величину $\frac{\tilde{f}_1 - \tilde{f}_0}{h}$ и полагаем

$$f'(x_0) \approx \frac{\tilde{f}_1 - \tilde{f}_0}{h}, \qquad (17')$$

где $\tilde{f}_i$ — табличные значения $f(x)$ и $|f_i - \tilde{f}_i| < \varepsilon$, $i = 1, 2$. Погрешность формулы (17′) уже вместо (18) оценивается другим неравенством

$$\left| f'(x_0) - \frac{\tilde{f}_1 - \tilde{f}_0}{h} \right| \leqslant \frac{M_2 h}{2} + \frac{2\varepsilon}{h}, \qquad (19)$$

из которого следует, что с уменьшением h формула (17′) может давать сколь угодно плохое приближение к искомой производной $f'(x_0)$ (наименьшее значение $2\sqrt{M_2\varepsilon}$ правой части (19) достигается при $h = 2\sqrt{\varepsilon/M_2}$).

Если же вместо формулы (17) применять формулу численного дифференцирования более высокого порядка точности, то положение несколько исправляется. Многоточечные формулы численного дифференцирования являются своего рода «регуляризатором» и дают значительно лучшие результаты.

Глава 3

ПРИБЛИЖЕНИЕ ФУНКЦИЙ

В данной главе рассматриваются вопросы, связанные с приближением функций в нормированных пространствах. Общая постановка задачи построения для заданного элемента f из линейного нормированного пространства $\mathbf{R}$ элемента наилучшего приближения $\Phi_0 \in M_n$

(M_n — линейная оболочка линейно-независимых элементов $\{\varphi_i\}_{i=0}^n$) была рассмотрена в § 1, гл. 1. Там же была доказана теорема о существовании хотя бы одного элемента наилучшего приближения (см. теорему 2, § 1, гл. 1).

Ограничимся здесь рассмотрением случая, когда $\mathbf{R} = L_{p,\alpha}[a, b]$, т. е. когда $\mathbf{R}$ совпадает с классом функций, суммируемых с p-й степенью:

$$f \in L_{p,\alpha}[a, b], \text{ если } \|f\|_{L_{p,\alpha}} = \{\int_a^b |f(x)|^p d\alpha(x)\}^{\frac{1}{p}} < \infty,$$

где $\alpha(x)$ — неубывающая функция, $p \geqslant 1$ и интеграл понимается как интеграл Лебега — Стильтьеса. Причем, основное внимание будет уделено случаям $p = 2$ и $p = \infty$, $f \in C[a, b]$.

Если $p = 2k$, $k = 1, 2, \ldots$, то элемент наилучшего приближения будет обладать одним общим свойством, которое сформулируем в виде следующего утверждения:

Теорема А. *Пусть функции $\varphi_i(x) \in C[a, b]$ $(i = 0, 1, \ldots, n)$ образуют систему Чебышева на $[a, b]$ и линейная оболочка их обозначена через M_n. Пусть $\Phi_0(x) \in M_n$ — наилучшее приближение к $f(x) \in L_{2k,\alpha}[a, b]$ в M_n, тогда существует $(n+2)$ точки x_i $(i = 1, 2, \ldots, n+2)$ такие, что*

$$a \leqslant x_1 < x_2 < \ldots < x_{n+1} < x_{n+2} \leqslant b;$$
$$\operatorname{sign} \delta(x_i) = -\operatorname{sign} \delta(x_{i+1}), \quad i = 1, 2, \ldots, n+1,$$

где $\delta(x) = f(x) - \Phi_0(x)$.

Доказательство. Пусть Φ_0 — обобщенный многочлен наилучшего приближения и таких точек, о которых идет речь в теореме $(k < n+2)$. Тогда в каждом из промежутков $[x_i, x_{i+1}]$ $(i = 1, 2, \ldots, k-1)$ найдется точка ξ_i, при переходе через которую функция $\delta(x)$ меняет знак. Построим обобщенный многочлен, удовлетворяющий условиям

$$\Phi_1(\xi_i) = 0, \quad i = 1, 2, \ldots k, \quad \|\Phi_1\|_{L_{2k,\alpha}} = 1,$$

что всегда возможно, ибо множество M_n удовлетворяет условию Хаара. Тогда произведение $\delta(x)\Phi_1(x) = [f(x) - \Phi_0(x)]\Phi_1(x)$ будет сохранять постоянный знак на $[a, b]$, и, не уменьшая общности, можно считать его положительным.

Введем обобщенный многочлен

$$\Phi(x) = \Phi_0(x) + \lambda\Phi_1(x)$$

и рассмотрим выражение

$$\|f - \Phi\|^{2k}_{L_{2k,\alpha}} = \int_a^b [f(x) - \Phi_0(x) - \lambda\Phi_1(x)]^{2k} d\alpha(x) = \|f -$$

$$- \Phi_0\|^{2k}_{L_{2k,\alpha}} - \sum_{r=1}^{2k-1} (-1)^{r-1} C_{2k}^r \lambda^r \int_a^b [f(x) - \Phi_0(x)]^{2k-r} [\Phi_1(x)]^r d\alpha(x) + \lambda^{2k}.$$

Выберем λ как положительный корень уравнения

$$P_{2k-1}(\lambda) = \lambda^{2k-1} - \sum_{r=1}^{2k-1} (-1)^{r-1} [C_{2k}^{r} - kC_{2k-2}^{r-1}] b_r \lambda^{2k-1-r} = 0,$$

где $b_r = \int_a^b [f(x) - \Phi_0(x)]^{2k-r} [\Phi_1(x)]^r d\alpha(x) > 0$, что всегда можно сделать, ибо многочлен $P_{2k-1}(\lambda)$ имеет нечетную степень и

$$P_{2k-1}(0) = -b_{2k-1}[C_{2k}^{2k-1} - kC_{2k-2}^{2k-2}] = -b_{2k-1}k < 0, \quad P_{2k-1}(+\infty) = +\infty.$$

Обозначим этот корень через λ_1, тогда будем иметь

$$\|f - \Phi\|_{L_{2k,\alpha}}^{2k} = \|f - \Phi_0\|_{L_{2k,\alpha}}^{2k} - \lambda_1 k \int_a^b [f(x) - \Phi_0(x)] \Phi_1(x) [f(x) - \Phi_0(x) - \lambda_1 \Phi_1(x)]^{2k-2} d\alpha(x) < \|f - \Phi_0\|_{L_{2k,\alpha}}^{2k},$$

т. е. обобщенный многочлен Φ_0 не является наилучшим приближением к f в M_n. Противоречие доказывает теорему.

§ 1. СРЕДНЕКВАДРАТИЧЕСКИЕ ПРИБЛИЖЕНИЯ

В данном параграфе рассмотрены вопросы приближения функций $f(x)$, принадлежащих банаховому пространству **B**, функциями $\varphi(x) \in M_n$, где M_n — конечномерное подпространство пространства **B**. Причем **B** = **H** — гильбертово пространство со скалярным произведением (u, v), норма и расстояние для которого определяются формулами:

$$\|u\| = (u, u)^{\frac{1}{2}}; \quad \Delta(u, v) = \|u - v\|. \tag{1}$$

Основное внимание будет уделено тому случаю, когда под скалярным произведением (u, v) понимается выражение

$$(u, v) = \int_a^b u(x) v(x) d\alpha(x), \tag{2}$$

где $\alpha(x)$ — неубывающая функция, и интеграл понимается в смысле Лебега — Стильтьеса. Тогда расстояние между функциями в терминах этого скалярного произведения есть среднеквадратическое уклонение

$$(u - v, u - v)^{\frac{1}{2}} = \Delta(u, v) = \left\{\int_a^b [u(x) - v(x)]^2 d\alpha(x)\right\}^{\frac{1}{2}}. \tag{3}$$

Если $\alpha(x)$ является функцией скачков, т. е. если она кусочно-постоянна и имеет скачки величины ρ_i в точках $x_i \in [a, b]$, то скалярное произведение (2) сводится к сумме

$$(u, v) = \sum_i \rho_i u(x_i) v(x_i), \tag{2'}$$

которая задает скалярное произведение функций дискретного аргумента. Расстояние (3) для этого случая принимает вид

$$(u-v,\ u-v)^{\frac{1}{2}} = \Delta(u,\ v) = \{\sum_i \rho_i [u(x_i) - v(x_i)]^2\}^{\frac{1}{2}}. \qquad (3')$$

Построение среднеквадратических приближений функций оправдано в силу следующих причин. Часто приближаемая функция $f(x)$ известна неточно, ее значения $\tilde{f}(x)$ получены из эксперимента и содержат случайные ошибки, поэтому нет смысла требовать равномерной близости $\varphi(x)$ к $\tilde{f}(x)$. Целесообразно добиваться «интегральной» близости. Как показывает практика, приближающие функции, построенные по методу среднеквадратического приближения, значительно лучше представляют реальную функцию $f(x)$, чем интерполяционные многочлены. Кроме того, среднеквадратические приближения позволяют расширить класс **B** приближаемых функций. При рассмотрении равномерного приближения ограничиваются классом $C(S)$ непрерывных функций на компакте S и это требование существенно, если ставить задачу равномерного приближения функций $f(x)$ многочленами. В случае же среднеквадратического приближения нужно лишь требовать существования $\int_a^b [f(x)]^2\, d\alpha(x)$, т. е. можно приближать функции из класса $L_{2,\alpha}[a,\ b]$.

Приведем формулировку задачи среднеквадратического приближения в терминах общей задачи приближения линейных операторов (см. § 1, гл. 1). Здесь $F(f) = f(x)$, $\forall f \in \mathbf{B}$, $F_n : \mathbf{B} \to M_n$ и $F_n(f) = \sum_{k=0}^{n} C_k \varphi_k(x)$, где $\{\varphi_k(x)\}_{k=0}^{n}$ — система линейно-независимых функций из **B**, замыкание линейной оболочки (над полем комплексных чисел) которых образует M_n. В качестве банахова пространства **B** взято гильбертово пространство $L_{2,\alpha}[a,\ b]$. Необходимо по заданной функции f и заданному $\varepsilon > \theta$ найти такой оператор F_n, что

$$\|F(f) - F_n(f)\|_{L_{2,\alpha}} = \|f - \sum_{k=0}^{n} C_k \varphi_k\|_{L_{2,\alpha}} = \Delta(f,\ \sum_{k=0}^{n} C_k \varphi_k) \leqslant \varepsilon, \qquad (4)$$

где $\Delta(u,\ v)$ определяется формулой (3) либо (3′).

Понятно, что самое лучшее, что можно сделать, оперируя с подпространством M_n, это найти элемент наилучшего приближения (см. определение 3, гл. 1). Существование и единственность такого элемента гарантируется теоремой 3, гл. 1.

Любой элемент $f \in \mathbf{H} = L_{2,\alpha}[a,\ b]$ можно единственным образом представить в виде

$$f = \varphi + v, \qquad (5)$$

где $\varphi \in M_n$ и является проекцией f на M_n и $v \perp M_n$ (приложение, § 1). Тогда будем иметь

$$(v,\ u) = (f - \varphi,\ u) = 0,\ \forall u \in M_n \qquad (6)$$

и, следовательно, φ — элемент наилучшего приближения (см. § 1, гл. 1), т. е. элемент наилучшего приближения для f в M_n является проекцией этого элемента на M_n.

1. Построение элемента наилучшего приближения

Поскольку элемент наилучшего приближения $\varphi \in M_n$, то имеет место представление

$$\varphi(x) = \sum_{i=0}^{n} C_i \varphi_i(x) \tag{7}$$

и задача построения элемента наилучшего приближения сводится к определению комплексных коэффициентов C_i $(i = 0, 1, \ldots, n)$. На основании (6) получаем следующую систему:

$$(f - \varphi, \varphi_k) = 0, \quad k = 0, 1, \ldots, n \tag{8}$$

или с учетом (7)

$$\sum_{i=0}^{n} C_i (\varphi_i, \varphi_k) = (f, \varphi_k), \quad k = 0, 1, \ldots, n. \tag{8'}$$

Определитель этой системы есть определитель Грамма

$$G(\varphi_0, \varphi_1, \ldots, \varphi_n) = |(\varphi_i, \varphi_k)|_{i=\overline{0,n}}^{k=\overline{0,n}}, \tag{9}$$

построенный по системе линейно-независимых функций. Следовательно, $G(\varphi_0, \varphi_1, \ldots, \varphi_n) \neq 0$ и система (8′) имеет единственное решение C_i $(i = 0, 1, \ldots, n)$, найдя которое, можно сразу определить $\varphi(x)$.

Найдем теперь отклонение элемента f от φ, т. е. величину

$$\Delta^2(f) = \|f - \varphi\|^2 = (f - \varphi, f - \varphi) = (f - \varphi, f) - (f - \varphi, \varphi) =$$
$$= (f - \varphi, f) = \|f\|^2 - (\varphi, f).$$

Из последнего соотношения вместе с (7) получаем

$$\Delta^2(f) = \|f\|^2 - \sum_{i=0}^{n} C_i (\varphi_i, f). \tag{10}$$

Рассмотрим следующую систему уравнений:

$$\sum_{i=0}^{n} d_i (\varphi_i, \varphi_k) + d_{n+1} (f, \varphi_k) = 0, \quad k = 0, 1, \ldots, n;$$

$$\sum_{i=0}^{n} d_i (\varphi_i, f) - d_{n+1} [\Delta^2(f) - \|f\|^2] = 0. \tag{11}$$

Сравнивая (11) с (8′) и (10), легко заметить, что однородная система (11) относительно d_i $(i = 0, 1, \ldots, n + 1)$ имеет нетривиальное решение

$$d_i = C_i, \quad i = 0, 1, \ldots, n; \quad d_{n+1} = -1. \tag{12}$$

Следовательно, определитель ее

$$\begin{vmatrix} (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \dots & (\varphi_n, \varphi_0) & (f, \varphi_0) \\ \dots & \dots & \dots & \dots & \dots \\ (\varphi_0, \varphi_n) & (\varphi_1, \varphi_n) & \dots & (\varphi_n, \varphi_n) & (f, \varphi_n) \\ (\varphi_0, f) & (\varphi_1, f) & \dots & (\varphi_n, f) & (f, f) - \Delta^2(f) \end{vmatrix} = 0,$$

отсюда получаем

$$\Delta^2(f) = \frac{G(\varphi_0, \varphi_1, \dots, \varphi_n, f)}{G(\varphi_0, \varphi_1, \dots, \varphi_n)}, \tag{13}$$

где использовано обозначение (9).

Поскольку соотношение (13) имеет место $\forall f \in \mathbf{H}$, индукцией устанавливается справедливость следующей леммы:

Лемма 1. *Определитель Грамма* $G(\varphi_0\ \varphi_1, \dots, \varphi_n)$, *построенный по любой системе линейно-независимых функций* $\varphi_i \in \mathbf{H}$ $(i = 0, 1, \dots, n)$, *— положительный.*

Если $\{\varphi_i\}_{i=0}^n$ — ортонормальная система в $\mathbf{H}$, т. е.

$$(\varphi_i, \varphi_j) = \delta_{i,j}, \quad i, j = 0, 1, \dots, n, \tag{14}$$

то система (8′) становится системой с диагональной матрицей и имеет следующее решение:

$$C_i = (f, \varphi_i), \quad i = 0, 1, \dots, n. \tag{14′}$$

Формулы (14′) показывают, что в этом случае C_i являются коэффициентами Фурье элемента f по ортонормальной системе $\{\varphi_i\}_{i=0}^n$, элемент наилучшего приближения φ является отрезком ряда Фурье и формула (10) принимает вид

$$\Delta^2(f) = \|f\|^2 - \sum_{i=0}^n C_i \overline{(f, \varphi_i)} = \|f\|^2 - \sum_{i=0}^n |C_i|^2, \tag{15}$$

т. е.

$$\Delta(f) = \{\|f\|^2 - \sum_{i=0}^n |C_i|^2\}^{\frac{1}{2}}.$$

Из (15) вытекает *неравенство Бесселя*

$$\sum_{i=0}^\infty |C_i|^2 \leqslant \|f\|^2, \tag{16}$$

следствием которого является сходимость ряда $\sum\limits_{n=0}^\infty |C_i|^2$.

Заметим, что с помощью процесса ортогонализации Шмидта любую линейно-независимую систему элементов $\{\varphi_i\}$ в $\mathbf{H}$ можно привести к ортонормированной системе.

Обозначим через $\varphi^{(n)}$ элемент наилучшего приближения f в подпространстве M_n^-. Возникает вопрос: будет ли $\varphi^{(n)} \to f$ при $n \to \infty$. Ответ дает следующая теорема:

Теорема 1. *В гильбертовом пространстве* $\mathbf{H}$ *последовательность наилучших приближений* $\varphi^{(n)}$, *построенная по полной ортонормальной*

системе $\{\varphi_i\}_{i=0}^{\infty}$ *(отрезок ряда Фурье), сходится к этому элементу.*

Доказательство. Докажем сначала, что последовательность $\varphi^{(n)}$ $(n = 0, 1, \ldots)$ — фундаментальная. Действительно, в силу ортонормальности системы $\{\varphi_i\}$ будем иметь

$$\|\varphi^{(n)} - \varphi^{(m)}\|^2 = \|\sum_{i=m+1}^{n} C_i\varphi_i\|^2 = \sum_{i=m+1}^{n} |C_i|^2.$$

Ввиду сходимости ряда $\sum_{i=0}^{\infty} |C_i|^2$ правая часть последнего выражения может быть сделана сколь угодно малой за счет выбора достаточно больших m и n. Обозначим через $\tilde{f}$ выражение $\lim_{n\to\infty} \varphi^{(n)}$, которое существует вследствие полноты пространства **H**, и рассмотрим скалярное произведение

$$(f - \tilde{f}, \varphi_k) = (f - \sum_{i=0}^{\infty} C_i\varphi_i, \varphi_k) =$$

$$= (f - \sum_{i=0}^{n} C_i\varphi_i, \varphi_k) - (\sum_{i=n+1}^{\infty} C_i\varphi_i, \varphi_k) = 0, \quad k = 0, 1, \ldots, n \tag{17}$$

(равенство нулю следует из (8) и (14)). Из полноты ортонормальной системы $\{\varphi_i\}_{i=0}^{\infty}$ получаем, что (17) может иметь место только тогда, когда

$$f = \tilde{f},$$

что и требовалось доказать.

При выполнении условий теоремы 1 будем иметь

$$\|f - \varphi^{(n)}\|^2 = \|f\|^2 - \sum_{i=0}^{n} |C_i|^2,$$

откуда после предельного перехода в обеих частях по $n \to \infty$, получим так называемое *равенство Парсеваля*

$$\sum_{i=0}^{\infty} |C_i|^2 = \|f\|^2. \tag{18}$$

2. Среднеквадратические приближения функций алгебраическими многочленами

Пусть $\mathbf{H} = L_{2,\alpha}[a, b]$ — вещественное гильбертово пространство со скалярным произведением (2), в котором функции, отличающиеся друг от друга на множестве меры нуль, считаются равными. Сходимость в этом пространстве есть сходимость в среднем относительно распределения $d\alpha(x)$. В качестве M_n возьмем линейную оболочку многочленов с вещественными коэффициентами степени не выше n, т. е. $M_n = \bigcup_{i=0}^{n} \pi_i$. Для многочлена наилучшего приближения $P_n(x)$ функции $f(x)$ имеет место следующая теорема (частный случай теоремы A):

Теорема 2. *Пусть $P_n(x)$ — многочлен наилучшего приближения к $f(x) \in L_{2,\alpha}[a, b]$, тогда $\exists x_i \in [a, b]$ (не равные между собой), в которых*

$$\operatorname{sign}[f(x_i) - P_n(x_i)] = (-1)\operatorname{sign}[f(x_{i+1}) - P_n(x_{i+1})],$$
$$i = 0, 1, \ldots, n+1. \tag{19}$$

При определении коэффициентов многочлена $P_n(x)$ наилучшего приближения в смысле среднеквадратического, как следует из пункта 1, целесообразно в качестве системы линейно-независимых функций $\{\varphi_i\}_{i=0}^{\infty} \in L_{2,\alpha}[a, b]$ взять последовательность многочленов $\{p_i(x)\}_{i=0}^{\infty}$, образующих ортонормальную систему в $L_{2,\alpha}[a, b]$:

$$(p_i, p_j) = \int_a^b p_i(x)\, p_j(x)\, d\alpha(x) = \delta_{ij}. \tag{20}$$

Тогда многочлен наилучшего приближения $P_n(x)$ запишется в виде

$$P_n(x) = \sum_{i=0}^{n} C_i p_i(x), \tag{21}$$

а коэффициенты C_i будут согласно (14′) определяться формулой

$$C_i = (f, p_i) = \int_a^b f(x)\, p_i(x)\, d\alpha(x). \tag{22}$$

Имеет место следующее обобщение теоремы Вейерштрасса:

Теорема 3. *Пусть $f(x) \in L_{2,\alpha}[a, b]$, интервал $[a, b]$ — конечный. Тогда $\forall \varepsilon > 0$ существует такой многочлен $P_n(x)$, что*

$$\|f - P_n\| = \left\{\int_a^b [f(x) - P_n(x)]^2\, d\alpha(x)\right\}^{\frac{1}{2}} < \varepsilon,$$

т. е. система ортонормальных многочленов $\{p_i(x)\}_{i=0}^{\infty}$ является замкнутой, а следовательно, и полной в $L_{2,\alpha}[a, b]$.

Заметим, что ограничение на конечность интервала $[a, b]$ является существенным.

Доказательство. В силу свойств интеграла Лебега — Стильтьеса каждую функцию $f(x) \in L_{2,\alpha}[a, b]$ можно сколь угодно точно приблизить в смысле метрики пространства $L_{2,\alpha}[a, b]$ последовательностью непрерывных функций $f_n(x)$. В силу теоремы Вейерштрасса для всех $f_n(x)$ и всех $\varepsilon_n > 0$ существует $P_n(x)$ — многочлен с рациональными коэффициентами и такой, что $\|f_n - P_n\|_C \leqslant \varepsilon_n$. Это значит, что

$$\|f - P_n\|_{L_{2,\alpha}[a,b]} \leqslant \|f - f_n\|_{L_{2,\alpha}[a,b]} + \|f_n - P_n\|_{L_{2,\alpha}[a,b]} \leqslant$$
$$\leqslant \|f - f_n\|_{L_{2,\alpha}[a,b]} + K\|f_n - P_n\|_C \to 0,$$

где $K = \left[\int_a^b d\alpha(x)\right]^{\frac{1}{2}}$, т. е. множество многочленов с рациональными коэффициентами плотно в $L_{2,\alpha}[a, b]$. Отсюда и следует утверждение теоремы.

В отношении неограниченного интервала $[a, b]$ справедлива следующая теорема:

Теорема 4. *Система* $\{e^{-\frac{x}{2}} x^{\frac{\alpha}{2}} L_n^{\alpha}(x)\}_{n=0}^{\infty}$ — *полная в* $L_{2,x}(0, \infty)$; *система* $\{e^{-\frac{x^2}{2}} H_n(x)\}_{n=0}^{\infty}$ — *полная в* $L_{2,x}(-\infty, \infty)$, *где* $L_n^{\alpha}(x)$ — *многочлены Лагерра,* $H_n(x)$ — *многочлены Эрмита* (приложение, § 2).

Доказательство теоремы опускаем.

Объединение теорем 1, 3 и 4 приводит к следующему результату: $\forall f(x) \in L_{2,\alpha}[a, b]$, где распределение $d\alpha(x)$ и $[a, b]$ соответствуют классическим ортогональным многочленам; ряд Фурье, построенный по нормированной системе классических ортогональных многочленов, сходится в $L_{2,\alpha}[a, b]$ к $f(x)$.

Для случая, когда $\alpha(x)$ — функция скачков и скалярное произведение в гильбертовом пространстве $L_{2,\alpha}[a, b]$ задается формулой (2′), в качестве линейно-независимой системы функций $\{\varphi_k\}$ можно брать ортонормальную систему многочленов дискретного аргумента (приложение, § 2). Рассмотрим следующий пример:

Пример 1. Пусть в скалярном произведении (2′) $\rho_i \equiv 1$ и $i = 0, 1, \ldots, N-1$, т. е.

$$(u, v) = \sum_{i=0}^{N-1} u(x_i)\, v(x_i), \tag{23}$$

и необходимо построить многочлен наилучшего приближения степени $m < N-1$ для функции $f(x)$, значения которой $f(i)$ $(i = 0, 1, \ldots, N-1)$ — известны.

Этот случай довольно часто встречается на практике при уравновешивании результатов наблюдений. В качестве линейно-независимой системы функций $\{\varphi_k\}$ возьмем нормированные многочлены Чебышева дискретного аргумента

$$\bar{t}_{k,N}(x) = k!\, \Delta^k \left[\binom{x}{k}\binom{x-N}{k}\right] / N_k = t_k(x)/N_k, \tag{24}$$

$$N_k = \left[\frac{N(N^2-1^2)(N^2-2^2)\ldots(N^2-k^2)}{(2k+1)}\right]^{\frac{1}{2}}, \quad k = 0, 1, \ldots, N-1, \tag{25}$$

которые образуют ортонормальную систему в смысле скалярного произведения (23).

В качестве подпространства $M_m \subset L_{2,\alpha}[0, N-1]$ $(m < N-1)$ вводится линейная оболочка многочленов m-й степени. Тогда согласно формуле (22) для коэффициентов элемента наилучшего приближения

$$P_m(x) = \sum_{p=0}^{m} C_p \bar{t}_{p,N-1}(x), \tag{26}$$

построенного для функции $f(x) \in L_{2,\alpha}[0, N-1]$, получаем формулу

$$C_p = \sum_{i=0}^{N-1} f(i)\, \bar{t}_{p,N-1}(i), \quad p = 0, 1, \ldots, m. \tag{27}$$

Если значения функции $f(x)$ известны в равноотстоящих узлах $x_i \in [a, b]$, $i = 0, 1, \ldots, N-1$, $x_i - x_{i-1} = h$, $i = 1, 2, \ldots, N-1$ и $x_i \neq i$, то линейной заменой независимого переменного

$$t = \frac{x - x_0}{h}$$

всегда можно прийти к рассмотренному выше случаю.

3. Среднеквадратические приближения функций тригонометрическими многочленами

Пусть приближаемая функция $f(x) \in L_{2,x}[0, 2\pi]$ и в качестве системы линейно-независимых функций $\{\varphi_i\}$ выбрана тригонометрическая система

$$\frac{1}{\sqrt{2\pi}}, \quad \frac{1}{\sqrt{\pi}}\cos x, \quad \frac{1}{\sqrt{\pi}}\sin x, \quad \frac{1}{\sqrt{\pi}}\cos 2x, \quad \frac{1}{\sqrt{\pi}}\sin 2x, \ \ldots, \tag{28}$$

которая, как известно, является полной ортонормальной системой в $L_{2,x}[0, 2\pi]$. Полнота системы (28) вытекает из второй теоремы Вейерштрасса о плотности системы тригонометрических многочленов вида

$$T_n(x) = \frac{a_0}{2} + \sum_{i=0}^{n} (a_i \cos ix + b_i \sin ix) \tag{29}$$

в пространстве $C[0, 2\pi]$ и неравенства

$$\|f - T_n\|_{L_{2,x}} \leqslant (2\pi)^{\frac{1}{2}} \|f - T_n\|_C.$$

Согласно результатам п.1, тригонометрический многочлен (29) будет элементом наилучшего приближения для $f(x)$ в том случае, если его коэффициенты являются коэффициентами Фурье функции $f(x)$, т. е. определяются формулами:

$$\alpha_i = \frac{1}{\pi 2^{\delta_{0i}}} \int_0^{2\pi} f(x) \cos ix dx; \quad \beta_i = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin ix dx, \tag{30}$$

$$i = 0, 1, 2, \ \ldots, \ n.$$

Функцию $f(x)$ также можно приближать в смысле среднего квадратического тригонометрическими многочленами, построенными по системам функций

$$\frac{1}{\sqrt{\pi}}, \quad \sqrt{\frac{2}{\pi}}\cos x, \quad \sqrt{\frac{2}{\pi}}\cos 2x, \ \ldots; \tag{31}$$

$$\sqrt{\frac{2}{\pi}}\sin x, \quad \sqrt{\frac{2}{\pi}}\sin 2x, \quad \sqrt{\frac{2}{\pi}}\sin 3x, \ \ldots, \tag{32}$$

являющимися полными ортонормированными системами в $L_{2,x}[0, \pi]$.

Если $\alpha(x)$ является функцией скачков и величины скачков ρ_i в точках $x_i \in (0, 2\pi]$ $(i = 1, 2, \ldots, N)$ все равны 1, то, естественно, возникает вопрос: при каком расположении узлов x_i система тригонометрических функций

$$1, \ \cos x, \ \sin x, \ \ldots, \ \cos nx, \ \sin nx \quad (N > 2n + 1) \tag{28'}$$

будет ортогональной в $L_{2,\alpha}[0, 2\pi]$?

Легко убедиться, что ортогональность достигается в случае, когда

$$x_i = i\frac{2\pi}{N}, \quad i = 1, 2, \ \ldots, \ N, \tag{33}$$

т. е. являются равноотстоящими с шагом $h = \frac{2\pi}{N}$.

Многочлен $T_n(x)$ будет многочленом наилучшего приближения для $f(x)$ в $L_{2,\alpha}[0, 2\pi]$, если коэффициенты его определяются формулами:

$$a_i = \frac{2}{N}{}^{1-\delta_{i0}} \sum_{k=1}^{N} f(x_k) \cos ix_k, \quad b_i = \frac{2}{N} \sum_{k=1}^{N} f(x_k) \sin ix_k, \tag{34}$$

$$i = 0, 1, \ldots, n,$$

где x_k определяется из (33). Формулы (34) носят название *формул Бесселя.*

4. Применение метода наименьших квадратов в смежных вопросах

Сглаживание результатов наблюдения. Пусть в результате наблюдений для значений аргумента $x_0, x_1, \ldots, x_k$ получена таблица значений функции $f(x)$. На практике обычно $x_0, x_1, \ldots, x_k$ находятся точно или во всяком случае значительно точнее, чем $f(x_0), \ldots, f(x_k)$. Будем предполагать, что систематические ошибки, а также грубые ошибки при определении $f(x_i)$ исключены. С целью уменьшения случайных ошибок и получения более плавной функции $f(x)$ применяют процесс сглаживания, состоящий в том, что полученные в результате наблюдений значения $f(x_i)$ заменяют значениями $\bar{f}(x_i)$, которые дает выбранный нами способ сглаживания.

Опишем способ сглаживания, основанный на методе наименьших квадратов. Предположим, что $x_i - x_{i-1} = h$, $i = 1, 2, \ldots, k$, все измеренные значения $f(x_i)$ имеют одинаковую точность и функция $f(x)$ на каждом участке, охватывающем N узлов, может быть достаточно хорошо приближена многочленом m-й степени, $m \leqslant N - 1$. Чтобы найти сглаженное значение $f(x_i)$ в точке x_i, выбирают N четным и узлы $x_{i-\frac{N}{2}+p}$ $(p = 0, 1, \ldots, N)$, для которых x_i является средним узлом. По имеющимся из наблюдений значениям $f(x_{i-\frac{N}{2}+p})$, $p = 0, 1, \ldots, N$ с помощью многочленов Чебышева дискретного аргумента строят многочлен наилучшего приближения $P_m(x)$, коэффициенты которого определяются формулой:

$$C_p = \sum_{r=0}^{N-1} f(x_{i-\frac{N}{2}+r})\, \bar{t}_{p,N-1}(r), \quad p = 0, 1, \ldots, m.$$

После чего полагают

$$\bar{f}(x_i) = P_m\left(\frac{N-1}{2}\right).$$

Для практического использования можно заранее найти при заданных m и N выражения $\bar{f}(x_i)$ через $f(x_{i-\frac{N}{2}+p})$, $p = 0, 1, \ldots, N$,

т. е. коэффициенты разложения

$$\bar{f}(x_i) = P_m\left(\frac{N-1}{2}\right) = \sum_{p=0}^{m}\left(\sum_{r=0}^{N-1} f(x_{i-\frac{N}{2}+r})\,\bar{t}_{p,N-1}(r)\right)\bar{t}_{p,N-1}\left(\frac{N-1}{2}\right) =$$

$$= \sum_{r=0}^{N-1}\left[\sum_{p=0}^{m}\bar{t}_{p,N-1}(r)\,\bar{t}_{p,N-1}\left(\frac{N-1}{2}\right)\right] f(x_{i-\frac{N}{2}+r}).$$

Приведем несколько таких выражений, где для кратности введем обозначения $f(x_i) = f_i$:

$$m = 1$$

$$N = 2 \quad \bar{f}(x_i) = \frac{1}{3}[f_{i-1} + f_i + f_{i+1}];$$

$$N = 4 \quad \bar{f}(x_i) = \frac{1}{5}[f_{i-2} + f_{i-1} + f_i + f_{i+1} + f_{i+2}];$$

$$m = 3$$

$$N = 4 \quad \bar{f}(x_i) = \frac{1}{35}[-3f_{i-2} + 12f_{i-1} + 17f_i + 12_{i+1} - 3f_{i+2}];$$

$$N = 6 \quad \bar{f}(x_i) = \frac{1}{21}[-2f_{i-3} + 3f_{i-2} + 6f_{i-1} + 7f_i + 6f_{i+1} + 3f_{i+2} - 2f_{i+3}];$$

$$m = 5$$

$$N = 6 \quad \bar{f}(x_i) = \frac{1}{231}[5f_{i-3} - 30f_{i-2} + 75f_{i-1} + 131f_i + 75f_{i+1} - 30f_{i+2} + 5f_{i+3}];$$

$$N = 8 \quad \bar{f}(x_i) = \frac{1}{429}[15f_{i-4} - 55f_{i-3} + 30f_{i-2} + 135f_{i-1} + 179f_i + 135f_{i+1} + 30f_{i+2} - 55f_{i+3} + 15f_{i+4}].$$

Иногда сглаживание производят несколько раз, однако при этом следует отметить, что многократное сглаживание может сильно исказить истину.

Построение эмпирических формул и решение систем нелинейных алгебраических уравнений. Пусть в результате измерений функции $y(x)$ при $x = x_1$, $x = x_2$, ..., $x = x_n \in [a, b]$ получена таблица величин y_i и мы хотим на основании этих данных построить аналитическую функцию

$$\bar{y}(x) = f(x, a_1, a_2, \ldots, a_m), \tag{35}$$

зависящую от m ($m < n$) параметров a_i, $i = 1, 2, \ldots, m$, достаточно простого вида и которая достаточно хорошо приближала бы функцию $y(x)$ на всем промежутке $[a, b]$.

Вид функции f и число параметров в некоторых случаях известны из каких-либо дополнительных соображений, в других случаях эти функции определяют из графика, построенного по известным значениям

$y(x_i)$, а число параметров подбирают так, чтобы эмпирическая формула наилучшим образом отображала результаты наблюдений и была достаточно проста. Поставим задачу: по известным $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ найти значения параметров $a_1, a_2, \ldots, a_m$ $(n > m)$. Если бы значения y_i и x_i были известны точно, то задача принципиально решалась бы просто: нужно было бы из системы

$$\begin{aligned} y_1 &= f(x_1, a_1, \ldots, a_m); \\ y_2 &= f(x_2, a_1, \ldots, a_m); \\ &\ldots\ldots\ldots\ldots \\ y_n &= f(x_n, a_1, \ldots, a_m) \end{aligned} \tag{36}$$

взять m уравнений и найти из них значения a_i (предполагается, что эта система имеет единственное решение). При этих значениях параметров остальные уравнения тоже удовлетворялись бы. Если же y_i являются приближенными (что и бывает на практике), то, найдя из каких-либо m уравнений значения a_i $(i = 1, 2, \ldots, m)$, убеждаемся, что эти значения не будут удовлетворять остальным уравнениям, т. е. система (36) чаще всего будет несовместной. Поэтому постараемся так определить значения параметров, чтобы в некотором смысле все уравнения системы были удовлетворены с наименьшей погрешностью.

Параметры $a_1, \ldots, a_m$ можно определять из условия минимума функции

$$S_\alpha(a_1, \ldots, a_m) = \sum_{i=1}^{n} |y_i^* - f(x_i, a_1, \ldots, a_m)|^\alpha.$$

Если погрешности в значениях y_i^* по отношению к y_i считать случайными величинами с нормальным законом распределения Гаусса, то эмпирические формулы, в которых $a_1, \ldots, a_m$ определены из условия минимума функции $S_2(a_1, \ldots, a_m)$ при $n \gg m$, наилучшим образом согласуются с результатами эксперимента. А практический опыт показывает, что случайные погрешности значений y_i^* чаще всего имеют закон распределения, близкий к нормальному распределению Гаусса. Поэтому рассмотрим метод определения параметров $a_1, \ldots, a_m$ исходя из минимизации функции

$$S_2(a_1, \ldots, a_m) = \sum_{i=1}^{n} (y_i^* - f(x_i, a_1, \ldots, a_m))^2.$$

Этот метод называется *методом наименьших квадратов*.

Рассмотрим некоторые способы решения поставленной задачи.

Если функция $S_2(a_1, \ldots, a_m)$ достигает абсолютного минимума внутри области изменения параметров $a_1, a_2, \ldots, a_m$, то для определения точек, подозрительных на экстремум, запишем систему

$$\frac{\partial S_2}{\partial a_k} = -2\sum_{i=1}^{n} [y_i^* - f(x_i, a_1, \ldots, a_m)] \frac{\partial f(x_i, a_1, \ldots, a_m)}{\partial a_k} = 0$$
$$(k = 1, 2, \ldots, m). \tag{37}$$

Определив все решения системы (37), принадлежащие области изменения параметров $a_1, a_2, \ldots, a_m$, и выбрав то решение, для которого

$S_2(a_1, \ldots, a_m)$ имеет абсолютный минимум, получим искомые значения параметров.

Пусть $f(x, a_1, \ldots, a_m)$ — линейна относительно параметров $a_1, \ldots, a_m$. При прежних обозначениях в этом случае требуется найти такой вектор

$$r' = a_1' c_1 + a_2' c_2 + \cdots + a_m' c_m,$$

чтобы

$$\|h'\|_3 = \|r' - y^*\|_3 = \sqrt{(h', h')} = \min_{r \in R} \|r - y^*\|_3 = \min_{r \in R} \sqrt{(h, h)}.$$

Известно, что если h' — элемент наилучшего приближения, то для любого $r \in R_m$

$$(h', r) = 0$$

или

$$(h', c_i) = 0 \quad (i = 1, 2, \ldots, m).$$

Из последнего условия для отыскания $a_1', \ldots, a_m'$ получим систему линейных алгебраических уравнений

$$\sum_{j=1}^{m} a_j' (c_j, c_k) = (y^*, c_k) \quad (k = 1, 2, \ldots, m) \tag{38}$$

с симметричной матрицей

$$S = \begin{pmatrix} (c_1, c_1) & (c_2, c_1) & \ldots & (c_m, c_1) \\ (c_1, c_2) & (c_2, c_2) & \ldots & (c_m, c_2) \\ \ldots & \ldots & \ldots & \ldots \\ (c_1, c_m) & (c_2, c_m) & \ldots & (c_m, c_m) \end{pmatrix}. \tag{39}$$

Систему (38) называют *системой нормальных уравнений*. Если векторы $c_1, \ldots, c_m$ линейно-независимы, то определитель матрицы как определитель Грамма положителен и система (38) имеет единственное решение. Если же $c_1, \ldots, c_m$ линейно-зависимы, то $|S| = 0$ и система (38) имеет не единственное решение. Существуют приемы решения систем вида (38) и в этом случае, причем некоторые естественные условия на $a_1', \ldots, a_m'$ вполне определяют нужное решение этих систем.

Заметим, что система нормальных уравнений может быть записана в виде

$$c^* c a' = c^* y^*,$$

где

$$c = \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1m} \\ c_{21} & c_{22} & \ldots & c_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & \ldots & c_{nm} \end{pmatrix},$$

c^* — матрица, транспонированная к c. Следовательно, систему нормальных уравнений можно получить, если систему условных уравнений $ca' = y^*$ умножить слева на c^*.

Рассмотрим теперь случай нелинейной зависимости функции $f(x, a_1, \ldots, a_m)$ от параметров, но предположим, что известны приближенные

значения параметров $a_1^0, \ldots, a_m^0$, отличающиеся от искомых значений $a_1, \ldots, a_m$ малыми поправками $\alpha_1, \ldots, \alpha_m$ и функция $f(x, a_1, \ldots, a_m)$ дифференцируема по $a_1, a_2, \ldots, a_m$. Тогда имеют место приближенные равенства

$$f(x_i, a_1, \ldots, \alpha_m) = f(x_i, a_1^0, \ldots, a_m^0) + \sum_{k=1}^{m} f'_{a_k}(x_i, a_1^0, \ldots, a_m^0)\,\alpha_k.$$

Если ввести обозначения

$$f'_{a_k}(x_i, a_1^0, \ldots, a_m^0) = C_{ik} \quad (i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, m);$$

$$y_i^* - f(x_i, a_1^0, \ldots, a_m^0) = y_i^{**} \quad (i = 1, 2, \ldots, n),$$

то для $\alpha_1, \alpha_2, \ldots, \alpha_m$ получим систему условных уравнений

$$C\alpha = y^{**}, \tag{40}$$

где $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, $y^{**} = (y_1^{**}, y_2^{**}; \ldots; y_m^{**})$. Обозначив решение системы нормальных уравнений, соответствующей системе (40), через $\alpha_1^0, \alpha_2^0, \ldots, \alpha_m^0$, получим следующее приближение параметров:

$$a_1^1 = a_1^0 + \alpha_1^0; \quad a_2^1 = a_2^0 + \alpha_2^0; \ \ldots; \ a_m^1 = a_m^0 + \alpha_m^0.$$

Принимая их за новое начальное приближение параметров, можно продолжить процесс уточнения до тех пор, пока с заданной точностью поправки не будут равны нулю.

Изложенный выше метод легко распространить на случай определения параметров $a_1, a_2, \ldots, a_m$ из системы r эмпирических формул с n независимыми переменными:

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n, \quad a_1, \ldots, a_m); \\ y_2 &= f_2(x_1, \ldots, x_n, \quad a_1, \ldots, a_m); \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_r &= f_r(x_1, \ldots, x_n, \quad a_1, \ldots, a_m), \end{aligned}$$

если для точек $(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$ $(i = 1, 2, \ldots, N)$ известны приближенные значения $y_j^{(i)*} = f_j(x_1^{(i)}, \ldots, x_n^{(i)}, a_1, \ldots, a_m)$, причем $N \gg m$.

В заключение отметим, что на задачу приближения функций, заданных таблицей значений, с помощью алгебраического многочлена степени m по методу наименьших квадратов можно смотреть как на задачу построения эмпирических формул в виде многочлена степени m. Роль параметров в этом случае играют коэффициенты многочлена, причем система условных уравнений имеет вид

$$f(x_i) = \sum_{k=0}^{m} a_k x_i^k \quad (i = 1, 2, \ldots, n),$$

а коэффициенты многочлена наилучшего приближения в смысле метода наименьших квадратов будут находиться как решение системы нормальных уравнений.

§ 2. РАВНОМЕРНЫЕ ПРИБЛИЖЕНИЯ

Пусть в соответствии с общей постановкой задачи о приближении функций (гл. 1, § 1) $f(x) \in C(S)$, т. е. $f(x)$ принадлежит банаховому пространству функций, непрерывных на компакте S, и пусть $\{\varphi_i(x)\}_{i=0}^n \in C(S)$ — последовательность линейно-независимых функций. Замыкание линейной оболочки над полем R_1 функций $\{\varphi_i(x)\}$ обозначаем, как и раньше, через M_n. Задача равномерного приближения состоит в нахождении для $f(x)$ элемента наилучшего приближения $\Phi_0(x) \in M_n$ в смысле метрики пространства $C(S)$, т. е. для $\Phi_0(x)$ должно выполняться соотношение

$$\Delta_n(f) = \Delta_n(f, \Phi_0) = \inf_{\Phi \in M_n} \|f - \Phi\|_C = \inf_{\Phi \in M_n} \max_{x \in S} |f(x) - \Phi(x)|. \quad (1)$$

Существование хотя бы одного элемента (обобщенного многочлена) наилучшего приближения гарантируется теоремой 2, § 1, гл. 1. Для единственности обобщенного многочлена наилучшего приближения согласно теореме 3, § 1, гл. 1 необходимо и достаточно, чтобы система $\{\varphi_i(x)\}_{i=0}^n$ была системой Чебышева на S.

1. Свойства наилучшего приближения

Как было показано в § 1, свойства наилучшего приближения очень просто получаются из геометрических свойств гильбертовых пространств. В случае равномерных приближений дело обстоит значительно сложнее и тем не менее и здесь возможен «геометрический» подход.

Пусть S_1 имеет тот же смысл, что и в лемме 1, § 1, гл. 1, т. е.

$$S_1 = \{s \in S : |\delta(s)| = \Delta_n(f)\}, \quad (2)$$

где

$$\delta(s) = \Phi_0(s) - f(s). \quad (3)$$

Разобьем S_1 на два подмножества S_{+1} и S_{-1}, которые определим следующим образом:

$$S_k = \{s \in S_1 : k \operatorname{sign} \delta(s) = 1\}, \quad k = \pm 1. \quad (4)$$

Рассмотрим совокупность векторов вида

$$\varphi(s) = (\varphi_i(s))_{i=0}^n, \quad s \in S, \quad (5)$$

которые являются элементами R_{n+1}. Выделим из множества векторов (5) те, для которых $s \in S_1$, и введем обозначения

$$L_k = \{\operatorname{sign} k\varphi(s) : s \in S_k\}, \quad k = \pm 1; \quad L = L_{+1} \cup L_{-1}. \quad (6)$$

По построению все множества S_1, $S_{\pm 1}$ являются компактными, а следовательно, будут компактными и L, $L_{\pm 1}$. Для любого множества L через L° будем обозначать наименьшее выпуклое множество, содержащее L, т. е. для всех $m = 1, 2, \ldots, n$, $l_i \in L$, $i = 1, 2, \ldots, m$,

$$\sum_{i=1}^m \rho_i l_i \in L^0, \ \rho_i > 0, \ \sum_{i=1}^m \rho_i = 1.$$

Прежде чем привести геометрическую формулировку леммы 1, § 1, гл. 1, приведем следующую вспомогательную лемму:

Лемма 1. *Пусть L — компакт в R_{n+1}. Выпуклое множество*

$$C = \{x \in R_{n+1} : (l, x) < 0, \ \forall l \in L\} = \varnothing \tag{7}$$

тогда и только тогда, когда $0 \in L^\circ$, где (l, x) — скалярное произведение в R_{n+1} : $(l, x) = \sum_{i=0}^{n} l_i x_i$.

Доказательство. *Необходимость.* Пусть $0 \in L^0$, т. е. существуют m; $l_i \in L$, $i = 1, 2, \ldots, m$; $\rho_i > 0$, $\sum_{i=1}^{m} \rho_i = 1$ такие, что выполняется соотношение

$$\sum_{i=1}^{m} \rho_i l_i = 0.$$

Если бы множество C было не пусто, т. е. не удовлетворяло (7), то нашелся бы хоть один элемент x, для которого выполнялось бы неравенство $(l, x) < 0 \ \forall l \in L$ и, в частности, $(l_i, x) < 0$ $(i = 1, 2, \ldots, m)$. Но тогда $\left(\sum_{i=1}^{m} \rho_i l_i, x\right) < 0$, пришли к противоречию.

Достаточность. Достаточность будем доказывать от противного. Пусть $C = \varnothing$ и $0 \bar{\in} L^\circ$. Поскольку наименьшая выпуклая оболочка компакта есть компакт (приложение, § 1) и $0 \bar{\in} L^\circ$, то в R_{n+1} найдется такой открытый шар

$$B = \{x : \|x\| < \varepsilon, \ \varepsilon > 0\}$$

с центром в точке 0, что $L^\circ \cap B = \varnothing$. Следовательно, найдется такая гиперплоскость

$$\Gamma = \{x \in R_{n+1} : (x, y) = \lambda, \ y \in R_{n+1}\}, \tag{8}$$

которая разделит пространство R_{n+1} на две части так, что в каждую из частей попадет одно и только одно из множеств L°, B. Точнее, по следствию из теоремы Хана — Банаха (приложение, § 1) существует такая гиперплоскость (8), что

$$(x, y) \leqslant \lambda, \ \forall x \in L^0; \tag{9}$$

$$(x, y) > \lambda, \ \forall x \in B. \tag{9'}$$

Так как $0 \in B$, то из последнего неравенства (9′) вытекает, что $\lambda < 0$. Тогда для фиксированного таким образом элемента $y \in R_{n+1}$ из (9) следует

$$(x, y) \leqslant \lambda < 0, \ \forall x \in L,$$

т. е. $y \in C$ в противоположность допущению, что $C = \varnothing$. Лемма доказана.

Теперь мы можем дать геометрическую интерпретацию леммы 1, § 1, гл. 1.

Лемма 2. *Для того чтобы элемент $\Phi_0 \in M_n$ был наилучшим приближением к f в M_n, необходимо и достаточно, чтобы $0 \in L^0$, где L° — наименьшая выпуклая оболочка множества L из* (6).

Доказательство. Согласно введенным обозначениям (6), условия леммы 1, § 1, гл. 1, можно записать в виде

$$C = \{x \in R_{n+1} : (l, x) < 0, \ \forall l \in L\} = \varnothing,$$

где $x = (x_i)_{i=0}^n$; $l = [\operatorname{sign} \delta(s)] \varphi(s)$, $s \in L_1$. Используя лемму 1, получаем доказательство требуемого утверждения.

Приведем вспомогательную лемму.

Лемма 3. *Пусть последовательность $\{\varphi_i(s)\}_{i=0}^n \in C(S)$ является системой Чебышева на S и M_n — ее линейная оболочка. Пусть $g : M_n \to R_1$ непрерывный линейный функционал, задаваемый формулой*

$$g(\Phi) = \sum_{i=1}^{k} \lambda_i \Phi(x_i), \quad \left(\sum_{i=1}^{k} \lambda_i^2 \neq 0\right), \quad \forall \Phi \in M_n, \tag{10}$$

где $x_i \neq x_j$ при $i \neq j$ и $x_i \in S$, $i = 1, 2, \ldots, k$. Тогда, если

$$g(\Phi) = 0, \quad \forall \Phi \in M_n, \tag{11}$$

то $k = n + 2$ и числа λ_i определяются с точностью до мультипликативной постоянной.

Доказательство. Пусть $k \leqslant n + 1$. Подставляя в условие (11) вместо $\Phi(x)$ последовательно элементы системы Чебышева $\{\varphi_i(x)\}_{i=0}^n$, получим

$$g(\varphi_j) = \sum_{i=1}^{k} \lambda_i \varphi_j(x_i) = 0, \quad j = 0, 1, \ldots, n, \tag{12}$$

причем не все λ_i равны нулю. Полагая $\lambda_i = 0$ $(i = k + 1, \ldots, n + 1)$, если $k < n + 1$, систему (12) можно записать в виде

$$\sum_{i=1}^{n+1} \lambda_i \varphi_j(x_i) = 0, \quad j = 0, 1, \ldots, n.$$

Поскольку последняя система имеет нетривиальное решение, ее определитель должен равняться нулю. Отсюда приходим к противоречию, что $\{\varphi_i(x)\}_{i=0}^n$ является системой Чебышева на S. Итак, $k = n + 2$.

Чтобы найти все λ_i, достаточно взять произвольное λ_{n+2} и решить систему

$$\sum_{i=1}^{n+1} \lambda_i \varphi_j(x_i) = -\lambda_{n+2} \varphi_j(x_{n+2}), \quad j = 0, 1, \ldots, n,$$

которая имеет единственное решение. В результате получаем решение, пропорциональное параметру λ_{n+2}. Лемма доказана.

Теперь можно привести теорему, играющую фундаментальную роль в теории равномерных приближений. Для случая, когда ищется многочлен наилучшего приближения в $C[a, b]$, т. е. когда система линейно-независимых функций $\{\varphi_i\}$ совпадает с последовательностью многочленов $(\varphi_i(x) \in \pi_i)$, эта теорема была впервые доказана П. Л. Чебышевым. Поэтому будем ее называть *обобщенной теоремой Чебышева.*

Теорема 1. *Пусть $\{\varphi_i(x)\}_{i=0}^n$ и M_n те же, что и в лемме 3. Тогда для того чтобы элемент $\Phi_0(x) \in M_n$ был наилучшим приближением*

к $f(x) \in C(S)$ ($f \notin M_n$) в M_n^*, необходимо и достаточно, чтобы существовали $x_i \in S$ и $\rho_i > 0$ $i = 1, 2, \ldots, n+2$, удовлетворяющие условиям:

$$[\operatorname{sign} \delta(x_i)]\, \delta(x_i) = \Delta_n(f), \quad i = 1, 2, \ldots, n+2, \tag{13}$$

$$\sum_{i=1}^{n+2} \rho_i \operatorname{sign} \delta(x_i)\, \Phi(x_i) = 0, \quad \forall \Phi \in M_n. \tag{14}$$

Доказательство. По лемме 2 элемент Φ_0 будет наилучшим приближением к f в M_n тогда и только тогда, если $\mathbf{0} \in L^0$. А согласно лемме Каратеодори (приложение, § 1), $\mathbf{0} \in L^0$, тогда и только тогда, когда существуют $\boldsymbol{l}_i \in L$ и $\rho_i > 0$, $\sum_{i=0}^{k} \rho_i = 1$, где $k \leqslant n+2$, такие, что $\mathbf{0} = \sum_{i=0}^{k} \rho_i \boldsymbol{l}_i$.

По определению множества L (см. (6)), имеем:

$$\mathbf{0} = \sum_{i=1}^{k} \rho_i \operatorname{sign} \delta(x_i)\, \varphi(x_i), \quad x_i \in S_1$$

или в координатной форме

$$0 = \sum_{i=1}^{k} \rho_i \operatorname{sign} \delta(x_i)\, \varphi_j(x_i), \quad j = 0, 1, \ldots, n. \tag{15}$$

Из (15) и леммы 3 следует, что $\mathbf{0} \in L^0$ тогда и только тогда, если имеет место (14), а следовательно, и (13).

2. Равномерные приближения на отрезке

В том случае, когда компакт S совпадает с отрезком $[a, b]$, тогда обобщенная теорема Чебышева значительно упрощается. Предварительно докажем лемму, уточняющую лемму 3, для рассматриваемого случая $S = [a, b]$.

Лемма 4. *Пусть выполнены условия леммы 3 с заменой S на $[a, b]$, тогда функционал g, задающийся формулой* (10), *имеет вид*

$$g(\Phi) = \sum_{i=1}^{n+2} \lambda_i \Phi(x_i), \quad \forall \Phi \in M_n, \tag{16}$$

где $a \leqslant x_1 < x_2 < \ldots < x_{n+2} \leqslant b$, $\lambda_i \lambda_{i+1} < 0$, $i = 1, 2, \ldots, n+1$.

Доказательство. Согласно лемме 3, необходимо доказать только чередование знаков в последовательности λ_i, $i = 1, 2, \ldots, n+2$. В виду того что $\{\varphi_i(x)\}_{i=0}^{n}$ — система Чебышева на $[a, b]$, по теореме 1, § 1, гл. 1, существует единственная функция $\Phi \in M_n$ и такая, что

$$\Phi(x_i) = 0, \quad i = 1, 2, \ldots, n+2, \quad i \neq j, j+1,$$
$$\Phi(x_j) = 1.$$

Легко видеть, что $\Phi(x_{j+1}) > 0$, иначе, будучи непрерывной функцией, на отрезке $[x_j, x_{j+1}]$ она должна была бы иметь $(n+1)$-й нуль, что

противоречило бы условию Хаара. Подстановка функции $\Phi(x)$ в (11) с учетом (16) приводит к соотношению

$$g(\Phi) = \lambda_j + \lambda_{j+1}\Phi(x_{j+1}) = 0,$$

из которого и следует требуемое утверждение.

Теорема 2. *Пусть выполнены условия теоремы 1 для $S = [a, b]$. Функция $\Phi_0(x) \in M_n$ будет элементом наилучшего приближения к $f(x) \in C[a, b]$ тогда и только тогда, если существуют точки x_i, удовлетворяющие условиям:*

$$a \leqslant x_1 < x_2 < \cdots < x_{n+2} \leqslant b;$$

$$|\delta(x_i)| = \Delta_n(f), \quad i = 1, 2, \ldots, n+2; \tag{17}$$

$$\delta(x_i) = -\delta(x_{i+1}), \quad i = 1, 2, \ldots, n+1. \tag{18}$$

Доказательство. *Необходимость.* Выполнение условий теоремы 1 приводит к справедливости (17), а лемма 4 с $\lambda_i = \operatorname{sign}\delta(x_i)\rho_i$ $(i = 1, 2, \ldots, n+2)$ — к справедливости (18).

Достаточность. Пусть $\Phi \in M_n$ является такой функцией, что выполняются условия (17) и (18). Построим по точкам x_i $(i = 1, 2, \ldots, n+2)$, также как и при доказательстве леммы 3, функционал (10) со свойством (11). Изменив в случае необходимости знак у всех λ_i, можно считать, что $\operatorname{sign}(\lambda_1) = \operatorname{sign}\delta(x_1)$.

По лемме 4 $\lambda_i\lambda_{i+1} < 0$, $i = 1, 2, \ldots, n+1$. Тогда согласно (18) $\operatorname{sign}\lambda_i = \operatorname{sign}\delta(x_i)$ $(i = 1, 2, \ldots, n+2)$ и, взяв $\rho_i = |\lambda_i|$ $(i = 1, 2, \ldots, n+2)$, получим

$$g(\Phi) = \sum_{i=1}^{n}\lambda_i\Phi(x_i) = \sum_{i=1}^{n}\rho_i \operatorname{sign}\delta(x_i)\,\Phi(x_i) = 0, \ \forall\,\Phi \in M_n.$$

Тогда по теореме 1 $\Phi = \Phi_0$ — наилучшее приближение к f в M_n.

Определение 1. Набор точек x_i $(i = 1, 2, \ldots, n+2)$, удовлетворяющих условиям (17) и (18), называется *чебышевским альтернансом.*

Приведем некоторые оценки для величин $\Delta_n(f)$, когда $M_n = \pi_n$.

Пусть $l : C[a, b] \to R_1$ — непрерывный линейный функционал, являющийся разделенной разностью $(n+1)$-го порядка,

$$l(f) = f(x_1;\ x_2;\ \ldots\ x_{n+2}) = \sum_{i=1}^{n+2}\frac{f(x_i)}{\omega'(x_i)} \tag{19}$$

(приложение, § 3). Очевидно

$$\|l\| = \sum_{i=1}^{n+2}\frac{1}{|\omega'(x_i)|}, \tag{20}$$

где

$$\omega(x) = \prod_{i=1}^{n+2}(x - x_i).$$

Преобразуем функционал l так, чтобы норма его стала равной единице. Для этого достаточно умножить его на соответствующую постоянную, после чего будем иметь:

$$\bar{l}(f) = Nl(f) = \sum_{i=1}^{n+2} \lambda_i f(x_i), \text{ где } \lambda_i = [\omega'(x_i)]^{-1} \|l\|^{-1}; \tag{21}$$

$$N = \left[\sum_{i=0}^{n+2} \frac{1}{|\omega'(x_i)|}\right]^{-1} = \|l\|^{-1}.$$

Оценка 1. $\forall f(x) \in C[a, b]$ и $x_i \in [a, b]$ $(i = 1, 2, \ldots, n+2)$ будем иметь:

$$\Delta_n(f) \geqslant |\bar{l}(f)|, \tag{22}$$

где наилучшее приближение ищется в классе многочленов n-й степени, т. е. $M_n = \pi_n$.

Доказательство. Так как на основании свойств разделенных разностей $\bar{l}(f) = 0$, $\forall f \in \pi_n$ и, кроме того, $\|\bar{l}\| = 1$, то

$$|\bar{l}(f)| = |\bar{l}(\Phi - f)| \leqslant \|\bar{l}(\Phi - f)\| \leqslant \|\Phi - f\|, \quad \forall \Phi \in \pi_n.$$

Отсюда получаем

$$|\bar{l}(f)| \leqslant \inf_{\Phi \in \pi_n} \|\Phi - f\| = \Delta_n(f)$$

и оценка (1) доказана.

Оценка 2. Если для некоторого многочлена $\Phi(x) \in \pi_n$ и различных точек $x_i \in [a, b]$, $i = 1, 2, \ldots, n+2$,

$$\operatorname{sign} \delta(x_i) = -\operatorname{sign} \delta(x_{i+1}) \neq 0, \quad i = 1, 2, \ldots, n+1,$$

то

$$\min_i |\delta(x_i)| \leqslant \Delta_n(f), \tag{23}$$

где $\delta(x) = \Phi(x) - f(x)$.

Доказательство. Из формул (21) видно, что $\lambda_i \lambda_{i+1} < 0$ для всех $i = 1, 2, \ldots, n+1$ (см. доказательство леммы 3). Следовательно,

$$|\bar{l}(f)| = |\bar{l}(\Phi - f)| = \left|\sum_{i=1}^{n+2} \lambda_i [\Phi(x_i) - f(x_i)]\right| = \sum_{i=1}^{n+2} |\lambda_i| |\Phi(x_i) - f(x_i)| \geqslant$$

$$\geqslant \min_i |\Phi(x_i) - f(x_i)| \sum_{i=1}^{n+2} |\lambda_i| = \min_i |\delta(x_i)|$$

и оценка (2) доказана.

Оценка 3. Если функции $f(x)$, $g(x) \in C^{(n+1)}[-1, 1]$ и $|f^{(n+1)}(x)| \leqslant g^{(n+1)}(x)$ $\forall x \in [-1, 1]$, то $\Delta_n(f) \leqslant \Delta_n(g)$.

Доказательство. Пусть сначала имеет место строгое неравенство $|f^{(n+1)}(x)| < g^{(n+1)}(x)$ $\forall x \in [-1, 1]$. Предположим противное, т. е. $\Delta_n(f) > \Delta_n(g)$. Возьмем многочлены наилучшего приближения $\Phi_{01}(x)$ и $\Phi_{02}(x)$ соответственно к функции $f(x)$ и $g(x)$ в π_n и положим

$$\begin{aligned} h_1(x) &= [\Phi_{01}(x) - f(x)] + [\Phi_{02}(x) - g(x)] = \delta_1(x) + \delta_2(x); \\ h_2(x) &= \delta_1(x) - \delta_2(x). \end{aligned} \tag{24}$$

Пусть x_i $(i = 1, 2, \ldots, n+2)$ — точки чебышевского альтернанса для $\delta_1(x) = \Phi_{01}(x) - f(x)$. Тогда в силу нашего предположения

$$\begin{aligned} &h_1(x_i) h_2(x_i) > 0, && i = 1, 2, \ldots, n+2; \\ &h_k(x_i) h_k(x_{i+1}) < 0, && i = 1, 2, \ldots, n+1; \quad k = 1, 2. \end{aligned} \tag{25}$$

Вычислим разделенные разности $(n+1)$-го порядка от введенных функций (24):

$$h_k(x_1;\ x_2;\ \ldots;\ x_{n+2}) = \sum_{i=1}^{n+2} \frac{h_k(x_i)}{\omega'(x_i)}, \quad k = 1, 2,$$

которые в силу (25) и свойств разделенных разностей будут иметь одинаковые знаки, а вместе с ними и выражения:

$$h_k^{(n+1)}(\xi_k) = h_k(x_1;\ x_2;\ \ldots;\ x_{n+2}) = -f^{(n+1)}(\xi_k) + (-1)^k g^{(n+1)}(\xi_k), \quad k = 1, 2,$$

где $\xi_k \in [-1, 1]$ $(k = 1, 2)$ — некоторые промежуточные точки. Пришли к противоречию, предположив, что

$$|f^{(n+1)}(x)| < g^{(n+1)}(x), \quad \forall x \in [-1, 1].$$

Случай нестрогого неравенства $|f^{(n+1)}(x)| \leqslant g^{(n+1)}(x)$, $\forall x \in [-1, 1]$ можно свести к уже рассмотренному, если вместо функции $g(x)$ ввести функцию

$$\tilde{g}(x) = g(x) + \varepsilon \frac{x^{n+1}}{(n+1)!}, \quad \varepsilon > 0, \tag{26}$$

для которой уже будет выполняться строгое неравенство

$$|f^{(n+1)}(x)| < g^{(n+1)}(x) + \varepsilon.$$

Поэтому

$$\Delta_n(f) \leqslant \Delta_n\left(g + \varepsilon \frac{x^{n+1}}{(n+1)!}\right) \leqslant \Delta_n(g) + \frac{\varepsilon}{(n+1)!} \Delta_n(x^{n+1}), \tag{27}$$

а так как ε — произвольное положительное число, то оценка 3 полностью доказана.

Заметим, что величина $\Delta_n(x^{n+1})$ может быть легко вычислена. Имеет место следующая лемма:

Лемма 5. *Для функции x^{n+1} элементом наилучшего приближения в π_n является многочлен*

$$\Phi_0(x) = x^{n+1} - \frac{1}{2^n} T_{n+1}(x) \in \pi_n,$$

где $T_{n+1}(x)$ — многочлен Чебышева первого рода. Причем

$$\Delta_n(x^{n+1}) = \frac{1}{2^n}.$$

Доказательство. Поскольку в точках $x_{i+1} = \cos\frac{i\pi}{n+1}$ $(i = 0, 1, \ldots, n+1)$ выполняются соотношения:

$$\delta(x_{i+1}) = \Phi_0(x_{i+1}) - x_{i+1}^{n+1} = -\frac{1}{2^n} T_{n+1}(x_{i+1}) = -\frac{1}{2^n}\cos i\pi = -\frac{(-1)^i}{2^n},$$

$$i = 0, 1, \ldots, n+1;$$

$$\Delta_n(\Phi_0, x^{n+1}) = \|\Phi_0(x) - x^{n+1}\| = \left\|\frac{1}{2^n} T_{n+1}(x)\right\| = \left\|\frac{1}{2^n}\cos(n+1)\arccos x\right\| = \frac{1}{2^n},$$

то эти точки являются чебышевским альтернансом и можно воспользоваться теоремой 2. Отсюда сразу получаем то, что требовалось доказать.

Следствие 1. Пусть $f(x) \in C^{(n+1)}[-1, 1]$ и

$$0 \leqslant m_{n+1} \leqslant f^{(n+1)}(x) \leqslant M_{n+1}, \quad \forall x \in [-1, 1],$$

тогда

$$\frac{m_{n+1}}{2^n(n+1)!} \leqslant \Delta_n(f) \leqslant \frac{M_{n+1}}{2^n(n+1)!}. \tag{28}$$

Доказательство. Воспользуемся оценкой 3 и положим $g(x) = \frac{M_{n+1}x^{n+1}}{(n+1)!}$, тогда, применяя лемму 5, получим правую часть оценки (28). Аналогично получаем левую часть.

3. Алгоритм построения алгебраических многочленов наилучшего приближения

Пусть M_n — подпространство $C(S)$, порожденное линейно-независимыми элементами $\varphi_i(x)$ $(i = 0, 1, \ldots, n)$, образующими систему Чебышева на компакте S. Рассмотрим сначала наилучшее приближение к f в M_n на дискретном множестве различных точек $x_i \in S$ $(i = 1, 2, \ldots, n+2)$, которое обозначим через S_1.

Определение 2. Элемент $\overline{\Phi}_0 \in M_n$ называется *наилучшим приближением* к f в M_n на дискретном множестве точек $S_1 = \{x_i : i = 1, 2, \ldots, n+2\}$, если

$$\overline{\Delta}_n(f) = \min_{\Phi \in M_n} \Delta_n(\Phi; f) = \min_{\Phi \in M_n} \max_{x \in S_1} |\Phi(x) - f(x)|.$$

Теорема 3. *$\forall f \in C(S)$ существует единственное наилучшее приближение $\overline{\Phi}_0 \in M_n$ к f относительно S_1. При этом, если*

$$\overline{\Delta}_n(f) > 0,$$

то на $C(S)$ существует единственный непрерывный линейный функционал

$$l(\Phi) = \sum_{i=1}^{n+2} \rho_i \varepsilon_i \Phi(x_i), \quad \rho_i > 0, \quad \varepsilon_i = \pm 1, \quad \sum_{i=1}^{n+2} \rho_i = 1, \tag{29}$$

для которого

$$l(\Phi) = 0, \ \forall \Phi \in M_n, \ l(f) < 0. \tag{30}$$

Кроме того, наилучшее приближение $\overline{\Phi}_0$ удовлетворяет условию

$$\overline{\Phi}_0(x_i) - f(x_i) = \varepsilon_i \overline{\Delta}_n(f), \quad i = 1, 2, \ldots, n+2. \tag{31}$$

Доказательство. Для доказательства воспользуемся результатами теоремы 1. Нужно показать только, что $l(f) < 0$. Действительно,

$$-l(f) = -\sum_{i=1}^{n+2} \rho_i \varepsilon_i (x_i) = \sum_{i=1}^{n+2} \rho_i \varepsilon_i [\overline{\Phi}_0(x_i) - f(x_i)] =$$

$$= \sum_{i=1}^{n+2} \rho_i |\overline{\Phi}_0(x_i) - f(x_i)| = \overline{\Delta}_n(f).$$

Для практического построения $\overline{\Phi}_0(x)$ поступают следующим образом. Ищем $\overline{\Phi}_0(x)$ в виде

$$\overline{\Phi}_0(x) = \sum_{i=0}^{n} C_i \varphi_i(x)$$

и для отыскания коэффициентов C_i исходя из (31) составляем систему линейных алгебраических уравнений

$$\sum_{i=0}^{n} C_i\varphi_i(x_j) - \varepsilon_j\bar{\Delta}_n(f) = f(x_j), \quad j = 1, 2, \ldots, n+2. \tag{32}$$

Причем в системе (32) величину $\bar{\Delta}_n^{-}(f)$ также считаем неизвестной. Этот метод эффективен в том случае, когда все ε_j известны заранее (хотя бы с точностью до знака), что будет иметь место, когда $S = [a, b]$ и

$$a \leqslant x_1 < x_2 < \cdots < x_{n+2} \leqslant b.$$

В этом случае известно, что $\varepsilon_j\varepsilon_{j+1} = -1$ $(j = 1, 2, \ldots, n+1)$ и система (32) принимает вид

$$\sum_{i=0}^{n} C_i\varphi_i(x_j) + \varepsilon(-1)^j\bar{\Delta}_n(f) = f(x_j); \quad j = 1, 2, \ldots, n+2, \tag{32'}$$

где ε равно либо $+1$, либо -1.

Пусть $M_n = \pi_n$ и пусть $\Phi_1, \Phi_2 \in \pi_{n+1}$ — многочлены, удовлетворяющие условиям:

$$\Phi_1(x_i) = f(x_i); \quad \Phi_2(x_i) = (-1)^i, \quad i = 1, 2, \ldots, n+2.$$

Возьмем многочлен

$$\Phi(x) = \Phi_1(x) + \lambda\Phi_2(x) \tag{33}$$

и выберем параметр λ так, чтобы $\Phi(x) \in \pi_n^{-}$, тогда

$$\Phi(x_i) - f(x_i) = \lambda(-1)^i, \quad i = 1, 2, \ldots, n+2.$$

Следовательно, $\Phi(x) = \bar{\Phi}_0(x)$ и $\bar{\Delta}_n^{-}(f) = |\lambda|$.

$$\Phi_1(x) = f(x_1) + \sum_{i=1}^{n+1} \omega_i(x) f(x_1;\ x_2;\ \ldots;\ x_{i+1});$$

$$\Phi_2(x) = -1 + \sum_{i=1}^{n+1} \omega_i(x) \sum_{k=1}^{i} \frac{(-1)^k}{\omega_i'(x_k)};$$

$$\omega_i(x) = \prod_{p=1}^{i} (x - x_p), \quad i = 1, 2, \ldots.$$

Для того чтобы многочлен $\Phi(x) \in \pi_n$, полагаем в (33)

$$\lambda = -f(x_1;\ x_2;\ \ldots;\ x_{n+1}) \left[\sum_{k=1}^{n+1} \frac{(-1)^k}{\omega_{n+1}'(x_k)}\right]^{-1}. \tag{33'}$$

Теперь перейдем к описанию алгоритма, который был предложен советским математиком Е. Я. Ремезом.

Пусть на ν-й итерации найдено наилучшее приближение $\Phi_0^{(\nu)}$ к f на множестве $S_1^{(\nu)} = \{x_i^{(\nu)} : i = 1, 2, \ldots, n+2\}$ и

$$\Lambda_n^{(\nu)}(f) = \min_{\Phi \in M_n} \Delta_n^{(\nu)}(\Phi, f) = \min_{\Phi \in M_n} \max_{x \in S_1^{(\nu)}} |\Phi(x) - f(x)|.$$

Предположим, $\Delta_n^{(\nu)}(f) > 0$. Обозначим через $l^{(\nu)}$ непрерывный линейный функционал (29) из теоремы 3:

$$l^{(\nu)}(\Phi) = \sum_{i=1}^{n+2} \rho_i^{(\nu)} \varepsilon_i^{(\nu)} \Phi(x_i^{(\nu)}), \quad \rho_i^{(\nu)} > 0; \quad \sum_{i=1}^{n+2} \rho_i^{(\nu)} = 1,$$

который согласно (30) удовлетворяет условиям:

$$l^{(\nu)}(\Phi) = 0; \quad \forall \Phi \in M_n; \quad l^{(\nu)}(f) = -\Delta_n^{(\nu)}(f),$$

причем

$$\varepsilon_i^{(\nu)} = \operatorname{sign}[\Phi_0^{(\nu)}(x_i^{(\nu)}) - f(x_i^{(\nu)})].$$

Очевидно,

$$\Delta_n^{(\nu)}(f) \leqslant \Delta_n(f) \leqslant \Delta_n(\Phi_0^{(\nu)}, f). \tag{34}$$

Если $\Delta_n^{(\nu)}(f) = \Delta_n(\Phi_0^{(\nu)}, f)$, то согласно теореме 2 $\Phi_0^{(\nu)}(x) = \Phi_0(x)$ и задача решена. Пусть

$$\Delta_n^{(\nu)}(f) < \Delta_n(\Phi_0^{(\nu)}, f).$$

Возьмем такую точку $x_0 \in S$, что $\varepsilon_0[\Phi_0^{(\nu)}(x_0) - f(x_0)] = \Delta_n(\Phi_0^{(\nu)}, f)$, где $\varepsilon_0 = \operatorname{sign}[\Phi_0^{(\nu)}(x_0) - f(x_0)]$ (таких точек может оказаться несколько, тогда берем любую из них). Очевидно, $x_0 \notin S_1^{(\nu)}$.

Заменим множество $S_1^{(\nu)}$ множеством $S_1^{(\nu+1)} = \{x_i^{(\nu+1)} : i = 1, 2, \ldots \ldots, n+2\}$, где

$$x_i^{(\nu+1)} = \begin{cases} x_i^{(\nu)}, & i \neq j; \\ x_0, & i = j. \end{cases}$$

Произведем также замену $\varepsilon_i^{(\nu)}$ на $\varepsilon_i^{(\nu+1)}$ по правилу

$$\varepsilon_i^{(\nu+1)} = \begin{cases} \varepsilon_i^{(\nu)}, & i \neq j; \\ \varepsilon_0, & i = j. \end{cases}$$

Пусть индекс j выбран так, что существуют положительные числа $\rho_i^{(\nu+1)}$, для которых

$$l^{(\nu+1)}(\Phi) = \sum_{i=1}^{n+2} \rho_i^{(\nu+1)} \varepsilon_i^{(\nu+1)} \Phi(x_i^{(\nu+1)}) = 0, \quad \forall \Phi \in M_n,$$

где

$$\rho_i^{(\nu+1)} > 0; \quad \sum_{i=1}^{n+2} \rho_i^{(\nu+1)} = 1; \quad l^{(\nu+1)}(f) < 0. \tag{35}$$

Пусть $\Phi_0^{(\nu+1)} \in M_n$ — наилучшее приближение к f в M_n относительно $S_1^{(\nu+1)}$ и

$$\Delta_n^{(\nu+1)}(f) = \min_{\Phi \in M_n} \Delta_n^{(\nu)}(\Phi, f) = \min_{\Phi \in M_n} \max_{x \in S_1^{(\nu+1)}} |\Phi(x) - f(x)|,$$

тогда, как и на предыдущей итерации,

$$\Delta_n^{(\nu+1)}(f) \leqslant \Delta_n(f) \leqslant \Delta_n(\Phi_0^{(\nu+1)}, f).$$

Лемма 6. Имеют место неравенства

$$\Delta_n^{(\nu)}(f) < \Delta_n^{(\nu+1)}(f) \leqslant \Delta_n(f). \tag{36}$$

Более того,

$$\Delta_n^{(\nu+1)}(f) = \Delta_n^{(\nu)} + \rho_j^{(\nu+1)}[\Delta_n(\Phi_0^{(\nu)}, f) - \Delta_n^{(\nu)}(f)]. \tag{37}$$

Доказательство. Действительно, на основании свойств функционала $l^{(\nu+1)}$ получаем:

$$\Delta_n^{(\nu+1)}(f) = -l^{(\nu+1)}(f) = l^{(\nu+1)}(\Phi_0^{(\nu)} - f) =$$

$$= \sum_{i=1}^{n+2} \rho_i^{(\nu+1)}\varepsilon_i^{(\nu+1)}[\Phi_0^{(\nu)}(x_i^{(\nu+1)}) - f(x_i^{(\nu+1)})] = \sum_{\substack{i=1\\ j\neq 1}}^{n+2} \rho_i^{(\nu+1)}\varepsilon_i^{(\nu)}[\Phi_0^{(\nu)}(x_i^{(\nu)}) -$$

$$- f(x_i^{(\nu)})] + \rho_j^{(\nu+1)}\varepsilon_0^{(\nu+1)}[\Phi_0^{(\nu)}(x_0) - f(x_0)] =$$

$$= \Delta_n^{(\nu)}(f)\sum_{i=1}^{n+2} \rho_i^{(\nu+1)} - \rho_j^{(\nu+1)}\Delta_n^{(\nu)}(f) + \rho_j^{(\nu+1)}\Delta_n(\Phi_0^{(\nu)}, f) =$$

$$= \Delta_n^{(\nu)}(f) + \rho_j^{(\nu+1)}[\Delta_n(\Phi_0^{(\nu)}, f) - \Delta_n^{(\nu)}(f)]$$

и соотношение (37) доказано. Отсюда на основании (35) и (34) сразу получаем (36).

Для случая, когда компакт S совпадает с отрезком $[a, b]$, функционалы $l^{(\nu)}$ для всех ν являются просто разделенными разностями и нахождение постоянных $\rho_i^{(\nu+1)}$, удовлетворяющих условиям (35), не представляет труда и производится по формулам (21).

Относительно перехода от множества $S_1^{(\nu)}$ к множеству $S_1^{(\nu+1)}$ для $S = [a, b]$ существует следующее простое правило:

Точкой x_0 следует заменить ту из точек $x_j^{(\nu)} \in S_1^{(\nu)}$, чтобы после упорядочивания полученного нового множества точек $x_1^{(\nu+1)}$ $(i = 1, 2, \ldots, n+2)$, образующих $S_1^{(\nu+1)}$, имело место соотношение

$$[\Phi_0^{(\nu)}(x_i^{(\nu+1)}) - f(x_i^{(\nu+1)})][\Phi_0^{(\nu)}(x_{i+1}^{(\nu+1)}) - f(x_{i+1}^{(\nu+1)})] < 0,$$
$$i = 1, 2, \ldots, n+1.$$

Для доказательства сходимости $\Phi_0^{(\nu)} \to \Phi_0$ построенного итерационного процесса рассмотрим предварительно две леммы.

Лемма 7. *Существует* $\varepsilon > 0$ *такое, что*

$$\rho(x_i^{(\nu)}, x_j^{(\nu)}) \geqslant \varepsilon, \quad i \neq j, \ i, j = 1, 2, \ldots, n+2, \ \nu = 1, 2, \ldots, \tag{38}$$

где $x_i^{(\nu)} \in S_1^{(\nu)}$ *и* $\rho(x, y)$ *— расстояние между элементами* x *и* y *компакта* S.

Доказательство. Возьмем две произвольные точки $x_i^{(\nu)} \neq x_j^{(\nu)}$ из $S_1^{(\nu)}$ и предположим, что условие леммы для этих точек не выполняется. Если ввести обозначения

$$\lim_{k\to\infty} x_r^{(\nu_k)} = \bar{x}_r, \quad r = 1, 2, \ldots, n+2, \tag{39}$$

которые будут иметь смысл в силу принадлежности $x_r^{(\nu)}$ компакту S, то $\bar{x}_i = \bar{x}_j$ и среди предельных точек $\bar{x}_r$ $(r = 1, 2, \ldots, n+2)$ будет не более $n+1$ различных. Следовательно, можно найти такой многочлен $\Phi(x) \in M_n$, что

$$\Phi(\bar{x}_i) = f(\bar{x}_i), \quad i = 1, 2, \ldots, n+2. \tag{40}$$

На основании леммы 6 можно записать

$$|f(x_i^{(\nu)}) - \Phi_0^{(\nu)}(x_i^{(\nu)})| = \Delta_n^{(\nu)}(f) > \Delta_n^{(1)}(f) > 0, \quad i = 1, 2, \ldots, n+2$$

(простейший случай окончания итерационного процесса за конечное число шагов исключается).

Выберем $\Delta_n^{(1)}(f) > \delta > 0$ и такие окрестности V_r точек x_r, что

$$|f(x) - \Phi(x)| < \delta, \ \forall x \in \bigcup_{r=1}^{n+2} V_r. \tag{41}$$

Последнее неравенство будет иметь место в силу (40).

Далее ввиду (39) и (41) существует номер N такой, что при $\nu_k > N$ будет иметь место соотношение

$$\operatorname{sign}[\Phi_0^{(\nu_k)}(x_r^{(\nu_k)}) - \Phi(x_r^{(\nu_k)})] = \operatorname{sign}\{[\Phi_0^{(\nu_k)}(x_r^{(\nu_k)}) - f(x_r^{(\nu_k)})] -$$

$$- [\Phi(x_r^{(\nu_k)}) - f(x_r^{(\nu_k)})]\} = \operatorname{sign}[\Phi_0^{(\nu_k)}(x_r^{(\nu_k)}) - f(x_r^{(\nu_k)})], \ r = 1, 2, \ldots, n+2.$$

Следовательно, функционал

$$l^{(\nu_k)}(\Phi^{(\nu_k)} - \Phi) = \sum_{r=1}^{n+2} \rho_r^{(\nu_k)} \varepsilon_r^{(\nu_k)} [\Phi_0^{(\nu_k)}(x_r^{(\nu_k)}) - \Phi(x_r^{(\nu_k)})] =$$

$$= \sum_{r=1}^{n+2} \rho_r^{(\nu_k)} |\Phi_0^{(\nu_k)}(x_r^{(\nu_k)}) - \Phi(x_r^{(\nu_k)})|$$

будет положительным на элементе $\Phi^{(\nu_k)} - \Phi \in M_n$, что противоречит его свойствам. Если бы случайно произошло, что $\Phi = \Phi^{(\nu_k)}$, то тогда бы следовало вместо ν_k взять ν_{k+1}. Лемма доказана.

Лемма 8. *$\exists q$, $1 > q > 0$ и такое, что все числа $\rho_i^{(\nu)}$, входящие в функционал $l^{(\nu)}$, удовлетворяют неравенству*

$$\rho_i^{(\nu)} \geqslant q, \quad i = 1, 2, \ldots, n+2; \quad \nu = 1, 2, \ldots.$$

Доказательство. Рассмотрим компакт $S^{n+2} = \underbrace{S \times S \times \cdots \times S}_{n+2}$.

Полученные в процессе итераций векторы $\boldsymbol{x}^{(\nu)} = (x_i^{(\nu)})_{i=1}^{n+2}$ лежат в компакте

$$S_1^{n+2} = \{\boldsymbol{x} \in S^{n+2} : \rho(x_i, x_j) \geqslant \varepsilon > 0, \quad i \neq j, \ i, j = 1, 2, \ldots, n+2\}.$$

Коэффициенты $\rho_i^{(\nu)}$ функционала $l^{(\nu)}$ согласно его свойствам являются непрерывными функциями от $\boldsymbol{x}^{(\nu)}$. Тогда на компакте S_1^{n+2} они достигают своей точной нижней границы

$$q_i = \inf_{\nu} \rho_i^{(\nu)}, \quad i = 1, 2, \ldots, n+2. \tag{42}$$

Используя (29), легко показать, что все q_i в (42) положительны, и в качестве q можно взять $\min\limits_i q_i$, причем

$$0 < q < 1.$$

Теорема 4. *Имеют место предельные соотношения:*

$$\lim_{\nu\to\infty} \Delta_n^{(\nu)}(f) = \Delta_n(f) = \Delta_n(\Phi_0, f); \tag{43}$$

$$\lim_{\nu\to\infty} \Phi_0^{(\nu)} = \Phi_0. \tag{44}$$

Доказательство. Используя леммы 6 и 8, будем иметь

$$\Delta_n^{(\nu+1)}(f) - \Delta_n^{(\nu)}(f) = \rho_f^{(\nu+1)}[\Delta_n(\Phi_0^{(\nu)}, f) - \Delta_n^{(\nu)}(f)] \geqslant$$
$$\geqslant q[\Delta_n(\Phi_0^{(\nu)}, f) - \Delta_n^{(\nu)}(f)] \geqslant q[\Delta_n(f) - \Delta_n^{(\nu)}(f)],$$

откуда следует неравенство

$$\Delta_n(f) - \Delta_n^{(\nu+1)}(f) \leqslant (1-q)[\Delta_n(f) - \Delta_n^{(\nu)}(f)],$$

следствием которого является (43). Точно так же получаем

$$\|\Phi_0^{(\nu)} - f\| - \Delta_n(f) \leqslant \|\Phi_0^{(\nu)} - f\| - \Delta_n^{(\nu)}(f) \leqslant$$
$$\leqslant \frac{1}{q}[\Delta_n^{(\nu+1)}(f) - \Delta_n^{(\nu)}(f)]_{\nu\to\infty} \to 0.$$

Из последнего соотношения уже несложно получить (44).

Рассмотрим один из возможных подходов к построению приближений для многочлена $\Phi_0(x) \in \pi_n$, равномерно приближающего функцию $f(x) \in C[a, b]$. Предпосылкой этого подхода служит теорема A и тот факт, что

$$\lim_{k\to\infty} \|f\|_{L_k} = \|f\|_C, \quad \forall f \in C[a, b]. \tag{45}$$

Введем обозначения, пусть $\Phi_0^{(k)}(x)$ — многочлен наилучшего приближения к $f(x)$ в M_n в смысле нормы в $L_k[a, b]$, т. е.

$$\Delta_n^{(k)}(f) = \inf_{\Phi\in M_n} \|f - \Phi\|_{L_k} = \Delta_n^{(k)}(\Phi_0^{(2k)}, f), \quad k = 1, 2, \ldots. \tag{46}$$

Имеет место следующая теорема:

Теорема 5. *Последовательность* $\Phi_0^{(m_k)}(x) \in \pi_n$, $m_{k+1} = m_k(m_k + 1)$, $k = 0, 1, \ldots, m_0 = 1$, *построенная согласно соотношениям* (46), *является сходящейся, причем*

$$\lim_{k\to\infty} \Phi_0^{(m_k)}(x) = \Phi_0(x).$$

Доказательство. Воспользовавшись неравенством Гельдера, будем иметь:

$$\|f\|_{L_{m_k}} \leqslant (b-a)^{\frac{1}{m_k+1}} \|f\|_{L_{m_{k+1}}}, \quad \forall f \in C[a, b]. \tag{47}$$

Не уменьшая общности, можно в (47) считать, что $b - a = 1$. Тогда (47) приводит к неравенству

$$\|f - \Phi_0^{(m_k)}\|_{L_{m_k}} \leqslant \|f - \Phi_0^{(m_{k+1})}\|_{L_{m_k}} \leqslant \|f - \Phi_0^{(m_{k+1})}\|_{L_{m_{k+1}}}$$

или в соответствии с обозначениями (46)

$$\Delta_n^{(m_k)}(f) \leqslant \Delta_n^{(m_{k+1})}(f), \quad k = 0, 1, \ldots. \tag{48}$$

Используя очевидное неравенство

$$\|f - \Phi_0^{(m_k)}\|_{L_{m_k}} \leqslant \|f - \Phi_0\|_{L_{m_k}} \leqslant \Delta(f), \quad (b - a = 1)$$

и соотношение (48), получаем, что последовательность $\{\Delta_n^{(m_k)}(f)\}_{k=0}^{\infty}$ является монотонно возрастающей и ограниченной сверху. Следовательно, она имеет предел и

$$\lim_{k\to\infty} \Delta_n^{(m_k)}(f) = \lim_{k\to\infty} \|f - \Phi_0^{(m_k)}\|_{L_{m_k}} = \|f - \overline{\Phi}_0\|_C.$$

Покажем, что $\overline{\Phi}_0 = \Phi_0$. Действительно, из неравенства

$$\|f - \Phi_0^{(m_k)}\|_{L_{m_k}} \leqslant \|f - \Phi_0\|_{L_{m_k}}$$

при $k \to \infty$ получаем

$$\|f - \overline{\Phi}_0\|_C \leqslant \|f - \Phi_0\|_C,$$

что может быть только в том случае, когда $\overline{\Phi}_0 = \Phi_0$.

§ 3. ИНТЕРПОЛЯЦИОННЫЕ И СГЛАЖИВАЮЩИЕ СПЛАЙН-ФУНКЦИИ

В этом параграфе ограничимся некоторыми аспектами теории сплайн-функций, связанными с минимизацией функционалов в гильбертовых пространствах.

1. Интерполяционные сплайн-функции

Пусть $W_2^l = W_2^l\ [a, b]$, $l \geqslant 1$, — гильбертово пространство вещественных функций, определенных на отрезке $[a, b]$, с абсолютно непрерывной $(l - 1)$-й производной и суммируемой с квадратом l-й производной.

Скалярное произведение в этом пространстве задается формулой

$$(u, v) = \sum_{i=0}^{l} \int_a^b u^{(i)}(x)\, v^{(i)}(x)\, dx, \tag{1}$$

с помощью которой вводится норма

$$\|u\|_{W_2^l} = (u, u)^{\frac{1}{2}}. \tag{2}$$

Определение 1. Пространство вещественных функций $s(x)$, определенных на отрезке $[a, b]$ и удовлетворяющих условиям:

1) $s(x) \in \pi_{2q-1}\ \forall x \in (x_i, x_{i+1}), \quad i = 1, 2, \ldots, n-1;$

2) $s(x) \in \pi_{q-1}\ \forall x \in [a, x_1), (x_n, b];$

3) $s(x) \in C^{(2q-2)}[a, b],$

где $n \geqslant q$, называется *пространством сплайн-функций* порядка q относительно точек x_i и обозначается через S.

Из определения ясно, что $S \subset W_2^q$ и что многочлены $s(x)$ из каждого участка склеиваются с соседними многочленами в точках стыка x_i по своим значениям и значениям своих производных до порядка $(2q-2)$ включительно. Производная порядка $(2q-1)$ может претерпевать разрыв.

Используя обозначение

$$(x)^+ = \frac{x + |x|}{2}, \tag{3}$$

которым будем также пользоваться и в дальнейшем, получим, что функция

$$\delta_i(x) = [(x - x_i)^+]^{2q-1}/(2q-1)! \in C^{(2q-2)}[a, b] \tag{4}$$

и

$$\delta_i^{(2q-1)}(x) = \frac{\operatorname{sign}(x - x_i) + 1 + \delta_{x,\, x_i}}{2}.$$

Пусть $p_i(x) \in \pi_{2q-1}$ и $p_i(x) = s(x)$, $\forall x \in (x_i, x_{i+1})$, т. е.

$$p_i(x) = p_{i-1}(x) + d_i \delta_i(x), \quad \forall x \in (x_i, x_{i+1}), \tag{5}$$

где

$$d_i = s^{(2q-1)}(x_i + 0) - s^{(2q-1)}(x_i - 0). \tag{5'}$$

Легко видеть, что $\forall s(x) \in S$ имеет место представление

$$s(x) = p_0(t) + \sum_{i=1}^{n} d_i \frac{[(x - x_i)^+]^{2q-1}}{(2q-1)!}, \tag{6}$$

где $p_0(x) \in \pi_{q-1}$. Для того чтобы выполнялось условие 2) из определения 1, необходимо, чтобы имело место тождество

$$\sum_{i=1}^{n} d_i \frac{(x - x_i)^{2q-1}}{(2q-1)!} \equiv 0, \tag{7}$$

ибо левая часть (7) должна принадлежать классу π_{q-1}. Приравнивая в (7) коэффициенты при степенях x, получаем

$$\sum_{i=1}^{n} d_i (x_i)^k = 0, \qquad k = 0, 1, \ldots, q-1. \tag{8}$$

Итак $s(x) \in S$ тогда и только тогда, когда имеет место представление

$$s(x) = \sum_{j=0}^{q-1} \alpha_j x^j + \sum_{i=1}^{n} d_i \frac{[(x - x_i)^+]^{2q-1}}{(2q-1)!}, \tag{9}$$

в котором d_i удовлетворяют условию (8).

Лемма 1. *$\forall f \in W_2^q$ и $\forall s(x) \in S$ справедливо соотношение*

$$\int_a^b s^{(q)}(x) f^{(q)}(x)\,dx = (-1)^q \sum_{i=1}^{n} [s^{(2q-1)}(x_i+0) - s^{(2q-1)}(x_i-0)]\, f(x_i). \quad (10)$$

Доказательство. Используя условие 2) определения 1 и интегрируя по частям левую часть (10), получаем

$$\int_a^b s^{(q)}(x) f^{(q)}(x)\,dx = [s^{(q)} f^{(q-1)}(x)]\,\Big|_{x=a}^{x=b} - \int_a^b s^{(q+1)}(x) f^{(q-1)}(x)\,dx =$$

$$= -\int_a^b s^{(q+1)}(x) f^{(q-1)}(x)\,dx = \cdots = (-1)^{q-1} \int_a^b s^{(2q-1)}(x) f'(x)\,dx =$$

$$= (-1)^{q-1} \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} s^{(2q-1)}(x) f'(x)\,dx =$$

$$= (-1)^{q-1} \sum_{i=1}^{n-1} \sum_{j=1}^{i} d_j\,[f(x_{i+1}) - f(x_i)]. \quad (11)$$

Применяя к последнему выражению в (11) первую разностную формулу Грина (приложение, § 3), получаем требуемое соотношение (10).

Можно ввести фундаментальные интерполяционные сплайн-функции $s_j(x)$ со свойствами: $s_j(x) \in S$, $j = 1, 2, \ldots, n$,

$$s_j(x_i) = \delta_{ij}, \qquad i, j = 1, 2, \ldots, n,$$

тогда

$$s(x) = \sum_{i=1}^{n} f_i s_i(x).$$

Теорема 1. *Для любых чисел f_i $(i=1, 2, \ldots, n)$ существует единственная функция $s(x) \in S$ и такая, что $s(x_i) = f_i$ $(i = 1, 2, \ldots, n)$, т. е. $s(x)$ — интерполяционная функция.*

Доказательство. Возьмем $s(x)$ в виде выражения (9), которое содержит $n + q$ параметров α_j $(j = 0, 1, \ldots, q-1)$ и d_i $(i = 1, 2, \ldots, n)$, причем d_i должны удовлетворять системе уравнений (8). Таким образом, для определения неизвестных параметров получаем следующую систему:

$$\begin{aligned} &s(x_i) = f_i, \qquad i = 1, 2, \ldots, n; \\ &\sum_{j=1}^{n} d_j (x_j)^k = 0, \qquad k = 0, 1, \ldots, q-1. \end{aligned} \quad (12)$$

Покажем, что однородная система (12) имеет только тривиальное решение. Действительно, пусть функция $s(x) \not\equiv 0$ удовлетворяет однородной системе (12), тогда из (10) будем иметь

$$\int_a^b [s^{(q)}(x)]^2\,dx = 0,$$

т. е. $s^{(q)}(x) \equiv 0$; $s(x) \in \pi_{q-1}$ и $s(x_i) = 0$, $i = 1, 2, \ldots, n$ при $q \leqslant n$, что может быть лишь в том случае, когда $s(x) \equiv 0$. Противоречие доказывает теорему.

Введем обозначение

$$I_f = \{u \in W_2^q\colon\ u(x_i) = f_i, \quad i = 1, 2, \ldots, n\}, \tag{13}$$

тогда имеет место следующая теорема:

Теорема 2. *Пусть $s(x) \in S$ и $s(x_i) = f_i$ $(i = 1, 2, \ldots, n)$, тогда*

$$\int_a^b [s^{(q)}(x) - f^{(q)}(x)]^2\,dx = \min_{\sigma \in S} \int_a^b [\sigma^{(q)}(x) - f^{(q)}(x)]^2\,dx, \quad \forall f \in I_f \tag{14}$$

и всякая другая функция $\tilde{s}(x) \in S$, обладающая этим свойством, отличается от s на многочлен $(q-1)$-й степени;

$$\int_a^b [s^{(q)}(x) - \sigma^{(q)}(x)]^2\,dx = \min_{u \in I_f} \int_a^b [u^{(q)}(x) - \sigma^{(q)}(x)]^2\,dx, \quad \forall \sigma \in S \tag{15}$$

и $s(x)$ — единственная функция, принадлежащая I_f и обладающая этим свойством.

Доказательство. $\forall\ \sigma(x) \in S$ будем иметь

$$\int_a^b [\sigma^{(q)}(x) - f^{(q)}(x)]^2\,dx = \int_a^b [s^{(q)}(x) - f^{(q)}(x) + \sigma^{(q)}(x) - s^{(q)}(x)]^2\,dx =$$

$$= \int_p^b [s^{(q)}(x) - f^{(q)}(x)]^2\,dx + \int_a^b [\sigma^{(q)}(x) - s^{(q)}(x)]^2\,dx +$$

$$+ 2\int_a^b [s^{(q)}(x) - f^{(q)}(x)]\,[\sigma^{(q)}(x) - s^{(q)}(x)]\,dx. \tag{16}$$

К третьему интегралу в правой части (16) можно применить лемму 1, ибо $\sigma(x) - s(x) \in S$, $s(x) - f(x) \in W_2^q$, но поскольку $s(x)$ — интерполяционный сплайн для $f(x)$, то на основании (10) этот интеграл равен нулю. Следовательно, из (16) получаем

$$\int [\sigma^{(q)}(x) - f^{(q)}(x)]^2\,dx = \int_a^b [s^{(q)}(x) - f^{(q)}(x)]^2\,dx +$$

$$+ \int_a^b [\sigma^{(q)}(x) - s^{(q)}(x)]^2\,dx \geqslant \int_a^b [s^{(q)}(x) - f^{(q)}(x)]^2\,dx,$$

что и доказывает (14). Доказательство (15) производится аналогично.

Следствие 1. Пусть выполнены условия теоремы 2, тогда

$$\int_a^b [s^{(q)}(x)]^2\,dx = \min_{u \in I_f} \int_a^b [u^{(q)}(x)]^2\,dx \tag{17}$$

и $s(x)$ — единственная функция из I_f, обладающая этим свойством.

2. Сглаживающие сплайн-функции

Если значения функции $f(x)$ в узлах x_i $(i = 1, 2, \ldots, n)$ получены из эксперимента и, следовательно, являются неточными, то нет смысла строить приближающий сплайн интерполяционного типа, описанный в п. 1.

Поэтому вместо отыскания минимума функционала, стоящего в правой части (17), целесообразно решать другую экстремальную задачу:

$$\min_{u \in W_2^q} \left\{ \int_a^b [u^{(q)}(x)]^2 dx + \rho \sum_{i=1}^n [u(x_i) - f_i]^2 \right\}, \qquad n \geqslant q, \tag{18}$$

$$(\rho > 0).$$

Имеет место следующая теорема:

Теорема 3. *Для любых чисел $f_i (i=1,2, \ldots, n)$ существует единственная сплайн-функция $s(x) \in S$ такая, что*

$$s(x_i) + \frac{(-1)^q}{\rho} [s^{(2q-1)}(x_i + 0) - s^{(2q-1)}(x_i - 0)] = f_i, \tag{19}$$

$$i = 1, 2, \ldots, n.$$

Доказательство. Возьмем $s(x)$ в виде (9), тогда с учетом (8) для определения $(n + q)$ параметров сплайн-функции $s(x)$ будем иметь следующую систему уравнений:

$$\begin{gathered} \sum_{i=1}^n d_i (x_i)^k = 0, \qquad k = 0, 1, \ldots, q-1; \\ s(x_i) + \frac{(-1)^q}{\rho} d_i = f_i \qquad i = 1, 2, \ldots, n, \end{gathered} \tag{20}$$

где d_i имеет тот же смысл, что и в (5′). Предположим, что однородная система (20) с $f_i = 0$ имеет нетривиальное решение, соответствующее сплайну $\tilde{s}(x)$. Тогда, полагая в (10) $s(x) = f(x) = \tilde{s}(x)$, будем иметь

$$\int_a^b [\tilde{s}^{(q)}(x)]^2 dx + \frac{1}{\rho} \sum_{i=1}^n [d_i]^2 = \int_a^b [\tilde{s}^{(q)}(x)]^2 dx + \rho \sum_{i=1}^n [\tilde{s}(x_i)]^2 = 0,$$

откуда $\tilde{s}^{(q)}(x) \equiv 0$; $\tilde{s}^{(q)}(x) \in \pi_{q-1}$; $\tilde{s}(x_i) = 0$, $i = 1, 2, \ldots, n$. Следовательно, $\tilde{s}(x) \equiv 0$. Противоречие доказывает существование и единственность решения системы (20).

Можно ввести фундаментальные сглаживающие сплайн-функции $s_j(x) \in S$, т. е. функции, удовлетворяющие условиям

$$s_j(x_i) + \frac{(-1)^q}{\rho} [s_j^{(2q-1)}(x_i + 0) - s_j^{(2q-1)}(x_i - 0)] = \delta_{ij}, \tag{21}$$

тогда имеет место формула

$$s(x) = \sum_{i=1}^n f_i s_i(x). \tag{22}$$

Теорема 4. Пусть $s(x)$ — единственная сплайн-функция, удовлетворяющая условиям теоремы 3. Тогда, если ввести обозначение

$$R(\sigma(x)-u(x))=\int_a^b[\sigma^{(q)}(x)-u^{(q)}(x)]^2\,dx+$$

$$+\rho\sum_{i=1}^{n}\left\{\frac{(-1)^q}{\rho}[\sigma^{(2q-1)}(x_i+0)-\sigma^{(2q-1)}(x_i-0)]+u(x_i)-f_i\right\}^2=$$

$$=\int_a^b[\sigma^{(q)}(x)-u^{(q)}(x)]^2\,dx+\rho\sum_{i=0}^{n}[\sigma(x_i)-u(x_i)]^2, \tag{23}$$

то

$$R(s(x)-f(x))=\min_{\sigma\in S}R(\sigma(x)-f(x)),\quad \forall f\in W_2^q \tag{24}$$

и всякая функция $\tilde{s}(x)\in S$, обладающая этим свойством, отличается от $s(x)$ на многочлен $(q-1)$-й степени;

$$R(\sigma(x)-s(x))=\min_{u\in W_2^q}R(\sigma(x)-u(x)),\quad \forall\sigma\in S \text{ и} \tag{25}$$

$s(x)$ — единственная функция из W_2^q, обладающая этим свойством.

Доказательство. С учетом обозначения (23), $\forall\,\sigma(x)\in S$ будем иметь

$$R(\sigma(x)-f(x))=R(s(x)-f(x)+\sigma(x)-s(x))=R(s(x)-f(x))+$$

$$+R(\sigma(x)-s(x))-2\int_a^b[s^{(q)}(x)-f^{(q)}(x)][\sigma^{(q)}(x)-s^{(q)}(x)]\,dx-$$

$$-2\frac{1}{\rho}\sum_{i=1}^{n}[\sigma^{(2q-1)}(x_i+0)-\sigma^{(2q-1)}(x_i-0)-d_i]\,d_i.$$

Воспользовавшись леммой 1 и заменив предварительно $f(x)$ на $s(x)-f(x)$ и $s(x)$ — на $\sigma(x)-s(x)$, последнее выражение преобразуется к виду

$$R(\sigma(x)-f(x))=R(s(x)-f(x))+R(\sigma(x)-s(x))+$$

$$+2(-1)^q\sum_{i=1}^{n}[\sigma^{(2q-1)}(x_i+0)-\sigma^{(2q-1)}(x_i-0)-d_i]\frac{(-1)^q}{\rho}d_i-$$

$$-2\frac{1}{\rho}\sum_{i=1}^{n}[\sigma^{(2q-1)}(x_i+0)-\sigma^{(2q-1)}(x_i-0)-d_i]\,d_i=$$

$$=R(s(x)-f(x))+R(\sigma(x)-s(x))\geqslant R(s(x)-f(x)),$$

из которого следует (24). Соотношение (25) доказывается аналогично.

Следствие 2. Пусть сплайн-функция $s(x)$ та же, что и в теореме 4. Тогда

$$\int_a^b [s^{(q)}(x)]^2\,dx + \rho \sum_{i=1}^{n} [s(x_i) - f_i]^2 =$$

$$= \min_{u \in W_2^q} \left\{ \int_a^b [u^{(q)}(x)]^2\,dx + \rho \sum_{i=1}^{n} [u(x_i) - f_i]^2 \right\} \tag{26}$$

и $s(x)$ — единственная функция из W_2^q, обладающая этим свойством.

Глава 4

ПРИБЛИЖЕННОЕ ВЫЧИСЛЕНИЕ ОПРЕДЕЛЕННЫХ ИНТЕГРАЛОВ

В соответствии с общей задачей приближения линейных операторов, рассмотренной в главе 1, аппроксимируем линейный оператор J, задаваемый формулой

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

где $\rho(x) \geqslant 0$, $\forall x \in [a, b]$, а $f(x)$ принадлежит некоторому банаховому пространству, линейным оператором *(квадратурной формулой)* вида

$$J_n(f) = \sum_{k=1}^{m} \sum_{i=1}^{\alpha_k} A_{k,i}^{(n)} f^{(i)}(x_k^{(n)}), \quad \sum_{k=1}^{m} \alpha_k = n.$$

Здесь $x_k^{(n)} \in [a, b]$, $k = 1, 2, \ldots, m$,— *узлы* (абсциссы) *квадратурной формулы;* $A_{k,i}^{(n)}$ — *коэффициенты (веса)*. Тогда оператор $J(f)$ можно представить в виде

$$J(f) = J_n(f) + R_n(f),$$

где $R_n(f)$ — *остаточный член квадратурной формулы.*

Ниже приводятся и исследуются квадратурные формулы конкретного вида, которые наиболее часто употребляются на практике.

§ 1. ФОРМУЛЫ НЬЮТОНА—КОТЕСА

Построим квадратурную формулу вида

$$\int_c^d f(x)\,dx = \sum_{i=1}^{n} A_i^{(n)} f(x_i) + R(f). \tag{1}$$

Веса $A_i^{(n)}$ квадратурной формулы (1) будем находить из условия

$$R(f) = 0, \quad \forall f \in \bigcup_{i=0}^{n-1} \pi_i, \tag{1'}$$

которое согласно теории интерполирования (§ 1, гл. 2) приводит к соотношению

$$A_i^{(n)} = \int_c^d Q_{n-1,i}(x)\,dx, \qquad i = 1, 2, \ldots, n, \tag{2}$$

где $Q_{n-1,i}(x)$ — фундаментальные многочлены Лагранжа (см. формулу (17), § 1, гл. 2). Алгебраическая точность формулы (1) равна $n - 1$.

Если в квадратурной формуле (1), (2) узлы $x_i \in [c, d]$, $i = 1, 2, \ldots, n$, являются *равноотстоящими*, т. е. $x_{i+1} - x_i = h$, $i = 1, 2, \ldots, n - 1$, то такая формула называется *формулой Ньютона — Котеса*.

Будем предполагать, что расстояние между узлами x_i задается формулой

$$\begin{cases} \dfrac{d-c}{n+1}, & \text{если } x_1 = c + h \text{ (случай 1);} \\ \dfrac{d-c}{n-1}, & \text{если } x_1 = c \qquad \text{(случай 2).} \end{cases} \tag{3}$$

В первом случае узлы квадратурной формулы не содержат точек c и d и промежуток интегрирования разбивается этими узлами на $n + 1$ равных частей. Во втором случае концы промежутка интегрирования являются узлами интерполирования и промежуток интегрирования разбивается узлами на $n - 1$ равных частей. Формулы численного интегрирования, которые получаются в первом случае, называются *формулами открытого типа*, а во втором случае — *формулами замкнутого типа*. Если положить

$$x_1 = c + kh, \quad x_n = d - kh, \tag{3'}$$

то для формул открытого типа $k = 1$, а для формул замкнутого типа $k = 0$.

Сделаем в интеграле (1) замену: $x = x_0 + hy = c + h(y + k - 1)$, полагая $f(x_0 + hy) = F(y)$, тогда

$$\int_c^d f(x)\,dx = h \int_{1-k}^{n+k} F(y)\,dy. \tag{4}$$

В последнем интеграле заменим функцию $F(y)$ интерполяционным многочленом Лагранжа с узлами в точках 1, 2, ..., n (см. формулу (18), § 1, гл. 2), в результате чего получим

$$\int_c^d f(x)\,dx = h \int_{1-k}^{n+k} F(y)\,dy = h \sum_{i=1}^{n} \tilde{J}_{i,k}^{(n)} F(i) + R_{n,k}(f), \tag{5}$$

где

$$\tilde{J}_{i,k}^{(n)} = \frac{(-1)^{n-i}}{(i-1)!\,(n-i)!} \int_{1-k}^{n+k} \frac{(y-1)(y-2)\ldots(y-n)}{y-i}\,dy; \tag{6}$$

$$R_{n,k}(f) = h \int_{1-k}^{n+k} (y-1)(y-2)\ldots(y-n)\,F(y;\ 1;\ 2;\ \ldots;\ n)\,dy. \tag{7}$$

Учитывая замену, запишем равенство (5) в виде

$$\int_c^d f(x)\,dx = (d-c)\sum_{i=1}^{n} J_{i,k}^{(n)} f(x_0+ih) + R_{n,k}(f), \tag{8}$$

где в обозначениях формулы (1) $A_i^{(n)} = (d-c)\,J_{i,k}^{(n)}$ и

$$J_{i,k}^{(n)} = \frac{\tilde{J}_{i,k}^{(n)}}{n-1+2k}. \tag{9}$$

Величины $J_{i,k}^{(n)}$ не зависят от промежутка интегрирования и могут быть вычислены раз и навсегда. Вычисление облегчается еще и тем обстоятельством, что равноотстоящие от концов коэффициенты формулы Ньютона — Котеса равны, т. е.

$$J_{i,k}^{(n)} = J_{n-i+1,k}^{(n)}. \tag{10}$$

В самом деле,

$$J_{n-i+1,k}^{(n)} = \frac{(-1)^{i-1}}{(n-1+2k)(n-i)!\,(i-1)!}\int_{1-k}^{n+k} \frac{(y-1)(y-2)\ldots(y-n)}{y-n+i-1}\,dy.$$

Заменяя под знаком интеграла y на $n-z+1$, получим

$$J_{n-i+1,k}^{(n)} = \frac{(-1)^{i}}{(n-1+2k)(n-i)!\,(i-1)!}\int_{n+k}^{1-k} \frac{(n-z)(n-z-1)\ldots(-z+1)}{i-z}\,dz =$$

$$= \frac{(-1)^{n-i}}{(n-1+2k)(n-i)!\,(i-1)!}\int_{1-k}^{n+k} \frac{(z-1)(z-2)\ldots(z-n)}{z-i}\,dz = J_{i,k}^{(n)},$$

что и требовалось доказать.

Изучение веса $J_{i,0}^{(n)}$ показало, что величины $|J_{i,0}^{(n)}|$ с возрастанием n неограниченно возрастают и $\lim\limits_{n\to\infty}\sum\limits_{i=1}^{n} |J_{i,0}^{(n)}| = \infty$. Отсюда, в частности, следует, что при больших n малые ошибки в значениях функции $f(x_0+ih)$ могут дать большую погрешность в квадратурной сумме. На основании указанной особенности на практике квадратурные формулы Ньютона — Котеса при большом n не используются. Предпочтение отдается формулам с малым значением n. Для уменьшения же погрешности результата предварительно разбивают отрезок $[c, d]$ на достаточно большое число малых интервалов и к каждому из них применяют квадратурную формулу с малым числом узлов.

Рассмотрим формулу открытого типа при $n = 1$ и формулы замкнутого типа при $n = 2$ и $n = 3$.

Положим в формуле (5) $k = 1$, $n = 1$, тогда

$$\int_c^d f(x)\,dx = (d-c)\,f\left(\frac{c+d}{2}\right) + R_{1,1}(f). \tag{11}$$

Разобьем отрезок $[c, d]$ на r равных частей длины $h = \frac{d-c}{r}$ и применим формулу вида (11) к каждому из интегралов, на которые разобьется

исходный интеграл точками деления $x_i = c + ih$, $i = \overline{1, r-1}$. Получим

$$\int_c^d f(x)\,dx = h\sum_{i=1}^{r} f\left(c + \frac{2i-1}{2}h\right) + R_1(f), \tag{12}$$

где $R_1(f) = \sum_{i=1}^{n} R_{1,1}^{(i)}(f)$, $R_{1,1}^{(i)}(f)$ — остаточный член квадратурной формулы на i-м интервале. Формула (12) называется *формулой средних прямоугольников*.

Если положить $k = 0$, $n = 2$, то формула (5) для этих значений примет вид

$$\int_c^d f(x)\,dx = \frac{d-c}{2}[f(c) + f(d)] + R_{2,0}(f). \tag{13}$$

Взяв в (13) $c + ih$ вместо c и $c + (i+1)h$ вместо d и суммируя обе части по i от 0 до r, получим *формулу трапеций:*

$$\int_c^d f(x)\,dx = h\left[\frac{1}{2}f(c) + \sum_{i=1}^{r} f(c+ih) + \frac{1}{2}f(d)\right] + R_2(f), \tag{14}$$

где $h = (d-c)/(r+1)$; $R_2(f) = \sum_{i=1}^{r+1} R_{2,0}^{(i)}(f)$, $R_{2,0}^{(i)}(f)$ — остаточный член квадратурной формулы на i-м интервале.

Пусть в формуле (5) $k = 0$, $n = 3$, тогда имеем:

$$\int_c^d f(x)\,dx = \frac{d-c}{6}\left[f(c) + 4f\left(\frac{c+d}{2}\right) + f(d)\right] + R_{3,0}(f). \tag{15}$$

Если теперь в (15) заменить c на $c + 2ih$, d — на $c + 2(i+1)h$ и просуммировать обе части по i от 0 до r, то получим *формулу Симпсона,* или формулу парабол

$$\int_c^d f(x)\,dx = \frac{h}{3}\left[f(c) + 4\sum_{i=1}^{r+1} f(c + (2i-1)h) + \right.$$

$$\left. + 2\sum_{i=1}^{r} f(c+2ih) + f(d)\right] + R_3(f), \tag{16}$$

где $h = \frac{d-c}{2(r+1)}$, $R_3(f) = \sum_{i=1}^{r} R_{3,0}^{(i)}(f)$, $R_{3,0}^{(i)}(f)$ — остаточный член квадратурной формулы на интервале $[c + 2ih,\ c + 2(i+1)h]$.

Лемма 1. *Пусть в квадратурной формуле* (1), *точной* $\forall f \in \bigcup_{i=0}^{n-1} \pi_i$, *узлы* x_i $i = 1, 2, \ldots, n$ *расположены симметрично относительно середины отрезка* $[a, b]$. *Тогда, если* $n = 2m + 1$ $(m = 1, 2, \ldots)$, *то алгебраическая точность формулы* (1) *будет равна* n, *т. е. повышается на единицу.*

Доказательство. Построим для функции $f(x)$ интерполяционный многочлен n-й степени по узлам $x_1, x_2, \ldots, x_n, x_i$, т. е. с одним кратным узлом (i-м). Воспользовавшись формулой (24), § 1, гл. 2, получим

$$L_n(x) = f(x_1) + (x - x_1) f(x_1;\ x_2) + \cdots + (x - x_1) \ldots$$
$$\ldots (x - x_{n-1}) f(x_1;\ x_2;\ \ldots;\ x_n) + (x - x_1) \ldots$$
$$\ldots (x - x_n) f(x_1;\ x_2;\ \ldots;\ x_n;\ x_i) =$$
$$= L_{n-1}(x) + (x - x_1) \ldots (x - x_n) f(x_1;\ x_2;\ \ldots;\ x_n;\ x_i) \qquad (17)$$

(относительно разделенных разностей с кратными узлами см. приложение, § 3).

Поскольку n у нас нечетное и все узлы x_k, $k = 1, 2, \ldots, n$, расположены симметрично точке $\frac{a+b}{2}$, то

$$\int_a^b (x - x_1)(x - x_2) \ldots (x - x_n)\, dx = 0$$

и из (17) получаем $\int_a^b L_n(x)\, dx = \int_a^b L_{n-1}(x)\, dx$. Но $R(L_{n-1}) = 0$, следовательно, и $R(L_n) = 0$, что и требовалось доказать.

На основании доказанной леммы формула средних прямоугольников имеет алгебраическую точность 1, а формула Симпсона — 3.

Названия квадратурных формул (12), (14) и (15) происходят из геометрических соображений: криволинейные трапеции, сумма площадей которых равна точному значению интеграла, заменяются соответственно прямоугольниками, трапециями, криволинейными трапециями, верхней стороной которых является парабола.

Отметим еще, что при $k = 0$, $n = 4$ формула (8) принимает вид так называемого *правила трех восьмых:*

$$\int_c^d f(x)\, dx = (d - c)\left[\frac{1}{8} f(c) + \frac{3}{8} f\left(c + \frac{d-c}{3}\right) + \right.$$
$$\left. + \frac{3}{8} f\left(c + \frac{2(d-c)}{3}\right) + \frac{1}{8} f(d)\right] + R_{4,0}(f). \qquad (18)$$

Поскольку алгебраическая точность формулы (18) совпадает с алгебраической точностью формулы Симпсона при количестве узлов на 1 больше, чем в формуле Симпсона, то правило трех восьмых не получило широкого применения.

§ 2. КВАДРАТУРНЫЕ ФОРМУЛЫ НАИВЫСШЕЙ АЛГЕБРАИЧЕСКОЙ СТЕПЕНИ ТОЧНОСТИ

Пусть формула численного интегрирования

$$\int_a^b f(x)\, \rho(x)\, dx = \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) + R(f), \qquad (1)$$

где $]a, b]$ — любой конечный или бесконечный интервал; $\rho(x) \geqslant 0$ $\forall x \in [a, b]$ — весовая функция, удовлетворяющая неравенствам

$$\left| \int_b^a \rho(x) x^i dx \right| < \infty, \qquad i = 0, 1, \ldots,$$

является формулой интерполяционного типа, т. е. $R(f) = 0$ $\forall f \in \bigcup_{i=0}^{n-1} \pi_i$ и

$$C_k^{(n)} = \int_a^b Q_{n-1,k}(x) \rho(x) dx = \int_a^b \rho(x) \frac{\omega(x)}{(x - x_k^{(n)}) \omega'(x_k^{(n)})} dx; \tag{2}$$

$$\omega(x) = \prod_{i=1}^{n} (x - x_i^{(n)}), \quad x_i^{(n)} \in [a, b].$$

Рассматривая узлы $x_i^{(n)}$, $i = 1, 2, \ldots, n$, как произвольные точки из $[a, b]$, поставим задачу определения этих узлов из условия, чтобы формула (1) имела наивысшую степень алгебраической точности. Сразу заметим, что наивысшая степень алгебраической точности меньше или равна $2n - 1$. Действительно, рассматривая формулу (1) для многочлена $f(x) = \prod_{i=0}^{n} (x - x_i^{(n)})^2 \in \pi_{2n}$, получаем

$$R(f) = \int_a^b \prod_{i=1}^{n} (x - x_i^{(n)})^2 \rho(x) dx > 0.$$

Если удастся найти такое распределение узлов $x_i^{(n)}$, что в (1) $R(f) = 0$ $\forall f \in \bigcup_{i=0}^{2n-1} \pi_i$, то такую формулу будем называть *квадратурной формулой наивысшей алгебраической степени точности.*

Теорема 1. *Для того чтобы формула численного интегрирования* (1), (2) *была квадратурной формулой наивысшей алгебраической степени точности, необходимо и достаточно, чтобы* $\omega(x) = \frac{P_n(x)}{k_{0,n}}$, *т. е. чтобы узлы* $x_i^{(n)}$, $i = 1, 2, \ldots, n$, *совпадали с нулями многочлена* $P_n(x)$, *входящего в последовательность многочленов* $\{P_i(x)\}_{i=0}^{\infty}$, *образующих ортогональную систему на* $[a, b]$ *с весом* $\rho(x)$. *Причем такая квадратурная формула будет единственной.*

Доказательство. *Необходимость.* Предположим, что в (1) $R(f) = 0$, $\forall f \in \bigcup_{i=0}^{2n-1} \pi_i$ Тогда $\forall Q_m(x) \in \pi_m$, $m \leqslant n - 1$, произведение $Q_m(x) \omega(x) \in \bigcup_{i=0}^{2n-1} \pi_i$ и из (1) следует, что

$$\int_a^b Q_m(x) \omega(x) \rho(x) dx = \frac{1}{k_{0,n}} \int_a^b Q_m(x) P_n(x) \rho(x) dx = R(Q_m \omega) = 0, \tag{3}$$

$$m = 0, 1, \ldots, n - 1.$$

Но соотношение (3) как раз и является определением ортогональной системы многочленов $\{P_i(x)\}$.

Достаточность. Пусть $\omega(x) = \frac{P_n(x)}{k_{0,n}}$ и $f(x) \in \bigcup_{i=0}^{2n-1} \pi_i$, тогда имеет место представление

$$f(x) = \omega(x) Q_m(x) + r_p(x), \tag{4}$$

где многочлены $Q_m(x)$ и $r_p(x)$ имеют степень меньше n, и

$$\int_a^b f(x) \rho(x)\, dx = \int_a^b \omega(x) Q_m(x) \rho(x)\, dx + \int_a^b r_p(x) \rho(x)\, dx. \tag{5}$$

Первое слагаемое в правой части согласно предположению обращается в нуль. Поскольку из (4) следует, что

$$f(x_i^{(n)}) = r_p(x_i^{(n)}) \text{ и } r_p(x) \in \pi_m, \quad m \leqslant n-1,$$

то $r_p(x)$ совпадает с интерполяционным многочленом функции $f(x)$. Но формула (1) является квадратурной формулой интерполяционного типа, следовательно,

$$\int_a^b f(x) \rho(x)\, dx = \int_a^b r_p(x) \rho(x)\, dx = \sum_{i=1}^{n} C_i^{(n)} f(x_i^{(n)}), \quad \forall f \in \bigcup_{i=1}^{2n-1} \pi_i$$

и достаточность доказана. Единственность квадратурной формулы наивысшей алгебраической степени точности вытекает из единственности (с точностью до мультипликативных постоянных) системы многочленов $\{P_i(x)\}_{i=0}^{\infty}$, ортогональных с весом $\rho(x)$ на $[a, b]$ (приложение, § 2). Теорема доказана полностью.

Квадратурная формула (1), (2) наивысшей алгебраической степени точности называется еще *формулой механических квадратур Гаусса.*

Весовые коэффициенты $C_k^{(n)}$ в квадратуре Гаусса (1), (2) обычно называются *коэффициентами Кристоффеля* и для них справедлива такая теорема:

Теорема 2. *Коэффициенты Кристоффеля* $C_k^{(n)} > 0$, $k = 1, 2, \ldots, n$, *и удовлетворяют соотношениям:*

$$\sum_{k=1}^{n} C_k^{(n)} = \int_a^b \rho(x)\, dx; \tag{6}$$

$$C_k^{(n)} = \int_a^b \left[\frac{P_n(x)}{P_n'(x_k^{(n)})(x - x_k^{(n)})} \right]^2 \rho(x)\, dx; \tag{7}$$

$$C_k^{(n)} = \frac{k_{0,n+1}}{k_{0,n}} \cdot \frac{-h_n}{P_{n+1}(x_k^{(n)}) P_n'(x_k^{(n)})} = \frac{k_{0,n}}{k_{0,n-1}} \cdot \frac{h_{n-1}}{P_{n-1}(x_k^{(n)}) P_n'(x_k^{(n)})}; \tag{8}$$

$$[C_k^{(n)}]^{-1} = \sum_{p=0}^{n} h_p^{-1} [P_p(x_k^{(n)})]^2. \tag{9}$$

Доказательство. Справедливость соотношений (6) и (7) следует из того, что формула Гаусса (1) является точной для любого многочлена степени не выше $2n-1$ и, в частности, для $f(x)=1$ и

$$f(x)=\left[\frac{P_n(x)}{P_n'(x_k^{(n)})(x-x_k^{(n)})}\right]^2\in\pi_{2n-2}.$$

Для доказательства формулы (8) воспользуемся тождеством Кристоффеля — Дарбу (приложение, § 2)

$$\sum_{p=0}^{n} h_p^{-1}P_p(x)P_p(y)=\frac{k_{0,n}h_n}{k_{0,n+1}}\cdot\frac{P_{n+1}(x)P_n(y)-P_n(x)P_{n+1}(y)}{x-y}. \tag{10}$$

Положив $y=x_i^{(n)}$, умножим обе части этого тождества на выражение $-\frac{k_{0,n+1}}{k_{0,n}h_n}\cdot\frac{1}{P_{n+1}(x_i^{(n)})P_n'(x_i^{(n)})}$ и проинтегрируем с весом $\rho(x)$ на интервале $[a, b]$. Согласно формуле (2) и соотношению $\omega(x)=\frac{P_n(x)}{k_{0,n}}$, мы получим справа $C_k^{(n)}$, а слева вторую часть формулы (8), ибо все слагаемые, кроме первого, обратятся в нуль вследствие ортогональности последовательности $\{P_i(x)\}$. Третья часть формулы (8) получается из второй с использованием рекуррентного соотношения для последовательности ортогональных многочленов $\{P_i(x)\}$ (приложение, § 2). Наконец, справедливость формулы (9) устанавливается путем предельного перехода в тождестве Кристоффеля — Дарбу при $y\to x=x_i^{(n)}$ и использования формулы (8).

Приведем веса и абсциссы (узлы) квадратурной формулы (8) для весовых функций $\rho(x)$, связанных с классическими ортогональными многочленами (табл. 1). Интегралы такого типа довольно часто встречаются на практике, что и объясняет выделение их из всего множества интегралов с другими весовыми функциями.

В табл. 1 строка под номером 3 выделена из строки 2 особо, так как все весовые коэффициенты в этой квадратурной формуле постоянны, но тем не менее порядок алгебраической точности ее $2n-1$. Эта квадратурная формула называется *квадратурной формулой Чебышева.*

П. Л. Чебышевым была поставлена задача о построении квадратурных формул с равными весовыми коэффициентами для любой весовой функции $\rho(x)$. Эта задача будет рассмотрена в следующем параграфе.

Приведем таблицу абсцисс и весов квадратурной формулы Гаусса для случая $\rho(x)\equiv 1$, т. е. для весовой функции, соответствующей многочленам Лежандра (табл. 2).

Рассмотрим теперь, как решается рассмотренная выше задача построения квадратурной формулы наивысшей алгебраической точности, когда классы π_n многочленов n-й степени заменяются классами тригонометрических многочленов n-й степени.

Пусть $f(x)$ есть произвольная периодическая функция. Ее период можно считать приведенным к 2π. Рассмотрим интеграл $\int_0^{2\pi} f(x)\,dx$ и для

Таблица 1

№ п/п	Интегралы	Классические ортонормированные многочлены	Абсциссы	Узлы
1	$\int_{-1}^{1} f(x)dx,$ $\rho = 1$	Лежандра $P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2-1)^n$	Совпадают с нулями многочлена $P_n(x)$	$C_k^{(n)} = \frac{2}{n^2} \frac{1-(x_k^{(n)})^2}{P_{n-1}^2(x_k^{(n)})}$
2	$\int_{-1}^{1} f(x)(1-x)^\alpha (1+x)^\beta dx,$ $\rho(x) = (1-x)^\alpha (1+x)^\beta,$ $\alpha, \beta > -1$	Якоби $P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha} \times$ $\times (1+x)^{-\beta} \frac{d^n}{dx^n}\left[(1-x)^{\alpha+n} \times \right.$ $\left. \times (1+x)^{\beta+n}\right]$	x_k^n совпадают с нулями многочлена $P_n^{(\alpha,\beta)}(x)$	$C_k^{(n)} = \frac{2^{\alpha+\beta+1}\Gamma(\alpha+n+1)}{n!\,\Gamma(\alpha+\beta+n+1)} \times$ $\times \frac{\Gamma(\beta+n+1)}{(1-x_k^2)\left[P_n^{(\alpha,\beta)\prime}(x_k^{(n)})\right]^2}$
3	$\int_{-1}^{1} f(x)(1-x^2)^{-\frac{1}{2}} dx,$ $\rho(x) = (1-x^2)^{-\frac{1}{2}}$	Чебышева 1-го рода $P_n\left(-\frac{1}{2}, -\frac{1}{2}\right)(x) =$ $= C_n \cos(n \arccos x)$	$x_k^{(n)} = \cos\frac{2k-1}{2n}\pi,$ $k = 1, 2, \ldots, n$	$C_k^{(n)} = \frac{\pi}{n}$
4	$\int_{0}^{\infty} x^\alpha e^{-x} f(x)\,dx,$ $\rho(x) = x^\alpha e^{-x},\ \alpha > -1$	Лагерра $L_n^\alpha(x) = (-1)^n x^{-\alpha} e^x \frac{d^n}{dx^n} \times$ $\times (x^{\alpha+n} e^{-x})$	$x_k^{(n)}$ совпадают с нулями многочлена $L_n^\alpha(x)$	$C_k^{(n)} = \frac{\Gamma(n+1)\Gamma(n+\alpha+1)}{x_k^{(n)}\left[L_n^{\alpha\prime}(x_k^{(n)})\right]^2}$
5	$\int_{-\infty}^{\infty} e^{-x^2} f(x)\,dx,$ $\rho(x) = e^{-x^2}$	Эрмита $H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}$	$x_k^{(n)}$ совпадают с нулями многочлена $H_n(x)$	$C_k^{(n)} = \frac{2^{n+1} n! \sqrt{\pi}}{\left[H_n'(x_k^{(n)})\right]^2}$

Таблица 2

№ п/п	Абсциссы	Веса
$n=1$	$x_1 = 0$	$C_1^1 = 1$
$n=2$	$-x_1 = x_2 = 0{,}577\,350\,269\,189\,625\,8$	$C_1^{(2)} = C_2^{(2)} = \frac{1}{2}$
$n=3$	$-x_1 = x_3 = 0{,}774\,596\,669\,241\,483\,4$	$C_1^{(3)} = C_3^{(3)} = \frac{5}{18}$,
	$x_2 = 0$	$C_2^{(3)} = \frac{4}{9}$
$n=4$	$-x_1 = x_4 = 0{,}861\,136\,311\,594\,049\,2$	$C_1^{(4)} = C_4^{(4)} = 0{,}173\,927\,422\,568\,728\,4$
	$-x_2 = x_3 = 0{,}339\,981\,043\,584\,864\,6$	$C_2^{(4)} = C_3^{(4)} = 0{,}326\,072\,577\,431\,271\,6$
$n=5$	$-x_1 = x_5 = 0{,}906\,179\,845\,938\,664\,0$	$C_1^{(5)} = C_5^{(5)} = 0{,}118\,463\,442\,528\,094\,5$
	$-x_2 = x_4 = 0{,}538\,469\,310\,105\,683\,0$	$C_2^{(5)} = C_4^{(5)} = 0{,}239\,314\,335\,249\,683\,2$
	$x_3 = 0$	$C_3^{(5)} = \frac{64}{225} = 0{,}284\,444\,444\,444\,444\,4$
$n=6$	$-x_1 = x_6 = 0{,}932\,469\,514\,203\,152\,0$	$C_1^{(6)} = C_6^{(6)} = 0{,}085\,662\,246\,189\,585\,2$
	$-x_2 = x_5 = 0{,}661\,209\,386\,466\,264\,4$	$C_2^{(6)} = C_5^{(6)} = 0{,}180\,380\,786\,524\,069\,3$
	$-x_3 = x_4 = 0{,}238\,619\,186\,083\,197\,0$	$C_3^{(6)} = C_4^{(6)} = 0{,}233\,956\,967\,286\,345\,5$
$n=7$	$-x_1 = x_7 = 0{,}949\,107\,912\,342\,759\,6$	$C_1^{(7)} = C_7^{(7)} = 0{,}064\,742\,483\,084\,434\,8$
	$-x_2 = x_6 = 0{,}741\,531\,185\,599\,394\,4$	$C_2^{(7)} = C_6^{(7)} = 0{,}139\,852\,695\,744\,638\,4$
	$-x_3 = x_5 = 0{,}405\,845\,151\,377\,397\,0$	$C_3^{(7)} = C_5^{(7)} = 0{,}190\,915\,025\,252\,559\,5$
	$x_4 = 0$	$C_4^{(7)} = \frac{256}{1225} = 0{,}208\,979\,591\,836\,734\,7$

его вычисления построим формулу вида

$$\int_0^{2\pi} f(x)\,dx \approx \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}), \quad 0 \leqslant x_k^{(n)} \leqslant 2\pi, \qquad (k = 1, 2, \ldots, n). \tag{11}$$

Веса $C_k^{(n)}$ и узлы $x_k^{(n)}$ квадратурной формулы (11) будем выбирать из условия, чтобы формула (11) была точной для тригонометрического многочлена

$$T_m(x) = a_0 + \sum_{k=1}^{m} (a_k \sin kx + b_k \cos kx) \tag{12}$$

максимально высокой степени. Можно проверить, что для любых $C_k^{(n)}$ и $x_k^{(n)}$ формула (11) не может быть точной для всех тригонометрических многочленов степени n. Чтобы показать это, возьмем функцию

$$T_n(x) = \prod_{k=1}^{n} \sin^2 \frac{x - x_k^{(n)}}{2}, \tag{13}$$

которая, как не трудно видеть, будет тригонометрическим многочленом n-й степени. Но для него формула (11) не может быть точной, так как $\int\limits_0^{2\pi} T_n(x)\,dx > 0$, а $\sum\limits_{k=1}^{n} C_k^{(n)} T_n(x_k^{(n)}) = 0$. Тригонометрическая степень точности квадратурной формулы (11) всегда меньше n и при помощи выбора $C_k^{(n)}$ и $x_k^{(n)}$ можно сделать ее самое большее равной $n-1$. Как оказывается, наибольшая степень точности $n-1$ достигается квадратурной формулой с равными весовыми коэффициентами и равноотстоящими узлами.

Пусть α — любое число, удовлетворяющее неравенству $0 \leqslant \alpha < h = \frac{2\pi}{n}$. За узлы $x_k^{(n)}$ примем точки $x_k^{(n)} = \alpha + ih$, $i = 0, 1, \ldots, n-1$. Тогда квадратурная формула (11) имеет вид

$$\int\limits_0^{2\pi} f(x)\,dx \approx C \sum_{k=1}^{n} f\left[\alpha + (k-1)\frac{2\pi}{n}\right]. \tag{14}$$

Значение постоянной C находим из условия, чтобы формула (14) была точной для функции $f(x) \equiv 1$, откуда $C = \frac{2\pi}{n}$. Убедимся в том, что квадратурная формула (14) будет точной для любого тригонометрического многочлена степени $n-1$. Поскольку любой тригонометрический многочлен степени $n-1$ на основании формул Эйлера есть линейная комбинация системы функций Чебышева e^{imx}, $m = 0, 1, \ldots, n-1$, на интервале $(0, 2\pi]$, достаточно проверить формулу (14) на системе функций

$$e^{imx}, \quad m = 0, 1, 2, \ldots, n-1.$$

Имеем, с одной стороны,

$$\int\limits_0^{2\pi} e^{imx}dx = \frac{e^{imx}}{im}\Bigg|_0^{2\pi} = 0, \qquad m = 1, 2, \ldots, n-1,$$

с другой стороны,

$$\frac{2\pi}{n}\sum_{k=1}^{n} e^{im\left[\alpha+(k-1)\frac{2\pi}{n}\right]} = \frac{2\pi}{n} e^{im\alpha} \frac{e^{imn\frac{2\pi}{n}} - 1}{e^{im\frac{2\pi}{n}} - 1} = 0.$$

Следовательно, учитывая то, что формула (14) точна при $n = 1$ (т. е. для функции $f(x) \equiv 1$), доказано точное выполнение (14) для любого тригонометрического многочлена.

§ 3. ФОРМУЛЫ ЧЕБЫШЕВА

В этом параграфе изучим квадратурные формулы вида

$$\int\limits_{-1}^{1} f(x)\,\rho(x)\,dx = C\sum_{k=1}^{n} f(x_k^{(n)}) + R(f), \quad \rho(x) > 0, \tag{1}$$

определяя весовой множитель C и узлы $x_k^{(n)} \in [-1, 1]$, $k = 1, 2, \ldots, n$, из условия, чтобы

$$R(f) = 0 \quad \forall f \in \bigcup_{i=0}^{m} \pi_i \tag{2}$$

для максимально большого m. Поскольку (1) содержит $n+1$ параметр: C, $x_k^{(n)}$, $k = 1, 2, \ldots, n$, то $m \geqslant n$.

Применение квадратурных формул вида (1) является наиболее целесообразным тогда, когда значения $f(x_k^{(n)})$ находятся путем измерений и содержат случайные ошибки. Предположим, что эти случайные ошибки независимы с одинаковыми дисперсией σ и математическим ожиданием, равным нулю. Тогда дисперсия приближенного значения интеграла $\int_{-1}^{1} f(x)\rho(x)\,dx$, вычисляемого по формуле

$$\int_{-1}^{1} f(x)\rho(x)\,dx \approx \sum_{i=1}^{n} C_i^{(n)} f(x_i^{(n)}), \tag{3}$$

будет равна

$$D\left(\sum_{i=1}^{n} C_i^{(n)} f(x_i^{(n)})\right) = \sum_{i=1}^{n} [C_i^{(n)}]^2 Df(x_i^{(n)}) = \sigma \sum_{i=1}^{n} [C_i^{(n)}]^2. \tag{4}$$

Потребовав, чтобы формула (3) была точной для $f(x) = \text{const}$, получаем

$$\sum_{i=1}^{n} C_i^{(n)} = \int_{-1}^{1} \rho(x)\,dx = \mu_0. \tag{5}$$

Нетрудно видеть, что минимум правой части (4) (минимум дисперсии) при условии (5) достигается при

$$C_i^{(n)} = \frac{\mu_0}{n}, \qquad i = 1, 2, \ldots, n, \tag{6}$$

т. е. для квадратурных формул вида (1).

Ввиду того что формула (1) должна быть квадратурной формулой интерполяционного типа, то согласно формуле (2), § 2, должно выполняться условие

$$C = \int_{-1}^{1} \rho(x) \frac{\omega(x)}{(x - x_k^{(n)})\,\omega'(x_k^{(n)})}\,dx, \qquad k = 1, 2, \ldots, n. \tag{7}$$

Осталось выяснить, существует ли для заданного $\rho(x)$ такое распределение абсцисс $x_k^{(n)} \in [-1, 1]$, $k = 1, 2, \ldots, n$, для которого квадратурная формула (1) обладает свойством (2) при $m \geqslant n$ и свойством (7).

То, что существует хотя бы одна весовая функция $\rho(x)$, для которой ответ на поставленный выше вопрос — положительный, следует из § 2, табл. 1. Такой весовой функцией является

$$\rho(x) = (1 - x^2)^{-\frac{1}{2}}.$$

Рассмотрим теперь общий случай.

Задачу отыскания абсцисс $x_k^{(n)}$, $k = 1, 2, \ldots, n$, заменяем задачей отыскания многочлена

$$\omega(x) = \prod_{k=1}^{n} (x - x_k^{(n)}) = \sum_{k=0}^{n} b_k x^{n-k}, \quad b_0 = 1. \tag{8}$$

Так как для формулы (1) должно выполняться условие (2) при $m = n$, то отсюда приходим к системе

$$\sum_{k=1}^{n} [x_k^{(n)}]^i = \int_{-1}^{1} \rho(x)\, x^i dx = \mu_i, \qquad i = 0, 1, \ldots, n. \tag{9}$$

Левые части системы (9) являются симметричными функциями узлов $x_k^{(n)}$, но через эти симметричные функции выражаются и коэффициенты производной от многочлена $\omega(x)$. Действительно,

$$\omega'(x) = \sum_{i=1}^{n} \frac{\omega(x)}{x - x_i^{(n)}}; \tag{10}$$

$$\omega'(x) = nx^{n-1} + b_1(n-1)x^{n-2} + \cdots + b_{n-1}. \tag{11}$$

Далее, по схеме Горнера имеем:

$$\frac{\omega(x)}{x - x_i^{(n)}} = x^{n-1} + (b_1 + x_i^{(n)})x^{n-2} + (b_2 + b_1 x_i^{(n)} + (x_i^{(n)})^2)x^{n-3} + \cdots$$

$$\cdots + (b_{n-1} + b_{n-2}x_i^{(n)} + \cdots + (x_i^{(n)})^{n-1}). \tag{12}$$

Подставляя (12) в (10) и сравнивая с (11), получаем следующую систему для определения коэффициентов b_i:

$$\sum_{k=0}^{i} b_{i-k}\mu_k = (n - i)\, b_i, \qquad i = 1, 2, \ldots, n-1, \tag{13}$$

из которой последовательно находим все b_i, $i = 1, 2, \ldots, n+1$ (система (13) имеет нижнюю треугольную матрицу). Для определения b_n добавляем к (13) еще одно уравнение

$$\sum_{k=1}^{n} \omega(x_k^{(n)}) = 0 = \sum_{i=0}^{n} b_i \mu_{n-i}. \tag{14}$$

Итак, показано, что для каждого n можно найти такой многочлен $\omega(x)$, что приняв его корни за абсциссы квадратурной формулы численного интегрирования, все весовые коэффициенты этой формулы будут равны величине (6).

Пример 1. Рассмотрим частный случай формул Чебышева, когда $\rho(x) \equiv 1$. Тогда согласно формулам (6) и (9) имеем:

$$C = \frac{2}{n}; \quad \mu_i = \int_{-1}^{1} x^i dx = \frac{1 - (-1)^{i+1}}{i+1}, \qquad i = 0, 1, \ldots, n. \tag{15}$$

Система уравнений (13), (14) для определения коэффициентов многочлена $\omega(x)$ принимает вид

$$\sum_{k=0}^{i} \frac{1 - (-1)^{k+1}}{k+1} b_{i-k} = (n - i)\, b_i, \qquad i = 1, 2, \ldots, n. \tag{16}$$

Из системы (16) вытекает, что все b_l с нечетными индексами равны нулю и многочлен $\omega(x)$ приобретает вид

$$\omega(x) = x^n + b_2 x^{n-2} + \cdots + b_{2\left[\frac{n}{2}\right]} x^{2\left\{\frac{n}{2}\right\}}, \tag{17}$$

где [ ] и { } обозначают соответственно целую и дробную часть числа. При нечетном n один нуль многочлена $\omega(x)$ равен 0, а остальные размещены симметрично относительно начала координат. Следовательно, при нечетном n, кроме системы (16), будет выполняться равенство

$$\sum_{k=1}^{n} (x_k^n)^{n+1} = 0,$$

т. е. формулы Чебышева будут точными для многочленов степени $n+1$, а не только степени n, как это предполагалось ранее.

Таблица 3

n	Абсциссы
1	$x_1 = 0$
2	$-x_1 = x_2 = 0{,}577\,350\,269\,1$
3	$-x_1 = x_3 = 0{,}707\,106\,781\,2$ $x_2 = 0$
4	$-x_2 = x_3 = 0{,}187\,592\,474\,1$ $-x_1 = x_4 = 0{,}794\,654\,472\,3$
5	$x_3 = 0$ $-x_2 = x_4 = 0{,}374\,541\,409\,6$ $-x_1 = x_5 = 0{,}832\,497\,487\,0$
6	$-x_3 = x_4 = 0{,}266\,635\,401\,5$ $-x_2 = x_5 = 0{,}422\,518\,653\,8$ $-x_1 = x_6 = 0{,}866\,246\,818\,1$
7	$x_4 = 0$ $-x_3 = x_5 = 0{,}323\,911\,810\,5$ $-x_2 = x_6 = 0{,}529\,656\,775\,3$ $-x_1 = x_7 = 0{,}883\,861\,700\,8$
9	$x_5 = 0$ $-x_4 = x_6 = 0{,}167\,906\,184\,2$ $-x_3 = x_7 = 0{,}528\,761\,783\,1$ $-x_2 = x_8 = 0{,}601\,018\,655\,4$ $-x_1 = x_9 = 0{,}911\,589\,300\,7$

Приведем таблицу узлов формулы Чебышева (табл. 3) для

$$n = 1\,(1)\,7,\ 9.$$

При $n = 8$, как показали вычисления, среди $x_k^{(n)}$ будут два комплексных числа.

Тот факт, что при $n = 8$ в квадратуре Чебышева для весовой функции $\rho(x) \equiv 1$ оказались комплексные узлы, является не случайным фактом. Проведенные расчеты для $n > 9$ показали, что всякий раз обязательно появляются комплексные узлы. В общем виде вопрос о возможности или невозможности построения квадратуры Чебышева с действительными узлами для $n > 9$ был решен в 30-х годах академиком С. Н. Бернштейном. Перед тем как привести формулировку его теоремы, введем следующее определение:

Определение 1. Если система уравнений

$$\frac{1}{n}\sum_{i=1}^{n} (x_i^{(n)})^k = \int_{-1}^{1} x^k \rho(x)\,dx, \qquad k = 1,\ 2,\ \ldots,\ n, \tag{18}$$

имеет действительные решения для всех натуральных n, то весовую функцию $\rho(x)$ будем называть *весом, допускающим квадратуру Чебышева.*

Теорема 1 (С. Н. Бернштейна). *При $n \geqslant 10$ в квадратуре Чебышева* (1) *для $\rho(x) \equiv 1$ среди абсцисс $x_k^{(n)}$ есть комплексные, т. е. вес $\rho(x) \equiv 1$ не допускает квадратуры Чебышева.*

До недавнего времени все результаты по отысканию весовых функций, отличных от $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ и допускающих квадратуру Чебышева,

были отрицательны, т. е. для исследуемых весов показывалось, что система (18), начиная с некоторого n, имеет комплексные решения. В 1966 году американским математиком И. Л. Алменом был получен первый положительный результат, более того, им было доказано существование целого семейства весовых функций, допускающих квадратуру Чебышева. Точнее, им была установлена теорема.

Теорема 2. *Если* $|a| \leqslant \frac{1}{4}$, *то весовая функция*

$$\rho(x) = \frac{1}{\pi\sqrt{1-x^2}} \cdot \frac{1+2ax}{1+4a^2+4ax} \tag{19}$$

допускает квадратуру Чебышева.

§ 4. КВАДРАТУРНЫЕ ФОРМУЛЫ С ИСПОЛЬЗОВАНИЕМ ПРОИЗВОДНЫХ ОТ ПОДЫНТЕГРАЛЬНОЙ ФУНКЦИИ

Для построения квадратурных формул с использованием значений производных в узлах можно поступать так же, как и при выводе формул Ньютона — Котеса (§ 1), только вместо интерполяционного многочлена Лагранжа использовать, например, интерполяционный многочлен Эрмита.

Так, если наложить условие, чтобы в квадратурной формуле

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} \sum_{j=0}^{\alpha_i-1} A_{i,j}^{(n)} f^{(j)}(x_i^{(n)}) + R(f), \tag{1}$$

$R(f) = 0$, $\forall f \in \bigcup_{k=0}^{m} \pi_k$, $m = \sum_{i=0}^{n} \alpha_i - 1$, то легко видеть, что веса $A_{ij}^{(n)}$ определяются при помощи фундаментальных многочленов Эрмита $H_{ij}(x)$ (см. § 1, гл. 2) по формулам

$$A_{ij}^{(n)} = \sum_{k=0}^{\alpha_i-j-1} \frac{1}{k!}\frac{1}{j!}\left[\frac{(x-x_i^{(n)})^{\alpha_i}}{\Omega(x)}\right]_{x=x_i}^{(k)} \int_a^b \frac{\Omega(x)}{(x-x_i)^{\alpha_i-j-k}}\,dx. \tag{2}$$

Однако интерполяционная формула Эрмита оказывается непригодной, если при построении квадратурных формул с использованием производных в узлах $x_i^{(n)}$ известны не все производные подряд до определенного порядка, а с пропусками, т. е. если вместо (1) строить формулу вида

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} \sum_{j=0}^{\alpha_i-1} A_{i,j}^{(n)} f^{(\beta_{ij})}(x_i^{(n)}) + R(f), \tag{3}$$

где $\beta_{i,0} < \beta_{i,1} < \cdots < \beta_{i,\alpha_i-1}$, $i = 0, 1, \ldots, n$.

Несмотря на различие квадратурных формул (1) и (3), покажем, что и для последнего случая возможен подход, опирающийся на теорию интерполирования. Во избежание очень громоздких выкладок

ограничимся случаем, при котором $n = 1$; $x_0 = a$; $x_1 = b$;

$$\beta_{0,j} = \beta_{1,j} = 2j - 1; \quad \alpha_0 = \alpha_1 = r + 1; \quad \beta_{0,0} = \beta_{1,0} = 0, \tag{3'}$$

т. е. $\int\limits_a^b f(x)\,dx = A_{0,0}f(a) + A_{1,0}f(b) + \sum\limits_{j=1}^{r} [A_{0,j}f^{(2j-1)}(a) + A_{1,j}f^{(2j-1)}(b)] +$

$$+ R(f). \tag{4}$$

Поскольку

$$\int\limits_a^b f(x)\,dx = \int\limits_a^b f(a + b - x)\,dx, \tag{5}$$

то из (4) и (5) следует

$$\int\limits_a^b f(x)\,dx - \int\limits_a^b f(a + b - x)\,dx = 0 =$$

$$= (A_{0,0} - A_{1,0})[f(a) - f(b)] + \sum_{j=1}^{r} (A_{0,j} + A_{1,j})[f^{(2j-1)}(a) + f^{(2j-1)}(b)]. \tag{6}$$

Ввиду того что (6) должно иметь место для всех $f(x)$ из класса достаточно гладких функций, получаем

$$A_{0,0} = A_{1,0}; \quad A_{0,j} = -A_{1,j}, \qquad j = 1, 2, \ldots, r,$$

и искомая квадратурная формула (4) принимает вид

$$\int\limits_a^b f(x)\,dx = \int\limits_a^b \frac{1}{2}[f(x) + f(a + b - x)]\,dx =$$

$$= A_0[f(a) + f(b)] + \sum_{j=1}^{r} A_j[f^{(2j-1)}(a) + f^{(2j-1)}(b)] + R(f). \tag{7}$$

Введем функцию

$$\Phi(x) = \frac{1}{2}[f(a) + f(a + b - x)]. \tag{8}$$

Она обладает следующим свойством:

$$\begin{gathered} \Phi(a) = \Phi(b) = \frac{1}{2}[f(a) + f(b)]; \\ \Phi^{(2j-1)}(a) = -\Phi^{(2j-1)}(b) = \frac{1}{2}[f^{(2j-1)}(a) - f^{(2j-1)}(b)], \\ j = 1, 2, \ldots, r. \end{gathered} \tag{9}$$

Построим для функции $\Phi(x)$ обобщенный интерполяционный многочлен Эрмита $P_{2r}(x)$, который удовлетворяет условию (9):

$$P_{2r}(x) = \frac{1}{2}[f(a) + f(b)]\,p_0(x) + \sum_{j=1}^{r} \frac{1}{2}[f^{(2j-1)}(a) - f^{(2j-1)}(b)]\,p_j(x), \tag{10}$$

где для многочленов $p_j(x)$, $j = 0, 1, \ldots, r$, должны выполняться соотношения:

$$p_0(x_k) = 1; \quad p_j^{(2s-1)}(x_k) = (-1)^{k+1}\delta_{s,j}; \tag{11}$$
$$j = 0, 1, \ldots, r, \quad s = 1, 2, \ldots, r, \quad k = 1, 2.$$

Здесь $x_1 = a$; $x_2 = b$; $\delta_{s,j}$ — символ Кронекера.

Воспользовавшись свойствами чисел и многочленов Бернулли (приложение, § 2):

$$B_k(1) = B_k(0) = B_k;$$
$$B_{2n+1}(0) = B_{2n+1} = 0, \qquad n = 1, 2, \ldots; \tag{12}$$
$$B_n'(x) = nB_{n-1}(x),$$

легко видеть, что

$$p_0(x) = 1;$$
$$p_j(x) = -2\frac{(b-a)^{2j-1}}{(2j)!}\left[B_{2j}\left(\frac{b-x}{b-a}\right) - B_{2j}\right], \quad j = 1, 2, \ldots, r. \tag{13}$$

Заметим, что степень интерполяционного многочлена $P_{2r}(x)$ не может быть меньше $2r$, так как в противном случае не будет выполняться условие

$$P_{2r}^{(2r-1)}(a) = -P_{2r}^{(2r-1)}(b) = \frac{1}{2}[f^{(2r-1)}(a) - f^{(2r-1)}(b)] \neq 0.$$

Для нахождения остаточного члена интерполяционной формулы

$$\Phi(x) = P_{2r}(x) + R_{2r}(x) \tag{14}$$

нам понадобится следующая теорема:

Теорема 1. *Пусть* $\Phi(x) \in C^{(2r+2)}[a, b]$ *и обладает свойством*

$$\Phi(a) = \Phi(b) = \varphi_0, \ \Phi^{(2j-1)}(a) = -\Phi^{(2j-1)}(b) = \varphi_j, \ j = 1, 2, \ldots, r$$

и пусть $P_{2r}(x)$ *— ее обобщенный интерполяционный многочлен Эрмита:*

$$P_{2r}(a) = P_{2r}(b) = \varphi_0;$$
$$P_{2r}^{(2j-1)}(a) = -P_{2r}^{(2j-1)}(b) = \varphi_j, \qquad j = 1, 2, \ldots, r. \tag{15}$$

Если существует функция $g(x) \in C^{(2r+2)}[a, b]$ *такая, что*

$$g(a) = g(b) = 0; \quad g^{(2j-1)}(a) = g^{(2j-1)}(b) = 0, \quad j = 1, 2, \ldots, r;$$
$$g^{(2r+2)}(x) \neq 0, \quad \forall x \in [a, b], \tag{16}$$

то остаточный член в интерполяционной формуле (14) *может быть представлен в виде*

$$R_{2r}(x) = \Phi(x) - P_{2r}(x) = \frac{\Phi^{(2r+2)}(\xi)}{g^{(2r+2)}(\xi)}g(x), \quad \xi \in (a, b), \quad \forall x \in (a, b). \tag{17}$$

Доказательство почти повторяет доказательство теоремы 1, § 3, гл. 2. Введем в рассмотрение функцию

$$F(t) = \Phi(t) - P_{2r}(t) - \lambda g(t),$$

которая по построению принадлежит $C^{(2r+2)}[a, b]$ и удовлетворяет тем же условиям (16), что и функция $g(x)$. Выберем параметр λ из условия $F(x) = 0$. Для этого положим

$$\lambda = \frac{\Phi(x) - P_{2r}(x)}{g(x)}.$$

Тогда, применяя последовательно теорему Ролля и используя условия (15) и (16), получим:

1) $F(a) = F(b) = F(x) = 0$, следовательно, существуют точки $\xi_1^{(1)} = a < \xi_2^{(1)} < \xi_3^{(1)} < \xi_4^{(1)} = b$ такие, что $F'(\xi_i^{(1)}) = 0$, $i = \overline{1, 4}$;

2) существуют точки $\xi_1^{(2)} = a < \xi_2^{(2)} < \xi_3^{(2)} < \xi_4^{(2)} = b$ такие, что $F''(\xi_i^{(2)}) = 0$, $i = \overline{1, 4}$, и т. д.;

r) существуют точки $\xi_1^{(r)} = a < \xi_2^{(r)} < \xi_3^{(r)} < \xi_4^{(r)} = b$ такие, что $F^{(2r-1)}(\xi_i^{(r)}) = 0$, $i = \overline{1, 4}$. Отсюда заключаем, что существует такая точка ξ, для которой $F^{(2r+2)}(\xi) = 0$; $\xi \in (a, b)$.

Теорема доказана.

Следствие 1. В качестве функции $g(x)$, о которой шла речь в теореме 1, можно взять многочлен вида

$$g(x) = \Omega_{2r}(x) = (b-a)^{2r+2}\left[B_{2r+2}\left(\frac{x-a}{b-a}\right) - B_{2r+2}\right] \tag{18}$$

и формула (17) для остаточного члена запишется следующим образом:

$$R_{2r}(x) = \Phi(x) - P_{2r}(x) =$$
$$= \frac{\Phi^{(2r+2)}(\xi)}{(2r+2)!}(b-a)^{2r+2}\left[B_{2r+2}\left(\frac{x-a}{b-a}\right) - B_{2r+2}\right]. \tag{19}$$

Это утверждение непосредственно вытекает из свойств чисел и многочленов Бернулли (12).

Лемма 1. *Для каждой функции $\Phi(x)$, удовлетворяющей условиям теоремы 1, существует единственный обобщенный интерполяционный многочлен Эрмита в смысле* (15).

Доказательство леммы оставляем в качестве упражнения для читателя.

Вернемся теперь к построению квадратурной формулы (7), наложив условия, чтобы

$$R(f) = 0 \qquad \forall f \in \bigcup_{k=0}^{2r+1} \pi_k.$$

Из последнего условия и из формул (11), (13) получаем выражения для весов:

$$A_0 = \frac{b-a}{2};$$

$$A_j = \frac{1}{2}\int_a^b p_j(x)\,dx = -\frac{(b-a)^{2j-1}}{(2j)!}\int_a^b\left[B_{2j}\left(\frac{b-x}{b-a}\right) - B_{2j}\right]dx =$$
$$= \frac{(b-a)^{2j}}{(2j)!}B_{2j}, \qquad j = 1, 2, \ldots, r, \tag{20}$$

и квадратурная формула (7) будет иметь вид

$$\int_a^b f(x)\,dx = \int_a^b \Phi(x)\,dx = \frac{b-a}{2}[f(a)+f(b)] +$$

$$+\sum_{j=1}^{r} \frac{(b-a)^{2j}}{(2j)!} B_{2j}[f^{(2j-1)}(a) - f^{(2j-1)}(b)] + R(f), \tag{21}$$

где согласно (8) и (19)

$$R(f) = \frac{(b-a)^{2r+2}}{(2r+2)!} \int_a^b \frac{1}{2}[f^{(2n+2)}(\xi) + f^{(2n+2)}(a+b-\xi)] \times$$

$$\times \left[B_{2r+2}\left(\frac{x-a}{b-a}\right) - B_{2r+2}\right] dx, \qquad \xi = \xi(x) \in (a, b). \tag{22}$$

При получении формул (20) было использовано соотношение

$$\int_a^b B_{2j}\left(\frac{b-x}{b-a}\right) dx = -\frac{b-a}{2j+1} \int_a^b \frac{d}{dx} B_{2j+1}\left(\frac{b-x}{b-a}\right) dx =$$

$$= -\frac{b-a}{2j+1}[B_{2j+1}(0) - B_{2j+1}(1)] = 0,$$

что следует из (12).

Если положить в формуле (21) $a = c + ih$; $b = c+(i+1)h$ и просуммировать ее по i от 0 до $n-1$, то в результате получим *формулу Эйлера*

$$\int_c^{c+nh} f(x)\,dx = h\left[\frac{1}{2}f(c) + \sum_{i=1}^{n-1} f(c+ih) + \frac{1}{2}f(c+nh)\right] + \tag{23}$$

$$+\sum_{j=1}^{r} \frac{h^{2j}B_{2j}}{(2j)!}[f^{(2j-1)}(c) - f^{(2j-1)}(c+nh)] + R_{2r},$$

где

$$R_{2r} = \frac{h^{2r+2}}{(2r+2)!} \sum_{i=1}^{n} \int_{c+(i-1)h}^{c+ih} \frac{1}{2}[f^{(2n+2)}(\xi_i) + f^{(2n+2)}(2c+(2i-1)h-\xi_i)] \times$$

$$\times \left[B_{2r+2}\left(\frac{x-c-(i-1)h}{h}\right) - B_{2r+2}\right] dx, \tag{24}$$

где $\xi_i \in (c+(i-1)h,\ c+ih)$, $i = 1, 2, \ldots, n$,— некоторые средние точки.

Отметим, что первое слагаемое в правой части формулы (23) представляет собой формулу трапеций. В этом смысле на остальные слагаемые можно смотреть как на поправку к формуле трапеций.

Кроме непосредственного своего назначения формула Эйлера применяется для нахождения сумм значений функции в равноотстоящих точках.

§ 5. ОСТАТОЧНЫЙ ЧЛЕН КВАДРАТУРНЫХ ФОРМУЛ

Все квадратурные формулы, рассмотренные ранее, являются квадратурными формулами интерполяционного типа и поэтому остаточный член их самым тесным образом связан с остаточным членом соответствующей формулы интерполирования.

Будем исследовать остаточный член квадратурной формулы интерполяционного типа

$$\int_a^b f(x)\rho(x)\,dx = \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) + R(f), \quad \rho(x) \geqslant 0, \tag{1}$$

алгебраический порядок точности которой $m \geqslant n-1$. Тогда, заменяя функцию $f(x)$ интерполяционным многочленом m-й степени, построенным по узлам $x_k^{(n)}$, $k = 1, 2, \ldots, n$, для $R(f)$ получим выражение

$$R(f) = \int_a^b \rho(x) \frac{f^{(m+1)}(\xi)}{(m+1)!} \Omega(x)\,dx, \quad \xi \in (a, b), \tag{2}$$

если остаточный член интерполяционной формулы взят в форме Коши (см. § 3, гл. 2). При $m = n-1$ используем интерполяционный многочлен Лагранжа и $\Omega(x)$ в этом случае совпадает с $\omega_{n-1}(x) = \prod_{k=1}^{n} (x - x_k^{(n)})$, при $m > n-1$ — интерполяционный многочлен Эрмита и для этого случая $\Omega(x) = \prod_{k=1}^{n} (x - x_k^{(n)})^{\alpha_k}$, где $\sum_{k=1}^{n} \alpha_k = m$.

Изучать остаточный член по формуле (2) не совсем удобно и поэтому формулу (2) обычно приводят к виду

$$R(f) = f^{(m+1)}(\eta) B_m, \quad \eta \in (a, b), \tag{3}$$

где B_m — монотонная постоянная, не зависящая от $f(x)$. В том случае, когда $\Omega(x)$ — знакопостоянная функция, такое преобразование выполнить легко: достаточно к интегралу в (2) применить теорему о среднем, в результате чего получим

$$B_m = \frac{1}{(m+1)!} \int_a^b \rho(x)\Omega(x)\,dx. \tag{4}$$

Отсюда можно сделать такой вывод:

Для случая $m > n-1$ среди $x_k^{(n)}$, $k = 1, 2, \ldots, n$, следует выбирать кратные узлы и порядок кратности таким образом, чтобы многочлен $\Omega(x)$ был знакопостоянным на $[a, b]$, если это возможно.

Заметим, что если преобразование (2) в (3) существует, то постоянная B_m определяется формулой

$$B_m = \int_a^b \frac{(x-a)^{m+1}}{(m+1)!} \rho(x)\,dx - \sum_{k=1}^{n} C_k^{(n)} \frac{(x_k^{(n)} - a)^{m+1}}{(m+1)!}, \tag{5}$$

т. е. совпадает с остаточным членом квадратурной формулы (1) для функции $f(x) = \frac{(x-a)^{m+1}}{(m+1)!}$. Доказательство этого факта очевидно.

Остается открытым вопрос о существовании такого преобразования. Имеет место следующая лемма:

Лемма 1. *Пусть функция*

$$K_n(y) = \frac{1}{m!}\left[\int_y^b \rho(x)(x-y)^m\,dx - \sum_{i=1}^{n} C_i^{(n)}\left[(x_i^{(n)}-y)^+\right]^m\right], \tag{6}$$

где $(t)^+ = \frac{1}{2}(t+|t|)$, *сохраняет знак на* $[a, b]$. *Тогда будут иметь место формулы* (3), (4).

Доказательство. Запишем для функции $f(x)$ ряд Тейлора с остаточным членом в интегральной форме

$$f(x) = \sum_{j=0}^{m} \frac{f^{(j)}(a)}{j!}(x-a)^j + \int_a^x \frac{f^{(m+1)}(y)}{m!}(x-y)^m\,dy$$

или с использованием символа $(\ )^+$:

$$f(x) = \sum_{j=0}^{m} \frac{f^{(j)}(a)}{j!}(x-a)^j + \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x-y)^+\right]^m dy. \tag{7}$$

Применим к обеим частям (7) линейный оператор R. Тогда, принимая во внимание, что $R(f) = 0$, $\forall f \in \bigcup_{i=0}^{m} \pi_i$, получим

$$R(f) = R\left(\int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x-y)^+\right]^m dy\right). \tag{8}$$

Ввиду того что

$$\int_a^b \rho(x)\int_a^b \frac{f^{(m+1)}(y)}{m!}\left[(x-y)^+\right]^m dy\,dx = \int_a^b \frac{f^{(m+1)}(y)}{m!}\left[\int_y^b \rho(x)(x-y)^m\,dx\right]dy,$$

формула (8) приобретает вид

$$R(f) = \int_a^b K_n(y) f^{(m+1)}(y)\,dy. \tag{8'}$$

Учитывая условия леммы и применяя теорему о среднем к последнему интегралу, получаем формулы (3), (4), что и требовалось доказать.

Найдем теперь остаточные члены простейших формул Ньютона — Котеса (см. § 1).

Формула средних прямоугольников ($n = 1$, алгебраическая точность $m = 1$). Имеем

$$R(f) = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{f''(\xi)}{2!}\,dx = \frac{f''(\zeta)}{2!}\int_a^b \left(x - \frac{a+b}{2}\right)^2 dx =$$

$$= \frac{f''(\zeta)}{24}(b-a)^3. \tag{9}$$

Здесь $\Omega(x) = \left(x - \frac{a+b}{2}\right)^2 > 0$, $\forall x$ и использован интерполяционный многочлен Эрмита с одним двукратным узлом $x_1^{(2)} = x_2^{(2)} = \frac{a+b}{2}$.

Формула трапеций ($n = 2$, алгебраическая точность $m = 1$). Имеем

$$R(f) = \int_a^b (x-a)(x-b)\frac{f''(\xi)}{2!}\,dx = -\frac{f''(\zeta)}{12}(b-a)^3. \tag{10}$$

Здесь $\Omega(x) = (x-a)(x-b) < 0$, $\forall x \in (a, b)$ и использован интерполяционный многочлен Лагранжа с двумя узлами $x_1^{(2)} = a$; $x_2^{(2)} = b$.

Формула Симпсона ($n = 3$, алгебраическая точность $m = 3$). Имеем

$$R(f) = \int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\frac{f^{\mathrm{IV}}(\xi)}{4!}\,dx = -\frac{f^{\mathrm{IV}}(\zeta)}{2880}(b-a)^5. \tag{11}$$

Здесь $\Omega(x) = (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b) < 0$, $\forall x \in (a, b)$ и использован интерполяционный многочлен Эрмита с узлами $x_1^{(4)} = a$; $x_2^{(4)} = x_3^{(4)} = \frac{a+b}{2}$; $x_4^{(4)} = b$.

Правило трех восьмых ($n = 4$, алгебраическая точность $m = 3$). Имеем

$$R(f) = \int_a^b (x-a)\left(x - \frac{2a+b}{3}\right)\left(x - \frac{a+2b}{3}\right)(x-b)\frac{f^{\mathrm{IV}}(\xi)}{4!}\,d\xi =$$

$$= \int_a^b K_4(y) f^{\mathrm{IV}}(y)\,dy. \tag{12}$$

Здесь $\Omega(x) = (x-a)\left(x - \frac{2a+b}{3}\right)\left(x - \frac{a+2b}{3}\right)(x-b)$ не является знакопостоянной на $[a, b]$ и можно применять лемму 1. В данном случае, согласно (6)

$$K_4(y) = \frac{1}{4!}(b-y)^4 - \frac{1}{48}\left\{[(a-y)^+]^3 + 3\left[\left(\frac{2a+b}{3} - y\right)^+\right]^3 + \right.$$

$$\left. + 3\left[\left(\frac{a+2b}{3} - y\right)^+\right]^3 + [(b-y)^+]^3\right\}$$

и, разбив промежуток $[a, b]$ на три интервала $\left[a, \frac{2a+b}{3}\right]$; $\left[\frac{2a+b}{3}, \frac{a+2b}{3}\right]$; $\left[\frac{a+2b}{3}, b\right]$, нетрудно доказать, что в каждом из них $K_4(y) \leqslant 0$. Следовательно, будут иметь место формулы (3), (5), которые приводят к результату

$$R(f) = -\frac{f^{(\mathrm{IV})}(\zeta)}{6480}(b-a)^5, \tag{13}$$

который несколько лучше, чем для формулы Симпсона (11).

Для случая общих формул Ньютона — Котеса ((8), § 1) после несложных, но довольно громоздких вычислений, не связанных непосредственно с использованными выше приемами, можно показать, что имеют место формулы:

$$\int_c^d f(x)\,dx = (d-c)\sum_{i=1}^{2m-1} J_{i,k}^{(2m-1)} f(a+ih) - \frac{h^{2m+1} f^{(2m)}(\zeta)}{(2m)!}\int_{1-k}^{2m+k-1} \varphi(y)\,dy \tag{14}$$

$$n = 2m - 1;$$

$$\int_c^d f(x)\,dx = (d-c)\sum_{i=1}^{2m} J_{i,k}^{(2m)} f(a+ih) - \frac{h^{2m+1} f^{(2m)}(\zeta)}{2m!}\left[\int_{1-k}^{2m+k-1} \varphi(y)\,dy - \right.$$

$$\left. - \int_{2m+k-1}^{2m+k} (y-1)(y-2)\ldots(y-2m)\,dy\right], \qquad n = 2m, \tag{15}$$

где $\varphi(x) = \int_{1-k}^{x} (y-1)(y-2)\ldots(y-2m+1)\,dy$ и $n = \dfrac{d-c}{n+2k-1}$.

Рассмотрим вопрос об остаточном члене квадратурных формул наивысшей алгебраической степени точности (см. (1), (2), § 2).

Теорема 1. *Пусть весовая функция* $\rho(x)$ в (1), § 2, *неотрицательная на* $[a, b]$ *и* $f(x) \in C^{(2n)}[a, b]$, *тогда* $\exists \xi \in [a, b]$ *и такая, что для остаточного члена* $R(f)$ *квадратурной формулы* (1), § 2, *наивысшей алгебраической степени точности справедливо равенство*

$$R(f) = \frac{f^{(2n)}(\zeta)}{(2n)!}\int_a^b \rho(x)\,\omega^2(x)\,dx. \tag{16}$$

Доказательство. Возьмем интерполяционный многочлен Эрмита $H_{2n-1}(x)$, удовлетворяющий условиям:

$$H_{2n-1}^{(\mu)}(x_i^{(n)}) = f^{(\mu)}(x_i^{(n)}), \qquad \mu = 0, 1; \quad i = 1, 2, \ldots, n, \tag{17}$$

и подставим в (1), § 2, вместо функции $f(x)$ выражение

$$f(x) = H_{2n-1}(x) + \frac{f^{(2n)}(\xi)}{(2n)!}\,\Omega(x), \quad \xi \in [a, b], \tag{18}$$

где

$$\Omega(x) = \prod_{i=1}^{n} (x - x_i^{(n)})^2.$$

В результате получим

$$\int_a^b \rho(x) f(x)\,dx = \int_a^b \rho(x) H_{2n-1}(x)\,dx + \int_a^b \rho(x)\frac{f^{(2n)}(\xi)}{(2n)!}\,\Omega(x)\,dx =$$

$$= \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) + \int_a^b \rho(x)\frac{f^{(2n)}(\xi)}{(2n)!}\,\Omega(x)\,dx, \tag{19}$$

так как квадратурная формула наивысшей алгебраической степени точности является точной для многочлена степени меньшей или равной $2n - 1$. На основании теоремы о среднем из (19) получаем утверждение теоремы.

Приведем теперь к виду (3) остаточный член формулы Эйлера ((24), § 4). Из свойств чисел и многочленов Бернулли вытекает, что функция

$$B_{2r+2}(t) - B_{2r+2}$$

знакопостоянна на отрезке $[0, 1]$. Применяя к (24), § 4, теорему о среднем и используя непрерывность функции $f^{(2n+2)}(x)$ на $[a, b]$, получаем

$$R_{2r} = \frac{-h^{2r+3}n}{(2r+2)!}\,B_{2r+2} f^{(2n+2)}(\xi), \qquad \xi \in [c, d]. \tag{20}$$

По приведенным формулам для остаточных членов удобно производить их оценки, по которым можно судить о точности численного интегрирования.

Пусть $f(x) \in W^{(m+1)}(M;\ a,\ b)$, т. е. $f(x)$ принадлежит классу функций, непрерывно дифференцируемых до порядка m включительно и имеющих кусочно-непрерывную производную порядка $m+1$, удовлетворяющую неравенству

$$|f^{(m+1)}(x)| < M. \tag{21}$$

Тогда из (1) и (8) следует неравенство

$$\left| \int_a^b \rho(x) f(x)\, dx - \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) \right| \leqslant M \int_a^b \rho(x) |K_n(x)|\, dx = MN_n, \tag{22}$$

где постоянная $N_n = \int_a^b \rho(x) |K_n(x)|\, dx$ не зависит от $f(x)$ и может быть вычислена с любой степенью точности для каждой конкретной квадратурной формулы.

В неравенстве (22) правую часть уменьшить нельзя, так как для $\forall g(x) \in W^{(m+1)}(M;\ a,\ b)$ и такой, что $g^{(m+1)}(x) = M \operatorname{sign} K_n(x)$, неравенство (22) превращается в равенство, т. е.

$$\sup_{f \in W^{(m+1)}(M;a,b)} \left| \int_a^b f(x) \rho(x)\, dx - \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) \right| = MN_n. \tag{23}$$

§ 6. КВАДРАТУРНЫЕ ФОРМУЛЫ С НАИЛУЧШЕЙ ОЦЕНКОЙ ОСТАТОЧНОГО ЧЛЕНА НА КЛАССАХ ФУНКЦИЙ

В конце предыдущего параграфа была решена следующая задача: задан класс функций F, определенных на отрезке $[a,\ b]$, и множество квадратурных формул S, требовалось для конкретной формулы из S определить величину

$$\sup_{f \in F} |R(f)|,$$

где $R(f)$ — остаточный член квадратурной формулы. Естественно теперь поставить такую задачу: найти квадратурную формулу из S, на которой достигается

$$\inf_{S} \sup_{f \in F} |R(f)|.$$

Важность такой задачи вытекает из следующих соображений. При вычислении интеграла $\int_a^b f(x)\, dx$ по какой-либо квадратурной формуле основной объем работы падает на вычисление значений функции в узлах квадратурной формулы. Естественно при заданной точности вычисления интеграла выбрать ту формулу, которая требует минимального объема вычислительной работы. Целесообразно искать решение этой проблемы не для конкретной функции, а сразу для некоторого класса функций. При этом от квадратурной формулы нужно требовать, чтобы она имела наилучшую оценку остаточного члена на рассматриваемом классе функций по сравнению с другими квадратурными формулами из S, требующими примерно одинакового объема вычислительной работы. Это особенно важно для классов функций малой

гладкости, так как на таких классах функций наилучшими иногда оказываются простейшие квадратурные формулы.

Рассмотрим некоторые примеры решения задачи об отыскании квадратурной формулы с наилучшей оценкой остаточного члена на классе функций.

Возьмем сначала класс $C^{(1)}(L)$ функций $f(x)$, непрерывных вместе с их первыми производными на отрезке $[0, 1]$, для которых $|f'(x)| \leqslant L$, $L > 1$, и множество S квадратурных формул вида

$$\int_0^1 f(x)\,dx \approx \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}), \quad x_k^{(n)} \in [0, 1] \tag{1}$$

с фиксированным n, удовлетворяющих требованию, что они дают точное значение интеграла, если $f(x) = \text{const}$, т. е.

$$\sum_{k=1}^{n} C_k^{(n)} = 1. \tag{1'}$$

Точно так же, как это делалось в предыдущем параграфе, можно показать, что на классе функций $C^{(1)}(L)$ для остаточного члена $R_n(f)$ имеет место соотношение

$$\sup_{C^{(1)}(L)} |R_n(f)| = L\int_0^1 |K_n(t)|\,dt,$$

где

$$K_n(t) = t + \sum_{i=1}^{k} C_i^{(n)}, \quad t \in (x_k^{(n)}, x_{k+1}^{(n)}), \qquad k = 0, 1, \ldots, n. \tag{2}$$

Здесь $x_0^{(n)} = 0$, $x_{n+1}^{(n)} = 1$ и сумма считается равной 0 при $k = 0$. Итак, для того чтобы построить наилучшую квадратурную формулу вида (1) на классе функций $C^{(1)}(L)$, нужно $x_k^{(n)}$ и $C_k^{(n)}$ выбрать так, чтобы $\int_0^1 |K_n(t)|\,dt$ имел минимальное значение при $x_k^{(n)} \in (0, 1)$, $\sum_{k=1}^{n} C_k^{(n)} = 1$.

Полагая $\xi_k = \sum_{i=1}^{k} C_i^{(n)}$ и используя (2), имеем

$$\int_0^1 |K_n(t)|\,dt = \frac{(x_1^{(n)})^2}{2} + \sum_{k=1}^{n-1} \int_{x_k^{(n)}}^{x_{k+1}^{(n)}} |\xi_k - t|\,dt + \frac{(1 - x_n^{(n)})^2}{2}.$$

Можно показать (предоставляем это читателю), что

$$\int_{x_k^{(n)}}^{x_{k+1}^{(n)}} |\xi_k - t|\,dt \geqslant \frac{(x_{k+1}^{(n)} - x_k^{(n)})^2}{4}$$

и знак равенства достигается при $\xi_k = \frac{x_k^{(n)} + x_{k+1}^{(n)}}{2}$. Поэтому

$$\int_0^1 |K_n(t)|\,dt \geqslant \frac{1}{4}\left[2(x_1^{(n)})^2 + \sum_{k=1}^{n-1}(x_{k+1}^{(n)} - x_k^{(n)})^2 + 2(1 - x_n^{(n)})^2\right].$$

Если $0 \leqslant x_1^{(n)} < x_2^{(n)} < \cdots < x_n^{(n)} \leqslant 1$, то правая часть последнего неравенства достигает минимального значения $\frac{1}{4n}$ при $x_1^{(n)} = \frac{1}{2n}$; $x_{k+1}^{(n)} - x_k^{(n)} = 2x_1^{(n)} = \frac{1}{n}$ $(k = 1, 2, \ldots, n-1)$; $1 - x_n^{(n)} = x_1^{(n)} = \frac{1}{2n}$. Значит,

$$\int_0^1 |K_n(t)|\,dt \geqslant \frac{1}{4n}$$

и знак равенства имеет место при $x_k^{(n)} = \frac{2k-1}{n}$;

$$\xi_k = \frac{x_k^{(n)} + x_{k+1}^{(n)}}{2} = \frac{k}{n}; \quad C_k^{(n)} = \xi_k - \xi_{k-1} = \frac{1}{n}.$$

Таким образом, наилучшей на классе $C^{(1)}(L)$ квадратурной формулой вида (1) при условиях (1′) является формула средних прямоугольников

$$\int_0^1 f(x)\,dx \approx \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{2k-1}{2n}\right) \tag{3}$$

и оценка остаточного члена этой формулы на классе $C^{(1)}(L)$ будет иметь вид

$$\left|\int_0^1 f(x)\,dx - \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{2k-1}{2n}\right)\right| \leqslant \frac{M}{4n}. \tag{4}$$

Рассмотрим теперь класс $W_2^q(L)$ функций $f(x)$, непрерывных на отрезке $[0, 1]$ вместе со своими производными до порядка $(q-1)$ включительно и имеющих q-ю производную, интегрируемую с квадратом,

$$\int_0^1 [f^{(q)}(x)]^2\,dx \leqslant L^2. \tag{5}$$

Каждую функцию этого класса можно представить в виде

$$f(x) = \sum_{p=0}^{q-1} \frac{f^{(p)}(0)}{p!} x^p + \int_0^1 \frac{[(x-t)^+]^{q-1}}{(q-1)!} f^{(q)}(t)\,dt. \tag{6}$$

Найдем среди формул (1) наилучшую квадратурную формулу для класса $W_2^q(L)$.

Остаточный член квадратурной формулы (1) для функций этого класса имеет вид (8), § 5, с ядром $K_n(t)$ ((6), § 5). Используя

неравенство Коши-Буняковского, получаем

$$|R_n(f)| = \left| \int_0^1 f^{(q)}(t)\, K_n(t)\, dt \right| \leqslant \|f^{(q)}\|_{L_2} \|K_n\|_{L_2} = L\psi_n, \tag{7}$$

где

$$\psi_n = \|K_n(x)\|_{L_2} = \left\{ \int_0^1 [K_n(t)]^2\, dt \right\}^{\frac{1}{2}}.$$

Ввиду того что

$$\varphi^{(q)}(t) = L \frac{|K_n(t)| \operatorname{sign}[K_n(t)]}{\|K_n\|_{L_2}} \in L_2[0, 1]$$

и $\|\varphi^{(q)}\|_{L_2} = L$, то для такой функции $\varphi(x)$ формула (7) приводит к соотношению

$$|R_n(\varphi)| = \|K_n\|_{L_2}.$$

Следовательно,

$$\sup_{f \in W_2^q(L)} |R_n(f)| = L \|K_n\|_{L_2} \tag{8}$$

и целесообразно ввести следующее определение:

Определение 1. Квадратурная формула (1) называется *оптимальной квадратурной формулой* на классе $W_2^q(L)$, если для ее погрешности имеет место оценка (7), с константой ψ_n, удовлетворяющей условию

$$\psi_n^2 = \inf_{x_i, C_i^{(n)}} \int_0^1 [K_n(t)]^2\, dt.$$

Справедлива следующая теорема:

Теорема 1. *Среди всех квадратурных формул вида* (1), *точных для многочленов степени меньшей или равной* $q - 1$, *оптимальной квадратурной формулой на классе* $W_2^q(L)$ *будет формула, точная на сплайн-функциях порядка* q.

Доказательство. Пусть квадратурная формула

$$\overline{Q}_n(f) = \sum_{k=1}^{n} \overline{C}_k^{(n)} f(x_k^{(n)}) \tag{9}$$

является точной на сплайн-функциях порядка q, т. е.

$$\overline{R}_n(s) = Q(s) - \overline{Q}_n(s) = 0, \quad \forall s \in S. \tag{10}$$

Построим какую-нибудь другую квадратурную формулу

$$Q_n(f) = \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}), \tag{11}$$

для которой

$$R_n(f) = Q(f) - Q_n(f) = 0, \quad \forall f \in \bigcup_{r=0}^{q-1} \pi_r. \tag{12}$$

Используя ядро $K_n(t)$, остаточные члены квадратурных формул (9) и (11) могут быть записаны в виде

$$\overline{R}_n(f) = \int_0^1 \overline{K}_n(t) f^{(q)}(t)\,dt; \quad R_n(f) = \int_0^1 K_n(t) f^{(q)}(t)\,dt. \tag{13}$$

Следовательно,

$$Q_n(f) - \overline{Q}_n(f) =$$

$$= \sum_{i=1}^{n} [C_i^{(n)} - \overline{C}_i^{(n)}] f(x_i^{(n)}) = \overline{R}_n(f) - R_n(f) = \int_0^1 M_n(t) f^{(q)}(t)\,dt,$$

где

$$M_n(t) = \overline{K}_n(t) - K_n(t) = \sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{[(x_i^{(n)} - t)^+]^{q-1}}{(q-1)!}. \tag{13'}$$

Поскольку обе квадратурные формулы (9) и (11) точные на множестве многочленов степени меньшей или равной $q-1$, то

$$Q_n(P) - \overline{Q}_n(P) = \sum_{i=1}^{n} [C_i^{(n)} - \overline{C}_i^{(n)}] P(x_i^{(n)}) = 0, \quad \forall P \in \bigcup_{j=0}^{q-1} \pi_j \tag{14}$$

и, кроме того, будет иметь место соотношение

$$\sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{[(x_i^{(n)} - t)^+]^{q-1}}{(q-1)!} = (-1)^q \sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{[(t - x_i^{(n)})^+]^{q-1}}{(q-1)!}. \tag{15}$$

Для доказательства (15) положим $t = \xi \in [x_k^{(n)}, x_{k+1}^{(n)}]$, тогда равенство (15) примет вид

$$\sum_{i=k+1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{(x_i^{(n)} - t)^{q-1}}{(q-1)!} = (-1)^q \sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{(t - x_i^{(n)})^{q-1}}{(q-1)!},$$

или

$$\sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{(x_i^{(n)} - t)^{q-1}}{(q-1)!} = \sum_{k=0}^{q-1} \frac{(-1)^k}{k!} \sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{(x_i^{(n)})^{q-k-1}}{(q-k-1)!} t^k \equiv 0,$$

ибо коэффициенты при степенях t на основании (14) все равны нулю. Полученное тождество доказывает справедливость соотношения (15).

С учетом (15), формулу (13) можно привести к виду

$$M_n(t) = (-1)^q \sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{[(t - x_i^{(n)})^+]^{q-1}}{(q-1)!}, \tag{16}$$

причем согласно (14) должны выполняться соотношения:

$$\sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] [x_i^{(n)}]^j = 0, \qquad j = 0, 1, \ldots, q-1. \tag{17}$$

Следовательно, функция

$$s(x) = (-1)^q \sum_{i=1}^{n} [\overline{C}_i^{(n)} - C_i^{(n)}] \frac{[(t - x_i^{(n)})^+]^{2q-1}}{(2q-1)!} \tag{18}$$

удовлетворяет определению сплайн-функции порядка q (см. § 3, гл. 3) и, кроме того,

$$s^{(q)}(x) = M_n(x). \tag{19}$$

Поскольку квадратурная формула (9) точная на всех сплайн-функциях q-го порядка, то

$$\overline{R}_n(s) = \int_0^1 \overline{K}_n(t)\, s^{(q)}(t)\, dt = \int_0^1 \overline{K}_n(t)\, M_n(t)\, dt = 0. \tag{20}$$

Отсюда следует

$$\int_0^1 [K_n(t)]^2\, dt = \int_0^1 [K_n(t) - \overline{K}_n(t) + \overline{K}_n(t)]^2\, dt =$$

$$= \int_0^1 [\overline{K}_n(t) - K_n(t)]^2\, dt + 2\int_0^1 [K_n(t) - \overline{K}_n(t)]\, \overline{K}_n(t)\, dt + \int_0^1 [\overline{K}_n(t)]^2\, dt =$$

$$= \int_0^1 [\overline{K}_n(t) - K_n(t)]^2\, dt + \int_0^1 [\overline{K}_n(t)]^2\, dt \geqslant \int_0^1 [\overline{K}_n(t)]^2\, dt,$$

здесь учтены (13) и (20). Причем равенство достигается тогда и только тогда, когда $K_n(t) = \overline{K}_n(t)$, т. е. при $C_i^{(n)} = \overline{C}_i^{(n)}$, $i = 1, 2, \ldots, n$. Теорема доказана полностью.

Отметим, что теорема 1 указывает только класс квадратурных формул, среди которых находится оптимальная квадратурная формула на классе $W_2^q(L)$. Для отыскания явного вида ее необходимо еще минимизировать $\int_0^1 [K_n(t)]^2\, dt$ по всевозможным наборам узлов $x_i^{(n)} \in [0, 1]$, $i = 1, 2, \ldots, n$.

Пример 1. Пусть $q = 1$, тогда, согласно теореме 1, оптимальную квадратурную формулу на классе функций из $W_2^1(L)$ нужно искать среди формул вида (1), точных на сплайн-функциях первого порядка.

Ввиду того, что фундаментальными сплайн-функциями первого порядка будут выражения

$$S_j(t) = \begin{cases} 0 & t \in [0, x_{j-1}^{(n)}], \\ (t - x_{j-1}^{(n)})/(x_j^{(n)} - x_{j-1}^{(n)}), & t \in [x_{j-1}^{(n)}, x_j^{(n)}], \\ (t - x_{j+1}^{(n)})/(x_j^{(n)} - x_{j+1}^{(n)}), & t \in [x_j^{(n)}, x_{j+1}^{(n)}], \\ 0, & t \in [x_{j+1}^{(n)}, 1], \end{cases} \tag{21}$$

$$j = 1, 2, \ldots, n,$$

то коэффициенты $C_i^{(n)}$ оптимальной квадратурной формулы будут задаваться формулами:

$$C_j^{(n)} = \int_0^1 s_j(t)\, dt = \int_{x_{j-1}^{(n)}}^{x_j^{(n)}} \frac{t - x_{j-1}^{(n)}}{x_j^{(n)} - x_{j-1}^{(n)}}\, dt + \int_{x_j^{(n)}}^{x_{j+1}^{(n)}} \frac{t - x_{j+1}^{(n)}}{x_j^{(n)} - x_{j+1}^{(n)}}\, dt = \frac{x_{j+1}^{(n)} - x_{j-1}^{(n)}}{2}, \tag{22}$$

$$j = 1, 2, \ldots, n; \quad x_0^{(n)} = 0, \quad x_{n+1}^{(n)} = 1.$$

В данном случае

$$K_n(t) = 1 - t - \sum_{i=1}^{n} C_i^{(n)} [(x_i^{(n)} - t)^+]^0 = \sum_{i=1}^{j} C_i^{(n)} - t = \frac{1}{2} [x_{j+1}^{(n)} + x_j^{(n)} - x_1^{(n)}]^2 - t,$$

$$t \in [x_j^{(n)}, x_{j+1}^{(n)}],$$

и

$$\int_0^1 [K_n(t)]^2 dt = \sum_{j=0}^{n} \int_{x_j^{(n)}}^{x_{j+1}^{(n)}} [K_n(t)]^2 dt =$$

$$= \frac{[x_1^{(n)}]^3}{3} + \frac{1}{12} \sum_{j=1}^{n-1} [x_{j+1}^{(n)} - x_j^{(n)}]^3 + \frac{[1 - x_n^{(n)}]^3}{3}. \qquad (23)$$

Определяя минимум выражения (23), находим, что он равен $\frac{1}{12n^2}$ и достигается при $x_i^{(n)} = \frac{2i-1}{2n}$. Таким образом приходим к выводу, что оптимальной квадратурной формулой вида (1) на классе $W_2^1(L)$ будет формула

$$C_i^{(n)} = \frac{1}{n}, \quad x_i^{(n)} = \frac{2i-1}{2n}, \qquad i = 1, 2, \ldots, n,$$

т. е. формула средних прямоугольников.

Оценка остаточного члена для этой формулы будет

$$\left| \int_0^1 f(x)\, dx - \frac{1}{n} \sum_{i=1}^{n} C_i^{(n)} f\left(\frac{2i-1}{2n}\right) \right| \leqslant \frac{L}{2\sqrt{3}n}, \quad \forall f \in W_2^1(L).$$

Чтобы получить оптимальную квадратурную формулу на классе $W_2^q(L)$, нужно фактически построить только фундаментальные сплайн-функции порядка q и вычислить $C_i^{(n)}$ по формуле

$$C_i^{(n)} = \int_0^1 s_i(x)\, dx, \qquad i = 1, 2, \ldots, n,$$

что следует из теоремы 1.

§ 7. СХОДИМОСТЬ ОБЩЕГО КВАДРАТУРНОГО ПРОЦЕССА, НЕ СОДЕРЖАЩЕГО ПРОИЗВОДНЫХ

Будем рассматривать квадратурный процесс, определяемый бесконечными треугольными матрицами абсцисс $x_k^{(n)} \in [a, b]$ и весов $C_k^{(n)}$,

$$X = \begin{bmatrix} x_0^{(0)} & & & \\ x_0^{(1)} & x_1^{(1)} & & \\ \cdots & \cdots & \cdots & \cdots \\ x_0^{(n)} & x_1^{(n)} & \ldots & x_n^{(n)} \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \quad C = \begin{bmatrix} c_0^{(0)} & & & \\ c_0^{(1)} & c_1^{(1)} & & \\ \cdots & \cdots & \cdots & \cdots \\ c_0^{(n)} & c_1^{(n)} & \ldots & c_n^{(n)} \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \qquad (1)$$

Определение 1. Квадратурная формула, соответствующая n-й строке матриц X и C, имеющая вид

$$Q(f) = \int_a^b \rho(x) f(x)\, dx = \sum_{k=0}^{n} C_k^{(n)} (x_k^{(n)}) + R(f) = Q_n(f) + R(f), \qquad (2)$$

называется *сходящейся,* если

$$\lim_{n\to\infty} Q_n(f) = \lim_{n\to\infty} \sum_{k=0}^{n} C_k^{(n)} f(x_k^{(n)}) = Q(f). \tag{3}$$

Введенный выше линейный функционал $Q_n(f)$, который можно рассматривать в банаховом пространстве $C[a, b]$, является непрерывным и норма его равна

$$\|Q_n\| = \sum_{k=0}^{n} |C_k^{(n)}|.$$

Теорема 1. *Для того чтобы* $\forall f(x) \in [a, b]$ *имела место сходимость последовательности квадратурных формул* $Q_n(f)$, *построенных по треугольным матрицам* X *и* C, *необходимо и достаточно, чтобы*

1) $\lim\limits_{n\to\infty} Q_n(P) = Q(P)$, $P(x) \in \pi_n$, $n = 0,1, \ldots$;

2) $\exists M \in R_1$ такая, что

$$\sum_{k=0}^{n} |C_k^{(n)}| < M, \qquad n = 0, 1, \ldots. \tag{4}$$

Доказательство. Известно, что множество многочленов является всюду плотным подмножеством $C[a, b]$ (теорема Вейерштрасса). На этом подмножестве по условию 1) квадратурный процесс сходится и нормы линейных функционалов Q_n согласно (4) ограничены в совокупности, следовательно, утверждение теоремы 1 следует из общей теоремы сходимости для линейных операторов (приложение, § 1).

Приведем два простых следствия из теоремы 1.

Теорема 2. *Для того чтобы* $\forall f(x) \in C[a, b]$ *интерполяционный квадратурный процесс сходился, необходимо и достаточно выполнения неравенства* (4).

Доказательство. Первое условие теоремы 1 здесь выполнено, ибо из того, что квадратурный процесс — интерполяционный, следует, что для всякого многочлена $P_m(x)$ m-й степени существует $n > m$ такое, что

$$Q_n(P_m) = \int_a^b \rho(x) P_m(x)\, dx.$$

Второе условие теоремы 1 совпадает с условием доказываемой теоремы.

Теорема 3. *Если весовые коэффициенты* $C_k^{(n)}$ — *неотрицательны, то квадратурный процесс сходится* $\forall f(x) \in C[a, b]$ *тогда и только тогда, когда он сходится* $\forall P(x) \in \pi_n$, $n = 0, 1, \ldots$.

Доказательство. Необходимость очевидна. Нужно проверить лишь достаточность. Ввиду того что при $f(x) = 1$ должно выполняться предельное соотношение

$$\lim_{n\to\infty} Q_n(1) = \lim_{n\to\infty} \sum_{k=0}^{n} C_k^{(n)} = \lim_{n\to\infty} \sum_{k=0}^{n} |C_k^{(n)}| = \lim_{n\to\infty} \|Q\| = 1,$$

то множество норм $\|Q_n\|$ должно быть ограниченным для всех, $n = 0, 1, \ldots$, т. е. выполняется условие 2) теоремы 1.

Приведем две теоремы, указывающие на влияние выбора абсцисс квадратурных формул интерполяционного типа на сходимость.

Теорема 4. *Если абсциссы* $x_k^{(n)}$ *выбираются равноотстоящими на* $[-1, 1]$, *с* $x_0^{(n)} = -1$, $x_n^{(n)} = 1$, *то веса* $C_k^{(n)} = Q(Q_{in})$ *таковы, что величина* $\sum_{k=0}^{n} |C_k^{(n)}|$ *не ограничена по* n. *Следовательно,* $\exists\, g\ (x) \in C\,[a, b]$ *такая, что*

$$\lim_{n\to\infty} Q_n(g) \neq Q(g).$$

Здесь $Q_{in}(x)$, $i = 0, 1, \ldots, n$,— фундаментальный многочлен (см. § 2, гл. 1). Имеет место однако и положительный результат.

Теорема 5. *Если в качестве абсцисс выбрать нули многочленов Чебышева первого рода* $T_{n+1}(x)$, *т. е.*

$$x_k^{(n)} = \cos\frac{(2k+1)\pi}{2n+1}, \qquad k = 0, 1, \ldots, n,$$

или нули многочленов Чебышева второго рода $U_{n+1}(x)$, *т. е.*

$$x_k^{(n)} = \cos\frac{(k+1)\pi}{n+2}, \qquad k = 0, 1, \ldots, n,$$

то все веса $C_k^{(n)}$ *будут положительны и*

$$\lim_{n\to\infty} Q_n(f) = Q_n(f), \quad \forall f \in C\,[a, b].$$

Отметим, что сходимость квадратурного процесса наивысшей алгебраической степени точности является прямым следствием теоремы 3. Действительно, при $\rho(x) \geqslant 0$ квадратурная формула наивысшей алгебраической степени точности может быть построена для всех n (теорема 1, § 2, гл. 4) и все весовые коэффициенты положительны (теорема 2, § 2, гл. 4). Кроме того, квадратурный процесс будет сходиться для любого многочлена.

Часть II

ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ ОПЕРАТОРНЫХ УРАВНЕНИЙ

Глава 5

ПРОЕКЦИОННО-ВАРИАЦИОННЫЕ МЕТОДЫ

В практике инженерных расчетов большое распространение получили методы, укладывающиеся в общие рамки так называемых проекционных методов. Среди них метод Бубнова — Галеркина, метод Ритца, метод наименьших квадратов и ряд других. В данной главе изложены наиболее применяемые проекционные методы, при этом основное внимание уделяется вариационным методам, теория которых в настоящее время близка к своему завершению. В последнее время интерес к проекционно-вариационным методам значительно повысился в связи с появлением интересных результатов по применению этих методов для получения хороших разностных аппроксимаций задач математической физики (имеется в виду метод конечных элементов).

В начале главы изложены общие идеи проекционно-вариационных методов. Затем приведен метод наименьших квадратов и метод Ритца.

§ 1. МЕТОД МОМЕНТОВ

Почти все проекционные методы являются частными случаями метода моментов или как его еще называют метода Галеркина — Петрова.

Пусть $A : \mathbf{E} \to \mathbf{F}$ — линейный оператор, действующий из банахова пространства $\mathbf{E}$ в банахово пространство $\mathbf{F}$, с областью определения $D(A) \subset \mathbf{E}$ и областью значений $R(A) \subset \mathbf{F}$. Рассмотрим уравнение

$$Au = f. \tag{1}$$

Проекционный метод решения этого уравнения заключается в следующем. Задаются две последовательности подпространств $\{E_n\}$ и $\{F_n\}$; $E_n \subset D(A)$; $F_n \subset \mathbf{F}$, $n = 1, 2, \ldots$, а также линейные проекционные операторы (проекторы) $P_n : \mathbf{F} \to F_n$, т. е. операторы, удовлетворяющие условиям:

$$P_n^2 = P_n; \quad P_n\mathbf{F} = F_n, \qquad n = 1, 2, \ldots .$$

Уравнение (1) заменяем уравнением

$$P_n(Au_n - f) = 0, \tag{2}$$

решение которого ищем в E_n. В зависимости от того, каким образом строятся подпространства E_n, F_n и проекционные операторы P_n, получаются те или другие методы.

Пусть пространства **E** и **F** являются гильбертовыми. Выберем в $D(A)$ и в **F** две последовательности линейно-независимых элементов $\{\varphi_i\}_{i=1}^{\infty}$ и $\{\psi_i\}_{i=1}^{\infty}$ соответственно. Первую из них будем называть *координатной системой,* а вторую — *проекционной.* В качестве подпространств E_n и F_n возьмем линейные оболочки элементов φ_i, ψ_i, $i = 1, 2, \ldots, n$ соответственно, а в качестве проектора P_n — оператор ортогонального проектирования (ортопроектор **F** на F_n). Для дальнейших преобразований уравнения (2) воспользуемся следующей леммой:

Лемма 1. *Для любого элемента ψ из гильбертового пространства* **F** *равенство*

$$P_n\psi = 0 \tag{3}$$

и система равенств

$$(\psi, \psi_j) = 0, \qquad j = 1, 2, \ldots, n \tag{4}$$

эквивалентны. Здесь $(\cdot, \cdot)$ — скалярное произведение в **F**; *P_n — ортопроектор* **F** *на F_n-линейную оболочку ψ_i, $i = 1, 2, \ldots, n$.*

Доказательство. Пусть справедливо (3). Тогда в силу равенств $\psi_j = P_n\psi_j$, $j = 1, 2, \ldots, n$ и самосопряженности оператора P_n получаем

$$(\psi, \psi_j) = (\psi, P_n\psi_j) = (P_n\psi, \psi_j) = 0, \qquad j = 1, 2, \ldots, n,$$

так что верны соотношения (4). Допустим теперь, что справедливы равенства (4). Так как $P_n\psi \in F_n$, то $P_n\psi = \sum\limits_{i=1}^{n} a_i\psi_i$.

Тогда

$$(P_n\psi, P_n\psi) = \left(P_n\psi, \sum_{j=1}^{n} a_j\psi_j\right) = \sum_{j=1}^{n} \bar{a}_j (P_n\psi, \psi_j) = \sum_{j=1}^{n} \bar{a}_j (\psi, \psi_j) = 0$$

и лемма доказана.

В силу этой леммы уравнение (2) эквивалентно системе

$$(Au_n - f, \psi_j) = 0, \qquad j = 1, 2, \ldots, n, \tag{5}$$

где $u_n \in E_n$ и, следовательно, находится в виде

$$u_n = \sum_{i=1}^{n} c_i\varphi_i. \tag{6}$$

Подставляя (6) в (5) и используя свойства скалярного произведения и линейность оператора A, получаем

$$\sum_{i=1}^{n} (A\varphi_i, \psi_j)\, c_i = (f, \psi_j), \qquad j = 1, 2, \ldots, n. \tag{7}$$

Таким образом, задача определения u_n свелась к задаче нахождения решения системы линейных алгебраических уравнений (7) относительно неизвестных коэффициентов c_i, $i = 1, 2, \ldots, n$. Описанный метод

носит название *метода моментов* или *метода Галеркина — Петрова*. При приближенном решении конкретных уравнений этим методом координатную систему $\{\varphi_i\}_{i=1}^{\infty}$ и проекционную $\{\psi_i\}_{i=1}^{\infty}$ можно выбирать различными способами. От того, как выбраны эти системы, зависит быстрота сходимости применяемого метода и устойчивость счета.

Приведем два частных варианта метода моментов, которые довольно часто применяются на практике.

Если гильбертовы пространства **E** и **F** совпадают и если координатная система совпадает с проекционной $\varphi_i = \psi_i$, $i = 1, 2, \ldots$, то метод моментов носит название *метода Бубнова — Галеркина*.

Если элементы проекционной системы выбираются исходя из координатной системы по формулам:

$$\psi_j = A\varphi_j, \qquad j = 1, 2, \ldots,$$

то метод моментов называют *методом наименьших квадратов*.

Определение 1. Будем говорить, что последовательность пространств $\{E_n\}_{n=1}^{\infty}$ *предельно плотная* в банаховом пространстве **E**, если $\forall\, \varphi \in \mathbf{E}$ выполняется соотношение

$$\lim_{n\to\infty} \rho(\varphi, E_n) = \lim_{n\to\infty} \inf_{u\in E_n} \|\varphi - u\| = 0.$$

Теорема 1. *Пусть область определения $D(A)$ оператора A плотная в* **E**, *а область значений $R(A)$ — в* **F** *и пусть A переводит $D(A)$ в $R(A)$ взаимно однозначно. Пусть подпространства $A(E_n)$ и F_n замкнуты в* **F**. *Пусть, наконец, проекторы P_n равномерно ограничены по n*

$$\|P_n\| \leqslant c, \qquad n = 1, 2, \ldots. \tag{8}$$

Тогда для того, чтобы $\forall\, f \in \mathbf{F}$, начиная с некоторого номера n_0, существовало единственное решение u_n, $n \geqslant n_0$, уравнения (2) *и чтобы* $\lim\limits_{n\to\infty} \|Au_n - f\| = 0$, *необходимо выполнение условий:*

1) *последовательность подпространств $A(E_n)$ предельно плотная в* **F**;

2) *при $n \geqslant n_0$ оператор P_n переводит $A(E_n)$ в F_n взаимно однозначно;*

3) $\varliminf\limits_{n\to\infty} \tau_n = \varliminf\limits_{n\to\infty} \inf\limits_{\substack{z_n\in A(E_n)\\ \|z_n\|=1}} \|P_n z_n\| = \tau > 0.$

При этом быстрота сходимости характеризуется неравенством

$$\rho(f, A(E_n)) \leqslant \|Au_n - f\| \leqslant \left(1 + \frac{c}{\tau_n}\right) \rho(f, A(E_n)). \tag{9}$$

В случае, когда подпространства E_n и F_n — конечномерны и размерности их совпадают, условие 2) *является следствием условия* 3).

Доказательство. Произведем в (2) замену $Au_n = x_n$, после чего получим

$$P_n x_n = P_n f, \quad x_n \in A(E_n). \tag{10}$$

Достаточность. Пусть выполнены условия 1) — 3) теоремы. Обозначим через $\tilde{P}_n$ сужение проектора P_n на подпространство $A(E_n)$. Согласно условию 2), при $n \geqslant n_0$ существует обратный ограниченный оператор $\tilde{P}_n^{-1}$, переводящий F_n на $A(E_n)$. Таким образом, существует единственный элемент $x_n \in A(E_n)$, являющийся решением уравнения (10),

$$x_n = \tilde{P}_n^{-1} P_n f.$$

Но тогда $u_n = A^{-1} x_n$ — единственный элемент, удовлетворяющий уравнению (2). Существование A^{-1} следует из того, что A переводит $D(A)$ в $R(A)$ взаимно однозначно.

Далее из условия 3) вытекает, что $\|\tilde{P}_n^{-1}\| \leqslant \frac{1}{\tau_n} \leqslant \frac{1}{\tau}$. Тогда с учетом (8) получаем

$$\|\tilde{P}_n^{-1} P_n\| \leqslant \|\tilde{P}_n^{-1}\| \|P_n\| \leqslant \frac{c}{\tau_n} \leqslant \frac{c}{\tau} < \infty,$$

т. е. нормы операторов $\tilde{P}_n^{-1} P_n$ равномерно ограничены. $\forall f_n \in A(E_n)$ имеем $\tilde{P}_n^{-1} P_n f_n = f_n$, следовательно,

$$Au_n - f = x_n - f = \tilde{P}_n^{-1} P_n f - f = \tilde{P}_n^{-1} P_n (f - f_n) - (f - f_n),$$

откуда получаем

$$\|Au_n - f\| \leqslant \left(1 + \frac{c}{\tau_n}\right) \|f - f_n\|. \tag{11}$$

Из (11) в силу произвольности элемента $f_n \in A(E_n)$ следует правая часть неравенства (9). Левая часть (9) следует просто из определения расстояния элемента от подпространства. Соотношение $\lim\limits_{n\to\infty} \|Au_n - f\| = 0$ следует из предельной плотности последовательности пространств E_n.

Необходимость. Пусть $\forall f \in \mathbf{F}$ при $n \geqslant n_0$ существует единственное решение уравнения (2) или (10) и $\lim\limits_{n\to\infty} \|Au_n - f\| = 0$. Нужно показать, что выполняются условия 1) — 3). Поскольку

$$\lim_{n\to\infty} \|Au_n - f\| = \lim_{n\to\infty} \|x_n - f\| = 0, \quad \forall f \in \mathbf{F},$$

где $x_n \in A(E_n)$, то условие 1) выполняется. Далее, уравнение (2) при $n \geqslant n_0$ имеет единственное решение, следовательно, единственное решение имеет и уравнение (10), т. е. выполняется условие 2). Для установления справедливости 3) достаточно показать, что нормы операторов $\tilde{P}_n^{-1}$ равномерно ограничены при $n \geqslant n_0$. Ввиду того что при $n \geqslant n_0$ $x_n = \tilde{P}_n^{-1} P_n f$, то

$$\tilde{P}_n^{-1} P_n f \underset{n\to\infty}{\to} f, \quad \forall f \in \mathbf{F},$$

и на основании теоремы Банаха — Штейнгауза (приложение, § 1) получаем

$$\|\tilde{P}_n^{-1}P_n\| \leqslant c, \quad \forall n \geqslant n_0.$$

В частности, для $f_n \in F_n$ имеем:

$$\|\tilde{P}_n^{-1}f_n\| = \|\tilde{P}_n^{-1}P_nf_n\| \leqslant c\|f_n\|,$$

откуда следует выполнение условия 3). Теорема доказана.

Заметим, что в условиях теоремы 1 нет предположения о разрешимости уравнения (1), поэтому не затрагивался вопрос о сходимости приближений u_n. Если же $f \in R(A)$ и существует ограниченный обратный оператор A^{-1}, то из очевидного неравенства

$$\|u_n - A^{-1}f\| \leqslant \|A^{-1}\|\|Au_n - f\|$$

вытекает, что сходимость невязки к нулю влечет за собой сходимость приближений u_n к решению уравнения (1).

Отметим, что условие (8) сильно сужает область применимости теоремы 1. Дело в том, что во многих случаях даже для конечномерных пространств F_n не существует проекторов P_n с равномерно ограниченными нормами. Например, проектор $P_n : C[0, 1] \to \bigcup_{k=0}^{n} \pi_k$ имеет норму

$$\|P_n\| \geqslant c \ln n \qquad (c = \text{const} > 0),$$

т. е. $\lim\limits_{n\to\infty} \|P_n\| = \infty$ (см. § 4, гл. 2).

§ 2. ВАРИАЦИОННЫЕ МЕТОДЫ. ОБЩИЕ ПОЛОЖЕНИЯ

Пусть для уравнения

$$Au = f, \tag{1}$$

о котором известно, что $f \in R(A)$ и что оно обладает единственным решением u^*, построен функционал $\Phi : \mathbf{E} \to \mathrm{R}_1$, минимум которого достигается в единственной точке u^*. Тогда задачу решения уравнения (1) заменяем задачей минимизации функционала $\Phi(u)$. Методы, основанные на этой идее, носят название *вариационных*.

Вначале рассмотрим вопрос о минимизации квадратичных функционалов, не связывая его с решением уравнения (1). Далее, конкретизируя способы перехода от уравнения (1) к квадратичному функционалу, изложим метод наименьших квадратов и метод Ритца.

Пусть $G(u, v)$ — симметричная полуторалинейная форма, т. е.

$$G(\alpha_1u_1 + \alpha_2u_2, v) = \alpha_1G(u_1, v) + \alpha_2G(u_2, v);$$

$$G(u, \alpha_1v_1 + \alpha_2v_2) = \bar{\alpha}_1G(u,v_1) + \bar{\alpha}_2G(u, v_2);$$

$$G(u, v) = \overline{G(v, u)},$$

где u, u_i, $v_i \in D(G) \subset \mathbf{H}$, $i = 1, 2$, причем область определения формы $D(G)$ — линейна и плотна в $\mathbf{H}$. С формой $G(u, v)$ связан

вещественный квадратичный функционал $G(u, u)$ с той же областью определения, который мы будем считать положительным, т. е.

$$G(u, u) > 0, \quad u \neq 0. \tag{2}$$

Пусть, далее, $l(u)$ — линейный функционал с $D(l) \supset D(G)$.

Рассмотрим теперь следующий функционал:

$$\Phi(u) = G(u, u) - 2\,\mathrm{Re}\, l(u) + c, \quad u \in D(G), \tag{3}$$

где c — вещественная постоянная.

Лемма 1. *Если квадратичный функционал $G(u, u)$ положительно определен, т. е.*

$$G(u, u) \geqslant \mu(u, u), \quad \mu > 0, \quad \forall u \in D(G), \tag{4}$$

где (u, v) — скалярное произведение в $\mathbf{H}$, и если функционал $l(u)$ ограничен, то функционал $\Phi(u)$ ограничен снизу.

Доказательство в виде упражнения оставляем читателю.

Введем следующее определение:

Определение 1. Последовательность $\{u^{(n)}\}_{n=1}^{\infty}$ элементов из $D(G)$ называется *минимизирующей* для функционала $\Phi(u)$, если

$$\lim_{n\to\infty} \Phi(u^{(n)}) = \Phi_0 = \inf_{u \in D(G)} \Phi(u). \tag{5}$$

Метризуем пространство $D(G)$ с помощью введения следующей метрики:

$$\rho_G(u, v) = [G(u - v, u - v)]^{\frac{1}{2}}, \quad \forall u, v \in D(G) \tag{6}$$

(справедливость аксиом расстояния легко проверяется), тогда справедлива следующая теорема:

Теорема 1. *Пусть $\{u^{(n)}\}_{n=1}^{\infty}$ — минимизирующая последовательность для функционала $\Phi(u)$. Тогда последовательность $\{u^{(n)}\}_{n=1}^{\infty}$ — фундаментальная в $D(G)$. Если, кроме того, функционал $\Phi(u)$ достигает своего минимального значения Φ_0 в некоторой точке $u^* \in D(G)$, то*

$$\lim_{n\to\infty} \rho_G(u^{(n)}, u^*) = 0.$$

Доказательство. Ввиду определения минимизирующей последовательности для всех $\varepsilon > 0$ существует N такое, что

$$\Phi(u^{(n)}) < \Phi_0 + \varepsilon, \quad \forall n > N,$$

и, следовательно,

$$\Phi(u^{(m)}) + \Phi(u^{(n)}) - 2\Phi\left(\frac{u^{(m)} + u^{(n)}}{2}\right) < 2\varepsilon, \quad \forall n, m > N. \tag{7}$$

Докажем тождество

$$\Phi(u) + \Phi(v) - 2\Phi\left(\frac{u + v}{2}\right) = \frac{1}{2}\rho_G^2(u - v), \quad \forall u, v \in D(G). \tag{8}$$

Действительно,

$$\Phi(u)+\Phi(v)-2\Phi\left(\frac{u+v}{2}\right)=G(u,u)+G(v,v)-2G\left(\frac{u+v}{2},\ \frac{u+v}{2}\right)=$$

$$=\frac{1}{2}G(u,u)-\operatorname{Re}G(u,v)+\frac{1}{2}G(v,v)=\frac{1}{2}G(u-v,u-v)=$$

$$=\frac{1}{2}\rho_G^2(u-v).$$

Из (7) и (8) следует первая часть утверждения теоремы. Вторая часть следует из первой, если принять во внимание тот факт, что последовательность

$$u^{(1)},\ u^*,\ u^{(2)},\ u^*,\ \ldots$$

является также минимизирующей. Теорема доказана.

Следствие 1. Если квадратичный функционал G удовлетворяет условию (4), т. е. является положительно определенным, то имеют место утверждения теоремы 1 в метрике гильбертового пространства $\mathbf{H}:\rho\ (u,\ v)=(u-v,\ u-v)^{\frac{1}{2}}$.

Теорема 2. *Пусть $\Phi(u)$ достигает своего минимума Φ_0 в некоторой точке $u^*\in D(G)$, тогда*

1) $G(u^*,v)=l(v),\quad \forall v\in D(G)$

2) $\Phi(u^*+v)=\Phi(u^*)+G(v,v),\quad \forall v\in D(G)$.

Из 2), в частности, следует, что минимум $\Phi(u)$ является строгим.

Доказательство. В силу предположения теоремы $\forall v\in D(G)$ будем иметь:

$$\Phi(u^*+v)=G(u^*+v,\ u^*+v)-2\operatorname{Re}l(u^*+v)+c=\Phi(u^*)+$$
$$+2\operatorname{Re}G(u^*,\ v)-2\operatorname{Re}l(v)+G(v,\ v)\geqslant\Phi(u^*). \tag{9}$$

Отсюда получаем, что $\forall v\in D(G)$ должно выполняться неравенство

$$2\operatorname{Re}[G(u^*,\ v)-l(v)]+G(v,\ v)\geqslant 0 \tag{10}$$

и, в частности, оно должно иметь место при замене v на tv, где t — произвольное вещественное число. Тогда будем иметь:

$$2t\operatorname{Re}[G(u^*,\ v)-l(v)]+t^2G(v,\ v)\geqslant 0. \tag{10'}$$

Поскольку и величиной, и знаком t разрешается варьировать, то всегда t можно выбрать так, чтобы в левой части (10′) первое слагаемое было преобладающим и отрицательным. Полученное противоречие разрешается только тогда, когда выполняется условие

$$\operatorname{Re}G(u^*,\ v)=\operatorname{Re}l(v).$$

Аналогично, подставляя в (10′) itv вместо v, показываем, что должно выполняться также и такое условие:

$$\operatorname{Im}G(u^*,\ v)=\operatorname{Im}l(v),$$

тем самым первая часть теоремы доказана. Но тогда из (9) следует вторая часть теоремы и доказательство полностью закончено.

В предположении, что элемент $u^* \in D(G)$, минимизирующий функционал $\Phi(u)$, существует, возникает вопрос нахождения приближений к этому элементу. Чтобы ответить на этот вопрос, поступим следующим образом. Выберем в $D(G)$ линейно-независимую систему элементов $\{u_i\}_{i=1}^{\infty}$ и построим по этой системе линейную оболочку первых n элементов u_i, $i = 1, 2, \ldots, n$, которую обозначим через $H_n \subset D(G)$.

Имеют место следующие две леммы, которые ввиду их простоты приведем без доказательства:

Лемма 2. *Функционалы $G(u, u)$ и $l(u)$ непрерывны на каждом из пространств H_n.*

Лемма 3. *Квадратичный функционал $G(u, u)$ положительно определен на каждом из подпространств H_n, т. е. существуют $m_n > 0$, $n = 1, 2, \ldots$, и такие, что*

$$G(u, u) \geqslant m_n(u, u) = m_n \|u\|^2, \quad \forall u \in H_n. \tag{11}$$

Лемма 4. *Для всех n в H_n имеется единственный элемент $u^{(n)}$ такой, что*

$$\Phi_n = \Phi(u^{(n)}) = \inf_{u \in H_n} \Phi(u). \tag{12}$$

Доказательство. Ввиду того что ограниченность линейного оператора является необходимым и достаточным условием его непрерывности, то из леммы 2 вытекает, что

$$|l(u)| \leqslant d_n \|u\|, \quad \forall u \in H_n,$$

и, следовательно,

$$\Phi(u) \geqslant m_n \|x\|^2 - 2d_n \|x\|^2 = m_n \left[\|x\| - \frac{d_n}{m_n}\right]^2 - \frac{d_n}{m_n} + c.$$

Из последнего неравенства получаем, что $\Phi(u) > c$, если $\|x\| > \frac{2d_n}{m_n}$.

Пусть $S_n = \left\{x : \|x\| \leqslant \frac{2d_n}{m_n}\right\}$, тогда будут иметь место соотношения

$$\inf_{H_n \setminus S_n} \Phi(u) > c = \Phi(0) \geqslant \inf_{S_n} \Phi(u),$$

откуда следует, что

$$\inf_{H_n} \Phi(u) = \inf_{S_n} \Phi(u).$$

Но S_n — компакт и $\Phi(u)$ непрерывен на этом компакте, следовательно, существует $u_n \in S_n$ и такой, что будет справедливо (12). Единственность u_n следует из второго утверждения теоремы 2. Лемма доказана.

Описанный выше процесс нахождения последовательности $\{u^{(n)}\}_{n=1}^{\infty}$ называют обычно *процессом Ритца*. Утверждать, что полученная в результате этого процесса последовательность $\{u^{(n)}\}_{n=1}^{\infty}$ минимизирующая, конечно нельзя. Нужны дополнительные предположения. Рассмотрим, к чему сводится задача определения $u^{(n)}$ для случая, когда **H** — вещественное гильбертово пространство.

Ввиду того что $u^{(n)}$ отыскивается в H_n, то

$$u^{(n)} = \sum_{i=1}^{n} a_i^{(n)} u_i, \tag{13}$$

где $a_i^{(n)}$ — пока неизвестные вещественные коэффициенты. Подставляя (13) в функционал $\Phi(u)$, получаем функцию от n вещественных переменных $a_i^{(n)}$, $i = 1, 2, \ldots, n$:

$$\Phi\left(\sum_{i=1}^{n} a_i^{(n)} u_i\right) = \sum_{i,j=1}^{n} a_i^{(n)} a_j^{(n)} G(u_i, u_j) - 2\sum_{i=1}^{n} a_i^{(n)} l(u_i) + c.$$

На основании леммы 4 эта функция имеет единственный минимум, который находится из следующей системы:

$$0 = \frac{\partial \Phi}{\partial a_i^{(n)}} = 2\sum_{j=1}^{n} a_j^{(n)} G(u_i, u_j) - 2l(u_i), \qquad i = 1, 2, \ldots, n \tag{14}$$

или

$$\sum_{i=1}^{n} a_i^{(n)} G(u_i, u_j) = l(u_j), \qquad j = 1, 2, \ldots, n. \tag{14'}$$

Определителем системы (14′) является определитель Грамма, если рассматривать $G(u, v)$ как скалярное произведение в $D(G)$ (все аксиомы скалярного произведения, как нетрудно убедиться, выполнены). Известно, что он отличен от нуля. Тем самым получено доказательство леммы 4 с другой стороны и причем конструктивное.

Таким образом, процесс Ритца для каждого n сводится к решению системы линейных алгебраических уравнений вида (14′). Заметим, что все проекционные методы, и в том числе вариационные, не являются итерационными в том смысле, что для отыскания каждого последующего приближения $u^{(n)}$ не используются предыдущие приближения $u^{(i)}$, $i = 1, 2, \ldots, n-1$. Поэтому на практике обычно ограничиваются отысканием лишь одного какого-нибудь приближения $u^{(k)}$.

§ 3. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Пусть $A : \mathbf{H} \to \mathbf{H}$ — линейный оператор, действующий в вещественном гильбертовом пространстве $\mathbf{H}$, для которого существует обратный оператор A^{-1}. Рассмотрим вопрос о решении уравнения

$$Au = f, \tag{1}$$

где $u \in D(A)$, $f \in R(A)$ с точки зрения вариационного подхода, изложенного ранее. Отправным моментом является построение подходящего функционала, минимум которого достигается на единственном решении уравнения (1). Под словами «подходящий функционал» следует понимать такой функционал, к которому можно было бы применить общие теоретические положения § 2. В методе наименьших квадратов такой функционал строится чрезвычайно просто:

$$\Phi(u) = \|Au - f\|^2 = (Au - f, Au - f) = (Au, Au) - 2(Au, f) + \|f\|^2. \tag{2}$$

В силу сделанных предположений относительно оператора A и вида функционала $\Phi(u)$ убеждаемся, что

$$\inf_{u \in D(A)} \Phi(u) = 0 = \Phi(u^*), \tag{3}$$

где u^* — решение уравнения (1). Сравнивая функционал (2) с общим функционалом (3), § 2, видно, что, если в нем положить $D(G) = D(A)$; $G(u, v) = (Au, Av)$, $l(u) = (Au, f)$; $c = \|f\|^2$, то получим функционал (2).

Будем находить последовательность приближений к решению уравнения (1) $\{u^{(n)}\}_{n=1}^{\infty}$ с помощью процесса Ритца, описанного в конце § 2. Пусть $\{u_i\}_{i=1}^{\infty} \subset D(A)$ — система линейно-независимых элементов. Строим по ней линейную оболочку H_n, натянутую на u_i, $i = 1, 2, \ldots, n$, и ищем $u^{(n)}$ в виде

$$u^{(n)} = \sum_{i=1}^{n} a_i^{(n)} u_i. \tag{4}$$

Система линейных алгебраических уравнений (14′), § 2, для определения постоянных $a_i^{(n)}$, $i = 1, 2, \ldots, n$, принимает в данном случае вид

$$\sum_{i=1}^{n} a_i^{(n)} (Au_i, Au_j) = (f, Au_j), \qquad j = 1, 2, \ldots, n, \tag{5}$$

т. е. приходим к методу, о котором уже упоминали в § 1, указывая, что он является частным случаем метода моментов.

Для исследования сходимости метода наименьших квадратов (доказательства того, что $\{u^{(n)}\}_{n=1}^{\infty}$ — минимизирующая последовательность для функционала (2)) нам понадобится следующая лемма:

Лемма 1. *Пусть $U = \{u_i\}_{i=1}^{\infty} \subset \mathbf{H}$ — некоторая линейно-независимая система элементов. Пусть H_n — линейная оболочка, натянутая на элементы u_i, $i = 1, 2, \ldots, n$, и через $\bar{U}$ обозначено замыкание линейной оболочки системы U. Тогда*

1) *для того чтобы $u \in \bar{U}$ $(u \in \mathbf{H})$, необходимо и достаточно, чтобы имело место соотношение*

$$\lim_{n \to \infty} u^{(n)} = \lim_{n \to \infty} P_{H_n} u = u; \tag{6}$$

2) *для того чтобы $\bar{U} = \mathbf{H}$, необходимо и достаточно, чтобы предельное соотношение* (6) *было справедливо $\forall\ u \in \mathbf{H}$.*

Здесь в (6) P_{H_n} — проектор из $\mathbf{H}$ в H_n.

Доказательство. Пусть $u \in \bar{U}$, тогда для всех $\varepsilon > 0$ существует номер N и линейная комбинация $\tilde{u}^{(N)} \in H_N$ такие, что

$$\|u - \tilde{u}^{(N)}\| < \varepsilon.$$

Пусть

$$\|u - u_{(n)}\| = \inf_{v \in H_n} \|u - v\|,$$

т. е. $u^{(n)}$ — наилучшее приближение элемента u на подпространстве H_n. Тогда на основании свойств элемента наилучшего приближения

в гильбертовых пространствах (см. § 1, гл. 4) будем иметь:

$$(u - u^{(n)}, v) = 0, \quad \forall v \in H_n,$$

т. е. $u^{(n)} = P_{H_n} u$ — проекция u на подпространство H_n. В силу приведенных рассуждений будут иметь место неравенства:

$$\| u^{(n)} - u \| = \| P_{H_n} u - u \| \leqslant \| P_{H_N} u - u \| \leqslant \| \tilde{u}^{(N)} - u \| < \varepsilon, \quad \forall\, n \geqslant N$$

и соотношение (6) доказано. Наоборот, если имеет место (6), то последовательность элементов $u^{(n)} = P_{H_n} u \in H_n \subset \overline{U}$ и сходится к некоторому элементу u. Поскольку все предельные элементы последовательностей из $\overline{U}$ принадлежат $\overline{U}$, то и $u \in \overline{U}$. Первое утверждение леммы доказано. Второе утверждение следует из первого.

Теорема 1. *Для того чтобы последовательность $\{u^{(n)}\}_{n=1}^{\infty}$, построенная с помощью процесса Ритца, была минимизирующей для функционала* (2), *необходимо и достаточно, чтобы $f \in \overline{V}$, где $\overline{V}$ — замкнутая линейная оболочка системы элементов $\{v_i\}_{i=1}^{\infty} = \{Au_i\}_{i=1}^{\infty}$.*

Доказательство. Согласно определению, последовательность $\{u^{(n)}\}_{n=1}^{\infty}$ будет минимизирующей для функционала (2), если

$$\lim_{n \to \infty} \| Au^{(n)} - f \| = 0.$$

Обозначим через V_n линейную оболочку элементов $v_i = Au_i$, $i = 1, 2, \ldots, n$, тогда по построению $v^{(n)} = Au^{(n)}$ является элементом наилучшего приближения для f в подпространстве V_n гильбертового пространства **H**. Следовательно,

$$v^{(n)} = Au^{(n)} = P_{V_n} v = AP_{H_n} u$$

и использование предыдущей леммы доказывает теорему.

Приведенная теорема 1 характерна именно для метода наименьших квадратов. В других вариантах метода моментов стремление к нулю невязки $\| Au^{(n)} - f \|$, вообще говоря, не имеет места, если даже $u^{(n)} \to u^*$.

Следствие 1. Если система элементов $\{Au_i\}_{i=1}^{\infty}$ — полная в **H**, то $\forall f \in R(A)$ $\lim\limits_{n \to \infty} \| Au^{(n)} - f \| = 0$.

Доказательство следует из теоремы 1.

На основании следствия 1 из § 2 будет справедлива следующая теорема:

Теорема 2. *Если $f \in \overline{V}$ и оператор A^*A — положительно определенный, где A^* — оператор, сопряженный к A, то*

$$\lim_{n \to \infty} \| u^{(n)} - u^* \| = 0.$$

Доказательство. Поскольку f принадлежит замкнутой линейной оболочке $\overline{V}$ элементов $\{Au_i\}_{i=1}^{\infty}$, то по теореме 1 последовательность

$u^{(n)}$, $n = 1, 2, \ldots$, минимизирующая для функционала (2). Из положительной определенности оператора $A^* A$ вытекает, что

$$G(u, u) = (Au, Au) = (A^*Au, u) \geqslant \mu \|u\|^2, \quad \forall u \in D(A),$$

где $\mu > 0$. Применяя далее следствие 1, § 2, получаем утверждение теоремы.

Замечание. Теорема 2 будет справедлива, если условие положительной определенности оператора A^*A заменить условием существования ограниченного оператора A^{-1}.

§ 4. МЕТОД РИТЦА

Из всех проекционно-вариационных методов метод Ритца пожалуй получил самое широкое распространение. Хотя он применим при условиях более ограничительных, чем большинство других проекционных методов, зато более прост с вычислительной точки зрения.

Перейдем к изложению метода Ритца. Пусть ищется решение операторного уравнения

$$Au = f, \tag{1}$$

где $u \in D(A) \subset \mathbf{H}$, $f \in R(A) \subset \mathbf{H}$. Здесь под $\mathbf{H}$ будем подразумевать вещественное гильбертово пространство. Относительно линейного оператора A введем следующие ограничения:

$$A = A^*, \text{ т. е. } (Au, v) = (u, Av), \quad \forall u, v \in D(A), \tag{2}$$

$$(Au, u) \geqslant \mu \|u\|^2, \quad \mu > 0, \quad \forall u \in D(A). \tag{3}$$

Согласно идее вариационных методов, заменим задачу решения уравнения (1) задачей минимизации функционала

$$\Phi(u) = (Au, u) - 2(f, u), \quad u \in D(A), \tag{4}$$

который называется *функционалом энергии*. Если ввести обозначения $G(u, v) = (Au, v)$; $l(u) = (f, u)$ $\forall u, v \in D(A)$ и, наконец, положить $D(G) = D(A)$, то получим функционал, рассмотренный в § 2 ($c = 0$). Ввиду того что функционал $l(u) = (u, f)$ ограничен на $D(G)$, а квадратичный функционал $G(u, u) = (Au, u)$ положительно определен, то функционал $\Phi(u)$ ограничен снизу. Следовательно, существует элемент $u^* \in D(G)$ и такой, что

$$\Phi_0 = \Phi(u^*) = \min_{u \in D(G)} \Phi(u). \tag{5}$$

Согласно теореме 2, § 2, этот минимум является строгим и

$$G(u^*, v) = l(v), \quad \forall v \in D(G)$$

или

$$G(u^*, v) - l(v) = (Au^*, v) - (f, v) = (Au^* - f, v) = 0, \quad \forall v \in D(G). \tag{6}$$

Если множество $D(G) = D(A)$ всюду плотно в $\mathbf{H}$, то из (6) следует, что $Au^* - f = 0$ (приложение, § 1), т. е. элемент, на котором достигается минимум функционала $\Phi(u)$, является одновременно решением

уравнения (1). Обратное утверждение следует из (9), § 2. Таким образом, доказана следующая лемма:

Лемма 1. *Пусть линейный оператор A обладает свойствами* (2), (3) *и имеет область определения $D(A)$, всюду плотную в* **H**. *Тогда для того, чтобы элемент $u^* \in D(A)$ был решением уравнения* (1), *необходимо и достаточно, чтобы имело место соотношение* (5).

Эта лемма лежит в основе метода Ритца. Вместо решения уравнения (1) займемся минимизацией функционала энергии (4), для чего применим процесс Ритца, описанный в § 2. Выберем координатную систему $\{u_i\}_{i=1}^{\infty} \in D(A)$ и построим последовательность подпространств H_n, являющихся линейными оболочками элементов u_i, $i = 1, 2, \ldots, n$. Далее с помощью процесса Ритца находим последовательность $u^{(n)} = \sum_{i=1}^{n} a_i^{(n)} u_i$. Коэффициенты $a_i^{(n)}$, $i = 1, 2, \ldots, n$, согласно § 2 определяются из системы

$$\sum_{i=1}^{n} a_i^{(n)} (Au_i, u_j) = (f, u_j), \qquad j = 1, 2, \ldots, n. \tag{7}$$

Существование и единственность элемента $u^{(n)}$, доставляющего минимум функционала $\Phi(u)$ на H_n, следует из леммы 4, § 2.

Сравнивая (7) с расчетными формулами метода моментов (см. § 1), видим, что они будут совпадать, если в методе моментов выбрать проекционную систему, совпадающую с координатной. Однако наличие дополнительных предположений (2) и (3) в методе Ритца позволяет получить более сильные результаты, которые нельзя получить в общем методе моментов.

Для дальнейшего изучения метода Ритца удобно ввести в рассмотрение *энергетическое пространство* $\mathbf{H}_A$ (приложение, § 1). Введем в $D(A)$ *энергетическое скалярное произведение*

$$(u, v)_A = (Au, v), \quad \forall u, v \in D(A) \tag{8}$$

(проверку справедливости аксиом скалярного произведения оставляем читателю) и *энергетическую норму*

$$\|u\|_A = [(u, u)_A]^{\frac{1}{2}} = [(Au, u)]^{\frac{1}{2}}. \tag{9}$$

Пополнение $D(A)$ в норме $\|\cdot\|_A$ обозначается через $\mathbf{H}_A$. Из определения $\mathbf{H}_A$ легко видеть, что $D(A)$ является всюду плотным множеством в $\mathbf{H}_A$.

Имеют место неравенства:

$$\|u\| \leqslant \frac{\|u\|_A}{\sqrt{\mu}}, \qquad \forall u \in \mathbf{H}_A; \tag{10}$$

$$\|u\|_A \leqslant \frac{\|Au\|}{\sqrt{\mu}}, \quad \forall u \in D_A. \tag{11}$$

Неравенство (10) следует из определения энергетической нормы и неравенства (3) для всех $u \in D(A)$. Для остальных элементов из $\mathbf{H}_A$ оно устанавливается предельным переходом. Неравенство (11) следует из неравенства Коши — Буняковского

$$[\|u\|_A]^2 = (Au, u) \leqslant \|Au\| \|u\|$$

и уже доказанного соотношения (10).

Рассмотрим вопрос о том, чем будут являться элементы последовательности $\{u^{(n)}\}_{n=1}^{\infty}$ в терминах введенного энергетического пространства $\mathbf{H}_A$. Имеем:

$$\Phi(u) = (Au, u) - 2(f, u) = \|u\|_A^2 - 2(Au^*, u) = \\ = \|u\|_A^2 - 2(u^*, u)_A, \quad \forall u \in D(A),$$

или

$$\Phi(u) = \|u - u^*\|_A^2 - \|u^*\|_A^2, \quad \forall u \in D(A). \tag{12}$$

Поскольку $u^{(n)}$ доставляет минимум функционалу (12) на подпространстве H_n, то $u^{(n)} = P_{H_n}^{\mathbf{H}_A} u^*$ является проекцией u^* на H_n в энергетическом пространстве $\mathbf{H}_A$.

Приведем теорему, указывающую условия, при выполнении которых последовательность $\{u^{(n)}\}$ будет минимизирующей.

Теорема 1. *Для того чтобы последовательность $\{u^{(n)}\}$ была минимизирующей, необходимо и достаточно, чтобы $u^* \in K^{(\infty)}$, где $K^{(\infty)}$ — замкнутая в метрике $\mathbf{H}_A$ линейная оболочка координатной системы $\{u_i\}_{i=1}^{\infty}$.*

Доказательство. На основании леммы 1, § 3, для того чтобы $u^* \in K^{(\infty)}$, необходимо и достаточно, чтобы $P_{H_n}^{\mathbf{H}_A} u^* \to u^*$. Так как $P_{H_n}^{\mathbf{H}_A} u^* = u^{(n)}$, то теорема доказана.

Замечание. Из сходимости $u^{(n)}$ к u^* в метрике энергетического пространства $\mathbf{H}_A$ на основании неравенства (10) вытекает сходимость $u^{(n)}$ к u^* в метрике $\mathbf{H}$.

Теорема 2. *Пусть координатная система $\{u_i\}_{i=1}^{\infty}$ полна в $\mathbf{H}_A$.*

Тогда
$$\lim_{n\to\infty} \|u^{(n)} - u^*\|_A = \lim_{n\to\infty} \|u^{(n)} - u^*\| = 0.$$

Доказательство следует из теоремы 1.

Отметим, что метод Ритца для уравнения

$$A^*Au = A^*f \tag{13}$$

эквивалентен методу наименьших квадратов. Действительно,

$$\Phi(u) = (A^*Au, u) - 2(A^*f, u) = (Au, Au) - 2(f, Au) = \\ = (Au - f, Au - f) - (f, f) = \|Au - f\|^2 - \|f\|^2,$$

т. е. функционалы в методе Ритца и в методе наименьших квадратов для случая уравнения (13) отличаются на постоянное слагаемое.

Глава 6

РАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Среди приближенных методов решения задач математической физики особое положение занимают разностные методы, в которых дифференциальная краевая задача заменяется разностной краевой задачей (разностной схемой).

§ 1. ОБЩИЕ ВОПРОСЫ МЕТОДА СЕТОК

Пусть A — линейный оператор с областью определения $D(A)$ и областью значений $R(A)$ соответственно из банаховых пространств $\tilde{\mathbf{B}}_1$ и $\tilde{\mathbf{B}}_2$. Рассмотрим задачу нахождения решения уравнения

$$Au = f, \quad u \in \mathbf{B}_1, \quad f \in \mathbf{B}_2 \ (\mathbf{B}_1 \subseteq \tilde{\mathbf{B}}_1, \quad \mathbf{B}_2 \subseteq \tilde{\mathbf{B}}_2). \tag{1}$$

Говорят, что задача (1) поставлена корректно (на $\tilde{\mathbf{B}}_1$, $\tilde{\mathbf{B}}_2$), если для любого $f \in \mathbf{B}_2$ она имеет единственное решение, непрерывно зависящее от f.

Основной целью всякого приближенного метода является получение решения исходной задачи с заданной точностью $\varepsilon > 0$ за конечное число действий.

Конечно-разностный (сеточный) метод решения будем характеризовать заданием:

1) множества некоторых нормированных пространств с нормой $\|h\|$ такого, что оно содержит сходящуюся к нулю последовательность, но не содержит нуля;

2) семейства операторных уравнений

$$A_h v_h = \varphi_h, \quad v_h \in \mathbf{B}_{1h}, \quad \varphi_h \in \mathbf{B}_{2h}, \tag{2}$$

где $A_h : \mathbf{B}_{1h} \to \mathbf{B}_{2h}$ — линейные операторы;

3) способа сравнения элементов пространств $\mathbf{B}_1$ и $\mathbf{B}_{1h}$.

В случае, если для приближенного решения уравнения (1) применяется конечно-разностный метод, то за h можно принять вектор шагов сетки; операторное уравнение (2), зависящее от параметра h, называют *разностной схемой;* $\mathbf{B}_{1h}$ — пространством сеточных функций v_h, определенных на множестве $\overline{\Omega}_h$ узлов сетки.

1. Погрешность метода. Сходимость

Способ сравнения элементов $\mathbf{B}_1$ и $\mathbf{B}_{1h}$ определим при помощи семейства линейных операторов P_{1h}, отображающих $\mathbf{B}_1$ в $\mathbf{B}_{1h}$.

Пусть существуют линейные операторы $P_{1h} : \mathbf{B}_1 \to \mathbf{B}_{1h}$ такие, что

$$P_{1h} u = u_h \in \mathbf{B}_{1h}, \text{ если } u \in \mathbf{B}_1, \tag{3}$$

и выполнены условия согласования норм $\|\cdot\|_1$ и $\|\cdot\|_{1h}$ соответственно пространств $\mathbf{B}_1$ и $\mathbf{B}_{1h}$, т. е.

$$\lim_{\|h\| \to 0} \|P_{1h} u\|_{1h} = \|u\|_1. \tag{4}$$

Тогда можно говорить о сходимости элементов v_h пространства $\mathbf{B}_{1h}$ к элементу u из пространства $\mathbf{B}_1$, если

$$\|y_h\|_{1h} = \|v_h - P_{1h} u\|_{1h} = \|v_h - u_h\|_{1h} \to 0 \text{ при } \|h\| \to 0. \tag{5}$$

Таким образом, если u — решение уравнения (1), v_h — решение уравнения (2), то *под погрешностью разностной схемы* (2) понимают величину

$$y_h = v_h - P_{1h} u = v_h - u_h. \tag{6}$$

Говорят, что решение разностной схемы (2) сходится к решению задачи (1) со скоростью $O(|h|^m)$ (или имеет порядок точности $m > 0$), если при достаточно малом $|h| < h_0$ выполняется неравенство

$$\|y_h\|_{1h} = \|v_h - u_h\| = O(|h|^m), \text{ или } \|y_h\|_{1h} \leqslant \varkappa |h|^m, \tag{7}$$

где $\varkappa > 0$ — постоянная, не зависящая от h.

Таким образом, порядок точности m зависит от скорости сходимости $\|y_h\|_{1h}$ к нулю.

Следует отметить, что это не единственный способ сравнения элементов, принадлежащих пространствам $\mathbf{B}_1$ и $\mathbf{B}_{1h}$. Иногда, наоборот, сеточную функцию v_h доопределяют во все точки области Ω определения функции u. Доопределенную таким образом функцию v обозначим через $v(x)$. Очевидно, разность $v(x) - u(x) \in \mathbf{B}_1$. Тогда можно говорить о сходимости элементов v_h пространства $\mathbf{B}_{1h}$ к элементу u пространства $\mathbf{B}_1$, если

$$\|v(x) - u(x)\|_{\mathbf{B}_1} \to 0 \text{ при } |h| \to 0.$$

Однако чаще для исследования сходимости разностных схем применяется первый подход.

2. Корректность

Определение. Разностная схема (2) *корректна,* если она однозначно разрешима и устойчива или:

1) для любых $\varphi_h \in \mathbf{B}_{2h}$ существует единственное решение v_h уравнения (2) (схема (2) однозначно разрешима);

2) решение v_h уравнения (2) непрерывно (и притом равномерно по h), зависит от φ_h (схема (2) устойчива на $(\mathbf{B}_{1h}, \mathbf{B}_{2h})$).

Для линейных операторов A_h однозначная разрешимость схемы означает, что существует оператор A_h^{-1} и

$$v_h = A_h^{-1}\varphi_h, \tag{8}$$

а устойчивость схемы означает, что $A_h^{-1} : \mathbf{B}_{2h} \to \mathbf{B}_{1h}$ равномерно по h ограничен, т. е.

$$\|A_h^{-1}\| \leqslant \varkappa, \tag{9}$$

где $\varkappa > 0$ — постоянная, не зависящая от h. Поэтому из (8) и (9) получаем, что для устойчивых разностных схем имеет место оценка (при любых $\varphi_h \in \mathbf{B}_{2h}$)

$$\|v_h\|_{1h} \leqslant \|A_h^{-1}\| \|\varphi_h\|_{2h} \leqslant \varkappa \|\varphi_h\|_{2h},$$

или

$$\|v_h\|_{1h} \leqslant \varkappa \|\varphi_h\|_{2h}. \tag{10}$$

Неравенство вида (10) называют *априорной оценкой* для схемы (2). Получение таких оценок и составляет основное содержание теории устойчивости разностных схем.

Корректность разностной схемы является, вообще говоря, внутренним свойством схемы (2), не связанным с корректностью исходной задачи (1). Но на самом деле эти понятия взаимосвязаны. Например,

если кроме условий (3), (4) будут выполняться условия существования линейных операторов $P_{2h} : \mathbf{B}_2 \to \mathbf{B}_{2h}$ таких, что

$$P_{2h}f = f_h \in \mathbf{B}_{2h}, \text{ если } f \in \mathbf{B}_2, \tag{11}$$

и условия согласования норм пространств $\mathbf{B}_2$ и $\mathbf{B}_{2h}$

$$\lim_{\|h\| \to 0} \| P_{2h}f \|_{2h} = \| f \|_2, \tag{12}$$

то из (10) следует справедливость неравенства

$$\| u \|_1 \leqslant \varkappa \| f \|_2, \tag{13}$$

т. е. устойчивость исходной задачи.

3. Аппроксимация

Пусть определены семейства линейных операторов вида (3) и (11) и выполнены условия согласования норм (4) и (12). Запишем уравнение, которому удовлетворяет погрешность

$$y_h = v_h - u_h \tag{14}$$

разностной схемы (2). Подставляя $v_h = y_h + u_h$ в (2), получим

$$A_h y_h = \psi_h, \quad y_h \in \mathbf{B}_{1h}, \quad \psi_h \in \mathbf{B}_{2h}, \tag{15}$$

где $\psi_h = \varphi_h - A_h u_h$ — погрешность аппроксимации уравнения (1) разностной схемой (2) на решении u исходной задачи (1).

Определения: 1. Разностная схема (2) обладает m-м порядком аппроксимации на элементе $u \in \mathbf{B}_1$, если при всех достаточно малых $|h| \leqslant h_0$

$$\| \psi_h \|_{2h} = \| \varphi_h - A_h u_h \|_{2h} = O(|h|^m). \tag{16}$$

2. Оператор A_h аппроксимирует оператор A с порядком $m > 0$, если для любого $u \in \mathbf{B}_1$ справедлива оценка

$$\| A_h (P_{1h} u) - P_{2h} (Au) \|_{2h} = O(|h|^m). \tag{17}$$

3. φ_h аппроксимирует f с порядком $m > 0$, если для любого $f \in \mathbf{B}_2$ при всех достаточно малых $|h| \leqslant h_0$ имеет место оценка

$$\| \varphi_h - P_{2h} f \|_{2h} \leqslant O(|h|^m). \tag{18}$$

Если выполнены условия (17) и (18), то разностная схема (2) будет иметь m-й порядок аппроксимации на решении u уравнения (1).

В самом деле,

$$\psi_h = \varphi_h - A_h u_h = \varphi_h - A_h u_h + P_{2h}(Au) - P_{2h} f,$$

откуда, если выполняются условия (16) и (17), то

$$\| \psi_h \|_{2h} \leqslant \| \varphi_h - P_{2h} f \|_{2h} + \| A_h (P_{1h} u) - P_{2h} (Au) \|_{2h} = O(|h|^m). \tag{19}$$

Отметим, что в зависимости от выбора $\mathbf{B}_{2h}$ порядок аппроксимации может изменяться или даже аппроксимации может и не быть.

Если разностная схема (2) корректна, а y_h — решение задачи (15), то из (10) имеем:

$$\| y_h \|_{1h} \leqslant \varkappa \| \psi_h \|_{2h}. \tag{20}$$

Из неравенств (20), (16) и (7) следует, что решение v_h задачи (2) будет сходиться к решению u задачи (1), причем порядок сходимости будет совпадать с порядком аппроксимации разностной схемы (2) на решении u уравнения (1). Таким образом, из неравенства (20) следует такая теорема:

Теорема 1. *Если схема* (2) *корректна и аппроксимирует задачу* (1), *то решение задачи* (2) *при* $|h| \to 0$ *сходится к решению и задачи* (1), *причем порядок точности схемы совпадает с порядком аппроксимации на решении и уравнения* (1), *или, как принято говорить, из устойчивости и аппроксимации схемы* (2) *следует ее сходимость к решению и уравнения* (1).

Таким образом, опираясь на теорему 1, задачу исследования сходимости разностной схемы (2) (оценки скорости сходимости) можно разбить на две более простые задачи: 1) исследование порядка аппроксимации разностной схемой исходного операторного уравнения; 2) исследование корректности разностной схемы (2).

Отметим, что доказательство корректности разностной схемы связано не только с исследованием сходимости решения v_h задачи (2) к решению u исходного уравнения (1), но также с физической детерминированностью задачи (2) и возможностью ее решения по приближенным исходным данным. Поэтому, если нужно выяснить пригодность той или иной разностной схемы для решения задачи, мало знать, что схема устойчива. Вообще желательно знать примерно величину коэффициента $\varkappa$. Кроме того, в связи с необходимостью получения для решения второй задачи априорных оценок вида (10) встает вопрос не только о выборе согласованных норм соответственно в пространствах $\mathbf{B}_1$ и $\mathbf{B}_{1h}$, $\mathbf{B}_2$ и $\mathbf{B}_{2h}$, но также о связи между выбором норм в $\mathbf{B}_{2h}$ ($\mathbf{B}_2$) сравнительно с нормами в $\mathbf{B}_{1h}$ ($\mathbf{B}_1$). Для такого выбора норм в $\mathbf{B}_{2h}$ нет общего правила, однако их нужно стремиться выбирать так, чтобы порядок аппроксимации оказался как можно выше, но устойчивость при этом еще не утерялась.

Решение задачи (1) (исследование порядка аппроксимации разностной схемой исходного операторного уравнения), как правило, не связано с принципиальными трудностями и для широкого круга задач сводится фактически к разложению решения в ряд Тейлора.

Для доказательства корректности разностной схемы нельзя указать единого подхода.

Некоторые приемы исследования корректности основаны на использовании информации о свойствах оператора A_h разностной задачи (2) (например, положительной определенности, самосопряженности), на изучении спектра разностных операторов, на использовании разностного принципа максимума, использовании неравенств между операторными коэффициентами разностных схем.

§ 2. О ПОСТРОЕНИИ СЕТОК, СЕТОЧНЫХ ФУНКЦИЙ И СОГЛАСОВАННЫХ НОРМ

Разностную схему вида (2) строят на сетке Ω_h, покрывающей область Ω определения функции u. Замена области Ω некоторым конечным множеством точек, лежащих в этой области и являющихся областью определения функций дискретного аргумента, называется разностной сеткой. Рассмотрим примеры наиболее применяемых видов сеток Ω_h, методов проектирования функций $u \in \mathbf{B}_1$ на пространства сеточных функций $\mathbf{B}_{1h}$ с областью определения Ω_h и способов выбора норм, удовлетворяющих условию согласования (4), § 1.

Примеры наиболее употребительных видов сеток.

Пример 1. Равномерная сетка на отрезке $\Omega = \{a \leqslant x \leqslant b\}$.

Разобьем отрезок $[a, b]$ на n равных частей точками $x_i = a + ih$ — узлами сетки. При этом $h = \dfrac{b-a}{n}$ — шаг сетки (параметр сетки), характеризует плотность распределения узлов

$$\Omega_h = \left\{x_i, \quad x_i = a + ih, \quad h = \frac{b-a}{n}, \quad i = \overline{0, n}\right\}. \tag{1}$$

Пример 2. Неравномерная сетка на отрезке $\Omega = \{a \leqslant x \leqslant b\}$.

Разобьем отрезок на n частей точками (рис. 1):

$$x_0 = a, \quad x_i = x_{i-1} + h_i \quad (i = \overline{1, n-1}), \quad x_n = b.$$

При этом шаги сетки h_i удовлетворяют условию

$$\sum_{i=1}^{n} h_i = b - a;$$

$$\Omega_h = \{x_i, \quad x_i = x_{i-1} + h_i, \quad x_0 = a, \quad i = \overline{0, n}\}. \tag{2}$$

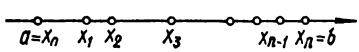


Рис. 1

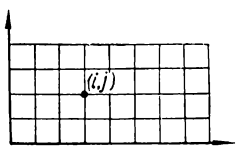


Рис. 2

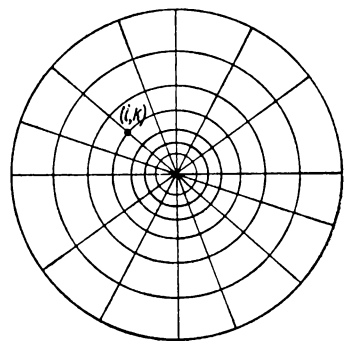


Рис. 3

Пример 3. Равномерная сетка на плоскости.

Пусть $\Omega = \{0 \leqslant x \leqslant a, \ 0 \leqslant y \leqslant b\}$. Разобьем отрезки $[0, a]$ и $[0, b]$ соответственно на n и m частей и через точки

$$x_i = ih \quad \left(i = \overline{0, n}, \quad h = \frac{a}{n}\right), \quad y_j = j\tau \quad \left(j = \overline{0, m}, \quad \tau = \frac{b}{m}\right) \tag{3}$$

проведем прямые, параллельные координатным осям. В результате пересечения прямых получим точки (узлы) (x_i, y_i), которые и образуют сетку (рис. 2).

Пример 4. Неравномерная сетка на плоскости (изометрическая).

Пусть $\Omega = \{r \leqslant \rho \leqslant R,\ 0 \leqslant \theta \leqslant 2\pi\}$. Покроем область Ω лучами $\theta_k = kh_1$ $(h_1 = \frac{2\pi}{n},\ k = \overline{0, n-1})$ и окружностями

$$\rho_i = re^{ih}\left(h = \frac{1}{m}\ln\frac{R}{r},\quad i = \overline{0, m}\right). \tag{4}$$

Точки пересечения окружностей ρ_i и лучей θ_k назовем узлами сетки $(\rho_i, \theta_k) \sim (i, k)$ (рис. 3).

Пример 5. Треугольная сетка на плоскости.

Покроем область Ω прямыми (рис. 4):

$$y = \frac{\sqrt{3}}{2}jh;\quad y = \sqrt{3}x + \frac{\sqrt{3}}{2}jh;\quad y = -\sqrt{3}x + \frac{\sqrt{3}}{2}jh, \tag{5}$$

$$h > 0,\quad j = 0,\ \pm 1,\ \pm 2,\ \ldots.$$

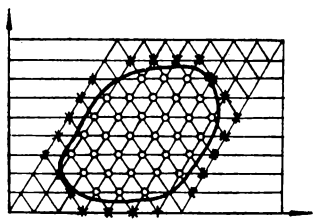


Рис. 4

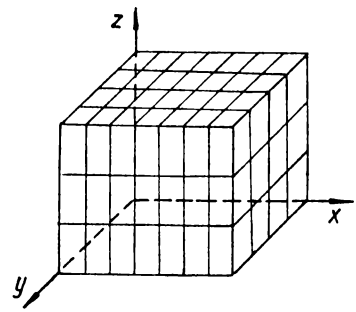


Рис. 5

Пример 6. Произвольная сетка.

Область Ω покрывается произвольным множеством точек, расположенных внутри Ω.

Пример 7. Сеточный параллелепипед.

В трехмерном пространстве построим равномерную по каждой из пространственных переменных ортогональную сетку, узлы которой определяются по формулам:

$$x_l = x_0 + ih_1;\quad y_k = y_0 + kh_2;\quad z_j = z_0 + jh_3, \tag{6}$$

где h_1, h_2, h_3 — шаги сетки соответственно по осям Ox, Oy, Oz. Сеточный параллелепипед Ω_h будет определяться совокупностью узлов x_l, y_k, z_j, $(i = \overline{0, n},\ k = \overline{0, m},\ j = \overline{0, l})$ (рис. 5).

Обычно рассматривается множество сеток, зависящих от параметра h, причем в случае неравномерной сетки или многомерной сетки под h понимают вектор $h = (h_1, h_2, \ldots, h_p)$.

Множество сеточных функций v_h, областью определения которых является множество сеточных областей Ω_h, образуют множество пространств $\mathbf{B}_{1h}$. Для сравнения функций $u \in \mathbf{B}_1$ с функциями $v_h \in \mathbf{B}_{1h}$ в пространствах $\mathbf{B}_1$ и $\mathbf{B}_{1h}$ нужно строить согласованные нормы (см. формула (4), § 1), выбор которых обычно характеризуется классом функций, которому принадлежит решение исходной задачи, способом выбора оператора проектирования, свойствами оператора A. Приведем простейшие примеры согласованных норм, широко используемых в теории разностных схем.

Пример 1. Рассмотрим пространство $\mathbf{C}[a, b]$ (пространство непрерывных на $[a, b]$ функций) и определим норму соотношением

$$\|u\|_{\mathbf{C}} = \max_{a \leqslant x \leqslant b} |u(x)|. \tag{7}$$

Тогда, если положить

$$P_h u = u_h = u(x) \quad \forall x \in \Omega_h, \tag{8}$$

то в качестве сеточного аналога нормы (7) можно принять величину

$$\|u\|_{\mathbf{C}_h} = \max_{x \in \Omega_h} |u(x)| \quad \text{(сеточный аналог нормы в } \mathbf{C}). \tag{9}$$

Очевидно, условие согласования норм (7) и (9) выполняется.

Пример 2. Рассмотрим множество непрерывно дифференцируемых на $[a, b]$ функций и введем норму при помощи соотношения

$$\|u\| = \max \left[\max_{a \leqslant x \leqslant b} |u(x)|, \ \max_{a \leqslant x \leqslant b} |u'(x)|\right]. \tag{10}$$

Норму в пространстве сеточных функций, определенных на сетке Ω_h вида (1), можно определить следующим образом:

$$\|u_h\|_h = \max \left[\max_{x \in \Omega_h} |u_h(x)|, \ \max_{x \in \Omega_h} \frac{|u_h(x+h) - u_h(x)|}{h}\right]. \tag{11}$$

Если оператор проектирования определить при помощи соотношения (8), то норма (11) будет согласована с нормой (10). Норма вида (11) применяется в основном при исследовании разностных схем, связанных с обыкновенными дифференциальными уравнениями.

Пример 3. Рассмотрим пространство $\mathbf{L}_2[a, b]$. Построим сетку Ω_h вида (1) и положим

$$P_h u = u_h(x) = \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} u(t)\, dt \quad \forall x \in \Omega_h. \tag{12}$$

Тогда норма

$$\|u_h\|_h = \left(\sum_{x \in \Omega_h} h u^2(x)\right)^{\frac{1}{2}} \tag{13}$$

будет согласованным сеточным аналогом нормы в $\mathbf{L}_2[a, b]$ вида

$$\|u\| = \left[\int_a^b u^2(t)\, dt\right]^{\frac{1}{2}}. \tag{14}$$

Очевидно, нормы, удовлетворяющие условию (4), § 1, можно ввести в пространствах $\mathbf{B}_1$ и $\mathbf{B}_{1h}$ различными способами, причем их выражение часто зависит, как уже указывалось, и от свойств оператора A. Так, в случае, если оператор A положительный и самосопряженный, то на $D(A) \in \mathbf{B}_1$ можно построить энергетическое гильбертово пространство $\mathbf{H}_A$, построив замыкание $D(A)$ в смысле сходимости по так называемой энергетической норме, которая определяется из соотношения

$$\|u\|_A = \sqrt{(Au, u)}. \tag{15}$$

В пространстве $\mathbf{H}_A$ в этом случае можно построить сеточный аналог нормы (15), удовлетворяющий условию (4), § 1.

Если оператор A положительный, самосопряженный и A^{-1} существует, то для исследования корректности разностной задачи часто пользуются негативной нормой

$$\|u\|_{A^{-1}} = \sqrt{(A^{-1}u,\ u)}.$$

Выбор вида нормы во всех случаях определяется одной целью: исследованием сходимости построенной разностной схемы.

Однако, если исследование сходимости разностной схемы сводится к построению оценок вида (20), § 1, то вопрос выбора нормы в пространстве $\mathbf{B}_{2h}$ тоже весьма существенный.

В самом деле, пусть в пространстве $\mathbf{B}_{2h}$ введена некоторым образом норма $\|\cdot\|_{2h}$ и определен порядок аппроксимации разностной схемы (2), § 1, равный m, т. е.

$$\|\varphi_h - A_h u_h\|_{2h} \leqslant c\,|h|^m \tag{16}$$

($c > 0$ — постоянная, не зависящая от h). Если разностная схема при этом будет устойчива, т. е.

$$\|v_h\|_{1h} \leqslant \varkappa \|\varphi_h\|_{2h}, \tag{17}$$

то ее решение будет сходиться к решению уравнения (1), § 1, со скоростью $O(|h|^m)$. Очевидно, если вместо нормы $\|\cdot\|_{2h}$ ввести норму $\|\cdot\|_{2h}^{(1)} = |h|^l\ \|\cdot\|_{2h}$ ($l > 0$), то порядок аппроксимации разностной схемы увеличится:

$$\|\varphi_h - A_h u_h\|_{2h}^{(1)} \leqslant c\,|h|^{m+l}. \tag{18}$$

Однако при таком выборе нормы из неравенства (17) следует, что

$$\|v_h\|_{1h} \leqslant \frac{\varkappa}{h^l} \|\varphi_h\|_{2h}^{(1)}, \tag{19}$$

т. е. при новом выборе нормы устойчивости может не быть.

Если вместо нормы $\|\cdot\|_{2h}$ ввести норму

$$\|\cdot\|_{2h}^{(2)} = \frac{1}{h^l}\|\cdot\|_{2h} \qquad (m > l > 0), \tag{20}$$

то при этом будет иметь место устойчивость исходной разностной схемы, так как

$$\|v_h\|_{1h} \leqslant c_1 \|\varphi_h\|_{2h}^{(2)} \quad (c_1 > h^l \varkappa), \tag{21}$$

но порядок аппроксимации уменьшается и будет равен $m - l$, т. е. уменьшится порядок точности разностной схемы. Таким образом, выбирать норму в $\mathbf{B}_{2h}$ нужно так, чтобы сохранялась устойчивость и имела место аппроксимация как можно более высокого порядка. Однако выбрать норму так, чтобы имела место и аппроксимация, и устойчивость, не всегда удается, ибо тогда всякая разностная схема была бы сходящейся.

§ 3. ВОПРОСЫ КОНСТРУИРОВАНИЯ РАЗНОСТНЫХ СХЕМ

Метод сеток решения операторного уравнения (1), § 1, прежде всего связан с построением разностной схемы и исследованием порядка ее аппроксимации.

Разностная схема, аппроксимирующая дифференциальную задачу, может быть построена неединственным образом. Поэтому возникает задача построения разностных схем, оптимальных в определенном смысле. По существу требуется построить такую разностную схему, которая на сравнительно грубых сетках (а не при стремлении шага к нулю) гарантировала бы разумный уровень точности для найденного приближенного решения. Чтобы добиться этого, разностная схема должна отражать основные свойства исходного операторного уравнения (такие, например, как симметричность, кососимметричность, знакопеременность, принцип максимума, разностные аналоги типичных априорных оценок для дифференциальных уравнений и т. д.). Однако построение таких разностных схем сопряжено со значительными трудностями. Поэтому в основу различных подходов при конструировании разностных схем положены те свойства операторных уравнений, которые считаются важнейшими и должны быть сохранены у их разностных аналогов.

Отметим некоторые особенности аппроксимации линейных операторов, связанных с дифференциальными краевыми задачами.

Пусть в некоторой области Ω с границей Γ поставлена дифференциальная краевая задача. Это значит, что требуется найти функцию $u(p)$, которая в каждой точке P области Ω удовлетворяет дифференциальному уравнению

$$Lu = \varphi_0(X) \tag{1}$$

и одному или нескольким условиям на границе

$$l_i u|_{\Gamma_i} = \varphi_i(X) \quad (i = \overline{1, s}), \tag{2}$$

где Γ_i — часть границы Γ; l_i u — дифференциальные операторы краевых условий; $u \in \mathbf{H}$, $\varphi_i(P) \in \mathbf{H}_i$ $(i = \overline{0, s})$; $\mathbf{H}$, $\mathbf{H}_0$, $\mathbf{H}_i$ $(i = \overline{1, s})$ — гильбертовы пространства функций с областями определения элементов соответственно в $\Omega + \Gamma$, Ω, Γ_i $(i = \overline{1, s})$. Условия (2) носят название граничных (краевых) условий. Если задачу (1), (2) записать в виде

$$Au = f(X), \tag{3}$$

то

$$Au = \begin{cases} Lu, & \text{если } X \in \Omega; \\ l_i u, & \text{если } X \in \Gamma_i, \quad (i = \overline{1, s}); \end{cases}$$

$$f = \begin{cases} \varphi_0(X), & \text{если } X \in \Omega; \\ \varphi_i(X), & \text{если } X \in \Gamma_i, \ (i = \overline{1, s}). \end{cases} \tag{4}$$

Таким образом, исходная задача как бы разбивается на несколько подсистем, т. е. $Au = f$ представима в виде

$$\begin{cases} l_0 u = \varphi_0, \\ l_1 u = \varphi_1, \\ l_2 u = \varphi_2, \\ \cdot\ \cdot\ \cdot\ \cdot \\ l_s u = \varphi_s, \end{cases} \tag{5}$$

где для удобства введено обозначение $Lu = l_0 u$.

Аналогично разностную схему

$$A_h u_h = f_h \quad (u_h \in \mathbf{H}_h, \quad f_h \in \mathbf{H}_{f_h}) \tag{6}$$

в этом случае удобно разбить на несколько подсхем и записать в виде

$$\begin{aligned} l_{0h} u_h &= \varphi_{0h}, \\ l_{1h} u_h &= \varphi_{1h}, \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ l_{sh} u_h &= \varphi_{sh}. \end{aligned} \tag{7}$$

Если правую часть каждой подсхемы (7) считать элементом гильбертова пространства $\mathbf{H}_{\varphi_{ih}}$, то, выбирая согласовано нормы в пространствах $\mathbf{H}_{\varphi_{ih}}$ и $\mathbf{H}_{f_h}$

$$\| f_h \|_{\mathbf{H}_{f_h}} = \max_k \| \varphi_{kh} \|_{\mathbf{H}_{\varphi_{kh}}}, \tag{8}$$

можно говорить о порядке аппроксимации разностной схемы (7) на решении u исходной задачи (3). Для этого исходя из аппроксимации каждой подсхемой (7) соответствующего операторного уравнения (5) на решении u уравнения (3) находим оценку для

$$\| \psi_{kh} \|_{\mathbf{H}_{\varphi_{kh}}} = \| \varphi_{kh} - [l_k u]_h \|_{\mathbf{H}_{\varphi_{kh}}} \quad (k = \overline{0, s}). \tag{9}$$

Порядок аппроксимации разностной схемы (7) на решении u задачи (3) будет равен минимальному из порядков аппроксимаций разностных схем

$$l_{kh} u_h = \varphi_{kh}$$

на решении u исходного операторного уравнения.

Таким образом, порядок аппроксимации разностной схемы в целом зависит не только от порядка аппроксимации дифференциального оператора Lu, но также и порядков аппроксимаций операторов краевых условий $l_i u$ $(i = \overline{1, s})$. Методы аппроксимации оператора Lu могут отличаться от методов аппроксимации граничных операторов. Более того, часто различными бывают даже методы аппроксимации оператора Lu в зависимости от того, лежит ли точка X, в которой строится локальная аппроксимация оператора Lu, вблизи границы Γ_h области Ω_h или внутри области Ω_h. Методы аппроксимации граничных условий могут отличаться как по характеру расположения узлов, лежащих вблизи границы сеточной области относительно граничных узлов Γ,

так и по порядку погрешности аппроксимации. При этом значительно различаются случаи, когда оператор граничных условий содержит производные и когда он их не содержит. Таким образом, анализ краевых условий (особенно для многомерных задач) требует особого внимания. Поэтому наряду с методами аппроксимации операторов дифференциальных уравнений отдельно остановимся на некоторых методах аппроксимации краевых условий.

Выделим несколько конструктивных подходов к построению конечно-разностных аппроксимаций для дифференциальных операторов:

1) метод формальной замены производных конечно-разностными выражениями (использование формул численного дифференцирования);

2) метод неопределенных коэффициентов;

3) метод интегральных тождеств (интегро-интерполяционный метод);

4) вариационный метод построения разностных схем.

Первые два метода построения разностных схем основаны на использовании ряда Тейлора для достаточно гладких функций и приводят, как правило, к сохранению «локальных» свойств дифференциальных уравнений. Метод интегральных тождеств выделяет некоторые «интегральные» свойства дифференциальной задачи (например, соответствующий закон сохранения). Такие схемы получили название консервативных, и они оказываются пригодными и для дифференциальных уравнений с разрывными коэффициентами. В вариационных методах построения разностных схем на первое место ставится требование, чтобы разностная схема минимизировала соответствующий дискретный функционал.

1. Метод формальной замены производных конечно-разностными выражениями. Метод неопределенных коэффициентов

Для построения разностных схем, связанных с дифференциальными краевыми задачами, чаще всего применяются формулы численного дифференцирования и метод неопределенных коэффициентов. Оба эти метода принадлежат к числу наиболее простых методов конструирования разностных схем.

Если функции, являющиеся решением исходной задачи (1), (2), достаточно гладкие, то аппроксимация при таких подходах в норме **C** обычно устанавливается без труда, чего нельзя вообще сказать в этом случае об исследовании корректности разностной задачи.

Поскольку задачей данного учебного пособия является ознакомление читателя с некоторыми принципиальными вопросами метода сеток, то мы будем рассматривать весьма простые постановки задач, на примере которых исследуемые методы удобно иллюстрируются.

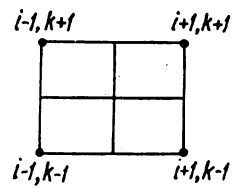
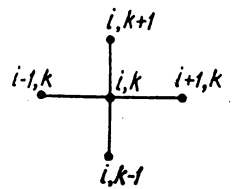
Проиллюстрируем первые два подхода на ряде примеров, помещенных в табл. 4, предполагая, что областью определения оператора L является множество достаточно гладких в области Ω функций. Пространственные переменные условимся обозначать через $x_1, x_2, \ldots$ $\ldots, x_p$, а временную переменную через t. Говорят, что разностный

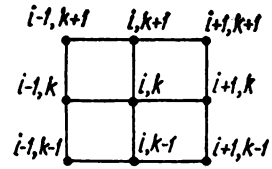
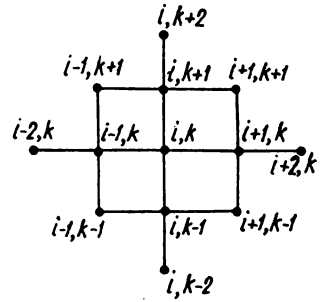
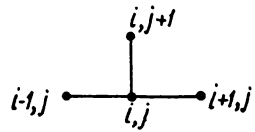
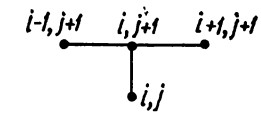
Таблица 4

Вид дифференциального оператора	Формула и принятые обозначения	Ближайший не обращающийся в нуль член разложения в ряд Тейлора	Порядок локальной аппроксимации по переменным: t	Порядок локальной аппроксимации по переменным: X	Шаблон разностной схемы	Номер формулы
$u^{\mathrm{I}}(x_i)$	Аппроксимация на равномерной сетке $(x_i = ih)$ $u(x_i) = u_i$					
	$u_{x,i} = \dfrac{u_{i+1} - u_i}{h}$	$\dfrac{h}{2} u^{\mathrm{II}}(x_i) + O(h^2)$		1	i $i+1$	(16)
	$u_{\bar{x},i} = \dfrac{u_i - u_{i-1}}{h}$	$-\dfrac{h}{2} u^{\mathrm{II}}(x_i) + O(h^2)$		1	$i-1$ i	(17)
	$u_{\sigma x,i} = \sigma u_{x,i} + (1-\sigma) u_{\bar{x},i}$ σ — любое вещественное число	$(\sigma - \dfrac{1}{2}) h u^{\mathrm{II}}(x_i) + O(h^2)$		1	$i-1$ i $i+1$	(18)
	$u_{\overset{\circ}{x},i} = \dfrac{u_{i+1} - u_{i-1}}{2h}$	$\dfrac{h^2}{6} u^{\mathrm{III}}(x_i) + O(h^4)$		2	$i-1$ $i+1$	(19)
	$u_{x,i+0} = \dfrac{-3u_i + 4u_{i+1} - u_{i+2}}{2h}$	$-\dfrac{h^2}{3} u^{\mathrm{III}}(x_i) + O(h^3)$		2	i $i+1$ $i+2$	(20)
	$u_{x,i-0} = \dfrac{3u_i - 4u_{i-1} + u_{i-2}}{2h}$	$\dfrac{h^2}{3} u^{\mathrm{III}}(x_i) + O(h^3)$		2	$i-2$ $i-1$ i	(21)
$u^{\mathrm{II}}(x_i)$	$u_{\bar{x}x,i} = \dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$ или $\Lambda u = \dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$	$\dfrac{h^2}{12} u^{\mathrm{IV}}(x_i) + O(h^4)$		2	$i-1$ i $i+1$	(22)
$u^{\mathrm{III}}(x_i)$	$u_{\bar{x}xx,i} = \dfrac{u_{i+2} - 2u_{i+1} + 2u_{i-1} - u_{i-2}}{2h^3}$	$-\dfrac{1}{4} h^2 u^{\mathrm{V}}(x_i) + O(h^4)$		2	$i-2$ $i-1$ $i+1$ $i+2$	(23)

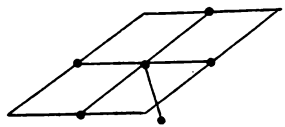
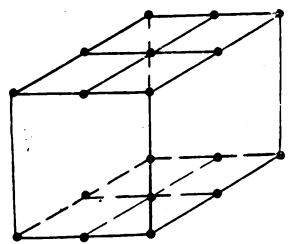
$u^{IV}(x_i)$	$u_{\bar{x}x\bar{x}x,i} = \dfrac{u_{i+2} - 4u_{i+1} + 6u_i - 4u_{i-1} + u_{i-2}}{h^4}$	$-\dfrac{1}{6} h^2 u^{VI}(x_i) + O(h^4)$	2	i-2 i-1 i i+1 i+2	(24)
	Аппроксимация на неравномерной сетке $(x_i = x_{i-1} + h_i)$				
$u^{I}(x_i)$	$u_{x,i} = \dfrac{u_{i+1} - u_i}{h_{i+1}}$	$\dfrac{h_{i+1}}{2} u^{II}(x_i) + O(h^2_{i+1})$	1	i i+1	(25)
	$u_{\bar{x},i} = \dfrac{u_i - u_{i-1}}{h_i}$	$-\dfrac{h_i}{2} u^{II}(x_i) + O(h^2_i)$	1	i-1 i	(26)
	$u_{\sigma x,i} = \sigma u_{x,i} + (1-\sigma) u_{\bar{x},i}$	$\dfrac{1}{2}[(\sigma-1) h_{i+1} + \sigma h_i] u^{II}_i + O(\hbar^2_{i+1})$	1	i-1 i i+1	(27)
		$\hbar_{i+1} = \dfrac{1}{2}(h_i + h_{i+1})$			
$u^{II}(x_i)$	$u_{\bar{x}\hat{x},i} = \dfrac{1}{\hbar_{i+1}}\left(\dfrac{u_{i+1} - u_i}{h_{i+1}} - \dfrac{u_i - u_{i-1}}{h_i}\right)$	$\dfrac{h_{i+1} - h_i}{3} u^{III}(x_i) + O(h^2_{i+1})$	1		(28)
	$\hbar_{i+1} = \dfrac{1}{2}(h_i + h_{i+1})$				
	$\tilde{u}_{\bar{x}x\hat{x},i} = \dfrac{1}{h_i}\left(\dfrac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{\hbar_{i+1}} - \dfrac{u_{i-\frac{1}{2}} - u_{i-\frac{3}{2}}}{\hbar_i}\right)$	$\dfrac{h_{i+1} - 2h_i + h_{i-1}}{4h_i} u^{II}(x_i) + O(\tilde{h})$			
	$u_{i+\frac{1}{2}} = u\left(x_i + \dfrac{h_{i+1}}{2}\right)$	$\tilde{h}$ — некоторая средняя величина шага сетки			(29)

Продолжение табл. 4

Вид дифференциального оператора	Формула и принятые обозначения	Ближайший не обращающийся в нуль член разложения в ряд Тейлора	Порядок локальной аппроксимации по переменным: t	x	Шаблон разностной схемы	Номер формулы
$Lu = \left.\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)\right\|_{x=x_i}$	$x_i = ih, \quad u(x_i) = u_i$ $L_h u_h = (a u_{\bar{x}})_{x,i} =$ $= \frac{1}{h^2}[a_{i+1}(u_{i+1} - u_i) - a_i(u_i -$ $- u_{i-1})]$ $a_i = p(x_i - 0{,}5h)$ или $a_i = \frac{1}{2}(p(x_i) + p(x_{i-1}))$	$\left(\frac{a_{i+1} - a_i}{h} - p^{\mathrm{I}}(x_i)\right) u^{\mathrm{I}}(x_i) +$ $+ \left(\frac{a_{i+1} + a_i}{2} - p(x_i)\right) u^{\mathrm{II}}(x_i) +$ $+ \frac{a_{i+1} - a_i}{6} h u^{\mathrm{III}}(x_i) + O(h^2)$		2		(33)
$Lu = \left(\frac{\partial^2 u}{\partial x_1 \partial x_2}\right)_{x_1 x_2 k}$	$u(x_{1,i}, x_{2,k}) = u_{i,k}$ $x_{1,i} = ih_1, \; x_{2,k} = kh_2$ $(i,k = 0, \pm 1, \pm 2, \ldots)$ $L_h u_h = \frac{u_{i+1,k+1} - u_{i-1,k+1} - u_{i+1,k-1} + u_{i-1,k-1}}{4h_1 h_2}$	$\frac{h_1^3}{12h_2} \cdot \frac{\partial^3 u(x_{1,i}, x_{2,k})}{\partial x_1^3} +$ $+ \frac{h_2^2}{6} \cdot \frac{\partial^3 u(x_{1,i}, x_{2,k})}{\partial x_1 \partial x_2^2} +$ $+ O(h_1^4 + h_2^4)$		2 при $\frac{h_1}{h^2} = \mathrm{const}$		(41)
$+ \frac{\partial^2 u}{\partial x_2^2}$	$\Delta_h u_h = \frac{u_{i+1,k} - 2u_{i,k} + u_{i-1,k}}{h_1^2} +$ $+ \frac{u_{i,k+1} - 2u_{i,k} + u_{i,k-1}}{h_2^2}$	$\frac{h_1^2}{12} \frac{\partial^4 u(x_{1,i}, x_{2,k})}{\partial x_1^4} +$ $+ \frac{h_2^2}{12} \frac{\partial^4 u(x_{1,i}, x_{2,k})}{\partial x_2^4} + O(h_1^4 + h_2^4)$		2		(46)

$\Delta u = \frac{\partial^2 u}{\partial x_1^2} +$	$L_h u_h = \Delta_h u_h + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \Lambda_2 u_h$	$O(h_1^4 + h_2^4)$		4		(154)
$\Delta^2 u = \frac{\partial^4 u}{\partial x_1^4} + 2\frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4 u}{\partial x_2^4}$	$\Delta_h^2 u_h = \frac{1}{h^4} [20u_{i,k} - 8(u_{i+1,k} + u_{i-1,k} + u_{i,k+1} + u_{i,k-1}) + 2(u_{i+1,k+1} + u_{i-1,k+1} + u_{i-1,k-1} + u_{i+1,k-1}) + u_{i+2,k} + u_{i-2,k} + u_{i,k+2} + u_{i,k-2}]$	$\frac{5}{9}h^2 \left(\frac{\partial^6 u(x_{1,i}, x_{2,k})}{\partial x_1^6} + \frac{\partial^6 u(x_{1,i}, x_{2,k})}{\partial x_2^6} \right) + O(h^4)$		2		(49)
$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}$	$L_{h\tau} u = u_t^j - \Lambda u^j, \quad u_t^j = \frac{u_i^{j+1} - u_i^j}{\tau}$, $\Lambda u^j = \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2}$	$\frac{\tau}{2} \left(\frac{\partial^2 u(x_i, t_j)}{\partial t^2} + 2\frac{\partial^3 u(x_i, t_j)}{\partial x^2 \partial t} \right) + O(\tau^2 + h^2)$	1	2		(54)
	$L_{h\tau} u = u_t^j - \Lambda u^{j+1}$, $\Lambda u^{j+1} = \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{h^2}$					(57)

Продолжение табл. 4

Вид дифференциального оператора	Формула и принятые обозначения	Ближайший не обращающийся в нуль член разложения в ряд Тейлора	Порядок локальной аппроксимации по переменным: t	Порядок локальной аппроксимации по переменным: X	Шаблон разностной схемы	Номер формулы
$\sum_{m=1}^{q} \frac{\partial^2 u(t,x_1,\ldots,x_q)}{\partial x_m^2}$	$x_{l,i_k} = i_k h_l,\ t_j = j\tau\ (l = \overline{1,q}, i_k = \overline{1,n_k})$ 1) $L_{h\tau}u = u_t^j - \sum_{m=1}^{q} \Lambda_m u^{j+1}$, где $u_t^j = \frac{u^{j+1} - u^j}{\tau}$, $\Lambda_m u^{j+1} = u_{x_m x_m}^{j+1}$ $u(x_{1,i_1}, x_{2,i_2}, x_{3,i_3}, \ldots, x_{q,i_q}, t_j) = u^j(X)$	$\frac{\tau}{2}\left(\frac{\partial^2 u}{\partial t^2} - 2\sum_{m=1}^{q} \frac{\partial^3 u(x,t)}{\partial x_m^2 \partial t}\right) + O(h^2 + \tau^2)$	1	2	при $q=2$ 	(67)
$\frac{\partial u(t,x_1,\ldots,x_q)}{\partial t}$	2) $L_{h\tau}u = u_t^j - (\Lambda + \frac{h^2}{6}\sum_{l=1}^{q}\sum_{m=l}^{q}\Lambda_l\Lambda_m \frac{u^{j+1} + u^j}{2} + \left[\frac{h^2}{12}\Lambda + \frac{\tau^2}{4}\left(1 + \frac{h^4}{36\tau^2}\right)\sum_{l=1}^{q}\sum_{m=l}^{q}\Lambda_l\Lambda_m\right]u_t^j$	$O(\tau^2 + h^4)$	2	4	при $q=2$ 	(211)

оператор $L_h u_h$ аппроксимирует дифференциальный оператор Lu с порядком $m > 0$ в точке X (т. е. локально), если

$$L_h u_h(X) - Lu(x) = O(|h|^m).$$

В графе порядок «локальной» аппроксимации указывается число m, причем на первом месте стоит порядок «локальной» аппроксимации дифференциального оператора разностным оператором по переменной t.

Для облегчения запоминаний разностных аппроксимаций дифференциальных операторов их принято сопоставлять с некоторым «шаблоном» (мнемоническим изображением). На «шаблоне» изображено взаимное расположение точек сетки, значения в которых связывает разностный оператор, отнесенный к некоторой фиксированной точке $x \in \Omega_h$.

Использование формул численного дифференцирования. Этот метод основывается на локальной аппроксимации функций, обладающих ограниченными производными достаточно высокого порядка, рядом Тейлора, который в окрестности узла $(x_1, x_2, \ldots, x_p)$ можно записать в виде

$$u(x_1 \pm h_1, x_2 \pm h_2, \ldots, x_p \pm h_p = u(x_1, \ldots, x_p) +$$
$$+ \sum_{i=1}^{p} \left(\pm h_i \frac{\partial}{\partial x_i}\right) u(x_1, \ldots, x_p) + \frac{1}{2!} \sum_{i=1}^{p} \left(h_i \frac{\partial}{\partial x_i}\right)^2 u(x_1, \ldots, x_p) +$$
$$+ \frac{1}{3!} \sum_{i=1}^{p} \left(\pm h_i \frac{\partial}{\partial x_i}\right)^3 u(x_1, \ldots, x_p) + \ldots. \tag{10}$$

Описание метода формальной замены производных конечно-разностными выражениями начнем с примеров, приведенных в табл. 4, для аппроксимации простейших дифференциальных операторов на равномерной и неравномерной сетках.

Пример 1. Рассмотрим дифференциальный оператор вида

$$Lu = \frac{du}{dx}, \tag{11}$$

определенный на множестве непрерывных в области $\Omega = \{a < x < b\}$ функций, имеющих ограниченные производные до третьего порядка включительно,

$$u \in C_3(a, b). \tag{12}$$

Пусть Ω_h — сеточная область вида

$$\Omega_h = \left\{x_i, \quad x_i = ih, \quad 0 < i < n, \quad h = \frac{b-a}{n}\right\}. \tag{13}$$

Так как $u(x)$ — непрерывная функция, то оператор проектирования функции u $(P_h u = u_h)$ можно положить равным

$$P_h u = u_h = u(x) \quad \forall x \in \Omega_h. \tag{14}$$

Рассмотрим множество сеточных функций u_h с областью определения Ω_h (13).

Запишем формулу Тейлора (10) в каждом внутреннем узле x_i сетки Ω_h (13):

$$u(x_i \pm h) = u(x_i) \pm hu'(x_i) + \frac{h^2}{2!} u''(x_i) \pm \frac{h^3}{3!} u'''(x_i) + O(h^4). \tag{15}$$

Тогда оператор (11) на сетке Ω_h можно аппроксимировать следующим образом:

а) $u_{x,i} = \dfrac{u_{i+1} - u_i}{h}$ (правая разностная производная); (16)

б) $u_{\bar{x},i} = \dfrac{u_i - u_{i-1}}{h}$ (левая разностная производная); (17)

в) используя линейную комбинацию (16) и (17), получим

$$u_{\sigma x,i} = \sigma u_{x,i} + (1-\sigma) u_{\bar{x},i} \quad (\sigma \text{ — любое вещественное число}); \tag{18}$$

г) при $\sigma = \dfrac{1}{2}$ имеем:

$$u_{\overset{\circ}{x},i} = \frac{u_{i+1} - u_{i-1}}{2h} \quad \text{(центральная разностная производная)}; \tag{19}$$

д) $u_{x,i+0} = \dfrac{-3u_i + 4u_{i+1} - u_{i+2}}{2h}$; (20)

е) $u_{x,i-0} = \dfrac{3u_i - 4u_{i-1} + u_{i-2}}{2h}$. (21)

Исходя из (16) — (18) для локальной погрешности аппроксимации

$$\psi_h(x) = L_h u_h - (Lu)_h$$

в точке x_i получим соответственно следующие выражения:

$$\psi_h(x_i) = u_{x,i} - u'(x_i) = \frac{h}{2} u''(x_i) + O(h^2); \tag{16'}$$

$$\psi_h(x_i) = u_{\bar{x},i} - u'(x_i) = \frac{h}{2} u''(x_i) + O(h^2); \tag{17'}$$

$$\psi_h(x_i) = \sigma u_{x,i} + (1-\sigma) u_{\bar{x},i} - u'(x_i) = \left(\sigma - \frac{1}{2}\right) h u''(x_i) + O(h^2), \tag{18'}$$

т. е. порядок локальной аппроксимации оператора (11) разностными операторами (16) — (18) в точке x_i равен единице. Разностные операторы вида (19) — (21) в точке x_i имеют второй порядок аппроксимации, так как

$$u_{\overset{\circ}{x},i} - u'(x_i) = \frac{h^2}{6} u'''(x_i) + O(h^4) = O(h^2); \tag{19'}$$

$$u_{x,i+0} - u'(x_i) = -\frac{h^2}{3} u'''(x_i) + O(h^3) = O(h^2); \tag{20'}$$

$$u_{x,i-0} - u'(x_i) = \frac{h^2}{3} u'''(x_i) + O(h^3) = O(h^2). \tag{21'}$$

Два последних соотношения следуют из разложения в окрестности точки x_i в ряд Тейлора соответственно функций $u(x_i + h)$, $u(x_i + 2h)$ и $u(x_i - h)$, $u(x_i - 2h)$. Из локальной аппроксимации разностными операторами вида (16) — (21) следует аппроксимация на сетке Ω_h соответственно с теми же порядками точности.

Очевидно, для аппроксимации оператора вида (11) даже на равномерной сетке ω_h можно построить бесчисленное множество разностных операторов. Мы привели примеры наиболее употребительных аппроксимаций простейшего дифференциального оператора.

Отметим, что аналогично могут быть построены разностные операторы вида (25) — (27) (см. табл. 4), аппроксимирующие дифференциальный оператор (11) на неравномерной сетке, и разностные

операторы (22) — (24), аппроксимирующие соответственно дифференциальные операторы

$$Lu = \frac{d^2u}{dx^2}; \quad Lu = \frac{d^3u}{dx^3}; \quad Lu = \frac{d^4u}{dx^4}$$

на равномерной сетке Ω_h.

На неравномерной сетке порядок погрешности аппроксимации, вообще говоря, понижается.

Для иллюстрации рассмотрим разностные операторы второй производной на неравномерной сетке.

Пример 2. Пусть

$$Lu = \frac{d^2u}{dx^2} \qquad (u \in C_4(a, b)). \tag{28}$$

На неравномерной сетке Ω_h (см. (2), § 2) рассмотрим разностные операторы вида

$$u_{\bar{x}\hat{x},i} = \frac{1}{\hbar_{i+1}}\left[\frac{u_{i+1}-u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}\right], \quad \hbar_{i+1} = \frac{1}{2}(h_i + h_{i+1}); \tag{29}$$

$$\tilde{u}_{\bar{x}\hat{x},i} = \frac{1}{h_i}\left[\frac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{\hbar_{i+1}} - \frac{u_{i-\frac{1}{2}} - u_{i-\frac{3}{2}}}{\hbar_i}\right], \quad u_{i+\frac{1}{2}} = u\left(x_i + \frac{h_{i+1}}{2}\right). \tag{30}$$

Исходя из (29) — (30) для локальной погрешности аппроксимации получим

$$\psi_h(x_i) = u_{\bar{x}\hat{x},i} - u''(x_i) = \frac{h_{i+1} - h_i}{3} u'''(x_i) + O(\hbar_{i+1}^2); \tag{29'}$$

$$\psi_h(x_i) = \tilde{u}_{\bar{x}\hat{x},i} - u''(x_i) = \frac{h_{i+1} - 2h_i + h_{i-1}}{4h_i} u''(x_i) + O(\tilde{h}), \tag{30'}$$

где $\tilde{h}$ — некоторая средняя величина шага сетки. Из соотношения (30′) следует, что локальная аппроксимация в точке x_i на неравномерной сетке отсутствует, так как

$$c_i = \frac{h_{i+1} - 2h_i + h_{i-1}}{4h_i} = O(1). \tag{31}$$

Поэтому желательно так выбирать неравномерную сетку, чтобы величины c_i вида (31) были невелики. Один из часто применяемых на практике способов связан с построением неравномерной сетки по закону геометрической прогрессии $h_{i+1} = qh_i$, где q величина, близкая к единице. Тогда

$$c_i = \frac{q - 2 + \frac{1}{q}}{4} = \frac{(q-1)^2}{4q}.$$

Пример 3. Рассмотрим дифференциальный оператор с переменными коэффициентами

$$Lu = \frac{d}{dx}\left(p(x)\frac{du}{dx}\right) = p'(x)u'(x) + p(x)u''(x) \tag{32}$$

в области $\Omega = \{0 < x < 1\}$.

На сетке (13) построим следующую аппроксимацию дифференциального оператора разностным

$$L_h u_h = (au_{\bar{x}})_{x,i} = \frac{1}{h}(a_{i+1}u_{\bar{x},i+1} - a_i u_{\bar{x},i}) = \frac{1}{h}(a_{i+1}u_{x,i} - a_i u_{\bar{x},i}). \tag{33}$$

Учитывая (16) и (17), получим

$$L_h u_h - (Lu)_h = \frac{a_{i+1} - a_i}{h} u'(x_i) + \frac{a_{i+1} + a_i}{2} u''(x_i) + O(h^2) - p'(x_i) u'(x_i) -$$

$$- p(x_i) u''(x_i) = \left(\frac{a_{i+1} - a_i}{h} - p'(x_i)\right) u'(x_i) + \left(\frac{a_{i+1} + a_i}{2} - p(x_i)\right) u''(x_i) +$$

$$+ \frac{a_{i+1} - a_i}{6} h u'''(x_i) + O(h^2). \tag{34}$$

Из (34) следует, что если a_i будут выбраны так, что

$$\begin{gathered} \frac{a_{i+1} + a_i}{2} - p(x_i) = O(h^2); \\ \frac{a_{i+1} - a_i}{h} - p'(x_i) = O(h^2), \\ \frac{a_{i+1} - a_i}{6} h u'''(x_i) = O(h^2), \end{gathered} \tag{35}$$

то

$$L_h u_h - (Lu)_h = O(h)^2. \tag{36}$$

Для выполнения соотношений (35) достаточно положить

$$a_i = p(x_i - 0{,}5h), \tag{37}$$

или

$$a_i = \frac{1}{2}(p_i + p_{i-1}), \tag{38}$$

или

$$a_i = \left(\frac{1}{p_i} + \frac{1}{p_{i-1}}\right)^{-1}. \tag{39}$$

Пример 4. При аппроксимации дифференциального оператора

$$Lu = \frac{\partial^2 u(x_1, x_2)}{\partial x_1 \partial x_2} \text{ (смешанная производная)} \tag{40}$$

на сетке Ω_h (§ 2, (3)) разностным оператором вида

$$L_h u_h = \frac{u_{i+1,k+1} - u_{i-1,k+1} - u_{i+1,k-1} + u_{i-1,k-1}}{4h_1 h_2} \tag{41}$$

погрешность локальной аппроксимации при $\frac{h_1}{h_2} = \text{const}$ будет равна двум. В самом деле,

$$\left(\frac{\partial^2 u}{\partial x_1 \partial x_2}\right)_{ik} = \left[\frac{\partial}{\partial x_2}\left(\frac{\partial u}{\partial x_1}\right)\right]_{ik} = \frac{\left(\frac{\partial u}{\partial x_1}\right)_{i,k+1} - \left(\frac{\partial u}{\partial x_1}\right)_{i,k-1}}{2h_2} + O(h_2^2) =$$

$$= \frac{1}{2h_2}\left\{\left[\frac{u_{i+1,k+1} - u_{i-1,k+1}}{2h_1} + O(h_1^2)\right] - \left[\frac{u_{i+1,k-1} - u_{i-1,k-1}}{2h_1} + O(h_1^2)\right]\right\} + O(h_2^2) =$$

$$= L_h u_h + O(h_1^2 + h_2^2). \tag{42}$$

Пример 5. Рассмотрим дифференциальный оператор Лапласа

$$Lu = \Delta u = \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right). \tag{43}$$

Пусть $u(x_1, x_2) \in C_4(\Omega)$,

$$\Omega = \{0 < x_1 < c, \quad 0 < x_2 < d\}. \tag{44}$$

На прямоугольной сетке

$$\Omega_h = \left\{(ih_1, kh_2) \in \Omega, \quad h = (h_1, h_2), \quad h_1 = \frac{c}{n}, \quad h_2 = \frac{d}{m}\right\} \tag{45}$$

разностный оператор

$$\Delta_h u_h = \Lambda_1 u + \Lambda_2 u, \tag{46}$$

где

$$\Lambda_1 u = u_{\bar{x}_1 x_1} = \frac{u_{i+1,k} - 2u_{ik} + u_{i-1,k}}{h_1^2}, \quad u_{ik} = u(x_{1i}, x_{2k});$$

$$\Lambda_2 u = u_{\bar{x}_2 x_2} = \frac{u_{i,k+1} - 2u_{ik} + u_{i,k-1}}{h_2^2}, \tag{46'}$$

будет иметь второй порядок аппроксимации в точке (x_{1i}, x_{2k}), так как из (22) (см. табл. 4) следует, что

$$\Delta_h u_h - (\Delta u)_h = \frac{h_1^2}{12}\frac{\partial^4 u(x_{1i}, x_{2k})}{\partial x_1^4} + \frac{h_2^2}{12}\frac{\partial^4 u(x_{1i}, x_{2k})}{\partial x_2^4} + O(|h|^4). \tag{47}$$

Соотношения (46) и (47) на квадратной сетке (45) при $h_1 = h_2 = h$ принимают соответственно вид

$$\Delta_h u_h = \frac{u_{i+1,k} + u_{i,k+1} + u_{i-1,k} + u_{i,k-1} - 4u_{ik}}{h^2}; \tag{46''}$$

$$\Delta_h u_h - (\Delta u)_h = \frac{h^2}{12}\left(\frac{\partial^4 u(x_{1i}, x_{2k})}{\partial x_1^4} + \frac{\partial^4 u(x_{1i}, x_{2k})}{\partial x_2^4}\right) + O(h^4). \tag{47'}$$

Пример 6. Рассмотрим бигармонический оператор

$$Lu = \Delta^2 u = \frac{\partial^4 u}{\partial x_1^4} + 2\frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4 u}{\partial x_2^4} \tag{48}$$

в области Ω (44). Для получения разностного оператора, аппроксимирующего оператор (48) на квадратной сетке (45) при $h_1 = h_2 = h$, можно положить

$$L_h u_h = \Delta_h^2 u_h = \Delta_h(\Delta_h u).$$

Тогда, если воспользоваться соотношением (46″), получим

$$L_h u_h = \frac{1}{h^4}[20u_{ik} - 8(u_{i+1,k} + u_{i-1,k} + u_{i,k+1} + u_{i,k-1}) + 2(u_{i+1,k+1} +$$

$$+ u_{i-1,k+1} + u_{i-1,k-1} + u_{i+1,k-1}) + u_{i+2,k} + u_{i-2,k} + u_{i,k+2} + u_{i,k-2}]. \tag{49}$$

Легко показать, что в классе $C_6(\Omega)$ порядок локальной аппроксимации оператора (48) разностным оператором (49) будет равен двум и оценивается величиной

$$\frac{5}{9}h^2\left(\frac{\partial^6 u(x_{1i}, x_{2k})}{\partial x_1^6} + \frac{\partial^6 u(x_{1i}, x_{2k})}{\partial x_2^6}\right) + O(h^4) = O(h^2). \tag{50}$$

Пример 7. Рассмотрим дифференциальный оператор

$$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} \tag{51}$$

в классе функций $C_4^2(\Omega)$, где

$$\Omega = \{0 < x < c, \quad 0 < t \leqslant T\}. \tag{52}$$

В сеточной области

$$\Omega_{h\tau}=\left\{(x_i,\ t_j),\quad x_i=ih,\quad t_j=j\tau,\quad 0<i<n,\quad 0<j\leqslant m,\quad \tau=\frac{T}{m}\right\} \tag{53}$$

рассмотрим следующую аппроксимацию дифференциального оператора (51):

$$L_{h\tau}u_{h\tau}=u_t^j-\Lambda u^j, \tag{54}$$

где

$$u_t^j=\frac{u_i^{j+1}-u_i^j}{\tau},\quad \Lambda u^j=u_{\bar{x}x,i}^j=\frac{u_{i+1}^j-2u_i^j+u_{i-1}^j}{h^2},\quad u_i^j=u(x_i,\ t_j).$$

Если оператор проектирования функции $u(x,\ t)$ на $\Omega_{h,\tau}$ положить равным

$$P_{h\tau}u=u_{h\tau}=u(x,\ t),$$

где $(x,\ t)\in\Omega_{h\tau}$, то в соответствии с формулами (16) и (22) (см. табл. 4) имеем:

$$\begin{aligned} u_{t,i}^j&=\frac{\partial u(x_i,\ t_j)}{\partial t}+\frac{\tau}{2}\frac{\partial^2 u(x_i,\ t_j)}{\partial t^2}+O(\tau^2);\\ u_{\bar{x}x,i}^j&=\frac{\partial^2 u(x_i,\ t_j)}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4 u(x_i,\ t_j)}{\partial x^4}+O(h^4)\end{aligned} \tag{55}$$

и погрешность локальной аппроксимации оператора (51) разностным оператором (54) оценивается величиной

$$\begin{aligned} L_{h\tau}u_{h\tau}-(Lu)_{h\tau}&=\frac{\tau}{2}\frac{\partial^2 u(x_i,\ t_j)}{\partial t^2}+O(\tau^2)-\frac{h^2}{12}\frac{\partial^4 u(x_i,\ t_j)}{\partial x^4}-O(h^4)=\\ &=O(\tau+|h|^2).\end{aligned} \tag{56}$$

Очевидно, на сетке $\Omega_{h,\tau}$ (53) можно построить и следующую аппроксимацию дифференциального оператора (51):

$$L_{h\tau}u_{h\tau}=u_{t,\,i}^j-\Lambda u^{j+1}, \tag{57}$$

где

$$\Lambda u^{j+1}=u_{\bar{x}x,i}^{j+1}=\frac{u_{i+1}^{j+1}-2u_i^{j+1}+u_{i-1}^{j+1}}{h^2},\quad u_i^{j+1}=u(x_i,\ t_{j+1}). \tag{57'}$$

Так как (см. (22) табл. 4).

$$\Lambda u^{j+1}=\frac{\partial^2 u(x_i,\ t_{j+1})}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4 u(x_i,\ t_{j+1})}{\partial x^4}+O(h^4), \tag{58}$$

то

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=\{u_{t,i}^j-\Lambda u^{j+1}\}-\left\{\frac{\partial u(x_i,\ t_j)}{\partial t}-\frac{\partial^2 u(x_i,\ t_j)}{\partial x^2}\right\}=O(\tau+|h|^2). \tag{59}$$

Используя линейную комбинацию разностных аппроксимаций вида (54), (57), рассмотрим следующую аппроксимацию дифференциального оператора (51):

$$L_{h\tau}u_{h\tau}=u_t^j-\sigma\Lambda u^{j+1}-\sigma_1\Lambda u^j. \tag{60}$$

Тогда

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=\frac{\partial u_i^j}{\partial t}+\frac{\tau}{2}\frac{\partial^2 u_i^j}{\partial t^2}+O(\tau^2)-\sigma\left[\frac{\partial^2 u_i^j}{\partial x^2}+\tau\frac{\partial^3 u_i^j}{\partial x^2\partial t}+\right.$$

$$+O(\tau^2)+O(h^2)\Bigg]-\sigma_1\left[\frac{\partial^2 u_i^j}{\partial x^2}+O(h^2)\right]-(Lu)_{h\tau}=(1-\sigma-\sigma_1)\frac{\partial^2 u_i^j}{\partial x^2}+$$

$$+\tau\left(\frac{1}{2}\frac{\partial^2 u_i^j}{\partial t^2}-\sigma\frac{\partial^3 u_i^j}{\partial x^2\partial t}\right)+O(\tau^2+h^2).$$

При $\sigma_1=1-\sigma$

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=O(\tau+h^2), \tag{61}$$

причем при $\sigma=\frac{1}{2}$

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=\frac{\tau}{2}\frac{\partial}{\partial t}(Lu)_{h\tau}+O(\tau^2+h^2). \tag{62}$$

Очевидно, при любом σ ($\sigma_1=1-\sigma$) разностный оператор (60) имеет локальную аппроксимацию порядка $O(\tau+h^2)$.

Для дифференциального оператора (51) можно построить разностную аппроксимацию вида

$$L_{h\tau}u_{h\tau}=u_t^j+\frac{\alpha\tau}{h}u_{t\bar{x}}^j-\Lambda u^j, \tag{63}$$

откуда

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=O(\tau+h^2)+O\left(\frac{\alpha\tau}{h}\right).$$

Следовательно, разностный оператор (63) будет аппроксимировать дифференциальный оператор (51) только при выполнении некоторых соотношений между шагами τ и h (условная аппроксимация). Например, при $\tau=ch^2$ будем иметь:

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=O(h).$$

Пример 8. Рассмотрим дифференциальный оператор

$$Lu=\frac{\partial u}{\partial t}-\sum_{m=1}^{q}\frac{\partial^2 u}{\partial x_m^2},\quad u=u(t,x_1,\ldots,x_q) \tag{64}$$

в цилиндре $\Omega=\{0<x_l<1,\ 0<t\leqslant T,\ i=\overline{1,q}\}$.

На сеточной области

$$\Omega_{h\tau}=\{(x_{1,i_1},x_{2,i_2},\ldots,x_{q,i_q},t_j),\quad t_j=j\tau,\quad x_{l,i_k}=$$

$$=i_kh_l,\quad 0<j\leqslant m,\quad l=\overline{1,q},\quad 0<k<n_k,\quad h=(h_1,h_2,\ldots,h_q)\} \tag{65}$$

построим аппроксимацию дифференциального оператора (64) разностным в виде

$$L_{h\tau}u_{h\tau}=\frac{u^{j+1}-u^j}{\tau}-\sum_{m=1}^{q}\Lambda_m u^{j+1}, \tag{66}$$

где

$$\Lambda_m u^{j+1}=u^{j+1}_{\bar{x}_m x_m}=\frac{\partial^2 u(x,t+\tau)}{\partial x_m^2}+\frac{h^2}{12}\frac{\partial^4 u(x,t+\tau)}{\partial x^4}+O(h^4).$$

Тогда

$$L_{h\tau}u_{h\tau}-(Lu)_{h\tau}=\frac{\tau}{2}\frac{\partial^2 u(x,t)}{\partial t^2}+O(\tau^2)-\sum_{m=1}^{q}\left\{\frac{\partial^2}{\partial x_m^2}\left[u(x,t)+\right.\right.$$

$$\left.\left.+\tau\frac{\partial u(x,t)}{\partial t}\right]\right\}+O(h^2)+\sum_{m=1}^{q}\frac{\partial^2 u(x,t)}{\partial x_m^2}=O(\tau+h^2).$$

Таким образом, разностный оператор (66) аппроксимирует дифференциальный оператор (64) со вторым порядком по пространственным переменным x и первым порядком по t.

Метод неопределенных коэффициентов. Этот метод построения основывается на записи «шаблона» с неопределенными коэффициентами для аппроксимации дифференциального оператора в целом в некоторой фиксированной точке $X_0 \in \Omega_h$ (центре «шаблона»)

$$L_h u_h = \sum_{k=0}^{q} a_k u(X_k). \tag{67}$$

Неопределенные коэффициенты находят из условия получения наивысшего относительно $|h|$ порядка аппроксимации в данной точке X_0 разностным оператором дифференциального оператора, или, что эквивалентно, из условия совпадения значений дифференциального и разностных операторов от многочленов как можно более высокой степени m. Если полученная при этом система уравнений относительно неопределенных коэффициентов a_k имеет решение, то

$$L_h u_h - Lu(X_0) = \sum_{k=0}^{q} a_k u(X_k) - Lu(X_0) = O(h^m).$$

Для дифференциальных операторов, обладающих каким-либо свойством симметрии, неопределенные коэффициенты при значениях сеточной функции в точках «шаблона» выбирают обычно таким образом, чтобы это свойство симметрии сохранилось. Число неопределенных параметров в этом случае значительно уменьшается (см. пример 10).

Пример 9. Рассмотрим дифференциальный оператор n-го порядка вида

$$Lu(x) = \sum_{k=0}^{n} p_k(x) u^{(k)}(x), \tag{68}$$

где $p_k(x)$ — заданные непрерывные функции. Тогда, если $u(x) \in C_{n+m+1}(\Omega)$, методом неопределенных коэффициентов по «шаблону», состоящему из q различных точек $x_i + \alpha_k h$ $(k = \overline{1, q})$, можно построить разностный оператор вида

$$L_h u_h = \sum_{k=1}^{q} a_k u(x_i + \alpha_k h), \tag{69}$$

обладающий в точке x_i $(m+1)$-м порядком аппроксимации $(n+1 \leqslant q \leqslant n+1+m)$, т. е. всегда можно так подобрать a_k, что

$$L_h u_h - (Lu)_h = \sum_{k=1}^{q} a_k u(x_i + \alpha_k h) - Lu(x_i) = O(h^{m+1}). \tag{70}$$

Покажем, что при выбранном $q \geqslant n+1$ и разных α_k, коэффициенты a_k «шаблона» (69) определятся однозначно, причем

$$L_h u_h - (Lu)_h = R(u^{(m+n+1)}(x_i + \theta_k \alpha_k h), a_1, a_2, \ldots, a_q, \alpha_1, \ldots, \alpha_q). \tag{71}$$

Для оценки величины $|R|$ имеет место следующее неравенство:

$$|R| \leqslant h^{m+1} \frac{|u^{(m+n+1)}|_{\max}}{(m+n+1)!} |P_n(h)|, \tag{72}$$

где $|u^{(n+m+1)}|_{\max}$ — максимальное значение $(n+m+1)$-й производной на отрезке, содержащем все точки «шаблона» $(x_i + \alpha_k h,\ k = \overline{1, q})$; $P_n(h)$ — полином от h степени не выше n, не зависящий от u.

В самом деле, разложим (69) в окрестности точки x_l в ряд Тейлора с остаточным членом в форме Лагранжа

$$L_h u_h = a_1 \sum_{k=0}^{n+m} \frac{(h\alpha_1)^k}{k!} u^{(k)}(x_l) + a_1 \frac{(\alpha_1 h)^{n+m+1}}{(n+m+1)!} u^{(n+m+1)}(x_l + \theta_1 \alpha_1 h) +$$

$$+ a_2 \sum_{k=0}^{n+m} \frac{(h\alpha_2)^k}{k!} u^{(k)}(x_l) + a_2 \frac{(\alpha_2 h)^{n+m+1}}{(n+m+1)!} u^{(n+m+1)}(x_l + \theta_2 \alpha_2 h) +$$

$$+ \cdots + a_q \sum_{k=0}^{n+m} \frac{(h\alpha_q)^k}{k!} u^{(k)}(x_l) + a_q \frac{(\alpha_q h)^{n+m+1}}{(n+m+1)!} u^{(n+m+1)}(x_l + \theta_q \alpha_q h) =$$

$$= u(x_l) \sum_{k=1}^{q} a_k + \frac{h}{1!} u'(x_l) \sum_{k=1}^{q} \alpha_k a_k + \cdots + \frac{h^{n+m}}{(n+m)!} u^{(n+m)}(x_l) \sum_{k=1}^{q} \alpha_k^{n+m} a_k +$$

$$+ R, \quad 0 < \theta_k < 1, \quad k = \overline{1, q}, \tag{73}$$

$$R = \frac{h^{m+1}}{(n+m+1)!} \sum_{k=1}^{q} h^n \alpha_k^{n+m+1} a_k u^{(n+m+1)}(x_l + \theta_k \alpha_k h). \tag{74}$$

Если потребовать, чтобы выполнялось соотношение (70), то для определения a_k получим следующую систему:

$$\begin{cases} \dfrac{h^l}{l!} \displaystyle\sum_{k=1}^{q} \alpha_k^l a_k = p_l(x_l), & l = \overline{0, n}; \\ \displaystyle\sum_{k=1}^{q} \alpha_k^l a_k = 0, & l = \overline{n+1, n+m}. \end{cases} \tag{75}$$

Система уравнений (75) всегда разрешима, так как ее определитель (определитель Вандермонда) отличен от нуля. При выполнении соотношений (75)

$$L_h u_h - Lu(x_l) = R \tag{76}$$

и, учитывая (74), для оценки $|R|$ получим (72).

Разностных аппроксимаций вида (69) для дифференциального оператора (68) существует бесконечно много, но получаемая при этом система (75) относительно a_k всякий раз будет иметь единственное решение.

Пример 10. Пусть

$$Lu = \Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}. \tag{77}$$

Составим выражение для сеточного оператора в точке X_0 в виде «шаблона», построенного по пяти точкам,

$$L_h u_h = \Delta_h u_h = a_0 u(X_0) + \sum_{k=1}^{4} a_k u(X_k),$$

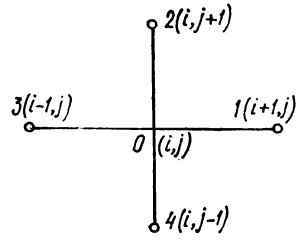


Рис. 6

где X_k — узлы квадратного «шаблона», представленного на рис. 6. В силу инвариантности оператора Δ относительно любого поворота системы координат можно искать $\Delta_h uh$ в виде

$$\Delta_h u_h = a_0 u(X_0) + a_1 \sum_{k=1}^{4} u(X_k).$$

Определим точки X_k:

$$X_0 = X_0(x_{1i}, x_{2j}); \quad X_1 = X_1(x_{1i} + h, x_{2j}); \quad X_2 = X_2(x_{1i}, x_{2j} + h);$$
$$X_3 = X_3(x_{1i} - h, x_{2j}); \quad X_4 = X_4(x_{1i}, x_{2j} - h). \tag{78}$$

Используя разложение в ряд Тейлора $u(X_k)$ в окрестности точки X_0 и требуя, чтобы

$$\Delta_h u_h - \Delta u(x_{1i}, x_{2j}) = O(h^m), \tag{79}$$

где m по возможности большее число, для определения $a_k (k = 0, 1)$ получим следующую систему:

$$\begin{cases} a_0 + 4a_1 = 0, \\ h^2 a_1 = 1 \end{cases} \tag{80}$$

или

$$a_1 = \frac{1}{h^2}, \quad a_0 = -\frac{4}{h^2}$$

и $\Delta_h u_h$ будет определяться формулой (46″).

Для определения порядка «локальной» аппроксимации будем иметь:

$$\Delta_h u_h - \Delta u(x_{1i}, x_{2j}) = \frac{h^4}{4!} a_1 \left[\frac{\partial^4 u(x_{1i} + \theta_1 h, x_{2j})}{\partial x_1^4} + \frac{\partial^4 u(x_{1i} - \theta_2 h, x_{2j})}{\partial x_1^4} + \right.$$

$$\left. + \frac{\partial^4 u(x_{1i}, x_{2j} + \theta_3 h)}{\partial x_2^4} + \frac{\partial^4 u(x_{1i}, x_{2j} - \theta_3 h)}{\partial x_2^4} \right] = O(h^2), \tag{81}$$

если $u(X) \in C_4(\Omega)$.

Метод неопределенных коэффициентов может быть использован для того, чтобы среди аппроксимаций данного порядка найти наилучшую по какому-либо признаку. Этот метод также широко используется для построения разностных схем повышенного порядка точности (см. [18] и [30]).

2. Аппроксимация граничных условий

Условимся узел X называть внутренним ($X \in \Omega_h$), если при построении разностной схемы, аппроксимирующей исходную дифференциальную задачу, он используется только для замены дифференциального уравнения. Все остальные узлы отнесем к граничным. Следовательно, узел X будет граничным, если он используется как для замены дифференциального уравнения, так и для замены граничных условий. В дальнейшем, для определенности, в качестве примеров рассмотрим в двумерных областях аппроксимацию операторов, не зависящих от производных (граничные условия первого рода), и операторов, зависящих от производных.

1. Граничные условия первого рода

$$u|_\Gamma = \varphi(x_1, x_2). \tag{82}$$

Простой снос граничных условий. Пусть X_i — граничный узел области $\Omega_h + \Gamma_h$, а Q_i — ближайшая к X_i в каком-то смысле (по расстоянию от точки X_i или в направлении координатных осей) точка границы Γ. Тогда полагают

$$l_h u_h \equiv u(X_i) = \varphi(Q_i). \tag{83}$$

Погрешность аппроксимации при этом будет $O(h)$. Это следует из разложения $u(X_i)$ в ряд Тейлора в окрестности точки Q_i. Если

взаимное расположение точек X_i и Q_i будет такое, как на рис. 7, то

$$u(X_i) = u(Q_i) \pm \delta_i \frac{\partial u(Q_i)}{\partial x} + \frac{\delta_i^2}{2!} \cdot \frac{\partial^2 u(Q_i)}{\partial x^2} \pm \ldots, \tag{84}$$

где $0 \leqslant \delta_i < h$.

Следовательно,

$$u(X_i) = \varphi(Q_i) + O(h), \tag{85}$$

так как δ_i соизмеримы с h. Если $\delta_i = 0$, то выражение

$$u(X_i) = \varphi(Q_i)$$

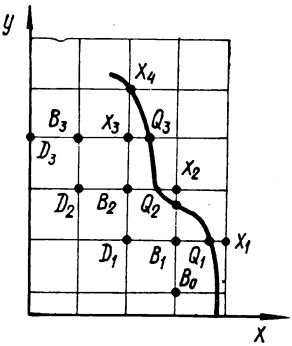


Рис. 7

будет являться точной аппроксимацией граничного условия (82). Значит, если узлы $\Gamma_h \in \Gamma$, то погрешность аппроксимации граничных условий первого рода будет равна нулю. Поэтому при построении области Ω_h желательно, чтобы все узлы Γ_h принадлежали Γ.

Линейная интерполяция. От точки Q_i будем двигаться вдоль координатной линии внутрь области Ω до некоторой точки B_i, которая является ближайшей к Q_i (после X_i, если $X_i \in \Omega$) точкой, принадлежащей области Ω (см. рис. 7).

Для определения значений $u(X_i)$ построим линейный интерполяционный полином по точкам Q_i и B_i.

Тогда

$$u(X_i) = u(Q_i) + (X_i - Q_i)\, u(Q_i, B_i) + R, \tag{86}$$

где

$$u(Q_i, B_i) = \frac{u(Q_i) - u(B_i)}{Q_i - B_i}; \quad |R| \leqslant \frac{M_2}{2} |(X_i - Q_i)(X_i - B_i)|;$$
$$M_2 = \max \left| \frac{\partial^2 u(\tilde{Q}_i)}{\partial x^2} \right|; \quad B_i < \tilde{Q}_i < X_i. \tag{87}$$

Откуда, если через δ_i по-прежнему обозначить расстояние между точками Q_i и X_i ($\delta_i = r_{X_i Q_i}$) и шаги сетки по оси x и y соответственно через h и h_1, то

$$u(X_1) = \frac{h\varphi(Q_1) - \delta_1 u(B_1)}{h - \delta_1} + O(h^2);$$
$$u(X_2) = \frac{h_1\varphi(Q_2) - \delta_2 u(B_1)}{h_1 - \delta_2} + O(h_1^2); \tag{88}$$
$$u(X_3) = \frac{h\varphi(Q_3) + \delta_3 u(B_3)}{h + \delta_3} + O(h^2);$$
$$u(X_4) = \varphi(Q_4).$$

Интерполяция более высокого порядка. С помощью большего числа внутренних узлов, лежащих на одной координатной линии, можно построить интерполяционный полином более высокого порядка, что позволит повысить погрешность аппроксимации до третьего порядка и выше.

Уравнение для сноса краевого условия первого рода в точке X_i, имеющее погрешность аппроксимации $O(h^3)$, имеет вид

$$u(X_i) = \varphi(Q_i) + (X_i - Q_i)\, u(Q_i, B_i) + \\ + (X_i - Q_i)(X_i - B_i)\, u(Q_i, B_i, D_i) + O(h^3), \tag{89}$$

где $u(Q_i, B_i, D_i)$ — разделенные разности второго порядка, построенные по точкам Q_i, B_i и D_i.

Например,

$$u(X_3) = \varphi(Q_3) - \frac{\delta_3}{h+\delta_3}[\varphi(Q_3) - u(B_3)] - \delta_3 h\left[\frac{u(Q_3)}{(2h+\delta_3)(h+\delta_3)} + \right. \\ \left. + \frac{u(D_3)}{h(2h+\delta_3)} - \frac{u(B_3)}{h(h+\delta_3)}\right] + O(h^3). \tag{90}$$

Для построения разностных аппроксимаций в граничных узлах X^*, являющихся центром некоторого «шаблона», содержащего точки, которые не принадлежат Ω_h, существуют различные подходы. Большинство из них связаны с интерполяцией функции на отрезке, содержащем точку X^*, и записью разностной аппроксимации дифференциального оператора $Lu = \varphi_0$ в этой точке путем использования найденных из интерполяционного полинома значений для $u(X^*)$. Для определенности рассмотрим пример, когда внутри области, ограниченной контуром Γ, выполняется уравнение Лапласа, а граничный узел X^*, являющийся центром пятиточечного шаблона (см. рис. 7), совпадает с B_1. По значениям $u(B_1)$, $u(D_1)$ и $u(Q_1)$ составим интерполяционный полином и его вторую производную примем за приближенное значение $\frac{\partial^2 u(B_1)}{\partial x^2}$. Аналогично находится приближенное значение $\frac{\partial^2 u(B_1)}{\partial y^2}$.

Для удобства записи обозначим через r_{BQ} расстояние между точками B и Q. Тогда $\frac{\partial^2 u(B_1)}{\partial x^2} \approx 2u(B_1, D_1, Q_1)$, или

$$\frac{\partial^2 u(B_1)}{\partial x^2} \approx 2\left(\frac{u(D_1)}{r_{D_1B_1}r_{D_1Q_1}} + \frac{u(Q_1)}{r_{Q_1B_1}r_{Q_1D_1}} - \frac{u(B_1)}{r_{B_1Q_1}r_{B_1D_1}}\right). \tag{91}$$

Аналогично

$$\frac{\partial^2 u(B_1)}{\partial y^2} \approx 2\left(\frac{u(B_0)}{r_{B_0Q_2}r_{B_0B_1}} + \frac{u(Q_2)}{r_{Q_2B_0}r_{Q_2B_1}} - \frac{u(B_1)}{r_{B_1Q_2}r_{B_1B_0}}\right) \tag{92}$$

и пятиточечная аппроксимация оператора Лапласа в граничном узле B_1, являющемся центром шаблона, будет иметь вид

$$\frac{2u(D_1)}{r_{D_1B_1}r_{D_1Q_1}} + \frac{2u(Q_1)}{r_{Q_1B_1}r_{Q_1D_1}} + \frac{2u(B_0)}{r_{B_0Q_2}r_{B_0B_1}} + \frac{2u(Q_2)}{r_{Q_2B_0}r_{Q_2B_1}} - \\ - 2\left(\frac{u(B_1)}{r_{B_1Q_1}r_{B_1D_1}} + \frac{u(B_1)}{r_{B_1Q_2}r_{B_1B_0}}\right) = 0. \tag{93}$$

Порядок аппроксимации при этом будет $O(h)$. Для повышения порядка аппроксимации можно воспользоваться интерполяционным полиномом более высокой степени или методом неопределенных коэффициентов построения разностных аппроксимаций для дифференциальных операторов, или пятиточечной аппроксимацией дифференциального

оператора Лапласа с привлечением точек, находящихся вне области (так называемых фиктивных точек, т. е. точек X_1, X_2 для граничной точки B_1). Используя последний из указанных подходов, будем иметь в точке B_1:

$$\frac{u(D_1)}{r_{D_1B_1}r_{D_1X_1}}+\frac{u(X_1)}{r_{X_1D_1}r_{X_1B_1}}+\frac{u(X_2)}{r_{X_2B_1}r_{X_2B_0}}+\frac{u(B_0)}{r_{B_0B_1}r_{B_0X_2}}-$$
$$-\left(\frac{u(B_1)}{r_{B_1D_1}r_{B_1X_1}}+\frac{u(B_1)}{r_{B_1B_0}r_{B_1X_2}}\right)=0, \qquad (94)$$

где значения $u(X_1)$ и $u(X_2)$ находятся из (88).

2. Граничные операторы, зависящие от производных

$$lu=a(X)\,u(X)+b(X)\frac{\partial u(X)}{\partial s}, \quad X\in\Gamma, \qquad (95)$$

где $\frac{\partial u(x)}{\partial s}$ — производная по направлению s в точке $X\in\Gamma$ (вектор s направлен внутрь области Ω).

Граница области составлена из отрезков прямых, параллельных координатным осям. В этом случае можно воспользоваться односторонними разностными аппроксимациями оператора $\frac{\partial u}{\partial s}=\frac{\partial u}{\partial n}$ (см. табл. 4, (21), (20)).

Например, в точке X_i (рис. 8):

а) если положить

$$l_hu_h=a_iu_i+b_i\frac{u_{i+1}-u_i}{h}, \qquad (96)$$

то

$$l_hu_h-(lu)_h=O(h);$$

б) если положить

$$l_hu_h=a_iu_i+b_i\frac{-3u_i+4u_{i+1}-u_{i+2}}{2h}, \qquad (97)$$

то

$$l_hu_h-(lu)_h=O(h^2).$$

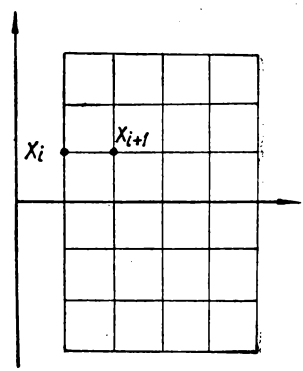


Рис. 8

Используя метод неопределенных коэффициентов, можно построить аппроксимацию достаточно высокого порядка (см. пример 9), но при этом увеличивается число узлов сетки, которые используются для аппроксимации граничных условий.

в) Повысить порядок аппроксимации краевых условий можно, воспользовавшись дифференциальным уравнением для рассматриваемой краевой задачи.

Например, пусть требуется решить задачу:

$$\frac{\partial^2u}{\partial t^2}=\frac{\partial^2u}{\partial x^2}+f(x,t), \quad 0<x<a, \quad 0<t\leqslant T; \qquad (98)$$

$$u(x,0)=\gamma_0(x); \quad \frac{\partial u(x,0)}{\partial t}=\gamma_1(x);$$
$$u(0,t)=\gamma_2(t); \quad u(a,t)=\gamma_3(t). \qquad (99)$$

Рассмотрим равномерную сетку $\Omega_{h\tau}$ с шагами h и τ.

Аппроксимируя краевое условие $\frac{\partial u(x,\ 0)}{\partial t} = \gamma_1(x)$ разностным уравнением вида

$$l_h u_h = u_t(x,\ 0) = \gamma_{1h}, \tag{100}$$

получим, что погрешность аппроксимации будет равна $O(\tau)$, так как

$$u_t(x,\ 0) = \frac{\partial u(x,\ 0)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x,\ 0)}{\partial t^2} + O(\tau^2).$$

Если предположить, что на границе области ($t = 0$) существуют и непрерывны производные, входящие в уравнение (98), то

$$\frac{\partial^2 u(x,0)}{\partial t^2} = \frac{\partial^2 u(x,\ 0)}{\partial x^2} + f(x,\ 0) \tag{101}$$

и вместо (100) можно построить разностный оператор

$$l_h u_h = u_t(x,\ 0) - \frac{\tau}{2}[\gamma''_{0h} + f(x,\ 0)] = \gamma_{1h}. \tag{102}$$

Тогда погрешность аппроксимации начального условия $\frac{\partial u(x,\ 0)}{\partial t} = \gamma_1(x)$ на решении задачи (98), (99) будет иметь порядок $O(\tau^2)$.

г) Для повышения порядка аппроксимации граничных условий вида (95) иногда используют фиктивные узлы. Для этого рассматривают сетку, сдвинутую на $\frac{h}{2}$, т. е. такую сетку, в которой граница Г проходит посредине между линиями сеточной области (рис. 9).

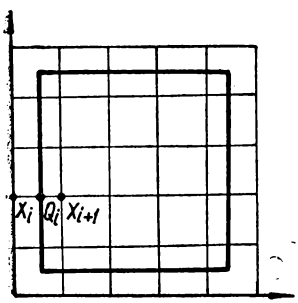


Рис. 9

Используя «фиктивный» (внешний) узел, производную по нормали в точке Q_i можно аппроксимировать центральными разностями (см. табл. 4, (19)):

$$\frac{\partial u(Q_i)}{\partial n} = \frac{u(X_{i+1}) - u(X_i)}{h} + O(h^2), \tag{103}$$

а значения $u(Q_i)$ полусуммой значений u в соседних точках:

$$u(Q_i) = \frac{u(X_i) + u(X_{i+1})}{2} + O(h^2). \tag{104}$$

Разностный оператор, аппроксимирующий граничный оператор (95), в точке Q_i можно записать следующим образом:

$$l_h u_h = a(Q_i)\frac{u(X_{i+1}) + u(X_i)}{2} + b(Q_i)\frac{u(X_{i+1}) - u(X_i)}{h} \tag{105}$$

и

$$l_h u_h - lu(Q_i) = O(h^2). \tag{106}$$

Случай криволинейной границы. Приведем пример простейшего подхода для аппроксимации граничного оператора вида (95), если граница области Ω криволинейная и сетка, покрывающая область Ω, квадратная (рис. 10). Из точки Q_i, лежащей на пересечении границы Г

и координатной линии, проведем вектор s до пересечения с каким-либо отрезком, соединяющим два ближайших узла сеточной области Ω_h.

Тогда, полагая

$$\frac{\partial u(Q_i)}{\partial s} = \frac{u(P_i) - u(Q_i)}{r_{Q_iP_i}}, \tag{107}$$

для аппроксимации граничного оператора (95) имеем:

$$l_h u_h = a(Q_i)\, u(Q_i) + b(Q_i) \frac{u(P_i) - u(Q_i)}{r_{Q_iP_i}}. \tag{108}$$

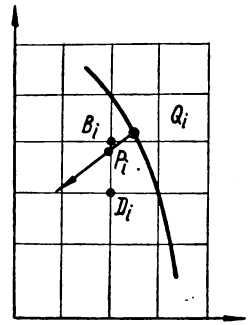


Рис. 10

Точка $P_i \in \Omega_h + \Gamma_h$. Поэтому значение $u(P_i)$ с помощью интерполяционного полинома выражают через значения $u(B_i)$ и $u(D_i)$:

$$u(P_i) = u(B_i) \frac{r_{D_iP_i}}{r_{B_iD_i}} + u(D_i) \frac{r_{P_iB_i}}{r_{B_iD_i}} \tag{109}$$

и окончательно получим

$$l_h u_h = a(Q_i)\, u(Q_i) + \frac{b(Q_i)}{r_{Q_iP_i}} \left[u(B_i) \frac{r_{D_iP_i}}{r_{B_iD_i}} + u(D_i) \frac{r_{P_iB_i}}{r_{B_iD_i}} - u(Q_i) \right], \tag{110}$$

$$l_h u_h - lu(Q_i) = O(h). \tag{111}$$

Разностный аналог дифференциального оператора в граничных точках (B_i, D_i), лежащих внутри области Ω, записываем, используя ближайшие граничные точки (типа Q_i), лежащие на пересечении координатных линий и границы Γ (см., например, аппроксимацию вида (93) оператора Лапласа в граничных точках).

Алгоритмы для аппроксимаций граничных условий с погрешностью высшего порядка описаны в [16], [19], [52], [65], [72], [73], [79], [86], [90] и др.

3. Примеры простейших разностных схем

Приведем примеры разностных схем, конструирование которых связано с использованием формул численного дифференцирования и метода неопределенных коэффициентов.

Стационарные задачи. 1. Рассмотрим третью краевую задачу на отрезке $[a, b]$ для обыкновенного дифференциального уравнения второго порядка

$$\begin{cases} -\dfrac{d^2u}{dx^2} + g(x)\, u(x) = \varphi(x); \\ -u'(a) + \alpha_0 u(a) = \gamma_0; \\ \quad u'(b) + \alpha_1 u(b) = \gamma_1, \end{cases} \tag{112}$$

где $\alpha_0 > 0$, $\alpha_1 > 0$, $g(x) \geqslant 0$ $\forall\, x \in [a, b]$.

На сетке $\Omega_h + \Gamma_h = \left\{x_i : x_i = x_0 + ih,\ x_0 = a,\ h = \frac{b-a}{n},\ i = \overline{0,\ n}\right\}$ построим разностные схемы:

$$\text{а)}\quad \begin{cases} -v_{\bar{x}x,i} + g_i v_i = \varphi_i, \quad i = \overline{1,\ n-1}; \\ -\dfrac{-3v_0 + 4v_1 - v_2}{2h} + \alpha_0 v_0 = \gamma_0; \\ \dfrac{3v_n - 4v_{n-1} + v_{n-2}}{2h} + \alpha_1 v_n = \gamma_1; \end{cases} \tag{113}$$

$$\text{б)}\quad \begin{cases} -v_{\bar{x}x,i} + g_i v_i = \varphi_i, \quad i = \overline{1,\ n-1}; \\ -v_{x,0} + \dfrac{h}{2}(g_0 v_0 - \varphi_0) + \alpha_0 v_0 = \gamma_0; \\ v_{\bar{x},n} + \dfrac{h}{2}(g_n v_n - \varphi_n) + \alpha_1 v_n = \gamma_1. \end{cases} \tag{114}$$

На сетке $\tilde{\Omega}_h + \tilde{\Gamma}_h = \left\{x_i : x_i = x_0 + ih,\ x_0 = a - \frac{h}{2},\ h = \frac{b-a}{n},\ i = \overline{0,\ n}\right\}$ с фиктивными узлами $x_0 = a - \frac{h}{2}$ и $x_n = b + \frac{h}{2}$ можно построить разностную схему

$$\text{в)}\quad \begin{cases} -v_{\bar{x}x,i} + g_i v_i = \varphi_i, \quad i = \overline{1,\ n-1}; \\ -v_{x,0} + \alpha_0 \dfrac{v_0 + v_1}{2} = \gamma_0; \\ v_{\bar{x},n} + \alpha_1 \dfrac{v_n + v_{n-1}}{2} = \gamma_1. \end{cases} \tag{115}$$

Системы разностных уравнений (113) — (115), если ввести в рассмотрение $(n+1)$-мерный вектор

$$V = (v_0,\ v_1,\ \ldots,\ v_n)', \tag{116}$$

можно записать в матричной форме

$$AV = F, \tag{117}$$

где A и F определяются соответственно:
в случае а):

$$A = \frac{1}{h^2}\begin{pmatrix} 3 + 2\alpha_0 h & -4 & 1 & & & 0 \\ -1 & 2 + g_1 h^2 & -1 & & & \\ & -1 & 2 + g_2 h^2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 + g_{n-1} h^2 & -1 \\ 0 & & & 1 & -4 & 3 + 2\alpha_1 h \end{pmatrix};$$

$$F = \left(\frac{2\gamma_0}{h},\ \varphi_1,\ \varphi_2,\ \ldots,\ \varphi_{n-1},\ \frac{2\gamma_1}{h}\right)';$$

в случае б):

$$A=\frac{1}{h^2}\begin{pmatrix} 1+\frac{h^2g_0}{2}+\alpha_0 h & -1 & & & & & 0 \\ -1 & 2+g_1h^2 & -1 & & & & \\ & -1 & 2+g_2h^2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2+g_{n-1}h^2 & -1 & \\ 0 & & & & -1 & 1+\frac{h^2g_n}{2}+\alpha_1 h \end{pmatrix};$$

$$F=\left(\frac{\gamma_0}{h}+\frac{1}{2}\varphi_0,\ \varphi_1,\ \varphi_2,\ \ldots,\ \varphi_{n-1},\ \frac{\gamma_1}{h}+\frac{1}{2}\varphi_n\right)';$$

в случае в):

$$A=\frac{1}{h^2}\begin{pmatrix} 2+h\alpha_0 & -2+h\alpha_0 & & & & & \\ -1 & 2+g_1h^2 & -1 & & & & 0 \\ & -1 & 2+g_2h^2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2+g_{n-1}h^2 & -1 & \\ 0 & & & & -2+h\alpha_1 & 2+h\alpha_1 \end{pmatrix};$$

$$F=\left(\frac{2\gamma_0}{h},\ \varphi_1,\ \varphi_2,\ \ldots,\ \varphi_{n-1},\ \frac{2\gamma_1}{h}\right)'.$$

Введем в R_{n+1} скалярное произведение

$$[v,\ z]=\sum_{i=0}^{n} hv_iz_i$$

и норму

$$|[v_h]|_h=\sqrt{[v_h,\ v_h]}.$$

Пусть $\varphi(x)$ и $g(x)$ достаточно гладкие функции, причем такие, что $u(x)\in \mathbf{C}_4[a,\ b]$. Тогда все построенные разностные схемы будут обладать вторым порядком аппроксимации на решении задачи (112).

Оператор краевой задачи (112) является симметричным и положительно-определенным. Однако из вида матриц разностных схем (113) — (115) следует, что только матрица разностной схемы (114) является симметричной. Этим свойством будет обладать и разностная схема (115) при условии, что $2-h\alpha_0\neq 0$ и $2-h\alpha_1\neq 0$. Матрица A разностной схемы (113) не обладает свойством симметрии. Поэтому более внимательного изучения среди указанных разностных схем заслуживает схема (114). Покажем, что оператор разностной схемы (114) при $\alpha_0,\ \alpha_1\geqslant\delta>0$ является положительно-определенным, т. е.

$$[A_hv_h,\ v_h]\geqslant\delta\,|[v_h]|_h^2.$$

Если предварительно умножить обе части двух последних уравнений системы (114) на $\frac{1}{h}$, получим

$$[A_h v_h,\ v_h] = (-v_{\bar{x}x},\ v_h) + (g_h v_h,\ v_h) - v_{x,0} v_0 + \frac{h}{2} g_0 v_0^2 +$$

$$+ \alpha_0 v_0^2 + v_{\bar{x},n} v_n + \frac{h}{2} g_n v_n^2 + \alpha_1 v_n^2$$

и, пользуясь первой разностной формулой Грина (приложение, § 3), получим

$$[A_h v_h,\ v_h] = (v_{\bar{x}},\ v_{\bar{x}}] + \alpha_0 v_0^2 + \alpha_1 v_n^2 + \sum_{i=1}^{n-1} h g_i v_i^2 + \frac{h}{2} g_n v_n^2 + \frac{h}{2} g_0 v_0^2, \tag{118}$$

где скалярные произведения $(\cdot)$ и $(\cdot]$ определяются по формулам (37) (см. приложение, § 1).

Из (118) следует, что

$$[A_h v_h,\ v_h] \geqslant (v_{\bar{x}},\ v_{\bar{x}}] + \alpha_0 v_0^2 + \alpha_1 v_n^2, \tag{119}$$

так как $g_i \geqslant 0$ $(i = \overline{0,\ n})$.

Оценим величину $|(v_{\bar{x}}]|^2$. Для этого заметим, что

$$v_i^2 = \left(v_0 + \sum_{j=1}^{i} h v_{\bar{x},j}\right)^2 = \left(v_n - \sum_{j=i+1}^{n} h v_{\bar{x},j}\right)^2. \tag{120}$$

Из (120), используя неравенство $(p \pm s)^2 \leqslant 2(p^2 + s^2)$, получим

$$v_i^2 \leqslant 2\left\{v_0^2 + \left(\sum_{j=1}^{i} h v_{\bar{x},j}\right)^2\right\}, \tag{121}$$

$$v_i^2 \leqslant 2\left\{v_n^2 + \left(\sum_{j=i+1}^{n} h v_{\bar{x},j}\right)^2\right\}. \tag{122}$$

Из неравенства Коши — Буняковского имеем:

$$\left(\sum_{j=1}^{i} h v_{\bar{x},j}\right)^2 \leqslant x_i \sum_{j=1}^{i} h v_{\bar{x},j}^2 \leqslant x_i (v_{\bar{x}},\ v_{\bar{x}}], \tag{123}$$

$$\left(\sum_{j=i+1}^{n} h v_{\bar{x},j}\right)^2 \leqslant (b - a - x_{i+1}) \sum_{j=i+1}^{n} h v_{\bar{x},j}^2 \leqslant (b - a)(v_{\bar{x}},\ v_{\bar{x}}]. \tag{123'}$$

Поэтому

$$v_i^2 \leqslant 2\{v_0^2 + (b - a)(v_{\bar{x}},\ v_{\bar{x}}]\}, \tag{124}$$

$$v_i^2 \leqslant 2\{v_n^2 + (b - a)(v_{\bar{x}},\ v_{\bar{x}}]\}, \tag{125}$$

или

$$v_i^2 \leqslant v_0^2 + v_n^2 + 2(b - a)(v_{\bar{x}},\ v_{\bar{x}}]. \tag{126}$$

Тогда

$$|[v_h]|_h^2 = \sum_{i=0}^{n} h v_i^2 \leqslant \{v_0^2 + v_n^2 + 2(b - a)(v_{\bar{x}},\ v_{\bar{x}}]\}(2b - a) =$$

$$= (2b-a)\left\{\frac{\alpha_0 v_0^2}{\alpha_0} + \frac{\alpha_1 v_1^2}{\alpha_1} + 2(b-a)(v_{\bar{x}}, v_{\bar{x}}]\right\} \leqslant$$

$$\leqslant \varkappa \{\alpha_0 v_0^2 + \alpha_1 v_n^2 + (v_{\bar{x}}, v_{\bar{x}}]\}, \tag{127}$$

где

$$\varkappa = (2b-a)\max\left\{\frac{1}{\alpha_0}, \frac{1}{\alpha_1}, 2(b-a)\right\}. \tag{128}$$

Таким образом, из (119) и (127) получаем оценку

$$[A_h v_h, v_h] \geqslant \frac{1}{\varkappa} |[v_h]|_h^2, \tag{129}$$

откуда и следует положительная определенность оператора A_h. Разностные схемы (113) — (115) имеют одинаковый порядок аппроксимации на решениях исходной задачи, однако разностная схема (114) обладает еще свойствами симметричности и положительной определенности.

2. *Разностные схемы для уравнения Пуассона.* Для простоты изложения ограничимся случаем двух независимых переменных.

В прямоугольнике Ω с границей Γ: $\Omega + \Gamma = \{0 \leqslant x \leqslant a, 0 \leqslant y \leqslant c\}$ требуется найти решение задачи Дирихле для уравнения Пуассона

$$\Delta u = f(x_1, x_2); \quad \Delta u = \sum_{k=1}^{2} L_k u; \quad L_k u = \frac{\partial^2 u}{\partial x_k^2}; \tag{130}$$

$$u|_\Gamma = \gamma(x_1, x_2). \tag{131}$$

Рассмотрим на сетке

$$\bar{\Omega}_h = \Omega_h + \Gamma_h = \left\{\{ih_1, kh_2\}, \quad h = (h_1, h_2), \quad 0 \leqslant i \leqslant \right.$$

$$\left. \leqslant n+1, \ 0 \leqslant k \leqslant m+1, \ h_1 = \frac{a}{n+1}, \ h_2 = \frac{c}{m+1}\right\} \tag{132}$$

разностную схему

$$\begin{cases} \Delta_h v_h = \varphi_h, \quad \Delta_h v_h = \Lambda_1 v_h + \Lambda_2 v_h; \\ \Lambda_1 v_h = \dfrac{v_{i-1,k} - 2v_{ik} + v_{i+1,k}}{h_1^2}, \ \Lambda_2 v_h = \dfrac{v_{i,k-1} - 2v_{ik} + v_{i,k+1}}{h_2^2}; \end{cases} \tag{133}$$

$$\begin{cases} v_{0k} = \gamma_{0k}, \quad v_{n+1,k} = \gamma_{n+1,k}; \\ v_{i0} = \gamma_{i0}, \quad v_{i,m+1} = \gamma_{i,m+1}; \\ i = \overline{1, n}; \quad k = \overline{1, m}. \end{cases} \tag{134}$$

Будем предполагать, что функции $f(x_1, x_2)$, $\gamma(x_1, x_2)$ таковы, что решение задачи (130), (131) является достаточно гладкой функцией. Краевые условия (131) аппроксимируются точно, а поэтому для исследования порядка аппроксимации разностной схемы достаточно исследовать аппроксимацию уравнения (130) в точках $(ih_1, kh_2) \in \Omega$.

Уравнение для погрешности $y_h = v_h - u_h$ будет иметь вид

$$\Delta_h y_h = \psi_h, \tag{135}$$

где

$$\psi_h = \varphi_h - (f)_h.$$

В линейном пространстве сеточных функций с областью определения $\bar{\Omega}_h$ введем норму при помощи формулы (9), § 2.

Так как $(\Delta u)_h = (f)_h = (f_{ik})$ ($i = \overline{1, n}$, $k = \overline{1, m}$), то, если положить

$$\varphi_{ik} = f_{ik} \tag{136}$$

или

$$\varphi_{ik} = f_{ik} + O(h^2) \tag{137}$$

и учесть соотношение (47), получим, что разностная схема (133), (134) будет аппроксимировать задачу (130), (131) с погрешностью $O(h^2)$, $h = \max(h_1, h_2)$.

Для того чтобы разностное уравнение (133), (134) можно было записать в матричной форме

$$AV = F, \tag{138}$$

введем в рассмотрение матрицу Λ_1 размерности $(n \times n)$

$$h_1^2\Lambda_1 = \begin{pmatrix} -2 & 1 & 0 & \dots\dots & 0 \\ 1 & -2 & 1 & \dots\dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots\dots & 1 & -2 & 1 \\ 0 & \dots\dots & 0 & 1 & -2 \end{pmatrix} \tag{139}$$

и n-мерные векторы $\boldsymbol{V}_k$, $\boldsymbol{F}_k$:

$$\boldsymbol{V}_k = (v_{1k}, v_{2k}, \dots, v_{nk})', \tag{140}$$

$$\boldsymbol{F}_k = \left(\varphi_{1k} - \frac{\gamma_{0k}}{h_1^2}, \varphi_{2k}, \dots, \varphi_{n-1,k}, \varphi_{nk} - \frac{\gamma_{n+1,k}}{h_1^2}\right)'. \tag{141}$$

Тогда систему (133), (134) можно записать в виде

$$\begin{gathered} \frac{1}{h_2^2}[\boldsymbol{V}_{k+1} + (h_2^2\Lambda_1 - 2I_n)\boldsymbol{V}_k + \boldsymbol{V}_{k-1}] = \boldsymbol{F}_k \quad (k = \overline{1, m}), \\ \boldsymbol{V}_0 = (\gamma_{i0})_{i=\overline{1,n}}, \quad \boldsymbol{V}_{m+1} = (\gamma_{i,m})_{i=\overline{1,n}}. \end{gathered} \tag{142}$$

Исключим из системы (142) граничные условия $\boldsymbol{V}_0$, $\boldsymbol{V}_{m+1}$:

$$\begin{gathered} \frac{1}{h_2^2}[(h_2^2\Lambda_1 - 2I_n)\boldsymbol{V}_1 + \boldsymbol{V}_2] = \boldsymbol{F}_1 - \frac{1}{h_2^2}\boldsymbol{V}_0, \\ \frac{1}{h_2^2}[\boldsymbol{V}_{k+1} + (h_2^2\Lambda_1 - 2I_n)\boldsymbol{V}_k + \boldsymbol{V}_{k-1}] = \boldsymbol{F}_k \quad (k = \overline{2, m-1}), \\ \frac{1}{h_2^2}[\boldsymbol{V}_{m-2} + (h_2^2\Lambda_1 - 2I_n)\boldsymbol{V}_{m-1}] = \boldsymbol{F}_m - \frac{1}{h_2^2}\boldsymbol{V}_{m+1}. \end{gathered} \tag{143}$$

Систему (143) можно записать в виде

$$AV = F, \tag{144}$$

где A — блочная матрица порядка $(n)(m) \times (n)(m)$; V, F — матрицы размерности $(n) \times (m)$:

$$A = \frac{1}{h_2^2}\begin{pmatrix} h_2^2\Lambda_1 - 2I_n & I_n & 0 & \dots & 0 \\ I_n & h_2^2\Lambda_1 - I_n & I_n & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & I_n & h_2^2\Lambda_1 - I_n & I_n \\ 0 & \dots & \dots & I_n & h_2^2\Lambda_1 - I_n \end{pmatrix}, \quad (145)$$

$$V = (V_1, V_2, \dots, V_m)', \quad F = \left(F_1 - \frac{1}{h_2^2}V_0, F_2, \dots, F_{m-1}, F_m - \right.$$
$$\left. - \frac{1}{h_2^2}V_{m+1}\right)'. \quad (146)$$

Матрицу A системы (144) представим в виде суммы двух матриц

$$A = A_1 + A_2, \quad (147)$$

где

$$A_1 = \begin{pmatrix} \Lambda_1 & & 0 \\ & \ddots & \\ 0 & & \Lambda_1 \end{pmatrix}; \quad A_2 = \frac{1}{h_2^2}\begin{pmatrix} -2I_n & I_n & 0 & \dots & 0 \\ I_n & -2I_n & I_n & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & I_n & -2I_n \end{pmatrix}. \quad (148)$$

Если воспользоваться тензорным произведением матриц, то

$$A_1 = I_m \otimes \tilde{\Lambda}_1; \quad A_2 = \tilde{\Lambda}_2 \otimes I_n, \quad (149)$$

где $h_1^2\tilde{\Lambda}_1$, $h_2^2\tilde{\Lambda}_2$— трехдиагональные матрицы соответственно размерности $(n) \times (n)$ и $(m) \times (m)$, элементы которых определяются правой частью соотношения (139); I_k — единичная матрица порядка $(k) \times (k)$.

Система (138) может быть представлена в виде

$$(I_m \otimes \tilde{\Lambda}_1 + \tilde{\Lambda}_2 \otimes I_n)\, V = F. \quad (150)$$

Решение системы (144) может быть найдено прямым или итерационным методом.

Схема повышенной точности для уравнения Пуассона. При построении разностной схемы (133) использован разностный оператор, «локальная» погрешность аппроксимации которого определяется соотношением (47), § 2,

$$\Lambda_1 v_h + \Lambda_2 v_h - (\Delta u)_h = \frac{h_1^2}{12}(L_1L_1u)_h + \frac{h_2^2}{12}(L_2L_2u)_h + O(|h|^4). \quad (151)$$

Если воспользоваться уравнением (130), то

$$\begin{aligned} L_1L_1u &= L_1f - L_1L_2u, \\ L_2L_2u &= L_2f - L_2L_1u. \end{aligned} \quad (152)$$

Тогда

$$\Lambda_1 v_h + \Lambda_2 v_h - (\Delta u)_h = -\frac{h_1^2 + h_2^2}{12}(L_1 L_2 u)_h +$$

$$+ \frac{h_1^2}{12}(L_1 f)_h + \frac{h_2^2}{12}(L_2 f)_h + O(|h|^4). \tag{153}$$

Поэтому естественно ожидать, что разностная схема

$$\Lambda_1 v_h + \Lambda_2 v_h + \frac{h_1^2 + h_2^2}{12}\Lambda_1 \Lambda_2 v_h = \frac{h_1^2}{12}(L_1 f)_h + \frac{h_2^2}{12}(L_2 f)_h + (f)_h, \tag{154}$$

$$V_h|_{\Gamma h} = (\gamma)_h$$

будет иметь четвертый порядок аппроксимации $O(|h|^4)$ на решениях $u(x_1, x_2)$ задачи (130), (131) при условии достаточной гладкости функций $f(x_1, x_2)$ и $\gamma(x_1, x_2)$.

Так как краевые условия (131) аппроксимируются точно, то достаточно исследовать погрешность аппроксимации уравнения (130) разностным уравнением (154). Функция погрешности аппроксимации y_h будет удовлетворять уравнению

$$\Delta_h y_h + \frac{h_1^2 + h_2^2}{12}\Lambda_1 \Lambda_2 y_h = \psi_h, \tag{155}$$

$$\psi_h = \frac{h_1^2 + h_2^2}{12}(L_1 L_2 u)_h - \frac{h_1^2 + h_2^2}{12}(\Lambda_1 \Lambda_2 u)_h + O(|h|^4). \tag{156}$$

Из (22) табл. 4 имеем:

$$(\Lambda_2 u)_h = (L_2 u)_h + \frac{h_2^2}{12}(L_2 L_2 \tilde{u})_h,$$

$$\tilde{u}_h = u(x_{1i}, \tilde{x}_{2k}), \quad \tilde{x}_{2k} \in (x_{2k} - h_2, x_{2k} + h_2);$$

$$(\Lambda_1 \Lambda_2 u)_h = \Lambda_1 (L_2 u)_h + \frac{h_2^2}{12}\Lambda_1 (L_2 L_2 \tilde{u})_h =$$

$$= (L_1 L_2 u)_h + \frac{h_2^2}{12}(L_1 L_2 L_2 \tilde{u})_h + O(|h|^2); \tag{157}$$

$$\tilde{u}_h = u(\tilde{x}_{1i}, \tilde{x}_{2k}); \quad \tilde{x}_{1i} \in (x_{1i} - h_1, x_{1i} + h_1).$$

Подставляя (157) в (156), получим

$$\psi_h = O(|h|^4).$$

Справедлива такая лемма:

Лемма 1. *Разностная схема* (154) *имеет на решении* $u = u(x_1, x_2)$ *уравнения* (130), (131) *четвертый порядок аппроксимации.*

Разностное уравнение (154) можно записать в матричной форме

$$\tilde{A}V = \tilde{F}, \tag{158}$$

где

$$\tilde{A} = A_1 + A_2 + \frac{h_1^2 + h_2^2}{12}A_1 A_2, \tag{159}$$

а матрицы A_1, A_2 определяются по формулам (148):

$$\tilde{F} = \begin{pmatrix} \tilde{F}_1 \\ \tilde{F}_2 \\ \vdots \\ \tilde{F}_m \end{pmatrix}, \quad \begin{aligned} \tilde{F}_1 &= \varphi_1 - \frac{1}{h_2^2} v_0 - \frac{h_1^2 + h_2^2}{12} \Lambda_2 v_0, \\ \tilde{F}_2 &= \varphi_2, \\ &\dots\dots\dots\dots\dots\dots \\ \tilde{F}_{m-1} &= \varphi_{m-1}, \\ \tilde{F}_m &= \varphi_m - \frac{1}{h_2^2} v_m - \frac{h_1^2 + h_2^2}{12} \Lambda_2 v_m, \end{aligned} \tag{160}$$

$$\varphi_k = \left(\varphi_{1k} - \frac{\gamma_{0k}}{h_1^2} - \frac{h_1^2 + h_2^2}{2} \Lambda_1 \gamma_{0k}, \ \varphi_{2k}, \ \dots, \ \varphi_{n-1,k}, \ \varphi_{nk} - \right.$$

$$\left. - \frac{\gamma_{n+1,k}}{h_1^2} - \frac{h_1^2 + h_2^2}{12} \Lambda_1 \gamma_{m+1,k} \right), \tag{161}$$

$$\varphi_{ik} = f_{ik} + \frac{h_1^2}{12} \cdot \frac{\partial^2 f(x_{1i}, x_{2k})}{\partial x_1^2} + \frac{h_2^2}{12} \cdot \frac{\partial^2 f(x_{1i}, x_{2k})}{\partial x_2^2},$$

$$v_0 = (\gamma_{i0})_{i=\overline{1,n}}, \quad v_{m+1} = (\gamma_{i,m+1})_{i=\overline{1,n}} \quad (i = \overline{1, n}; \ k = \overline{1, m}). \tag{162}$$

3. *Разностная схема для бигармонического уравнения с использованием схем для гармонических уравнений.* В квадрате $\bar{\Omega}$: $\{0 \leqslant x_1, x_2 \leqslant a\}$ требуется найти решение бигармонического уравнения

$$\Delta\Delta u = \varphi(x_1, x_2), \quad x \in \Omega, \tag{163}$$

удовлетворяющее на границе квадрата Γ следующим условиям:

$$u|_\Gamma = \gamma_1(x_1, x_2); \quad \Delta u|_\Gamma = \gamma_2(x_1, x_2). \tag{164}$$

Для некоторых задач теории упругости часто оказывается более удобно сводить решение исходной задачи к эквивалентной системе двух уравнений

$$\Delta\omega = \varphi(x_1, x_2), \tag{165}$$

$$\omega|_\Gamma = \gamma_2(x_1, x_2), \tag{166}$$

$$\Delta u = \omega(x_1, x_2), \tag{167}$$

$$u|_\Gamma = \gamma_1(x_1, x_2). \tag{168}$$

Для решения системы (165) — (168) на квадратной сетке вида (132) при $h_1 = h_2 = h$ может быть построена разностная схема

$$\begin{gathered} \Delta_h w_h + \frac{h^2}{6} \Lambda_1 \Lambda_2 w_h = \varphi_h + \frac{h^2}{12} \Delta_h \varphi_h, \\ w_h|_{\Gamma_h} = \gamma_{2h}, \\ \Delta_h v_h + \frac{h^2}{6} \Lambda_1 \Lambda_2 v_h = w_h + \frac{h^2}{12} \varphi_h, \\ v_h|_{\Gamma_h} = \gamma_{1h}. \end{gathered} \tag{169}$$

Легко показать, как это делалось для схемы (154), что схема (169) будет иметь четвертый порядок аппроксимации.

Нестационарные задачи. 4. *Разностные схемы для уравнения теплопроводности.* В области

$$\overline{\Omega} \cup \{0 \leqslant t \leqslant T\}, \tag{170}$$

где

$$\overline{\Omega} = \{0 \leqslant x \leqslant a\}, \tag{171}$$

требуется найти решение первой краевой задачи для уравнения теплопроводности

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + f(x, t), & (172) \\ u(x, 0) = \gamma_1(x), \\ u(0, t) = \gamma_2(t), \quad u(a, t) = \gamma_3(t). & (173) \end{cases}$$

Простейшими разностными схемами на сетке

$$\overline{\Omega}_{h\tau} = \left\{(x_i, t_j);\ x_i = ih;\ t_j = j\tau;\ i = \overline{0, n+1};\ j = \overline{0, m};\ h = \right.$$
$$\left. = \frac{a}{n+1};\ \tau = \frac{T}{m}\right\} \tag{174}$$

являются схемы вида:

а)
$$v_t^j = \Lambda_1 v^j + \varphi^j, \tag{175}$$

$$v_i^0 = \gamma_{1i}^0,$$
$$v_0^j = \gamma_2^j, \quad v_{n+1}^j = \gamma_3^j, \tag{176}$$

где $\varphi_i^j = f(x_i, t_j)$; $\gamma_{l,i}^j = \gamma_l(x_i, t_j)$ $(l = \overline{1, 3})$, а уравнение (175) в индексной форме записывается в виде

$$\frac{v_i^{j+1} - v_i^j}{\tau} = \frac{v_{i-1}^j - 2v_i^j + v_{i+1}^j}{h^2} + \varphi_i^j \quad (i = \overline{1, n};\ j = \overline{0, m}); \tag{177}$$

б)
$$v_t^j = \Lambda_1 v^{j+1} + \varphi_i^j;\ \Lambda_1 v^{j+1} = \frac{v_{i-1}^{j+1} - 2v_i^{j+1} + v_{i+1}^{j+1}}{h^2}, \tag{178}$$

а краевые условия записываются в виде (176).

Очевидно, погрешность аппроксимации разностных схем (175), (176) и (178), (176) будет $Q(\tau + h^2)$, так как краевые условия первого рода и функция $f(x, t)$ аппроксимируются точно, а порядок аппроксимации исходного уравнения разностными уравнениями (175) или (178) на решениях $u(x, t)$ задачи (170), (171) на основании (56), § 2, и, соответственно, (59), § 2, будет $O(\tau + h^2)$.

Схемы вида (175) и (178) называют однородными, так как аппроксимация независимо от точки (x, t) носит единообразный характер.

Определение. *Однородными разностными схемами* называют такие схемы, вид которых не зависит ни от выбора конкретной задачи из данного класса, ни от выбора узла разностной сетки.

Разностная схема (175), (176) называется явной двухслойной разностной схемой, так как значение v_i^{j+1} в каждой точке x_i $(j+1)$-го временного слоя $t = t_{j+1}$ может быть определено через значения v_i^j на предыдущем слое по явным формулам.

Разностная схема (178), (176) называется *неявной двухслойной разностной схемой*, так как для определения значений v_i^{j+1} на $(j+1)$-м временном слое получаем систему алгебраических уравнений с трехдиагональной матрицей порядка $(n \times n)$

$$Av_h^{j+1} = F^j; \tag{179}$$

$$A = \frac{\tau}{h^2}\begin{pmatrix} 2+\frac{h^2}{\tau} & -1 & & & 0 \\ -1 & 2+\frac{h^2}{\tau} & -1 & & \\ & \cdot & \cdot & & \\ 0 & & & -1 & 2+\frac{h^2}{\tau} \end{pmatrix}. \tag{180}$$

Правые части F^j системы (179) зависят от значений v_i^j на предыдущем временном слое:

$$v_h^{j+1} = (v_i^{j+1})_{i=\overline{1,n}},$$

$$F^j = \Big(\tau\varphi_1^j + v_1^j + \frac{\tau}{h^2}\gamma_2^{j+1},\ \tau\varphi_2^j + v_2^j,\ \ldots,\ \tau\varphi_{n-1}^j + v_{n-1}^j,\ \tau\varphi_n^j + $$
$$+ v_n^j + \frac{\tau}{h^2}\gamma_3^{j+1}\Big).$$

Решение системы (179) може быть найдено методом прогонки.

Схема с весами для одномерного уравнения теплопроводности. Рассмотрим разностную схему для однородного уравнения теплопроводности, в которой для аппроксимации оператора $Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}$ используется разностный оператор (60), § 2:

в) $$v_t^j = \sigma\Lambda_1 v^{j+1} + (1-\sigma)\Lambda_1 v^j + \varphi^j. \tag{181}$$

Краевые условия записываются в виде (176).

Для функции $y_h = v_h - u_h$ — погрешности аппроксимации — получим следующую задачу:

$$y_t^j = \sigma\Lambda_1 y_i^{j+1} + (1-\sigma)\Lambda_1 y_i^j + \psi_i^j \quad (i = \overline{1, n};\ j = \overline{1, m}), \tag{182}$$

$$y_i^0 = 0 \quad (i = \overline{0, n+1}),$$
$$y_0^j = 0; \quad y_{n+1}^j = 0 \quad (j = \overline{0, m}), \tag{183}$$

где

$$\psi_i^j = \varphi^j + \sigma\Lambda_1 u_i^{j+1} + (1-\sigma)\Lambda_1 u_i^j - u_{t,i}^j. \tag{184}$$

Исследуем погрешность аппроксимации уравнения (172) разностным уравнением (181), используя асимптотическое разложение ψ_i^j

в окрестности точки $\left(x_i,\ t_{j+\frac{1}{2}}\right)$, $\left(t_{j+\frac{1}{2}} = \left(j+\frac{1}{2}\right)\tau\right)$ до величин четвертого порядка относительно h и второго относительно τ.

Обозначим

$$\tilde{u} = u\left(x_i,\ t_{j+\frac{1}{2}}\right),\quad \tilde{f} = f\left(x_i,\ t_{j+\frac{1}{2}}\right),\quad Lu = \frac{\partial^2 u}{\partial x^2},$$

$$\psi_i^j = \varphi_i^j + \sigma\left[L\left(\tilde{u} + \frac{\tau}{2}\frac{\partial\tilde{u}}{\partial t}\right) + \frac{h^2}{12}LL\left(\tilde{u} + \frac{\tau}{2}\frac{\partial\tilde{u}}{\partial t}\right)\right] +$$

$$+ (1-\sigma)\left[L\left(\tilde{u} - \frac{\tau}{2}\frac{\partial\tilde{u}}{\partial t}\right) + \frac{h^2}{12}LL\left(\tilde{u} - \frac{\tau}{2}\frac{\partial\tilde{u}}{\partial t}\right)\right] -$$

$$- \frac{\partial\tilde{u}}{\partial t} + O(\tau^2 + h^4) = \varphi_i^j - \tilde{f} + \tau\left(\sigma - \frac{1}{2}\right)\times$$

$$\times\left[L\frac{\partial\tilde{u}}{\partial t} + \frac{h^2}{12}LL\frac{\partial\tilde{u}}{\partial t}\right] + \frac{h^2}{12}LL\tilde{u} + O(\tau^2 + h^4). \tag{185}$$

Так как $L\frac{\partial u}{\partial t} = LLu + Lf$, то (185) можно записать в виде

$$\psi_i^j = \varphi^j - \tilde{f} + \left(\tau\left(\sigma - \frac{1}{2}\right) + \frac{h^2}{12}\right)LL\tilde{u} + \tau\left(\sigma - \frac{1}{2}\right)L\tilde{f} +$$

$$+ \frac{h^2}{12}\tau\left(\sigma - \frac{1}{2}\right)LL\frac{\partial\tilde{u}}{\partial t} + O(\tau^2 + h^4). \tag{186}$$

Отсюда следует, если положить

$$\sigma = \frac{1}{2};\quad \varphi_i^j = \tilde{f}, \tag{187}$$

то

$$\psi_i^j = O(\tau^2 + h^2),$$

и, если положить

$$\tau\left(\sigma - \frac{1}{2}\right) + \frac{h^2}{12} = 0, \tag{188}$$

$$\varphi_i^j = \tilde{f} + \frac{h^2}{12}L\tilde{f}, \tag{189}$$

то

$$\psi_i^j = O(\tau^2 + h^4).$$

Таким образом, справедлива следующая лемма:

Лемма 2. *Неявная разностная схема с весами вида* (181), (176) *имеет на решении* $u(x,\ t)$ *уравнения* (172) *порядок аппроксимации*:

$$1)\ O(\tau^2 + h^4)\ \text{при}\ \sigma = \frac{1}{2} + \frac{h^2}{12\tau},\ \varphi_i^j = f_i^{j+\frac{1}{2}} + \frac{h^2}{12}\frac{\partial^2 f_i^{j+\frac{1}{2}}}{\partial x^2},$$

$$\left(i = \overline{1,\ n};\quad j = \overline{1,\ m}\right),\quad u \in C_6^3; \tag{190}$$

2) $O(\tau^2+h^2)$ при $\sigma=\frac{1}{2}$, $\varphi_i^j=f_i^{j+\frac{1}{2}}$ $(i=\overline{1,n};\ j=\overline{1,m})$, $u\in C_4^3$; (191)

3) $O(\tau+h^2)$ при $\sigma\neq\frac{1}{2}$, $\sigma\neq\frac{1}{2}+\frac{h^2}{12\tau}$, $\varphi_i^j=f_i^{j+\frac{1}{2}}+O(\tau+h^2)$

$$(i=\overline{1,n};\ j=\overline{1,m}),\quad u\in C_4^2. \tag{192}$$

Разностное уравнение (181) при φ^j и σ, определяющихся по формулам (190), можно заменить уравнением вида

$$v_t^j=\Lambda\frac{v^{j+1}+v^j}{2}-\frac{h^2}{12}\Lambda v_t^j+\frac{h^2}{12}\Lambda f^{j+\frac{1}{2}}+f^{j+\frac{1}{2}}, \tag{193}$$

которое имеет тот же порядок аппроксимации, что и уравнение (181) при достаточно гладкой функции $f(x,t)$. Схему (193) можно записать в виде

$$\left(I+\frac{h^2}{12}\Lambda\right)v_t^j=\Lambda\frac{v^{j+1}+v^j}{2}+\left(I+\frac{h^2}{12}\Lambda\right)f^{j+\frac{1}{2}} \tag{194}$$

или

$$Dv_t^j=\Lambda\frac{v^{j+1}+v^j}{2}+Df^{j+\frac{1}{2}}, \tag{195}$$

где

$$D=I+\frac{h^2}{12}\Lambda,\quad \Lambda=\Lambda_1.$$

Рассмотрим неявную схему для уравнения (172) несимметричного вида

г) $$(1+\sigma)v_t^j-\sigma v_{\bar t}^j=\Lambda_1 v^{j+1}+\varphi_i^{j+1} \tag{196}$$

или в индексной форме

$$(1+\sigma)\frac{v_i^{j+1}-v_i^j}{\tau}-\sigma\frac{v_i^j-v_i^{j-1}}{\tau}=\frac{v_{i-1}^{j+1}-2v_i^{j+1}+v_{i+1}^{j+1}}{h^2}+\varphi_i^{j+1}. \tag{196'}$$

Разностная схема (196) содержит три слоя $(t_{j+1},\ t_j,\ t_{j-1})$ и поэтому нужно дополнительно, кроме условий (176), задать значения $v_i^1=v(x_i,\tau)$ $(i=\overline{1,n})$. Функция v_i^1 выбирается таким образом, чтобы обеспечить нужный порядок аппроксимации разностной схемы в целом. Поэтому вначале исследуем погрешность аппроксимации уравнения (172) разностной схемой (196). Для погрешности аппроксимации

$$\psi_i^{j+1}=\varphi_i^{j+1}+\Lambda_1 u^{j+1}-(1+\sigma)u_t^{j+1}+\sigma u_{\bar t}^j$$

в окрестности точки $(x_i,\ t_{j+1})$ получим следующее асимптотическое разложение:

$$\psi_i^{j+1}=\varphi_i^{j+1}+\left(\frac{\partial^2 u_i^{j+1}}{\partial x^2}+\frac{h^2}{12}\cdot\frac{\partial^4 u_i^{j+1}}{\partial x^4}\right)-(1+\sigma)\times$$

$$\times\left(\frac{\partial u_i^{j+1}}{\partial t}-\frac{\tau}{2}\cdot\frac{\partial^2 u_i^{j+1}}{\partial t^2}\right)+\sigma\left(\frac{\partial u_i^{j+1}}{\partial t}-\frac{3}{2}\tau\frac{\partial^2 u_i^{j+1}}{\partial t}\right)+O(\tau^2+h^4)=$$

$$=\varphi_i^{j+1}-f_i^{j+1}+\frac{h^2}{12}\cdot\frac{\partial^4 u_i^{j+1}}{\partial x^4}+\tau\left(\frac{1}{2}-\sigma\right)\frac{\partial^2 u_i^{j+1}}{\partial t^2}+O(\tau^2+h^4). \quad (197)$$

Так как

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^4 u}{\partial x^4}+\frac{\partial^2 f}{\partial x^2}+\frac{\partial f}{\partial t},$$

то

$$\psi_i^{j+1}=\varphi_i^{j+1}-f_i^{j+1}+\left[\frac{h^2}{12}+\tau\left(\frac{1}{2}-\sigma\right)\right]\times$$

$$\times\frac{\partial^2 u^{j+1}}{\partial t^2}-\frac{h^2}{12}\left(\frac{\partial^2 f_i^{j+1}}{\partial x^2}+\frac{\partial f_i^{j+1}}{\partial t}\right)+O(\tau^2+h^4). \quad (198)$$

Из (198) имеем:

1) $\psi_i^{j+1}=O(\tau^2+h^4)$ при $\sigma=\frac{1}{2}+\frac{h^2}{12\tau}$,

$$\varphi_i^{j+1}=f_i^{j+1}+\frac{h^2}{12}\left(\frac{\partial^2 f_i^{j+1}}{\partial x^2}+\frac{\partial f_i^{j+1}}{\partial t}\right),\quad u\in C_6^3;$$

2) $\psi_i^{j+1}=O(\tau^2+h^2)$ при $\sigma=\frac{1}{2}$, $\varphi_i^{j+1}=f_i^{j+1}$, $u\in C_4^3$;

3) $\psi_i^{j+1}=O(\tau+h^2)$ при σ любом $\left(\sigma\neq\frac{1}{2},\ \sigma\neq\frac{1}{2}+\frac{h^2}{12\tau}\right)$,

$$\varphi_i^{j+1}=f_i^{j+1},\quad u\in C_4^2.$$

Для задания функции v_i^1 в случае уравнения теплопроводности (172) можно воспользоваться соотношением

$$\frac{\partial u_i^1}{\partial t}=\frac{\partial u_i^0}{\partial t}+\tau\frac{\partial^2 u_i^0}{\partial t^2}+O(\tau^2)=\frac{\partial^2 u_i^0}{\partial x^2}+f_i^0+$$

$$+\tau\left(\frac{\partial^4 u_i^0}{\partial x^4}+\frac{\partial^2 f_i^0}{\partial x^2}+\frac{\partial f_i^0}{\partial t}\right)+O(\tau^2). \quad (199)$$

Если положить

$$v_i^1=\gamma_{1i}^0+\tau\gamma_{1i}^{\mathrm{II}}+\tau^2\gamma_{1i}^{\mathrm{IV}}+\tau f_i^0+\tau^2\left(\frac{\partial^2 f_i^0}{\partial x^2}+\frac{\partial f_i^0}{\partial t}\right)\quad (i=\overline{1,n}), \quad (200)$$

то трехслойная разностная схема (196), (176), (200) при $\varphi_i^{j+1}=f_i^{j+1}$ и $\sigma=\frac{1}{2}$ будет иметь порядок аппроксимации $O(\tau^2+h^2)$, а при $\sigma=\frac{1}{2}+\frac{h^2}{12\tau}$, $\varphi_i^{j+1}=f_i^{j+1}+\frac{h^2}{12}\left(\frac{\partial^2 f_i^{j+1}}{\partial x^2}+\frac{\partial f_i^{j+1}}{\partial t}\right)$ порядок $O(\tau^2+h^4)$.

Значение v_i^{j+1} $(i=\overline{1,n})$ на $(j+1)$-м временном слое находится из системы трехточечных уравнений.

Например, при $\sigma = \frac{1}{2}$, $\varphi_i^{j+1} = f_i^{j+1}$ схема (196) будет иметь вид

$$v_{i-1}^{j+1} - \left(2 + \frac{3}{2}\frac{h^2}{\tau}\right) v_i^{j+1} + v_{i+1}^{j+1} = F_i^{j+1} \quad (i = \overline{1, n};\ j = \overline{1, m-1}),$$

$$F_i^{j+1} = -h^2 f_i^{j+1} - \frac{2h^2}{\tau} v_i^j + \frac{h^2}{2\tau} v_i^{j-1}, \tag{201}$$

где v_i^0, v_0^j, v_{n+1}^j, v_i^1, $(i = \overline{1, n})$ находится соответственно по формулам (176) и (200).

Решение системы может быть найдено методом прогонки, причем для определения F_i^{j+1} используются значения $v_{h\tau}$ на двух предыдущих слоях.

Приведем примеры явных двухслойных и трехслойных разностных схем для однородного уравнения теплопроводности (172) ($\varphi = 0$), которые обладают условной аппроксимацией:

д) схема В. К. Саульева [73]

$$v_t^j + \frac{\alpha\tau}{h} v_{t\bar{x}}^j = \Lambda_1 v^j, \tag{202}$$

е) схема «ромб» [87]

$$v_{\mathring{t}}^j + \frac{\tau^2}{h^2} v_{\bar{t}t}^j = \Lambda_1 v^j \tag{203}$$

или в индексной форме записи

$$v_i^{j+1} = \frac{1}{\alpha + \beta^{-1}} [\alpha v_{i-1}^{j+1} + (1-\alpha) v_{i-1}^j + (\beta^{-1} + \alpha - 2) v_i^j + v_{i+1}^j], \tag{202'}$$

$$\beta = \frac{\tau}{h^2}, \quad \alpha > 0,$$

$$v_i^{j+1} = \frac{1}{1+2\beta} [2\beta (v_{i+1}^j + v_{i-1}^j) + (1-2\beta) v_i^{j-1}]. \tag{203'}$$

Схемы (202), (203) принадлежат к числу явных схем. Счет по формулам (202′), (203′) схематически можно представить соответственно в виде следующих «шаблонов»:

$$\begin{pmatrix} \cdot & \cdot & \\ & \cdot & \\ \cdot & \cdot & \cdot \end{pmatrix}, \quad \begin{pmatrix} & \cdot & \\ \cdot & \cdot & \cdot \end{pmatrix}.$$

Из (202) имеем

$$\psi_h = O\left(\tau + h^2 + \frac{\alpha\tau}{h}\right).$$

Аналогично из (203)

$$\psi_h = O\left(\tau^2 + h^2 + \frac{\tau^2}{h^2}\right).$$

Следовательно, схемы (202), (203) аппроксимируют исходное уравнение условно при $\frac{\tau}{h} \to 0$. Однако можно показать, что схема (203) является безусловно устойчивой. Этим же свойством (безусловной устойчивости) при $\alpha \geqslant 1$ обладает и схема (202).

5. *Разностные схемы для многомерного уравнения теплопроводности.* В цилиндре:

$$\bar{\Omega} \cup \{0 \leqslant t \leqslant T\}, \tag{204}$$

где

$\bar{\Omega} = \{0 \leqslant x_k \leqslant a,\ k = \overline{1,\ q}\}$ — q-мерный параллелепипед с границей Γ ($\bar{\Omega} = \Omega + \Gamma$), требуется найти численное решение первой краевой задачи для многомерного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = Lu + \varphi(X,\ t), \quad Lu = \sum_{k=1}^{q} L_k u, \tag{205}$$

$$L_k u = \frac{\partial^2 u}{\partial x_k^2}, \quad X = (x_1,\ x_2,\ \ldots,\ x_q),$$

$$u(X,\ 0) = \gamma_0(X), \quad u|_\Gamma = \gamma_1(X,\ t).$$

Построим разностную сетку (для простоты изложения равномерную по X и t)

$$\bar{\Omega}_{h\tau} = \Big\{(i_1 h,\ i_2 h,\ \ldots,\ i_q h,\ j\tau);\ i_k =$$

$$= \overline{0,\ n+1};\ k = \overline{1,\ q};\ j = \overline{0,\ m};\ h = \frac{a}{n+1};\ \tau = \frac{T}{m}\Big\} \tag{206}$$

с границей $\Gamma_h = \{X_l \in \Gamma\}$ и рассмотрим на множестве точек сетки $\bar{\Omega}_{h\tau}$ семейство разностных схем.

а) Явная двухслойная разностная схема

$$v_t^j = \Lambda v^j + f^j, \quad \left(\Lambda = \sum_{k=1}^{q} \Lambda_k\right), \tag{207}$$

$$v_h^0 = \gamma_{0h}, \quad v_h|_{\Gamma_h} = \gamma_{1h}. \tag{208}$$

б) Неявная двухслойная разностная схема

$$v_t^j = \Lambda v^{j+1} + f^j, \quad \left(\Lambda = \sum_{k=1}^{q} \Lambda_k\right), \tag{209}$$

$$v_h^0 = \gamma_{0h}, \quad v_h|_{\Gamma_h} = \gamma_{1h}.$$

в) Схема повышенной точности

$$v_t^j = \left(\Lambda + \frac{h^2}{6} \sum_{l=1}^{q-1} \sum_{m=l+1}^{q} \Lambda_l \Lambda_m\right) \frac{v^{j+1} + v^j}{2} -$$

$$- \left[\frac{h^2}{12} \Lambda + \frac{\tau^2}{4} (1 + \sigma^2) \sum_{l=1}^{q-1} \sum_{m=l+1}^{q} \Lambda_l \Lambda_m\right] v_t^j = \varphi^{j+\frac{1}{2}}. \tag{210}$$

Для оценки порядка аппроксимации разностных схем (207) — (210) воспользуемся формулами:

$$\Lambda u = Lu + \frac{h^2}{12} \sum_{k=1}^{q} L_k L_k u + O(h^4),$$

$$u_t = \frac{\partial u}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2} + O(\tau^2), \tag{211}$$

$$\frac{u^{j+1}+u^j}{2} = u^j + \frac{\tau}{2}\frac{\partial u^j}{\partial t} + O(\tau^2).$$

Тогда схемы (207), (209) аппроксимируют уравнение (205) со вторым порядком по пространственным переменным и первым порядком по временной переменной. Для погрешности аппроксимации схемы (210) при определенном выборе параметра σ и $\varphi^{j+\frac{1}{2}}$ справедлива следующая лемма:

Лемма 3. *Погрешность аппроксимации разностной схемы* (210), (208) *на решении* $u(X, t)$ *уравнения* (205), *если* $u(X, t)$ *обладает ограниченными производными вида*

$$\frac{\partial^{\beta_l+\gamma} u}{\partial x_l^{\beta_l}\partial t^{\gamma}} \quad (l = \overline{1, q},\ \beta_l \leqslant 6,\ \gamma \leqslant 3) \tag{212}$$

при

$$\varphi^{j+\frac{1}{2}} = f^{j+\frac{1}{2}} + \frac{h^2}{12}\Lambda f^{j+\frac{1}{2}}, \quad \sigma = \frac{h^2}{6\tau}, \tag{213}$$

будет $O(\tau^2 + h^4)$.

В самом деле, для погрешности аппроксимации разностной схемы (210) при $\varphi^{j+\frac{1}{2}} = f^{j+\frac{1}{2}} + \frac{h^2}{12}\Lambda f^{j+\frac{1}{2}}$, $\sigma = \frac{h^2}{6\tau}$ имеем:

$$\psi_h = \left(\Lambda + \frac{h^2}{6}\sum_{l<m}\Lambda_l\Lambda_m\right)\frac{u^{j+1}+u^j}{2} - \left[\frac{h^2}{12}\Lambda + \frac{\tau^2}{4}\left(1 + \frac{h^4}{36\tau^2}\right)\times\right.$$
$$\left.\times\sum_{l<m}\Lambda_l\Lambda_m\right]u_t^j + \left(I + \frac{h^2}{12}\Lambda\right)f^{j+\frac{1}{2}} - u_t^j. \tag{214}$$

Пользуясь разложениями (157) и (211), получим:

$$\psi_h = \left(L + \frac{h^2}{12}LL\right)\left(u + \frac{\tau}{2}\frac{\partial u}{\partial t}\right) - \frac{h^2}{12}L\left(\frac{\partial u}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}\right) + f +$$
$$+ \frac{\tau}{2}\frac{\partial f}{\partial t} + \frac{h^2}{12}Lf + \frac{h^2\tau}{24}L\frac{\partial f}{\partial t} - \frac{\partial u}{\partial t} - \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2} + O(|h|^4 + \tau^2).$$

Откуда

$$\psi_h = Lu + f - \frac{\partial u}{\partial t} + \frac{\tau}{2}\frac{\partial}{\partial t}\left(Lu + f - \frac{\partial u}{\partial t}\right) + \frac{h^2}{12}L\times$$
$$\times\left(Lu + f - \frac{\partial u}{\partial t}\right) + \frac{h^2\tau}{24}L\frac{\partial}{\partial t}\left(Lu + f - \frac{\partial u}{\partial t}\right) + O(\tau^2 + |h|^4),$$

или

$$\psi_h = \left(I + \frac{\tau}{2}\frac{\partial}{\partial t} + \frac{h^2}{12}L + \frac{h^2\tau}{24}L\frac{\partial}{\partial t}\right)\left(Lu + f - \frac{\partial u}{\partial t}\right) + O(\tau^2 + |h|^4).$$

Так как $Lu + f - \frac{\partial u}{\partial t} = 0$, то

$$\psi_h = O(\tau^2 + |h|^4),$$

т. е. лемма доказана.

Численная реализация явной разностной схемы (207), (208) не вызывает затруднений, чего нельзя сказать о неявных разностных схемах для многомерных нестационарных задач ($q \geqslant 2$) вида (209), (210). Однако, как будет показано ниже, явные разностные схемы являются условно устойчивыми, т. е. они устойчивы при определенных соотношениях между шагами сетки (например, при $\frac{\tau}{h^2} \leqslant \frac{1}{4}$ для схемы (207), (208) при $q = 2$).

Неявные разностные схемы обычно безусловно устойчивы. Поэтому остановимся на вопросе численной реализации неявных разностных схем для многомерных нестационарных задач более подробно.

4. Метод факторизации

Неявное многомерное разностное уравнение (209) запишем в виде

$$Dv^{j+1} = v^j + \tau f^j, \tag{215}$$

где

$$Dv^{j+1} = \left(I - \tau \sum_{k=1}^{q} \Lambda_k\right) v^{j+1}, \tag{216}$$

и наряду со схемой (215) рассмотрим схему

$$Bv^{j+1} = v^j + \tau f^j, \tag{217}$$

где

$$Bv^{j+1} = \prod_{k=1}^{q} (I - \tau\Lambda_k) f^{j+1}. \tag{218}$$

Схема (217) аппроксимирует уравнение (205) с тем же порядком, что и схема (209). В самом деле, записывая схему (218) в виде

$$\left(I - \tau \sum_{k=1}^{q} \Lambda_k + \tau^2\psi\right) v^{j+1} = v^j + \tau f^j, \tag{219}$$

где

$$\psi v^{j+1} = \left(\sum_{l=1}^{q-1} \sum_{m=l+1}^{q} \Lambda_l\Lambda_m - \tau \sum_{l=1}^{q-2} \sum_{m=l+1}^{q-1} \sum_{k=m+1}^{q} \Lambda_l\Lambda_m\Lambda_k + \cdots + \right.$$
$$\left. + (-1)^{q-2}\tau^{q-2} \prod_{k=1}^{q} \Lambda_k\right) v^{j+1}, \quad (q \geqslant 2),$$

замечаем, что (218) имеет порядок аппроксимации $O(\tau + h^2)$.

Решение системы (217), (208) может быть сведено к последовательному решению q уравнений:

$$B_1 v^{j+\frac{1}{q}} = v^j + \tau f^j,$$
$$B_k v^{j+\frac{k}{q}} = v^{j+\frac{k-1}{q}}$$
$$(B_k = I - \tau\Lambda_k, \quad k = \overline{1, q}), \tag{220}$$

дополненных граничными условиями (208). Здесь $v^{j+\frac{k}{q}}$ — промежуточные значения. Если для решения уравнения вида

$$B_k v^j = \Phi^j$$

требуется затратить Q арифметических действий, пропорциональное числу узлов сетки на j-м слое $\left(Q = O\left(\frac{1}{h^q}\right)\right)$, то такие схемы принято называть экономическими. Очевидно, схемы $Bv^{j+1} = w^j$ с оператором

$$B = \prod_{k=1}^{q} B_k \tag{221}$$

в этом случае также будут экономичными. Таким образом, предложенный метод реализации разностной схемы (217) будет экономичным. Такой подход для реализации неявной схемы (209) получил название *метода факторизации*:

По исходной схеме (209) строят так называемую производящую схему (217), имеющую тот же порядок аппроксимации, что и исходная схема (209). Для построения производящей схемы обычно многомерный оператор на верхнем слое (оператор D в (215)) заменяется факторизованным (расщепляющимся) оператором B, т. е. оператором, который может быть записан в виде произведения операторов B_k.

Производящую схему (217) можно записать в виде системы (220) q одномерных уравнений. Любое одномерное уравнение системы (220) может быть представлено как трехточечное уравнение

$$a_i v_{i-1}^{j+\frac{k}{q}} - c_i v_i^{j+\frac{k}{q}} + b_i v_{i+1}^{j+\frac{k}{q}} = -F_i^{i+\frac{k-1}{q}} \quad (i = \overline{1, n}) \tag{222}$$

с краевыми условиями

$$v_0^{j+\frac{k}{q}} = \alpha_1 v_1^{j+\frac{k}{q}} + \beta_1, \quad v_{n+1}^{j+\frac{k}{q}} = \alpha_2 v_n^{j+\frac{k}{q}} + \beta_2, \tag{223}$$

где a_i, c_i, b_i, α_1, β_1, α_2, β_2 — заданные числа.

Для решения систем вида (222), (223) может быть использован метод прогонки.

Несколько замечаний о выборе краевых условий для $v^{j+\frac{k}{q}}$.

Для того чтобы схема (220) имела порядок аппроксимации $O(\tau + h^2)$, нужно, чтобы формулы (220) для вспомогательной функции $v^{j+\frac{k}{q}}$ имели место в $\bar{\Omega}_h$, поэтому приходится вводить поправки в краевые условия для вспомогательной функции.

Поясним это на примере для случая $q = 2$.

Тогда система (220) будет иметь вид

$$(I - \tau\Lambda_1) v^{j+\frac{1}{2}} = v^j + \tau\varphi^j \quad (\varphi^j = f^j), \tag{224}$$

$$(I - \tau\Lambda_2) v^{j+1} = v^{j+\frac{1}{2}}, \tag{225}$$

где вспомогательная функция $v^{j+\frac{1}{2}} = (I - \tau\Lambda_2)\, v^{j+1}$ определена во всех точках $(x_{1i},\ x_{2k}) \in \bar{\Omega}_h$ $(0 \leqslant i \leqslant n+1,\ 0 \leqslant k \leqslant n+1)$. Для узлов сетки, принадлежащих Ω_h, имеем:

$$(1 - \tau\Lambda_1)\, v^{j+\frac{1}{2}} = v^j + \tau\varphi^j \quad (x_{1i},\ x_{2k}) \in \Omega_h, \tag{226}$$

а для граничных точек $(x_{10},\ x_{2k})$, $(x_{1,n+1},\ x_{2k})(k = \overline{1,\ n})$ она должна удовлетворять следующим граничным условиям:

$$\begin{gathered} v_{0,k}^{j+\frac{1}{2}} = (I - \tau\Lambda_2)\, \gamma_1^{j+1}(x_{10},\ x_{2k}), \\ v_{n+1,k}^{j+\frac{1}{2}} = (I - \tau\Lambda_2)\, \gamma_1^{j+1}(x_{1,n+1},\ x_{2k}) \quad (k = \overline{1,\ n}). \end{gathered} \tag{227}$$

Таким образом, система для определения вспомогательной функции $v^{j+\frac{1}{2}} \in \bar{\Omega}_h$ записывается в виде (226), (227) и состоит из n уравнений, которые могут быть сведены к виду (222), (223). Определив $v^{j+\frac{1}{2}}$, решаем n систем

$$\begin{gathered} (I - \tau\Lambda_2)\, v^{j+1} = v^{j+\frac{1}{2}}, \quad X \in \Omega_h, \\ v_{i0}^{j+1} = \gamma_1^{j+1}(x_{1i},\ x_{20}), \quad v_{i,n+1}^{j+1} = \gamma_1^{j+1}(x_{1i},\ x_{2,n+1}). \end{gathered} \tag{228}$$

Возможен и другой подход: используя граничные условия (208), из системы (217) исключим значения v_h в граничных точках. В этом случае изменятся значения правых частей уравнения в точках, лежащих вблизи границы, и решение задачи (217), (218) сведется к решению уравнения вида

$$Bv^{j+1} = v^j + \tau\tilde{\varphi}^j, \tag{229}$$

которое может быть решено методом расщепления.

В случае разностной схемы повышенной точности (210), (213) для многомерной задачи теплопроводности исходную систему записываем в виде

$$D_1 v^{j+1} = C_1 v^j + \tau\varphi^{j+\frac{1}{2}}, \tag{230}$$

где

$$\begin{gathered} D_1 = I - \frac{\tau}{2}(1 - \sigma)\Lambda + \frac{\tau^3}{4}(1 - \sigma)^2 \sum_{l<m} \Lambda_l \Lambda_m, \\ C_1 = I + \frac{\tau}{2}(1 + \sigma)\Lambda + \frac{\tau^2}{4}(1 + \sigma)^2 \sum_{l<m} \Lambda_l \Lambda_m, \quad \left(\sigma = \frac{h^2}{6\tau}\right). \end{gathered} \tag{231}$$

Факторизуя оператор верхнего слоя схемы (230), построим производящую схему

$$Rv^{j+1} = C_1 v^j + \tau\varphi^{j+\frac{1}{2}}, \tag{232}$$

где

$$R = \prod_{k=1}^{q} \left[I - \frac{\tau}{2}(1 - \sigma)\Lambda_k \right]. \tag{233}$$

Уравнение (232) можно переписать в виде

$$\left[I-\frac{\tau}{2}(1-\sigma)\sum_{k=1}^{q}\Lambda_k+\frac{\tau^2}{4}(1-\sigma)^2\sum_{l<m}\Lambda_l\Lambda_m\right]v^{j+1}-$$

$$-\frac{\tau^3}{8}(1-\sigma)^3\psi v^{j+1}=c_1v^j+\tau\varphi^{j+\frac{1}{2}}, \tag{234}$$

где

$$\psi v^{j+1}=\sum_{l<m<r}\Lambda_l\Lambda_m\Lambda_r+\cdots+(-1)^{q-3}\left(\frac{\tau}{2}\right)^{q-3}(1-\sigma)^{q-3}\prod_{k=1}^{q}\Lambda_k. \tag{235}$$

Отсюда видно, что уравнение (232) отличается от уравнения (231) или (210) наличием в левой части дополнительного слагаемого

$$\frac{\tau^3}{8}(1-\sigma)^3\psi v^{j+1},$$

где ψv^{j+1} определяется по формуле (235). Поэтому схема (210), (213), (208) при достаточной гладкости функции $u(X, t)$ будет иметь порядок аппроксимации $O(\tau^2+|h|^4)$.

Оператор разностной схемы (232), (233) будет факторизованным оператором с порядком аппроксимации $O(\tau^2+|h|^4)$. Для его численной реализации может быть применен метод расщепления.

Производящие схемы можно строить в результате факторизации операторов как на верхнем, так и на нижнем слоях исходной схемы.

Например, такой производящей схемой для разностной схемы (210), (213) является схема вида

$$Rv^{j+1}=Sv^j+\tau\varphi^{j+\frac{1}{2}},\quad \varphi^{j+\frac{1}{2}}=f^{j+\frac{1}{2}}+\frac{h^2}{12}\Lambda f^{j+\frac{1}{2}}, \tag{236}$$

где

$$S=\prod_{k=1}^{q}\left(I+\frac{\tau}{2}(1+\sigma)\Lambda_k\right). \tag{237}$$

Очевидно, производящая схема (236), (237) имеет тот же порядок аппроксимации ($O(\tau^2+|h|^4)$), что и исходная схема (210), (213), (208).

Решение разностной задачи (236), (208) может быть сведено к следующей эквивалентной системе одномерных уравнений:

$$R_1v^{j+\frac{1}{q}}=Sv^j+\tau\varphi^{j+\frac{1}{2}},$$

$$R_kv^{j+\frac{k}{q}}=v^{j+\frac{k-1}{q}}\quad(k=\overline{2,q}), \tag{238}$$

где

$$R_k=I-\frac{\tau}{2}(1-\sigma)\Lambda_k\quad(k=\overline{1,q}), \tag{239}$$

с краевыми условиями

$$v^{j+\frac{k}{q}}=\prod_{l=k+1}^{q}\left(I-\frac{\tau}{2}(1-\sigma)\Lambda_l\right)\gamma_1^{j+1} \tag{240}$$

при $X_k=0$, $X_k=a$.

Таким образом, для одной и той же схемы можно построить несколько производящих схем. Выбор той или другой производящей схемы для счета зависит от многих причин и, в частности, от используемой ЦВМ.

Иногда производящие схемы имеют порядок аппроксимации ниже, чем исходная схема. Например, для многомерного уравнения колебаний построена производящая схема порядка $Q(\tau^3 + |h|^4)$ по исходной схеме, имеющей порядок аппроксимации $Q(\tau^4 + |h|^4)$ (см. [17]).

Факторизованные схемы применимы лишь для областей Ω, представляющих собой q-мерный параллелепипед $q \geqslant 2$, так как в этом случае операторы Λ_k попарно перестановочны, и производящие схемы будут эквивалентны исходным разностным схемам.

В схемах метода факторизации возникает желание заменить краевые условия для промежуточных функций $v^{j+\frac{k}{q}}$ вида (227), (240) краевыми условиями более простого вида, например,

$$v_1^{j+\frac{k}{q}} = \gamma_1^{j+\frac{k}{q}} \quad \text{при } X_k = 0,\ X_k = a \quad (k = \overline{1, q-1}),$$

т. е. чтобы промежуточные шаги обладали теми же правами, что и целые шаги для v^j. Оказалось, что такой подход можно осуществить, но порядок точности разностной схемы при этом, как правило, понижается.

Математическое обоснование такого подхода связано с понятием суммарной аппроксимации разностной схемы, введенной А. А. Самарским [4], [64], [67]. Это позволило не только обосновать уже известные схемы метода расщепления, но и построить более общие схемы для достаточно общего класса областей.

Замечание 1. Разностные схемы вида (207) — (210) могут быть использованы для аппроксимации уравнений теплопроводности с постоянными коэффициентами, т. е. для уравнений

$$\frac{\partial u}{\partial t} = \sum_{k=1}^{q} c_k \frac{\partial^2 u}{\partial x_k^2} + f(X, t) \qquad (c_k > 0 \text{ — постоянные}), \tag{241}$$

так как вводя новые переменные

$$x_k' = \frac{x_k}{\sqrt{c_k}}, \tag{242}$$

придем к уравнению вида (205).

Замечание 2. При построении разностных схем для нестационарного уравнения с комплексными коэффициентами может быть использован тот же подход, что и в случае вещественных коэффициентов. Например, для нестационарного уравнения Шредингера

$$\begin{gathered} i\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad (i = \sqrt{-1}), \\ u(0, t) = u(a, t) = 0, \quad u(x, 0) = \gamma_1(x), \end{gathered} \tag{243}$$

двухслойная разностная схема с весами

$$\begin{gathered} iv_t = \sigma \Lambda v^{j+1} + (1-\sigma)\Lambda v^j, \quad (\sigma \text{ — любое комплексное число}), \\ v_0^j = v_{n+1}^j = 0, \quad v_k^0 = \gamma_{1k} \end{gathered} \tag{244}$$

будет иметь порядок погрешности аппроксимации не ниже $O(\tau + |h|^2)$ при любом комплексном числе σ, $O(\tau^2 + |h|^2)$ — при $\sigma = \frac{1}{2}$ и $O(\tau^2 + |h|^4)$ — при $\sigma = \frac{1}{2} - \frac{h^2}{12\tau} i$.

5. Разностные схемы для уравнений колебаний

В области (170) требуется найти решение уравнения колебаний струны

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \tag{245}$$

удовлетворяющее условиям:

$$\begin{gathered} u|_{t=0} = \gamma_0(x), \quad \left.\frac{\partial u}{\partial t}\right|_{t=0} = \gamma_1(x), \\ u(0, t) = \gamma_2(t), \quad u(a, t) = \gamma_3(t). \end{gathered} \tag{246}$$

Разностная схема для уравнения колебаний струны, если воспользоваться разностными операторами (22), § 2, для аппроксимации дифференциального уравнения и (102), § 2, для аппроксимации краевых условий, на сетке $\overline{\Omega}_{h\tau}$ (174), будет иметь вид:

$$v_{\bar{t}t} = v_{\bar{x}x} + f_i^j, \tag{247}$$

$$\begin{gathered} v_i^0 = \gamma_{0,i}, \quad v_{t,i}^0 - \frac{1}{2}\tau(\gamma_{0,i}'' + f_i^0) = \gamma_{1,i}, \\ v_0^j = \gamma_2^j, \quad v_{n+1}^j = \gamma_3^j. \end{gathered} \tag{248}$$

Погрешность аппроксимации разностной схемы (247), (248) имеет порядок $O(h^2 + \tau^2)$. Схема (247) может быть записана, например, следующим образом:

$$V_{\bar{t}t}^j = \Lambda V^j + F^j, \tag{249}$$

где

$$\begin{gathered} V^j = (v_1^j, v_2^j, \ldots, v_n^j)', \quad F^j = (\tilde{f}_1^j, f_2^j, \ldots, f_{n-1}^j, \tilde{f}_n^j)', \\ \Lambda \text{ — матрица вида (139)}, \quad \tilde{f}_1^j = f_1^j - \frac{\gamma_2^j}{h^2}, \quad \tilde{f}_n^j = f_n^j - \frac{\gamma_3^j}{h^2}, \end{gathered} \tag{250}$$

или

$$V^{j+1} = (2I + \tau^2\Lambda)V^j - V^{j-1} + \tau^2 F^j, \tag{251}$$

или

$$V_{\bar{t}}^{j+1} = V_{\bar{t}}^j + \tau\Lambda V^j + \tau F^j. \tag{252}$$

Схемы повышенной точности для многомерного уравнения колебаний. В цилиндре (204) требуется найти численное решение следующей задачи:

$$\begin{gathered} \frac{\partial^2 u}{\partial t^2} = Lu + f(X, t), \quad Lu = \sum_{k=1}^{q} L_k u, \\ L_k u = \frac{\partial^2 u}{\partial x_k^2}, \quad X = (x_1, \ldots, x_q), \\ u(X, 0) = \gamma_0(X), \quad \frac{\partial u(X, 0)}{\partial t} = \gamma_1(X), \\ u|_\Gamma = \gamma_2(X, t). \end{gathered} \tag{253}$$

На разностной сетке (206) рассмотрим семейство разностных схем:

$$[I-(\alpha-\sigma)\tau^2\Lambda]\, v^j_{\bar{t}t} = \left(\Lambda + \frac{h^2}{6}\sum_{l<m}\Lambda_l\Lambda_m\right)v^j + \varphi^j, \tag{254}$$

$$v_i^0 = \gamma_{0i}, \quad v_i^1 = \gamma_{0,i} + \frac{1}{2}\tau^2(L\gamma_{0,i} + f_i^0) + \tau\gamma_{1,i},$$
$$v_h|_{\Gamma_h} = \gamma_{2h}, \tag{254'}$$

где

$$\Lambda = \sum_{k=1}^{q}\Lambda_k, \quad \varphi^j = f^j + \frac{h^2}{12}Lf^j, \quad \sigma = \frac{h^2}{12\tau^2}, \quad \alpha \text{ — параметр}$$
$$\left(0<\alpha\ll\frac{1}{\tau}\right).$$

Для исследования порядка аппроксимации воспользуемся разложениями (211). Тогда аналогично, как это осуществлялось при исследовании порядка аппроксимации схемы (210), для схемы (254) получим:

$$\psi_h = f^j + \frac{h^2}{12}Lf^j + \left(\Lambda + \frac{h^2}{6}\sum_{l<m}\Lambda_l\Lambda_m\right)u^j - [I-(\alpha-\sigma)\tau^2\Lambda]\, u^j_{\bar{t}t} =$$
$$= \left(I + \frac{h^2}{12}L\right)\left(Lu + f - \frac{\partial^2 u}{\partial t^2}\right) + O(\tau^2 + h^4) = O(\tau^2 + h^4). \tag{255}$$

Схема (254) может быть записана следующим образом:

$$\tilde{D}v^j_{\bar{t}t} = \tilde{S}v^j + \varphi^j, \tag{256}$$

где

$$\tilde{D} = I - \left(\alpha - \frac{h^2}{12\tau^2}\right)\tau^2\Lambda, \quad \tilde{S} = \Lambda + \frac{h^2}{6}\sum_{l<m}\Lambda_l\Lambda_m, \tag{257}$$

или

$$\tilde{D}v^{j+1} = (2\tilde{D} + \tau^2\tilde{S})\, v^j + \tau^2\varphi^j - \tilde{D}v^{j-1}, \tag{258}$$

или

$$\tilde{D}v^{j+1}_{\bar{t}} = \tilde{D}v^j_{\bar{t}} + \tau\tilde{S}v^j + \tau\varphi^j. \tag{259}$$

Рассмотрим факторизованный оператор

$$\tilde{B} = \prod_{k=1}^{q}\left[I - \left(\alpha - \frac{h^2}{12\tau^2}\right)\tau^2\Lambda_k\right]. \tag{260}$$

Очевидно, разностная схема

$$\tilde{B}v^j_{\bar{t}t} = \tilde{S}v^j + \varphi^j \tag{261}$$

будет производящей для разностной схемы (256). Аналогично факторизованные разностные схемы вида

$$\tilde{B}(v^{j+1} + v^{j-1}) = (2\tilde{D} + \tau^2\tilde{S})\, v^j + \tau^2\varphi^j,$$

$$\tilde{B}v^{j+1}_{\bar{t}} = \tilde{D}v^j_{\bar{t}} + \tau\tilde{S}v^j + \tau\varphi^j$$

будут являться производящими для схемы (256). Они имеют тот же порядок аппроксимации, что и исходная схема.

6. Интегро-интерполяционный метод

Этот метод основан на использовании интегральных тождеств для краевой задачи, которые могут быть получены из физических законов, положенных в основу вывода данного дифференциального уравнения (например, законов сохранения количества тепла, движения, массы, импульса, энергии), связаны с некоторыми «интегральными» характеристиками задачи (например, с определением обобщенного решения краевой задачи, симметричностью оператора и т. п.), построены путем интегрирования дифференциального уравнения по ячейке разностной схемы и т. д.

При построении разностных схем интегро-интерполяционным методом применяют методы интерполяции интегрального соотношения, записанного относительно элементарной ячейки сетки. Изменяя интерполяцию искомого решения и коэффициентов уравнения, можно получить различные схемы.

Интегро-интерполяционный метод позволяет строить *однородные разностные схемы сквозного счета*, т. е. такие разностные схемы, коэффициенты которых вычисляются во всех узлах произвольной сетки для любой задачи из данного класса по одним и тем же формулам. Это особенно важно при рассмотрении краевых задач с разрывными коэффициентами и таких краевых задач, в которых нерегулярность разностной схемы имеет разностное происхождение (например, за счет аппроксимации решения в граничных точках).

Разностные схемы, выражающие на сетке законы сохранения, называют *консервативными схемами*. Приведем пример построения консервативной разностной схемы, использующей закон сохранения количества тепла.

Пример. Рассмотрим первую краевую задачу для стационарного уравнения теплопроводности (или задачу одномерной диффузии):

$$-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+q(x)\,u=f(x), \tag{262}$$

$$u(0)=u(a)=0, \tag{263}$$

где $p(x)\geqslant p_0>0$, $q(x)\geqslant 0$, $f(x)\in Q^0[0,\ a]$ ($Q^0[a,\ b]$ — класс кусочно-непрерывных на $[a,\ b]$ функций).

Если $p(x)$ имеет разрыв первого рода в некоторой точке ζ, то дифференциальное уравнение (262) рассматривается лишь в областях гладкости $p(x)$ $(0<x<\zeta,\ \zeta<x<a)$. Для того чтобы выделить единственное решение исходной задачи, на разрыве нужно задать дополнительные условия

$$u(\zeta+0)=u(\zeta-0),$$
$$p\left.\frac{du}{dx}\right|_{x=\zeta+0}=p\left.\frac{du}{dx}\right|_{x=\zeta-0}. \tag{264}$$

Рассмотрим неравномерную сетку

$$\bar{\Omega}_h\equiv\left\{x_i,\ x_{i+1}=x_i+h_{i+1},\ x_0=0,\ i=\overline{0,\ n},\ \sum_{l=1}^{n+1}h_l=a\right\} \tag{265}$$

и запишем уравнение баланса тепла на отрезке $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, где $x_{i+\frac{1}{2}} = x_i + \frac{1}{2} h_{i+1}$.

Пусть через сечение $x_{i-\frac{1}{2}}$ поступает поток тепла

$$Q_{i-\frac{1}{2}} = \left(-p \frac{du}{dx}\right)_{i-\frac{1}{2}}, \tag{266}$$

($p(x)$ — коэффициент теплопроводности), а через сечение $x_{i+\frac{1}{2}}$ выделяется количество тепла

$$-Q_{i+\frac{1}{2}} = \left(p \frac{du}{dx}\right)_{i+\frac{1}{2}}.$$

Если на отрезке $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ имеются источники тепла с плотностью распределения $f(x)$, то за счет них выделится количество тепла, равное

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx.$$

Величина

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} qu\,dx,$$

где q — коэффициент теплоотдачи, qu — мощность стоков тепла, будет характеризовать теплоотдачу с боковой поверхности.

Уравнение баланса тепла на отрезке $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ имеет вид

$$Q_{i-\frac{1}{2}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx = Q_{i+\frac{1}{2}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} qu\,dx \tag{267}$$

и служит основным интегральным соотношением для построения разностной схемы.

Для построения разностной схемы, основывающейся на применении простейших интерполяционных полиномов, желательно выразить значение $Q_{i\pm\frac{1}{2}}$ через решение задачи и известные функции. Для этого можно проинтегрировать равенство $\frac{du}{dx} = -\frac{Q}{p(x)}$ на отрезках $[x_{i-1}, x_i]$,

$[x_i, x_{i+1}]$ и применить простейшие интерполяционные формулы

$$u_i - u_{i-1} = -\int\limits_{x_{i-1}}^{x_i} \frac{Q}{p(x)}\,dx \approx -Q_{i-\frac{1}{2}} \int\limits_{x_{i-1}}^{x_i} \frac{dx}{p(x)},$$
$$u_{i+1} - u_i = -\int\limits_{x_i}^{x_{i+1}} \frac{Q}{p(x)}\,dx \approx -Q_{i+\frac{1}{2}} \int\limits_{x_i}^{x_{i+1}} \frac{dx}{p(x)}. \tag{268}$$

Подставляя (268) в (267), получим однородную разностную схему [75]:

$$L_h u_h = \frac{1}{\hbar_{i+1}}\left[a^*_{i+1}\frac{u_{i+1}-u_i}{\hbar_{i+1}} - a^*_i \frac{u_i - u_{i-1}}{h_i}\right] - b^*_i u_i + f^*_i = 0 \tag{269}$$
$$(i = \overline{1, n}),$$
$$u_0 = u_{n+1} = 0,$$

где введены обозначения:

$$a^*_i = \frac{1}{\frac{1}{h_i}\int\limits_{x_{i-1}}^{x_i} \frac{dx}{p(x)}}, \quad b^*_i = \frac{1}{\hbar_{i+1}} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx,$$

$$f^*_i = \frac{1}{\hbar_{i+1}} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)\,dx, \quad \hbar_{i+1} = \frac{1}{2}(h_{i+1} + h_i) \tag{270}$$

и при вычислении квадратуры $\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} qu\,dx$ использован полином нулевой степени $P_0(x) = u_i$ для интерполирования $u(x)$ при $x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$:

$$\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} qu\,dx \approx u_i \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)\,dx.$$

В классе непрерывных коэффициентов $p(x) \in \mathbf{C}_3[0, a]$, $q(x) \in \mathbf{C}_2[0, a]$ разностная схема (269) будет иметь вид

$$L_h v_h = (a u_{\bar{x}})_{\hat{x}} - q_i u_i + f_i = 0 \quad (i = \overline{1, n}),$$
$$u_0 = u_{n+1} = 0, \tag{271}$$
$$a_i = p_{i-\frac{1}{2}}, \quad q_i = q(x_i), \quad f_i = f(x_i).$$

Из консервативной разностной схемы (271) следует, что закон сохранения тепла выполняется в суммарном смысле во всей сеточной области. Для этого обозначим

$$\tilde{Q}_{i-\frac{1}{2}} = -a_i \frac{u_i - u_{i-1}}{h_i}$$

и просуммируем равенство (271) по $i = \overline{1, n}$

$$\tilde{Q}_{\frac{1}{2}} - \tilde{Q}_{n+\frac{1}{2}} + \sum_{i=1}^{n} \hbar_{i+1} q_i u_i - \sum_{i=1}^{n} \hbar_{i+1} f_i = 0. \qquad (272)$$

Равенство (272) можно интерпретировать как разностный аналог интегрального закона сохранения энергии. Вообще при построении разностных схем для отдельных уравнений или системы уравнений их нужно строить таким образом, чтобы они допускали преобразования, аналогичные преобразованиям в дифференциальном случае. Иными словами, при построении разностных схем должны выполняться не только разностные аналоги основных законов сохранения, но и все соотношения, которые диктуются физическими законами данной задачи. В этом случае схемы называются *полностью консервативными*. Полностью консервативные схемы позволяют вести расчеты на сравнительно грубых сетках (см. [72]).

Разностные схемы (269), (271) могут быть записаны в виде трехточечного шаблона

$$\begin{cases} d_i v_{i-1} - c_i v_i + b_i v_{i+1} = F_i, \\ v_0 = v_{n+1} = 0, \end{cases}$$

т. е. их решение может быть сведено к решению системы линейных уравнений с трехдиагональной матрицей.

Если точки разрыва функций $p(x)$, $q(x)$, $f(x)$ будут принадлежать множеству узлов сетки Ω_h, то соотношение баланса (267) можно получить в результате интегрирования уравнения (262) по ячейке $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, если учесть формулу (266). Проинтегрируем исходное уравнение (262) в пределах $(x_{i-\frac{1}{2}}, x)$

$$Q_{i-\frac{1}{2}} - p\frac{du}{dx} + \int_{x_{i-\frac{1}{2}}}^{x} (qu - f)\, ds = 0 \qquad (273)$$

и полученное соотношение (273), предварительно поделенное на $p(x)$ в пределах (x_{i-1}, x_i)

$$Q_{i-\frac{1}{2}} \int_{x_{i-1}}^{x_i} \frac{dx}{p(x)} - u_i + u_{i-1} + \int_{x_{i-1}}^{x_i} \frac{dx}{p(x)} \int_{x_{i-\frac{1}{2}}}^{x} (qu - f)\, ds = 0. \qquad (274)$$

Из (274) определяется $Q_{i-\frac{1}{2}}$ через решение задачи и известные функции

$$Q_{i-\frac{1}{2}} = \frac{1}{\int\limits_{x_{i-1}}^{x_i} \frac{dx}{p(x)}} \left[u_i - u_{i-1} - \int\limits_{x_{i-1}}^{x_i} \frac{dx}{p(x)} \int\limits_{x_{i-\frac{1}{2}}}^{x} (qu - f)\, ds \right]. \quad (275)$$

Аналогично

$$Q_{i+\frac{1}{2}} = \frac{1}{\int\limits_{x_i}^{x_{i+1}} \frac{dx}{p(x)}} \left[u_{i+1} - u_i - \int\limits_{x_i}^{x_{i+1}} \frac{dx}{p(x)} \int\limits_{x_{i+\frac{1}{2}}}^{x} (qu - f)\, ds \right]. \quad (276)$$

Подставляя (275), (276) в (267), получим основное интегральное тождество

$$\tilde{a}_{i+1}(u_{i+1} - u_i) - \tilde{a}_i(u_i - u_{i-1}) - \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (qu - f)\, ds =$$

$$= \tilde{a}_{i+1} \int\limits_{x_i}^{x_{i+1}} \int\limits_{x_{i+\frac{1}{2}}}^{x} \frac{q(s)\,u(s) - f(s)}{p(x)}\, ds dx - \tilde{a}_i \int\limits_{x_{i-1}}^{x_i} \int\limits_{x_{i-\frac{1}{2}}}^{x} \frac{q(s)\,u(s) - f(s)}{p(x)}\, ds dx, \quad (277)$$

где

$$\tilde{a}_i = \frac{1}{\int\limits_{x_{i-1}}^{x_i} \frac{dx}{p(x)}}. \quad (278)$$

Интегральное соотношение (277), (278) может быть использовано для получения конечно-разностных уравнений. В частности, применяя простейшие квадратурные формулы к левой части уравнения (277) и пренебрегая правой частью этого уравнения, приходим к конечно-разностному уравнению (269). С помощью разложения в соотношении (277) подынтегральных функций в ряд Тейлора легко показать, что уравнение (269) аппроксимирует (277) с погрешностью не ниже первого порядка в классе разрывных коэффициентов $p(x)$, $q(x)$, $f(x) \in \mathbf{Q}^2[0, a]$, если точки разрыва указанных коэффициентов принадлежат сетке Ω_h. Тогда

$$\| v_h - (u)_h \|_{c_h} = O(h).$$

Здесь $\mathbf{Q}^s[0, a]$ — пространство кусочно-непрерывных функций вместе с производными до s-го порядка с возможными точками разрыва первого рода.

Используя более тонкий анализ, А. Н. Тихонову и А. А. Самарскому удалось показать, что разностная схема (269), если

$$a_i^* = p_{i-\frac{1}{2}}, \quad b_i^* = \frac{h_i q_i^- + h_{i+1} q_i^+}{2\hbar_i}, \quad f_i^* = \frac{h_i f_i^- + h_{i+1} f_i^+}{2\hbar_i},$$

$$\hbar_i = \frac{1}{2}(h_{i+1} + h_i), \quad q_i^{\pm} = q_i(x_i \pm 0), \quad f_i^{\pm} = f(x_i \pm 0), \tag{279}$$

$$p(x),\ q(x),\ f(x) \in \mathbf{Q}^2[0,\ a] \tag{280}$$

и точки разрыва функций p, q и f принадлежат Ω_h, имеет второй порядок точности, а именно: если в пространство сеточных функций ввести негативную норму

$$\|\psi_h\|_{-1} = \sum_{i=1}^{n-1} h_i \left| \sum_{k=i}^{n-1} \hbar_k \psi_k \right|, \tag{281}$$

то для погрешности аппроксимации разностной схемы (269), (279)

$$\psi_h = L_h u_h - (Lu)_h \tag{282}$$

будет выполняться неравенство

$$\|\psi_h\|_{-1} \leqslant \varkappa \tilde{h}^2, \tag{283}$$

где $\varkappa = \text{const} > 0$, не зависящая от h; $\tilde{h} = (1,\ h^2]^{\frac{1}{2}}$, $h = \max|h_i|$.

Очевидно, в зависимости от условий, накладываемых на функции $p(x)$, $q(x)$, $f(x)$, выбора коэффициентов a_i^*, b_i^*, f_i^* и выбора сетки Ω_h можно строить разностные схемы, обладающие различным порядком точности. Более подробные результаты таких исследований можно найти в [4], [49], [75].

7. Вариационный метод построения разностных схем

Этот подход применяется при построении разностных схем для тех задач, которым может быть поставлена в соответствие задача минимизации некоторого функционала.

Так, например, если исходная задача

$$Au = f \tag{284}$$

может быть интерпретирована как линейное операторное уравнение в гильбертовом пространстве $\mathbf{H}$ с симметричным и положительным оператором, то задаче (279) ставится в соответствие задача отыскания $\min\limits_{u \in D(A)} \Phi(u)$, где

$$\Phi(u) = (Au,\ u) - 2(f,\ u), \tag{285}$$

причем $\min\limits_{U \in D(A)} \Phi(u) = -(Au,\ u)$ и достигается на решениях задачи (279) (метод Ритца).

Один из простейших подходов построения вариационных разностных схем связан с дискретной аппроксимацией исходного функционала $(\Phi_h(u_h))$ на некоторой сетке $\overline{\Omega}_h$ и определения значений u_h в точках сетки $\overline{\Omega}_h$ из условия минимизации функционала $\Phi_h(u_h)$.

Проиллюстрируем указанный метод построения разностной схемы для дифференциального уравнения вида (262) с краевыми условиями (263).

Задача (262), (263) может быть сведена к задаче минимизации следующего функционала:

$$\Phi(u) = \int_0^a \left[p\left(\frac{du}{dx}\right)^2 + qu^2 - 2uf \right] dx, \tag{286}$$

где допустимыми функциями являются $u \in \mathbf{W}_2^1$ и обращающиеся в нуль на концах интервала.

Введем сетку

$$\Omega_h = \left\{ x_i,\ x_i = ih,\ i = \overline{0,\ n+1},\ h = \frac{a}{n+1} \right\}. \tag{287}$$

Если функции p, q, f, u достаточно гладкие, то дискретный функционал

$$\Phi_h(u_h) \equiv h \sum_{i=0}^{n} [p_i (u_{x,i})^2 + q_i u_i^2 - 2u_i f_i] \tag{288}$$

отличается от $\Phi(u)$ на величину порядка $O(h)$. Значения u_i $(i = \overline{1, n})$ находим таким образом, чтобы они минимизировали функционал (288)

$$\frac{\partial \Phi_h(u_h)}{\partial u_i} = 0, \quad \frac{\partial^2 \Phi_h(u_h)}{\partial u_i^2} > 0 \quad (i = \overline{1, n}). \tag{289}$$

Отсюда

$$-p_i u_{x,i} + p_{i-1} u_{\bar{x},i} + hq_i u_i - hf_i = 0 \tag{290}$$

или

$$\begin{cases} -(pu_x)_{\bar{x}} + q_i u_i = f_i, & 1 \leqslant i \leqslant n, \\ u_0 = u_{n+1} = 0. \end{cases}$$

Очевидно,

$$\psi_h = -(pu_x)_{\bar{x}} + q_i u_i - f_i + \frac{d}{dx}\left(p\,\frac{du}{dx}\right)_i - q_i u_i + f_i = O(h).$$

Если вместо дискретного функционала (288) рассмотрим функционал

$$\tilde{\Phi}_h(u_h) \equiv h \sum_{i=0}^{n} [p_{i+\frac{1}{2}} (u_{x,i})^2 + q_i u_i^2 - 2u_i f_i] \tag{291}$$

над всеми функциями, определенными на сетке (287) и равными нулю при $i = 0$, $i = n + 1$, то функция, минимизирующая $\tilde{\Phi}_h(u_h)$, должна удовлетворять системе

$$\begin{gathered} -(p_{i+\frac{1}{2}} u_x)_{\bar{x},i} + q_i u_i = f_i, \\ u_0 = 0, \quad u_{n+1} = 0. \end{gathered} \tag{292}$$

Система разностных уравнений (292), как следует из (34), (37), аппроксимирует задачу (262), (263) в случае гладких функций p, q, f, u со вторым порядком точности.

Более общий подход построения разностных схем на основе вариационных принципов связан с минимизацией функционала в некотором

конечномерном подпространстве $\mathbf{F}_h$, состоящем из функций, принадлежащих $W_2^{(s+1)}(\Omega)$, если $u(x) \in \mathbf{W}_2^{(s+1)}(\Omega)$.

Наиболее известный способ построения подпространства $\mathbf{F}_h$ заключается в следующем: область $\overline{\Omega}$ покрывают конечным числом непересекающихся сеточных элементов Δ_h (триангуляторов) так, чтобы

$$\overline{\Omega} = \bigcup_{k=1}^{N} \Delta_h.$$

Если область Ω одномерная, то в качестве Δ_h выбирают отрезки, в двумерном случае — треугольники, в трехмерном — тетраэдры, h — означает максимальную длину соответственно отрезка, стороны, ребра.

На каждом из триангуляторов строят интерполяционные полиномы степени m для искомой функции таким образом, чтобы они имели кусочно-определенные производные до порядка s. $\mathbf{F}_h$ — конечное измеримое подпространство, состоящее из всех функций, которые на триангуляторах равны построенным интерполяционным полиномам. Каждая функция из $\mathbf{F}_h$ будет принадлежать $\mathbf{W}_2^{(s+1)}(\Omega)$. Затем функционал заменяется суммой по всем триангуляторам, на каждом из которых значения искомой функции заменяется построенным интерполяционным полиномом. Из условия минимизации функционала в узлах интерполяции строится система разностных уравнений.

Такой метод построения разностных схем получил название *метода конечных элементов* (МКЭ) или конечных функций, так как при этом используются функции с конечной областью ненулевых значений. Применим МКЭ к построению разностной схемы для (262) при условии, что $p(x) \geqslant p_0 > 0$, $q(x) \geqslant 0$, $f(x)$ — кусочно-непрерывные функции с возможными разрывами первого рода в некоторых точках x_k. Введем сетку

$$\overline{\overline{\Omega}} = \{x_i,\ i = \overline{0,\ n+1};\ x_0 = 0,\ x_i = x_{i-1} + h_i\}$$

таким образом, чтобы возможные точки разрывов x_k совпадали с узлами сетки. На каждом из отрезков $x_i \leqslant x \leqslant x_{i+1}$ построим линейный интерполяционный полином таким образом, чтобы на концах отрезка он принимал значения u_i.

Тогда

$$u_h(x) = \frac{x - x_i}{h_{i+1}} u_{i+1} + \frac{x_{i+1} - x}{h_{i+1}} u_i \quad (x_i \leqslant x \leqslant x_{i+1}),$$
$$u_0 = 0, \quad u_{n+1} = 0. \tag{293}$$

Функция $u_h(x)$ будет кусочно-линейной функцией на $[0, a]$, принимающей в граничных точках нулевые значения

$$u_h(x) \in \mathbf{W}_2^1(0, a).$$

Интерполяционный полином (293) удобно записать в виде

$$u_h(x) = \omega_{i1}(x) u_i + \omega_{i2}(x) u_{i+1},$$

где

$$\omega_{i1} = \frac{x_{i+1} - x}{h_{i+1}}, \quad \omega_{i2} = \frac{x - x_i}{h_{i+1}},$$

или в матричной форме

$$u_h(x) = (\omega_{i1}\omega_{i2})\, V_i, \quad V_i = (u_i,\ u_{i+1})'. \tag{294}$$

Функционал $\Phi(u)$ представим в виде

$$\Phi(u) = \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} \left[p\left(\frac{du}{dx}\right)^2 + qu^2 - 2uf \right] dx. \tag{295}$$

На каждом из отрезков $[x_i,\ x_{i+1}]$ функцию $u(x)$ заменим интерполяционным полиномом вида (294).

Тогда

$$\Phi(u_h) = \sum_{i=0}^{n} w_i,$$

где

$$w_i = \int_{x_i}^{x_{i+1}} \left[p\left(\frac{du_h(x)}{dx}\right)^2 + qu_h^2(x) - 2u_h(x) f(x) \right] dx.$$

Далее учитывая, что

$$\int_{x_i}^{x_{i+1}} p\left(\frac{du_h(x)}{dx}\right)^2 dx = \frac{1}{h_{i+1}^2} V_i' \int_{x_i}^{x_{i+1}} p(x) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} dx V_i,$$

$$\int_{x_i}^{x_{i+1}} qu_h^2(x)\, dx = V_i' \int_{x_i}^{x_{i+1}} q(x) \begin{pmatrix} \omega_{i1}^2 & \omega_{i1}\omega_{i2} \\ \omega_{i1}\omega_{i2} & \omega_{i2}^2 \end{pmatrix} dx V_i,$$

$$\int_{x_i}^{x_{i+1}} fu_h(x)\, dx = V_i' \int_{x_i}^{x_{i+1}} f(x) \begin{pmatrix} \omega_{i1} \\ \omega_{i2} \end{pmatrix} dx,$$

функционал $\Phi(u_h)$ можно записать в следующей матричной форме:

$$\Phi(u_h) = \sum_{i=0}^{n} \left(V_i' S_i V_i - 2V_i F_i \right), \tag{296}$$

где $S_i = (s_{kj}^i)_{k=\overline{1,2}}^{j=\overline{1,2}}$ — симметричная матрица; $F_i = (f_k^i)_{i=\overline{1,2}}$ — вектор-столбец, элементы которого будут определяться соответственно по формулам:

$$s_{12}^{(i)} = s_{21}^{(i)} = \frac{-1}{h_{i+1}^2} \int_{x_i}^{x_{i+1}} p(x)\, dx + \int_{x_i}^{x_{i+1}} q(x)\, \omega_{i1}\omega_{i2} dx,$$

$$s_{kk}^{(i)} = \frac{1}{h_{i+1}^2} \int_{x_i}^{x_{i+1}} p(x)\, dx + \int_{x_i}^{x_{i+1}} q(x)\, \omega_{ik}^2 dx \quad (k = 1,\ 2), \tag{296'}$$

$$f_k^i = \int_{x_i}^{x_{i+1}} f(x)\, \omega_{ik} dx.$$

Матрицу S_i иногда называют матрицей жесткости элемента $[x_i, x_{i+1}]$, вектор F_i — вектором нагрузок на элемент $[x_i, x_{i+1}]$.

Значения u_i находим так, чтобы выполнялось соотношение (289). Для этого удобно предварительно в (296) выделить слагаемые, содержащие u_i. Тогда для определения u_i получим разностную схему с трехдиагональной матрицей

$$s_{12}^{i} u_{i+1} + (s_{11}^{i} + s_{22}^{i-1}) u_i + s_{12}^{i-1} u_{i-1} = f_1^i + f_2^{i-1} \quad (i = \overline{1, n}),$$
$$u_0 = u_{n+1} = 0. \tag{297}$$

Очевидно, если ввести в рассмотрение вектор U искомых значений приближенного решения

$$U = (u_i)_{i=\overline{1,n}}, \tag{298}$$

то значения u_i будут определяться из системы

$$SU = F, \tag{299}$$

где S — симметричная трехдиагональная матрица, составленная из элементов матриц жесткости

$$S = \begin{pmatrix} s_{11}^{(1)} + s_{22}^{0} & s_{12}^{(1)} & & & \\ s_{12}^{(1)} & s_{11}^{(2)} + s_{22}^{(1)} & s_{12}^{(2)} & & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & & & s_{12}^{n-1} & s_{11}^{(n)} + s_{22}^{(n-1)} \end{pmatrix},$$

$F = (f_1^i + f_2^{i-1})_{i=\overline{1,n}}$ — n-мерный вектор.

Функция погрешности ψ_h разностной схемы (297) будет удовлетворять уравнению

$$\psi_h = -L_h u_h - \frac{d}{dx}\left(p \frac{du}{dx}\right)_i + q u_i - f_i, \quad \psi_0 = \psi_{n+1} = 0, \tag{300}$$

где $L_h U_h$ — левая часть уравнения (297). С помощью разложения в ряд Тейлора получим, что

$$\|\psi_h\| = O(h^{\alpha}), \tag{301}$$

где $\alpha = 2$ в случае равномерной сетки с отдельными точками разрыва первого рода для функций p, q, f и $\alpha = 1$ в остальных случаях (точки разрыва первого рода функций p, q, f совпадают с узловыми точками).

Пусть операторное уравнение (284) связано с многомерной краевой задачей. Для определенности рассмотрим двумерную задачу диффузии

$$-\Delta p \Delta u + qu = f \text{ в } \Omega,$$
$$u|_{\Gamma} = 0. \tag{302}$$

Коэффициенты p, q, и f будем считать кусочно-непрерывными функциями с возможными разрывами первого рода вдоль координатных линий внутри области Ω. Область Ω покроем сеткой координатных линий так, чтобы возможные разрывы функций попали на координатные линии, а затем покроем область Ω треугольной сеткой, проводя разбиение элементарных прямоугольников на треугольники (рис. 11).

На каждом из треугольников строим интерполяционный полином для искомой функции, например, второй степени

$$P_k(x_1, x_2) = a_{00} + a_{10}x_1^2 + a_{20}x_2^2 + a_{01}x_1 + a_{02}x_2 + a_{11}x_1x_2; \quad k = \overline{1, N}. \tag{303}$$

Для определения коэффициентов интерполяционного полинома в качестве узлов интерполяции можно выбрать точки вершин треугольника и точки, в которых отрезок, соединяющий две вершины треугольника, делится пополам. Коэффициенты полинома (303) определятся однозначно. Полагая $u_h(x_1, x_2) = 0$ для $(x_1, x_2) \in \Gamma$, получим функцию

$$u_h(x_1, x_2) = \begin{cases} P_1(x_1, x_2), \\ P_2(x_1, x_2), \\ \cdots\cdots \\ P_N(x_1, x_2), \end{cases} \tag{304}$$

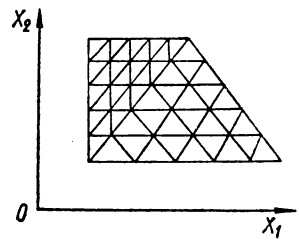


Рис. 11

которая будет непрерывной и обращается в нуль на Г, т. е.

$$u_h(x_1, x_2) \in \mathbf{W}_2^{(1)}(\Omega).$$

Непрерывность функции $u_h(x_1, x_2)$ следует из однозначного определения ее на сторонах треугольников (так как на каждой стороне треугольника интерполяционный полином $P_k(x_1, x_2)$ будет интерполяционным полиномом второй степени одной переменной, а он однозначно определяется значениями функций, заданными в трех точках).

Решение задачи в каждой из треугольных областей ищут с помощью представления (304).

Если в качестве исходного полинома (303) воспользоваться полиномом третьей степени, то число неизвестных параметров будет равно 10 и для их определения задают одно из следующих условий:

1) значения функции в вершинах треугольника, центре тяжести и точках, делящих каждую сторону треугольника на три равные части;

2) значения функции в вершинах и в центре тяжести треугольника, а также значения ее первых частных производных в вершинах.

При применении МКЭ обычно стараются использовать интерполяционные полиномы, которые имеют наибольшее из возможного числа условий, определяющих полиномиальную концентрацию вершин треугольников. Сокращение условий, которые наложены в вершинах, предполагает уменьшение порядка точности в соответствующей процедуре МКЭ.

Построение разностной схемы на основе использования интерполяционных полиномов дальше проводится так же, как и в одномерном случае,— путем минимизации функционала.

Отметим, что разностные схемы МКЭ могут быть построены на основе минимизации функционала по методу Галеркина. Но в этом случае предварительно нужно строить систему базисных функций ω_k подпространства $\mathbf{F}_h$ размерности N. Обычно каждая из базисных функций подпространств $\mathbf{F}_h$ определяется таким образом, что только один параметр из N равен единице, а остальные $N - 1$ параметров равны

нулю. Тогда любая функция $u_h \in F_h$ может быть единственным образом представлена в форме

$$u_h = \sum_{k=1}^{N} u_k \omega_k,$$

где u_k — действительные числа. Минимизируя функционал метода Галеркина по u_k, получаем вариационную разностную схему. Хотя построение системы координатных функций ω_k не всегда тривиально (простейшие примеры построения базисных функций можно найти в [50], [54], [57], [94]), однако применение метода Галеркина значительно расширяет круг задач, для которых могут быть построены вариационным методом разностные схемы.

Отдельные результаты, касающиеся сходимости приближенных решений, построенных на основе использования разностных схем МКЭ, можно найти в [49], [54], [94]. В частности отметим, что при использовании МКЭ построения разностных схем для оператора Лапласа, если использовать многочлен третьей степени, получаемая точность имеет третий порядок.

В настоящее время этот метод интенсивно развивается. К недостаткам нужно отнести недостаточно хорошую обоснованность вопросов сходимости, аппроксимации и устойчивости.

§ 4. ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ РАЗНОСТНЫХ СХЕМ[1]

Изучение устойчивости — одна из основных задач теории разностных схем. Определение устойчивости может быть введено различными способами, но суть этих определений связана с непрерывной зависимостью решения разностной задачи от входных данных, равномерной относительно шагов сетки.

Существуют различные подходы для исследования устойчивости разностных схем, большинство из которых связано с исследованием спектра разностных операторов. С обзором работ по устойчивости разностных схем можно познакомиться, например, по книгам [4], [5], [24], [28], [49], [63], [84].

Среди методов исследования устойчивости разностных схем выделим мажорантный метод, основанный на использовании принципа максимума, метод разделения переменных, метод энергетических неравенств.

Принцип максимума может быть использован для исследования устойчивости разностных схем, связанных как со стационарными разностными задачами, так и с нестационарными разностными задачами. Однако эффективное применение этого принципа связано с построением некоторых мажорантных функций, что не всегда тривиально.

Метод разделения переменных применяется для исследования самосопряженных краевых задач. В случае несамосопряженных краевых задач может быть использован спектральный метод Годунова — Рябенького [24], связанный с изучением спектра семейства операторов

[1] Основные результаты по исследованию устойчивости разностных схем, приведенные в § 4, впервые были получены А. А. Самарским.

перехода. С помощью этого метода получаются необходимые условия устойчивости.

Метод энергетических неравенств и априорных оценок применяется для исследования устойчивости конкретных классов разностных схем. Он позволяет эффективно учитывать граничные условия и переменность коэффициентов. Необходимые и достаточные условия устойчивости записываются в виде неравенств между операторными коэффициентами разностных схем. Метод априорных оценок может быть применен как для исследования устойчивости стационарных, так и нестационарных задач ([4], [5]).

В случае разностных задач с положительно определенным оператором можно воспользоваться следующими априорными оценками для нормы решения через правую часть.

Теорема 1. *Пусть разностная задача*

$$A_h v_h = \varphi_h \tag{1}$$

аппроксимирует дифференциальную задачу

$$Au = f$$

с погрешностью ψ. *Тогда, если* A_h — *положительно определенный оператор, т. е.*

$$(A_h y_h,\ y_h) \geqslant \delta (y_h,\ y_h), \quad \delta > 0, \quad y_h \in \mathbf{H}_h,$$
$$y_h = v_h - (u)_h,$$

то для функции ошибки справедлива оценка

$$\| y_h \| \leqslant \frac{1}{\delta} \| \psi \|.$$

Доказательство следует из неравенств

$$\delta^2 (y_h,\ y_h)^2 \leqslant (A_h y_h,\ y_h)^2 = (\psi,\ y_h)^2 \leqslant (\psi,\ \psi)(y_h,\ y_h).$$

Исходя из теоремы 1, можно утверждать, что разностная схема (1) с положительно определенным оператором устойчива в смысле выполнения оценки вида

$$\| v_h \| \leqslant \frac{1}{\delta} \| \varphi_h \|.$$

Например, из теоремы 1, учитывая положительную определенность оператора разностной схемы (114), § 3, получаем, что разностная задача (114), § 3, аппроксимирующая задачу (112), § 3, с погрешностью $O(h^2)$, сходится к решению задачи (112) со скоростью $O(h^2)$.

1. Принцип максимума

Для ряда разностных операторов в ограниченных сеточных областях эвклидова пространства имеет место свойство, которое называют *сеточным аналогом принципа максимума.* Принцип максимума и его следствия могут быть использованы для доказательства корректности разностных уравнений.

Пусть каждое уравнение разностной схемы (1) может быть представлено в виде

$$Rv_h = a(X)\, v_h(X) - \sum_{\zeta \in Ш(x)} b(X,\ \zeta)\, v_h(\zeta) = \varphi(X), \quad X \in \Omega_h, \tag{2}$$

где коэффициенты $a(x)$, $b(x, \zeta)$ удовлетворяют условиям

$$a(X) > 0, \quad b(X, \zeta) > 0, \quad X, \zeta \in \Omega_h, \tag{3}$$

$$d(X) = a(X) - \sum_{\zeta \in Ш(X)} b(X, \zeta) \geqslant 0, \quad X \in \Omega_h. \tag{4}$$

Здесь $Ш(X)$ — шаблон (звезда) точки X, причем суммирование проводится по всем точкам ζ, принадлежащим шаблону точки X, не включая узел X (центр звезды), $\Omega_h \in R_n$.

Пусть сетка Ω_h такова, что значение $v_h(X)$ из любой точки $X_l \in \Omega_h$ может быть перенесено в любую точку $X_m \in \Omega_h$ (Ω_h — связная сетка). Для этого достаточно предположить, что из точки $X_l \in \Omega_h$ можно попасть в точку $X_m \in \Omega_h$, используя такие точки $X_p, X_{p+1}, \ldots, X_{p+k}$, что

$$X_l \in Ш(X_p),\ X_p \in Ш(X_{p+1}), \ldots, X_{p+k-1} \in Ш(X_{p+k}),\ X_{p+k} \in Ш(X_m), \tag{5}$$

и

$$b(X_{p+j+1}, X_{p+j}) \neq 0,\ b(X_l, X_p) \neq 0,\ b(X_{p+k}, X_m) \neq 0 \quad (j = \overline{1, k-1}).$$

Теорема 2 (принцип максимума). *Пусть $v_h(x) \neq \text{const}$ задана на Ω_h и коэффициенты разностного оператора Rv_h удовлетворяют условиям* (3) — (4), *тогда, если*

$$Rv_h \leqslant 0 \quad (Rv_h \geqslant 0) \quad \forall X \in \Omega_h, \tag{6}$$

то

$$v_h(X) \leqslant 0 \quad (v_h(X) \geqslant 0) \quad \forall X \in \Omega_h. \tag{7}$$

Иными словами, при выполнении условий (3), (4), (6) *и связности Ω_h $v_h(x)$ не может принимать положительного максимума (отрицательного минимума) в точках $X \in \Omega_h$.*

Доказательство. Пусть $Rv_h \leqslant 0$. Предположим, что найдется такая точка $X_0 \in \Omega_h$, что $v_h(X_0) > 0$. Так как $v_h(x) \neq \text{const}$ и область Ω_h связная, то найдется такая точка $X_1 \in \Omega_h$ (X_1 — центр шаблона), в которой $v_h(X)$ принимает положительный максимум и, кроме того, данному шаблону будет принадлежать точка X_2 такая, что

$$v_h(X_2) < v_h(X_1).$$

Рассмотрим

$$Rv_h(X_1) = a(X_1)\, v(X_1) - \sum_{\zeta \in Ш(X_1)} b(X_1, \zeta)\, v_h(X_1) +$$

$$+ \sum_{\zeta \in Ш(X_1)} b(X_1, \zeta)\, [v_h(X_1) - v_h(\zeta)] \geqslant d(X_1)\, v(X_1) +$$

$$+ b(X_1, X_2)\, [v_h(X_1) - v_h(X_2)] > 0.$$

Узел $X_1 \in \Omega_h$ и $Rv_h(X_1) > 0$, что противоречит условию теоремы, т. е. $v_h(X) \leqslant 0$ для всех $X \in \Omega_h$.

Следствие 1. Если $R_h v_h(X) = 0$, $X \in \Omega_h$, выполнены условия (3), (4) и $d(X) \not\equiv 0$, то

$$v_h(X) = 0, \quad X \in \Omega_h. \tag{8}$$

Доказательство. На основании теоремы 2 решением уравнения

$$R_h v_h(X) = 0 \tag{9}$$

является функция $v_h(X) = \text{const}$, но так как $d(X) \not\equiv 0$, то

$$v_h(X) = 0, \quad X \in \Omega_h.$$

Следствие 2. Система (2), если выполняются условия (3), (4) и $d(X) \not\equiv 0$, имеет единственное решение при любых $\varphi(X)$.

Доказательство. На основании следствия 1 однородная система (9) имеет только тривиальное решение, т. е. система (2) будет однозначно разрешима при любых $\varphi(X)$.

Теорема 3 (сравнения). *Пусть сеточные функции $u(X)$, $v(X)$ являются соответственно решениями уравнений*

$$Ru_h(X) = F(X) \geqslant 0, \quad X \in \Omega_h, \tag{10}$$

$$Rv_h(X) = \varphi(X), \quad X \in \Omega_h, \tag{11}$$

тогда, если

$$|\varphi(X)| \leqslant F(X) \tag{12}$$

и коэффициенты разностного оператора Rv_h удовлетворяют условиям (3), (4), *то*

$$|v_h(X)| \leqslant u_h(X), \quad X \in \Omega_h. \tag{13}$$

Доказательство. Из (10), (12) имеем: $Ru_h \geqslant 0$, $R(u_h - v_h) \geqslant 0$, $R_h(u_h + v_h) \geqslant 0$, т. е.

$$u_h \geqslant 0, \quad u_h - v_h \geqslant 0, \quad u_h + v_h \geqslant 0$$

или

$$|v_h(X)| \leqslant u_h(X) \quad \forall X \in \Omega_h.$$

Функция $u_h(X)$ называется в этом случае мажорантной по отношению к функции $v_h(X)$.

Если известна мажорантная функция, то теорема сравнения может быть использована для построения априорных оценок вида (10), § 1. Однако построить мажорантную функцию не всегда удается.

Построить оценки для решения неоднородного уравнения (2) в области Ω_h, если наложены дополнительные ограничения на коэффициенты уравнения (2) в различных точках области Ω_h, можно на основании следующей теоремы.

Теорема 4. *Пусть коэффициенты и правая часть уравнения* (2) *удовлетворяют условиям*

$$a(X) > 0, \quad b(X, \zeta) > 0, \quad d(X) = \varphi(X) = 0, \quad X, \zeta \in \Omega_h^*, \tag{14}$$

$$a(X) > 0, \quad b(X, \zeta) \geqslant 0, \quad d(X) > 0, \quad X, \zeta \in \Omega_h^{**}, \tag{15}$$

где Ω_h^* — некоторое связное подмножество множества Ω_h, Ω_h^{**} — дополнение Ω_h^* до Ω_h. Тогда

$$\|v(X)\|_{C_h} \leqslant \left\| \frac{\varphi(X)}{d(X)} \right\|_{C_h^{**}}, \tag{16}$$

где

$$\|v(X)\|_{C_h} = \max_{X \in \Omega_h} |v(X)|, \quad \left\| \frac{\varphi(X)}{d(X)} \right\|_{C_h^{**}} = \max_{X \in \Omega_h^{**}} \left| \frac{\varphi(X)}{d(X)} \right|. \tag{17}$$

Доказательство. Рассмотрим уравнение

$$Ru_h(X) = |\varphi(X)|, \quad X \in \Omega_h.$$

Очевидно, $u_h(X) \geqslant 0$ будет являться мажорантной функцией для $v_h(X)$, $X \in \Omega_h$. Обозначим через $X_0 \in \Omega_h$ точку, в которой функция $u_h(X)$ достигает максимального значения. Можно считать, что точка $X_0 \in \Omega_h^{**}$. В самом деле, если $X_0 \in \Omega_h^*$, то из (14) имеем:

$$\varphi(X_0) = 0, \quad d(X_0) = a(X_0) - \sum_{\zeta \in Ш(X_0)} b(X_0, \zeta) = 0,$$

$$a(X_0)\, u_h(X_0) - \sum_{\zeta \in Ш(X_0)} b(X_0, \zeta)\, u_h(\zeta) = 0,$$

откуда

$$\sum_{\zeta \in Ш(X_0)} b(X_0, \zeta)\, [u_h(X_0) - u_h(\zeta)] = 0$$

и

$$u_h(X_0) = u_h(\zeta), \quad \zeta \in Ш(X_0).$$

Значит, $u_h(X)$ постоянная на Ω_h^* (в силу связности Ω_h^*). Пусть $u_h(X^*) = u_h(X_0)$, где $X^* \in \Omega_h^*$, а $Ш(X^*)$ содержит хотя бы одну точку X^{**} из Ω_h^{**}. Тогда $u_h(X^{**}) = u_h(X_0)$, но $X^{**} \in \Omega_h^{**}$, т. е. можно считать, что $X_0 \in \Omega_h^{**}$.

Тогда $Ru_h(X_0) = |\varphi(X_0)|$, если $X_0 \in \Omega_h^{**}$ можно записать в виде

$$d(X_0)\, u_h(X_0) - \sum_{\zeta \in Ш(X_0)} b(X_0, \zeta)\, (u_h(\zeta) - u_h(X_0)) = |\varphi(X_0)|,$$

$$u_h(X_0) \leqslant \frac{|\varphi(X_0)|}{d(X_0)},$$

$$\|u\|_{C_h} = \max_{X \in \Omega_h} |u(X)| = |u(X_0)| \leqslant \max_{X \in \Omega_h^{**}} \left| \frac{\varphi(X)}{d(X)} \right| = \left\| \frac{\varphi(X)}{d(X)} \right\|_{C_h^{**}},$$

но $u(x)$ мажорантная функция для $v(X)$, следовательно,

$$\|v_h\|_{C_h} \leqslant \|u\|_{C_h} \leqslant \left\| \frac{\varphi(X)}{d(X)} \right\|_{C_h^{**}}.$$

Следствие 1. Пусть коэффициенты и правая часть уравнения (2) удовлетворяют условию

$$a(X) > 0, \quad b(X, \zeta) > 0, \quad d(X) = \varphi(X) = 0, \quad X, \zeta \in \Omega_h^*, \tag{18}$$

$$a(X) = 1, \quad b(X, \zeta) = 0, \quad d(X) = 1, \quad \varphi(X) = \gamma(X), \quad X, \zeta \in \Omega_h^{**}, \tag{19}$$

тогда

$$\|v_h\|_{C_h} \leqslant \|\gamma\|_{C_h^{**}} = \max_{X \in \Omega_h^{**}} |\gamma(X)|. \tag{20}$$

Неравенство (20) следует из (16), если учесть (19).

З а м е ч а н и е. Очевидно, если множество точек $\Omega_h^{**} = \Gamma_h$ принять за граничные точки области Ω_h^*, то из (18) — (20) следует, что для решения однородного уравнения

$$Rv_h(X) = 0, \quad X \in \Omega_h^* \tag{21}$$

с граничными условиями

$$v_h(X) = \gamma(X), \quad X \in \Gamma_h \tag{22}$$

справедлива оценка

$$\| v_h \|_{\Omega_h^* + \Gamma_h} \leqslant \| \gamma(X) \|_{\Gamma_h}, \tag{23}$$

где

$$\| v_h \|_{\Omega_h^* + \Gamma_h} = \max_{X \in \Omega_h^* + \Gamma_h} | v_h(X) |,$$
$$\| \gamma(X) \|_{\Gamma_h} = \max_{X \in \Gamma_h} | \gamma(X) |. \tag{24}$$

Для получения оценок вида (16) в случае двуслойных разностных схем может быть использован следующий подход.

Шаблон любого узла $X_{j+1} = X(x, t_{j+1}) \in \Omega_{h\tau}$ удобно разбить на узлы, принадлежащие $j+1$ слою $Ш_{j+1}(X_{j+1})$, и узлы, принадлежащие j-му слою $Ш_j(X_{j+1})$. Здесь $\Omega_{h\tau}$ — совокупность узлов $X_j = X(x, t_j)$, где x — узел сетки Ω_h (Ω_h — состоит из конечного числа узлов, ограниченной области пространства $\mathbf{R}_n$), t_j — узел сетки $\Omega_\tau \in [0, T]$. Уравнение (2) можно записать в виде

$$a(X_{j+1}) v_h(X_{j+1}) - \sum_{\zeta \in Ш_{j+1}(X_{j+1})} b(X_{j+1}, \zeta) v_h(\zeta) = F(X_{j+1}), \tag{25}$$

где

$$F(X_{j+1}) = \varphi(X_{j+1}) + \sum_{\zeta \in Ш_j(X_{j+1})} b(X_{j+1}, \zeta) v_h(\zeta).$$

Теорема 5. *Пусть коэффициенты и правая часть уравнения* (25) *удовлетворяют условиям*

$$a(X_{j+1}) > 0 \quad \forall X_{j+1} \in \Omega_{h\tau}, \quad b(X_{j+1}, \zeta) > 0$$
$$\forall \zeta \in Ш_{j+1}(X_{j+1}), \quad Ш_j(X_{j+1}), \tag{26}$$

$$d_{j+1}(X_{j+1}) = a(X_{j+1}) - \sum_{\zeta \in Ш_{j+1}(X_{j+1})} b(X_{j+1}, \zeta) > 0 \quad \forall X_{j+1} \in \Omega_{h\tau}, \tag{27}$$

$$e_j(X_{j+1}) = \sum_{\zeta \in Ш_j(X_{j+1})} b(X_{j+1}, \zeta) > 0, \tag{28}$$

$$\frac{e_j(X_{j+1})}{d_{j+1}(X_{j+1})} \leqslant 1 + c\tau, \quad c = \text{const} > 0 \tag{29}$$

и существует такая функция $\Phi_{j+1} = \Phi(x, t_{j+1})$, *что*

$$\max_{x \in \Omega_h} \left| \frac{\varphi(x, t_{j+1})}{\tau d_{j+1}(x, t_{j+1})} \right| \leqslant \max_{x \in \Omega_h} | \Phi(x, t_{j+1}) |, \tag{30}$$

то для решения разностной задачи (25) *имеет место следующая оценка:*

$$\| v_h^{j+1} \|_{C_h} = \max_{x \in \Omega_h} | v_h(x, t_{j+1}) | \leqslant e^{ct_j} \left(\| v_h^0 \|_{C_h} + \sum_{k=1}^{j+1} \tau \| \Phi(x, t_k) \|_{C_h} \right). \tag{31}$$

Доказательство. На основании теоремы 4, предполагая, что Ω_h^* — пустое множество, а в Ω_h^{**} выполняются условия (26), (27), получим:

$$\|v^{j+1}\|_{C_h} \leqslant \left\| \frac{F(X_{j+1})}{d_{j+1}(X_{j+1})} \right\|_{C_h}. \tag{32}$$

Из (30) и (29) имеем:

$$\left\| \frac{F(X_{j+1})}{d_{j+1}(X_{j+1})} \right\|_{C_h} \leqslant \left\| \frac{\varphi(X_{j+1})}{d_{j+1}(X_{j+1})} \right\|_{C_h} + \left\| \frac{e_j(X_{j+1})}{d_{j+1}(X_{j+1})} \right\|_{C_h} \|v_h^j\|_{C_h} \leqslant$$
$$\leqslant \tau \|\Phi_{j+1}\|_{C_h} + (1 + c\tau) \|v_h^j\|_{C_h} \leqslant \tau \|\Phi_{j+1}\|_{C_h} + e^{c\tau} \|v_h^j\|_{C_h}.$$

Если учесть (32), то получим рекуррентное соотношение, из которого следует, что

$$\|v^{j+1}\|_{C_h} \leqslant \left(\sum_{k=1}^{j+1} \tau \|\Phi_k\|_{C_h} + \|v_h^0\|_{C_h} \right) e^{cj\tau}.$$

Примеры построения априорных оценок вида (10), § 1, полученных на основе использования принципа максимума.

Пример 1. Разностная задача Дирихле для уравнения Пуассона.

Рассмотрим разностную схему (133), (136), § 3, и запишем ее в виде

$$\left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) v_{ik} - \left(\frac{v_{i+1,k}}{h_1^2} + \frac{v_{i-1,k}}{h_1^2} + \frac{v_{i,k+1}}{h_2^2} + \frac{v_{i,k-1}}{h_2^2}\right) = -f_{ik}, \tag{33}$$

$$v_{0k} = \gamma_{0k}, \quad v_{n+1,k} = \gamma_{n+1,k},$$

$$v_{i0} = \gamma_{i0}, \quad v_{i,m+1} = \gamma_{i,m+1}, \quad (i = \overline{1, n}, \quad k = \overline{1, m})$$

$$a(x_{1i}, x_{2k}) = \begin{cases} \dfrac{2}{h_1^2} + \dfrac{2}{h_2^2}, & i = \overline{1, n}; \quad k = \overline{1, m}, \\ 1, & i = 0, n+1; \quad k = \overline{1, m}, \\ 1, & i = \overline{1, n}; \quad k = 0, m+1, \end{cases}$$

$$b(X, \zeta) = \left\{ \frac{1}{h_1^2}, \frac{1}{h_1^2}, \frac{1}{h_2^2}, \frac{1}{h_2^2} \right\},$$

$$d(x_{1i}, x_{2k}) = \begin{cases} 0 & i = \overline{2, n-1}; \quad k = \overline{2, m-1}, \\ \dfrac{1}{h_1^2} & i = 1, n; \quad k = \overline{1, m}, \\ \dfrac{1}{h_2^2} & i = \overline{1, n}; \quad k = 1, m, \\ 1 & i = 0, n+1; \quad k = \overline{1, m}, \\ 1 & i = \overline{1, n}, \quad k = 0, m+1. \end{cases}$$

т. е. коэффициенты разностной схемы (33) удовлетворяют условиям (3), (4). Область Ω_h — связная, поэтому на основании следствия 2, теоремы 2 существует единственное решение разностного уравнения (33).

Представим решение неоднородного разностного уравнения (33) в виде

$$v_h = \tilde{v}_h + \underset{\sim}{v}_h, \tag{34}$$

где $\tilde{v}_h$ — решение неоднородного уравнения (33) с однородными краевыми условиями, $\underset{\sim}{v}_h$ — решение соответственно однородного уравнения с неоднородными краевыми условиями. Тогда в соответствии с замечанием к следствию 1, теоремы 4 будем иметь:

$$\|\underset{\sim}{v}_h\|_{C_h} \leqslant \|\gamma\|_{\Gamma_h}, \quad \|v\|_{C_h} = \max_{X \in \Omega_h + \Gamma_h} |v(X)|. \tag{35}$$

Для оценки $\tilde{v}_h$ в $\overline{\Omega}_h$ воспользуемся мажорантной функцией вида

$$u(x_1, x_2) = \frac{\delta}{4}[\rho^2 - (x_1 - b_1)^2 - (x_2 - b_2)^2], \tag{36}$$

где $\delta = \max\limits_{X \in \Omega_h} |f(X)|$, причем для точек $X \in \Omega_h$, в которых правые части разностной системы (33) для $\tilde{v}_h$ совпадают с однородными граничными условиями, можно считать $f(X) = 0$, ρ — радиус окружности, содержащей область Ω_h внутри, b_1, b_2 — координаты центра этой окружности.

Тогда

$$-\Delta_h u_h(X) = \delta, \quad X = X(x_1, x_2) \in \Omega_h,$$
$$u_{0k} = u_{n+1,k} = 0, \quad k = \overline{1, m},$$
$$u_{i0} = u_{i,m+1} = 0, \quad i = \overline{1, n},$$

т. е. функция $u(x_1, x_2)$ будет мажорантной для функции v_h и

$$\|\tilde{v}_h\|_{C_h} \leqslant \|u_h\|_{C_h} \leqslant \frac{\rho^2\delta}{4} = \frac{\rho^2}{4}\|f_h\|_{\tilde{C}_h}, \quad \|f_h\|_{\tilde{C}_h} = \max_{X \in \Omega_h} |f(X)|.$$

Поэтому

$$\|v_h\|_{C_h} \leqslant \|\tilde{v}_h\|_{C_h} + \|\underset{\sim}{v}_h\|_{C_h} \leqslant \frac{\rho^2}{4}\|f_h\|_{\tilde{C}_h} + \|\gamma\|_{\Gamma_h}, \tag{37}$$

что означает устойчивость разностной схемы (133), § 3. Следовательно, разностная схема (133), § 3, сходится со скоростью $O(|h|^2)$ к решению задачи (130), (131), § 3.

Пример 2. Рассмотрим неявную разностную схему для уравнения теплопроводности (179), § 3. Запишем её в виде

$$\begin{gathered} v_i^{j+1}\left(\frac{1}{\tau} + \frac{2}{h^2}\right) - \frac{1}{h^2}(v_i^{j+1} + v_{i-1}^{j+1}) - \frac{1}{\tau} v_i^j = f_i^j, \\ v_i^0 = \gamma_{0i}, \quad v_0^j = \gamma_1^j, \quad v_{n+1}^j = \gamma_2^j \quad (i = \overline{1, n}; \quad j = \overline{0, m}). \end{gathered} \tag{38}$$

В этом случае

$$a(x_i, t_j) = \begin{cases} \dfrac{1}{\tau} + \dfrac{2}{h^2}, & i = \overline{1, n}, \quad j = 1, 2, \ldots \\ 1, & i = \overline{1, n}, \quad j = 0, \\ 1, & i = 0, n+1, \quad j = 1, 2, \ldots \end{cases} \tag{39}$$

$$b(x, \zeta) = \left\{\frac{1}{h^2}, \frac{1}{h^2}, \frac{1}{\tau}\right\},$$

$$d(x_i, t_j) = \begin{cases} 0, & i = \overline{1, n}; \quad j = 1, 2, \ldots \\ 1, & i = \overline{1, n}; \quad j = 0 \\ 1, & i = 0, n+1; \quad j = 1, 2, \ldots \end{cases} \tag{40}$$

Представим решение неоднородного разностного уравнения (38) в виде

$$v_h = \underset{\sim}{v}_h + \tilde{v}_h, \tag{41}$$

где $\underset{\sim}{v_h}$— решение однородного разностного уравнения с неоднородными начальными условиями ($v_i^0 = 0$), $\tilde{v}_h$ — решение неоднородного уравнения с однородными краевыми условиями ($\tilde{v}_0^j = 0$, $\tilde{v}_{n+1}^j = 0$, $j = 1, 2$). Так как условия (18), (19) для функции $\underset{\sim}{v_h}$ выполнены, то из (20) получим:

$$\| \underset{\sim}{v_h} \|_{C_h} \leqslant \max_{0<j\leqslant m+1} (|\gamma_1^j| + |\gamma_2^j|). \tag{42}$$

Для оценки $\| \tilde{v}_h \|_{C_h}$ воспользуемся теоремой 5, записав уравнение для $\tilde{v}_h$ в виде

$$\begin{gathered} \tilde{v}_i^{j+1}\left(\frac{1}{\tau} + \frac{1}{h^2}\right) - \frac{1}{h^2}(\tilde{v}_i^{j+1} + \tilde{v}_{i-1}^{j+1}) = f_i^j + \frac{1}{\tau} v_i^j, \\ \tilde{v}_0^j = 0, \quad \tilde{v}_{n+1}^j = 0, \quad \tilde{v}_i^0 = \gamma_{0i} \quad (i = \overline{1, n}; \quad j = \overline{0, m}). \end{gathered} \tag{43}$$

Очевидно, в этом случае

$$d_{j+1}(x_i, t_{j+1}) = \frac{1}{\tau} > 0, \quad i = \overline{1, n}, \quad j = \overline{1, m},$$

$$e_j = \frac{1}{\tau}, \quad \frac{e_j}{d_{j+1}} = 1$$

и

$$\left\| \frac{\varphi_{j+1}}{\tau d_{j+1}} \right\|_{C_h} = \| \varphi_{j+1} \|_{C_h}, \quad \text{где} \quad \| \varphi_{j+1} \|_{C_h} = \max_{X \in \Omega_h} \| f(x, t_j) \|.$$

Поэтому, используя (31), получим:

$$\| \tilde{v}_h^{j+1} \|_{C_h} = \| \tilde{v}_h^0 \|_{C_h} + \sum_{k=1}^{j+1} \tau \| \varphi_k \|_{C_h}.$$

Итак,

$$\| v_h \|_{C_h} \leqslant \| \tilde{v}_h^{j+1} \|_{C_h} + \| \underset{\sim}{v_h^{j+1}} \|_{C_h} \leqslant$$

$$\leqslant \| \gamma_0 \|_{C_h} + \sum_{k=1}^{j+1} \tau \| \varphi_k \|_{C_h} + \max_{0\leqslant k\leqslant j+1} (|\gamma_1(t_k)| + |\gamma_2(t_k)|),$$

т. е. неявная разностная схема (179), § 3, при любых τ, h абсолютно аппроксимирует уравнение (172), (173), § 3, с порядком $O(\tau + h^2)$ и сильно устойчива.

В случае явной разностной схемы (175), § 3, для уравнения теплопроводности, проводя совершенно аналогичные рассуждения, как и в случае неявной схемы, замечаем, что для того чтобы выполнялись условия теоремы 4 при оценке $\underset{\sim}{v_h}$, должно иметь место соотношение

$$d = \frac{1}{\tau} - \frac{2}{h^2} \geqslant 0,$$

или

$$\tau \leqslant \frac{h^2}{2}. \tag{44}$$

Итак, явная двухслойная схема (175), § 3, аппроксимирует уравнение (172), (173), § 3, с порядком $O(\tau + h^2)$, но условно устойчива $\left(\tau \leqslant \frac{h^2}{2}\right)$.

2. Операторно-разностные схемы. Корректность операторно-разностных схем

К разностным методам решения нестационарных задач полностью применима общая теория метода сеток. Однако в нестационарных задачах роль временной переменной отличается от роли пространственных переменных. Это прежде всего связано с тем фактом, что состояние процесса, описываемого нестационарным уравнением, в данный момент времени зависит от его состояния в прошедшее время и не зависит от состояния в следующий момент времени. Поэтому в нестационарных разностных задачах значения функции на j-м временном слое определяются только через значения функции на предыдущих временных слоях и не зависят от значений функции на временных слоях, следующих за j-м. Сеточную функцию, определенную на j-м временном слое, обозначим через v_h^j. Линейное пространство, образуемое такими функциями,— через $\mathbf{H}_h$. Разностные методы приводят к системам уравнений, которые содержат неизвестную функцию v_h^j на некотором конечном числе слоев временной сетки

$$\overline{\Omega}_\tau = \{t_j = j\tau, \quad j = 0, 1, \ldots q\}. \tag{45}$$

Линейная разностная нестационарная задача может быть представлена в виде m-слойной разностной схемы

$$\sum_{k=0}^{m-1} B_k(t)\, v_h^{j+1-k} = F_h^j \quad (j = m-2, m-1, m, \ldots;\ m = 2, 3, \ldots), \tag{46}$$

где $v_h^0, v_h^1, \ldots, v_h^{m-2}$ — заданные значения, B_k — линейные операторы ($k = \overline{0, m-1}$), $B_k : \mathbf{H}_h \to \mathbf{H}_h$, B_0 — обратимый линейный оператор, F_h — правая часть уравнения (46) определяется исходными данными и возможно значениями функции на предыдущих слоях. Разностные схемы (175), (178), (181), (194), (202), § 3, являются примерами двухслойных разностных схем, разностные схемы (196), (203), (249), (256), (261), § 3,— примерами трехслойных разностных схем.

Если переход от слоя $t = t_j$ к слою $t = t_{j+1}$ в свою очередь осуществляется при помощи операторной матрицы перехода, то такие разностные схемы принадлежат к числу составных схем. Составная разностная схема называется m-слойной разностной схемой q-го ранга, если она может быть представлена в виде

$$\sum_{k=1}^{q} B_{\alpha k}(t_j)\, v_h^{j+\frac{k}{q}} = \sum_{\beta=0}^{m-2} D_{\alpha\beta} v^{j-\beta} + F_h^{j+\frac{\alpha}{q}}, \tag{47}$$

$$\alpha = \overline{1, q},\ j = m-2, m-1, m, \ldots;\ m = 2, 3, \ldots,$$

где $B_{\alpha k}$, $D_{\alpha\beta} : \mathbf{H}_h \to \mathbf{H}_h$, $v^{j+\frac{k}{q}}$ $(k = \overline{1, q-1})$ — промежуточные функции.

В частности, двухслойная разностная схема ранга q будет иметь вид

$$\sum_{k=1}^{q} B_{\alpha k} v_h^{j+\frac{k}{q}} = D_{\alpha 0} v_h^j + F_h^{j+\frac{\alpha}{q}} \quad (\alpha = \overline{1, q}). \tag{48}$$

Чтобы по заданному v_h^j найти v_h^{j+1}, нужно решить систему уравнений с операторной матрицей перехода $B = (B_{\alpha k})(\alpha,\ k = \overline{1,\ q})$.

Примером составных двухслойных разностных схем ранга q являются производящие схемы вида (217), (236), § 3.

Назовем решением задачи (46) в момент времени t_j вектор

$$v^j = (v_h^j,\ v_h^{j-1},\ \ldots,\ v_h^{j-(m-2)}, \tag{49}$$

компоненты которого удовлетворяют уравнению (46) и каждая из компонент является элементом пространства $\mathbf{H}_h$.

Рассмотрим линейное пространство $\mathbf{H}_h^{(m-1)}$, образованное векторами вида (49), в котором сложение векторов и умножение на число α определяется покоординатно, т. е.

$$v^j + z^j = (v_h^j + z_h^j,\ v_h^{j-1} + z_h^{j-1},\ \ldots,\ v_h^{j-(m-2)} + z_h^{j-(m-2)}),$$

$$\alpha v^j = (\alpha v_h^j,\ \alpha v_h^{j-1},\ \ldots,\ \alpha v_h^{j-(m-2)}),$$

$$z^j = (z_h^j,\ \ldots,\ z_h^{j-m-2}).$$

Нулевым элементом пространства $\mathbf{H}_h^{(m-1)}$ будет вектор, каждая компонента которого есть нуль пространства $\mathbf{H}_h$.

Пространство $\mathbf{H}_h^{(m-1)}$ будет называться прямой суммой пространств $\mathbf{H}_h$:

$$\mathbf{H}_h^{(m-1)} = \underbrace{\mathbf{H}_h \oplus \mathbf{H}_h \oplus \cdots \oplus \mathbf{H}_h}_{m-1}. \tag{50}$$

Если в $\mathbf{H}_h$ заданы операторы $D_{\alpha k}$ $(\alpha,\ k = \overline{1,\ m-1})$, $D_{\alpha k} : \mathbf{H}_h \to \mathbf{H}_h$, то оператор $D = (D_{\alpha k}) : \mathbf{H}_h^{(m-1)} \to \mathbf{H}_h^{(m-1)}$ в пространстве $\mathbf{H}_h^{(m-1)}$ будет определяться как квадратная матрица порядка $(m-1)$ с компонентами $D_{\alpha k}$. Для таких операторных матриц справедливы обычные правила сложения матриц и умножения матриц на число. Если учесть, что пространство $\mathbf{H}_h^{(m-1)}$ образовано векторами вида (49), то m-слойную операторно-разностную схему (46) можно записать следующим образом:

$$v^{j+1} = \mathrm{T}v^j + F^j, \quad j = 0,\ 1,\ \ldots, \tag{51}$$

где $v^0 = (v_h^0,\ \ldots,\ v_h^{m-2})$ — задано,

$$\mathrm{T} = \begin{pmatrix} -B_0^{-1}B_1 & -B_0^{-1}B_2 & \ldots & -B_0^{-1}B_{m-1} \\ I & 0 & \ldots & 0 \\ 0 & & \ldots\ I & 0 \end{pmatrix}, \tag{52}$$

$$F^j = (B_0^{-1}F_h^j,\ 0,\ \ldots,\ 0). \tag{53}$$

Запись уравнения (46) в виде (51) равносильна введению новых зависимых переменных:

$$v^{j-1} = \omega^j,$$

$$v^{j-2} = u^j,$$

$$v^{j-3} = w^j,$$

$$\cdots\cdots$$

Уравнение (46) можно тогда записать в виде системы

$$\begin{cases} v^{j+1} = -B_0^{-1}B_1 \quad v^j - B_0^{-1}B_2 \quad \omega^j - B_0^{-1}B_3 \quad u^j - \cdots + B_0^{-1}F_n^j \\ \omega^{j+1} = v^j \\ u^{j+1} = \omega^j \\ w^{j+1} = u^j \\ \cdots\cdots, \end{cases} \tag{51'}$$

у которой индексы принимают значения j и $j+1$ (т. е. только 2 значения).

Таким образом, m-слойную операторно-разностную схему в пространстве $\mathbf{H}_h^{(m-1)}$ можно представить как двухслойную разностную схему с оператором перехода T. Устойчивость m-слойной схемы (46) определяется как устойчивость эквивалентной ей двухслойной схемы (51) — (53). Поэтому более подробно остановимся на вопросах исследования устойчивости двухслойных разностных схем. Однако следует отметить, что исследование устойчивости 3, 4, 5-слойных разностных схем во многих случаях удобнее проводить непосредственно исходя из определения этих схем, чем путем сведения их к двухслойным разностным схемам.

Двухслойную разностную схему можно записать в виде

$$B_0 v_h^{j+1} + B_1 v_h^j = \varphi_h^j, \quad j = 0, 1, \ldots, \tag{54}$$
$$v_h^0 = \gamma,$$

где B_0 — обратимый оператор из $\mathbf{H}_h \to \mathbf{H}_h$, B_1 — линейный оператор из $\mathbf{H}_h \to \mathbf{H}_h$, φ_h^j, v_h^0 — заданные функции, называемые правыми частями разностного уравнения и начальными условиями. Пусть в $\mathbf{H}_h$ определены некоторые нормы $\| v_h^j \|_{(h,j)}$, $\| \varphi_h^j \|_{(2h,j)}$ (нормы на слое).

Разностная схема (54) называется *корректной,* если существует единственное решение разностной схемы при любых входных данных и оно непрерывно зависит от входных данных (условие устойчивости). Эта зависимость равномерна по h и τ, т. е. при всех достаточно малых h и τ ($|h| \leqslant h_0$, $\tau \leqslant \tau_0$) и любых v_h^0, $\Phi(t)$ выполняется неравенство

$$\| v_h^{j+1} \|_{(h,j+1)} \leqslant \varkappa_1 \| v_h^0 \|_{(h,0)} + \varkappa_2 N(\Phi(t)), \tag{55}$$

где $N(\Phi(t))$ — функционал от вектора $\Phi(t) = (\varphi_h^0, \varphi_h^1, \ldots, \varphi_h^j)$, обладающий свойствами нормы; $\varkappa_1$, $\varkappa_2$ — положительные постоянные, не зависящие от h, τ, v_h^0, $\Phi(t)$. Соответственно устойчивость m-слойной схемы (46) может быть определена следующим образом.

Разностную схему (46) будем называть *устойчивой*, если при любых начальных данных $v_h^0 = (v_h^0, \ldots, v_h^{m-2})$ и любых правых частях выполняется неравенство

$$\| v^{j+1} \|_{(h,j+1)} \leqslant \varkappa_1 \| v^0 \|_{(1,0)} + \varkappa_2 N(F(t)),$$

где $\| v^{j+1} \|_{(h,j+1)}$ — какая-либо норма в пространстве $\mathbf{H}_h^{m-1}$; $N(F(t))$ — функционал от $F(t) = (F_h^0, F_h^{(1)}, \ldots, F_h^j)$, обладающий свойствами нормы; $\varkappa_1 = \text{const} > 0$, $\varkappa_2 = \text{const} > 0$, $\varkappa_1$, $\varkappa_2$ не зависят от h, τ, v^0, $F(t)$.

Схема, устойчивая при любых допустимых h и τ, называется *абсолютно устойчивой.*

Разностная схема, устойчивая при дополнительных условиях, связывающих h и τ, называется *условно устойчивой.*

Двухслойная разностная схема (54) может быть записана в следующей эквивалентной форме (каноническая форма двухслойной схемы):

$$\begin{cases} B(t)v_t^j + A(t)v^j = \varphi, \quad 0 \leqslant t_j = t_{j-1} + \tau_j \leqslant T_0, \\ v^0 = \gamma(x), \end{cases} \tag{56}$$

где операторы B, A связаны с операторами B_0, B_1 соотношениями

$$B = \tau B_0, \quad A = B_1 + B_0, \qquad v_t^j = \frac{v_h^{j+1} - v_h^j}{\tau}. \tag{57}$$

Решение разностной задачи (54) можно представить в виде $v_h^j = \underset{\sim}{v}_h^j + \tilde{v}_h^j$, где $\underset{\sim}{v}_h^j$ — решение однородного разностного уравнения с неоднородными начальными условиями

$$B(t)\underset{\sim}{v}_t^j + A(t)\underset{\sim}{v}^j = 0, \quad 0 \leqslant t_j \leqslant T_0, \qquad \underset{\sim}{v}^0 = \gamma, \tag{58}$$

$\tilde{v}_h^j$ — решение неоднородного уравнения с однородными начальными условиями

$$B(t)\tilde{v}_t^j + A(t)\tilde{v}^j = \varphi, \quad 0 \leqslant t_j \leqslant T_0, \qquad \tilde{v}^0 = 0. \tag{59}$$

Поэтому естественно различают устойчивость разностной схемы (54) по начальным данным и правой части.

Если для решения задачи (59) имеет место оценка (55) при $\tilde{v}_h^0 = 0$, то схема (54) называется *устойчивой по правой части,*

$$\| v_h^{j+1} \|_{(h,j+1)} \leqslant \varkappa_2 N(\Phi(t)). \tag{60}$$

Если для решения задачи (58) верна оценка (55) при $\varphi = 0$, то схема (54) называется *устойчивой по начальным данным,*

$$\| v_h^{j+1} \|_{(h,j+1)} \leqslant \varkappa_1 \| v_h^0 \|_{(h,0)} \tag{61}$$

и *равномерно устойчиеа по начальным данным,* если она устойчива относительно возмущений, вносимых на каждом слое, т. е. для решения задачи (58) для любого $v_h^k \in \mathbf{H}_h$ $(k = \overline{0, j})$ выполняется оценка

$$\| v_h^{j+1} \|_{(h,j+1)} \leqslant \varkappa_1 \| v_h^k \|_{(h,k)} \quad (k = \overline{0, j},\ j = 1, 2, \ldots), \tag{62}$$

где $\varkappa_1$ не зависит от выбора h, τ, v_h^k $(k = \overline{0, j},\ j = 1, 2, \ldots)$. При $\varkappa_1 = 1$ соотношение (62) гарантирует ненарастание погрешностей, внесенных на каждом промежуточном временном слое.

Устойчивость двухслойной разностной схемы (54) по начальным данным во многих случаях удается исследовать, используя энергетические нормы.

Рассмотрим семейство двухслойных разностных схем (58) в вещественном гильбертовом пространстве $\mathbf{H}_M$ со скалярным произведением

$$(u, v)_M = (Mu, v) \tag{63}$$

и нормой

$$\|v\|_M = \sqrt{(Mv, v)}, \tag{64}$$

где

$$M = M^* > 0, \quad M : \mathbf{H}_h \to \mathbf{H}_h. \tag{65}$$

Введем понятие устойчивости схемы (58) в пространстве $\mathbf{H}_M$ с постоянной ρ, которое будет соответствовать равномерной устойчивости схемы (58) по начальным данным в $\mathbf{H}_M$.

Схема (58) в пространстве $\mathbf{H}_M$ будет называться *устойчивой с постоянной* $\rho > 0$ (ρ-устойчива), если при любых $v_h^j \in \mathbf{H}_h$, являющихся решением уравнения (58), для всех j справедлива оценка

$$\|v_h^{j+1}\|_M \leqslant \rho \|v_h^j\|_M, \tag{66}$$

причем $\rho^j \leqslant \varkappa_1$ для всех $j\tau \leqslant T_0$, $\varkappa_1 = \text{const} > 0$, $\varkappa_1$ не зависит от h, τ, j. Из (66) следует

$$\|v_h^{j+1}\|_M \leqslant \rho^{j+1} \|v_h^0\|_M \leqslant \varkappa_1 \|v_h^0\|_M, \tag{67}$$

т. е. из ρ-устойчивости схемы (58) в $\mathbf{H}_M$ следует устойчивость по начальным данным в $\mathbf{H}_M$.

Если

$$A = A^* > 0,$$

то устойчивость схемы (58) можно исследовать в нормах пространства $\mathbf{H}_A$, а при $B = B^* > 0$ в нормах пространства $\mathbf{H}_B$. Очевидно, схема, неустойчивая в $\mathbf{H}_M$, может оказаться устойчивой в $\mathbf{H}_{\tilde{M}}$.

Постоянная $\rho > 0$ в неравенстве (67) вообще может зависеть от h и τ. В частности, если $\rho = 1$, то разностная схема оказывается устойчивой при любых $\tau > 0$, $h > 0$, поэтому она будет абсолютно устойчива в $\mathbf{H}_M$.

Если $\rho = 1 + c\tau$ ($c = \text{const} > 0$, не зависящая от h и τ), то при малых τ в этом случае допускается экспоненциальное возрастание со временем погрешностей округлений. Такие разностные схемы называют слабо устойчивыми. Если $\rho = 1 - \beta(h, \tau, T_0)$, где $\beta > 0$ и $\lim\limits_{h\to 0, \tau\to 0} \beta(h, \tau, T_0) \to 0$, то такие схемы называют *сильно устойчивыми по начальным данным* в $\mathbf{H}_M$. Если $\rho = e^{-\tau\delta(h,\tau)}$, где $\delta(h, \tau) > 0$ при всех h и τ и $\lim\limits_{h\to 0, \tau\to 0} \delta(h, \tau) = \delta_0 > 0$, то разностную схему (58) называют *асимптотически устойчивой* в $\mathbf{H}_M$ при $t_j \to \infty$.

Необходимые и достаточные условия устойчивости операторно-разностных схем. Рассмотрим семейство двухслойных разностных схем

$$\begin{gathered} Bv_t^j + Av_h^j = \varphi^j, \quad t \leqslant t_j = j\tau \leqslant T_0, \\ v_h^0 = \gamma(x) \end{gathered} \tag{68}$$

с операторами A и B, не зависящими от t_j и удовлетворяющими условиям

$$A = A^* > 0, \tag{69}$$

$$B > 0. \tag{70}$$

Двухслойную неявную схему (68) можно записать в виде явной двухслойной схемы. В самом деле, для рассматриваемого семейства схем оператор B^{-1} существует, и область его определения $D(B^{-1})$ будет совпадать со всем пространством $\mathbf{H}_h$. Поскольку $A = A^* > 0$, то существует $A^{\frac{1}{2}}$ — корень квадратный из оператора A и

$$A^{\frac{1}{2}} = (A^{\frac{1}{2}})^*. \tag{71}$$

Применим к (68) оператор $A^{\frac{1}{2}}B^{-1}$ и обозначим:

$$u_h = A^{\frac{1}{2}} v_h, \tag{71'}$$

тогда

$$u_h^{j+1} = (I - \tau A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}})\, u_h^j + \tau A^{\frac{1}{2}} B^{-1} \varphi_h^j,$$

или

$$u_h^{j+1} = T u_h^j + \tau \psi_h^j, \quad j = 0, 1, \ldots, \tag{72}$$

где

$$T = I - \tau S \text{ (оператор перехода)}, \tag{73}$$

$$S = A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}, \tag{74}$$

$$\psi_h = A^{\frac{1}{2}} B^{-1} \varphi_h^j. \tag{75}$$

Если двухслойная разностная схема (68) принадлежит семейству схем с постоянными операторами A и B, удовлетворяющими условиям

$$A \geqslant 0, \tag{76}$$

$$B = B^* > 0, \tag{77}$$

то, обозначив

$$u_h = B^{\frac{1}{2}} v_h, \quad \tilde{S} = B^{\frac{1}{2}} A B^{-\frac{1}{2}}, \quad \psi_h = B^{-\frac{1}{2}} \varphi_h, \tag{78}$$

вновь придем к явной схеме (72).

Для исследования устойчивости двухслойных разностных схем можно воспользоваться оценкой нормы оператора перехода T соответствующей явной схемы или методом энергетических неравенств.

Необходимые и достаточные условия устойчивости двухслойных разностных схем могут быть сформулированы в виде неравенств между операторными коэффициентами разностных схем. Эти условия позволяют не только исследовать схемы на устойчивость, но и строить новые устойчивые схемы. Остановимся на некоторых из них.

а) *Необходимые и достаточные условия устойчивости по начальным данным.* Пусть схема

$$Bv_t^j + Av_h^j = 0,$$

$$v_h^0 = \gamma(x) \tag{79}$$

принадлежит семейству двухслойных схем с операторами A и B, удовлетворяющими условиям (69), (70). Тогда (79) можно записать в виде явной схемы

$$u_h^{j+1} = Tu_h^j, \tag{80}$$

где u_h^j и T определяются соответственно по формулам (71′), (73). Очевидно, если

$$\|T\| \leqslant \rho, \tag{81}$$

то

$$\|u_h^{j+1}\| = \|Tu_h^j\| \leqslant \rho \|u_h^j\|,$$

$$(A^{\frac{1}{2}} v_h^{j+1},\ A^{\frac{1}{2}} v_h^{j+1})^{\frac{1}{2}} \leqslant \rho (A^{\frac{1}{2}} v_h^j,\ A^{\frac{1}{2}} v_h^j)^{\frac{1}{2}},$$

или, принимая во внимание (71), получим:

$$\|v_h^{j+1}\|_{\mathbf{H}_A} \leqslant \rho \|v_h^j\|_{\mathbf{H}_A}. \tag{82}$$

Таким образом, если имеет место неравенство (81), то разностная схема (79), (69), (70) устойчива в $\mathbf{H}_A$ с постоянной $\rho > 0$.

Очевидно, если разностная схема (79) принадлежит семейству схем с постоянными операторами A и B, удовлетворяющими условиям (76), (77), и для оператора

$$T = I - \tau B^{-\frac{1}{2}} A B^{\frac{1}{2}}$$

выполняется условие (81), тогда, учитывая (78), получим:

$$\|v_h\|_{\mathbf{H}_B} \leqslant \rho \|v_h^j\|_{\mathbf{H}_B}, \tag{83}$$

т. е. разностная схема (79), (76), (77) будет устойчива в $\mathbf{H}_B$. Поэтому, если сформулировать условия, накладываемые на операторные коэффициенты разностной схемы (79), при которых будет выполняться неравенство (81), то тем самым будут указаны условия, при которых имеет место устойчивость разностной схемы (79), (69), (70) в $\mathbf{H}_A$ и схемы (79), (76), (77) в $\mathbf{H}_B$.

Теорема 6. *Пусть операторы A и B разностной схемы (79) не зависят от t и удовлетворяют условиям $A = A^* > 0$, $B > 0$. Тогда операторное неравенство*

$$B \geqslant \frac{\tau}{1+\rho} A \tag{84}$$

при $\rho \leqslant 1$ необходимо и при $\rho \geqslant 1$ достаточно для устойчивости разностной схемы (79) в $\mathbf{H}_A$.

Доказательство. Докажем необходимость условия (84) при $\rho \leqslant 1$. Необходимость условия (84) понимается в том смысле, что если будет иметь место (81), то для операторов A и B будет выполняться неравенство (84).

Из (81) следует, что для любого $u_h \in \mathbf{H}_h$

$$\|Tu_h\|^2 \leqslant \rho^2 \|u_h\|^2.$$

Учитывая (73), получим:

$$\|Tu_h\|^2 = \|u_h\|^2 - 2\tau (Su_h, u_h) + \tau^2 (Su_h, Su_h) \leqslant \rho^2 \|u_h\|^2,$$

$$(1-\rho^2)\|u_h\|^2 - 2\tau (Su_h, u_h) + \tau^2 \|Su_h\|^2 \leqslant 0. \tag{85}$$

Далее, так как $\rho \leqslant 1$ и

$$(Su_h, u_h) \leqslant \|Su_h\| \|u_h\|, \tag{86}$$

из (85) находим:

$$(1-\rho^2)\frac{(Su_h, u_h)^2}{\|Su_h\|^2} - 2\tau (Su_h, u_h) + \tau^2 (Su_h, Su_h) \leqslant 0$$

или

$$\frac{(Su_h, u_h)^2}{\|Su_h\|^2}\left[1-\rho^2 - 2\tau \frac{\|Su_h\|^2}{(Su_h, u_h)} + \tau^2 \frac{\|Su_h\|^4}{(Su_h, u_h)^2}\right] \leqslant 0. \tag{87}$$

Обозначим

$$\eta = \tau \frac{\|Su_h\|^2}{(Su_h, u_h)}.$$

Тогда из (87) имеем:

$$\eta^2 - 2\eta + 1 - \rho^2 \leqslant 0,$$

или

$$1-\rho \leqslant \eta \leqslant 1+\rho.$$

Следовательно,

$$\frac{1-\rho}{\tau} \leqslant \frac{(Su_h, Su_h)}{(Su_h, u_h)} \leqslant \frac{1+\rho}{\tau}. \tag{88}$$

Второе из неравенств (88) дает оценку

$$(Su_h, Su_h) \leqslant \frac{1+\rho}{\tau}(Su_h, u_h). \tag{89}$$

Обозначим

$$w = B^{-1}A^{\frac{1}{2}}u_h.$$

Из (89) получим:

$$\frac{\tau}{1+\rho}(A^{\frac{1}{2}}w, A^{\frac{1}{2}}w) \leqslant (A^{\frac{1}{2}}w, A^{-\frac{1}{2}}Bw)$$

или, учитывая (71), последнее неравенство можно записать в виде

$$(Bw, w) \geqslant \frac{\tau}{1+\rho}(Aw, w), \tag{90}$$

или

$$B \geqslant \frac{\tau}{1+\rho}A.$$

Используя первое из неравенств (88), при $\rho < 1$ можно получить следующее необходимое условие устойчивости разностной схемы (79):

$$(Bw, w) \leqslant \frac{\tau}{1-\rho}(Aw, w). \tag{91}$$

Докажем достаточность условий (84) для устойчивости разностной схемы (79) в $\mathbf{H}_A$ при $\rho \geqslant 1$.

Пусть выполняется неравенство (90). Тогда имеет место неравенство (89). Поэтому

$$\|Tu_h\|^2 = \|u_h\|^2 - 2\tau(Su_h, u_h) + \tau^2(Su_h, Su_h) \leqslant$$
$$\leqslant \|u_h\|^2 + \tau(\rho - 1)(Su_h, u_h).$$

Из (89) и (86) имеем:

$$\|Su_h\|^2 \leqslant \frac{1+\rho}{\tau}(Su_h, u_h) \leqslant \frac{1+\rho}{\tau}\|Su_h\|\|u_h\|,$$

$$(Su_h, u_h) \leqslant \frac{1+\rho}{\tau}\|u_h\|^2.$$

Значит,

$$\|Tu_h\|^2 \leqslant \|u_h\|^2 + (\rho^2 - 1)\|u_h\|^2 = \rho^2\|u_h\|^2,$$

т. е. схема (79) устойчива в $\mathbf{H}_A$.

Для решения задачи (79) будет справедлива оценка

$$\|v_h^{j+1}\|_{H_A} \leqslant \rho^{j+1}\|v_h^0\|_{\mathbf{H}_A}. \tag{92}$$

Следствие 1. Пусть A и B — постоянные операторы и $A = A^* > 0$, $B > 0$. Тогда условие

$$B \geqslant \frac{\tau}{2}A \tag{93}$$

необходимо и достаточно для устойчивости разностной схемы (79) в H_A с постоянной $\rho = 1$ (схема (79) будет абсолютно устойчивой в $\mathbf{H}_A$).

Норма оператора перехода для исходного семейства схем (79), (69), (70) тогда и только тогда меньше единицы, если выполнено условие (93).

Отметим, что условие устойчивости (69), (70), (93) для двухслойной разностной схемы, записанной в виде (54), принимает вид

$$B_0 > 0, \quad B_0 + B_1 = (B_0 + B_1)^* > 0, \quad B_0 - B_1 \geqslant 0.$$

Теорема 7. *Пусть операторы разностной схемы* (79) *не зависят от* t_j *и удовлетворяют условиям*

$$A = A^* > 0, \quad B = B^* > 0. \tag{94}$$

Тогда условие

$$B \geqslant \frac{\tau}{1+\rho}A \tag{95}$$

при $\rho \geqslant 1$ *необходимо и достаточно для устойчивости разностной схемы* (79) *в* $\mathbf{H}_A$ *с постоянной* $\rho \geqslant 1$.

Доказательство. Достаточность условия (95) при $\rho \geqslant 1$ и его необходимость при $\rho = 1$ для устойчивости разностной схемы (79) следует из теоремы 6. Докажем необходимость условия (95) при $\rho > 1$.

Пусть

$$\|T\| = \|I - \tau A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}\| \leqslant \rho. \tag{96}$$

Так как оператор $T = T^*$, то неравенство (96) эквивалентно следующим неравенствам:

$$
\begin{gathered}
-\rho I \leqslant T = I - \tau S \leqslant \rho I, \\
\frac{1-\rho}{\tau} I \leqslant S = A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} \leqslant \frac{1+\rho}{\tau} I, \\
\frac{1-\rho}{\tau} A^{-1} \leqslant B^{-1} \leqslant \frac{1+\rho}{\tau} A^{-1}.
\end{gathered}
\tag{97}
$$

Из неравенства

$$B^{-1} \leqslant \frac{1+\rho}{\tau} A^{-1}$$

в силу самосопряженности операторов A и B следует неравенство (95); неравенство

$$\frac{1-\rho}{\tau} A^{-1} \leqslant B^{-1},$$

так как $\rho > 1$, $\tau > 0$ и операторы A и B удовлетворяют условиям (94), тривиально. Теорема доказана.

Теорема 8. *Пусть операторы A и B разностной схемы* (79) *не зависят от t_j и удовлетворяют условиям*

$$A = A^* \geqslant 0, \quad B = B^* > 0.$$

Тогда неравенство

$$B \geqslant \frac{\tau}{1+\rho} A$$

при $\rho \geqslant 1$ необходимо и достаточно для устойчивости разностной схемы (79) *в* $\mathbf{H}_B$.

Доказательство этой теоремы проводится аналогично теореме 7, если учесть, что из (81) следует (83).

б) *Необходимые и достаточные условия устойчивости двухслойных разностных схем по правой части.* Рассмотрим двухслойную разностную схему вида (68) с однородными начальными условиями и постоянными операторами A и B.

$$
\begin{gathered}
B v_t^j + A v_h^j = \varphi_h^j, \quad t \leqslant t_j \leqslant T_0, \\
v_h^0 = 0.
\end{gathered}
\tag{98}
$$

Если оператор B^{-1} существует, то двухслойную неявную схему (68) можно представить в виде следующей явной схемы:

$$v_h^{j+1} = T v_h^j + \tau B^{-1} \varphi_h^j, \tag{99}$$

$$T = I - \tau B^{-1} A, \tag{99'}$$

откуда для (98), так как $v_h^0 = 0$, имеем:

$$v_h^{j+1} = \sum_{k=0}^{j} \tau T^{j-k} B^{-1} \varphi_h^k. \tag{100}$$

Пусть для нормы оператора перехода выполняется неравенство

$$\|T\| \leqslant \rho, \tag{101}$$

где $\rho = \text{const} > 0$, $\rho^j \leqslant \varkappa_1$ для всех $j\tau \leqslant T_0$, $\varkappa_1 = \text{const} > 0$, не зависящая от h, τ, j. Тогда из (100) для решения задачи (98) получим оценку

$$\|v_h^{j+1}\|_{(h,j+1)} \leqslant \sum_{k=0}^{j} \tau\rho^{j-k} \|B^{-1}\varphi_h^k\|_{(h,k)},$$

или

$$\|v_h^{j+1}\|_{(h,j+1)} \leqslant \varkappa_1 \sum_{k=0}^{j} \tau \|B^{-1}\varphi_h^k\|_{(h,k)}. \tag{102}$$

Оценка (102) означает устойчивость схемы (68) по правой части в смысле (60). Имеет место теорема.

Теорема 9. *Для устойчивости двухслойной разностной схемы* (98), *достаточно, чтобы для нормы оператора перехода выполнялось неравенство* (101), *при этом верна оценка* (102).

Если выполнено условие (101) для нормы оператора перехода, то двухслойная разностная схема (68) будет равномерно устойчива по начальным условиям.

Таким образом, существует связь между равномерной устойчивостью по начальным данным и устойчивостью по правой части.

Теорема 10. *В случае постоянного оператора перехода T от неявной двухслойной схемы* (68) *к явной двухслойной схеме* (72) *условие равномерной устойчивости по начальным данным является необходимым и достаточным для устойчивости вида* (102) *по правой части или, как принято говорить, из равномерной устойчивости по начальным данным следует устойчивость по правой части.*

Доказательство. *Необходимость.* Пусть имеет место оценка (102). Выберем φ_h^k так, чтобы

$$\tau B^{-1}\varphi_h^k = \beta\delta_{kl},$$

где

$$\delta_{kl} = \begin{cases} 0, & k \neq l, \\ 1, & k = l. \end{cases}$$

Тогда из (100) будем иметь:

$$v_h^{j+1} = \beta T^{j-l},$$

и так как имеет место оценка (102), то

$$\|v_h^{j+1}\|_{(h,j+1)} = \|\beta T^{j-l}\|_{(h,j+1)} \leqslant \varkappa_1 \|\beta\|_{(h,l)}.$$

В силу произвольности j, l из последнего неравенства имеем:

$$\|T^i\| \leqslant \varkappa_1, \quad i = \overline{1, j},$$

т. е. схема (68) равномерно устойчива по начальным данным.

Достаточность. Если для схемы (79) выполняется оценка

$$\|v_h^{j+1}\|_{(h,j+1)} \leqslant \varkappa_1 \|v_h^k\|_{(h,k)}, \quad k = \overline{0, j}; \quad j = 1, 2, \ldots, \tag{103}$$

то справедливо неравенство

$$\|T\| \leqslant \varkappa_1. \tag{104}$$

Из (100), учитывая (104), получим (102).

Для решения задачи (68), если выполняются условия (101), имеет место следующая оценка:

$$\|v_h^{j+1}\|_{(h,j+1} \leqslant \varkappa_1 \left(\|v_h^0\|_{(h,0)} + \sum_{k=0}^{j} \tau \|B^{-1}\varphi_h^k\|_{(h,k)}\right). \tag{105}$$

Если операторы A и B постоянны и

$$A = A^* > 0, \quad B = B^* > 0, \tag{106}$$

то схему (68) можно записать в виде

$$u_h^{j+1} = Tu_h^j + \tau B^{\frac{1}{2}} \varphi_h^j, \quad T = I - \tau B^{\frac{1}{2}} A B^{\frac{1}{2}}.$$

Тогда, если выполняется неравенство

$$\|T\| = \|I - \tau B^{\frac{1}{2}} A B^{\frac{1}{2}}\| \leqslant \rho,$$

($\rho = \text{const} > 0$, $\rho^j \leqslant \varkappa_1$ для всех $j\tau \leqslant T_0$), то

$$\|u_h^{j+1}\|_{(h,j+1)} \leqslant \varkappa_1 \left(\|u_{h(h,0)}^0 + \sum_{k=0}^{j} \tau \|B^{-\frac{1}{2}} \varphi_h^k\|_{(h,k)}\right) \tag{107}$$

или для решения задачи (68) будет иметь место оценка

$$\|v_h^{j+1}\|_{\mathbf{H}_B} \leqslant \varkappa_1 \left(\|v_h^0\|_{\mathbf{H}_B} + \sum_{k=0}^{j} \tau \|\varphi_h^k\|_{\mathbf{H}_{B^{-1}}}\right), \tag{108}$$

означающая устойчивость разностной схемы (68) в смысле (55).

Если разностная схема (68) принадлежит семейству схем с постоянными операторами A и B, удовлетворяющими условиям (69), (70), (84), то из теоремы 9 и соотношений (72), (75), (92) следует, что для решения задачи (68) справедлива оценка

$$\|v_h^{j+1}\|_{\mathbf{H}_A} \leqslant \rho^{j+1} \|v_h^0\|_{\mathbf{H}_A} + \sum_{k=0}^{j} \tau \rho^{j-k} \|A^{\frac{1}{2}} B^{-1} \varphi_h^k\|_{(h,k)} \tag{109}$$

или

$$\|v_h^{j+1}\|_{\mathbf{H}_A} \leqslant \varkappa_1 \left(\|v_h^0\|_{\mathbf{H}_A} + \sum_{k=0}^{j} \tau \|B^{-1} \varphi_h^k\|_{\mathbf{H}_A}\right), \tag{110}$$

т. е. двухслойная разностная схема (68) будет устойчивой.

Таким образом, имеет место следующая теорема:

Теорема 11. *Если* $A = A^* > 0$, $B = B^* > 0$ — *постоянные операторы и выполняется условие*

$$B \geqslant \frac{\tau}{1+\rho} A,$$

то для решения задачи (68) *имеют место априорные оценки* (108), (110).

Используя метод энергетических неравенств, метод выделения стационарных неоднородностей, можно получить оценки вида (108), (110), в которые входят более простые нормы вектора правой части.

В частности, значительно упрощается исследование сходимости разностных схем, если воспользоваться негативной нормой

$$\|\varphi\|_{A^{-1}} = (A^{-1}\varphi,\ \varphi)^{\frac{1}{2}}, \tag{111}$$

которая при $A = A^* > 0$ совпадает с нормой

$$\|\varphi\|_{A^{-1}} = \sup_{\|u\|_{H_A} \neq 0} \frac{(\varphi,\ u)}{\|u\|_{H_A}}. \tag{112}$$

Так, представляя решение задачи (68) в виде

$$v_h^{j+1} = z_h^{j+1} + x_h^{j+1},$$

где z_h^j удовлетворяет уравнению

$$Az_h^{j+1} = \varphi_h^j$$

и условиям

$$z(0) = z(\tau),$$

для x_h^{j+1} будем иметь:

$$Bx_t^j + Ax_h^j = f_h^j, \quad f_h^j = -(B - \tau A)\, z_t^j, \quad f_h^0 = 0,$$

$$x_h(0) = v_h(0) - z_h(0).$$

Тогда в соответствии с (109)

$$\|x_h^{j+1}\| \leqslant \rho^{j+1}\|x_h^0\|_{H_A} + \sum_{k=1}^{j} \tau\rho^{j-k}\|A^{\frac{1}{2}}B^{-1}f_h^{(k)}\|.$$

Заметим, что

$$A^{\frac{1}{2}}B^{-1}f_h^k = -A^{\frac{1}{2}}B^{-1}(B - \tau A)\, z_t^k = -(I - \tau A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}})\, A^{-\frac{1}{2}}\varphi_{\bar{t}}^k,$$

так как

$$z_t^k = A^{-1}\varphi_t^{k-1} = A^{-1}\varphi_{\bar{t}}^k.$$

Далее, если $B \geqslant \frac{\tau}{1+\rho} A$, то

$$\|I - \tau A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}\| = \|T\| < \rho.$$

Поэтому

$$\|A^{\frac{1}{2}}B^{-1}f_h^k\| \leqslant \rho\|A^{-\frac{1}{2}}\varphi_{\bar{t}}^k\| \leqslant \rho\|\varphi_{\bar{t}}^k\|_{A^{-1}}.$$

Так как

$$\|v_h(0)\|_{H_A} \leqslant \|x_h(0)\|_{H_A} + \|z_h(0)\|_{H_A} = \|v_h^0\|_{H_A} + \|\varphi_h^0\|_{A^{-1}},$$

то

$$\|v_h^{j+1}\| \leqslant \rho^{j+1}\|v_h^0\|_A + \rho^{j+1}\|\varphi_h^0\|_{A^{-1}} +$$

$$+ \sum_{k=1}^{j} \tau\rho^{j+1-k}\|\varphi_{\bar{t}}^k\|_{A^{-1}} + \|\varphi_h^j\|_{A^{-1}}. \tag{113}$$

Оценки, аналогичные (113), могут быть получены и в случае переменных операторов $B = B(t) > 0$, $A = A(t) = A^*(t) > 0$, если оператор $A(t) > 0$ удовлетворяет условию Липшица по переменной t и $B(t) \geqslant \frac{\tau}{2} A(t)$.

Аналогично можно показать, что для трехслойных разностных схем

$$B_0 v^{j+1} + B_1 v^j + B_2 v^{j-1} = F^j \quad (j = 1, 2, \ldots), \tag{114}$$

(v^0, v^1 — заданы, v^0, $v^1 \in \mathbf{H}_h$), записанных в каноническом виде

$$B v_{\overset{\circ}{t}} + \tau^2 R v_{\bar{t}t} + A v = \varphi, \tag{115}$$

достаточным условием устойчивости в энергетической норме вида

$$\| v^j \|_э = \left(\frac{1}{4} \| v^{j+1} + v^j \|_A^2 + \| v^{j+1} - v^j \|^2_{R - \frac{1}{4} A} \right)^{\frac{1}{2}} \tag{116}$$

является выполнение неравенств (см. [4], [5]):

$$A > 0, \quad R \geqslant \frac{1}{4} A. \tag{117}$$

Примеры исследования устойчивости двухслойных разностных схем.

Пример 1. Устойчивость разностной схемы с весами для уравнения теплопроводности.

Рассмотрим разностную схему (181), § 3. Схема имеет канонический вид (56) с $B = I - \sigma\tau\Lambda_1$, $A = -\Lambda_1$. Так как $-\Lambda_1 v_h = -v_{\bar{x}x}$ — самосопряженный положительно определенный оператор (см. (129), § 3), то на основании теоремы 11 схема будет устойчива в $\mathbf{H}_A$, если выполнено условие (84). Пусть $\rho = 1$, тогда (84) запишется в виде

$$B - \frac{\tau}{2} A = I + \left(\sigma - \frac{1}{2} \right) \tau A \geqslant 0.$$

Замечая, что

$$0 < A \leqslant \| A \| I, \tag{118}$$

получим:

$$I + \left(\sigma - \frac{1}{2} \right) \tau A \geqslant \left[\frac{1}{\| A \|} + \left(\sigma - \frac{1}{2} \right) \tau \right] A,$$

поэтому при

$$\sigma \geqslant \frac{1}{2} - \frac{1}{\tau \| A \|} \tag{119}$$

разностная схема (181), § 3, будет устойчивой. Если учесть, что собственные значения оператора $A v_h = -v_{\bar{x}x}$ $v(0) = v(1) = 0$ находятся по формулам

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{\pi k h}{2} \quad (k = \overline{1, n}) \tag{120}$$

и

$$\| A \| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} < \frac{4}{h^2},$$

то разностная схема (181), § 3, будет устойчивой при

$$\sigma \geqslant \frac{1}{2} - \frac{h^2}{4\tau}. \tag{121}$$

Разностные схемы (175), (178), § 3, являются частными случаями разностной схемы (181), § 3, соответственно при $\sigma = 0$ и $\sigma = 1$. Поэтому разностная схема (175),

§ 3, будет устойчивой при $\frac{\tau}{h^2} \leqslant \frac{1}{2}$ (условно устойчива), а разностная схема (178), § 3, будет абсолютно устойчивой.

Пример 2. Рассмотрим разностную схему (202), § 3. Введем в рассмотрение операторы

$$Dv = v_{\bar{x}}, \quad Av = -\Lambda_1 v = -v_{\bar{x}x}.$$

Тогда схему (202), § 3, можно представить в виде

$$\left(I + \frac{\alpha\tau}{h} D\right) v_t + Av = 0,$$

или

$$Bv_t + Av = 0,$$

где

$$B = I + \frac{\alpha\tau}{h} D.$$

Заметим, что $v_{\bar{x}} - v_x = -hv_{\bar{x}x}$. В силу формул суммирования по частям $(v_{\bar{x}}, v) = -(v_x, v)$. Поэтому

$$(v_{\bar{x}}, v) = -(v_x, v) = -\frac{h}{2}(v_{\bar{x}x}, v) > 0, \tag{122}$$

так как оператор $Av = -v_{\bar{x}x}$ — положительно определенный. Следовательно, операторы

$$B = I + \frac{\alpha\tau}{h} D > 0, \quad (\alpha > 0), \quad A = -\Lambda_1 = -\Lambda_1^* > 0$$

будут удовлетворять условиям теоремы 6 с $\rho = 1$, если

$$I + \frac{\alpha\tau}{h} D - \frac{\tau}{2} A \geqslant 0.$$

Учитывая (122), имеем:

$$I + \frac{\alpha\tau}{h} D - \frac{\tau}{2} A \geqslant I + \frac{\tau}{2}(\alpha - 1) A. \tag{123}$$

Последнее неравенство можно усилить, если воспользоваться (118)

$$I + \frac{\alpha\tau}{h} D - \frac{\tau}{2} A \geqslant \left(\frac{1}{\|A\|} + \frac{1}{2}\tau(\alpha - 1)\right) A, \tag{124}$$

откуда разностная схема (202), § 3, будет устойчивой в $\mathbf{H}_A$ при

$$\alpha \geqslant 1 - \frac{2}{\tau\|A\|}, \tag{125}$$

или

$$\alpha \geqslant 1 - \frac{h^2}{2\tau}. \tag{126}$$

При $\alpha \geqslant 1$ разностная схема (202), § 3, будет безусловно устойчивой в $\mathbf{H}_A$.

Принцип регуляризации. Достаточные условия устойчивости разностных схем, записанные в виде операторных неравенств, могут быть использованы для построения новых разностных схем, обладающих нужными качествами.

Пусть разностная схема (68), принадлежащая семейству двухслойных разностных схем вида (69), (70), является устойчивой, т. е.

$$B \geqslant \frac{\tau}{2} A.$$

Очевидно, если построить двухслойную разностную схему, принадлежащую этому же семейству схем с оператором $\tilde{B} \geqslant B$, то она также будет устойчивой. Пусть оператор B представлен в виде

$$B = I + \tau R, \tag{127}$$

где $R = R^* > 0$, то, выбирая

$$R \geqslant \left(\frac{1}{2} - \frac{1}{\tau \| A \|}\right) A,$$

получим устойчивую разностную схему, так как

$$B = I + \tau R \geqslant \frac{\tau}{2} A.$$

Оператор R называют *регуляризатором*. Очевидно, если выбрать $\tilde{R} = \tilde{R}^* \geqslant R$, то получим устойчивую разностную схему вида

$$(I + \tau \tilde{R}) v_t + A v_h = \varphi_h, \quad v_h(0) \text{ — задано.}$$

Выбор оператора регуляризации осуществляется таким образом, чтобы построенная разностная схема обладала лучшими в каком-то смысле качествами по сравнению с исходной схемой, например, имела бы аппроксимацию заданного порядка, была устойчивой и экономичной. Последнее условие обычно достигается за счет выбора в качестве оператора B факторизованного оператора, например,

$$B = (I + \tau R_1)(I + \tau R_2), \tag{128}$$

где

$$R_1 = R_2^*, \quad R_1 + R_2 = R. \tag{129}$$

При таком способе факторизации операторы $B_1 = I + \tau R_1$, $B_2 = I + \tau R_2$ имеют треугольные матрицы и являются экономичными.

Способ факторизации разностной схемы (217), § 3, можно трактовать как прием регуляризации, при котором по оператору

$$D = I - \tau \sum_{k=1}^{q} \Lambda_k = I + \tau R$$

строится факторизованный оператор

$$B = \prod_{k=1}^{q} (I - \tau \Lambda_k).$$

Оператор B можно представить в виде

$$B = I + \tau (R + \tau \psi) = I + \tau \tilde{R}, \tag{130}$$

$$\tilde{R} = R + \tau \psi,$$

где ψ определяется по формуле (219), § 3, при этом приходим к экономичной разностной схеме с тем же порядком аппроксимации.

Приведем еще один пример, в котором за счет выбора регуляризирующего оператора удается построить разностную схему с конечным числом узлов для неограниченной области с сохранением порядка аппроксимации.

Пусть требуется найти решение следующей задачи:

$$u'' + q(x)\,u = f(x), \quad x \in (-\infty, \infty). \tag{131}$$

Предположим, что функции $q(x)$ и $f(x)$ таковы, что уравнение (131) имеет решение, обладающее следующим свойством:

$$|xu'| < c, \quad \forall\, x \in (-\infty, \infty). \tag{132}$$

Поставим в соответствие уравнению (131) следующее конечно-разностное уравнение вида:

$$\Lambda v = v_{\bar{x}x} + q(x)\,v_h = f_h(x), \quad x \in \Omega_h, \tag{133}$$

где

$$\Omega_h = \{x_i = ih; \quad i = 0, \pm 1, \pm 2, \ldots\}.$$

Разностная схема (133) имеет второй порядок аппроксимации и представляет собой бесконечную систему линейных алгебраических уравнений.

Введем в схему (133) регуляризатор R:

$$\begin{gathered} \Lambda v + Rv = f_h(x), \quad x \in \tilde{\Omega}_h, \\ u(x_{-n-1}) \neq \infty, \quad u(x_{n+1}) \neq \infty, \end{gathered} \tag{134}$$

где

$$\Omega_h = \{x_i = ih: \quad i = 0, \pm 1, \ldots, \pm n\}, \tag{135}$$

R — разностный оператор не выше второго порядка и $Ru = O(h^2)$, а в остальном пока оператор R произвольный. Выберем регуляризатор R таким образом, чтобы коэффициенты при $u(x_{-n-1})$ и $u(x_{n+1})$ в (134) обращались в нуль. Если оператор R с указанными выше свойствами существует, то схема (134), (135), представляющая собой конечную систему линейных алгебраических уравнений, аппроксимирует исходную задачу (131), (132) со вторым порядком. При этом в граничных точках области $\tilde{\Omega}_h$ краевые условия (135) выполняются точно и порядок аппроксимации не нарушается. Остающийся выбор регуляризатора R можно использовать в том плане, чтобы разностная схема (134), (135) имела заранее известные собственные значения.

В нашем случае в качестве регуляризатора R можно взять оператор вида

$$Rv = -2xh^2 v_{\overset{\circ}{x}} \left(h = n^{-\frac{1}{4}}\right).$$

Тогда конечно-разностный оператор $\Lambda + R$ будет иметь собственные значения, определяющиеся формулой

$$\lambda_i^h = \frac{2i}{\sqrt{n}}, \quad i = \overline{0,2n}.$$

Изложенный подход решения краевых задач для бесконечных областей имеет очевидное преимущество перед обычным методом редукции (методом решения бесконечных систем линейных алгебраических уравнений), ибо допускает только одну погрешность — погрешность аппроксимации и не содержит погрешности за счет усечения области.

3. Аддитивные схемы

Рассмотрим составные разностные схемы q-го ранга вида (47), (48).

В частности, двухслойная разностная схема q-го ранга может быть записана в виде следующей канонической системы:

$$B\frac{v^{j+\frac{k}{q}}-v^{j+\frac{k-1}{q}}}{\tau}+\sum_{\alpha=1}^{q}D_{k\alpha}\cdot v^{j+\frac{\alpha}{q}}=\varphi_k^j,\quad k=1,\overline{q},\qquad (136)$$

$$j=0,1,2,\ldots,$$

v^0 — задано. Здесь B, $D_{k\alpha}$ — линейные операторы, $B\,D_{k\alpha}$: $\mathbf{H}_h\to\mathbf{H}_h$, $v^{j+\frac{\alpha}{q}}=\omega^\alpha$ — промежуточные значения функции.

Аддитивной схемой называется такая составная разностная схема, погрешность аппроксимации которой определяется величиной

$$\Psi=\sum_{k=1}^{q}\psi_k,\qquad (137)$$

где ψ_k — погрешность аппроксимации отдельного уравнения системы с номером k на решении u исходного уравнения.

Аддитивная схема обладает суммарной аппроксимацией, если

$$\|\Psi\|\to 0 \text{ при } |h|\to 0,\quad \tau\to 0.\qquad (138)$$

Впервые понятие суммарной аппроксимации для составной разностной схемы было введено А. А. Самарским.

Принцип аддитивности разностной схемы позволил обосновать известные методы (в частности, методы переменных направлений) и провести исследования новых разностных методов для более широкого круга задач. Следует, однако, отметить, что при построении аддитивных разностных схем наблюдается понижение порядка точности разностной схемы.

Пример (построение аддитивной разностной схемы). Рассмотрим следующую задачу:

$$\frac{\partial u}{\partial t}=Lu+f(x,t),\quad 0\leqslant t\leqslant T,\quad Lu=\sum_{k=1}^{q}L_k u\qquad (139)$$

$$u(x,0)=\gamma_0(x),$$

$$u|_\Gamma=\gamma_1(x,t).$$

Уравнение (139) можно записать в виде

$$\sum_{k=1}^{q}W_k(u)=0,$$

где

$$W_k(u) = \frac{1}{q}\frac{\partial u}{\partial t} - L_k u - f_k(x, t)$$

и

$$\sum_{k=1}^{q} f_k(x, t) = f(x, t), \quad f(x, t), \ f_k(x, t) \in C_D^N(\Omega).$$

На отрезке изменения переменной t $(0 \leqslant t \leqslant T_0)$ нанесем две системы узлов: основную $t_j = j\tau$, $j = \overline{0, m}$, $\tau = \frac{T_0}{m}$ и вспомогательную

$$t_{j+\frac{l}{q}} = t_j + \frac{l\tau}{q}, \quad l = \overline{1, q}, \tag{140}$$

т. е. внутри каждого из полуинтервалов $(t_j, t_{j+1}]$ находится $q - 1$ точка вспомогательной системы узлов (140).

Уравнение (139) на каждом из отрезков $[t_j, t_{j+1}]$ заменяется системой дифференциальных уравнений

$$W_l(u) = 0, \quad l = \overline{1, q} \tag{141}$$

таким образом, чтобы на l-м вспомогательном полуинтервале $(t_{j+\frac{l-1}{q}}, t_{j+\frac{l}{q}}]$ выполнялось l-е уравнение системы (141), причем

$$\begin{gathered} u_{l-1}(x, t_{j+\frac{l-1}{q}}) = u_l(x, t_{j+\frac{l-1}{q}}), \quad l = \overline{2, q}, \quad j = 0, 1, \ldots \\ u_q(x, t_j) = u_1(x, t_j), \quad j = 1, 2, \ldots \\ u_1(x, 0) = \gamma_0(x), \quad j = 0. \end{gathered} \tag{142}$$

При $t^* \in [t_j, t_{j+1}]$

$$W_l(u(x, t^*)) = W_l(u(x, t_{j+\frac{1}{2}})) + O(\tau), \quad l = \overline{1, q}, \ j = 0, 1, \ldots \tag{143}$$

Погрешность аппроксимации Ψ уравнения (139) системой (141) будет равна:

$$\Psi = \sum_{l=1}^{q} \psi_l(t), \quad t \in [t_j, t_{j+1}],$$

где $\psi_l(t)$ — погрешность аппроксимации для уравнения $W_l(u(x, t)) = 0$ номера l на решениях уравнения (139). Очевидно,

$$\psi_l(t) = W_l(u(x, t_{j+\frac{1}{2}})) + O(\tau).$$

Поэтому

$$\Psi = \sum_{l=1}^{q} \psi_l(t) = \sum_{l=1}^{q} [W_l(u(x, t_{j+\frac{1}{2}}) + O(\tau)] = O(\tau).$$

Таким образом, аддитивная система дифференциальных уравнений (141) аппроксимирует уравнение (139) в суммарном смысле с первым порядком по τ.

Система (141) на каждом из вспомогательных полуинтервалов

$$\left(t_{j+\frac{l-1}{q}},\ t_{j+\frac{l}{q}}\right]$$

заменяется разностной схемой, аппроксимирующей уравнение (141) с номером l, т. е. строится разностная схема

$$A_{hl}(v^j_{hl}) = 0, \quad l = \overline{1, q}, \quad j = \overline{0, m}. \tag{144}$$

Покажем, что аддитивная разностная схема (144) будет обладать суммарной аппроксимацией, если каждое l-е уравнение системы (141) аппроксимируется l-м разностным уравнением схемы (144) в обычном смысле, т. е. если

$$\|\tilde{\psi}^j_{hl}\| = \|-(W_l(u^j))_{hl} + A_{hl}(u^j)_{hl}\| \to 0, \quad |h| \to 0, \quad \tau \to 0. \tag{145}$$

В самом деле, пусть Ψ^j_{hl} — погрешность, с которой аппроксимирует уравнение номера l исходное уравнение (139) на его решении,

$$\Psi^j_{hl} = A_{hl}(u^j)_{hl}. \tag{146}$$

Здесь под $u^j = u(x, t^*)$ понимается решение уравнения (139) при $t^* \in [t_j, t_{j+1}]$. Очевидно,

$$\Psi^j_{hl} = \tilde{\psi}^j_{hl} + (W_l(u^j))_{hl}. \tag{147}$$

Если учесть соотношение (143), то

$$\Psi^j_{hl} = \tilde{\psi}^j_{hl} + (W_l(u^{j+\frac{1}{2}}))_{hl} + O(\tau),$$

или

$$\Psi^j_{hl} = \Psi^{*j}_{hl} + (W_l(u^{j+\frac{1}{2}}))_{hl},$$

где

$$\|\Psi^*_{hl}\| \to 0, \quad \text{при} \quad |h| \to 0, \quad \tau \to 0.$$

Следовательно,

$$\Psi^j_h = \sum_{l=1}^{q} \Psi^j_{hl} = \sum_{l=1}^{q} [\Psi^{*j}_{hl} + (W_l(u^{j+\frac{1}{2}}))_{hl}] = \sum_{l=1}^{q} \Psi^{*j}_{hl},$$

так как

$$\sum_{l=1}^{q} W_l(u(x, t^*)) = 0, \quad t^* \in [t_j, t_{j+1}].$$

Поэтому

$$\|\Psi^j_h\| \to 0 \quad \text{при} \quad |h| \to 0, \quad \tau \to 0, \quad j = 0, 1, \ldots,$$

т. е. аддитивная схема (144) аппроксимирует уравнение (139) в суммарном смысле, хотя каждое из уравнений (144) номера l может не аппроксимировать уравнение (139).

Для аддитивных разностных схем понятие устойчивости по правой части должно быть введено таким образом, что из условий суммарной аппроксимации следовало стремление к нулю решения разностной задачи (136) с нулевыми начальными условиями.

Достаточные условия устойчивости аддитивных схем.

Теорема 12. *Если постоянный оператор B и матрица-оператор $D=(D_{\alpha k})_{\alpha=\overline{1,q}}^{k=\overline{1,q}}$ разностной схемы* (136) *удовлетворяют условиям*

$$B=B^*>0, \tag{148}$$

$$\sum_{\alpha,k=1}^{q}(D_{k\alpha}\eta_\alpha,\ \eta_k)\geqslant 0 \quad \forall\ \eta_\alpha\in \mathbf{H}_h, \tag{149}$$

то для решения задачи (136) *справедлива оценка*

$$\|v^{j+1}\|_B^2\leqslant \mathrm{e}\left\{\|v^0\|_B^2+\sum_{m=1}^{j}\tau t_{j+1}\left\|\sum_{k=1}^{q}\varphi_k^m\right\|_{B^{-1}}^2+\sum_{m=1}^{j}q\,\frac{\tau^2}{2}\sum_{k=1}^{q}\|\varphi_k^m|_{B^{-1}}^2\right\}. \tag{150}$$

Из соотношения (150) следует, что если $v^0=0$, $\|\varphi\|=\left\|\sum\limits_{k=1}^{q}\varphi_k^m\right\|_{B^{-1}}\to$ $\to 0$ при $|h|\to 0$, $\tau\to 0$ и $\sum\limits_{k=1}^{q}\|\varphi_k^m\|_{B^{-1}}=O(1)$ — ограничена, то

$$\|v^{j+1}\|_B\to 0,\quad j=1,\ 2,\ \ldots.$$

Значит, из суммарной аппроксимации и устойчивости аддитивной схемы будет следовать ее сходимость.

Доказательство. Обозначим

$$v^{j+\frac{k}{q}}=\omega^k,\quad \varphi_k^j=\Phi_k. \tag{151}$$

Тогда разностную схему (136) можно записать в виде

$$B\omega_{\bar t}^k+\sum_{\alpha=1}^{q}D_{k\alpha}\omega^\alpha=\Phi_k,\quad k=\overline{1,q}. \tag{152}$$

Построим энергетическое тождество. Для этого умножим скалярно (152) на $\tau\omega^k$

$$\tau(B\omega_{\bar t}^k,\ \omega^k)+\tau\sum_{\alpha=1}^{q}D_{k\alpha}(\omega^\alpha,\ \omega^k)=\tau(\Phi_k,\ \omega^k). \tag{153}$$

Очевидны следующие тождества:

$$\tau(B\omega_{\bar t}^k,\ \omega^k)=\tau\left(B\omega_{\bar t}^k,\ \frac{\tau}{2}\omega_{\bar t}^k+\frac{\omega^k+\omega^{k-1}}{2}\right)=$$

$$=\frac{\tau^2}{2}(B\omega_{\bar t}^k,\ \omega_{\bar t}^k)+\frac{\tau}{2}(B\omega_{\bar t}^k,\ \omega^k+\omega^{k-1})=$$

$$=\frac{1}{2}[\tau^2(B\omega_{\bar t}^k,\ \omega_{\bar t}^k)+(B\omega^k,\ \omega^k)-(B\omega^{k-1},\ \omega^{k-1})], \tag{154}$$

$$\tau(\Phi_k,\ \omega^k)=\tau\left(\Phi_k,\ \omega^0+\sum_{i=1}^{k}(\omega^i-\omega^{i-1})\right)=$$

$$=\tau(\Phi_k,\ \omega^0)+\tau^2\sum_{i=1}^{k}(\Phi_k,\ \omega_{\bar t}^i). \tag{155}$$

Подставив (154) и (155) в (153) и просуммировав по всем $k = \overline{1, q}$, получим основное энергетическое тождество

$$\sum_{k=1}^{q} \frac{\tau^2}{2} (B\omega_{\bar{t}}^k, \omega_{\bar{t}}^k) + (B\omega^q, \omega^q) - (B\omega^0, \omega^0) +$$

$$+ \tau \sum_{k=1}^{q} \sum_{\alpha=1}^{q} (D_{k\alpha}\omega^\alpha, \omega^k) = \tau \sum_{k=1}^{q} (\Phi_k, \omega^0) + \tau^2 \sum_{k=1}^{q} \sum_{i=1}^{k} (\Phi_k \omega_{\bar{t}}^i). \qquad (156)$$

Оценим правую часть тождества (156). Для этого заметим, что

$$\tau^2 \sum_{k=1}^{q} \sum_{i=1}^{k} (\Phi_k, \omega_{\bar{t}}^i) = \tau^2 \sum_{k=1}^{q} \left(\omega_{\bar{t}}^k, \sum_{i=k}^{q} \Phi_i\right) = \tau^2 \sum_{k=1}^{q} \left(\omega_{\bar{t}}^k, BB^{-1} \sum_{i=k}^{q} \Phi_i\right) =$$

$$= \tau^2 \sum_{k=1}^{q} \left(B\omega_{\bar{t}}^k, B^{-1} \sum_{i=k}^{q} \Phi_i\right) \leqslant \frac{\tau^2}{2} \sum_{k=1}^{q} \left(\| \omega_{\bar{t}}^k \|_B^2 + \left\| \sum_{i=k}^{q} \Phi_i \right\|_{B^{-1}}^2\right). \qquad (157)$$

Учитывая обобщенное неравенство Коши — Буняковского и ε-неравенство

$$ab \leqslant \frac{\varepsilon}{2} a^2 + \frac{1}{2\varepsilon} b^2, \ \varepsilon > 0,$$

получим оценку для первого слагаемого правой части (156)

$$\tau \sum_{k=1}^{q} (\Phi_k, \omega^0) = \tau \left(B\omega^0, B^{-1} \sum_{k=1}^{q} \Phi_k\right) \leqslant$$

$$\leqslant \frac{\tau\varepsilon}{2} \| \omega^0 \|_B^2 + \frac{\tau}{2\varepsilon} \left\| \sum_{k=1}^{q} \Phi_k \right\|_{B^{-1}}^2. \qquad (158)$$

Подставляя (157) и (158) в (156) и учитывая (152), получим:

$$\frac{\tau^2}{2} \sum_{k=1}^{q} \| \omega_{\bar{t}}^k \|_B + \| \omega^q \|_B^2 \leqslant \| \omega^0 \|_B^2 + \frac{\tau\varepsilon}{2} \| \omega^0 \|_B^2 +$$

$$+ \frac{\tau}{2\varepsilon} \left\| \sum_{k=1}^{q} \Phi_k \right\|_{B^{-1}}^2 + \frac{\tau^2}{2} \sum_{k=1}^{q} \left(\| \omega_{\bar{t}}^k \|_B^2 + \frac{\tau^2}{2} \left\| \sum_{i=k}^{q} \Phi_i \right\|_{B^{-1}}^2 \right),$$

или, учитывая введенные обозначения (151), получим:

$$\| v^{j+1} \|_B^2 \leqslant \left(1 + \frac{\tau\varepsilon}{2}\right) \| v^j \|_B^2 + \frac{\tau}{2} \left[q\tau \left\| \sum_{i=k}^{q} \varphi_i^j \right\|_{B^{-1}}^2 + \frac{1}{\varepsilon} \left\| \sum_{k=1}^{q} \varphi_k^j \right\|_{B^{-1}}^2 \right]. \qquad (159)$$

Обозначим

$$q_{j+1} = \| v^{j+1} \|_B^2 > 0, \ \psi_j = \frac{\tau}{2} \left[q\tau \left\| \sum_{i=k}^{q} \varphi_i^j \right\|_{B^{-1}}^2 + \frac{1}{\varepsilon} \left\| \sum_{k=1}^{q} \varphi_k^j \right\|_{B^{-1}}^2 \right] > 0,$$

$$\rho = 1 + \frac{\tau\varepsilon}{2} > 0,$$

тогда (159) можно записать в виде

$$q_{j+1} = \rho q_j + \psi_j = \rho^{j+1} q_0 + \sum_{l=1}^{j} \rho^{j+1-l} \psi_l \leqslant \rho^{j+1} \left(q_0 + \sum_{l=1}^{j} \psi_l\right),$$

откуда

$$\|v^{j+1}\|_B^2 \leqslant \left(1+\frac{\tau\varepsilon}{2}\right)^{j+1}\left(\|v^0\|_B^2+\frac{\tau}{2}\sum_{l=1}^{j} q\tau\left\|\sum_{i=k}^{q}\varphi_i^l\right\|_{B^{-1}}^2+\right.$$

$$\left.+\frac{\tau}{2\varepsilon}\sum_{l=1}^{j}\left\|\sum_{k=1}^{q}\varphi_k^l\right\|_{B^{-1}}^2\right.$$

и при $\varepsilon=\frac{2}{t_{j+1}}$, если учесть, что $\left(1+\frac{\tau\varepsilon}{2}\right)^{j+1}\leqslant e^{\frac{\tau\varepsilon(j+1)}{2}}$, получим:

$$\|v^{j+1}\|_B^2 \leqslant e\left[\|v^0\|_B^2+\sum_{l=1}^{j}\frac{q\tau^2}{2}\sum_{k=1}^{q}\|\varphi_k^l\|^2+\sum_{j=1}^{j}\tau t_{j+1}\left\|\sum_{k=1}^{q}\varphi_k^l\right\|_{B^{-1}}^2\right]. \quad (160)$$

4. Сеточный метод Фурье и его применение для исследования устойчивости разностных схем

Пусть E_h — совокупность всех вещественных сеточных функций v_h, определенных в точках сетки

$$\Omega_h=\left\{x_j=jh;\quad j=\overline{0,n},\ h=\frac{a}{n}\right\} \quad (161)$$

и удовлетворяющих условиям

$$v_h(0)=v_h(a)=0. \quad (162)$$

В пространстве сеточных функций с обычными операциями сложения и умножения на вещественные числа при помощи равенства

$$(v_h,u_h)=\sum_{j=1}^{n-1}hv_h(jh)\,u_h(jh) \quad (163)$$

введем скалярное произведение. Тогда пространство сеточных функций E_h будет представлять собой эвклидово вещественное пространство R размерности $n-1$. Если система функций $\omega_k(jh)$ образует ортонормированный базис в R_{n-1}, то любая функция $v_h\in R_{n-1}$ может быть разложена по этому базису:

$$v_h(jh)=\sum_{k=1}^{n-1}c_k\omega_k(jh),\quad j=\overline{1,n-1}. \quad (164)$$

Так как $\omega_k(jh)$ система ортогональных в R_{n-1} функций, т. е.

$$(\omega_k(jh),\ \omega_l(jh))=\sum_{k=1}^{n-1}h\omega_k(jh)\,\omega_l(jh)=\delta_{kl}, \quad (165)$$

где δ_{kl}—символ Кронекера, то c_k в (164) будут определяться по формуле

$$c_k=(v_h(jh),\ \omega_k(jh))=\sum_{j=1}^{n-1}hv_h(jh)\,\omega_k(jh). \quad (166)$$

Последовательность c_k можно рассматривать как спектральный образ функции $v_h(jh)$ $(j=\overline{1,n-1})$. Если c_k и b_k — спектральные

образы v_h, u_h, то

$$(v_h, u_h) = \sum_{k=1}^{n-1} hv_h(jh)\, u_h(jh) = \sum_{k=1}^{n-1} c_k b_k,$$

или

$$(v_h, v_h) = \sum_{k=1}^{n-1} c_k^2 \tag{167}$$

— разностный аналог равенства Парсеваля.

Примером ортонормированного базиса в пространстве R_{n-1} является система функций

$$\omega_k(x) = \sqrt{\frac{2}{a}} \sin \frac{k\pi x}{a}, \quad x = jh. \tag{168}$$

Система (168) удовлетворяет условиям (162) и (165). Поэтому произвольная функция v_h, заданная на сетке (161) и удовлетворяющая условиям (162), представима в виде «суммы Фурье»

$$v_h(jh) = \sqrt{\frac{2}{a}} \sum_{k=1}^{n-1} c_k \sin \frac{k\pi jh}{a}, \tag{169}$$

где

$$c_k = \sqrt{\frac{2}{a}} \sum_{j=1}^{n-1} hv_h(jh) \sin \frac{k\pi j}{n}, \quad k = \overline{1, n-1}. \tag{170}$$

Если каждую из функций v_h доопределить во все точки jh ($j = 0, \pm 1, \pm 2, \ldots$) оси x, то получим a-периодическую сеточную функцию, удовлетворяющую условиям (162). В этом случае конечная «сумма Фурье» является аналогом бесконечного ряда Фурье для нечетной функции v_h.

Отметим, что система (168) является системой собственных функций разностной задачи (приложение, § 3):

$$\begin{aligned} v_{x\bar{x}} &= -\lambda v, \\ v(0) &= v(a) = 0, \end{aligned} \tag{171}$$

собственные значения λ_k которой определяются по формулам:

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2a}, \quad k = \overline{1, n-1}. \tag{172}$$

Аналогично на сетке (161) можно рассмотреть совокупность всех сеточных комплекснозначных функций v_h и каждую из них доопределить во все точки оси x с координатами $x_j = jh$, $j = 0, \pm 1, \pm 2, \ldots$ так, чтобы получилась a-периодическая сеточная функция. Совокупность всех таких функций обозначим M_h и введем в M_h скалярное произведение

$$(v_h, u_h) = \sum_{k=0}^{n-1} hv_h(jh)\, \bar{u}_h(jh). \tag{173}$$

Примером системы линейно-независимых a-периодических ортогональных в M_h функций являются функции:

$$\hat{\omega}_k = e^{ik\frac{2\pi}{a}x}, \quad i = \sqrt{-1}; \quad x = jh, \; k = 0, \pm 1, \pm 2, \ldots, \frac{n-1}{2}. \tag{174}$$

В самом деле,

$$(\hat{\omega}_k(x), \hat{\omega}_l(x)) = \sum_{j=0}^{n-1} h e^{\frac{i2\pi}{a}(k-l)jh} = nh\delta_{kl} = a\delta_{kl}. \tag{175}$$

Следовательно, любую функцию из M_h можно разложить в «сумму Фурье»

$$v_h(x) = \sum_{k=-\frac{n-1}{2}}^{k=\frac{n-1}{2}} A_k \hat{\omega}_k(x), \tag{176}$$

где

$$A_k = \frac{1}{a}(v_h, \hat{\omega}_k). \tag{177}$$

Функции $\hat{\omega}_k(x) = e^{ik\frac{2\pi}{a}x}\left(k = 0, \pm 1, \pm 2, \ldots, \pm \frac{n-1}{2}\right)$ являются собственными функциями оператора

$$\frac{i}{2}(\hat{\omega}_{\bar{x}} + \hat{\omega}_x) = \lambda\hat{\omega},$$

собственные значения которого

$$\lambda_k = \frac{n}{a}\sin\frac{2\pi k}{n}.$$

Функции $\hat{\omega}_k(x)$, $k = 0, \pm 1, \ldots, \pm \frac{n-1}{2}$, являются также собственными функциями оператора $(\cdot)_{\bar{x}}$ и $(\cdot)_x$ на M_h, причем

$$(\hat{\omega}_k(x))_x = \frac{e^{ik\frac{2\pi}{a}(j+1)h} - e^{ik\frac{2\pi}{a}jh}}{h} = e^{ik\frac{2\pi}{a}x}\frac{(e^{ik\frac{2\pi}{a}h} - 1)}{h} = \hat{\omega}_k(x)\tilde{\lambda}_k(h),$$

$$(\hat{\omega}_k(x))_{\bar{x}} = -\tilde{\lambda}_k(h)\hat{\omega}_k(x), \tag{178}$$

где

$$\tilde{\lambda}_k(h) = \frac{e^{ik\frac{2\pi}{a}h} - 1}{h}.$$

Поэтому $\hat{\omega}_k(x) = e^{ik\frac{2\pi}{a}x}\left(k = 0, \pm 1, \ldots, \pm \frac{n-1}{2}\right)$ образуют полную систему собственных функций для любого оператора A_h в виде

$$A_h v_h = \sum_{sq} b_{sq} D^s \overline{D}^q v_h, \tag{179}$$

где $D^s v_h = v_{xx\ldots x}$, $\overline{D}^q v_h = v_{\bar{x}\bar{x}\ldots\bar{x}}$, b_{sq} — постоянные коэффициенты.

Это позволяет вопросы исследования устойчивости разностных схем связать с изучением спектра разностных операторов, по собственным функциям которых могут быть разложены в ряд Фурье искомые решения разностной схемы.

Примеры исследования устойчивости разностных схем с использованием сеточного анализа Фурье.

Пример 1. Рассмотрим простейшую разностную схему (175), § 3, для однородного уравнения теплопроводности (172), § 3, при нулевых граничных условиях, т. е. при $\gamma_2(t) = \gamma_3(t) = 0$, $f(x, t) = 0$.

Схема будет иметь вид:

$$v_t^j = \Lambda_1 v_h^j,$$
$$v_{h0}^j(0) = v_{hn}^j(a) = 0, \tag{180}$$
$$v_{hi}^0 = \gamma_{1i}.$$

Будем искать решение схемы (180) в виде разложения в ряд Фурье по собственным функциям задачи (171)

$$v_{hi}^j = \sum_{k=1}^{n} c_k^j \omega_k(jh), \tag{181}$$

причем c_k^j определим так, чтобы разложение вида (181) имело место и для начальных условий, т. е. при $c_k^0 = (\gamma_1, \omega_k)$.

Подставляя предполагаемую формулу решения в искомое уравнение (180) и учитывая линейную независимость ω_k, получим рекуррентные соотношения для определения c_k^j:

$$c_k^{j+1} = q_k c_k^j, \quad q_k = 1 - \tau\lambda_k, \quad \lambda_k = \frac{4}{h^2}\sin^2\frac{\pi k h}{2a}$$

— k-е собственное значение оператора (171). Очевидно, $|q_k| < 1$ при $\frac{\tau}{h^2} \leqslant \frac{1}{2}$.

Если $|q_k| < 1$, то

$$\| v_h^{j+1} \|^2 = (v_h^{j+1}, v_h^{j+1}) = \sum_{i=1}^{n} h (v_{hi}^{j+1})^2 = \sum_{k=1}^{n} (c_k^{j+1})^2 \leqslant \sum_{k=1}^{n} (c_k^j)^2 = \| v_h^j \|^2,$$

т. е. схема (180) будет устойчивой по начальным данным в сеточной норме L_2.

Сеточный метод Фурье можно применить для исследования устойчивости двухслойных явных разностных схем вида:

$$v_t^j + Pv^j = F^j, \quad j = 0, 1, \ldots, \tag{182}$$
$$v^0 = \gamma(x),$$

если

$$P = P^* > 0 \tag{183}$$

и известен отрезок $[\alpha, \beta]$, которому принадлежат собственные значения λ_k оператора P:

$$P\omega_k(x) = \lambda_k \omega_k(x), \quad 0 < \alpha \leqslant \lambda_k \leqslant \beta. \tag{184}$$

Очевидно, при $A = A^* > 0$, $B = B^* > 0$ разностная схема (68), § 4, может быть приведена к виду (182) с оператором P, удовлетворяющим условию (183).

В дальнейшем будем предполагать, что система собственных функций оператора P образует ортонормированный базис n-мерного

пространства R_n. Введем в рассмотренные ряды Фурье

$$v_h^j = \sum_k c_k^j \omega_k(x), \quad F_h^j = \sum_k b_k^j \omega_k(x), \tag{185}$$

$$v_h^0 = \sum_k c_k^0 \omega_k(x).$$

Подставляя (185) в (182), получим для коэффициентов c_k^j, b_k^j следующее рекуррентное соотношение:

$$\frac{c_k^{j+1} - c_k^j}{\tau} + \lambda_k c_k^j = b_k^j, \quad k = \overline{1, n}, \quad j \geqslant 0$$

или

$$c_k^{j+1} = (1 - \tau\lambda_k) c_k^j + \tau b_k^j.$$

Обозначим

$$q_k = 1 - \tau\lambda_k.$$

Тогда, если

$$|q_k| = |1 - \tau\lambda_k| < 1, \tag{186}$$

то

$$\|v_h^{j+1}\| = \left\| \sum_{k=1}^n c_k^{j+1} \omega_k(x) \right\| = \left(\sum_k (c_k^{j+1})^2 \right)^{\frac{1}{2}} = \left(\sum_k (q_k c_k^j + \tau b_k^j)^2 \right)^{\frac{1}{2}} \leqslant$$

$$\leqslant \left(\sum_k (c_k^j)^2 \right)^{\frac{1}{2}} + \tau \left(\sum_k (b_k^j)^2 \right)^{\frac{1}{2}} = \|v_h^j\| + \tau \|F_h^j\|$$

или при выполнении условия (186) для решения задачи (182) справедлива оценка

$$\|v_h^{j+1}\| \leqslant \|v_h^0\| + \sum_{k=0}^{j} \tau \|F_h^k\|. \tag{187}$$

Очевидно, оценка (187) будет иметь место при

$$0 < \tau\beta \leqslant 2.$$

Сеточный метод Фурье может быть применен для исследования устойчивости трехслойных разностных схем.

Пример 2. Рассмотрим разностную схему (247), § 3, аппроксимирующую однородное уравнение (245), (246), § 3, с точностью $O(h^2 + \tau^2)$. Разностное уравнение (247), § 3, перепишем в виде

$$Av_h^{j+1} - Bv_h^j + Cv_h^{j-1} = 0, \tag{188}$$

где

$$A = I, \quad B = 2I + \tau^2\Lambda, \quad C = I. \tag{189}$$

Операторы A, B, C имеют общую систему собственных функций, образующих ортонормированный базис. Для доказательства сходимости разностной схемы можно применить метод разложения искомого решения в ряд Фурье по этой системе функций.

Если искать решение уравнения (188) в виде

$$v_h = v_0 e^{ikx + \beta t}, \quad v_0 = \text{const}, \quad k \text{ — целое}, \tag{190}$$

то для того чтобы функция (190) удовлетворяла уравнению (188) необходимо и достаточно, чтобы β и k удовлетворяли характеристическому уравнению

$$\rho_s^2 - (2 - \tau^2\lambda_s)\,\rho_s + 1 = 0. \tag{191}$$

Здесь $\rho = e^{\beta t}$, $\lambda_s = \dfrac{4}{h^2}\sin^2\dfrac{\pi h s}{2a}$. (192)

Очевидно, при $|\rho_s| < 1$ разностная схема (188) будет устойчива. Уравнение (191) представляет собой квадратный трехчлен. Имеет место следующая лемма:

Лемма (Критерий Гурвица). *Корни квадратного трехчлена $x^2 + ax + b = 0$ по модулю меньше или равны единице тогда и только тогда, когда выполнены следующие два условия:*

$$|a| \leqslant 1 + b, \quad |b| \leqslant 1.$$

Убедиться в справедливости леммы можно непосредственной проверкой.

Таким образом, разностная схема (188) будет устойчива при $|2 - \tau^2\lambda_s| \leqslant 2$. Следовательно, трехслойная разностная схема (188) абсолютно устойчива.

Аналогично можно доказать устойчивость производящей разностной схемы вида (261), (254′), § 3, при $q \leqslant 4$, если параметр α удовлетворяет условию

$$\alpha - \beta \geqslant \max\left\{0,\ \frac{1}{4} - 3\beta,\ \max_{k=\overline{2,q}} \frac{-3\beta\sqrt{2k} + \sqrt{18\beta^2(2-k) + (k-4)(1-k)\beta}}{(k-1)\sqrt{2k}}\right\}.$$

Здесь введено обозначение $\beta = \dfrac{h^2}{12\tau^2}$ ([17], [18]).

§ 5. ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ РАЗНОСТНЫХ УРАВНЕНИЙ

Метод сеток решения линейных задач математической физики приводит к системам линейных алгебраических уравнений. Эти системы обычно имеют следующие характерные особенности:

а) большое число неизвестных,

б) специфическое расположение ненулевых элементов и их редкость по отношению к нулевым элементам матриц.

Для решения таких систем уравнений обычно применяются итерационные методы (см. гл. 8), причем они оказываются наиболее эффективными, если известны границы спектра собственных чисел матрицы исходной системы.

Однако для большинства систем линейных алгебраических уравнений, соответствующих разностным методам решения задач математической физики, характерен большой разброс спектра. Следствием большой разбросанности спектра матрицы системы является медленная сходимость итерационных методов.

Для ряда разностных операторов специального типа можно построить прямые методы решения разностных уравнений, которые оказываются более эффективными по сравнению с самыми быстрыми итерационными методами. Эти прямые методы — быстрое дискретное преобразование Фурье (БДПФ), метод тензорных произведений (ТП), метод суммарных представлений (СП), метод биортогонализации и другие — в той или иной мере используют явное представление для собственных чисел и собственных функций матрицы системы разностных уравнений. Наиболее эффективными эти методы оказываются, когда исходная краевая задача допускает разделение переменных.

Метод суммарных представлений [59]. Одним из основных положений этого метода является построение формул суммарных представлений (ФСП), выражающих общее решение конечно-разностного уравнения через значения его на границе области или через небольшое число произвольных постоянных, которые подлежат определению из дополнительных (краевых или других) условий.

С точки зрения аналогии с классическими методами математической физики получение ФСП можно рассматривать как дискретный аналог метода конечных интегральных преобразований или метода функций Грина. Аналогично, как в непрерывном случае, с помощью конечного интегрального преобразования (если это возможно) двумерная задача сводится к распадающейся системе обыкновенных дифференциальных уравнений, точно также в методе СП применяют так называемую P-трансформацию, которая позволяет свести систему в частных конечных разностях к распадающейся системе обыкновенных разностных уравнений. Затем ищут аналитическое решение обыкновенных разностных уравнений, из которого обратным преобразованием (P^{-1}-трансформацией) получают ФСП. Основная идея метода СП на примере решения уравнения в частных разностях второго порядка может быть истолкована следующим образом.

Пусть требуется найти решение конечно-разностной задачи

$$L_h v_{ij} = f_{ij}, \quad (i, j) \in \Omega_h, \quad L_h = L_{1h} + L_{2h}, \tag{1}$$

$$l_h v_{jl} = \gamma_{ij}, \quad (i, j) \in \Gamma_h, \tag{2}$$

где

$$L_{1h} v_{ij} = a_i^{(1)} v_{i+1,j} - d_i^{(1)} v_{ij} + b_i^{(1)} v_{i-1,j}, \tag{3}$$

$$L_{2h} v_{ij} = a_j^{(2)} v_{i,j+1} - d_j^{(2)} v_{ij} + b_j^{(2)} v_{i,j-1},$$

$\Omega_h = \{(i, j),\ i = \overline{1, n},\ j = \overline{1, m}\}$ — сеточный прямоугольник с границей $\Gamma_h = \{(0, j),\ j = \overline{1, m}\} \cup \{(n+1, j),\ j = \overline{1, m}\} \cup$

$$\cup \{(i, 0),\ i = \overline{1, n}\} \cup \{(i, m+1),\ i = \overline{1, n}\}.$$

Введем две матрицы $\Pi^{(1)}$, $\Pi^{(2)}$, соответствующие разностным операторам L_{1h}, L_{2h},

$$\Pi^{(k)} = \begin{pmatrix} -d_1^{(k)} & a_1^{(k)} & 0 & \dots & 0 \\ b_2^{(k)} & -d_2^{(k)} & a_2^{(k)} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & b_{s_k-1}^{(k)} & -d_{s_k-1}^{(k)} & a_{s_k-1}^{(k)} \\ 0 & \dots & 0 & b_{s_k}^{(k)} & -d_{s_k}^{(k)} \end{pmatrix}, \tag{4}$$

$$k = 1,\ 2,\ s_1 = n,\ s_2 = m.$$

Пусть хотя бы одна из матриц $\Pi^{(k)}$ ($k = 1, 2$) является матрицей простой структуры (подобна диагональной). Для определенности пусть это будет матрица $\Pi^{(1)}$, т. е.

$$\Pi^{(1)} = PMP^{-1}, \tag{5}$$

где P-фундаментальная матрица, столбцами которой являются собственные векторы матрицы $\Pi^{(1)}$; $M = (\mu_i)_{i=\overline{1,n}}$ — диагональная матрица собственных чисел. Свойство матрицы $\Pi^{(1)}$ позволяет краевую задачу для уравнения в частных конечных разностях привести к краевой задаче для обыкновенного разностного уравнения.

Действительно, запишем краевую задачу (1) — (3) в векторной форме

$$\Pi^{(1)} v_j + L_{2h} v_j = f_j - \omega_j, \tag{6}$$

$$v_0 = \gamma_0, \quad v_{m+1} = \gamma_{m+1}, \tag{7}$$

где

$$v_j = (v_{ij})_{i=\overline{1,n}}, \quad f_j = (f_{ij})_{i=\overline{1,n}}, \quad \omega_j = (\gamma_{0j}, 0, \ldots, 0, \gamma_{n+1,j}) \tag{8}$$

n-мерные векторы.

Для любого n-мерного вектора v введем обозначения

$$\hat{v} = P^{-1} v = (\hat{v}_i)_{i=\overline{1,n}}, \tag{9}$$

тогда, производя P-трансформацию задачи (6), (7) (умножение слева уравнения (8) и условий (9) на матрицу P^{-1}) с учетом (5), (9), получаем:

$$M\hat{v}_j + L_{2h}\hat{v}_j = \hat{f}_j - \hat{\omega}_j, \tag{10}$$

$$\hat{v}_0 = \hat{\gamma}_0, \ \hat{v}_{m+1} = \hat{\gamma}_{m+1} \tag{11}$$

или в скалярном виде

$$a_j^{(2)} \hat{v}_{i,j+1} - (-\mu_i + d_j^{(2)}) \hat{v}_{ij} + b_j^{(2)} \hat{v}_{i,j-1} = \hat{f}_{ij} - \hat{\omega}_{ij}, \tag{10'}$$

$$\hat{v}_{i0} = \hat{\gamma}_{i0}, \ \hat{v}_{i,m+1} = \hat{\gamma}_{i,m+1} \quad (i, = \overline{1, n}; \ j = \overline{1, m}). \tag{11'}$$

Система вида (10′), (11′) — система с трехдиагональной матрицей и поэтому вообще может быть эффективно решена методом прогонки, но метод СП преследует цель построения ФСП, т. е. аналитического представления для решения исходной задачи. Поэтому предположим, что для каждого фиксированного i краевая задача (10′), (11′) имеет разностную функцию Грина, которую для определенности обозначим через $G_{pj}(\mu_i)$. Тогда решение краевой задачи (10′), (11′) можно представить в виде

$$\hat{v}_{ij} = b_1^{(2)} G_{1j}(\mu_i) \hat{\gamma}_{i0} + a_m^{(2)} G_{mj}(\mu_i) \hat{\gamma}_{i,m+1} - \\ - \sum_{p=1}^{m} G_{pj}(\mu_i) [\hat{f}_{ip} - \hat{\omega}_{ip}] \quad (j = \overline{1, m}). \tag{12}$$

Переходя в (12) к векторной форме записи и производя обратную P-трансформацию (умножение обеих частей (12) слева на матрицу P), получаем явное решение исходной разностной краевой задачи (1), (2)

$$v_j = b_1^{(2)} P G_{1j} P^{-1} \gamma_0 + a_m^{(2)} G_{mj} P^{-1} \gamma_{m+1} - \\ - P \sum_{p=1}^{m} G_{pj} P^{-1} [f_p - \omega_p] \quad (j = \overline{1, m}), \tag{13}$$

где

$$G_{pj} = [G_{pj}(\mu_i)]_{i=\overline{1,n}}$$

— диагональная матрица.

Формула (13) называется *формулой суммарных представлений (первого типа)*.

Таким образом, для построения формул суммарных представлений вида (13) необходимо:

1) изучить свойства матриц $\Pi^{(k)}$ $(k = 1, 2)$. Если хотя бы одна из матриц является матрицей простой структуры, то нужно построить фундаментальную матрицу P, матрицу собственных чисел M и задачу для уравнений в частных конечных разностях свести к задаче для обыкновенных конечных разностей;

2) решить задачу для обыкновенного конечно-разностного уравнения. Произвести обратную P-трансформацию найденного решения и получить формулу суммарных представлений.

Если обе матрицы $\Pi^{(k)}$ $(k = 1, 2)$ являются матрицами простой структуры, то при построении формул суммарных представлений можно построить фундаментальные матрицы и матрицы собственных чисел для матриц $\Pi^{(1)}$, $\Pi^{(2)}$, т. е. найти разложение

$$\Pi^{(k)} = P^{(k)} M^{(k)} (P^{(k)})^{-1} \quad (k = 1, 2). \tag{14}$$

Вводя обозначения

$$V = (v_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}}, \quad F = (f_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}}, \tag{15}$$

$$W = (b_i^{(1)} \delta_{1i} \gamma_{0j} + a_n^{(1)} \delta_{ni} \gamma_{n+1,j} + b_j^{(2)} \delta_{1j} \gamma_{i0} + a_m^{(2)} \delta_{mj} \gamma_{i,m+1})_{i=\overline{1,n}}^{j=\overline{1,m}},$$

конечно-разностную краевую задачу (1) — (3) записываем в матричном виде

$$\Pi^{(1)} V + V \Pi^{(2)} = F - W. \tag{16}$$

Используя соотношения (14), произведем двойную P-трансформацию матричного уравнения (16) (умножение обеих частей (16) слева на матрицу $(P^{(1)})^{-1}$ и справа на матрицу $P^{(2)}$), получим:

$$M^{(1)} \hat{V} + \hat{V} M^{(2)} = \hat{F} - \hat{W} \tag{17}$$

или в координатной форме

$$(\mu_i^{(1)} + \mu_j^{(2)}) \hat{v}_{ij} = \hat{f}_{ij} - \hat{W}_{ij} \quad (i = \overline{1, n}; \ j = \overline{1, m}), \tag{18}$$

где для любой матрицы $V = (v_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}}$ порядка $n \times m$ выражение $\hat{V}$ означает:

$$\hat{V} = (P^{(1)})^{-1} V P^{(2)} = (\hat{v}_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}}. \tag{19}$$

Если

$$\mu_i^{(1)} + \mu_j^{(2)} \neq 0 \quad (i = \overline{1, n}; \ j = \overline{1, m}), \tag{20}$$

то система (18) легко решается. Записав решение системы (18) и перейдя к матричной форме записи, произведем обратную P-трансформацию

(умножение обеих частей матричного уравнения слева на матрицу $P^{(1)}$ и справа на матрицу $(P^{(2)})^{-1}$), тогда получим искомое решение в виде формулы суммарных представлений второго типа

$$V = P^{(1)} \{M_1^{(-1)} \otimes [(P^{(1)})^{-1} (F + W) P^{(2)}]\} (P^{(2)})^{-1}. \tag{21}$$

Здесь знак $\otimes$ означает поэлементное умножение матриц

$$M_1^{-1} = \left[\frac{1}{\mu_i^{(1)} + \mu_j^{(2)}}\right]_{i=\overline{1,n}}^{j=\overline{1,n}}. \tag{22}$$

Примеры построения формул суммарных представлений для конкретных разностных уравнений.

Пример 1. Пусть в области $\overline{\Omega}_h$ (132), § 3, требуется найти решение конечно-разностного уравнения Пуассона (133), (134), § 3.

Запишем разностную задачу (133), (134), § 3, в виде

$$L_{1h} v_{ij} + \alpha^2 L_{2h} v_{ij} = h_1^2 f_{ij}, \quad (i, j) \in \Omega_h, \tag{23}$$

$$v_{ij} = \gamma_{ij} \quad (i, j) \in \Gamma_h, \tag{24}$$

где

$$L_{1h} = v_{i+1,j} + v_{i-1,j},$$

$$L_{2h} = v_{i,j+1} - 2\beta v_{ij} + v_{i,j-1}, \quad \beta = 1 + \alpha^2; \quad \alpha^2 = \frac{h_1^2}{h_2^2}. \tag{25}$$

Для построения формул суммарных представлений первого типа воспользуемся обозначениями (8), тогда уравнение в конечных разностях (23) можно записать в векторной форме

$$\Pi^{(1)} \boldsymbol{v}_j + \alpha^2 \boldsymbol{v}_{j+1} - 2\beta \boldsymbol{v}_j + \alpha^2 \boldsymbol{v}_{j-1} = h_1^2 \boldsymbol{f}_j - \boldsymbol{\omega}_j, \tag{26}$$

$$\boldsymbol{v}_0 = \boldsymbol{\gamma}_0, \quad \boldsymbol{v}_{m+1} = \boldsymbol{\gamma}_{m+1}, \tag{27}$$

где матрица $\Pi^{(1)}$ имеет вид

$$\Pi^{(1)} = [t_{ij}]_{i=\overline{1,n}}^{j=\overline{1,n}}; \quad t_{ij} = \begin{cases} 1, & i = j + 1, \quad i = j - 1, \\ 0, & \text{во всех остальных случаях.} \end{cases} \tag{28}$$

Задача о собственных числах и о собственных векторах матрицы $\Pi^{(1)}$

$$\Pi^{(1)} \boldsymbol{v} - 2\mu \boldsymbol{v} = 0 \tag{29}$$

при переходе к скалярной форме записи эквивалентна следующей конечно-разностной задаче Штурма — Лиувилля:

$$v_{k+1} - 2\mu v_k + v_{k-1} = 0, \quad v_0 = v_{n+1} = 0, \tag{29'}$$

собственные значения $2\mu_k$ которой находятся по формуле

$$2\mu_k = 2 \cos \frac{k\pi}{n+1}, \quad k = \overline{1, n}. \tag{30}$$

Система ортонормированных собственных векторов матрицы $\Pi^{(1)}$ будет иметь вид

$$p_{kj} = \frac{\sqrt{2}}{\sqrt{n+1}} \sin \frac{jk\pi}{n+1} \quad (k, j = \overline{1, n}). \tag{31}$$

Фундаментальная матрица

$$P = \sqrt{\frac{2}{n+1}} \left(\sin \frac{jk\pi}{n+1}\right)_{k=\overline{1,n}}^{j=\overline{1,n}} \tag{32}$$

матрицы $\Pi^{(1)}$ является матрицей ортогонального преобразования, т. е.

$$P = P^{-1} = P^*, \tag{33}$$

поэтому

$$\Pi^{(1)} = PMP, \tag{34}$$

где

$$M = (2\mu_i) = \left(2\cos\frac{i\pi}{n+1}\right)_{i=\overline{1,n}}$$

— диагональная матрица.

Заметим, что если матрица Π представима в виде

$$\Pi = a\Pi^{(1)} - bI, \tag{35}$$

где a и b — некоторые постоянные, то фундаментальная матрица для матрицы Π будет совпадать с фундаментальной матрицей для матрицы $\Pi^{(1)}$, а матрица собственных чисел Π будет определяться равенством

$$\widetilde{M} = (2\eta_j)_{j=\overline{1,n}} = \left(2\left(a\cos\frac{j\pi}{n+1} - \frac{b}{2}\right)\right)_{j=\overline{1,n}}. \tag{36}$$

Произведем P-трансформацию векторного уравнения (26), (27), тогда получим:

$$M\hat{v}_j + \alpha^2[\hat{v}_{j+1} - 2\beta\hat{v}_j + \hat{v}_{j-1}] = h_1^2\hat{f}_j - \hat{\omega}_j, \quad \hat{v}_j = Pv_j,$$

$$\hat{v}_0 = \hat{\gamma}_0, \quad \hat{v}_{m+1} = \hat{\gamma}_{m+1},$$

или в скалярной форме

$$\hat{v}_{i,j+1} - 2(\beta - \alpha^{-2}\mu_i)\hat{v}_{ij} + \hat{v}_{i,j-1} = h_2^2\hat{f}_{ij} - \alpha^{-2}\hat{\omega}_{ij}, \tag{37}$$

$$\hat{v}_{i0} = \hat{\gamma}_{i0}, \quad \hat{v}_{i,m+1} = \hat{\gamma}_{i,m+1} \quad (i = \overline{1, n}, \ j = \overline{1, m}).$$

Решение краевой задачи (37) может быть записано через функцию Грина

$$\hat{v}_{ij} = G_{1j}(\eta_i)\hat{\gamma}_{i0} + G_{mj}(\eta_i)\hat{\gamma}_{i,m+1} - \sum_{p=1}^{m} G_{pj}(\eta_i)[h_2^2\hat{f}_{ip} - \alpha^{-2}\hat{\omega}_{ip}], \tag{38}$$

где разностная функция Грина имеет вид

$$G_{pj}(\eta_i) = \frac{U_{|p,j|-1}(\eta_i)U_{m-(p,j)}(\eta_i)}{U_m(\eta_i)} \quad (p = \overline{1, m}). \tag{39}$$

Здесь $U_i(\eta_i)$ — многочлен Чебышева второго рода,

$$|p, j| = \min(p, j) = \frac{p + j - |p - j|}{2},$$

$$(p, j) = \max(p, j) = \frac{p + j + |p - j|}{2}, \tag{40}$$

η_i в соответствии с формулами (35), (36) будут иметь вид

$$\eta_i = 1 + \alpha^{-2}\left(1 - \cos\frac{i\pi}{n+1}\right). \tag{41}$$

Переходя в равенстве (38) к векторной форме записи и производя обратную P^{-1}-трансформацию, получим формулу суммарных представлений первого типа решения разностной задачи Дирихле для уравнения Пуассона в прямоугольнике $\overline{\Omega}_h$ (132), § 3,

$$v_j = PG_{1j}P\gamma_0 + PG_{mj}P\gamma_{m+1} - P\sum_{p=1}^{m} G_{pj}P[h_2^2\hat{f}_p - \alpha^{-2}\omega_p]. \tag{42}$$

Здесь P — матрица размерности $n \times n$, элементы которой имеют вид (32); $G_{pj} = (G_{pj}(\eta_i))_{i=\overline{1,n}}$ — диагональные матрицы размерности n, элементы которых определяются по формулам (39) — (41).

Для построения ФСП второго типа для разностной схемы (133), (134), § 3, в области $\overline{\Omega}_h$ (см. (132), § 3) используем обозначение (15), матрицы

$$H = (h_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}} = \begin{pmatrix} \gamma_{01} & \gamma_{02} & \dots & \gamma_{0m} \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ \gamma_{n+1,1} & \gamma_{n+1,2} & \dots & \gamma_{n+1,m} \end{pmatrix},$$

$$Q = (q_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}} = \begin{pmatrix} \gamma_{10} & 0 \dots 0 & \gamma_{1,m+1} \\ \gamma_{20} & 0 \dots 0 & \gamma_{2,m+1} \\ \dots & \dots & \dots \\ \gamma_{n0} & 0 \dots 0 & \gamma_{n,m+1} \end{pmatrix} \tag{43}$$

и квадратные матрицы $\Pi^{(1)}$ и $\Pi^{(2)}$ соответственно размерностей $n \times n$ и $m \times m$, элементы которых определяются из соотношения (28). Тогда задача (133), (134), § 3, может быть записана следующим образом:

$$\Pi^{(1)}V + \alpha^2 V\Pi^{(2)} - 2\beta V = h_1^2 F - H - \alpha^2 Q. \tag{44}$$

Для матриц $\Pi^{(1)}$ и $\Pi^{(2)}$ имеют место соотношения (см. (34)):

$$\begin{aligned} \Pi^{(1)} &= P^{(1)}M^{(1)}P^{(1)}, \\ \Pi^{(2)} &= P^{(2)}M^{(2)}P^{(2)}, \end{aligned} \tag{45}$$

где

$$P^{(1)} = (p_{ik}^{(1)})_{i=\overline{1,n}}^{k=\overline{1,n}} = \sqrt{\frac{2}{n+1}}\left(\sin\frac{ik\pi}{n+1}\right)_{i=\overline{1,n}}^{k=\overline{1,n}},$$

$$P^{(2)} = (p_{ik}^{(2)})_{i=\overline{1,m}}^{k=\overline{1,m}} = \sqrt{\frac{2}{m+1}}\left(\sin\frac{ik\pi}{m+1}\right)_{i=\overline{1,m}}^{k=\overline{1,m}} \tag{46}$$

— квадратные матрицы соответственно n-го и m-го порядков, причем $P^{(i)}P^{(i)} = I$, $i = 1, 2$,

$$M^{(1)} = (2\mu_i^{(1)})_{i=\overline{1,n}} = \left(2\cos\frac{i\pi}{n+1}\right)_{i=\overline{1,n}},$$

$$M^{(2)} = (2\mu_i^{(2)})_{i=\overline{1,m}} = \left(2\cos\frac{i\pi}{m+1}\right)_{i=\overline{1,m}}$$

— диагональные матрицы n-го и m-го порядков.

Введем P-трансформированные матрицы

$$\begin{aligned} \hat{V} &= P^{(1)}VP^{(2)}, \ \hat{F} = P^{(1)}FP^{(2)}, \\ \hat{H} &= P^{(1)}HP^{(2)}, \ \hat{Q} = P^{(1)}QP^{(2)}. \end{aligned} \tag{45'}$$

После умножения равенства (44) слева на $P^{(1)}$ и справа на $P^{(2)}$ получим:

$$M^{(1)}\hat{V} + \alpha^2\hat{V}M^{(2)} - 2\beta\hat{V} = h_1^2\hat{F} - \hat{H} - \alpha^2\hat{Q}$$

или в скалярной форме

$$2\mu_i^{(1)}\hat{v}_{ij} + 2\alpha^2\mu_j^{(2)}\hat{v}_{ij} - 2\beta\hat{v}_{ij} = h_1^2\hat{f}_{ij} - \alpha^2\hat{q}_{ij} - \hat{h}_{ij}. \tag{47}$$

Здесь $\hat{v}_{ij}$, $\hat{f}_{ij}$, $\hat{h}_{ij}$, $\hat{q}_{ij}$ — элементы матриц $\hat{V}$, $\hat{F}$, $\hat{H}$, $\hat{Q}$.

Из (47) имеем:

$$2(\beta - \mu_i^{(1)} - \alpha^2\mu_j^{(2)})\hat{v}_{ij} = -h_1^2\hat{f}_{ij} + \alpha^2\hat{q}_{ij} + \hat{h}_{ij}$$

$$(i = \overline{1, n}; \quad j = \overline{1, m}).$$

Пусть $\beta - \mu_i^{(1)} - \alpha^2\mu_j^{(2)} \neq 0$, тогда

$$\hat{v}_{ij} = \frac{-h_1^2\hat{f}_{ij} + \alpha^2\hat{q}_{ij} + \hat{h}_{ij}}{2(\beta - \mu_i^{(1)} - \alpha^2\mu_j^{(2)})}. \tag{48}$$

Из (45′), (46) следует, что

$$v_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{m} p_{ik}^{(1)}p_{jl}^{(2)}\hat{v}_{kl},$$

или, учитывая (48),

$$v_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{m} \frac{p_{ik}^{(1)}p_{jl}^{(2)}}{2(\beta - \mu_k^{(1)} - \alpha^2\mu_l^{(2)})} \sum_{\mu=1}^{n}\sum_{\nu=1}^{m} p_{\mu k}^{(1)}p_{\nu l}^{(2)}(-h_1^{(2)}f_{\mu\nu} + h_{\mu\nu} + \alpha^2 q_{\mu\nu}). \tag{49}$$

Формулу (49) можно записать в виде (21). Скалярная форма записи формулы суммарных представлений (49) совпадает с формулой решения задачи (133), (134), § 3, полученной методом тензорного произведения.

Формулы метода тензорного произведения получаются из представления разностной схемы (133), (134), § 3, в виде (150), § 3. Тогда, если учесть, что треугольные матрицы $\tilde{\Lambda}_1$ и $\tilde{\Lambda}_2$ являются матрицами простой структуры вида (35), причем

$$\tilde{\Lambda}_1 = \tilde{P}^{(1)}\tilde{M}^{(1)}\tilde{P}^{(1)}, \quad \Lambda_2 = \tilde{P}^{(2)*}\tilde{M}^{(2)}\tilde{P}^{(2)},$$

то из (150), § 3, получаем, что решение разностной задачи Дирихле можно представить в виде

$$V = \tilde{P}^{(2)} \otimes \tilde{P}^{(1)}(I_m \otimes \tilde{M}^{(1)} + \tilde{M}^{(2)} \otimes I_n)^{-1}\tilde{P}^{(2)} \otimes \tilde{P}^{(1)}F. \tag{50}$$

Скалярная форма записи формулы (50) имеет вид (49).

По аналогии с формулами (42), (49) могут быть построены формулы суммарных представлений решения разностного уравнения (133), § 3, при условии задания на сторонах прямоугольника (132), § 3, периодических краевых условий.

Например, если на горизонтальных сторонах сеточного прямоугольника (132), § 3, заданы краевые условия периодичности, т. е.

$$v_{i0} = v_{im}, \quad v_{i,m+1} = v_{i1} \quad (i = \overline{1, n}), \tag{51}$$

а на вертикальных — краевые условия первого рода

$$v_{0k} = \gamma_{0k}, \quad v_{n+1,k} = \gamma_{n+1,k} \quad (k = \overline{1, m}), \tag{52}$$

то с помощью прямоугольных матриц (43), квадратной матрицы (28) и квадратной матрицы $\Pi^{(3)}$ порядка m

$$\Pi^{(3)} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & & \cdot & \cdots & 1 & 0 & 1 \\ 1 & 0 & & \cdot & \cdots & 0 & 1 & 0 \end{pmatrix} \tag{53}$$

разностную краевую задачу (133), § 3, (51), (52) можно записать в виде

$$\Pi^{(1)} V + \alpha^2 V \Pi^{(3)} - 2\beta\, V = h_1^2 F - H. \tag{54}$$

Фундаментальная матрица $P^{(3)}$ и матрица собственных чисел $M^{(3)}$ матрицы $\Pi^{(3)}$ может быть найдена в результате решения задачи (29′) с краевыми условиями периодичности $v_0 = v_m$, $v_{m+1} = v_1$:

$$P^{(3)} = (p_{ij}^{(3)})_{i=\overline{1,m}}^{j=\overline{1,m}} = \sqrt{\frac{2}{n}} \left(\zeta_j + \delta_j \sin\left(\frac{(-1)^j + 1}{2} \frac{\pi}{2} + i\xi_j \frac{2\pi}{m} \right) \right)_{i=\overline{1,m}}^{j=\overline{1,m}},$$

где

$$\zeta_j = \begin{cases} \dfrac{1}{\sqrt{2}} & \text{при } j = 1, \\ 0 & \text{при } j = \overline{2, n}; \end{cases} \qquad \delta_k = \begin{cases} \dfrac{1}{\sqrt{2}} & \text{при } j = m, \text{ если } m \text{ — четное}, \\ 1 & \text{во всех остальных случаях}, \end{cases}$$

$$\xi_j = \operatorname{entier} \frac{j}{2} \quad (j = \overline{1, m}), \tag{55}$$

$$P^{(3)} P^{(3)*} = I,$$

$$M^{(3)} = 2\left[1, \cos\alpha, \cos\alpha, \cos 2\alpha, \cos 2\alpha, \ldots, \cos\frac{m-1}{2}\alpha, \cos\frac{m-1}{2}\alpha\right], \tag{56}$$

диагональная матрица порядка m, $\alpha = \dfrac{2\pi}{m}$.

Учитывая, что

$$\Pi^{(3)} = P^{(3)} M^{(3)} P^{(3)*}, \tag{57}$$

и проводя аналогичные преобразования, как и при выводе формулы (49), получим формулу суммарных представлений решения разностного уравнения Пуассона с периодическими краевыми условиями, заданными на горизонтальных сторонах прямоугольника [61]

$$v_{ij} = \sum_{k=1}^{n} \sum_{l=1}^{m} \frac{p_{ik}^{(1)} p_{jl}^{(3)}}{2(\beta - \mu_k^{(1)} - \alpha^2 \mu_l^{(3)})} \sum_{\mu=1}^{n} \sum_{\nu=1}^{m} p_{\mu k}^{(1)} p_{\nu l}^{(3)} (-h_1^2 f_{\mu\nu} + h_{\mu\nu}), \tag{58}$$

если $\beta - \mu_k^{(1)} - \alpha^2 \mu_l^{(3)} \neq 0$, $k = \overline{1, n}$; $l = \overline{1, m}$. Аналогично могут быть рассмотрены всевозможные комбинации краевых условий первого, второго и третьего родов и условий периодичности на сторонах сеточного прямоугольника.

Указанный прямой метод решения разностных уравнений может быть обобщен на случай большего числа переменных и других типов дифференциальных уравнений как с постоянными, так и переменными коэффициентами [48], [60], [59].

Метод суммарных представлений можно рассматривать как аналитический аппарат решения задач математической физики в дискретной постановке. Как известно, одним из наиболее эффективных методов получения аналитического решения краевых задач математической физики для односвязных (двухсвязных) областей является метод конформных отображений данной области на круг (круговое кольцо). Если воспользоваться аппаратом конформных отображений и методом суммарных представлений, то во многих случаях удается расширить вид областей, для которых можно найти решение соответствующей конечно-разностной задачи в явном виде или в виде формулы, содержащей небольшое число параметров.

Так, если рассмотреть уравнение Пуассона

$$\Delta_{\sigma\tau}\omega = F(\sigma, \tau), \quad \Delta_{\sigma\tau} = \frac{\partial^2}{\partial\sigma^2} + \frac{\partial^2}{\partial\tau^2} \tag{59}$$

в декартовой системе координат σ, τ, то в переменных

$$x = \ln\rho, \quad y = \theta\left(\rho = \sqrt{\sigma^2 + \tau^2}, \quad \theta = \operatorname{arctg}\frac{\tau}{\sigma}\right)$$

уравнение (59) можно записать в виде

$$\Delta_{xy}u = f(x, y), \tag{60}$$

где

$$u(x, y) = \omega(\rho, \theta), \quad f(x, y) = e^{2x}F(x, y).$$

Узлам прямоугольной сетки плоскости x, y:

$$x_i = x_0 + ih_1, \quad y_j = y_0 + jh_2 \tag{61}$$

$$(i = 0, \pm 1, \pm 2, \ldots; \quad j = 0, \pm 1, \pm 2, \ldots)$$

в плоскости σ, τ соответствует дискретное множество точек

$$\rho_i = \rho_0 e^{ih_1}, \quad \vartheta_j = jh_2 \quad (\rho_0 = e^{x_0}, \quad i, k = 0, \pm 1, \pm 2, \ldots). \tag{62}$$

Пусть $\omega = \omega(\rho_i, \vartheta_j)$ — функция дискретных переменных ρ_i, ϑ_j, определенная на множестве точек (62). Если положить

$$\omega(\rho_i, \vartheta_j) = u(x_i, y_j),$$

то $\omega(\rho_i, \vartheta_j)$, как функция от i, j на множестве точек (62), будет удовлетворять уравнению (133), § 3.

Поэтому результаты, известные для уравнения (133), § 3, можно применить к нахождению значений $\omega(\rho_i, \vartheta_j)$. Так, формула (58) будет давать решение конечно-разностного уравнения Пуассона в сеточном кольце, определяющемся совокупностью точек (ρ_i, ϑ_j) изометрической сетки

$$\rho_i = \rho_0 e^{ih_1}, \quad \vartheta_j = jh_2 = j\frac{2\pi}{m},$$

$$\left(h_1 > 0, \quad i = \overline{0, n+1}, \quad j = \overline{1, m}, \quad n = \operatorname{entier}\left(\frac{1}{n_1}\ln\frac{R}{\rho_0} - 1\right)\right),$$

если на границе (R, ϑ_j) сеточного кольца заданы краевые условия Дирихле.

В случае краевых задач, связанных с эллиптическими уравнениями более высокого порядка, в частности краевых задач, связанных с бигармоническими уравнениями, для построения ФСП можно использовать представление решения бигармонического уравнения через гармонические функции. Используя метод конформных отображений, удается построить ФСП решения основной бигармонической задачи в областях не только прямоугольной формы, но и для произвольной односвязной области Ω, если известна функция, дающая конформное отображение круга на область Ω.

Формулы суммарных представлений могут быть использованы для решения краевых задач в области более сложной конфигурации, составленных из канонических областей.

Продемонстрируем применение метода суммарных представлений для решения краевых задач в областях, составленных из прямоугольников. Для простоты рассмотрим L-образную область $\Omega = \Omega_1 + \Omega_2$, которая состоит из двух прямоугольников Ω_1 и Ω_2. Пусть в области Ω требуется найти решение задачи Дирихле (130), (131), § 3.

Дискретизируя эту задачу так, как это делалось ранее для прямоугольника, записываем формулу суммарных представлений (42) для каждого из сеточных прямоугольников:

$$\Omega_{1h} = \{(x_i, y_j), \quad i = \overline{0, n+1}, \quad j = \overline{0, m+1}\}, \tag{63}$$

$$\Omega_{2h} = \{(x_i, y_j), \quad i = \overline{n+1, n'+n+2}, \quad j = \overline{0, m'+1}\}.$$

В формуле (42) значения искомого решения в m' точках сетки на стыке прямоугольников будут выступать в качестве неизвестных параметров. Чтобы определить эти параметры в каждой точке, лежащей на стыке прямоугольников, записываем разностное уравнение (133), (136), § 3,

$$h_1^{-2}[v_{n+2,j} - 2v_{n+1,j} + v_{n,j}] + h_2^{-2}[v_{n+1,j+1} - 2v_{n+1,j} + \\ + v_{n+1,j-1}] = f_{n+1,j} \quad (j = \overline{1, m'}). \tag{64}$$

Подставляя в (64) вместо $v_{n,j}$ $(j = \overline{1, m'})$ значения, найденные по формуле суммарных представлений для прямоугольника Ω_{1h}, вместо $v_{n+2,j}$ $(j = \overline{1, m'})$ — значения, найденные по формуле суммарных представлений для прямоугольника Ω_{2h}, получим систему линейных алгебраических уравнений m'-го порядка. Эта система может быть приведена к виду

$$\sum_{l=1}^{m'} c_{l,j} v_{n+1,l} = F_{n+1,j}, \quad j = \overline{1, m'} \tag{65}$$

и для своего решения требует $O(m'^3)$ арифметических операций. Общее число арифметических операций при использовании формул суммарных представлений для областей, составленных из прямоугольников будет порядка $O\left(\frac{1}{h^3}\right)$. Экономичные итерационные методы в этом случае требуют операций порядка $O\left(\frac{1}{h^3} \ln \frac{1}{h}\right)$. Явные методы решения разностных уравнений будут обладать определенными преимуществами перед итерационными методами, если решение исходной задачи нужно знать не во всей области, а лишь в определенной ее части или нужно вычислить некоторые характеристики (интегральные, дифференциальные), связанные с решением краевой задачи. Это обусловлено тем, что запись решения в виде формулы суммарных представлений позволяет проводить выборочный счет и производить необходимые для получения данной характеристики преобразования над самой формулой. Метод суммарных представлений применяется для решения различных классов задач теории фильтрации, теории пластин и оболочек, задач на собственные значения [14], [22], [44], [45], [47].

§ 6. МЕТОД ПРЯМЫХ. МЕТОД ИНТЕГРАЛЬНЫХ СООТНОШЕНИЙ

Метод прямых и метод интегральных соотношений принадлежат к группе приближенных методов решения краевых задач, основная идея которых связана с понижением размерности решаемой задачи. В этих методах размерность понижается за счет аппроксимации исходной задачи для дифференциального уравнения в частных производных или системы дифференциальных уравнений также системой диффренциальных уравнений, но с меньшим числом непрерывных переменных.

В классическом методе прямых приближенное решение задачи для уравнений в частных производных в случае двух независимых переменных сводится к решению задачи для обыкновенных дифференциальных уравнений. Для этого область интегрирования разбивается на полосы обычно фиксированными прямыми линиями, и производные по одному из направлений заменяются конечно-разностными уравнениями. Полученная при этом система обыкновенных дифференциальных уравнений обычно решается численными методами. В методе интегральных соотношений разбиение области проводится кривыми линиями, форма которых определяется видом границы области интегрирования. Кроме того, в методе интегральных соотношений широко используются различные классы аппроксимирующих и весовых функций, выбор которых ставится в соответствие с поведением ожидаемого решения. В результате можно получить хорошую точность при малом числе полос. В этом плане метод интегральных соотношений является более гибким. Однако, если вопросы сходимости приближенных решений, полученных по методу прямых, для основных разностных уравнений математической физики изучены довольно полно, то этого нельзя сказать в настоящее время о методе интегральных соотношений.

1. Метод прямых

В методе прямых, в отличие от метода сеток, конечно-разностными выражениями аппроксимируются производные по какой-либо одной переменной.

Основную идею метода прямых изложим на примере решения краевой задачи Дирихле в прямоугольнике

$$\overline{\Omega} = \{0 \leqslant x \leqslant a, \quad 0 \leqslant y \leqslant b\} \tag{1}$$

для уравнения эллиптического типа в предположении, что коэффициенты, правая часть уравнения и граничные функции удовлетворяют требованиям существования и единственности решения поставленной задачи

$$c(x, y)\frac{\partial^2 u}{\partial x^2} + d(x, y)\frac{\partial^2 u}{\partial y^2} + e(x, y)\frac{\partial u}{\partial x} + q(x, y)u = f(x, y), \tag{2}$$

$$u(x, 0) = \gamma_0(x), \quad u(x, b) = \gamma_1(x), \tag{3}$$

$$u(0, y) = \gamma_2(y), \quad u(a, y) = \gamma_3(y) \tag{4}$$

при

$$c(x, y) \geqslant c^0 > 0, \quad d^0 \geqslant d(x, y) \geqslant \hat{d}^0 > 0$$
$$q(x, y) \leqslant -q^0 < 0. \tag{5}$$

Построим решетчатую область Ω_m с границей Γ_m, разбив прямоугольник Ω на $(m+1)$ полосу прямыми $y_j = jh$

$$\Omega_m = \left\{0 < x < a, \quad y_j = jh, \quad h = \frac{b}{m+1}, \quad j = \overline{1, m}\right\}. \tag{6}$$

Граница Γ_m состоит из отрезков прямых

$$0 \leqslant x \leqslant a, \quad y = 0; \quad 0 \leqslant x \leqslant a, \quad y = b \tag{6'}$$

и точек $(0, y_j), (a, y_j)$, $j = \overline{1, m}$.

На решетке Ω_m заменим производную $\left.\frac{\partial^2 u}{\partial y^2}\right|_{y=y_j}$ разностным выражением вида

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{y=y_j} = \frac{1}{h^2}[u(x, y_{j-1}) - 2u(x, y_j) + u(x, y_{j+1})] +$$
$$+ \frac{h^2}{12} \frac{\partial^4 u(x, y_j + \theta h)}{\partial x^4}, \quad |\theta| < 1 \tag{7}$$

Тогда на решетке (6), (6′) систему, аппроксимирующую исходную задачу (2) — (4) с погрешностью $O(h^2)$, можно записать в виде

$$L_j(v_j) = c_j(x) \frac{d^2 v_j(x)}{dx^2} + e_j(x) \frac{dv_j(x)}{dx} +$$
$$+ d_j(x) \frac{v_{j-1}(x) - 2v_j(x) + v_{j+1}(x)}{h^2} + q_j(x) v_j(x) = f_j(x), \tag{8}$$
$$v_0(x) = \gamma_0(x), \quad v_{m+1}(x) = \gamma_1(x)$$
$$v_j(0) = \gamma_2(y_j), \quad v_j(a) = \gamma_3(y_j), \quad j = \overline{1, m}.$$

Здесь использованы следующие обозначения:

$$c_j(x) = c(x, y_j), \quad e_j(x) = e(x, y_j),$$
$$d_j(x) = d(x, y_j), \quad q_j(x) = q(x, y_j), \quad f_j(x) = f(x, y_j).$$

Разрешимость и равномерная сходимость системы (8) к решению исходной задачи может быть доказана с помощью принципа максимума.

Теорема 1. *Пусть $v_j(x)$ является решением задачи* (8) *при* $\gamma_0(x) = \gamma_1(x) = 0$, $\gamma_2(y_j) = \gamma_3(y_j) = 0$ *и* $f_j(x) \geqslant 0$ *для всех* j *и* x. *Тогда*

$$v_j(x) \leqslant 0 \quad (j = \overline{0, m+1}). \tag{9}$$

Доказательство проводится от противного. Предположим, что при $j = l$ $(1 \leqslant l \leqslant m)$ и $x = \eta \in (0, a)$ функция $v_l(x)$ принимает максимальное положительное значение, т. е.

$$v_l(\eta) > 0, \quad v_l'(\eta) = 0, \quad v_l''(\eta) \leqslant 0.$$

Тогда

$$L_l(v_l(\eta)) = c_l(x) v_l''(\eta) + d_l(\eta) \frac{v_{l-1}(\eta) - 2v_l(\eta) + v_{l+1}(\eta)}{h^2} +$$
$$+ q_l(\eta) v_l(\eta) = f_l(\eta) < 0,$$

что противоречит условию теоремы. Аналогично показывается, что при $f_j(x) \leqslant 0$ выполняется соотношение $v_j(x) \geqslant 0$.

Иными словами, сетчатая функция $v_j(x)$, являющаяся решением задачи (8), с коэффициентами $c(x, y)$, $d(x, y)$, $q(x, y)$, удовлетворяющими условиям (5), при $f_j(x) \geqslant 0$ не может принимать в Ω_m положительного максимума, а при $f_j(x) \leqslant 0$ отрицательного минимума.

Из принципа максимума следует единственность решения системы уравнений (8). В самом деле, если существует два решения системы (8), удовлетворяющие одинаковым краевым условиям, то их разность будет удовлетворять соответствующей однородной системе с нулевыми граничными условиями и в силу принципа максимума тождественно равна нулю.

Для исследования сходимости решения системы (8) к решению исходной задачи (2) — (4) докажем теорему сравнения.

Теорема 2. *Пусть функции* $z_j(x)$ и $\omega_j(x)$ *удовлетворяют уравнениям:*

$$L_j z_j(x) = \psi_j(x) \quad L_j \omega_j(x) = \varphi_j(x) \quad (j = \overline{1, m}, \quad 0 < x < a) \tag{10}$$

в Ω_m *и условиям:*

$$\omega_j(0) \geqslant |z_j(0)|, \quad \omega_j(a) \geqslant |z_j(a)| \quad (j = \overline{1, m}), \tag{11}$$

$$\omega_0(x) \geqslant |z_0(x)|, \quad \omega_{m+1}(x) \geqslant |z_{m+1}(x)|, \quad 0 \leqslant x \leqslant a \tag{12}$$

на Γ_m, *где*

$$L_j z_j(x) = c_j(x) z_j''(x) + d_j(x) \frac{z_{j+1}(x) - 2z_j(x) + z_{j-1}(x)}{h^2} + \\ + e_j(x) z_j'(x) + q_j(x) z_j(x). \tag{13}$$

Тогда, если $\varphi_j(x) \leqslant -|\psi_j(x)|$ $(j = \overline{1, m})$ *и коэффициенты в* (13) *удовлетворяют условиям (5), то*

$$|z_j(x)| \leqslant \omega_j(x) \text{ в } \Omega_m + \Gamma_m. \tag{14}$$

Доказательство. Рассмотрим две системы функций:

$$\omega_j(x) \pm z_j(x) \quad j = \overline{0, m+1}, \quad 0 \leqslant x \leqslant a. \tag{15}$$

Очевидно,

$$L_j(\omega_j(x) \pm z_j(x)) = \varphi_j \pm \psi_j \leqslant 0 \text{ в } \Omega_m.$$

Из (11) и (12) имеем:

$$\omega_j(x) \pm z_j(x) \geqslant 0 \quad \text{на} \quad \Gamma_m,$$

т. е.

$$\omega_j(x) \pm z_j(x) \geqslant 0 \quad \text{в} \quad \Omega_m + \Gamma_m$$

или

$$\omega_j(x) > |z_j(x)| \quad j = \overline{0, m+1}, \quad 0 \leqslant x \leqslant a. \tag{16}$$

Обозначим через

$$z_j(x) = u(x, y_j) - v_j(x)$$

разность между точным решением исходной задачи и решением системы (8) на прямой $y = y_j$. Тогда

$$L_j(z_j(x)) = \psi_j(x), \quad j = \overline{1, m}, \quad 0 < x < a, \tag{17}$$

$$z_0(x) = z_{m+1}(x) = 0,$$

$$z_j(0) = z_j(a) = 0, \quad j = \overline{1, m}. \tag{18}$$

Для величины $\psi_j(x)$ можно указать следующую мажорантную оценку:

$$|\psi_j(x)| \leqslant \frac{h^2}{12} M_4 d^0, \quad j = \overline{1, m}, \quad 0 \leqslant x \leqslant a, \tag{19}$$

где

$$M_4 = \max_{\Omega} \left| \frac{\partial^4 u(x, y)}{\partial y^4} \right|.$$

Рассмотрим систему функций:

$$\omega_j(x) = \frac{h^2}{12q^0} M_4 d^0, \quad j = \overline{0, m+1}, \quad 0 \leqslant x \leqslant a. \tag{20}$$

Очевидно,

$$L_j(\omega_j(x)) = q_j(x) \frac{h^2}{12q^0} M_4 d^0 \leqslant -\frac{h^2}{12} M_4 d^0 \leqslant -|\psi_j(x)|,$$

$$j = \overline{1, m}, \quad 0 < x < a,$$

где $|\psi_j(x)|$ оценивается неравенством (19).

Для функций $\omega_j(x)$ и $z_j(x)$ выполняются условия теоремы сравнения, а поэтому

$$|z_j(x)| \leqslant \frac{h^2}{12q^0} M_4 d^0, \quad j = \overline{0, m+1}, \quad 0 \leqslant x \leqslant a.$$

Таким образом, имеет место следующая теорема:

Теорема 3. *Если существует единственное решение задачи Дирихле* (1) — (5), *обладающее непрерывной производной* $\frac{\partial^4 u(x, y)}{\partial y^4}$, *то для погрешности* $z(x)$ *метода прямых вида* (8) *справедлива оценка*

$$|z_j(x)| \leqslant \frac{h^2 M_4 d^0}{12q^0}, \quad j = \overline{0, m+1}, \quad 0 \leqslant x \leqslant a. \tag{21}$$

Метод прямых дает возможность построить алгоритм решения краевой задачи в прямоугольной области, так как каждое из уравнений системы (8) определено на пересечении $D = \bigcap_{-k \leqslant l \leqslant \nu} \eta_{j+l}$, где η_{j+l} — проекция на ось x внутреннего сечения области Ω прямой y_{j+l}. В случае областей неправильной формы, например, для криволинейной трапеции граничная задача для системы обыкновенных дифференциальных уравнений (8) будет неразрешима.

Предложены специальные подходы для реализации метода прямых в этом случае (см., например, [1], [36]), но при этом значительно усложняется вычислительный алгоритм.

Изложенная выше методика может быть применена к эллиптическим уравнениям вида

$$c(x) \frac{\partial^2 u}{\partial x^2} + b(x) \frac{\partial^2 u}{\partial y^2} + d(x) \frac{\partial u}{\partial x} + q(x) u = f(x, y),$$

$$c(x), \quad b(x) > 0,$$

но аппроксимирующая задача в случае уравнений с переменными коэффициентами требует для своего решения, как правило, больших вычислительных затрат, чем метод сеток в этом же случае.

Аналогичные построения метода прямых и исследование его сходимости можно провести для краевых задач, связанных с уравнениями гиперболического и параболического типов. При этом могут быть построены так называемые продольные и поперечные системы метода прямых.

Так, в полуполосе

$$\Omega = (0 \leqslant x \leqslant a, \quad 0 \leqslant t < \infty) \tag{22}$$

для граничной задачи, связанной с уравнениями параболического типа

$$Lu = c(x, t)\frac{\partial^2 u(x, t)}{\partial x^2} + d(x, t)\frac{\partial u(x, t)}{\partial x} + q(x, t)u(x, t) - \\ - \frac{\partial u}{\partial t} = f(x, t), \tag{23}$$

$$\begin{gathered} u(x, 0) = \gamma_0(x), \\ u(0, t) = \gamma_2(t), \quad u(a, t) = \gamma_3(t), \end{gathered} \tag{24}$$

где $c(x, t) \geqslant c^0 > 0$, $q(x) \geqslant 0$, можно построить приближенный алгоритм как на поперечной решетке

$$\Omega_m + \Gamma_m = \{0 \leqslant x \leqslant a, \quad t_j = j\tau, \quad j = \overline{0, m+1}, \quad \tau > 0\}, \tag{25}$$

так и на продольной решетке

$$\Omega_n + \Gamma_n^- = \left\{x_i = ih, \quad h = \frac{a}{n+1}, \quad i = \overline{0, n+1}, \quad 0 \leqslant t \leqslant T\right\}. \tag{26}$$

Поперечная схема метода прямых на решетке (25), имеющая погрешность аппроксимации $O(\tau)$, может быть представлена следующим образом:

$$\begin{gathered} L_j(v_j(x)) = c_j(x)v_j''(x) + d_j(x)v_j'(x) + q_j(x)v_j(x) - \\ - \frac{v_j(x) - v_{j-1}(x)}{\tau} = f_j(x), \quad v_0(x) = \gamma_0(x), \\ v_j(0) = \gamma_2(t_j), \quad v_j(a) = \gamma_3(t_j). \end{gathered} \tag{27}$$

Исследование разрешимости системы (27) и ее сходимости к решению задачи (23) проводится аналогично разобранному выше примеру, причем если решение задачи (23) будет обладать в прямоугольнике Ω непрерывной производной $\frac{\partial^2 u(x, t)}{\partial t^2}$, то для погрешности системы вида (27) справедлива оценка

$$|z_j(x)| \leqslant O(\tau), \quad j = \overline{1, m}. \tag{28}$$

Продольная схема на решетке (26), если производные u_x, $u_{\bar{x}x}$ на каждой из прямых $x = x_i$ заменить соответственно по формулам (16), (22) табл. 4, будет иметь вид

$$L_k(v_k(t)) = c_k(t)\frac{v(x_{k+1}, t) - 2v(x_k, t) + v(x_{k-1}, t)}{h^2} +$$

$$+ d_k(t)\frac{v(x_{k+1}, t) - v(x_{k-1}, t)}{2h} + q_k(t)\, v(x_k, t) - \frac{dv(x_k, t)}{dt} = f(x_k, t), \quad (29)$$

$$v_0(t) = \gamma_2(t), \quad v_{n+1}(t) = \gamma_3(t), \quad v_k(0) = \gamma_0(x_k), \quad k = \overline{1, n}.$$

Здесь $c_k(t) = c(x_k, t)$. Аналогичные обозначения приняты для функций $d_k(t)$, $q_k(t)$.

Погрешность $z_k(t) = u(x_k, t) - v(x_k, t)$ продольной схемы (29) удовлетворяет следующей задаче Коши:

$$L_k(z_k(t)) = r_k(t),$$

$$z_0(t) = 0, \quad z_{n+1}(t) = 0, \quad z_k(0) = 0, \quad k = \overline{1, n}, \quad (30)$$

причем

$$|r_k(t)| \leqslant \varkappa h^2, \quad k = \overline{1, n}, \ \varkappa = \frac{1}{12}\Big(\max_{\Omega_n} c(x, t) \times$$

$$\times \max_{\Omega_n}\left|\frac{\partial^4 u(x)}{\partial x^4}\right| + 2\max_{\Omega_n}\left|\frac{\partial^3 u(x, t)}{\partial x^3}\right| \max_{\Omega_n}|d(x, t)|\Big\}.$$

Задачу Коши (30) можно записать в векторно-матричной форме:

$$\frac{dz(t)}{dt} = B(t)\, z(t) - r(t),$$

$$z(0) = 0, \quad (31)$$

где

$$B(t) = (b_{ij}(t))_{i=\overline{1,n}}^{j=\overline{1,n}},$$

$$b_{ii} = -\frac{2}{h^2} c_i(t) + g_i(t),$$

$$b_{i,i-1} = \frac{1}{h^2}\left(c_i(t) - \frac{h}{2} d_i(t)\right),$$

$$b_{i,i+1} = \frac{1}{h^2}\left(c_i(t) + \frac{h}{2} d_i(t)\right),$$

$$b_{ij} = 0 \quad (i, j = \overline{1, n}, \quad i \neq j, \quad j \neq i - 1, \quad j \neq i + 1).$$

$$z(t) = (z_i(t))'_{i=\overline{1,n}}, \quad r(t) = (r_i(t))'_{i=\overline{1,n}}.$$

Для решения задачи (29) имеет место следующая оценка:

$$\|z(t)\|_1 \leqslant \int_0^t \|r(\tau)\|_1 \exp \int_\tau^t \max_{1 \leqslant i \leqslant n}\left(b_{ii}(s) + \sum_{\substack{j=1 \\ i \neq j}}^n |b_{ij}(s)|\, ds\right) d\tau. \quad (32)$$

Если h удовлетворяет неравенству

$$h \leqslant \frac{2c}{\max\limits_{\Omega_n} |d(t)|}, \quad \text{то} \quad b_{ij} \geqslant 0 \quad (i, j = \overline{1, n}, \ i \neq j). \quad (33)$$

Учитывая, что $b_{ii}(t) \leqslant q_i(t)$, имеем:

$$\max_{1 \leqslant i \leqslant n}\left(b_{ii}(t) + \sum_{\substack{j=1 \\ i \neq j}}^n |b_{ij}(t)|\right) = \max_{i=\overline{1,n}} \sum_{j=1}^n b_{ij} \leqslant \max_{1 \leqslant i \leqslant n} q_j(t) \leqslant q,$$

$$\| z(t) \|_1 \leqslant \int_0^t \| r(\tau) \|_1 e^{q(t-\tau)} d\tau = h^2 \frac{\varkappa}{q} (e^{qt} - 1) \text{ при } q \neq 0, \tag{34}$$

$$\| z(t) \|_1 \leqslant h^2 \varkappa t \quad \text{при} \quad q = 0. \tag{35}$$

Построение аппроксимирующей задачи метода прямых для уравнений гиперболического типа

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(c(x) \frac{\partial u}{\partial x}\right) - q(x) u(x) + f(x, t) \tag{36}$$

$$(c(x) \geqslant c_0 > 0, \quad 0 < x < a, \quad 0 < t < T),$$

$$u(x, 0) = \gamma_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = \gamma_1(x), \quad 0 \leqslant x \leqslant a,$$

$$u(0, t) = \gamma_2(t), \quad u(a, t) = \gamma_3(t), \quad 0 \leqslant t < T$$

проводится аналогично рассмотренным выше примерам для уравнений эллиптического и параболического типов.

Например, продольный вариант указанного метода приведет к системе обыкновенных дифференциальных уравнений с постоянными коэффициентами:

$$\frac{d^2 v_j}{dt^2} = \frac{c_j (v_{j+1}(t) - v_j(t)) - c_{j-1}(v_j(t) - v_{j-1}(t))}{h^2} - q_j v_j(t) = f_j(t),$$

$$v_j(0) = \gamma_0(x_j), \quad \frac{dv_j(0)}{dt} = \gamma_1(x_j), \tag{37}$$

$$v_0(t) = \gamma_2(t), \quad v_{n+1}(t) = \gamma_3(t).$$

Здесь $c_j = c(x_j)$.

Не останавливаясь на доказательстве сходимости построенной дифференциально-разностной схемы (см. [1], [39]), отметим, что решение задачи (37) равномерно сходится к решению задачи (36) со скоростью $O(h)$.

Для построения схем, обладающих более высоким порядком аппроксимации, можно применить те же приемы, которые используются и в методе сеток.

Например, рассмотрим уравнение колебаний струны

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < a, \quad 0 < t < T,$$

$$u(0, x) = \gamma_0(x), \quad \frac{\partial u(0, x)}{\partial t} = \gamma_1(x), \quad 0 \leqslant x \leqslant a, \tag{38}$$

$$u(0, t) = \gamma_2(t), \quad u(a, t) = \gamma_3(t), \quad 0 \leqslant t < T.$$

Предполагая достаточную гладкость решения задачи (38), из разложения в ряд Тейлора в окрестности точки (t, x_i) функции $u(x, t)$ и ее вторых производных имеем:

$$u(x_{i+1}, t) - 2u(x_i, t) + u(x_{i-1}, t) = h^2 \frac{\partial^2 u(t, x_i)}{\partial x^2} + \\ + \frac{h^4}{12} \frac{\partial^4 u(t, x_i)}{\partial x^4} + O(h^6), \tag{39}$$

$$\frac{\partial^2 u(x_{i+1}, t)}{\partial x^2} - 2 \frac{\partial^2 u(x_i, t)}{\partial x^2} + \frac{\partial^2 u(x_{i-1}, t)}{\partial x^2} = h^2 \frac{\partial^4 u(x_i, t)}{\partial x^4} + O(h^4). \tag{40}$$

Подставляя в (39) $\frac{\partial^4 u(x_i, t)}{\partial x^4}$ из (40) и заменяя $\frac{\partial^2 u(x, t)}{\partial x^2}$ по формулам

$$\frac{\partial^2 u(x, t)}{\partial x^2} = \frac{\partial^2 u(x, t)}{\partial t^2} - f(x, t), \tag{41}$$

получим следующую систему дифференциальных уравнений:

$$\frac{5}{6}\frac{d^2 u(x, t)}{dt^2} + \frac{1}{12}\left[\frac{d^2 u(x_{i+1}, t)}{dt^2} + \frac{d^2 u(x_{i-1}, t)}{dt^2}\right] -$$
$$- \frac{1}{h^2}[u(x_{i+1}, t) - 2u(x_i, t) + u(x_{i-1}, t)] =$$
$$= \frac{5}{6} f(x_i, t) + \frac{1}{12}[f(x_{i+1}, t) + f(x_{i-1}, t)] + O(h^4).$$

Система уравнений метода прямых, имеющая аппроксимацию порядка $O(h^4)$, приводит к решению смешанной граничной задачи для обыкновенных дифференциальных уравнений второго порядка:

$$v_0(t) = \gamma_2(t), \quad v_{n+1}(t) = \gamma_3(t),$$
$$\frac{5}{6}\frac{d^2 v_i(t)}{dt^2} + \frac{1}{12}\left[\frac{d^2 v_{i+1}(t)}{dt^2} + \frac{d^2 v_{i-1}(t)}{dt^2}\right] - \frac{1}{h^2}[v_{i+1}(t) -$$
$$- 2v_i(t) + v_{i-1}(t)] = \frac{5}{6} f_i(t) + \frac{1}{12}[f_{i+1}(t) + f_{i-1}(t)], \quad i = \overline{1, n}, \tag{42}$$
$$v_k(0) = \gamma_{0k}, \quad \frac{dv_k(0)}{dt} = \gamma_{1k}, \quad k = \overline{1, n}.$$

Метод прямых, как правило, вносит меньшую аппроксимирующую погрешность по сравнению с методом сеток, но требует больших затрат вычислительного труда для доведения результата до числа.

2. Метод интегральных соотношений

Основную идею метода интегральных соотношений изложим на примере системы дифференциальных уравнений в частных производных, записанной в виде (см. [12], [26])

$$\frac{\partial}{\partial x} P_i(x, y, u_1, \ldots, u_k) + \frac{\partial}{\partial y} Q_i(x, y, u_1, \ldots, u_k) =$$
$$= F_i(x, y, u_1, u_2, \ldots, u_k) \quad (i = \overline{1, k}). \tag{43}$$

Здесь $u_1, u_2, \ldots, u_k$ — искомые функции.

При решении методом интегральных соотношений исходную систему можно записать в любой системе координат, однако желательно пользоваться такой системой координат, чтобы некоторая часть контура области $\overline{\Omega}$, в которой рассматривается решение задачи, совпадал с координатной линией.

Для определенности рассмотрим решение уравнения (1) в двумерной области $\overline{\Omega}$, представляющей собой криволинейный четырехугольник с границами

$$x = 0, \quad x = a, \quad y = 0, \quad y = \eta(x) > 0, \tag{44}$$

и предположим, что на границах $x = 0$, $x = a$ задано k условий и k условий задано на границах $y = 0$, $y = \eta(x)$. Если граница $\eta(x)$ заранее не известна, то требуется еще одно дополнительное условие.

Каждое из уравнений (43) умножим на некоторую сглаживающую функцию $\rho_i(y)$, $i = \overline{1, m}$ и проинтегрируем по переменной y от $y = 0$ до $y = \eta(x)$.

$$\int_0^{\eta(x)} \rho_i(y) \frac{\partial P_i}{\partial x} dy + \int_0^{\eta(x)} \rho_i(y) \frac{\partial Q_i}{\partial y} dy = \int_0^{\eta(x)} \rho_i(y) F_i dy. \tag{45}$$

Первый интеграл в левой части (45) дифференцируем по параметру, а второй — интегрируем по частям, тогда

$$\frac{d}{dx} \int_0^{\eta(x)} \rho_i(y) P_i dy - \eta'(x) \rho_i(\eta) P_i(\eta) + Q_i(\eta) \rho_i(\eta) - Q_i(0) \rho_i(0) -$$

$$- \int_0^{\eta(x)} \rho_i'(y) Q_i dy = \int_0^{\eta(x)} \rho_i(y) F_i dy \quad (i = \overline{1, k}). \tag{46}$$

Здесь приняты следующие обозначения:

$$\begin{gathered} \rho_i(\eta) = \rho_i(y)|_{y=\eta(x)}, \quad \rho_i(0) = \rho_i(y)|_{y=0}, \\ Q_i(\eta) = Q_i(x, y, u_1(x, y), \ldots, u_k(x, y))|_{y=\eta(x)}, \\ Q_i(0) = Q_i(x, y, u_1(x, y), \ldots, u_k(x, y))|_{y=0}. \end{gathered} \tag{47}$$

Аналогичные обозначения приняты для $P_i(\eta)$.

В дальнейшем область интегрирования разбивается на криволинейные полосы при помощи кривых

$$y_j = \frac{1}{n} \eta(x) \quad (j = \overline{0, n}), \tag{48}$$

и функции P_i, Q_i, F_i заменяются обобщенными интерполяционными полиномами по некоторой системе базисных функций $\varphi_j(y)$, т. е.

$$\begin{gathered} P_i(x, y, u_1, \ldots, u_k) = \sum_{j=0}^{n} \varphi_j(y) P_{ij} + R_{in}^{(1)}(P), \\ Q_i(x, y, u_1, \ldots, u_k) = \sum_{j=0}^{n} \varphi_j(y) Q_{ij} + R_{in}^{(2)}(Q), \\ F_i(x, y, u_1, \ldots, u_k) = \sum_{j=0}^{n} \varphi_j(y) F_{ij} + R_{in}^{(3)}(F). \end{gathered} \tag{49}$$

Здесь

$$P_{ij} = P_i(x, y_j(x), u_1(x, y_j(x)), \ldots, u_k(x, y_j(x)),$$

а $R_{in}^{(1)}(P)$ — остаток интерполирования; для Q_{ij}, F_{ij}, $R_{in}^{(2)}(Q)$, $R_{in}^{(3)}(F)$ приняты аналогичные обозначения.

Для каждого i введем в рассмотрение систему линейно-независимых сглаживающих функций

$$(\rho_{ij}(y))_{i=\overline{1,k}}^{j=\overline{1,n}} = (\rho_{i1}(y), \rho_{i2}(y), \ldots, \rho_{in}(y))_{i=\overline{1,k}}.$$

Вообще система $\rho_{ij}(y)$ может и не зависеть от номера i, т. е. для всех значений i эти функции могут совпадать. Функции $\rho_{ij}(y)$ выбираются так, чтобы была обеспечена сходимость всех интегралов в интегральных соотношениях (46).

Уравнение (46) запишем для каждой из $k \times n$ функций ρ_{ij} и в полученную систему подставим вместо P_i, Q_i, F_i их выражения по формуле (49), отбрасывая остаточные члены интерполяционных формул:

$$\frac{d}{dx}\sum_{j=0}^{n} P_{ij}\int_0^{\eta(x)} \rho_{ij}(y)\,\varphi_j(y)\,dy - \eta'(x)\,\rho_{ij}(\eta)\,P_{in} +$$

$$+\rho_{ij}(\eta)\,Q_{in} - \rho_{ij}(0)\,Q_{i0} - \sum_{i=0}^{n} Q_{ij}\int_0^{\eta(x)} \rho'_{ij}(y)\,\varphi_j(y)\,dy =$$

$$= \sum_{j=0}^{n} F_{ij}\int_0^{\eta(x)} \varphi_j(y)\,\rho_{ij}(y)\,dy \quad (i=\overline{1,k};\quad j=\overline{1,n}). \tag{50}$$

К полученной системе обыкновенных уравнений (50) нужно добавить условия, заданные на границе области Ω.

Заметим, что если в качестве $\rho_{ij}(y)$ выбрать «ступенчатую» функцию, т. е. пусть

$$\rho_{ij}(y) = \begin{cases} 0, & y < y_{j-1}, \\ 1, & y_{j-1} \leqslant y \leqslant y_j, \\ 0, & y_j < y, \end{cases} \tag{51}$$

то система (46) запишется в виде

$$\frac{d}{dx}\int_{\eta_{j-1}x}^{\eta_j(x)} P_i dy - \eta'_j(x)\,P_{ij} + \eta'_{j-1}P_{i,j-1} + Q_{ij} = \int_{\eta_{j-1}}^{\eta_j(x)} F_i dy \tag{52}$$

$$(i=\overline{1,k},\ j=\overline{1,n}).$$

В методе интегральных соотношений аппроксимируется интеграл, что способствует повышению точности аппроксимации. Кроме того, введение весовых функций $\rho_{ij}(y)$ дает возможность их подобрать таким образом, чтобы допускалось непрерывное представление интеграла и в том случае, когда подынтегральная функция имеет разрыв первого рода.

Вопросы сходимости метода интегральных соотношений исследовались для ряда линейных задач математической физики. В частности, рассмотренный выше метод прямых является частным случаем метода интегральных соотношений. В случае нелинейной системы дифференциальных уравнений метод интегральных соотношений приводит к необходимости решения краевых задач или задач Коши для системы обыкновенных дифференциальных уравнений. Практическая сходимость метода интегральных соотношений обычно проверяется по убыванию разности решений в двух последовательных приближениях.

Глава 7

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Приближенные методы решения краевых задач, рассмотренные в предыдущей главе, используются при решении линейных уравнений математической физики. Большая же часть задач, с которыми приходится встречаться на практике, связана с нелинейными уравнениями. Теория разностных схем для линейных уравнений имеет много общих приемов с методами, применяемыми для построения приближенных решений нелинейных задач математической физики. Наиболее законченные результаты в этом направлении относятся к нелинейным задачам, связанным с обыкновенными дифференциальными уравнениями.

§ 1. ЗАДАЧА КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Пусть на некотором отрезке $x_0 \leqslant x \leqslant \xi$ требуется найти решение дифференциального уравнения n-го порядка

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}), \tag{1}$$

которое в точке $x = x_0$ принимает заданные начальные значения

$$y^{(l)}(x_0) = y_0^l \quad (l = \overline{0, n-1}). \tag{2}$$

Предполагается, что существует единственное решение $y(x)$ задачи Коши для уравнения (1) на отрезке $[x_0, \xi]$. Последнее будет иметь место, если функция f непрерывна в Ω по всем аргументам и удовлетворяет условию Липшица по переменным $y, y', \ldots, y^{(n-1)}$.

Задачу Коши для дифференциального уравнения n-го порядка вида (1) можно свести к эквивалентной ему нормальной системе вида

$$\frac{dY}{dx} = F(x, Y), \quad Y(x_0) = Y_0, \tag{3}$$

если положить $y'(x) = y_1(x)$, $y''(x) = y_2(x)$, $\ldots$, $y^{(n-1)}(x) = y_{n-1}(x)$, $y^{(n)}(x) = f(x, y(x), y_1(x), y_2(x), \ldots, y_{n-1}(x))$ и ввести обозначения

$$\begin{gathered} Y(x) = (y(x), y_1(x), \ldots, y_{n-1}(x))', \\ F(x, Y) = (y_1(x), y_2(x), \ldots, y_{n-1}(x), f(x, y(x), \ldots, y_{n-1}(x)))', \\ Y_0 = (y_0, y_0', y_0'', \ldots, y_0^{(n-1)})'. \end{gathered} \tag{4}$$

Задача Коши для системы n уравнений первого порядка

$$\begin{gathered} y_l' = f_l(x, y_1(x), y_2(x), \ldots, y_n(x)), \\ y_l(x_0) = y_l^0, \quad l = \overline{1, n} \end{gathered} \tag{5}$$

также может быть записана в виде (3), если ввести следующие обозначения:

$$Y(x) = (y_1(x), y_2(x), \ldots, y_n(x))',$$

$$F(x, Y) = (f_1, f_2, \ldots, f_n)', \quad f_l = f_l(x, Y(x)), \tag{6}$$
$$Y^0 = (y_1^0, y_2^0, \ldots, y_n^0).$$

Большинство приближенных методов, применяемых при решении задачи Коши для уравнения первого порядка

$$\begin{aligned} y' &= f(x, y), \\ y(x_0) &= y_0, \end{aligned} \tag{7}$$

без всяких изменений переносится на случай систем уравнений вида (3). Поэтому вначале остановимся на методах решения простейшей задачи Коши вида (7) на отрезке $[x_0, \xi]$.

1. Разложение решения в ряд Тейлора

Рассмотрим задачу Коши вида (7) и предположим, что $f(x, y)$ является аналитической функцией в точке x_0, y_0, т. е. в некоторой окрестности этой точки разлагается в степенной ряд вида

$$f(x, y) = \sum_{\alpha_0\alpha_1} C_{\alpha_0\alpha_1} (x - x_0)^{\alpha_0} (y - y_0)^{\alpha_1},$$

где α_i $(i = \overline{0, 1})$ — целые неотрицательные числа, $C_{\alpha_0\alpha_1}$ — постоянные коэффициенты. Тогда интеграл уравнения (7) с начальными условиями $y(x_0) = y_0$ является аналитическим в точке x_0. Пользуясь рядом Тейлора, можно написать приближенное равенство

$$y(x) \approx \sum_{k=0}^{n} \frac{y^{(k)}(x_0)}{k!} (x - x_0)^k \tag{8}$$

для $|x - x_0| < \rho$ — радиуса сходимости ряда Тейлора, так как в этом случае погрешность формулы (8) будет стремиться к нулю при $n \to \infty$.

Для определения $y^{(k)}(x_0)$ дифференцируем по x соотношение (7)

$$\begin{aligned} y''(x_0) &= f_x(x_0, y_0) + f_y(x_0, y_0) f(x_0, y_0), \\ y'''(x_0) &= f_{xx}(x_0, y_0) + 2f_{xy}(x_0, y_0) f(x_0, y_0) + \\ &+ f_{yy}(x_0, y_0)(f(x_0, y_0))^2 + f_y(x_0, y_0) y''(x_0), \ \ldots \end{aligned} \tag{9}$$

Метод разложения в ряд Тейлора требует вычисления значения $f(x, y)$ и производных $f_{x^l y^{l-l}}$ для $l < n$, а поэтому применение этого метода можно считать целесообразным для случая решения одного и того же дифференциального уравнения вида (7) при различных начальных условиях с использованием программ, по которым осуществляется формальное дифференцирование функций.

Грубая оценка погрешности метода разложения решения в ряд Тейлора может быть получена, если воспользоваться остаточным членом разложения решения вспомогательного дифференциального уравнения, правая часть которого представляет собой оценку неучтенных в приближенном решении (8) степеней разложения в ряд Тейлора.

В случае системы дифференциальных уравнений вида (3), (6) решение задачи Коши имеет вид

$$Y(x) = \sum_{k=0}^{N} \frac{Y^{(k)}(x_0)}{k!}(x - x_0)^k. \tag{10}$$

Векторное соотношение (10) означает, что каждая компонента вектора $Y(x)$ разлагается в соответствующий ряд Тейлора.

Очевидно,

$$Y'(x_0) = F(x_0, Y_0),$$
$$Y''(x) = F_x + F_y Y' = F_x + F_y F,$$

где

$$F_y = \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \dots & \frac{\partial f_1}{\partial y_n} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \dots & \frac{\partial f_2}{\partial y_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_n}{\partial y_1} & \frac{\partial f_n}{\partial y_2} & \dots & \frac{\partial f_n}{\partial y_n} \end{pmatrix}, \quad F_x = \begin{pmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \\ \dots \\ \frac{\partial f_n}{\partial x} \end{pmatrix},$$

$$Y'''(x) = F_{xx} + 2F_{xy}Y' + F_{yy}Y'Y' + F_y Y'',$$

где

$$F_{x^l y^k} = \frac{\partial^{l+k} f_i}{\partial x^l \partial y_{j_1} \dots \partial y_{j_k}} \text{ — тензор.}$$

При использовании методов степенных рядов на отрезке $[x_0, \xi]$ иногда оказывается целесообразным разбить этот отрезок точками x_j $(j = \overline{0, n})$ и искать решения по формуле (8) соответственно на отрезках

$$[x_0, x_1], \quad [x_1, x_2], \quad \dots, \quad [x_{n-1}, \xi],$$

принимая за начальное значение на отрезке $[x_j, x_{j+1}]$ значение $y(x_j)$, найденное при применении метода разложения в ряд Тейлора на предыдущем отрезке $[x_{j-1}, x_j]$. В этом случае метод рядов Тейлора будет принадлежать к числу *одношаговых методов решения задачи Коши,* т. е. таких, которые позволяют найти приближенное решение задачи в узле x_{j+1} по информации об этом решении в одной предыдущей узловой точке.

Наиболее часто употребляемыми одношаговыми разностными схемами являются схемы Рунге — Кутта.

2. Одношаговые методы типа Рунге — Кутта

Рассмотрим задачу Коши

$$y' = f(x, y), \quad y(x_0) = y_0, \quad (x_0, y_0) \in \Omega.$$

Точное решение задачи (7) обозначим через $y(x)$. При достаточной гладкости функции $f(x, y)$, определенной в области Ω, имеет место

разложение ($x_1 = x_0 + h$, $h > 0$)

$$y(x_1) - y(x_0) = \sum_{k=1}^{s} \frac{h^k}{k!} y^{(k)}(x_0) + 0(h^{s+1}). \tag{11}$$

Принимая во внимание соотношения (9), приходим к выводу, что коэффициенты при степенях h в правой части равенства (11), исключая остаточный член, выражаются через значения функции f и ее производных в точке (x_0, y_0).

Рассмотрим линейную комбинацию значений функции f в точках (ξ_i, η_i) $(i = \overline{1, r})$:

$$\sum_{k=1}^{r} P_k h f(\xi_k, \eta_k), \tag{12}$$

где

$$\xi_i = x_0 + \alpha_i h, \quad \alpha_1 = 0, \tag{13}$$

$$\eta_i = y_0 + \beta_{i1} k_1 + \beta_{i2} k_2 + \ldots + \beta_{i,i-1} k_{i-1}, \tag{14}$$

$$k_i(h) = hf(\xi_i, \eta_i), \tag{15}$$

P_i, α_i, β_{ij} $(0 < j < i \leqslant r)$ — постоянные коэффициенты.

Очевидно, если каждое из выражений (15) разложить по формуле Тейлора в окрестности точки (x_0, y_0), то линейная комбинация (12) запишется в виде полинома от h, коэффициенты которого будут выражаться через значения функции f, ее производные, взятые в точке (x_0, y_0), и неопределенные коэффициенты P_i, α_i, β_{ij} с прибавлением некоторого остаточного члена. Поэтому представим правую часть равенства (11) в виде линейной комбинации функций вида (15)

$$y(x_0 + h) - y(x_0) \approx \sum_{i=1}^{r} P_i k_i(h) \tag{16}$$

или

$$y_1 - y_0 = \sum_{i=1}^{r} P_i k_i(h), \tag{17}$$

где через y_1 обозначено приближенное значение искомого решения в точке $x_1 = x_0 + h$.

Введем в рассмотрение функцию

$$R(h) = y(x_0 + h) - y_1 = y(x_1) - y(x_0) - \sum_{i=1}^{r} P_i k_i(h). \tag{18}$$

Согласно формуле Тейлора имеем:

$$R(h) = \sum_{k=0}^{s} \frac{h^k R^k(0)}{k!} + h^{s+1} \frac{R^{(s+1)}(\theta h)}{(s+1)!}. \tag{19}$$

Подберем коэффициенты P_i, α_i, β_{ij} таким образом, чтобы

$$R(0) = R'(0) = \cdots = R^{(s)}(0) = 0 \tag{20}$$

до возможно более высокого порядка s.

Тогда величина

$$R(h) = \frac{h^{s+1} R^{(s+1)}(\theta h)}{(s+1)!} \tag{21}$$

будет называться погрешностью метода Рунге — Кутта на шаге, s — порядком (степенью) точности формул Рунге — Кутта, а формулы вида (17) — формулами метода Рунге — Кутта решения задачи Коши.

Предположим, что главная (относительно h) часть погрешности (21) формулы (17) при выбранных значениях параметров P_i, α_i, β_{ij} $(0 < j < i \leqslant r)$ может быть представлена в виде

$$R(h) = \gamma h^{s+1} \Psi [f]_0 + O(h^{s+2}), \tag{22}$$

где γ — некоторый параметр; $\Psi [f]_0$ — вполне определенный оператор, зависящий от f и вычисляемый в точке (x_0, y_0). Формула (22) будет иметь место, если f имеет непрерывные производные вплоть до $(s+1)$-го порядка. Так как

$$R(h) = y(x_1) - y_1,$$

то формулы вида (17) для двух отличающихся только знаком значений параметра γ $(\gamma \neq 0)$ составят формулы, которые будут давать верхние и нижние приближения к искомому решению $y(x_1)$ (если $\Psi [f]_0 \neq 0$). Таким образом, если главный член погрешности формул метода Рунге — Кутта представить в виде (22), то формулы вида (17), отвечающие значениям $\pm \gamma$, составят формулы так называемого двустороннего метода Рунге — Кутта.

Примеры построения формул метода Рунге — Кутта.

а) Пусть $r = 1$. Тогда, согласно (13) — (15), (17), (19), (20), имеем:

$$y_1 = y_0 - P_1 k_1(h),$$

$$k_1 = hf(x_0, y_0),$$

$$R(h) = y(x_1) - y(x_0) - P_1 hf(x_0, y_0),$$

$$R'(0) = (y'(x_0 + h) - P_1 f(x_0, y_0))|_{h=0} = (1 - P_1) f(x_0, y_0),$$

$$R''(0) = y''(x_0),$$

причем $R(0) = 0$, $R'(0) = 0$ при любых значениях $f(x, y)$, если $P_1 = 1$. Формула вида

$$y_1 = y_0 - hf(x_0, y_0) \tag{23}$$

будет иметь первый порядок точности, так как

$$R(h) = \frac{h^2}{2!} y''(x_0 + \theta h), \quad |\theta| < 1. \tag{24}$$

Метод решения задачи Коши (7) по формулам вида (23) соответствует приближенным формулам метода Эйлера.

Таблица 5

Односторонние формулы Рунге — Кутта вида $y_1 = y_0 + \sum_{i=1}^{r} P_i k_i,\ k_i = hf(x_0 + \alpha_i h,\ y_0 + \beta_{i1} k_1 + \ldots + \beta_{i,\,i-1}\ k_{i-1})$

решения задачи Коши для уравнения $y' = f(x, y)$

r	№ п/п	i	P_i	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}	$x_i = x_0 + \alpha_i h$	$y_i = y_0 + \beta_{i1} k_1 + \ldots + \beta_{i,i-1}\ k_{i-1}$	$k_i = hf(x_i, y_i)$	Вид формулы	Порядок погрешности на шаге
1	1	1	1	0						x_0	y_0	k_1	$y_1 = y_0 + k_1$	1
2	2	1	0	0						x_0	y_0	k_1		2
		2	1	1/2	1/2					$x_0 + \frac{1}{2}h$	$y_0 + \frac{1}{2}k_1$	k_2	$y_1 = y_0 + k_2$	
	3	1	1/2	0						x_0	y_0	k_1		2
		2	1/2	1	1					$x_0 + h$	$y_0 + k_1$	k_2	$y_1 = y_0 + \frac{1}{2}(k_1 + k_2)$	
	4	1	1/4	0						x_0	y_0	k_1		2
		2	3/4	2/3	2/3					$x_0 + \frac{2}{3}h$	$y_0 + \frac{2}{3}k_1$	k_2	$y_1 = y_0 + \frac{1}{4}(k_1 + 3k_2)$	
3	5	1	1/6	0						x_0	y_0	k_1		3
		2	2/3	1/2	1/2					$x_0 + \frac{1}{2}h$	$y_0 + \frac{1}{2}k_1$	k_2	$y_1 = y_0 + \frac{1}{6} \times$	
		3	1/6	1	—1	2				$x_0 + h$	$y_0 - k_1 + 2k_2$	k_3	$\times (k_1 + 4k_2 + k_3)$	

	6	1	1/4	0						x_0	y_0	k_1		
		2	0	1/3	1/3					$x_0+\frac{1}{3}h$	$y_0+\frac{1}{3}k_1$	k_2	$y_1=y_0+\frac{1}{4}(k_1+3k_3)$	3
		3	3/4	2/3	0	2/3				$x_0+\frac{2}{3}h$	$y_0+\frac{2}{3}k_2$	k_3		
	7	1	2/9							x_0	y_0	k_1		3
		2	1/3	1/2	1/2					$x_0+\frac{1}{2}h$	$y_0+\frac{1}{2}k_1$	k_2	$y_1=y_0+\frac{1}{9}\times$	
		3	4/9	3/4	0	3/4				$x_0+\frac{3}{4}h$	$y_0+\frac{3}{4}k_2$	k_3	$\times(2k_1+3k_2+4k_3)$	
4	8	1	1/6	0						x_0	y_0	k_1		
		2	1/3	1/2	1/2					$x_0+\frac{h}{2}$	$y_0+\frac{k_2}{2}$	k_2	$y_1=y_0+\frac{1}{6}\times$	4
		3	1/3	1/2	0	1/2				$x_0+\frac{h}{2}$	$y_0+\frac{k_2}{2}$	k_3	$\times(k_1+2k_2+2k_3+k_4)$	
		4	1/6	1	0	0	1			x_0+h	y_0+k_3	k_4		
	9	1	1/6	0						x_0	y_0	k_1		
		2	0	1/4	1/4					$x_0+\frac{h}{4}$	$y_0+\frac{k_1}{4}$	k_2	$y_1=y_0+\frac{1}{6}\times$	4
		3	2/3	1/2	0	1/2				$x_0+\frac{h}{2}$	$y_0+\frac{k_2}{2}$	k_3	$\times(k_1+4k_3+k_4)$	
		4	1/6	1	1	—2	2			x_0+h	$y_0+k_1-2k_2+k_3$	k_4		

Продолжение табл. 5

r	№ п/п	i	P_i	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}	$x_i = x_0 + \alpha_i h$	$y_i = y_0 + \beta_{i1} k_1 + \ldots + \beta_{i,\, i-1} k_{i-1}$	$k_i = hf(x_i, y_i)$	Вид формулы	Порядок погрешности на шаге
	10	1	1/8	0						x_0	y_0	k_1		
		2	3/8	1/3	1/3					$x_0 + \frac{h}{3}$	$y_0 + \frac{1}{3}k_1$	k_2	$y_1 = y_0 + \frac{1}{8} \times$	4
		3	1/8	2/3	—1/3	1				$x_0 + \frac{2}{3}h$	$y_0 - \frac{1}{3}k_1 + k_2$	k_3	$\times (k_1 + 3k_2 + 3k_3 + k_4)$	
		4	1/8	1	1	—1	1			$x_0 + h$	$y_0 + k_1 - k_2 + k_3$	k_4		
6	11	1	23/192	0						x_0	y_0	k_1		
		2	0	1/3	1/3					$x_0 + \frac{1}{3}h$	$y_0 + \frac{k_1}{3}$	k_2		5
		3	125/192	2/5	4/25	6/25				$x_0 + \frac{2}{5}h$	$y_0 + \frac{4}{25}k_1 + \frac{6}{25}k_2$	k_3		
		4	0	1	1/4	—3	15/4			$x_0 + h$	$y_0 + \frac{1}{4}k_1 - 3k_2 + \frac{15}{4}k_3$	k_4		
		5	—81/192	2/3	2/17	10/9	—50/81	8/81		$x_0 + \frac{2}{3}h$	$y_0 + \frac{2}{17}k_1 + \frac{10}{9}k_2 - \frac{50}{81}k_3 + \frac{8}{84}k_4$	k_5	$y_1 = y_0 + \frac{1}{192} \times$ $\times (23k_1 + 125k_3 - 81k_5 + 125k_6)$	
		6	125/192	4/5	2/25	12/25	2/15	8/75	0	$x_0 + \frac{4}{5}h$	$y_0 + \frac{2}{25}k_1 + \frac{12}{25}k_2 + \frac{2}{15}k_3 + \frac{8}{75}k_4$	k_6		

б) Пусть $r = 2$

$$y_1 = y_0 - P_1 k_1(h) - P_2 k_2(h), \tag{25}$$

$$k_1(h) = hf(x_0, y_0) = hf_0, \quad k_2(h) = hf(x_0 + \alpha_2 h, y_0 + \beta_{21} k_1) = hf(x^*, y^*), \quad x^* = x_0 + \alpha_2 h, \quad y^* = y_0 + \beta_{21} hf(x_0, y_0);$$

$$R(h) = y(x_1) - y(x_0) - P_1 k_1(h) - P_2 k_2(h);$$
$$R(0) = 0; \tag{26}$$

$$R'(0) = \{y'(x_0 + h) - P_1 f_0 - P_2 f(x^*, y^*) - P_2 h[\alpha_2 f_x(x^*, y^*) + \beta_{21} f_y(x^*, y^*) y'(x_0)]\}_{h=0} = (1 - P_1 - P_2) f_0; \tag{27}$$

$$R''(0) = \{y''(x_0 + h) - 2P_2[\alpha_2 f_x(x^*, y^*) + \beta_{21} f_y(x^*, y^*) y'(x_0)] - P_2 h[\alpha_2^2 f_{xx}(x^*, y^*) + 2\alpha_2\beta_{21} f_{xy}(x^*, y^*) y'(x_0) + \beta_{21} f_{yy}(x^*, y^*) y'^2(x_0)]\}_{h=0} = [(1 - 2P_2\alpha_2) f_x + (1 - 2P_2\beta_{21}) f_y y']_0; \tag{28}$$

$$R'''(0) = \{f_{xx}(x_1) + 2f_{xy}(x_1) f(x_1) + f_{yy}(x_1) f^2(x_1) + f_y(x_1) y''(x_1) - 3P_2[\alpha_2^2 f_{xx}(x^*, y^*) + 2\alpha_2\beta_{21} f_{xy}(x^*, y^*) y'(x_0) + \beta_{21}^2 f_{yy}(x^*, y^*)(y'(x_0))^2 + O(h)]\}_{h=0} = [(1 - 3P_2\alpha_2^2) f_{xx} + (2 - 6P_2\alpha_2\beta_{21}) f_{xy} f + (1 - 3P_2\beta_{21}^2) f_{yy} f^2 + f_y(f_x + f_y f)]_0. \tag{29}$$

Из равенств (26) — (28) следует, что при любой допустимой функции $f(x, y)$, $R(0) = 0$, $R'(0) = 0$, $R''(0) = 0$ при условии $P_2 \neq 0$

$$1 - p_1 - p_2 = 0, \tag{30}$$

$$1 - 2p_2\alpha_2 = 0, \tag{31}$$

$$1 - 2p_2\beta_{21} = 0. \tag{32}$$

Для определения параметров p_1, p_2, α_{21}, β_{21} из системы (30) — (32) следует положить $\alpha_2 = \beta_{21}$. Тогда система уравнений

$$1 - p_1 - p_2 = 0,$$
$$1 - 2p_2\alpha_2 = 0$$

будет определять однопараметрическое семейство формул Рунге — Кутта второго порядка точности. Задавая произвольно один из параметров, например p_1, можно построить различные формулы метода Рунге — Кутта вида (25). Пусть $p_1 = 0$, тогда $p_2 = 1$, $\alpha_2 = \beta_{21} = \frac{1}{2}$ и формула метода Рунге — Кутта второго порядка точности запишется в виде

$$y_1 = y_0 + hf\left(x_0 + \frac{h}{2}, \; y_0 + \frac{k_1}{2}\right);$$

$$k_1 = hf(x_0, y_0), \quad R(h) = \frac{h^3}{3!}\left[\frac{1}{4}(f_{xx} + 2f_{xy} + f_{yy} f^2) + f_y(f_x + f_y f)\right]_0 + O(h^4).$$

Примеры формул Рунге — Кутта второго порядка точности на шаге приведены в табл. 5.

Погрешность шага односторонних формул Рунге — Кутта второго порядка точности в соответствии с (29) будет иметь вид

$$R(h) = \frac{h^3}{3}[(1 - 3p_2\alpha_2^2) f_{xx} + (2 - 6p_2\alpha_2\beta_{21}) f_{xy} f + (1 - 3p_2\beta_{21}^2) f_{yy} f^2 + f_y(f_x + f_y f)]_0 + O(h^4). \tag{33}$$

Легко видеть, что если параметры p_2, α_2, β_{21} удовлетворяют условиям

$$\beta_{21} = \alpha_2 = \frac{2}{3}; \quad p_2 = \frac{3}{4}; \quad p_1 = \frac{1}{4} \tag{34}$$

то на классе функции f, для которых

$$f_x f_y + f f_y^2 \equiv 0, \tag{35}$$

выражение, стоящее в квадратных скобках соотношения (33), обратится в нуль и $R(h) = O(h^4)$. Таким образом, формула Рунге — Кутта вида

$$y_1 = y_0 + \frac{1}{4} hf(x_0, y_0) + \frac{3}{4} hf\left(x_0 + \frac{2}{3}h, \quad y_0 + \frac{2}{3} hf_0\right) \tag{36}$$

на классе функций f, удовлетворяющих условию (35), будет иметь порядок точности $s = 3$. К классу функций f, удовлетворяющих условию (35), принадлежат все функции f, не зависящие от y. Это означает, что формулу (36) можно использовать для вычисления квадратур. Ее можно использовать также, если значения величин $f_y y''$ малы, так как погрешность формулы (36) определяется соотношением

$$R(h) = \frac{h^3}{3!} (f_y y'')_0 + O(h^4). \tag{37}$$

Двусторонние формулы типа Рунге — Кутта при $r = 2$ можно построить, если предположить, что $R''(0) \neq 0$, где $R''(0)$ определяется по формуле (28).

Считая, что $\alpha_2 = \beta_{21}$ и вводя параметр γ

$$\gamma = \frac{1}{2}(1 - 2p_2\alpha_2) = \frac{1}{2} - p_2\alpha_2, \tag{38}$$

получим, что погрешность шага $R(h)$ формул Рунге — Кутта вида (25) при условии

$$1 - p_1 - p_2 = 0 \tag{39}$$

будет иметь вид

$$R(h) = \gamma h^2 [f_x + f_y f]_0 + O(h^3),$$

т. е. главная (относительно h) часть погрешности содержит параметр γ.

Относительно неизвестных p_1, p_2, α_2, γ получается система двух уравнений (38), (39). Выразим α_2 и p_2 через значения параметров p_1 и γ:

$$p_2 = 1 - p_1, \tag{40}$$

$$\alpha_2 = \frac{1 - 2\gamma}{2(1 - p_1)} \quad (\text{при } 1 - p_1 \neq 0), \tag{41}$$

$$\beta_{21} = \frac{1 - 2\gamma}{2(1 - p_1)}. \tag{42}$$

Таким образом, соотношения (40) — (42), (25) определяют двупараметрическое семейство двусторонних формул типа Рунге — Кутта первого порядка точности.

Если потребовать, чтобы

$$0 \leqslant \alpha_2 \leqslant 1,$$

то на выбор параметра γ будут наложены следующие ограничения:

$$p_1 - \frac{1}{2} \leqslant \gamma \leqslant \frac{1}{2} \text{ при } 1 - p_1 > 0, \tag{43}$$

$$\frac{1}{2} \leqslant \gamma \leqslant p_1 - \frac{1}{2} \text{ при } 1 - p_1 < 0. \tag{44}$$

Примеры двусторонних формул типа Рунге — Кутта первого порядка точности приведены в табл. 6.

Очевидно, на классе функций f, удовлетворяющих условию (35) при $r = 2$, можно построить двусторонние формулы Рунге — Кутта второго порядка точности, если выбрать $\alpha_2 = \beta_{21}$, и параметр γ ввести следующим образом:

$$\gamma = \frac{1 - 3p_2\alpha_2^2}{3} = \frac{1}{3} - p_2\alpha_2^2.$$

Из (26) — (29) следует, что при любой допустимой функции $f(x, y)$

$$R = 0, \quad R'(0) = 0, \quad R''(0) = 0 \text{ и } R(h) = \frac{\gamma h^3}{2}[f_{xx} + 2f_{xy}f_0 + f_{yy}f_0]_0 + O(h^4) \quad (45)$$

при условии, что

$$1 - p_1 - p_2 = 0; \quad (46)$$

$$1 - 2p_2\alpha_2 = 0; \quad (47)$$

$$\alpha_2 = \beta_{21}; \quad (48)$$

$$\gamma = \frac{1}{3} - p_2\alpha_2^2. \quad (49)$$

Разрешим (47), (49) соответственно относительно $p_2\alpha_2^2$ и $\alpha_2^2 p_2$, а затем разделим первое равенство на второе:

$$\alpha_2 = \frac{2}{3}(1 - 3\gamma) \qquad (p_2 \neq 0, \quad \alpha_2 \neq 0).$$

Из (47) находим:

$$p_2 = \frac{3}{4(1 - 3\gamma)} \quad (1 - 3\gamma \neq 0),$$

а из (46)

$$p_1 = \frac{1 - 12\gamma}{4(1 - 3\gamma)}.$$

Если потребовать, чтобы $0 < \alpha_2 \leqslant 1$, то

$$\frac{1}{6} \leqslant \gamma < \frac{1}{3}.$$

Например, при $\gamma = \pm\frac{1}{6}$ получим следующие формулы двустороннего метода типа Рунге — Кутта второго порядка точности на классе функций f, удовлетворяющих условию (35),

$$y_1^{(1)} = y_0 - \frac{h}{2}f_0 + \frac{3}{2}hf\left(x_0 + \frac{1}{3}h, \quad y_0 + \frac{1}{3}hf_0\right); \quad R^{(1)}(h) =$$

$$= \frac{h^3}{12}[f_{xx} + 2f_{xy}f_0 + f_{yy}f_0^2]_0 + O(h^4);$$

$$y_1^{(2)} = y_0 + \frac{h}{2}f_0 + \frac{h}{2}f(x_0 + h, \quad y_0 + hf_0);$$

$$R^{(2)}(h) = -\frac{h^3}{12}[f_{xx} + 2f_{xy}f_0 + f_{yy}f_0^2]_0 + O(h^4).$$

В качестве приближенного решения задачи (7) в точке $x_1 = x_0 + h$ можно принять полусумму верхних и нижних приближений, полученных по двусторонним формулам типа Рунге — Кутта,

$$y_1 = \frac{y_1^{(1)} + y_1^{(2)}}{2}. \quad (50)$$

Двусторонние формулы Рунге-Кутта вида $y_1^{(l)} = y_0 + \sum_{i=1}^{r} P_i k_i;\ k_i = hf(x_0 + \alpha_i h,$

$y^1 = f(x, y)$

r	№ п/п	l	γ	c	P_l	α_l	β_{l1}	β_{l2}	β_{l3}	$x_i = x_0 + \alpha_i h$
2	1	1	1/2	1	0	0				x_0
				2	1	0	0			x_0
		2	—1/2	1	0	0				x_0
				2	1	1	1			$x_0 + h$
	2	1	1/4	1	0	0				x_0
				2	1	1/4	1/4			$x_0 + \frac{h}{4}$
		2	—1/4	1	0	0				x_0
				2	1	3/4	3/4			$x_0 + \frac{3}{4}h$
	3	1	1/6	1	0	0				x_0
				2	1	1/3	1/3			$x_0 + \frac{h}{3}$
		2	—1/6	1	0	0				x_0
				2	1	2/3	2/3			$x_0 + \frac{2}{3}h$
3	4	1	1	1	1/6	0				x_0
				2	1/6	1	1			$x_0 + h$
				3	2/3	1/2	7/4	—5/4		$x_0 + \frac{h}{2}$
		2	—1	1	1/6	0				x_0
				2	1/6	1	1			$x_0 + h$
				3	2/3	1/2	—5/4	7/4		$x_0 + \frac{h}{2}$
	5	1	1/24	1	1/4	0				x_0
				2	0	1/4	—1/4			$x_0 + 1/4\ h$
				3	3/4	2/3	0	2/3		$x_0 + 2/3\ h$
		2	—1/24	1	1/4	0				x_0

Таблица 6

$y_0 + \beta_{i1}\ k_1 + \ldots + \beta_{i,\ i-1}\ k_{i-1})$, $l = 1{,}2$ **решения задачи Коши для уравнения**

$y_l = y_0 = \beta_{i1} k_1 + \ldots + \beta_{i,i-1} k_{i-1}$	$k_i = hf(x_i, y_i)$	Вид формул для $y_1^{(l)}$, $y_1 = \frac{y_1^{(1)} + y_1^{(2)}}{2}$	Погрешность на шаге	Порядок погрешности на шаге
y_0 y_0	k_1 k_1	$y_1^{(1)} = y_0 + k_1$	$\frac{h^2}{2}[f_x + ff_y]_0 + O(h^3)$	1
y_0 $y_0 + k_1$	k_1 k_2	$y_1^{(2)} = y_0 + k_2$	$-\frac{h_2}{2}[f_x + ff_y]_0 + O(h^3)$	1
y_0 $y_0 + \frac{k_1}{4}$	k_1 k_2	$y_1^{(1)} = y_0 + k_2$	$\frac{h^2}{4}[f_x + ff_y]_0 + O(h^3)$	1
y_0 $y_0 + \frac{3}{4}k_1$	k_1 k_2	$y_1^{(2)} = y_0 + k_2$	$-\frac{h_2}{4}[f_x + ff_y]_0 + O(h^3)$	1
y_0 $y_0 + \frac{k_1}{3}$	k_1 k_2	$y_1^{(1)} = y_0 + k_2$	$\frac{h^2}{6}[f_x + ff_y]_0 + O(h^3)$	1
y_0 $y_0 + \frac{2}{3}k_1$	k_1 k_2	$y_1^{(2)} = y_0 + k_2$	$-\frac{h^2}{6}[f_x + ff_y]_0 + O(h^3)$	1
y_0 $y_0 + k_1$ $y_0 + \frac{7}{4}k_1 - \frac{5}{4}k_2$	k_1 k_2 k_3	$y_1^{(1)} = y_0 + \frac{1}{6} \times (k_1 + k_2 + 4k_3)$	$h^3[f_x f_y + ff_y]_0 + O(h^4)$	2
y_0 $y_0 + k_1$ $y_0 - \frac{5}{4}k_1 + \frac{7}{4}k_2$	k_1 k_2 k_3	$y_1^{(2)} = y_0 + \frac{1}{6} \times (k_1 + k_2 + 4k_3)$	$-h^3[f_x f_y + ff_y^2]_0 + O(h^4)$	2
y_0 $y_0 + \frac{1}{4}k_1$ $y_0 + \frac{2}{3}k_2$	k_1 k_2 k_3	$y_1^{(1)} = y_0 + \frac{1}{4} \times (k_1 + 3k_3)$	$\frac{h^2}{24}[f_x f_y + ff_y^2]_0 + O(h^3)$	2
y_0	k_1			2

r	№ п/п	l	γ	o	P_l	α_l	β_{l1}	β_{l2}	β_{l3}	$x_l = x_0 + \alpha_l h$
				2	0	5/12	5/12			$x_0 + \frac{5}{12}h$
				3	3/4	2/3	0	2/3		$x_0 + \frac{2}{3}h$
	6	1	1/2	1	—3/2	0				x_0
				2	3	1/3	1/3			$x_0 + \frac{h}{3}$
				3	—1/2	1	2	—1		$x_0 + h$
		2	—1/2	1	3/2	0				x_0
				2	—2/3	1/3	0			$x_0 + \frac{h}{3}$
				3	1	1	1/2	1/2		$x_0 + h$
	7	1	1/2	1	0	0				x_0
				2	0	1/3	1/3			$x_0 + \frac{h}{3}$
				3	1	1/2	0	1/2		$x_0 + \frac{h}{2}$
		2	—1/2	1	2/5	0				x_0
				2	0	1/3	1/3			$x_0 + \frac{h}{3}$
				3	3/5	5/6	0	5/6		$x_0 + \frac{5}{6}h$
4	8	1	1/12	1	1/6	0				x_0
				2	0	2/3	2/3			$x_0 + \frac{2}{3}h$
				3	1/6	1	1/4	3/4		$x_0 + h$
				4	2/3	1/2	1/4	3/8	—1/8	$-x_0 + \frac{1}{2}h$
		2	—1/12	1	2/9	0				x_0
				2	9/10	2/3	2/3			$x_0 + \frac{2}{3}h$
				3	—1/6	1	13/4	—9/4		$x_0 + h$
				4	2/45	3/2	27/18	0	15/8	$x_0 + \frac{3}{2}h$

$y_l = y_0 + \beta_{l1} k_1 + \dots + \beta_{l\,l-1} k_{l-1}$	$k_i = hf(x_i, y_i)$	Вид формул для $y_1^{(l)}$, $y_1 = \frac{y_1^{(1)} + y_1^{(2)}}{2}$	Погрешность на шаге	Порядок погрешности на шаге
$y_0 + \frac{5}{12}k_1$ $y_0 + \frac{2}{3}k_2$	k_2 k_3	$y_1^{(2)} = y_0 + \frac{1}{4} \times (k_1 + 3k_3)$	$-\frac{h^2}{24}\left[f_x f_y + f f_y^2\right]_0 + O(h^3)$	
y_0 $y_0 + \frac{k_1}{3}$ $y_0 + 2k_1 - k_2$	k_1 k_2 k_3	$y_1^{(1)} = y_0 - \frac{1}{2} \times (3k_1 - 6k_2 + k_3)$	$\frac{h^3}{4}[f_{xx} + 2ff_{xy} + f^2 f_{yy}]_0 + O(h^4)$	2
y_0 y_0 $y_0 + \frac{1}{2}(k_1 + k_2)$	k_1 k_2 k_3	$y_1^{(2)} = y_0 + \frac{3}{2}k_1 - \frac{2}{3}k_2 + k_3$	$-\frac{h^3}{4}[f_{xx} + 2ff_{xy} + f^2 f_{yy}]_0 + O(h^4)$	2
y_0 $y_0 + \frac{k_1}{3}$ $y_0 + \frac{k_2}{2}$	k_1 k_2 k_3	$y_1^{(1)} = y_0 + k_3$	$-\frac{1}{24}h^3[f_{xx} + 2f_{yx} + f^2 f_{yy}]_0 + O(h^3)$	2
y_0 $y_0 + \frac{k_1}{3}$ $y_0 + \frac{5}{6}k_3$	k_1 k_2 k_3	$y_1^{(2)} = y_0 + \frac{1}{3} \times (2k_1 + 3k_3)$	$\frac{h^3}{24}[f_{xx} + 2ff_{xy} + f^2 f_{yy}]_0 + O(h^3)$	2
y_0 $y_0 + \frac{2}{3}k_1$ $y_0 + \frac{1}{4}k_1 + \frac{3}{4}k_2$ $y_0 + \frac{1}{4}k_1 + \frac{3}{8}k_2 - \frac{1}{8}k_3$	k_1 k_2 k_3 k_4	$y_1^{(1)} = y_0 + \frac{1}{6} \times (k_1 + k_3 + 4k_4)$	$\frac{h^4}{12}(f_x f_y^2 + f f_y^3)_0 + O(h^5)$	3
y_0 $y_0 + \frac{2}{3}k_1$ $y_0 + \frac{13}{4}k_1 - \frac{9}{4}k_2$ $y_0 + \frac{1}{8}(27k_1 + 15k_3)$	k_1 k_2 k_3 k_4	$y_1^{(2)} = y_0 + \frac{1}{45} \times (10k_1 + 2k_4) + \frac{1}{30}(27k_2 - 5k_3)$	$-\frac{h^4}{12}(f_x f_y^2 + f f_y^3)_0 + O(h^5)$	3

Так как главные (относительно h) части погрешности шага двусторонних формул типа Рунге — Кутта равны по абсолютной величине и противоположны по знаку, то

$$|y(x_1) - y_1| = \left| y(x_1) - \frac{y_1^{(1)} + y_1^{(2)}}{2} \right| = O(h^3)$$

или порядок погрешности на шаге приближенного решения (50) возрастает на единицу по сравнению с порядком погрешности для приближенных значений $y_1^{(1)}$ и $y_1^{(2)}$.

в) Аналогично можно построить семейство формул типа Рунге — Кутта при $r = 3$

$$\begin{gathered} y_1 = p_1k_1 + p_2k_2 + p_3k_3; \\ k_1 = hf(x_0, y_0), \quad k_2 = hf(x_0 + \alpha_2, \quad y_0 + \beta_{21}k_1), \\ k_3 = hf(x_0 + \alpha_3, y_0 + \beta_{31}k_1 + \beta_{32}k_2). \end{gathered} \tag{51}$$

В случае односторонних формул Рунге — Кутта восемь параметров p_i, α_i, β_{ij} $(0 < j < i \leqslant 3)$ определяются из системы $(s = 3)$

$$R(0) = 0, \quad R'(0) = 0, \quad R''(0) = 0, \quad R'''(0) = 0, \tag{52}$$

где $R(h) = y(x_0 + h) - y(x_0) - \sum_{k=1}^{3} p_ik_i$, причем для достижения возможно более высокого порядка $(s = 3)$ следует положить

$$\alpha_2 = \beta_{21}, \quad \alpha_3 = \beta_{31} + \beta_{32}, \tag{53}$$

$$\begin{gathered} p_1 + p_2 + p_3 = 1, \quad p_2\alpha_2 + p_3\alpha_3 = \frac{1}{2}, \\ p_2\alpha_2^2 + p_3\alpha_3^2 = \frac{1}{3}, \quad p_3\beta_{32}\alpha_2 = \frac{1}{6}. \end{gathered} \tag{54}$$

Система шести уравнений (53), (54) содержит восемь неизвестных. Из бесчисленного множества решений этой системы выделяют формулу 5 из табл. 5.

Погрешность формулы 5 из табл. 5 на шаге будет иметь вид

$$R(h) = \frac{h^4}{4!} R^{(IV)}(\theta h).$$

Если $f(x, y)$ не зависит от y, то формула 5 из табл. 5 превращается в квадратурную формулу Симпсона 4-го порядка точности

$$\begin{gathered} y_1 = y_0 + \frac{h}{6}\left[f(x) + f\left(x + \frac{h}{2}\right) + f(x + h)\right], \\ R(h) = O(h^5). \end{gathered} \tag{55}$$

Для построения двусторонних формул типа Рунге — Кутта в случае $r = 3$ заметим, что при выполнении условий (53)

$$\begin{gathered} R(h) = h(1 - p_1 - p_2 - p_3) f_0 + \frac{h^2}{2}(1 - 2p_2\alpha_2 - 2p_3\alpha_3)[f_x + ff_y]_0 + \\ + \frac{h^3}{6}(1 - 3p_2\alpha_2^2 - 3p_3\alpha_3^2)[f_{xx} + 2ff_{xy} + f^2f_{yy}]_0 + \\ + \frac{h^3}{6}(1 - 6p_3\alpha_2\beta_{32})[f_xf_y + ff_y^2]_0 + O(h^4). \end{gathered} \tag{56}$$

Поэтому, если положить

$$1 - p_1 - p_2 - p_3 = 0, \tag{57}$$

$$\frac{1}{2} - p_2\alpha_2 - p_3\alpha_3 = 0; \quad 1 - 3p_2\alpha_2^2 - 3p_3\alpha_3^2 = 0,$$

то

$$R(h) = h^3\gamma\,[f_x f_y + f f_y^2]_0 + O(h^4),$$

где

$$\gamma = \frac{1}{6} - p_3\beta_{32}\alpha_2. \tag{58}$$

Из (53), (57), (58) можно определить p_i, α_i, β_{ij} $(0 < j < i \leqslant 3)$ при $\alpha_2 \neq 0$, $\alpha_3 \neq 0$, $\alpha_3 - \alpha_2 \neq 0$, $3\alpha_2 - 2 \neq 0$:

$$p_2 = \frac{3\alpha_3 - 2}{6\alpha_2(\alpha_3 - \alpha_2)}; \quad p_3 = \frac{2 - 3\alpha_2}{6\alpha_3(\alpha_3 - \alpha_2)}; \tag{59}$$

$$p_1 = 1 - \frac{2 - 3\alpha_2}{6\alpha_3(\alpha_3 - \alpha_2)} - \frac{3\alpha_3 - 2}{6\alpha_2(\alpha_3 - \alpha_2)};$$

$$\beta_{32} = \frac{\alpha_3(6\gamma - 1)(\alpha_3 - \alpha_2)}{\alpha_2(3\alpha_2 - 2)}; \quad \beta_{31} = \alpha_3\left[1 - \frac{6\gamma(\alpha_3 - \alpha_2)}{\alpha_2(3\alpha_2 - 2)}\right].$$

Таким образом соотношения (59) определяют трехпараметрическое семейство двусторонних формул типа Рунге — Кутта второго порядка точности (примеры таких формул приведены в табл. 6). Очевидно, если положить в (56)

$$1 - p_1 - p_2 - p_3 = 0, \quad \frac{1}{2} - p_2\alpha_2 - p_3\alpha_3 = 0, \quad \frac{1}{6} - p_3\alpha_2\beta_{32} = 0, \tag{60}$$

то

$$R(h) = \frac{\tilde{\gamma}}{2} h^3\,[f_{xx} + 2ff_{xy} + f^2 f_{yy}]_0 + O(h^4), \tag{61}$$

где

$$\tilde{\gamma} = \frac{1}{3} - p_2\alpha_2^2 - p_3\alpha_3^2. \tag{62}$$

Из системы (53), (60), (62) при условии $6\tilde{\gamma} + 3\alpha_2 - 2 \neq 0$, $\alpha_2 \neq 0$, $\beta_{32} \neq 0$, $\tilde{\gamma} \neq 0$ определятся параметры:

$$p_3 = \frac{1}{6\beta_{32}\alpha_2}; \quad p_2 = \frac{1}{\alpha_2}\left(\frac{1}{2} - \frac{\alpha_3}{6\beta_{32}\alpha_2}\right);$$

$$p_1 = 1 - \frac{1}{\alpha_2}\left(\frac{1}{2} - \frac{\alpha_3}{6\beta_{32}\alpha_2}\right) - \frac{1}{6\beta_{32}\alpha_2}; \tag{63}$$

$$\beta_{32} = \frac{\alpha_3(\alpha_2 - \alpha_3)}{\alpha_2(6\tilde{\gamma} + 3\alpha_2 - 2)}; \quad \beta_{31} = \alpha_3\left[1 + \frac{\alpha_3 - \alpha_2}{\alpha_2(6\tilde{\gamma} + 3\alpha_2 - 2)}\right].$$

Соотношения (63) определяют еще одно трехпараметрическое семейство двусторонних формул типа Рунге — Кутта второго порядка точности (примеры таких формул приведены в табл. 6).

Аналогично можно построить формулы типа Рунге — Кутта при $r = 4$.

Примеры односторонних формул Рунге — Кутта четвертого порядка точности на шаге приведены в табл. 5, а двусторонних формул третьего порядка точности в табл. 6.

При $r = 5$ увеличения порядка точности формул Рунге — Кутта не достигается, а при $r \geqslant 6$ получаются формулы, имеющие пятый порядок точности, но они требуют на каждом шаге не менее 5-кратного вычисления правой части уравнения (см. табл. 5), а поэтому неудобны для практических вычислений.

Формулы типа Рунге — Кутта принадлежат к числу одношаговых методов приближенного решения задачи Коши. Счет по любой из формул Рунге — Кутта начинается с вычисления y_1, которое принимается за приближенное значение решения $y(x_0 + h_0)$ в точке $x_0 + h_0$. Затем

точку $x_1 = x_0 + h_0$ берут за начальную, а значение y_1 за начальное значение в этой точке и находят значение y_2 в точке $x_2 = x_1 + h_1$ и т. д.

В случае использования двусторонних формул типа Рунге — Кутта иногда поступают следующим образом. На каждом j-м шаге вычисляют два значения: $y_j^{(1)}$ и $y_j^{(2)}$. Пусть для определенности $y_j^{(1)} < y_j^{(2)}$. Исходя из $y_j^{(1)}$, производят вычисления по обеим формулам двустороннего метода и находят $y_{j+1}^{(1),(1)}$ и $y_{j+1}^{(1),(2)}$. Наименьшее значение из $y_{j+1}^{(1),(1)}$ и $y_{j+1}^{(1),(2)}$ принимают за $y_{j+1}^{(1)}$ на $(j+1)$-м шаге. Аналогично, исходя из $y_j^{(2)}$, вычисляют $y_{j+1}^{(2),(1)}$, $y_{j+1}^{(2),(2)}$ и наибольшее из получившихся значений принимают за $y_{j+1}^{(2)}$; если $y_{j+1}^{(1)} > y_{j+1}^{(2)}$, то считается, что двусторонний метод Рунге — Кутта неприменим.

В качестве приближенного значения $y(x_{j+1})$ принимают величину

$$y_{j+1} = \frac{y_{j+1}^{(1)} + y_{j+1}^{(2)}}{2}. \tag{64}$$

Оценка абсолютной и относительной погрешности приближенного решения (64) на шаге оценивается соответственно по формулам:

$$\frac{|y_j^{(1)} - y_j^{(2)}|}{2}, \quad \frac{|y_j^{(1)} - y_j^{(2)}|}{|y_j^{(1)} + y_j^{(2)}|}. \tag{65}$$

Очевидно, использование двустороннего метода Рунге — Кутта приводит к увеличению объема вычислений, но в большинстве случаев позволяет судить о границах изменения решения задачи Коши, если вычислительная погрешность не оказывает существенного влияния на достоверность получаемых границ.

Для практической оценки погрешности на шаге в случае использования односторонних формул Рунге — Кутта порядка s применяется принцип Рунге: приближенные значения $y_2^{(1)}$, $y_2^{(2)}$ в точке $x_2 = x_0 + 2h$ находят по односторонней формуле с шагом $2h$ и применяя дважды ту же формулу — с шагом h. Тогда

$$y(x_2) - y_2^{(1)} \approx \frac{R^{(s+1)}(0)(2h)^{s+1}}{(s+1)!}$$

и

$$y(x_2) - y_2^{(2)} \approx \frac{2R^{(s+1)}(0)h^{s+1}}{(s+1)!}.$$

Откуда

$$y_2^{(2)} - y_2^{(1)} \approx \frac{2h^{(s+1)}R^{(s+1)}(0)}{(s+1)!}[2^s - 1]. \tag{66}$$

Абсолютная погрешность на шаге оценивается величиной

$$\frac{|y_2^{(2)} - y_2^{(1)}|}{2^s - 1}.$$

Если значение $\left|\frac{y_2^{(2)} - y_2^{(1)}}{2^s - 1}\right|$ окажется большим, то нужно уменьшить шаг и вновь провести вычисления.

В случае задачи Коши (5) для системы обыкновенных дифференциальных уравнений построение формул типа Рунге — Кутта осуществляется следующим образом: для каждого l вводят параметры

$$\alpha_{l,i};\ \beta_{l,ij};\ p_{l,i} \qquad (0 < j < i \leqslant r_l,\ l = \overline{1,\ n})$$

и строят приближенные равенства вида

$$y_l(x+h) - y_l(x) \approx \sum_{m=1}^{r} p_{l,m} k_{l,m}, \tag{67}$$

где

$$k_{l,1} = hf_l(x,\ y_1,\ \ldots,\ y_n);$$

$$k_{l,m} = hf_l(x + \alpha_{l,m},\ y_1 + \beta_{1,m1}k_{1,1} + \beta_{1,m2}k_{1,2} + \cdots + \beta_{1,mm-1}k_{1,m-1},\ y_2 + \\ + \beta_{2,m1}k_{2,1} + \beta_{2,m2}k_{2,2} + \cdots + \beta_{2,mm-1}k_{2,m-1},\ \ldots,\ y_n + \beta_{n,m1}k_{n,1} + \\ + \beta_{n,m2}k_{n,2} + \cdots + \beta_{n,mm-1}k_{n,m-1}) \tag{68}$$

$$(l = \overline{1,\ n},\quad m = \overline{2,\ r}).$$

Число свободных параметров при построении формул метода Рунге — Кутта в случае системы обыкновенных дифференциальных уравнений значительно увеличивается, а значит, и возрастают возможности получения формул Рунге — Кутта одного и того же порядка точности на шаге. Если выбрать $\alpha_{l,i}$, $\beta_{l,ij}$, $p_{l,i}$ одинаковыми для всех l ($l = \overline{1,\ n}$), то в этом случае односторонние и двусторонние формулы типа Рунге — Кутта, построенные для одного уравнения, легко переносятся на случай системы уравнений.

Сделаем некоторые замечания, касающиеся устойчивости счета по формулам типа Рунге — Кутта на примере простейшей модельной задачи Коши

$$y' = \mu y, \quad y(0) = y_0. \tag{69}$$

Если использовать расчетные формулы метода Рунге — Кутта вида (17), (13) — (15) и последовательно вычислить значения $y_1, y_2, \ldots, y_k$, то получим:

$$y_1 = y_0 + \sum_{i=1}^{r} p_i k_i, \quad k_1 = \mu h y_0, \quad k_2 = y_0 + \beta_{21} h k_1 = P_2(h)\, y_0, \quad k_3 = \\ = y_0 + \beta_{31} h k_1 + \beta_{32} h k_2 = P_3(h)\, y_0, \\ k_r = P_r(h)\, y_0,$$

$$y_1 = y_0 + \sum_{i=1}^{r} p_i P_i(h)\, y_0 = Q_r(h)\, y_0. \tag{70}$$

Откуда

$$y_k = Q_r(h)\, y_{k-1} = (Q_r(h))^k\, y_0. \tag{71}$$

Следовательно, для того чтобы счет по формулам типа Рунге — Кутта был устойчивым, шаг интегрирования должен быть выбран так, чтобы

$$|Q_r(h)| < 1. \tag{72}$$

Например, в случае применения формул Рунге — Кутта четвертого порядка точности вида (8) табл. 5 имеем:

$$y_1 = \left(1 + \mu h + \frac{\mu^2 h^2}{2} + \frac{\mu^3 h^3}{6} + \frac{\mu^4 h^4}{24}\right) y_0,$$

и

$$|Q_4(h)| = \left|1 + \mu h + \frac{\mu^2 h^2}{2} + \frac{\mu^3 h^3}{6} + \frac{\mu^4 h^4}{24}\right| \leqslant 1$$

при $\mu < 0$, если

$$|\mu h| \leqslant C, \text{ где } C \approx 2{,}8. \tag{73}$$

В случае интегрирования задачи Коши для линейной системы с постоянными коэффициентами

$$Y' = AY, \quad Y(0) = Y_0, \quad A = (a_{ij})_{i=\overline{1,n}}^{j=\overline{1,n}} \tag{74}$$

расчетные формулы для вычисления Y_1 приводят к соотношению

$$Y_1 = Y_0 + \sum_{i=1}^{r} p_i K_i = Q_r(Ah)\, Y_0. \tag{75}$$

Если предположить, что матрица A является матрицей простой структуры, то ее собственные векторы $\varphi_1, \ldots, \varphi_n$ образуют базис, по которому можно разложить векторы Y_1 и Y_0

$$Y_1 = \sum_{j=1}^{n} c_j^{(1)} \varphi_j, \quad Y_0 = \sum_{j=1}^{n} c_j^0 \varphi_j. \tag{76}$$

Подставив (76) в (75), получим:

$$Y_1 = \sum_{j=1}^{n} c_j^0 Q_r(Ah)\, \varphi_j = \sum_{j=1}^{n} c_j^0 Q_r(\mu_j h)\, \varphi_j,$$

где μ_i — собственные значения матрицы A.

Из соотношения

$$\sum_{j=1}^{n} c_j^{(1)} \varphi_j = \sum_{j=1}^{n} c_j^0 Q_r(\mu_j h)\, \varphi_j$$

получаем:

$$c_j^{(1)} = Q_r(\mu_j h)\, c_j^0$$

и так как

$$Y_k = \sum_{j=1}^{n} (Q_r(\mu_j h))^k c_j^{(k)} \varphi_j,$$

то

$$c_j^{(k)} = (Q_r(\mu_j h))^k c_j^0. \tag{77}$$

Из (77) видно, что малые возмущения коэффициентов c_j^0 при $|Q_r(\mu_j h)| > 1$ могут привести к большим возмущениям решения Y_k.

Формулы преобразования коэффициентов $c_j^{(k)}$ совпадают с формулами численного интегрирования методами Рунге — Кутта уравнения $y' = \mu_p y$, поэтому для получения приемлемой точности при интегрировании системы необходимо выполнение условия вида (73)

$$\max_j |\mu_j| h < \text{const}, \tag{78}$$

где μ_j — собственные значения матрицы A.

В случае нелинейной системы

$$Y' = f(x, Y) \tag{79}$$

представим ее в окрестности искомого решения Y^* в виде

$$Y' = f(x, Y^*) + f_y(x, Y^*)(Y - Y^*). \tag{80}$$

Из представления (80) приходим к выводу, что в окрестности каждой точки x с высокой точностью должна интегрироваться линейная система вида

$$v' = f_y(x, Y^*(x))\, v, \tag{81}$$

а значит, на выбор шага приходится накладывать условия вида (78), где μ_i — собственные значения матрицы $f_y(x, Y^*(x))$.

Ограничения вида (78) на шаг уравнения являются весьма обременительными даже для вычислений на ЭВМ. Однако многие задачи проблемы регулирования, управления, кинетики, электроники приводят к решению таких систем, у которых матрица $f_y(x, Y^*(x))$ имеет большой разброс собственных значений, причем большие по модулю собственные значения имеют большую отрицательную вещественную часть. Такие системы дифференциальных уравнений получили даже специальное название «жестких» систем. Решение «жестких» систем методами типа Рунге — Кутта становится практически невозможным из-за жестких требований устойчивости.

Применительно к «жестким» системам в последнее время предложены различные модификации одношаговых и многошаговых методов (см., например, [85] [89]). Таким образом, несмотря на достаточную универсальность методов типа Рунге — Кутта для решения задач Коши нельзя сбрасывать со счета его недостатки, которые в основном связаны с вопросами устойчивости счета по формулам и значительным машинным временем, затрачиваемым на его реализацию особенно в тех случаях, когда правая часть уравнения (7) довольно сложная функция.

Оценка погрешности одношаговых методов. Пусть y_k — приближенное значение решения задачи Коши (7) в точке x_k, полученное в результате применения какого-либо одношагового метода s-го порядка,

$$y_k = \Phi(x_k, x_{k-1}, y_k, y_{k-1}, h_{k-1}, f), \quad h_{k-1} = x_k - x_{k-1}. \tag{82}$$

Оценим величину $\varepsilon_k = y(x_k) - y_k + \Delta_k$, где Δ_k — погрешность округления при вычислении y_k. Одношаговые методы решения задачи Коши применяются рекуррентно, т. е. y_k $(k \geqslant 1)$ находятся в результате применения одношагового метода (82) к решению задачи Коши для уравнения

$$v'(x) = f(x, v(x)), \tag{83}$$

$$v(x_{k-1}) = y_{k-1}. \tag{84}$$

Поэтому величину $R_k = y(x_k) - y_k$ можно представить в виде

$$R_k = y(x_k) - v(x_k) + v(x_k) - y_k = y(x_k) - v(x_k) + O(h_{k-1}^{s+1}). \tag{85}$$

Выразим R_k через $R_{k-1} = y(x_{k-1}) - y_{k-1}$. Так как функции $y(x)$ и $v(x)$ удовлетворяют уравнению $y' = f(x, y)$, то функция

$w(x) = y(x) - v(x)$ будет удовлетворять уравнению

$$w'(x) = \frac{\partial f(x, \tilde{v}(x))}{\partial y} w(x), \tag{86}$$

где $\tilde{v}(x) = v(x) + \theta(x)(y(x) - v(x))$, $0 < \theta(x) < 1$.

Для того чтобы в этом убедиться, достаточно почленно вычесть из уравнения $y' = f(x, y)$ уравнение (83) и применить формулу Лагранжа о конечном приращении. Проинтегрировав уравнение (86) при начальном условии

$$w(x_{k-1}) = R_{k-1},$$

находим:

$$w(x) = R_{k-1} e^{\int\limits_{x_{k-1}}^{x} \frac{\partial f(x, \tilde{v}(x))}{\partial y} dx} \tag{87}$$

или

$$w(x_k) = y(x_k) - v(x_k) = R_{k-1} e^{\int\limits_{x_{k-1}}^{x_k} \frac{\partial f(x, \tilde{v}(x))}{\partial y} dx}. \tag{88}$$

Подставляя (88) в (85), получим разностное уравнение для R_k

$$R_k = R_{k-1} e^{\int\limits_{x_{k-1}}^{x_k} f_y(x, \tilde{v}) dx} + O(h_{k-1}^{s+1}). \tag{89}$$

Обозначим через L константу Липшица и пусть при $x_l \in [x_0, \xi]$, $|f_y| \leqslant L$ для всех y и $|h_l| < h$.

Из (89) получим:

$$|R_k| \leqslant |R_{k-1}| e^{Lh} + O(h^{s+1}) \leqslant O(h^{s+1}) \sum_{j=0}^{k-1} e^{Ljh} = O(h^{s+1}) \frac{e^{Lkh} - 1}{e^{Lh} - 1}. \tag{90}$$

Если учесть асимптотическое равенство

$$h \sum_{j=0}^{k-1} e^{Ljh} = \int\limits_{x_0}^{x_k} e^{Lx} dx + O(h), \tag{91}$$

то

$$|R_k| \leqslant O(h^s) \left[\int\limits_{x_0}^{x_k} e^{Lx} dx + O(h) \right]. \tag{92}$$

Аналогично можно построить мажорантную оценку погрешности одношаговых методов решения задачи Коши для системы вида (3). Если обозначить через $\boldsymbol{R}_k = \boldsymbol{Y}(x_k) - \boldsymbol{Y}_k$, а через L константу Липшица для системы (3), то

$$\|\boldsymbol{R}_k\|_1 \leqslant O(h^{s+1}) \sum_{j=0}^{k-1} \exp Lhjn \tag{92'}$$

или

$$\|\boldsymbol{R}_k\|_1 \leqslant O(h^s) \left[\int\limits_{x_0}^{x_k} \exp xnLdx + O(h) \right].$$

Для оценки полной погрешности одношаговых методов решения задачи Коши рассмотрим

$$\varepsilon_k = y(x_k) - y_k + \Delta_k = R_k + \Delta_k,$$

где Δ_k — погрешность округления на шаге с номером k.

Исходя из (89), (90), имеем:

$$|\varepsilon_k| \leqslant |\varepsilon_0| e^{Lhk} + \sum_{j=0}^{k-1} (O(h^{s+1}) + \Delta) e^{Ljh} \leqslant |\varepsilon_0| e^{Lhk} + \\ + (O(h^{s+1}) + \Delta) \frac{e^{Lkh} - 1}{e^{Lh} - 1} \tag{93}$$

или

$$|\varepsilon_k| \leqslant e^{L(x_k - x_0)} [|\varepsilon_0| + O(h^{s+1}) + k\Delta], \tag{94}$$

где $\Delta = \max\limits_l |\Delta_l|$.

Если учесть равенство (91), то можно получить для ε_k следующую оценку:

$$|\varepsilon_k| \leqslant |\varepsilon_0| e^{Lhk} + \left(O(h^s) + \frac{\Delta}{h}\right) \left[\int_{x_0}^{x_k} e^{Lx} dx + O(h)\right]. \tag{95}$$

Очевидно, $\max |\varepsilon_k| \to 0$ при $h \to 0$, если одновременно $\frac{\Delta}{h} \to 0$, $|\varepsilon_0| \to 0$.

Иными словами, приближенное решение задачи Коши (7), построенное по какому-либо одношаговому методу, будет близко к точному решению при малой вычислительной погрешности и достаточно мелком шаге интегрирования. Следует отметить, что величина $\frac{\Delta}{h}$ при малом h $\left(h \ll \frac{\Delta}{h}\right)$ может оказать значительное влияние на погрешность результата.

3. Многошаговые методы решения задачи Коши

Рассмотренные в предыдущем пункте одношаговые методы являются частным случаем более общих m-шаговых методов, которые для вычисления приближенного значения в точке x_j используют информацию о значениях y_i в m точках, лежащих в окрестности точки x_j.

МЕТОД НЕОПРЕДЕЛЕННЫХ КОЭФФИЦИЕНТОВ

Метод неопределенных коэффициентов является одним из способов построения m-шаговых методов решения задачи Коши.

Производную $y'(x_j)$ и значение $f(x_j, y(x_j))$ приближают соответственно выражениям вида

$$\sum_{k=0}^{m} \frac{a_k y(x_j - kh)}{h}, \quad \sum_{k=0}^{m} b_k f(x_j - kh, y(x_j - kh))$$

с неопределенными параметрами a_k и b_k $(k = \overline{0, m})$.

Значения точного решения дифференциальной задачи (7) в точке x_j будут удовлетворять соотношению

$$\sum_{k=0}^{m} \frac{a_k}{h} y(x_{j-k}) - \sum_{k=0}^{m} b_k f(x_{j-k},\ y(x_{j-k})) = R_j(h). \tag{96}$$

Соотношению (96) ставят в соответствие конечно-разностную схему вида

$$\sum_{k=0}^{m} \frac{a_k}{h} y_{j-k} = \sum_{k=0}^{m} b_k f_{j-k}. \tag{97}$$

Величина

$$R_j(h) = \sum_{k=0}^{m} \frac{a_k y(x_j - kh)}{h} - \sum_{k=0}^{m} b_k f(x_j - kh,\ y(x_j - kh)) \tag{98}$$

называется погрешностью аппроксимации дифференциального уравнения (7) разностной схемой (97).

Выбор параметров a_k и b_k подчиняют следующим условиям:

$$\lim_{h\to 0} \sum_{k=0}^{m} \frac{a_k y(x_j - kh)}{h} = y'(x_j), \quad x_0 \leqslant x_j \leqslant \zeta; \tag{99}$$

$$\lim_{h\to 0} \sum_{k=0}^{m} b_k f(x_j - kh,\ y(x_j - kh)) = f(x_j,\ y(x_j)), \quad x_0 \leqslant x_j \leqslant \zeta; \tag{100}$$

$$\max_{x_0 \leqslant x_j < \zeta} |R_j(h)| \to 0 \text{ при } h \to 0. \tag{101}$$

Если при выполнении условий (99) — (101) $R_j(h) = O(h^p)$, то говорят, что схема (97) имеет p-й порядок аппроксимации.

Коэффициенты a_k, b_k $(k = \overline{0, m})$ подбирают так, чтобы выполнялись условия (99) — (101) и $R_j(h) = O(h^p)$. Для этого значения $y(x_j - kh)$ и $f(x_j - kh,\ y(x_j - kh))$, входящие в выражение погрешности аппроксимации (98), представим по формуле Тейлора

$$y(x_j - kh) = \sum_{l=0}^{p} \frac{(-kh)^l y^{(l)}(x_j)}{l!} + r_j^k, \tag{102}$$

$$y'(x_j - kh) = \sum_{l=0}^{p-1} \frac{(-kh)^l y^{(l+1)}(x_j)}{l!} + \rho_j^k, \tag{103}$$

где r_j^k, ρ_j^k — остаточные члены рядов Тейлора (102) и (103)

$$|r_j^k| \leqslant \frac{c_{p+1}(kh)^{p+1}}{(p+1)!}, \quad |\rho_j^k| \leqslant \frac{c_{p+1}(kh)^p}{p!}, \tag{104}$$

$$|y^{(p+1)}(x)| \leqslant c_{p+1} < \infty, \quad x_0 \leqslant x \leqslant \zeta.$$

Подставляя (102), (103) в (98), получим:

$$R_j(h) = \sum_{k=0}^{m} \frac{a_k}{h} y(x_j) + \sum_{k=0}^{m} (-ka_k - b_k) y'(x_j) + \sum_{k=0}^{m} \Big[\frac{k^2 a_k}{2!} +$$

$$+\frac{kb_k}{1!}\Big]hy''(x_j)+\cdots+\sum_{k=0}^{m}\left[\frac{(-k)^p a_k}{p!}-\frac{(-k)^{p-1}b_k}{(p-1)!}\right]h^{p-1}y^{(p)}(x_j)+$$

$$+\sum_{k=0}^{m}\frac{a_k r_j^k}{h}-\sum_{k=0}^{m}b_k\rho_j^k. \tag{105}$$

Обозначим

$$\sum_{k=0}^{m}a_k=\varphi_0,\quad \sum_{k=0}^{m}\left[\frac{(-k)^l}{l!}a_k-\frac{(-k)^{l-1}}{(l-1)!}b_k\right]=\varphi_l,\quad l\geqslant 1, \tag{106}$$

тогда, если

$$\varphi_0=\varphi_1=\varphi_2=\cdots=\varphi_p=0, \tag{107}$$

то

$$|R_j(h)|\leqslant c_{p+1}h^p\sum_{k=0}^{m}\left(|a_k|\frac{k^{p+1}}{(p+1)!}+|b_k|\frac{k^p}{p!}\right). \tag{108}$$

Соотношения (107) образуют алгебраическую систему линейных однородных уравнений с $(2m+2)$ неизвестными a_k, b_k $(k=\overline{0,m})$. Эта система будет иметь ненулевое решение при $2m+2>p+1$, а для $|R_j(h)|$ будет иметь место оценка (108)

$$R_j(h)=O(h^p).$$

Проверим выполнение условий (99), (100).

Выражения (99), (100), пользуясь разложением в ряд Тейлора, можно записать в виде

$$\lim_{h\to 0}\left(\sum_{k=0}^{m}\frac{a_k}{h}y(x_j)-\sum_{k=0}^{m}ka_k y'(x_j)+O(h)\right)=y'(x_j), \tag{109}$$

$$\lim_{h\to 0}\left(\sum_{k=0}^{m}b_k f(x_j,\ y(x_j))+O(h)\right)=f(x_j,\ y(x_j)),\quad x_0\leqslant x_j\leqslant\zeta. \tag{110}$$

Соотношения (109), (110) будут выполняться, если

$$\sum_{k=0}^{m}a_k=0,\quad -\sum_{k=0}^{m}ka_k=1,\quad \sum_{k=0}^{m}b_k=1 \tag{111}$$

или при $\varphi_0=0$, $\varphi_1=\sum_{k=0}^{m}(-ka_k-b_k)=0$,

$$\sum_{k=0}^{m}b_k=1. \tag{112}$$

Таким образом, если коэффициенты a_k, b_k $(k=\overline{0,m})$ определяются из однородной системы линейных алгебраических уравнений (107) при $0\leqslant p\leqslant 2m$ и удовлетворяют условию нормировки (112), то разностная схема (97) будет иметь p-й порядок аппроксимации.

Если строить разностные формулы вида (97) с заранее фиксированными значениями некоторых параметров, например, $b_0=0$, т. е. решать систему вида $b_0=0$, $\varphi_0=0$, $\varphi_1=0$, ..., $\varphi_{2m}=0$, то порядок аппроксимации будет понижаться.

Примеры m-шаговых формул решения задачи Коши:

а) При $m = 1$ для определения a_k, b_k ($k = 0, 1$) получим систему

$$\begin{aligned} &a_0 + a_1 = 0 \qquad (\varphi_0 = 0),\\ &a_1 + b_0 + b_1 = 0 \quad (\varphi_1 = 0),\\ &\frac{a_1}{2} + b_1 = 0 \quad (\varphi_2 = 0). \end{aligned} \tag{113}$$

Общее решение системы (113) имеет вид

$$a_0 = \varkappa, \quad a_1 = -\varkappa, \quad b_0 = b_1 = \frac{\varkappa}{2}.$$

Из условия нормировки (112) находим, что $\varkappa = 1$.

Разностная схема (97) запишется следующим образом:

$$y_j = y_{j-1} + \frac{h}{2}(f_j + f_{j-1})$$

и будет иметь второй порядок аппроксимации.

б) При $m = 2$ система для определения коэффициентов a_k, b_k ($k = \overline{0, 2}$) имеет вид

$$\begin{aligned} &a_0 + a_1 + a_2 = 0 \qquad (\varphi_0 = 0),\\ &a_1 + 2a_2 + b_0 + b_1 + b_2 = 0 \qquad (\varphi_1 = 0),\\ &\frac{a_1}{2} + 2a_2 + b_1 + 2b_2 = 0 \qquad (\varphi_2 = 0),\\ &-\frac{a_1}{6} - \frac{4}{3}a_2 - \frac{b_1}{2} - 2b_2 = 0 \qquad (\varphi_3 = 0),\\ &\frac{a_1}{24} + \frac{2}{3}a_2 + \frac{b_1}{6} + \frac{4}{3}b_2 = 0 \qquad (\varphi_4 = 0). \end{aligned} \tag{114}$$

Запишем общее решение системы (114)

$$a_0 = \varkappa, \quad a_1 = 0, \quad a_2 = -\varkappa, \quad b_0 = b_2 = \frac{\varkappa}{3}, \quad b_1 = \frac{4}{3}\varkappa.$$

Из (112) находим $\varkappa = \frac{1}{2}$.

Расчетная схема четвертого порядка аппроксимации будет иметь вид

$$y_j = y_{j-2} + \frac{h}{3}(f_j + 4f_{j-1} + f_{j-2}). \tag{115}$$

в) Рассмотрим систему вида (114) при $m = 2$ и фиксированном значении коэффициента $b_0 = 0$, т. е. систему

$$b_0 = 0, \quad \varphi_0 = 0, \quad \varphi_1 = 0, \quad \varphi_2 = 0, \quad \varphi_3 = 0. \tag{116}$$

Запишем решения системы (116), удовлетворяющие условию нормировки (112):

$$a_0 = \frac{1}{6}; \quad a_1 = \frac{2}{3}; \quad a_2 = -\frac{5}{6}; \quad b_0 = 0; \quad b_1 = \frac{2}{3}; \quad b_2 = \frac{1}{3}.$$

Следовательно, разностная схема третьего порядка аппроксимации ($\varphi_4 \neq 0$) будет иметь вид

$$y_j = -4y_{j-1} + 5y_{j-2} + h(4f_{j-1} + 2f_{j-2}). \tag{117}$$

г) Пусть $m = 2$, $b_0 = 0$, $a_2 = 0$, т. е. рассмотрим систему вида

$$\begin{aligned} &b_0 = 0, \quad a_2 = 0, \quad a_0 + a_1 = 0 \quad (\varphi_0 = 0), \quad a_1 + b_1 + b_2 = 0 \quad (\varphi_1 = 0),\\ &\frac{a_1}{2} + b_1 + 2b_2 = 0 \qquad (\varphi_2 = 0). \end{aligned} \tag{118}$$

Решениями системы (118), удовлетворяющими условию (112), будут числа $a_0 = 1$, $a_1 = -1$, $a_2 = 0$, $b_0 = 0$, $b_1 = \frac{3}{2}$, $b_2 = -\frac{1}{2}$.

Расчетная схема второго порядка аппроксимации ($\varphi_3 \neq 0$) будет иметь вид

$$y_j = y_{j-1} + \frac{h}{2}(3f_{j-1} - f_{j-2}),$$

или

$$y_j = y_{j-1} + h\left(f_{j-1} + \frac{1}{2}\Delta f_{j-2}\right), \tag{119}$$

где $\Delta^k f_j$ — конечная разность k-го порядка.

Формула (119) носит название *экстраполяционной формулы Адамса.*

Экстраполяционные формулы Адамса могут быть построены, если проинтегрировать уравнение (7) на отрезке $[x_{j-1}, x_j]$

$$y(x_j) = y(x_{j-1}) + \int_{x_{j-1}}^{x_j} f(x, y(x))\,dx \tag{120}$$

и подынтегральную функцию заменить интерполяционным полиномом Ньютона интерполирования назад, построенным по $m+1$ точке

$$x_{j-1},\ x_{j-2},\ \ldots,\ x_{j-m-1},$$

не принадлежащих отрезку $(x_{j-1}, x_j]$.

Тогда

$$y(x_j) = y(x_{j-1}) + h\int_0^1 f(x_j + th)\,dt, \quad t = \frac{x - x_{j-1}}{h},$$

$$f(x_{j-1} + th) = f(x_{j-1}) + \frac{t}{1!}\Delta f(x_{j-2}) + \frac{t(t+1)}{2!}\Delta^2 f(x_{j-3}) + \cdots +$$

$$+ \frac{t(t+1)\ldots(t+m-1)}{m!}\Delta^m f(x_{j-m-1}) + \tilde{r}_m(t), \tag{121}$$

$$\tilde{r}_m(t) = \frac{h^{m+1}}{(m+1)!}t(t+1)\ldots(t+m)\,y^{(m+2)}(\eta),$$

$$x_{j-m-1} < \eta < x_j, \qquad 0 \leqslant t \leqslant 1$$

и расчетная схема экстраполяционного метода Адамса будет иметь вид

$$y_j = y_{j-1} + h\left(f_{j-1} + \frac{1}{2}\Delta f_i + \frac{5}{12}\Delta^2 f_{j-3} + \cdots + \beta_m \Delta^m f_{j-m-1}\right), \tag{122}$$

$$\beta_m = \int_0^1 \frac{t(t+1)\ldots(t+m-1)}{m!}\,dt > 0.$$

Погрешность аппроксимации экстраполяционной схемы метода Адамса оценивается величиной

$$|R_j(h)| \leqslant h^{m+2}\int_0^1 \frac{t(t+1)\ldots(t+m)}{(m+1)!}\,|y^{m+2}(\eta)|\,dt, \tag{123}$$

или

$$|R_j(h)| \leqslant h^{m+2}c_{m+2}\beta_{m+1}. \tag{124}$$

Отметим, что оценки вида (123), (124) являются более точными по сравнению с оценками вида (108).

д) Рассмотрим систему вида (107), (112) при $m = 2$ и фиксированном значении коэффициента $a_2 = 0$;

$$a_2 = \varphi_0 = \varphi_1 = \varphi_2 = \varphi_3 = 0.$$

Тогда $a_0 = 1$; $a_1 = -1$; $b_0 = \frac{5}{12}$; $b_1 = \frac{2}{3}$; $b_2 = -\frac{1}{12}$ и расчетная схема третьего порядка аппроксимации имеет вид

$$y_j = y_{j-1} + h\left(\frac{5}{12} f_j + \frac{2}{3} f_{j-1} - \frac{1}{12} f_{j-2}\right),$$

или

$$y_j = y_{j-1} + h\left(f_j - \frac{1}{2} \Delta f_{j-1} - \frac{1}{12} \Delta^2 f_{j-2}\right). \tag{125}$$

Формула (125) носит название *интерполяционной формулы метода Адамса*. Они могут быть построены, если воспользоваться выражением (120) и подынтегральную функцию заменить интерполяционным полиномом Ньютона интерполирования назад, построенным по точкам

$$x_j, \quad x_{j-1}, \quad x_{j-2}, \ \ldots, \ x_{j-m}.$$

Тогда

$$y_j = y_{j-1} + h\left(f_j - \frac{1}{2} \Delta f_{j-1} - \frac{1}{12} \Delta^2 f_{j-2} + \cdots + \tilde{\beta}_m \Delta^m f_{j-m}\right) \tag{126}$$

$$\tilde{\beta}_m = \int_{-1}^{0} \frac{u(u+1)\ldots(u+m-1)}{m!} du; \quad u = \frac{x - x_j}{h}, \quad -1 \leqslant u \leqslant 0.$$

Для погрешности аппроксимации интерполяционной схемы метода Адамса имеет место следующая оценка:

$$|R_j(h)| \leqslant h^{m+2} \int_{-1}^{0} \frac{u(u+1)\ldots(u+m)}{(m+1)!} \| y^{(m+2)}(\tilde{\eta}) | du, \tag{127}$$

$$x_{j-m} < \tilde{\eta} = \tilde{\eta}(u) < x_j.$$

В табл. 7 приведены значения коэффициентов для некоторых m-шаговых разностных методов вида (97) решения задачи Коши.

В m-шаговых разностных схемах вида (97) с $b_0 = 0$, $a_0 \neq 0$, к которым, в частности, принадлежат экстраполяционные формулы метода Адамса, значения $y(x)$ в точке x_j определяются по явной формуле, если известны приближенные значения $y(x)$ в m предыдущих точках (в m начальных точках).

Для построения начала таблицы могут быть использованы одношаговые методы (типа Рунге — Кутта, разложения в ряд Тейлора). При этом шаг рекомендуется выбрать так, чтобы разность m-го порядка в формуле вида (122) была постоянной в пределах нескольких единиц последней цифры. Для проверки последнего условия обычно определяются начальные значения в большем числе точек (на одну, две), чем это требует разностная схема.

Если в разностной схеме вида (97) $b_0 \neq 0$, $a_0 \neq 0$ (например, случай интерполяционной формулы Адамса), то для нахождения y_j получаем нелинейное уравнение вида

$$y_j = \psi(y_j), \tag{128}$$

где

$$\psi(y_j) = -\sum_{l=1}^{m} \tilde{a}_l y_{j-l} + h \sum_{l=0}^{m} \tilde{b}_l f_{j-l}.$$

Значения коэффициентов m-шаговых разностных методов вида (97), решения задачи Коши для уравнения $y'=f(x, y)$

Таблица 7

m	a_i	b_i	Порядок аппроксимации	Примечание
1	$a_0=1$ $a_1=-1$	$b_0=0$ $b_1=1$	1	Формула Эйлера
	$a_0=1$ $a_1=-1$	$b_0=1/2$ $b_1=1/2$	2	Метод трапеций
2	$a_0=1$ $a_1=-1$ $a_2=0$	$b_0=0$ $b_1=3/2$ $b_2=-1/2$	2	Экстраполяционная формула Адамса
	$a_0=1$ $a_1=-1$ $a_2=0$	$b_0=5/12$ $b_1=2/3$ $b_2=-1/12$	3	Интерполяционная формула Адамса
	$a_0=1$ $a_1=0$ $a_2=-1$	$b_0=1/3$ $b_1=4/3$ $b_2=1/3$	4	
3	$a_0=1$ $a_1=-1$ $a_2=0$ $a_3=0$	$b_0=0$ $b_1=23/12$ $b_2=-4/3$ $b_3=5/12$	3	Экстраполяционная формула Адамса
	$a_0=1$ $a_1=-1$ $a_2=0$ $a_3=0$	$b_0=3/8$ $b_1=19/24$ $b_2=-5/12$ $b_3=1/24$	4	Интерполяционная формула Адамса

Для решения уравнения (128) обычно строится итерационный процесс вида

$$y_j^{k+1}=\psi(y_j^k), \qquad k=0,\ 1,\ \ldots \tag{129}$$

За начальное приближение y^0 выбирают значение y_j, найденное по явному m-шаговому или одношаговому методу.

Многошаговые разностные методы без труда распространяются на случай решения задачи Коши для системы уравнений, а значит, и на уравнения высших порядков. Наряду с односторонними многошаговыми методами можно строить и двусторонние многошаговые разностные методы.

В многошаговых методах при переходе от шага к шагу в значительной мере уменьшается, по сравнению с одношаговыми методами, необходимость дополнительной информации о правой части уравнения. Имеется возможность повторно использовать одну и ту же информацию. Основной недостаток этих методов связан с проблемой построения начала таблицы.

Оценка погрешности многошаговых методов. Прежде чем переходить к оценке погрешности многошаговых методов, сделаем некоторые замечания относительно устойчивости счета по формулам многошагового метода вида (97).

Рассмотрим задачу Коши для уравнения

$$y' = 0, \quad y(x_0) = y_0. \tag{130}$$

Будем применять к решению задачи (130) m-шаговый метод.

Пусть

$$\eta_k = y(x_k) - y_k, \qquad k = 0, 1, \ldots, \tag{131}$$

где $y(x_k)$ — точное решение задачи Коши, y_k — приближенное значение, найденное по формуле (97). Будем пренебрегать погрешностями округления, которые возникают при счете по формуле (97). Тогда, вычитая из (96) соотношение (97), получим разностное уравнение для η_k

$$\sum_{k=0}^{m} a_k \eta_{i-k} = 0 \tag{132}$$

($R_i(h) = 0$ при $f(x, y) = 0$).

Решение однородного разностного уравнения (132) с постоянными коэффициентами выражается через корни характеристического уравнения

$$\Lambda(z) = \sum_{k=0}^{m} a_k z^{i-k} = 0, \tag{133}$$

и для того чтобы η_k были ограниченными при $k \to \infty$, все корни уравнения (133) должны быть расположены в единичном круге $|z| \leqslant 1$ плоскости комплексной переменной z и на границе единичной окружности $|z| = 1$ не должно быть кратных корней. Это условие называют α-условием. Таким образом, из m-шаговых разностных методов имеет смысл рассматривать только такие формулы вида (97), для которых корни характеристического уравнения (133) удовлетворяют указанным выше условиям.

С этой точки зрения разностная схема вида (117) должна быть забракована, так как квадратный трехчлен

$$\Lambda(z) = z^2 + 4z - 5$$

имеет корни $z_1 = 1$, $z_2 = 5 > 1$. Все разностные схемы, приведенные в табл. 7, удовлетворяют α-условию. Это условие накладывает определенные ограничения на построение m-шаговых разностных схем высокого порядка аппроксимации p, так как в этом случае среди корней характеристического уравнения появляются корни по модулю больше единицы. Доказано, что при $p > m + 2$ в случае неявной схемы и при $p > m$ в случае явной схемы среди корней характеристического уравнения (133) имеется корень по модулю больше единицы. В дальнейшем условимся рассматривать такие разностные схемы, которые удовлетворяют α-условию.

Для оценки величины $\eta_k = y(x_k) - y_k$ в результате вычитания из (96) соотношения (97) при $f(x, y) \not\equiv 0$ получаем следующую

разностную схему:

$$\sum_{k=0}^{m} a_k \eta_{j-k} - h \sum_{k=0}^{m} b_k [f(x_{j-k}, y(x_{j-k}) - f_{j-k}] = hR_j(h).$$

На основании теоремы Лагранжа о конечном приращении последнее равенство можно записать в виде

$$\sum_{k=0}^{m} a_k \eta_{j-k} - h \sum_{k=0}^{m} b_k F_{j-k} \eta_{j-k} = hR_j(h), \tag{134}$$

где

$$F_{j-k} = \frac{\partial f(x_{j-k}, y_{j-k} + \theta_{j-k}\eta_{j-k})}{\partial y}, \quad 0 < \theta_{j-k} < 1.$$

Разрешим (134) относительно η_j

$$\eta_j = \sum_{k=1}^{m} \frac{hb_k F_{j-k} - a_k}{a_0 - hb_0 F_j} \eta_{j-k} + \frac{hR_j}{a_0 - hb_0 F_j}. \tag{135}$$

Соотношение (135) перепишем в виде

$$\eta_j = \sum_{k=1}^{m} \left(\frac{hb_k F_{j-k} - a_k}{a_0 - hb_0 F_j} + \frac{a_k}{a_0} - \frac{a_k}{a_0} \right) \eta_{j-k} + \frac{hR_j}{a_0 - hb_0 F_j}. \tag{136}$$

В векторно-матричном виде (136) можно представить следующим образом:

$$\boldsymbol{H}_j = hC\boldsymbol{H}_{j-1} + A\boldsymbol{H}_{j-1} + \boldsymbol{q}_j, \tag{137}$$

где

$$\boldsymbol{H}_j = (\eta_j, \eta_{j-1}, \ldots, \eta_{j-m+1})', \quad \boldsymbol{q}_j = \left(\frac{hR_j}{a_0 - hb_0 F_j}, 0, \ldots, 0 \right)' - \tag{138}$$

m-мерные векторы,

$$C = \begin{pmatrix} c_{1j} & c_{2j} & \ldots & c_{mj} \\ 0 & 0 & \ldots & 0 \\ 0 & \ldots & \ldots & 0 \end{pmatrix}, \quad c_{kj} = \frac{hb_k F_{j-k} - a_k}{h(a_0 - hb_0 F_j)} + \frac{a_k}{ha_0},$$

$$A = \begin{pmatrix} -\frac{a_1}{a_0} & -\frac{a_2}{a_0} & \ldots & -\frac{a_m}{a_0} \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 \end{pmatrix} \tag{139}$$

матрицы порядка $m \times m$.

Характеристический многочлен матрицы A,

$$P(\lambda) = |A - \lambda E| \tag{140}$$

с точностью до множителя $(-1)^m$ записывается в виде

$$\lambda^m + \frac{a_1}{a_0}\lambda^{m-1} + \frac{a_2}{a_0}\lambda^{m-2} + \cdots + \frac{a_m}{a_0} = 0$$

и с точностью до const пропорционален характеристическому уравнению (133) разностной схемы (97). По предположению все корни уравнения (140) лежат в единичном круге $|z| \leqslant 1$ и на границе единичного круга нет кратных корней. Поэтому существует такая неособая матрица V, что $V^{-1}AV = D$ и $\|D\|_1 \leqslant 1$.

Умножим (137) на V^{-1} и обозначим

$$V^{-1}H_j = \hat{H}_j, \quad V^{-1}g_j = \hat{g}_j,$$
$$\hat{H}_j = hV^{-1}CV\hat{H}_{j-1} + D\hat{H}_{j-1} + \hat{g}_j \tag{141}$$

или

$$\hat{H}_j = (hV^{-1}CV + D)\,\hat{H}_{j-1} + \hat{g}_j.$$

Откуда

$$\|\hat{H}_j\|_1 \leqslant (h\|V^{-1}\|_1\|C\|_1\|V\|_1 + 1)\|\hat{H}_{j-1}\|_1 + \|\hat{g}_j\|_1. \tag{142}$$

Обозначим

$$\|V^{-1}\|_1\|C\|_1\|V\|_1 = \beta, \quad \|\hat{g}_j\|_1 = \gamma_j. \tag{143}$$

Если $|f_y| \leqslant L$ при $x_0 \leqslant x \leqslant \zeta$, то

$$\|C\|_1 = \sum_{k=1}^{m} |c_{kj}| = \sum_{k=1}^{m} \left| \frac{b_k F_{j-k} a_0 - a_k b_0 F_j}{(a_0 - hb_0 F_j)\, a_0} \right| \leqslant \sum_{k=1}^{m} 2L\, \frac{|b_k a_0| + |a_k b_0|}{|a_0|^2}$$

и

$$\beta \leqslant 2\|V^{-1}\|_1\|V\|_1 L \sum_{k=1}^{m} \frac{|a_k b_0| + |b_k a_0|}{|a_0|^2}, \tag{144}$$

$$\gamma_j = \|\hat{g}_j\|_1 \leqslant \|V^{-1}\|_1\, 2\, \frac{h|R_j|}{|a_0|}. \tag{145}$$

Из рекуррентного соотношения (142), учитывая (144), (145), имеем:

$$\|\hat{H}_j\|_1 \leqslant (1 + h\beta)\|\hat{H}_{j-1}\|_1 + \gamma_j \leqslant (1 + h\beta)\,[(1 + h\beta)\|\hat{H}_{j-2}\|_1 + \gamma_{j-1}] +$$
$$+ \gamma_j = (1 + h\beta)^2\|\hat{H}_{j-2}\|_1 + (1 + h\beta)\,\gamma_{j-1} + \gamma_j \leqslant \ldots$$
$$\ldots \leqslant (1 + h\beta)^{j-m+1}\|\hat{H}_{m-1}\|_1 + \sum_{l=m}^{j} (1 + h\beta)^{j-l}\gamma_l,$$

или

$$\|\hat{H}_j\|_1 \leqslant \exp \beta h\,(j - m + 1)\|\hat{H}_{m-1}\|_1 + \sum_{l=m}^{j} \exp\,[\beta h\,(j - l)]\,\gamma_l \leqslant$$
$$\leqslant \exp \beta\,(\xi - x_0)\left(\|\hat{H}_{m-1}\|_1 + \sum_{l=m}^{j} \gamma_l\right).$$

Далее

$$|\eta_j| \leqslant \|H_j\|_1$$

и, учитывая (141), имеем:

$$\|\hat{H}_{m-1}\|_1 \leqslant \|V^{-1}\|_1\|H_{m-1}\|_1 = \|V^{-1}\|_1 \max_{0 \leqslant k < m-1} |\eta_k|,$$
$$\|H_j\|_1 \leqslant \|V\|_1\|\hat{H}_j\|_1.$$

Поэтому

$$|\eta_j| \leqslant \|V\|_1 \|\hat{H}_j\|_1 \leqslant \|V\|_1 \exp\beta(\xi - x_0)\left[\|V^{-1}\|_1 \max_{0\leqslant k<m-1} |\eta_k| + \sum_{l=m}^{j} \gamma_l\right],$$

$$|\eta_j| \leqslant \exp\beta(\xi - x_0)\, \varkappa_1 \left(\max_{0\leqslant k<m-1} |\eta_k| + 2h \sum_{l=m}^{j} \frac{|R_l|}{|a_0|}\right), \tag{146}$$

где

$$\|V\|_1 \|V^{-1}\|_1 \leqslant \varkappa_1.$$

Если принять во внимание погрешности округления, которые возникают при счете по разностной схеме (97), и обозначить их через Δ_j, то вместо (97) будем иметь:

$$\sum_{k=0}^{m} a_k y_{j-k} - h \sum_{k=0}^{m} b_k f_{j-k} = \Delta_j \tag{97'}$$

и поэтому правая часть разностного уравнения (134) для η_j будет равна $hR_j(h) + \Delta_j$ и для γ_j вместо (145) будет иметь место оценка

$$\gamma_j \leqslant \|V^{-1}\| \left(\frac{2h|R_j|}{|a_0|} + |\Delta_j|\right),$$

поэтому неравенство (146) примет вид

$$|\eta_j| \leqslant \exp\beta(\zeta - x_0)\, \varkappa_1 \left[\max_{0\leqslant k\leqslant m-1} |\eta_k| + \frac{2h}{|a_0|} \sum_{l=m}^{j} |R_l| + \sum_{l=m}^{j} |\Delta_l|\right]. \tag{147}$$

Из соотношения (147) следует, что $|\eta_j| \to 0$ $(j > m)$, если $|f_y| < L$ при $x_0 \leqslant x \leqslant \xi$, разностная схема удовлетворяет условию α и выполняются условия:

$$\max_{0\leqslant k<m-1} |\eta_k| \to 0, \quad h \sum_{l=m}^{j} |R_l| \to 0, \quad \sum_{l=m}^{j} |\Delta_l| \to 0. \tag{148}$$

Хотя оценка (148) является грубой, однако она довольно ясно указывает, что погрешности вычисления начальных данных, округления, длина отрезка, на котором ищется решение, оказывают существенное влияние на величину η_j. Причем, если величины

$$\max_{0\leqslant k<m-1} |\eta_k|, \quad h \sum_{l=m}^{j} |R_l|$$

с уменьшением шага h стремятся к нулю, то поведение величины $\sum_{l=m}^{j} |\Delta_l|$ при $h \to 0$ определяется поведением величины $\sum_{l=m}^{j} \frac{|\Delta l|}{h}$, а поэтому с уменьшением h точность вычислений Δ_l по формуле (97) нужно повышать.

§ 2. ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Под краевой задачей для системы обыкновенных дифференциальных уравнений

$$\boldsymbol{Y}'(x) = \boldsymbol{F}(x, \boldsymbol{Y}(x)) \tag{1}$$

понимают задачу интегрирования системы (1) на отрезке $[a, b]$ (отрезок $[a, b]$ может быть и бесконечным) при дополнительных условиях

$$\varphi(Y(x_1), Y(x_2), \ldots, Y(x_k)) = d, \tag{2}$$

заданных в k ($k \geqslant 2$) различных точках отрезка $[a, b]$,

$$a \leqslant x_1 < x_2 < \cdots < x_k \leqslant b. \tag{3}$$

Здесь $Y(x)$, $F(x, Y)$, $\varphi(Y(x_1), \ldots, Y(x_n))$ — вектор-функция размерности n

$$Y(x) = (y_i(x))'_{i=\overline{1,n}}, \quad F(x, Y) = (F_i(x, Y))'_{i=\overline{1,n}},$$

$$\varphi(Y(x_1), Y(x_2), \ldots, Y(x_k)) = (\varphi_i(x, y_1(x_1), \ldots, y_n(x_1), \ldots$$

$$\ldots, y_1(x_k), \ldots, y_n(x_k)))'_{i=\overline{1,n}},$$

где $d = (d_i)'_{i=\overline{1,n}}$ — заданный вектор.

Из многоточечных задач вида (1), (2) особо выделяется группа двухточечных линейных задач вида

$$Y' - A(x)Y = F(x), \tag{4}$$

$$B_1 Y(a) + B_2 Y(b) = d. \tag{5}$$

Здесь $Y(x)$, $F(x)$, d — n-мерные векторы,

$$A(x) = (a_{ij}(x))_{i=\overline{1,n}}^{j=\overline{1,n}}, \quad B_1 = (b_{ij}^{(1)})_{i=\overline{1,n}}^{j=\overline{1,n}}, \quad B_2 = (b_{ij}^{(2)})_{i=\overline{1,n}}^{j=\overline{1,n}}$$

— матрицы размерности $n \times n$.

Очевидно, любая двухточечная краевая задача для линейного дифференциального уравнения n-го порядка

$$y^{(n)}(x) = \sum_{i=1}^{n} p_i(x) y^{(n-i)}(x) + f(x), \tag{6}$$

$$s_i = \gamma_i \quad (i = \overline{1, n}), \tag{7}$$

где

$$s_i = \sum_{j=0}^{n-1} (\alpha_{ij} y^{(j)}(a) + \beta_{ij} y^{(j)}(b)), \tag{8}$$

может быть записана в виде (4), (5).

Последнее утверждение становится очевидным, если ввести следующие обозначения:

$$Y(x) = (y(x), y'(x), y''(x), \ldots, y^{(n-1)}(x))',$$

$$F(x) = (0, 0, \ldots, 0, f(x))', \quad d = (\gamma_i)'_{i=\overline{1,n}},$$

$$A(x) = \begin{pmatrix} 0 & 1 & & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \\ p_n & p_{n-1} & p_{n-2} & \ldots & p_1 \end{pmatrix},$$

$$B_1 = (\alpha_{ij})_{i=\overline{1,n}}^{j=\overline{1,n}}, \quad B_2 = (\beta_{ij})_{i=\overline{1,n}}^{j=\overline{1,n}}.$$

Отметим также, что краевая задача (4), (5) может быть сведена к решению однородной линейной системы дифференциальных уравнений с неоднородными краевыми условиями, т. е. к решению следующей задачи:

$$Z'(x) = M(x)\,Z(x), \tag{9}$$

$$\hat{B}_1 Z(a) + \hat{B}_2 Z(b) = \hat{d}. \tag{10}$$

Здесь $Z(x) = (Y(x),\ 1)'$, $\hat{d} = (d,\ 1)$ — $(n+1)$-мерные векторы.

$$M(x) = \begin{pmatrix} A(x) & F(x) \\ 0 & 0 \end{pmatrix}, \quad \hat{B}_1 = \begin{pmatrix} B_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{B}_2 = \begin{pmatrix} B_2 & 0 \\ 0 & 1 \end{pmatrix}$$

— матрицы размерности $(n+1) \times (n+1)$.

Для многоточечных ($k \geqslant 2$) краевых задач по сравнению с задачей Коши значительно сложнее проводятся исследования, связанные с вопросами существования решения и построения приближенных методов его определения.

Большинство методов решения основано на сведении нелинейной задачи к нелинейной системе алгебраических уравнений. Для решения полученной нелинейной системы могут быть использованы различные приближенные алгоритмы. Сходимость этих алгоритмов существенно зависит от выбора начального приближения. Поэтому при практическом решении конкретных нелинейных задач предполагается специальный метод получения начального приближения.

1. Метод редукции к задачам Коши

Некоторые сведения о разрешимости краевой задачи (1), (2) можно получить из разрешимости задачи Коши для уравнения (1) с начальным условием

$$Y(a) = Y_0. \tag{11}$$

В самом деле, пусть решение задачи Коши (1), (11) существует и единственно при любом $Y_0 \in \Omega$, где Ω — некоторая область n-мерного векторного пространства. Тогда при каждом фиксированном $x_i \in [a,\ b]$ решение задачи Коши (1), (11) будет определять некоторый вектор

$$Y(x_i) = \omega(x_i,\ Y_0). \tag{12}$$

Подставив (12) в краевое условие (2), получим:

$$\varphi(Y_0,\ \omega(x_2,\ Y_0),\ \omega(x_3,\ Y_0),\ \ldots,\ \omega(x_k,\ Y_0)) = d. \tag{13}$$

Соотношение (13) представляет собой систему нелинейных уравнений относительно Y_0. Следовательно, решение нелинейной задачи (1), (2) эквивалентно решению системы (13) и задачи Коши для уравнения (1) при начальном условии

$$Y(a) = Y_0,$$

где Y_0 — решение системы (13). Поэтому количество решений нелинейной краевой задачи (1), (2) будет совпадать с количеством решений

системы (13). Если система (13) неразрешима, то нелинейная краевая задача не будет иметь решения.

Основная трудность при таком подходе связана с определением Y_0, ибо система нелинейных уравнений (13) не выписывается явно, так как не известно явно выражение (12).

Построены различные алгоритмы приближенного вычисления левой части уравнения (13). Однако эти алгоритмы дают сходимость к решению только при достаточно хорошем начальном приближении.

Линейные системы. Сравнительно просто такой подход реализуется в случае линейной краевой задачи (9), (10). В этом случае решение задачи Коши для уравнения (9) с начальным условием $Z(a)$ представляется в виде

$$Z(x) = \Gamma(x) Z(a), \tag{14}$$

где $\Gamma(x)$ — фундаментальная матрица системы (9), т. е. решение задачи Коши является линейной функцией относительно начального вектора $Z(a)$. Подставляя (14) в краевое условие (10), получим:

$$[\hat{B}_1 + \hat{B}_2\Gamma(b)] Z(a) = \hat{d}. \tag{15}$$

Соотношение (15) представляет собой систему линейных алгебраических уравнений относительно вектора $Z(a)$ с определителем, отличным от нуля (предположив противное, получили бы, что однородная краевая задача имеет ненулевое решение). Если $Z(a)$ — решение системы (15), то вектор-функция (14) будет удовлетворять уравнению (9), граничным условиям (10), т. е. являться решением задачи (9), (10). Этот метод определения решения линейной задачи (9), (10) получил название *метода стрельбы* (или метода дополнительных функций).

Таким образом, на основании изложенного, вычислительную схему метода стрельбы можно описать следующим образом.

Пусть e_i означает i-й координатный орт, т. е. $(n+1)$-мерный вектор, у которого i-я компонента равна единице, а все остальные нули

$$e_i = (0,\ 0,\ \ldots,\ 0,\ 1,\ 0,\ \ldots,\ 0)'. \tag{16}$$

Для построения фундаментальной матрицы $\Gamma(x)$ численным интегрированием находим решение $(n+1)$-й задачи Коши вида

$$Z_i'(x) = M(x) Z_i(x), \tag{17}$$

$$Z_i(a) = e_i \quad (i = \overline{1,\ n}) \tag{18}$$

на отрезке $[a,\ b]$.

Здесь $(n+1)$-мерный вектор-функция $Z_i(x)$ является решением i-й начальной задачи Коши (17), (18). Тогда

$$\Gamma(b) = (Z_1(b),\ Z_2(b),\ \ldots,\ Z_{n+1}(b)). \tag{19}$$

Решая систему (15), находим $Z(a)$, и, подставляя $\Gamma(x)$ и $Z(a)$ в (14), находим решение задачи (9), (10).

Метод стрельбы несколько упрощается, если краевые условия (10) разделены, т. е. рассмотрим задачу вида

$$Y'(x) = A(x) Y(x) + F(x),$$

$$D_1 Y(a) = \omega_1, \tag{20}$$

$$D_2 Y(b) = \omega_2, \tag{21}$$

где $D_1 = (d_{ij}^{(1)})_{i=\overline{1,n-k}}^{j=\overline{1,n}}$, $D_2 = (d_{ij}^{(2)})_{i=\overline{1,k}}^{i=\overline{1,n}}$ — заданные прямоугольные матрицы соответственно порядков $(n-k) \times n$ и $k \times n$ $(k < n)$, причем предполагается, что ранг D_1 равен $n-k$, а ранг D_2 равен k; ω_1, ω_2 — заданные n-мерные векторы, остальные обозначения имеют тот же смысл, что и раньше.

Тогда метод стрельбы заключается в следующем.

Находим вектор Y_0 такой, что

$$D_1 Y_0 = \omega_1, \tag{22}$$

т. е. Y_0 удовлетворяет краевому условию на левом конце, и строим полную систему линейно-независимых векторов ${}^{(i)}Y$ $(i = \overline{1, k})$ — решений уравнения

$$D_1^{(i)} Y = 0. \tag{23}$$

Их будет k, так как ранг матрицы D_1, по предположению, равен $n-k$.

Затем численным интегрированием решаем k задач Коши для однородного уравнения

$$Y_i'(x) = A(x)\, Y_i(x), \tag{24}$$

$$Y_i(a) = {}^{(i)}Y, \quad i = \overline{1, k} \tag{25}$$

и задачу Коши для неоднородного уравнения

$$Y'(x) = A(x)\, Y(x) + F(x), \tag{26}$$

$$Y(a) = Y_0. \tag{27}$$

Тогда всякое решение уравнения (4), удовлетворяющее граничному условию (20), можно представить в виде

$$Y(x) = \sum_{i=1}^{k} c_i Y_i(x) + Y_{k+1}(x), \tag{28}$$

где c_i — постоянные интегрирования, которые подлежат определению из краевых условий (21); $Y_i(x)$ $(i = \overline{1, k})$ — решение задач (24), (25); $Y_{k+1}(x)$ — решение задачи (26), (27). Для определения c_i получаем систему линейных алгебраических уравнений

$$D_2\left(\sum_{i=1}^{k} c_i Y_i(b) + Y_{k+1}(b)\right) = \omega_2, \tag{29}$$

с определителем, отличным от нуля.

Если c_i $(i = \overline{1, k})$ — решение системы (29), то вектор-функция (28) будет являться решением исходной задачи (4), (20), (21). Для того чтобы не хранить всю информацию о $Y_i(x)$ $(i = \overline{1, k+1})$, можно поступить следующим образом: находим

$$Y(a) = \sum_{i=1}^{k} c_i Y_i(a) + Y_{k+1}(a) \tag{30}$$

и для определения искомого вектора $Y(x)$ численно решаем задачу Коши (4), (30). Искомое решение $Y(x)$ принадлежит многообразию L_k, которое определяется заданием векторов $Y_i(x)$, являющихся решением однородной системы (24). Если среди $Y_i(x)$ есть быстрорастущие с ростом x, то при продвижении к точке b происходит как бы «сплющивание» базисных векторов $Y_i(x)$, на которые натянуто многообразие L_k, а значит, значение $Y(b)$ не может быть определено с достаточной точностью. Иными словами, в методе стрельбы вычислительная погрешность, допущенная при определении $Y_i(x)$, может оказать большое влияние на результат вычислений. Чтобы избежать «сплющивания» векторов $Y_i(x)$, на которые натянуто многообразие L_k, их можно ортогонализировать. Суть метода ортогонализации состоит в следующем.

Разбиваем отрезок $[a, b]$ на части точками интегрирования x_i $(i = \overline{0, N})$, среди которых выберем точки ортогонализации X_i $(i = \overline{1, M})$, $X_M = b$. Выбор указанных точек X_i обычно обусловливается степенью требуемой точности решения задачи. Решение задач Коши (24) — (27) в точке X_i, получаемое при «прямом» ходе, обозначим через $U_r(x_i)$ $(r = \overline{1, k+1})$. Очевидно, в точке $x = a$ векторы $U_1, U_2, \ldots, U_{k+1}$ совпадают с векторами $Y_i(a)$ из (28). В каждой точке X_i ортонормируем векторы

$$U_r \quad (r = \overline{1, k})$$

и обозначим их через Z_r $(r = \overline{1, k})$. Очевидно,

$$Z_r = \frac{1}{\delta_{rr}}\left(U_r - \sum_{j=1}^{r-1} \delta_{rj} Z_j\right), \tag{31}$$

$$\delta_{rj} = (U_r, Z_j) \quad j < r, \tag{32}$$

$$\delta_{rr} = \sqrt{(U_r, U_r) - \sum_{j=1}^{r-1} \delta_{rj}}\,. \tag{33}$$

Вектор Z_{k+1} не нормируется и вычисляется по формуле

$$Z_{k+1} = U_{k+1} - \sum_{j=1}^{k} \delta_{k+1,j} Z_j. \tag{34}$$

Значения $Z_r(X_i)$ являются начальными для получения решений задачи Коши при $x \in [X_i, X_{i+1}]$.

При проведении ортогонализации на каждом шаге вырабатывается матрица треугольного вида

$$\Delta^{(s)} = \begin{Vmatrix} \delta_{11}^s & 0 & \ldots & \ldots & 0 \\ \delta_{21}^s & \delta_{22}^s & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \delta_{M1}^s & \ldots & \ldots & \delta_{MM}^s & 0 \\ \delta_{M+1,1}^s & \ldots & \ldots & \delta_{M+1,M}^s & 1 \end{Vmatrix}. \tag{35}$$

Для определения k-мерного вектора $C^{(M)} = (C_j^{(M)})_{j=\overline{1,k}} = (c_1(b), c_2(b), \ldots, c_k^{\cdot}(b))$ воспользуемся соотношением (29)

$$D_2\left(\sum_{j=1}^{k} c_j^{(M)} Z_j(b) + Z_{k+1}(b)\right) = \omega_2. \tag{36}$$

Из системы (36) находим $C^{(M)}$.

Если нас интересует значение $Y(x)$ в точке X_s, то его можно представить в виде

$$Y(X_s) = \sum_{j=1}^{k} c_j^{(s)} Z_j(X_s) + Z_{k+1}(X_s).$$

Значит, если будут известны $C_j^{(s)}$, то легко определить и $Y(X_s)$. Но $c_j^{(s)}$ определяется из рекуррентного соотношения

$$(\Delta^{s+1})' \hat{C}^{(s)} = \hat{C}^{(s+1)}, \tag{37}$$

где $\hat{C}^{(s)} = (c_1^{(s)}, c_2^{(s)}, \ldots, c_k^{(s)}, 1)$ — $(k+1)$-мерный вектор или

$$\hat{C}^{s+l} = \left(\prod_{i=1}^{l} \Delta^{s+j}\right)' \hat{C}^{(s)}.$$

Следовательно, для определения $\hat{C}^{(s)}$ достаточно решить систему линейных алгебраических уравнений (37) с треугольной матрицей. Обозначим через

$$\Phi(X_s) = (Z_1(X_s), Z_2(X_s), \ldots, Z_{k+1}(X_s)) \tag{38}$$

матрицу размерности $n \times (k+1)$. Тогда после определения вектора $\hat{C}^{(s)}$ искомый вектор-функция $Y(X_s)$ в точке X_s находится по формуле

$$Y(X_s) = \Phi(X_s)\hat{C}^{(s)}. \tag{39}$$

С точки зрения реализации описанного алгоритма следует заметить:

1) с целью экономии времени счета и памяти желательно в качестве начальной точки интегрирования выбирать такую, в которой задано большее число граничных условий;

2) хранить матрицы $\Phi(X_s)$ и $\Delta^{(s)}$ нужно не во всех точках ортогонализации, а лишь в тех, в которых выдаются результаты;

3) число точек ортогонализации в зоне ожидаемых локальных эффектов следует увеличить. Метод ортогональной прогонки хорошо зарекомендовал себя при решении практических задач.

Для решения линейных краевых задач вида (9), (10) с одной и той же левой частью и различными столбцами свободных членов $F(x)$ может быть использован приближенный метод, который получил название *метода сопряженных уравнений*.

Предположим, что соотношение

$$\hat{B}_1 Y(a) + \hat{B}_2(x) Y(x) = \hat{d} \tag{40}$$

выполняется во всех точках промежутка $[a, b]$. Тогда, определив $\hat{B}_2(a)$, можем свести краевую задачу (9), (10) к задаче Коши

$$Y' - M(x)Y = 0,$$
$$Y(a) = Y_0, \tag{41}$$

где Y_0 определяется из уравнения

$$[\hat{B}_1 + \hat{B}_2(a)]Y(a) = \hat{d}. \tag{42}$$

Дифференцируя (40), получим:

$$\frac{d\hat{B}_2(x)}{dx}Y(x) + \hat{B}_2(x)\frac{dY(x)}{dx} = 0.$$

Если учесть (9), то

$$\left(\frac{d\hat{B}_2(x)}{dx} + \hat{B}_2(x)M(x)\right)Y(x) = 0 \tag{43}$$

или, так как соотношение (43) должно выполняться в любой точке промежутка $[a, b]$, то

$$\frac{d\hat{B}_2(x)}{dx} + \hat{B}_2(x)M(x) = 0, \quad \frac{d\hat{B}_2^*}{dx} = M^*(x)\hat{B}_2^*(x). \tag{44}$$

Равенство (44) можно представить в виде

$$\frac{d\hat{B}_{2i}(x)}{dx} = -M^*(x)\hat{B}_{2i}(x), \tag{45}$$

где $\hat{B}_{2i}(x)$ — i-я строка матрицы $\hat{B}_2$. Система (45) называется сопряженной с уравнением (9).

Так как соотношение (40) выполняется в точке $x = b$, то

$$\hat{B}_2(b) = \hat{B}_{2i}(b) \quad (i = \overline{1, n+1}), \tag{46}$$

а значит, известно значение $\hat{B}_{2i}(b)$.

Поэтому, решив $(n+1)$ задачу Коши (45), (46), на отрезке $[b, a]$ можно определить значение $\hat{B}_2(a)$. Из (42) находим $Y(a)$, и вместо задачи (9), (10) решаем задачу Коши для уравнения (9) с начальным условием $Y(a)$, найденным из уравнения (42).

Очевидно, для решения линейных краевых задач может быть применен метод сеток (см. гл. 6).

Нелинейные системы. Укажем на некоторые приближенные методы, позволяющие решение задачи (1), (2) свести к решению задачи Коши. Для простоты изложения рассмотрим случай двухточечной краевой задачи

$$Y'(x) = F(x, Y(x))$$
$$\varphi(Y(x_1), Y(x_2)) = d. \tag{47}$$

Среди приближенных методов, редуцирующих нелинейную краевую задачу (1), (47) к решению задач Коши, отметим методы, основанные

на линеаризации либо системы (1), (47), либо системы (13) при $k = 2$.

Метод линеаризарии. Основная идея метода линеаризации заключается в следующем. Производится линеаризация системы (1) в пространстве вектор-функций $Y(x)$, а именно решение представляют в виде

$$Y(x) = Y^{(i)}(x) + U^{(i)}(x),$$

где $Y^{(i)}(x)$ — некоторое приближение к искомому решению двухточечной задачи (1), (47). Тогда, предполагая достаточную гладкость функций F и φ и используя формулу Лагранжа, получим:

$$\frac{d\,(Y^{(i)}(x) + U^{(i)}(x))}{dx} = F\,(x,\ Y^{(i)}(x)) + \Phi_F\,(x,\ Z_1^{(i)},\ \ldots,\ Z_n^{(i)})\,U^{(i)}(x),$$

$$\varphi\,(Y^{(i)}(a) + U^{(i)}(a),\ Y^{(i)}(b) + U^{i}(b)) = \varphi\,(Y^{(i)}(a),\ Y^{(i)}(b)) +$$
$$+ \Phi_0\,(\omega_1^{(i)}(a),\ \ldots,\ \omega_n^{(i)}(a),\ \omega_1^{(i)}(b),\ \omega_2^{(i)}(b),\ \ldots,\ \omega_n^{(i)}(b))\,U^{(i)}(a) +$$
$$+ \Phi_1\,(\omega_1^{(i)}(a),\ \ldots,\ \omega_n^{(i)}(a),\ \omega_1^{(i)}(b),\ \omega_2^{(i)}(b),\ \ldots,\ \omega_n^{(i)}(b))U^{(i)}(b).$$

Здесь $\Phi_F\,(x,\ Z_1^{(i)},\ \ldots,\ Z_n^{(i)}) = \left(\dfrac{\partial F_i\,(x_i,\ z_{i1},\ z_{i2},\ \ldots,\ z_{in})}{\partial y_j}\right)_{i=\overline{1,n}}^{j=\overline{1,n}}$ — матрица Якоби вектор-функции $F\,(x,\ Y(x))$;

$$\Phi_0\,(\omega_1^{(i)}(a),\ \ldots,\ \omega_n^{(i)}(b)) =$$
$$= \left(\frac{\partial\varphi\,(\omega_{i1}(a),\ \omega_{i2}(a),\ \ldots,\ \omega_{in}(a),\ \omega_{i1}(b),\ \ldots,\ \omega_{in}(b))}{\partial y_k(a)}\right)_{k=\overline{1,n}},$$

$$\Phi_1\,(\omega_1^{(i)}(a),\ \ldots,\ \omega_n^{(i)}(b)) = \left(\frac{\partial\varphi\,(\omega_{i1}(a),\ \ldots,\ \omega_{in}(a),\ \omega_{i1}(b),\ \ldots,\ \omega_{in}(b))}{\partial y_k(b)}\right)_{k=\overline{1,n}}$$

— матрицы Якоби вектор-функции $\varphi\,(Y(a),\ Y(b))$, а векторы $Z_k^{(i)}(x)$, $\omega_k^{(i)}(a)$, $\omega_k^{(i)}(b)$ удовлетворяют неравенствам:

$$\|Z_k^{(i)}(x) - Y^{(i)}(x)\| \leqslant \|U^{(i)}(x)\|,$$
$$\|\omega_k^{(i)}(a) - Y^{(i)}(a)\| \leqslant \|U^{(i)}(a)\|,$$
$$\|\omega_k^{(i)}(b) - Y^{(i)}(b)\| \leqslant \|U^{(i)}(b)\|.$$

Краевая задача (1), (47) приближенно заменяется следующей линейной краевой задачей:

$$\frac{dU^{(i)}(x)}{dx} = \Phi_F\,(x,\ Y^{(i)}(x))\,U^{(i)}(x) + F\,(x,\ Y^{(i)}(x)) - \frac{dY^{(i)}(x)}{dx}\ ; \quad (48)$$

$$\Phi_0\,(Y^{(i)}(a),\ Y^{(i)}(b))\,U^{(i)}(a) + \Phi_1\,(Y^{(i)}(a),\ Y^{(i)}(b))\,U^{(i)}(b) =$$
$$= d - \varphi\,(Y^{(i)}(a),\ Y^{(i)}(b)). \quad (49)$$

Решение задачи (48), (49) может быть найдено путем редукции к задаче Коши. Значение $U^{(i)}(x)$ будет являться поправкой к исходному начальному вектору $Y^{(i)}(x)$, а значит,

$$Y^{(i+1)}(x) = Y^{(i)}(x) + U^{(i)}(x),$$

где $U^{(i)}(x)$— решение задачи (48), (49), будет являться новым приближением к искомому решению.

Аналогично можно линеаризировать уравнение (13). Однако, если при линеаризации соотношений (1), (47) получаем матрицы Якоби от известных вектор-функций $F(x, Y, (x))$ и $\varphi(Y(a), Y(b))$, то при линеаризации (13) значения

$$\frac{\partial\varphi(Y(a)\,\omega(b, Y(a))}{\partial y_i(a)} \quad (i=\overline{1, n})$$

будут неизвестны. В этом случае поступают следующим образом. Обозначим

$$\varphi(Y(a), \omega(b, Y(a)) = \psi(Y(a)) \tag{50}$$

и значения $\frac{\partial\psi}{\partial y_i(a)}$ заменим разностными выражениями вида

$$\frac{\partial\psi(Y^{(i)}(a))}{\partial y_k(a)} = \frac{\psi(Y^{(i)}(a)+h_i e_k)-\psi(Y^{(i)}(a))}{h_i}, \tag{51}$$

где e_k $(k=\overline{1, n})$ — k-й координатный орт, $e_0=0$, $0<h_i\leqslant h_0$ — некоторые постоянные.

Так как

$$\psi(Y^{(i)}(a)+h_i e_k) = \varphi(Y^{(i)}(a)+h_i e_k, \omega(b, Y^{(i)}(a)+h_i e_k)) \tag{52}$$

и вектор-функция $\omega(b, Y^{(i)}(a)+h_i e_k)$ — есть решение задачи Коши для уравнения (1) в точке b при начальном условии $Y^{(i)}(a)+h_i e_k$, то для определения $\omega(b, Y^{(i)}(a)+h_i e_k)$ нужно решить $(n+1)$ задачу Коши вида

$$\begin{cases} \dfrac{dY_k^{(i)}(x)}{dx} = F(x, Y_k^{(i)}(x)), & (53)\\ Y_k^{(i)}(a) = Y^{(i)}(a)+h_i e_k, \quad k=\overline{0, n}, & (54)\end{cases}$$

где $Y^{(i)}(a)$ — некоторое приближение к $Y(a)$, выбор которого находится в нашем распоряжении. Определив $Y_k^{(i)}(a)$ как решение задач Коши (53), (54), обратимся к (50), положив

$$\begin{gathered} Y(a) = Y^{(i)}(a)+U^{(i)}(a),\\ \psi(Y(a)) \approx \psi(Y^{(i)}(a)) + \Phi_\psi(Y^{(i)}(a))\,U^{(i)}(a), \end{gathered} \tag{55}$$

где $\Phi_\psi = \left(\frac{\partial\psi(Y^{(i)}(a))}{\partial y_l^i(a)}\right)_{l=\overline{1,n}}$ — матрица Якоби вектор-функции $\psi(Y(a)$. Или, если принять во внимание (51), из (55) получим следующий приближенный алгоритм для определения $(i+1)$ приближения к $Y(a)$:

$$Q_\psi(Y_k^{(i)}(a))\,U^{(i)}(a) = -\varphi(Y_k^{(i)}(a), Y_k^{(i)}(b)) + d, \tag{56}$$

$$Y^{(i+1)}(a) = Y^{(i)}(a) + U^{(i)}(a), \tag{57}$$

$$Q_\psi(Y_k^{(i)}(a)) = \left(\frac{\varphi(Y_k^{(i)}(a), Y_k^{(i)}(b)) - \varphi(Y_{k-1}^{(i)}(a), Y_{k-1}^{(i)}(b))}{h_i}\right)_{k=\overline{1,n}} \tag{58}$$

матрица размерности $n \times n$, $Y_k^{(l)}(x)$ — решение задач Коши (53), (54) на отрезке $[a, b]$.

Рассмотрим еще один из подходов редукции нелинейной краевой задачи (1), (47) к задаче Коши.

Продолжение решения по параметру. Суть метода заключается в следующем.

Выбираем произвольный вектор Y_0 и решаем задачу Коши для уравнения (1) при следующих начальных условиях:

$$Y(a) = Y_0. \tag{59}$$

Обозначим решение задачи Коши (1), (59) через $\hat{Y}(x)$. По этому решению находим

$$d_0 = \varphi(\hat{Y}(a), \hat{Y}(b)). \tag{60}$$

Вместо задачи (1), (47) рассмотрим целое семейство задач вида

$$Y' = F(x, Y),$$

$$\varphi(Y(a), Y(b)) = \lambda(d - d_0) + d_0. \tag{61}$$

Если для каждого $\lambda \in [0, 1]$ существует решение $Y(\lambda)$ задачи (1), (61), непрерывно зависящее от λ, то $Y(1)$ совпадает с решением исходной задачи. При $\lambda = 0$ решение задачи (1), (61) совпадает с решением задачи Коши (1), (59). Для того чтобы найти $Y(1)$, поступают следующим образом.

Дифференцируя (1), (61) по λ, получаем:

$$\frac{d}{dx}\left(\frac{\partial Y}{\partial \lambda}\right) = \sum_{j=1}^{n} \frac{\partial F}{\partial y_j} \frac{\partial Y_j}{\partial \lambda}, \tag{62}$$

$$\sum_{j=1}^{n} \frac{\partial \varphi}{\partial y_{aj}} \frac{\partial y_{aj}}{\partial \lambda} + \sum_{j=1}^{n} \frac{\partial \varphi}{\partial y_{bj}} \frac{\partial y_{bj}}{\partial \lambda} = d - d_0. \tag{63}$$

Обозначим

$$\frac{\partial Y}{\partial \lambda} = U(x, \lambda), \quad U(a, \lambda) = U_a, \quad U(b, \lambda) = U_b,$$

$$\Phi_F = \left(\frac{\partial F}{\partial y_1}, \frac{\partial F}{\partial y_2}, \ldots, \frac{\partial F}{\partial y_n}\right),$$

$$\Phi_a = \left(\frac{\partial \varphi}{\partial y_{a1}}, \frac{\partial \varphi}{\partial y_{a2}}, \ldots, \frac{\partial \varphi}{\partial y_{an}}\right),$$

$$\Phi_b = \left(\frac{\partial \varphi}{\partial y_{b1}}, \frac{\partial \varphi}{\partial y_{b2}}, \ldots, \frac{\partial \varphi}{\partial y_{bn}}\right),$$

тогда соотношения (62), (63) запишутся следующим образом:

$$\frac{dU}{dx} = \Phi_F(x, Y)\, U, \tag{64}$$

$$\Phi_a(Y(a), Y(b))\, U_a + \Phi_b(Y(a), Y(b))\, U_b = d - d_0. \tag{65}$$

Задача (64), (65) для $U(x, \lambda)$ является линейной краевой задачей. Если решить линейную краевую задачу (64), (65) при $x \in [a, b]$ и задачу Коши

$$\frac{\partial Y}{\partial \lambda} = U(x, \lambda), \tag{66}$$

$$Y(x, a) = \hat{Y}(x), \tag{67}$$

то при $\lambda = 1$ получим решение исходной задачи (1), (47).

Так как правая часть уравнения (67) является функцией параметра λ и для ее определения при каждом значении параметра нужно решать задачу (64), (65), то при решении задачи Коши (66), (67) обычно используются формулы Эйлера.

Рассмотренные алгоритмы приближенного решения нелинейных краевых задач являются модификациями некоторых методов решения нелинейных операторных уравнений (в частности, методов Ньютона и продолжения решения по параметру), примененных к решению операторных уравнений конкретного вида (1), (47). Некоторые результаты по вопросам сходимости рассмотренных в данном пункте приближенных алгоритмов можно найти, например, в [58], [82], [83].

Очевидно, разностные методы решения нелинейных краевых задач (1), (47) будут приводить к системам нелинейных уравнений. Системы такого вида содержат большое число неизвестных, а поэтому их решение обычно очень трудоемко. Для решения задач специального вида, которые приводят к простым вычислительным схемам (типа метода прогонки), можно рекомендовать разностные методы.

Глава 8

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ОПЕРАТОРНЫХ УРАВНЕНИЙ

§ 1. МЕТОД ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

Пусть R — полное метрическое пространство; $\rho(x, y)$ — расстояние между элементами $x, y, \in R$, $\mathfrak{M}$ — замкнутое множество в R. Пусть на $\mathfrak{M}$ задан некоторый нелинейный оператор Q, преобразующий это множество в себя: $Q\mathfrak{M} \subset \mathfrak{M}$. Назовем точку $u \in \mathfrak{M}$ неподвижной точкой оператора Q, если

$$u = Q(u). \tag{1}$$

Таким образом, неподвижные точки оператора Q являются решениями уравнения (1). Рассмотрим задачу нахождения решения уравнения (1) при $u \in \mathfrak{M}$ — некоторому априори задаваемому множеству из R; в частности, $\mathfrak{M}$ может совпадать с R, но обычно в роли $\mathfrak{M}$ выступают сферы $S = \{u : \rho(u, v_0) \leqslant r\}$, $v_0 \in \mathfrak{M}$. Одним из основных методов нахождения решения уравнения (1) являются итерационные методы. Они позволяют найти какое-либо решение исходного уравнения в результате некоторого бесконечного процесса, который называется процессом итераций. Однако итерационные методы не являются

универсальными, так как они позволяют найти решение задачи (1), если выделено множество $\mathfrak{M} \subseteq R$, которое содержит единственное решение уравнения (1), и если можно гарантировать построение некоторой последовательности, сходящейся к решению уравнения (1).

Наиболее простым итерационным методом нахождения решения уравнения (1) с точки зрения его реализации является метод последовательных приближений, или метод простых итераций, который состоит в том, что, задавшись произвольным элементом $u^0 \in \mathfrak{M}$, называемым начальным приближением, строят последовательность $\{u^k\}$ по формуле

$$u^{k+1} = Q(u^k) \quad (k = 0, 1, 2, \ldots). \tag{2}$$

Итерационный процесс такого вида называется методом простой итерации. При этом нужно выяснить вопрос сходимости последовательности (2) к решению уравнения (1) (принадлежности $\mathfrak{M}$).

1. Принцип сжатых отображений

Этот принцип лежит в основе построения сходящихся итерационных методов, а также используется для доказательства существования единственного решения уравнения (1) в $\mathfrak{M}$.

Будем говорить, что оператор Q является оператором сжатия на $\mathfrak{M}$, если для любых $x, y \in \mathfrak{M}$ выполнено условие

$$\rho(Q(x), Q(y)) \leqslant q\rho(x, y), \tag{3}$$

где $0 \leqslant q < 1$ не зависит от x, y.

Теорема 1 (принцип сжатых отображений). *Пусть оператор Q — оператор сжатия на $\mathfrak{M}$ и преобразует замкнутое множество $\mathfrak{M} \subset \mathbf{R}$ само в себя: $Q\mathfrak{M} \subset \mathfrak{M}$, то в $\mathfrak{M}$ существует единственное решение u уравнения* (1), *которое может быть получено как предел последовательности*

$$u^{k+1} = Q(u^k) \quad (k = 0, 1, \ldots), \tag{4}$$

где u^0 — произвольный элемент из $\mathfrak{M}$. Быстрота сходимости последовательности $\{u^k\}$ к решению оценивается неравенствами

$$\rho(u, u^k) \leqslant \frac{q^k}{1-q}\rho(u^1, u^0) \quad (k = 1, 2, \ldots) \tag{5}$$

$$\rho(u, u^k) \leqslant q^k\rho(u, u^0) \quad (k = 1, 2, \ldots). \tag{5'}$$

Доказательство. Согласно (3), (4),

$$\rho(u^{k+1}, u^k) = \rho(Q(u^k), Q(u^{k-1})) \leqslant q\rho(u^k, u^{k-1}) \leqslant \ldots \leqslant q^k\rho(u^1, u^0).$$

При $p > 0$, в силу неравенства треугольников, имеем:

$$\rho(u^{k+p}, u^k) \leqslant \rho(u^{k+p}, u^{k+p-1}) + \rho(u^{k+p-1}, u^{k+p-2}) + \ldots +$$
$$+ \rho(u^{k+1}, u^k) \leqslant (q^{k+p-1} + q^{k+p-2} + \ldots + q^k)\rho(u^1, u^0) \leqslant$$
$$\leqslant \frac{q^k - q^{k+p}}{1-q}\rho(u^1, u^0) \leqslant \frac{q^k}{1-q}\rho(u^1, u^0) \underset{k\to\infty}{\to} 0. \tag{6}$$

Согласно критерию Больцано — Коши, последовательность $\{u^k\}$ сходится к некоторому элементу $u^* \in \mathrm{R}$. Но $\{u^k\} \in \mathfrak{M}$ ($\mathfrak{M}$ — замкнуто), следовательно, $u^* \in \mathfrak{M}$ и $Q(u^*)$ имеет смысл. Переходя к пределу в (6) при $p \to \infty$, получаем:

$$\rho(u^*, u^k) \leqslant \frac{q^k}{1-q} \rho(u^1, u^0).$$

Очевидно,

$$0 \leqslant \rho(u^{k+1}, Q(u^*)) = \rho(Q(u^k), Q(u^*)) \leqslant q\rho(u^k, u^*),$$

но $\rho(u^k, u^*) \to 0$ при $k \to \infty$, $\lim\limits_{k\to\infty} Q(u^k) = Q(u^*)$, $Q(u^k) = u^{(k+1)} \to$ $\to u^*$ при $k \to \infty$. Следовательно,

$$u^* = Qu^*.$$

Но u также является решением уравнения (1), принадлежащим $\mathfrak{M}$. Пусть $\rho(u, u^*) \neq 0$, тогда

$$\rho(u, u^*) = \rho(Q(u), Q(u^*)) \leqslant q\rho(u, u^*),$$

что невозможно, так как $0 \leqslant q < 1$, т. е. $\rho(u, u^*) = 0$, $u = u^*$. Оценка (5) получается из (6) переходом к пределу при $p \to \infty$. Неравенство (5′) вытекает из (3)

$$\rho(u, u^k) = \rho(Q(u), Q(u^{k-1})) \leqslant q\rho(u, u^{k-1}) \leqslant \ldots \leqslant q^k\rho(u, u^0).$$

Отметим, что условие (3) не является необходимым для сходимости простого итерационного процесса.

Если в роли $\mathfrak{M}$ выступает некоторая сфера S, то теорему (1) удобно применять в следующей частной форме.

Теорема 2. *Пусть Q является оператором сжатия на замкнутой сфере $S = \{\rho(u, v_0) \leqslant r\}$ полного метрического пространства* R *и пусть*

$$\rho(Q(v_0), v_0) \leqslant (1-q)r.$$

Тогда в S существует одно и только одно решение уравнения (1), *которое может быть получено как предел последовательности* (4), *где u^0 — произвольный элемент из S. Быстрота сходимости последовательности* (4) *к решению оценивается неравенствами* (5), (5′).

Для доказательства теоремы 2 достаточно проверить включение $QS \subset S$, т. е. оператор $Q(u)$ осуществляет отображение S в себя. Последнее следует из цепочки очевидных неравенств:

$$\rho(Q(u), v_0) \leqslant \rho(Q(u), Q(v_0)) + \rho(Q(v_0), v_0) \leqslant q\rho(u, v_0) + (1-q)r \leqslant r,$$

т. е. $Q(u) \in S$.

Оценки (5), (5′) в общем случае не могут быть улучшены, как показывает пример уравнения (1) с оператором $Q(u) \equiv qu$.

Если справедливо неравенство вида (5), то говорят, что последовательность $\{u^k\}$ сходится к u со скоростью геометрической прогрессии.

О количестве операций и скорости сходимости итерационных процессов. Качество итерационного процесса часто характеризуют при помощи числа арифметических действий $N(\varepsilon)$, необходимых для достижения заданной точности ε, скорости убывания начальной ошибки после проведения k итераций. Большой практический интерес представляет собой также характер зависимости $N(\varepsilon)$ от точности задания априорной информации об исходном операторе, а также вычислительная устойчивость итерационного метода.

Неравенство (5) позволяет определить, сколько нужно найти последовательных приближений, чтобы решить уравнение (1) с заданной точностью ε. Например, неравенство $\rho(u, u^{(k)}) < \varepsilon$ заведомо будет выполняться, если

$$k > k(\varepsilon) = \frac{1}{\ln q} \ln \frac{\varepsilon(1-q)}{\rho(u', u^0)}. \tag{7}$$

Из неравенства (5′) следует, что первоначальная ошибка после k итераций уменьшится в $\frac{1}{\delta}$ раз, если $q^k \leqslant \delta$. Таким образом, для уменьшения первоначальной ошибки в $\frac{1}{\delta}$ раз при достаточно малом δ и $q \neq 0$ требуется провести по порядку

$$k_1 > k(\delta) = \frac{\ln(\delta^{-1})}{\ln(q^{-1})} = -\frac{\ln \delta}{v} \tag{8}$$

итераций.

Величину $v = \ln q^{-1} = -\ln q$ называют скоростью сходимости итерационного процесса. Она характеризует быстроту экспоненциального подавления начальной невязки. Последнее становится очевидным, если неравенство (5′) записать в виде

$$\rho(u^k, u) \leqslant e^{k \ln q} \rho(u^0, u) = e^{-kv} \rho(u^0, u). \tag{9}$$

Величина

$$\delta_k = \sup_{\rho(u, u^0) \neq 0} \frac{\rho(u, u^k)}{\rho(u, u^0)} \tag{10}$$

характеризует скорость убывания начальной ошибки после проведения k итераций.

О единственности решения операторного уравнения. В условиях принципа сжатых отображений решение уравнения (1) в $\mathfrak{M}$ единственно. Однако из единственности решения в $\mathfrak{M}$ не следует единственности решения вообще. Одно и то же уравнение можно рассматривать как уравнение с оператором, действующим в различных множествах $\mathfrak{M}$ разных пространств. Если каким-либо способом (например, при помощи теоремы принципа сжатых отображений) доказана единственность решения рассматриваемого уравнения на множестве $\mathfrak{M}$ некоторого пространства R, то отсюда не следует, что уравнение не имеет других решений в пространстве R и что оно не имеет решений, которые не принадлежат R.

2. Нестационарные итерационные процессы

Пусть итерационный процесс для решения уравнения (1) имеет вид

$$u^{k+1} = Q_k(u^k) \quad (k = 0, 1, 2, \ldots), \tag{11}$$

где каждому $k \geqslant 0$ соответствует оператор Q_k, удовлетворяющий условиям:

$$Q_k \mathfrak{M} \subset \mathfrak{M}, \ u = Q_k(u), \tag{12}$$

$$\rho(Q_k(u), Q_k(v)) \leqslant q_k \rho(u, v), \ q_k = e^{-\gamma_k}, \ \sum_{k=0}^{\infty} \gamma_k = \infty. \tag{13}$$

Тогда итерационный процесс (11) сходится к элементу u для любого начального приближения $u^0 \in \mathfrak{M}$.

Действительно,

$$\rho(u^k, u) = \rho(Q_{k-1}(u^{k-1}), Q_{k-1}(u)) \leqslant$$

$$\leqslant q_{k-1}\rho(u^{k-1}, u) \leqslant \cdots \leqslant \exp\left(-\sum_{i=0}^{k-1} \gamma_k\right) \rho(u^{(0)}, u) \underset{k\to\infty}{\to} 0.$$

Таким образом, нестационарный итерационный процесс (11) может сходиться даже в том случае, если имеется конечное число q_k больших единицы, но выполняются условия (13).

§ 2. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ЛИНЕЙНЫХ ОПЕРАТОРНЫХ УРАВНЕНИЙ

В основе последующих изложений лежат понятия и определения банаховых (**B**) и гильбертовых (**H**) пространств.

1. Операторные уравнения первого и второго рода

Операторным уравнением первого рода назовем уравнение вида

$$Au = f, \tag{1}$$

где $u \in \mathbf{E}$, $f \in \mathbf{F}$, $\mathbf{E}$, $\mathbf{F}$ — некоторые **B**-пространства, A — линейный оператор, действующий из $\mathbf{E}$ в $\mathbf{F}$. Если существует A^{-1}, то уравнение (1) имеет единственное решение для любого $f \in \mathbf{F}$.

Уравнение вида

$$u = Tu + f \tag{2}$$

назовем операторными уравнениями второго рода. Формально уравнение (2) можно свести к уравнению (1), если положить $A = I - T$. Возможны также следующие формальные преобразования уравнения первого рода в уравнения второго рода. Например, если:

а) добавить один и тот же элемент к обеим частям исходного (или частично преобразованного) уравнения (1),

б) применить к обеим частям уравнения (1) (или частично преобразованного уравнения) один и тот же оператор. Если этот оператор будет линейным ограниченным, то всякое решение исходного уравнения

будет решением нового уравнения. Вообще это преобразование может добавить лишние решения, так как оператор преобразования некоторые ненулевые элементы может перевести в нулевые.

в) ввести замену неизвестного u по формуле $u = P\omega$, где P — некоторый оператор, причем в этом случае необходимо, чтобы решение нового уравнения принадлежало области определения оператора P. Решения исходного уравнения, которые не представимы в виде $u = P\omega$, могут быть потеряны.

Существуют и другие подходы. Один из простейших видов преобразования уравнения (1) к уравнению (2) можно записать следующим образом:

$$Bu = Bu - H(Au - f). \tag{3}$$

Оказывается, что во многих известных эффективных итерационных методах при их построении использованы преобразования вида (3).

Возможность применения численного метода для решения операторного уравнения связана с устойчивостью задачи на (**E**, **F**).

Если A, $A^{-1} \in$ **B** (**E**), то для качественной характеристики связи между погрешностями правой части и решениями может быть использовано понятие обусловленности оператора A.

Числом обусловленности оператора A называют величину

$$\beta(A) = \|A\|\|A^{-1}\|. \tag{4}$$

Число обусловленности зависит от используемой нормы, но очевидно, $\beta(A) \geqslant 1$. Пусть

$$\varepsilon = u - u^0; \quad r = f - f^0 = f - Au^0, \tag{5}$$

где u^0 — приближенное решение уравнения (1). Тогда $A\varepsilon = r$ и

$$\|\varepsilon\| \leqslant \|A^{-1}\|\|r\|, \quad \|f^0\| \leqslant \|A\|\|u^0\|, \tag{6}$$

$$\|r\| \leqslant \|A\|\|\varepsilon\|, \quad \|u^0\| \leqslant \|A^{-1}\|\|f^0\|. \tag{7}$$

Если ввести в рассмотрение относительную погрешность $\|\varepsilon\|/\|u^0\|$, то для нее будут иметь место следующие неравенства:

$$\beta^{-1}(A)\frac{\|r\|}{\|f^0\|} \leqslant \frac{\|\varepsilon\|}{\|u^0\|} \leqslant \beta(A)\frac{\|r\|}{\|f^0\|}. \tag{8}$$

В самом деле, оценка сверху следует из неравенств (6), а оценка снизу — из неравенств (7). Неравенство (8) означает, что величина $\beta(A)$ ограничивает сверху, а $\beta^{-1}(A)$ — снизу отношение относительной неопределенности решения к относительной неопределенности правой части уравнения (1). Операторное уравнение с большими значениями мер обусловленности $\beta(A)$ принято называть плохо обусловленными, а с малыми — хорошо обусловленными. Следует подчеркнуть, что, например, при решении задач линейной алгебры значение $\beta(A)$ является гораздо более важным критерием трудности решения системы линейных алгебраических уравнений $Au = f$ по сравнению с теми трудностями, которые связаны с высоким порядком решаемой системы или малостью ее определителя.

Например, матрица

$$A = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 2 & 1 & \cdot & \cdot & \cdot & \cdot & 0 \\ & 2 & 1 & & & & \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & & & & & 2 & 1 \end{pmatrix} \tag{9}$$

имеет определитель, равный 1, однако эта матрица плохо обусловлена, так как у матрицы A^{-1} есть элемент 2^{n-1}.

Если операторное уравнение окажется плохо обусловленным, то необходимо использовать дополнительную информацию количественного или качественного характера о решении рассматриваемого операторного уравнения, чтобы свести решение задачи к решению операторного уравнения, имеющего устойчивые решения.

2. Итерационные методы решения линейных операторных уравнений первого рода

Итерационные процессы решения линейных операторных уравнений вида

$$Au = f$$

заключаются в построении последовательности $\{u^k\} \in \mathbf{E}$:

$$u^{k+1} = \Phi_k(A, f, u^k, u^{k-1}, \ldots, u^{k-m+1}) \quad (k = 0, 1, \ldots), \tag{10}$$

начиная с некоторого начального приближения $u^0 \in \mathbf{E}$, такой, что $u^k \to A^{-1}f$ при $k \to \infty$.

Итерационный процесс вида (10), когда функция u^{k+1} считается зависящей от A, f, u^k, u^{k-1}, ..., u^{k-m+1}, называется *m-шаговым итерационным процессом*. Если Φ_k не зависит от k, то итерационный процесс (10) называют *стационарным*. Если вид функции Φ_k по k меняется циклически с некоторым периодом N, то итерации называются *циклическими с периодом N*. Всякая циклическая итерация с периодом N эквивалентна некоторой стационарной итерации относительно приближенных значений u^0, u^N, u^{2N}, u^{3N}, ...

Наиболее полно изучены и чаще всего используются в практике вычислений одношаговые и двухшаговые итерационные процессы вида (10).

Простейшими среди итерационных процессов (10) являются одношаговые итерационные процессы

$$u^{k+1} = \Phi_k(A, f, u^k). \tag{11}$$

Если Φ_k — линейная функция от u^k, то итерационный процесс называют *линейным одношаговым итерационным процессом,*

$$B_k u^{k+1} = C_k u^k + \psi_k \quad (k = 0, 1, \ldots), \tag{12}$$

$u^0 \in \mathbf{E}$ — задано. Требуя, чтобы точное решение задачи (1) оставалось при этом неподвижной точкой, т. е. чтобы

$$B_k A^{-1} f = C_k A^{-1} f + \psi_k,$$

получим, что B_k, C_k, ψ_k должны быть связаны соотношением

$$\psi_k = H_k f, \tag{13}$$

где

$$H_k = (B_k - C_k) A^{-1}. \tag{14}$$

Таким образом, линейный одношаговый неявный итерационный процесс (12), для которого решение исходного уравнения (1) является неподвижной точкой, можно записать в виде

$$B_k (u^{k+1} - u^k) + H_k A u^k = H_k f \quad (k = 0, 1, \ldots). \tag{15}$$

Очевидно, если выбрать в итерационном процессе $B_k = I$ и соответственно $H_k = A^{-1}$, то уже за одну итерацию получим точное решение исходного уравнения. Однако такой выбор операторов B_k и H_k еще не говорит о том, что для получения решения уравнения (1) с точностью ε нам придется затратить меньшее число действий по сравнению с тем случаем, если вести счет по формуле (15) с $H_k \neq A^{-1}$ и $B_k \neq I$.

Таким образом, задача состоит в выборе итерационного метода, позволяющего найти решение уравнения (1) с заданной точностью ε за минимальное число действий. Эта задача обычно сводится к двум более простым задачам: о минимизации числа итераций и о минимизации числа действий для нахождения решения каждой итерации. Поскольку A и f фиксированы, то произвольно можно выбирать только операторы B_k и H_k.

При оценках качества итерационного метода величины, характеризующие число итераций и число действий для нахождения решения каждой итерации, должны быть взаимосвязаны.

Вид априорной информации о свойствах оператора A существенным образом влияет на использование и конструкцию того или другого итерационного метода. Особенно большое значение при построении итерационных методов имеет полная информация о спектре оператора A или хотя бы информация о некотором множестве, включающем в себя этот спектр, информация о возможности представления оператора A в виде $A = \sum_{\alpha=1}^{m} A_\alpha$, если известна некоторая априорная информация об операторах A_α $(\alpha = \overline{1, m})$ и т. д.

При конструировании итерационного метода операторы B_k выбираются из условия минимизации числа арифметических действий, затрачиваемых на одну итерацию, а H_k из условия построения быстросходящихся итерационных процессов. При этом обычно задается некоторый класс $\mathfrak{M}^0$ операторов H_k и на нем строятся эффективные итерационные методы. В качестве $\mathfrak{M}^0$ чаще всего выбирают $\mathfrak{M}^0$: $\{\tau_k I\}$, где τ_k — некоторые числовые параметры. Очевидно, по мере расширения $\mathfrak{M}^0$ представляется возможным получить более быстросходящиеся

итерационные процессы. Однако множества $\mathfrak{M}^0$, из которых выбирают операторы H_k, должны быть такими, чтобы ускорение сходимости итерационного процесса не очень усложняло счет по формуле (15) на каждом шаге. Выбор операторов B_k обычно производят таким образом, чтобы B_k^{-1} легко находились. Тогда неявная линейная одношаговая итерационная схема вообще может быть сведена к явной схеме

$$u^{k+1} = u^k - \tilde{H}_k(Au^k - f), \quad \tilde{H}_k = B_k^{-1}H_k \quad (k = 0, 1, 2, \ldots).$$

Наиболее полно исследованы итерационные схемы вида (15), в которых последовательность $\{u^k\}$ может быть найдена из явных рекуррентных соотношений вида

$$u^{k+1} = u^k - H_k(Au^k - f) \quad (k = 0, 1, 2, \ldots), \tag{16}$$

где $u^0 \in \mathbf{E}$ — задано, а H_k — некоторая последовательность операторов, характеризующая тип итерационного метода. Для того чтобы существовали операторы H_k, дающие сходящиеся итерационные методы (16) для решения уравнения (1), необходимо и достаточно, чтобы это уравнение имело единственное решение при любом $f \in \mathbf{E}$ (приложение теорема 16, § 1). На вопрос о том, из какого множества $\mathfrak{M}^0 \subseteq \mathbf{B}(\mathbf{E})$ можно выбрать операторы H_k в сходящихся итерационных методах, отвечает теорема о возмущениях (приложение, теорема 15, § 1), согласно которой достаточно, чтобы $H_k \in \mathfrak{M}^0 = \{H : \| H - A^{-1} \| \leqslant \leqslant \| A^{-1} \|^{-1}\}$.

Пусть $\varepsilon^k = u - u^k$ — погрешность k-й итерации. Тогда

$$\varepsilon^{k+1} = T_k\varepsilon^k, \tag{17}$$

где $T_k = I - H_kA$. Оператор T_k называется разрешающим оператором. Из (17) следует, что

$$\varepsilon^{k+1} = C_k\varepsilon^0; \quad C_k = T_kT_{k-1} \ldots T_0.$$

Следовательно, сходимость итераций для данной начальной ошибки ε^0 зависит от поведения $C_k\varepsilon^0$ при $k \to \infty$. Очевидно, итерационный процесс (16) будет сходиться для данной начальной ошибки ε^0 тогда и только тогда, когда в рассматриваемом пространстве $C_k\varepsilon^0 \to 0$ при $k \to \infty$. Значение ε^0, как правило, неизвестно, поэтому вообще легче исследовать поведение невязки

$$r^k = f - Au^k = A\varepsilon^k; \quad r^{k+1} = AT_kA^{-1}r^k, \quad r^{k+1} = AC_kA^{-1}r^0.$$

При $H_k \subseteq \mathfrak{M}^0 = \{\tau_{k+1}I\}$ итерационную схему (15) можно записать в следующей канонической форме:

$$B_k\frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k = f \quad (k = 0, 1, \ldots), \tag{18}$$

где $u^0 \in \mathbf{E}$ — произвольный элемент. Такую одношаговую итерационную схему называют двухслойной итерационной схемой, так как по форме она совпадает с двухслойной разностной схемой для нестационарных задач.

Для погрешности k-й итерации $\varepsilon^k = u - u^k$ имеем следующую задачу:

$$B_k \frac{\varepsilon^{k+1} - \varepsilon^k}{\tau_{k+1}} + A\varepsilon^k = 0 \quad (k = 0, 1, 2, \ldots), \ \varepsilon^0 = u - u^0. \tag{19}$$

Задача (19) — задача об устойчивости по начальным данным.

§ 3. МЕТОД ПРОСТЫХ ИТЕРАЦИЙ РЕШЕНИЯ ЛИНЕЙНЫХ УРАВНЕНИЙ

1. Метод простых итераций для линейного уравнения второго рода

Метод простых итераций (последовательных приближений) для решения линейного уравнения второго рода

$$u = Tu + f, \ u \in \mathbf{E} \tag{1}$$

имеет вид

$$u^{k+1} = Tu^k + f \quad (k = 0, 1, 2, \ldots), \tag{2}$$

где $u^0 \in \mathbf{E}$ — произвольный элемент.

Итерационный процесс вида (2) может быть применен непосредственно для решения уравнений второго рода, либо для решения уравнений первого рода после их преобразования к виду (1). Если последовательность $\{u^k\}$ сходится к некоторому элементу u, то $u = \lim\limits_{k\to\infty} u^k$ будет являться решением уравнения (2). Чтобы в этом убедиться, достаточно перейти к пределу при $k \to \infty$ в соотношении (2). Сходимость метода последовательных приближений для уравнения (2) связана с принципом сжатых отображений.

Если

$$\|T\| < 1, \tag{3}$$

то оператор уравнения (1) будет оператором сжатия и в соответствии с теоремой 1, § 1, его решение можно найти методом последовательных приближений (2), § 1.

С точки зрения линейных уравнений вида

$$\lambda u = Tu + f, \tag{4}$$

сходимость метода последовательных приближений

$$\lambda u^{k+1} = Tu^k + f \quad (k = 0, 1, \ldots) \tag{5}$$

связывается со сходимостью по норме ряда

$$\sum_{n=0}^{\infty} \lambda^{-(n+1)} T^n f, \tag{6}$$

сумма которого (в случае сходимости) есть $(\lambda I - T)^{-1} f$ (приложение, теорема 11, § 1).

Теорема 1. Если

$$|\lambda| > r(T) \tag{7}$$

($r(T)$ — спектральный радиус оператора T), то уравнение (4) *имеет единственное решение, которое может быть получено как предел последовательности* (5) *при любом* $u^0 \in \mathbf{E}$.

Доказательство. Пусть $|\lambda| > r$, тогда ряд

$$\sum_{n=0}^{\infty} \lambda^{-(n+1)} T^n f \tag{8}$$

(приложение, теорема 11, § 1) сходится и его сумма равна $(\lambda I - T)^{-1} f$. Применяя последовательно формулу (5), получим:

$$u^k = \lambda^{-1}(Tu^{k-1} + f) = \lambda^{-2}T^2u^{k-2} + \lambda^{-2}Tf + \lambda^{-1}f =$$

$$= \lambda^{-k}T^k u^{(0)} + \sum_{i=0}^{k-1} \lambda^{-(i+1)} T^i f. \tag{9}$$

Так как ряд (8) сходится по норме, то $\lim\limits_{k\to\infty} \|\lambda^{-(k+1)}T^k\| < 1$, т. е. $\lambda^{-k}T^k u^{(0)} \to 0$ при $k \to \infty$.

Тогда из (9) имеем:

$$\lim_{k\to\infty} u^k = \sum_{i=0}^{\infty} \lambda^{-(i+1)} T^i f = (\lambda I - T)^{-1} f = u,$$

т. е. u есть решение уравнения (4)

$$\lambda u = Tu + f.$$

Покажем единственность решения. Пусть кроме u существует v такое, что $u \neq v$ и $\lambda v = Tv + f$. Рассмотрим $z = u - v \neq 0$

$$z = \lambda^{-1}Tz = \lambda^{-2}T^2 z = \ldots = \lambda^{-k}T^k z = \ldots$$

Из существования предела

$$\sum_{n=0}^{k} \lambda^{-n} T^n z = (k+1)\, z$$

следует, что $z \equiv 0$, т. е. $u = v$.

Таким образом, из теоремы 1, § 1, и соотношения (3) следует, что метод простых итераций (5) сходится при любом начальном приближении $u^0 \in \mathbf{E}$ и при любой фиксированной $f \in \mathbf{E}$ к единственному решению уравнения (2), если выполняется одно из следующих условий:

а) $$r(T) = \lim_{n\to\infty} \sqrt[n]{\|T\|^n} < 1; \tag{10}$$

б) $$\|T\| < 1. \tag{11}$$

В случае сходимости последовательности $\{u^k\}$ к решению уравнения (1) имеют место следующие оценки для $\varepsilon^k = u - u^k$:

$$\|\varepsilon^k\| \leqslant \|T\| \|\varepsilon^{k-1}\|, \tag{12}$$

$$\|\varepsilon^k\| = \|T^k\varepsilon^0\| \leqslant \|T^k\|\|\varepsilon^0\|, \tag{13}$$

$$\|\varepsilon^k\| \leqslant \|T^k(I-T)^{-1}\|\|u^1-u^0\|. \tag{14}$$

Если выполнено условие (3), то имеют место следующие неравенства:

$$\|\varepsilon^k\| \leqslant \frac{\|T\|^k}{1-\|T\|}\|u^1-u^0\|, \tag{15}$$

$$\|\varepsilon^k\| \leqslant \|T\|^k\|u^0\| + \frac{\|T\|^k\|f\|}{1-\|T\|}. \tag{16}$$

Оценки вида (12), (13) следуют из неравенства

$$\varepsilon^k = T\varepsilon^{k-1} = T^k\varepsilon^0, \tag{17}$$

которое получается, если из уравнения (1) вычесть соотношение (2). Далее

$$u^1-u^0 = Tu^0+f-u^0 = Tu^0-Tu+u-u^0 = \varepsilon^0-T\varepsilon^0 = (I-T)\varepsilon^0$$

и $\varepsilon^0 = (I-T)^{-1}(u^1-u^0)$. Подставляя выражение для ε^0 в (17), получаем неравенство (14). Из (14), если $\|T\| \leqslant 1$, получаем неравенство (15) (приложение, теорема 14, § 1). Оценка (16) может быть получена следующим образом.

Применяя последовательно формулу (2), находим:

$$u^k = T^ku^0 + \sum_{i=0}^{k-1} T^if. \tag{18}$$

Так как $\|T\| < 1$, то $(I-T)^{-1}$ существует (приложение, теорема 13, § 1)

$$(I-T)^{-1} = \sum_{i=0}^{\infty} T^i,$$

$$u = (I-T)^{-1}f = \sum_{i=0}^{\infty} T^if. \tag{19}$$

Вычитая из (19) (18), получим:

$$\varepsilon^k = u-u^k = \sum_{i=k}^{\infty} T^if - T^ku^0, \tag{20}$$

откуда

$$\|\varepsilon^k\| \leqslant \sum_{i=k}^{\infty}\|T\|^i\|f\| + \|T\|^k\|u^0\| \leqslant \|T\|^k\|u^0\| + \frac{\|T\|^k\|f\|}{1-\|T\|}.$$

Заметим, что если положить $u^0 = f$, то из (20) получим:

$$\|\varepsilon^k\| \leqslant \frac{\|T^{k+1}\|\|f\|}{1-\|T\|}. \tag{21}$$

Выбирая в качестве u^0 значение f, итерационный процесс (2) можно осуществлять по формуле

$$u^k = f + Tf + T^2f + \ldots + T^kf \quad (k = 0, 1, 2, \ldots). \tag{22}$$

При такой организации итерационный процесс носит довольно единообразный характер и новое приближение получается из предыдущего

путем поправки к найденному приближению, но алгоритм не будет самоисправляющимся. Поэтому итерационный процесс лучше вести по формуле (2), где каждое новое приближение можно рассматривать как исходное и алгоритм носит самоисправляющийся характер.

2. Влияние ошибок исходных данных на сходимость метода простых итераций

Пусть оператор T и элемент f в сходящемся итерационном процессе (2) заданы приближенно с ошибками δ_1 и δ_2 соответственно, т. е.

$$\|T-\tilde{T}\| \leqslant \delta_1, \quad \|f-\tilde{f}\| \leqslant \delta_2. \tag{23}$$

Тогда итерационный процесс (2) будет осуществляться по формуле

$$\tilde{u}^{k+1} = \tilde{T}\tilde{u}^k + \tilde{f} \quad (k = 0, 1, \ldots), \quad u^0 \in \mathbf{E}. \tag{24}$$

Последовательные приближения (24) могут не сходиться к решению u уравнения (1), так как за счет приближенного задания исходных данных будет возникать неустранимая погрешность, и поэтому при помощи метода простой итерации можно получить приближенное решение $\tilde{u}$.

Для последовательных приближений (24) будет иметь место неравенство

$$\lim_{k\to\infty} \|\tilde{u}^k - u\| \leqslant \delta_3 = \frac{\delta_1 \|\tilde{f}\| (1-\|\tilde{T}\|)^{-1} + \delta_2}{1-\delta_1-\|\tilde{T}\|}. \tag{25}$$

Будем считать, что все последовательные приближения $\{\tilde{u}^k\}$ принадлежат множеству $\mathfrak{M}^0$, на котором T является оператором сжатия, $u^0 = \tilde{f}$. Из (2) и (24) имеем

$$\|u^k - \tilde{u}^k\| \leqslant \|Tu^{k-1} - \tilde{T}\tilde{u}^{k-1}\| + \|f-\tilde{f}\|. \tag{26}$$

Откуда, учитывая (23), получим:

$$\begin{aligned}
\|u^{k+1} - \tilde{u}^{k+1}\| &\leqslant \|T\| \|u^k - \tilde{u}^k\| + \delta_1 \|\tilde{u}^k\| + \delta_2 \leqslant \\
&\leqslant (\|\tilde{T}\| + \delta_1) \|u^k - \tilde{u}^k\| + \delta_1 (1-\|\tilde{T}\|)^{-1} \|\tilde{f}\| + \delta_2 = \\
&= \beta \|u^k - \tilde{u}^k\| + \omega = \beta^2 \|u^{k-1} - \tilde{u}^{k-1}\| + \beta\omega + \omega = \\
&= (\beta^k + \beta^{k-1} + \ldots + 1)\,\omega = \omega \frac{1-\beta^{k+1}}{1-\beta},
\end{aligned}$$

где $\beta = \|\tilde{T}\| + \delta_1$, $\omega = \delta_1 (1 - \|\tilde{T}\|^{-1} \|\tilde{f}\| + \delta_2$.

Итак,

$$\|u^{k+1} - \tilde{u}^{k+1}\| \leqslant \frac{\delta_1 \|\tilde{f}\| (1-\|\tilde{T}\|)^{-1} + \delta_2}{1-\delta_1-\|\tilde{T}\|}. \tag{27}$$

Таким образом, оценка полной ошибки $\| u - \tilde{u}^{k+1} \|$ в соответствии с (27) и (26) имеет вид

$$\| u - \tilde{u}^{k+1} \| \leqslant \| u - u^{k+1} \| + \| u^{k+1} - \tilde{u}^{k+1} \| \leqslant$$

$$\leqslant \frac{(\| \tilde{T} \| + \delta_1)^{k+1} (\| \tilde{f} \| + \delta_2)}{1 - \delta_1 - \| \tilde{T} \|} + \frac{\delta_1 \| \tilde{f} \| (1 - \| \tilde{T} \|)^{-1} + \delta_2}{1 - \delta_1 - \| \tilde{T} \|}. \tag{28}$$

Откуда при $k \to \infty$ следует (25).

При практической работе информация о порядке полной погрешности итерационного метода часто полезна для получения качественных выводов о том, с какой точностью разумно решать задачу. Из оценки (25) следует, что не имеет смысла стремление получить решение задачи (1) методом простой итерации с погрешностью ε, существенно меньшей δ_3.

§ 4. МЕТОДЫ УСКОРЕНИЯ СХОДИМОСТИ ПРОЦЕССОВ, ОСНОВАННЫЕ НА ИСПОЛЬЗОВАНИИ ЭНЕРГЕТИЧЕСКИ ЭКВИВАЛЕНТНЫХ ОПЕРАТОРОВ

Рассмотрим операторное уравнение

$$Au = f, \tag{1}$$

где A — линейный оператор, заданный в вещественном гильбертовом пространстве $\mathbf{H}$, $A : \mathbf{H} \to \mathbf{H}$, $u \in \mathbf{H}$, $f \in \mathbf{H}$.

Остановимся на некоторых методах ускорения сходимости двухслойных неявных итерационных процессов, которые существенно зависят от типа информации известной относительно оператора A.

Пусть A — самосопряженный положительно определенный оператор, энергетически эквивалентный оператору B с постоянными γ_1, γ_2, т. е. пусть выполняются соотношения

$$A = A^* > 0, \quad B = B^* > 0, \tag{2}$$

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_2 \geqslant \gamma_1 > 0, \tag{3}$$

где (3) означает, что

$$\gamma_1 (Bv, v) \leqslant (Av, v) \leqslant \gamma_2 (Bv, v) \text{ для всех } v \in \mathbf{H}.$$

Рассмотрим итерационный процесс вида

$$B \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k = f \quad (k = 0, 1, 2, \ldots) \tag{4}$$

с произвольным начальным вектором $u^0 \in \mathbf{H}$. Для исследования сходимости итерационных процессов вида (4) обычно неявную схему сводят к явной и используют результаты § 2. Неявная схема (4) при выполнении условий (2), (3) эквивалентна явной схеме

$$\omega^{k+1} = T_{k+1} \omega^k + \tau_{k+1} \psi, \quad T_{k+1} = I - \tau_{k+1} D, \quad k = 0, 1, 2, \ldots, \tag{5}$$

где $\omega^0 \in \mathbf{H}$ — произвольный элемент, D и ψ определяются соответственно равенствами:

$$D = D_1 = B^{-\frac{1}{2}} A B^{-\frac{1}{2}}, \quad \psi = \psi_1 = B^{-\frac{1}{2}} f \text{ при } \omega^k = B^{\frac{1}{2}} u^k, \tag{6}$$

или

$$D = D_2 = A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}, \quad \psi = \psi_2 = A^{\frac{1}{2}} B^{-1} f \text{ при } \omega^k = A^{\frac{1}{2}} u^k. \tag{7}$$

В самом деле, так как B — самосопряженный положительно определенный оператор, то существуют операторы $B^{\frac{1}{2}}$, $B^{-\frac{1}{2}}$, которые являются также самосопряженными положительно определенными. Уравнение (4) эквивалентно следующему:

$$B^{\frac{1}{2}} u^{k+1} = B^{\frac{1}{2}} u^k - \tau_{k+1} B^{-\frac{1}{2}} A B^{-\frac{1}{2}} B^{\frac{1}{2}} u^k + \tau_{k+1} B^{-\frac{1}{2}} f.$$

Обозначая $D_1 = B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$, $\omega^k = B^{\frac{1}{2}} u^k$, $\psi = B^{-\frac{1}{2}} f$, сведем схему (4) к явной итерационной схеме вида

$$\omega^{k+1} = \omega^k - \tau_{k+1} D_1 \omega^k + \tau_{k+1} \psi \quad (k = 0, 1, 2, \ldots). \tag{8}$$

Из (8) получим схему (5) при $D = D_1$, $\psi = \psi_1$.

Аналогично, записывая уравнение (4) в виде

$$A^{\frac{1}{2}} u^{k+1} = A^{\frac{1}{2}} u^k - \tau_{k+1} (A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}) A^{\frac{1}{2}} u^k + \tau_{k+1} A^{\frac{1}{2}} B^{-1} f,$$

от схемы (4) придем к явной схеме (5) при $D = D_2$, $\psi = \psi_2$. В силу условий (2), (3) оператор D будет самосопряженным и удовлетворять условию

$$\gamma_1 I \leqslant D \leqslant \gamma_2 I. \tag{9}$$

Таким образом, если оператор A удовлетворяет условиям (2), (3), то неявная схема (4) эквивалентна явной схеме (5), которая соответствует уравнению $D\omega = \psi$ при условии, что известны верхняя и нижняя границы спектра оператора D.

Изучение сходимости итерационного процесса, осуществляемого по схеме (4), может быть сведено к оценке (при $k \to \infty$) решения однородного уравнения

$$B \frac{\varepsilon^{k+1} - \varepsilon^k}{\tau_{k+1}} + A\varepsilon^k = 0 \quad (k = 0, 1, 2, \ldots) \tag{10}$$

для погрешности $\varepsilon^k = u - u^k$. При выполнении условий (2), (3) неявная схема (10) эквивалентна следующей явной схеме:

$$z^{k+1} = T_{k+1} z^k, \quad T_{k+1} = I - \tau_{k+1} D, \tag{11}$$

где $D = D_1$ при $z^k = B^{\frac{1}{2}} \varepsilon^k$, или $D = D_2$ при $z^k = A^{\frac{1}{2}} \varepsilon^k$.
Оценим $\|z^k\|$:

$$\|z^k\| \leqslant \|P_k\| \|z^0\|, \text{ где } P_k = \prod_{i=1}^{k} T_i. \tag{12}$$

Если $\|P_k\| \leqslant q_k < 1$, то для решения задачи (11) справедлива оценка

$$\|z^k\| \leqslant q_k \|z^0\| \tag{13}$$

или

$$(B^{\frac{1}{2}}\varepsilon^k, B^{\frac{1}{2}}\varepsilon^k)^{\frac{1}{2}} \leqslant q_k (B^{\frac{1}{2}}\varepsilon^0, B^{\frac{1}{2}}\varepsilon^0),$$

$$(B\varepsilon^k, \varepsilon^k)^{\frac{1}{2}} \leqslant q_k (B\varepsilon^0, \varepsilon^0)^{\frac{1}{2}} \quad \text{при } z^k = B^{\frac{1}{2}}\varepsilon^k$$

и

$$(A\varepsilon^k, \varepsilon^k)^{\frac{1}{2}} \leqslant q_k (A\varepsilon^0, \varepsilon^0) \quad \text{при } z^k = A^{\frac{1}{2}}\varepsilon^k.$$

Таким образом, из (13) получаем следующую априорную оценка для решения задачи (10):

$$\|\varepsilon^k\|_M \leqslant q_k \|\varepsilon^0\|_M, \text{ где } M = B \text{ или } M = A. \tag{14}$$

Отсюда видно, что через k итераций начальная погрешность ε^0 уменьшается в $\frac{1}{q_k}$ раза.

1. Стационарные итерационные процессы

Пусть оператор A уравнения $Au = f$ удовлетворяет условиям (2), (3). Для решения уравнения (1) построим итерационный процесс

$$B\frac{u^{k+1} - u^k}{\tau} + Au^k = f \quad (k = 0, 1, 2, \ldots), \tag{15}$$

где $u^0 \in \mathbf{H}$ — произвольный элемент, $\tau = \text{const}$. Для погрешности $\varepsilon^k = u - u^k$ k-й итерации получаем однородное уравнение (10) при $\tau_{k+1} = \tau = \text{const}$, которое в свою очередь эквивалентно явной схеме (11) с оператором перехода $T = I - \tau D$. Так как T — самосопряженный оператор в гильбертовом пространстве, то

$$\|T\| = r(T) = \sup_{\gamma_1 \leqslant \lambda \leqslant \gamma_2} |1 - \tau\lambda|. \tag{16}$$

Очевидно, при $0 < \tau < \frac{2}{\gamma_2}$ оператор перехода будет сжимающим оператором ($\|T\| < 1$). Параметр τ выберем так, чтобы оператор $T = I - \tau D$ был оператором сжатия с минимальной нормой, т. е. найдем такое τ_0, которое доставляет минимальное значение $\sup\limits_{\gamma_1 \leqslant \lambda \leqslant \gamma_2} |1 - \tau\lambda|$. Поскольку $\varphi(\tau, \lambda) = 1 - \tau\lambda$ — линейная функция от λ, то максимальное значение она достигает на одном из концов отрезка $[\gamma_1, \gamma_2]$. Функция $\varphi(\tau, \lambda) \geqslant 1$ при $\tau \leqslant 0$ на отрезке $[\gamma_1, \gamma_2]$ и монотонно убывает при $\tau > 0$. Далее $\varphi(\tau, \gamma_1) \geqslant 0$ при $0 < \tau \leqslant \gamma_2^{-1}$ и $\varphi(\tau, \gamma_2) < 0$ при $\tau > \gamma_2^{-1}$. Следовательно, при некотором τ_0 наступит момент, когда

$$1 - \tau_0\gamma_1 = -(1 - \tau_0\gamma_2). \tag{16'}$$

Решая уравнение (16′) относительно τ_0, получим:

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \tag{17}$$

В этом случае, как легко видеть,

$$\|T\| = \sup_{\gamma_1 \leqslant \lambda \leqslant \gamma_2} \|\varphi(\tau_0, \lambda)| = |1 - \tau_0\gamma_1| = |1 - \tau_0\gamma_2| = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = q < 1. \tag{18}$$

Так как $T = T^*$, то $\|P_k\| = \|T^k\| = \|T\|^k$, и для погрешности ε^k k-й итерации будет иметь место неравенство

$$\|\varepsilon^k\|_M \leqslant q^k \|\varepsilon^0\|_M, \text{ где } q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1}, \ M = B \text{ или } M = A. \tag{19}$$

В частности, если оператор A уравнения (1) удовлетворяет условиям

$$A = A^*, \ \gamma_1 I \leqslant A \leqslant \gamma_2 I, \quad \gamma_2 \geqslant \gamma_1 > 0, \tag{20}$$

то для решения уравнения $Au = f$ может быть применен сходящийся итерационный процесс

$$u^{k+1} = u^k - \frac{2}{\gamma_1 + \gamma_2}(Au^k - f) \quad (k = 0, 1, 2, \ldots), \tag{21}$$

который называют *методом смещений* или *оптимальным линейным итерационным процессом*. Скорость сходимости метода смещения будет оцениваться формулой

$$v = -\ln\frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = -\ln\left(1 - \frac{2}{1 + \frac{\gamma_2}{\gamma_1}}\right). \tag{22}$$

Если число обусловленности оператора A $\beta(A) = \frac{\gamma_2}{\gamma_1} \to \infty$, то асимптотическая скорость сходимости метода смещений будет оцениваться величиной

$$v_a \approx \frac{2}{\beta} = \frac{2\gamma_1}{\gamma_2}, \tag{23}$$

а число итераций k, при котором первоначальная ошибка уменьшится в $\frac{1}{\delta}$ раз,— величиной

$$k > k(\varepsilon) = O\left(\frac{\beta}{2}\ln\frac{1}{\delta}\right) = O\left(\frac{\gamma_2}{2\gamma_1}\ln\left(\frac{1}{\delta}\right)\right). \tag{24}$$

Если для самосопряженного положительно определенного оператора A известна только верхняя граница спектра $\gamma_2 = \sup_k |\lambda_k|$, то, выбирая $\tau_0 = \frac{1}{\gamma_2}$, можно построить сходящийся итерационный процесс вида

$$u^{k+1} = u^k - \frac{1}{\gamma_2}(Au^k + f) \ (k = 0, 1, 2, \ldots), \tag{25}$$

который называют *простейшим итерационным процессом*. Скорость сходимости последнего будет равна $\ln\left(1 - \frac{\gamma_1}{\gamma_2}\right)$ и при $\beta = \frac{\gamma_2}{\gamma_1} \to \infty$ v_a и k будут оцениваться соответствующими величинами

$$v_a \approx \frac{1}{\beta} = \frac{\gamma_1}{\gamma_2}, \ k > k(\varepsilon) = O\left(\beta \ln\frac{1}{\delta}\right), \tag{26}$$

из которых следует, что метод смещений (21) сходится приблизительно в два раза быстрее простейшего итерационного процесса (25).

Пусть оператор A уравнения (1) не является самосопряженным. Одношаговый стационарный неявный итерационный процесс вида (4) будет сходиться, если выполнены условия следующей теоремы.

Теорема 1. *Пусть A — линейный несамосопряженный оператор, энергетически эквивалентный оператору B, с оценками эквивалентности γ_1, γ_2, γ_3, т. е. для оператора A выполняются условия*

$$\gamma_1 B \leqslant A_c \leqslant \gamma_2 B \quad \left(A_c = \frac{1}{2}(A + A^*)\right),$$

$$|(B^{-1}A_k u, A_k u)| \leqslant \gamma_3 (Bu, u), \quad \left(A_k = \frac{1}{2}(A - A^*)\right).$$

Тогда стационарный итерационный процесс вида

$$B\frac{u^{k+1} - u^k}{\tau} + Au^k = f \quad (k = 0, 1, 2, \ldots)$$

при значениях параметра

$$\tau = \tau_0 \frac{1 - \gamma_3/(\gamma_1\gamma_2 + \gamma_3)}{1 + q\sqrt{\gamma_3/(\gamma_1\gamma_2 + \gamma_3)}}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} \tag{27}$$

будет сходиться и имеет место следующая оценка быстроты сходимости:

$$\|u - u^k\|_B \leqslant \tilde{q}^k \|\varepsilon^0\|_B, \tag{28}$$

где

$$\tilde{q} = \frac{q + \sqrt{\gamma_3/(\gamma_1\gamma_2 + \gamma_3)}}{1 + q\sqrt{\gamma_3/(\gamma_1\gamma_2 + \gamma_3)}} < 1. \tag{29}$$

Доказательство. Для погрешности $\varepsilon^k = u - u^k$ будет выполняться соотношение

$$\varepsilon^k = (I - \tau B^{-1}A)\,\varepsilon^{k-1} \quad (k = 0, 1, 2, \ldots). \tag{30}$$

Покажем, что оператор $T = I - \tau B^{-1}A$ является оператором сжатия в $\mathbf{H}_B$ при τ, определяемом формулой (27). Представим T в виде

$$T = \theta I + (1 - \theta)\, I - \tau B^{-1}(A_c + A_k),$$

где $0 < \theta < 1$ — произвольное число. Пусть x — произвольный элемент из $\mathbf{H}_B$, используя неравенство треугольников, получим:

$$\|(I - \tau B^{-1}A)\,x\|_B = \|(\theta I - \tau B^{-1}A_c)\,x\|_B + \|((1 - \theta)\, I - \tau B^{-1}A_k)\,x\|_B. \tag{31}$$

Для оценки первого слагаемого в правой части неравенства (31) имеем:

$$\|(\theta I - \tau B^{-1}A_c)\,x\|_B \leqslant \theta \left\| I - \frac{\tau}{\theta} B^{-1}A_c \right\|_B \|x\|_B. \tag{32}$$

Так как $T_1 = I - \frac{\tau}{\theta} B^{-1}A_c$ — самосопряженный оператор, то $\|T_1\| = \sup\limits_{\lambda(T_1) \in \mathrm{Sp}\, T_1} |\lambda(T_1)|$, и на основании (16), (17) при $\frac{\tau}{\theta} = \frac{2}{\gamma_1 + \gamma_2}$

норма T_1 принимает минимальное значение

$$\inf_{\tau,\theta} \| T_1 \| = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = q. \tag{33}$$

Для оценки второго слагаемого правой части неравенства (31) имеем:

$$(B((1-\theta)I - \tau B^{-1}A_k)x,\ ((1-\theta)I - \tau B^{-1}A_k)x) =$$
$$= (1-\theta)^2 (Bx, x) + \tau^2 (B^{-1}A_k x, A_k x),$$

так как $(A_k x, x) = 0$.

Поэтому

$$\| ((1-\theta)I - \tau B^{-1}A_k)x \|_B \leqslant ((1-\theta)^2 + \tau^2\gamma_3)^{1/2} \| x \|_B. \tag{34}$$

Итак, (31) — (34) приводят к неравенству

$$\| T \|_B = \| I - \tau B^{-1}A \|_B \leqslant \varphi(\theta), \text{ где } \varphi(\theta) = \theta q + ((1-\theta)^2 + \tau_0^2\gamma_3\theta^2)^{\frac{1}{2}}.$$

Найдем минимум функции $\varphi(\theta)$ при $\tau_0 = \dfrac{2}{\gamma_1 + \gamma_2}$

$$\varphi'(\theta) = q - \frac{1 - \theta - \tau_0^2\gamma_3\theta}{((1-\theta)^2 + \tau_0^2\gamma_3\theta^2)^{\frac{1}{2}}}.$$

Решая полученное квадратное уравнение относительно θ, найдем

$$\theta_0 = \frac{1 - \gamma_3/(\gamma_1\gamma_2 + \gamma_3)}{1 + q\sqrt{\gamma_3/(\gamma_1\gamma_2 + \gamma_3)}} \tag{35}$$

и при $\tau = \tau_0\theta_0$

$$\| T \|_B \leqslant \varphi(\theta_0) = \frac{q + \sqrt{\gamma_3/(\gamma_1\gamma_2 + \gamma_3)}}{1 + q\sqrt{\gamma_3/(\gamma_1\gamma_2 + \gamma_3)}} < 1. \tag{36}$$

2. Чебышевский циклический итерационный процесс

Пусть A удовлетворяет условиям (2), (3). Рассмотрим итерационный процесс (4), когда параметры τ_k изменяются на каждом шаге итерационного процесса.

Тогда в соответствии с (12)

$$\| z^k \| \leqslant \| P_k(D) \| \| z^0 \|,$$

где $P_k(D) = \prod_{i=1}^{k} (I - \tau_i D)$ — операторный полином степени k. Так как оператор D самосопряженный и удовлетворяет условиям (9), то

$$\| P_k \| \leqslant \max_{t \in [\gamma_1, \gamma_2]} | P_k(t) |, \quad P_k(t) = \prod_{i=1}^{k} (1 - \tau_i t). \tag{37}$$

Параметры τ_i $(i = 1, 2, \ldots)$ находим из условия минимума $\| P_k \|$. Таким образом, нужно найти полином $P_k(t)$ такой, чтобы

$$P_k(0) = 1, \tag{38}$$

и он наименее уклонялся от нуля при $t \in [\gamma_1, \gamma_2]$. Линейная замена

$$t = \frac{1}{2}[(\gamma_1 + \gamma_2) + (\gamma_2 - \gamma_1)x] \tag{39}$$

позволяет свести эту задачу к построению полинома $P_k(x)$, наименее уклоняющегося от нуля в промежутке $[-1, 1]$ и удовлетворяющего условию нормировки

$$P_k\left(\frac{\gamma_1 + \gamma_2}{\gamma_1 - \gamma_2}\right) = 1. \tag{40}$$

Эту задачу решает многочлен

$$R_k(x) = \frac{T_k(x)}{T_k\left(\frac{\gamma_1 + \gamma_2}{\gamma_1 - \gamma_2}\right)}, \quad |x| \leqslant 1, \tag{41}$$

где $T_k(x)$ — полином Чебышева первого рода. Корни многочлена $R_k(x)$ совпадают с корнями $T_k(x)$ и расположены в точках

$$x_i = \cos\frac{(2i-1)\pi}{2k} \quad (i = \overline{1, k}). \tag{42}$$

Корни многочлена $P_k(t)$ расположены в точках τ_i^{-1}

$$\tau_i = 2[\gamma_1 + \gamma_2 + (\gamma_2 - \gamma_1)x_i]^{-1}, \tag{43}$$

или

$$\tau_i = \frac{2}{(\gamma_1 + \gamma_2)\left[1 + \frac{\gamma_2 - \gamma_1}{\gamma_1 + \gamma_2}x_i\right]} = \frac{\tau_0}{(1 + qx_i)} \quad (i = \overline{1, k}), \tag{44}$$

где $\tau_0 = \frac{2}{\gamma_1 + \gamma_2}$, $q = \frac{\gamma_2 - \gamma_1}{\gamma_1 + \gamma_2}$, а x_i находится по формуле (42), причем $\tau_{i+k} = \tau_i$.

Максимум отклонения $R_k(x)$ при $|x| \leqslant 1$ будет равен:

$$\max_{-1 \leqslant x \leqslant 1} |R_k(x)| = \frac{1}{\left|T_k\left(\frac{\gamma_1 + \gamma_2}{\gamma_1 - \gamma_2}\right)\right|} =$$

$$= \frac{2}{\left(\frac{1}{q} + \sqrt{\frac{1}{q^2} - 1}\right)^k + \left(\frac{1}{q} - \sqrt{\frac{1}{q^2} - 1}\right)^k}.$$

Обозначая через $\beta = \frac{\gamma_2}{\gamma_1}$, преобразуем выражения

$$\frac{1}{q} + \sqrt{\frac{1}{q^2} - 1} = \frac{\beta + 1}{\beta - 1} + \frac{2\sqrt{\beta}}{\beta - 1} = \frac{\sqrt{\beta} + 1}{\sqrt{\beta} - 1} = \frac{1}{q_1} \tag{45}$$

и

$$\frac{1}{q} + \sqrt{\frac{1}{q^2} - 1} = \frac{\sqrt{\beta} - 1}{\sqrt{\beta} + 1} = q_1 < 1.$$

Следовательно,

$$\max_{-1 \leqslant x \leqslant 1} |R_k(x)| = \frac{2}{\frac{1}{q_1^k} + q_1^k} = \frac{2q_1^k}{1 + q_1^{2k}} < 1 \tag{46}$$

т. е. имеет место следующая теорема.

Теорема 2. Пусть оператор перехода T_i итерационной схемы (5) *на каждом шагу определяется из соотношения* $T_i = I - \tau_i D$ $(i = \overline{1, n})$ *и выполнены условия* $D = D^*$, $\gamma_1 I \leqslant D \leqslant \gamma_2 I$, $\gamma_2 \geqslant \gamma_1 > 0$. *Тогда при значениях* $\tau_i = \frac{\tau_0}{1 + q x_i}$, *где* $\tau_0 = \frac{2}{\gamma_1 + \gamma_2}$, $q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1}$, $x_i = \cos \frac{2i - 1}{2k} \pi$ $(i = \overline{1, n})$, $\min\limits_{\tau_i} \| T_k \| \leqslant q_k < 1$, *где*

$$q_k = \frac{2q_1^k}{1 + q_1^{2k}}, \quad q_1 = \frac{\sqrt{\gamma_2} - \sqrt{\gamma_1}}{\sqrt{\gamma_2} + \sqrt{\gamma_1}}.$$

Для величины $\varepsilon^k = u - u^k$ будет справедлива следующая оценка:

$$\| \varepsilon^k \|_M \leqslant q_k \| \varepsilon^0 \|_M, \text{ где } M = B \text{ или } M = A, \tag{47}$$

а q_k определяется по формулам (45), (46). Заметим, что при $k \to \infty$

$$q_k = \frac{2}{q_1^k + \frac{1}{q_1^k}} \leqslant 2q_1^k$$

и первоначальная ошибка уменьшится в $\frac{1}{\delta}$ раза, если число итераций

$$k > k(\delta) = \frac{\ln\left(\frac{2}{\delta}\right)}{\ln\left(\frac{1}{q_1}\right)}. \tag{48}$$

При $\beta = \frac{\gamma_2}{\gamma_1} \gg 1$
получим асимптотическое равенство

$$v_a \approx -\ln q_1 = \ln \frac{1}{q_1} = \ln\left(1 + \frac{2}{\sqrt{\beta - 1}}\right) \approx \frac{2}{\sqrt{\beta}}. \tag{49}$$

Поэтому при $\beta \to \infty$ получаем следующую оценку для числа итераций $k(\delta)$, при которой $q_k < \delta$:

$$k(\delta) = O\left(\frac{\sqrt{\beta} \ln \frac{2}{\delta}}{2}\right). \tag{50}$$

Схему (4) с последовательностью параметров τ_i $(= \overline{1, k})$, определяющихся по формуле (44), иногда называют схемой Ричардсона.

Отметим, что для плохо обусловленных операторных уравнений (при больших $\beta = \gamma_2/\gamma_1$) в рассматриваемом методе (4) с набором параметров τ_i, определяющихся по формуле (44), может происходить быстрая потеря значащих цифр в промежуточных и окончательных результатах, а также выход за разряды значений промежуточных итераций. Это можно проследить, если обратиться к неявной схеме (5) с набором параметров τ_i, определяющихся по формуле (44).

Оценим

$$\| T_i \| = \| I - \tau_i D \| = \sup_{\gamma_1 \leqslant \lambda \leqslant \gamma_2} | 1 - \tau_i \lambda |,$$

где $\tau_i = \tau_0/(1 + qx_i)$. Рассмотрим значение линейной функции $\varphi(\tau_i\lambda) = (1 - \tau_i\lambda)$ на интервале $[\gamma_1, \gamma_2]$

$$1 - \tau_i\gamma_1 = 1 - \frac{\tau_0\gamma_1}{1 + qx_i} = 1 - \frac{1 - q}{1 + qx_i} = \frac{q(x_i + 1)}{1 + qx_i}$$

$$1 - \tau_i\gamma_2 = 1 - \frac{\tau_0\gamma_2}{1 + qx_i} = 1 - \frac{1 + q}{1 + qx_i} = \frac{q(x_i - 1)}{1 + qx_i}.$$

Отсюда,

$$\text{при } x_i > 0, \ \|T_i\| = \frac{q(x_i + 1)}{1 + qx_i};$$

$$\text{при } x_i < 0, \ \|T_i\| = -|1 - \tau_i\gamma_2| = \frac{q(1 + |x_i|)}{1 - q|x_i|} = 1 + \frac{q(1 + 2|x_i|) - 1}{1 - q|x_i|}$$

Поэтому, если $x_i < 0$, $i \geqslant k_0$, где k_0 — наименьшее число, для которого $x_{k0} < 0$, и выполнено условие $q(1 + 2|x_{k0}|) > 1$, то $1 \leqslant \|T_{k0}\| < \|T_{k0+1}\|$. Значит, $\prod_{i=k_0}^{k} \|T_i\| > \|T_{k0}\|^{k-k_0} > 1$ и погрешность округления, появившаяся при определении u^{k_0}, будет увеличиваться с ростом i от k_0 до k. Для того чтобы сделать этот метод устойчивым по отношению к погрешностям округления, нужно упорядочить выбор параметров τ_i путем их перемешивания.

Один из способов упорядочения параметров τ_i в формуле (44) заключается в следующем [41].

Пусть при фиксированном $k < 2^r$ (r — наименьшее целое число и такое, что $k < 2^r$) число $i - 1$ ($1 \leqslant i \leqslant 2^r$) записано в двоичной системе, $i = \delta_1, \delta_2, \ldots, \delta_r$ (здесь δ_j равны либо нулю, либо единице). Образуем число $\sigma(i) = \delta_r\delta_{r-1} \ldots \delta_2\delta_1 + 1$ и считаем, что τ_l предшествует τ_m, если $\sigma(l) < \sigma(m)$. Это специальная перенумерация параметров как бы «равномерно» размещает их по величине, или операторы с нормой, большей единицы, «равномерно» размещаются среди операторов, уменьшающих норму ошибки.

Второй способ упорядочения параметров τ_i в (44) при $k = 2^r$ определяется по следующему рекуррентному способу. Пусть при $k/2$ установлен порядок номеров τ_i $(\tau_{\sigma_1}, \tau_{\sigma_2}, \ldots, \tau_{\sigma_{\frac{k}{2}}})$ в соответствии с предыдущей процедурой. Тогда для $k = 2^r$ порядок номеров τ_i определяется следующим образом:

$$(b_1, b_2, \ldots, b_k) = (\sigma_1, k + 1 - \sigma_1, \sigma_2, k + 1 - \sigma_2, \ldots, k + 1 - \sigma_{\frac{k}{2}}).$$

Обычно при $k \leqslant 10$ вычислительный процесс не требует специальной организации выборки параметров τ_i. Возможны и другие способы (см. [4], [5], [40], [42]) выбора параметров в чебышевских циклических методах.

Отметим, что при $k = 1$, $\tau = \frac{2}{\gamma_1 + \gamma_2}$ итерационный процесс метода смещений является частным случаем нестационарного циклического чебышевского итерационного процесса.

Из оценок асимптотической скорости сходимости рассмотренных в § 4 итерационных процессов видно, что оператор B следует выбирать так, чтобы отношение γ_2/γ_1 было по возможности меньшим.

§ 5. МЕТОДЫ РАСЩЕПЛЕНИЯ ОПЕРАТОРА

При построении линейных одношаговых итерационных процессов вида (15), § 2, операторы B_k выбираются из условия минимизации числа арифметических действий, затрачиваемых на одну итерацию, т. е. должны быть экономичными по числу арифметических действий (например, по порядку k относительно $\beta(A)$ при $\beta \to \infty$, или по числу арифметических действий на каждом шаге итерационного процесса) и такими, чтобы отношение чисел эквивалентности γ_2/γ_1 в соотношении (3), § 4, было по возможности минимальным.

В итерационных методах расщеплений сложился следующий конструктивный способ выбора B_k:

$$B_k = \prod_{\alpha=1}^{n} B_{k\alpha}, \tag{1}$$

где $B_{k\alpha}$ $(\alpha = \overline{1, n})$ обладают более простыми свойствами по сравнению с оператором A исходного уравнения $Au = f$ (операторы $B_{k\alpha}$, $\alpha = \overline{1, n}$, $k = 0, 1, 2, \ldots$ легко обратимы и экономичны).

Итерационные процессы методов расщеплений записываются в виде

$$\prod_{\alpha=1}^{n} B_{k\alpha} \frac{u^{k+1} - u^k}{\tau_{k+1}} = -Au^k + f, \quad k = 0, 1, 2, \ldots, \tag{2}$$

где τ_k — числовые параметры.

Схема реализации, соответствующая каждому шагу итерационного процесса (1), если предположить, что $B_{k\alpha}^{-1}$ $(\alpha = \overline{1, n}, k = 0, 1, 2, \ldots)$ достаточно просто находится, может быть представлена в виде последовательного решения уравнений с экономичными операторами $B_{k\alpha}$:

$$\begin{aligned}
&B_{k1}\xi^{k+\frac{1}{n}} = F_k, \quad F_k = \prod_{\alpha=1}^{n} B_{k\alpha}u^k - \tau_{k+1}(Au^k - f),\\
&B_{k2}\xi^{k+\frac{2}{n}} = \xi^{k+\frac{1}{n}},\\
&B_{k3}\xi^{k+\frac{3}{n}} = \xi^{k+\frac{2}{n}},\\
&\ldots\ldots\ldots\ldots\\
&B_{kn}\xi^{k+1} = \xi^{k+\frac{n-1}{n}}, \quad \xi^{k+1} \equiv u^{k+1}, \quad k = 0, 1, 2, \ldots
\end{aligned} \tag{3}$$

Отметим, что для одного и того же итерационного метода (2) можно построить целое семейство алгоритмов, реализующих этот метод, и тем самым осуществить выбор наиболее удачного алгоритма с точки зрения объема вычислений на каждом шаге, лучшей устойчивости к ошибкам округлений, удобства программирования и т. п.

Особо большое значение уделяется специальному выбору операторов B_k, когда оператор A уравнения $Au = f$ положительно определен и представим в виде

$$A = \sum_{\alpha=1}^{n} A_\alpha. \tag{4}$$

Тогда полагают, что

$$B_k = \prod_{\alpha=1}^{n} (I + \sigma_{k\alpha} A_\alpha) \quad (k = 0, 1, 2, \ldots),$$

где $\sigma_{k\alpha}$ $(\alpha = \overline{1, n})$ — некоторые числовые параметры, или

$$B_k = \prod_{\alpha=1}^{n} (P + \sigma_{k\alpha} A_\alpha) \quad (k = 0, 1, 2 \ldots),$$

где P — некоторые положительные операторы.

Итерационная схема метода расщеплении записывается в виде

$$\prod_{\alpha=1}^{n} (I + \sigma_{k\alpha} A_\alpha)(u^{k+1} - u^k) = -\tau_{k+1}(Au^k - f), \quad k = 0, 1, 2, \ldots \tag{5}$$

и содержит $(n + 1)$ произвольных параметров τ_k, $\sigma_{k\alpha}$ $(\alpha = \overline{1, n})$, которыми можно распорядиться для оптимизации вычислительного алгоритма на каждом шагу $(k = 0, 1, 2, \ldots)$.

Рассмотрим важный частный случай метода расщеплений, когда

и

$$A = A_1 + A_2 > 0 \tag{6}$$

$$\tau_{k+1} = \sigma_{k+1,1} + \sigma_{k+1,2}. \tag{7}$$

Итерационный процесс метода расщепления будет иметь вид

$$\prod_{\alpha=1}^{2} (I + \sigma_{k+1,\alpha} A_\alpha) \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k = f \quad (k = 0, 1, 2, \ldots). \tag{8}$$

Уравнение (8) эквивалентно уравнению

$$\prod_{\alpha=1}^{2} (I + \sigma_{k+1,\alpha} A_\alpha) u^{k+1} = (I - \sigma_{k+1,2} A_1)(I - \sigma_{k+1,1} A_2) u^k + \\ + (\sigma_{k+1,1} + \sigma_{k+1,2}) f \tag{9}$$

или, если операторы $(I + \sigma_{k+1,\alpha} A_\alpha)^{-1}$ $(\alpha = 1, 2)$, $(k = 0, 1 \ldots)$ существуют, то

$$u^{k+1} = (I + \sigma_{k+1,2} A_2)^{-1} (I + \sigma_{k+1,1} A_1)^{-1} [(I - \sigma_{k+1,2} A_1)(I - \sigma_{k+1,1} A_2) + \\ + (\sigma_{k+1,1} + \sigma_{k+1,2}) f] \quad (k = 0, 1, 2, \ldots). \tag{10}$$

Возможны различные схемы реализации итерационного процесса (9). Приведем некоторые из них при $\sigma_{k+1,1} = \sigma_{k+1,2} = \frac{\tau_{k+1}}{2} = \sigma_{k+1}$ (см. [4], [15], [27], [50], [84], [95], [96]).

а)

$$(I + \sigma_{k+1} A_1) u^{k+\frac{1}{2}} = (I - \sigma_{k+1} A_2) u^k + \sigma_{k+1} f,$$

$$(I+\sigma_{k+1}A_2)\,u^{k+1}=(I-\sigma_{k+1}A_1)\,u^{k+\frac{1}{2}}+\sigma_{k+1}f, \tag{11}$$
$$u^0\in\mathbf{H},\quad k=0,\,1,\,\ldots;$$

б) обобщение схемы (11) при

$$\sigma_{k+1,1}+\sigma_{k+1,2}=\tau_{k+1},$$
$$(I+\sigma_{k+1,1}A_1)\,u^{k+\frac{1}{2}}=(I-\sigma_{k+1,1}A_2)\,u^k+\sigma_{k+1,1}f,$$
$$(I+\sigma_{k+1,2}A_2)\,u^{k+1}=(I-\sigma_{k+1,2}A_1)\,u^{k+\frac{1}{2}}+\sigma_{k+1,2}f; \tag{12}$$

в)

$$(I+\sigma_{k+1}A_1)\,u^{k+\frac{1}{2}}=[I-\sigma_{k+1}(A_1+A_2)+\sigma_{k+1}^2A_1A_2]\,u^k+2\sigma_{k+1}f, \tag{13}$$
$$(I+\sigma_{k+1}A_2)\,u^{k+1}=u^{k+\frac{1}{2}};$$

г)

$$(I+\sigma_{k+1}A_1)\,\omega^k=Au^k-f,$$
$$(I+\sigma_{k+1}A_2)\,\omega^{k+\frac{1}{2}}=\omega^k, \tag{14}$$
$$u^{k+1}=u^k-2\sigma_{k+1}\omega^{k+\frac{1}{2}};$$

д)

$$(I+\sigma_{k+1}A_1)\,u^{k+\frac{1}{2}}=(I-\sigma_{k+1}A_1)\,u^k+2\sigma_{k+1}f_1,$$
$$(I+\sigma_{k+1}A_2)\,u^{k+1}=(I-\sigma_{k+1}A_2)\,u^{k+\frac{1}{2}}+2\sigma_{k+1}f_2, \tag{15}$$
$$f=f_1+f_2,\quad k=0,\,1,\,2,\,\ldots\ u^0\in\mathbf{H};$$

Схема вида (11) является неявной относительно вспомогательной неизвестной $u^{k+\frac{1}{2}}$ и искомой величины u^{k+1}, т. е. переход от k-й итерации к $(k+1)$-й осуществляется при помощи двух шагов. Исключая в (11) $u^{k+\frac{1}{2}}$ и учитывая перестановочность операторов $(I+\sigma_{k+1}A_1)^{-1}$ и $I-\sigma_{k+1}A_1$, получим (9).

Схемы (12) и (13) после подстановки промежуточных величин приводятся к виду (9), а схема (14) к схеме вида (8). Схема (15) эквивалентна схеме (11) относительно новой переменной, которая равна полусумме соседних приближений $\dfrac{u^{k+\frac{1}{2}}+u^{k+1}}{2}$.

Если обозначить $\dfrac{u^{k+1}+u^{k+\frac{1}{2}}}{2}$ через $\omega^{k+\frac{1}{2}}$, а $\dfrac{u^{k+1}+u^{k+\frac{3}{2}}}{2}$ через ω^{k+1}, то, сложив уравнения (15), а затем второе из уравнений (15) с аналогичным уравнением для $u^{k+\frac{3}{2}}$ получим схему вида (11) относительно ω^{k+1}.

1. Сходимость метода расщеплений для уравнений с самосопряженными и несамосопряженными операторами

а) *Анализ сходимости в коммутативном случае.*

Проведем исследования схемы расщеплений вида

$$\prod_{\alpha=1}^{n} (I + \sigma_{k+1,\alpha} A_\alpha) \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k = f \quad (k = 0, 1, 2, \ldots), \tag{5'}$$

когда самосопряженный оператор A уравнения $Au = f$ представим в виде

$$A = \sum_{\alpha=1}^{n} A_\alpha,$$

где A_α $(\alpha = \overline{1, n})$ — самосопряженные положительно определенные перестановочные операторы, определенные в гильбертовом пространстве **H** и отображающие элементы этого пространства в элементы этого же пространства. Таким образом для A_α $(\alpha = \overline{1, n})$ имеем:

$$A_\alpha = A_\alpha^*, \ m_\alpha I \leqslant A_\alpha \leqslant M_\alpha I, \ m_\alpha > 0 \quad (\alpha = \overline{1, n}), \tag{16}$$

$$A_s A_l = A_l A_s \quad (l, s = \overline{1, n}). \tag{17}$$

Из условий (16) и (17) следует, что

$$A_\alpha \varphi = \lambda^{(\alpha)} \varphi \ (\alpha = \overline{1, n}), \ A\varphi = \lambda\varphi, \ \sum_{\alpha=1}^{n} \lambda^{(\alpha)} = \lambda, \tag{18}$$

т. е. A_α $(\alpha = \overline{1, n})$ и A имеют общую систему собственных функций, образующих в **H** ортонормированный базис.

Функция погрешности $\varepsilon^k = u - u^k$ итерационного процесса (5) удовлетворяет уравнению

$$\prod_{\alpha=1}^{n} (I + \sigma_{k+1,\alpha} A_\alpha) (\varepsilon^{k+1} - \varepsilon^k) = -\tau_{k+1} A \varepsilon^k \quad (k = 0, 1, 2, \ldots)$$

или

$$\varepsilon^{k+1} = T_{k+1} \varepsilon^k = T_{k+1} T_k \varepsilon^{k-1} = \ldots = T_{k+1} T_k \ldots T_1 \varepsilon^0, \tag{19}$$

где

$$T_i = I - \tau_i \Big(\prod_{\alpha=1}^{n} (I + \sigma_{i\alpha} A_\alpha)\Big)^{-1} A \quad (i = 1, 2, \ldots). \tag{20}$$

Собственные значения оператора T_i $(i = 1, 2, \ldots)$ будут задаваться формулой

$$\lambda(T_i) = 1 - \frac{\tau_i \lambda}{\prod_{\alpha=1}^{n} (1 + \sigma_{i\alpha} \lambda^{(\alpha)})} \tag{21}$$

и так как T_i — самосопряженный оператор, то

$$\|T_i\| = \sup_{\lambda(T_i) \in \mathrm{Sp} T_i} |\lambda(T_i)| = \sup_{\lambda(T_i) \in \mathrm{Sp} T_i} \left| 1 - \frac{\tau_i \lambda}{\prod_{\alpha=1}^{n} (1 + \sigma_{i\alpha} \lambda^{\alpha})} \right|. \tag{22}$$

Очевидно, при $\sigma_{i\alpha} > 0$, $\tau_i > 0$ $(\alpha = \overline{1, n},\ i = 1, 2, \ldots)$ оператор T_i будет оператором сжатия, если

$$\prod_{\alpha=1}^{n} (1 + \sigma_{i\alpha}\lambda^{(\alpha)}) \geqslant \frac{\tau_i \lambda}{2} \quad (i = 1, 2, \ldots). \tag{23}$$

Достаточным условием выполнения неравенства (23) будет следующее:

$$\sigma_{i\alpha} \geqslant \frac{\tau_i}{2} > 0 \quad (\alpha = \overline{1, n},\ i = 1, 2, \ldots), \tag{24}$$

так как

$$\prod_{\alpha=1}^{n} \left(1 + \frac{\tau_i}{2} \lambda^{(\alpha)}\right) > \frac{\tau_i}{2} \sum_{\alpha=1}^{n} \lambda^{(\alpha)}.$$

Таким образом, имеет место следующая лемма.

Лемма 1. *Если коммутирующие операторы A_α $(\alpha = \overline{1, n})$ удовлетворяют условию $A_\alpha = A_\alpha^* > 0$, то все собственные числа операторов T_i вида*

$$T_i = I - \tau_i \left(\prod_{\alpha=1}^{n} (I + \sigma_{i\alpha} A_\alpha)\right)^{-1} A \tag{25}$$

при

$$\sigma_{i\alpha} \geqslant \frac{\tau_i}{2} > 0$$

по модулю будут меньше единицы.

Таким образом, при выполнении условий (16), (17), (24) итерационный процесс (5′) является сходящимся. Условия (24) дают широкие возможности для поисков оптимальных с точки зрения быстроты сходимости итерационного процесса (5) значений параметров $\sigma_{i\alpha}$, τ_i. Наиболее полно задача оптимального выбора параметров $\sigma_{i\alpha}$ и τ_i исследована для частного случая итерационного процесса (5), когда $n = 2$.

Пусть

$$A = A_1 + A_2,\ A_\alpha = A_\alpha^*,\ m_\alpha I \leqslant A_\alpha \leqslant M_\alpha I,\ m_\alpha > 0 \ (\alpha = 1, 2), \tag{26}$$
$$\sigma_{i1} + \sigma_{i2} = \tau_i > 0 \quad (i = 1, 2, 3, \ldots).$$

Для погрешности $\varepsilon^k = u - u^k$ итерационной схемы (5), исходя из (10), получим:

$$\varepsilon^k = T_k \varepsilon^{k-1} = T_k T_{k-1} \ldots T_1 \varepsilon^0, \tag{27}$$

где

$$T_i = (I + \sigma_{i2} A_2)^{-1} (I + \sigma_{i1} A_1)^{-1} (I - \sigma_{i2} A_1)(I - \sigma_{i1} A_2)$$
$$(i = \overline{1, k}). \tag{28}$$

Учитывая, что

$$\lambda(T_i) = \frac{(1 - \sigma_{i2}\lambda^{(1)})(1 - \sigma_{i1}\lambda^{(2)})}{(1 + \sigma_{i1}\lambda^{(1)})(1 + \sigma_{i2}\lambda^{(2)})} \quad (i = \overline{1, k}),$$

из (27) для ε^k получим оценку

$$\|\varepsilon^k\| \leqslant q_k \|\varepsilon^0\|,$$

где

$$q_k = \max_{\substack{\lambda^{(1)}\in[m_1M_1]\\ \lambda^{(2)}\in[m_2M_2]}} |\Phi_k(\lambda^{(1)}, \lambda^{(2)})|; \Phi_k(\lambda^{(1)}, \lambda^{(2)}) = \prod_{i=1}^{k} \frac{(1-\sigma_{i2}\lambda^{(1)})(1-\sigma_{i1}\lambda^{(2)})}{(1+\sigma_{i1}\lambda^{(1)})(1+\sigma_{i2}\lambda^{(2)})}. \tag{29}$$

Параметры σ_{i1}, σ_{i2} выбирают так, чтобы функция $\Phi_k(\lambda^{(1)}, \lambda^{(2)})$ была функцией, наименее уклоняющейся от нуля. В результате приходим к следующей задаче минимакса: найти такие положительные числа σ_{i1}, σ_{i2} $(i = \overline{1, k})$ и число $k = k(\varepsilon)$ $(0 < \varepsilon < 1$ — любое число), при которых достигается

$$\min_{\{\sigma_{i1}\},\{\sigma_{i2}\}} \max_{\substack{\lambda^{(1)}\in[m_1M_1]\\ \lambda^{(2)}\in[m_2M_2]}} \left| \prod_{i=1}^{k} \frac{(1-\sigma_{i2}\lambda^{(1)})(1-\sigma_{i1}\lambda^{(2)})}{(1+\sigma_{i1}\lambda^{(1)})(1+\sigma_{i2}\lambda^{(2)})} \right|. \tag{30}$$

Один из способов решения этой задачи заключается в следующем.

Заменой переменных

$$\lambda^{(1)} = \frac{1}{2}(M_1 - m_1)t + \frac{M_1+m_1}{2}, \quad \lambda^{(2)} = \frac{M_2-m_2}{2}z + \frac{M_2+m_2}{2}$$

изменение аргументов функции $\Phi_k(\lambda^{(1)}, \lambda^{(2)})$ приводится к отрезку $[-1, 1]$.

Тогда

$$\Phi_k(\lambda^{(1)}, \lambda^{(2)}) = \prod_{i=1}^{k} \frac{(r_i - t)(s_i - z)}{(\varkappa s_i + \gamma + \delta\varkappa + t)(r_i\varkappa^{-1} + \delta + \gamma\varkappa^{-1} + z)}, \tag{31}$$

где

$$\gamma = \frac{M_1+m_1}{M_1-m_1}, \quad \delta = \frac{M_2+m_2}{M_2-m_2}, \quad \varkappa = \frac{M_2-m_2}{M_1-m_1},$$
$$r_i = \frac{2}{(M_1-m_1)\sigma_{i2}} - \gamma, \quad s_i = \frac{2}{(M_2-m_2)\sigma_{i1}} - \delta, \quad -1 \leqslant t, \ z \leqslant 1. \tag{32}$$

Если принять r и s за независимые случайные величины, а r_i и соответственно s_i — за значения этих величин, получающихся в результате независимых испытаний, то задача (30) может быть сведена к отысканию дробно-линейной функции (31), у которой вероятность отклонения от нуля минимальна. Для вычисления оптимальных параметров σ_{i1}, σ_{i2} получены следующие формулы [20]:

$$\sigma_{i1} = \frac{2\{\varkappa + \gamma + \varkappa\delta - 1 - 2\varkappa \operatorname{sn}^2[\xi_i K(p')]\}}{(M_2-m_2)\{2(\gamma-1)\operatorname{sn}^2[\xi_i K(p')] - \varkappa - \gamma + 1 + \delta\gamma + \varkappa\delta^2 - \delta\}}, \tag{33}$$

$$\sigma_{i2} = \frac{2\{\gamma + \sigma\varkappa + 1 - \varkappa - 2\varkappa \operatorname{sn}^2[\xi_i K(p')]\}}{(M_1-m_1)\{2\varkappa(\delta-1)\operatorname{sn}^2[\xi_i K(p')] + \varkappa - 1 + \varkappa\gamma\delta - \varkappa\delta - \varkappa\gamma + \gamma^2\}}, \tag{34}$$

где $\varkappa$, γ, δ определяются по формулам (32), $\xi_i = \frac{2i+1}{2k}$ (i — целые числа, циклически меняющиеся от 1 до k), $p' = 2\sqrt{\varkappa}\,[(\gamma+\delta\varkappa)^2 -$

$-(1-\varkappa)^2]^{-\frac{1}{2}}$, $K(p')$ — полный эллиптический интеграл, $\operatorname{sn}(\varkappa)$ — эллиптический синус.

Попутно в [20] установлен следующий факт: если границы спектров операторов A_α ($\alpha = 1, 2$) совпадают ($M_1 = M_2$, $m_1 = m_2$), то $\sigma_{i1} = \sigma_{i2} = \sigma_i$,

$$\sigma_i = \frac{2}{M_1 - m_1} \frac{\gamma - \operatorname{sn}^2[\xi_i K(p')]}{\gamma^2 + (\gamma - 1)\operatorname{sn}^2[\xi_i K(p')] - \gamma} \quad \left(\gamma = \frac{M_1 + m_1}{M_1 - m_1}\right), \tag{35}$$

$$p' = \frac{M_1 - m_1}{M_1 + m_1}$$

и значения σ_i, определяющиеся по формуле (35), будут теми значениями, при которых наименее уклоняется от нуля функция

$$\prod_{i=1}^{k} \frac{1 - \sigma_i \lambda^{(1)}}{1 + \sigma_i \lambda^{(1)}} \tag{36}$$

при $m_1 \leqslant \lambda^{(1)} \leqslant M_1$.

Определение параметров $\sigma_{i\alpha}$ по формулам (33), (34) связано с довольно громоздкими вычислениями, а поэтому оправдывает себя, если приходится решать много задач одного и того же класса. Кроме того, строя другие более грубые способы выбора параметров $\sigma_{i\alpha}$ в итерационных процессах вида (5), можно провести сравнение с итерационным процессом при оптимальном наборе параметров (33), (34).

Остановимся на некоторых более грубых способах выбора параметров $\sigma_{i\alpha}$ ($i = 1, 2, \ldots$; $\alpha = 1, 2$).

Пусть границы спектров операторов A_1 и A_2 совпадают ($m_1 = m_2$, $M_1 = M_2$). Для выбора параметров σ_{i1}, σ_{i2} при условии, что $\sigma_{i1} + \sigma_{i2} = \tau_i > 0$, нужно решить задачу на минимакс для функции

$$\prod_{i=1}^{k} |\Phi_i(\lambda^{(1)})|, \tag{37}$$

где

$$\Phi_i(\lambda^{(1)}) = \frac{1 - \sigma_{i1}(\lambda^{(1)})}{1 + \sigma_{i1}\lambda^{(1)}} \frac{1 - \sigma_{i2}\lambda^{(1)}}{1 + \sigma_{i2}\lambda^{(1)}}. \tag{38}$$

Параметры σ_{i1}, σ_{i2} входят в (37) симметрично. Если положить $\sigma_{i1} = \sigma_{i2} = \sigma_i$, то задача (37) может быть заменена следующей задачей минимакса:

$$\min_{\{\sigma_i\}} \max_{\lambda^{(1)} \in [m_1 M_1]} \Phi_k(\lambda^{(1)}), \quad \Phi_k(\lambda^{(1)}) = \prod_{i=1}^{k} \left(\frac{1 - \sigma_i \lambda^{(1)}}{1 + \sigma_i \lambda^{(1)}}\right)^2, \tag{39}$$

где $0 < m_1 < M_1 < \infty$.

Частное решение задачи (39) содержится в работе [15], [95] и основано на задании некоторой величины q_k так, чтобы $\max \Phi_k(\lambda^{(1)}) \leqslant q_k$, а затем определяются возможные значения числа k и параметров $\sigma_1, \sigma_2, \ldots, \sigma_k$, соответствующие этому условию. Исследование

быстроты сходимости итерационного процесса (8) основано в этом случае на следующей лемме:

Лемма 2. *Пусть* $\varphi(\sigma, x) = \frac{1-\sigma x}{1+\sigma x}$ — *функция вещественных переменных* σ, x *с областью определения* $0 < a \leqslant x \leqslant b < \infty$, $\sigma > 0$ *и пусть*

$$\bar{q} = \min_{\sigma} \max_{x[a,b]} |\varphi(\sigma, x)|. \tag{40}$$

Тогда

$$\bar{q} = \varphi(\sigma_0, a) = |\varphi(\sigma_0, b)| = \frac{1-\sqrt{\frac{a}{b}}}{1+\sqrt{\frac{a}{b}}}, \tag{41}$$

где

$$\sigma_0 = \frac{1}{\sqrt{ab}}. \tag{42}$$

Доказательство. Для любого $\sigma > 0$ функция $\varphi(\sigma, x)$ строго монотонна при $x \in [a, b]$, причем $\varphi(\sigma, 0) = 1$, $\varphi(\sigma, +\infty) = -1$, следовательно, $\varphi(\sigma, x) < 1$. Если σ возрастает, то $\varphi(\sigma, x)$ убывает для всех $x > 0$; $\varphi(\sigma, x) = 0$, если $\sigma x = 1$. Отсюда ясно, что $\max\limits_{x \in [a,b]} |\varphi(\sigma, x)|$ будет минимизироваться для таких σ, при которых

$$\varphi(\sigma, a) = -\varphi(\sigma, b). \tag{43}$$

Основываясь на (43), получим уравнение для определения σ

$$\frac{1-\sigma a}{1+\sigma a} = -\frac{1-\sigma b}{1+\sigma b}, \tag{44}$$

т. е.

$$\sigma_0 = \frac{1}{\sqrt{ab}}. \tag{45}$$

Из (44) и (45) получаем (41).

Возвратимся к задаче (39) при $\lambda^{(1)} \in [m_1, M_1]$, $m_1 > 0$.

Разобьем интервал $[m_1, M_1]$ на N подынтервалов $[t_0, t_1]$, $[t_1, t_2]$, ..., $[t_{N-1}, t_N]$ ($t_0 = m_1$, $t_N = M_1$) таким образом, чтобы на каждом из подынтервалов при оптимальном значении σ_j величина $\max\limits_{t_{j-1} \leqslant x \leqslant t_j} \left|\frac{1-\sigma_j\lambda}{1+\sigma_j\lambda}\right|$ не превосходила заданной величины $q_k' < 1$. Для этого, как следует из леммы 2, должно выполняться соотношение

$$\sigma_j = \frac{1}{\sqrt{t_{j-1}t_j}}. \tag{46}$$

Тогда

$$q_k' = (1-\sqrt{t_{j-1}/t_j})/(1+\sqrt{t_{j-1}/t_j})$$

или $\frac{t_{j-1}}{t_j} = \left(\frac{1-q_k'}{1+q_k'}\right)^2$.

Откуда

$$t_j = \left(\frac{1+q_k'}{1-q_k'}\right)^2 t_{j-1} = \left(\frac{1+q_k'}{1-q_k'}\right)^4 t_{j-2} = \ldots = \left(\frac{1+q_k'}{1-q_k'}\right)^{2j} m_1. \tag{47}$$

Пусть k — минимальное число, при котором

$$t_k = \left(\frac{1+q_k'}{1-q_k'}\right)^{2k} m_1 \geqslant M_1.$$

Тогда

$$k \geqslant \frac{\ln \frac{M_1}{m_1}}{2 \ln \frac{1+q_k'}{1-q_k'}}. \tag{48}$$

Если ввести в рассмотрение среднюю скорость сходимости $v_{\text{ср}} = \frac{1}{k} \ln \frac{1}{q_k'}$, то

$$v_{\text{ср}} = -\frac{1}{k} \ln q_k' \geqslant \frac{2 \ln \frac{1-q_k'}{1+q_k'}}{\ln \frac{M_1}{m_1}} \ln q_k'. \tag{49}$$

Функция $\ln \frac{1-q_k'}{1+q_k'} \ln q_k'$ достигает максимума при $\frac{1-q_k'}{1+q_k'} = q_k'$. Отсюда $q_k' = \sqrt{2} - 1 \approx 0{,}414$. Значение $q_k' \approx 0{,}414$ является наилучшим для предложенного алгоритма, причем в этом случае

$$v_{\text{ср}} \geqslant \frac{2}{\ln \frac{M_1}{m_1}} \ln^2 (\sqrt{2} - 1) \tag{50}$$

и

$$\max_{\lambda^{(1)} \in [m_1 M_1]} \Phi_k(\lambda^{(1)}) \leqslant \sqrt{2} - 1.$$

Таким образом, использование данного алгоритма предполагает циклический с периодом $k \geqslant \frac{\ln m_1 - \ln M_1}{\ln (3 - 2\sqrt{2})}$ перебор параметров σ_j ($j = \overline{1, k}$) в итерационном процессе вида

$$(I + \sigma_{j+1} A_1)(I + \sigma_{j+1} A_2) u^{j+1} = (I - \sigma_{j+1} A_1)(I - \sigma_{j+1} A_2) u^j + 2\sigma_{j+1} f, \tag{51}$$
$$j = 0, 1, \ldots$$

Параметры σ_j определяются по формулам (46).

Для реализации итерационного процесса (51) можно использовать схемы методов расщеплений (например, одну из схем вида (11) — (14)).

Отметим, что если коммутирующие операторы A_1, A_2 имеют различные границы спектра, то путем дробно-линейного преобразования переменных $\lambda^{(1)}$ и $\lambda^{(2)}$ можно свести к случаю, когда спектры операторов A_1, A_2 принадлежат одному и тому же отрезку. Так, например, вводя новые переменные x и y, по формулам [4]:

$$\lambda^{(1)} = \frac{x - p}{q - rx}, \quad \lambda^{(2)} = \frac{y + p}{q + ry}, \tag{52}$$

где

$$p = \frac{\delta - \xi}{\delta + \xi}, \quad r = \frac{M_1 - M_2 + (M_1 + M_2)\,p}{2M_1M_2}, \quad q = r + \frac{1-p}{M_1},$$

$$\delta = \frac{M_2\,(M_1 - m_1)}{M_1\,(M_2 + m_1)}, \quad \xi = \sqrt{\frac{(M_1 - m_1)\,(M_2 - m_2)}{(M_1 + m_2)\,(M_2 + m_1)}},$$

получим, что $a \leqslant x,\ y \leqslant 1$ $\left(a = \frac{1-\xi}{1+\xi} > 0\right)$, и задача сводится к рассмотренному выше случаю (39).

Разные авторы [4], [71], [98] предлагали другие приемы частичного решения проблемы выбора оптимальных параметров $\sigma_{i\alpha}$, τ_i. Один из наиболее простых приемов заключается в следующем.

Считают параметры $\sigma_{i\alpha}$, τ_i постоянными ($\sigma_{i\alpha} = \sigma$, $\tau_i = \tau$) и полагают, что

$$\sigma = \frac{\tau}{2} = \frac{1}{\sqrt{Mm}}, \tag{53}$$

где

$$0 < m \leqslant \lambda(A) \leqslant M \ (M = M_1 + M_2,\ m = m_1 + m_2). \tag{54}$$

Тогда итерационный процесс вида

$$\prod_{\alpha=1}^{2} (I + \sigma A_\alpha)\,\frac{u^{k+1} - u^k}{2\sigma} + Au^k = f \quad (k = 0, 1, 2, \ldots) \tag{55}$$

на основании леммы 1 будет сходиться. Оценим скорость сходимости. Для этого рассмотрим функцию

$$\varphi(x, y) = \frac{1-x}{1+x}\,\frac{1-y}{1+y},$$

где $x = \sigma\lambda^{(1)}$, $y = \sigma\lambda^{(2)}$,

$$a_1 = \frac{m_1}{\sqrt{Mm}} \leqslant x \leqslant \frac{M_1}{\sqrt{Mm}} = b_1; \quad a_2 = \frac{m_2}{\sqrt{Mm}} \leqslant y \leqslant \frac{M_2}{\sqrt{Mm}} = b_2.$$

Тогда, исходя из (28),

$$\|T\| \leqslant \sup_{\substack{x \in [a_1 b_1] \\ y \in [a_2 b_2]}} |\varphi(x, y)|.$$

Функция $\varphi(x, y)$ монотонна при $x > 0$, $y > 0$ и может достигать своего экстремального значения на границе области определения x и y. Найдем $\varphi\left(\frac{m_1}{\sqrt{Mm}}, \frac{m_2}{\sqrt{Mm}}\right)$ и $\varphi\left(\frac{M_1}{\sqrt{Mm}}, \frac{M_2}{\sqrt{Mm}}\right)$:

$$\varphi\left(\frac{m_1}{\sqrt{Mm}}, \frac{m_2}{\sqrt{Mm}}\right) = \frac{1 - \sqrt{\frac{m}{M}} + \frac{m_1 m_2}{mM}}{1 + \sqrt{\frac{M}{m}} + \frac{m_1 m_2}{mM}},$$

$$\varphi\left(\frac{M_1}{\sqrt{Mm}}, \frac{M_2}{\sqrt{Mm}}\right) = \frac{1 - \sqrt{\frac{M}{m}} + \frac{M_1 M_2}{mM}}{1 + \sqrt{\frac{M}{m}} + \frac{M_1 M_2}{mM}}.$$

При $\beta_i = \frac{M_i}{m_i} \gg 1$

$$\varphi\left(\frac{m_1}{\sqrt{Mm}}, \frac{m_2}{\sqrt{Mm}}\right) \approx 1 - 2\sqrt{\frac{m}{M}},$$

$$\varphi\left(\frac{M_1}{\sqrt{Mm}}, \frac{M_2}{\sqrt{mM}}\right) \approx 1 - 2\sqrt{\frac{m}{M}}\,\frac{M^2}{M_1M_2},$$

так как $M = M_1 + M_2$, то $M^2 \geqslant 4M_1M_2$ и асимптотическая скорость сходимости итерационного процесса (55) при $\sigma = \frac{1}{\sqrt{Mm}}$ будет не меньше чем $\frac{2}{\sqrt{\beta}}$, $\beta = \frac{M}{m}$

$$v_a \approx \frac{2}{\sqrt{\beta}} = 2\sqrt{\frac{m}{M}}. \tag{56}$$

б) *Анализ сходимости метода расщепления в некоммутативном случае.*

Проведем исследование схемы метода расщеплений вида

$$\prod_{\alpha=1}^{2} (I + \sigma_\alpha A_\alpha)\frac{u^{k+1} - u^k}{\tau} + Au^k = f \quad (k = 0, 1, \ldots), \quad \sigma_1 + \sigma_2 = \tau > 0 \tag{57}$$

при условии, что операторы A_α не коммутируют, но удовлетворяют условиям (16), т. е. оператор A будет самосопряженным и положительно определенным.

При выполнении условий (16) уравнение (57) эквивалентно уравнению вида (9) и (10) с $\sigma_{k+1,\alpha} = \sigma_\alpha$ ($\alpha = 1, 2$), $\tau_{k+1} = \tau$. Для доказательства сходимости итерационного процесса (57), исходя из (10), запишем уравнение для $\varepsilon^k = u - u^k$

$$\varepsilon^{k+1} = (I + \sigma_2 A_2)^{-1}(I + \sigma_1 A_1)^{-1}(I - \sigma_2 A_1)(I - \sigma_1 A_2)\,\varepsilon^k \quad (k = 0, 1, 2, \ldots). \tag{58}$$

Пусть $\sigma_1 = \sigma_2 = \frac{\tau}{2}$.

Уравнение (58) запишется в виде

$$\varepsilon^{k+1} = T(\sigma)\,\varepsilon^k = T^{k+1}(\sigma)\,\varepsilon^0 \quad (k = 0, 1, 2, \ldots), \tag{59}$$

где $T(\sigma) = (I + \sigma A_2)^{-1}(I + \sigma A_1)^{-1}(I - \sigma A_1)(I - \sigma A_2)$.
Оператор $T(\sigma)$ подобен оператору

$$\tilde{T}(\sigma) = (I + \sigma A_2)\,T(\sigma)\,(I + \sigma A_2)^{-1} = T_1(\sigma)\,T_2(\sigma), \tag{60}$$

где $T_\alpha(\sigma) = (I + \sigma A_\alpha)^{-1}(I - \sigma A_\alpha)$ $(\alpha = 1, 2)$.

Для спектрального радиуса $r(T(\sigma))$ справедлива оценка

$$r(T(\sigma)) \leqslant \|\tilde{T}(\sigma)\| \leqslant \|T_1(\sigma)\|\,\|T_2(\sigma)\|. \tag{61}$$

Рассмотрим схему

$$z^{k+1} = T_\alpha z^k \quad (\alpha = 1, 2;\ k = 0, 1, 2, \ldots). \tag{62}$$

Запишем каноническую форму схемы (62)

$$B_\alpha \frac{z^{k+1}-z^k}{2\sigma} + A_\alpha z^k = 0 \quad (k = 0, 1, 2, \ldots), \tag{63}$$

где $B_\alpha = (I + \sigma A_\alpha)$ $(\alpha = 1, 2)$.

Представим B_α в виде $B_\alpha = (A_\alpha^{-1} + \sigma I) A_\alpha$ $(\alpha = 1, 2)$. Учитывая (16), получим:

$$\left(\frac{1}{m_\alpha} + \sigma\right) A_\alpha \leqslant B_\alpha \leqslant \left(\frac{1}{M_\alpha} + \sigma\right) A_\alpha$$

или

$$\gamma_{1\alpha} B_\alpha \leqslant A_\alpha \leqslant \gamma_{2\alpha} B_\alpha \quad (\alpha = 1, 2), \tag{64}$$

$$\gamma_{2\alpha} = \frac{m_\alpha}{1+\sigma m_\alpha}, \quad \gamma_{1\alpha} = \frac{M_\alpha}{1+\sigma M_\alpha}. \tag{65}$$

Из неравенства (64) следует (см. § 4, (15), (17)), что $\inf\limits_\sigma \| T_\alpha(\sigma) \|$ достигается при

$$2\sigma_{0\alpha} = \frac{2}{\gamma_{1\alpha} + \gamma_{2\alpha}} \quad (\alpha = 1, 2). \tag{66}$$

Откуда

$$\sigma_{0\alpha}\left(\frac{m_\alpha}{1+\sigma_{0\alpha} m_\alpha} + \frac{M_\alpha}{1+\sigma_{0\alpha} M_\alpha}\right) = 1,$$

$$\sigma_{0\alpha} = \frac{1}{\sqrt{m_\alpha M_\alpha}} \tag{67}$$

и

$$\inf_\sigma \| T_\alpha(\sigma) \| = \inf_\sigma \| (I + \sigma A_\alpha)^{-1} (I - \sigma A_\alpha) \| = \frac{1 - \sqrt{m_\alpha / M_\alpha}}{1 + \sqrt{m_\alpha / M_\alpha}}. \tag{68}$$

Следовательно, для итерационной схемы (57) при $\sigma_1 = \sigma_2 = \frac{\tau}{2} = \frac{1}{\sqrt{M_1 m_1}} = \frac{1}{\sqrt{m_2 M_2}}$ имеет место оценка

$$\| \varepsilon^k \| \leqslant q^k \| \varepsilon^0 \| \quad (k = 0, 1, 2, \ldots), \tag{69}$$

где

$$q = \frac{1 - \sqrt{m_1/M_1}}{1 + \sqrt{m_1/M_1}} \cdot \frac{1 - \sqrt{m_2/M_2}}{1 + \sqrt{m_2/M_2}} < 1.$$

Теорема 1. *Пусть выполнены условия* (16). *Для итерационной схемы* (57) *при* $\sigma_1 = \sigma_2 = \frac{\tau}{2} = \frac{1}{\sqrt{M_1 m_1}} = \frac{1}{\sqrt{M_2 m_2}}$ *имеет место оценка* (69).

В частности, если операторы A_α $(\alpha = 1, 2)$ удовлетворяют условиям (17) и $m_1 = m_2$, $M_1 = M_2$, то для погрешности ε^k итерационной схемы (57) при

$$\sigma_1 = \sigma_2 = \frac{\tau}{2} = \frac{1}{\sqrt{M_1 m_1}} \tag{70}$$

будет выполняться неравенство

$$\|\varepsilon^k\| \leqslant \tilde{q}^k \|\varepsilon^0\| \quad (k = 0, 1, 2, \ldots), \tag{71}$$

$$\tilde{q} = \left(\frac{1 - \sqrt{\frac{m_1}{M_1}}}{1 + \sqrt{\frac{m_1}{M_1}}} \right)^2. \tag{72}$$

Скорость сходимости схемы (57) будет оцениваться величиной

$$v = -\ln\left(\frac{1 - \sqrt{\frac{m_1}{M_1}}}{1 + \sqrt{\frac{m_1}{M_1}}} \right)^2 = -2\ln\left(1 - \frac{2\sqrt{\frac{m_1}{M_1}}}{1 + \sqrt{\frac{m_1}{M_1}}} \right) \tag{73}$$

и при $\beta_1 = \frac{M_1}{m_1} \gg 1$

$$v_a \approx \frac{4}{\sqrt{\beta_1}} = 4\sqrt{\frac{m_1}{M_1}}. \tag{74}$$

Для операторов A_α ($\alpha = 1, 2$) с сильно отличающимися границами спектров m_1, m_2, M_1, M_2 можно получить следующую оценку для величины q в неравенстве (69):

$$q \leqslant \frac{1 - \sqrt{\frac{m_{i0}}{M_{i0}}}}{1 + \sqrt{\frac{m_{i0}}{M_{i0}}}}. \tag{75}$$

В самом деле, при $\sigma_i = \frac{1}{\sqrt{m_i M_i}}$ $(i = 1, 2)$

$$\|T_i(\sigma_i)\| = \|(I + \sigma_i A_i)^{-1}(I - \sigma_i A_i)\| \leqslant q_i = \frac{1 - \sqrt{\frac{m_i}{M_i}}}{1 + \sqrt{\frac{m_i}{M_i}}}.$$

Пусть i_0 означает номер, для которого $q_{i0} = \min(q_1, q_2)$. Тогда $q_i \leqslant 1$ для всех i, σ, m_i, M_i при

$$\sigma = \frac{1}{\sqrt{m_{i0} M_{i0}}}. \tag{76}$$

Значит $\|T_1(\sigma)\| \|T_2(\sigma)\| \leqslant \frac{1 - \sqrt{\frac{m_{i0}}{M_{i0}}}}{1 + \sqrt{\frac{m_{i0}}{M_{i0}}}}$, т. е. для величины q справедлива оценка вида (75).

Отсюда следует, что метод будет сходиться, если даже один из операторов положительно полуопределен ($m_i = 0$), а второй — положительно определен. Асимптотическая скорость сходимости итерационного процесса (57) будет характеризоваться величиной

$$v_a \approx 2\sqrt{\frac{m_{i0}}{M_{i0}}}. \tag{77}$$

в) *Анализ сходимости метода расщеплений для уравнений с несамосопряженным оператором.*

Пусть в уравнении $Au = f$ оператор A удовлетворяет условиям $A = A_1 + A_2$, где

$$A_\alpha \geqslant m_\alpha I, \ m_\alpha > 0 \quad (\alpha = 1, 2), \tag{78}$$

$$\| A_\alpha u \|^2 \leqslant M_\alpha (A_\alpha u, u), \ M_\alpha > 0 \quad (\alpha = 1, 2).$$

Исходя из (58) для погрешности $\varepsilon^k = u - u^k$ итерационного процесса (57), если предположить существование операторов $(I + \sigma A_1)^{-1}$, $(I + \sigma A_2)^{-1}$, получим уравнение (59).

Вводя в рассмотрение

$$z^k = (I + \sigma A_2) \varepsilon^k, \tag{79}$$

из (59) имеем:

$$z^k = T_1 T_2 z^{k-1}, \tag{80}$$

где $T_\alpha = (I + \sigma A_\alpha)^{-1} (I - \sigma A_\alpha) \quad (\alpha = 1, 2)$.

Оценим $\| T_\alpha \| (\alpha = 1, 2)$ при $\sigma > 0$. Для этого рассмотрим

$$\| (I + \sigma A_\alpha) u \|^2 = \| u \|^2 + 2\sigma (A_\alpha u, u) + \sigma^2 \| A_\alpha u \|^2 \leqslant$$

$$\leqslant (m_\alpha^{-1} + 2\sigma + \sigma^2 M_\alpha) (A_\alpha u, u), \tag{81}$$

$$\| (I - \sigma A_\alpha) u \|^2 = \| u \|^2 - 2\sigma (A_\alpha u, u) + \sigma^2 \| A_\alpha u \|^2. \tag{82}$$

Вычитая (82) из (81), получим:

$$4\sigma (A_\alpha u, u) = \| (I + \sigma A_\alpha) u \|^2 - \| (I - \sigma A_\alpha) u \|^2$$

или

$$\| (I - \sigma A_\alpha) u \|^2 = \| (I + \sigma A_\alpha) u \|^2 - 4\sigma (A_\alpha u, u) \leqslant$$

$$\leqslant \left(1 - \frac{4\sigma}{(m_\alpha^{-1} + 2\sigma + M_\alpha \sigma^2)}\right) \| (I + \sigma A_\alpha) u \|^2.$$

Если обозначить $(1 + \sigma A_\alpha) u = v$, тогда

$$(I - \sigma A_\alpha) u = (I - \sigma A_\alpha) (I + \sigma A_\alpha)^{-1} v,$$

т. е.

$$\| T_\alpha v \|^2 \leqslant \left(1 - \frac{4\sigma}{m_\alpha^{-1} + 2\sigma + M_\alpha \sigma^2}\right) \| v \|^2$$

или

$$\| T_\alpha \|^2 \leqslant 1 - \frac{4\sigma}{m_\alpha^{-1} + 2\sigma + M_\alpha \sigma^2} = f_\alpha (\sigma) \quad (\alpha = 1, 2)$$

и

$$\| z^k \| \leqslant \sqrt{f_1 (\sigma) f_2 (\sigma)} \| z^{k-1} \|.$$

Учитывая, что $f_\alpha (\sigma)$ достигает max при $\sigma = \sigma_{0\alpha} = \frac{1}{\sqrt{m_\alpha M_\alpha}}$ и $f_\alpha (\sigma_{0\alpha}) = \frac{1 - \sqrt{\xi_\alpha}}{1 + \sqrt{\xi_\alpha}}$, $\xi_\alpha = \frac{m_\alpha}{M_\alpha}$, для оценки $\| z^k \|$ имеем:

$$\| z^k \| \leqslant \left(\frac{1 - \sqrt{\xi_1}}{1 + \sqrt{\xi_1}} \cdot \frac{1 - \sqrt{\xi_2}}{1 + \sqrt{\xi_2}} \right)^{\frac{1}{2}} \| z^{k-1} \|. \tag{83}$$

При $m_1 = m_2$, $M_1 = M_2$ и $\sigma_0 = \frac{1}{\sqrt{m_1 M_1}}$ асимптотическая скорость сходимости будет:

$$v_a \approx 2\sqrt{\frac{m_1}{M_1}}, \tag{84}$$

т. е. в два раза медленнее случая самосопряженных операторов A_α ($\alpha = 1, 2$) (см. 74)).

2. Принцип регуляризации в построении итерационных методов расщеплений

При построении итерационных методов расщеплений оператор B можно выбрать из некоторого допустимого семейства операторов, энергетически эквивалентных оператору A, но так чтобы при этом выполнялись принципы, положенные в основу выбора операторов B в методах расщеплений: экономичность и максимум отношения оценок эквивалентности операторов A и B, т. е. чтобы величины $\frac{\gamma_1}{\gamma_2}$ достигала по возможности максимального значения.

Пусть оператору $A = A^* > 0$ поставлен в соответствие некоторый энергетически эквивалентный оператор R, самосопряженный и положительно определенный

$$\hat{m}R \leqslant A \leqslant \hat{M}R, \ \hat{M} \geqslant \hat{m} > 0, \ R = R^* > c_0 I, \ c_0 > 0. \tag{85}$$

Тогда оператор R можно представить в виде

$$R = R_1 + R_2, \tag{86}$$

где R_1 и R_2 неперестановочны, но сопряжены друг с другом ($R_1 = R_2^*$), т. е.

$$(R_1 u, u) = (R_2 u, u) = \frac{1}{2}(Ru, u). \tag{87}$$

(Например, если R — симметричная матрица, то это соответствует представлению симметричной матрицы в виде суммы верхней и нижней треугольных матриц, а поэтому неявная схема вида (89) получила название попеременно-треугольного метода [4], [70]).

Если R_1 и R_2 экономичные операторы, то полагают, что

$$B = (I + \sigma R_1)(I + \sigma R_2) \tag{88}$$

и схема метода расщеплений будет иметь вид

$$(I + \sigma R_1)(I + \sigma R_2)\frac{u^{k+1} - u^k}{\tau} + Au^k = f. \tag{89}$$

Предположим, что удалось указать такую постоянную c_1, что

$$\| R_2 u \|^2 \leqslant \frac{c_1}{4}(Ru, u) \tag{90}$$

для всех $u \in \mathbf{H}$.

Для оценки быстроты сходимости итерационной схемы вида (89) следует определить постоянные эквивалентности операторов A и B.

Для этого сначала определим постоянные эквивалентности операторов R и B.

Оператор $B = (I + \sigma R_1)(I + \sigma R_2)$ можно представить в виде

$$B = (I - \sigma R_1)(I - \sigma R_2) + 2\sigma (R_1 + R_2).$$

Так как

$$(I - \sigma R_1)(I - \sigma R_2) = (I - \sigma R_1)(I - \sigma R_1)^* \geqslant 0,$$

то

$$B \geqslant 2\sigma R. \tag{91}$$

С другой стороны,

$$B = I + \sigma R + \sigma^2 R_1 R_2,$$

причем

$$(R_1 R_2 u, u) = (R_2 u, R_2 u) \leqslant \frac{c_1}{4}(Ru, u),$$

Следовательно,

$$(Bu, u) \leqslant \left(\frac{1}{c_0} + \sigma + \sigma^2 \frac{c_1}{4}\right)(Ru, u). \tag{92}$$

Из неравенств (91), (92) находим, что

$$\left(\frac{1}{c_0} + \sigma + \frac{\sigma^2 c_1}{4}\right)^{-1} B \leqslant R \leqslant \frac{0{,}5}{\sigma} B.$$

Параметр σ определяется из условия максимума отношения оценок эквивалентности операторов A и B, т. е. из условия минимума функции

$$\beta(\sigma) = \frac{\hat{M}\left(1 + c_0\sigma + \frac{\sigma^2 c_0 c_1}{4}\right)}{2\hat{m}c_0\sigma}. \tag{93}$$

Так как

$$\beta'(\sigma) = \frac{\hat{M}}{8\hat{m}c_0} \cdot \frac{c_0 c_1 \sigma^2 - 4}{\sigma^2},$$

то при $\sigma_0 = \frac{2}{\sqrt{c_0 c_1}}$, $\beta'(\sigma_0) = 0$, $\beta''(\sigma_0) \geqslant 0$.

Таким образом,

$$\gamma_1(\sigma_0) B \leqslant A \leqslant \gamma_2(\sigma_0) B, \tag{94}$$

где

$$\gamma_1(\sigma) = \frac{\hat{m}c_0}{2(1 + \sqrt{\eta})}, \quad \gamma_2(\sigma) = \frac{\hat{M}c_0}{4\sqrt{\eta}}, \quad \eta = \frac{c_0}{c_1}. \tag{95}$$

В частности, если рассмотреть итерационный процесс (89) при

$$\sigma = \frac{2}{\sqrt{c_0 c_1}}, \quad \tau = \tau_0 = \frac{2}{\gamma_1(\sigma_0) + \gamma_2(\sigma_0)}, \tag{96}$$

то быстрота его сходимости будет оцениваться при помощи следующего неравенства:

$$\|\varepsilon^k\|_M = \|u - u^k\|_M \leqslant q^k \|\varepsilon^0\|_M, \quad M = A \quad \text{или} \quad M = B, \tag{97}$$

$$q = \frac{\gamma_2(\sigma_0) - \gamma_1(\sigma_0)}{\gamma_2(\sigma_0) + \gamma_1(\sigma_0)} = \frac{(\hat{M} - 2\hat{m})\sqrt{\eta} + M}{(\hat{M} + 2\hat{m})\sqrt{\eta} + M} < 1. \tag{98}$$

Имеет место следующая теорема.

Теорема 2. *Если выполнены условия* (85), (86), (90), *то итерационная схема* (89) *при* $\sigma = \frac{2}{\sqrt{c_0 c_1}}$, $\tau = \frac{8\sqrt{\eta}/(1+\sqrt{\eta})}{c_0[(2\hat{m}+\hat{M})\sqrt{\eta}+\hat{M}]}$, $\eta = \frac{c_0}{c_1}$, $(\hat{M}-2\hat{m})\sqrt{\eta}+\hat{M}>0$ *сходится в* $\mathbf{H}_A$ *и* $\mathbf{H}_B$ *так, что имеет место оценка* (97), (98).

Переход от u^k к u^{k+1} для схемы (89) с факторизованным оператором можно осуществить с помощью одного из алгоритмов вида (11) — (14). Различные модификации метода регуляризации содержатся в работах [4], [70] и др.

§ 6. ОДНОШАГОВЫЕ ИТЕРАЦИОННЫЕ МЕТОДЫ, ОСНОВАННЫЕ НА ИСПОЛЬЗОВАНИИ КВАДРАТИЧНОГО ФУНКЦИОНАЛА

В вариационных методах задача отыскания решения уравнения

$$Au = f \tag{1}$$

сводится к задаче отыскания минимума некоторого функционала $\Phi_M(v)$, который строится на основе использования квадратичного функционала

$$\Omega_M(v) = (M(u-v),\ u-v) \tag{2}$$

с самосопряженным положительно определенным оператором M и отличается от него лишь неизвестным постоянным слагаемым.

Заметим, что

$$\Omega_M(v) = (M\varepsilon,\ \varepsilon), \tag{3}$$

где $\varepsilon = u - v$, u — решение уравнения (1), v — произвольный элемент из $\mathbf{H}$, достигает минимума, равного нулю при $v = u$. Квадратичный функционал (3) называют функционалом ошибок и его можно рассматривать как квадрат унитарной нормы для элементов вида $u - v$.

Различные итерационные методы отличаются выбором оператора M, способом минимизации $\Omega_M(v)$ на каждом шаге итерационного процесса.

Пусть $v = \omega + \tau t$, где t, $\omega \in \mathbf{H}_M$, τ — вещественное число. Рассмотрим производную от функционала $\Omega_M(v)$ по направлению t в точке ω, т. е.

$$\frac{\partial\Omega(v)}{\partial t} = \frac{1}{\|t\|_M}\lim_{\tau\to 0}\frac{\Omega(\omega+\tau t)-\Omega(\omega)}{\tau} = \frac{1}{\|t\|_M}\left.\frac{\partial\Omega(v)}{\partial\tau}\right|_{\tau=0}. \tag{4}$$

Пусть производная существует и не равна нулю в точке ω^0 для каждого направления t и имеется направление t^0, по которому эта производная наименьшая. Это направление назовем направлением антиградиента функционала в точке ω^0. Вычислим значение градиента от $\Omega_M(\omega+\tau t)$ в пространствах $\mathbf{H}_M$ и $\mathbf{H}$ (при $M = I$). Для этого запишем $\Omega_M(v) = \Omega_M(\omega+\tau t)$ в виде

$$\Omega_M(\omega+\tau t) = (M(u-\omega-\tau t),\ u-\omega-\tau t) = \tau^2[t]_M^2 - 2\tau[t,\varepsilon]_M + [\varepsilon]_M^2, \tag{5}$$

где

$$\varepsilon = u - \omega, \quad [t, \varepsilon]_M = (Mt, \varepsilon). \tag{6}$$

Тогда

$$\frac{\partial \Omega_M (\omega + \tau t)}{\partial t} = \frac{1}{\|t\|_M} \frac{\partial \Omega_M (\omega + \tau t)}{\partial \tau}\bigg|_{\tau=0} = -\frac{2[t, \varepsilon]_M}{\|t\|_M}. \tag{7}$$

Таким образом, направление антиградиента задается элементом

$$t^0 = u - \omega, \tag{8}$$

т. е. в пространстве $\mathbf{H}_M$ антиградиент в любой точке ω^0 направлен на решение u. В пространстве $\mathbf{H}$:

$$\frac{\partial \Omega_M (\omega + \tau t)}{\partial t} = \frac{1}{\|t\|} \frac{\partial \Omega_M (\omega + \tau t)}{\partial \tau}\bigg|_{\tau=0} = -2[t, \varepsilon]_M (t, t)^{\frac{1}{2}} =$$

$$= 2(t, M(u - \omega))(t, t)^{\frac{1}{2}} \tag{9}$$

или

$$t^0 = M(u - \omega). \tag{10}$$

Значение u неизвестно, а значит, неизвестно и значение t^0. Однако элемент t^0 можно выразить через невязку $r = f - A\omega$

$$t^0 = M(A^{-1}f - \omega) = MA^{-1}(f - A\omega) = MA^{-1}r. \tag{11}$$

Поэтому, если положить $M = CA$, где оператор C выбран таким образом, чтобы оператор M был самосопряженным и положительно определенным, то

$$t^0 = Cr \tag{12}$$

и при $\omega = u^0$ величина $t^0 = Cr^0$ является известной.

Найдем

$$\Omega_M (\omega + \tau t^0) = \Omega_M (v) - 2\tau (t^0, t^0) + \tau^2 (Mt^0, t^0) \tag{13}$$

и вычислим то значение τ_0, при котором $\Omega_M (\omega + \tau t^0)$ достигает максимума по направлению t^0, т. е.

$$\tau_0 = \frac{(t^0, t^0)}{(Mt^0, t^0)}, \tag{14}$$

и элемент $\omega^1 = \omega^0 + \tau_0 t^0$ минимизирует функционал Ω_M на некотором подпространстве $\mathbf{H}_0$: $\{\omega + \tau t^0\}$ пространства $\mathbf{H}$. Таким образом, итерационный процесс для нахождения решения уравнения (1) можно строить в виде

$$u^{k+1} = u^k + t^k \tau_k, \quad t^k = Cr^k, \quad r^k = f - Au^k; \quad \tau_k = \frac{(t^k, t^k)}{(Mt^k, t^k)}. \tag{15}$$

Следовательно, элемент

$$u^{k+1} = u^k + \tau_k t^k \tag{16}$$

минимизирует функционал $\Omega_M (u + \tau t)$.

1. Метод наискорейшего спуска

Различные варианты методов наискорейшего спуска отличаются друг от друга выбором оператора M, задающего метрику пространства $\mathbf{H}_M$. Пусть A — положительно определенный самосопряженный оператор

$$A : \mathbf{H} \to \mathbf{H}.$$

Полагая в (15) $M = A$, $C = I$, получим итерационный процесс метода скорейшего спуска

$$u^{k+1} = u^k + \tau_k r^k, \quad r^k = f - Au^k, \tag{17}$$

$$\tau_k = \frac{(r^k, r^k)}{(Ar^k, r^k)} \qquad (k = 0, 1, 2, \ldots), \tag{18}$$

u^0 — произвольный элемент из $\mathbf{H}$.

Для построения формул итерационного процесса неявного метода скорейшего спуска исходное уравнение $Au = f$ заменяется эквивалентным уравнением

$$Dv = \psi, \tag{19}$$

где $D = B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$, $v = B^{\frac{1}{2}}u$; $\psi = B^{-\frac{1}{2}}f$, $B = B^* > 0$.

Для решения уравнения (19) применяется итерационный процесс наискорейшего спуска:

$$v^{k+1} = v^k + \tilde{\tau}_k \tilde{r}^k, \quad \tilde{r}^k = \psi - Dv^k, \tag{20}$$

$$\tilde{\tau}_k = \frac{(\tilde{r}^k, \tilde{r}^k)}{(D\tilde{r}^k, \tilde{r}^k)}. \tag{21}$$

Выразим $\tilde{r}^k$ через $r^k = f - Au^k$:

$$\tilde{r}^k = B^{-\frac{1}{2}}f - B^{-\frac{1}{2}}AB^{-\frac{1}{2}}B^{\frac{1}{2}}u^k = B^{-\frac{1}{2}}(f - Au^k) = B^{-\frac{1}{2}}r^k. \tag{22}$$

Действуя на (20) оператором $B^{-\frac{1}{2}}$, получим:

$$u^{k+1} = u^k + \tilde{\tau}_k w^k, \quad w^k = B^{-1}r^k \tag{23}$$

и, учитывая (22), из (21) найдем:

$$\tilde{\tau}_k = \frac{(w^k, r^k)}{(Aw^k, w^k)}. \tag{24}$$

Итерационный процесс (23), (24) называют *неявным методом наискорейшего спуска*. Для доказательства сходимости итерационных градиентных методов решения операторных уравнений первого рода воспользуемся следующей леммой.

Лемма 1. *Пусть D — несамосопряженный положительно определенный оператор*

$$\gamma_1 I \leqslant D \leqslant \gamma_2 I, \quad \gamma_i > 0 \qquad (i = 1, 2) \tag{25}$$

и существуют числа $\tau_* > 0$, $q_* > 0$, $q_* \in (0, 1)$ *такие, что для них справедливы неравенства:*

$$\gamma_1\tau_* \leqslant 1 - q_*^2 \leqslant \gamma_2\tau_*, \tag{26}$$

$$\| I - \tau_* D \| \leqslant q_* < 1. \tag{27}$$

Тогда имеет место неравенство:

$$(Dx, x)^2 \geqslant (1 - q_*^2) \| Dx \|^2 \| x \|^2. \tag{28}$$

Доказательство. Рассмотрим

$$\| (I - \tau_* D) x \|^2 \leqslant q_*^2 \| x \|^2. \tag{29}$$

С другой стороны,

$$\| (I - \tau_* D) x \|^2 = \| x \|^2 - 2\tau_* (Dx, x) + \tau^2 \| Dx \|^2. \tag{30}$$

Из (26) и (27) имеем:

$$\| Dx \|^2 \leqslant \frac{2}{\tau_*} (Dx, x) - \frac{1 - q_*^2}{\tau_*^2} \| x \|^2 =$$

$$= \frac{(Dx, x)^2}{\tau_* \| x \|^2} \left(\frac{2 \| x \|^2}{(Dx, x)} - \frac{1 - q_*^2}{\tau_*} \cdot \frac{\| x \|^4}{(Dx, x)^2} \right) = \frac{(Dx, x)^2}{\tau_* \| x \|^2} \psi(\beta), \tag{31}$$

где

$$\psi(\beta) = 2\beta - \frac{1 - q_*^2}{\tau_*} \beta^2; \quad \beta = \frac{\| x \|^2}{(Dx, x)} \quad \text{и} \quad \gamma_2^{-1} \leqslant \beta \leqslant \gamma_1^{-1},$$

$$\psi'(\beta) = 2 \left(1 - \frac{1 - q_*^2}{\tau_*} \beta \right).$$

При $\beta_0 = \dfrac{\tau_*}{1 - q_*^2}$

$$\psi'(\beta) = 0, \quad \psi''(\beta) = - \frac{2(1 - q_*^2)}{\tau_*} < 0,$$

т. е.

$$\max_{\beta \in [\gamma_2^{-1}, \gamma_1^{-1}]} \psi(\beta) = \beta_0 \left(2 - \frac{1 - q_*^2}{\tau_*} \beta_0 \right) = \beta_0. \tag{32}$$

Подставим (32) в (31)

$$\| Dx \|^2 \leqslant \frac{(Dx, x)^2}{(1 - q_*^2) \| x \|^2}.$$

Для оценки скорости сходимости итерационного процесса (17), (18) имеет место теорема.

Теорема 1. *Пусть* $A = A^*$, $\gamma_1 I \leqslant A \leqslant \gamma_2 I$, $\gamma_2 \geqslant \gamma_1 > 0$. *Тогда метод скорейшего спуска* (17), (18) *сходится со скоростью геометрической прогрессии:*

$$\| \varepsilon^{k+1} \|_A \leqslant q^k \| \varepsilon^0 \|_A, \tag{33}$$

где

$$q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} < 1.$$

Доказательство. Рассмотрим итерационную формулу метода скорейшего спуска

$$u^{k+1} = u^k + \tau_k r^k, \quad \tau_k = \frac{(r^k, r^k)}{(Ar^k, r^k)}, \quad r^k = f - Au^k \qquad (k = 0, 1, 2, \ldots),$$

тогда

$$r^{k+1} = r^k - \tau_k A r^k. \tag{17'}$$

Введем в рассмотрение величину

$$\eta^k = A^{-\frac{1}{2}} r^k.$$

Применив к обеим частям равенств (17′) оператор $A^{-\frac{1}{2}}$, получим:

$$\eta^{k+1} = \eta^k - \tau_k A\eta^k, \quad \tau_k = \frac{(A\eta^k, \eta^k)}{(A\eta^k, \eta^k)}.$$

Вычислим квадрат нормы $\|\eta^{k+1}\|^2$

$$\|\eta^{k+1}\|^2 = \|\eta^k\|^2 - 2\tau_k (A\eta^k, \eta^k) + \tau_k^2 \|A\eta^k\|^2 = \|\eta^k\|^2 - \frac{(A\eta^k, \eta^k)^2}{\|A\eta^k\|^2} =$$

$$= \left(1 - \frac{(A\eta^k, \eta^k)}{\|\eta^k\|^2 \|A\eta^k\|^2}\right) \|\eta^k\|^2. \tag{34}$$

Введем в рассмотрение величины

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1}.$$

Тогда $\| I - \tau_0 A \| \leqslant q < 1$, $\tau_0\gamma_1 \leqslant 1 - q^2 \leqslant \tau_0\gamma_2$. Из леммы 1 следует, что

$$\frac{(A\eta^k, \eta^k)^2}{\|\eta^k\|^2 \|A\eta^k\|^2} \geqslant 1 - q^2. \tag{35}$$

Используя оценку (35), из (34) получим: $\|\eta^{k+1}\|^2$

$$\|\eta^{k+1}\|^2 \leqslant (1 - (1 - q^2)) \|\eta^k\|^2 = q^2 \|\eta^k\|^2. \tag{34'}$$

Так как

$$\|\eta^{k+1}\|^2 = (A^{-1} r^{k+1}, r^{k+1}) = (A^{-1}(Au - Au^{k+1}), Au - Au^{k+1}) =$$

$$= (A(u - u^{k+1}), u - u^{k+1}) = \|u - u^{k+1}\|_A^2 = \|\varepsilon^{k+1}\|_A^2,$$

то, учитывая (34′), имеем:

$$\|\varepsilon^{k+1}\|_A \leqslant q \|\varepsilon^k\|_A \leqslant q^{k+1} \|\varepsilon^0\|_A.$$

Применяя результаты теоремы 1 к схеме (19) и переходя от v^k к u^k при помощи подстановки $v^k = B^{\frac{1}{2}} u^k$, для неявной схемы скорейшего спуска при $A = A^*$, $B = B^*$ получим оценку

$$\|u^k - u\|_A \leqslant q^k \|u^0 - u\|_A.$$

Следовательно, для метода наискорейшего спуска верна та же оценка скорости сходимости, что и для явной двухшаговой схемы с

$\tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2}$. Метод наискорейшего спуска не использует явно информацию о спектре, но требует дополнительной затраты арифметических действий на подсчет скалярных произведений вида (18). Из-за нелинейного характера метода получить более точные характеристики быстроты сходимости метода пока не удалось. Однако в ряде работ (см. напр. [49]) отмечено, что в первых итерациях стремление приближенного решения к точному существенно быстрее, чем в других сравнимых с ним по асимптотической скорости сходимости методах, т. е. гармоники ряда Фурье для невязки внутри спектрального интервала подавляются быстро, а подавление гармоник, соответствующих окрестностям собственных чисел γ_1 и γ_2, происходит медленнее и на некотором этапе процесс выходит на асимптотический режим, оценивающийся неравенством вида (33).

2. Метод минимальных невязок

Рассмотрим неградиентные методы построения итерационных процессов, в которых производится спуск по отношению к одному функционалу в направлении антиградиента другого функционала. Итерационный метод невязок строится в результате минимизации функционала невязки $\Phi(u) = (f - Au, f - Au)$ в направлении антиградиента функционала ошибок в пространстве $\mathbf{H}_A$, где A — положительно определенный оператор; $A : \mathbf{H} \to \mathbf{H}$.

Пусть итерационный процесс имеет вид

$$u^{k+1} = u^k + \tau_k r^k, \tag{36}$$

$$r^k = f - Au^k, \tag{37}$$

где τ_k определяется из условия минимизации функционала

$$\Phi(u) = (f - Au, f - Au).$$

Рассмотрим $r^{k+1} = f - Au^{k+1} = r^k - \tau_k A r^k$.

$$\Phi_A(r^{k+1}, r^{k+1}) = \|r^k\|^2 - 2\tau_k (Ar^k, r^k) + \tau_k^2 \|Ar^k\|^2 = \|r^k\|^2 - \frac{(Ar^k, r^k)^2}{\|Ar^k\|^2} +$$

$$+ \|Ar^k\|^2 \left[\tau_k - \frac{(Ar^k, r^k)}{\|Ar^k\|^2}\right]^2. \tag{38}$$

Очевидно, функционал $\Phi_A(r^{k+1})$ принимает минимальное значение при

$$\tau_k = \frac{(Ar^k, r^k)}{\|Ar^k\|^2}. \tag{39}$$

Таким образом, итерационный процесс метода минимальных невязок для решения операторного уравнения $Au = f$ с положительно определенным оператором осуществляется по формулам (36), (37), (39). Применяя к уравнению (19) формулы метода минимальных невязок

(36), (37), (39) и переходя от v^k к u^k, получим:

$$u^{k+1} = u^k + \tau_k w^k, \quad w^k = B^{-1} r^k, \tag{40}$$

$$\tau_k = \frac{(Aw^k, w^k)}{(B^{-1}Aw^k, Aw^k)} \tag{41}$$

— формулы неявного метода минимальных невязок.

Заменим уравнение $Au = f$ эквивалентным уравнением

$$D_1 v = f, \tag{42}$$

где

$$D_1 = AB^{-1}, \quad v = Bu, \quad B = B^* > 0. \tag{43}$$

Применяя к (42) формулы метода минимальных невязок и переходя от v^k к u^k, можно получить формулы так называемого метода расщеплений с вариационной оптимизацией

$$u^{k+1} = u^k + \tau_k w^k, \quad r^k = f - Au^k, \quad w^k = B^{-1} r^k, \tag{44}$$

$$\tau_k = \frac{(Aw^k, r^k)}{\| Aw^k \|^2}. \tag{45}$$

Для оценки скорости сходимости метода минимальных невязок имеет место следующая теорема.

Теорема 2. *Пусть A — положительно определенный оператор, $A : \mathbf{H} \to \mathbf{H}$ и выполнены условия*

$$\gamma_1 I \leqslant A \leqslant \gamma_2 I, \quad \gamma_2 > \gamma_1 > 0, \tag{46}$$

$$\| A_k \| = \frac{1}{2} \| A - A^* \| \leqslant \gamma_3, \quad \gamma_3 \geqslant 0. \tag{47}$$

Тогда метод минимальных невязок

$$u^{k+1} = u^k + \tau_k r^k, \quad \tau_k = \frac{(Ar^k, r^k)}{\| Ar^k \|^2} \tag{48}$$

сходится со скоростью геометрической прогрессии, т. е. справедлива оценка

$$\| r^k \| = \| f - Au^k \| \leqslant \bar{q}^k \| f - Au^0 \| \tag{49}$$

или

$$\| \varepsilon^k \|_{A^*A} = \| u - u^k \|_{A^*A} \leqslant \bar{q}^k \| \varepsilon^0 \|_{A^*A},$$

где

$$\bar{q} = \frac{q + \varkappa}{1 + \varkappa q} < 1, \quad q = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} < 1, \quad \varkappa = \gamma_3 / \sqrt{\gamma_3^2 + \gamma_2 \gamma_1} < 1. \tag{50}$$

Доказательство. По аналогии с (34) имеем:

$$\| r^{k+1} \| = \left(1 - \frac{(Ar^k, r^k)^2}{\| r^k \|^2 \| Ar^k \|^2} \right) \| r^k \|^2.$$

Полагая в (26), (27) $D = A$

$$\tau^* = \frac{2}{(\gamma_1 + \gamma_2)(1 + \varkappa q)}, \quad q_* = \bar{q}$$

и учитывая (28), получим (49). Если оператор A — самосопряжен, то $\varkappa = 0$, $\bar{q} = q$ и

$$\|r^k\| = \|f - Au^k\| \leqslant q^k \|r^0\|. \tag{51}$$

Асимптотическая скорость сходимости метода минимальных невязок

$$v_a \approx \frac{2}{\beta}, \quad \beta = \frac{\gamma_2}{\gamma_1}. \tag{52}$$

Аналогично используя результаты теоремы 2 применительно к уравнению (19), будем иметь:

$$\|Dv^k - \psi\| \leqslant \bar{q}^k \|Dv^0 - \psi\|$$

или

$$\|Au^k - f\|_{B^{-1}} \leqslant \bar{q}^k \|Au^0 - f\|_{B^{-1}} \quad \text{при} \quad B = B^*, \ A \neq A^*.$$

В неявных итерационных методах скорейшего спуска (23), (24) метода минимальных поправок (40), (41) и метода вида (44), (45) для определения w^k требуется решить уравнение

$$Bw^k = r^k, \quad r^k = f - Au^k, \tag{54}$$

а в случае неявного метода минимальных поправок наряду с решением уравнения (54) требуется еще решить и уравнение

$$Bv^k = Aw^k \tag{55}$$

для определения по формуле (41) параметра τ_k. Решение уравнений (54), (55) можно осуществить либо прямым методом, либо при помощи некоторого итерационного метода с начальными значениями $w^0 = 0$ и соответственно $v^0 = 0$.

В частности, внутренние итерации для нахождения поправки можно проводить, используя метод наискорейшего спуска или метод минимальных поправок. Внутренние итерации для решения уравнений (54), (55) должны проводиться по возможности более экономичными методами, при этом оператор B задается либо в явном виде (например, B — факторизованный оператор)

$$B = (I + \sigma A_1)(I + \sigma A_2), \quad A_2 = A_1^*), \tag{56}$$

либо строится в результате некоторого вычислительного процесса (см., например [27], [28]). Схема реализации алгоритма (23), (24) при задании оператора B в виде (56) может быть следующая:

$$(I + \sigma A_1)\, w^{k-\frac{1}{2}} = r^k, \quad r^k = f - Au^k,$$

$$(I + \sigma A_2)\, w^k = w^{k-\frac{1}{2}};$$

$$z^k = Aw^k;$$

$$\tau_k = \frac{(w^k, r^k)}{(z^k, w^k)}; \quad u^{k+1} = u^k + \tau_k w^k. \tag{57}$$

Схема реализации алгоритма расщепления с вариационной оптимизацией (44), (45) при $B = \prod_{\alpha=1}^{s} (I + \sigma_k A_\alpha)$ имеет следующий вид:

$$
\begin{gathered}
(I + \sigma_k A_1)\, w^{k+\frac{1-s}{s}} = r^k, \quad r^k = f - Au^k, \\
(I + \sigma_k A_2)\, w^{k+\frac{2-s}{s}} = w^{k+\frac{1-s}{s}}, \\
\cdots\cdots\cdots\cdots\cdots \\
(I + \sigma_k A_s) w^k = w^{k-\frac{1}{s}}; \\
z^k = Aw^k, \quad \tau_k = \frac{(z^k, r^k)}{(z^k, z^k)}, \\
u^{k+1} = u^k + \tau_k w^k.
\end{gathered}
\tag{58}
$$

Задача оптимального выбора параметров σ_k в схеме (58) пока не решена. При $s = 2$ и коммутирующих операторах A_i $(i = 1, 2)$ полагают

$$\sigma_k = 2\tau_{k-1}.$$

Подчеркнем, что вариационные итерационные методы не требуют сведения о спектре оператора. В этом смысле их можно считать самонастраивающимися на оптимальный процесс. Однако, не используя явно информацию о спектре оператора, они требуют дополнительной затраты действий на подсчет некоторых скалярных произведений вида (18), (24), (39), (45). Следует отметить также комбинированные способы итераций, основанные на использовании сходящихся итерационных методов вида

$$\tilde{u}^k = u^k - H_k (Au^k - f) \qquad (k = 0, 1, 2, \ldots) \tag{59}$$

и минимизации функционала ошибок $\Omega_M(\tilde{u}^k)$ на некотором подпространстве $\mathbf{H}_k$ пространства $\mathbf{H}$. Тогда, если $\Omega_M(\tilde{u}^k) \leqslant \Omega_M(u^k)$, то полагают, что

$$u^{k+1} = \tilde{u}^k + w^k, \tag{60}$$

где w^k находится из условия:

$$\min_{w^k \in \mathbf{H}_k} \Omega_M(\tilde{u}^k + w^k). \tag{61}$$

При таком подходе к построению итерационных методов для ускорения их сходимости обычно увеличивают размерность подпространства $\mathbf{H}_k$, т. е. вместо линейного подпространства рассматривают некоторое подпространство размерности n, ибо при фиксированном n методы типа (59), (60) перестают быть оптимальными при $\varepsilon \to 0$. Но с увеличением размерности подпространства не только возрастает сходимость, но и значительно усложняется оператор для нахождения w^k.

Однако в целом ряде работ Н. С. Бахвалова, В. И. Лебедева, Р. П. Федоренко, В. П. Ильина на примере решения разностными методами эллиптических, кинетических и интегральных уравнений показано, что значительное усложнение оператора для ускорения сходимости итерационного процесса все же дает возможность построить быстросходящиеся итерационные процессы, которые в то же время являются неулучшаемыми по порядку затраченных действий (т. е. когда общее число действий в итерационном методе равно по порядку числу действий в одной итерации).

§ 7. ДВУХШАГОВЫЕ ИТЕРАЦИОННЫЕ МЕТОДЫ

Рассмотрим уравнение

$$Au = f \tag{1}$$

(A — линейный оператор, действующий в гильбертовом пространстве **H**).

Двухшаговый неявный итерационный процесс, оставляющий неподвижной точку решения уравнения (1), можно записать следующим образом:

$$C_k(u^{k+1} - u^k) + H_k(Au^k - f) - D_k(u^k - u^{k-1}) = 0 \tag{2}$$
$$(k = 0, 1, 2, \ldots).$$

Здесь C_k, H_k, D_k — некоторые операторы, $D_0 \equiv 0$, u^0 — произвольный элемент, принадлежащий **H**.

В самом деле, общий вид итерационного процесса, связывающего три итерации u^{k+1}, u^k, u^{k-1}, следующий:

$$C_k u^{k+1} + S_k u^k + D_k u^{k-1} = \psi_k \quad (k = 0, 1, \ldots), \quad D_0 = 0, \tag{3}$$

где C_k, S_k, D_k — некоторые операторы, действующие в гильбертовом пространстве **H**.

Итерационный процесс вида (3) должен оставлять неподвижной точку u ($u \equiv A^{-1}f$) решения уравнения (1), т. е.

$$(C_k + S_k + D_k) A^{-1} f = \psi_k.$$

Если обозначить через H_k оператор

$$(C_k + S_k + D_k) A^{-1} = H_k,$$

то

$$\psi_k = H_k f, \tag{4}$$

$$S_k = -C_k - D_k + H_k A. \tag{5}$$

Подставив в (3) соотношения (4) и (5), получим (2).

Итерационные процессы вида (2) иногда называют трехчленными итерационными процессами или итерационными схемами второго порядка. Двухшаговая итерационная схема (2) с постоянными операторами и параметрами может быть записана в следующем каноническом виде:

$$B\frac{u^{k+1} - u^{k-1}}{2\tau} + K(u^{k+1} - 2u^k + u^{k-1}) + Au^k = f \quad (k = 1, 2, \ldots). \tag{6}$$

Здесь $C_k = 2\tau K + B$, $D_k = 2\tau K - B$, $H_k = 2\tau I$, u^0, u^1, ... — некоторые произвольные элементы, принадлежащие $\mathbf{H}$, либо u^0 — произвольный элемент, принадлежащий $\mathbf{H}$, а u^1 — решение двухслойной схемы

$$B\,\frac{u^1 - u^0}{2\tau} + Au^0 = f. \tag{7}$$

Если в (2) выбрать

$$C_k = I, \quad H_k = a_k I, \quad D_k = -d_k I, \tag{8}$$

то u^{k+1} будет находиться из соотношений

$$u^{k+1} = u^k - a_k(Au^k - f) - d_k(u^k - u^{k-1}), \quad d_0 = 0, \quad k = 0, 1, 2, \ldots \tag{9}$$

Для погрешности $\varepsilon^{k+1} = u - u^{k+1}$ имеем:

$$\varepsilon^{k+1} = \varepsilon^k - a_k A\varepsilon^k - d_k(\varepsilon^k - \varepsilon^{k-1}) = (I - a_k A - d_k I)\,\varepsilon^k + d_k\varepsilon^{k-1}. \tag{10}$$

Последовательно применяя равенство (10), получим:

$$\varepsilon^1 = (I - a_0 A)\,\varepsilon^0 = P_1(A)\,\varepsilon^0;$$

$$\varepsilon^2 = [(1 - d_1)I - a_1 A]\,\varepsilon^1 + d_1\varepsilon^0 = P_2(A)\,\varepsilon^0;$$

где

$$P_2(A) = [(1 - d_1)I - a_1 A]\,P_1(A) + d_1 P_0(A),$$

$$I = P_0(A);$$

$$\varepsilon^3 = [(1 - d_2)I - a_2 A]\,P_2(A)\,\varepsilon^0 + d_2 P_1(A)\,\varepsilon^0 = P_3(A)\,\varepsilon^0;$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \tag{11}$$

$$\varepsilon^{k+1} = [(1 - d_k)I - a_k A]\,\varepsilon^k + d_k\varepsilon^{k-1} = P_{k+1}(A)\,\varepsilon^0.$$

Здесь $P_k(t)$ — последовательность многочленов, подчиненных рекуррентным соотношениям

$$P_{k+1}(t) = (1 - d_k - a_k t)\,P_k(t) + d_k P_{k-1}(t), \quad d_0 = 0, \quad P_0(t) = 1, \quad k = 0, 1, \ldots \tag{12}$$

и условиям нормировки

$$P_{k+1}(0) = 1. \tag{13}$$

Для операторных уравнений второго рода вида

$$u = Ku + \psi \tag{14}$$

трехчленная итерационная схема (9) может быть представлена в виде

$$u^{k+1} = (1 - a_k - d_k)\,u^k + a_k K u^k + d_k u^{k-1} + a_k\psi, \quad k = 0, 1, 2, \ldots, \tag{15}$$

где $d_0 = 0$, $u^0 \in \mathbf{H}$ — произвольный элемент.

Для погрешности $\varepsilon^{k+1} = u - u^{k+1}$ итерационного процесса (15) будет выполняться соотношение

$$\varepsilon^{k+1} = [(1 - a_k - d_k)I + a_k K]\,\varepsilon^k + d_k\varepsilon^{k-1}. \tag{16}$$

Последовательно применяя равенство (16) аналогично (11), (12), получим:

$$\varepsilon^{k+1} = Q_{k+1}(K)\,\varepsilon^0, \tag{17}$$

где $Q_{k+1}(t)$ — многочлен степени $k+1$, для которого имеет место следующая рекуррентная формула:

$$Q_{k+1}(t) = (1 - a_k - d_k + a_k t)\, Q_k(t) + d_k Q_{k-1}(t),$$
$$Q_0(t) = 1, \quad d_0 = 0, \quad k = 0, 1, 2, \ldots \tag{18}$$

и

$$Q_k(1) = 1. \tag{19}$$

Таким образом, исследование сходимости итерационных методов вида (9), (15) связано с исследованием поведения многочленов, подчиненных соответственно рекуррентным соотношениям (12), (18).

Для определения коэффициентов многочленов вида (12), (18) можно воспользоваться последовательностью многочленов ортогональных в некоторой среднеквадратической метрике с весом $\rho(t) > 0$, который порождается последовательностью чисел a_k, d_k.

В частности, при наличии априорной информации о спектре оператора A можно определить коэффициенты a_k, d_k в итерационных процессах (9), (15) через полиномы Чебышева первого рода и достаточно эффективно оценить скорость сходимости итерационного процесса.

Пусть оператор K уравнения (14) является самосопряженным и все собственные значения оператора K принадлежат интервалу $[-q, q]$, $q < 1$, т. е. $K = K^*$,

$$\|K\| = q < 1.$$

В качестве последовательности многочленов $Q_k(t)$, удовлетворяющих рекуррентным соотношениям (18) и условию нормировки (19), выберем полиномы Чебышева, определенные на отрезке $[-q, q]$ (приложение, § 2).

Тогда

$$Q_{k+1}(t) = \frac{T_{k+1}\left(\frac{t}{q}\right)}{T_{k+1}\left(\frac{1}{q}\right)}. \tag{20}$$

Учитывая рекуррентные соотношения для полиномов Чебышева первого рода $(k+1)$-го порядка, определенных на отрезке $[-q, q]$,

$$T_{k+1}\left(\frac{t}{q}\right) = 2\frac{t}{q} T_k\left(\frac{t}{q}\right) - T_{k-1}\left(\frac{t}{q}\right), \tag{21}$$

трехчленные соотношения (18) и формулу (20), получаем выражение для коэффициентов a_k, d_k при $k \geqslant 1$

$$a_k = \frac{2T_k\left(\frac{1}{q}\right)}{qT_{k+1}\left(\frac{1}{q}\right)}, \quad d_k = 1 - a_k. \tag{22}$$

Для оценки скорости сходимости итерационного процесса

$$u^{k+1} = a_k(Ku^k - u^{k-1} + \psi) + u^{k-1}, \tag{23}$$

где a_k определяются по формуле (22), будет иметь место следующее неравенство:

$$\|\varepsilon^{k+1}\| \leqslant$$

$$\leqslant \frac{\|\varepsilon^0\|}{\left|T_{k+1}\left(\frac{1}{q}\right)\right|} = \frac{2\|\varepsilon^0\|}{\left|\left(\frac{1}{q}+\sqrt{\frac{1}{q^2}-1}\right)^{k+1}+\left(\frac{1}{q}-\sqrt{\frac{1}{q^2}-1}\right)^{k+1}\right|}. \quad (24)$$

Очевидно, если оператор A уравнения (1) удовлетворяет условиям

$$A = A^*, \quad \gamma_1 I \leqslant A \leqslant \gamma_2 I, \quad \gamma_2 \geqslant \gamma_1 > 0, \quad (25)$$

то, полагая, что

$$K = I - \frac{2}{\gamma_1+\gamma_2} A, \quad \psi = \frac{2}{\gamma_1+\gamma_2} f,$$

уравнение (1) приводится к виду (14) с самосопряженным оператором K, причем $\mathrm{Sp}K \in [-q, q]$, где $q = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}$

$$\|K\| = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1} = q < 1.$$

Двухшаговый итерационный процесс (23) будет иметь вид

$$u^{k+1} = a_k\left[u^k - u^{k-1} - \frac{2}{\gamma_1+\gamma_2}(Au^k - f)\right] + u^{k-1},$$

$$a_k = \frac{2T_k\left(\frac{1}{q}\right)}{qT_{k+1}\left(\frac{1}{q}\right)}, \quad q = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1} \quad (26)$$

и носит название трехчленного чебышевского итерационного процесса. Для оценки величины $\|\varepsilon^{k+1}\|$ трехчленного чебышевского итерационного процесса (26) можно воспользоваться неравенством (24), причем

$$v_a \approx 2\sqrt{\frac{\gamma_1}{\gamma_2}}. \quad (27)$$

При $k \to \infty$, $a_k \to \frac{2}{1+\sqrt{1-q^2}}$ и метод (26) переходит в так называемый метод верхней релаксации.

Пусть в (15)

$$a_k = 1 + \alpha_k, \quad d_k = -\alpha_k,$$

т. е. рассмотрим подкласс итерационных методов (15)

$$u^{k+1} = (1+\alpha_k)(Ku^k + \psi) - \alpha_k u^{k-1} \quad (28)$$

с оператором

$$K = I - \frac{2}{\gamma_1+\gamma_2} A, \quad \psi = \frac{2}{\gamma_1+\gamma_2} f,$$

предполагая наличие априорной информации вида (25) об операторе A. Если в итерационном процессе (28) α_k выбрать из некоторой последовательности, состоящей из нулей и единиц, то получим двухшаговый итерационный процесс (см. напр., [78]). В этом процессе после k-го

шага выбирается следующая последовательность:

$$\alpha_k = \alpha_{k+1} = \cdots = \alpha_{k+i} = 0, \quad \alpha_{k+i+1} = \alpha_{k+i+2} = \cdots = \alpha_{k+N-1} = 1,$$
$$\alpha_{k+N} = 0, \quad i \geqslant 0. \tag{29}$$

Тогда для погрешности $\varepsilon^l = u - u^l$ $(l = k+1, \ldots, k+N)$ имеем:

$$\varepsilon^{k+1} = K\varepsilon^k; \ \ldots; \ \varepsilon^{k+i} = K^i\varepsilon^k,$$
$$\varepsilon^{k+i+1} = K\varepsilon^{k+i} = T_1(K)\,\varepsilon^{k+i},$$
$$\varepsilon^{k+i+2} = 2K\varepsilon^{k+i+1} - \varepsilon^{k+i} = T_2(K)\,\varepsilon^{k+i},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\varepsilon^{k+N} = T_{N-1}(K)\,\varepsilon^{k+i} = T_{N-i}(K)\,K^i\varepsilon^k,$$

где $T_l(t) = 2tT_{l-1}(t) - T_{l-2}(t)$ — последовательность многочленов Чебышева. Так как $\|K\| \leqslant 1$, то $\|T_l(K)\| \leqslant \max\limits_{|t|\leqslant 1} |T_l(t)| \leqslant 1$.

Итерационный процесс (28) с параметрами α_l, которые находятся по формуле (29), будет сходиться, причем имеет место следующая оценка быстроты сходимости:

$$\|\varepsilon^{k+N}\| \leqslant \|K\|^i \|T_{N-i}(K)\| \|\varepsilon^k\|. \tag{30}$$

Введение наряду с коэффициентами $\alpha_l = 1$ коэффициентов $\alpha_m = 0$ служит в основном для погашения ошибок от округлений.

Асимптотическая скорость сходимости трехчленного итерационного процесса (28), (29) при $K = I - \dfrac{2}{\gamma_1+\gamma_2}A$, $\psi = \dfrac{2}{\gamma_1+\gamma_2}f$ будет характеризоваться величиной (27).

Рассмотрим трехчленный итерационный процесс (28) при $\alpha_k = \alpha = \mathrm{const}$ и произвольных $u^0, u^1 \in \mathbf{H}$ и оценим быстроту его сходимости.

Для погрешности $\varepsilon^{k+1} = u - u^{k+1}$ будем иметь:

$$\varepsilon^{k+1} = (1+\alpha)K\varepsilon^k - \alpha\varepsilon^{k-1}$$

или

$$\varepsilon^2 = (1+\alpha)K\varepsilon^1 - \alpha\varepsilon^0 = \Phi_1(K)\,\varepsilon^1 - \alpha\Phi_0(K)\,\varepsilon^0,$$
$$\Phi_1(K) = (1+\alpha)K, \quad \Phi_0(K) = I,$$
$$\varepsilon^3 = [(1+\alpha)K\Phi_1(K) - \alpha\Phi_0(K)]\,\varepsilon^1 - \alpha\Phi_0(K)\,\varepsilon^0 = \Phi_2(K)\,\varepsilon^1 - \alpha\Phi_0(K)\,\varepsilon^0,$$
$$\Phi_2(K) = (1+\alpha)K\Phi_1(K) - \alpha\Phi_0(K), \tag{31}$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\varepsilon^{k+1} = \Phi_k(K)\,\varepsilon^1 - \alpha\Phi_{k-1}(K)\,\varepsilon^0,$$

где

$$\Phi_k(t) = (1+\alpha)\,t\Phi_{k-1}(t) - \alpha\Phi_{k-2}(t), \quad \Phi_0(t) = 1, \quad k = 1, 2, \ldots \tag{32}$$

Пусть $K = I - \dfrac{2}{\gamma_1+\gamma_2}A$, $\psi = \dfrac{2}{\gamma_1+\gamma_2}f$, где оператор A удовлетворяет условиям (25).

Положим

$$\Phi_{k+1}(t) = \eta^{k+1}\Psi_{k+1}\left(\frac{t}{q}\right), \quad \eta \neq 0, \quad q = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}.$$

Тогда получим:

$$\eta^{k+1}\Psi_{k+1}\left(\frac{t}{q}\right)=(1+\alpha)\eta^k\frac{qt}{q}\Psi_k\left(\frac{t}{q}\right)-\alpha\eta^{k-1}\Psi_k\left(\frac{t}{q}\right)$$

или

$$\Psi_{k+1}\left(\frac{t}{q}\right)=(1+\alpha)\frac{q}{\eta}\cdot\frac{t}{q}\Psi_k\left(\frac{t}{q}\right)-\frac{\alpha}{\eta^2}\Psi_k\left(\frac{t}{q}\right), \tag{33}$$

где

$$-1\leqslant t_1=\frac{t}{q}\leqslant 1.$$

Учитывая рекуррентные соотношения для полиномов Чебышева второго рода

$$U_{k+1}(t_1)=2t_1U_k(t_1)-U_{k-1}(t_1),\quad |t_1|\leqslant 1,\quad k\geqslant 1,$$

где

$$U_{k+1}(t_1)=\frac{\sin(k+1)\arccos t_1}{\sin(\arccos t_1)},\quad U_0(t_1)=1,\quad U_1(t_1)=2t_1,$$

получим, что при

$$\frac{1+\alpha}{2\eta}q=\frac{\alpha}{\eta^2}=1\quad\left(\text{т. е. при}\quad \alpha=\left(\frac{1-\sqrt{\frac{\gamma_1}{\gamma_2}}}{1+\sqrt{\frac{\gamma_1}{\gamma_2}}}\right)^2\right) \tag{34}$$

$$\Psi_0(t_1)=1,\quad \Psi_1(t_1)=2t_1,$$

$$\Psi_{k+1}(t_1)=U_{k+1}(t_1),\quad |t_1|\leqslant 1,\quad k=1,2,\ldots$$

Далее, учитывая рекуррентные соотношения для полиномов Чебышева первого и второго рода (приложение, § 2)

$$T_k(t_1)=U_k(t_1)-t_1U_{k-1}(t_1),\quad |t_1|\leqslant 1,$$

из (31) получаем:

$$\varepsilon^{k+1}=\eta^{k+1}\left[U_k(\tilde{K})\left(\frac{1}{\eta}\varepsilon^1-\tilde{K}\varepsilon^0\right)+T_{k+1}(\tilde{K})\varepsilon^0\right],\quad \tilde{K}=\frac{1}{q}K.$$

Так как

$$\|U_k(\tilde{K})\|\leqslant\max_{|t_1|<1}|U_k(t_1)|\leqslant k+1\quad\text{и}\quad\frac{\varepsilon^1}{\eta}=\left(\frac{q}{\eta}-1\right)\frac{\varepsilon^1}{q}+\frac{1}{q}\varepsilon^1,$$

то

$$\|\varepsilon^{k+1}\|\leqslant\eta^{k+1}\left\{\frac{k+1}{q}\left(\frac{q}{\eta}-1\right)\|\varepsilon^1\|+\frac{k+1}{q}\|\varepsilon^1-K\varepsilon^0\|+\|\varepsilon^0\|\right\}.$$

Выберем $u^1=Ku^0+\frac{2}{\gamma_1+\gamma_2}f$, $u^0\in\mathbf{H}$ — произвольный элемент, тогда

$$\|\varepsilon^{k+1}\|\leqslant q_{k+1}\|\varepsilon^0\|,$$

$$q_{k+1}=\left(\frac{1-\sqrt{\frac{\gamma_1}{\gamma_2}}}{1+\sqrt{\frac{\gamma_1}{\gamma_2}}}\right)^{k+1}\cdot\left(1+\frac{2\sqrt{\frac{\gamma_1}{\gamma_2}}}{1+\frac{\gamma_1}{\gamma_2}}(k+1)\right). \tag{35}$$

Таким образом, имеет место следующая теорема.

Теорема 1. *Пусть A — самосопряженный оператор и выполнены условия* (25). *Тогда для двухшаговой итерационной схемы* (28) *при*

$$\alpha_k = \alpha = \left(\frac{1-\sqrt{\frac{\gamma_1}{\gamma_2}}}{1+\sqrt{\frac{\gamma_1}{\gamma_2}}}\right)^2, \quad K = I - \frac{2}{\gamma_1+\gamma_2}A, \quad \psi = \frac{2}{\gamma_1+\gamma_2}f, \tag{36}$$

произвольном $u^0 \in \mathbf{H}$ и $u^1 = Ku^0 + \psi$ имеет место априорная оценка вида (35).

Очевидно, для неявной схемы (6) при $\tau = \frac{1}{\sqrt{\gamma_1\gamma_2}}$, $K = \frac{\gamma_1+\gamma_2}{4}B$ при произвольном u^0 из $\mathbf{H}$ и u^1, удовлетворяющим уравнению

$$B\frac{u^1-u^0}{\tau} + Au^0 = f,$$

если $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $A = A^* > 0$, $B = B^* > 0$, будет верна оценка

$$\|u^{k+1} - u\|_M \leqslant q_{k+1}\|u^0 - u\|_M, \tag{37}$$

где $M = A$ или $M = B$.

Оценка (37) следует из (35), если неявную схему трактовать как явную схему (28) для уравнения

$$C\omega = \varphi,$$

при $C = A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}$; $\varphi = A^{\frac{1}{2}}B^{-1}f$, $\omega = A^{\frac{1}{2}}u$,

или $C = B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$; $\varphi = B^{-\frac{1}{2}}f$, $\omega = B^{\frac{1}{2}}u$.

Таким образом, априорная информация вида (25) об операторе A позволяет построить двухшаговые итерационные процессы вида (26) или (28) с набором параметров вида (29), (35) и эффективно оценить быстроту их сходимости соответственно по формулам (24), (30), (35). Однако при наличии априорной информации об операторе A вида (25) можно также построить одношаговые итерационные процессы вида (21), § 4; (4), (44), § 4; (17), § 5. Если сравнить итерационные процессы (26) и (4), (44), § 4, то приближения u^{k+1} у этих методов совпадают, однако метод (26) обладает устойчивым счетом и если в (4), (44), § 4, ошибка оптимально гасится на $k+1$ итерации, то в (26) это происходит для каждого i.

Сравнивая трехслойные схемы (28) (при наборе параметров по формулам (29) или (34)) с двухслойными схемами (4), (44), § 4 (с устойчивым чебышевским набором параметров), легко убедиться, что последние сходятся быстрее.

В работе [55] показано, что двухслойная схема слабее зависит от возмущения величины γ_1 по сравнению с трехслойной схемой.

Если спектральные свойства оператора A не известны, но сохранено предположение о положительной определенности оператора A, то решение задачи об определении коэффициентов a_k, d_k так, чтобы получить оптимальную сходимость на каждом шаге, дают методы минимальных итераций или сопряженных градиентов. Коэффициенты a_k, d_k в этом случае определяются из условий ортогональности многочленов

$P_k(t)$ с некоторым распределением вида

$$\int P_k(t) P_m(t)\, d\sigma = \begin{cases} 0, & k \neq m \\ 1, & k = m \end{cases},$$

$$d\sigma = t\,(\varepsilon^0(t))^2 \rho(t)\, dt,$$

где $\rho(t)$ — положительная функция на некотором отрезке $[a, b]$ и равная нулю вне этого отрезка. Выбирая различным образом $\rho(t)\, dt$, можно построить различные итерационные процессы метода сопряженных градиентов.

§ 8. ИТЕРАЦИОННЫЕ МЕТОДЫ ДВУХСТОРОННИХ ПРИБЛИЖЕНИЙ

В итерационных методах двухсторонних приближений строится алгоритм, который заключает истинное решение в вилку, а значит, позволяет контролировать точность полученного приближенного решения. Однако на каждом шаге такие итерационные методы требуют удвоенного по порядку числа действий и удвоенной памяти по сравнению с одношаговыми итерационными методами.

Рассмотрим в гильбертовом пространстве **H** уравнение

$$Au = f. \tag{1}$$

Пусть оператор

$$K = I - HA \tag{2}$$

является самосопряженным

$$\operatorname{Sp} K \in [m, M], \quad 0 < m \leqslant M < 1. \tag{3}$$

На основе линейного одношагового явного итерационного процесса с оператором перехода

$$K_{\sigma_k} = \frac{K - \sigma_k I}{1 - \sigma_k} = (1 - \sigma_k)^{-1} [(1 - \sigma_k I) - HA] = I - \frac{HA}{1 - \sigma_k}, \tag{4}$$

где σ_k — некоторый скаляр ($0 \leqslant \sigma_k < 1$), можно построить двухсторонний итерационный процесс. Для этого оператор K_{σ_k} представим в виде

$$K_{\sigma_k} = K_{1\sigma_k} - K_{2\sigma_k}, \tag{5}$$

где

$$K_{1\sigma_k} = \left(\frac{|K_{\sigma_k}| + K_{\sigma_k}}{2}\right), \quad K_{2\sigma_k} = \frac{|K_{\sigma_k}| - K_{\sigma_k}}{2}, \qquad |K_{\sigma_k}| = \sqrt{K_{\sigma_k}^2}. \tag{6}$$

Итерационный метод решений определим по формулам:

$$\begin{aligned} W^k &= K_{1\sigma_k} W^{k-1} - K_{2\sigma_k} V^{k-1} + \frac{Hf}{1 - \sigma_k}, \\ V^k &= -K_{2\sigma_k} W^{k-1} + K_{1\sigma_k} V^{k-1} + \frac{Hf}{1 - \sigma_k}. \end{aligned} \tag{7}$$

Покажем, что (7) будет давать двухсторонний метод последовательных приближений, если W^0 и V^0 выбрать так, чтобы

$$V^0 \leqslant u \leqslant W^0. \tag{8}$$

Для этого полуупорядочим элементы W, $V \in \mathbf{H}$ следующим образом:

$$W = \sum_i w_i \varphi_i, \quad V = \sum_i v_i \varphi_i,$$

где φ_i — полная система собственных элементов оператора K, $K\varphi_i = \nu_i \varphi_i$.

Будем говорить, что $W \geqslant V$, если для всех i $w_i \geqslant v_i$.

Введем в рассмотрение

$$\begin{aligned} \varepsilon^k &= u - W^k = \sum_i \varepsilon_i^k \varphi_i, \\ \eta^k &= u - V^k = \sum_i \eta_i^k \varphi_i. \end{aligned} \tag{9}$$

Из (7) имеем:

$$\begin{aligned} \varepsilon^k &= K_{1\sigma_k} \varepsilon^{k-1} - K_{2\sigma_k} \eta^{k-1}, \\ \eta^k &= -K_{1\sigma_k} \varepsilon^{k-1} + K_{2\sigma_k} \eta^{k-1}. \end{aligned}$$

Если $\varepsilon^{k-1} \leqslant 0$, $\eta^{k-1} \geqslant 0$, то

$$\begin{aligned} 0 \geqslant \varepsilon^k &\geqslant -|K_{\sigma_k}| \max\{-\varepsilon^{k-1}, \eta^{k-1}\}, \\ 0 \leqslant \eta^k &\leqslant |K_{\sigma_k}| \max\{-\varepsilon^{k-1}, \eta^{k-1}\}. \end{aligned} \tag{10}$$

Значит, если выбрать начальные приближения так, чтобы выполнялось неравенство (8), то для всех $k \geqslant 1$

$$V^k \leqslant u \leqslant W^k.$$

Остановимся на вопросе выбора σ_k. Для этого можно потребовать, чтобы сумма квадратов ошибок уменьшилась в наибольшее число раз. Из (7) имеем

$$W^k - V^k = |K_{\sigma_k}| (W^{k-1} - V^{k-1})$$

$$W^k + V^k - 2u = K_{\sigma_k} (W^{k-1} + V^{k-1} - 2u).$$

После n итераций, учитывая (9), будем иметь:

$$(\varepsilon_i^n)^2 + (\eta_i^n)^2 = \left(\prod_{k=1}^{n} \frac{\nu_i - \sigma_k}{1 - \sigma_k} \right)^2 [(\varepsilon_i^0)^2 + (\eta_i^0)^2].$$

Поэтому, если σ_l выбрать из множества чисел

$$\sigma_l = \frac{1}{2} (2 - m - M - (M - m) \cos \frac{(2l-1)\pi}{2n}, \quad l = \overline{1, n},$$

то сумма квадратов ошибок уменьшится в $T_n \left(\frac{2 - M - m}{M - m} \right)$ раз.

§ 9. МЕТОД ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ ОБРАТНОГО ОПЕРАТОРА

Этот итерационный процесс основан на построении последовательности операторов, которые приближаются к оператору A^{-1}.

Пусть оператор $H_0 \in \mathfrak{M}$ и такой, что норма оператора

$$K_0 = I - H_0 A \tag{1}$$

удовлетворяет условию

$$\|K_0\| = \|I - H_0 A\| = q < 1. \tag{2}$$

Тогда, если положить

$$u^0 = H_0 f, \tag{3}$$

то можно ожидать, что приближение

$$u^1 = u^0 + H_0 r^0, \tag{4}$$

где $r^0 = f - Au^0$, лучше аппроксимирует u, чем u^0,

$$u^1 = (2H_0 - H_0 A H_0) f = (I + K_0) H_0 f.$$

Обозначим

$$(I + K_0) H_0 = H_1, \tag{5}$$

тогда

$$u^1 = H_1 f. \tag{6}$$

Образуем последовательности:

$$\begin{aligned} K_1 &= I - H_1 A, & H_2 &= (I + K_1) H_1, \\ K_2 &= I - H_2 A, & H_3 &= (I + K_2) H_2, \\ &\cdots\cdots\cdots & &\cdots\cdots\cdots\cdots \\ K_k &= I - H_k A, & H_{k+1} &= (I + K_k) H_k. \end{aligned} \tag{7}$$

Тогда

$$K_{k+1} = I - H_{k+1} A = I - (I + K_k) H_k A = I - (I + K_k)(I - K_k) = K_k^2,$$

т. е.

$$K_k = K_0^{2^k}.$$

Очевидно, $K_k \to 0$, а значит, $H_k \to A^{-1}$:

$$\|H_k - A^{-1}\| = \|A^{-1}(I - K_k) - A^{-1}\| \leqslant \|A^{-1} K_k\| \leqslant \|K_0\| \frac{q^{2^k}}{1 - q}.$$

Метод последовательных приближений обратного оператора (7) дает итерационный метод решения уравнения $Au = f$:

$$\begin{gathered} u^0 = H_0 f, \quad u^{k+1} = (I + K_0^{2^k}) u^k \quad (k = 0, 1, \ldots), \\ K_0 = I - H_0 A, \end{gathered} \tag{8}$$

в котором оператор $K_0^{2^k}$ образуется путем умножения оператора $K_0^{2^k - 1}$ на себя,

$$\|\varepsilon^{k+1}\| = \|(I + K_0^{2^k}) \varepsilon^k\| = \|(I + K_0^{2^k})(I + K_0^{2^k - 1}) \ldots (I + K_0\| \|\varepsilon^0\| \leqslant q^{2^k} \|\varepsilon^0\|.$$

В частности, если $A > 0$, то всегда можно, не уменьшая общности, считать, что $0 < A < I$ и в итерационном процессе (8) положить $H_0 = I$. Средняя скорость сходимости итерационного процесса (8) равна $+\infty$. Для уменьшения начальной ошибки в $\frac{1}{\delta}$ раз требуется

$$k \approx \log_2 \frac{\log_2 \delta}{\log_2 q} \tag{9}$$

итераций.

Для построения итерационного процесса можно воспользоваться также следующими соотношениями:

$$u^0 = H_0 f, \quad r^0 = f - Au^0 = Kf,$$
$$W^{k+1} = H_0 r^k, \quad r^{k+1} = r^k - AW^{k+1} = Kr^k = K^{k+1} f$$

и вычислительная схема будет иметь вид

$$\begin{gathered} u^k = u^0 + W^1 + \cdots + W^k, \\ W^{k+1} = H_0 r^k, \quad r^k = K^k f, \qquad k = 0, 1, 2, \ldots \end{gathered} \tag{10}$$

Схема (10) состоит в построении серии добавок к приближенному значению u^0 и сходится со скоростью геометрической прогрессии.

§ 10. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

1. Метод Ньютона—Канторовича и некоторые его модификации

Одним из наиболее известных итерационных процессов, применяемых при построении приближенного решения нелинейного операторного уравнения

$$P(u) = 0, \tag{1}$$

является метод Ньютона. Этот метод отличается от метода последовательных приближений более быстрой сходимостью.

В дальнейшем будем предполагать, что оператор P отображает банахово пространство $\mathbf{E}$ в банахово пространство $\mathbf{F}$, причем отображение P сильно дифференцируемо (дифференцируемо по Фреше) в некотором шаре $S(u^0, r)$ радиуса r с центром в точке u^0 и производная $P'(u)$ удовлетворяет условию Липшица с const L, т. е.

$$\|P'(\tilde{u}) - P'(\tilde{\tilde{u}})\| \leqslant L\|\tilde{u} - \tilde{\tilde{u}}\|, \quad \forall \tilde{u}, \tilde{\tilde{u}} \in S(u^0, r). \tag{2}$$

Примем центр шара за нулевое приближение к искомому решению и рассмотрим выражение

$$P(u^0) - P(u). \tag{3}$$

Заменив (3) его главной линейной частью, т. е. близким ему выражением

$$P(u^0) - P(u) \approx P'(u^0)(u^0 - u), \tag{4}$$

получим относительно u линейное уравнение

$$P(u^0) - P'(u^0)(u^0 - u) = 0. \tag{5}$$

Решение уравнения (5) можно рассматривать как приближение к решению $P(u) = 0$. Тогда

$$u^1 = u^0 - (P'(u^0))^{-1} P(u^0). \tag{6}$$

Итерационный процесс, в котором последовательные приближения к решению операторного уравнения (1) определяются по рекуррентной формуле

$$u^{k+1} = u^k - (P'(u^k))^{-1} P(u^k), \quad k = 0, 1, 2, \ldots, \tag{7}$$

получил название итерационного процесса Ньютона (или Ньютона — Канторовича).

Вычислительная схема итерационного процесса Ньютона записывается обычно следующим образом:

$$P'(u^k) z^k = P(u^k), \tag{8}$$

$$u^{k+1} = u^k - z^k, \quad k = 0, 1, 2, \ldots \tag{9}$$

Отметим некоторые модификации итерационного процесса (7):

$$u^{k+1} = u^k - (P'(u^0))^{-1} P(u^k), \quad k = 0, 1, 2, \ldots, \tag{10}$$

$$u^{k+1} = u^k - (P'(u^k) + \beta_k I)^{-1} P(u^k), \quad k = 0, 1, 2, \ldots, \tag{11}$$

$$u^{k+1} = u^k - \gamma_k (P'(u^k))^{-1} P(u^k), \quad k = 0, 1, 2, \ldots \tag{12}$$

В модифицированном процессе (10) обратный оператор $(P'(u^0))^{-1}$ на каждом шаге вычисляется при одном и том же значении аргумента. Такая модификация метода Ньютона уменьшает скорость сходимости, но оказывается целесообразной с вычислительной точки зрения, так как вычисление на каждом шаге $(P'(u^k))^{-1}$ является сложной задачей.

В итерационном процессе (11) β_k подбирается таким образом, чтобы оператор $(P'(u^k) + \beta_k I)^{-1}$ на каждом шаге был невырожденным; в (12) параметр γ_k выбирается из условия ускорения сходимости итерационного процесса.

При решении систем нелинейное уравнение $P(\boldsymbol{u}) = 0$ представляет собой n-мерный вектор-функцию

$$P(\boldsymbol{u}) = (P_i(\boldsymbol{u}))_{i=\overline{1,n}}, \quad \boldsymbol{u} = (u_i)_{i=\overline{1,n}},$$

под $P'(\boldsymbol{u})$ следует понимать матрицу Якоби системы функций $P_1(\boldsymbol{u})$, $P_2(\boldsymbol{u})$, ..., $P_n(\boldsymbol{u})$ относительно переменных $u_1, u_2, \ldots, u_n$ — компонент n-мерного вектора $\boldsymbol{u}$. При решении систем нелинейных уравнений, в которых n-мерный вектор-функция $P(\boldsymbol{u})$ сложно зависит от u_i ($i = \overline{1, n}$) или не задан в явном виде, а только указан алгоритм, позволяющий вычислять значение вектор-функции $P(\boldsymbol{u})$ при заданном значении аргумента, матрица Якоби в вычислительной схеме (8)

заменяется разностным аналогом $R(u, h)$:

$$P'(u^k) \approx R(u^k, h) =$$

$$= \begin{pmatrix} \frac{P_1(u^k + he_1) - P_1(u^k)}{h} & \frac{P_1(u^k + he_2) - P_1(u^k)}{h} & \dots & \frac{P_1(u^k + he_n) - P_1(u^k)}{h} \\ \frac{P_2(u^k + he_1) - P_2(u^k)}{h} & \frac{P_2(u^k + he_2) - P_2(u^k)}{h} & \dots & \frac{P_2(u^k + he_n) - P_2(u^k)}{h} \\ \dots & \dots & \dots & \dots \\ \frac{P_n(u^k + he_1) - P_n(u^k)}{h} & \frac{P_n(u^k + he_2) - P_n(u^k)}{h} & \dots & \frac{P_n(u^k + he_n) - P_n(u^k)}{h} \end{pmatrix}.$$

Здесь $0 < h \leqslant h_0$ — параметр шага, e_k — k-й единичный орт.

Вычислительная схема дискретной модификации итерационного процесса Ньютона будет иметь следующий вид:

$$\begin{gathered} R(u^k, h)\, z^k = P(u^k), \\ u^{k+1} = u^k - z^k, \quad k = 0, 1, 2, \dots \end{gathered} \tag{13}$$

Отметим, что метод Ньютона и его модификации совпадают с обычным методом последовательных приближений, примененным к уравнению $u = Q(u)$, где

$$Q(u) = u - H(u)\, P(u), \tag{14}$$

$$H(u) = \begin{cases} (P'(u))^{-1} & \text{в методе Ньютона,} \\ (P'(u^0))^{-1} & \text{в модифицированном методе Ньютона (10),} \\ (P'(u) + \beta I)^{-1} & \text{в (11),} \\ \gamma (P'(u))^{-1} & \text{в (12).} \end{cases}$$

Поэтому результаты, полученные при исследовании сходимости обычного метода последовательных приближений, могут быть использованы при исследовании сходимости метода Ньютона.

Теорема 1 (о сходимости метода Ньютона). *Пусть оператор $P(u)$ определен в шаре $S(u^0, r)$, сильно дифференцируемый в нем и $P'(u)$ удовлетворяет условию Липшица с константой L, т. е.*

$$\| P'(\tilde{u}) - P'(\tilde{\tilde{u}}) \| \leqslant L \| \tilde{u} - \tilde{\tilde{u}} \|, \quad \forall \tilde{u}, \tilde{\tilde{u}} \in S(u^0, r).$$

Пусть, кроме того, оператор $P'(u^0)$ имеет обратный $\Gamma_0 = (P'(u^0))^{-1}$ и известна оценка его нормы

$$\| \Gamma_0 \| \leqslant M,$$

а на начальном элементе u^0 выполняется неравенство

$$\| \Gamma_0 P(u^0) \| \leqslant N. \tag{15}$$

Тогда, если

$$\begin{gathered} \eta = 2LMN < \frac{1}{2}, \\ r \leqslant x_0 N, \end{gathered} \tag{16}$$

где x_0 — меньший корень уравнения

$$\eta x^2 - 2x + 2 = 0 \qquad \left(x_0 = \frac{1-\sqrt{1-2\eta}}{\eta}\right),$$

то уравнение (1) имеет в сфере $S(u^0, r)$ единственное решение u, и последовательность $\{u^k\}$, начатая с u^0 и определяемая рекуррентной формулой (7), сходится к этому решению. Быстрота сходимости итерационного процесса Ньютона (7) характеризуется неравенством

$$\|u^k - u\| \leqslant \frac{1}{2^{k-1}} (2\eta)^{2^k-1} N. \tag{17}$$

Доказательство. Рассмотрим отображение

$$\Phi(u) = u - (P'(u^0))^{-1} P(u) = u - \Gamma_0 P(u) \tag{18}$$

и покажем, что оператор $\Phi(u)$ отображает сферу $S(u^0, r)$ саму в себя. Действительно,

$$\begin{gathered}\Phi(u) - u^0 = u - u^0 - \Gamma_0 P(u) = \\ = \Gamma_0 (P'(u^0)(u - u^0) - P(u) + P(u^0)) - \Gamma_0 P(u^0), \\ \|\Phi(u) - u^0\| \leqslant \|\Gamma_0\| \|P'(u^0)(u - u^0) - P(u) + P(u^0)\| + \\ + \|\Gamma_0 P(u^0)\| \leqslant M \|P'(u^0)(u - u^0) - P(u) + P(u^0)\| + N.\end{gathered} \tag{19}$$

Отображение

$$\Psi(u) = P(u) - P(u^0) - P'(u^0)(u - u^0) \tag{20}$$

по предположению дифференцируемо, т. е.

$$\Psi'(u) = P'(u) - P'(u^0), \tag{21}$$

$$\|\Psi'(u)\| = \|P'(u) - P'(u^0)\| \leqslant L \|u - u^0\| \leqslant LNx_0. \tag{22}$$

Из (20) имеем:

$$\|\Psi(u)\| = \|\Psi(u) - \Psi(u^0)\| \leqslant LNx_0 \|u - u^0\| \leqslant LN^2 x_0^2 \tag{23}$$

и для $\|\Phi(u) - u^0\|$ будет иметь место следующая оценка:

$$\|\Phi(u) - u^0\| \leqslant MLx_0^2 N^2 + N = N\left(\frac{\eta}{2} x_0^2 + 1\right) = Nx_0,$$

т. е. оператор $\Phi(u)$ осуществляет отображение сферы $S(u^0, Nx_0)$ саму в себя. Более того, оператор $\Phi(u)$ является оператором сжатия. Действительно,

$$\begin{gathered}\Phi'(u) = I - \Gamma_0 P'(u) = \Gamma_0 (P'(u^0) - P'(u)), \\ \|\Phi'(u)\| \leqslant M \|P'(u^0) - P'(u)\| \leqslant LM \|u - u^0\| \leqslant MLNx_0 = \\ = \frac{\eta}{2} x_0 = \frac{\eta}{2} \frac{1-\sqrt{1-2\eta}}{\eta} = q < \frac{1}{2}.\end{gathered} \tag{24}$$

Поэтому

$$\|\Phi(u^1) - \Phi(u^2)\| \leqslant \frac{1}{2} \|u^1 - u^2\|. \tag{25}$$

На основании принципа сжатых отображений (теорема 1, гл. 8, § 1) $\Phi(u)$ имеет на сфере $S(u^0, r)$ одну неподвижную точку u такую, что

$$u = \Phi(u),$$

т. е.

$$P(u) = 0.$$

Для доказательства сходимости итерационного процесса (7) покажем, что последовательные приближения u^k будут такими, что

$$\begin{gathered}\|(P'(u^k))^{-1}\| \leqslant M_k, \\ \|(P'(u^k))^{-1} P(u^k)\| \leqslant N_k, \\ M_k = 2M_{k-1}, \quad M_0 = M, \\ N_k = \left(\frac{1}{2}\right)^k (2\eta)^{2^k-1} N, \\ \eta_k = 2M_k N_k L < \frac{1}{2}, \quad k = 1, 2, \ldots\end{gathered} \tag{26}$$

Из (24) на основании теоремы 14 (приложение, § 1) следует, что оператор

$$I - (I - \Gamma_0 P'(u^1)) = \Gamma_0 P'(u^1)$$

имеет линейный обратный, причем

$$\|(\Gamma_0 P'(u^1))^{-1}\| \leqslant \frac{1}{1-q} < 2,$$

значит,

$$\|(P'(u^1))^{-1}\| \leqslant \|(\Gamma_0 P'(u^1))^{-1}\| \|\Gamma_0\| < 2M. \tag{27}$$

Далее, учитывая (5), (20) и (15), имеем:

$$\begin{gathered}\|P(u^1)\| = \|P(u^1) - P(u^0) - P'(u^0)(u^1 - u^0)\| \leqslant \\ \leqslant \|\Psi(u^1) - \Psi(u^0)\| \leqslant L\|u^1 - u^0\|^2 \leqslant LN^2.\end{gathered} \tag{28}$$

Из (7), (27) и (28) получаем следующее неравенство:

$$\|u^2 - u^1\| = \|(P'(u^1))^{-1} P(u^1)\| \leqslant 2MLN^2 = \frac{1}{2} 2\eta N = N_1 < N,$$

$$\eta_1 = 2M_1 N_1 L = 2 \cdot 2M\eta NL = 2\eta^2 < \frac{1}{2}.$$

Таким образом, условия (15), (16) теоремы 1 будут выполняться для сферы $S(u^1, r)$, вложенной в сферу $S(u^0, r)$. Повторяя аналогичные рассуждения, получим:

$$\|(P'(u^k))^{-1}\| \leqslant M_k,$$

$$\|u^{k+1} - u^k\| = \|(P'(u^k))^{-1} P(u^k)\| \leqslant N_k, \quad k = 1, 2, \ldots,$$

где постоянные M_k, N_k связаны рекуррентными соотношениями (26).

При $q > 1$

$$\|u^{k+q} - u^k\| \leqslant \sum_{l=k}^{k+q-1} \|u^{l+1} - u^l\| \leqslant \sum_{l=k}^{k+q-1} \left(\frac{1}{2}\right)^l (2\eta)^{2^l-1} N =$$

$$= \left(\frac{1}{2}\right)^k (2\eta)^{2^k-1} N \left[1 + \frac{1}{2}(2\eta)^{2^k} + \cdots + \left(\frac{1}{2}\right)^{q-1} (2\eta)^{2^k(2^{q-1}-1)}\right].$$

Так как $2\eta < 1$, то

$$\|u^{k+q} - u^k\| \leqslant \left(\frac{1}{2}\right)^k (2\eta)^{2^k-1} N \left(1 + \frac{1}{2} + \dots + \left(\frac{1}{2}\right)^{q-1}\right) \leqslant$$

$$\leqslant \left(\frac{1}{2}\right)^{k-1} (2\eta)^{2^k-1} N.$$

При $q \to \infty$ имеем:

$$\|u - u^k\| \leqslant \left(\frac{1}{2}\right)^{k-1} (2\eta)^{2^k-1} N.$$

В частности,

$$\|u - u^0\| \leqslant 2N.$$

Из неравенства (24) получаем следующую оценку быстроты сходимости модифицированного метода Ньютона вида (10)

$$\|u^k - u\| \leqslant \frac{q^k}{1-q} \|(P'(u^0))^{-1} P(u^0)\|.$$

Модифицированный метод Ньютона (10) сходится со скоростью геометрической прогрессии.

Метод Ньютона в случае решения алгебраического или трансцендентного уравнения

$$f(u) = 0 \tag{29}$$

имеет вид

$$u^{k+1} = u^k - \frac{f(u^k)}{f'(u^k)}, \qquad k = 0, 1, 2, \dots \tag{30}$$

Геометрически формула (30) эквивалентна определению абсциссы точки пересечения с осью u касательной, проведенной к кривой $y = f(u)$ в точке $(u^k, f(u^k))$.

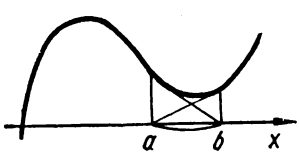


Рис. 12

Последовательность (30) даже для одного скалярного уравнения не всегда сходится (рис. 12).

Проблема выбора начального приближения, обеспечивающего сходимость итерационного процесса, имеет существенное значение для применения метода Ньютона. Способ выбора начального приближения в случае решения скалярного уравнения (29) методом (30), если выделен отрезок $[a, b]$, содержащий единственный корень уравнения, указывает следующая теорема.

Теорема 2. *Пусть $f(a) f(b) < 0$, причем $f(x)$ и $f''(x)$ отличны от нуля и сохраняют определенные знаки при $x \in [a, b]$. Тогда исходя из $u^0 \in [a, b]$ и удовлетворяющего неравенству*

$$f(u^0) f''(u^0) > 0,$$

можно вычислить методом Ньютона единственный корень $u \in [a, b]$ уравнения (29) *с любой точностью.*

Доказательство. Пусть для определенности $f(a) < 0$, $f(b) > 0$ и $f'(u) > 0$, $f''(u) > 0$, $\forall u \in [a, b]$.
Положим $u^0 = b$, тогда $u^0 > u$, $f(u^0) > 0$ и $u^1 > u$.

Покажем, что $f(u^k) > 0$ и $u^{k+1} > u$.

Рассмотрим

$$f(u) = f(u^k) + f'(u^k)(u - u^k) + \frac{1}{2} f''(c)(u - u^k)^2, \text{ где } c \in (u, u^k).$$

Очевидно,

$$f(u^k) > f'(u^k)(u^k - u) > 0,$$

так как $f''(c) > 0$.

Следовательно,

$$u^{k+1} = u^k - \frac{f(u^k)}{f'(u^k)} > u.$$

Последовательность $\{u^k\}$ будет ограниченной монотонно убывающей:

$$u^{k+1} < u^k < u^{k-1} < \ldots,$$

а значит, существует

$$\lim_{k\to\infty} u^k = \tilde{u}.$$

Покажем, что $\tilde{u} = u$.

Переходя к пределу в (30), получим:

$$\tilde{u} = \tilde{u} - \frac{f(\tilde{u})}{f'(\tilde{u})},$$

т. е. $f(\tilde{u}) = 0$, или $\tilde{u} = u$.

Если два последовательные приближения u^k и u^{k+1}, полученные по методу Ньютона, совпадают с точностью до ε, то этот факт еще не гарантирует совпадения с той же точностью u^{k+1} с u (см. рис. 13).

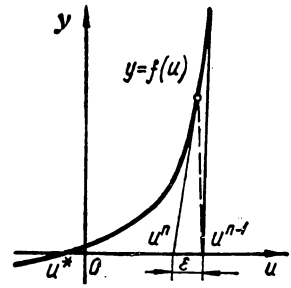


Рис. 13

Для оценки погрешности k-го приближения, найденного по методу Ньютона, можно воспользоваться следующим неравенством:

$$|u^k - u| \leqslant \frac{M_2}{2m}(u^k - u^{k-1})^2, \tag{31}$$

где

$$M_2 = \max_{u\in[a,b]} |f''(u)|, \quad m = \min_{u\in[a,b]} |f'(u)|.$$

В самом деле, применяя теорему Лагранжа, имеем:

$$f(u^k) = f(u) + (u^k - u) f'(\xi), \quad \xi \in (u^k, u)$$

или

$$|u^k - u| \leqslant \frac{|f(u^k)|}{m}. \tag{32}$$

По формуле Тейлора находим:

$$f(u^k) = f(u^{k-1}) + (u^k - u^{k-1}) f'(u^{k-1}) + \frac{1}{2} f''(\eta)(u^k - u^{k-1})^2,$$

$$\eta \in (u^{k-1}, u^k),$$

откуда

$$|f(u^k)| \leqslant \frac{1}{2} M_2 (u^k - u^{k-1})^2. \tag{33}$$

Подставляя (33) в (32), получим (31). Из (31) следует, что при $\frac{M_2}{2m} < 1$ из совпадения u^k, u^{k-1} с точностью ε следует совпадение u^k с u до ε^2, т. е. количество верных десятичных знаков как бы удваивается, если процесс Ньютона сходится.

2. Метод продолжения решения по параметру

Метод продолжения решения по параметру можно рассматривать как один из способов получения начальных приближений, достаточно близких к решению уравнения

$$P(u) = 0.$$

Здесь P — отображение, переводящее открытое множество Ω одного **B**-пространства **E** в другое **B**-пространство **F**.

Суть метода заключается в том, что вместо отображения $P(u)$ вводится в рассмотрение целое семейство отображений $W: \Omega \times [0, 1]$ таких, что

$$W(u, 0) = P_0(u), \quad \forall u \in \Omega, \tag{34}$$

$$W(u, 1) = P(u), \quad \forall u \in \Omega, \tag{35}$$

причем решение $u(0)$ уравнения

$$W(u, 0) = P_0(u) = 0 \tag{36}$$

известно или легко находится, а решение уравнения

$$W(u, 1) = 0 \tag{37}$$

требуется определить, и оно будет совпадать с искомым решением уравнения (1).

Существуют различные формальные приемы построения семейства $W(u, \lambda)$. Например:

а) $$W(u, \lambda) = P(u) + (\lambda - 1) P(\tilde{u}), \quad u, \tilde{u} \in \Omega, \quad \lambda \in [0, 1], \tag{38}$$

$\tilde{u}$ — фиксировано и выбирается таким образом, чтобы решение уравнения

$$P_0(u) = P(u) - P(\tilde{u}) = 0$$

было известно или легко находилось;

б) $$W(u, \lambda) = \lambda P(u) + (1 - \lambda) P_0(u), \quad u \in \Omega, \quad \lambda \in [0, 1]. \tag{39}$$

Здесь в нашем распоряжении находится выбор отображения $P_0(u)$. При его построении прежде всего должны быть соблюдены условия, накладываемые на соотношение (36).

Основная цель при построении уравнения

$$W(u, \lambda) = 0, \quad \lambda \in [0, 1] \tag{40}$$

заключается в получении решения u (1).

Способы получения решения уравнения (40) также могут быть различными. Один из методов получения решения уравнения (40) заключается в следующем: интервал $[0, 1]$ разбивается точками

$$0 = \lambda_0 < \lambda_1 < \cdots < \lambda_n = 1$$

и рассматриваются уравнения

$$W(u, \lambda_l) = 0, \quad l = \overline{1, n}. \tag{41}$$

Для решения уравнения (41) применяется какой-либо итерационный метод, причем за начальное приближение при решении l-го уравнения используется решение $(l - 1)$-го уравнения. Если величина $h_l = \lambda_{l+1} - \lambda_l$ мала, то можно ожидать, что $u(\lambda_{l-1})$ будет таким начальным приближением для $u(\lambda_l)$, при котором имеет место сходимость применяемого итерационного процесса.

Например, пусть $P(\boldsymbol{u}) = 0$ представляет собой систему нелинейных уравнений и $W(\boldsymbol{u}, \lambda)$ определяется при помощи соотношения (38).

Пусть для решения каждой l-й задачи (41) строится один шаг по методу Ньютона. Тогда вычислительная схема построенного приближенного метода будет иметь вид

$$\begin{gathered} u^{k+1} = u^k - (P'(u^k))^{-1}(P(u^k) + (\lambda_k - 1)P(\tilde{u})), \quad k = \overline{0, n-1}; \\ u^{k+1} = u^k - (P'(u^k))^{-1}P(u^k), \quad k = n, n+1, \ldots, \end{gathered} \tag{42}$$

где $P'(u^k)$ — матрица Якоби вектор-функции $P(u^k)$.

Ясно, что для решения l-й задачи (41) можно применить какой-либо другой итерационный метод или метод Ньютона с конечным числом шагов $m_i \geqslant 1$.

Если предположить, что параметр λ в уравнении

$$W(u, \lambda) = 0, \quad \forall \lambda \in [0, 1]$$

введен таким образом, что уравнение (40) удовлетворяет условиям (34)—(37), имеет решение $u(\lambda)$, непрерывно зависящее от параметра λ, и отображение $W(u, \lambda)$ имеет непрерывные частные производные по u и λ, то для решения уравнения (40) может быть предложен следующий способ, который для простоты изложения проиллюстрируем на примере решения системы нелинейных уравнений.

Продифференцируем (40) по λ. В результате получим:

$$W_u(u, \lambda)\frac{\partial u(\lambda)}{\partial \lambda} = -W_\lambda(u, \lambda), \tag{43}$$

где $W_u(u, \lambda)$ — матрица Якоби вектор-функции $W(u, \lambda)$. Система (43) удовлетворяет начальному условию $u(0)$:

$$W(u(0), 0) = 0. \tag{44}$$

Таким образом, задача отыскания $W(u, 1)$ сводится к решению задачи Коши (43), (44) на отрезке $[0, 1]$. Приближенное решение задачи Коши может быть найдено любым из рассмотренных в гл. 7 методов, но при этом возникают вопросы о существовании единственного решения $u(\lambda)$, $\lambda \in [0, 1]$ задачи Коши (43), (44) и выборе численного метода решения, обладающего устойчивым счетом. Эти вопросы исследовались рядом авторов для конкретных классов нелинейных операторных уравнений (см., например, [58], [82], [83]). Для иллюстрации численных алгоритмов такого подхода к решению уравнения (1) рассмотрим расчетную схему, построенную на основе использования метода Эйлера для решения задачи Коши (43), (44), если найденное значение $u(1)$ используется как начальное приближение в методе Ньютона.

Вычислительная схема итерационного процесса для определения решения уравнения (1) будет иметь вид

$$\begin{gathered} u^{k+1} = u^k - h_k W_u^{-1}(u^k, \lambda_k) W_\lambda(u^k, \lambda_k), \quad k = \overline{0, n-1}; \\ h_k = \lambda_{k+1} - \lambda_k, \quad u^0 = u(0), \end{gathered}$$

где $u(0)$ — решение уравнения

$$\begin{gathered} P_0(u) = 0; \\ u^{k+1} = u^k - (P'(u^k))^{-1}P(u^k), \quad k = n, n+1, \ldots \end{gathered} \tag{45}$$

В частности, если $W(u, \lambda)$ имеет вид (38), то вычислительная схема (45) эквивалентна следующей:

$$u^{k+1} = u^k - h_k (P'(u^k))^{-1} P(\tilde{u}),$$
$$u^0 = u(0), \quad k = \overline{0, n-1}, \tag{46}$$
$$u^{k+1} = u^k - (P'(u^k))^{-1} P(u^k), \quad k = n, n+1, \ldots$$

Здесь $u(0)$ — решение уравнения

$$P(u) - P(\tilde{u}) = 0,$$

$\tilde{u}$ — некоторое фиксированное значение, принадлежащее Ω.

3. Методы минимизации

Методы минимизации при решении нелинейных (а также линейных уравнений) вида

$$P(u) = 0$$

основаны на сведении задачи (1) к задаче минимизации функционала $G(P(u))$. С этой точки зрения проекционные методы (Ритца, Галеркина, наименьших квадратов, моментов, а также их некоторые видоизменения и обобщения) можно истолковывать как методы минимизации решения уравнения (1).

Проекционные методы имеют довольно широкую область применения. Однако нахождение достаточно точных приближений при помощи проекционных методов в случае нелинейных уравнений связано с необходимостью решения систем уравнений высокого порядка. Весьма сложным в проекционных методах является вопрос выбора последовательности базисных функций в подпространстве E_n, на которое проектируется пространство E, а также каким надо выбрать n, чтобы получить приближенное решение с требуемой точностью. Отметим также, что в проекционных методах приближение, полученное при $n = k$, обычно не используется для нахождения приближений при $n > k$.

Алгоритмы итерационных методов минимизации, как правило, обладают более простой вычислительной схемой. Однако вопрос выбора начального приближения и быстрота сходимости итерационных процессов минимизации значительно затрудняют их использование для практического решения задач.

Как в проекционных, так и в итерационных методах, основанных на идее минимизации, общий вид записи минимизирующего функционала не позволяет учесть специфических свойств конкретных алгоритмов и рассматриваемых операторов. Поэтому вопросы, связанные с построением алгоритмов и исследованием их сходимости, относятся, как правило, к конкретным видам функционалов $G(P(u))$ и операторов $P(u)$.

Алгоритмы, основанные на уменьшении функционала $G(P, (u))$ на каждом шаге итераций

$$G(P(u^{k+1})) \leqslant G(P(u^k)), \quad k = 0, 1, 2, \ldots, \tag{47}$$

получили название методов спуска.

Остановимся на одном из вариантов методов спуска, когда в качестве $G(P(u))$ выбран функционал вида

$$G(P(u)) = \|P(u)\|^2 \tag{48}$$

с эвклидовой нормой, а $P(u) = 0$ представляет собой систему нелинейных уравнений n-го порядка $u = (u_i)_{i=\overline{1,n}}$. Очевидно, (48) достигает минимального значения (равного нулю) при $u = u^*$, где u^* — решение уравнения

$$P(u) = 0. \tag{49}$$

Верно и обратное: любой нулевой минимум функционала (48) достигается в точке, являющейся решением уравнения (49).

Итерационный процесс вида (47) для функционала (48) строится следующим образом: задаются некоторой начальной точкой u^0 и двигаются от нее по направлению убывания функционала до тех пор, пока точка u^k не достигнет минимума $G(P(u))$. Если этот минимум окажется нулевым, то находим решение системы (49), если он окажется некоторой положительной величиной, то спуск следует начинать с другой начальной точки u^0. При этом обычно предполагается, что поверхность $G(P(u)) = \|P(u)\|^2$ замкнутая и ограничивает область Ω, во внутренних точках которой $G(P(u)) < G(P(u^0))$. Таким образом, в методах спуска выбор начального приближения u^0 также имеет очень большое значение. В качестве направления спуска выбирается направление градиента функционала $G(P(u^0))$, т. е. направление наибыстрейшего изменения функционала в точке u^0, причем положительное направление вектора $\operatorname{grad} G(P(u^0))$ соответствует возрастанию функции $G(P(u^0))$, а отрицательное — убыванию $G(P(u^0))$

$$\operatorname{grad} G(P(u)) = \left(\frac{\partial G}{\partial u_i}\right)_{i=\overline{1,n}}$$

и

$$\operatorname{grad} G(P(u^*)) = 0.$$

Дифференцируя (48), найдем:

$$\frac{\partial G}{\partial u_i} = \sum_{k=1}^{n} \frac{\partial P_k}{\partial u_i} P_k(u), \quad i = \overline{1, n} \tag{50}$$

или

$$\operatorname{grad} G(P(u)) = \Phi_P^T(u)\, P(u), \tag{51}$$

где $\Phi_P^T(u)$ — транспонированная матрица Якоби вектор-функции $P(u)$. Линия, по которой происходит движение точки u к минимуму $G(P(u))$, называют *линией спуска*. Построение линий спуска может осуществляться различными способами. Один из методов построения линии спуска заключается в следующем: через точку u^0 проводят прямую

$$u = u^0 - \gamma \Phi_P^T(u^0)\, P(u^0) \tag{52}$$

и находят ближайший к точке u^0 локальный минимум функции $G(P(u))$, рассматриваемый как функция переменного γ, т. е. находят:

$$\min \omega(\gamma) = \min G(u^0 + \gamma \Phi_P^T(u^0)\, P(u^0)). \tag{53}$$

Пусть γ_1 то значение, при котором достигается этот минимум. Тогда в качестве следующей начальной точки выбирается u^1:

$$u^1 = u^0 - \gamma_1 \Phi_P^T (u^0) P (u^0).$$

Аналогично строится точка

$$u^2 = u^1 - \gamma_2 \Phi_P^T (u^1) P (u^1)$$

и вообще

$$u^{k+1} = u^k - \gamma_{k+1} \Phi_P^T (u^k) P (u^k). \tag{54}$$

Точки u^k можно рассматривать как узловые точки ломаной спуска. Движение происходит до тех пор, пока точка не достигнет минимума $G (P (u))$.

Для определения γ_k $(k = 1, 2, \ldots)$ в итерационном процессе (54) на каждом шаге нужно решить задачу вида (53), т. е. нужно строить какой-либо алгоритм минимизации функции одной переменной. Следует отметить, что эту задачу не обязательно решать до конца. Иногда достаточно достигнуть некоторого смещения в сторону минимума функции $\omega (\gamma)$ и перейти к следующему шагу итерационного процесса (54). При этом существенное значение имеет вопрос трудоемкости вычислительного алгоритма для решения задачи (53) по отношению к алгоритму (54). Для решения задачи (53) могут быть использованы различные приближенные алгоритмы. В частности, можно использовать следующий подход: значение γ, доставляющее минимум функции $\omega (\gamma)$, находится среди решений уравнения

$$\omega' (\gamma) = 0. \tag{55}$$

Для решения уравнения (55) может быть применен итерационный процесс Ньютона (с начальным значением $\gamma^0 = 0$) или метод последовательных приближений. После определения значения γ_k по этим алгоритмам нужно проверить, является ли это значение точкой минимума для функции $\omega (\gamma)$.

Построение линий спуска иногда осуществляют по некоторой заранее выбранной совокупности направлений. Это так называемый *метод спуска по направлениям*. Чаще всего в качестве таких направлений выбирают координатные орты e_i $(i = \overline{1, n})$. В этом случае итерационный процесс называют методом покоординатного спуска.

Исходя из точки u^0 и двигаясь вдоль координатного орта e_1, выбирают коэффициент γ_1 так, чтобы величина $\tilde{\omega} (\gamma) = G (u^0 - \gamma_1 e_1)$ достигала локального минимума. Тогда точку

$$u^1 = u^0 - \gamma_1 e_1$$

принимают за узловую точку первого звена ломаной спуска. Второе звено строится аналогично, но в качестве начальной точки выбирается u^1 и спуск проводится по направлению e_2.

Вычисление n звеньев ломаной спуска эквивалентно одному циклу итерационного процесса. Затем проводят вычисления второго цикла, причем порядок направлений второго цикла может не совпадать с порядком направлений первого цикла. Если из проведенных ранее

вычислений видно, что спуск вдоль каких-то координатных линий обеспечивает наибольшее убывание функционала $\tilde{\omega}(\gamma)$, то считается целесообразным более частый спуск по направлению этих координат. Иногда номер очередной координаты, по которой осуществляется спуск, выбирается недетерминированно, т. е. строят случайный покоординатный спуск.

Практически итерационный процесс продолжается до тех пор, пока с заданной точностью не совпадут минимумы функционала на каждом из направлений последнего цикла.

Однако методы спуска позволяют находить стационарные точки функционала $G(P(u))$, которые вообще не обязательно являются решениями системы $P(u) = 0$. Для отыскания решений системы (49) обычно проводят многократную реализацию одного из методов спуска при различных значениях начального приближения u^0. Выбор начального значения u^0 производят либо упорядоченно, путем покрытия области Ω, в которой разыскивается решение, достаточно плотной сеткой, либо выбирая начальные значения случайно.

Основные результаты, касающиеся сходимости рассмотренных в этом параграфе итерационных процессов, можно найти, например, в [58].

ПРИЛОЖЕНИЕ

Для изучения курса «Методы вычислений» читатель должен ознакомиться с отдельными разделами из функционального анализа, теории специальных функций, дискретного анализа и т. д.

Некоторые сведения из этих областей математики, которые, по мнению авторов, являются наиболее важными при изучении вопросов, изложенных в предлагаемом учебном пособии, помещены в данном приложении. Авторы считают, что помещенный в пособии вспомогательный материал может быть использован как справочный. Для более полного ознакомления с затронутыми в пособии вопросами рекомендуется соответствующая литература.

§ 1. НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ ФУНКЦИОНАЛЬНОГО АНАЛИЗА

Большинство результатов, помещенных в настоящем параграфе, приводятся без доказательств, так как согласно существующим учебным программам доказательства приведенных здесь результатов уже известны студентам из курса функционального анализа. С этими результатами можно подробно ознакомиться, например, по следующим учебникам: [13], [34], [43].

1. Линейные метрические пространства

Множество **L** элементов x, y, z,... произвольной природы называется *линейным* над полем комплексных чисел C, если в нем определены операции сложения и умножения на числа, не выводящие за пределы **L** и удовлетворяющие условиям:

1) $(x + y) + z = x + (y + z)$ (ассоциативность сложения);
2) $x + 0 = 0 + x = x$ (существование нулевого элемента);
3) $x + y = y + x$ (коммутативность сложения);
4) для всякого x существует элемент $-x$ такой, что $x + (-x) = 0$ (существование противоположного элемента);
5) $(\alpha + \beta)\, x = \alpha x + \beta x$ (первый закон дистрибутивности умножения);
6) $\alpha\,(x + y) = \alpha x + \alpha y$ (второй закон дистрибутивности умножения);
7) $\alpha\,(\beta x) = (\alpha\beta)\; x$ (ассоциативность умножения);
8) $1 \cdot x = x$;
9) если $\alpha x = 0$ и $x \neq 0$, то $\alpha = 0$.

Линейные многообразия. Непустое множество $\mathbf{L}_1$ элементов линейного множества **L** называется *линейным многообразием,* если вместе с элементами $x_1, x_2, \ldots, x_n$ множество содержит любую линейную комбинацию

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

этих элементов, где $a_i \in C$, $i = \overline{1, n}$.

Линейное многообразие $\mathbf{L}_1$ называется *максимальным,* если оно не совпадает со всем линейным множеством **L** и не содержится ни в каком другом линейном многообразии, кроме **L**.

Пусть $\mathbf{L}$ — линейное множество и $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_m$ — принадлежащие ему линейные многообразия. Если каждый элемент $x \in \mathbf{L}$ однозначно представим в виде

$$x = \sum_{i=1}^{m} x_i, \quad x_i \in \mathbf{L}_i, \quad i = \overline{1, m}, \tag{1}$$

то говорят, что линейное множество $\mathbf{L}$ есть прямая сумма линейных многообразий $\mathbf{L}_i$ $(i = \overline{1, n})$:

$$\mathbf{L} = \mathbf{L}_1 \oplus \mathbf{L}_2 \oplus \cdots \oplus \mathbf{L}_m. \tag{2}$$

Выпуклые множества. Под *отрезком,* определяемым элементами x и y линейного множества $\mathbf{L}$, понимается совокупность элементов вида $\alpha x + (1 - \alpha) y$, где $0 \leqslant \alpha \leqslant 1$. Множество $\mathbf{S} \subset \mathbf{L}$ называется *выпуклым,* если оно полностью содержит отрезок, соединяющий два любых его элемента.

Для произвольного множества $\mathbf{S} \subset \mathbf{L}$ существует наименьшее выпуклое множество $\mathbf{S}^0$, содержащее $\mathbf{S}$, которое называется выпуклой *оболочкой* множества $\mathbf{S}$. Выпуклая оболочка $\mathbf{S}^0$ состоит из всевозможных элементов вида

$$x = \sum_{i=1}^{n} \alpha_k x_k,$$

где $\alpha_k \geqslant 0$, $\sum_{k=1}^{n} \alpha_k = 1$, $x_k \in \mathbf{S}$, n — любое натуральное число.

Множество $\mathbf{M}$ называется *метрическим пространством,* если каждой паре его элементов x, y поставлено в соответствие действительное число $\rho(x, y)$ (расстояние между элементами x и y), удовлетворяющее условиям (аксиомам):

1) $\rho(x, y) \geqslant 0$, $\rho(x, y) = 0 \Leftrightarrow x = y$;

2) $\rho(x, y) = \rho(y, x)$ (аксиома симметрии);

3) $\rho(x, y) \leqslant \rho(x, z) + \rho(z, y)$, $\forall z \in \mathbf{M}$ (неравенство треугольника).

Если $\{x_n\} \in \mathbf{M}$, $x \in \mathbf{M}$ и $\rho(x_n, x) \to 0$ при $n \to \infty$, то говорят, что $\{x_n\}$ сходится к x:

$$\lim_{n \to \infty} x_n = x \qquad (x_n \to x). \tag{3}$$

Линейное множество $\mathbf{L}$ называется *линейным метрическим* пространством, если на множестве $\mathbf{L}$ введена метрика, причем так, что алгебраические операции непрерывны в метрике множества $\mathbf{L}$, т. е.

1) Из того что $x_n \to x$, $y_n \to y$, следует

$$x_n + y_n \to x + y.$$

2) Если $x_n \to x$, $\alpha_n \to \alpha$, то $\alpha_n x_n \to \alpha x$.

Последовательность точек $\{x_n\}$ метрического пространства $\mathbf{M}$ называется *фундаментальной,* если для любого $\varepsilon > 0$ найдется натуральное число $N(\varepsilon)$ такое, что $\rho(x_m, x_i) < \varepsilon$ при $m, i \geqslant N(\varepsilon)$. Всякая сходящаяся последовательность фундаментальна, но обратное утверждение не имеет места. Например, если в метрическом пространстве $\mathbf{M}$, состоящем из рациональных чисел с метрикой $\rho(x, y) = |x - y|$, последовательность $\{x_n\}$ сходится к некоторому иррациональному числу, то она будет фундаментальной в $\mathbf{M}$, но в $\mathbf{M}$ не существует элемента, который бы явился ее пределом.

Если в метрическом пространстве $\mathbf{M}$ каждая фундаментальная последовательность сходится к некоторому элементу того же пространства, то $\mathbf{M}$ называется *полным пространством.*

Линейное множество $\mathbf{R}$ называется *нормированным пространством,* если каждому элементу $x \in \mathbf{R}$ поставлено в соответствие вещественное число $\|x\| > 0$ (называемое нормой элемента x), удовлетворяющее аксиомам:

1) $\|x\| \geqslant 0$ и $\|x\| = 0 \Leftrightarrow x = 0$;

2) $\|x + y\| \leqslant \|x\| + \|y\|$ (неравенство треугольника);

3) $\|\alpha x\| = |\alpha| \|x\|$ $\forall \alpha \in \mathbf{C}$ (однородность нормы).

Пусть L_n — линейное множество, состоящее из всевозможных n-мерных векторов $x = (\xi_1, \xi_2, \ldots, \xi_n)$. В L_n можно ввести норму по формуле

$$\|x\| = \left\{\sum_{i=1}^{n} \xi_i^2\right\}^{\frac{1}{2}}. \tag{4}$$

Линейное множество L_n с такой нормой называется *евклидовым пространством* R_n.

Теорема 1 (Каратеодори). Если $p \in A^0$ и A — подмножество в R_n, то найдутся такие элементы $p_1, p_2, \ldots, p_k \in A$, $k \leqslant n+1$, что $p = \sum_{i=1}^{k} \rho_i p_i$, $\rho_i \geqslant 0$, $\sum_{i=1}^{k} \rho_i = 1$, т. е. $p \in \{p_1, p_2, \ldots, p_k\}^0$.

Последовательность $\{x_n\}$ элементов нормированного пространства R называется *сходящейся* к элементу x_0, если

$$\|x_0 - x_n\| \to 0 \quad \text{при } n \to \infty.$$

Элемент $x_0 \in R$ называется *предельной точкой* множества $T \subset R$, если всякая окрестность x_0 содержит бесконечно много элементов из множества T. Для того чтобы элемент x_0 был предельной точкой множества T, необходимо и достаточно, чтобы последовательность $\{x_n\} \in T$ сходилась к $x_0 \in R$.

Множество $\overline{T} \subset R$ называется *замкнутым*, если все предельные точки T принадлежат T. Замкнутое линейное многообразие линейного нормированного пространства называют *линейным подпространством*. Замкнутое множество $\overline{T}$ состоит из точек трех типов: 1) изолированных точек множества T; 2) предельных точек множества T, принадлежащих T; 3) предельных точек множества T, не принадлежащих T.

Пусть T_1 и T_2 — два множества в метрическом пространстве M. Множество T_1 называется *плотным* в T_2, если $\overline{T}_1 \supset T_2$.

Метрические пространства, в которых имеется счетное всюду плотное множество, называют *сепарабельными*.

Теорема 2. Всякое множество сепарабельного пространства само является сепарабельным пространством.

Множество T метрического пространства M называется *компактным*, если из любой бесконечной последовательности $\{x_n\} \in T$ можно выделить подпоследовательность, сходящуюся к некоторому элементу $x \in M$.

Замкнутое множество T метрического пространства M называется *компактом*.

Теорема 3. Если A — компакт в R_n, то A^0 — также компакт в R_n.

Полное нормированное пространство называется *банаховым пространством*, или коротко, типа B.

Пусть R — линейное множество над C и пусть любым двум ее элементам x и y (в частности, может быть $x = y$) поставлено в соответствие число (x, y), обладающее следующими свойствами *(аксиомы скалярного произведения)*:

1) $(x, y) = \overline{(y, x)}$;

2) $(x + y, z) = (x, z) + (y, z)$;

3) $(\alpha x, y) = \alpha (x, y)$, $\forall \alpha \in C$;

4) $(x, x) \geqslant 0$, причем $(x, x) = 0 \Leftrightarrow x = 0$.

При выполнении условий 1) — 4) число (x, y) называется скалярным произведением.

Из аксиом скалярного произведения вытекают следствия:

$$\begin{gathered}(x, y + z) = (x, y) + (x, z);\\ (x, \beta y) = \bar{\beta}\,(x, y).\end{gathered} \tag{5}$$

Неравенство Коши — Буняковского

$$|(x, y)| \leqslant \sqrt{(x, x)} \cdot \sqrt{(y, y)}. \tag{6}$$

По скалярному произведению в **R** можно ввести норму

$$\|x\| = \sqrt{(x, x)}. \tag{7}$$

Гильбертовым пространством **H** называется совокупность элементов x, y, ... произвольной природы, удовлетворяющая следующим условиям:

1) **H** — линейное множество над полем C;

2) для всех $x, y \in$ **H** определено скалярное произведение (x, y), удовлетворяющее аксиомам 1)—4);

3) пространство **H** полно в смысле метрики

$$\rho(x, y) = \|x - y\| = \sqrt{(x - y, x - y)};$$

4) пространство **H** бесконечномерное, т. е. в нем для любого n можно найти n линейно-независимых элементов.

Если выполнены первые три условия, пространство **H** называется *унитарным*. Любое гильбертово пространство является банаховым.

В гильбертовых пространствах вводится понятие ортогональности ($\perp$). Элементы $x, y \in$ **H** называются ортогональными ($x \perp y$), если $(x, y) = 0$.

Из свойств скалярного произведения вытекает, что

1) $0 \perp x, \quad \forall\, x \in$ **H**,

2) $x \perp x \Leftrightarrow x = 0$,

3) если $x \perp y_1, y_2, \ldots, y_m$, то $x \perp \sum_{i=1}^{m} c_i y_i$, (8)

4) если множество **T** всюду плотно в **H** и $x \in$ **H** ортогонален каждому элементу из **T**, то $x = 0$.

Если **L** — подпространство пространства **H**, то совокупность всех элементов, ортогональных к **L**, образует подпространство, которое называется ортогональным дополнением к **L**, $\mathbf{H} = \mathbf{L} \oplus \mathbf{L}_1$.

Если **L** — подпространство пространства **H**, то для всякого $x \in$ **H** существует единственное представление

$$x = y + z,$$

где $y \in$ **L**, $z \perp \mathbf{L}_1$, (y — проекция x на **L**). Элемент y находится на наименьшем расстоянии от x по сравнению с другими элементами из **L**.

Для того чтобы линейное многообразие **L** было всюду плотно в пространстве **H**, необходимо и достаточно, чтобы в **H** не существовало элемента, отличного от нулевого и ортогонального всем элементам множества **L**.

2. Линейные операторы

Пусть **E** и **M** — два линейных нормированных пространства. Если каждому элементу $x \in \mathbf{D} \subset \mathbf{E}$ сопоставлен по определенному правилу элемент $y = Ax \in \mathbf{M}$, то говорят, что на множестве **D** или в **E** задан оператор A со значениями в **M**. Множество **D** называется *областью определения* оператора A и обозначается **D** (A). Совокупность всех элементов y вида $y = Ax$ ($x \in$ **D**) называется *областью значений* оператора A и обозначается **R** (A).

Оператор A называется *линейным*, если **D** (A) — линейное многообразие в **E** и для любых $x_1, x_2 \in$ **D** (A), $\alpha_1, \alpha_2 \in$ **C**

$$A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A x_1 + \alpha_2 A x_2. \tag{9}$$

Оператор A называется непрерывным в точке $x_0 \in$ **D** (A), если из $\|x_n - x_0\| \to 0$, $x_n \in$ **D** (A) следует, что $\|Ax_n - Ax_0\| \to 0$.

Линейный оператор называется *ограниченным*, если

$$\|Ax\|_{\mathbf{M}} \leqslant c\|x\|_{\mathbf{E}}, \tag{10}$$

где $c > 0$ — некоторая постоянная, которая не зависит от выбора $x \in$ **E**; $\|\cdot\|_{\mathbf{M}}$, $\|\cdot\|_{\mathbf{E}}$ — нормы соответственно в **M** и **E**. Наименьшее из чисел c, удовлетворяющее условию (10), называется *нормой оператора* A и обозначается $\|A\|_{\mathbf{E}\to\mathbf{M}}$.

Если **M** совпадает с **E**, то норму оператора A обозначают $\| A \|$.

Из определения следует, что

$$\| A \|_{\mathbf{E}\to\mathbf{M}} = \sup_{x\in\mathbf{E}} \frac{\| Ax \|_{\mathbf{M}}}{\| x \|_{\mathbf{E}}} = \sup_{\|x\|_{\mathbf{E}}=1} \| Ax \|_{\mathbf{M}}. \tag{11}$$

Теорема 4. Для того чтобы линейный оператор, действующий из **E** в **M**, был непрерывным, необходимо и достаточно, чтобы он был ограниченным.

Пусть $\{A_n\}$ — последовательность ограниченных линейных операторов, действующих из линейного нормированного пространства **E** в **M**. Последовательность $\{A_n\}$ называется *сходящейся по норме* к линейному ограниченному оператору $A_0 : \mathbf{E} \to \mathbf{M}$, если

$$\lim_{n\to\infty} \| A_0 - A_n \|_{\mathbf{E}\to\mathbf{M}} = 0.$$

Последовательность $\{A_n\}$ называется *сильно сходящейся* к оператору A_0, если

$$\lim_{n\to\infty} \| A_0 x - A_n x \|_{\mathbf{E}\to\mathbf{M}} = 0$$

при любом $x \in \mathbf{E}$.

Последовательность $\{A_n\}$ называется *слабосходящейся* к оператору A_0, если при любом $x \in \mathbf{E}$ последовательность $\{A_n x\}$ слабо сходится к $A_0 x$, т. е.

$$\lim_{n\to\infty} A_n x = A_0 x. \quad \forall\, x \in \mathbf{E}.$$

Из сходимости по норме следует сильная сходимость, из сильной — слабая.

Теорема 5. Для того чтобы последовательность ограниченных линейных операторов $\{A_n\}$, отображающих пространство **E** типа **B** в пространство **M** типа **B**, сильно сходилась к некоторому ограниченному линейному оператору, необходимо и достаточно, чтобы:

1) нормы операторов A_n были ограниченными в совокупности;

2) на всех элементах x всюду плотного в **E** множества **T** последовательность $\{A_n x\}$ была бы сходящейся.

Эта теорема имеет широкое применение в вопросах, связанных со сходимостью интерполяционных процессов, процессов механических квадратур и т. д. Необходимость первого условия называется *теоремой Банаха — Штейнгауза.*

Два оператора A и B называются *равными,* если области их определения совпадают и для всех $x \in \mathbf{D}(A) = \mathbf{D}(B)$ выполнено условие

$$Ax = Bx.$$

Если $\mathbf{D}(A_i) = \mathbf{R}(A_i) = \mathbf{E}$ $(i = 1, 2)$, то на множестве **E** можно ввести понятие *произведения* операторов A_1 и A_2:

$$Ax = (A_1 A_2)\, x = A_1 (A_2 x).$$

Для любых двух линейных ограниченных операторов A и B их произведение AB — линейный ограниченный оператор:

$$\| AB \| \leqslant \| A \| \| B \|.$$

Если $(AB)\, x = (BA)\, x$ для всех $x \in \mathbf{E}$, то A и B называются *коммутативными или перестановочными.*

Пусть $\mathbf{D}(A) = \mathbf{E}$, $\mathbf{R}(A) = \mathbf{M}$, (**E**, **M** — линейные нормированные пространства). Если каждому $y \in \mathbf{M}$ соответствует только один $x \in \mathbf{E}$, для которого $Ax = y$, то это соответствие можно рассматривать как оператор A^{-1}, определенный на $\mathbf{M} = \mathbf{R}(A)$ со значениями, заполняющими $\mathbf{E} = \mathbf{D}(A)$. Оператор A^{-1} называется *обратным оператором к* A. По определению

$$A^{-1}Ax = x \qquad (x \in \mathbf{E}) \quad \text{и} \quad AA^{-1}y = y \qquad (y \in \mathbf{M}).$$

Теорема 6. (С. Банаха). Если линейный ограниченный оператор A, отображающий банахово пространство **E** на банахово пространство **M**, имеет обратный A^{-1}, то оператор A^{-1} ограничен.

Из теоремы об обратном операторе вытекает, что из существования и единственности решения уравнения

$$Ax = y$$

при всякой правой части из **M** следует непрерывная зависимость решения $x = A^{-1}y$ от правой части.

Теорема 7. Пусть A — линейный оператор, действующий из банахова пространства **E** в банахово пространство **M**,

$$\mathbf{D}(A) = \mathbf{E}, \quad \mathbf{R}(A) = \mathbf{M}.$$

Для того чтобы обратный оператор A^{-1} существовал и был ограниченным ($\mathbf{D}(A^{-1}) = \mathbf{M}$, $\mathbf{R}(A^{-1}) = \mathbf{E}$), необходимо и достаточно, чтобы существовала такая постоянная $m > 0$, что $\forall\ x \in \mathbf{E}$

$$\| Ax \|_{\mathbf{M}} \geqslant m \| x \|_{\mathbf{E}}, \tag{12}$$

при этом будет выполняться неравенство

$$\| A^{-1} \| \leqslant \frac{1}{m}.$$

Линейные функционалы. Если значениями оператора являются вещественные числа, то оператор называется *функционалом*. Аддитивный однородный функционал, определенный в некотором линейном пространстве **L**, называется *линейным функционалом*, т. е. линейный функционал удовлетворяет условиям:

1) $F(x + y) = F(x) + F(y)$ (аддитивность);

2) $F(\alpha x) = \alpha F(x)$ (однородность).

Функционал F, аддитивный и сопряженно-однородный, определенный в комплексном линейном пространстве, называется *сопряженно-линейным*, т. е. он удовлетворяет условиям:

1) $F(x + y) = F(x) + F(y)$,

2) $F(\alpha x) = \bar{\alpha} F(x)$.

Так как множество **R** вещественных чисел есть пространство типа **B**, то для линейных функционалов сохраняются все определения и теоремы, приведенные выше для линейных операторов.

Совокупность элементов линейного множества **L**, для которых выполняется уравнение $F(x) = C$, называется *гиперплоскостью*. Очевидно, гиперплоскость — максимальное линейное многообразие.

Если функционал $F(x)$ непрерывен на **L**, то гиперплоскость $L_F = \{x\colon F(x) = c\}$ — замкнута.

Теорема 8. Пусть **T** — непустое выпуклое открытое множество в линейном нормированном пространстве **L** и **M** — линейное многообразие, не пересекающееся с **T**, т. е. $\mathbf{T} \cap \mathbf{M} = \varnothing$. Тогда существует замкнутая гиперплоскость, содержащая **M** и не пересекающаяся с **T**.

Теорема 5 называется *теоремой Хана — Банаха в геометрической форме*.

Пространство линейных ограниченных операторов в комплексном банаховом пространстве. Линейные ограниченные операторы в комплексном банаховом пространстве **E**, действующие в то же банахово пространство, сами образуют банахово пространство **B** (**E**) с банаховой алгеброй. Это означает, что они образуют полное нормированное пространство с нормой оператора $\| A \| = \sup\limits_{\| x \| = 1} \| Ax \|$, $x \in \mathbf{E}$, удовлетворяющей всем аксиомам нормы:

1) $\| A \| > 0$, если $\| A \| = 0$, то $\| Ax \| = 0$ для всех x, т. е. $A = 0$;

2) $\| \beta A \| = | \beta | \, \| A \|$; (13)

3) $\| A + B \| \leqslant \| A \| + \| B \|$,

и, кроме того, $\| AB \| \leqslant \| A \| \, \| B \|$.

Для элементов **B** (**E**) над полем комплексных чисел определено понятие сложения и умножения, причем выполняются следующие условия:

$$\begin{aligned} &1)\ (AB)\,D = A\,(BD);\\ &2)\ (\beta A)\,B = A\,(\beta B);\\ &3)\ A\,(\beta B + \mu D) = \beta AB + \mu AD;\\ &4)\ \text{существует единица } I \in \mathrm{B}\,(\mathrm{E})\\ &\qquad IA = AI = A. \end{aligned} \tag{14}$$

Резольвентным множеством оператора $A \in$ **B** (**E**) называется множество $\rho\,(A)$ всех комплексных чисел λ, для которых оператор $(A - \lambda I)^{-1}$ существует и является ограниченным оператором. Дополнение множества $\rho\,(A)$ до поля всех комплексных чисел называется *спектром* оператора A и обозначается $\mathrm{Sp}\,(A)$. Оператор $R\,(\lambda, A) = (A - \lambda I)^{-1}$, где $\lambda \in \rho\,(A)$, называется *резольвентой* оператора A.

Пусть **B** (**E**) — пространство типа **B**, образованное линейными ограниченными операторами, действующими из комплексного банахова пространства **E** в то же пространство. Справедливы следующие теоремы.

Теорема 9. Если пространство **E** типа **B** содержит более одного элемента, то $\mathrm{Sp}\,(A)$ — ограниченное замкнутое множество, лежащее в замкнутом круге с центром в точке нуль и радиусом, равным $\| A \|$.

Теорема 10. Для всякого оператора $A \in$ **B** (**E**) существует предел

$$r\,(A) = \lim_{n\to\infty} \sqrt[n]{\| A^n \|}, \tag{15}$$

называемый *спектральным радиусом*. Очевидно, $r\,(A) \leqslant \| A \|$.

Теорема 11. Если $|\lambda| > r\,(A)$, то резольвента $R\,(\lambda, A)$ существует и записывается рядом вида

$$R\,(\lambda, A) = (A - \lambda I)^{-1} = -\sum_{n=0}^{\infty} \lambda^{-(n+1)} A^n, \tag{16}$$

который сходится по норме операторов.

Теорема 12. При $A \in$ **B** (**E**) имеет место формула

$$r\,(A) = \sup_{\lambda \in \mathrm{Sp} A} |\lambda|. \tag{17}$$

Теорема 13. При $A \in$ **B** (**E**) операторный ряд

$$I + A + A^2 + \cdots$$

сходится, если $r\,(A) < 1$ и его сумма равна $(I - A)^{-1}$, и расходится при $r\,(A) > 1$.

Теорема 14. Если $A \in$ **B** (**E**) и $\| A \| \leqslant q < 1$, то оператор $I - A$ имеет линейный обратный, причем

$$\| (I - A)^{-1} \| \leqslant (1 - q)^{-1}. \tag{18}$$

Теорема 15. Если A, $A^{-1} \in$ **B** (**E**), то множество **G** элементов **B** (**E**), имеющих в **B** (**E**) обратные, содержит вместе с оператором A сферу радиуса

$$\| A - D \| < \| A^{-1} \|^{-1}.$$

Если оператор D лежит в этой сфере, то его обратный представим рядами:

$$D^{-1} = A^{-1} \sum_{n=0}^{\infty} [(A - B)\,A^{-1}]^n, \tag{19}$$

или

$$D^{-1} = \sum_{n=0}^{\infty} [A^{-1}\,(A - B)]^n\,A^{-1}. \tag{20}$$

Если $D_\varepsilon \in$ **G** и $\| D_\varepsilon - A \| \to 0$ при $\varepsilon \to 0$, то и

$$\| A^{-1} - D_\varepsilon^{-1} \|_{\varepsilon\to 0} \to 0.$$

Теорема 15 называется *теоремой о возмущениях* и является фундаментальной при обосновании многих вопросов вычислительной математики.

Пусть в **B** задано уравнение

$$Au = f. \tag{21}$$

Имеет место следующая теорема.

Теорема 16. Уравнение (21) имеет единственное решение $\forall f \in \mathbf{B}$ тогда и только тогда, когда существует оператор D, имеющий обратный D^{-1}, обладающий таким свойством, что ряд

$$\sum_{k=0}^{\infty} T^k g,$$

где $T = I - DA$, сходится для любого $g \in$ B. Решение уравнения (21) в этом случае дается формулой

$$u = \sum_{k=0}^{\infty} T^k Df. \tag{22}$$

Линейные ограниченные операторы в вещественном гильбертовом пространстве. Пусть в вещественном гильбертовом пространстве **H** задан линейный ограниченный оператор A с $\mathbf{D}(A) = \mathbf{H}$.

Будем называть оператор A *положительно полуопределенным* (неотрицательным), если

$$(Ax, x) \geqslant 0, \quad \forall x \in \mathbf{H}, \tag{23}$$

причем возможность равенства нулю скалярного произведения (Ax, x) допускается на элементе x, тождественно не равном нулю. Такие операторы обычно обозначают $A \geqslant 0$.

Если равенство нулю исключается и $(Ax, x) > 0$ для всех $x \in \mathbf{H}$, кроме $x = 0$, то оператор называют *положительным* и обозначают $A > 0$.

Если

$$(Ax, x) \geqslant \gamma^2 \|x\|, \quad \forall x \in \mathbf{H}, \tag{24}$$

где $\gamma^2 > 0$ — число, единое для всех $x \in \mathbf{H}$, то оператор называют *положительно определенным* и обозначают $A \geqslant \gamma^2 I$.

Если

$$(Ax, x) \geqslant -\delta^2 \|x\|^2, \quad \forall x \in \mathbf{H}, \tag{25}$$

где δ^2 — положительное число, оператор A называют *полуограниченным снизу* и обозначают $A \geqslant -\delta^2 I$.

Говорят, что $A \geqslant B$, если $A - B$ неотрицательный оператор, $\mathbf{D}(A) = \mathbf{D}(B) = \mathbf{H}$ и для всех $x \in \mathbf{H}$ имеет место неравенство

$$((A - B)x, x) \geqslant 0. \tag{26}$$

Пусть A и A^* — линейные операторы, заданные на **H**.

Оператор A^* называется *сопряженным* оператору A, если для всех $x, y \in \mathbf{H}$ выполнено равенство

$$(Ax, y) = (x, A^*y). \tag{27}$$

Линейный ограниченный оператор A называется *самосопряженным*, если для всех $x, y \in \mathbf{H}$ имеет место равенство

$$(Ax, y) = (x, Ay), \tag{28}$$

т. е. $A = A^*$.

Если A — линейный ограниченный оператор, то сопряженный оператор A^* также является линейным ограниченным и $\|A\| = \|A^*\|$.

Для любого линейного оператора A с $\mathbf{D}(A) = \mathbf{H}$ операторы A^*A и AA^* — самосопряженные неотрицательные операторы. Заметим также, что $(A^*)^* = A$, $(A^*)^{-1} = (A^{-1})^*$.

Для нормы самосопряженного оператора A с $\mathbf{D}(A) = \mathbf{H}$ имеет место формула

$$\|A\| = \sup_{\|x\| \neq 0} \frac{|(Ax, x)|}{\|x\|^2} = \sup_{\|x\|=1} |(Ax, x)| \tag{29}$$

и

$$\|A\| = r(A).$$

Для произвольного неотрицательного оператора A, заданного на $\mathbf{H}$. можно определить число (Ax, x), которое называют энергией оператора A.

Если A — положительный самосопряженный линейный оператор, то при помощи числа $(Ax, y) = (x, y)_A$ можно на линейной системе $\mathbf{D}(A) \subset \mathbf{H}$ ввести скалярное произведение $(x, y)_A$ и норму $\|x\|_A = \sqrt{(Ax, x)}$. Замыкание $\mathbf{D}(A)$ в смысле сходимости по норме $\|\cdot\|_A$ образует энергетическое гильбертово пространство $\mathbf{H}_A$. Очевидно, для положительного самосопряженного оператора A в $\mathbf{H}$ имеет место обобщенное неравенство Коши — Буняковского

$$(Ax, y)^2 \leqslant (Ax, x)\,(Ay, y). \tag{30}$$

Если A — положительный самосопряженный оператор и A^{-1} существует, то можно ввести на линейном множестве $\mathbf{R}(A) \in \mathbf{H}$ скалярное произведение при помощи соотношений

$$(x, y)_{A^{-1}} = (A^{-1}x, y) \tag{31}$$

и «*негативную*» норму

$$\|x\|_{A^{-1}} = (A^{-1}x, x)^{\frac{1}{2}}. \tag{32}$$

Замыкание $\mathbf{R}(A)$ в смысле сходимости по норме $\|\cdot\|_{A^{-1}}$ образует гильбертово пространство $\mathbf{H}_{A^{-1}}$ с негативной нормой $\|\cdot\|_{A^{-1}}$.

Теорема 17. Пусть A — положительно определенный ограниченный линейный оператор с $\mathbf{D}(A) = \mathbf{H}$. Тогда существует ограниченный обратный оператор A^{-1} с $\mathbf{D}(A^{-1}) = \mathbf{H}$.

Оператор B называется *квадратным корнем* из оператора A, если $B^2 = A$.

Теорема 18. Если $A = A^* > 0$, то существует единственный неотрицательный самосопряженный квадратный корень из оператора A, обозначаемый через $A^{\frac{1}{2}}$, перестановочный со всяким оператором, перестановочным с A.

Оператор, перестановочный со своим сопряженным, называется *нормальным оператором*

$$AA^* = A^*A.$$

Оператор A называется *унитарным*, если $AA^* = I$.

Оператор A называется *кососимметрическим*, если $A^* = -A$. Любой оператор A можно представить в виде суммы самосопряженного и кососимметрического

$$A = A_0 + A_1, \quad A_0 = A_0^* = \frac{1}{2}(A + A^*), \quad A_1 = -A_1^* = \frac{1}{2}(A - A^*).$$

Пусть B — линейный самосопряженный положительно определенный оператор, действующий из $\mathbf{H}$ в $\mathbf{H}$. Будем говорить, что линейный оператор $A : \mathbf{H} \to \mathbf{H}$ *эквивалентен по спектру* (энергетически эквивалентен) оператору B и записывать

$$B \sim A,$$

если существуют числа $\gamma_1, \gamma_2, \gamma_3$ ($\gamma_2 \geqslant \gamma_1 > 0, \gamma_3 \geqslant 0$) такие, что для любого элемента $x \in \mathbf{H}$ справедливы неравенства

$$\gamma_1 \|x\|_B^2 \leqslant (A_0x, x) \leqslant \gamma_2 \|x\|_B^2, \tag{33}$$

$$|(B^{-1}A_1x, A_1x)| \leqslant \gamma_3 \|x\|_B^2, \tag{34}$$

где

$$A_0 = \frac{1}{2}(A + A^*), \quad A_1 = \frac{1}{2}(A - A^*).$$

Числа γ_1, γ_2 и γ_3 называют оценками эквивалентности. Если $A = A^*$, то, полагая что $\gamma_3 = 0$, получим следующие условия эквивалентности операторов по спектру

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x). \tag{35}$$

Линейные операторы в конечном вещественном эвклидовом пространстве. Пусть A — линейный оператор, заданный в конечномерном линейном нормированном пространстве R_n с нормой $\|x\| = \sqrt{(x, x)}$ и ортонормированными базисными векторами $(\varphi_k)_{k=\overline{1,n}}$. Тогда каждому оператору A в пространстве R_n соответствует матрица $(a_{ij})_{i=\overline{1,n}}^{j=\overline{1,n}}$ размерности $n \times n$, причем

$$(a_{ik})_{i=\overline{1,n}} = A\varphi_k \qquad (k = \overline{1, n})$$

и всякая матрица $(a_{ij})_{i=\overline{1,n}}^{j=\overline{1,n}}$ определяет линейный оператор A. Самосопряженный оператор A в R_n имеет n взаимно ортогональных собственных векторов φ_k $(k = \overline{1, n})$ и произвольный вектор $x \in R_n$ можно разложить по этим собственным векторам

$$x = \sum_{k=1}^{n} c_k\varphi_k, \quad c_k = (x, \varphi_k), \quad \|x\|^2 = \sum_{k=1}^{n} c_k^2.$$

Если A и B самосопряженные ($A = A^*$, $B = B^*$) перестановочные ($AB = BA$) операторы в R_n, то они имеют общую систему собственных векторов φ_k $(k = \overline{1, n})$ и для собственных значений $\lambda_{AB}^{(k)}$, $\lambda_{A+B}^{(k)}$ соответственно операторов AB и $A + B$ имеют место равенства:

$$\lambda_{AB}^{(k)} = \lambda_A^{(k)}\lambda_B^{(k)}, \quad \lambda_{A+B}^{(k)} = \lambda_A^{(k)} + \lambda_B^{(k)} \qquad (k = \overline{1, n}),$$

где $\lambda_A^{(k)}$, $\lambda_B^{(k)}$, $\lambda_{AB}^{(k)}$, $\lambda_{A+B}^{(k)}$ — собственные значения номера k соответственно операторов A, B, AB, $A + B$.

Для собственных значений самосопряженного неотрицательного линейного оператора A ($A = A^* \geqslant 0$) имеют место следующие соотношения:

$$\begin{gathered} \lambda_A^{(1)}(x, x) \leqslant (Ax, x) \leqslant \lambda_A^{(n)}(x, x), \\ \lambda_A^{(1)}\|x\| \leqslant \|Ax\| \leqslant \lambda_A^{(n)}\|x\|, \\ \lambda_A^{(k)} > 0, \quad (k = \overline{1, n}), \quad \lambda_A^{(1)} \leqslant \cdots \leqslant \lambda_A^{(n)}. \end{gathered} \tag{36}$$

В вещественном конечно-мерном пространстве R будем рассматривать скалярные произведения и нормы вида

$$\begin{gathered} (x, y) = \sum_{i=1}^{n-1} x_i y_i h, \quad (x, y] = \sum_{i=1}^{n} h x_i y_i, \quad [x, y) = \sum_{i=0}^{n} h x_i y_i, \\ \|x\| = \sqrt{(x, x)}, \quad \|x]| = \sqrt{(x, x]}, \quad |[x\| = \sqrt{[x, x)}, \end{gathered} \tag{37}$$

$$\begin{aligned} &\|x\|_1 = \max|x_i| && \text{(кубическая норма или } \|x\|_\infty\text{)}, \\ &\|x\|_2 = \sum_{i=1}^{n}|x_i| && \text{(октаэдрическая норма)}, \\ &\|x\|_3 = \sqrt{(x, x)} = \sqrt{\sum_{i=1}^{n} x_i^2} && \text{(сферическая норма, или евклидова длина вектора)}. \end{aligned} \tag{38}$$

Векторным нормам $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_3$ соответствуют следующие *подчиненные* им матричные нормы:

$$\|A\|_1 = \max_{1 \leqslant i \leqslant n} \sum_{k=1}^{n} |a_{ik}|,$$

$$\| A \|_2 = \max_{1 \leqslant k \leqslant n} \sum_{i=1}^{n} | a_{ik} |, \tag{39}$$

$$\| A \|_3 = \max_{1 \leqslant i \leqslant n} \sqrt{\lambda_{AA'}^{(i)}},$$

где $\lambda_{AA'}^{(i)}$ — собственное число матрицы AA'.

Тензорное произведение матриц.

Пусть $A = (a_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}}$ и $B = (b_{ij})_{i=\overline{1,l}}^{j=\overline{1,k}}$ — две матрицы соответственно порядка $n \times m$ и $l \times k$.

Тензорным (иначе прямым или кронекеровским) произведением матриц A и B ($A \oplus B$) называется матрица порядка $nl \times mk$ вида

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1m}B \\ a_{21}B & a_{22}B & \dots & a_{2m}B \\ \dots & \dots & \dots & \dots \\ a_{n1}B & a_{n2}B & \dots & a_{nm}B \end{pmatrix}.$$

Из определения следует, что

$(A + D) \otimes B = A \otimes B + D \otimes B,$

$(A \otimes B)' = A' \otimes B',$

$(A \otimes B)(D \otimes C) = AD \otimes BC,$

$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$

Собственное число $\lambda_{A \otimes B}$ и собственный вектор v матрицы $A \otimes B$ выражаются через собственные числа λ_A, λ_B и собственные векторы φ, ψ соответственно матриц A и B, а именно:

$$\lambda_{A \otimes B} = \lambda_A \lambda_B,$$
$$v = \varphi \otimes \psi.$$

§ 2. ОРТОГОНАЛЬНЫЕ МНОГОЧЛЕНЫ И НЕКОТОРЫЕ СПЕЦИАЛЬНЫЕ ФУНКЦИИ

1. Ортогональные многочлены непрерывного аргумента

Общие свойства ортогональных многочленов непрерывного аргумента. Явное выражение для многочленов $P_n(x)$, ортогональных с весом $\rho(x)$, на интервале $[a, b]$ задается с помощью формулы

$$P_n(x) = A_n \begin{vmatrix} c_0 & c_1 & \dots & c_n \\ c_1 & c_2 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ 1 & x & \dots & x^n \end{vmatrix}, \tag{1}$$

где $c_n = \int_a^b x^n \rho(x)\, dx$ — момент весовой функции, A_n — нормирующая постоянная.

Любые три последовательных многочлена из ортогонального семейства $\{P_n(x)\}_{n=0}^{\infty}$ связаны трехчленным рекуррентным соотношением

$$xP_n(x) = \frac{k_{n,n}}{k_{n+1,n+1}} P_{n+1}(x) + \left(\frac{k_{n-1,n}}{k_{n,n}} - \frac{k_{n,n+1}}{k_{n+1,n+1}}\right) P_n(x) + \\ + \frac{k_{n-1,n-1}}{k_{n,n}} \cdot \frac{h_n^2}{h_{n-1}^2} P_{n-1}(x), \tag{2}$$

где $h_n^2 = \int_a^b P_n^2(x)\,\rho(x)\,dx$ — квадрат нормы многочлена $P_n(x)$, $k_{n-i,n}$ — коэффициенты при степенях x^{n-i} многочлена $P_n(x)$:

$$P_n(x) = \sum_{i=0}^{n} k_{n-i,n} x^{n-i}.$$

Имеет место *тождество Кристоффеля — Дарбу*

$$\sum_{k=0}^{n} \frac{P_k(x)\,P_k(y)}{h_k^2} = \frac{1}{h_n^2} \cdot \frac{k_{n,n}}{k_{n+1,n+1}} \cdot \frac{P_{n+1}(x)\,P_n(y) - P_n(x)\,P_{n+1}(y)}{x-y}, \tag{3}$$

следствием которого является соотношение

$$\sum_{k=0}^{n} \frac{P_k^2(x)}{h_k^2} = \frac{1}{h_n^2} \cdot \frac{k_{n,n}}{k_{n+1,n+1}} [P'_{n+1}(x)\,P_n(x) - P_{n+1}(x)\,P'_n(x)]. \tag{4}$$

Любой многочлен степени $m < n$ является линейной комбинацией многочленов $P_0(x)$, $P_1(x)$, ..., $P_m(x)$ и, следовательно, ортогонален к $P_n(x)$. Это приводит к следующему утверждению о нулях ортогональных многочленов.

Теорема 1. Все нули $P_n(x)$ являются простыми и расположены строго внутри отрезка ортогональности $[a, b]$, между двумя последовательными нулями $P_n(x)$ расположен в точности один нуль $P_{n+1}(x)$ и по крайней мере один нуль $P_m(x)$, для которого $m > n$.

В дальнейшем нули ортогональных многочленов $P_n(x)$, как правило, будем обозначать следующим образом:

$$x_{1,n}^P > x_{2,n}^P > \cdots > x_{n,n}^P,$$

где верхний индекс P показывает, к какой именно системе ортогональных многочленов относятся эти нули.

Классические ортогональные многочлены. Среди всех систем ортогональных многочленов особое место как по степени исследования, так и по широте использования занимают классические ортогональные многочлены. Классификация классических ортогональных многочленов приведена в следующей таблице:

Классические ортогональные многочлены	Интервал ортогональности (a, b)	Весовая функция, $\rho(x)$
Лагранжа $P_n(x)$	$[-1, 1]$	1
Гегенбауэра (или ультрасферические) $C_n^{\tau}(x)$	$[-1, 1]$	$(1-x^2)^{\tau-\frac{1}{2}}$
Якоби $P_n^{(\alpha,\beta)}(x)$	$[-1, 1]$	$(1-x)^{\alpha}(1+x)^{\beta}$
Лагерра $L_n^{\alpha}(x)$	$[0, \infty)$	$x^{\alpha} e^{-x}$
Эрмита $H_n(x)$	$(-\infty, \infty)$	e^{-x^2}

Все эти многочлены обладают целым рядом общих свойств, наиболее важными из которых являются:

1) многочлены $\{P_n'(x)\}_{n=0}^{\infty}$ образуют ортогональную систему;

2) $P_n(x)$ удовлетворяют дифференциальному уравнению вида

$$\sigma(x)\,y'' + \tau(x)\,y' + \lambda_n\,y = 0, \tag{5}$$

где $\sigma(x)$ и $\tau(x)$ — многочлены не выше второй и первой степени соответственно, а λ_n не зависит от x;

3) имеет место обобщенная формула Родрига

$$P_n(x) = \frac{B_n}{\rho(x)} \cdot \frac{d^n}{dx^n}\left[\sigma^n(x)\,\rho(x)\right], \tag{6}$$

где B_n — постоянная;

4) вес $\rho(x)$ удовлетворяет дифференциальному уравнению Пирсона

$$\frac{d}{dx}\left[\sigma(x)\,\rho(x)\right] = \tau(x)\,\rho(x). \tag{7}$$

Здесь в зависимости от типа интервала ортогональности (a, b) многочлен второй степени $\sigma(x)$ имеет вид

$$\sigma(x) = \begin{cases} (x-a)(b-x), & a,\ b \neq \pm\infty, \\ (x-a), & a \neq -\infty,\ b = \infty, \\ (b-x), & a = -\infty,\ b \neq \infty, \\ 1, & -a = b = \infty. \end{cases} \tag{8}$$

Основные характеристики классических ортогональных многочленов приведены в следующей таблице:

Многочлены / Основные характеристики	$P_n^{(\alpha,\beta)}(x),\ \alpha, \beta > -1$	$L_n^{\alpha}(x),\ \alpha > -1$	$H_n(x)$
$\sigma(x)$	$1-x^2$	x	1
$\tau(x)$	$\beta-\alpha-(\alpha+\beta+2)x$	$1+\alpha-x$	$-2x$
λ_n	$n(n+\alpha+\beta+1)$	n	$2n$
B_n	$\dfrac{(-1)^n}{2^n\,n!}$	$\dfrac{1}{n!}$	$(-1)^n$
h_n^2	$\dfrac{2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\,\Gamma(n+\beta+1)}{n!\,(2n+\alpha+\beta+1)\,\Gamma(n+\alpha+\beta+1)}$	$\dfrac{\Gamma(n+\alpha+1)}{n!}$	$2^n n!\sqrt{\pi}$
$k_{n,\,n}$	$\dfrac{\Gamma(2n+\alpha+\beta+1)}{2^n n!\,\Gamma(n+\alpha+\beta+1)}$	$\dfrac{(-1)^n}{n!}$	2^n
$k_{n-1,\,n}$	$\dfrac{(\alpha-\beta)\,\Gamma(2n+\alpha+\beta)}{2^n\,(n-1)!\,\Gamma(n+\alpha+\beta+1)}$	$(-1)^{n-1}\dfrac{n+\alpha}{(n-1)!}$	0

Классические ортогональные многочлены, как и многие другие специальные функции, являются решениями дифференциального уравнения гипергеометрического типа

$$\sigma(z)\, y'' + \tau(z)\, y' + \lambda y = 0,$$

где $\sigma(z)$ и $\tau(z)$ — произвольные многочлены не выше второй и первой степеней соответственно, λ — произвольное комплексное число.

Уравнение гипергеометрического типа имеет следующий канонический вид:

$$z(1-z)\, y'' + [\gamma - (\alpha + \beta + 1)\, z]\, y' - \alpha\beta y = 0, \tag{9}$$

вырожденное гипергеометрическое уравнение имеет вид

$$zy'' + (\gamma - z)\, y' - \alpha y = 0 \tag{10}$$

и, наконец, *уравнение для функций Эрмита*

$$y'' - 2zy' + 2\nu y = 0. \tag{11}$$

Частными решениями уравнения (9) являются выражения:

$$y_1 = F(\alpha, \beta; \gamma; z), \quad y_2 = z^{1-\gamma} F(\alpha - \gamma + 1, \beta - \gamma + 1; 2 - \gamma; z), \tag{12}$$

где $y = F(\alpha, \beta; \gamma; z) = \sum\limits_{n=0}^{\infty} \dfrac{(\alpha)_n (\beta)_n}{n!\,(\gamma)_n} z^n$, $|z| < 1$ — *гипергеометрический ряд.*

Здесь $(a)_n = a(a+1)\ldots(a+n-1)$, $n = 1, 2, \ldots$, $(a)_0 = 0$ — символ Похгаммера.

Приведем конкретный вид формул для рассмотренных выше классических ортогональных многочленов.

Многочлены Якоби. Формула Родрига:

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)}{2^n n!} (1-x)^{-\alpha} (1+x)^{-\beta} \frac{d^n}{dx^n} [(1-x)^{\alpha+n} (1+x)^{\beta+n}].$$

Рекуррентная формула:

$$2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)\, P_{n+1}^{(\alpha,\beta)}(x) =$$
$$= (2n+\alpha+\beta+1)[(2n+\alpha+\beta)(2n+\alpha+\beta+2)\, x + \alpha^2 - \beta^2]\, P_n^{(\alpha,\beta)}(x) -$$
$$- 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)\, P_{n-1}^{(\alpha,\beta)}(x),$$
$$n = 1, 2, \ldots$$

Дифференциальное уравнение:

$$(1-x^2)\, y'' + [\beta - \alpha - (\alpha+\beta+2)\, x]\, y' + n(n+\alpha+\beta+1)\, y = 0.$$

Многочлены Гегенбауэра, или ультрасферические многочлены. Связь с многочленами Якоби:

$$C_n^{\tau}(x) = \frac{(2\tau)_n}{\left(\tau + \frac{1}{2}\right)_n} P_n^{\left(\tau - \frac{1}{2},\, \tau - \frac{1}{2}\right)}(x).$$

Формула Родрига:

$$C_n^{\tau}(x) = (-1)^n (1-x^2)^{\frac{1}{2} - \tau} \frac{(2\tau)_n}{2^n n! \left(\tau + \frac{1}{2}\right)_n} \cdot \frac{d^n}{dx^n} [(1-x^2)^{n + \tau - \frac{1}{2}}].$$

Рекуррентная формула:

$$(n+1)\, C_{n+1}^{\tau}(x) = 2(n+\tau)\, x C_n^{\tau}(x) - (n + 2\tau - 1)\, C_{n-1}^{\tau}(x), \qquad n = 0, 1, 2, \ldots;$$
$$C_{-1} = 0; \quad C_0 = 1.$$

Дифференциальное уравнение:

$$(1-x^2)\,y''-(2\tau+1)\,xy'+n\,(n+2\tau)\,y=0.$$

Многочлены Лежандра. Связь с другими ортогональными многочленами:

$$P_n(x)=C_n^{\frac{1}{2}}(x)=P_n^{(0,0)}(x).$$

Формула Родрига:

$$P_n(x)=\frac{1}{2^n n!}\cdot\frac{d^n}{dx^n}\,[(x^2-1)^n].$$

Рекуррентная формула:

$$(n+1)\,P_{n+1}(x)=(2n+1)\,xP_n(x)-nP_{n-1}(x),\qquad n=0,\ 1,\ \ldots;$$

$$P_{-1}=0;\quad P_0=1.$$

Дифференциальное уравнение:

$$(1-x^2)\,y''-2xy'+n\,(n+1)\,y=0.$$

Многочлены Чебышева первого и второго рода. Многочлены, задаваемые формулами

$$T_n(x)=(g_n)^{-1}\,P_n^{\left(-\frac{1}{2},\,-\frac{1}{2}\right)}(x),\quad U_n(x)=(2g_{n+1})^{-1}P_n^{\left(\frac{1}{2},\,\frac{1}{2}\right)}(x),$$

называются многочленами Чебышева первого и второго рода соответственно. Здесь $g_n=\dfrac{\left(\frac{1}{2}\right)_n}{n!}$.

Формула Родрига:

$$T_n(x)=\frac{(-1)^n}{2^n\left(\frac{1}{2}\right)_n}\sqrt{1-x^2}\,\frac{d^n}{dx^n}\,[(1-x^2)^{n-\frac{1}{2}}],$$

$$U_n(x)=\frac{(-1)^n\,(n+1)}{2^{n+1}\left(\frac{1}{2}\right)_{n+1}}\,(1-x^2)^{-\frac{1}{2}}\,\frac{d^n}{dx^n}\,[(1-x^2)^{n+\frac{1}{2}}].$$

Рекуррентная формула одна и та же и для $T_n(x)$ и для $U_n(x)$:

$$z_{n+1}(x)=2xz_n(x)-z_{n-1}(x),$$

где $z_n(x)$ является либо $T_n(x)$, либо $U_n(x)$.

Дифференциальные уравнения:

$$(1-x^2)\,y''-xy'+n^2y=0,\quad y=T_n(x),$$

$$(1-x^2)\,y''-3xy'+n\,(n+2)\,y=0,\quad y=U_n(x).$$

Связь с тригонометрическими функциями

$$T_n(\cos\theta)=\cos n\theta,\quad U_n(\cos\theta)=\frac{\sin(n+1)\,\theta}{\sin\theta}.$$

Многочлены Лагерра. Формула Родрига:

$$L_n^{\alpha}(x)=\frac{e^x x^{-\alpha}}{n!}\cdot\frac{d^n}{dx^n}\,(e^{-x}x^{n+\alpha}).$$

Рекуррентное соотношение:

$$(n+1)\,L_{n+1}^{\alpha}(x)-(2n+\alpha+1-x)\,L_n^{\alpha}(x)+(n+\alpha)\,L_{n-1}^{\alpha}(x)=0,\qquad n=0,\,1,\,\ldots;$$

$$L_{-1}=0;\quad L_0=1.$$

Дифференциальное уравнение:

$$xy''+(\alpha+1-x)\,y'+ny=0.$$

Многочлены Эрмита. Формула Родрига:

$$H_n(x)=(-1)^n\,e^{x^2}\frac{d^n}{dx^n}(e^{-x^2}).$$

Рекуррентное соотношение:

$$H_{n+1}(x)-2xH_n(x)+2nH_{n-1}(x)=0,\qquad n=0,\,1,\,\ldots;$$

$$H_{-1}=0;\quad H_0=1.$$

Дифференциальное уравнение:

$$y''-2xy'+2ny=0.$$

В заключении приведем две асимптотические формулы для нулей ультрасферических многочленов и многочленов Лагерра:

$$x_{i,n}^{C^{\tau}}=\frac{x_{i,n}^{H}}{\sqrt{\tau-\frac{1}{2}}}\left\{1-\frac{1}{8}\left[2n+1+2\,(x_{i,n}^{H})^2\right]\left(\tau-\frac{1}{2}\right)^{-1}+O\,(\tau^{-2})\right\},$$

$$x_{i,n}^{L^{\alpha}}=\alpha-\sqrt{2\alpha}\left\{x_{i,n}^{H}-\frac{\sqrt{2}}{6}\left[2\,(x_{i,n}^{H})^2+2_n+1\right]\alpha^{-\frac{1}{2}}+O\,(\alpha^{-1})\right\},$$

где $x_{i,n}^{H}$ — нули многочлена Эрмита $H_n(x)$. Эти формулы могут быть полезными в качестве начальных приближений в алгоритмах отыскания нулей классических ортогональных многочленов, зависящих от параметра.

2. Ортогональные многочлены дискретного аргумента

Если под скалярным произведением двух функций $f(x)$ и $g(x)$ понимать выражение

$$(f,\,g)=\sum_i \rho\,(x_i)\,f\,(x_i)\,g\,(x_i),\tag{13}$$

где $\rho\,(x_i)>0$ и суммирование производится по всем x_i, которые удовлетворяют неравенству $a\leqslant x_i\leqslant b$, то многочлены, ортогональные в смысле введенного выше скалярного произведения, будут называться *ортогональными многочленами дискретного аргумента.*

Чаще всего выбирают точки x_i целыми $(x_i=i)$. Так же, как и в непрерывном случае, среди всех классов ортогональных многочленов дискретного аргумента особое место занимают классические ортогональные многочлены дискретного аргумента, которые обладают следующими свойствами:

1) разностные производные $\Delta p_n(x)$ классических ортогональных многочленов дискретного аргумента $p_n(x)$ являются классическими ортогональными многочленами дискретного аргумента;

2) $p_n(x)$ удовлетворяют разностному уравнению

$$\Delta\,[\sigma\,(x)\,\rho\,(x)\,\nabla p_n\,(x)]+\lambda_n\rho\,(x)p_n\,(x)=0,\tag{14}$$

где

$$\lambda_n=-n\left[\tau'\,(x)+(n-1)\,\frac{\sigma''\,(x)}{2}\right],\tag{15}$$

$\sigma\,(x)$ и $\tau\,(x)$ — многочлены не выше второй и первой степеней соответственно;

3) имеет место дискретный аналог формулы Родрига:

$$p_n(x) = \frac{A_n}{\rho(x)} \nabla^n [\rho_n(x)], \quad \rho_n(x) = \rho(x+n) \prod_{k=1}^{n} \tau(x+k), \tag{16}$$

где A_n — постоянная;

4) вес $\rho(x)$ удовлетворяет разностному аналогу уравнения Пирсона:

$$\Delta[\sigma(x)\rho(x)] = \tau(x)\rho(x). \tag{17}$$

Приведем типы классических ортогональных многочленов дискретного аргумента в виде такой таблицы:

(a, b)	$\rho(x)$	A_n	Многочлен	Обозначение
$[0, N-1]$	1	$\frac{1}{n!}$	Чебышева	$t_n(x)$
$[0, N]$	$p^x q^{N-x} \binom{N}{x}$, $p+q=1$; $p, q>0$	$\frac{(-1)^n}{n!}$	Кравчука	$K_n(x)$
$[0, \infty)$	$\frac{e^{-a} a^x}{\Gamma(x+1)}$, $0<a<1$	e^{-a}	Шарлье	$C_n(x; a)$
$[0, \infty)$	$c^x \frac{(\beta)_x}{x!}$, $0<c<1$; $\beta<0$	c^{-n}	Маикснера	$m_n(x; \beta; c)$
$[0, \infty)$	$\frac{(\beta)_x(\gamma)_x}{x!\,(\delta)_x}$	$\frac{1}{n!}$	В. Гана	$p_n(x; \beta, \gamma, \delta)$

Дадим конкретизацию некоторых приведенных выше формул.

Многочлены В. Гана. Аналог формулы Родрига:

$$p_n(x; \beta, \gamma, \delta) = \frac{1}{n!} \frac{x!\,(\delta)_x}{(\beta)_x(\gamma)_x} \Delta^n \left[\frac{(\beta)_x(\gamma)_x}{(x-n)!\,(\delta)_{x-n}} \right].$$

Разностное уравнение:

$$\Delta \left[x(N-x) \frac{\Gamma(\gamma+x)\Gamma(N-x)}{\Gamma(x+1)\Gamma(\delta+1-x)} \nabla p_n(x) \right] +$$
$$+ \frac{\Gamma(\gamma+x)\Gamma(N-x)}{\Gamma(x+1)\Gamma(\delta+1-x)} n(N+\gamma-\delta+n-1)\, p_n(x) = 0.$$

Многочлены Чебышева. Многочлены Чебышева дискретного аргумента являются частным случаем многочленов В. Гана:

$$t_n(x) = p_n(x;\ 1;\ 1-N,\ 1-N).$$

Аналог формулы Родрига:

$$t_n(x) = n!\,\Delta^n \left[\binom{x}{n} \binom{x-N}{n} \right].$$

Разностное уравнение:

$$(x+2)(x-N+2)\,\Delta^2 t_n(x) + [2x-N+3-n(n+1)]\,\Delta t_n(x) - n(n+1)\, t_n(x) = 0.$$

3. Некоторые специальные функции

Гамма-функция. Одним из способов определения гамма-функции может служить формула

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt = \int_0^1 \left(\ln \frac{1}{t} \right)^{z-1} dt, \quad \operatorname{Re} z > 0. \tag{18}$$

Имеет место функциональное уравнение

$$\Gamma(z+1) = z\Gamma(z), \tag{19}$$

которое для z, совпадающего с натуральным числом n, приводит к формуле

$$\Gamma(n+1) = n!$$

Справедливо соотношение

$$\Gamma(z)\,\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}, \tag{20}$$

из которого следует, что

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Числа и многочлены Бернулли. Числа Бернулли B_n определяются равенством

$$z(e^z - 1)^{-1} = \sum_{n=0}^{\infty} B_n \frac{z^n}{n!}, \quad |z| < 2\pi, \tag{21}$$

а многочлены Бернулли $B_n(x)$ — равенством

$$ze^{xz}(e^z - 1)^{-1} = \sum_{n=0}^{\infty} B_n(x) \frac{z^n}{n!}, \quad |z| < 2\pi. \tag{22}$$

Из (21) и (22) в качестве следствия имеем:

$$B_n(x) = \sum_{k=0}^{n} C_n^k B_k x^{n-k} \tag{23}$$

и, в частности,

$$B_0(x) = 1, \quad B_1(x) = x - \frac{1}{2}, \quad B_2(x) = x^2 - x + \frac{1}{6}, \ \ldots \tag{23'}$$

Очевидно,

$$B_n(0) = B_n.$$

Имеют место формула дифференцирования

$$B_n'(x) = nB_{n-1}(x) \tag{24}$$

и функциональное соотношение

$$B_n(x+1) - B_n(x) = nx^{n-1}, \qquad n = 0, 1, \ldots, \tag{25}$$

из которого вытекает, что

$$B_n(1) = B_n(0) = B_n. \tag{26}$$

Из (24) и (25) следует:

$$\int_x^y B_n(t)\,dt = \frac{B_{n+1}(y) - B_{n+1}(x)}{n+1}, \quad \int_x^{x+1} B_n(t)\,dt = x^n.$$

Многочлены Бернулли удовлетворяют соотношению

$$B_n(1-x) = (-1)^n B_n(x).$$

Для чисел Бернулли справедливы рекуррентные формулы

$$\sum_{k=0}^{n-1} C_n^k B_k = 0, \qquad n = 2, 3, \ldots, \tag{27}$$

из которых с учетом (23′) и (26) следует:

$$B_0 = 1,\ B_1 = -\frac{1}{2},\ B_2 = \frac{1}{6},\ B_4 = -\frac{1}{30},\ B_6 = \frac{1}{42},\ B_8 = -\frac{1}{30},$$

$$B_{10} = \frac{5}{66},\ B_{12} = -\frac{691}{2730},\ B_{14} = \frac{7}{6},\ B_{16} = -\frac{3617}{510},\ \ldots$$

Имеет место следующая теорема.

Теорема 2. Если функция на $[a, b]$ имеет непрерывную производную порядка $\nu \geqslant 1$, тогда при $x \in [a, b]$ справедливо равенство

$$f(x) = \frac{1}{h}\int_a^b f(t)\,dt + \sum_{k=1}^{\nu-1} \frac{h^{k-1}}{k!} B_k\left(\frac{x-a}{h}\right)[f^{(k-1)}(b) - f^{(k-1)}(a)] -$$

$$-\frac{h^{(\nu-1)}}{\nu!}\int_a^b f^{(\nu)}(t)\left[B_\nu^*\left(\frac{x-t}{h}\right) - B_\nu^*\left(\frac{x-a}{h}\right)\right]dt, \quad h = b - a,$$

где $B_k^*(x)$ являются периодическими функциями с периодом 1, которые для $x \in (-\infty, \infty)$ определяются формулами

$$B_n^*(x) = B_n(x), \quad \forall\, x \in [0, 1), \quad B_n^*(x+1) = B_n^*(x), \quad \forall\, x \in (-\infty, \infty).$$

Функции Бесселя. Цилиндрические функции $u = z_\nu(z)$ порядка ν определяются как частные решения *уравнения Бесселя*

$$z^2u'' + zu' + (z^2 - \nu^2)\,u = 0, \tag{28}$$

где z — комплексная переменная; ν — параметр, который может принимать любые вещественные или комплексные значения.

Уравнение (28) путем замены $y(z) = \varphi(z)\,u(z)$ может быть приведено к уравнению гипергеометрического типа, например, при $\varphi(z) = e^{-iz}z^{-\nu}$ к уравнению

$$zy'' + \tau(z)\,y' + i(2\nu + 1)\,y = 0, \tag{29}$$

где $\tau(z) = 2iz + 2\nu + 1$.

Решения уравнения (28), представляющиеся в виде интегральных представлений (Пуассона)

$$J_\nu(z) = \frac{\left(\frac{z}{2}\right)^\nu}{\sqrt{\pi}\,\Gamma\left(\nu + \frac{1}{2}\right)}\int_{-1}^{1}(1-t^2)^{\nu-\frac{1}{2}}\cos zt\,dt, \quad \operatorname{Re}\nu > -\frac{1}{2},$$

$$H_\nu^{(k)}(z) = \sqrt{\frac{2}{\pi z}}\,\frac{e^{(-1)^{(k-1)}i\left(z-\nu\frac{\pi}{2}-\frac{\pi}{4}\right)}}{\Gamma\left(\nu+\frac{1}{2}\right)}\times$$

$$\times\int_0^\infty e^{-t}t^{\nu-\frac{1}{2}}\left(1 + (-1)^{k-1}\frac{it}{2z}\right)^{\nu-\frac{1}{2}}dt, \quad k = 1, 2;\ \operatorname{Re}\nu > -\frac{1}{2},$$

называются *функциями Бесселя первого рода* и *функциями Ханкеля* соответственно.

Связь между различными цилиндрическими функциями:

$$J_\nu(z) = \frac{1}{2}[H_\nu^{(1)}(z) + H_\nu^{(2)}(z)],$$

$$Y_\nu(z) = \frac{1}{2i}[H_\nu^{(1)}(z) - H_\nu^{(2)}(z)],$$

где $Y_\nu(z)$ называются *функциями Бесселя второго рода:*

$$Y_\nu(z) = \frac{\cos \pi\nu J_\nu(z) - J_{-\nu}(z)}{\sin \pi\nu},$$

$$J_{-n}(z) = (-1)^n J_n(z).$$

Имеет место следующее разложение в ряд:

$$J_\nu(z) = \sum_{k=0}^{\infty} \frac{(-1)^k \left(\frac{z}{2}\right)^{\nu+2k}}{k!\,\Gamma(\nu+k+1)}.$$

§ 3. ЭЛЕМЕНТЫ ДИСКРЕТНОГО АНАЛИЗА

1. Разделенные разности и их свойства

В самых различных вопросах вычислительной математики важную роль играют разделенные разности, являющиеся в некотором смысле обобщением понятия производной.

Возьмем некоторую функцию $f(x)$, определенную на отрезке $[a, b]$, и совокупность точек (узлов) $x_0, x_1, \ldots, x_n$, $(x_i \neq x_j,\ i \neq j)$, $x_i \in [a, b]$. Разделенные разности первого порядка определяются равенством

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

разделенные разности второго порядка — равенством

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

Разности высших порядков определяются рекуррентно:

$$f(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}) = \frac{f(x_{i+1};\ x_{i+2};\ \ldots;\ x_{i+k}) - f(x_i;\ x_{i+1};\ \ldots;\ x_{i+k-1})}{x_{i+k} - x_i}. \quad (1)$$

Имеет место следующая лемма.

Лемма 1. Разделенные разности k-го порядка являются симметричными функциями своих аргументов и для них справедлива формула

$$f(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}) = \sum_{j=1}^{i+k} \frac{f(x_j)}{\omega'(x_j)},$$

где $\omega(x) = (x - x_i)(x - x_{i+1}) \ldots (x - x_{i+k})$.

Непосредственно из леммы вытекает ряд следствий:

1. Разделенная разность — линейный оператор, т. е.

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}) =$$
$$= \alpha_1 f_1(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}) + \alpha_2 f_2(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}).$$

2. Если x_i и t_i связаны линейной зависимостью

$$x_i = \alpha t_i + \beta, \qquad i = 0, 1, \ldots, n,$$

то

$$f(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}) = \frac{1}{\alpha^k} g(t_i;\ t_{i+1};\ \ldots;\ t_{i+k}),$$

где $g(t) = f(\alpha t + \beta)$; $t_j = \dfrac{x_j - \beta}{\alpha}$.

3. Разделенная разность k-го порядка от многочлена $f(x) = x^n$ является однородным многочленом относительно своих аргументов степени $n - k$; при $k = n$ она равна 1, при $k > n$ равна 0.

Разделенные разности удобно располагать в виде такой таблицы:

x_0	$f(x_0)$			
x_1	$f(x_1)$	$f(x_0;\ x_1)$	$f(x_0;\ x_1;\ x_2)$	
x_2	$f(x_2)$	$f(x_1;\ x_2)$	$f(x_1;\ x_2;\ x_3)$	$f(x_0;\ x_1;\ x_2;\ x_3)$
x_3	$f(x_3)$	$f(x_2;\ x_3)$	$f(x_2;\ x_3;\ x_4)$	$f(x_1;\ x_2;\ x_3;\ x_4)$
x_4	$f(x_4)$	$f(x_3;\ x_4)$		

На основании следствий 1,3 заключаем, что разделенные разности порядка n от многочлена n-й степени — постоянные, а разности порядка, больше n — равны 0.

Приведенное замечание дает возможность обнаружить ошибки в таблицах многочленов или близких к ним функций.

Определим разделенные разности с кратными узлами с помощью следующей формулы:

$$f(\underbrace{x_1;\ x_1;\ \ldots;\ x_1}_{k_1+1\text{ раз}};\ \underbrace{x_2;\ x_2;\ \ldots;\ x_2}_{k_2+1\text{ раз}};\ \ldots;\ \underbrace{x_n;\ x_n;\ \ldots;\ x_n}_{k_n+1\text{ раз}}) =$$

$$= \lim_{\varepsilon\to 0} f(x_1;\ x_1+\varepsilon;\ \ldots;\ x_1+k_1\varepsilon;\ \ldots;\ x_n;\ x_n+\varepsilon;\ \ldots;\ x_n+k_n\varepsilon), \qquad (2)$$

где для существования такого предела предполагается, что функция $f(x)$ обладает непрерывными производными до порядка $m = \sum_{i=1}^{n} k_i$ включительно. Имеет место следующее соотношение:

$$f(\underbrace{x_1;\ x_1;\ \ldots;\ x_1}_{p+1\text{ раз}}) = \frac{f^{(p)}(x_1)}{p!}. \qquad (3)$$

Разделенные разности с повторяющимися аргументами (кратными узлами) любого порядка могут быть выражены через разности низших порядков, а именно:

$$f(\underbrace{x_1;\ x_1;\ \ldots;\ x_1}_{k_1+1\text{ раз}};\ \ldots;\ \underbrace{x_n;\ x_n;\ \ldots;\ x_n}_{k_n+1\text{ раз}}) =$$

$$= \frac{1}{x_n - x_0}\,[f(\underbrace{x_1;\ x_1;\ \ldots;\ x_1}_{k_1\text{ раз}};\ \ldots;\ \underbrace{x_n;\ x_n;\ \ldots;\ x_n}_{k_n+1\text{ раз}}) -$$

$$- f(\underbrace{x_1;\ x_1;\ \ldots;\ x_1}_{k_1+1\text{ раз}};\ \ldots;\ \underbrace{x_n;\ x_n;\ \ldots;\ x_n}_{k_n\text{ раз}})]. \qquad (4)$$

Формула (4) играет ту же роль, что и формула (1).

2. Конечные разности и их свойства

Пусть для функции $y = f(x)$, заданной на отрезке $[a, b]$, известны ее значения в равноотстоящих узлах, $x_i = x_0 + ih \in [a, b]$ $(i = 0, 1, \ldots, n)$, которые мы обозначим через $y_i = f(x_i)$. Разности $y_{i+1} - y_i$ называют разностями первого порядка и для

них употребляются обозначения Δy_i — разность вперед; ∇y_{i+1} — назад; $\delta y_{i+\frac{1}{2}} = y'_{i+\frac{1}{2}}$ — центральная, т. е.

$$y_{i+1} - y_i = \Delta y_i = \nabla y_{i+1} = \delta y_{i+\frac{1}{2}}.$$

Если обозначить через ω какой-либо из операторов Δ, ∇, δ, то разности высших порядков образуются при помощи рекуррентной формулы

$$\omega^m y_i = \omega(\omega^{m-1} y_i).$$

Конечные разности удобно располагать в виде следующей таблицы (для определенности взяты разности вперед):

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
x_0	y_0			
x_1	y_1	Δy_0	$\Delta^2 y_0$	
x_2	y_2	Δy_1	$\Delta^2 y_1$	$\Delta^3 y_0$
x_3	y_3	Δy_2		

На практике приходится контролировать вычислительный процесс. Контроль очень просто осуществляется при составлении таблицы разностей. Очевидно,

$$\sum_{i=0}^{n-1} \Delta y_i = y_n - y_0, \quad \sum_{i=0}^{n-2} \Delta^2 y_i = \Delta y_{n-1} - \Delta y_0, \ \ldots,$$

т. е. сумма чисел каждого столбца таблицы конечных разностей равна разности крайних чисел предыдущего столбца. Целесообразно поэтому ввести в дополнение к таблице еще две строки: строку Σ, состоящую из сумм чисел в столбцах, и строку s, состоящую из разностей крайних чисел в столбцах.

Лемма 2. Разности вперед m-го порядка выражаются через значения функции по формуле

$$\Delta^m y_i = \sum_{j=0}^{m} (-1)^j C_m^j y_{i+m-j}. \tag{5}$$

Аналогичные формулы имеют место для операторов Δ, δ.

Лемма 3. При $x_i = x_0 + ih$ справедлива формула

$$f(x_i;\ x_{i+1};\ \ldots;\ x_{i+k}) = \frac{\Delta^k y_i}{k!\, h^k}. \tag{6}$$

На основании последней леммы можно заключить, что для конечных разностей будут справедливы следствия 1—3, имеющие место для разделенных разностей.

3. Конечно-разностные уравнения

Уравнение вида

$$y(x+k) + p_1(x)\, y(x+k+1) + \cdots + p_k(x)\, y(x) = 0 \tag{7}$$

называется однородным конечно-разностным уравнением k-го порядка; здесь x приобретает значения 0, 1, 2, ...

Теорема 1. Если $y_m(x)$ $(m = 1, 2, \ldots, k)$ — решение уравнения, причем определитель

$$D[y(0), \ldots, y_k(0)] = \begin{vmatrix} y_1(0) & y_2(0) & \ldots & y_k(0) \\ \cdots & \cdots & \cdots & \cdots \\ y_1(k-1) & y_2(k-1) & \ldots & y_k(k-1) \end{vmatrix} \neq 0, \tag{8}$$

то общее решение уравнения (7) имеет вид

$$y(x) = \sum_{m=1}^{k} c_m y_m(x), \tag{9}$$

где c_m $(m = 1, 2, \ldots, k)$ — произвольные постоянные.

Теорема 2. Общее решение линейного неоднородного уравнения

$$\sum_{i=0}^{k} p_i(x)\, y(x+k-i) = Q(x), \qquad p_0 \equiv 1 \tag{10}$$

представляется в виде суммы его частного решения $y^*(x)$ и общего решения (9) однородного уравнения (7)

$$y(x) = y^*(x) + \sum_{m=1}^{k} c_m y_m(x),$$

причем должно выполняться условие (8).

Условия, при которых (8) не имеет места, заключены в следующей теореме.

Теорема 3. Если функции $y_m(x)$ $(m = 1, 2, \ldots, k)$ — линейно-зависимы, то

$$D[y_1(x), \ldots, y_k(x)] = \begin{vmatrix} y_1(x) & y_2(x) & \ldots & y_k(x) \\ \cdot & \cdot & \cdot & \cdot \\ y_1(x+k-1) & y_2(x+k-1) & \ldots & y_k(x+k-1) \end{vmatrix} = 0,$$

$\forall x = 0, 1, \ldots$. Если же $y_m(x)$ $(m = 1, 2, \ldots, k)$ — линейно-независимы, то определитель $D[y_1(x), \ldots, y_m(x)]$ не может равняться нулю $\forall x = 0, 1, \ldots$.

Если $y_m(x)$ $(m = 1, 2, \ldots, k)$ — линейно-независимые решения уравнения (7), коэффициенты которого определены и конечны для всякого целого $x \geqslant 0$ и $p_k(x) \neq 0$, $\forall x \geqslant 0$, то частное решение $y^*(x)$ неоднородного уравнения (10) может быть легко получено методом вариации постоянных и имеет вид

$$y^*(x) = \sum_{i=0}^{x-1} \frac{\begin{vmatrix} y_1(i+1) & y_2(i+1) & \ldots & y_k(i+1) \\ \cdot & \cdot & \cdot & \cdot \\ y_1(i+k-1) & y_2(i+k-1) & \ldots & y_k(i+k-1) \\ y_1(x) & y_2(x) & \ldots & y_k(x) \end{vmatrix} Q(i)}{D[y_1(i+1), y_2(i+1), \ldots, y_k(i+1)]} \tag{11}$$

Пусть коэффициенты уравнения (7) являются постоянными, т. е. $p_i(x) = a_i$ $(i = 1, 2, \ldots, k)$, тогда система линейно-независимых решений $y_m(x)$ $(m = 1, 2, \ldots, k)$ уравнения (7) строится в явном виде и, в зависимости от корней характеристического уравнения

$$\lambda^k + a_1\lambda^{k-1} + \cdots + a_k = 0, \tag{12}$$

они могут быть представлены следующим образом:

а) случай различных корней, тогда

$$y_m(x) = \lambda_m^x, \qquad m = 1, 2, \ldots, k,$$

где λ_m — корни уравнения (12);

б) случай кратных корней. Пусть для определенности λ_1 являются корнем уравнения (12) кратности s, тогда группа линейно-независимых решений, соответствующих этому корню, будет иметь вид

$$y_j(x) = x^{j-1}\lambda_1^x, \qquad j = 1, 2, \ldots, s.$$

4. Некоторые разностные формулы

Основные тождества. Для любых двух функций $u(x)$, $v(x)$, заданных на произвольной равномерной сетке Ω_h, имеют место следующие формулы дифференцирования произведения:

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

Из этих тождеств выводятся *формулы суммирования по частям*

$$(u, v_x) = -(u_{\bar{x}}, v] + u_N v_N - u_0 v_1, \tag{2}$$

$$(u, v_{\bar{x}}) = -[u_x, v) + u_N v_{N-1} - u_0 v_0. \tag{3}$$

Частным случаем формулы (2) является *первая разностная формула Грина*

$$(u, (av_{\bar{x}})_x) = -(u_{\bar{x}}, av_{\bar{x}}] + a_N u_N v_{\bar{x},N} - a_1 u_0 v_{\bar{x},1}. \tag{4}$$

Если поменять в (4) u и v местами, а затем вычесть полученное выражение из (4), то получим *вторую разностную формулу Грина*

$$(u, (av_{\bar{x}})_x) - ((au_{\bar{x}})_x, v) = a_N (uv_{\bar{x}} - vu_{\bar{x}})_N - a_1 (uv_x - vu_x)_0. \tag{5}$$

Разностная функция Грина. Одним из способов решения конечно-разностных краевых задач вида

$$\Lambda v = (av_{\bar{x}})_x - dv = -f(x), \quad 0 < x = ih < 1, \tag{6}$$

$$v(0) = v(1) = 0, \quad a \geqslant c_1 > 0, \quad d \geqslant 0$$

является использование разностной функции Грина.

Функция $G(x, \xi)$, $x = x_i = ih$, $\xi = \xi_j = jh$ называется разностной функцией Грина для задачи (6), если она удовлетворяет условиям:

$$\begin{gathered} \Lambda G(x, \xi) = -\frac{\delta(x, \xi)}{h}, \quad x, \xi \in \Omega_h; \\ G(0, \xi) = G(1, \xi) = 0, \quad \xi \in \Omega_h, \end{gathered} \tag{7}$$

где $\delta(x, \xi)$ — символ Кронекера.

Если функция, удовлетворяющая условиям (7), найдена, то решить краевую задачу (6) можно при помощи формулы

$$v(x) = (G(x, \xi), f(\xi)). \tag{8}$$

Пусть $\alpha(x) = \alpha(x, h)$ и $\beta(x) = \beta(x, h)$ — решения следующих конечно-разностных задач Коши:

$$\begin{gathered} \Lambda\alpha = (a\alpha_{\bar{x}})_x - d\alpha = 0, \quad x \in \Omega_h, \quad \alpha(0) = 0, \quad a_1 \alpha_{\bar{x},1} = 1; \\ \Lambda\beta = (a\beta_{\bar{x}})_x - d\beta = 0, \quad x \in \Omega_h, \quad \beta(1) = 0, \quad -a_N \beta_{\bar{x},N} = 1, \end{gathered} \tag{9}$$

тогда для функции $G(x, \xi)$ имеет место представление

$$G(x, \xi) = \frac{\alpha(|x, \xi|)\, \beta((x, \xi))}{\alpha(1)}, \tag{10}$$

где

$$(x, \xi) = \frac{x + \xi + |x - \xi|}{2} = \max(x, \xi), \quad |x, \xi| = \frac{x + \xi - |x - \xi|}{2} = \min(x, \xi). \tag{11}$$

Простейшая задача на собственные значения. Рассмотрим задачу нахождения сеточных функций $v(x) \not\equiv 0$ и чисел λ, которые удовлетворяют уравнению

$$\begin{gathered} \Lambda v + \lambda v = -(\mathring{A}v)_i + \lambda v_i = v_{\bar{x}x,i} + \lambda v_i = 0, \\ i = 1, 2, \ldots, N-1, \quad hN = l, \quad v_0 = v_N = 0. \end{gathered} \tag{12}$$

Решением задачи (12) являются собственные функции

$$v_k(x_i) = \sqrt{\frac{2}{l}} \sin \frac{\pi k x_i}{l}, \qquad k = 1, 2, \ldots, N-1, \tag{13}$$

совпадающие со следами на сетке первых $N-1$ собственных функций дифференциальной задачи

$$u''(x)+\lambda u(x)=0, \quad x\in(0, l), \quad u(0)=u(l)=0. \tag{14}$$

Собственными значениями задачи (12) являются выражения

$$\lambda_k^h=\frac{4}{h^2}\sin^2\frac{\pi kh}{2l}, \qquad k=1, 2, \ldots, N-1, \tag{15}$$

которые уже не совпадают с первыми $N-1$ собственными значениями задачи (14)

$$\lambda_k=\left(\frac{\pi k}{l}\right)^2, \qquad k=1, 2, \ldots \tag{16}$$

Функции $v^{(k)}(x)$ образуют ортонормальную систему:

$$(v^{(k)}, v^{(m)})=\delta_{k,m}.$$

Первые разностные производные от собственных функций, имеющие вид

$$(v^{(k)}(x))_{\bar{x}}=\sqrt{\frac{2\lambda_k^h}{l}}\cos\frac{k\pi(x-0{,}5h)}{l}, \tag{17}$$

ортогональны в смысле скалярного произведения $(\cdot, \cdot]$ и, кроме того,

$$\|v_{\bar{x}}^{(k)}]\,|^2=\lambda_k^h.$$

Для собственных значений (15) справедливы следующие оценки:

$$\frac{8}{l^2}<\lambda_1^h<\lambda_2^h<\cdots<\lambda_{N-1}^h<\frac{4}{h^2} \quad \text{при } N\geqslant 2. \tag{18}$$

§ 4. МЕТОД ПРОГОНКИ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Рассмотрим систему линейных алгебраических уравнений с трехдиагональной матрицей

$$\begin{aligned} -c_1x_1+b_1x_2&=f_1\\ d_ix_{i-1}-c_ix_i+b_ix_{i+1}&=f_i \qquad (i=\overline{2, n-1}).\\ d_nx_{n-1}-c_nx_n&=f_n \end{aligned} \tag{1}$$

Если ввести матричные операторы сдвига

$$I^+(x_i)=(x_{i+1}), \quad I^-(x_i)=x_{i-1}, \tag{2}$$

то систему уравнений (1) можно записать в виде

$$(DI^- - C+BI^+)X=F, \tag{3}$$

где D, C, B — диагональные матрицы соответственно с элементами

$$d_i, \ c_i, \ b_i; \quad F=(f_i)_{i=\overline{1,n}}; \quad X=(x_i)_{i=\overline{1,n}}.$$

Будем искать решение системы (3) в виде

$$I^-X=MX+W, \tag{4}$$

где $M=(m_i)_{i=\overline{1,n}}$ — диагональная матрица, $W=(w_i)_{i=\overline{1,n}}$ — вектор-столбец. Умножим соотношение (4) на диагональную матрицу D и вычтем из уравнения (2)

$$\begin{aligned} (-C+DM)X&=-BI^+X-DW+F,\\ X&=-(DM-C)^{-1}BI^+X+(DM-C)^{-1}(F-DW). \end{aligned} \tag{5}$$

Последнее уравнение по форме совпадает с (7). Для того чтобы оно было совместно с задачей (3), необходимо, чтобы выполнялись почленные равенства:

$$I^{+}M = (C - DM)^{-1} B; \quad I^{+}W = (DM - C)^{-1} (F - DW). \tag{6}$$

Записывая (4) и (6) в координатной форме, получим следующие рекуррентные соотношения для прогоночных коэффициентов:

$$m_{i+1} = \frac{b_i}{c_i - d_i m_i} \quad (i = \overline{2, n-1}), \quad w_{i+1} = \frac{f_i - d_i w_i}{d_i m_i - c_i} \quad (i = \overline{2, n}),$$
$$m_2 = \frac{b_1}{c_1}, \quad w_2 = -\frac{f_1}{c_1}; \tag{7}$$

$$x_n = w_{n+1},$$
$$x_{i-1} = m_i x_i + w_i \quad (i = \overline{n, 2}). \tag{8}$$

По формулам (7) осуществляется прямая прогонка и находятся значения m_i, w_i $(i = \overline{2, n})$, а по формулам (8) находят x_i. Формулы (7), (8) называют формулами правой прогонки, так как x_i находятся последовательно, начиная с x_n. Аналогично можно построить формулы левой прогонки (см. (8′)).

Схема метода прогонки очень проста и экономична; для решения системы порядка $n \times n$ требует $O(n)$ арифметических операций.

Очевидно, прогоночные формулы (8) будут обладать устойчивым счетом, если $|m_i| < 1$. Условия

$$d_i > 0, \; b_i > 0, \; c_i \geqslant d_i + b_i \quad (i = \overline{2, n-1}); \quad \left|\frac{b_1}{c_1}\right|, \left|\frac{d_n}{c_n}\right| < 1$$

обеспечивают устойчивость прогоночных формул (8).

Схема матричной прогонки. Если в системе (1) d_i, c_i, b_i — считать блочными матрицами порядка $k_i \times k_i$, а f_i — блочным вектором порядка $k_i \times k_i$, то можно построить аналогично формулам (7), (8) формулы матричной прогонки [24]. Запишем формулы левой матричной прогонки

$$p_{n-1} = c_n^{-1} d_n, \quad p_{i-1} = (c_i - b_i d_i)^{-1} d_i \quad (i = \overline{n-1, 2}),$$
$$y_{n-1} = -c_n^{-1} f_n, \quad y_{i-1} = (b_i p_i - c_i)^{-1} (f_i - b_i y_i) \quad (i = \overline{n-1, 1}), \tag{8′}$$
$$x_1 = y_0, \quad x_{i+1} = p_i x_i + y_i \quad (i = \overline{2, n-1}).$$

Здесь предполагается, что блочные матрицы $(b_i p_i - c_i)^{-1}$ $(i = \overline{1, n})$ — невырожденные.

Отметим, что к решению систем вида (1) приводят многие задачи, возникающие при аппроксимации краевых задач математической физики разностными методами.

Схема циклической прогонки. Метод прогонки может быть применен для решения систем линейных алгебраических уравнений с трехдиагональной матрицей и отличными от нуля элементами в верхнем правом и нижнем левом углах матрицы, т. е. для систем вида

$$AX = F, \tag{9}$$

где

$$A = \begin{pmatrix} -c_1 & b_1 & 0 & \dots & 0 & d_1 \\ d_2 & -c_2 & b_2 & 0 \dots & 0 & 0 \\ 0 & d_3 & -c_3 & b_3 \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b_n & 0 & \dots & 0 & d_n & -c_n \end{pmatrix}. \tag{10}$$

Запишем систему (9), (10) в виде

$$A_{n-1} X^{n-1} + u^{n-1} x_n = F^{n-1},$$
$$(v^{n-1}, X^{n-1}) - c_n x_n = f_n, \tag{11}$$

где

$$A_{n-1} = \begin{pmatrix} -c_1 & b_1 & 0 \ldots\ldots & 0 \\ d_2 & -c_2 & b_2 \ldots\ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 \ldots\ldots & & d_{n-1} & -c_{n-1} \end{pmatrix}; \tag{12}$$

$$u^{n-1} = (d_1 \quad 0 \ldots 0 \quad b_{n-1})'; \quad v^{n-1} = (b_n \quad 0 \ldots 0 \quad d_n)';$$

$$F^{n-1} = (f_i)_{i=\overline{1,n-1}}; \quad X^{n-1} = (x_i)_{i=\overline{1,n-1}}.$$

Пусть векторы $Z^{n-1} = (z_i)_{i=\overline{1,n-1}}$ и $Y^{n-1} = (y_i)_{i=\overline{1,n-1}}$ — решения задач

$$A_{n-1}Z^{n-1} = F^{n-1}, \quad A_{n-1}Y^{n-1} = -u^{n-1}. \tag{13}$$

Решение уравнения (13) будем искать в виде

$$X^{n-1} = Z^{n-1} + x_n Y^{n-1}. \tag{14}$$

Подставим (14) в (11) и найдем:

$$x_n = \frac{f_n - (v^{n-1}, Z^{n-1})}{(v^{n-1}, Y^{n-1}) - c_n} = \frac{f_n - b_n z_1 - d_n z_{n-1}}{d_n y_{n-1} + b_n y_1 - c_n}. \tag{15}$$

Для решения задачи (13) могут быть применены прогоночные формулы вида (7), (8). Тогда алгоритм циклической прогонки будет иметь вид

$$m_2 = \frac{b_1}{c_1}, \quad w_2 = -\frac{f_1}{c_1}, \quad \gamma_2 = \frac{d_1}{c_1};$$

$$m_{i+1} = \frac{b_i}{c_i - m_i d_i}, \quad w_{i+1} = \frac{f_i - d_i w_i}{d_i m_i - c_i}, \tag{16}$$

$$\gamma_{i+1} = \frac{d_i \gamma_i}{c_i - m_i d_i}, \qquad i = \overline{2, n-1};$$

$$z_n = w_n, \quad y_n = \gamma_n + m_n;$$

$$z_{i-1} = m_i z_i + w_i, \quad y_{i-1} = m_i y_i + \gamma_i \qquad (i = \overline{n, 2});$$

$$x_n = \frac{f_n - b_n z_1 - d_n z_{n-1}}{d_n y_{n-1} + b_n y_1 - c_n}, \quad x_i = z_i + x_n y_i \qquad (i = \overline{n-1, 1}).$$

Количество операций в методе циклической прогонки будет иметь порядок $O(n)$. Условия $d_i > 0$, $b_i > 0$, $c_i > a_i + b_i$ обеспечивают устойчивость прогоночных формул (16).

К решению систем вида (9) приводят разностные линейные задачи с периодическими решениями.

СПИСОК ЛИТЕРАТУРЫ

Обязательная

1. Березин И. С., Жидков Н. П. Методы вычислений, т. 1. М., «Наука», 1966, т. 2. М., Физматгиз, 1963.
2. Лоран П.-Ж. Аппроксимация и оптимизация. М., «Мир», 1975.
3. Михлин С. Г. Вариационные методы в математической физике. М., Гостехиздат, 1957.
4. Самарский А. А. Введение в теорию разностных схем. М., «Наука», 1971.
5. Самарский А. А., Гулин А. В. Устойчивость разностных схем. М., «Наука», 1973.

Дополнительная

6. Абрамов А. А., Андреев В. Б. О применении метода прогонки к нахождению периодических решений дифференциальных и разностных уравнений.— «Журнал вычислительной математики и математической физики», 1963, т. 3, № 2, с. 377—381.
7. Аткинсон Ф. Дискретные и непрерывные граничные задачи. М., «Мир», 1968.
8. Бабенко К. И. Некоторые вопросы теории приближенного задания и вычисления функций. М., 1970. (Препринт Ин-та приклад. матем. АН СССР).
9. Бабушка И., Витасек Э., Прагер М. Численные процессы решения дифференциальных уравнений. М., «Мир», 1969.
10. Бахвалов Н. С. Численные методы, т. 1. М., «Наука», 1973.
11. Бейтмен Г. А., Эрдейи А. Высшие трансцендентные функции, т. 1. М., «Наука», 1973, т. 2. М., «Наука», 1974.
12. Белоцерковский О. М., Чушкин П. И. Численный метод интегральных соотношений.— «Журнал вычислительной математики и математической физики», 1962, т. 2, № 5, с. 731—760.
13. Бирман М. Ш., Виленкин Н. Я. и др. Функциональный анализ. М., «Наука», 1972.
14. Бублик Б. Н. Численное решение задач динамики пластин и оболочек. Киев, Изд-во Киев. ун-та, 1969.
15. Вазов В., Форсайт Дж. Разностные методы решения дифференциальных уравнений в частных производных. М., Изд-во иностр. лит., 1963.
16. Валиуллин А. Н., Яненко Н. Н. Экономичные разностные схемы повышенной точности для полигармонического уравнения.— «Известия Сибирского отделения АН СССР», 1967, вып. 3, № 13, серия техн. наук.
17. Валиуллин А. Н., Паасонен В. И. Экономичные разностные схемы повышенного порядка точности для многомерного уравнения колебаний.— Информ. бюл. «Численные методы механики сплошной среды», т. 1, 1970, № 1, Новосибирск.
18. Валиуллин А. Н. Схемы повышенной точности для задач математической физики. Новосибирск, 1973. (Новосиб. ун-т).

19. Волков Е. А. О применении интерполяционного многочлена Лагранжа при решении методом сеток задачи Дирихле для уравнения Пуассона.— «Журнал вычислительной техники и математической физики», 1964, т. 4, № 3, с. 466—473.
20. Воробьев Ю. В. Случайный итерационный процесс.— «Журнал вычислительной математики и математической физики», 1964, т. 4, № 6, 1965, т. 5, № 5.
21. Гаврин М. К. Лекции по методам вычислений. М., «Наука», 1971.
22. Глущенко А. А. Некоторые пространственные задачи теории фильтрации. Киев, Изд-во Киев. ун-та, 1970.
23. Годунов С. К. Метод ортогональной прогонки для решения систем разностных уравнений.— «Журнал вычислительной математики и математической физики», 1962, т. 2, № 6, с. 972—982.
24. Годунов С. К., Рябенький В. С. Разностные схемы. М., «Наука», 1973.
25. Горбунов А. Д., Шахов Ю. А. О приближенном решении задачи Коши для обыкновенных дифференциальных уравнений с наперед заданным числом верных знаков. ч. 1,— «Журнал вычислительной математики и математической физики», т. 3, 1963, № 2, с. 239—253, ч. 2.— «Журнал вычислительной математики и математической физики», т. 6, 1964, № 3, с. 426—433.
26. Дородницын А. А. Метод интегральных соотношений для численных решений уравнений в частных производных. («Труды конференции по вычислительной технике»). М., Изд-во АН СССР, 1958.
27. Дьяконов Е. Г. Итерационные методы решения разностных аналогов краевых задач для уравнений эллиптического типа. Материалы Международной летней школы по численным методам. Киев, 1966. (Ин-т кибернетики АН УССР).
28. Дьяконов Е. Г. Разностные методы решения краевых задач. М., вып. 1, 1971; вып. 2, 1972 (Москов. ун-т).
29. Заворин А. Н., Хесина И. Я. О некоторых численных методах решения жестких систем обыкновенных дифференциальных уравнений.— «Журнал вычислительной математики и математической физики», т. 13, 1973, № 1, с. 71—79.
30. Ильин В. П. Разностные методы решения эллиптических уравнений. Новосибирск, 1970 (Новосиб. ун-т).
31. Канторович Л. В., Крылов В. И. Приближенные методы высшего анализа. М., Физматгиз, 1962.
32. Коллатц Л. Численные методы решения дифференциальных уравнений. М., Изд-во иностр. лит., 1953.
33. Коллатц Л. Функциональный анализ и вычислительная математика. М., «Мир», 1969.
34. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. М., «Наука», 1968.
35. Коробов Н. М. Теоретико-числовые методы в приближенном анализе. М., Физматгиз, 1963.
36. Костюкович Е. Х. О сходимости метода прямых.— Докл. АН СССР, т. 118. 1958, № 3.
37. Красносельский М. А., Вайникко Г. М. и др. Приближенное решение операторных уравнений. М., «Наука», 1969.
38. Красносельский М. А., Крейн С. Г. Итеративный процесс с минимальными невязками. Матем. сб., т. 31, 1952.
39. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы высшей математики, т. 1. Минск, «Вышэйшая школа», 1972; т. 2. Минск, «Вышэйшая школа», 1975.
40. Лебедев В. И. Об оптимизации в итерационных методах.— В сб.: Математическое обеспечение ЭЦВМ. Киев, 1972, с. 109—135 (Ин-т кибернетики АН УССР).
41. Лебедев В. И., Финогенов С. А. О порядке выбора итерационных параметров в чебышевском циклическом итерационном методе.— «Журнал вычислительной математики и математической физики», 1971, т. 11, № 2.
42. Лебедев В. И., Финогенов С. А. Об одном алгоритме выбора параметров чебышевских циклических методов.— В сб.: Вычислительные методы линейной алгебры. Под ред. акад. Г. И. Марчука. Новосибирск, 1972.
43. Люстерник Л. А., Соболев В. И. Элементы функционального анализа. М., «Наука», 1975.

44. Ляшенко И. Н. Задачи на собственные значения для уравнений второго порядка в частных конечных разностях. К., Изд-во Киев. ун-та, 1970.
45. Ляшко И. И. Решение фильтрационных задач методом суммарных представлений. Киев, Изд-во Киев. ун-та, 1963.
46. Ляшко И. И., Великоиваненко И. М. Численно-аналитическое решение краевых задач теории фильтрации. Киев, «Наукова думка», 1973.
47. Ляшко И. И., Макаров В. Л. Об основах метода суммарных представлений.— В сб.: Математическое обеспечение ЭЦВМ. Киев, 1972, с. 64—79. (Ин-т кибернетики АН УССР).
48. Макаров В. Л. Разностные схемы с точными и явными спектрами, ч. 1, Киев, 1974. (Препринт 74—12 Ин-та матем. АН УССР), ч. 2, Киев, 1974. (Препринт 74—13 Ин-та матем. АН УССР).
49. Марчук Г. И. Методы вычислительной математики. М., «Наука», 1973.
50. Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов. М., Атомиздат, 1971.
51. Математическое обеспечение ЭЦВМ. (Сборник статей). Киев, 1972. (Ин-т кибернетики АН УССР).
52. Микеладзе Ш. Е. К вопросу численного интегрирования дифференциальных уравнений с частными производными при помощи сеток. 1940, т. 1, № 4, с. 249—254. (Груз. фил. АН СССР).
53. Михлин С. Г. Численная реализация вариационных методов. М., «Наука», 1966.
54. Молчанов И. Н., Николенко Л. Д. Метод конечных элементов и его применение для решения некоторых одномерных краевых задач. Киев, 1976. (Препринт Ин-та кибернетики АН УССР).
55. Николаев Е. С., Самарский А. А. О вычислительной устойчивости двуслойных и трехслойных итерационных схем.— «Журнал вычислительной математики и математической физики», 1972, т. 12, № 5.
56. Островский А. М. Решение уравнений и систем уравнений. Изд-во иностр. лит. М., 1963.
57. Оганесян Л. А., Ривкинд В. Я., Руховец Л. А. Вариационно-разностные методы решения эллиптических уравнений.— В сб.: Дифференциальные уравнения и их применение. Вильнюс, Изд-во ин-та физики и математики АН ЛитССР, вып. 5, ч. 1, 1973, ч. 2, вып. 8, 1974.
58. Ортега Дж., Рейнболот В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. М., «Мир», 1975.
59. Положий Г. Н. Численное решение двумерных и трехмерных краевых задач математической физики и функции дискретного аргумента. Киев, Изд-во Киев. ун-та, 1962.
60. Положий Г. Н., Макаров В. Л. Обобщение формул суммарных представлений осесимметричного потенциала и специальные функции дискретного аргумента первого и второго рода. ч. 1.— В сб.: Вычислительная и прикладная математика. Киев, 1969, вып. 9.
61. Положий Г. Н., Скоробогатько А. А. Об одном классе формул суммарных представлений.— В сб.: Вычислительная математика. Киев, 1965, вып. 1.
62. Положий Г. Н., Пахарева Н. А., Бондаренко П. С. Математический практикум. М., Физматгиз, 1960.
63. Рихтмайер Р., Мортон К. Разностные методы решения краевых задач. М., «Мир», 1972.
64. Самарский А. А. Об одном экономичном разностном методе решения многомерного параболического уравнения в произвольной области.— «Журнал вычислительной математики и математической физики», 1962, т. 2, № 5, с. 787—811.
65. Самарский А. А. Схемы повышенного порядка точности для многомерного уравнения теплопроводности.— «Журнал вычислительной математики и математической физики», 1963, т. 3, № 15, с. 812 — 840.
66. Самарский А. А. Об одной разностной схеме повышенного порядка точности для уравнения теплопроводности с несколькими пространственными переменными.— «Журнал вычислительной математики и математической физики», 1964, т. 4, № 1, с. 161—165.

67. Самарский А. А. Аддитивные схемы Тезисы докладов, Секц. 14, М., 1966, с. 46—47.
68. Самарский А. А. Итерационные разностные схемы.— В сб.: Математическое обеспечение ЭЦВМ. Киев, 1972,с. 47—59. (Ин-т кибернетики АН УССР).
69. Самарский А. А. Некоторые вопросы теории разностных схем.— В сб.: Математическое обеспечение ЭЦВМ. Киев, 1972, с. 18—46. (Ин-т кибернетики АН УССР).
70. Самарский А. А. Итерационные методы для сеточных уравнений.— «Труды, посвященные 60-летию акад. Л. Илиева». София, 1975, с. 153—164.
71. Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений. М., «Наука», 1976.
72. Самарский А. А., Попов Ю. П. Разностные схемы газовой динамики. М., «Наука», 1975.
73. Саульев В. К. Интегрирование уравнений параболического типа методом сеток. М., Физматгиз, 1960.
74. Сеге Г. Ортогональные многочлены. М., Физматгиз, 1968.
75. Тихонов А. Н., Самарский А. А. Об однородных разностных схемах.— «Журнал вычислительной математики и математической физики», 1961, т. 1, № 3, с. 425—440.
76. Тихонов А. Н., Самарский А. А. Об однородных разностных схемах высокого порядка точности на неравномерных сетках.— «Журнал вычислительной математики и математической физики», 1963, т. 3, № 1.
77. Тихонов А. Н., Горбунов А. Д. Оценки погрешности метода типа Рунге — Кутта и выбор оптимальных сеток.— «Журнал вычислительной математики и математической физики», 1964, т. 4, № 2, с. 232—241.
78. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. М., Физматгиз, 1960.
79. Фрязинов И. В. О разностной аппроксимации граничных условий для третьей краевой задачи.— «Журнал вычислительной математики и математической физики», 1964, т. 4, № 6, с. 1106—1112.
80. Фрязинов И. В. Априорные оценки для одного семейства экономичных схем.— «Журнал вычислительной математики и математической физики», 1969, т. 9, № 3, с. 595—604.
81. Хемминг Р. В. Численные методы. М., «Наука», 1972.
82. Шаманский В. Е. Методы численного решения краевых задач на ЭЦВМ. т. 1. Киев, Изд-во АН УССР, 1963.
83. Шаманский В. Е. Методы численного решения краевых задач на ЭЦВМ. т. 2. Киев, «Наукова думка», 1966.
84. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. Новосибирск, Изд-во Сибирского отделения АН СССР, 1967.
85. Chu Sherwood C., Berman M. An expotential method for the solution of the systems of ordinary differential equations, Commun. of the ACM, 17, N 12, 1974, 699—702.
86. Collatz L. Bemerkungen zur Fehlerabschätzung für das Differenzverfahren bei partiellen Differentialgleichungen, Z. angew. Math. Mech., 13, 1933, 56—57.
87. DuFort E. C., Frankel S. P. Stability conditions in the numerical solving of parabolic differential equations, Math. Tables and other Aids Comput., 7.43, 1953, 135—152.
88. Harlow F. H. The particle—in — cell method for numerical solution of problems in fluid dynamics, «Experimental arithmetics high speed computing and mathematics», v. 15, 1963.
89. Rosenbrock H. H. Some general implicit processes for the numerical solution of differential equations, Comp. J., v. 5, N 4, 1963, 329—330.
90. Shortley C. H., Weller H. The numerical solution of Laplase's equation, J. Appl. Phys., 9, 1938, 334—344.
91. Ullman J. L. A class of weight functions that admit Tchebycheff quadrature, The Michigan Math. J., v. 13, N 4, 1966.
92. Prüfer H. Neue Herleitung der Sturm—Liovilleshen Reihenentwicklung stetiger Funktionen, Math. Ann., 95 (1926), 499—518.
93. Varga R. Functional analysis and approximation theory in numerical analyses, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1971.

94. Zlamal M. On the finite element method, Numer. Math., 12, 5, 1968.
95. Peaceman D. W., Rachford H. H. The numerical solution of parabolic and elliptic differential equations, J. Soc. Indust. Appl. Math., N 1, 1955, 28—41.
96. Kellog U. Another alternating — direction — implicit method, J. Soc. Indust. Appl. Math., v. 11, N 4, 1963, 976—979.
97. Douglas J., Rachford H. H. On the numerical solution of heat conduction problems in two and three space variables, Trans. Amer. Soc., 82, 1956, 421—439.
98. Waschpress E. L. Iterative solution of elliptic systems and applications to the neutron diffusion equations of reactor physics, Prentice — Hall, Inc. Englewood Cliffs, N. J., 1966.

Принятые условные обозначения

$\mathbf{B}, \mathbf{B}_1, \mathbf{B}_2, \ldots$ — пространства типа Банаха

$\mathbf{H}$ — гильбертово пространство

$C_l^k(\Omega)$ — множество функций с непрерывными производными порядка l по пространственным переменным и порядка k по временной переменной

$W_2^s(\Omega)$ — пространство Соболева

A^* — оператор, сопряженный оператору A

$A_c = \frac{1}{2}(A + A^*)$ — симметрическая составляющая оператора

$A_к = \frac{1}{2}(A - A^*)$ — кососимметрическая составляющая оператора A

$\operatorname{Sp} A$ — спектр оператора A

$r(A) = \lim\limits_{n\to\infty} \| A^n \|^{\frac{1}{n}}$ — спектральный радиус линейного ограниченного оператора

$A > 0$ — положительный оператор

$A \geqslant 0$ — неотрицательный оператор

$A \geqslant \gamma^2 I$ — положительно определенный оператор, γ^2 — число

$A = (a_{ij})_{i=\overline{1,n}}^{j=\overline{1,m}}$ — матрицы порядка $n \times m$

$A' = (a_{ji})_{j=\overline{1,m}}^{i=\overline{1,n}}$ — матрица, транспонированная к матрице A

$\| u \|$ — норма элемента $u \in \mathbf{H}$

$\Omega_h = \{x_i = ih,\ h > 0,\ i = \overline{1,\ n-1}\}$ — равномерная сетка на интервале $(0,\ a)$

$u_i = u(x_i)$ — функция, заданная на Ω_h

$u_x = u_{x,i} = \frac{u_{i+1} - u_i}{h}$ — правая разностная производная

$u_{\bar{x}} = u_{\bar{x},i} = \dfrac{u_i - u_{i-1}}{h}$ — левая разностная производная

$u_{\overset{\circ}{x}} = u_{\overset{\circ}{x},i} = \dfrac{u_{i+1} - u_{i-1}}{2h}$ — центральная разностная производная

$\Lambda u = u_{x\bar{x}} = \dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$ — вторая разностная производная в точке

$\hat{\Omega}_h = \{x_i \in (0, a),\ x_i = x_{i-1} + h_i,\ \ i = \overline{1, n-1}\}$ — неравномерная сетка на интервале $(0, a)$

$h_i = x_i - x_{i-1}$ — шаг сетки $\hat{\Omega}_h$

$$\left.\begin{aligned} u_{\hat{x}} &= u_{\hat{x},i} = \frac{u_{i+1} - u_i}{\hbar}, \\ \hbar_i &= \frac{1}{2}(h_i + h_{i+1}) \end{aligned}\right\}$$ — правая разностная производная на неравномерной сетке

$\Omega_\tau = \{t_j = j\tau,\ \tau > 0,\ \ n = 0, 1, 2, \ldots\}$ — временная сетка

$u_x^\sigma = \sigma u^{j+1} + (1 - \sigma) u^j$, σ — числовой параметр

$$\left.\begin{aligned} (u, v) &= \sum_{j=0}^{n-1} u_j v_j h \\ (u, v] &= \sum_{j=1}^{n} u_j v_j h \end{aligned}\right\}$$ — скалярные произведения

$$\left.\begin{aligned} \|[u\| &= \sqrt{[u, u)} \\ \|u]\| &= \sqrt{(u, u]} \\ \|u\|_c &= \max_{x_j \in \bar{\Omega}_h} |u(x_j)| \end{aligned}\right\}$$ — нормы

$\varnothing$ — пустое множество

$\Leftrightarrow$ — тогда и только тогда

$\forall x$ — для всех x

$\exists$ — существует такое

$\rightarrow$ — следует

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Абсциссы квадратурной формулы 14
Аксиомы скалярного произведения 368
Алгоритм Ремеза 74
Аппроксимация граничного условия 156
— — — зависящего от производных 159
— — — первого рода 156
— оператора 133
— разностная 131, 140
— разностной схемы 175
— суммарная 222
Априорная оценка разностной схемы 132

Вариационные методы 121
Веса квадратурной формулы 13
Выпуклая оболочка множества 367

Гамма-функция 382
Гипергеометрический ряд 379
Гиперплоскость 371
Грина вторая разностная формула 389
— первая разностная формула 389
— разностная функция 389

Допустимая перестановка 30

Задача Коши для обыкновенных дифференциальных уравнений 253
— краевая для обыкновенных дифференциальных уравнений 235
— приближения функций 7
— теории интерполирования 5
— численного дифференцирования 14
— — интегрирования 13

Интерполирующая функция 34
Интерполяционная схема Эйткена 27
Интерполяционные сплайн-функции 79
Интерполяционный многочлен Бесселя 30
— — Джексона 44
— — Лагранжа 27
— — Стирлинга 31
— — Эрмита 27
Итерационная схема 297
— — верхней релаксации 346
— — двухслойная 304
— — — каноническая 304
— — двухсторонних приближений 350
— — двухшаговая 343, 346
— — метода Ньютона 354
— — — расщеплений 318, 319
— — модифицированного метода Ньютона 355
— — одношаговая 304
— — оптимальная линейная 312
— — последовательного приближения обратного оператора 352
— — простейшая 312
— — чебышевская трехчленная 346
— — — циклическая 314
Итерационные методы решения линейных операторных уравнений второго рода 302
— — — — — — первого рода 300
— — — нелинейных операторных уравнений 296, 353

Квадратный корень из оператора 374
Квадратурная формула Чебышева 92, 95

Квадратурные формулы замкнутого типа 42, 86
— — интерполяционного типа 14
— — наивысшей алгебраической степени точности 90
— — наилучшей степени точности 14
— — с наилучшей оценкой на классе функций 13
Классические ортогональные многочлены 377
Компакт 368
Конечно-разностные уравнения 387
Конечные разности 386
Константа Лебега 42
Координатная система 118
Корректность операторно-разностной схемы 207
Краевая задача первая 165, 169, 170, 176, 183
— — с условиями периодичности 239
— — третья 161

Лемма Бернштейна 26
Линейное многообразие 366
— — максимальное 366
Линейное подпространство 368

Матрица жесткости 194
Метод Адамса интерполяционный 280
— — экстраполяционный 279
— аналитической замены 16
— Бубнова — Галеркина 119
— вариационный построения разностных схем 190
— дополнительных функций 288
— интегральных соотношений 250
— интегро-интерполяционный построения разностных схем 185
— конечных элементов (МКЭ) 192
— линеаризации 293
— минимальных невязок 339
— минимизации 362
— многошаговый решения задачи Коши 275, 281
— моментов (Галеркина — Петрова) 119
— наименьших квадратов 63, 119
— наискорейшего спуска 336
— неопределенных коэффициентов 154
— одношаговый решения задачи Коши 255
— ортогонализации 290
— последовательных приближений 296
— прогонки 390
— продолжения решения по параметру 295, 359
— простых итераций 305
— прямых 243
— разложения решения в ряд Тейлора 254
— расщепления оператора 318
— редукции к задачам Коши 287
— Ритца 128
— Рунге — Кутта 255
— смещений 312
— сопряженных уравнений 291
— спуска по направлениям 364
— стрельбы 288
— суммарных представлений 233
— ускорения сходимости итерационных процессов 309
— Фурье исследования устойчивости разностных схем 227
— Эйлера решения задачи Коши 257
Минимизирующая последовательность 122
Многочлен наименее отклоняющийся от нуля 37
— обобщенный 6
— — интерполяционный 6
Многочлены Бернулли 383
— Лагерра 380
— Лежандра 380
— ультрасферические (Гегенбауэра) 380
— Чебышева первого и второго рода 380
— Эрмита 381
— Якоби 380
Множество выпуклое 367
— компактное 368

Наилучшее приближение 73
Направление антиградиента функцио-

нала 334
Необходимые и достаточные условия устойчивости операторно-разностных схем 209, 211
Неравенство Бесселя 56
Неравенство Коши — Буняковского 348
Норма кубическая 375
— негативная 373
— октаэдрическая 375
— оператора 369
— подчиненная матричная 375
— сферическая 375
— энергетическая 129

Область значения оператора 369
— определения оператора 369
Обобщенная теорема Чебышева 68
— формула Родрига 377
Обратное интерполирование 46
Общий вид остаточного члена интерполяционной формулы 34
Оператор бигармонический 151
— кососимметрический 374
— Лапласа 150
— нормальный 374
— обратный 370
— ограниченный линейный 369
— положительно определенный 163, 373
— — полуопределенный 373
— положительный 373
— полуограниченный снизу 373
— расщепляющийся 179
— регуляризатор 220
— самосопряженный 373
— сопряженный 373
— унитарный 374
— факторизованный 179
— эквивалентный по спектру 374
Операторно-разностные схемы 205
Операторы энергетически эквивалентные 313, 309
Оптимальная квадратурная формула на классе 111
Остаточный член формулы интерполяционной 7
— — — квадратурной 13
— — — наивысшей алгебраической степени точности 107
— — — правила трех восьмых 106
— — — Симпсона 106
— — — средних прямоугольников 105
— — — трапеций 106
— — — Эйлера 107
Оценка быстроты сходимости двухстороннего итерационного процесса 351
— — — двухшагового итерационного процесса 346, 349
— — — итерационных процессов метода расщеплений 323, 330, 331
— — — метода минимальных невязок 339
— — — — наискорейшего спуска 337
— — — — Ньютона 356, 359
— — — — смещений 312
— — — модифицированного метода Ньютона 358
— — — простейшего итерационного процесса 312
— — — чебышевского циклического итерационного процесса 312
— погрешности многошаговых методов 282
— — одношаговых методов 273

Параметры итерационные 303, 315, 317, 325
Периодическая система Чебышева 23
Плотное множество 368
Погрешность аппроксимации разностной схемы 131
— численного дифференцирования 15
Порядок погрешности аппроксимации разностной схемы решения операторного уравнения 133
Построение эмпирических формул 62
Правило трех восьмых 69
Предельная точка множества 368
Предельно плотная последовательность пространств 119
Принцип максимума 197
— регуляризации построения итерационных процессов 332
— — — разностных схем 219

— сжатых отображений 297
Проекционный метод 117
Простейшая задача на собственные значения 389
Пространство банахово 368
— гильбертово 369
— линейное 367
— метрическое 367
— нормированное 367
— полное 367
— сепарабельное 368
— унитарное 369
— эвклидово 368
Прямые методы решения разностных уравнений 232

Равенство Парсеваля 57
Равномерно сходящийся интерполяционный процесс 42
Равные операторы 370
Разделенные разности 335
Разностная производная левая 148
— — правая 148
— — смешанная 150
— — центральная 148
— схема 131
— — абсолютно устойчивая 208
— — аддитивная 222
— — безусловно устойчивая 175
— — бигармонического уравнения 169
— — консервативная 185
— — корректная 207
— — многомерного уравнения теплопроводности 176
— — неявная 171
— — однородная 170
— — повышенной точности для уравнения колебаний 183
— — — — — — Пуассона 167
— — полностью консервативная 188
— — производящая 179, 180, 184
— — ρ-устойчивая 209
— — «ромб» 175
— — Саульева 175
— — с весами 171
— — сильно устойчивая 209
— — сквозного счета 185
— — уравнения колебаний 183
— — — Пуассона 165
— — — теплопроводности 170
— — условно устойчивая 208
— — устойчивая 207
— — явная 171
Резольвента оператора 372
Резольвентное множество оператора 372

Сглаживание результатов наблюдения 61
— сплайн-функциями 83
Сетка изометрическая 136
— неравномерная 135, 136
— равномерная на отрезке 135
— — на плоскости 135
— связная 198
— треугольная 136
Сильно сходящаяся последовательность операторов 370
Система нормальных уравнений 64
— Чебышева 6
Слабо сходящаяся последовательность операторов 370
Спектр оператора 372
Спектральный радиус 372
Схема матричной прогонки 391
— циклической прогонки 391
Сходимость итерационных процессов 306, 313, 316, 329
— метода Ньютона 355
— — расщеплений с несамосопряженным оператором 331
— — — с самосопряженным оператором 331
— разностной схемы 132
Сходимость общего квадратурного процесса 114
Сходящаяся по норме последовательность операторов 370
— последовательность 368
Сходящийся интерполяционный процесс 42

Таблица коэффициентов Лагранжа 27

Тензорное произведение матриц 376
Теорема Банаха 370
— Банаха — Штейнгауза 370
— Бернштейна 98
— о возмущениях 382
— сравнения 199
— Хаара 11
— Хана — Банаха в геометрической форме 371
Тождество Кристоффеля — Дарбу 377
Трипод 20

Узлы квадратурной формулы 13
Уравнение Бесселя 384
— вырожденное гипергеометрическое 379
— гипергеометрического типа 379
— для функций Эрмита 379
— Пирсона 377
Условие Хаара 6

Формулы Адамса 279, 280
— — для интерполирования вперед 30
— — — — назад 30
— — механических квадратур 91
— Ньютона для интерполирования вперед 30
— — — — назад 30
— Ньютона — Котеса 85
— парабол (Симпсона) 88
— Рунге — Кутта двусторонние 264
— — односторонние 258
— средних прямоугольников 88
— суммарных представлений 235
— суммирования по частям 389
— трапеций 88
— Эйлера 103, 257
Фундаментальная последовательность 367
Фундаментальные обобщенные многочлены 7
Функции Бесселя 384
— — второго рода 385
— — первого рода 384
— Ханкеля 384
Функционал 371
Функционал сопряженно-линейный 371
— энергии 128

Числа Бернулли 383
Число обусловленности оператора 301
— арифметических операций итерационных методов 298, 312, 316, 326, 330, 359

Элемент наилучшего приближения 8
Энергетическое скалярное произведение 129

Ядро Фейера 44

ОГЛАВЛЕНИЕ

Предисловие 3

Часть I. Аппроксимация линейных операторов

Глава 1. Общие вопросы аппроксимации линейных операторов

§ 1. Постановка задач аппроксимации линейных операторов 5
§ 2. Единый способ построения формул интерполяционного типа для приближения линейных функционалов 16
§ 3. Системы Чебышева и их свойства 19

Глава 2. Интерполирование

§ 1. Интерполирование алгебраическими многочленами 23
§ 2. Интерполирование периодических функций 32
§ 3. Анализ погрешности интерполяционных формул 34
§ 4. Сходимость интерполяционных формул 42
§ 5. Некоторые вопросы применения интерполяционных формул . . . 46

Глава 3. Приближение функций

§ 1. Среднеквадратические приближения 53
§ 2. Равномерные приближения 66
§ 3. Интерполяционные и сглаживающие сплайн-функции 79

Глава 4. Приближенное вычисление определенных интегралов

§ 1. Формулы Ньютона — Котеса 85
§ 2. Квадратурные формулы наивысшей алгебраической степени точности 89
§ 3. Формулы Чебышева 95
§ 4. Квадратурные формулы с использованием производных от подынтегральной функции 99
§ 5. Остаточный член квадратурных формул 104
§ 6. Квадратурные формулы с наилучшей оценкой остаточного члена на классах функций 108
§ 7. Сходимость общего квадратурного процесса, не содержащего производных 114

Часть II. Приближенные методы решения операторных уравнений

Глава 5. Проекционно-вариационные методы

§ 1. Метод моментов 117
§ 2. Вариационные методы. Общие положения 121
§ 3. Метод наименьших квадратов 125
§ 4. Метод Ритца 128

Глава 6. Разностные методы решения задач математической физики

§ 1. Общие вопросы метода сеток 131
§ 2. О построении сеток, сеточных функций и согласованных норм . . . 135
§ 3. Вопросы конструирования разностных схем 139
§ 4. Исследование устойчивости разностных схем 196
§ 5. Прямые методы решения разностных уравнений 232
§ 6. Метод прямых. Метод интегральных соотношений 243

Глава 7. Численные методы решения обыкновенных дифференциальных уравнений

§ 1. Задача Коши для обыкновенных дифференциальных уравнений 253
§ 2. Численные методы решения краевых задач для обыкновенных дифференциальных уравнений 285

Глава 8. Итерационные методы решения операторных уравнений

§ 1. Метод последовательных приближений 296
§ 2. Итерационные методы решения линейных операторных уравнений 300
§ 3. Метод простых итераций решения линейных уравнений 305
§ 4. Методы ускорения сходимости процессов, основанные на использовании энергетически эквивалентных операторов 309
§ 5. Методы расщепления оператора 318
§ 6. Одношаговые итерационные методы, основанные на использовании квадратичного функционала 334
§ 7. Двухшаговые итерационные методы 343
§ 8. Итерационные методы двухсторонних приближений 350
§ 9. Метод последовательных приближений обратного оператора . . . 352
§ 10. Итерационные методы решения нелинейных уравнений 353

Приложение 366

Список литературы 393

Принятые условные обозначения 398

Предметный указатель 400

Иван Иванович Ляшко
Владимир Леонидович Макаров
Агнесса Андреевна Скоробогатько

Методы вычислений

(Численный анализ.
Методы решения задач математической физики)

Редактор Л. П. Онищенко,
Обложка художника Г. М. Балюна
Художественный редактор С. П. Духленко
Технический редактор М. С. Чабан
Корректор И. П. Кривчикова

Информ. бланк № 196

Сдано в набор 21.01.1977 г. Подписано в печать 14.07.1977 г. Формат 60×90 1/16. Бумага типографская № 2. 25,5 печ. л. 24,78 уч.-изд. л. Тираж 8000 экз. Изд. № 3053. БФ 07815. Зак. № 7-394. Цена 1 р. 23 коп.

Головное издательство издательского объединения «Вища школа». 252054, Киев-54, ул. Гоголевская, 7

Книжная фабрика «Коммунист» РПО «Полиграфкнига» Госкомиздата УССР, Харьков, ул. Энгельса, 11.

1 РУБ. 23 КОП.