3. МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ

3.1. Понятие о сеточных методах

Наряду с аналитическими методами для решения задач математической физики активно используют численные методы. Роль этих методов возросла в связи с развитием вычислительной техники. Разумеется, богатые возможности современных компьютеров — серьезный аргумент в пользу численных методов. Но это не уменьшает роль аналитических методов. Скорее, каждое из направлений математической физики заняло свою естественную нишу: аналитические методы позволяют упростить математическое описание исследуемого процесса и провести качественный анализ его течения. Численные методы позволяют получить конкретное числовое описание протекающего процесса, но по этому описанию трудно, если вообще возможно, делать какие-либо заключения о качественных особенностях исследуемого процесса.

Чтобы довести решение задачи „до числа“, нужно вписаться в особенности процесса вычислений. Решить задачу численно можно лишь с конечным числом неизвестных. Поэтому для численного решения задачи математической физики, для которой характерно изменение некоторой величины в некоторой пространственно-временной области, ее необходимо приближенно заменить некоторым дискретным аналогом. Для этого в пространственно-временной области выбирают конечное число точек. Всю совокупность точек называют ***сеткой***, а каждую отдельную точку — ***узлом сетки***. Дифференциальное уравнение, граничные и начальные условия заменяют соотношениями между значениями искомой величины в узлах сетки. Например, в задаче о распространении тепла в объеме $V \subset \mathbb{R}^3$ в течение времени $0 \leqslant t \leqslant T$ необходимо найти функцию $u(x,t)$, описывающую температуру в точке $x \in V$ в момент времени t. В четырехмерной области $V \times [0,\, T]$ выбирают набор узлов $(x_i,\, t_i)$, $i = \overline{1,\, N}$, и вместо уравнений исходной задачи формируют уравнения, связывающие значения $u_i = u(x_i, t_i)$ температуры в выбранных узлах. Таким образом, краевая задача, содержащая дифференциальное (а возможно, и интегральное) уравнение, заменяется системой в общем случае нелинейных уравнений.

Описанная процедура, называемая ***дискретизацией***, позволяет в случае успеха получить значения неизвестной функции в узлах сетки. Если эти значения достаточно точны, а узлы сетки расположены достаточно часто в пространственно-временной области, то значения неизвестной функции в других точках области можно получить с помощью методов интерполяции, позволяющих воссоздать функцию по ее значениям в некотором конечном наборе точек.

Такова общая схема. Описанная схема не единственна. Возможны другие подходы к приближенному решению краевой задачи, не связанные с выбором сетки в пространственно-временной области. Если численный метод следует предложенной схеме, т.е. основан на выборе сетки, то его относят к ***сеточным методам***. Сеточные методы различаются по способу выбора сетки и правилам формирования уравнений, связывающих значения неизвестной функции в узлах сетки.

Не следует рассчитывать на то, что в результате дискретизации задачи математической физики мы получим систему уравнений, которая даст точные значения неизвестной функции в узлах сетки. Уравнения, связывающие значения искомой величины в узлах сетки, строят на основе каких-либо общих принципов, которые позволяют приближенно заменить дифференциальное уравнение и краевые условия соотношениями между значениями величины в близлежащих узлах. Эти принципы позволяют надеяться лишь на то, что при увеличении количества узлов и при уменьшении расстояний между соседними узлами ошибка, возникающая при дискретизации, будет неограниченно уменьшаться.

Это означает, что в конкретном сеточном методе речь идет не о выборе фиксированной сетки и формировании уравнений, связывающих значения величины в узлах, а о выборе бесконечной серии $\{S^N\}$ сеток, такой, что количество узлов в этих сетках неограниченно возрастает, а расстояние между ближайшими узлами сеток стремится к нулю. Метод также должен определять формирование сеточных уравнений для каждой сетки в выбранной серии. В таком случае мы будем говорить о ***сеточной схеме***, понимая под этим набор правил формирования серии сеток и соответствующих систем сеточных уравнений. Качество сеточной схемы (и сеточного метода вообще) определяется тем, как быстро решение дискретной задачи для сетки с номером N при $N \to \infty$ сходится к решению краевой задачи (и сходится ли вообще).

Исходную краевую задачу можно записать как операторное уравнение $Lu = f$, в котором u — неизвестная функция, L — оператор, объединяющий левые части дифференциального уравнения и краевых условий, а f — правые части дифференциального уравнения и краевых условий. Например, для одномерной краевой задачи

$$\begin{cases} u_t - a^2 u_{xx} = \gamma(x,t), \\ u(x,0) = \varphi(x), \\ u(0,t) = \mu(t), \\ u(l,t) = \nu(t) \end{cases}$$

можно рассмотреть линейное пространство дважды непрерывно дифференцируемых функций $u(x,t)$ и в нем линейный оператор L, который каждой функции $u(x,t)$ ставит в соответствие упорядоченный набор

$$L(u) = \big(u_t - a^2 u_{xx},\ u(x,0),\ u(0,t),\ u(l,t)\big),$$

являющийся элементом соответствующего линейного пространства. Тогда краевую задачу можно записать в виде $Lu = f$, где f — это упорядоченный набор

$$f = (\gamma(x,t),\ \varphi(x),\ \mu(t),\ \nu(t)).$$

Рассматриваемые линейные пространства бесконечномерны. При дискретизации операторное уравнение $Lu = f$ заменяется серией уравнений $\widehat{L}_N \widehat{u}_N = \widehat{f}_N$, в котором $\widehat{u}_N$ — ***сеточная функция***, т.е. некоторая функция, определенная в узлах сетки S^N, $\widehat{L}_N$ — векторная функция, описывающая левые части сеточных уравнений, а $\widehat{f}_N$ — вектор, описывающий правые части сеточных уравнений.

Сравнить решение исходной краевой задачи с решением ее дискретного аналога можно лишь в узлах сетки. Пусть u_N — сеточная функция, определенная на сетке S^N, значениями которой являются значения в узлах сетки решения u краевой задачи $L(u) = f$. Решение дискретного аналога тем точнее, чем меньше величина $\|\widehat{u}_N - u_N\|_N$, где $\|\cdot\|_N$ — какая-либо норма в линейном пространстве сеточных функций на сетке S^N. Если $\|\widehat{u}_N - u_N\|_N \to 0$ при $N \to \infty$, то говорят, что ***сеточная схема сходится*** к решению краевой задачи.

Описанный способ оценки точности дискретизации не является единственно возможным. Как правило, сравнение сложных математических объектов осуществляется в рамках того или иного линейного пространства, зачастую бесконечномерного, и строится на базе той или иной нормы. Чтобы сравнить функцию $u(x,t)$ непрерывных аргументов с сеточной функцией $\widehat{u}_N$, их надо привести к какому-то одному виду. Если есть оператор P, действующий из пространства функций $u(x,t)$ в пространство сеточных функций $\widehat{u}_N$, то отличие функции $u(x,t)$ от $\widehat{u}_n$ можно характеризовать величиной $\|P(u) - \widehat{u}_N\|_N$. Сужение функции $u(x,t)$ на узлы сетки как раз и играет роль оператора P.

Сходимость сеточной схемы к решению краевой задачи означает адекватность дискретной математической модели, получаемой в рамках этой схемы, соответствующей непрерывной математической модели, т.е. краевой задаче. Качество сеточной схемы характеризуется еще одним свойством — ее ***устойчивостью***. Под эти понимается непрерывная зависимость решения

сеточной задачи от исходных данных. Под исходными данными понимаются значения правых частей краевой задачи, левые части краевой задачи характеризуют закон развития соответствующего процесса и закон взаимодействия с окружающей средой. При дискретизации исходные данные включаются в систему нелинейных уравнений как значения правых частей и коэффициентов уравнений. Незначительное изменение исходных данных означает, что произошли незначительные изменения в начальном состоянии процесса или в состоянии окружающей среды. Такие изменения не должны приводить к резкому, скачкообразному изменению решения дискретной задачи. Устойчивость сеточной схемы — ее внутреннее свойство, никак не связанное с краевой задачей, для которой используется эта схема. Просто некоторые коэффициенты в системе нелинейных уравнений могут варьироваться, а устойчивость означает непрерывную зависимость решения системы от этих коэффициентов.

Исследовать сеточную схему на сходимость — достаточно сложное дело. В этом исследовании помогают ***аппроксимирующие свойства*** схемы. Под этим понимается следующее. Сужение u_N решения u исходной краевой задачи $Lu = f$ на сетку S^N не совпадает с решением $\widehat{u}_N$ дискретной задачи $\widehat{L}\widehat{u}_N = \widehat{f}_N$, но близко к нему. Подставив u_N в дискретное операторное уравнение, т.е. вычислив $\widehat{L}u_N$, мы не получим правую часть уравнения $\widehat{f}_N$, но значение будет близким к нему, т.е. величина $\|\widehat{L}u_N - \widehat{f}_N\|_N$ будет мала. Эту величину называют ***невязкой***. Невязка характеризует близость сеточных функций u_N и $\widehat{u}_N$. При некоторых дополнительных предположениях можно утверждать, что величина $\|u_N - \widehat{u}_N\|_N$ мала тогда и только тогда, когда мала невязка. Поэтому о сходимости сеточной схемы можно судить по степени убывания к нулю невязки, соответствующей этой схеме.

Невязка интересна тем, что ее во многих случаях можно оценить в зависимости от некоторого параметра h_N, характеризующего расстояния между близкими узлами сетки, причем такая оценка, как правило, имеет вид $o(h_N^k)$, т.е. характеризуется порядком малости невязки по отношению к параметру h_N. В этом случае k называют ***порядком аппроксимации*** сеточной схемы. Оценка невязки затем используется для доказательства сходимости сеточной схемы.

Итак, анализ различных сеточных схем базируется на трех китах: сходимости, устойчивости и порядке аппроксимации. Основными являются первые два свойства, а третье носит вспомогательную роль, помогая проверить наличие первых двух.

3.2. Разностная аппроксимация производных

Одним из способов формирования сеточной схемы является замена частных производных в дифференциальном уравнении их разностными аналогами. Например, рассмотрим функцию одного переменного $f(x)$. Согласно определению производной

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h}.$$

Поэтому в качестве приближенного значения производной $f'(x_0)$ в данной точке x_0 можно взять отношение

$$\frac{f(x_0+h) - f(x_0)}{h}$$

с достаточно малым значением h. В качестве приближенного значения $f'(x_0)$ можно также использовать отношения

$$\frac{f(x_0) - f(x_0-h)}{h} \quad \text{и} \quad \frac{f(x_0+h) - f(x_0-h)}{2h}.$$

Качество этих разностных аналогов производной можно оценить с помощью формулы Тейлора с остаточным членом в форме Пеано. Действительно, согласно этой формуле

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + o(h^2).$$

и

$$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + o(h^2).$$

Поэтому

$$\frac{f(x_0+h) - f(x_0)}{h} - f'(x_0) = \frac{1}{2}f''(x_0)h + o(h) = O(h).$$

Аналогично

$$\frac{f(x_0) - f(x_0-h)}{h} - f'(x_0) = -\frac{1}{2}f''(x_0)h + o(h) = O(h).$$

Оказывается, что симметричная разность дает больший порядок аппроксимации:

$$\frac{f(x_0+h) - f(x_0-h)}{h} - f'(x_0) = o(h) = O(h^2).$$

В этих оценках предполагается, что функция $f(x)$ по меньшей мере дважды дифференцируема. Это не жесткое требование, так как решения краевых задач, как правило, удовлетворяют ему.

По тем же правилам можно строить приближенные формулы для производных второго порядка. Наиболее употребительной является следующая формула:

$$f''(x_0) \approx= \frac{f(x_0-h) - 2f(x_0) + f(x_0+h)}{h^2}.$$

Оценим точность этой формулы, снова используя формулу Тейлора, но более высокой степени:

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \frac{1}{6}f'''(x_0)h^3 + O(h^4).$$

Заменяя h на $-h$ и затем складывая две формулы, получаем

$$f(x_0+h) + f(x_0-h) = 2f(x_0) + f''(x_0)h^2 + O(h^4).$$

Следовательно,

$$\frac{f(x_0-h) - 2f(x_0) + f(x_0+h)}{h^2} - f''(x_0) = O(h^2).$$

Вообще подобные формулы можно строить следующим образом. Выберем, например, два дополнительных узла x_1 и x_2 и образуем выражение $a_0f(x_0) + a_1f(x_1) + a_2f(x_2)$. Коэффициенты a_i попробуем выбрать так, что это выражение будет близко к $f'(x_0)$. Для этого положим $h_1 = x_1 - x_0$, $h_2 = x_2 - x_0$ и применим формулу Тейлора соответствующей степени:

$$f(x_1) = f(x_0+h_1) = f(x_0) + f'(x_0)h_1 + \frac{1}{2}f''(x_0)h_1^2 + o(h_1^2),$$
$$f(x_2) = f(x_0+h_2) = f(x_0) + f'(x_0)h_2 + \frac{1}{2}f''(x_0)h_2^2 + o(h_2^2).$$

Подставив эти представления в выражение, получим

$$a_0f(x_0) + a_1f(x_1) + a_2f(x_2) =$$
$$= (a_0+a_1+a_2)f(x_0) + (a_1h_1 + a_2h_2)f'(x_0) + \frac{1}{2}(a_1h_1^2 + a_2h_2^2)f''(x_0) + o(h_1^2) + o(h_2^2).$$

Наилучшая аппроксимация будет в том случае, когда коэффициенты a_i удовлетворяют системе уравнений

$$\begin{cases} a_0 + a_1 + a_2 = 0, \\ a_1h_1 + a_2h_2 = 1, \\ a_1h_1^2 + a_2h_2^2 = 0. \end{cases}$$

В этом случае

$$a_0f(x_0) + a_1f(x_1) + a_2f(x_2) - f'(x_0) = o(h_1^2) + o(h_2^2).$$

Записанная система при любых различных $h_1 \neq 0$ и $h_2 \neq 0$ имеет решение, так как определитель этой системы равен $h_1h_2(h_2 - h_1)$. В частном случае $h_1 = -h_2 = h > 0$ мы придем к симметрической разности.

3.3. Одномерное уравнение теплопроводности

Сеточный метод, основанный на замене в дифференциальном уравнении производных конечными разностями, называют ***методом конечных разностей***, а сеточную схему такого метода — ***разностной схемой***[*]. Метод конечных разностей накладывает определенные ограничения на структуру сетки, так как конечные разности должны составляться из значений функции в узлах сетки. Для метода конечных разностей характерно расположение узлов слоями по переменным. В простейших ситуациях (задачи с неподвижными границами) диапазон изменения времени t не связан с областью изменения пространственных переменных x, y, z. Поэтому естественно выбрать сетку следующим образом. В пространственной области D изменения переменных x, y, z выберем некоторое множество узлов $\boldsymbol{x}^1$, $\boldsymbol{x}^2$, …, $\boldsymbol{x}^k$. В качестве узлов пространственно-временной сетки выбираем точки $(\boldsymbol{x}^i,\, t_j)$, где $t_j = t_0 + \tau j$, $j = 1, \ldots, M$; t_0 — начальный момент времени, τ — шаг временно́й переменной, некоторое фиксированное число. Подобное расположение узлов упрощает аппроксимацию производной по времени. Совокупность узлов, соответствующих одному моменту времени, называют ***временны́м слоем сетки***. Расположение узлов $\boldsymbol{x}^i$ может определяться формой области D, видом дифференциального уравнения и граничных условий, в каких координатах они записаны (декартовых, цилиндрических, сферических).

Детали метода конечных разностей рассмотрим на конкретном примере

$$\begin{cases} c\rho\dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\Big(K(u(x))\dfrac{\partial u}{\partial x}\Big), \quad x \in (0,\, l), \quad t > 0; \\ u(x,0) = \varphi(x); \\ u(0,t) = \mu(t); \\ K(u)\dfrac{\partial u}{\partial x}\Big|_{x=l} = W(t). \end{cases}$$

Сформулированная краевая задача описывает процесс распространения тепла в однородном стержне, теплоизолированном с боковых сторон. В ней c — удельная линейная теплоемкость материала стержня, ρ — линейная плотность материала стержня, $K(u)$ — коэффициент теплопроводности, зависящий от температуры.

Мы имеем дело с нелинейным дифференциальным уравнением[**]. На левом конце стержня $(x = 0)$ поддерживается определенная температура $\mu(t)$. На правом конце задан режим теплообмена, при котором в стержень поступает поток тепла $W(t)$. Функция $\varphi(x)$ описывает начальное распределение температур.

Разностная схема. Для сформулированной задачи наиболее простой является прямоугольная сетка W_h^τ с узлами $w_i^j = (x_i,\, t_j)$, $i = \overline{0,\, N}$, $j = \overline{0,\, M}$, где $x_i = ih$, $t = j\tau$, параметры h и τ определяют шаг сетки по переменным x и t. В области определения должно уместиться целое число шагов сетки, т.е. должно выполняться соотношение $l = hN$, величина $T = \tau M$ — это период времени, на котором решается задача.

Сеточная функция $\boldsymbol{v}$ в данном случае представляет собой двумерный массив (матрицу) $\{v_i^j\}$, в котором значение v_i^j соответствует узлу w_i^j. Выбор сетки упрощает аппроксимацию дифференциального уравнения, так как узлы сетки расположены слоями по переменным. Аппроксимация производной по времени не вызывает затруднений — об этом чуть позже. Обсудим, как аппроксимировать дифференциальный оператор по переменному x. Сложность состоит в том, что в него входит неизвестная функция. Один из возможных вариантов — следующая формула:

$$\frac{\partial}{\partial x}\Big(K(u)\frac{\partial u}{\partial x}\Big)(x_i,t_j) \sim \frac{1}{h}\Big(K_{i+}^j\frac{u_{i+1}^j - u_i^j}{h} - K_{i-}^j\frac{u_i^j - u_{i-1}^j}{h}\Big),$$

[*]Часто термин „разностная схема“ отождествляют с термином „сеточная схема“, так как сеточные методы, основанные на конечных разностях, наиболее распространенные.

[**]Если K зависит от x, но не от u, то это линейное дифференциальное уравнение.

в которой задействованы три узла сетки. В ней величины K_{i+}^j и K_{i-}^j символизируют в некотором смысле среднее значение $K(u)$ на отрезках $[x_j,\, x_{j+1}]$ и $[x_{j-1},\, x_j]$. Про $K(u)$ ничего не известно, кроме ее значений в узлах сетки $K_i^j = K(u_i^j)$. Но, с другой стороны, речь идет о приближении. Поэтому естественно считать, что

$$K_{i+}^j = \frac{K_{i+1}^j + K_i^j}{2}, \qquad K_{i-}^j = \frac{K_i^j + K_{i-1}^j}{2}.$$

Частную производную по времени можно аппроксимировать одной из формул

$$\frac{\partial u}{\partial t}(x_i,t_j) \sim \frac{u_i^{j+1} - u_i^j}{\tau}, \qquad \frac{\partial u}{\partial t}(x_i,t_j) \sim \frac{u_i^j - u_i^{j-1}}{h},$$

В результате мы для каждого внутреннего узла сетки получим уравнение, связывающее значение сеточной функции в этом узле со значениями сеточной функции в трех соседних узлах. При этом в зависимости от выбора аппроксимации для производной по времени появляется две разностные схемы. В первом случае уравнение связывает три смежных узла j-го слоя сетки с одним узлом $(j+1)$-го слоя (рис. 3.1). Во втором случае сеточное уравнение связывает один узел j-го слоя сетки с тремя смежными узлами $(j+1)$-го слоя (рис. 3.2). Конфигурацию узлов, входящих в сеточное уравнение, называют ***шаблоном сетки***.

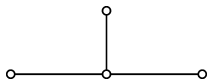

Рис. 3.1

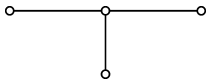

Рис. 3.2

Шаблон на рис. 3.1 соответствует сеточному уравнению

$$\frac{u_i^{j+1} - u_i^j}{\tau} = \frac{K_{i+}^j(u_{i+1}^j - u_i^j) - K_{i-}^j(u_i^j - u_{i-1}^j)}{h^2},$$

а шаблон на рис. 3.2 — сеточному уравнению

$$\frac{u_i^{j+1} - u_i^j}{\tau} = \frac{K_{i+}^{j+1}(u_{i+1}^{j+1} - u_i^{j+1}) - K_{i-}^{j+1}(u_i^{j+1} - u_{i-1}^{j+1})}{h^2},$$

Замена начального условия достаточно очевидна: $u_i^0 = \varphi(x_i)$, $i = \overline{0,\, N}$. При такой замене значения сеточной функции на нулевом временно́м слое оказываются известными. Точно так же можно заменить граничное условие первого рода, которое в рассматриваемой задаче поставлено на левом конце: $u_0^j = \mu(t_j)$, $j = \overline{0,\, M}$. В граничное условие второго рода (на правом конце) входит производная по x, которую естественно заменить конечной разностью, и коэффициент теплопроводности $K(u)$, который интерпретируем так же, как и при составлении сеточного уравнения. В результате приходим к „усеченным“ сеточным уравнениям

$$K_{N-}^j\frac{u_N^j - u_{N-1}^j}{h} = W(t_j).$$

Для аппроксимации производной по времени можно было бы использовать симметричную конечную разность, дающую бо́льшую точность. Однако выигрыш в точности аппроксимации перечеркивается проигрышем, который возникает из-за того, что в сеточное уравнение будет входить не два, а три временных слоя. В результате сеточное уравнение усложняется.

Послойный характер сеточных взаимосвязей подсказывает порядок решения системы уравнений. Она распадается на M систем, каждая из которых соответствует одному временному слою. Действительно, если известны значения сеточной функции на j-м слое, то сеточные урав-

нения для узлов этого слоя плюс два граничных условия для $(j+1)$-го слоя составляют полную систему уравнений для узлов $(j+1)$-го слоя, из которой можно найти значения сеточной функции во всех узлах этого слоя. Таким образом, процесс вычислений напоминает естественное течение процесса распространения тепла: от одного момента времени к другому. Наиболее привлекательна разностная схема с шаблоном на рис. 3.1. Для этой схемы значения сеточной функции в узлах $(j+1)$-го слоя находятся простым пересчетом. А в случае разностной схемы с шаблоном на рис. 3.2 мы имеем систему уравнений с M неизвестными, причем это нелинейная система, так как в сеточные уравнения входят параметры K_{i+}^{j+1} и K_{i-}^{j+1}, являющиеся функциями неизвестных значений $(j+1)$-го слоя. В теории разностных схем первую называют ***явной разностной схемой***, а вторую — ***неявной разностной схемой***. Поскольку сеточные уравнения и той, и другой разностной схемы связывают значения функции на двух слоях, о них говорят как о ***двуслойных разностных схемах***. Упомянутая выше разностная схема, в которой производная по времени аппроксимируется симметричной разностью, является ***трехслойной разностной схемой***.

Хотя кажется очевидным, что явная разностная схема имеет преимущество, на самом деле это не так: явная схема проигрывает неявной в вопросах устойчивости. Детали обсудим чуть позже, а здесь лишь подчеркнем, что ни той, ни другой разностной схеме нельзя отдать предпочтение: выигрыш в одном приводит к проигрышу в другом. Отметим также, что возможна смешанная разностная схема, которая получается, если в каждом внутреннем узле сетки сложить сеточные уравнения явной и неявной разностных схем с дополнительными масштабными коэффициентами (ее шаблон показан на рис. 3.3).

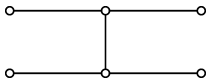

Рис. 3.3

Для дальнейшего обсуждения условимся о некоторых обозначениях. Пусть Ω_h — линейное пространство одномерных сеточных функций, принимающих значения в точках x_i, $i = \overline{0,\, N}$. Элемент $\boldsymbol{f}$ этого линейного пространства можно записать как $(N+1)$-мерный вектор $\boldsymbol{f} = (f_0,\, \ldots,\, f_N)$, компоненты которого — значения сеточной функции, т.е. $f_i = \boldsymbol{f}(x_i)$, $i = \overline{0,\, N}$. Совокупность сеточных уравнений для j-го слоя, включая аппроксимацию граничных условий, можно рассматривать как оператор Λ^j, который сеточной функции $\boldsymbol{f} = (f_0,\, \ldots,\, f_N)$ ставит в соответствие сеточную функцию $\boldsymbol{g} = (g_0,\, \ldots,\, g_N)$ с компонентами

$$
\begin{aligned}
&g_0 = \mu(t_{j+1}),\\
&g_i = \frac{K_{i+}^j(f_{i+1} - f_i) - K_{i-}^j(f_i - f_{i-1})}{h^2}, \quad i = \overline{1,\, N},\\
&g_N = K_{N-}^{j+1}\frac{f_N - f_{N-1}}{h}
\end{aligned}
$$

Видно, что граничные условия носят несколько инородный характер. Поэтому желательно их исключить (напомним, что в краевой задаче при рассмотрении ее как операторного уравнения граничные условия не включают в это уравнения, интерпретируя их как ограничение множества допустимых функций). Вместо Ω_h введем другое линейное пространство Ω_h^0 сеточных функций, которые определены во внутренних узлах пространственной сетки x_i, $i = \overline{1,\, N-1}$. Изменим сеточные уравнения для 1-го и $(N-1)$-го пространственных узлов, учтя граничные условия:

$$
g_1 = \frac{K_{1+}^j(f_2 - f_1) - K_{1-}^j(f_1 - \mu^j)}{h^2}, \quad g_{N-1} = \frac{hW^j - K_{(N-1)-}^j(f_{N-1} - f_{N-2})}{h^2},
$$

где $\mu^j = \mu(t_j)$, $W^j = W(t_j)$. В результате приходим к оператору Λ^j, действующему в линейном пространстве Ω_h^0. При этом связь сеточных функций $\boldsymbol{v}^j$, $\boldsymbol{v}^{j+1} \in \Omega_h^0$, соответствующих двум соседним временным слоям, можно записать в виде

$$
\frac{\boldsymbol{v}^{j+1} - \boldsymbol{v}^j}{\tau} = \Lambda^j \boldsymbol{v}^j
$$

для явной разностной семы и в виде

$$
\frac{\boldsymbol{v}^{j+1} - \boldsymbol{v}^j}{\tau} = \Lambda^{j+1} \boldsymbol{v}^j
$$

для неявной разностной схемы. Смешанную разностную схему можно записать в виде

$$
\frac{\boldsymbol{v}^{j+1} - \boldsymbol{v}^j}{\tau} = \alpha\Lambda^{j+1}\boldsymbol{v}^{j+1} + (1-\alpha)\Lambda^j\boldsymbol{v}^j,
$$

где $\alpha \in [0,\, 1]$, причем $\alpha = 0$ соответствует явной схеме, а $\alpha = 1$ — неявной.

Устойчивость. Для поставленной нами задачи требование устойчивости налагает ограничение на параметры сетки h и τ. Рассмотрим смешанную разностную схему, рассматривая явную и неявную схемы как частный случай смешанной. Основной вывод здесь состоит в том, что смешанная схема, определяемая параметром α, является устойчивой, если

$$
\Big(\frac{1}{2} - \alpha\Big)\frac{4\tau}{h^2}\max_{x\in[0,l]} K(u(x)) \leqslant 1.
$$

Отсюда, в частности, следует, что смешанная разностная схема при $\alpha \geqslant 1/2$, в том числе неявная схема, устойчивы при любых соотношениях h и τ. Смешанная разностная схема при $\alpha < 1/2$, в том числе явная схема, устойчивы лишь при определенных сочетаниях параметров h и τ.

Чтобы упростить выкладки, рассмотрим частный случай краевой задачи, когда коэффициент теплопроводности не зависит от температуры, т.е. $K(u) \equiv K$. В этом случае все операторы Λ^j одинаковы и определяются соотношениями

$$
(\Lambda\boldsymbol{f})_i = \frac{K}{h^2}\big(f_{i-1} - 2f_i + f_{i+1}\big), \quad i = \overline{1,\, N-1}.
$$

Смешанная разностная схема определяется операторным уравнением

$$
\frac{\boldsymbol{u}^{j+1} - \boldsymbol{u}^j}{\tau} = \alpha\Lambda\boldsymbol{u}^{j+1} + (1-\alpha)\Lambda\boldsymbol{u}^j, \tag{3.1}
$$

граничными условиями

$$
u_0^j = \mu^j, \quad K\frac{u_N^j - u_{N-1}^j}{h} = W^j, \quad j = \overline{1,\, M}, \tag{3.2}
$$

и начальным условием

$$
\boldsymbol{u}^0 = \boldsymbol{\varphi}.
$$

Эту задачу удобно свести к случаю однородных граничных условий, заменяя операторное уравнение неоднородным. Такое преобразование аналогично преобразованию непрерывной задачи: достаточно придумать функцию u_0, удовлетворяющую поставленным граничным условиям и провести замену $u = v + u_0$. Тогда функция v будет удовлетворять однородным граничным условиям, а в операторном уравнении появится дополнительное слагаемое, соответствующее функции u_0. В нашем случае пусть $\boldsymbol{z}$ удовлетворяет граничным условиям (3.2). В операторное уравнение подставим $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{z}$. Тогда

$$
\frac{\boldsymbol{v}^{j+1} - \boldsymbol{v}^j}{\tau} = \alpha\Lambda\boldsymbol{v}^{j+1} + (1-\alpha)\Lambda\boldsymbol{v}^j + \boldsymbol{f}^j,
$$

где

$$\boldsymbol{f}^j = \alpha\Lambda\boldsymbol{z}^{j+1} + (1-\alpha)\Lambda\boldsymbol{z}^j - \frac{\boldsymbol{z}^{j+1} - \boldsymbol{z}^j}{\tau}.$$

Функцию $\boldsymbol{z}$ легко выбрать. Например, можно положить $z_i^j = 0$ при $1 \leqslant i \leqslant N-1$, $z_0^j = \mu^j$ и $z_N^j = hW^j/K$. Множество одномерных сеточных функций, удовлетворяющих однородным граничным условиям, т.е. функций $\boldsymbol{f} \in \Omega_h$, для которых $f_0 = 0$ и $f_{N-1} = f_N$, есть подпространство в Ω_h. Поскольку такие функции однозначно определяются своими значениями во внутренних узлах, это подпространство можно отождествить с Ω_h^0.

Для оценки близости сеточных функций в Ω_h^0 необходимо в этом линейном пространстве ввести норму. Поскольку оно конечномерное, выбор нормы не является существенным и может проводиться из соображений удобства. В Ω_h^0 можно ввести скалярное произведение

$$(\boldsymbol{u},\, \boldsymbol{v}) = \sum_{i=1}^{N-1} u_i v_i h$$

и норму $|\boldsymbol{u}| = \sqrt{(\boldsymbol{u},\, \boldsymbol{u})}$, индуцированную скалярным произведением. Мы также будем оперировать нормой

$$\|\boldsymbol{u}\|_c = \max\{|u_1|,\ |u_2|,\ \ldots,\ |u_{N-1}|\}.$$

Ошибка, получаемая при вычислении $(j+1)$-го слоя формируется ошибками, входящими в значения функций $\boldsymbol{v}^j$ и $\boldsymbol{f}^j$. Будем считать, что эти функции имеют вид $\boldsymbol{v}^j + \delta\boldsymbol{v}^j$ и $\boldsymbol{f}^j + \delta\boldsymbol{f}^j$, где символ δ указывает на погрешность функции. Тогда основное уравнение будет иметь вид

$$\frac{(\boldsymbol{v}^{j+1} + \delta\boldsymbol{v}^{j+1}) - (\boldsymbol{v}^j + \delta\boldsymbol{v}^j)}{\tau} = \alpha\Lambda(\boldsymbol{v}^{j+1} + \delta\boldsymbol{v}^{j+1}) + (1-\alpha)\Lambda(\boldsymbol{v}^j + \delta\boldsymbol{v}^j) + \boldsymbol{f}^j + \delta\boldsymbol{f}^j.$$

Считая, что „истинные“ функции $\boldsymbol{v}^j$, $\boldsymbol{v}^{j+1}$ и $\boldsymbol{f}^j$ также связаны операторным уравнением, заключаем, что

$$\frac{\delta\boldsymbol{v}^{j+1} - \delta\boldsymbol{v}^j}{\tau} = \alpha\Lambda(\delta\boldsymbol{v}^{j+1}) + (1-\alpha)\Lambda(\delta\boldsymbol{v}^j) + \delta\boldsymbol{f}^j,$$

т.е. погрешности связаны тем же уравнением. Из последнего уравнения находим

$$\delta\boldsymbol{v}^{j+1} - \tau\alpha\Lambda(\delta\boldsymbol{v}^{j+1}) = \delta\boldsymbol{v}^j + \tau(1-\alpha)\Lambda(\delta\boldsymbol{v}^j) + \tau\delta\boldsymbol{f}^j,$$

или

$$(E - \tau\alpha\Lambda)\delta\boldsymbol{v}^{j+1} = (E + \tau(1-\alpha)\Lambda)\delta\boldsymbol{v}^j + \tau\delta\boldsymbol{f}^j.$$

Если оператор $E - \tau\alpha\Lambda$ имеет обратный, то

$$\delta\boldsymbol{v}^{j+1} = (E - \tau\alpha\Lambda)^{-1}(E + \tau(1-\alpha)\Lambda)\delta\boldsymbol{v}^j + \tau(E - \tau\alpha\Lambda)^{-1}\delta\boldsymbol{f}^j. \tag{3.3}$$

Как мы видим, в погрешность $\delta\boldsymbol{v}^{j+1}$ входят две компоненты. Первая связана с ошибкой предыдущего слоя, а вторая — с ошибками в граничных условиях. Отметим, что оператор Λ является самосопряженным, так как в стандартном базисе (т.е. базисе из сеточных функций $\boldsymbol{e}^k$, для которых $e_i^k = \delta_i^k$, $i,\ k = \overline{1,\ N-1}$, где δ_i^k — символ Кронекера) матрица этого оператора симметрическая:

$$[\Lambda]_e = \frac{K}{h^2}\begin{pmatrix} -2 & 1 & 0 & \ldots & 0 & 0 \\ 1 & -2 & 1 & \ldots & 0 & 0 \\ 0 & 1 & -2 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix}$$

Такой оператор приводится к диагональному виду, причем элементы соответствующей матрицы — собственные числа — могут быть найдены как решения характеристического уравнения $\det(\Lambda - \mu E) = 0$. Операторы

$$L = (E - \tau\alpha\Lambda)^{-1}(E + \tau(1-\alpha)\Lambda), \qquad F = (E - \tau\alpha\Lambda)^{-1}$$

также являются самосопряженными. Характеристическое уравнение для L имеет вид

$$\begin{aligned}\det(L - \mu E) &= \det\big((E - \tau\alpha\Lambda)^{-1}(E + \tau(1-\alpha)\Lambda) - \mu E\big) = \\ &= \det(E - \tau\alpha\Lambda)^{-1} \cdot \det\big((E + \tau(1-\alpha)\Lambda) - \mu(E - \tau\alpha\Lambda)\big) = \\ &= \det(E - \tau\alpha\Lambda)^{-1} \cdot \det\big(\tau(1 - \alpha + \alpha\mu)\Lambda - (\mu - 1)E\big) = 0.\end{aligned}$$

Следовательно, если оператор $E - \tau\alpha\Lambda$ обратим, то уравнение $\det(L - \mu E) = 0$ равносильно уравнению

$$\det\left(\Lambda - \frac{\mu - 1}{\tau(1 - \alpha + \alpha\mu)}E\right) = 0.$$

откуда следует, что собственные числа μ_i оператора L связаны с собственными числами λ_i оператора Λ соотношениями

$$\frac{\mu_i - 1}{\tau(1 - \alpha + \alpha\mu_i)} = \lambda_i.$$

или

$$\mu_i = 1 + \frac{\tau\lambda_i}{1 - \alpha\tau\lambda_i}.$$

Характеристическое уравнение для оператора F имеет вид

$$\det(F - \nu E) = \det\big((E - \tau\alpha\Lambda)^{-1} - \nu E\big) = \det(E - \tau\alpha\Lambda)^{-1} \cdot \det(\tau\alpha\nu\Lambda - (\nu - 1)E) = 0,$$

откуда

$$\det\left(\Lambda - \frac{\nu - 1}{\tau\alpha\nu}E\right) = 0.$$

Следовательно, собственные числа ν_i оператора F можно найти через собственные числа λ_i оператора Λ по формулам

$$\nu_i = \frac{1}{1 - \tau\alpha\lambda_i}.$$

Найдем собственные числа оператора Λ. Для этого необходимо найти все нетривиальные сеточные функции $\boldsymbol{v}$ из пространства Ω_h^0, удовлетворяющие операторному уравнению $\Lambda\boldsymbol{v} = \lambda\boldsymbol{v}$. Оператор Λ на сеточных функциях является аналогом оператора $\frac{d^2}{dx^2}$ на непрерывных функциях. Поэтому и решения операторного уравнения можно найти по аналогии. Ищем решения указанного операторного уравнения в виде

$$v_i = A\sin\omega i + B\cos\omega i, \quad i = \overline{1,\ N-1}.$$

Используя вид оператора Λ, находим

$$(\Lambda\boldsymbol{v})_i = \frac{K}{h^2}(v_{i-1} - 2v_i + v_{i+1}) = \frac{K}{h^2}(2\cos\omega - 2)(A\sin\omega i + B\cos\omega i).$$

Нам остается найти среди этих функций те, которые попадают в Ω_h^0, т.е. удовлетворяют однородным граничным условиям $v_0 = 0$ и $v_N = v_{N-1}$. Из первого равенства следует, что $B = 0$, а из второго заключаем, что $\sin\omega N = \sin\omega(N-1)$, или $2\sin\frac{\omega}{2}\cos\omega\left(N - \frac{1}{2}\right) = 0$. Решениями полученного уравнения будет последовательность

$$\omega_k = \frac{(2k-1)\pi}{2N-1}, \quad k \in \mathbb{Z}.$$

Они определяют $N-1$ независимых сеточных функций $\boldsymbol{v}^k$ в соответствии с формулой

$$v_i^k = \sin\frac{(2k-1)\pi i}{2N-1}, \quad i,\, k = \overline{1,\, N-1}.$$

Собственными числами найденных собственных фукнций будут числа

$$\lambda_k = \frac{K}{h^2}\Big(2\cos\frac{(2k-1)\pi}{2N-1} - 2\Big) = -\frac{4K}{h^2}\sin^2\frac{(2k-1)\pi}{2(2N-1)}, \quad k = \overline{1,\, N-1}$$

В частности, получаем $-4K/h^2 \leqslant \lambda_k \leqslant 0$, а максимальное по модулю собственное число равно

$$\lambda_{\max} = -\frac{4K}{h^2}\cos^2\frac{\pi}{2N-1}.$$

Оператор $E - \tau\alpha\Lambda$ обратим, так как уравнение $(E-\tau\alpha\Lambda)\boldsymbol{v} = 0$ имеет только тривиальное решение. Действительно, это уравнение равносильно уравнению $\Lambda\boldsymbol{v} = (\tau\alpha)^{-1}\boldsymbol{v}$, но линейный оператор Λ не имеет положительных собственных чисел, а значит, и $(\tau\alpha)^{-1}$ не является собственным числом Λ.

Поскольку оператор $E - \tau\alpha\Lambda$ обратим, все ранее выписанные равенства, базирующиеся на предположении обратимости этого оператора, корректны.

Перейдем к оценке ошибки $\boldsymbol{v}^{j+1}$. Из равенства (3.3) получаем

$$|\delta\boldsymbol{v}^{j+1}| \leqslant |L(\delta\boldsymbol{v}^j)| + \tau|F(\delta\boldsymbol{f}^j)| \leqslant \|L\|\,|\delta\boldsymbol{v}^j| + \tau\|F\|\,|\delta\boldsymbol{f}^j|. \tag{3.4}$$

Евклидовы нормы операторов L и F легко находятся через их собственные числа (собственно , именно для этого и определялся их спектр):

$$\|L\| = \max|\mu_i| = \max\left\{1 - \frac{\frac{4K\tau}{h^2}\sin^2\frac{\pi}{2(2N-1)}}{1+\alpha\frac{4K\tau}{h^2}\sin^2\frac{\pi}{2(2N-1)}},\ \left|1 - \frac{\frac{4K\tau}{h^2}\cos^2\frac{\pi}{2N-1}}{1+\alpha\frac{4K\tau}{h^2}\cos^2\frac{\pi}{2N-1}}\right|\right\},$$

$$\|F\| = \max|\nu_i| = \max\frac{1}{1-\tau\alpha\lambda_i} = \frac{1}{1+\alpha\frac{4K\tau}{h^2}\sin^2\frac{\pi}{2(2N-1)}} \leqslant 1.$$

Два слагаемых в правой части неравенства (3.4) имеют разный характер. Первое является итерационным и в простейшей ситуации $\boldsymbol{f}^j = 0$ получаем

$$|\delta\boldsymbol{v}^j| \leqslant \|L\|^j\,|\delta\boldsymbol{v}^0|$$

Чтобы погрешность оставалась ограниченной, необходимо выполнение условия $\|L\| \leqslant 1$. Но это выполняется, если

$$\frac{\frac{4K\tau}{h^2}}{1+\alpha\frac{4K\tau}{h^2}} \leqslant 2.$$

что равносильно неравенству

$$\Big(\frac{1}{2} - \alpha\Big)\frac{4K\tau}{h^2} \leqslant 1. \tag{3.5}$$

При этом условии, применяя итерационно неравенство (3.4) и учитывая, что $\|F\| \leqslant 1$, получаем

$$|\boldsymbol{v}^j| \leqslant |\boldsymbol{v}^0| + \tau\sum_{k=0}^{j-1}|\delta\boldsymbol{f}^k| \leqslant |\boldsymbol{v}^0| + \tau j\max_{k=\overline{0,\,j-1}}|\delta\boldsymbol{f}^k| \leqslant |\boldsymbol{v}^0| + T\max_{k=\overline{0,\,M-1}}|\delta\boldsymbol{f}^k|,$$

что и означает устойчивость разностной схемы.

В общем случае ($K = K(u)$) анализ разностной схемы на устойчивость усложняется, так как операторы Λ^j уже не будут линейными. Но в конечном счете все сводится к получению оценок типа условия Липшица $|f(x) - f(y)| \leqslant C|x-y|$. Причем такие оценки можно сперва получать в малом, когда K можно считать постоянным, а затем объединять оценки по всей области. В конечном счете условие устойчивости разностной схемы в нелинейном случае получается в следующем виде:

$$\Big(\frac{1}{2} - \alpha\Big)\frac{4K_{\max}\tau}{h^2} \leqslant 1, \tag{3.6}$$

где $K_{\max}$ — максимальное значение коэффициента теплопроводности в области $[0,\,l]\times[0,\,T]$.

Вместо значения $K_{\max}$, которое можно определить лишь, зная искомую функцию $u(x,t)$, следует использовать какую-либо оценку сверху этой величины. Отметим, что в силу послойного характера разностной схемы в условии устойчивости можно вместо $K_{\max}$ использовать на каждом j-м шаге максимальное значение K в области $t_j \leqslant t \leqslant t_{j+1}$. Это значение близко максимальному значению $K^j_{\max}$ на прямой $t = t_j$. Учитывая это, можно использовать, например, явную схему с переменным шагом по времени, который подбирается так, что выполняется условие устойчивости

$$\Big(\frac{1}{2} - \alpha\Big)\frac{4K^j_{\max}\tau}{h^2} \leqslant 1.$$

Это обеспечивает устойчивость разностной схемы, поскольку погрешность вычислений при переходе на очередной слой не возрастает.

Аппроксимация и сходимость. Для установления сходимости рассматриваемой разностной схемы к решению исходной непрерывной задачи необходимо оценить разность сеточной функции $\widehat{u}$ — решения разностной задачи $\widehat{L}\widehat{u} = \widehat{f}$ — и сеточной функции u_s — сужения функции u на сетку S (индекс N сетки опущен). Различие оценивается по норме, т.е. величиной $|u_s - \widehat{u}|$. Как отмечалось, об этой величине можно судить по невязке $\widehat{L}(u_s - \widehat{u}) = \widehat{L}u_s - \widehat{f}$. В самом деле, в данном контексте устойчивость означает существование обратного оператора $\widehat{L}^{-1}$, являющегося непрерывным. Поскольку в рассматриваемой ситуации (краевая задача для уравнения теплопроводности с $K = \mathrm{const}$) оператор $\widehat{L}$ является линейным, непрерывность равносильна ограниченности. Здесь мы имеем

$$|u_s - \widehat{u}| \leqslant \|\widehat{L}^{-1}\|\,|\widehat{L}u_s - \widehat{f}|,$$

и скорость сходимости к нулю невязки такая же, как и скорость сходимости разности $|u_s - \widehat{u}|$.

Учитывая метод дискретизации задачи, заключаем, что замена дифференциального уравнения разностным происходит с порядком аппроксимации $O(\tau + h^2)$. Это верно для любого смешанного варианта разностной схемы, но при $\alpha = 1/2$ из-за дополнительной симметрии порядок аппроксимации равен $O(\tau^2 + h^2)$.

Однако в краевой задаче кроме дифференциального уравнения имеются граничные условия. В данном случае на левом конце краевое условие дискретной задачи в точности соответствует условию непрерывной задачи. А на правом конце в граничное условие входит первая производная, которая при дискретизации заменяется разностью. Порядок такой аппроксимации $O(h)$. Это хуже, чем порядок аппроксимации дифференциального уравнения, что, разумеется ухудшает качество разностной схемы. На практике используют более сложные аппроксимации граничного условия второго рода, обеспечивающие порядок $O(h^2)$. Мы такие аппроксимации рассматривать не будем, тем более что погоня за высоким порядком аппроксимации — палка о двух концах: поскольку дискретизация позволяет считать значения искомой функции только в узловых точках, значения в других точках вычисляют методами интерполяции. Точность интерполяции напрямую зависит от того, насколько плавно меняется функция. Выигрывая в высоком порядке аппроксимации на разностной схеме, мы можем все потерять на интерполяции.

Решение разностной задачи. Метод решения разностной задачи зависит от того, какой вариант разностной схемы выбран. В случае явной схемы речь идет о простом пересчете, так

как значения сеточной функции на очередном слое явно выражаются через значения сеточной функции на предыдущем слое. Этим явная схема привлекательна, но при этом (как и для смешанных схем с $\alpha < 1/2$) необходимо обеспечивать устойчивость соответствующим подбором параметров h и τ. В нелинейном случае трудность состоит в том, что такой подбор нельзя обеспечить заранее, так как в условие устойчивости входит неизвестная функция (через K). Возможен вариант с меняющимся от слоя к слою значением τ: вычислив очередной слой, мы можем рассчитать значения K_i^j и найти среди них максимальное. Используя найденное значение, можно рассчитать предельное значение τ для текущего слоя, при котором разностная схема сохраняет устойчивость. Подобный подход может привести к слишком малым расстояниям между временными слоями и в конечном счете оказаться неэффективным. Его можно признать удовлетворительным, когда зависимость K от u невысокая.

В случае смешанной или неявной схемы решение разностной задачи приводит к решению системы уравнений при переходе от слоя к слою. Рассмотрим сначала случай $K = \text{const}$. В этом случае переход от слоя к слою связан с решением системы линейных уравнений, определяемых операторным уравнением

$$(E - \tau\alpha\Lambda)\boldsymbol{v}^{j+1} = (E + \tau(1-\alpha)\Lambda)\boldsymbol{v}^j$$

относительно вектора неизвестных $\boldsymbol{v}^{j+1}$. Это уравнение связывает значения сеточной функции во внутренних узлах сетки, и к нему необходимо добавить два граничных условия. В результате получаем систему линейных уравнений

$$\begin{cases} v_0^j = \mu^j, \\ \big((E - \tau\alpha\Lambda)\boldsymbol{v}^{j+1}\big)_i = \big(E + \tau(1-\alpha)\Lambda)\boldsymbol{v}^j\big)_i, \quad i = \overline{1,\, N-1}, \\ -v_{N-1} + V_N = \dfrac{h}{K} W^j, \end{cases}$$

где

$$(\Lambda \boldsymbol{f})_i = \frac{K}{h^2}\big(f_{i-1} - 2f_i + f_{i+1}\big).$$

Положим $\rho = \frac{K\tau}{h^2}$, $b_0^j = \mu^j$, $b_i^j = \big(E + \tau(1-\alpha)\Lambda)\boldsymbol{v}^j\big)_i$, $i = \overline{1,\, N-1}$, $B_N = \frac{h}{K} W^j$. Тогда система линейных уравнений будет иметь вид $A\boldsymbol{v}^{j+1} = \boldsymbol{b}^j$ с трехдиагональной матрицей

$$A = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -\alpha\rho & 1+2\alpha\rho & -\alpha\rho & \dots & 0 & 0 & 0 \\ 0 & -\alpha\rho & 1+2\alpha\rho & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\alpha\rho & 1+2\alpha\rho & -\alpha\rho \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}$$

Системы с трехдиагональной матрицей решают методом прогонки (он описан ниже).

В нашем случае K зависит от неизвестной функции. В разностное уравнение смешанной схемы для j-го слоя входят параметры K_{i+}^{j+1} и K_{i-}^{j+1}, выражающиеся через неизвестные значения v_i^{j+1}. Для решения получающейся системы нелинейных уравнений можно использовать следующую итерационную процедуру. На первом шаге в качестве приближений параметров K_{i+}^{j+1} и K_{i-}^{j+1} используем их аналоги j-го слоя K_{i+}^{j} и K_{i-}^{j}. Это превращает систему уравнений в линейную с соответствующей трехдиагональной матрицей, которую можно решить методом прогонки. Решением системы будет сеточная функция $\boldsymbol{v}^{j,1}$, близкая к $\boldsymbol{v}^{j+1}$. С помощью $\boldsymbol{v}^{j,1}$ уточним значения параметров K_{i+}^{j+1} и K_{i-}^{j+1} и снова решим линейную систему, получив сеточную функцию $\boldsymbol{v}^{j,2}$. Опять уточняем значения и решаем систему и т.д.

Поскольку в процессе решения разностной задачи мы получаем лишь приближенное решение, причем, как правило, невысокой точности (две-три значащие цифры), то в рамках этого

итерационного процесса также не следует добиваться высокой точности. Практически достаточно провести две-три итерации.

Метод прогонки. Систему линейных уравнений с трехдиагональной матрицей можно записать в следующем виде

$$\begin{cases} v_0 - \varkappa_1 v_1 = \mu_1, \\ a_i v_{i-1} - c_i v_i + b_i v_{i+1} = -f_i, \quad i = \overline{1,\, N-1}, \\ -\varkappa_2 v_{N-1} + v_n = \mu_2. \end{cases} \tag{3.7}$$

Для решения такой системы используют ***метод прогонки***, который представляет собой частный случай метода Гаусса.

Из первого уравнения выразим v_0 и подставим во второе уравнение. Получим уравнение, связывающее переменные v_1 и v_2. Из этого уравнения выразим v_1 и подставим в третье уравнение и т.д. На каждом шаге мы получаем уравнение

$$v_i = \alpha_{i+1} v_{i+1} + \beta_{i+1}. \tag{3.8}$$

Коэффициенты α_i и β_i можно определить по соответствующим формулам, которые можно вывести следующим образом. Предположим, что на $(i-1)$-м шаге получено уравнение $v_{i-1} = \alpha_i v_i + \beta_i$. Подставим это представление v_{i-1} в основное уравнение системы $a_i v_{i-1} - c_i v_i + b_i v_{i+1} = -f_i$. В результате получим

$$(\alpha_i a_i - c_i) v_i + b_i v_{i+1} + f_i + \beta_i a_i = 0,$$

откуда

$$v_i = \frac{b_i}{c_i - \alpha_i a_i} v_{i+1} + \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}.$$

Таким образом, сравнивая с (3.8), заключаем, что

$$\alpha_{i+1} = \frac{b_i}{c_i - \alpha_i a_i}, \qquad \beta_{i+1} = \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}, \qquad i = \overline{1,\, N-1}.$$

На первом шаге $\alpha_1 = \varkappa_1$ и $\beta_1 = \mu_1$.

На последнем $(N-1)$-м шаге получаем уравнение $v_{N-1} = \alpha_N v_N + \beta_N$. Значение v_{N-1} из него подставляем в последнее уравнение системы. Получим $-\varkappa_2(\alpha_N v_N + \beta_N) + v_N = \mu_2$, откуда

$$v_N = \frac{\mu_2 + \beta_N \varkappa_2}{1 - \alpha_N \varkappa_2}.$$

Найденное значение V_N затем позволяет по формулам (3.8) найти остальные значения неизвестных.

Система (3.7) определена, т.е. имеет и притом единственное решение, если выполняются соотношения

$$|\varkappa_i| \geqslant 1, \quad i = 1,\, 2, \quad |\varkappa_1| + |\varkappa_2| > 2,$$
$$|c_i| \geqslant |a_i| + |b_i|, \quad i = \overline{1,\, N-1}.$$

(в этом случае говорят, что матрица имеет диагональное преобладание).

3.4. Одномерное волновое уравнение

Рассмотрим задачу малых колебаний однородной струны длины l, для которой заданы начальное положение и начальные скорости, а также законы движения концов струны. В рамках математической физики такая задача формулируется следующим образом:

$$\begin{cases} u_{tt} = a^2 u_{xx}, \quad 0 < x < l, \ t > 0; \\ u\big|_{t=0} = \varphi(x), \quad u_t\big|_{t=0} = \psi(x), \\ u\big|_{t=0} = \mu(t), \quad u\big|_{x=l} = \nu(t). \end{cases}$$

Для решения этой задачи сеточным методом выберем равномерную прямоугольную сетку с узлами $(x_i,\, t_j)$, $i = \overline{0,\, N}$, $j = \overline{0,\, M}$, где $x_i = ih$, $t_j = j\tau$, $h = \dfrac{l}{N}$, $\tau = \dfrac{T}{M}$. Частные производные заменим соответствующими конечными разностями. В результате дифференциальное уравнение заменится разностным уравнением

$$\frac{u_i^{j+1} - 2u_i^j + u_i^{j-1}}{\tau^2} = a^2 \frac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{h^2}, \quad i = \overline{1,\, N-1},\ \ j = \overline{1,\, M-1},$$

а начальное условие $u_t\big|_{t=0} = \psi(x)$ — разностным соотношением

$$\frac{u_i^1 - u_i^0}{\tau} = \psi_i, \quad i = \overline{1,\, N-1}.$$

Другое начальное условие и граничные условия в разностной задаче реализуются точно:

$$u_i^0 = \varphi_i, \quad i = \overline{0,\, N}; \quad u_0^j = \mu^j,\ \ u_N^j = \nu^j,\ \ j = \overline{1,\, M}.$$

Получена полная система уравнений, связывающая значения сеточной функции u_i^j в узлах выбранной сетки. Эта система, как и в случае уравнения теплопроводности может решаться послойно. По начальному положению струны определяются значения сеточной функции на нулевом слое, т.е. при $j = 0$. По начальным скоростям определяются значения сеточной функции на первом слое. Наконец, по разностному уравнению можно вычислить значения сеточной функции во внутренних узлах $(j+1)$-го слоя по уже известным значениям двух предыдущих слоев. Значения в граничных узлах $(j+1)$-го слоя находятся из граничных условий.

Учитывая вид разностного уравнения, заключаем, что полученная разностная схема явная. Разностное уравнение связывает пять узлов сетки: узел $(x_i,\, t_j)$ и четыре прилегающих узла. Соответствующий шаблон называется „крестом“ (рис. 3.4). Как и в случае уравнения теплопроводности, явная разностная схема оказывается условно устойчивой, т.е. она устойчива только при определенных соотношениях шагов h и τ сетки.

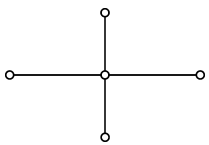

Рис. 3.4

Аппроксимация производных конечными разностями дает порядок аппроксимации $O(h^2+\tau^2)$ для дифференциального уравнения и $O(h^2 + \tau)$ для второго начального условия (начальных скоростей). Чтобы повысить порядок аппроксимации начального условия до такого же, как и порядок аппроксимации дифференциального уравнения, можно использовать два примыкающих узла сетки. В соответствии с формулой Тейлора

$$\frac{u_i^1 - u_i^0}{\tau} = \psi_i + \frac{\tau}{2} u_{tt}(x_i, t_j) + O(\tau^2).$$

Заменив неизвестное значение u_{tt} с помощью дифференциального уравнения на $a^2 u_{xx}$, а затем частную производную по x — второй конечной разностью, получим следующий вариант аппроксимации второго начального условия с порядком $O(h^2 + \tau^2)$:

$$\frac{u_i^1 - u_i^0}{\tau} = \psi_i + \frac{\tau}{2} \frac{u_{i-1}^0 - 2u_i^0 + u_{i+1}^0}{h^2}.$$

Устойчивость разностной схемы „крест“. Для исследования полученной разностной схемы модифицируем ее так, чтобы граничные условия были однородными. Это равносильно замене в разностном уравнении граничных узлов известными значениями. После этого в разностном уравнении граничные узлы будут отсутствовать, что равносильно нулевым значениям сеточной функции в них. Но при этом разностное уравнение станет неоднородным:

$$\frac{u_i^{j+1} - 2u_i^j + u_i^{j-1}}{\tau^2} = a^2 \frac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{h^2} + f_i^j, \quad i = \overline{1,\, N-1},\ \ j = \overline{1,\, M-1},$$

где

$$f_i^j = \begin{cases} \dfrac{a^2\mu^j}{h^2}, & i = 1; \\ 0, & 1 < i < N-1; \\ \dfrac{a^2\nu^j}{h^2}, & i = N-1. \end{cases}$$

Далее, введем обозначение $\boldsymbol{u}^j$ для j-го временного слоя. Тогда разностное уравнение примет вид

$$\frac{\boldsymbol{u}^{j+1} - 2\boldsymbol{u}^j + \boldsymbol{u}^{j-1}}{\tau^2} = \frac{a^2}{h^2}\Lambda\boldsymbol{u}^j + \boldsymbol{f}^j,$$

где $(\Lambda\boldsymbol{u})_i = u_{i-1} - 2u_i + u_{i+1}$, а $\boldsymbol{f}^j$ — временной слой сеточной функции f_i^j.

Поскольку разностное уравнение линейное, задача оценки погрешности решения разностной задачи распадается на две: задачу учета погрешности начальных условий, которые можно представить как ошибки $\delta\boldsymbol{u}^0$ и $\delta\boldsymbol{u}^1$ на первых двух слоях, вычисляемых с помощью начальных условий, и задачу учета ошибок $\delta\boldsymbol{f}^j$ сеточной функции $\boldsymbol{f}^j$, вытекающих из ошибок граничных условий. Рассмотрим первую из этих задач. Считаем, что сеточная функция задана точно, а начальные условия имеют ошибки $\delta\boldsymbol{u}^0$ и $\delta\boldsymbol{u}^1$. Тогда ошибки $\delta\boldsymbol{u}^j$ связаны однородным разностным уравнением

$$\frac{\delta\boldsymbol{u}^{j+1} - 2\delta\boldsymbol{u}^j + \delta\boldsymbol{u}^{j-1}}{\tau^2} = \frac{a^2}{h^2}\Lambda\delta\boldsymbol{u}^j, \quad j = \overline{1,\, M-1}. \tag{3.9}$$

Для решения поставленной задачи выберем такой базис $\boldsymbol{e}^i$ в линейном пространстве сеточных функций $\boldsymbol{u}^j$ с нулевыми значениями в граничных узлах, в котором оператор Λ имеет диагональную матрицу. Если ввести скалярное произведение

$$(\boldsymbol{u},\, \boldsymbol{v})_h = \sum_{i=1}^{N-1} u_i v_i h,$$

то оператор Λ будет самосопряженным, а выбранный базис $\boldsymbol{e}^i$ — ортогональным. При обсуждении уравнения теплопроводности было показано, как найти собственные векторы $\boldsymbol{e}^k$. В данном случае мы можем считать, что

$$(\boldsymbol{e}^k)_j = \sin\frac{k\pi j}{N}, \quad j = \overline{1,\, N-1}, \quad k = \overline{1,\, N-1}.$$

Непосредственный подсчет показывает, что

$$|\boldsymbol{e}^k|^2 = h\sum_{j=1}^{N-1} \sin^2\frac{k\pi j}{N} = \frac{Nh}{2} = \frac{l}{2},$$

где l — длина струны. При этом векторам $\boldsymbol{e}^k$, $k = \overline{1,\, N-1}$, соответствуют собственные значения $\lambda_k = -4\sin^2\dfrac{k\pi}{2N}$.

Разностное уравнение в силу диагональности матрицы оператора Λ распадается на независимые одномерные уравнения

$$\frac{v_i^{(j+1)} - 2v_i^{(j)} + v_i^{(j-1)}}{\tau^2} = \frac{a^2}{h^2}\lambda_i v_i^{(j)}, \quad i = \overline{1,\, N-1}, \tag{3.10}$$

в которых $v_i^{(j)}$, $i = \overline{1,\, N-1}$, — координаты сеточной функции $\delta\boldsymbol{u}^j$ в базисе $\boldsymbol{e}^i$. Решение одномерного разностного уравнения (3.10) ищем в виде сеточной функции q^j (степенная функция — по аналогии с линейным дифференциальным уравнением второго порядка). Подставив q^j вместо $v_i^{(j)}$, получим

$$\frac{q^{j+1} - 2q^j + q^{j-1}}{\tau^2} = \frac{a^2}{h^2}\lambda_i q^j,$$

или после сокращения на q^{j-1}

$$q^2 - \Big(2 + \frac{a^2\tau^2}{h^2}\lambda_i\Big)q + 1 = 0. \tag{3.11}$$

Полученное уравнение имеет два корня, дающее два независимых решения разностного уравнения второго порядка. Общее решение разностного уравнения получается в виде линейной комбинации $C_1 q_1^j + C_2 q_2^j$, где q_{i1}, q_{i2} — корни квадратного уравнения (3.11). Постоянные C_1 и C_2 определяются по первым двум известным значениям функции $v_i^{(j)}$ (т.е. при $j = 0$ и $j = 1$). Поскольку $v_i^{(j)}$ — координаты сеточной функции j-го слоя в ортогональном базисе, норма сеточной функции $\delta\boldsymbol{u}^j$ определяется по формуле

$$\|\delta\boldsymbol{u}^j\|^2 = \sum_{i=1}^{N-1} |v_i^j|^2 |\boldsymbol{e}^i|^2 = \frac{l}{2}\sum_{i=1}^{N-1} |v_i^j|^2.$$

Нетрудно заметить, что последовательность $\|\delta\boldsymbol{u}^j\|$ остается ограниченной при возрастании j в том и лишь в том случае, когда и q_{i1}, и q_{i2} для любого i по модулю не превышают единицы. Однако из квадратного уравнения (3.11) вытекает, что $q_1 q_2 = 1$. Оба условия будут выполняться тогда, когда корни q_{i1} и q_{i2} являются комплексными, т.е. при

$$-2 \leqslant 2 + a^2\frac{\tau^2}{h^2}\lambda_i \leqslant 2.$$

Правое неравенство выполняется, поскольку линейный оператор Λ имеет только отрицательные собственные значения. А левое неравенство означает, что

$$\frac{a^2\tau^2}{h^2} \leqslant \frac{4}{|\lambda_i|}$$

для каждого собственного значения λ_i оператора Λ. Чтобы это имело место, необходимо и достаточно выполнения неравенства

$$\frac{a^2\tau^2}{h^2} \leqslant \frac{4}{|\lambda_{\max}|},$$

где $\lambda_{\max}$ — максимальное по модулю собственное значение оператора Λ. Из анализа этого оператора, проведенного выше, следует, что $|\lambda_{\max}| = 4\cos^2\dfrac{\pi}{2N}$ близко к 4 и при росте N стремится к 4. Поэтому рассматриваемая разностная схема устойчива, если

$$\frac{a^2\tau^2}{h^2} \leqslant 1.$$

Найденное условие означает, что для любого собственного значения λ_i оба корня q_{i1} q_{i2} по модулю не превышают единицы. Это необходимое условие устойчивости: если оно не выполняется, то некоторые координаты, а следовательно, и норма, неограниченно возрастают от слоя к слою. Убедимся в том, что это условие является и достаточным.

Используя решения q_{i1} и q_{i2} уравнения (3.11), заключаем, что

$$v_i^j = C_{i1}q_{i1}^j + C_{i2}q_{i2}^j,$$

где постоянные C_{i1} и C_{i2} определяются из системы уравнений

$$\begin{cases} C_{i1} + C_{i2} = v_i^0, \\ C_{i1}q_{i1} + C_{i2}q_{i2} = v_i^1. \end{cases}$$

Из этой системы находим

$$C_{i1} = \frac{q_{i2}v_i^{(0)} - v_i^{(1)}}{q_{i2} - q_{i1}}, \quad C_{i2} = -\frac{q_{i1}v_i^{(0)} - v_i^{(1)}}{q_{i2} - q_{i1}}.$$

Следовательно,

$$v_i^{(j)} = v^{(1)}\frac{q_{i2}^j - q_{i1}^j}{q_{i2} - q_{i1}} - v^{(0)}\frac{q_{i2}^{j-1} - q_{i1}^{j-1}}{q_{i2} - q_{i1}}.$$

При большом количестве временных слоев величины $|q_{i2}^j - q_{i1}^j|$ и $|q_{i2}^{j-1} - q_{i1}^{j-1}|$ могут принимать практически любое значение от 0 до 2. Наилучшей оценкой в данном случае является

$$|v_i^{(j)}| \leqslant 2\frac{|v_i^{(1)}| + |v_i^{(0)}|}{|q_{i2} - q_{i1}|} = \frac{|v_i^{(1)}| + |v_i^{(0)}|}{\dfrac{2a\tau}{h}\sin\dfrac{i\pi}{2N}\sqrt{1 - \dfrac{a^2\tau^2}{h^2}\sin^2\dfrac{i\pi}{2N}}} \leqslant \frac{|v_i^{(1)}| + |v_i^{(0)}|}{\dfrac{2a\tau}{h}\sin\dfrac{i\pi}{2N}\sqrt{1 - \sin^2\dfrac{i\pi}{2N}}} \leqslant 2\frac{|v_i^{(1)}| + |v_i^{(0)}|}{\dfrac{a\tau}{h}\sin\dfrac{\pi}{N}}.$$

Суммируя с учетом неравенства $(x+y)^2 \leqslant 2x^2 + 2y^2$, заключаем, что

$$\|\delta\boldsymbol{u}^j\|^2 \leqslant \frac{2h^2\big(\|\delta\boldsymbol{u}^0\|^2 + \|\delta\boldsymbol{u}^1\|^2\big)}{a^2\tau^2\sin^2\dfrac{\pi}{N}}.$$

Если $\|\delta\boldsymbol{u}^0\| < \varepsilon$ и $\|\delta\boldsymbol{u}^1\| < \varepsilon$, то

$$\|\delta\boldsymbol{u}^j\| \leqslant \frac{2h\varepsilon}{a\tau\sin\dfrac{\pi}{N}}.$$

Мы видим, что с уменьшением шага сетки (увеличением N) погрешность, связанная с погрешностями начальных данных растет, хотя и не так быстро, как геометрическая прогрессия: рост составляет порядка N, или порядка $1/h$. С учетом порядка невязки $O(h^2 + \tau^2)$ приходим к выводу, что с измельчением сетки погрешность приближенного решения стремится к нулю. Отметим также, что занижение величины $\dfrac{a\tau}{h}$ приводит к увеличению погрешности. Оптимальный вариант — значение, равное единице или чуть меньше единицы. В этом случае можно пользоваться приближенной формулой

$$\|\delta\boldsymbol{u}^j\| \approx \frac{2\varepsilon}{\sin\dfrac{\pi}{N}} \approx \frac{2l\varepsilon}{\pi h}.$$

Теперь помимо погрешностей начальных условий учтем погрешности в граничных условиях. Уравнение на погрешности будет иметь вид

$$\frac{\delta\boldsymbol{u}^{j+1} - 2\delta\boldsymbol{u}^j + \delta\boldsymbol{u}^{j-1}}{\tau^2} = \frac{a^2}{h^2}\Lambda\delta\boldsymbol{u}^j + \delta\boldsymbol{f}^j.$$

Из этого уравнения находим

$$\delta\boldsymbol{u}^{j+1} = \Big(2E + \frac{\tau^2 a^2}{h^2}\Lambda\Big)\delta\boldsymbol{u}^j - \delta\boldsymbol{u}^{j-1} + \tau^2\delta\boldsymbol{f}^j.$$

Это уравнение отличается от (3.9) лишь дополнительным слагаемым, которое можно рассматривать как часть ошибки, имеющейся в сеточной функции на $(j-1)$-м слое. Используя сеточные функции $\delta\boldsymbol{u}^j$ и $\delta\boldsymbol{u}^{j-1} + \tau^2\delta\boldsymbol{f}^j$ как начальные условия для разностной задачи, заключаем, что на k-м слое погрешность $\tau^2\delta\boldsymbol{f}^j$ приведет к дополнительной погрешности порядка $\frac{h\tau\|\delta\boldsymbol{f}^j\|}{a\sin(\pi/N)} \approx \frac{\tau}{\pi}\|\delta\boldsymbol{f}^j\|$. На этом слое складываются погрешности, возникающие на всех предыдущих слоях, что (при условии $\|\delta\boldsymbol{f}^j\| < \varepsilon$) приводит к дополнительной суммарной погрешности порядка $\frac{T}{\pi}\varepsilon$ (T — период веремени, на котором рассматривается разностная аппроксимация).

Все проведенные подсчеты показывают, что, несмотря на фактическое отсутствие устойчивости разностной схемы, при $N \to \infty$ решение разностной задачи в узлах сетки стремится к решению краевой задачи, т.е. разностная схема обладает свойством сходимости.

3.5. Уравнение Пуассона

Рассмотрим краевую задачу

$$\begin{aligned}
&\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y} + f(x,y) = 0, \quad 0 < x < a,\ 0 < y < b,\\
&u\big|_{x=0} = \mu(y), \qquad u\big|_{x=a} = \nu(y),\\
&u\big|_{y=0} = \varphi(x), \qquad u\big|_{y=b} = \psi(x).
\end{aligned}$$

Выберем сетку $(x_i,\, y_j)$, где $x_i = ih_x$, $i = \overline{0,\, N}$, $y_j = jh_y$, $j = \overline{0,\, N}$. Разностную схему построим, аппроксимируя частные производные 2-го порядка вторыми разностями. Пусть $u_{ij} = u(x_i, y_j)$, $f_{ij} = f(x_i, y_j)$. Тогда

$$\begin{aligned}
&\frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h_x^2} + \frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{h_y^2} + f_{ij} = 0, \qquad (3.12)\\
&u_{0j} = \mu_j, \quad u_{Nj} = \nu_j, \quad u_{i0} = \varphi_i, \quad u_{iN} = \psi_i.
\end{aligned}$$

Здесь неизвестными являются значения сеточной функции во внутренних узлах, т.е. при $i,\, j = \overline{1,\, N-1}$, поскольку в граничных узлах значения сеточной функции даны. Порядок аппроксимации в данном случае равен $O(h_x^2 + h_y^2)$.

Каждое из уравнений разностной задачи связывает текущий узел $(i,\, j)$ сетки с четырьмя ближайшими узлами. Шаблон разностной схемы, называемой ***разностной схемой „крест“***, показан на рис. 3.4.

Этот шаблон совпадает с шаблоном разностной схемы для волнового уравнения, но эти разностные схемы различаются принципиально: в случае волнового уравнения решение можно было проводить послойно, в то время как в случае уравнения Пуассона это невозможно.

Анализ разностной схемы. Как и в других случаях, проведем анализ разностной схемы на устойчивость. Заключение о сходимости разностной схемы можно будет сделать исходя и ее устойчивости и порядка аппроксимации.

Так как аппроксимация привела к системе линейных уравнений, то анализ на устойчивость сводится к анализу свойств линейного оператора $L = \Lambda_x + \Lambda_y$, где

$$(\Lambda_x u)_{ij} = \frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h_x^2}, \quad (\Lambda_y u)_{ij} = \frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{h_x^2}.$$

Из системы линейных уравнений удалим граничные условия, подставив их непосредственно в уравнения для приграничных узлов (т.е. когда один из индексов 1 или $N-1$). В результате получим систему уравнений, соответствующую однородной разностной задаче. В этом случае ищется сеточная функция, принимающая в граничных узлах нулевые значения. Множество таких функций есть линейное пространство размерности $(N-1)^2$. На этом линейном пространстве введем скалярное произведение по формуле

$$(\boldsymbol{u},\, \boldsymbol{v}) = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} u_{ij}v_{ij}h_x h_y.$$

В этом случае стандартный базис (т.е. система сеточных функций, которые в одном узле имеют значение 1, а в остальных 0). будет представлять собой ортогональный базис, причем квадрат нормы любой базисной функции будет равен $h_x h_y$.

Можно показать, что в рассматриваемом евклидовом пространстве линейный оператор L является самосопряженным, непосредственно используя определение самосопряженного оператора. Проще однако записать матрицу этого оператора в стандартном базисе, установив порядок среди векторов базиса. Обозначим через $\boldsymbol{e}^{ij}$ базисную функцию, принимающую значение 1 в узле $(i,\, j)$. Установим следующий порядок базисных функций:

$$\boldsymbol{e}^{11},\ \boldsymbol{e}^{12},\ \ldots,\ \boldsymbol{e}^{1,N-1},\ \boldsymbol{e}^{21},\ \ldots,\ \boldsymbol{e}^{N-1,N-1}.$$

В матрице оператора в строке, соответствующей узлу $(i,\, j)$, т.е. в строке с номером $(i-1)*(N-1)+j$ диагональным элементом будет коэффициент, который в уравнении (3.12) соответствует узловому значению u_{ij}, коэффициенты при $u_{i,j-1}$ и $u_{i,j+1}$ будут расположены рядом с диагональным левее и правее, а коэффициенты при $u_{i-1,j}$ и $u_{i+1,j}$ будут отстоять от диагонального влево и вправо на $N-1$ мест. Такую матрицу удобно записать как блочно-трехдиагональную, состоящую из квадратных блоков порядка $N-1$:

$$[L] = \begin{pmatrix}
A & \frac{1}{h_x^2}E & 0 & \ldots & 0\\
\frac{1}{h_x^2}E & A & \frac{1}{h_x^2}E & \ldots & 0\\
0 & \frac{1}{h_x^2}E & A & \ldots & 0\\
\ldots & \ldots & \ldots & \ldots & \ldots\\
0 & 0 & 0 & \ldots & A
\end{pmatrix},$$

где

$$A = \begin{pmatrix}
-\frac{2}{h_x^2} - \frac{2}{h_y^2} & \frac{1}{h_y^2} & 0 & \ldots & 0\\
\frac{1}{h_y^2} & -\frac{2}{h_x^2} - \frac{2}{h_y^2} & \frac{1}{h_y^2} & \ldots & 0\\
0 & \frac{1}{h_y^2} & -\frac{2}{h_x^2} - \frac{2}{h_y^2} & \ldots & 0\\
\ldots & \ldots & \ldots & \ldots & \ldots\\
0 & 0 & 0 & \ldots & -\frac{2}{h_x^2} - \frac{2}{h_y^2}
\end{pmatrix},$$

а E — единичная матрица порядка $N-1$.

Поскольку полученная матрица линейного оператора в ортогональном базисе симметрична, этот оператор является самосопряженным*.

*Строго говоря, требуется, чтобы базис был ортонормированным. Однако в данном случае все векторы базиса имеют одинаковую норму. Нетрудно проверить, что правило „самосопряженный, если матрица симметрична“ остается верным.

Собственные функции рассматриваемого линейного оператора можно найти, проведя аналогию с непрерывным случаем. В данном случае собственные функции можно искать в виде $\boldsymbol{g}_{ij} = \sin\omega_x i \,\sin\omega_y j$. Учет однородных граничных условий позволяет найти частоты ω_x и ω_y. В результате получаем систему функций $\boldsymbol{g}^{kl}$, $k,\, l = \overline{1,\, N-1}$, со значениями

$$\boldsymbol{g}^{kl}_{ij} = \sin\frac{k\pi i}{N}\,\sin\frac{l\pi j}{N}.$$

Все эти функции попарно ортогональны, как собственные функции, отвечающие различным собственным значениям. Они имеют одинаковые нормы:

$$\|\boldsymbol{g}^{kl}\|^2 = h_x h_y \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\sin^2\frac{k\pi i}{N}\,\sin^2\frac{l\pi j}{N} = h_x h_y\left(\sum_{i=1}^{N-1}\sin^2\frac{k\pi i}{N}\right)\left(\sum_{j=1}^{N-1}\sin^2\frac{l\pi j}{N}\right) = \frac{N^2}{4}h_x h_y = \frac{ab}{4}.$$

Собственной функции $\boldsymbol{g}^{kl}$ отвечает собственное значение

$$\lambda_{kl} = -\frac{4}{h_x^2}\sin^2\frac{k\pi}{2N} - \frac{4}{h_y^2}\sin^2\frac{l\pi}{2N}.$$

Отметим минимальное $\lambda_{\min}$ и максимальное $\lambda_{\max}$ собственные значения:

$$\lambda_{\min} = -\left(\frac{4}{h_x^2}+\frac{4}{h_x^2}\right)\sin^2\frac{\pi}{2N}, \qquad \lambda_{\max} = -\left(\frac{4}{h_x^2}+\frac{4}{h_x^2}\right)\cos^2\frac{\pi}{2N}.$$

Рассматриваемую разностную задачу можно интерпретировать как операторное уравнение $L\boldsymbol{u} + \boldsymbol{f} = 0$. Для оценки устойчивости разностной схемы нужно оценить норму оператора L^{-1}, поскольку ошибка правой части $\delta\boldsymbol{f}$ и ошибка решения $\delta\boldsymbol{u}$ связаны соотношением $\delta\boldsymbol{u} = -L^{-1}\delta\boldsymbol{f}$, а их нормы — неравенством $\|\delta\boldsymbol{u}\| \leqslant \|L^{-1}\|\,\|\delta\boldsymbol{f}\|$.

Известно, что норма самосопряженного оператора L совпадает с максимумом модулей его собственных чисел, т.е. $\|L\| = \lambda_{\max}$. Линейный оператор L^{-1} также самосопряженный, а его максимальное по модулю собственное число есть $1/\lambda_{\min}$. Таким образом,

$$\|L^{-1}\| = 1/\lambda_{\min} = \frac{h_x^2 h_y^2}{4(h_x^2+h_y^2)\sin^2\frac{\pi}{2N}} = \frac{a^2b^2}{a^2+b^2}\,\frac{1}{4N^2\sin^2\frac{\pi}{2N}} \approx \frac{a^2b^2}{\pi^2(a^2+b^2)}.$$

Видно, что с ростом N норма оператора L^{-1} остается ограниченной. Это и означает, что построенная разностная схема является устойчивой. Поскольку порядок аппроксимации равен $O(1/N^2)$, заключаем, что при $N\to\infty$ разностная схема сходится со вторым порядком.

Замечание. Однако не все так хорошо, как кажется с первого взгляда. Схема абсолютно устойчива. Но приходится решать систему линейных уравнений $L\boldsymbol{u}+\boldsymbol{f}=0$. Качество системы линейных уравнений (ее чувствительность к ошибкам правых частей) определяется ***числом обусловленности*** $\operatorname{cond} L$, равным $\|L\|\,\|L^{-1}\|$. Увеличение числа обусловленности ведет к усилению влияния на результат ошибок округления, а в итерационном процессе — к увеличению числа итераций. В данном случае

$$\operatorname{cond} L = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} = \operatorname{ctg}^2\frac{\pi}{2N} \approx \frac{4N^2}{\pi^2}.$$

Методы решения системы линейных уравнений. Система линейных уравнений, возникающая при дискретизации краевой задачи, как правило, имеет большой порядок. Уже при $N = 50$ получаем систему с 2500 неизвестными. Для таких систем редко используют точные методы решения типа метода Гаусса. Предпочтение отдают итерационным методам.

Ряд итерационных методов решения системы $C\boldsymbol{u} = \boldsymbol{f}$, где $C = -L$, укладывается в единую схему. Составим так называемое эволюционное уравнение

$$B\frac{\boldsymbol{u}^{k+1}-\boldsymbol{u}^k}{\tau} + C\boldsymbol{u}^k = \boldsymbol{f}. \tag{3.13}$$

Нетрудно увидеть, что если $\boldsymbol{u}^k \to \boldsymbol{u}^*$ при $k\to\infty$, то левая часть уравнения стремится к $C\boldsymbol{u}^*$. Переход к пределу в равенстве приводит к соотношению $C\boldsymbol{u}^* = \boldsymbol{f}$, т.е. если ***итерационная последовательность*** $\{\boldsymbol{u}^k\}$ сходится, то ее пределом является решение системы $C\boldsymbol{u} = \boldsymbol{f}$.

Сходимость итерационной последовательности обеспечивается путем подходящего выбора матрицы B и коэффициента τ.

Из уравнения (3.13) можем в матричной форме выразить сеточную функцию $\boldsymbol{u}^{k+1}$:

$$\boldsymbol{u}^{k+1} = (E - \tau B^{-1}C)\boldsymbol{u}^k + \tau B^{-1}\boldsymbol{f}.$$

Пусть $\boldsymbol{u}^*$ — решение системы $C\boldsymbol{u} = \boldsymbol{f}$. Положим $\boldsymbol{x}^k = \boldsymbol{u}^k - \boldsymbol{u}^*$. Тогда $\boldsymbol{u}^k = \boldsymbol{x}^k + \boldsymbol{u}^*$. Подставив это представление в уравнение (3.13) и учитывая равенство $C\boldsymbol{u}^* = \boldsymbol{f}$, получаем

$$B\frac{\boldsymbol{x}^{k+1}-\boldsymbol{x}^k}{\tau} + C\boldsymbol{x}^k = 0,$$

откуда

$$\boldsymbol{x}^{k+1} = (E - \tau B^{-1}C)\boldsymbol{x}^k. \tag{3.14}$$

Из полученного равенства видно, что сходимость итерационного процесса связана со свойствами линейного оператора $T = E - \tau B^{-1}C$. Точнее, из равенства (3.14) вытекает, что

$$\boldsymbol{x}^k = T^k\boldsymbol{x}^0.$$

Достаточным условием сходимости к нулю последовательности $\{\boldsymbol{x}^k\}$ является сходимость к нулю последовательности норм $\{\|T^k\|\}$. Из курса линейной алгебры известно, что всегда существует предел

$$\rho(T) = \lim_{k\to\infty}\sqrt[k]{\|T^k\|},$$

называемый ***спектральным радиусом***, который совпадает с максимумом модулей корней характеристического уравнения линейного оператора* T. Следовательно, достаточным условием сходимости к нулю $\{\|\boldsymbol{x}^k\|\}$ является $\rho(T) < 1$, причем чем меньше $\rho(T)$, тем выше скорость сходимости. Таким образом, построение итерационных методов и их анализ сводится к анализу спектра линейного оператора $T = E - \tau B^{-1}C$, т.е. в данном случае совокупности его собственных значений.

Полагая $B = E$, приходим к ***методу простой итерации***. В этом методе $T_{\text{пр}} = E - \tau C$, а собственные значения μ_i оператора $T_{\text{пр}}$ связаны с собственными значениями λ_i оператора $-C = L$ равенствами $\mu_i = 1 + \tau\lambda_i$ (напомним, что все λ_i отрицательны). Следовательно, собственные значения μ_i располагаются левее 1 от $\mu_{\min}$, соответствующего собственному значению $\lambda_{\min}$ до самого левого значения $\mu_{\max}$, соответствующего значению $\lambda_{\max}$. Исходя из такого расположения собственных значений μ_i делаем вывод, что условие $\rho(T_{\text{пр}}) < 1$ равносильно неравенству $\mu_{\max} > -1$, или $1 + \tau\lambda_{\max} > -1$. таким образом, метод простой итерации сходится, если

$$\tau < \frac{2}{|\lambda_{\max}|}.$$

Варьируя параметр τ, можно не только добиться выполнения условия $\rho(T_{\text{пр}}) < 1$, но и оптимизировать метод простой итерации, обеспечив наименьшее значение спектрального радиуса.

*Это верно в случае евклидовой нормы и легко устанавливается для самосопряженного оператора.

Из простых соображений в данном случае нетрудно понять, что минимум $\rho(T_{\text{пр}})$ достигает при выполнении условия $-\mu_{\max} = \mu_{\min}$, равносильного равенству $-1 - \tau\lambda_{\max} = 1 + \tau\lambda_{\min}$. Отсюда находим

$$\tau = \frac{2}{|\lambda_{\max}| + |\lambda_{\min}|} = \frac{h_x^2 h_y^2}{2(h_x^2 + h_y^2)}.$$

Вычисления в методе простой итерации строятся согласно формуле $\boldsymbol{u}^{k+1} = (E - \tau C)\boldsymbol{u}^k + \tau\boldsymbol{f} = \boldsymbol{u}^k + \tau L\boldsymbol{u}^k + \tau\boldsymbol{f}$. В координатной форме это дает следующее равенство:

$$u_{ij}^{k+1} = \frac{h_x^2 h_y^2}{2(h_x^2 + h_y^2)}\left(\frac{u_{i-1,j}^k + u_{i+1,j}^k}{h_x^2} + \frac{u_{i,j-1}^k + u_{i,j+1}^k}{h_y^2}\right) + \frac{h_x^2 h_y^2}{2(h_x^2 + h_y^2)} f_{ij},$$

или

$$u_{ij}^{k+1} = \frac{h_y^2}{2(h_x^2 + h_y^2)}\big(u_{i-1,j}^k + u_{i+1,j}^k\big) + \frac{h_x^2}{2(h_x^2 + h_y^2)}\big(u_{i,j-1}^k + u_{i,j+1}^k\big) + \frac{h_x^2 h_y^2}{2(h_x^2 + h_y^2)} f_{ij}.$$

Матрицу C разделим на три части $C = S + D + R$, где S включает все элементы матрицы C под главной диагональю, D — элементы главной диагонали, $R = S^{\mathrm{T}}$ — элементы над главной диагональю. В ***методе Якоби*** полагают $B = D$ и $\tau = 1$. В рассматриваемом случае все диагональные элементы матрицы $C = -L$ одинаковы и равны $d = \frac{2}{h_x^2} + \frac{2}{h_y^2}$. Следовательно $D = dE$, а расчетная формула выглядит следующим образом:

$$\boldsymbol{u}^{k+1} = (E - d^{-1}C)\boldsymbol{u}^k + \boldsymbol{f}.$$

Видно, что эта формула есть частный случай метода простой итерации, в котором в качестве τ выбрано значение d^{-1}. Более того, в рассматриваемой разностной схеме d^{-1} совпадает с оптимальным значением $\tau_{\text{пр}}$ параметра метода простой итерации. Поэтому в данном случае метод Якоби совпадает с оптимизированным методом простой итерации.

В ***методе Зейделя*** полагают $B = S + D$, $\tau = 1$. Выбор в качестве матрицы B нижней треугольной облегчает ее обращение (на самом деле решение системы $B\boldsymbol{u}^{k+1} = (B - C)\boldsymbol{u}^k + \boldsymbol{f}$). Применение прямого хода метода Гаусса сразу дает решение и может быть реализовано в виде явных расчетных формул. В самом деле, имеем

$$S\boldsymbol{u}^{k+1} + D\boldsymbol{u}^{k+1} = -S^{\mathrm{T}}\boldsymbol{u}^k + \boldsymbol{f}.$$

Произведение $S\boldsymbol{u}^{k+1}$ содержит к моменту вычисления i-й компоненты вектора $\boldsymbol{u}^{k+1}$ только предыдущие компоненты. Перенося это произведение в правую часть как известное, получим

$$D\boldsymbol{u}_{k+1} = -S\boldsymbol{u}^{k+1} - S^{\mathrm{T}}\boldsymbol{u}^k + \boldsymbol{f}.$$

Сравним с формулой метода Якоби

$$D\boldsymbol{u}_{k+1} = -S\boldsymbol{u}^k - S^{\mathrm{T}}\boldsymbol{u}^k + \boldsymbol{f}.$$

Видно, что расчетные формулы метода Зейделя получаются из формул метода Якоби заменой k-го временного слоя во всех предшествующих узлах $(k+1)$-м. В исходной двухиндексной нумерации получаем

$$u_{ij}^{k+1} = \frac{h_y^2}{2(h_x^2 + h_y^2)}\big(u_{i-1,j}^{k+1} + u_{i+1,j}^k\big) + \frac{h_x^2}{2(h_x^2 + h_y^2)}\big(u_{i,j-1}^{k+1} + u_{i,j+1}^k\big) + \frac{h_x^2 h_y^2}{2(h_x^2 + h_y^2)} f_{ij}.$$

Метод Зейделя сходится быстрее, чем метод Якоби (он же оптимизированный метод простой итерации). Но оказывается, что метод Зейделя можно усилить, используя дополнительное значение u_{ij}^k, которое не вошло в формулы метода Зейделя. Это достигается введением в метод

Зейделя дополнительного параметра, позволяющего разделить матрицу D на две части, одна из которых относится к $(k+1)$-му слою (как в методе Зейделя), а вторая — к k-му слою. Положив $B = S + \frac{1}{\omega}D$, $\tau = 1$, придем к ***методу верхней релаксации***. Метод Зейделя вписывается в метод верхней релаксации как частный случай $\omega = 1$.

Подставив в уравнение $B\boldsymbol{u}^{k+1} - B\boldsymbol{u}^k + C\boldsymbol{u}^k = \boldsymbol{f}$ выбранный вариант матрицы B, получим

$$\left(S + \frac{1}{\omega}D\right)\boldsymbol{u}^{k+1} + \left(S^{\mathrm{T}} + \left(1 - \frac{1}{\omega}\right)D\right)\boldsymbol{u}^k = \boldsymbol{f}.$$

Учитывая, что произведение $S\boldsymbol{u}^{k+1}$ можно считать известным, выразим $\boldsymbol{u}^{k+1}$:

$$D\boldsymbol{u}^{k+1} = -\omega(S\boldsymbol{u}^{k+1} + S^{\mathrm{T}}\boldsymbol{u}^k) + \omega\boldsymbol{f} + (1 - \omega)D\boldsymbol{u}^k.$$

Таким образом,

$$\boldsymbol{u}^{k+1} = \omega D^{-1}\big(-S\boldsymbol{u}^{k+1} - S^{\mathrm{T}}\boldsymbol{u}^k + \boldsymbol{f}\big) + (1 - \omega)\boldsymbol{u}^k.$$

Первое слагаемое в правой части равенства (без ω) соответствует расчету по методу Зейделя. Расчет по методу верхней релаксации можно представить как двухступенчатый. На первой ступени вычисляем

$$\boldsymbol{u}^{k+1/2} = D^{-1}\big(-S\boldsymbol{u}^{k+1} - S^{\mathrm{T}}\boldsymbol{u}^k + \omega\boldsymbol{f}\big)$$

по формулам метода Зейделя, а затем на второй ступени получаем узловое значение на $(k+1)$-м слое:

$$\boldsymbol{u}^{k+1} = \omega\boldsymbol{u}^{k+1/2} + (1 - \omega)\boldsymbol{u}^k.$$

В координатной форме это дает следующие формулы:

$$u_{ij}^{k+1/2} = \frac{h_y^2}{2(h_x^2 + h_y^2)}\big(u_{i-1,j}^{k+1} + u_{i+1,j}^k\big) + \frac{h_x^2}{2(h_x^2 + h_y^2)}\big(u_{i,j-1}^{k+1} + u_{i,j+1}^k\big) + \frac{h_x^2 h_y^2}{2(h_x^2 + h_y^2)} f_{ij},$$

$$u_{ij}^{k+1} = \omega u_{ij}^{k+1/2} + (1 - \omega)u_{ij}^k.$$

Анализ итерационных методов. В методе Якоби (оптимизированном методе простой итерации) имеем

$$\rho(T_{\text{пр}}) = 1 + \tau_{\text{пр}}\lambda_{\min} = 1 - 2\sin^2\frac{\pi}{2N} = \cos\frac{\pi}{N}.$$

Отклонение $\boldsymbol{x}^k$ итерационной последовательности от точного решения оценивается неравенством $\|\boldsymbol{x}^k\| \leqslant \rho^k\|\boldsymbol{x}^0\|$. Записав $\rho^k\|\boldsymbol{x}^0\| \leqslant \varepsilon$, находим

$$k \geqslant \frac{\ln\big(\varepsilon/\|\boldsymbol{x}^0\|\big)}{\ln\rho}.$$

В методе Якоби

$$\ln\rho = \ln\cos\frac{\pi}{N} \approx -\frac{\pi^2}{2N^2}.$$

Значит, требуемое количество итераций составляет порядка $\frac{2K}{\pi^2}N^2$, где $K = \ln\big(\varepsilon/\|\boldsymbol{x}^0\|\big)$ отражает выбранные начальное приближение и требуемую точность, не связанные с конкретным итерационным методом.

Метод верхней релаксации (в том числе метод Зейделя) проанализировать можно, определив собственные числа матрицы $T_\omega = E - \tau B^{-1}C$ этого метода. Составим характеристическое уравнение $\det(T_\omega - \sigma E) = 0$ и пареобразуем его, используя вид матрицы T_ω:

$$\det(E - B^{-1}C - \sigma E) = \det\big((1 - \sigma)E - B^{-1}C\big) = \det B^{-1} \cdot \det\big((1 - \sigma)B - C\big) = 0.$$

Сомножитель $\det B^{-1}$ можно отбросить. Далее,

$$(1-\sigma)B - C = (1-\sigma)S + \frac{1-\sigma}{\omega}D - S - D - S^{\mathrm{T}} = \frac{1-\sigma-\omega}{\omega}D - \sigma S - S^{\mathrm{T}}.$$

Поэтому характеристическое уравнение имеет вид:

$$\det\left(\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}}D - \sqrt{\sigma}S - \frac{1}{\sqrt{\sigma}}S^{\mathrm{T}}\right) = 0.$$

Можно показать, что блочно-трехдиагональная матрица $\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}}D - \sqrt{\sigma}S - \frac{1}{\sqrt{\sigma}}S^{\mathrm{T}}$ подобна матрице $\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}}D - S - S^{\mathrm{T}}$. Поэтому характеристическое уравнение эквивалентно следующему:

$$\det\left(\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}}D - S - S^{\mathrm{T}}\right) = 0,$$

или

$$\det\left(\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}}E - D^{-1}(S + S^{\mathrm{T}})\right) = 0.$$

Отметим, что $-D^{-1}(S + S^{\mathrm{T}})$ есть матрица $T_{\text{пр}}$, поскольку

$$T_{\text{пр}} = E - d^{-1}C = d^{-1}(D - C) = D^{-1}(-S - S^{\mathrm{T}}).$$

Следовательно, число σ является корнем характеристического уравнения оператора T_ω тогда и только тогда, когда число $-\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}}$ является собственным значением оператора $T_{\text{пр}}$, т.е. совпадает с одним из μ_i:

$$\frac{1-\sigma-\omega}{\omega\sqrt{\sigma}} = -\mu_i.$$

Записанное уравнение является квадратным относительно $\sqrt{\sigma}$:

$$\sigma + \mu\omega\sqrt{\sigma} + \omega - 1 = 0$$

Решая его, получим зависимость σ от μ_i и ω. Для решения следует рассмотреть два случая: дискриминант Δ уравнения отрицательный и неотрицательный. Поскольку $\Delta = \mu^2\omega^2 - 4(\omega - 1)$, заключаем, что квадратное уравнение имеет действительные корни при $\mu^2\omega^2 \geqslant 4(\omega - 1)$ и комплексные при $\mu^2\omega^2 < 4(\omega - 1)$. Нас интересует не само значение σ, а его модуль. При отрицательном дискриминанте оба комплексных корня имеют одинаковые модули и при $\mu^2\omega^2 < 4(\omega - 1)$ имеем $|\sigma| = \omega - 1$. При $\mu^2\omega^2 \geqslant 4(\omega - 1)$ выбираем наибольший по модулю корень, что дает:

$$|\sigma| = \left(\frac{|\mu|\omega + \sqrt{\Delta}}{2}\right)^2.$$

Выясним, при каких значениях параметров μ и ω выполняется условие $|\sigma| < 1$. Если $\Delta < 0$, то $|\sigma| = (\omega - 1)^2$ и условие $|\sigma| < 1$ равносильно неравенствам $0 < \omega < 2$. Если $\Delta > 0$, то имеем неравенство

$$\left(\frac{|\mu|\omega + \sqrt{\Delta}}{2}\right)^2 < 1.$$

откуда $\sqrt{\Delta} < 2 - |\mu|\omega$. Возводя в квадрат с дополнительными условиями $\Delta > 0$ и $2 - |\mu|\omega > 0$ приходим к системе неравенств

$$\begin{cases} |\mu|\omega < 2, \\ |\mu|\omega < \omega, \\ \mu^2\omega^2 > 4(\omega - 1). \end{cases}$$

Из первого и третьего неравенств вытекает, что $\omega < 2$, второе неравенство оказывается сильнее первого, и мы приходим к следующим соотношениям:

$$\begin{cases} |\mu| < 1, \\ \mu^2\omega^2 > 4(\omega - 1). \end{cases}$$

В совокупности два случая (комплексных и действительных корней квадратного уравнения) приводят к следующему. Необходимыми и достаточными условиями выполнения неравенства $|\sigma| < 1$ являются $0 < \omega < 2$ и $|\mu| < 1$. На рис. 3.5 показана плоскость $(\omega,\, \mu)$, на которой затенена та область, в которой $\Delta < 0$, а пунктиром выделен квадрат $0 < \omega \leqslant 2$, $|\mu \leqslant 1$, в котором выполняется неравенство $|\sigma| < 1$. Обратим внимание на разделительную линию $\mu^2\omega^2 = 4(\omega - 1)$. Анализ показывает, что $|\sigma|(\omega, \mu)$ при фиксированном μ, $|\mu| \leqslant 1$, убывает слева от разделительной линии и возрастает справа. Значит, на каждой горизонтальной прямой точка разделительной линии есть точка минимума значения $|\sigma|$.

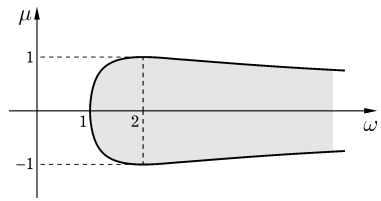

Рис. 3.5

По простой зависимости в затененной области заключаем, что на разделительной линии $|\sigma| = \omega - 1$. Наихудший вариант для $|\sigma|$ соответствует $\mu_{\max}$ (или $\mu_{\min}$). Поскольку $|\mu_{\max}| = \cos\frac{\pi}{N}$, из уравнения $\mu^2\omega^2 - 4\omega + 4 = 0$ находим (выбираем меньший корень, попадающий в выделенный прямоугольник)

$$\omega = \frac{2 - \sqrt{4 - 4\mu^2}}{\mu^2} = \frac{2}{1 + \sqrt{1 - \mu^2}} = \frac{2}{1 + \sin\frac{\pi}{N}}.$$

Этому значению ω соответствует значение

$$|\sigma| = \omega - 1 = \frac{1 - \sin\frac{\pi}{N}}{1 + \sin\frac{\pi}{N}} \approx 1 - \frac{2\pi}{N}.$$

Следовательно, количество итераций для метода верхней релаксации оценивается следующим образом:

$$k_{\text{рел}} = \frac{K}{2\pi}N.$$

Для метода Зейделя $\omega = 1$. Поэтому $|\sigma_{\max}| = \mu_{\max}^2 = \cos^2\frac{\pi}{N}$. Это дает

$$k_{\text{З}} = \frac{K}{\pi^2}N^2,$$

т.е. всего в два раза быстрее, чем метод Якоби. Метод верхней релаксации дает ускорение по сравнению с методом Якоби в N раз.

Условие останова. Хотя формулы для подсчета количества итераций во всех методах приведены, проще оценивать близость очередного приближения к точному решению по результатам последних двух итераций.

Поскольку

$$T\boldsymbol{x}^{k+1} - \boldsymbol{x}^{k+1} = T(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k),$$

то

$$\boldsymbol{x}^{k+1} = (T-E)^{-1}T(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k).$$

Следовательно,

$$\|\boldsymbol{x}^{k+1}\| \leqslant \|(E-T)^{-1}\|\,\|T\|\,\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$$

Остается оценить норму $\|(E-T)^{-1}\|\,\|T\|$. Так как

$$\|(E-T)\boldsymbol{x}\| \geqslant \|\boldsymbol{x}\| - \|T\|\,\|\boldsymbol{x}\| = (1-\|T\|)\,\|\boldsymbol{x}\|,$$

то $\|(E-T)^{-1}\| \leqslant \dfrac{1}{1-\|T\|}$. В результате получаем, что

$$\|\boldsymbol{x}^{k+1}\| \leqslant \frac{\|T\|}{1-\|T\|}\,\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$$

Оценивая норму $\|T\|$ спектральным радиусом, для метода верхней релаксации находим

$$\|\boldsymbol{x}^{k+1}\| \leqslant \frac{1-\sin\frac{\pi}{N}}{\sin\frac{\pi}{N}}\,\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|.$$

Учитывая, что $\sin\dfrac{\pi}{N} = \dfrac{2}{\omega} - 1$, можем записать

$$\|\boldsymbol{x}^{k+1}\| \leqslant \frac{2(\omega-1)}{2-\omega}\,\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|.$$

Отсюда заключаем, что неравенство $\|\boldsymbol{x}^{k+1}\| < \varepsilon$ будет выполняться, если

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| < \frac{2-\omega}{2(\omega-1)}\varepsilon.$$

Это неравенство и следует использовать в качестве условия останова. На практике используют упрощенный вариант условия останова

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| < (2-\omega)\varepsilon.$$

chapterno4

4. ИНТЕГРАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ

4.1. Общий подход к интегральным преобразованиям

Задача математической физики как операторное уравнение. Интерпретация решения как обращения линейного оператора. Преобразование линейного оператора при изоморфном отображении пространства. Общий вид интегрального преобразования. Условия применимости интегрального преобразования: существование обратного и его непрерывность.

В самом общем виде задачу математической физики можно представить как решение некоего операторного уравнения $Au = f$, рассматриваемого в некотором векторном пространстве. Мы полагаем, что векторное пространство снабжено нормой и является с этой нормой банаховым.

Эта интерпретация приводит к различным методам решения краевых задач математической физики. Один из таких методов. Связан со следующим. Любое конкретное банахово пространство может быть представлено различными способами. Идентичность векторных пространств устанавливается с помощью изоморфизма. В разных представлениях данного векторного пространства линейный оператор A может выглядеть по nf-разному. Можно попытаться найти такое представление векторного пространства, в котором оператор A имеет простой вид. Тогда и решение операторного уравнения упрощается.

Линейное пространство $L^2[0,\,l]$ со скалярным произведением

$$(\varphi,\,\psi) = \int\limits_0^l \varphi(x)\,\psi(x)\,dx$$

является гильбертовым. В таком линейном пространстве существуют полные ортогональные системы (например тригонометрическая). Пусть $\{e_n\}$ — одна из таких систем (считаем даже ортонормированная). Тогда каждый элемент f представим в виде $f = \sum \alpha_n e_n$, причем единственным образом. Возникает соответствие между элементом f и его коэффициентами Фурье по ортогональной системе $\{e_n\}$. При этом сумме элементов в L^2 соответствует сумма последовательностей коэффициентов Фурье, а при умножении элемента на число последовательность коэффициентов умножается на это число. Последовательность коэффициентов подчиняется неравенству $\sum \alpha_n^2 \leqslant \|f\|^2$. С другой стороны, если $\sum \alpha_n^2 \leqslant C^2$, то ряд $\alpha_n e_n$ в L^2 сходится к некоторому элементу f, причем $\sum \alpha_n^2 = \|f\|^2$.

Рассмотрим множество всех числовых последовательностей $\{\alpha_n\}$, подчиняющихся условию $\sum \alpha_n^2 < +\infty$. Это множество — линейное пространство. Введем в нем скалярное произведение

$$(\{\alpha_n\},\,\{\beta_n\}) = \sum \alpha_n\beta_n.$$

Нетрудно показать, что ряд справа сходится, и что эта формула действительно задает скалярное произведение (т.е. все аксиомы скалярного произведения выполнены). Полученное евклидово пространство оказывается гильбертовым, его обозначают l_2. А мы фактически показали, что $L^2[0,l]$ изоморфно l_2.

Пусть задано операторное уравнение $Au = f$, причем A — самосопряженный оператор, имеющий полную ортонормированную систему собственных векторов e_n, т.е. $Ae_n = \lambda_n e_n$. Каждой последовательности $\boldsymbol{\alpha} = \{\alpha_n\} \in l_2$ соответствует элемент $f \in L^2[0,l]$, представляющий собой сумму $f = \sum \alpha_n e_n$ ряда по ортогональной системе. Это соответствие биективно (в силу полноты $\{e_n\}$ и полноты L^2). Более того, это соответствие есть линейный оператор, причем непрерывный, поскольку в силу равенства Парсеваля $\|f\| = \|\boldsymbol{\alpha}\|$. Таким образом, установлен

изоморфизм между $L^2[0,l]$ и l_2. Правда, этот изоморфизм определяется выбором ортонормированной системы.

Возникает вопрос, как выглядит оператор A в изоморфном пространстве l_2? Подсчитаем коэффициенты β_n разложения вектора Au по системе $\{e_n\}$. Для этого воспользуемся формулами Эйлера — Фурье и учтем самосопряженность A:

$$\beta_n = (Au,\, e_n) = (u,\, Ae_n) = (u,\, \lambda_n e_n) = \lambda_n\,(u,\, e_n) = \lambda_n\alpha_n.$$

Получается так: действие линейного оператора заключается в том, что исходная функция $\{\alpha_n\}$ умножается на фиксированную функцию $\{\lambda_n\}$ (ведь последовательность есть функция натурального аргумента). Ясно, что представление оператора в таком виде позволяет легко найти обратный оператор: чтобы найти исходную функцию $\{\alpha_n\}$, надо результат $\{\beta_n\}$ разделить на функцию $\{\lambda_n\}$. Собственно, это и есть основная идея метода Фурье.

Переход от исходной функции f (оригинала) к ее представлению F в виде коэффициентов Фурье (изображению) позволяет легко обратить линейный оператор A и тем самым решить краевую задачу. Отметим, что переход от f к коэффициентам Фурье осуществляется согласно формулам Эйлера — Фурье:

$$F_n = (f,\, e_n) = \int\limits_0^l f(\xi)\, e_n(\xi)\, d\xi.$$

Это можно записать так:

$$F(n) = \int\limits_0^l f(\xi)\, e(n,\xi)\, d\xi.$$

В силу указанной записи преобразхование и называют интегральным. В общем случае аргумент n в силу разных причин не обязательно является натуральным.

Пусть на множестве $(a,\, b) \times P$, где $P \in \mathbb{C}$ — подмножество комплексной плоскости, определена функция $K(x,p)$. Интеграл

$$U(p) = \int\limits_a^b K(x,p)\, u(p)\, dx$$

определяет линейное преобразование, которое функцию u действительного переменного, определенную на интервале $(a,\, b)$, преобразует в функцию $U(p)$, определенную на множестве P. Это преобразование называют ***интегральным преобразованием***. Оно действует на множестве функций $u(x)$, для которых указанный интеграл определен при всех значениях $p \in P$. Это множество есть линейное пространство, называемое ***пространством оригиналов***. Совокупность образов $U(p)$ также есть линейное пространство, называемое ***пространством изображений***. При этом функцию $K(x,p)$ называют ***ядром интегрального преобразования***.

В рассмотренной выше интерпретации метода Фурье ядро интегрального преобразование определяется последовательностью собственных функций $e_n(x)$ оператора A краевой задачи, которая интерпретируется как функция $e(x,p)$, где $p \in \mathbb{N}$.

Определение интегрального преобразование дано для простейшего, одномерного случая. Однако оно легко обобщается на многомерный случай, достаточно вместо интервала $(a,\, b)$ рассмотреть некоторую область $D \subset \mathbb{R}^n$, а вместо определенного интеграла — n-мерный интеграл по области D.

Суть метода интегральных преобразований заключается в следующем. От уравнения $Au = f$ в пространстве оригиналов мы переходим к уравнению $\hat{A}U = F$ в пространстве изображений, в котором оператор $\hat{A}$ представляет собой просто умножение функции $U(p)$ на некоторую

функцию $\hat{A}(p)$. В пространстве изображений находим изображение $U(p)$ неизвестной функции, попросту деля правую часть на функцию $\hat{A}(p)$. Наконец, по изображению $U(p)$ восстанавливаем искомую функцию $u(x)$.

Для реализации описанного метода необходимо, чтобы выполнялись следующие требования: а) интегральное преобразование должно быть непрерывным, т.е. малые отклонения в функции u должны давать малые отклонения в изображении U (разумеется, для оценки этих отклонений в пространствах оригиналов и изображений должны быть заданы нормы); б) интегральное преобразование должно быть обратимо, т.е. разным функциям-оригиналам должны соответствовать разные функции-изображения; в) обратное преобразование должно быть непрерывным.

Описанные требования — это необходимые условия теоретической возможности применения интегрального преобразования. На практике все не так просто. Во-первых, интегральное преобразование должно упрощать конкретный линейный оператор, входящий в краевую задачу, т.е. в принципе под каждую краевую задачу нужно подбирать свое интегральное преобразование. Во-вторых, даже если для данной краевой задачи интегральное преобразование найдено, для решения конкретной задачи $Au = f$ требуется найти изображение правой части f, вычислив соответствующий интеграл, а затем по изображению неизвестной функции U найти оригинал, вычислив еще один интеграл.

Теоретически для каждой краевой задачи существует упрощающее интегральное преобразование. Однако условия упрощения приводят к некоторой краевой задачи для ядра интегрального преобразования. Не факт, что эта задача проще исходной. На практике бывает и наоборот, новая задача оказывается сложнее исходной. Это обстоятельство указывает на то, что от метода интегральных преобразований не следует ждать многого. Достаточно и того, что он помогает решать определенный класс задач.

Несколько упрощает проблему следующее соображение: если данное интегральное преобразование упрощает оператор A, то оно упрощает и любой оператор вида $P(A)$, где P — многочлен. В результате с помощью одного интегрального преобразования можно решить целый ряд задач. Кроме того, как и в случае метода Фурье можно рассчитывать, что применение интегральных преобразований по части переменных позволит попросту понизить размерность задачи. Именно так работает метод Фурье, который можно интерпретировать как применение интегрального преобразования, упрощающего дифференциальный оператор $\frac{d^2}{dx^2}$ на отрезке в классе функций, удовлетворяющих однородным граничным условиям.

4.2. Интегральное преобразование Фурье

Интеграл Фурье и его связь с рядами Фурье. Определение. Обратное преобразование. Свойства: 1) линейность; 2) убывание на бесконечности $\lim\limits_{p\to\infty} F(p) = 0$; 3) дифференцирование оригинала $F[f'] = i\omega F[f]$; 4) дифференцирование изображения $F[f]' = -iF[xf]$; 5) преобразование свертки функций. Применение преобразования Фурье к решению задач математической физики.

Любая функция $f(x)$, определенная на отрезке $[-l,\, l]$ и удовлетворяющая ***условиям Дирихле*** (т.е. функция кусочно непрерывна и кусочно монотонна на отрезке $[-l,\, l]$), может быть представлена рядом по тригонометрической системе

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty}\Big(A_n\cos\frac{n\pi x}{l} + B_n\sin\frac{n\pi x}{l}\Big).$$

При этом ряд сходится в каждой точке x отрезка к значению $\frac{f(x-0)+f(x+0)}{2}$ (к значению $\frac{f(l-0)+f(-l+0)}{2}$ на концах отрезка).

В этом известном результате из теории рядов Фурье (теорема Дирихле) речь фактически идет о функции с периодом $2l$. Подставляя вместо коэффициентов A_n и B_n их выражения с помощью формул Эйлера — Фурье и неограниченно увеличивая период l, приходим к формуле

$$f(x) \sim \frac{1}{\pi} \int_0^{\infty} d\omega \int_{-\infty}^{+\infty} f(x) \cos \omega(x-\xi)\, d\xi.$$

Интеграл справа называют интегралом Фурье. Если функция $f(x)$ абсолютно интегрируема на числовой оси и удовлетворяет условиям Дирихле, то интеграл Фурье в каждой точке $x \in \mathbb{R}$ сходится к значению $\frac{f(x-0)+f(x+0)}{2}$.

Уже в интеграле Фурье можно усмотреть пару интегральных преобразований: прямое (внутренний интеграл) и обратное (внешний интеграл). Однако традиционно тригонометрическое ядро прямого преобразования переписывают с помощью экспоенты мнимого аргумента. Это приводит к формуле

$$\frac{f(x-0)+f(x+0)}{2} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \int_{-\infty}^{+\infty} f(\xi) e^{i\omega(x-\xi)} d\xi,$$

где, правда, внешний интеграл понимается только в смысле главного значения, т.е. как предел $\int_{-T}^{T}$ при $T \to \infty$. Ядро внутреннего интеграла зависит от трех (а не двух) переменных. Однако оно разделяется в произведение, так что можно записать:

$$\frac{f(x-0)+f(x+0)}{2} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{i\omega x} d\omega \int_{-\infty}^{+\infty} f(\xi) e^{-i\omega\xi} d\xi.$$

В результате мы приходим к двум преобразованиям:
(прямому) ***преобразованию Фурье***

$$F(\omega) = \int_{-\infty}^{+\infty} f(x) e^{-i\omega x} dx$$

и ***обратному преобразованию Фурье***

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega x} d\omega.$$

Обратное преобразование Фурье, в котором интеграл понимается в смысле главного значения, позволяет восстанавливать оригинал по изображению, во всяком случае, для абсолютно интегрируемых функций, удовлетворяющих условиям Дирихле. В действительности сфера действия преобразований шире, но для этого требуется более общее понятие интеграла, чем несобственный интеграл Римана.

Хотя преобразование Фурье появилось как „разложение функций бесконечного периода“ по тригонометрической системе, оно оказалось тесно связанным с задачами математической физики, поскольку его истоки (тригонометрическая система) важнейшая составляющая математической физики. Оно является упрощающим для оператора дифференцирования, однако требует рассмотрения функции не на отдельном интервале, а на всей числовой оси.

В литературе распространено правило обозначать оригиналы строчными буквами, а изображения — соответствующими прописными. Однако в определенных ситуациях это не совсем

удобно и имеет смысл обозначать преобразование именно как линейный оператор. В связи с этим в записи $F[u](w)$ $F[u]$ обозначает (прямое) преобразование Фурье функции u, а приписывание еще одного аргумента в круглых скобках означает значение функции $F[u]$ в точке w.

Отметим основные свойства преобразования Фурье:

1. Линейность: $F[\alpha f + \beta g] = \alpha F[f] + \beta F[g]$, α, $\beta \in \mathbb{R}$.
2. Если $f \in L^1(\mathbb{R})$, то $F[f]$ непрерывна и ограничена на $\mathbb{R}$.
3. Теорема смещения: $F[e^{i\lambda x} f(x)](\omega) = F[f](\omega - \lambda)$.
4. Теорема запаздывания: $F[f(x-\lambda)](\omega) = F[f](\omega) e^{-i\omega\lambda}$.
5. Дифференцирование оригинала: $F[f'] = i\omega F[f]$.
6. Дифференцирование изображения: $F[f]' = -iF[xf(x)]$.
7. Теорема о свертке: $F[f * g] = F[f] \cdot F[g]$.

Замечание. Свертка двух функций, определенных и абсолютно интегрируемых на числовой оси определяется соотношением

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(\xi) g(x-\xi)\, d\xi.$$

Все свойства доказываются с помощью правил интегрирования. Рассмотрим, например, правило дифференцирования оригинала. Для начала отметим, что если функция f' абсолютно интегрируема, то существует предел $\lim\limits_{x\to\infty} f(x)$, равный нулю в случае, когда и f абсолютно интегрируема. С учетом этого

$$F[f'](\omega) = \int_{-\infty}^{+\infty} f'(x) e^{-i\omega x} dx = f(x) e^{-i\omega x}\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} f(x) de^{-i\omega x} = \int_{-\infty}^{+\infty} f(x) \cdot i\omega e^{-i\omega x} dx = i\omega F[f].$$

Пример 4.1. Рассмотрим задачу Коши для уравнения теплопроводности:

$$\begin{cases} u_t = a^2 u_{xx}, \quad t > 0, \quad x \in \mathbb{R}, \\ u\big|_{t=0} = \varphi(x). \end{cases}$$

Считая, что для каждого t функция $u(x,t)$ является оригиналом преобразования Фурье, можем перейти в задаче к изображениям Фурье:

$$\begin{cases} U_t = -a^2\omega^2 U, \\ U\big|_{t=0} = \Phi. \end{cases}$$

Получена задача Коши для обыкновенного дифференциального уравнения. Ее решение

$$U(\omega, t) = \Phi(\omega) e^{-a^2\omega^2 t}.$$

Переход к оригиналам дает представление решения исходной задачи в виде

$$u(x,t) = \varphi(x) * K(x,t) = \int_{-\infty}^{+\infty} G(x-\xi, t)\varphi(\xi)\, d\xi,$$

где $G(x,t)$ — оригинал изображения $e^{-a^2\omega^2 t}$ (ω — переменная, t — параметр).

Функцию $G(x,t)$ находим с помощью обратного преобразования Фурье:

$$G(x,t) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-\omega^2 a^2 t} e^{i\omega x} d\omega = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} \exp\Big[-a^2 t\Big(\omega - \frac{ix}{2a^2 t}\Big)^2 - \frac{x^2}{4a^2 t}\Big] d\omega =$$

$$= e^{-\frac{x^2}{4a^2 t}} \cdot \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-a^2 t\omega^2} d\omega = e^{-\frac{x^2}{4a^2 t}} \cdot \frac{1}{2\pi} \cdot \frac{1}{a\sqrt{t}}\int\limits_{-\infty}^{+\infty} e^{-\zeta^2} d\zeta = \frac{1}{2a\sqrt{\pi t}} e^{-\frac{x^2}{4a^2 t}}.$$

В вычислении использована формула для интеграла Пуассона:

$$\int\limits_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}.$$

4.3. Преобразование Лапласа

Операционное исчисление. Оригиналы и изображения. Преобразование Лапласа. Основные свойства. Теорема обращения. Применение операционного исчисления для решения задач математической физики.

Преобразование Лапласа имеет вид

$$L[f](p) = \int\limits_0^{\infty} f(t) e^{-pt} dt.$$

Оригиналами этого преобразования являются функции, определенные на числовой оси и удовлетворяющие условиям

1) $f(t) = 0$ при $t < 0$;
2) $f(t)$ кусочно непрерывна;
3) $|f(t)| \leqslant Me^{\sigma} t$.

Изображения преобразования Лапласа — аналитические функции, определенные в некоторой полуплоскости $\operatorname{Re} p > \sigma_0$.

6. ПРИБЛИЖЕННЫЕ АНАЛИТИЧЕСКИЕ МЕТОДЫ

6.1. Общий подход

Задача математической физики как операторное уравнение $Au = f$. Непрерывные и ограниченные операторы, норма оператора. Самосопряженные операторы. Положительно определенные операторы. Неограниченные операторы. Примеры. Область определения неограниченного оператора. Линейные подпространства и многообразия. Плотно определенные операторы. Симметричные и самосопряженные операторы. Расширение оператора. Классическое и обобщенное решение. Теорема: уравнение $Au = f$ для плотно определенного положительного оператора единственно. Теорема: если A — плотно определенный положительный, то u^* — решение $Au = f \iff u^*$ — тчк строгого глобального min функционала $(Au, u) - 2(f, u)$. Итерационные последовательности.

Многие задачи математической физики можно интепретировать как уравнение вида $Au = f$, где $A\colon U \to V$ — некоторый оператор, действующий из линейного пространства U в линейное пространство V, $u \in U$, $f \in V$. Сюда относятся краевые задачи, задачи на интегральные уравнения и т.п.

Остановимся на некоторых особенностях такого подхода. Для УМФ характерным является случай, когда U и V бесконечномерны. Мы полагаем, что в линейных пространствах U и V задана норма, а сами линейные пространства являются банаховыми. Норма позволяет ввести понятие непрерывности (линейного оператора). Отметим, что в конечномерном случае любой линейный оператор автоматически непрерывный, в то время как в бесконечномерном это не так. Хотя верно следующее: если линейный оператор непрерывен хотя бы в одной точке, то он неперывен и в любой другой, в частности в нуле, т.е. для любого $\varepsilon > 0$ существует такое $\delta > 0$, что $\|Au\| < \varepsilon$ при $\|u\| < \delta$.

Для нормированных линейных пространств можно также ввести другое понятие. Линейный оператор $A\colon U \to V$ называется ограниченным, если существует такое число $C > 0$, что

$$\|Au\| \leqslant C \|u\|.$$

При этом наименьшее такое C (а точнее, точная нижняя грань множества таких C) называется ***нормой линейного оператора***. Норму линейного оператора можно вычислить по формуле

$$\|A\| = \sup_{u \neq 0} \frac{\|Au\|}{\|u\|} = \sup_{\|u\|=1} \|Au\|.$$

Нетрудно убедиться в том, что ограниченный линейный оператор является непрерывным. Однако оказывается, что верно и обратное, т.е. любой линейный оператор ограничен.

Пример 6.1. Рассмотрим множество функций, дифференцируемых на отрезке $[a, b]$ с нормой

$$\|f\|_1 = \int\limits_a^b |f(x)|\, dx$$

В этом линейном пространстве определен линейный оператор $Au = u'$ (дифференцирования). Оказывается, что этот линейный оператор не является непрерывным.

Рассмотренное линейное пространство не является полным. В банаховом пространстве отсутствие непрерывности линейного оператора сопровождается и другой особенностью: такой

оператор определен не на всем линейном пространстве Так, оператор дифференцирования в $L^2[a,b]$ определен только на гладких функциях.

Область определения неограниченного линейного оператора. Линейные многообразия и подпространства. Плотно определенные операторы. Самосопряженные ограниченные операторы. Симметричные и самосопряженные неограниченные операторы.

Классическое и обобщенное решения. Симметрический положительно определенный оператор A в гильбертовом пространстве H. Понятие о квадратичном функционале.

Теорема 6.1. Пусть A — плотно определенный положительный оператор в H. Тогда уравнение $Au = f$ имеет не более одного решения.

◀ Если $Au_1 = f$ и $Au_2 = f$, то для $u = u_1 - u_2$ имеем $Au = 0$. Следовательно, $(Au,\, u) = 0$, откуда в силу положительной определенности получаем $u = 0$ и $u_1 = u_2$. ▶

Теорема 6.2. Пусть A — плотно определенный положительный оператор в H. Элемент $u^* \in H$ является решением уравнения $Au = f \iff u^*$ является точкой строгого глобального минимума квадратичного функционала $(Au,\, u) - 2\,(f,\, u)$.

◀ Пусть $u = u^* + v$. Тогда

$$(A(u^*+v),\, u^*+v) - 2\,(f,\, u^*+v) = (Au^*,\, u^*) + 2\,(Au^*,\, v) + (Av,\, v) - 2\,(f,\, u^*) - 2\,(f,\, v) = \\ = \big((Au^*,\, u^*) - 2\,(f,\, u^*)\big) + 2\,(Au^* - f,\, v) + (Av,\, v)\,.$$

Если $Au^* = f$, то $(Au^* - f,\, v) = 0$ и

$$(A(u^*+v),\, u^*+v) - 2\,(f,\, u^*+v) = \big((Au^*,\, u^*) - 2\,(f,\, u^*)\big) + (Av,\, v)\,.$$

Видно, что приращение функционала равно $(Av,\, v)$ и в силу положительной определенности этого функционала приращение при $v \neq 0$ положительно. Это означает, что u^* является точкой минимума функционала.

Пусть u^* — точка минимума функционала. Выберем произвольно $v_0 \neq 0$ и положим $v = tv_0$. Тогда приращение функционала будет равно

$$2t\,(Au^* - f,\, v_0) + t^2\,(Av_0,\, v_0)$$

и как функция переменного t будет достигать минимума при $t = 0$. Это возможно только при $(Au^* - f,\, v_0) = 0$. Так как элемент $Au^* - f$ ортогонален любому ненулевому вектору, то он сам равен нулю: $Au^* - f = 0$. ▶

Пример. Рассмотрим гильбертово пространство $L^2[a,b]$ и в нем линейный оператор $Au = -\dfrac{d^2u}{dx^2}$, определенный на многообразии всех дважды непрерывно дифференцируемых функций, удовлетворяющих однородным граничным условиям, например, I рода. Можно показать, что это положительный симметричный оператор. Доказанная теорема означает, что задача решения уравнения $-\dfrac{d^2}{dx^2}u = f(x)$ эквивалентно задаче поиска минимума функционала

$$\int_a^b \big((u'(x))^2 - 2f(x)u(x)\big)\,dx.$$

при (разумеется) дополнительных ограничениях $u(a) = u(b) = 0$.

Один из подходов к решению операторного уравнения $Au = f$ состоит в построении в нормированном пространстве последовательности $\{u_n\}$, сходящейся по норме к решению u^* операторного уравнения $Au = f$. Такую последовательность строят по шагам, называемым итерациями, причем на каждом шаге исходя из одного или нескольких построенных членов последовательности строят очередной член последовательности. В таком контексте последовательность $\{u_n\}$ называют итерационной. Примером здесь могут служить итерационные методы решения систем линейных уравнений. Ключевыми понятиями итерационных методов являются сходимость и скорость сходимости.

6.2. Общая схема приближенных методов

Общая схема приближенного метода: приближение $\boldsymbol{x}_n$ как решение „приближенного“ уравнения $A_n(\boldsymbol{x}) = \boldsymbol{y}_n$. Проекционные операторы P_n: $D(A) \to D(A_n)$ и Q_n: $R(A) \to R(A_n)$ ($D(A)$ и $R(A)$ — обл.опр. и обл.знач. оператора A). Условие $\boldsymbol{y}_n = Q_n(f)$. Оператор восстановления U_n: $D(A_n) \to D(A)$. Сходимость: $\|\boldsymbol{x}_n - P_n(u^*)\| \to 0$ при $n \to \infty$. Порядок аппроксимации: $\gamma_n(u) = \|A_n(\boldsymbol{x}_n) - A_nP_n(u^*)\| = \|Q_nA(u^*) - A_nP_n(u^*)\|$. Оценка $\|\boldsymbol{x}_n - P_n(u^*)\| \leqslant \|A_n^{-1}\|\gamma_n(u^*)$ в линейном случае. **Теорема.** Если $v_n = \|A_n^{-1}\|\gamma_n(u^*) \to 0$ и $U_nP_n(u^*) \to u^*$ при $n \to \infty$, посл-ть $\{\|U_n\|\}$ ограничена, то посл-ть $u_n = U_n(\boldsymbol{x}_n)$ сходится к u^*, причем

$$\|u_n - u^*\| \leqslant \|U_n\|v_n + \|U_nP_n(u^*) - u^*\|.$$

Построение итерационных последовательностей напрямую не всегда возможно. Можно предложить иной подход, в котором соотношения между элементами бесконечномерного пространства моделируются в конечномерном пространстве (именно такова ситуация в конечно-разностных методах, когда мы обычную функцию заменяем сеточной функцией, т.е. конечномерным вектором).

Описанный подход ведет к целой группе приближенных методов. Общая схема таких методов такова. Имеется оператор A с областью определения $D(A)$, и областью значений $R(A)$. Приближение $\boldsymbol{x}_n$ как решение „приближенного“ уравнения $A_n(\boldsymbol{x}) = \boldsymbol{y}_n$. Проекционные операторы P_n: $D(A) \to D(A_n)$ и Q_n: $R(A) \to R(A_n)$ ($D(A)$ и $R(A)$ — обл.опр. и обл.знач. оператора A). Условие $\boldsymbol{y}_n = Q_n(f)$. Оператор восстановления U_n: $D(A_n) \to D(A)$.

Сходимость: $\|\boldsymbol{x}_n - P_n(u^\circ)\| \to 0$ при $n \to \infty$. Порядок аппроксимации: $\gamma_n(u) = \|A_n(\boldsymbol{x}_n) - A_nP_n(u^\circ)\| = \|Q_nA(u^\circ) - A_nP_n(u^\circ)\|$. Оценка $\|\boldsymbol{x}_n - P_n(u^\circ)\| \leqslant \|A_n^{-1}\|\gamma_n(u^\circ)$ в линейном случае.

Теорема 6.3. Если $v_n = \|A_n^{-1}\|\gamma_n(u^\circ) \to 0$ и $U_nP_n(u^\circ) \to u^\circ$ при $n \to \infty$, посл-ть $\{\|U_n\|\}$ ограничена, то посл-ть $u_n = U_n(\boldsymbol{x}_n)$ сходится к u°, причем

$$\|u_n - u^\circ\| \leqslant \|U_n\|v_n + \|U_nP_n(u^\circ) - u^\circ\|.$$

◀ Так как $u_n = U_n(x_n)$, то

$$\|u_n - u^\circ\| = \|U_nx_n - u^\circ\| \leqslant \|U_n(x_n) - U_nP_nu^\circ\| + \|U_nP_nu^\circ - u^\circ\| \leqslant \|U_n\|\|x_n - P_nu^\circ\| + \|U_nP_nu^\circ - u^\circ\|.$$

Разность $x_n - P_nu^\circ$ оцениваем через невязку:

$$\|x_n - P_nu^\circ\| \leqslant \|A_n^{-1}\|\,\|A_nx_n - A_nP_nu^\circ\| = \|A_n^{-1}\|\,\|A_nP_nu^\circ - y_n\| = \|A_n^{-1}\|\gamma_n.$$

В результате:

$$\|u_n - u^\circ\| \leqslant \|U_n\|\|A_n^{-1}\|\gamma_n + \|U_nP_nu^\circ - u^\circ\| \qquad ▶$$

6.3. Метод малого параметра

Операторное уравнение $(B_0 + \lambda A)u = f$ при известном обратном операторе B_0^{-1}. Теорема: если B_0 обратим, то $\exists\delta > 0\ \forall\lambda,\ |\lambda < \delta$, оператор $B_) + \lambda A$ обратим. Степенной „операторный“ ряд и его сходимость. Пример: уравнение

$$u(t) - \lambda\int_a^b K(t,\tau)\,u(\tau)\,d\tau = f(t).$$

На практике возникают ситуации, когда требуется решить операторное уравнение $Au = f$, а известно решение уравнения $A_0u = f$ с некоторым близким оператором A_0. Предполагаем, что $A = A_0 + \lambda B$ при некотором значении λ. Тогда можно поступить так же, как и при вычислении функции по ряду Тейлора:

$$f(x + \lambda h) = \sum_{k=0}^{\infty}\frac{f^{(k)}(x)}{k!}\lambda^k.$$

Слагаемые ряда — поправки разной степени к базовому значению $f(x)$.

В данном случае решение уравнения $Au = f$ можно представить в виде:

$$u = A^{-1}f = (A_0 + \lambda B)^{-1}f = [A_0(E + \lambda A_0^{-1}B)]^{-1}f = (E + \lambda A_0^{-1}B)^{-1}A_0^{-1}f = (E + \lambda C)^{-1}\widetilde{f},$$

где $C = A_0^{-1}B$, $\widetilde{f} = A_0^{-1}f$. Запишем формальный ряд

$$(E + \lambda C)^{-1} = \sum_{k=0}^{\infty}(-1)^k\lambda^k C^k.$$

Теорема 6.4. Если ряд

$$D = \sum_{k=0}^{\infty}(-1)^k\lambda^k C^k \tag{6.1}$$

сходится, то $D = (E + \lambda C)^{-1}$.

◀ Обозначим через D_N частичную сумму ряда. Тогда

$$\begin{aligned}\|D(E + \lambda C) - E\| &= \|D(E + \lambda C) - D_N(E + \lambda C) + D_N(E + \lambda C) - E\| \leqslant \\ &\leqslant \|D(E + \lambda C) - D_N(E + \lambda C)\| + \|D_N(E + \lambda C) - E\| \leqslant \\ &\leqslant \|D - D_N\|\,\|E + \lambda C\| + \|D_N(E + \lambda C) - E\|.\end{aligned}$$

Первое слагаемое сходится к нулю, поскольку $\|D - D_N\| \to 0$ при $N \to \infty$. Во втором слагаемом имеем

$$\begin{aligned}D_N(E + \lambda C) = D_N + \lambda D_N C &= \sum_{k=0}^{N}(-1)^k\lambda^k C^k + \sum_{k=0}^{N}(-1)^k\lambda^{k+1}C^{k+1} = \\ &= \sum_{k=0}^{N}(-1)^k\lambda^k C^k - \sum_{k=1}^{N+1}(-1)^k\lambda^k C^k = E + (-1)^{N+1}\lambda^{N+1}C^{N+1}.\end{aligned}$$

Следовательно, $\|D_N(E + \lambda C) - E\| = \|\lambda^{N+1}C^{N+1}\| \to 0$ при $N \to \infty$. ▶

Теорема 6.5. Если нормированное пространство полное и $|\lambda| < \|C\|^{-1}$, то ряд (6.1) сходится.

В соответствии с теоремой решение операторного уравнения можно представить как сумму ряда. В качестве приближенного решения можно взять частичную сумму ряда.

Пример 6.2. Рассмотрим интегральное уравнение

$$u(t) - \lambda\int_a^b K(t,\tau)\,u(\tau)\,d\tau = f(t).$$

При $\lambda = 0$ это уравнение тривиально и имеет решение $u(t) = f(t)$. В данном случае $A_0 = I$ (единичный оператор), B представлен интегралом. Значит, $C = A_0^{-1}B$ и

$$C[u](t) = \int_a^b K(t,\tau)\,u(\tau)\,d\tau.$$

Первое приближение содержит два слагаемых ряда. Получаем:

$$u = A_0(I - \lambda C)f, \quad u(t) = f(t) - \lambda\int_a^b K(t,\tau)\,f(\tau)\,d\tau.$$

Второе приближение получаем добавлением еще одного слагаемого: $u = A_0(I - \lambda C + \lambda^2 C^2)f$. При этом

$$C^2[u](t) = \int_a^b K(t,\tau)d\tau\int_a^b K(\tau,\xi)u(\xi)\,d\xi = \int_a^b u(\xi)\,d\xi\int_a^b K(t,\tau)\,K(\tau,\xi)\,d\tau = \int_a^b K_2(t,\xi)\,u(\xi)\,d\xi,$$

где

$$K_2(t,\xi) = \int_a^b K(t,\tau)\,K(\tau,\xi)\,d\tau.$$

В результате

$$u(t) = f(t) - \lambda\int_a^b K(t,\tau)\,f(\tau)\,d\tau + \lambda^2\int_a^b K_2(t,\xi)\,f(\xi)\,d\xi.$$

6.4. Метод ортогональных проекций

Рассмотрим операторное уравнение $Au = f$, действующее из банахова пространства U в гильбертово пространство V. Выберем в V некоторую систему $\{v_k\}$, линейная оболочка которой всюду плотна в V (например, полная ортогональная система). Тогда уравнение $Au = f$ равносильно системе уравнений $(Au - f,\, v_k) = 0$, $k \in \mathbb{N}$.

Действительно, пусть $(Au - f,\, v_k) = 0$, $k \in \mathbb{N}$. Для любого числа $\varepsilon > 0$ можно найти такую линейную комбинацию $w = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_k v_k$, что $\|(Au - f) - w\| < \varepsilon$. Тогда, с одной стороны, $(Au - f,\, w) = 0$, а с другой

$$\begin{aligned}\|Au - f\|^2 = (Au - f,\, Au - f) &= (Au - f,\, (Au - f) - w) + (Au - f,\, w) = \\ &= (Au - f,\, Au - f - w) \leqslant \|Au - f\|\,\|(Au - f) - w\| \leqslant \|Au - f\|\varepsilon.\end{aligned}$$

Следовательно, $\|Au - f\| \leqslant \varepsilon$. Так как ε выбрано произвольно, заключаем, что $\|Au - f\| = 0$ и $Au = f$.

Выберем в банаховом пространстве U какой-либо базис, т.е. такую систему $\{e_n\}$, что любой элемент $u \in U$ представляется в виде ряда $u = \sum\limits_{n=1}^{\infty}\alpha_n e_n$, и притом единственным способом. В качестве приближения к решению u_* уравнения $Au = f$ выберем вектор u_N, удовлетворяющий условию $(Au - f,\, v_k) = 0$, $k = \overline{1,\, N}$, или $(Au,\, v_k) = (f,\, v_k)$, $k = \overline{1,\, N}$. Этот выбор неоднозначен. Дополнительно потребуем, чтобы u_N представлялся линейной комбинацией векторов e_i, $i = \overline{1,\, N}$. Записав $u_N = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_N e_N$, получим:

$$(Au_N,\, v_k) = \sum_{i=1}^{N}\alpha_i\,(Ae_i,\, v_k)\,.$$

В результате приходим к системе уравнений относительно коэффициентов α_i:

$$\sum_{i=1}^{N}(Ae_i,\, v_k)\,\alpha_i = (f,\, v_k)\,, \quad k = \overline{1,\, N}.$$

Матрица коэффициентов $(Ae_i,\, v_k)$, i, $k = \overline{1,\, N}$, квадратная.

Метод ортогональных проекций вписывается в общую схему приближенных методов. При этом оператор P_N в данном случае элементу u ставит в соответствие частичную сумму его

разложения в ряд, U_N — единичный оператор, поскольку $D(A_N)$ — подпространство $D(A)$, Q_N элементу f ставит в соответствие кортеж скалярных произведений $(f,\, v_n)$.

В этой связи возникают вопросы: а) о невырожденности матрицы A; б) о величине $\|A_N^{-1}\|$; в) о невязке метода. Если невязка метода стремится к нулю, а матрица A_N невырождена и $\|A_N^{-1}\| \leqslant C$, то метод ортогональных проекций дает последовательность, сходящуюся к решению задачи.

Пример 6.3. Рассмотрим линейное пространство l_2 последовательностей действительных чисел, сходящихся в среднем квадратичном. Это гильбертово пространство. Рассмотрим линейный оператор A: $l_2 \to l_2$, который элемент $x = \{x_1,\, x_2,\, \ldots,\, x_n,\, \ldots\}$ переводит в элемент $Ax = \{0,\, x_1,\, x_2,\, \ldots,\, x_n,\, \ldots\}$ (это оператор сдвига). Очевидно, что если $\|x\| = 1$, то и $\|Ax\| = 1$. Поэтому линейный оператор A ограничен и имеет норму, равную 1.

Выберем стандартный базис в качестве систем $\{e_n\}$ и $\{v_k\}$. Тогда $(Ae_i,\, v_k) = Spe_{i+1}v_k = 1$ при $i = k-1$, а иначе 0. В результате

$$((Ae_i,\, v_k)) = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \end{pmatrix}$$

Мы видим, что при любом N оператор A_N, полученный по методу ортогональных проекций, является вырожденным. Пример показывает, что выбор систем $\{e_i\}$ и $\{v_k\}$ важен для корректной работы метода ортогональных проекций.

Пример 6.4. Рассмотрим частный случай метода ортогональных проекций, когда последовательности $\{u_k\}$ и $\{v_n\}$ совпадают — так называемый ***метод Галеркина — Бубнова***. Применим его к задаче на уравнение теплопроводности

$$u_t = a^2 u_{xx} + f,$$
$$u\big|_{t=0} = u\big|_{x=0} = u\big|_{x=l} = 0.$$

В качестве функций выберем

$$v_{kn}(t,x) = te^{-kt}\sin\frac{n\pi x}{l}, \quad k = 0,\, 1,\, 2,\, \ldots, \quad n = 1,\, 2,\, \ldots$$

Полагаем $Au = u_t - a^2 u_{xx}$. Тогда

$$Av_{kn}(t,x) = (1-\gamma_{kn}t)e^{-kt}\sin\frac{n\pi x}{l}, \quad \gamma_{kn} = k - \frac{n^2\pi^2 a^2}{l^2}.$$

и

$$(Av_{kn},\, v_{pq}) = \int\limits_0^\infty t(1-\gamma_{kn}t)e^{-(k+p)t}dt \int\limits_0^l \sin\frac{n\pi x}{l}\,\sin\frac{q\pi x}{l}\,dx.$$

Так как

$$\int\limits_0^\infty t(1-\gamma_{kn}t)e^{-(k+p)t}dt = \frac{1}{k+p}\int\limits_0^l (1-2\gamma_{kn}t)e^{-(k+p)t}dt = \frac{1}{(k+p)^2} + \frac{2\gamma_{kn}}{(k+p)^3},$$

то

$$(Av_{kn},\, v_{pq}) = \begin{cases} 0, & n \neq q; \\ \dfrac{l}{2}\Big(\dfrac{1}{(k+p)^2} + \dfrac{2\gamma_{kn}}{(k+p)^3}\Big), & n = q. \end{cases}$$

7. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

Метод конечных элементов играет важную роль в решении многих технических задач. Его идея, как и идея метода конечных разностей, достаточно проста. Непрерывная задача, в частности краевая задача математической физики, сводится к дискретной задаче. При этом область изменения переменных заменяется совокупностью элементарных областей, а уравнения краевой задачи — соотношениями между параметрами, связанными с элементарными областями. эти параметры есть значения некоторой функции, приближающей искомое решение, в некоторых точках элементарной области. Приближающая функция строится по этим узловым значениям с помощью различных методов интерполяции и аппроксимации.

В некотором роде метод конечных элементов можно рассматривать как развитие метода коллокации, в котором приближение строится в виде кусочно постоянной функции. Метод конечных элементов можно также ассоциировать с методом конечных разностей, но в методе конечных разностей аппроксимация идет с помощью сеточных функций, определенных в конечном числе точек, а в методе конечных элементов приближением является функция с той же областью определения, что и исходная.

Поскольку приближающая функция строится интерполяционными методами, она имеет невысокую степень гладкости. Разумеется, аппроксимирующую функцию можно строить с любой степенью точности и гладкости, но за достижение таких показателей приходится расплачиваться сложностью процесса интерполяции, большим количеством промежуточных параметров и возрастанием объема вычислений, что делает такой подход невыгодным с вычислительной точки зрения. В основе лежит сплайн-интерполяция.

В теории метода конечных элементов элементарные области с фиксированными точками, в которых рассматриваются значения функции, называются ***конечными элементами***, фиксированные точки — ***узлами конечного элемента***, а значения функции в них — ***узловыми параметрами***. Совокупность всех конечных элементов, заменяющую область определения краевой задачи, называют ***сеткой конечных элементов***.

7.1. Об интегральной формулировке задачи

Поскольку аппроксимирующая функция имеет невысокую степень гладкости, исходная формулировка краевой задачи, включающая частные производные плохо подходит для построения аппроксимирующих соотношений. Необходимо изменить задачу так, чтобы в ней вместо операции дифференцирования использовалось интегрирование. В основе такого преобразования, как правило лежит возможность различного представления физических и механических законов. Например, в статике вместо баланса различных сил, который приводит к системе дифференциальных уравнений, описывающих процесс, можно использовать принцип минимума потенциальной энергии, что приводит к совершенного другого вида математической задаче — вариационной. В такого рода задаче требуется найти функцию в некотором заданном классе функций, на которой достигает минимума определенный функционал (в приведенном примере это потенциальная энергия системы).

С общей точки зрения (точки зрения функционального анализа) эту ситуацию можно описать так. Изначальную, дифференциальную формулировку задачи можно сформулировать как решение операторного уравнения $Au = f$, рассматриваемого в некотором банаховом (а как правило, гильбертовом) пространстве H. При этом линейный оператор A связан с дифференцированием (в частности, это может быть линейный дифференциальный оператор) и определен не на всем линейном пространстве H, а на некотором его подпространстве $D(A)$, всюду плотном

в H. Областью значений этого оператора является некоторое линейное пространство. Но чтобы обеспечить необходимые свойства, желательно, чтобы пространство также было банаховым или содержалось в таком пространстве. В общем случае предполагается, что область значений A — подпространство $R(A)$, всюду плотное в банаховом (или гильбертовом) пространстве Y.

Решением операторного уравнения при указанных предположениях является элемент $u_* \in D(A)$, для которого выполняется равенство $Au_* = f$. В ряде случаев можно построить такой линейный оператор $\widetilde{A}$, что его область определения включает область определения A, т.е. $D(\widetilde{A}) \supset D(A)$, а на множестве $D(A)$ операторы совпадают, т.е. $Au = \widetilde{A}U$, $u \in D(A)$. Дело в том, что область определения линейного оператора может быть ограничена „объективными“ факторами типа наличия особых точек, а может быть, и „субъективными“, в частности связанными со способом его описания. Простейший пример. Дифференциальное уравнение $\dot{x} = ax$ с заданным начальным условием $x(0) = x_0$ можно проинтегрировать и прийти к уравнению

$$x(t) = x_0 + a\int\limits_0^t x(t)\,dt.$$

Но если первое уравнение имеет смысл только для дифференцируемых функций, то второе — для любых интегрируемых. Если дифференцируемая функция удовлетворяет интегральному уравнению, то она удовлетворяет и дифференциальному. В данном случае линейный оператор A — это оператор $Au = \dot{u} - au$, а оператор $\widetilde{A}$ определяется равенством $\widetilde{A}u = u - \int\limits_0^t u(t)\,dt$.

На самом деле, если оператор A, связанный с дифференцированием, имеет расширение $\widetilde{A}$ на более широкий класс функций (элементов гильбертова пространства), то самой операции дифференцирования A можно придать расширительное толкование, рассматривая $\widetilde{A}$ тоже в качестве дифференцирования — так называемого обобщенного дифференцирования. В этом контексте решения уравнения $\widetilde{A}u = f$, не попадающие в $D(A)$, т.е. не являющиеся решениями уравнения $Au = f$, называют ***обобщенными решениями***, в то время как решения уравнения $Au = f$, автоматически являющиеся и решениями уравнения $\widetilde{A}u = f$, называют ***классическими решениями***.

Расширение области определения оператора достигают путем замены дифференциальных уравнений интегральными. Есть разные пути осуществления такой замены. Наиболее типичны две.

Если линейное пространство H гильбертово, а $Y = H$, то можно использовать общую схему приближенных методов. Выберем счетную последовательность элементов $v_k \in H$, линейная оболочка которой всюду плотна в H. Тогда равенство $Au = f$ равносильно системе равенств $(Au,\, v_k) = (f,\, v_k)$, $k \in \mathbb{N}$. Если линейный оператор A самосопряженный, а элементы v_k все принадлежат $D(A)$, то можно записать $(u,\, Av_k) = (f,\, v_k)$, $k \in \mathbb{N}$. В новой системе равенств требование $u \in D(A)$ (чтобы был определен элемент Au) можно опустить.

Если линейный оператор A в гильбертовом пространстве H самосопряженный и положительно определенный (т.е. $(Au,\, u) > 0$ при $u \in D(A)$ и $u \neq 0$), то выражение $(u,\, v)_A = (Au,\, v)$ определяет в $D(A)$ скалярное произведение. Функционал $F(u) = (Au,\, u) - 2\,(f,\, u)$ является выпуклым (т.е. удовлетворяет условию $F(\alpha u + \beta v) \leqslant \alpha F(u) + \beta F(v)$, $\alpha,\, \beta \in (0,\, 1)$, $\alpha + \beta = 1$), а точнее, строго выпуклым. Как и в конечномерном случае, строго выпуклого функционала максимум одна точка минимума. Но в данном случае можно показать, что квадратичный функционал имеет точку минимума. Если u_* — точка минимума F, то для достаточно малого по норме элемента Δu имеем $F(u_* + \Delta u) \geqslant F(u_*)$. Но

$$F(u_* + \Delta u) = (A(u_* + \Delta u),\, u_* + \Delta u) - 2\,(f,\, u_* + \Delta u) = F(u_*) + 2\,(Au_* - f,\, \Delta u) + (A\Delta u,\, \Delta u)\,.$$

Третье слагаемое по величине можно сделать много меньшим второго. Поэтому неравенство $F(u_* + \Delta u) \geqslant F(u_*)$ означает, что для любого вектора Δu выполняется неравенство

$(Au_* - f,\, \Delta u) \geqslant 0$. В этом неравенстве при изменении знака элемента Δu вся левая часть также меняет знак. Поэтому $(Au_* - f,\, \Delta u) = 0$. Ну, а отсюда уже следует, что $Au_* = f$. Наоборот, если $Au_* = f$, то $(Au_* - f,\, \Delta u) = 0$. Следовательно,

$$F(u_* + \Delta u) - F(u_*) = (A\Delta u,\, \Delta u)\,.$$

Правая часть в этом равенстве неотрицательна, а потому u_* является точкой минимума функционала $F(u)$.

Функционал $F(u)$ в вариационном исчислении называют ***функционалом энергии***.

7.2. Простейший пример

Рассмотрим одномерную краевую задачу

$$-(pu')' + qu = f,$$
$$u(0) = 0, \quad u'(1) = 0.$$

Здесь p, q, f — некоторые функции, определенные на отрезке $[0,\, 1]$. Для корректной интерпретации предполагаем, что q и f непрерывны, а p непрерывно дифференцируема.

В данном случае $Au = -(pu')' + qu$, а областью определения оператора A является класс непрерывно дифференцируемых функций, удовлетворяющих однородным граничным условиям. Можно показать, что этот класс всюду плотен в линейном пространстве $L_2[0,1]$ функций, интегрируемых с квадратом, т.е. в данном случае $H = L_2[0,1]$. Это линейное пространство гильбертово со скалярным произведением

$$(f,\, g) = \int\limits_0^1 f(x)\,g(x)\,dx.$$

Нетрудно убедиться в том, что линейный оператор A самосопряженный В самом деле,

$$(Au,\, v) = \int\limits_0^1 \big(-(pu')' + qu\big)v\,dx = -\int\limits_0^1 (pu')'v\,dx + \int\limits_0^1 quv\,dx =$$
$$= -pu'v\Big|_0^1 + \int\limits_0^1 pu'v'\,dx + \int\limits_0^1 quv\,dx = \int\limits_0^1 pu'v'\,dx + \int\limits_0^1 quv\,dx.$$

Видно, что $(Au,\, v) = (Av,\, u)$. Значит, A самосопряженный. Поскольку

$$(Au,\, u) = \int\limits_0^1 p(u')^2 dx + \int\limits_0^1 qu^2 dx,$$

нетрудно заключить, что оператор A положительно определенный при условии, что $p(x) \geqslant p_0 > 0$, $q(x) \geqslant 0$. Такие условия, как правило, вытекают из физических соображений. Будем считать, что они выполнены.

Итак, при сделанных предположениях оператор A является самосопряженным положительно определенным. Поэтому можно перейти к вариационной формулировке задачи, т.е. поиску точки минимума функционала энергии. В данном случае функционал энергии имеет вид

$$F[u] = \int\limits_0^1 \big(p(u')^2 + qu^2 - 2fu\big)dx.$$

Пусть $\{u_n\}$ — некоторый счетный базис в H. Предположим, что мы ищем приближенное решение в виде

$$\widetilde{u}_n = \sum_{k=1}^{n} \alpha_k u_k.$$

В качестве такого приближения естественно выбрать элемент, на котором функционал энергии достигает минимума на множестве всех линейных комбинаций функций $u_1, \ldots, u_n$. Но это множество (обозначим его H_n) — конечномерное линейное пространство с базисом $u_1, \ldots, u_n$. Поэтому поиск минимума функционала энергии на H_n — это задача конечномерной оптимизации. Функционал энергии относится к ***квадратичным функционалам***. Его сужение на конечномерное пространство H_n оказывается квадратичной функцией с положительно определенной квадратичной формой. Точка минимума такой функции единственна и может быть найдена решением системы линейных уравнений. Эта система имеет вид $Q\alpha = b$, где $Q = (Au_i,\, u_j)$, $b = ((f,\, u_i))$. Такой подход известен как ***метод Ритца***.

Отметим, что в данном случае можно ориентироваться не на счетный базис, а на последовательность конечномерных подпространств H_n, в каком-то смысле дающих в пределе гильбертово пространство H. При определенных условиях можно рассчитывать, что последовательность $\{\widetilde{u}_n\}$ точек минимума функционала $F[u]$ в этих подпространствах сходится к точке минимума функционала в H.

Последовательность подпространств $\{H_n\} \subset H$ назовем предельно плотной в H, Если для любого $u \in H$ имеем $\rho(u, H_n) \to 0$ при $n \to \infty$, где $\rho(u, H_n) = \inf\limits_{v \in H_n} \|u - v\|$.

Приступим к построению конечных элементов. Разобьем отрезок $[0,\, 1]$ на части точками $0 = x_0 < x_1 < \ldots < x_n = 1$. Для каждого узла x_i построим кусочно линейную функцию $u_i^{(n)}(x)$, равную 1 в точке x_i и нулю вне отрезка $[x_{i-1},\, x_{i+1}]$ (для крайних точек рассматриваем отрезки $[x_0,\, x_1]$ и $[x_{n-1},\, x_n]$). Такую функцию можно задать следующим образом:

$$u_i^{(n)}(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1},\, x_i]; \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & x \in [x_i,\, x_{i+1}]; \\ 0, & x \notin [x_{i-1},\, x_{i+1}]. \end{cases}$$

Каждый отрезок $[x_{i-1},\, x_i]$ выбранного разбиения вместе с его концами x_{i-1} и x_i и будет в данном случае конечным элементом. Концы отрезка будут его узлами. Две функции, связанные с узлами конечного элемента называют ***функциями формы***. Выбранные функции формы позволяют представить любую линейную на $[x_{i-1},\, x_i]$ функцию в виде линейной комбинации. Например, линейная функция $v(x)$, принимающая в точках x_{i-1} и x_i значения a_{i-1} и a_i, может быть представлена в виде

$$v(x) = a_{i-1}u_{i-1}^{(n)}(x) + a_i u_i^{(n)}(x), \quad x \in [x_{i-1},\, x_i].$$

Линейная комбинация $v(x) = \sum\limits_{i=0}^{n} a_i u_i^{(n)}(x)$ всех функций формы для данного разбиения представляет собой кусочно линейную функцию с узлами в точках x_i и значениями $v(x_i) = a_i$. Множество всех кусочно линейных функций изоморфно n-мерному линейному пространству. Найдем среди них ту, на которой функционал энергии принимает наименьшее значение. Для этого находим матрицу Q с элементами $\left(Au_i^{(n)},\, u_j^{(n)}\right)$:

$$\left(Au_i^{(n)},\, u_j^{(n)}\right) = \int_0^1 (pu_i'u_j' + qu_iu_j)\, dx.$$

Этот интеграл при произвольных функциях p и q аналитически вычислить нельзя. Но можно утверждать, что интеграл не равен нулю только для одинаковых или соседних узлов. Таким образом, матрица системы $Q\alpha = b$ является трехдиагональной. Один из простых способов вычисления интегралов — замена функций p, q, f их аппроксимациями того же рода. Например, представив $f \approx \sum\limits_{i=0} f_i u_i$, где $f_i = f(x_i)$, получаем

$$(f,\, u_i) \approx \int_0^1 \sum_{k=0} f_k u_k u_i\, dx = \sum_{k=0} f_k \int_0^1 u_k u_i\, dx = f_{i-1}\,(u_{i-1},\, u_i) + f_i\,(u_i,\, u_i) + f_{i+1}\,(u_{i+1},\, u_i)\,.$$

Короче говоря, для вычисления интегралов при такой аппроксимации достаточно иметь ***матрицу Грама*** $(u_i,\, u_j)$ базиса $u_1, \ldots, u_n$ в H_n. Здесь полезны формулы

$$\int_0^1 u_i^r dx = \frac{x_{i+1} - x_{i-1}}{r+1}, \quad r \in \mathbb{N};$$

$$\int_0^1 u_{i-1}^r u_i^s dx = \frac{r!s!}{(r+s+1)!}(x_i - x_{i-1}), \quad r,\, s \in \mathbb{N};$$

$$\int_0^1 u_i^r u_j^s dx = 0, \quad |i - j| > 1.$$

В этих формулах нужно уточнение при $i = 0$ и $i = n$:

$$\int_0^1 u_0^r dx = \frac{x_1 - x_0}{r+1}, \qquad \int_0^1 u_n^r dx = \frac{x_n - x_{n-1}}{r+1}.$$

7.3. Типы конечных элементов

Рассмотренный пример показывает главные моменты метода конечных элементов. Область краевой задачи разбивается на элементарные области. С каждой областью связывается набор узлов, а искомая функция заменяется в элементарной области некоторой интерполяционной функцией: эта функция определяется своими значениями в узлах конечного элемента. Переход от непрерывной краевой задачи к аппроксимации осуществляется на основе интегральной формулировки.

Из этих соображений вытекает, что элементарные области должны быть относительно простыми по структуре. Рассмотренный пример этого не показывает, но из теории интерполяции известно, что для аппроксимации могут учитываться не только значения самой функции в узлах, но и значения ее первой, а возможно и высших производных.

Классифицируя используемые конечные элементы, выделим два признака:

– используются производные функции или нет;

– по виду аппроксимирующих многочленов.

По первому признаку конечные элементы разделяются на ***лагранжевы*** и ***эрмитовы***. В первых не используются производные функции. В эрмитовых конечных элементах производные используются.

Терминология метода конечных элементов тесно связана с теорией многомерной интерполяции. Тот или иной конечный элемент должен обеспечивать эффективное решение интерполяционной задачи. Поэтому типы конечных элементов определяются видом областей и типов

многочленов, при которых обеспечивается удобная и эффективная интерполяция. Термины „лагранжев“ и „эрмитов“ связаны с понятиями „лагранжев сплайн“ и „эрмитов сплайн“.

Остановимся на лагранжевых конечных элементах. Среди них выделяются:

– ***симплексные конечные элементы***: используется линейная интерполяция, а элементарной областью является симплекс;

– ***комплексные конечные элементы***: в них используются полные многочлены степени два и выше (полный — значит, присутствуют все члены такого многочлена);

– ***мультиплексные конечные элементы***: используются неполные многочлены. Такие многочлены имеют одинаковую степень по каждой переменной, а элементарной областью является параллелепипед.

Остановимся подробнее на каждом типе конечных элементов.

Симплексные конечные элементы. Примером симплексного элемента является одномерные элементы из разобранного примера. Одномерным симплексом является отрезок и для такого симплекса было использовано два узла. В двумерном случае симплексом является треугольник, а в трехмерном — тетраэдр (треугольная пирамида). В общем случае в $\mathbb{R}^n$ ***симплексом*** называют многогранник, имеющий $n+1$ вершину.

Такое определение корректно и может употребляться только после того, как в арифметическом пространстве будут определены понятия, привычные для элементарной геометрии: плоскость, многогранник, грани и т.п. Впрочем, для наших целей достаточно считать, что симплекс — произвольный набор точек $x_1, \ldots, x_{n+1}$ в арифметическом пространстве $\mathbb{R}^n$, удовлетворяющий условию: векторы $x_2 - x_1$, $x_3 - x_1$, $\ldots$, $x_n - x_1$ линейно независимы. Отметим, что критерием этого является условие

$$\begin{vmatrix} x_1 & x_2 & \ldots & x_n \\ 1 & 1 & \ldots & 1 \end{vmatrix} \neq 0$$

(определитель записан в блочном виде, каждый символ x_i представляет собой столбец высоты n).

Узлами симплексного элемента являются его вершины. Рассмотрим в начале простой случай — двумерный. Тогда симплекс — это треугольник. Его узлами являются три вершины. Каждая функция в пределах треугольника аппроксимируется линейной, совпадающей с аппроксимируемой функцией в узлах. Из геометрических соображений легко заключить, что для любого треугольника с вершинами x_1, x_2, x_3 для любых значений f_1, f_2, f_3 существует, и притом единственная, линейная функция $\varphi(x)$, удовлетворяющая условиям $\varphi(x_i) = f_i$, $i = 1, 2, 3$. Функцию $\varphi(x)$ несложно представить в виде

$$\varphi(x) = f_1\varphi_1(x) + f_2\varphi_2(x) + f_3\varphi_3(x),$$

где линейная функция $\varphi_i(x)$ принимает значение 1 в узле x_i и значение 0 в двух остальных узлах. Вот эти три функции и будут функциями формы для двумерного симплекса.

Как построить функции формы $\varphi_i(x)$? Можно опять апеллировать к геометрии. Площадь треугольника с вершинами, например, x_1, x_2 и x является линейной функцией от координат точки x (проще всего это увидеть, если сторона x_1x_2 параллельна одной из осей координат; тогда одна из координат точки x есть высота треугольника x_1x_2x, опущенная на сторону x_1x_2). Следовательно, функция $\psi(x) = \dfrac{S_{x_1x_2x}}{S_{x_1x_2x_3}}$ является линейной и принимает значения: 1 в точке x_3; 0 в точках x_1 и $x - 2$. Другими словами, эта функция есть функция формы, отвечающая узлу x_3. Отметим, что площади треугольников $x_1x +_2 x$, x_1x_3x, x_2x_3x (в сумме они равны площади треугольника $x_1x_2x_3$), отнесенные к площади треугольника $x_1x_2x_3$, известны как ***барицентрические координаты точки x***.

Определение функций формы через барицентрические координаты открывает возможность определить функции формы аналитически. Для этого достаточно вычислить площади четырех

треугольников. Можно воспользоваться формулой

$$S_{x_1x_2x_3} = |\det(x_2 - x_1,\ x_3 - x_1)|,$$

которая получается, если воспользоваться для вычисления площади векторным произведением. Но, во-первых, раздражает знак модуля, а, во-вторых, надо еще проверить, попадает ли точка x внутрь треугольника (иначе функция формы некорректна).

Отметим, что

$$\det(x_2 - x_1,\ x_3 - x_1) = \det\begin{pmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{pmatrix}$$

(достаточно во втором определителе вычесть первый столбец из второго и третьего, а затем разложить по последней строке). Второй определитель более удобен, так как точки x_1, x_2, x_3 входят в него симметричным образом.

Положим

$$S_{x_1x_2x_3} = \begin{vmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{vmatrix}.$$

Если точка x принадлежит треугольнику $x_1x_2x_3$, то величины $S_{x_1x_2x_3}$ и $S_{x_1x_2x}$ имеют одинаковый знак. Точнее говоря, величина $S_{x_1x_2x_3}$ положительна, если вершины следуют в порядке, соответствующем движению против часовой стрелки. Таким образом, функции

$$\varphi_1(x) = \frac{S_{xx_2x_3}}{S_{x_1x_2x_3}}, \quad \varphi_2(x) = \frac{S_{x_1xx_3}}{S_{x_1x_2x_3}}, \quad \varphi_1(x) = \frac{S_{x_1x_2x}}{S_{x_1x_2x_3}}$$

представляют собой функции формы для треугольника $x_1x_2x_3$.

Полученные в двумерном случае результаты несложно экстраполировать на общий случай. Для симплекса с вершинами $x_1, \ldots, x_{n+1}$ полагаем

$$S = \begin{vmatrix} x_1 & x_2 & \ldots & x_{n+1} \\ 1 & 1 & \ldots & 1 \end{vmatrix}, \quad \varphi_i(x) = \frac{1}{S}\begin{vmatrix} x_1 & x_2 & \ldots & x_{i-1} & x & x_{i+1} & \ldots & x_{n+1} \\ 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 \end{vmatrix}.$$

Из общих свойств определителей вытекает, что функция $\varphi_i(x)$ является линейной и принимает значения: 1 в вершине x_i; 0 во всех остальных вершинах. Значит, она является функцией формы для рассматриваемого симплекса. Внутренними точками симплекса называют точки вида $\alpha_1x_1 + \ldots + a_{n+1}x_{n+1}$, где $\alpha_i \geqslant 0$, $i = \overline{1,\ n+1}$, и $\alpha_1 + \ldots + \alpha_{n+1} = 1$. Довольно легко показать, что для внутренних точек симплекса, и только для них, все функции формы положительны. В самом деле, любую точку $x \in \mathbb{R}^n$ можно представить в виде линейной комбинации $x = \alpha_1x_1 + \ldots + a_{n+1}x_{n+1}$, где α_i — произвольные действительные числа, в сумме составляющие единицу*, причем такое представление единственно. Точка x внутренняя для симплекса, когда все коэффициенты α_i этого представления положительны. Подставив указанное представление в функцию $\varphi_i(x)$, заключаем, что

$$\varphi_i(x) = \varphi_i\left(\sum_{j=1}^{n+1}\alpha_jx_j\right) = \sum_{i=j}^{n=1}\alpha_j\varphi_i(x_j) = \alpha_i.$$

Поэтому положительность коэффициентов разложения для x равносильна положительности значений всех функций формы на векторе x.

Наконец, уточним, что функция формы строится для каждого узла одна. При аппроксимации область разбивается на симплексы так, что симплексы прилегают друг к другу всей

*Записанное равенство равносильно равенству

$$x - x_1 = \alpha_2(x_2 - x - 1) + \ldots + \alpha_n(x_{n+1} - x_1),$$

представляющему собой разложение вектора $x - x_1$ по базису $x_2 - x_1, \ldots, x_{n+1} - x_1$.

гранью. В этом случае каждый узел является общей вершиной нескольких симплексов, причем совокупность всех симплексов, для которых точка x является вершиной, образует некоторую окрестность точки x. Функция формы строится одна для всей этой совокупности симплексов, причем в каждом симплексе она принимает значение 1 в вершине x. В двумерном случае графиком функции формы является пирамида, в основании которой находится многоугольник, образованный совокупностью треугольников, примыкающих к узлу.

В n-мерном случае скалярное произведение выражается n-мерным интегралом по n-мерной области от произведения двух функций. Для вычисления различных параметров, в том числе скалярных произведений полезна формула

$$\int\limits_V \varphi_{i_1}^{p_1}(x)\dots\varphi_{i_{n+1}}^{p_{n+1}}(x)\,dx = \frac{n!p_1!p_2!\dots p_{n+1}!}{(p_1+\dots+p_{n+1}+n)!}.$$

В этой формуле степени p_i могут быть любыми неотрицательными. При этом при $p_i = 0$ считаем, что $p_i! = 1$.

Комплексные конечные элементы. Все употребляемые типы конечных элементов определяются двумя обстоятельствами: а) методом разделения области на элементарные области; б) методом интерполяции функции в пределах каждой элементарной области по значениям в узлах. Симплексным конечным элементам соответствует разделение области на симплексы (в двумерном случае на треугольники; такое разделение называют триангуляцией). Но можно также разделять область и на другие элементарные области. Например, параллелепипеды.

В одномерном случае область краевой задачи есть промежуток. Его естественно разделять на отрезки. Поэтому все типы одномерных конечных элементов базируются на отрезках. Различие состоит в количестве узлов и их типе.

Выбрав на отрезке $N+1$ узел, можем функцию на этом отрезке интерполировать с помощью интерполяционного полинома Лагранжа степени N. Такой многочлен разлагается в сумму элементарных многочленов, каждый из которых в одном узле принимает значение 1, а в остальных — значение 0. Таких многочленов $N+1$, и они играют роль функций формы.

По расположению узлов описанные конечные элементы подразделяют на ***регулярные*** и ***сингулярные***. У регулярных конечных элементов оба конца отрезка являются узлами, у сингулярных один или оба конца не являются узлами.

При построении двумерных комплексных элементов аппроксимирующая функция должна быть полным многочленом степени s. Такой многочлен имеет $\frac{(s+1)(s+2)}{2}$ коэффициентов. Это значит, что для построения двумерного комплексного конечного элемента степени s в элементарной области нужно выбрать $\frac{(s+1)(s+2)}{2}$ узлов. Например, при $s=2$ нужно выбрать 6 узлов, при $s=3$ — 10 узлов. В качестве элементарной области рассмотрим треугольник. Тогда при $s=2$ узлы можно расположить в вершинах треугольника (три узла) и серединах его сторон (три узла). Такое расположение можно ассоциировать с разбиением базового треугольника на четыре подобных треугольника серединными линиями. Перенумеруем узлы следующим образом: *1*, *2*, *3* — вершины базового треугольника; *4*, *5*, *6* — середины сторон, противоположных вершинам *1*, *2*, *3*. Если φ_1, φ_2, φ_3 — функции формы симплексного элемента с данным треугольником, то функции формы $\varphi_i^{(2)}$ комплексного элемента степени 2 можно выразить через симплексные. В самом деле, легко убедиться, что функция $\varphi_1(2\varphi_1-1)$ обращается в нуль во всех узлах, кроме узла *1*, в котором эта функция равна единице. В остальных двух угловых узлах (в вершинах треугольника) функцию формы можно задать аналогично. В узле *4* функцию формы можно задать в виде $4\varphi_1\varphi_2$. Аналогично строятся функции формы в оставшихся двух узлах.

Комплексный треугольный элемент степени $s=3$ можно построить. разделив базовый треугольник на девять подобных прямыми, параллельными сторонам. При этом получим три узла в вершинах, три пары узлов на сторонах (они делят стороны на три равные части) и один

узел в точке пересечения медиан треугольника. Функции формы для таких конечных элементов также можно выразить через линейные функции формы для базового треугольника.

Мультиплексные конечные элементы. Мультиплексные лагранжевы элементы удобно строить на параллелепипедах со сторонами, параллельными осям координат. В этом случае можно реализовать разделение переменных в следующем смысле. Параллелепипед указанного типа — это область в пределах которой диапазоны изменения переменных не зависят друг от друга. Рассмотрим двумерный случай, $G = [a,\,b]\times[c,\,d]$. Тогда, выбрав на отрезке $[a,\,b]$ M узлов $x_1,\,\dots,\,x_M$, а на отрезке $[c,\,d]$ N узлов $y_1,\,\dots,\,y_N$, мы получим на прямоугольнике MN узлов $(x_i,\,y_j)$. При этом функции формы можно задать как произведение одномерных функций форм, т.е. интерполяционных многочленов Лагранжа.

Эрмитовы конечные элементы. Эрмитовы конечные элементы характеризуются тем, что в качестве узловых параметров используются не только значения функции в узлах, но и значения ее производных.

Одномерный конечный элемент строится на отрезке $[x_{i-1},\,x_i]$, $i=\overline{1,\,n}$. В эрмитовом случае можно в качестве узловых параметров использовать значения u_{i-1}, u'_{i-1}, u_i, u'_i функции и ее первой производной в концах отрезка. Как и в случае лагранжевых конечных элементов, каждому узловому параметру соответствует одна функция формы. Таким образом, на отрезке будет четыре функции формы: $\varphi_{i-1,0}$, $\varphi_{i-1,1}$, $\varphi_{i,0}$, $\varphi_{i,1}$. Например, последняя функция формы $\varphi_{i,1}$ имеет нулевое значение в узлах x_{i-1} и x_i, нулевую производную в узле x_{i-1} и единичную производную в узле x_i. С помощью четырех функций формы можно записать многочлен третьей степени, имеющий заданные значения функции и производной в узлах:

$$u(x) = u_{i-1}\varphi_{i-1,0} + u'_{i-1}\varphi_{i-1,1} + u_i\varphi_{i,0} + u'_i\varphi_{i,1}.$$

Запишем четыре функции формы. Для этого сперва удобно рассмотреть стандартный отрезок $[0,1]$, а затем линейным преобразованием распространить полученный вид функций на произвольный отрезок.

Для отрезка $[0,\,1]$ нетрудно выписать соответствующие функции:

$$\omega_{0,0}(\xi) = (1-\xi)^2(1+2\xi),\quad \omega_{0,1}(\xi) = (1-\xi)^2\xi,\quad \omega_{1,0}(\xi) = \xi^2(3-2\xi),\quad \omega_{1,1}(\xi) = \xi^2(\xi-1).$$

Для произвольного отрезка $[x_{i-1},\,x_i]$ достаточно выполнить замену $\xi = \frac{x-x_{i-1}}{h}$, $h = x_i - x_{i-1}$, и учесть, что такая замена приводит к тому, что производные по x получаются умножением производных по ξ на h. Получим

$$u_{i-1,0}(x) = \omega_{0,0}(\xi),\quad u_{i-1,1}(x) = h\omega_{0,1}(\xi),\quad u_{i,0}(x) = \omega_{1,0}(\xi),\quad u_{i,1}(x) = h\omega_{1,1}(\xi).$$

Как и в случае симплексных конечных элементов, реальная функция формы занимает две соседних элементарных области, примыкающих к соответствующему узлу:

$$u_{i,0}(x) = \begin{cases} \omega_{1,0}\Big(\frac{x-x_{i-1}}{x_i-x_{i-1}}\Big), & x\in[x_{i-1},\,x_i];\\ \omega_{0,0}\Big(\frac{x-x_i}{x_{i+1}-x_i}\Big), & x\in[x_i,\,x_{i+1}];\\ 0, & x\notin[x_{i-1},\,x_{i+1}]. \end{cases}$$

Примером двумерного эрмитова конечного элемента является треугольник с узлами в трех вершинах и в точке пересечения медиан. Для построения полного кубического многочлена от двух переменных необходимо 10 условий (по количеству коэффициентов). Такими условиями могут быть значения в четырех узлах (4 условия) и по две частные производные в угловых узлах (6 условий).

Соответствующие функции формы на самом деле могут быть выражены через три функции формы симплексного элемента.

7.4. Пример: уравнение теплопроводности

Рассмотрим задачу стационарной теплопроводности в некоторой области V. Такую задачу можно сформулировать следующим образом:

$$\begin{aligned}
&\operatorname{div}(\varkappa \operatorname{grad} u) + f = 0, \quad x \in V;\\
&u(x) = \mu_1, \quad x \in S_1;\\
&\varkappa \frac{\partial u}{\partial \boldsymbol{n}} + \beta u = \mu_2, \quad x \in S_2,
\end{aligned}$$

где $\varkappa$ — коэффициент теплопроводности, зависящий от точки пространства; S_1, S_2 — участки границы области V с разными граничными условиями (1-го рода на S_1, 3-го рода на S_2); β — коэффициент теплообмена.

Рассматриваемая задача связана с линейным оператором $Au = -\operatorname{div}(\varkappa \operatorname{grad} u)$. Можно утверждать, что при $\mu_1 = \mu_2 = \beta = 0$ этот оператор является самосопряженным положительно определенным. В этом случае задачу можно перевести в задачу поиска минимума функционала энергии $F(u) = \frac{1}{2}(Au,\, u) - (f,\, u)$. В общем случае линейный оператор A перестает быть самосопряженным, и такой переход уже сделать нельзя. Однако техника, использованная для получения функционала энергии позволяет и в этом случае от операторного уравнения перейти к задаче минимизации некоторого функционала, который в частном случае сводится к функционалу энергии. Вкратце техника эта такова.

Если функционал $F(u)$ достигает минимума в некоторой точке u_0, то для любого приращения δu, достаточно малого по норме, выполняется неравенство $F(u_0+\delta u) \leqslant F(u_0)$. Введем параметр t:

$$F(u_0 + t\delta u) \leqslant F(u_0).$$

Это соотношение при фиксированных u_0 и δu можно рассматривать как условие минимума в точке $t = 0$ функции одного переменного $\varphi(t) = F(u_0 + t\delta u)$. Если эта функция дифференцируема в точке $t = 0$, значение ее производной $\delta F(u_0, \delta u)$ при $t = 0$ должно обращаться в нуль. Величину $\delta F(u_0, \delta u)$ называют ***вариацией*** функционала $F(u)$ в точке u_0, соответствующей независимой вариации δu. В ряде случаев вариация функционала линейно зависит от δu и ее можно представить в виде $(dF(u_0),\, \delta u)$, где оператор $dF(u_0)$ представляет собой дифференциал функционала $F(u)$ в точке u_0. Например, для функционала энергии $F(u) = \frac{1}{2}(Au,\, u) - (f,\, u)$ дифференциал представляет собой оператор $Au - f$. Необходимым условием минимума является равенство $(dF(u_0),\, \delta u)$, верное для любого δu. Отсюда следует равенство $dF(u_0) = 0$. Наша задача обратная. Мы можем преобразовать операторное уравнение $Au = f$ в вариационное $(Au - f,\, \delta u) = 0$. Для вариационного уравнения, может быть, после некоторых преобразований, надо найти функционал, вариацией которого является выражение $(Au - f,\, \delta u)$ (или другое, полученное преобразованием). Если такой функционал найти удается, то задача сводится к поиску стационарных точек функционала. Если функционал выпуклый, то стационарные точки оказываются точками минимума. Именно эта схема в своем простейшем варианте и была использована для получения функционала энергии.

Опуская подробности, отметим, что в рассматриваемой задаче для оператора

$$Au = -\operatorname{div}(\varkappa \operatorname{grad} u),$$

несмотря на то, что он не является самосопряженным (при отсутствии однородных граничных условий), можно получить аналог функционала энергии, который выглядит следующим образом:

$$F(u) = \int\limits_V \Big(\frac{\varkappa}{2}(\operatorname{grad} u)^2 - fu\Big)\, dv + \int\limits_{S_2} \Big(\frac{\beta}{2}u^2 - \mu_2 u\Big)\, dS.$$

Сведение операторного уравнения к оптимизационной задаче позволяет использовать метод Ритца, состоящий в том, что искомая функция заменяется конечной линейной комбинацией базисных функций, коэффициенты которой ищутся из условия минимума. В качестве базисных функций выбираются функции формы.

Предъявленный функционал также определен не на всем гильбертовом пространстве, поскольку в нем используются производные первого порядка. Базисные функции должны попадать в область определения функционала. Это накладывает ограничения на выбор конечных элементов: функции формы должны быть по крайней мере „кусочно дифференцируемыми“, т.е. они должны быть непрерывными всюду в V и дифференцируемыми в V всюду, кроме, может быть некоторого „тощего“ множества (такого, значения на котором не влияют на значение интегралов). В данном случае можно использовать симплексные конечные элементы, а также какие-либо другие лагранжевы конечные элементы.

ОГЛАВЛЕНИЕ