РАЗДЕЛ III

Приближенное вычисление интегралов

Дополнительная литература:

1. Крылов В.И. Приближенное вычисление интегралов. М.: Наука, 1967.
2. Крылов В.И., Шульгина Л.Т. Справочная книга по численному интегрированию. М.: Наука, 1966.

ГЛАВА VI

Вычисление определенного интеграла

§ 1. Задача численного интегрирования: постановка, основные понятия

Пусть $f(x)$ – интегрируемая на отрезке $[a;b]$ функция. Ставится задача вычисления определенного интеграла $I=\int_a^b f(x)dx$. Если для функции $f(x)$ можно найти аналитическое выражение первообразной $F(x)$, то интеграл I можно вычислить, используя формулу Ньютона-Лейбница:

$$I=\int_a^b f(x)dx=F(b)-F(a).$$

Однако, как правило, выразить первообразную $F(x)$ через элементарные функции не удается. Поэтому приходиться прибегать к приближенному вычислению интеграла. Очевидно, одним из простейших приближенных алгоритмов, которые теоретически можно использовать для этих целей, является вычисление интеграла непосредственно по определению, с помощью интегральных сумм, в качестве одной из которых можно взять, например, такую:

$$S_n=\sum_{i=0}^{n} f(x_i)\Delta x_i\ .$$

Как следует из теории, таким образом интеграл I можно найти с любой наперед заданной точностью. Однако практически этот прием мало пригоден из-за медленной сходимости.

Поэтому для построения формул приближенного вычисления интеграла, как правило, используют следующий прием: функцию $f(x)$ заменяют близкой к ней функцией $\varphi(x)$, интеграл от которой просто может быть вычислен аналитически (с помощью формулы Ньютона-Лейбница) (например, алгебраическим многочленом). Успех приближения, как мы помним, зависит от свойств гладкости приближаемой функции. В силу этого подынтегральную функцию чаще всего представляют в виде произведения двух сомножителей, один из которых – достаточно гладкая функция (подлежащая в дальнейшем упрощающей замене), а вторая содержит основные особенности подынтегрального выражения и легко интегрируется, т.е. рассматривают интегралы вида $I=\int_a^b p(x)f(x)dx$. В этом вы-

ражении $p(x)$ – фиксированная, не эквивалентная нулю функция (ее мы далее будем называть весовой или просто весом), а $f(x)$ – достаточно гладкая функция, которую далее будем называть интегрируемой.
Пример:

$$\int_{-1}^{1} \frac{dx}{\sqrt{1-x^4}} = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} \frac{1}{\sqrt{1+x^2}} dx .$$

Здесь $p(x) = \frac{1}{\sqrt{1-x^2}}$, и $f(x) = \frac{1}{\sqrt{1+x^2}}$.

Если использовать интерполяционный способ замены $f(x)$ (причем линейный по параметрам), то приближенная формула для вычисления интеграла может выглядеть следующим образом:

$$I = \int_a^b p(x) f(x) dx \approx \sum_{k=0}^{n} A_k f(x_k). \tag{1.1}$$

При этом линейную комбинацию, стоящую в левой части соотношения (1.1), будем называть ***квадратурной суммой***, A_k – ее ***коэффициентами***, а x_k – ***узлами***.

При фиксированном n квадратурная сумма зависит от $2(n+1)$ параметров A_k и x_k ($k = \overline{0, n}$). Их выбор может осуществляться из следующих двух основных соображений:
1⁰. Повышение степени точности квадратурного правила.

Поясним, что имеется в виду. Поскольку формула (1.1) получается путем замены интегрируемой функции $f(x)$ некоторым обобщенным многочленом

$$Q_n(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \cdots + c_n\varphi_n(x)$$

с последующим его интегрированием, то можно ожидать, что если мы выбором узлов x_k и коэффициентов A_k в (1.1) достигнем хорошей точности в интегрировании функций $\varphi_i(x)$, то формула (1.1) должна будет также дать хороший результат (по точности) при вычислении интеграла от всякой функции $f(x)$ из рассматриваемого класса. Эти несложные соображения имеют, разумеется, только наводящее значение и погрешность построенной квадратурной формулы должна быть подвергнута точному анализу и оценке. Но они позволяют указать простой принцип выбора x_k и A_k: будем стремиться выбором x_k и A_k добиваться того, чтобы формула (1.1) давала точный результат для возможно большего числа первых функций $\varphi_i(x)$.

***Определение* 1.** Говорят, что квадратурная формула (1.1) имеет степень точности m относительно системы функций $\{\varphi_i(x)\}$, если она точна на первых m функциях $\varphi_0(x),...,\varphi_m(x)$ и не точна на функции $\varphi_{m+1}(x)$, т.е. выполняются соотношения

$$\begin{cases} \int_a^b p(x)\varphi_i(x)dx = \sum_{k=0}^{n} A_k\varphi_i(x_k), \ \ i = \overline{0, m}, \\ \int_a^b p(x)\varphi_{m+1}(x)dx \neq \sum_{k=0}^{n} A_k\varphi_{m+1}(x_k). \end{cases} \tag{1.2}$$

Если в качестве системы $\{\varphi_i(x)\}$ взять систему алгебраических многочленов, в частности, систему степеней, т.е. положить $\varphi_i(x) = x^i$, то из предыдущего определения следует определение ***алгебраической степени точности*** квадратурной суммы:

***Определение* 2.** Говорят, что квадратурная формула (1.1) имеет алгебраическую степень точности, равную m, если она точна для всевозможных многочленов степени m и существует хотя бы один многочлен степени $(m+1)$, для которого формула точной не является.

Требование точности для всевозможных многочленов степени m равносильно требованию точности на любой базисной системе пространства многочленов степени m, и в частности, на системе $\{x^i\}$. Поэтому из *Определения* 2 с необходимостью следует выполнение системы соотношений

$$\begin{cases} \int\limits_a^b p(x)x^i dx = \sum\limits_{k=0}^n A_k x_k^i, \ \ i = \overline{0, m}, \\ \int\limits_a^b p(x)x^{m+1} dx \neq \sum\limits_{k=0}^n A_k x_k^{m+1}. \end{cases} \tag{1.3}$$

Систему (1.3) можно использовать как для отыскания алгебраической степени точности заданной квадратурной формулы, так и для построения квадратурной формулы методом неопределенных коэффициентов. В последнем случае первые $(m+1)$ соотношений системы (1.3) рассматриваются как система уравнений (в общем случае нелинейных) относительно неизвестных A_k, x_k (или части их).

2^0. Минимизация остатка квадратурной формулы на классах функций.

Остатком квадратурной формулы естественно называть величину

$$R_n(f) = \int\limits_a^b p(x)f(x)dx - \sum_{k=0}^n A_k f(x_k) \tag{1.4}$$

Если при этом $f(x)$ принадлежит заданному классу F функций, то, очевидно, можно получить такую характеристику остатка: $\sup\limits_{f\in F}|R_n(f)|$, а в качестве задачи поставить задачу поиска минимума этой характеристики.

При этом необходимо иметь в виду следующие соображения: как правило, значения функции $f(x)$ в узлах x_k известны с некоторой погрешностью ε_k, $k = \overline{0, n}$. Эти погрешности при вычислении квадратурной суммы повлекут за собой ошибку $\sum\limits_{k=0}^n A_k\varepsilon_k$. Если считать, что для всех значений k погрешности ε_k ограничены по модулю сверху, т.е. $|\varepsilon_k| \le \varepsilon$, то для погрешности квадратурной суммы получим оценку

$$\left|\sum_{k=0}^n A_k\varepsilon_k\right| \le \varepsilon\sum_{k=0}^n |A_k|.$$

Поэтому необходимо подбирать коэффициенты A_k таким образом, чтобы величина $\sum\limits_{k=0}^n |A_k|$ была по возможности меньшей.

Пусть $p(x) \geq 0$ на отрезке $[a;b]$ и квадратурная формула (1.1) имеет алгебраическую степень точности, не меньшую нуля. Тогда

$$\sum_{k=0}^{n} A_k = \int_a^b p(x)dx > 0 .$$

С другой стороны,

$$\sum_{k=0}^{n} |A_k| \geq \sum_{k=0}^{n} A_k ,$$

причем равенство имеет место только в том случае, когда все A_k положительны. Поэтому условие знакопостоянства коэффициентов квадратурной суммы обеспечивает наименьшую оценку вычислительной погрешности квадратурной формулы, т.е. ее устойчивость.

Помимо отмеченных двух основных соображений в основу выбора параметров квадратурных формул могут быть положены и некоторые другие (например, распределение той же погрешности по заданному закону и т.п.).

§ 2. Интерполяционные квадратурные формулы

Как мы уже отмечали, для построения квадратурных формул чаще всего пользуются интерполированием интегрируемой функции, при этом наиболее употребительный класс приближающих функций – алгебраические многочлены.

Выберем на отрезке интегрирования $[a;b]$ $(n+1)$ произвольных различных точек $x_0, x_1, ..., x_n$ и проинтерполируем функцию $f(x)$ по ее значениям в этих точках. Интерполяционный многочлен в данном случае удобнее брать в форме Лагранжа:

$$f(x) = P_n(x) + r_n(x),$$

где

$$P_n(x) = \sum_{k=0}^{n} \Phi_k(x) f(x_k) = \sum_{k=0}^{n} \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)} f(x_k),$$

а $r_n(x)$ – остаток интерполирования.

Отсюда получим:

$$I = \int_a^b p(x)f(x)dx = \sum_{k=0}^{n} A_k f(x_k) + R_n(f), \qquad (2.1)$$

причем

$$A_k = \int_a^b p(x)\Phi_k(x)dx = \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)} dx , \qquad (2.2)$$

а

$$R_n(f) = \int_a^b p(x) r_n(x) dx .$$

Если остаток интерполирования $r_n(x)$ мал, то и величина $R_n(f)$ также будет малой и, следовательно, в (2.1) ей можно пренебречь. В итоге получим квадратурную формулу

$$I = \int_a^b p(x) f(x) dx \approx \sum_{k=0}^{n} A_k f(x_k). \qquad (2.3)$$

Квадратурные формулы (2.3), коэффициенты которых вычисляются по формулам (2.2), называют ***интерполяционными***.

Интерполяционные квадратурные формулы могут быть охарактеризованы следующей простой теоремой.

Теорема. Для того чтобы квадратурная формула (2.3) была интерполяционной, необходимо и достаточно, чтобы она была точной для всевозможных многочленов степени не выше n (т.е. имела алгебраическую степень точности, равную n).

Доказательство.

Необходимость. Пусть рассматриваемая квадратурная формула является интерполяционной, а $f(x)$ – алгебраический многочлен степени не выше n. Тогда, очевидно, $r_n(x) \equiv 0$ (см., например, представление остатка интерполирования в форме Лагранжа). Следовательно, и $R_n(f) = 0$, т.е. квадратурная формула точна для таких $f(x)$.

Достаточность. Пусть квадратурная формула точна для всех многочленов до степени n включительно. Докажем, что ее коэффициенты вычисляются по формулам (2.2), т.е. что

$$A_k = \int_a^b p(x) \Phi_k(x) dx, \quad k = \overline{0,n}.$$

Рассмотрим $\int_a^b p(x) \Phi_k(x) dx$. Так как $\Phi_j(x)$ – многочлен степени n, то для него квадратурная формула точна, т.е.

$$\int_a^b p(x) \Phi_j(x) dx = \sum_{k=0}^{n} A_k \Phi_j(x_k),$$

а поскольку $\Phi_j(x_k) = \delta_j^k$, то отсюда следует, что

$$\int_a^b p(x) \Phi_j(x) dx = \sum_{k=0}^{n} A_k \Phi_j(x_k) = A_j, \quad j = \overline{0,n}.$$

⊠

Из доказанной теоремы следует, что любая квадратурная формула, точная для многочленов степени не выше n и имеющая $(n+1)$ узел, является интерполяционной.

С помощью имеющихся представлений для остатка интерполирования $r_n(x)$ мы можем получить различные представления для остатка квадратурной формулы $R_n(f)$.

Пусть, например, $f(x) \in C^{n+1}[a;b]$. Тогда остаток интерполирования в форме Лагранжа имеет вид

$$r_n(x) = \omega_{n+1}(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad \xi \in [a;b].$$

Отсюда получаем:

$$R_n(f)=\frac{1}{(n+1)!}\int_a^b p(x)\omega_{n+1}(x)f^{(n+1)}(\xi)dx\,. \tag{2.4}$$

Если допустить, что $\left|f^{(n+1)}(x)\right|\le M$ для всех $x\in[a;b]$, то для остатка квадратурной формулы получим оценку сверху

$$|R_n(f)|\le\frac{M}{(n+1)!}\int_a^b|p(x)\omega_{n+1}(x)|dx\,, \tag{2.5}$$

которая является вычислимой и может быть, следовательно, использована для практической (априорной) оценки погрешности численного интегрирования. Заметим, что, как и в случае интерполирования, оценка (2.5) является достижимой.

2.1. Квадратурные формулы Ньютона-Котеса

Среди интерполяционных квадратурных формул ранее других были построены формулы Ньютона-Котеса. Они относятся к случаю равноотстоящих узлов.

Отрезок $[a;b]$ разделим на n равных частей длины $h=\frac{b-a}{n}$ и точки деления $x_k=a+kh$, $k=\overline{0,n}$, примем за узлы интерполяционной квадратурной формулы. Саму формулу запишем в виде

$$\int_a^b p(x)f(x)dx\approx(b-a)\sum_{k=0}^{n}B_k^n f(a+kh)\,, \tag{2.6}$$

где

$$B_k^n=\frac{1}{b-a}A_k=\frac{1}{b-a}\int_a^b p(x)\frac{\omega_{n+1}(x)}{(x-a-kh)\omega'_{n+1}(a+kh)}dx\,,\ k=\overline{0,n}\,. \tag{2.7}$$

Введем новую переменную по формуле $x=a+th$, т.е. $t=\frac{x-a}{h}$. Тогда имеем:

$0\le t\le n$; $b-a=nh$; $x-a-kh=th-kh=h(t-k)$; $dx=hdt$;

$\omega_{n+1}(x)=\omega_{n+1}(a+th)=h^{n+1}t(t-1)\cdots(t-n)$;

$\omega'_{n+1}(x_k)=\omega'_{n+1}(a+kh)=(x_k-a)(x_k-a-h)\cdots(x_k-a-(k-1)h)\cdot(x_k-a-(k+1)h)\cdot\ldots\cdot$

$\cdot(x_k a-nh)=kh\cdot(k-1)h\cdot\ldots\cdot h\cdot(-1)\cdot h\cdot(-2)\cdot h\cdot\ldots\cdot(-1)(n-k)\cdot h=(-1)^{n-k}h^n k!\cdot(n-k)!$.

Подставляя эти выражения в формулу (2.7), получим:

$$B_k^n=\frac{(-1)^{n-k}}{k!\cdot(n-k)!\cdot(b-a)}\int_0^n p(a+th)\frac{h^{n+1}t(t-1)\cdots(t-n)}{h(t-k)\cdot h^n}\cdot hdt=$$

$$=\frac{(-1)^{n-k}}{n\cdot k!\cdot(n-k)!}\int_0^n p(a+th)\frac{t(t-1)\cdots(t-n)}{(t-k)}dt\,,\ \ k=\overline{0,n}\,. \tag{2.8}$$

Отметим некоторые дополнительные свойства коэффициентов (2.8) и квадратурной формулы (2.6) в случае постоянной весовой функции ($p(x) \equiv 1$).

1^0. $B_k^n = B_{n-k}^n$, т.е. равноотстоящие от концов суммы коэффициенты равны.

Действительно,

$$B_{n-k}^n = \frac{(-1)^k}{n \cdot k! \cdot (n-k)!} \int_0^n \frac{t(t-1)\cdots(t-n)}{t-n+k}\,dt = [\text{делаем замену переменной } t = n - z] =$$

$$= \frac{(-1)^k}{n \cdot k! \cdot (n-k)!} \int_n^0 \frac{(n-z)(n-z-1)(n-z-2)\cdots(-z)}{-z+k}(-dz) =$$

$$= \frac{(-1)^k}{n \cdot k! \cdot (n-k)!} \int_0^n \frac{(-1)^{n+1} z(z-1)\cdots(z-n)}{-(z-k)}\,dz = \frac{(-1)^{n+k}}{n \cdot k! \cdot (n-k)!} \int_0^n \frac{z(z-1)\cdots(z-n)}{z-k}\,dz = B_k^n .$$

2^0. Квадратурная формула (2.6) точна для любой функции $f(x)$, нечетной относительно середины отрезка $[a;b]$ (т.е. функции, для которой выполняется соотношение $f\left(x - \frac{a+b}{2}\right) = \ = -f\left(\frac{a+b}{2} - x\right)$).

В самом деле, выполняя в интеграле замену переменной по формуле $t = x - \frac{a+b}{2}$ (относительно переменной t функция f будет просто нечетной), имеем:
с одной стороны

$$\int_a^b f(x)dx = \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} f(t)dt = 0 ,$$

а с другой

$$\sum_{k=0}^{n} B_k^n f(a+kh) = \left(B_0^n f(a) + B_n^n f(b)\right) + \left(B_1^n f(a+h) + B_{n-1}^n f(b-h)\right) + \cdots = 0$$

(причем в каждой скобке сумма равна нулю как сумма противоположных слагаемых, а то слагаемое, которому нет пары (если такое имеется), само равно нулю, поскольку $f\left(\frac{a+b}{2}\right) = 0$). Таким образом, $\int_a^b f(x)dx = (b-a)\sum_{k=0}^{n} B_k^n f(a+kh) = 0$.

3^0. При четном значении n (тогда общее количество узлов нечетно и, как следствие, точка $x = \frac{a+b}{2}$ также является узлом) квадратурная формула (2.6) имеет алгебраическую степень точности, равную $(n+1)$.

Действительно, так как формула (2.6) является интерполяционной, то она точна для всех алгебраических многочленов до степени n включительно, а в силу свойства **2^0** она точна и для многочлена $c\left(x - \frac{a+b}{2}\right)^{n+1}$. Следовательно, она будет точной и для базисной функции x^{n+1}.

Замечание. Свойство $\mathbf{3^0}$, по сути, означает, что мы проводим интерполирование по n простым узлам x_k и одному двукратному $x^* = \dfrac{a+b}{2}$, поскольку квадратурные формулы, получающиеся в означенном случае и при интерполировании по $(n+1)$ простому узлу, получаются одинаковыми. Это имеет, следовательно, значение при исследовании остатка.

Было установлено также, что при $n = 8$ в формуле (2.6) появляются отрицательные коэффициенты, а при $n \geq 10$ среди B_k^n ***обязательно*** будут отрицательные, причем при больших n – сколь угодно большие по модулю (в силу того, что сумма всех таких коэффициентов постоянна и равна единице). Поэтому

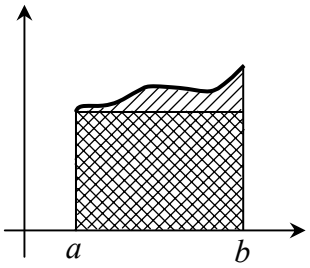
$$\sum_{k=0}^{n} \left| B_k^n \right| \underset{n\to\infty}{\to} \infty ,$$

что приводит к катастрофическому накоплению вычислительной погрешности. В силу этого квадратурные формулы Ньютона-Котеса с большим количеством узлов, как правило, не используются. Для достижения же высокой точности процедуры приближенного интегрирования исходный отрезок $[a;b]$ разбивают на отрезки небольшой длины. Тогда на каждом из них можно получить хороший результат уже при небольших значениях n. Получаемые таким образом квадратурные формулы называют ***составными*** или ***обобщенными***.

2.2. Примеры квадратурных формул с равноотстоящими узлами

1^0. Начнем рассмотрение со случая $n = 0$ (один узел интерполирования), который по формальным причинам не охватывается рассмотренными выше формулами Ньютона-Котеса. Получающиеся квадратурные формулы, исходя из геометрических соображений, носят название ***формул прямоугольников***.

а) *формула левых прямоугольников*



Единственным узлом интерполирования в данном случае предполагается левый конец отрезка интегрирования, т.е. $x_0 = a$. Интерполяционным многочленом для функции $f(x)$ по этим данным будет $P_0(x) = f(a)$. Подставляя это выражение вместо $f(x)$ под знак интеграла и выполняя интегрирование, получим искомую квадратурную формулу:

$$\int_a^b f(x)dx \approx (b-a)f(a). \qquad (2.9)$$

Заметим, что с геометрической точки зрения в правой части формулы (2.9) мы имеем площадь прямоугольника, одна сторона которого равна длине отрезка интегрирования, а вторая – значению интегрируемой функции $f(x)$ на левом конце данного отрезка (отсюда – название квадратурной формулы).

Если $f(x) \in C^1[a;b]$, то остаток интерполирования в форме Лагранжа имеет вид $r_0(x) = (x-a)f'(\xi)$, где $\xi \in [a;b]$ (и зависит от x (!)).

Следовательно, $R_0^{Л}(f) = \int_a^b r_0(x)dx = \int_a^b (x-a)f'(\xi)dx$. Чтобы упростить это выражение, воспользуемся теоремой о среднем. Поскольку функция $(x-a)$ сохраняет знак на отрезке интегрирования (неотрицательна), то ($\eta \in [a;b]$)

$$R_0^{Л}(f) = f'(\eta)\int_a^b (x-a)dx = \frac{(b-a)^2}{2} f'(\eta). \qquad (2.10)$$

Таким образом, (2.9) – квадратурная формула левых прямоугольников, а (2.10) – ее остаток. Легко видеть, что алгебраическая степень точности формулы (2.9) равна нулю, а с помощью (2.10) можно получить оценку сверху для величины погрешности численного интегрирования. Однако управлять величиной этой погрешности, очевидно, не представляется возможным. Поэтому построим сейчас на базе (2.9) ***составную формулу левых прямоугольников***.

Разобьем отрезок $[a;b]$ на N частей длины $h = \frac{b-a}{N}$, воспользуемся свойством аддитивности интеграла и на каждом из отрезков получившегося разбиения применим квадратурную формулу левых прямоугольников (2.9):

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)dx \approx h\sum_{k=0}^{N-1} f(x_k) = h\sum_{k=0}^{N-1} f(a+kh). \qquad (2.11)$$

(2.11) – составная (обобщенная) формула левых прямоугольников. Очевидно, для ее остатка по аналогии с (2.10) можно записать представление

$$R_{Л}(f) = \frac{h^2}{2}\sum_{k=0}^{N-1} f'(\xi_k), \;\; \xi_k \in [x_k; x_{k+1}].$$

Упростим полученное представление остатка. Поскольку функция $f'(x)$ непрерывна на отрезке $[a;b]$, то в соответствии с теоремой Вейерштрасса она достигает на нем своих наименьшего и наибольшего значений, т.е. $m = \min\limits_{x\in[a;b]} f'(x), \;\; M = \max\limits_{x\in[a;b]} f'(x)$, причем

$$m \le f'(x) \le M \;\text{ для всех }\; x \in [a;b].$$

Следовательно,

$$Nm \le \sum_{k=0}^{N-1} f'(\xi_k) \le NM \;\text{ и }\; m \le \frac{1}{N}\sum_{k=0}^{N-1} f'(\xi_k) \le M\,.$$

Тогда по теореме о промежуточном значении непрерывной функции существует точка $\eta \in [a;b]$, в которой выполняется соотношение $f'(\eta) = \frac{1}{N}\sum_{k=0}^{N-1} f'(\xi_k)$. Поэтому

$$R_{Л}(f) = \frac{h^2}{2}\sum_{k=0}^{N-1} f'(\xi_k) = \frac{h}{2}\cdot\frac{b-a}{N}\sum_{k=0}^{N-1} f'(\xi_k) = h\frac{b-a}{2} f'(\eta) = \frac{(b-a)^2}{2N} f'(\eta), \;\; \eta \in [a;b]. \qquad (2.12)$$

Очевидно, степень точности не повысилась, но $R_{Л}(f) \underset{N\to\infty}{\to} 0$ и, таким образом, у нас появляется рычаг, с помощью которого можно воздействовать на величину погрешности.

Замечание. Точки разбиения не обязаны быть, вообще говоря, равноотстоящими. В этом более общем случае составная формула левых прямоугольников будет иметь вид

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)dx \approx \sum_{k=0}^{N-1} (x_{k+1} - x_k) f(x_k) = \sum_{k=0}^{N-1} h_{k+1} f(x_k), \qquad (2.11')$$

а ее остаток будет таким:

$$R_{Л}(f) = \sum_{k=0}^{N-1} \frac{h_{k+1}^2}{2} f'(\xi_k)$$

и упростить его до вида, аналогичного представлению (2.12), не удается.

В дальнейшем замечаний по поводу неравномерной сетки более делать не будем.

б) *формула правых прямоугольников*

Здесь единственный узел интерполирования $x_0 = b$. Если $f(x) \in C^1[a;b]$, то по аналогии с рассуждениями пункта а) получим аналогичную по точности квадратурной формуле левых прямоугольников формулу ***правых прямоугольников***:

$$\int_a^b f(x)dx \approx (b-a) f(b). \qquad (2.13)$$

и ее остаток

$$R_0^{П}(f) = f'(\eta) \int_a^b (x-b)dx = -\frac{(b-a)^2}{2} f'(\eta), \ \eta \in [a;b], \qquad (2.14)$$

а также ***составную формулу правых прямоугольников***

$$\int_a^b f(x)dx = \sum_{k=1}^{N} \int_{x_k-1}^{x_k} f(x)dx \approx h \sum_{k=1}^{N} f(x_k) = h \sum_{k=1}^{N} f(a+kh) \qquad (2.15)$$

и ее остаток

$$R_{П}(f) = -h \frac{b-a}{2} f'(\eta) = -\frac{(b-a)^2}{2N} f'(\eta), \ \eta \in [a;b]. \qquad (2.16)$$

Заметим, что в случае, когда производная $f'(x)$ сохраняет знак на отрезке $[a;b]$, квадратурные формулы правых и левых прямоугольников дают оценку значения нужного значения интеграла с двух сторон.

в) *формула средних прямоугольников*

Как и в двух предыдущих вариантах, для приближения интегрируемой функции используется единственный узел. В этом случае – это середина отрезка интегрирования, т.е. $x_0 = \frac{a+b}{2}$. Интегрирование соответствующего интерполяционного многочлена дает квадратурную формулу

$$\int_a^b f(x)dx \approx (b-a) f\left(\frac{a+b}{2}\right). \qquad (2.17)$$

Для вывода же формулы остатка воспользуемся замечанием из предыдущего пункта. Симметричное расположение единственного узла дает возможность повысить алгебраическую степень точности формулы (2.17) до единицы и считать интерполяционную замену интегрируемой функции интерполированием с кратным узлом. Остаток соответствующей формулы Эрмита в случае $f(x) \in C^2[a;b]$ имеет вид

$$r_0(x)=\left(x-\frac{a+b}{2}\right)^2\frac{f''(\xi)}{2}.$$

Поэтому

$$R_0^C(f)=\int_a^b\left(x-\frac{a+b}{2}\right)^2\frac{f''(\xi)}{2}dx=\frac{f''(\eta)}{2}\int_a^b\left(x-\frac{a+b}{2}\right)^2dx=\frac{(b-a)^3}{24}f''(\eta),\ \eta\in[a;b], \quad (2.18)$$

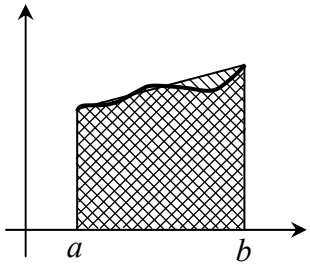
По аналогии с предыдущими двумя случаями получаем ***составную формулу средних прямоугольников***

$$\int_a^b f(x)dx=\sum_{k=0}^{N-1}\int_{x_k}^{x_{k+1}}f(x)dx\approx h\sum_{k=0}^{N-1}f\left(\frac{x_k+x_{k+1}}{2}\right)=h\sum_{k=0}^{N-1}f\left(a+kh+\frac{h}{2}\right) \quad (2.19)$$

и ее остаток

$$R_C(f)=h^2\frac{b-a}{24}f''(\eta)=\frac{(b-a)^3}{24N^2}f''(\eta),\ \eta\in[a;b]. \quad (2.20)$$

2^0. *Формула трапеций*



Квадратурная формула трапеций (малая) получается как частный случай формулы Ньютона-Котеса (2.6) (или непосредственно) при $n=1$ и имеет вид

$$\int_a^b f(x)dx\approx\frac{b-a}{2}(f(a)+f(b)). \quad (2.21)$$

Остаток интерполирования многочленом первой степени при $f(x)\in C^2[a;b]$ имеет вид

$$r_1(x)=(x-a)(x-b)\frac{f''(\xi)}{2}.$$

Поэтому (вновь используя факт знакопостоянства одного из сомножителей подынтегральной функции) имеем

$$R_1^T(f)=\int_a^b(x-a)(x-b)\frac{f''(\xi)}{2}dx=\frac{f''(\eta)}{2}\int_a^b(x-a)(x-b)dx=-\frac{(b-a)^3}{12}f''(\eta),\ \eta\in[a;b], \quad (2.22)$$

По аналогии с предыдущими случаями получаем ***составную формулу трапеций***

$$\int_a^b f(x)dx=\sum_{k=0}^{N-1}\int_{x_k}^{x_{k+1}}f(x)dx\approx\frac{h}{2}\sum_{k=0}^{N-1}[f(x_k)+f(x_{k+1})]=h\left[\frac{f(a)+f(b)}{2}+\sum_{k=1}^{N-1}f(x_k)\right] \quad (2.23)$$

и ее остаток

$$R_T(f)=-h^2\frac{b-a}{12}f''(\eta)=-\frac{(b-a)^3}{12N^2}f''(\eta),\ \eta\in[a;b]. \quad (2.24)$$

Пара квадратурных формул трапеций и средних прямоугольников также при определенных условиях (знакопостоянство второй производной интегрируемой функции на отрезке интегрирования) дает двустороннее приближение.

3^0. *Формула Симпсона* (*парабол*)

Указанная квадратурная формула получается как частный случай квадратурной формулы Ньютона-Котеса при $n = 2$. Таким образом, интерполирование интегрируемой функции $f(x)$ в этом случае осуществляется по трем узлам, равномерно, включая концы, расположенным на отрезке интегрирования: $x_0 = a,\ x_1 = \dfrac{a+b}{2},\ x_2 = b$. По формуле (2.8) при $p(x) \equiv 1$ (учитывая свойство $\mathbf{1^0}$) найдем:

$$B_0^2 = B_2^2 = \frac{(-1)^{2-0}}{2\cdot 0!\cdot 2!}\int_0^2 \frac{t(t-1)(t-2)}{t}dt = \frac{1}{4}\int_0^2 \left(t^2 - 3t + 2\right)dt = \frac{1}{6},$$

$$B_1^2 = \frac{(-1)^{2-1}}{2\cdot 1!\cdot 1!}\int_0^2 \frac{t(t-1)(t-2)}{t-1}dt = -\frac{1}{2}\int_0^2 \left(t^2 - 2t\right)dt = \frac{4}{6}.$$

Следовательно, ***квадратурная формула Симпсона*** имеет вид

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \tag{2.25}$$

Ее алгебраическая степень точности равна 3. Остаток формулы Симпсона получим, используя замечание о влиянии симметричного расположения узлов интерполировании. Остаток многочлена Эрмита в нашем случае имеет вид

$$r_2(x) = (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\frac{f^{(4)}(\xi)}{4!},$$

если $f(x) \in C^4[a;b]$. При этом, очевидно, многочленный сомножитель остатка знакопостоянен. Поэтому интегрирование данного остатка с применением теоремы о среднем дает

$$R_2^C(f) = \int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\frac{f^{(4)}(\xi)}{4!}dx = \frac{f^{(4)}(\eta)}{4!}\int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)dx = \tag{2.26}$$

$$= -\left(\frac{b-a}{2}\right)^5 \frac{f^{(4)}(\eta)}{90} = -\frac{(b-a)^5}{2880} f^{(4)}(\eta).$$

Составная формула Симпсона в литературе встречается в двух вариантах:

а) отрезок $[a;b]$ разбивается на произвольное количество (N) частей и на каждой из этих частей интеграл заменяется формулой типа (2.25). В итоге получим квадратурную формулу

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1}\int_{x_k}^{x_{k+1}} f(x)dx \approx \frac{h}{6}\sum_{k=0}^{N-1}\left[f(x_k) + 4f\left(\frac{x_k + x_{k+1}}{2}\right) + f(x_{k+1})\right]. \tag{2.27}$$

Ее остаток по аналогии с рассмотренными выше случаями с использованием формулы (2.26) и имеет вид

$$R_C(f) = -h^4 \frac{b-a}{2880} f^{(4)}(\eta) = -\frac{(b-a)^5}{2880N^4} f^{(4)}(\eta),\ \eta \in [a;b]. \tag{2.28}$$

б) Во второй версии отрезок интегрирования разбивается на ***четное*** количество $(2N)$ частей, и интеграл заменяется формулой Симпсона на ***каждой паре*** отрезков разбиения (т.е. на отрезке длиной $2h$). Такой вариант составной формулы Симпсона выглядит следующим образом:

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_{2k}}^{x_{2k+2}} f(x)dx \approx \frac{2h}{6} \sum_{k=0}^{N-1} [f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2})] =$$

$$= \frac{h}{3}[f(a) + f(b)] + \frac{4h}{3}[f(a+h) + f(a+3h) + \cdots + f(b-h)] + \tag{2.29}$$

$$+ \frac{2h}{3}[f(a+2h) + f(a+4h) + \cdots + f(b-2h)],$$

а ее остаток примет вид

$$R_C(f) = -h^4 \frac{b-a}{180} f^{(4)}(\eta) = -\frac{(b-a)^5}{2880N^4} f^{(4)}(\eta),\ \eta \in [a;b]. \tag{2.30}$$

Упражнение. Рассмотреть случай $n = 3$ (***квадратурная формула «трех восьмых»***).

2.3. Оценка погрешности квадратурных формул

Заметим, что все полученные выше представления остатков квадратурных формул (как простых, так и составных) позволяют решить задачу об априорной оценке погрешности приближенного значения интеграла, доставляемого соответствующей квадратурной формулой. Кроме того, в случае составных формул с их помощью можно выбрать параметры заданной квадратурной формулы (число N частей, на которое необходимо разбивать отрезок интегрирования или шаг разбиения) таким образом, чтобы модуль остатка был не более некоторой заданной величины ε (пользовательского требования к точности).

Так, например, если исходная квадратурная формула есть составная формула левых или правых прямоугольников, то $|R_0(f)| \le \frac{(b-a)^2 M_1}{2N} \le \varepsilon$, где $M_k = \max\limits_{x\in[a;b]} \left| f^{(k)}(x) \right|$, ε – величина пользовательского требования к точности, откуда для величины N получаем оценку

$$N_П \ge \frac{(b-a)^2 M_1}{2\varepsilon}. \tag{2.31}$$

Аналогично для составной формулы средних прямоугольников

$$N_C \ge \sqrt{\frac{(b-a)^3 M_2}{24\varepsilon}}, \tag{2.32}$$

для формулы трапеций

$$N_T \geq \sqrt{\frac{(b-a)^3 M_2}{12\varepsilon}}, \tag{2.32}$$

для формулы Симпсона (2.28)

$$N_S \geq \sqrt[4]{\frac{(b-a)^5 M_4}{2880\varepsilon}}, \tag{2.32}$$

Однако существенным недостатком указанного подхода к вычислению интеграла с заданной точностью является необходимость аналитического вычисления оценок норм производных от интегрируемой функции (в том числе достаточно высоких порядков), что далеко не всегда представляет собой простую задачу.

2.3.1. Учет избыточной гладкости интегрируемой функции

Все приведенные выше представления для остатков квадратурных формул были получены в предположении, что интегрируемая функция принадлежит вполне определенному классу гладкости. Естественно, если гладкость оказывается недостаточной, то соответствующее представление также не имеет места. В то же время, если функция обладает большей по сравнению с минимально необходимой степенью гладкости, то это позволяет при сохранении порядка выделить из погрешности составной квадратурной формулы некоторое количество последовательных главных частей.

Покажем, как это можно сделать, на примере составной формулы средних прямоугольников. Пусть, скажем, $f(x) \in C^4[a;b]$. Тогда остаток формулы средних прямоугольников есть величина порядка $O(h^2)$ (см. (2.20)). В то же время, раскладывая интегрируемую функцию на отрезке $[x_k ; x_{k+1}]$ в ряд Тейлора в точке $x_{k+\frac{1}{2}}$ и учитывая, что интеграл по указанному отрезку от нечетных степеней разности $\left(x - x_{k+\frac{1}{2}}\right)$ равен нулю, можем записать:

$$\int_{x_k}^{x_{k+1}} f(x)dx =$$

$$= \int_{x_k}^{x_{k+1}} \left[f\left(x_{k+\frac{1}{2}}\right) + \left(x - x_{k+\frac{1}{2}}\right) f'\left(x_{k+\frac{1}{2}}\right) + \frac{\left(x - x_{k+\frac{1}{2}}\right)^2}{2} f''\left(x_{k+\frac{1}{2}}\right) + \frac{\left(x - x_{k+\frac{1}{2}}\right)^3}{6} f'''\left(x_{k+\frac{1}{2}}\right) + \frac{\left(x - x_{k+\frac{1}{2}}\right)^4}{24} f^{IV}(\xi) \right] d\xi =$$

$$= hf\left(x_{k+\frac{1}{2}}\right) + \frac{h^3}{24} f''\left(x_{k+\frac{1}{2}}\right) + \frac{h^5}{1920} f^{IV}(\eta_k).$$

Тогда

$$I = \int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)dx = h\sum_{k=0}^{N-1} f\left(x_{k+\frac{1}{2}}\right) + \frac{h^2}{24}\left[h\sum_{k=0}^{N-1} f''\left(x_{k+\frac{1}{2}}\right)\right] + \frac{h^4(b-a)}{1920} f^{IV}(\eta).$$

Заметим теперь, что стоящее в квадратных скобках в правой части последнего равенства выражение есть не что иное, как квадратурная формула средних прямоугольников для вычисления интеграла $\int_a^b f''(x)dx$, т.е.

$$h\sum_{k=0}^{N-1} f''\left(x_{k+\frac{1}{2}}\right) = \int_a^b f''(x)dx + O\left(h^2\right).$$

Следовательно,

$$I = \int_a^b f(x)dx = h\sum_{k=0}^{N-1} f\left(x_{k+\frac{1}{2}}\right) + \frac{h^2}{24}\left(\int_a^b f''(x)dx + \frac{h^2}{24} f^{IV}\left(\eta_1\right)\right) + \frac{h^4(b-a)}{1920} f^{IV}(\eta) = \tag{2.33}$$

$$= h\sum_{k=0}^{N-1} f\left(x_{k+\frac{1}{2}}\right) + \frac{h^2}{24}\int_a^b f''(x)dx + O\left(h^4\right)$$

или

$$I = Q^C(h,f) + Ch^2 + O\left(h^4\right),$$

где через $Q^C(h,f)$ обозначена составная формула средних прямоугольников (2.19), а

$$C = \int_a^b f''(x)dx\,.$$

Таким образом, из остатка квадратурной формулы выделена первая главная часть. С помощью аналогичных рассуждений указанный процесс, вообще говоря, можно продолжить, предполагая еще более высокую гладкость интегрируемой функции. Аналогичные разложения также могут быть получены и для других квадратурных формул.

Упражнение. Получить указанное представление для составной формулы левых прямоугольников.

Представление (2.33) можно достаточно просто использовать с целью аналитического улучшения составной формулы средних прямоугольников. Вычисляя константу C, можем записать квадратурную формулу

$$I = \int_a^b f(x)dx \approx Q_1(h,f) = h\sum_{k=0}^{N-1} f\left(x_{k+\frac{1}{2}}\right) + \frac{h^2}{24}\left(f'(b) - f'(a)\right) \tag{2.34}$$

погрешность которой является величиной порядка $O\left(h^4\right)$ (а алгебраическая степень точности равна трем).

2.3.2. Правило Рунге

Выше мы показали, что погрешность составных квадратурных формул при условии дополнительной гладкости интегрируемой функции может быть разложена в ряд по последовательным главным частям. Этот факт можно использовать для апостериорных оценок погрешности. Способ, о котором пойдет речь ниже, носит название ***правила Рунге*** и может быть применен не только для оценки погрешности приближенного вычисления интегралов, но и для оценки погрешности любых других алгоритмов, допускающих упомянутое разложение погрешности по последовательным главным частям. Главная его идея заключается в том, чтобы по нескольким приближенным значениям искомой величины, полученным при различных значениях параметров вычислительного процесса (например, шага сетки h) вычислить параметры главных частей разложения остатка (такие как кон-

станты одной или нескольких последовательных главных частей, показатели степени шага сетки в этих главных частях и т.п.).

Покажем, как это можно технически реализовать в простейшем случае на примере оценки погрешности квадратурной формулы.

Итак, пусть имеет место разложение

$$R(h,f) = Ch^m + O\left(h^{m+p}\right).$$

Тогда, если I – искомый интеграл, а $Q(h,f)$ – аппроксимирующая его квадратурная сумма, то

$$I = Q(h,f) + R(h,f) \approx Q(h,f) + Ch^m .$$

Выполнив вычисление интеграла с помощью рассматриваемой квадратурной суммы с двумя различными значениями параметра h, можем записать приближенную систему, состоящую из двух уравнений, неизвестными которой являются константа C и точное значение искомого интеграла I:

$$\begin{cases} I \approx Q\left(h_1,f\right) + Ch_1^m, \\ I \approx Q\left(h_2,f\right) + Ch_2^m. \end{cases}$$

Отсюда, исключая I, находим:

$$C \approx \frac{Q(h_2,f) - Q\left(h_1,f\right)}{h_1^m - h_2^m}.$$

Следовательно,

$$R\left(h_1,f\right) \approx h_1^m \frac{Q(h_2,f) - Q\left(h_1,f\right)}{h_1^m - h_2^m}. \qquad (2.35)$$

Таким образом, мы получили выражение для вычисления главной части остатка квадратурной формулы, по которой можно проводить практическую оценку погрешности полученного приближенного значения интеграла $Q\left(h_1,f\right)$.

Вычисленное значение предполагает дальнейшую реакцию (программным путем) на его абсолютную величину в сравнении с задаваемой пользовательской величиной погрешности ε. В случае преобладания первой очевидным рецептом является дальнейшее измельчение сетки узлов (т.е. уменьшение величины шага h). При этом наиболее удобным способом организации работы служит выбор $h_2 = \frac{h_1}{2}$. Тогда в случае необходимости уменьшения шага полагают $h_1 = h_2$ и повторяют процесс вычисления интеграла. В этом случае значение $Q\left(h_1,f\right)$ оказывается полученным на предыдущем шаге процесса и, таким образом, оказывается возможной существенная экономия в объеме вычислений. Интеграл считается вычисленным с заданной пользователем точностью в том случае, когда вычисленная по формуле (2.35) величина остатка (ее модуль) оказывается меньше пользовательского требования ε.

Заметим, что вычисленная величина главной части остатка позволяет также и уточнять само приближенное значение интеграла. Это можно сделать, например, так:

$$I \approx Q(h_1, f) + h_1^m \frac{Q(h_2, f) - Q(h_1, f)}{h_1^m - h_2^m}. \qquad (2.36)$$

Также следует отметить, что правило Рунге может использоваться и в том случае, когда величина m априори неизвестна (например, в случае, когда интегрируемая функция не обладает достаточной для выписывания очередного члена разложения гладкостью). Кроме того, возможна и такая организация работы, которая позволяет вычислить сразу несколько последовательных главных частей. При этом, конечно же, необходимо выполнять расчеты не на двух, а на большем числе вложенных сеток.

§ 3. Квадратурные формулы наивысшей алгебраической степени точности (формулы Гаусса)

Выше мы получили следующий результат: в случае произвольного расположения узлов x_k квадратурную формулу

$$\int_a^b p(x) f(x) dx \approx \sum_{k=0}^{n} A_k f(x_k) \qquad (3.1)$$

за счет выбора коэффициентов A_k можно сделать точной для всех алгебраических многочленов до степени n включительно. Заметим, что использование свойства симметрии в расположении узлов (формулы средних прямоугольников и Симпсона) приводит к увеличению алгебраической степени точности на единицу.

Поставим задачу: выяснить, чего можно достичь в смысле повышения алгебраической степени точности за счет специального расположения узлов. Так как число узлов равно $(n+1)$, то можно надеяться за счет их выбора увеличить алгебраическую степень точности на $(n+1)$, т.е. увеличить ее до $n+(n+1) = 2n+1$.

Установим сейчас условия, при которых квадратурная формула будет иметь алгебраическую степень точности, равную $2n+1$. При этом вместо узлов x_k нам будет удобнее рассматривать многочлен $\omega_{n+1}(x) = (x - x_0) \cdots (x - x_n)$.

Справедлива следующая теорема (***критерий квадратурных формул наивысшей алгебраической степени точности***)

***Теорема* 1**. Для того чтобы рассматриваемая квадратурная формула (3.1) с $(n+1)$ узлами была точной для любых алгебраических многочленов до степени $(2n+1)$ включительно, необходимо и достаточно выполнение условий:

1) квадратурная формула должна быть интерполяционной, т.е.

$$A_k = \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)} dx, \quad k = \overline{0,n}; \qquad (3.2)$$

2) многочлен $\omega_{n+1}(x)$ должен быть ортогонален по данному весу $p(x)$ на отрезке $[a;b]$ ко всем многочленам степени не выше n, т.е.

$$\int_a^b p(x) \omega_{n+1}(x) Q_m(x) dx = 0, \ m \le n. \qquad (3.3)$$

Доказательство.
Необходимость. Так как квадратурная формула (3.1) точна для любых многочленов до степени $(2n+1)$ включительно, то она точна и для многочленов до степени n включительно, а тогда по теореме из **§ 2** она является интерполяционной, т.е. ее коэффициенты вычисляются по формулам (3.2).

Далее, рассмотрим произвольный многочлен $Q_m(x)$ степени $m \le n$. Тогда степень многочлена $\omega_{n+1}(x)Q_m(x)$ будет не выше $(2n+1)$. Следовательно, формула (3.1) точна для такого произведения, т.е. справедливо равенство

$$\int_a^b p(x)\omega_{n+1}(x)Q_m(x)dx = \sum_{k=0}^{n} A_k \omega_{n+1}(x_k)Q_m(x_k) = 0$$

(последнее равенство цепочки имеет место в силу того, что при любых $k = \overline{0, n}$ узлы x_k квадратурной суммы являются корнями многочлена $\omega_{n+1}(x)$, т.е. $\omega_{n+1}(x_k) = 0$). Таким образом, выполняется соотношение (3.3).
Достаточность. Рассмотрим произвольный алгебраический многочлен $f(x)$ степени не выше $(2n+1)$. Покажем, что для него квадратурная формула (3.1) точна. Выполнив деление $f(x)$ на $\omega_{n+1}(x)$ с остатком, получим:

$$f(x) = Q_m(x)\omega_{n+1}(x) + r(x),$$

где степень частного $Q_m(x)$ не превосходит n и степень остатка также будет не выше n. Из записанного равенства в силу того что точки x_k – корни многочлена $\omega_{n+1}(x)$, следует, что $f(x_k) = r(x_k),\ \ k = \overline{0, n}$. Тогда

$$\int_a^b p(x)f(x)dx = \int_a^b p(x)\omega_{n+1}(x)Q_m(x)dx + \int_a^b p(x)r(x)dx \overset{(3.3)}{=} 0 + \int_a^b p(x)r(x)dx \overset{(3.2)}{=}$$

$$\overset{(3.2)}{=} \sum_{k=0}^{n} A_k r(x_k) = \sum_{k=0}^{n} A_k f(x_k).$$

Таким образом, требуемое соотношение установлено. ⊠

Выясним теперь, когда требования, сформулированные в ***теореме* 1**, выполнимы. Вопрос фактически сводится к нахождению такого многочлена $\omega_{n+1}(x)$, который удовлетворяет соотношениям (3.3), причем его корни действительны, различны и все лежат на отрезке $[a;b]$. Справедлива

***Теорема* 2**. Если весовая функция $p(x)$ сохраняет знак на отрезке $[a;b]$, то приведенный многочлен $\omega_{n+1}(x)$ степени $(n+1)$, ортогональный на данном отрезке по весу $p(x)$ ко всем многочленам меньшей степени, существует и единственен для любого фиксированного n. При этом все его корни действительны, различны и лежат внутри данного отрезка.

Доказательство.
Запишем многочлен $\omega_{n+1}(x)$ с неопределенными коэффициентами:

$$\omega_{n+1}(x) = x^{n+1} + a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n .$$

Нахождение многочлена $\omega_{n+1}(x)$ эквивалентно нахождению его коэффициентов. Построим систему для нахождения величин a_i, $(i = \overline{0,n})$. Для этих целей воспользуемся условием ортогональности многочлена $\omega_{n+1}(x)$ многочленам 1, x, x^2,..., x^n. В итоге получим:

$$\int_a^b p(x)\left(x^{n+1} + a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n\right) x^k dx = 0, \quad k = \overline{0,n} . \tag{3.4}$$

Для однозначной разрешимости системы (3.4) достаточно показать, что соответствующая ей однородная система

$$\int_a^b p(x)\left(a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n\right) x^k dx = 0, \quad k = \overline{0,n} . \tag{3.5}$$

имеет лишь тривиальное решение $a_i = 0$, $(i = \overline{0,n})$.

Умножим k-е уравнение системы (3.5) на a_{n-k} и просуммируем получившиеся равенства по всем значениям $k = \overline{0,n}$. Тогда будем иметь:

$$\int_a^b p(x)\left(a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n\right)^2 dx = 0 .$$

Таким образом, в силу того, что весовая функция $p(x)$ сохраняет знак на отрезке $[a;b]$ и $p(x)$ отлична от тождественного нуля, следует, что $a_i = 0$, $(i = \overline{0,n})$.

Таким образом, многочлен $\omega_{n+1}(x)$ всегда может быть построен, причем единственным образом.

Рассмотрим теперь корни $\xi_1, \xi_2, ..., \xi_m$ многочлена $\omega_{n+1}(x)$ нечетной кратности, лежащие внутри отрезка $[a;b]$. Существование хотя бы одного такого корня следует из ортогональности многочлена $\omega_{n+1}(x)$ к $P_0(x) \equiv 1$. Действительно,

$$\int_a^b p(x)\omega_{n+1}(x)dx = 0 ,$$

а так как весовая функция $p(x)$ сохраняет знак, то $\omega_{n+1}(x)$ обязан на $[a;b]$ знак поменять, ибо в противном случае значение рассмотренного выше интеграла будет отличным от нуля (как интеграла от знакопостоянной функции).

В точках ξ_i многочлен $\omega_{n+1}(x)$, очевидно, меняет знак. Кроме того, $1 \le m \le n+1$. Пусть $m < n+1$. Тогда по корням $\xi_1, \xi_2, ..., \xi_m$ построим многочлен степени $m < n+1$ $Q_m(x) = (x - \xi_1) \cdots (x - \xi_m)$, к которому $\omega_{n+1}(x)$ должен быть ортогонален, т.е.

$$\int_a^b p(x)\omega_{n+1}(x)Q_m(x)dx = 0 .$$

Но это равенство невозможно в силу того, что $\omega_{n+1}(x)$ и $Q_m(x)$ имеют одни и те же точки перемены знаков и, таким образом, подынтегральное выражение сохраняет знак на отрезке $[a;b]$. Следовательно, $m = n+1$ и последнее утверждение теоремы доказано. ⊠

Следствие. Если весовая функция $p(x)$ знакопостоянна на отрезке интегрирования $[a;b]$, то квадратурная формула вида (3.1) с $(n+1)$ узлами, точная для любого многочлена до степени $2n+1$ включительно, существует для любого фиксированного значения n.

Возникает вопрос: будет ли степень $2n+1$ наивысшей? Ответ на него дает

***Теорема* 3.** Если $p(x)$ знакопостоянна на отрезке интегрирования, то ни при каком выборе узлов и коэффициентов квадратурной формулы с $(n+1)$ узлами рассматриваемого вида (3.1) не может быть точной для любого алгебраического многочлена степени $2n+2$.

Доказательство. Рассмотрим многочлен $\omega_{n+1}^2(x)$. Его степень равна $2n+2$. Очевидно, что $\int_a^b p(x)\omega_{n+1}^2(x)dx \neq 0$ ни при каком выборе x_k, так как подынтегральное выражение сохраняет знак на отрезке интегрирования. В то же время $\sum_{k=0}^{n} A_k\omega_{n+1}^2(x_k) = 0$, т.е.

$$\int_a^b p(x)\omega_{n+1}^2(x)dx \neq \sum_{k=0}^{n} A_k\omega_{n+1}^2(x_k).$$

⊠

Таким образом, если весовая функция $p(x)$ знакопостоянна на отрезке интегрирования $[a;b]$, то наивысшая алгебраическая степень точности квадратурной формулы с $(n+1)$ узлами равна $2n+1$. Такие квадратурные формулы называют ***квадратурными формулами типа Гаусса*** (или ***Гаусса-Кристоффеля***).

Отметим также, что справедлива

***Теорема* 4.** Если квадратурная формула вида (3.1) точна для всевозможных многочленов степени $2n$, то при знакопостоянной весовой функции весовой функции $p(x)$ все ее коэффициенты A_k имеют один и тот же знак (совпадающий со знаком $p(x)$).

Доказательство. Рассмотрим многочлен $\Phi_i^2(x)$ степени $2n$ Для него формула точна, поэтому

$$\int_a^b p(x)\Phi_i^2(x)dx = \sum_{k=0}^{n} A_k\Phi_i^2(x_k) = A_i, \quad i = \overline{0,n}.$$

Таким образом, знаки всех весовых коэффициентов совпадают со знаком весовой функции. ⊠

Следовательно, квадратурные формулы типа Гаусса имеют все коэффициенты одного знака, что равносильно их вычислительной устойчивости. Получим сейчас более удобные формулы для их вычисления.

***Лемма* (*тождество Кристоффеля-Дарбу*).** Для системы ортонормированных многочленов $Q_i(x),\ i = 0,1,...,n,...$ справедливо тождество

$$\sum_{i=0}^{n} Q_i(t)Q_i(x) = a_{n,n+1}\frac{Q_{n+1}(t)\cdot Q_n(x) - Q_{n+1}(x)\cdot Q_n(t)}{t-x}. \qquad (3.5)$$

Доказательство. Ранее (см. Гл. III, **§ 1**) мы получили трехчленное рекуррентное соотношение, связывающее ортонормированные многочлены:

$$a_{i,i+1}Q_{i+1}(x) + a_{i,i}Q_i(x) + a_{i-1,i}Q_{i-1}(x) = xQ_i(x), \qquad (*)$$

где

$$a_{i,k} = \int_a^b p(x)Q_i(x)Q_k(x)dx.$$

Умножим соотношение (*) на $Q_i(t)$. Получим:

$$xQ_i(x)\cdot Q_i(t) = a_{i,i+1}Q_{i+1}(x)\cdot Q_i(t) + a_{i,i}Q_i(x)\cdot Q_i(t) + a_{i,i-1}Q_{i-1}(x)\cdot Q_i(t). \qquad (**)$$

Поменяем в последнем равенстве ролями переменные x и t:

$$tQ_i(t)\cdot Q_i(x) = a_{i,i+1}Q_{i+1}(t)\cdot Q_i(x) + a_{i,i}Q_i(t)\cdot Q_i(x) + a_{i,i-1}Q_{i-1}(t)\cdot Q_i(x).$$

Теперь, вычитая из последнего равенства равенство (**), получим:

$$(t-x)Q_i(t)\cdot Q_i(x) = a_{i,i+1}[Q_{i+1}(t)\cdot Q_i(x) - Q_i(t)\cdot Q_{i+1}(x)] - a_{i-1,i}[Q_i(t)\cdot Q_{i-1}(x) - Q_i(x)\cdot Q_{i-1}(t)].$$

Просуммировав полученное равенство по i от 0 до n, будем иметь:

$$(t-x)\sum_{i=0}^{n} Q_i(x)Q_i(t) = a_{n,n+1}[Q_{n+1}(t)\cdot Q_n(x) - Q_{n+1}(x)\cdot Q_n(t)]$$

или

$$\sum_{i=0}^{n} Q_i(t)Q_i(x) = a_{n,n+1}\frac{Q_{n+1}(t)\cdot Q_n(x) - Q_{n+1}(x)\cdot Q_n(t)}{t-x}.$$

⊠

Заметим, что если записать $Q_n(x)$ в виде

$$Q_n(x) = c_n x^n + \cdots,$$

то равенство (*) примет вид

$$a_{i,i+1}\left(c_{i+1}x^{i+1} + \cdots\right) + a_{i,i}\left(c_i x^i + \cdots\right) + a_{i-1,i}\left(c_i x^{i-1} + \cdots\right) = x\left(c_i x^i + \cdots\right),$$

Откуда, приравнивая коэффициенты при x^{i+1}, получим:

$$a_{i,i+1}c_{i+1} = c_i,$$

т.е.

$$a_{i,i+1} = \frac{c_i}{c_{i+1}}, \quad i = 0, 1, \ldots.$$

С учетом полученного соотношения тождество (3.5) можно переписать в виде

$$\sum_{i=0}^{n} Q_i(t)Q_i(x) = \frac{c_n}{c_{n+1}} \frac{Q_{n+1}(t)\cdot Q_n(x) - Q_{n+1}(x)\cdot Q_n(t)}{t-x}. \qquad (3.6)$$

Теперь преобразуем формулу (3.2) для вычисления коэффициентов квадратурных формул наивысшей алгебраической степени точности, домножив ее числитель и знаменатель на c_{n+1} и учитывая, что $c_{n+1}\omega_{n+1}(x) = Q_{n+1}(x)$:

$$A_k = \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x-x_k)\omega'_{n+1}(x_k)} dx = \int_a^b p(x) \frac{Q_{n+1}(x)}{(x-x_k)Q'_{n+1}(x_k)} dx. \qquad (3.7)$$

Положим в (3.6) $t = x_k$ и учтем, что x_k – корень многочлена $Q_{n+1}(x)$. Тогда $Q_{n+1}(x_k) = 0$ и (3.6) перепишется в виде

$$\sum_{i=0}^{n} Q_i(x)Q_i(x_k) = \frac{c_n}{c_{n+1}} \cdot Q_n(x_k) \cdot \frac{Q_{n+1}(x)}{x-x_k}.$$

Домножим последнее равенство на $p(x)$ и проинтегрируем по отрезку $[a;b]$:

$$\sum_{i=0}^{n} Q_i(x_k) \int_a^b p(x)Q_i(x)dx = \frac{c_n}{c_{n+1}} \cdot Q_n(x_k) \cdot \int_a^b p(x) \frac{Q_{n+1}(x)}{x-x_k} dx. \qquad (***)$$

Так как система многочленов $Q_i(x)$ является ортонормированной, то

$$\int_a^b p(x)Q_i(x)dx = \begin{cases} 0, & \text{если } i \geq 1, \\ 1, & \text{если } i = 0. \end{cases}$$

Следовательно, равенство (***) с учетом (3.7) примет вид

$$1 = \frac{c_n}{c_{n+1}} Q_n(x_k) \int_a^b p(x) \frac{Q_{n+1}(x)}{x-x_k} dx = \frac{c_n}{c_{n+1}} \cdot Q_n(x_k) \cdot Q'_{n+1}(x_k) \cdot A_k.$$

Отсюда

$$A_k = \frac{c_{n+1}}{c_n} \cdot \frac{1}{Q_n(x_k) \cdot Q'_{n+1}(x_k)}, \quad k = \overline{0,n}. \qquad (3.8)$$

Формула (3.8) несколько более удобна для вычисления коэффициентов квадратурных формул типа Гаусса по сравнению с (3.2), поскольку не требует вычисления интегралов (однако требует знания систем ортонормированных многочленов).

Заметим, что в формуле (3.8) можно осуществить обратный переход от системы ортонормированных многочленов $Q_n(x)$ к приведенным ортогональным многочленам $\omega_n(x)$. Действительно, учитывая, что $c_k\omega_k(x) = Q_k(x)$, умножим числитель и знаменатель (3.8) на c_{n+1}. Тогда

$$A_k = \frac{1}{c_n^2 \cdot \omega_n(x_k) \cdot \omega'_{n+1}(x_k)}, \quad k = \overline{0,n},$$

а поскольку

$$1 = \|Q_n(x)\|^2 = c_n^2 \|\omega_n(x)\|^2,$$

то

$$A_k = \frac{\|\omega_n(x)\|^2}{\omega_n(x_k) \cdot \omega'_{n+1}(x_k)}, \quad k = \overline{0,n}. \tag{3.8′}$$

Наконец, установим формулу для вычисления остатка квадратурных формул типа Гаусса. Построим для функции $f(x)$ интерполяционный многочлен Эрмита степени не выше $2n+1$ с двукратными узлами $x_0, x_1, ..., x_n$:

$$f(x) = P_{2n+1}(x) + r_{2n+1}(x),$$

где

$$r_{2n+1}(x) = \omega_{n+1}^2(x) \cdot \frac{f^{(2n+2)}(\xi)}{(2n+2)!},$$

если $f(x) \in C^{2n+2}[a;b]$.

Тогда

$$\int_a^b p(x) f(x) dx = \int_a^b p(x) P_{2n+1}(x) dx + \int_a^b p(x) r_{2n+1}(x) dx.$$

С другой стороны, так как алгебраическая степень точности формулы равна $2n+1$, то

$$\int_a^b p(x) P_{2n+1}(x) dx = \sum_{k=0}^{n} A_k P_{2n+1}(x_k) = \sum_{k=0}^{n} A_k f(x_k).$$

Следовательно,

$$R_n(f) = \int_a^b p(x) r_{2n+1}(x) dx = \int_a^b p(x) \omega_{n+1}^2(x) \frac{f^{(2n+2)}(\xi)}{(2n+2)!} dx \overset{\text{т. о среднем}}{=}$$

$$\overset{\text{т. о среднем}}{=} \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \int_a^b p(x) \omega_{n+1}^2(x) dx = \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \cdot \|\omega_{n+1}(x)\|^2. \tag{3.9}$$

3.1. Некоторые частные случаи квадратурных формул типа Гаусса

Рассмотрим сейчас некоторые наиболее часто встречающиеся в приложениях случаи квадратурных формул типа Гаусса.

1^0. $p(x) \equiv 1$.

Отрезок интегрирования считаем конечным, а $f(x)$ – достаточно гладкой функцией (именно этот случай был подробно рассмотрен Гауссом).

Всякий конечный отрезок $[a;b]$ линейной заменой переменной может быть преобразован в отрезок $[-1;1]$, и мы будем считать, что интеграл приведен к виду

$$I = \int_{-1}^{1} f(x)dx\,. \qquad (*)$$

Построим в явном виде систему ортогональных многочленов. Пусть $\omega_n(x)$ – n-й приведенный многочлен данной системы, а $q(x)$ – произвольный многочлен степени не выше $n-1$. Введем обозначения

$$\varphi_1(x) = \int_{-1}^{x} \omega_n(x)dx\,,\ \ \varphi_{i+1}(x) = \int_{-1}^{x} \varphi_i(x)dx\,,\ \ i = 1,...,n-1\,. \qquad (**)$$

Тогда, последовательно интегрируя по частям, будем иметь:

$$0 = \int_{-1}^{1} \omega_n(x)q(x)dx = \varphi_1(x)q(x)\Big|_{-1}^{1} - \int_{-1}^{1} \varphi_1(x)q'(x)dx =$$

$$= \left[\varphi_1(x)q(x) - \varphi_2(x)q'(x)\right]\Big|_{-1}^{1} + \int_{-1}^{1} \varphi_2(x)q''(x)dx = \ldots =$$

$$= \left[\varphi_1(x)q(x) - \varphi_2(x)q'(x) + \cdots + (-1)^{n-1}\varphi_n(x)q^{(n-1)}(x)\right]\Big|_{-1}^{1} + (-1)^n \int_{-1}^{1} \varphi_n(x)q^{(n)}(x)dx\,.$$

Интегральное слагаемое в правой части полученного равенства, очевидно, равно нулю, поскольку согласно предположению $q(x)$ – многочлен степени не выше $n-1$. Точно так же равны нулю и все оставшиеся слагаемые на нижнем пределе двойной подстановки, так как в силу равенств (**) $\varphi_i(-1) = 0$, $i = \overline{1,\,n}$.

Отсюда в силу произвольности многочлена $q(x)$ следует, что

$$\varphi_i(1) = 0\,,\ i = \overline{1,\,n}\,.$$

Таким образом, многочлен степени $2n$ $\varphi_n(x)$ обладает корнями кратности n при $x = \pm 1$ и, значит,

$$\varphi_n(x) = C(x+1)^n(x-1)^n = C\left(x^2-1\right)^n\,,$$

где C – некоторая постоянная (старший коэффициент).

Тогда

$$\omega_n(x) = C\frac{d^n}{dx^n}\left(x^2-1\right)^n\,.$$

Постоянную C подбираем таким образом, чтобы многочлена $\omega_n(x)$ был равен единице. Так как

$$\omega_n(x) = C\frac{d^n}{dx^n}\left(x^2-1\right)^n = C\frac{d^n}{dx^n}\left(x^{2n}-\cdots\right) =$$

$$= C\cdot(2n)\cdot(2n-1)\cdot\ldots\cdot(n+1)\left(x^n-\cdots\right) = C\cdot\frac{(2n)!}{n!}\cdot\left(x^n-\cdots\right),$$

то

$$C = \frac{n!}{(2n)!},$$

т.е.

$$\omega_n(x) = \frac{n!}{(2n)!}\frac{d^n}{dx^n}\left(x^2-1\right)^n, \quad n = 0,1,\ldots. \tag{3.10}$$

Вычислим сейчас норму многочлена $\omega_n(x)$, применяя, как и выше, последовательное интегрирование по частям:

$$\int_{-1}^{1}\omega_n^2(x)dx = C^2\int_{-1}^{1}\frac{d^n}{dx^n}\left(x^2-1\right)^n\cdot\frac{d^n}{dx^n}\left(x^2-1\right)^n dx =$$

$$= C^2\left[\frac{d^n}{dx^n}\left(x^2-1\right)^n\cdot\frac{d^{n-1}}{dx^{n-1}}\left(x^2-1\right)^n\right]\Bigg|_{-1}^{1} - \int_{-1}^{1}\frac{d^{n-1}}{dx^{n-1}}\left(x^2-1\right)^n\cdot\frac{d^{n+1}}{dx^{n+1}}\left(x^2-1\right)^n dx =$$

$$= C^2\left[\frac{d^n}{dx^n}\left(x^2-1\right)^n\cdot\frac{d^{n-1}}{dx^{n-1}}\left(x^2-1\right)^n - \frac{d^{n+1}}{dx^{n+1}}\left(x^2-1\right)^n\cdot\frac{d^{n-2}}{dx^{n-2}}\left(x^2-1\right)^n\right]\Bigg|_{-1}^{1} +$$

$$+\int_{-1}^{1}\frac{d^{n-2}}{dx^{n-2}}\left(x^2-1\right)^n\cdot\frac{d^{n+2}}{dx^{n+2}}\left(x^2-1\right)^n dx = \cdots =$$

$$= C^2\left[\frac{d^n}{dx^n}\left(x^2-1\right)^n\cdot\frac{d^{n-1}}{dx^{n-1}}\left(x^2-1\right)^n - \frac{d^{n+1}}{dx^{n+1}}\left(x^2-1\right)^n\cdot\frac{d^{n-2}}{dx^{n-2}}\left(x^2-1\right)^n + \cdots + (-1)^{n-1}\frac{d^{2n-1}}{dx^{2n-1}}\left(x^2-1\right)^n\cdot\left(x^2-1\right)^n\right]\Bigg|_{-1}^{1} +$$

$$+(-1)^n\int_{-1}^{1}\frac{d^{2n}}{dx^{2n}}\left(x^2-1\right)^n\cdot\left(x^2-1\right)^n dx = (-1)^n C^2\cdot(2n)!\cdot\int_{-1}^{1}(x-1)^n(x+1)^n dx =$$

$$= (-1)^n C^2\cdot(2n)!\cdot\left[\frac{(x-1)^n(x+1)^{n+1}}{n+1}\Bigg|_{-1}^{1} - \frac{n}{n+1}\int_{-1}^{1}(x-1)^{n-1}(x+1)^{n+1}dx\right] =$$

$$= (-1)^n C^2\cdot(2n)!\cdot\left[\frac{(x-1)^n(x+1)^{n+1}}{n+1} - \frac{n}{n+1}\frac{(x-1)^{n-1}(x+1)^{n+2}}{n+2}\Bigg|_{-1}^{1} + \frac{n(n-1)}{(n+1)(n+2)}\int_{-1}^{1}(x-1)^{n-2}(x+1)^{n+2}dx\right] =$$

$$= (-1)^n C^2\cdot(2n)!\cdot\left[\frac{(x-1)^n(x+1)^{n+1}}{n+1} - \frac{n}{n+1}\frac{(x-1)^{n-1}(x+1)^{n+2}}{n+2} + \cdots + (-1)^{n-1}\frac{n!\cdot(x-1)(x+1)^{2n}}{(n+1)\cdot\ldots\cdot(2n)}\Bigg|_{-1}^{1}\right] +$$

$$+(-1)^n C^2\cdot(2n)!\cdot\frac{n!}{(n+1)\cdot\ldots\cdot(2n)}\int_{-1}^{1}(x+1)^{2n}dx = (-1)^n C^2\cdot(2n)!\cdot\frac{n!\cdot n!}{(2n)!}\cdot\frac{(x+1)^{2n+1}}{2n+1}\Bigg|_{-1}^{1} =$$

$$= \frac{(n!)^2}{((2n)!)^2}\cdot((n)!)^2\cdot\frac{2^{2n+1}}{2n+1}.$$

Таким образом,

$$\|\omega_n(x)\| = \frac{(n!)^2 \cdot 2^n}{(2n)!} \cdot \sqrt{\frac{2}{2n+1}}, \tag{3.11}$$

и, следовательно, многочлены $Q_n(x)$ ортонормированной системы будут иметь вид

$$Q_n(x) = \frac{\omega_n(x)}{\|\omega_n(x)\|} = \frac{n!}{(2n)!} \cdot \sqrt{\frac{2n+1}{2}} \cdot \frac{(2n)!}{(n!)^2 \cdot 2^n} \omega_n(x) = \sqrt{\frac{2n+1}{2}} \cdot \frac{1}{n! \cdot 2^n} \frac{d^n}{dx^n}\left(x^2-1\right)^n. \tag{3.12}$$

Окончательно можем сформулировать следующий результат: квадратурная формула наивысшей алгебраической степени точности для вычисления интеграла (*) имеет вид

$$I = \int_{-1}^{1} f(x)dx \approx \sum_{k=0}^{n} A_k f(x_k), \tag{3.13}$$

где x_k – корни многочлена степени $(n+1)$, определяемого формулой (3.10), а коэффициенты A_k могут быть вычислены по формулам (3.8) или (3.8′), в которых многочлены определяются по формулам (3.12) или (3.10) соответственно.

При этом, учитывая формулы (3.9) и (3.11), можем записать представление ее остатка:

$$R_n(f) = \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \cdot \|\omega_{n+1}(x)\|^2 = \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \cdot \frac{((n+1)!)^4}{((2n+2)!)^2} \cdot \frac{2^{2n+3}}{2n+3} =$$

$$= \frac{f^{(2n+2)}(\eta) \cdot 2^{2n+3}}{(2n+2)! \cdot (2n+3)} \cdot \left(\frac{((n+1)!)^2}{(2n+2)!}\right)^2. \tag{3.14}$$

***Замечание* 1.** В теории ортогональных многочленов чаще вместо системы многочленов $\omega_n(x)$ (или $Q_n(x)$), определенных выше, используют многочлены, отличающиеся от них постоянным множителем и имеющие вид

$$P_n(x) = \frac{1}{2^n \cdot n!} \frac{d^n}{dx^n}\left(x^2-1\right)^n, \tag{3.15}$$

норма которых равна $\|P_n(x)\| = \sqrt{\frac{2n+1}{2}}$. Эти многочлены называются ***многочленами Лежандра.*** В терминах многочленов Лежандра формулу для вычисления квадратурных коэффициентов можно записать в следующем виде

$$A_k = \frac{2}{(n+1)P_n(x_k)P'_{n+1}(x_k)}, \quad k = \overline{0,n}.$$

Если же при этом воспользоваться известным в теории многочленов Лежандра соотношением

$$\left(1-x^2\right)P'_n(x) = n[P_{n-1}(x) - xP_n(x)],$$

то формулы для вычисления коэффициентов можно еще более упростить и привести к виду

$$A_k = \frac{2}{\left(1-x_k^2\right)\left[P'_{n+1}(x_k)\right]^2},\ k=\overline{0,n}. \tag{3.16}$$

***Замечание* 2.** На «пользовательском» уровне узлы и коэффициенты квадратурной формулы Гаусса удобнее брать из соответствующих таблиц.

2^0. $p(x)=(b-x)^\alpha(x-a)^\beta$.

Как видно из записи интеграла с указанной весовой функцией, соответствующая квадратурная формула предназначена для приближенного интегрирования функций, на концах отрезка интегрирования имеющих степенные особенности.

Вновь, как и в предыдущем случае, можно ограничиться рассмотрением интеграла

$$I=\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta f(x)dx.$$

Здесь α и β – вещественные параметры, причем $\alpha,\beta>-1$ (последнее требование необходимо для сходимости интеграла).

Практически дословно повторяя рассуждения предыдущего пункта, можно было бы построить систему многочленов, ортогональных по весу $p(x)=(1-x)^\alpha(1+x)^\beta$ на отрезке $[-1;1]$. Однако мы этого больше делать не будем и воспользуемся готовыми результатами. Искомой системой многочленов является система ***многочленов Якоби***

$$P_n^{(\alpha,\beta)}(x)=\frac{(-1)^n}{2^n\cdot n!}\cdot(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left[(1-x)^{\alpha+n}(1+x)^{\beta+n}\right]. \tag{3.17}$$

Таким образом, квадратурная формула наивысшей алгебраической степени точности в этом случае будет иметь вид

$$I=\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta f(x)dx\approx\sum_{k=0}^{n}A_k f(x_k), \tag{3.18}$$

где узлы x_k будут корнями многочленов $P_{n+1}^{(\alpha,\beta)}(x)$, для коэффициентов A_k можно получить формулы

$$A_k=2^{\alpha+\beta+1}\cdot\frac{\Gamma(\alpha+n+2)\cdot\Gamma(\beta+n+2)}{(n+1)!\cdot\Gamma(\alpha+\beta+n+2)\cdot\left(1-x_k^2\right)\cdot\left[\frac{d}{dx}P_{n+1}^{(\alpha,\beta)}(x_k)\right]^2},\ k=\overline{0,n}, \tag{3.19}$$

а для остатка справедливо представление

$$R_n(f)=\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\cdot 2^{\alpha+\beta+2n+1}\cdot\frac{(n+1)!\Gamma(\alpha+n+2)\cdot\Gamma(\beta+n+2)\cdot\Gamma(\alpha+\beta+n+2)}{(\alpha+\beta+2n+3)\cdot\Gamma(\alpha+\beta+2n+3)}. \tag{3.20}$$

В формулах (3.19), (3.20) $\Gamma(x)$ обозначает Γ-функцию Эйлера.

Частными случаями многочленов Якоби являются рассмотренные выше многочлены Лежандра (соответствуют случаю $\alpha = \beta = 0$), а также многочлены – наименее отклоняющиеся от нуля – ***многочлены Чебышева первого рода*** $T_n(x)$. Эти последние соответствуют случаю $\alpha = \beta = -\frac{1}{2}$. В этом случае квадратурная формула наивысшей алгебраической степени точности выглядит особо просто. Поэтому рассмотрим ее несколько подробнее.

Ее узлы – корни многочлена Чебышева $T_{n+1}(x)$ – имеют вид (см. **§ 3, Гл. IV**)

$$x_k = \cos\frac{2k+1}{2n+2}\pi, \quad k = \overline{0,n}. \tag{3.21}$$

Помимо этого, квадратурная формула наивысшей алгебраической степени точности, соответствующая весовой функции $p(x) = \frac{1}{\sqrt{1-x^2}}$ обладает еще одним замечательным свойством. Пользуясь формулой (3.19), вычислим ее коэффициенты:

$$A_k = \frac{\Gamma^2\left(n+\frac{3}{2}\right)}{(n+1)!\cdot\Gamma(n+1)\cdot\left(1-x_k^2\right)\cdot\left[\frac{d}{dx}P_{n+1}^{\left(-\frac{1}{2},-\frac{1}{2}\right)}(x_k)\right]^2}.$$

Поскольку $P_{n+1}^{\left(-\frac{1}{2},-\frac{1}{2}\right)}(x) = C_{n+1}T_{n+1}(x)$ и

$$T'_{n+1}(x_k) = \left[\cos((n+1)\arccos x)\right]'\Big|_{x=x_k} = \sin[(n+1)\arccos x_k]\cdot\frac{n+1}{\sqrt{1-x_k^2}} = \frac{(-1)^k(n+1)}{\sqrt{1-x_k^2}},$$

то

$$\left(1-x_k^2\right)\left[\frac{d}{dx}P_{n+1}^{\left(-\frac{1}{2},-\frac{1}{2}\right)}(x_k)\right]^2 = C_{n+1}^2(n+1)^2$$

и, следовательно,

$$A_k = \frac{\Gamma^2\left(n+\frac{3}{2}\right)}{(n+1)!\cdot\Gamma(n+1)\cdot C_{n+1}^2\cdot(n+1)^2}, \quad k = \overline{0,n}.$$

Правая часть полученного равенства не зависит от номера k и поэтому все коэффициенты A_k будут одинаковы. Обозначим общую величину их буквой A. Численное значение A, конечно же, может быть найдено на основании последнего равенства, но проще это сделать, если воспользоваться тем, что квадратурная формула должна дать точный результат в случае $f(x) \equiv 1$ (т.е. иметь алгебраическую степень точности не менее нуля) и, стало быть, должно выполняться равенство

$$\sum_{k=0}^{n} A_k = (n+1)A = \int_{-1}^{1}\frac{dx}{\sqrt{1-x^2}} = \pi.$$

Отсюда имеем: $A = \frac{\pi}{n+1}$.

Таким образом, окончательно получаем: квадратурная формула имеет вид

$$I=\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}dx\approx\frac{\pi}{n+1}\sum_{k=0}^{n}f\left(\cos\frac{2k+1}{2n+2}\pi\right),\tag{3.22}$$

а ее остаток после несложных преобразований формулы (3.20) с использованием свойств функции Эйлера примет вид

$$R_n(f)=\frac{\pi}{2^{2n+1}\cdot(2n+2)!}f^{(2n+2)}(\eta).\tag{3.23}$$

3⁰. Интегралы вида $I=\int_0^{\infty}x^{\alpha}\,\mathrm{e}^{-x}\,f(x)dx$.

Ортогональными на полуоси $[0;\infty)$ по весу $p(x)=x^{\alpha}\,\mathrm{e}^{-x}$ $(\alpha>-1)$ являются ***многочлены Чебышева-Лягерра***

$$L_n^{(\alpha)}(x)=(-1)^n x^{-\alpha}\,\mathrm{e}^{x}\,\frac{d^n}{dx^n}\left(x^{\alpha+n}\,\mathrm{e}^{-x}\right).\tag{3.24}$$

Таким образом, узлы соответствующей квадратурной формулы наивысшей алгебраической степени точности будут корнями многочлена вида (3.24) $L_{n+1}^{(\alpha)}(x)$, коэффициенты будут вычисляться по формулам

$$A_k=\frac{\Gamma(n+2)\cdot\Gamma(n+\alpha+2)}{x_k\left[\dfrac{d}{dx}L_{n+1}^{(\alpha)}(x\ )\right]^2},\ \ k=\overline{0,n},\tag{3.25}$$

а остаток будет иметь вид

$$R_n(f)=\frac{\Gamma(n+2)\cdot\Gamma(n+\alpha+2)}{(2n+2)!}\cdot f^{(2n+2)}(\eta).\tag{3.26}$$

4⁰. Интегралы вида $I=\int_{-\infty}^{\infty}\mathrm{e}^{-x^2}\,f(x)dx$.

Систему многочленов, ортогональных на всей числовой оси $(-\infty;+\infty)$ по весу $p(x)=\mathrm{e}^{-x^2}$, образуют ***многочлены Чебышева-Эрмита***

$$H_n(x)=(-1)^n\,\mathrm{e}^{x^2}\,\frac{d^n}{dx^n}\mathrm{e}^{-x^2}.\tag{3.27}$$

Таким образом, узлы соответствующей квадратурной формулы наивысшей алгебраической степени точности являются корнями многочлена $H_{n+1}(x)$, коэффициенты могут быть вычислены по формуле

$$A_k=\frac{2^{n+1}\cdot(n+1)!\cdot\sqrt{\pi}}{\left[\dfrac{d}{dx}H_{n+1}^2(x_k)\right]^2},\ \ k=\overline{0,n},\tag{3.28}$$

а остаток имеет вид

$$R_n(f)=\frac{(n+1)!\cdot\sqrt{\pi}}{2^{n+1}\cdot(2n+2)!}\cdot f^{(2n+2)}(\eta)\,. \tag{3.29}$$

Еще раз напомним, что для всех рассмотренных выше случаев (а также и для многих других) составлены таблицы узлов и коэффициентов для различных значений n.

Упражнения.

1. Провести построение систем ортогональных многочленов для ***всех*** рассмотренных выше случаев.
2. Выписать по ***четыре первых*** многочлена каждой из рассмотренных выше ортогональных систем, а также найти узлы и коэффициенты соответствующих квадратурных формул типа Гаусса.

§ 4. Квадратурные формулы, содержащие наперед заданные узлы

В прикладных задачах иногда возникает необходимость построения таких квадратурных формул, часть узлов которых задается заранее, другая же часть может быть взята произвольно и выбором их можно распоряжаться для тех или иных целей.

Рассмотрим квадратурную формулу

$$\int_a^b p(x)f(x)dx\approx\sum_{l=1}^{m}B_l f(a_l)+\sum_{k=0}^{n-m}A_k f(x_k)\,, \tag{4.1}$$

в которой m узлов $a_1,...,a_m$ $(m\le n+1)$ фиксированы. Она содержит $2n-m+2$ параметров A_k, x_k $(k=0,...,n-m)$ и B_l $(l=1,...,m)$. Попытаемся их выбрать так, чтобы равенство (4.1) было точным для многочленов возможно более высокой степени.

Введем два многочлена, связанных с узлами a_l и x_k:

$$\Omega_m(x)=(x-a_1)\cdot\ldots\cdot(x-a_m),$$

$$\omega_{n-m+1}(x)=(x-x_0)\cdot\ldots\cdot(x-x_{n-m}).$$

За счет выбора коэффициентов A_k и B_l формулу (4.1) можно сделать точной для многочленов степени n. Для этого ее достаточно сделать интерполяционной. Достичь же того, чтобы (4.1) была точной для многочленов более высокой степени, можно только за счет специального подбора узлов x_k.

Справедлива теорема, аналогичная ***теореме* 1** из **§ 3** (причем смысл ее состоит в том, чтобы многочлен $\Omega_m(x)$ в «состав» весовой функции).

***Теорема* 1.** Для того чтобы квадратурная формула (4.1) была точной для многочленов степени $2n-m+1$, необходимо и достаточно, чтобы выполнялись следующие условия:

1) она была интерполяционной, т.е.

$$A_k=\int_a^b p(x)\frac{\omega_{n-m+1}(x)\Omega_m(x)}{(x-x_k)\omega'_{n-m+1}(x_k)\Omega_m(x_k)}dx\,,\ k=0,...,n-m\,,$$

$$B_l=\int_a^b p(x)\frac{\omega_{n-m+1}(x)\Omega_m(x)}{(x-a_l)\omega_{n-m+1}(a_l)\Omega'_m(a_l)}dx\,,\ l=1,...,m; \tag{4.2}$$

2) многочлен $\omega_{n-m+1}(x)$ был ортогонален на отрезке $[a;b]$ по весу $p(x)\Omega_m(x)$ ко всем многочленам степени не выше $n-m$, т.е.

$$\int_a^b p(x)\Omega_m(x)Q_p(x)\omega_{n-m+1}(x)dx = 0, \quad p \le n-m. \tag{4.3}$$

Доказательство (аналогично доказательству *теоремы* 1 из § 3).
Необходимость. Так как квадратурная формула (4.1) точна для всех многочленов до степени $2n-m+1$ включительно и $2n-m+1 \ge 2n-(n+1)+1 = n$, то она точна и для многочленов степени n и, таким образом, является интерполяционной (в силу критерия интерполяционных квадратурных формул).

Кроме того, положив $f(x) = \Omega_m(x)\omega_{n-m+1}(x)Q(x)$, где $Q(x)$ – произвольный многочлен степени не выше $n-m$, имеем: $f(x)$ – многочлен, степень которого не превосходит $2n-m+1$. Следовательно, для такой $f(x)$ квадратурная формула (4.1) точна. Поэтому

$$\int_a^b p(x)\Omega_m(x)\omega_{n-m+1}(x)Q(x)dx = \sum_{l=1}^{m} B_l\Omega_m(a_l)\omega_{n-m+1}(a_l)Q(a_l) + \sum_{k=0}^{n-m} A_k\Omega_m(x_k)\omega_{n-m+1}(x_k)Q(x_k) = 0.$$

Достаточность. Пусть теперь $f(x)$ – произвольный многочлен, степень которого не превосходит $2n-m+1$. Представим его в виде

$$f(x) = \Omega_m(x)\omega_{n-m+1}(x)Q(x) + r(x),$$

где $Q(x)$ – многочлен степени не выше $n-m$, а $r(x)$ – многочлен степени не выше n (остаток от деления $f(x)$ на $\Omega_m(x)\omega_{n-m+1}(x)$). При этом, очевидно,

$$f(a_l) = r(a_l), \ l = 1,...,m$$

$$f(x_k) = r(x_k), \ , k = 0,...,n-m.$$

Тогда имеем:

$$\int_a^b p(x)f(x)dx = \int_a^b p(x)\Omega_m(x)\omega_{n-m+1}(x)Q(x)dx + \int_a^b p(x)r(x)dx \overset{(4.3)}{=} 0 + \int_a^b p(x)r(x)dx \overset{(4.2)}{=}$$

$$\overset{(4.2)}{=} \sum_{l=1}^{m} B_l r(a_l) + \sum_{k=0}^{n-m} A_k r(x_k) = \sum_{l=1}^{m} B_l f(a_l) + \sum_{k=0}^{n-m} A_k f(x_k).$$

Полученное равенство завершает доказательство.

⊠

В то же время, другие теоремы, аналогичные *теоремам* 2 – 4 предыдущего параграфа, здесь, вообще говоря, не имеют места, поскольку в общем случае произведение $p(x)\Omega_m(x)$ не является знакопостоянным (даже при знакопостоянной функции $p(x)$). Поэтому далее будем предполагать, что многочлен $\omega_{n-m+1}(x)$ существует, т.е. квадратурная формула вида (4.1), обладающая алгебраической степенью точности, равной $2n-m+1$, может быть построена.

Получим в этом случае представление остатка. Для этого, как и выше, выполним интерполирование функции $f(x)$ на отрезке $[a;b]$ по следующим условиям:

$$P_{2n-m+1}(a_l) = f(a_l),\ l = 1,...,m,$$

$$P_{2n-m+1}(x_k) = f(x_k),\ \ P'_{2n-m+1}(x_k) = f'(x_k),\ k = 0,...,n-m.$$

Тогда остаток такого интерполирования может быть представлен в виде

$$r(x) = \Omega_m(x)\omega^2_{n-m+1}(x)\frac{f^{(2n-m+2)}(\xi)}{(2n-m+2)!},\ \ \xi \in [a;b].$$

Интегрируя равенство $f(x) = P_{2n-m+1}(x) + r(x)$, получим квадратурную формулу (4.1) и

$$R_n(f) = \int_a^b p(x)r(x)dx = \int_a^b p(x)\Omega_m(x)\omega^2_{n-m+1}(x)\frac{f^{(2n-m+2)}(\xi)}{(2n-m+2)!}dx\,. \qquad (4.4)$$

Точно таким же образом, как и в предыдущем параграфе, можно получить формулы для вычисления коэффициентов A_k, не требующие вычисления интегралов:

$$A_k = \frac{c_{n-m+1}}{c_{n-m}} \cdot \frac{1}{\Pi_{n-m}(x_k)\cdot\Pi'_{n-m+1}(x_k)\cdot\Omega_m(x_k)},\ k = 0,...,n-m, \qquad (4.5)$$

где $\{\Pi_i(x)\}$ – система многочленов, ортонормированная на отрезке $[a;b]$ по весу $p(x)\Omega_m(x)$.

Таким образом, мы вторично сталкиваемся с необходимостью находить систему многочленов, ортогональных на отрезке $[a;b]$ по весу $p(x)\Omega_m(x)$. В некоторых случаях здесь может оказаться полезной теорема о преобразовании ортогональной системы многочленов при умножении веса на многочлен.

***Теорема* 2.** Пусть $\Omega_m(x) = (x-a_1)\cdot\ldots\cdot(x-a_m)$ и существуют и единственны системы приведенных многочленов $\widetilde{P}_s(x)$ и $\widetilde{\Pi}_s(x)$, ортогональные на отрезке $[a;b]$ по весу $p(x)$ и $p(x)\Omega_m(x)$ соответственно. Тогда

$$\widetilde{\Pi}_{n-m}(x) = \frac{1}{\Delta\cdot\Omega_m(x)}\cdot\begin{vmatrix} \widetilde{P}_n(x) & \widetilde{P}_n(a_1) & \ldots & \widetilde{P}_n(a_m) \\ \widetilde{P}_{n-1}(x) & \widetilde{P}_{n-1}(a_1) & \ldots & \widetilde{P}_{n-1}(a_m) \\ \vdots & \vdots & \ldots & \vdots \\ \widetilde{P}_{n-m}(x) & \widetilde{P}_{n-m}(a_1) & \ldots & \widetilde{P}_{n-m}(a_m) \end{vmatrix}, \qquad (4.6)$$

где

$$\Delta = \begin{vmatrix} \widetilde{P}_{n-1}(a_1) & \ldots & \widetilde{P}_{n-1}(a_m) \\ \vdots & \ldots & \vdots \\ \widetilde{P}_{n-m}(a_1) & \ldots & \widetilde{P}_{n-m}(a_m) \end{vmatrix}.$$

Доказательство. Так как $\Omega_m(x)\widetilde{\Pi}_{n-m}(x)$ есть многочлен степени n со старшим членом x^n, то его можно разложить по многочленам системы $\widetilde{P}_s(x)$ в виде

$$\Omega_m(x)\widetilde{\Pi}_{n-m}(x) = \widetilde{P}_n(x) + c_1\widetilde{P}_{n-1}(x) + c_2\widetilde{P}_{n-2}(x) + \cdots .$$

Ортогональность $\widetilde{\Pi}_{n-m}(x)$ с весом $p(x)\Omega_m(x)$ ко всякому многочлену степени, меньшей $n-m$, равносильна тому, что в указанном разложении должны отсутствовать члены с $\widetilde{P}_s(x)$ для $s \le n-m-1$. Этот факт легко установить, если умножить записанное соотношение на произведение $p(x)\widetilde{P}_s(x)$ и проинтегрировать по отрезку $[a;b]$. Тогда получится равенство

$$0 = c_s \int_a^b p(x)\widetilde{P}_s^2 dx ,$$

из которого немедленно следует: $c_s = 0$.

Таким образом,

$$\Omega_m(x)\widetilde{\Pi}_{n-m}(x) = \widetilde{P}_n(x) + c_1\widetilde{P}_{n-1}(x) + c_2\widetilde{P}_{n-2}(x) + \cdots + c_m\widetilde{P}_{n-m}(x). \qquad (4.7)$$

Полагая в (4.7) x поочередно равным a_1, a_2,...,a_m, получим систему уравнений для определения коэффициентов разложения c_i:

$$\begin{cases} \widetilde{P}_n(a_1) + c_1\widetilde{P}_{n-1}(a_1) + c_2\widetilde{P}_{n-2}(a_1) + \cdots + c_m\widetilde{P}_{n-m}(a_1) = 0, \\ \cdots\cdots \\ \widetilde{P}_n(a_m) + c_1\widetilde{P}_{n-1}(a_m) + c_2\widetilde{P}_{n-2}(a_m) + \cdots + c_m\widetilde{P}_{n-m}(a_m) = 0 \end{cases} \qquad (4.8)$$

(эти равенства одновременно служат гарантом того, что правая часть равенства (4.7) делится на $\Omega_m(x)$ нацело, поскольку свидетельствуют о том, что у $\Omega_m(x)$ и правой части (4.7) одни и те же корни).

Так как по условию $\widetilde{\Pi}_{n-m}(x)$ существует, то система (4.8) должна иметь единственное решение, т.е. ее определитель, совпадающий с выписанным в условии теоремы Δ, отличен от нуля.

Чтобы получить формулу (4.6), добавим к (4.8) уравнение (4.7) и рассмотрим получившуюся систему

$$\begin{cases} \left(-\Omega_m(x)\cdot\widetilde{\Pi}_{n-m}(x) + \widetilde{P}_n(x)\right)\cdot 1 + c_1\widetilde{P}_{n-1}(x) + c_2\widetilde{P}_{n-2}(x) + \cdots + c_m\widetilde{P}_{n-m}(x) = 0, \\ \widetilde{P}_n(a_1)\cdot 1 + c_1\widetilde{P}_{n-1}(a_1) + c_2\widetilde{P}_{n-2}(a_1) + \cdots + c_m\widetilde{P}_{n-m}(a_1) = 0, \\ \cdots\cdots \\ \widetilde{P}_n(a_m)\cdot 1 + c_1\widetilde{P}_{n-1}(a_m) + c_2\widetilde{P}_{n-2}(a_m) + \cdots + c_m\widetilde{P}_{n-m}(a_m) = 0 \end{cases}$$

как однородную систему, состоящую из $(m+1)$ уравнений с неизвестными 1, c_1,...,c_m. Так как данная система имеет ненулевое решение, то ее определитель должен быть равен нулю, т.е.

$$\begin{vmatrix} -\Omega_m(x)\widetilde{\Pi}_{n-m}(x) + \widetilde{P}_n(x) & \widetilde{P}_{n-1}(x) & \dots & \widetilde{P}_{n-m}(x) \\ 0 + \widetilde{P}_n(a_1) & \widetilde{P}_{n-1}(a_1) & \dots & \widetilde{P}_{n-m}(a_1) \\ \vdots & \vdots & \dots & \vdots \\ 0 + \widetilde{P}_n(a_m) & \widetilde{P}_{n-1}(a_m) & \dots & \widetilde{P}_{n-m}(a_m) \end{vmatrix} = 0 .$$

Представляя данный определитель в виде суммы двух определителей, а затем разлагая один из них по первому столбцу, получим (4.6). ⊠

4.1. Формулы с предписанными узлами частного вида (Квадратурные формулы типа Маркова)

Ограничимся здесь рассмотрением следующих случаев:

1^0. $m=1$ и $a_1=a$;

2^0. $m=1$ и $a_1=b$;

3^0. $m=2$ и $a_1=a,\ a_2=b$.

Во всех этих случаях при знакопостоянной весовой функции $p(x)$ произведение $p(x)\Omega_m(x)$ также знакопостоянно на отрезке $[a;b]$ и, следовательно, квадратурные формулы, имеющие алгебраическую степень точности, равную $2n-m+1$, всегда могут быть построены (см. *теорему* 2 из § 3).

Итак, пусть имеем случай **1^0** (случай **2^0** сводится к нему линейной заменой переменной интегрирования $x=a+b-t$ и отдельно рассматриваться не будет).

Тогда

$$I=\int_a^b p(x)f(x)dx=Bf(a)+\sum_{k=0}^{n-1}A_k f(x_k)+R(f), \tag{4.9}$$

где

$$A_k=\frac{c_n}{c_{n-1}}\cdot\frac{1}{b(x_k-a)\Pi_{n-1}(x_k)\Pi_n'(x_k)},\ k=0,1,...,n-1,$$

$$B=\frac{1}{\Pi(a)}\int_a^b p(x)\Pi_n(x)dx. \tag{4.10}$$

При этом оказывается, что все коэффициенты положительны в случае положительной весовой функции.

Упражнение. Доказать положительность коэффициентов (4.10).

Алгебраическая степень точности квадратурной формулы (4.9) равна $2n$, а ее остаток имеет вид

$$R(f)=\frac{f^{(2n+1)}(\eta)}{(2n+1)!}\cdot\int_a^b p(x)(x-a)\omega_n^2(x)dx. \tag{4.11}$$

Рассмотрим сейчас несколько подробнее случай $p(x)\equiv 1$. Как и ранее, отрезок $[a;b]$ приведем к отрезку $[-1;1]$. Тогда квадратурная формула (4.9) примет вид

$$I=\int_{-1}^1 f(x)dx=Bf(-1)+\sum_{k=0}^{n-1}A_k f(x_k)+R(f). \tag{4.12}$$

При этом $\Omega_1(x)=1+x$ и, следовательно, многочлен $\omega_n(x)$ будет только коэффициентом отличаться от многочлена Якоби $P_n^{(0,1)}(x)$, а коэффициенты A_k только множителем $\frac{1}{\Omega_1(x_k)}$ будут отличаться от соответствующих коэффициентов формулы наивысшей алгебраической степени точности:

$$A_k = \frac{4}{(1+x_k)\left(1-x_k^2\right)\left[\dfrac{d}{dx}P_n^{(0,1)}(x_k)\right]^2},\ k=0,1,...,n-1,$$

$$B = \int_{-1}^{1}\frac{\omega_n(x)}{\omega_n(-1)}dx = \frac{1}{P_n^{(0,1)}(-1)}\int_{-1}^{1}P_n^{(0,1)}(x)dx = \frac{2}{(n+1)^2}, \tag{4.13}$$

$$R(f) = \frac{2}{n+1}\left[\frac{2^n\cdot n!\cdot(n+1)!}{(2n+1)!}\right]^2\cdot\frac{f^{(2n+1)}(\eta)}{(2n+1)!}. \tag{4.14}$$

3⁰. $\Omega_2(x) = (x-a)(b-x)$.

В соответствии с *теоремой* 2 имеем формулу для вычисления системы многочленов, ортогональных по весу $p(x)\Omega_2(x)$:

$$\Pi_{n-2}(x) = \frac{K_{n-2}}{(x-a)(x-b)}\begin{vmatrix} P_n(x) & P_n(a) & P_n(b) \\ P_{n-1}(x) & P_{n-1}(a) & P_{n-1}(b) \\ P_{n-2}(x) & P_{n-2}(a) & P_{n-2}(b) \end{vmatrix},$$

где $P_n(x)$ – многочлены, Образующие ортогональную систему по весу $p(x)$, а K_{n-2} – некоторая константа.

Все коэффициенты квадратурной формулы

$$\int_a^b p(x)f(x)dx \approx B_1 f(a) + B_2 f(b) + \sum_{k=0}^{n-2}A_k f(x_k)$$

знакопостоянны (положительны при положительном весе $p(x)$),

$$R(f) = \frac{f^{(2n)}(\eta)}{(2n)!}\cdot\int_a^b p(x)(x-a)(x-b)\omega_{n-1}^2(x)dx.$$

Вновь, как и выше, рассмотрим несколько подробнее случай $p(x)\equiv 1$ и $[a;b] = [-1;1]$. В этом случае квадратурная формула примет вид

$$I = \int_{-1}^{1}f(x)dx = B_1 f(-1) + B_2 f(1) + \sum_{k=0}^{n-2}A_k f(x_k) + R(f). \tag{4.15}$$

При этом, поскольку $\Omega_2(x) = 1-x^2$, то многочлен $\omega_{n-1}(x)$ будет только коэффициентом отличаться от многочлена Якоби $P_n^{(1,1)}(x)$, а коэффициенты A_k только множителем $\dfrac{1}{\Omega_2(x_k)}$ будут отличаться от соответствующих коэффициентов формулы наивысшей алгебраической степени точности:

$$A_k = 8\cdot\frac{n}{n+1}\cdot\frac{1}{\left(1-x_k^2\right)^2\left[\dfrac{d}{dx}P_{n-1}^{(1,1)}(x_k)\right]^2},\quad k=0,1,...,n-2, \tag{4.16}$$

$$B_1 = B_2 = \frac{2}{n(n+1)}, \tag{4.17}$$

$$R(f) = \frac{8n}{(2n+1)(n+1)} \cdot \left[\frac{2^{n-1} \cdot (n-1)! \cdot (n+1)!}{(2n)!} \right]^2 \cdot \frac{f^{(2n)}(\eta)}{(2n)!}. \tag{4.18}$$

§ 5. Квадратурные формулы с равными коэффициентами

В приложениях достаточно удобными могут оказаться квадратурные формулы, все коэффициенты которых одинаковы, т.е. квадратурные формулы, имеющие вид

$$I = \int_a^b p(x) f(x) dx \approx C_{n+1} \sum_{k=0}^{n} f(x_k). \tag{5.1}$$

Их называют ***квадратурными формулами Чебышева***. Требование точного выполнения равенства (5.1) при $f(x) \equiv 1$ приводит к уравнению

$$\int_a^b p(x) dx = (n+1) C_{n+1},$$

откуда

$$C_{n+1} = \frac{1}{n+1} \int_a^b p(x) dx. \tag{5.2}$$

Если, кроме того, потребовать, чтобы равенство (5.1) точно выполнялось для $f(x)$, равных $x, x^2, ..., x^{n+1}$, то для нахождения узлов x_i получим систему алгебраических уравнений

$$\begin{cases} x_0 + x_1 + \cdots + x_n = \dfrac{1}{C_{n+1}} \displaystyle\int_a^b p(x) x dx, \\ x_0^2 + x_1^2 + \cdots + x_n^2 = \dfrac{1}{C_{n+1}} \displaystyle\int_a^b p(x) x^2 dx, \\ \qquad\qquad \cdots\cdots \\ x_0^{n+1} + x_1^{N+1} + \cdots + x_n^{N+1} = \dfrac{1}{C_{n+1}} \displaystyle\int_a^b p(x) x^{N+1} dx. \end{cases} \tag{5.3}$$

Таким образом, для построения квадратурной формулы вида (5.1) достаточно по формуле (5.2) найти коэффициент C_{n+1}, а затем, решив систему (5.3), вычислить узлы искомой формулы. Однако, учитывая нелинейность системы (5.3), при практическом построении квадратурной формулы вида (5.1) удобнее искать не узлы x_i, а коэффициенты a_i многочлена $\omega_{n+1}(x)$:

$$\omega_{n+1}(x) = x^{n+1} + a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n.$$

Воспользуемся соотношениями Ньютона, связывающими коэффициенты a_i многочлена $\omega_{n+1}(x)$ и степенные суммы его корней $S_k = x_0^k + x_1^k + \cdots + x_n^k$. Тогда вместо формул (5.3) получим систему для определения величин a_i:

$$\begin{cases} S_1 + a_0 = 0, \\ S_2 + a_0 \cdot S_1 + 2a_1 = 0, \\ \quad \cdots\cdots \\ S_n + a_0 \cdot S_{n-1} + a_1 \cdot S_{n-2} + \cdots + na_{n-1} = 0, \\ S_{n+1} + a_0 \cdot S_n + a_1 \cdot S_{n-1} + \cdots + (n+1)a_n = 0. \end{cases} \tag{5.4}$$

Формулы (5.4) позволяют последовательно найти все коэффициенты a_i по известным (см. (5.3)) значениям S_k. Далее, решив уравнение $\omega_{n+1}(x) = 0$, найдем все узлы x_k, $k = \overline{0,n}$.

В соответствии с общей схемой алгебраическая степень точности квадратурной формулы (5.1) будет не менее чем $(n+1)$. Как показали исследования. При этом существует единственная квадратурная формула наивысшей алгебраической степени точности, имеющая равные коэффициенты (см. **§ 3**).

Сейчас рассмотрим несколько более подробно случай $p(x) \equiv 1$ и $[a;b] = [-1;1]$, т.е. квадратурные формулы вида

$$I = \int_{-1}^{1} f(x)dx \approx C_{n+1} \sum_{k=0}^{n} f(x_k). \tag{5.5}$$

В этом случае из (5.2) следует, что

$$C_{n+1} = \frac{1}{n+1} \int_{-1}^{1} dx = \frac{2}{n+1},$$

а так как

$$\int_{-1}^{1} x^k dx = \frac{1+(-1)^k}{k+1} = \begin{cases} 0, & \text{если } k \text{ нечетное}, \\ \dfrac{2}{k+1}, & \text{если } k \text{ четное}, \end{cases}$$

то система (5.3) примет вид

$$\begin{cases} S_1 = x_0 + x_1 + \cdots + x_n = 0, \\ S_2 = x_0^2 + x_1^2 + \cdots + x_n^2 = \dfrac{n+1}{3}, \\ S_3 = x_0^3 + x_1^3 + \cdots + x_n^3 = 0, \\ S_4 = x_0^4 + x_1^4 + \cdots + x_n^4 = \dfrac{n+1}{5}, \\ \quad \cdots\cdots \\ S_{n+1} = x_0^{n+1} + x_1^{n+1} + \cdots + x_n^{n+1} = \dfrac{n+1}{2} \cdot \dfrac{1+(-1)^{n+1}}{n+2}. \end{cases}$$

Следовательно, формулы для нахождения коэффициентов a_i примут вид

$$\begin{cases} a_0 = 0, \\ \dfrac{n+1}{3} + 2a_1 = 0, \\ a_2 = 0, \\ \dfrac{n+1}{5} + \dfrac{n+1}{3} a_1 + 4a_3 = 0, \\ \quad \ldots\ldots \end{cases}$$

Таким образом, все коэффициенты a_i четных номеров равны нулю и, следовательно, многочлен $\omega_{n+1}(x)$ будет иметь либо только четные, либо только нечетные степени x. Поэтому его корни располагаются на отрезке $[-1;1]$ симметрично относительно точки $x = 0$ (а значит, в случае, если $x = 0$ – узел, т.е. при $n = 2m$ алгебраическая степень точности формулы (5.5) будет не ниже $(n+2)$).

Рассмотрим примеры таких квадратурных формул.

1^0. $n = 0$.

Тогда $a_0 = 0$, $\omega_1(x) = x$. Следовательно, $x_0 = 0$ и, учитывая, что $C_1 = 2$, имеем квадратурную формулу

$$I = \int_{-1}^{1} f(x)dx \approx 2f(0),$$

алгебраическая степень точности которой равна 1.

2^0. $n = 1$.

Тогда $C_2 = 1$, для определения коэффициентов a_i имеем систему

$$\begin{cases} a_0 = 0, \\ 2a_1 + \dfrac{2}{3} = 0, \end{cases}$$

откуда $a_0 = 0$, $a_1 = -\dfrac{1}{3}$, т.е. $\omega_2(x) = x^2 - \dfrac{1}{3}$. Следовательно, $x_0 = -\dfrac{1}{\sqrt{3}}$, $x_1 = \dfrac{1}{\sqrt{3}}$ и квадратурная формула будет иметь вид

$$I = \int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Замечание. Было установлено, что при $n > 8$ для весовой функции $p(x) \equiv 1$ квадратурных формул Чебышева не существует, поскольку среди корней многочлена $\omega_{n+1}(x)$, построенного описанным выше способом, обязательно появляются комплексные. Лишь сравнительно недавно (в 1966 г.) были найдены весовые функции $p(x)$, для которых квадратурные формулы Чебышева существуют при любых n.

§ 6. Нестандартные приемы интегрирования

Как мы уже отмечали ранее, использование априорной информации о свойствах и характере поведения функции может существенно улучшить качество приближения. Так, периодические (или близкие к ним) функции более естественно приближать тригономет-

рическими многочленами, функции, имеющие экспоненциальное поведение, – многочленами от экспонент, и т.д.

Такой подход может оказать существенную помощь и при построении квадратурных формул.

6.1. Метод Филона

В радиотехнических задачах часто встречаются функции $f(x)$, описывающие высокочастотные колебания $\mathrm{e}^{i\omega x}$ с модулированной амплитудой. Это – быстропеременные функции и их производные $f^{(p)}(x) \sim \omega^p$ велики. Поэтому при интегрировании их по «штатным» квадратурным формулам приходится брать настолько мелкий шаг, чтобы выполнялось условие $\omega h << 1$, т.е. чтобы одна осцилляция содержала бы достаточное число узлов интегрирования. А это приводит к большому объему вычислений.

Для уменьшения объема вычислений используем априорные сведения о подынтегральной функции. Представим ее в виде $f(x) = y(x)\mathrm{e}^{i\omega x}$, где частота ω известна, а амплитуда $y(x)$ мало меняется за период основного колебания. Выбирая для $y(x)$ несложные полиномиальные аппроксимации, можем получить квадратурные формулы, называемые ***формулами Филона*** (по сути, речь идет о том, что мы рассматриваем $\mathrm{e}^{i\omega x}$ как весовую функцию).

Построим, например, аналог квадратурной формулы средних прямоугольников. Для этого при вычислении интеграла по отдельному интервалу сетки заменим амплитуду ее значением в середине интервала:

$$y(x) \approx y_{k-\frac{1}{2}}, \quad x \in [x_{k-1}; x_k].$$

При этом для остатка может быть записано приближенное представление (получаемое путем разложения в ряд Тейлора)

$$r(x) = y(x) - y_{k-\frac{1}{2}} \approx \left(x - x_{k-\frac{1}{2}}\right) y'_{k-\frac{1}{2}}.$$

Тогда для вычисления интеграла получим формулу

$$I = \int_a^b \mathrm{e}^{i\omega x}\, y(x)dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \mathrm{e}^{i\omega x}\, y(x)dx \approx \sum_{k=1}^{n} y_{k-\frac{1}{2}} \int_{x_{k-1}}^{x_k} \mathrm{e}^{i\omega x}\, dx = \tag{6.1}$$

$$= \sum_{k=1}^{n} y_{k-\frac{1}{2}} \frac{\mathrm{e}^{i\omega x_k} - \mathrm{e}^{i\omega x_{k-1}}}{i\omega} = \sum_{k=1}^{n} y_{k-\frac{1}{2}}\, \mathrm{e}^{i\omega x_{k-\frac{1}{2}}} \frac{\mathrm{e}^{i\frac{\omega}{2}h_k} - \mathrm{e}^{-i\frac{\omega}{2}h_k}}{2i\frac{\omega}{2}} = \frac{2}{\omega} \sum_{k=1}^{n} f_{k-\frac{1}{2}} \sin\left(\frac{\omega}{2} h_k\right),$$

а для ее остатка – выражение

$$R = \int_a^b r(x)\mathrm{e}^{i\omega x}\, dx \approx \sum_{k=1}^{n} y'_{k-\frac{1}{2}} \int_{x_{k-1}}^{x_k} \left(x - x_{k-\frac{1}{2}}\right) \mathrm{e}^{i\omega x}\, dx = \sum_{k=1}^{n} y'_{k-\frac{1}{2}} \left[\frac{x - x_{k-\frac{1}{2}}}{i\omega} \mathrm{e}^{i\omega x} \Big|_{x_{k-1}}^{x_k} - \frac{1}{i\omega} \int_{x_{k-1}}^{x_k} \mathrm{e}^{i\omega x}\, dx \right] = \tag{6.2}$$

$$= \sum_{k=1}^{n} y'_{k-\frac{1}{2}} \left[-\frac{2}{i\omega^2} \mathrm{e}^{i\omega x_{k-\frac{1}{2}}} \sin\left(\frac{\omega}{2} h_k\right) + \frac{h_k}{2} \cdot \frac{1}{i\omega} \left(\mathrm{e}^{i\omega x_k} + \mathrm{e}^{i\omega x_{k-1}}\right) \right] = \frac{2i}{\omega^2} \sum_{k=1}^{n} y'_{k-\frac{1}{2}} \left[\sin\frac{\omega h_k}{2} - \frac{h_k}{2} \cos\frac{\omega h_k}{2} \right] \mathrm{e}^{i\omega x_{k-\frac{1}{2}}}$$

Несложно видеть, что при $h \to 0$ (6.1) и (6.2) переходят в обобщенную формулу средних прямоугольников и ее остаток соответственно.

Для построения формул Филона высокого порядка приходится использовать более сложные многочленные аппроксимации.

Упражнение. Построить аналог квадратурной формулы трапеций.

6.2. Повышение гладкости интегрируемой функции

Как мы уже отмечали, как правило, все особенности подынтегрального выражения стараются включить в весовую функцию. В литературе этот способ носит название ***мультипликативного способа*** выделения особенностей**.** Пример такого рода мы приводили выше.

Вторым способом ослабления особенностей является ***аддитивный***. Его суть состоит в следующем. Функцию $f(x)$ представляют в виде $f_1(x)+f_2(x)$, где $f_1(x)$ содержит все или почти все особенности $f(x)$ и при этом интеграл $I_1 = \int_a^b p(x)f_1(x)dx$ вычисляется точно, а $f_2(x)$ имеет ослабленные особенности и для нее с бóльшим успехом применимы квадратурные формулы.

Рассмотрим пример:

$$I = \int_0^{\frac{\pi}{2}} \ln\sin x dx .$$

Подынтегральная функция имеет логарифмическую особенность на левом конце отрезка интегрирования. Поэтому представляет ее в виде

$$\ln\sin x = \ln x + \ln\frac{\sin x}{x} .$$

Тогда

$$I = I_1 + I_2 = \int_0^{\frac{\pi}{2}} \ln x dx + \int_0^{\frac{\pi}{2}} \ln\frac{\sin x}{x} dx .$$

При этом

$$I_1 = \frac{\pi}{2}\left(\ln\frac{\pi}{2} - 1\right),$$

а в I_2 подынтегральная функция не имеет особенностей и I_2 может быть вычислен, например, по квадратурной формуле Симпсона.

Рассмотрим далее несколько подробнее случай $p(x) \equiv 1$ и алгебраических особенностей. Пусть $f(x) = (x - x_0)^{\alpha}\varphi(x)$, где $\alpha > -1$, $x_0 \in [a;b]$, а $\varphi(x)$ – достаточно гладкая. При $\alpha < 0$ $f(x)$ имеет алгебраическую особенность, а при $\alpha > 0$ и нецелых производные от $f(x)$ начиная с некоторого порядка будут иметь особенности.

Разложим $\varphi(x)$ в ряд Тейлора:

$$\varphi(x) \approx \varphi(x_0) + \frac{x - x_0}{1!}\varphi'(x_0) + \cdots + \frac{(x - x_0)^{k-1}}{(k-1)!}\varphi^{(k-1)}(x_0)$$

и представим $f(x)$ в виде $f(x)=f_1(x)+f_2(x)$, где

$$f_1(x)=(x-x_0)^{\alpha}\left[\varphi(x_0)+\frac{x-x_0}{1!}\varphi'(x_0)+\cdots+\frac{(x-x_0)^{k-1}}{(k-1)!}\varphi^{(k-1)}(x_0)\right],$$

$$f_2(x)=f(x)-f_1(x).$$

Интеграл от $f_1(x)$ вычисляется точно, а вычисление интеграла от $f_2(x)$ с помощью квадратурных формул должно дать более хороший результат, так как $f_2(x)$ имеет более высокий порядок гладкости (конкретно: порядок выше на k единиц, поскольку справедливо соотношение $f_2(x)=(x-x_0)^{\alpha}\left[\frac{\varphi^{(k)}(\xi)}{k!}(x-x_0)^k\right]$).

6.3. Случай бесконечных пределов интегрирования

Основными подходами, предназначенными для решения указанной задачи, являются следующие:

1^0. Введение замены переменных, превращающей пределы интегрирования в конечные. Например, для интеграла $\int\limits_a^{\infty} f(x)dx$, $a>0$ замена $x=\frac{a}{1-t}$ превращает полупрямую $[a;+\infty)$ в отрезок $[0;1]$. Если после замены подынтегральная функция вместе с некоторым числом производных остается ограниченной, то интеграл можно найти стандартными (описанными выше) способами.

2^0. «Обрезание» бесконечного предела.
Пользуясь свойством аддитивности интеграла, представляем его в виде

$$\int\limits_a^{+\infty} f(x)dx=\int\limits_a^{b} f(x)dx+\int\limits_b^{+\infty} f(x)dx\approx\int\limits_a^{b} f(x)dx.$$

Очевидно, качество последнего приближенного равенства существенным образом зависит от выбора величины b. Поэтому данный подход требует корректной аналитической оценки и учета величины отброшенного слагаемого (при дальнейшем стандартном вычислении оставленного). Фактически данный прием хорошо комбинировать с применением асимптотических оценок для отбрасываемого члена.

3^0. Применение квадратурных формул наивысшей алгебраической степени точности со специальными весовыми функциями (см., например, п. **3.1**, формулы Чебышева-Лягерра и Чебышева-Эрмита).

4^0. Построение специальных ***нелинейных*** квадратурных формул, применимых на бесконечном интервале.

ГЛАВА VII

Вычисление кратных интегралов

При решении задачи о приближенном вычислении кратных интегралов чаще других применяются следующие два подхода:

1) сведение задачи к последовательному приближенному вычислению цепочки однократных интегралов. Его идеологической базой является известный в анализе прием сведения кратных интегралов к повторным. При этом, естественно, поскольку речь идет о приближенном вычислении однократных интегралов, то полностью работает вся изложенная в предыдущей главе теория квадратурных формул;
2) построение специальных приближенных формул, непосредственно решающих задачу о вычислении кратного интеграла минуя описанную выше промежуточную стадию.

С точки зрения общей теории при использовании второго подхода к решению задачи о приближенном вычислении кратных интегралов можно пользоваться практически теми же самыми основными понятиями, которые вводились в предыдущей главе. Рассмотрим постановку задачи подробнее.

Пусть необходимо в n-мерном пространстве E_n вычислить интеграл по некоторой области Ω

$$I = \int_{\Omega} p(x) f(x) dx , \tag{1}$$

где $x = (x_1, x_2, ..., x_n)$, $dx = dx_1 dx_2 \dots dx_n$; $p(x)$ – (как и ранее) весовая функция, в состав которой мы будем включать все или основные «неприятности» подынтегрального выражения. Более того, по опыту предыдущей главы будем $p(x)$ считать такой, что существуют интегралы (их мы будем называть моментами) $\mu_{\alpha_1 \dots \alpha_n} = \int_{\Omega} P(x) x_1^{\alpha_1} \dots x_n^{\alpha_n} dx$.

Тогда приближенная формула для вычисления интеграла (1) (по аналогии с одномерным случаем) будет иметь вид

$$I = \int_{\Omega} p(x) f(x) dx \approx \sum_{k=0}^{n} A_k f\left(x^{(k)}\right) \tag{2}$$

и называться ***кубатурной формулой***, A_k – ее коэффициенты, а $x^{(k)} = \left(x_i^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right)$ – узлы.

Если при подстановке в (2) вместо $f(x)$ любого алгебраического многочлена от n переменных до степени m включительно равенство превращается в точное (и уже не является таковым при степени многочлена, равной $m+1$), то m – алгебраическая степень точности кубатурной формулы (2).

Для построения кубатурных формул можно пользоваться всеми теми же приемами, которые мы рассматривали в Главе VI. Это:

1) непосредственное использование определения алгебраической степени точности;
2) интерполяционная замена интегрируемой функции с последующим точным вычислением интеграла.

При этом правомочна постановка задач о построении кубатурных формул с минимальным числом узлов и максимально возможной алгебраической степенью точности. Известные на настоящий момент результаты теории кубатурных формул аналогичны результатам, изложенным выше (см. Гл. VI). Так, имеет место понятие «***интерполяционная кубатурная формула***» и соответствующая теорема-критерий, а также теоремы о распределе-

нии узлов кубатурных формул, обладающих экстремальными характеристиками. При этом распределение узлов связано с поверхностями, определяемыми системами ортогональных многочленов.

Далее, учитывая, что все сказанное в той или иной мере справедливо для произвольного n, мы будем рассматривать вопросы, связанные с вычислением двукратных интегралов.

§ 1. Кубатурные формулы, основанные на сведении кратного интеграла к повторному

Как известно из курса анализа, вычисление кратных интегралов может быть осуществлено путем повторного вычисления однократных интегралов. Поэтому, как уже отмечалось выше, одним из простейших путей получения формул для приближенного вычисления кратных интегралов является повторное применение полученных нами ранее квадратурных формул для вычисления однократных интегралов.

Проиллюстрируем это на примере вычисления двойного интеграла по прямоугольнику:

$$I = \int_a^b \int_c^d f(x,y)dxdy\,. \tag{1.1}$$

Запишем интеграл (1.1) в виде

$$I = \int_a^b dx \int_c^d f(x,y)dy\,. \tag{1.2}$$

Применяя для вычисления внешнего интеграла квадратурную формулу средних прямоугольников, можем записать:

$$I = \int_a^b dx \int_c^d f(x,y)dy \approx (b-a)\int_c^d f\left(\frac{a+b}{2}, y\right)dy\,.$$

Вычислив теперь оставшийся интеграл также по формуле средних прямоугольников, окончательно получим:

$$I \approx (b-a)(d-c)f\left(\frac{a+b}{2}, \frac{c+d}{2}\right) = S \cdot f\left(\frac{a+b}{2}, \frac{c+d}{2}\right). \tag{1.3}$$

В качестве других примеров рассмотрим варианты кубатурных формул, получаемых на основе применения других известных вариантов квадратурных формул при повторном интегрировании.

1^0. Формула трапеций:

$$I = \int_a^b dx \int_c^d f(x,y)dy \approx \frac{b-a}{2}\left[\int_c^d f(a,y)dy + \int_c^d f(b,y)dy\right] \approx$$

$$\approx \frac{b-a}{2} \cdot \left[\frac{d-c}{2}(f(a,c)+f(a,d)) + \frac{d-c}{2}(f(b,c)+f(b,d))\right] \approx \tag{1.4}$$

$$\approx \frac{S}{4} \cdot [f(a,c)+f(a,d)+f(b,c)+f(b,d)].$$

2^0. Формула Симпсона:

$$I=\int_a^b dx\int_c^d f(x,y)dy \approx \frac{b-a}{6}\cdot\left[\int_c^d f(a,y)dy+4\int_c^d f\left(\frac{a+b}{2},y\right)dy+\int_c^d f(b,y)dy\right]\approx$$

$$\approx \frac{S}{36}\cdot[f(a,c)+f(a,d)+f(b,c)+f(b,d)]+ \qquad (1.5)$$

$$+\frac{S}{9}\cdot\left[f\left(a,\frac{c+d}{2}\right)+f\left(\frac{a+b}{2},c\right)+f\left(\frac{a+b}{2},d\right)+f\left(b,\frac{c+d}{2}\right)+4f\left(\frac{a+b}{2},\frac{c+d}{2}\right)\right].$$

Общая схема построения указанного типа кубатурных формул может быть получена, если воспользоваться формулами повторного интерполирования. Например, используя представление соответствующего интерполяционного многочлена в форме Лагранжа (см. формулу (8.11) Гл. IV)

$$P_{n,m}(x,y)=\sum_{i=0}^{n}\sum_{j=0}^{m}\frac{\omega_{n+1}(x)\omega_{m+1}(y)}{(x-x_i)(y-y)\omega'_{n+1}(x_i)\omega'_{m+1}(y_j)}f(x_i,y_j),$$

получим:

$$I=\int_a^b\int_c^d f(x,y)dxdy \approx \sum_{i=0}^{n}\sum_{j=0}^{m}f(x_i,y_j)\int_a^b\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}dx\int_c^d\frac{\omega_{m+1}(y)}{(y-y_j)\omega'_{m+1}(y)}dy. \qquad (1.6)$$

Рассмотрим теперь случай, когда область интегрирования Ω не является прямоугольником, но удовлетворяет условиям, при которых может быть осуществлено сведение к повторному интегралу без разбиения ее на подобласти (для этого достаточно, чтобы контур области пересекался прямыми, параллельными координатным осям, только в двух точках).

Тогда

$$I=\iint_\Omega f(x,y)dxdy=\int_a^b dx\int_{y_1(x)}^{y_2(x)}f(x,y)dy=\int_a^b F(x)dx.$$

Выбор правила для вычисления интеграла I, таким образом, должен быть согласован со свойствами функции $f(x,y)$ и, во-вторых, со свойствами области интегрирования Ω.

Если предположить, что $f(x,y)$ является достаточно гладкой всюду в Ω, то интеграл $F(x)$ может быть вычислен по одному из известных правил с постоянным весом, например, по правилу Гаусса, Симпсона и т.п. Форма области оказывает влияние только на границы интегрирования $y_1(x)$ и $y_2(x)$. Отрезок $[y_1(x);y_2(x)]$ можно привести к каноническому, например, к отрезку $[0;1]$ с помощью подстановки $y=y_1(x)+(y_2(x)-y_1(x))\eta$. Тогда

$$F(x)=(y_2(x)-y_1(x))\int_0^1 f(x,y_1(x)+(y_2(x)-y_1(x))\eta)d\eta=(y_2(x)-y_1(x))\Phi(x).$$

Выделившийся при замене в интеграле $I=\int_a^b F(x)dx$ множитель $y_2(x)-y_1(x)$ является естественной весовой функцией. Поэтому при вычислении интеграла $I=\int_a^b(y_2(x)-y_1(x))\Phi(x)dx$

можно воспользоваться любой квадратурной формулой, построенной для веса $p(x)=y_2(x)-y_1(x)$, например, квадратурной формулой наивысшей алгебраической степени точности.

Такой полный учет формы области, вероятно, неразумно делать, так как каждой области Ω будет отвечать свой вес $p(x)$ и поэтому пришлось бы использовать большое число узлов и коэффициентов.

Можно упростить задачу на основании следующих простых соображений. Рассмотрим две весовые функции, отличающиеся друг от друга достаточно гладким множителем $\rho(x)$, не обращающимся в нуль на отрезке интегрирования $[a;b]$: $q(x)=\rho(x)p(x)$. Тогда следует ожидать, что квадратурные формулы, соответствующие этим двум весовым функциям $p(x)$ и $q(x)$, будут близки по точности.

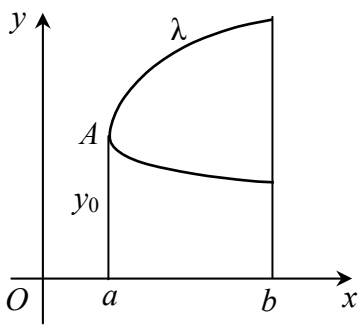
А теперь вспомним о весовой функции Якоби $q(x)=(x-a)^{\beta}(b-x)^{\alpha}$. Она зависит от двух параметров α и β и их часто можно подобрать таким образом, чтобы отношение

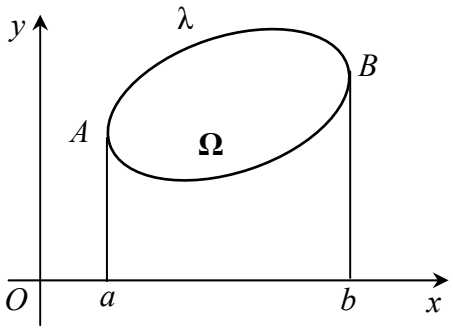
$$\rho(x)=\frac{y_2(x)-y_1(x)}{(b-x)^{\alpha}(x-a)^{\beta}},\ a\le x\le b,$$

было ограничено сверху и снизу положительными числами. В этом случае можно воспользоваться весом Якоби, преобразовав интеграл I к виду

$$I=\int_a^b (b-x)^{\alpha}(x-a)^{\beta}\psi(x)dx\ .$$

Например, если область интегрирования имеет форму, изображенную на рисунке слева, причем контур области λ имеет в точке A с прямой $x=a$ соприкосновение первого





порядка, то можно считать $\alpha=0$, $\beta=0{,}5$ и за весовую функцию принять $p(x)=\sqrt{x-a}$. Интеграл

$$I=\int_a^b \sqrt{x-a}\cdot\psi(x)dx$$

может быть вычислен с помощью формул (3.18), (3.19) главы VI.

Аналогично, в случае, если область интегрирования имеет вид, изображенный на рисунке справа и контур λ имеет с прямыми $x=a$ и $x=b$ соприкосновение первого порядка, то за весовую функцию можно принять $p(x)=\sqrt{(x-a)(b-x)}$, и к вычислению интеграла

$$I=\int_a^b \sqrt{(x-a)(b-x)}\cdot\psi(x)dx$$

также применить формулы (3.18), (3.19) главы VI.

§ 2. Простейшие кубатурные формулы

Несмотря на множество общих моментов, отмеченных во введении к данной главе, проблема вычисления кратных интегралов существенно сложнее по сравнению с аналогичной проблемой вычисления определенного интеграла, и в первую очередь, благодаря тому, что гораздо более сложной может быть область интегрирования (такой, например, является область с криволинейной, пусть и достаточно гладкой, границей). Поэтому изучение вопроса проведем на примерах наиболее простых областей.

2.1. Кубатурные формулы на прямоугольнике

Начнем с простейшей области интегрирования – прямоугольника: $\Omega = [a;b]\times[c;d]$ (весовую функцию $p(x,y)$ будем считать тождественной равной единице). Таким образом, речь идет о вычислении интеграла

$$I = \int_a^b \int_c^d f(x,y)dxdy\,. \tag{2.1}$$

Очевидно, замена независимых переменных $\begin{cases} x = \dfrac{b-a}{2}u + \dfrac{b+a}{2}, \\ y = \dfrac{d-c}{2}v + \dfrac{d+c}{2} \end{cases}$ переводит рассматриваемый интеграл в интеграл по квадрату $\Omega_1 = [-1;1]\times[-1;1]$:

$$I = \frac{(b-a)(d-c)}{4}\cdot\int_{-1}^{1}\int_{-1}^{1} f_1(u,v)dudv = \frac{S(\Omega)}{4}\cdot\int_{-1}^{1}\int_{-1}^{1} f_1(u,v)dudv\,.$$

Учитывая сказанное, далее более подробно рассмотрим вычисление интеграла

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy\,. \tag{2.2}$$

Самой простой кубатурной формулой будет, естественно, кубатурная формула с одним узлом, т.е. формула вида

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx A_0 f(x_0,y_0)\,. \tag{2.3}$$

Выберем узел (x_0,y_0) и коэффициент A_0 таким образом, чтобы обеспечить для рассматриваемой конструкции максимально возможную алгебраическую степень точности. Так как

$$\int_{-1}^{1}\int_{-1}^{1} dxdy = 4\,,\quad \int_{-1}^{1}\int_{-1}^{1} xdxdy = \int_{-1}^{1}\int_{-1}^{1} ydxdy = 0\,, \tag{2.3′}$$

то для определения параметров кубатурной формулы (2.3) получим, пользуясь определением алгебраической степени точности, систему уравнений

$$\begin{cases} A_0 = 4, \\ A_0 x_0 = 0, \\ A_0 y_0 = 0, \end{cases}$$

решив которую, найдем: $A_0 = 4$, $(x_0, y_0) = (0,0)$.

Следовательно, искомая кубатурная формула имеет вид

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx 4f(0,0). \tag{2.4}$$

Алгебраическая степень точности построенной формулы – не менее единицы (как несложно проверить – в точности равна 1). Найдем ее остаток.

Разлагая интегрируемую функцию $f(x,y)$ в ряд Тейлора в окрестности точки $(0,0)$, получим:

$$f(x,y) = f(0,0) + x\frac{\partial f(0,0)}{\partial x} + y\frac{\partial f(0,0)}{\partial y} + \frac{x^2}{2}\frac{\partial^2 f(0,0)}{\partial x^2} + xy\frac{\partial^2 f(0,0)}{\partial x \partial y} + \frac{y^2}{2}\frac{\partial^2 f(0,0)}{\partial y^2} + \cdots.$$

Тогда

$$R_0(f) = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy - 4f(0,0) =$$

$$= \int_{-1}^{1}\int_{-1}^{1}\left[x\frac{\partial f(0,0)}{\partial x} + y\frac{\partial f(0,0)}{\partial y} + \frac{x^2}{2}\frac{\partial^2 f(0,0)}{\partial x^2} + xy\frac{\partial^2 f(0,0)}{\partial x \partial y} + \frac{y^2}{2}\frac{\partial^2 f(0,0)}{\partial y^2} + \cdots \right] dxdy = \tag{2.5}$$

$$= \frac{2}{3}\left[\frac{\partial^2 f(0,0)}{\partial x^2} + \frac{\partial^2 f(0,0)}{\partial y^2} \right] + \cdots.$$

Заметим, что применительно к вычислению интеграла (2.1) кубатурная формула (2.4) будет выглядеть следующим образом:

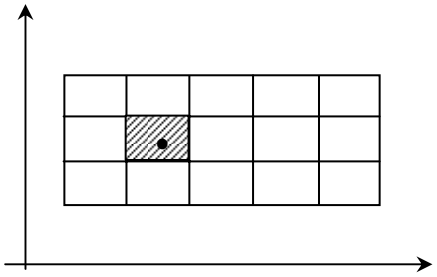
$$I = \int_{a}^{b}\int_{c}^{d} f(x,y)dxdy \approx S(\Omega)\cdot f\left(\frac{a+b}{2}, \frac{c+d}{2} \right), \tag{2.4′}$$

а ее остаток примет вид

$$R_0(f) = \frac{S(\Omega)}{24}\cdot\left[(b-a)^2 \frac{\partial^2 f\left(\frac{a+b}{2}, \frac{c+d}{2} \right)}{\partial x^2} + (d-c)^2 \frac{\partial^2 f\left(\frac{a+b}{2}, \frac{c+d}{2} \right)}{\partial y^2} \right] + \cdots. \tag{2.5′}$$

Учитывая, что точка $(\bar{x}, \bar{y}) = \left(\frac{a+b}{2}, \frac{c+d}{2} \right)$ является центром прямоугольника Ω, формулу (2.4′) (или (2.4)) (в точности совпадающую с формулой (1.3) предыдущего параграфа), называют ***кубатурной формулой средних***.

По аналогии с одномерным случаем легко получить ***составную* (*обобщенную*) *формулу средних***.



Разбивая область интегрирования на прямоугольные ячейки и применяя на каждой ячейке формулу средних (2.4′), получим:

$$I = \int_a^b \int_c^d f(x,y)dxdy \approx \sum_i S_i f(\bar{x}_i, \bar{y}_i). \qquad (2.6)$$

Здесь S_i – площадь i-й ячейки, а $(\bar{x}_i, \bar{y}_i)$ – координаты ее центра. Сетка, вообще говоря, не обязана быть равномерной по каждому направлению. Если же она таковой является, то формула (2.6) будет иметь несколько более простой вид:

$$I = \int_a^b \int_c^d f(x,y)dxdy \approx \sum_{i=-1}^{N_x} \sum_{j=1}^{N_y} h_x h_y f(\bar{x}_i, \bar{y}_i), \qquad (2.6')$$

где отрезок $[a;b]$ разбивается на N_x равных частей, $[c;d]$ – на N_y, т.е.

$$h_x = \frac{b-a}{N_x}, \quad h_y = \frac{d-c}{N_y}, \quad \bar{x}_i = a + \left(i - \frac{1}{2}\right) \cdot h_x, \quad \bar{y}_j = c + \left(j - \frac{1}{2}\right) \cdot h_y.$$

Для каждой ячейки Ω_{ij} остаток вычисляется по формуле (1.5′), т.е.

$$R_{ij}(f) = \frac{S(\Omega_{ij})}{24} \cdot \left[h_x^2 \frac{\partial^2 f(\bar{x}_i, \bar{y}_j)}{\partial x^2} + h_y^2 \frac{\partial^2 f(\bar{x}_i, \bar{y}_j)}{\partial y^2} \right] + \cdots.$$

Суммируя эти выражения по всем ячейкам сетки, получим погрешность составной кубатурной формулы средних:

$$R_0^C(f) = \frac{1}{24} \left[h_x^2 \int_a^b \int_c^d \frac{\partial^2 f(x,y)}{\partial x^2} dxdy + h_y^2 \int_a^b \int_c^d \frac{\partial^2 f(x,y)}{\partial y^2} dxdy \right] + \cdots = O\left(h_x^2 + h_y^2\right).$$

Таким образом, составная формула средних имеет второй порядок точности.

***Замечание* 1.** Формулу средних можно теоретически достаточно просто обобщить на случай более сложной области интегрирования Ω.

В этом случае (2.4′) примет вид

$$I = \iint_\Omega f(x,y)dxdy \approx S(\Omega) \cdot f(\bar{x}, \bar{y}), \qquad (2.7)$$

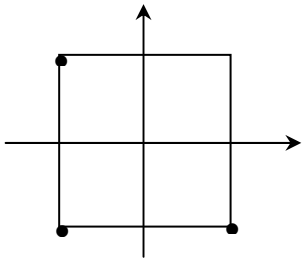
где $S(\Omega)$ – площадь области Ω, а $(\bar{x}, \bar{y})$ – координаты ее центра тяжести, т.е.

$$S(\Omega) = \iint_\Omega dxdy; \quad \bar{x} = \frac{1}{S(\Omega)} \iint_\Omega x dxdy; \quad \bar{y} = \frac{1}{S(\Omega)} \iint_\Omega y dxdy. \qquad (2.8)$$

Очевидно, формула (2.7) будет, как и в случае прямоугольной области, иметь алгебраическую степень точности, равную единице, что легко проверить непосредственно.

***Замечание* 2.** При любом другом расположении единственного узла кубатурная формула будет иметь алгебраическую степень точности, равную нулю.

Вновь возвращаясь к интегралу (2.2), рассмотрим другие примеры кубатурных формул. Естественно, целью поставим построение формул, имеющих более высокую алгебраическую степень точности по сравнению с формулой средних. Как уже отмечалось выше, важную роль играет распределение узлов кубатурной формулы. Вначале исследуем возможности простого количественного увеличения их числа с расположением в «естественных» местах области интегрирования.



Придерживаясь идеологии интерполяционной замены и вспоминая основные способы построения интерполяционного многочлена для функции двух независимых переменных, возьмем в качестве узлов интерполирования точки $(-1;-1)$, $(1;-1)$ и $(-1;1)$. Тогда

$$f(x,y) \approx P_1(x,y) = f(-1,-1) + (x+1)f(-1,1;-1) + (y+1)f(-1;-1,1)$$

Отсюда

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx 4f(-1,-1) + 2f(-1,1;-1) + 2f(-1;-1,1) = \tag{2.9}$$

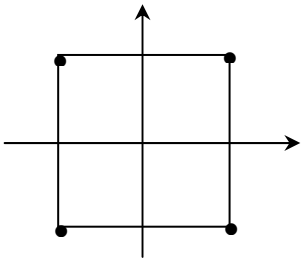
$$= 4f(-1,-1) + 4\frac{f(1,-1)-f(-1,-1)}{2} + 4\frac{f(-1,1)-f(-1,-1)}{2} = 2\cdot[f(-1,1)+f(1,-1)].$$

Таким образом, де-факто получилась кубатурная формула с ***двумя*** узлами, имеющая алгебраическую степень точности, равную единице.

Можно показать, что использование двух узлов не приводит к повышению алгебраической степени точности и в то же время выбор в качестве узлов любой пары точек, лежащих на прямой, проходящей через центр симметрии, позволяет построить кубатурную формулу с алгебраической степенью точности, равной единице.

Упражнение. Доказать сформулированное выше утверждение.

Точно так же и использование четырех узлов, расположенных в соответствии с «естественными эстетическими» соображениями, не приводит к повышению алгебраической степени точности. Убедимся в этом.

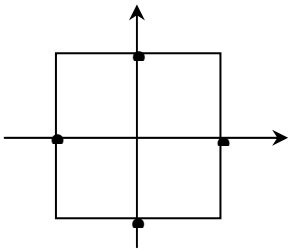


Вначале расположим узлы в вершинах квадрата. В этом случае кубатурную формулу проще всего строить, прибегая к процедуре повторного интерполирования. Очевидно, формула будет аналогична формуле (1.4) (поскольку интерполирование в каждом направлении проводится по двум узлам):

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx f(-1,-1) + f(-1,1) + f(1,-1) + f(1,1). \tag{2.10}$$

Несложно видеть, что алгебраическая степень точности построенной кубатурной формулы действительно будет равна единице, поскольку условие точности не выполняется ни для одного из элементарных многочленов второй степени, кроме xy.

В качестве второго примера рассмотрим случай, когда те же четыре узла кубатурной формулы расположены на серединах сторон квадрата, т.е. в точках $(-1,0),(0,1),(1,0)$ и $(0,-1)$. В данном случае интерполяционный многочлен строить несколько затруднительно ввиду его неоднозначности (по количеству узлов). Поэтому прибегнем непо-



средственно к определению алгебраической степени точности (и, как следствие, к методу неопределенных коэффициентов). Таким образом, будем искать кубатурную формулу вида

$$I=\int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx A_0 f(-1,0)+A_1 f(0,1)+A_2 f(1,0)+A_3 f(0,-1). \quad (2.11)$$

Учитывая формулы (2.3′), а также равенство $\int_{-1}^{1}\int_{-1}^{1} x^2 dxdy = \frac{4}{3}$, для определения коэффициентов формулы (2.11) получим систему уравнений

$$\begin{cases} A_0 + A_1 + A_2 + A_3 = 4, \\ -A_0 + A_2 = 0, \\ A_1 - A_3 = 0, \\ A_0 + A_2 = \frac{4}{3}, \end{cases}$$

из которой легко находим: $A_0 = A_2 = \frac{2}{3}$, $A_1 = A_3 = \frac{4}{3}$. Таким образом, (2.11) примет вид

$$I=\int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx \frac{2}{3}[f(-1,0)+2f(0,1)+f(1,0)+2f(0,-1)].$$

Заметим, что в этом случае выполняется также и уравнение, дающее точность на многочленах вида xy. Но в то же время $\int_{-1}^{1}\int_{-1}^{1} y^2 dxdy = \frac{4}{3}$, а $A_1 + A_3 = \frac{8}{3} \neq \frac{4}{3}$. Так что алгебраическая степень точности построенной кубатурной формулы остается равной единице, хотя и удовлетворяется на одно уравнение больше по сравнению с рассмотренным выше случаем.

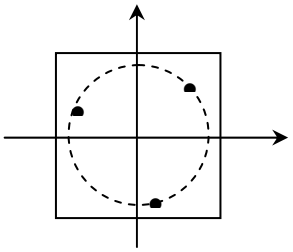
В общем случае, вспоминая, что многочлен второй степени от двух независимых переменных зависит от шести коэффициентов, третьей – от десяти и т.д., делаем вывод, что кубатурные формулы интерполяционного типа повышенной алгебраической степени точности будут достаточно трудоемкими. В то же время, пример формулы средних показывает, что возможно улучшение ситуации (формула средних вместо штатных трех узлов, расположенных почти произвольно (не на одной прямой (!)) содержит всего один, но специальный узел). В общем случае решение задачи минимизации количества узлов при заданной алгебраической степени точности связано с их расположением на некоторой алгебраической кривой [см., например, книгу Крылова В.И. «Приближенное вычисление интегралов»]. Рассмотрим некоторые частные случаи.

Вначале попытаемся построить кубатурную формулу с тремя узлами, имеющую алгебраическую степень точности, равную двум. Эта формула будет иметь вид

$$I=\int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx A_0 f(x_0,y_0)+A_1 f(x_1,y_1)+A_2 f(x_2,y_2). \quad (2.12)$$

Учитывая сказанное выше, расположим узлы специальным образом: равномерно на окружности некоторого радиуса r с центром в начале координат. Тогда, очевидно, для

описания положения узлов удобно воспользоваться полярной системой координат. В итоге получим:



$$x_0 = r\cos\varphi,\quad y_0 = r\sin\varphi;$$
$$x_1 = r\cos\left(\varphi+\frac{2\pi}{3}\right),\quad y_1 = r\sin\left(\varphi+\frac{2\pi}{3}\right);$$
$$x_2 = r\cos\left(\varphi-\frac{2\pi}{3}\right),\quad y_2 = r\sin\left(\varphi-\frac{2\pi}{3}\right).$$

Как следствие, система уравнений для определения параметров кубатурной формулы (2.12) примет вид

$$\begin{cases} A_0 + A_1 + A_2 = 4, \\ r\left(A_0\cos\varphi + A_1\cos\left(\varphi+\frac{2\pi}{3}\right) + A_2\cos\left(\varphi-\frac{2\pi}{3}\right)\right) = 0, \\ r\left(A_0\sin\varphi + A_1\sin\left(\varphi+\frac{2\pi}{3}\right) + A_2\sin\left(\varphi-\frac{2\pi}{3}\right)\right) = 0, \\ r^2\left(A_0\cos^2\varphi + A_1\cos^2\left(\varphi+\frac{2\pi}{3}\right) + A_2\cos^2\left(\varphi-\frac{2\pi}{3}\right)\right) = \frac{4}{3}, \\ r^2\left(A_0\sin^2\varphi + A_1\sin^2\left(\varphi+\frac{2\pi}{3}\right) + A_2\sin^2\left(\varphi-\frac{2\pi}{3}\right)\right) = \frac{4}{3}, \\ r^2\left(A_0\cos\varphi\sin\varphi + A_1\cos\left(\varphi+\frac{2\pi}{3}\right)\sin\left(\varphi+\frac{2\pi}{3}\right) + A_2\cos\left(\varphi-\frac{2\pi}{3}\right)\sin\left(\varphi+\frac{2\pi}{3}\right)\right) = 0. \end{cases} \tag{2.13}$$

Складывая пятое и четвертое уравнения данной системы, получим:

$$r^2\left(A_0 + A_1 + A_2\right) = \frac{8}{3},$$

откуда, с учетом первого уравнения, найдем: $r = \sqrt{\frac{2}{3}}$.

После этого исключим из второго и третьего уравнений неизвестное A_0. Умножая второе уравнение на $\sin\varphi$, а третье – на $\cos\varphi$ и вычитая полученные уравнения друг из друга, получим:

$$A_1\left[\sin\varphi\cos\left(\varphi+\frac{2\pi}{3}\right) - \cos\varphi\sin\left(\varphi+\frac{2\pi}{3}\right)\right] + A_2\left[\sin\varphi\cos\left(\varphi-\frac{2\pi}{3}\right) - \cos\varphi\sin\left(\varphi-\frac{2\pi}{3}\right)\right] = 0$$

или

$$-A_1\sin\frac{2\pi}{3} + A_2\sin\frac{2\pi}{3} = 0,$$

откуда следует, что $A_1 = A_2$. Подставляя это равенство во второе уравнение (2.13), будем иметь:

$$A_0\cos\varphi + A_1\left[\cos\left(\varphi+\frac{2\pi}{3}\right) + \cos\left(\varphi-\frac{2\pi}{3}\right)\right] = 0.$$

Полученное уравнение равносильно уравнению

$$A_0 \cos\varphi + 2A_1 \cos\varphi \cos\frac{2\pi}{3} = 0,$$

т.е. $A_0 = A_1 = A_2$. Тогда из первого уравнения (2.13) имеем: $A_0 = A_1 = A_2 = \frac{4}{3}$.

После этого остается заметить, что в силу тригонометрических соотношений

$$\cos\alpha + \cos\left(\alpha + \frac{2\pi}{3}\right) + \cos\left(\alpha - \frac{2\pi}{3}\right) = 0,$$

$$\sin\alpha + \sin\left(\alpha + \frac{2\pi}{3}\right) + \sin\left(\alpha - \frac{2\pi}{3}\right) = 0$$

все уравнения системы (2.13) при найденных значениях A_0, A_1, A_2 и r обращаются в тождества при любых значениях аргумента φ.

Таким образом, получаем однопараметрическое семейство кубатурных формул с тремя узлами второго порядка точности:

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx$$

$$\approx \frac{4}{3}\left(f\left(\sqrt{\frac{2}{3}}\cos\varphi, \sqrt{\frac{2}{3}}\sin\varphi\right) + f\left(\sqrt{\frac{2}{3}}\cos\left(\varphi + \frac{2\pi}{3}\right), \sqrt{\frac{2}{3}}\sin\left(\varphi + \frac{2\pi}{3}\right)\right) + f\left(\sqrt{\frac{2}{3}}\cos\left(\varphi - \frac{2\pi}{3}\right), \sqrt{\frac{2}{3}}\sin\left(\varphi - \frac{2\pi}{3}\right)\right)\right).$$

Придавая φ конкретные значения, получим частные случаи кубатурных формул:

1) $\varphi = 0$:

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx \frac{4}{3}\left(f\left(\frac{\sqrt{6}}{3}, 0\right) + f\left(-\frac{\sqrt{6}}{6}, \frac{\sqrt{2}}{2}\right) + f\left(-\frac{\sqrt{6}}{6}, -\frac{\sqrt{2}}{2}\right)\right), \qquad (2.14)$$

2) $\varphi = \frac{\pi}{4}$:

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx \frac{4}{3}\left(f\left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}+3}{6}, \frac{3-\sqrt{3}}{6}\right) + f\left(-\frac{3-\sqrt{3}}{6}, -\frac{3+\sqrt{3}}{6}\right)\right). \qquad (2.15)$$

Аналогичное расположение четырех узлов приводит к кубатурной формуле, алгебраическая степень точности которой равна 3:

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx f\left(\sqrt{\frac{2}{3}}\cos\varphi, \sqrt{\frac{2}{3}}\sin\varphi\right) + f\left(-\sqrt{\frac{2}{3}}\sin\varphi, \sqrt{\frac{2}{3}}\cos\varphi\right) +$$

$$+ f\left(-\sqrt{\frac{2}{3}}\cos\varphi, \sqrt{\frac{2}{3}}\sin\varphi\right) + f\left(\sqrt{\frac{2}{3}}\sin\varphi, -\sqrt{\frac{2}{3}}\cos\varphi\right). \qquad (2.16)$$

Действительно, в этом случае координаты узлов будут иметь вид

$$x_0 = r\cos\varphi,\ \ y_0 = r\sin\varphi;$$
$$x_1 = r\cos\left(\varphi+\frac{\pi}{2}\right) = -r\sin\varphi,\ \ y_1 = r\sin\left(\varphi+\frac{\pi}{2}\right) = r\cos\varphi;$$
$$x_2 = r\cos(\varphi+\pi) = -\cos\varphi,\ \ y_2 = r\sin(\varphi+\pi) = -r\sin\varphi;$$
$$x_3 = r\cos\left(\varphi+\frac{3\pi}{2}\right) = r\sin\varphi,\ \ y_3 = r\sin\left(\varphi+\frac{3\pi}{2}\right) = -r\cos\varphi.$$

Система уравнений для определения параметров кубатурной формулы с учетом соотношений

$$\int_{-1}^{1}\int_{-1}^{1} x^i y^j dxdy = 0 \text{ , если } i+j=3 \text{ и } i\ge 0,\ j\ge 0$$

примет вид

$$\begin{cases}
A_0 + A_1 + A_2 + A_3 = 4, \\
r\left(A_0\cos\varphi - A_1\sin\varphi - A_2\cos\varphi + A_3\sin\varphi\right) = 0, \\
r\left(A_0\sin\varphi + A_1\cos\varphi - A_2\sin\varphi - A_3\cos\varphi\right) = 0, \\
r^2\left(A_0\cos^2\varphi + A_1\sin^2\varphi + A_2\cos^2\varphi + A_3\sin^2\varphi\right) = \frac{4}{3}, \\
r^2\left(A_0\sin^2\varphi + A_1\cos^2\varphi + A_2\sin^2\varphi + A_3\cos^2\varphi\right) = \frac{4}{3}, \\
r^2\left(A_0\cos\varphi\sin\varphi - A_1\cos\varphi\sin\varphi + A_2\cos\varphi\sin\varphi - A_3\cos\varphi\sin\varphi\right) = 0, \\
r^3\left(A_0\cos^3\varphi - A_1\sin^3\varphi - A_2\cos^3\varphi + A_3\sin^3\varphi\right) = 0, \\
r^3\left(A_0\sin^3\varphi + A_1\cos^3\varphi - A_2\sin^3\varphi - A_3\cos^3\varphi\right) = 0, \\
r^3\left(A_0\cos^2\varphi\sin\varphi + A_1\cos\varphi\sin^2\varphi - A_2\cos^2\varphi\sin\varphi - A_3\cos\varphi\sin^2\varphi\right) = 0, \\
r^3\left(A_0\cos\varphi\sin^2\varphi - A_1\cos^2\varphi\sin\varphi - A_2\cos\varphi\sin^2\varphi + A_3\cos^2\varphi\sin\varphi\right) = 0.
\end{cases} \qquad (*)$$

Как и ранее, складывая четвертое и пятое уравнения данной системы, найдем: $r = \sqrt{\frac{2}{3}}$. Умножая второе уравнение на $\sin\varphi$, а третье – на $\cos\varphi$ и вычитая, получим:

$$-A_1 + A_3 = 0 ,$$

т.е.

$$A_1 = A_3 .$$

Аналогично, умножая второе уравнение системы на $\cos\varphi$, а третье – на $\sin\varphi$ и складывая, получим: $A_0 = A_2$.

Подставляя полученные соотношения в шестое уравнение системы, перепишем его в виде

$$r^2\cos\varphi\sin\varphi\left(2A_0 - 2A_1\right) = 0 ,$$

откуда

$$A_0 = A_1 ,$$

и, таким образом,

$$A_0 = A_1 = A_2 = A_3 = 1 .$$

Теперь остается проверить, что все уравнения системы (*) обращаются в тождество независимо от величины φ .

Вновь, как и выше выпишем некоторые частные случаи формулы (2.16):

1) $\varphi = 0$:

$$I=\int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx f\left(\frac{\sqrt{6}}{3},0\right)+f\left(0,\frac{\sqrt{6}}{3}\right)+f\left(-\frac{\sqrt{6}}{3},0\right)+f\left(0,-\frac{\sqrt{6}}{3}\right), \qquad (2.17)$$

2) $\varphi=\dfrac{\pi}{4}$:

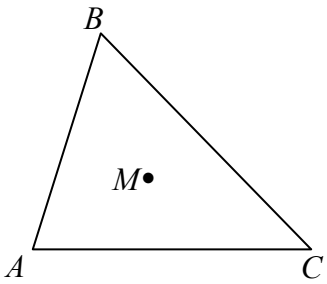
$$I=\int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx f\left(\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right)+f\left(-\frac{\sqrt{3}}{3},\frac{\sqrt{3}}{3}\right)+f\left(-\frac{\sqrt{3}}{3},-\frac{\sqrt{3}}{3}\right)+f\left(\frac{\sqrt{3}}{3},-\frac{\sqrt{3}}{3}\right). \qquad (2.18)$$

Замечание. Несмотря на то, что в полученных кубатурных формулах второго и третьего порядков точности параметр φ произволен, повысить степень точности за счет его выбора, как легко проверить, не удается.

Упражнение. Рассмотреть кубатурные формулы с пятью узлами, добавив к рассмотренным выше четырем узлам, равномерно расположенным на окружности, центр симметрии области.

2.2. Кубатурные формулы на треугольнике

Другим интересным типом области является треугольник. Он хорош не только тем, что для него можно явно (по координатам вершин) указать параметры соответствующей кубатурной формулы, но и как тип элементарной ячейки, которая может служить основой для построения составных кубатурных формул при интегрировании по произвольной области (при этом предполагается освоенным процесс ***триангуляции*** области).



Перейдем к рассмотрению простейших кубатурных формул, которые построим путем интерполирования подынтегральной функции с последующим интегрированием. Заметим, что площадь треугольника может быть вычислена по формуле

$$S=\frac{1}{2}|g(A,B,C)|, \text{ где } g(A,B,C)=\begin{vmatrix} x_A & y_A & 1 \\ x_B & y_B & 1 \\ x_C & y_C & 1 \end{vmatrix}, \qquad (2.19)$$

а координаты центра тяжести (это точка пересечения медиан) находятся по формуле

$$\bar{x}=x_M=\frac{x_A+x_B+x_C}{3};\ \ \bar{y}=y_M=\frac{y_A+y_B+y_C}{3}. \qquad (2.20)$$

Тогда (2.7), (2.19), (2.20) – кубатурная формула средних для треугольника. При этом обобщенную формулу (2.7) можно применять к областям, ограниченным ломаной линией.

Получим сейчас аналог формулы трапеций для треугольника. Выбирая в качестве узлов интерполирования вершины треугольника A,B и C и учитывая, что функция g из формулы (2.19) будет положительной, если расположение ее аргументов соответствует обходу вершин против часовой стрелки (например, $g(A,C,B)>0,\ g(A,B,C)<0$), заменим функцию $f(x,y)$ ее интерполяционным многочленом первой степени (см. (8.14) гл. IV), который запишем в виде (Q – произвольная точка плоскости, имеющая координаты (x,y))

$$P_1(x,y)=\frac{1}{g(A,C,B)}[f(A)g(Q,C,B)+f(B)g(A,C,Q)+f(C)g(A,Q,B)]. \qquad (2.21)$$

Так как $g(Q,C,B)$, $g(A,C,Q)$ и $g(A,Q,B)$ являются линейными функциями аргументов x и y, то для упрощения вычисления интегралов от них можно воспользоваться кубатурной формулой средних (2.7), (2.19), (2.20), которая имеет алгебраическую степень точности, равную единице, и, следовательно, в нашем случае будет давать точное значение каждого из интегралов. Поэтому (M – центр тяжести треугольника)

$$\iint\limits_{\Delta} f(x,y)dxdy \approx \iint\limits_{\Delta} P_1(x,y)dxdy =$$

$$= \frac{1}{g(A,C,B)} \iint\limits_{\Delta} [f(A)g(Q,C,B) + f(B)g(A,C,Q) + f(C)g(A,Q,B)]dxdy =$$

$$= \frac{1}{g(A,C,B)} \cdot S_{\Delta} \cdot [f(A)g(M,C,B) + f(B)g(A,C,M) + f(C)g(A,M,B)].$$

Теперь осталось заметить, что в силу (2.19) $g(A,C,B) = 2S_{\Delta}$, а по свойству медиан треугольника

$$g(M,C,B) = g(A,C,M) = g(A,M,B) = \frac{2}{3} S_{\Delta}.$$

Следовательно,

$$\iint\limits_{\Delta} f(x,y)dxdy \approx \frac{1}{2S_{\Delta}} \cdot S_{\Delta} \cdot [f(A) + f(B) + f(C)] \cdot \frac{2}{3} S_{\Delta} = \frac{S_{\Delta}}{3} [f(A) + f(B) + f(C)]. \qquad (2.22)$$

Как видим, здесь также нет принципиальных проблем, однако возникают значительные технические трудности при вычислении интегралов от многочленов по произвольному треугольнику. Эти трудности можно преодолеть, если от произвольного треугольника перейти к некоторому стандартному. Достичь этого можно, если ввести так называемые ***барицентрические*** (или ***симплексные***) координаты $(\lambda_1, \lambda_2, \lambda_3)$ произвольной точки (x, y) плоскости по формулам (A, B, C – вершины треугольника)

$$\begin{cases} \lambda_1 + \lambda_2 + \lambda_3 = 1, \\ x_A\lambda_1 + x_B\lambda_2 + x_C\lambda_3 = x, \\ y_A\lambda_1 + y_B\lambda_2 + y_C\lambda_3 = y. \end{cases} \qquad (*)$$

Отсюда следует, что если точки A, B, C не лежат на одной прямой (т.е. исходный треугольник не является вырожденным), то

$$\begin{vmatrix} 1 & 1 & 1 \\ x_A & x_B & x_C \\ y_A & y_B & y_C \end{vmatrix} = g(A,B,C) \neq 0$$

и, следовательно, барицентрические координаты $(\lambda_1, \lambda_2, \lambda_3)$ произвольной точки Q плоскости будут иметь вид

$$\lambda_1 = \frac{g(Q,B,C)}{g(A,B,C)},\ \ \lambda_2 = \frac{g(A,Q,C)}{g(A,B,C)},\ \ \lambda_3 = \frac{g(A,B,Q)}{g(A,B,C)}. \qquad (2.23)$$

Заметим, что введенные таким образом координаты обладают следующими свойствами:

1. Если точка Q лежит внутри треугольника, то $\lambda_i > 0,\ i = \overline{1,3}$;
2. Если точка Q лежит на стороне треугольника, то одна из координат равна нулю;
3. Вершины треугольника имеют координаты $A(1;0;0),\ B(0;1;0),\ C(0;0;1)$, центр тяжести – $M\left(\frac{1}{3};\frac{1}{3};\frac{1}{3}\right)$.

Первые три свойства означают, что исходный треугольник ABC общего положения взаимно однозначным образом отображается в плоскости, определяемой любыми двумя из трех переменных $\lambda_i > 0,\ i = \overline{1,3}$, в стандартный равнобедренный прямоугольный треугольник с единичными катетами и вершиной прямого угла, расположенного в начале координат. Доказательство их представляется достаточно очевидным и основывается на формулах (2.23).

4. $$\iint\limits_{\Delta} \lambda_1^{\alpha_1}(x,y)\cdot\lambda_2^{\alpha_2}(x,y)\cdot\lambda_3^{\alpha_3}(x,y)dxdy = S_{\Delta}\cdot\frac{\alpha_1!\cdot\alpha_2!\cdot\alpha_3!\cdot 2!}{(\alpha_1+\alpha_2+\alpha_3+2)!},\ \ \alpha_i \geq 0,\ \alpha_i \in Z,\ i = \overline{1,3}. \qquad (2.24)$$

Докажем (2.24). Опуская предположение $\alpha_i \in Z$, перейдем в интеграле к переменным λ_1, λ_2 (при этом помним, что $\lambda_3 = 1-\lambda_1-\lambda_2$ в силу первого из уравнений (*)). С учетом отмеченного второе и третье уравнения (*) примут вид

$$\begin{cases} x = x_C + (x_A - x_C)\lambda_1 + (x_B - x_C)\lambda_2, \\ y = y_C + (y_A - y_C)\lambda_1 + (y_B - y_C)\lambda_2. \end{cases}$$

Следовательно, для якобиана преобразования имеем:

$$J = \frac{D(x,y)}{D(\lambda_1,\lambda_2)} = \begin{vmatrix} \frac{\partial x}{\partial \lambda_1} & \frac{\partial x}{\partial \lambda_2} \\ \frac{\partial y}{\partial \lambda_1} & \frac{\partial y}{\partial \lambda_2} \end{vmatrix} = \begin{vmatrix} x_A - x_C & x_B - x_C \\ y_A - y_C & y_B - y_C \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ x_A & x_B & x_C \\ y_A & y_B & y_C \end{vmatrix} = g(A,B,C).$$

Таким образом, $|J| = 2!\cdot S_{\Delta}$, и в результате замены вычисляемый интеграл преобразуется в интеграл

$$I = S_{\Delta}\cdot 2!\cdot \iint\limits_{\substack{\lambda_1\geq 0,\ \lambda_2\geq 0 \\ \lambda_1+\lambda_2\leq 1}} \lambda_1^{\alpha_1}\cdot\lambda_2^{\alpha_2}\cdot(1-\lambda_1-\lambda_2)^{\alpha_3}\,d\lambda_1 d\lambda_2\ .$$

Переходя в последнем интеграле к повторному, запишем его в виде

$$I = S_{\Delta}\cdot 2!\cdot \iint\limits_{\substack{\lambda_1\geq 0,\ \lambda_2\geq 0 \\ \lambda_1+\lambda_2\leq 1}} \lambda_1^{\alpha_1}\cdot\lambda_2^{\alpha_2}\cdot(1-\lambda_1-\lambda_2)^{\alpha_3}\,d\lambda_1 d\lambda_2 = S_{\Delta}\cdot 2!\cdot \int\limits_0^1 \lambda_1^{\alpha_1}d\lambda_1 \int\limits_0^{1-\lambda_1} \lambda_2^{\alpha_2}(1-\lambda_1-\lambda_2)^{\alpha_3}\,d\lambda_2\ .$$

Сделав во внутреннем интеграле замену переменной по формуле $\lambda_2 = (1-\lambda_1)t$, в итоге получим:

$$I = S_\Delta \cdot 2! \cdot \int_0^1 \lambda_1^{\alpha_1} d\lambda_1 \int_0^{1-\lambda_1} \lambda_2^{\alpha_2} (1-\lambda_1-\lambda_2)^{\alpha_3} d\lambda_2 = S_\Delta \cdot 2! \cdot \int_0^1 \lambda_1^{\alpha_1} (1-\lambda_1)^{\alpha_1+\alpha_3+1} d\lambda_1 \int_0^1 t^{\alpha_2} (1-t)^{\alpha_3} dt =$$

$$= S_\Delta \cdot 2! \cdot \mathrm{B}(\alpha_1+1, \alpha_1+\alpha_3+2) \cdot \mathrm{B}(\alpha_2+1, \alpha_3+1) = S_\Delta \cdot 2! \cdot \frac{\Gamma(\alpha_1+1)\cdot\Gamma(\alpha_2+1)\cdot\Gamma(\alpha_3+1)}{\Gamma(\alpha_1+\alpha_2+\alpha_3+3)}.$$

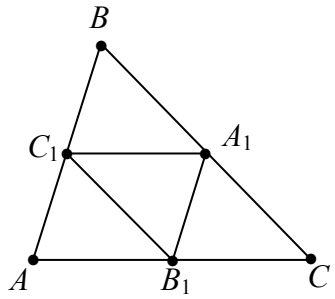
В случае целых значений α_i полученное значение интеграла совпадает с (2.24).

Использование барицентрических координат позволяет значительно упростить построение кубатурных формул. Действительно, в случае, рассмотренном выше, формула (2.21), дающая представление интерполяционного многочлена первой степени, примет вид

$$P_1(x,y) = P_1(\lambda_1,\lambda_2,\lambda_3) = \lambda_1 \cdot f(A) + \lambda_2 \cdot f(B) + \lambda_3 \cdot f(C).$$

После этого использование формул (2.24) при вычислении интегралов сразу же приводит к (2.22).

Построим сейчас кубатурную формулу более высокой алгебраической степени точности. Для этого заменим функцию $f(x,y)$ интерполяционным многочленом второй степени по узлам, расположенным в соответствии с теорией, изложенной в п. 8.4 главы IV: в вершинах треугольника и на серединах его сторон.



Согласно формулам (8.12), (8.13) (там же) (при этом функции $L_{k\,i}(Q)$ в (8.12) будем обозначать точками, через которые проходят соответствующие прямые) соответствующий интерполяционный многочлен примет вид

$$P_2(x,y) = P_2(\lambda_1,\lambda_2,\lambda_3) = \frac{BC(Q)}{BC(A)} \cdot \frac{B_1C_1(Q)}{B_1C_1(A)} \cdot f(A) + \frac{AC(Q)}{AC(B)} \cdot \frac{A_1C_1(Q)}{A_1C_1(A)} \cdot f(B) +$$

$$+ \frac{AB(Q)}{AB(C)} \cdot \frac{A_1B_1(Q)}{A_1B_1(C)} \cdot f(C) + \frac{AB(Q)}{AB(A_1)} \cdot \frac{AC(Q)}{AC(A_1)} \cdot f(A_1) + \qquad (2.25)$$

$$+ \frac{BC(Q)}{BC(B_1)} \cdot \frac{AB(Q)}{AB(B_1)} \cdot f(B_1) + \frac{BC(Q)}{BC(C_1)} \cdot \frac{AC(Q)}{AC(C_1)} \cdot f(C_1).$$

Уравнение прямой, проходящей через точки B и C, в барицентрических координатах имеет вид

$$\lambda_1 = 0.$$

Поэтому $BC(A) = 1;\ BC(B_1) = BC(C_1) = \frac{1}{2}$.

Аналогично

$$B_1C_1(Q) = \lambda_1 - \frac{1}{2} \text{ и } B_1C_1(A) = 1 - \frac{1}{2} = \frac{1}{2};$$

$$AC(Q) = \lambda_2 \text{ и } AC(B) = 1;\ AC(A_1) = AC(C_1) = \frac{1}{2};$$

$$A_1C_1(Q)=\lambda_2-\frac{1}{2}\ \text{и}\ A_1C_1(B)=1-\frac{1}{2}=\frac{1}{2};$$

$$AB(Q)=\lambda_3\ \text{и}\ AB(C)=1;\ AB\left(A_1\right)=AB\left(B_1\right)=\frac{1}{2};$$

$$A_1B_1(Q)=\lambda_3-\frac{1}{2}\ \text{и}\ A_1B_1(C)=1-\frac{1}{2}=\frac{1}{2}.$$

Учитывая эти равенства, (2.25) перепишем в виде

$$P_2\left(\lambda_1,\lambda_2,\lambda_3\right)=\lambda_1\left(2\lambda_1-1\right)\cdot f(A)+\lambda_2\left(2\lambda_2-1\right)\cdot f(B)+\lambda_3\left(2\lambda_3-1\right)\cdot f(C)+$$

$$+4\lambda_2\lambda_3\cdot f\left(A_1\right)+4\lambda_1\lambda_3\cdot f\left(B_1\right)+4\lambda_1\lambda_2\cdot f\left(C_1\right).$$

Остается проинтегрировать этот многочлен по треугольнику ABC. Используя (2.24), получим:

$$\iint\limits_{\Delta} f(x,y)dxdy\approx\frac{S_{\Delta}}{3}\left[f\left(A_1\right)+f\left(B_1\right)+f\left(C_1\right)\right]. \tag{2.26}$$

Используя формулы (2.24), легко показать, что данная кубатурная формула имеет алгебраическую степень точности, равную 2. Несмотря на то, что интерполирование велось по шести узлам, полученная кубатурная формула содержит всего три узла (и это количество узлов является минимально возможным для формул с алгебраической степенью точности, равной 2).

Замечание. Кубатурная формула с тремя узлами, обладающая алгебраической степенью точности, равной 2, не является единственной.

Действительно, несложно проверить, что в трехмерном (относительно переменных $\lambda_1,\lambda_2,\lambda_3$) пространстве узлы полученной кубатурной формулы лежат на сфере с центром в точке $M\left(\frac{1}{3};\frac{1}{3};\frac{1}{3}\right)$ (центре тяжести треугольника) радиуса $r=\frac{1}{\sqrt{6}}$. Используя гипотезу о симметрии (в частности, это означает, что для всех узлов, лежащих на сфере, коэффициенты кубатурной формулы одинаковы) можно непосредственно по определению алгебраической степени точности построить однопараметрическое семейство таких формул. Будем искать такую формулу в виде (опять-таки используя симплексные координаты)

$$\iint\limits_{\Delta} f(x,y)dxdy\approx S_{\Delta}\cdot\left[A_0f\left(\lambda_1^0,\lambda_2^0,\lambda_3^0\right)+A_1f\left(\lambda_1^1,\lambda_2^1,\lambda_3^1\right)+A_2f\left(\lambda_1^2,\lambda_2^2,\lambda_3^2\right)\right]. \tag{2.27}$$

Используя для проверки алгебраической степени точности многочлены $1,\lambda_1,\lambda_2,\lambda_1^2,\lambda_2^2,\lambda_1\lambda_2$ и формулы (2.24) для вычисления интегралов, получим для определения параметров формулы (2.27) систему уравнений

$$\begin{cases}A_0+A_1+A_2=1,\\ A_0\lambda_1^0+A_1\lambda_1^1+A_2\lambda_1^2=\frac{1}{3},\\ A_0\lambda_2^0+A_1\lambda_2^1+A_2\lambda_2^2=\frac{1}{3},\\ A_0\left(\lambda_1^0\right)^2+A_1\left(\lambda_1^1\right)^2+A_2\left(\lambda_1^2\right)^2=\frac{1}{6},\\ A_0\left(\lambda_2^0\right)^2+A_1\left(\lambda_2^1\right)^2+A_2\left(\lambda_2^2\right)^2=\frac{1}{6},\\ A_0\lambda_1^0\lambda_2^0+A_1\lambda_1^1\lambda_2^1+A_2\lambda_1^2\lambda_2^2=\frac{1}{12}.\end{cases} \tag{2.28}$$

Следуя гипотезе о равенстве коэффициентов, положим $A_0 = A_1 = A_2 = \frac{1}{3}$. Тогда оставшиеся уравнения перепишутся в виде

$$\begin{cases} \lambda_1^0 + \lambda_1^1 + \lambda_1^2 = 1, \\ \lambda_2^0 + \lambda_2^1 + \lambda_2^2 = 1, \\ \left(\lambda_1^0\right)^2 + \left(\lambda_1^1\right)^2 + \left(\lambda_1^2\right)^2 = \frac{1}{2}, \\ \left(\lambda_2^0\right)^2 + \left(\lambda_2^1\right)^2 + \left(\lambda_2^2\right)^2 = \frac{1}{2}, \\ \lambda_1^0\lambda_2^0 + \lambda_1^1\lambda_2^1 + \lambda_1^2\lambda_2^2 = \frac{1}{4}. \end{cases} \qquad (2.28')$$

Полагая в первом уравнении $\lambda_1^2 = \alpha$, выразим из него λ_1^1: $\lambda_1^1 = 1 - \alpha - \lambda_1^0$. Подставив это выражение в третье уравнение, получим:

$$\left(\lambda_1^0\right)^2 + \left(1 - \alpha - \lambda_1^0\right)^2 = \frac{1}{2} - \alpha^2,$$

откуда

$$\lambda_1^0 = \frac{1 - \alpha \pm \sqrt{2\alpha - 3\alpha^2}}{2}$$

при условии $0 \le \alpha \le \frac{2}{3}$.

Пусть для определенности $\lambda_1^0 = \frac{1 - \alpha + \sqrt{2\alpha - 3\alpha^2}}{2}$. Тогда с учетом введенных обозначений

$$\lambda_1^1 = \frac{1 - \alpha - \sqrt{2\alpha - 3\alpha^2}}{2}, \ \lambda_1^2 = \alpha.$$

Аналогично, используя второе и четвертое уравнения последней системы, найдем:

$$\lambda_2^0 = \frac{1 - \beta + \sqrt{2\beta - 3\beta^2}}{2}, \ \lambda_2^1 = \frac{1 - \beta - \sqrt{2\beta - 3\beta^2}}{2}, \ \lambda_2^2 = \beta, \ 0 \le \beta \le \frac{2}{3}.$$

Подставляя найденные выражения в последнее (пятое) уравнение системы, получим связь между параметрами α и β:

$$\frac{(1-\alpha)(1-\beta) + (1-\alpha)\sqrt{2\beta - 3\beta^2} + (1-\beta)\sqrt{2\alpha - 3\alpha^2} + \sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2}}{4} +$$

$$+ \frac{(1-\alpha)(1-\beta) - (1-\alpha)\sqrt{2\beta - 3\beta^2} - (1-\beta)\sqrt{2\alpha - 3\alpha^2} + \sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2}}{4} + \alpha\beta = \frac{1}{4}$$

или

$$2(1 - \alpha - \beta + \alpha\beta) + 2\sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2} + 4\alpha\beta = 1.$$

Уединяя радикал и приводя подобные, перепишем это уравнение в виде

$$2\sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2} = (2\alpha - 1) - 2(3\alpha - 1)\beta. \qquad (*)$$

Заметим также, что если выбрать в формулах для λ_1^0 и λ_2^0 (а значит, и для λ_1^1 и λ_2^1) знаки перед радикалами «в противофазе», то уравнение, связывающее α и β будет иметь вид

$$-2\sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2} = (2\alpha - 1) - 2(3\alpha - 1)\beta. \qquad (**)$$

Поскольку при избавлении от иррациональности уравнения (*) и (**) переходят в одно и то же уравнение:

$$4(2\alpha-3\alpha^2)(2\beta-3\beta^2)=(2\alpha-1)^2-4(2\alpha-1)(3\alpha-1)\beta+4(3\alpha-1)^2,$$

то это означает, что система (2.28′) имеет решения при любых отмеченных выше допустимых значениях переменных α и β.

Приводя подобные, получим квадратное уравнение относительно переменной β:

$$4\beta^2-4\beta(1-\alpha)+(2\alpha-1)^2=0.$$

Его корни –

$$\beta_1=\frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2},\ \beta_2=\frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2}.$$

Положим $\beta=\beta_1=\frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2}$. Тогда

$$2\beta-3\beta^2=\frac{4(1-\alpha)+4\sqrt{2\alpha-3\alpha^2}-3\left((1-\alpha)^2+2(1-\alpha)\sqrt{2\alpha-3\alpha^2}+2\alpha-3\alpha^2\right)}{4}=$$

$$=\frac{1-4\alpha+6\alpha^2-2(1-3\alpha)\sqrt{2\alpha-3\alpha^2}}{4}=\frac{\left(1-6\alpha+9\alpha^2\right)+2\alpha-3\alpha^2-2(1-3\alpha)\sqrt{2\alpha-3\alpha^2}}{4}=$$

$$=\left(\frac{1-3\alpha-\sqrt{2\alpha-3\alpha^2}}{2}\right)^2.$$

Поэтому (выбираем один из вариантов)

$$\lambda_2^0=\frac{1-\beta+\sqrt{2\beta-3\beta^2}}{2}=\frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}+1-3\alpha-\sqrt{2\alpha-3\alpha^2}}{4}=\frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2},$$

$$\lambda_2^1=\frac{1-\beta-\sqrt{2\beta-3\beta^2}}{2}=\frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}-\left(1-3\alpha-\sqrt{2\alpha-3\alpha^2}\right)}{4}=\alpha.$$

Таким образом, имеем следующее однопараметрическое семейство решений системы (2.28′) (в том, что надлежащая комбинация знаков подобрана удачно, убеждаемся непосредственной проверкой):

$$\begin{gathered}\lambda_1^0=\frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2};\ \lambda_1^1=\frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2};\ \lambda_1^2=\alpha;\\ \lambda_2^0=\frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2};\ \lambda_2^1=\alpha;\ \lambda_2^2=\frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2};\end{gathered}\qquad 0\le\alpha\le\frac{2}{3}. \tag{2.29}$$

Отсюда, в частности, полагая $\alpha=0$, находим:

$$\lambda_1^0=\frac{1}{2};\ \lambda_1^1=\frac{1}{2};\ \lambda_1^2=0;$$

$$\lambda_2^0=\frac{1}{2};\ \lambda_2^1=0;\ \lambda_2^2=\frac{1}{2}.$$

Таким образом, получаем формулу (2.26).

Аналогично, полагая $\alpha=\frac{2}{3}$, находим:

$$\lambda_1^0 = \frac{1}{6};\ \lambda_1^1 = \frac{1}{6};\ \lambda_1^2 = \frac{2}{3};$$

$$\lambda_2^0 = \frac{1}{6};\ \lambda_2^1 = \frac{2}{3};\ \lambda_2^2 = \frac{1}{6}.$$

В результате получаем кубатурную формулу

$$\iint\limits_{\Delta} f(x,y)dxdy \approx \frac{S_\Delta}{3}\left[f\left(\frac{1}{6};\frac{1}{6};\frac{2}{3}\right) + f\left(\frac{1}{6};\frac{2}{3};\frac{1}{6}\right) + f\left(\frac{2}{3};\frac{1}{6};\frac{1}{6}\right)\right]. \tag{2.30}$$

Заметим, что среди решений (2.29) существует бесконечно много с рациональными компонентами.

Несмотря на то, что, как показано выше, существует, по крайней мере, однопараметрическое семейство кубатурных формул с тремя узлами, обладающих алгебраической степенью точности, равной 2, среди них нет таких, степень точности которых была бы равной 3. В то же время добавление еще одного, четвертого, узла позволяет эту задачу решить.

Действительно, расположим три узла, как и выше, на рассмотренной там же сфере (при этом соответствующие коэффициенты кубатурной формулы будем считать, вновь следуя гипотезе о симметрии, равными), а в качестве четвертого узла рассмотрим центр указанной сферы. Таким образом, кубатурную формулу будем искать в виде

$$\iint\limits_{\Delta} f(x,y)dxdy \approx S_\Delta \cdot \left[A\left(f\left(\lambda_1^0,\lambda_2^0,\lambda_3^0\right) + f\left(\lambda_1^1,\lambda_2^1,\lambda_3^1\right) + f\left(\lambda_1^2,\lambda_2^2,\lambda_3^2\right)\right) + Bf\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)\right]. \tag{2.31}$$

Как и выше, для определения параметров формулы, добавив условия точности на многочленах $\lambda_1^3,\ \lambda_1^2\lambda_2,\ \lambda_1\lambda_2^2,\ \lambda_2^3$, получим систему уравнений

$$\begin{cases}
3A + B = 1, \\
A\left(\lambda_1^0 + \lambda_1^1 + \lambda_1^2\right) + \frac{1}{3}B = \frac{1}{3}, \\
A\left(\lambda_2^0 + \lambda_2^1 + \lambda_2^2\right) + \frac{1}{3}B = \frac{1}{3}, \\
A\left(\left(\lambda_1^0\right)^2 + \left(\lambda_1^1\right)^2 + \left(\lambda_1^2\right)^2\right) + \frac{1}{9}B = \frac{1}{6}, \\
A\left(\left(\lambda_2^0\right)^2 + \left(\lambda_2^1\right)^2 + \left(\lambda_2^2\right)^2\right) + \frac{1}{9}B = \frac{1}{6}, \\
A\left(\lambda_1^0\lambda_2^0 + \lambda_1^1\lambda_2^1 + \lambda_1^2\lambda_2^2\right) + \frac{1}{9}B = \frac{1}{12}, \\
A\left(\left(\lambda_1^0\right)^3 + \left(\lambda_1^1\right)^3 + \left(\lambda_1^2\right)^3\right) + \frac{1}{27}B = \frac{1}{10}, \\
A\left(\left(\lambda_1^0\right)^2\lambda_2^0 + \left(\lambda_1^1\right)^2\lambda_2^1 + \left(\lambda_1^2\right)^2\lambda_2^2\right) + \frac{1}{27}B = \frac{1}{30}, \\
A\left(\lambda_1^0\left(\lambda_2^0\right)^2 + \lambda_1^1\left(\lambda_2^1\right)^2 + \lambda_1^2\left(\lambda_2^2\right)^2\right) + \frac{1}{27}B = \frac{1}{30}, \\
A\left(\left(\lambda_2^0\right)^3 + \left(\lambda_2^1\right)^3 + \left(\lambda_2^2\right)^3\right) + \frac{1}{27}B = \frac{1}{10}.
\end{cases} \tag{2.32}$$

С учетом сделанного выше предположения о симметрии система будет переопределенной. Выразив из первого уравнения (2.32) B через A, рассмотрим отдельно две системы, первая из которых состоит из второго, четвертого и седьмого, а вторая – из третьего, пятого и десятого уравнений исходной. Неизвестными первой из них будут $\lambda_1^0,\ \lambda_1^1,\ \lambda_1^2$, а второй – $\lambda_2^0,\ \lambda_2^1,\ \lambda_2^2$. A и в том, и в другом случае считаем параметром. Поэтому указанные системы запишем в виде

$$\begin{cases} \lambda_1^0+\lambda_1^1+\lambda_1^2=1, \\ (\lambda_1^0)^2+(\lambda_1^1)^2+(\lambda_1^2)^2=\dfrac{1}{3}+\dfrac{1}{18A}, \\ (\lambda_1^0)^3+(\lambda_1^1)^3+(\lambda_1^2)^3=\dfrac{1}{9}+\dfrac{1}{270A} \end{cases} \quad \text{и} \quad \begin{cases} \lambda_2^0+\lambda_2^1+\lambda_2^2=1, \\ (\lambda_2^0)^2+(\lambda_2^1)^2+(\lambda_2^2)^2=\dfrac{1}{3}+\dfrac{1}{18A}, \\ (\lambda_2^0)^3+(\lambda_2^1)^3+(\lambda_2^2)^3=\dfrac{1}{9}+\dfrac{1}{270A}. \end{cases} \tag{2.33}$$

Обе эти системы как системы относительно указанных выше неизвестных имеют, очевидно, одно и то же множество решений (поскольку с точностью до обозначений неизвестных совпадают). Поэтому рассмотрим несколько подробнее первую из них. Переходя в ней к новым переменным, которые являются элементарными симметрическими многочленами от переменных $\lambda_1^0,\ \lambda_1^1,\ \lambda_1^2$ (т.е. $\sigma_1=\lambda_1^0+\lambda_1^1+\lambda_1^2$, $\sigma_2=\lambda_1^0\lambda_1^1+\lambda_1^0\lambda_1^2+\lambda_1^1\lambda_1^2$, $\sigma_3=\lambda_1^0\lambda_1^1\lambda_1^2$), получим:

$$\begin{cases} \sigma_1=1, \\ \sigma_1^2-2\sigma_2=\dfrac{1}{3}+\dfrac{1}{18A}, \\ \sigma_1^3-3\sigma_1\sigma_2+3\sigma_3=\dfrac{1}{9}+\dfrac{17}{270A}. \end{cases}$$

Отсюда $\sigma_1=1,\ \sigma_2=\dfrac{1}{3}-\dfrac{1}{36A},\ \sigma_3=\dfrac{1}{27}\left(1-\dfrac{11}{60A}\right)$. Таким образом, $\lambda_1^0,\ \lambda_1^1,\ \lambda_1^2$ являются корнями кубического уравнения

$$t^3-t^2+\left(\frac{1}{3}-\frac{1}{36A}\right)t-\frac{1}{27}\left(1-\frac{11}{60A}\right)=0 \tag{2.34}$$

(t может быть любым из неизвестных $\lambda_1^0,\ \lambda_1^1,\ \lambda_1^2$).

Как уже отмечалось выше, решения $\lambda_2^0,\ \lambda_2^1,\ \lambda_2^2$ второй из систем (2.33) также будут удовлетворять уравнению (2.34). В то же время понятно, что перестановка корней данного уравнения, определяющая значения $\lambda_1^0,\ \lambda_1^1,\ \lambda_1^2$, должна отличаться от аналогичной перестановки, определяющей $\lambda_2^0,\ \lambda_2^1,\ \lambda_2^2$ (т.е. эти решения должны быть отличными друг от друга), ибо в противном случае правая часть шестого из уравнений системы (2.32) будет совпадать с правой частью четвертого (или пятого) уравнения, а левая – от нее отличаться, и, таким образом, (2.32) окажется несовместной.

Таким образом, для того чтобы выяснить какими могут быть соответствующие решения систем (2.32), необходимо исследовать шестое, восьмое и девятое уравнения этой системы в следующих пяти случаях (указываем подстановки, верхняя из которых корни первой из систем (2.33), а нижняя – соответствующие им по номеру корни второй):

$$\begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^1 & \lambda_2^0 & \lambda_2^2 \end{pmatrix}, \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^1 & \lambda_2^2 & \lambda_2^0 \end{pmatrix}, \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^0 & \lambda_2^2 & \lambda_2^1 \end{pmatrix}, \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^2 & \lambda_2^0 & \lambda_2^1 \end{pmatrix}, \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^2 & \lambda_2^1 & \lambda_2^0 \end{pmatrix}.$$

Рассмотрим подробнее первый случай. Так как $\lambda_1^0=\lambda_2^1,\ \lambda_1^1=\lambda_2^0,\ \lambda_1^2=\lambda_2^2$, то система из шестого, восьмого и девятого уравнений (2.32) примет вид

$$\begin{cases} 2\lambda_1^0\lambda_1^1+(\lambda_1^2)^2=\dfrac{1}{3}-\dfrac{1}{36A}, \\ (\lambda_1^0)^2\lambda_1^1+\lambda_1^0(\lambda_1^1)^2+(\lambda_1^2)^3=\dfrac{1}{9}-\dfrac{1}{270A}, \\ \lambda_1^0(\lambda_1^1)^2+(\lambda_1^0)^2\lambda_1^1+(\lambda_1^2)^3=\dfrac{1}{9}-\dfrac{1}{270A}. \end{cases} \tag{2.35}$$

Как видим, второе и третье уравнения в данной системе совпадают. Поэтому в дальнейшем оставляем только одно из них. В то же время, правая часть первого уравнения совпадает с найденным ранее значением σ_2, т.е. справедливо соотношение, связывающее значения $\lambda_1^0,\ \lambda_1^1,\ \lambda_1^2$:

$$\lambda_1^0\lambda_1^1+\lambda_1^0\lambda_1^2+\lambda_1^1\lambda_1^2=2\lambda_1^0\lambda_1^1+\left(\lambda_1^2\right)^2$$

или

$$\left(\lambda_1^0-\lambda_1^2\right)\left(\lambda_1^2-\lambda_1^1\right)=0\,.$$

Таким образом, рассматриваемый случай распадается на два. Пусть вначале $\lambda_1^0=\lambda_1^2$. Тогда, во-первых, (2.35) примет вид

$$\begin{cases}2\lambda_1^0\lambda_1^1+\left(\lambda_1^0\right)^2=\dfrac{1}{3}-\dfrac{1}{36A},\\ \lambda_1^0\lambda_1^1\left(\lambda_1^0+\lambda_1^1\right)+\left(\lambda_1^0\right)^3=\dfrac{1}{9}-\dfrac{1}{270A}.\end{cases}\tag{2.36}$$

а во-вторых, уравнение (2.34) имеет двукратный корень, т.е. корень λ_1^0 является также корнем и производной многочлена, стоящего в левой части (2.34). Поэтому

$$3\left(\lambda_1^0\right)^2-2\lambda_1^0=-\left(\frac{1}{3}-\frac{1}{36A}\right).$$

Подставляя сюда вместо правой части левую часть первого из уравнений (2.36), имеем:

$$3\left(\lambda_1^0\right)^2-2\lambda_1^0=-2\lambda_1^0\lambda_1^1-\left(\lambda_1^0\right)^2,$$

откуда

$$\lambda_1^0=\frac{1-\lambda_1^1}{2}\,.$$

С учетом найденного соотношения (2.36) примет вид

$$\begin{cases}\lambda_1^1\left(1-\lambda_1^1\right)+\dfrac{\left(1-\lambda_1^1\right)^2}{4}=\dfrac{1}{3}-\dfrac{1}{36A},\\ \dfrac{\lambda_1^0\left(1-\lambda_1^1\right)\left(1+\lambda_1^1\right)}{4}+\dfrac{\left(1-\lambda_1^1\right)^3}{8}=\dfrac{1}{9}-\dfrac{1}{270A}\end{cases}$$

или

$$\begin{cases}\dfrac{\left(1-\lambda_1^1\right)\left(3\lambda_1^1+1\right)}{4}=\dfrac{1}{3}-\dfrac{1}{36A},\\ \dfrac{\left(1-\lambda_1^1\right)\left(3\left(\lambda_1^1\right)^2+1\right)}{8}=\dfrac{1}{9}-\dfrac{1}{270A}.\end{cases}$$

Умножая первое уравнение на $\frac{1}{15}$, а второе – на $-\frac{1}{2}$ и складывая, исключим неизвестное A:

$$\frac{\left(1-\lambda_1^1\right)\left(3\lambda_1^1+1\right)}{60}-\frac{\left(1-\lambda_1^1\right)\left(3\left(\lambda_1^1\right)^2+1\right)}{16}=-\frac{1}{30}\,.$$

Избавляясь от знаменателя, перепишем это уравнение в виде

$$\left(5\lambda_1^1-3\right)\left(3\lambda_1^1-1\right)^2=0\,.$$

Отсюда $\lambda_1^1=\frac{3}{5}$. Тогда $A=\frac{25}{48}$, $\lambda_1^0=\lambda_1^2=\frac{1}{5}$, $B=1-3A=-\frac{27}{48}$.

Таким образом, искомая кубатурная формула в данном случае имеет вид

$$\iint\limits_{\Delta}f(x,y)dxdy\approx\frac{S_\Delta}{48}\cdot\left[25\left(f\left(\frac{1}{5},\frac{3}{5},\frac{1}{5}\right)+f\left(\frac{3}{5},\frac{1}{5},\frac{1}{5}\right)+f\left(\frac{1}{5},\frac{1}{5},\frac{3}{5}\right)\right)-27f\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)\right].\tag{2.37}$$

Упражнения.

1. Исследовать оставшиеся случаи и показать, что с точностью до перестановки слагаемых кубатурные формулы будут иметь вид (2.37).
2. Исследовать возможность построения кубатурной формулы, обладающей алгебраической степенью точности 3, с четырьмя узлами, расположенными на окружности (например, с центром в точке $O\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$).

РАЗДЕЛ IV

Численное решение интегральных уравнений

Данный раздел является первым из разделов, посвященных численному решению задач, искомой величиной в которых является некоторая функция (функциональных уравнений).

В достаточно общем виде интегральное уравнение с одной неизвестной функцией может быть записано в форме

$$\Phi\left(x, u(x), \int_a^t f(x, s, u(s))ds \right) = 0, \tag{1}$$

где $u(x)$ – искомая функция, $t \in \{b, x\}$.

Если $t = b$, то говорят, что уравнение принадлежит виду Фредгольма, а если $t = x$, то виду Вольтерра.

В приложениях в столь общем виде интегральные уравнения встречаются достаточно редко. Гораздо чаще приходится иметь дело с линейными уравнениями.

В первом из случаев они будут иметь вид

$$A(x)u(x) + \int_a^b K(x, s)u(s)ds = F(x), \tag{2}$$

а во втором –

$$A(x)u(x) + \int_a^x K(x, s)u(s)ds = F(x). \tag{3}$$

Если коэффициент $A(x)$ тождественно равен нулю на отрезке $[a; b]$, то неизвестная функция $u(x)$ входит в соответствующее уравнение только под знаком интеграла. В этом случае уравнение носит название интегрального уравнения первого рода. Таким образом, имеем интегральное уравнение Фредгольма первого рода

$$\int_a^b K(x, s)u(s)ds = F(x) \tag{4}$$

и интегральное уравнение Вольтерра первого рода

$$\int_a^x K(x, s)u(s)ds = F(x). \tag{5}$$

Задача решения этих уравнений имеет принципиальные трудности, поскольку относится к разряду некорректных.

Если же $A(x) \neq 0$ для любого $x \in [a; b]$, то уравнения (2), (3) могут быть приведены соответственно к виду

$$u(x) - \lambda \int_a^b K(x, s)u(s)ds = f(x) \tag{6}$$

и

$$u(x)-\lambda\int_a^x K(x,s)u(s)ds=f(x). \qquad (7)$$

(6) – (7) – линейные интегральные уравнения Фредгольма и Вольтерра второго рода. Формально положив в (7) $K(x,s)\equiv 0$ при $s>x$, можно свести уравнение Вольтерра к уравнению Фредгольма, рассматривая его как частный случай последнего. Однако возникающая при этом разрывность ядра $K(x,s)$ при $x=s$ делает указанную процедуру нежелательной и, вообще говоря, теория интегральных уравнений Фредгольма и Вольтерра существенно отличаются, как и методы их численного решения, к изучению которых мы и переходим.

ГЛАВА VIII

Методы решения интегральных уравнений Фредгольма второго рода

§ 1. Метод механических квадратур

Рассмотрим интегральное уравнение Фредгольма второго рода

$$u(x)-\lambda\int_a^b K(x,s)u(s)ds=f(x). \qquad (1.1)$$

Выберем на отрезке $[a;\,b]$ $(n+1)$ точек $a\le x_0<x_1<\ldots<x_n\le b$ и заменим в уравнении (1.1) интеграл некоторой квадратурной суммой, причем точки x_k будут ее узлами. Тогда вместо (1.1) получим равенство

$$u(x)-\lambda\sum_{k=0}^{n}A_kK(x,x_k)u(x_k)=f(x)+\lambda\rho(x), \qquad (1.2)$$

где A_k – коэффициенты выбранной квадратурной формулы, а $\rho(x)$ – ее остаток. Соответствующие выражения для них мы получали ранее. Например, если в качестве квадратурной формулы используется составная формула средних прямоугольников на равномерной сетке, то

$$x_k=a+\left(k+\frac{1}{2}\right)h,\ k=\overline{0,n-1};\ h=\frac{b-a}{n},\ A_k=h,\ \rho(x)=\frac{h^2}{24}(b-a)\frac{\partial^2\left(K(x,\eta)u(\eta)\right)}{\partial s^2}.$$

Если теперь в равенстве (1.2) положить последовательно $x=x_i$, $i=\overline{0,n}$, то в результате получится система линейных алгебраических уравнений для нахождения точных значений решения в узлах:

$$u(x_i)-\lambda\sum_{k=0}^{n}A_kK(x_i,x_k)u(x_k)=f(x_i)+\lambda\rho(x_i),\ i=\overline{0,n}. \qquad (1.3)$$

Остаток $\rho(x)$ квадратурной формулы обычно мал по сравнению с самой величиной квадратурной суммы, поэтому, отбрасывая в (1.3) малые величины $\lambda\rho(x_i)$, получим систему линейных алгебраических уравнений

$$y_i - \lambda \sum_{k=0}^{n} A_k K_{ik} y_k = f_i, \ i = \overline{0,n} \tag{1.4}$$

(здесь использованы обозначения $y_i \approx u(x_i), \ K_{ik} = K(x_i, x_k), \ f_i = f(x_i)$).

Решения y_i системы (1.4) будут являться некоторыми приближениями к точным решениям задачи (1.1) в узлах $x_i, \ i = \overline{0,n}$ Найти их можно, используя для этих целей любой (или наиболее подходящий) из методов решения систем линейных алгебраических уравнений (например, метод Гаусса).

Замечание. Зная величины $y_i, \ i = \overline{0,n}$, можно легко восстановить приближенное решение во всех точках отрезка $[a; b]$. Для этих целей естественно воспользоваться формулой

$$y(x) = f(x) + \lambda \sum_{k=0}^{n} A_k K(x, x_k) y_k \ . \tag{1.5}$$

Как известно из теории систем линейных алгебраических уравнений, в случае, если определитель матрицы системы (1.4) отличен от нуля, т.е.

$$\Delta(\lambda) = \begin{vmatrix} 1 - \lambda A_0 K_{00} & -\lambda A_1 K_{01} & \dots & -\lambda A_n K_{0n} \\ -\lambda A_0 K_{10} & 1 - \lambda A_1 K_{11} & \dots & -\lambda A_n K_{1n} \\ & & \dots & \\ -\lambda A_0 K_{n0} & -\lambda A_1 K_{n1} & \dots & 1 - \lambda A_n K_{nn} \end{vmatrix} \neq 0 \, ,$$

то решение ее существует и единственно. Если же $\Delta(\lambda) = 0$, то значения λ, при которых это соотношение выполняется, будут приближениями к соответствующим собственным значениям ядра. Если для этих λ положить в (1.4) $f_i = 0, \ i = \overline{0,n}$, то отсюда можно найти и собственные функции ядра интегрального уравнения (1.1) (естественно, приближенно), т.е. метод механических квадратур может быть использован для решения проблемы собственных значений.

1.1. Оценка погрешности метода механических квадратур

Проведем исследование погрешности метода механических квадратур. Будем предполагать, что $\Delta(\lambda) \neq 0$. При численном решении системы (1.4) каждое уравнение удовлетворяется, вообще говоря, с некоторой погрешностью (округлений). Обозначим ее через $-\delta_i$. Тогда для величин $y_i, \ i = \overline{0,n}$ вместо (1.4) будут выполняться равенства

$$y_i - \lambda \sum_{k=0}^{n} A_k K_{ik} y_k = f_i - \delta_i, \ i = \overline{0,n} \, . \tag{1.6}$$

Отсюда, используя правило Крамера, можем записать:

$$y_i = \frac{1}{\Delta(\lambda)} \sum_{k=0}^{n} \Delta_{ki} (f_k - \delta_k), \ i = \overline{0,n} \, , \tag{1.7}$$

где Δ_{ki} – алгебраическое дополнение элемента $f_k - \delta_k$ определителя Δ_i.

Рассмотрим погрешность приближенного решения $\varepsilon_i = u(x_i) - y_i$ и получим для нее оценку. Ранее мы видели, что точные значения $u(x_i)$ удовлетворяют системе (1.3). Поэтому, аналогично (1.7), может записать:

$$u(x_i) = \frac{1}{\Delta(\lambda)} \sum_{k=0}^{n} \Delta_{ki} \left(f_k + \lambda \rho(x_k) \right), \; i = \overline{0,n}. \tag{1.8}$$

Вычитая из (1.8) (1.7), получим:

$$\varepsilon_i = \frac{1}{\Delta(\lambda)} \sum_{k=0}^{n} \Delta_{ki} \left[\lambda \rho(x_k) + \delta_k \right], \; i = \overline{0,n}. \tag{1.9}$$

Пусть теперь для погрешности $\rho(x)$ квадратурной формулы верна оценка $|\rho(x)| \le \rho = \rho(n)$ (ранее мы видели, что такая оценка всегда может быть получена и зависит от производных некоторого порядка интегрируемой функции). Предположим также, что для вычислительной погрешности также имеет место оценка $|\delta_i| \le \delta = \delta(n)$.

Тогда из (1.9) следует, что

$$|\varepsilon_i| \le \frac{|\lambda| \rho + \delta}{|\Delta(\lambda)|} \sum_{k=0}^{n} |\Delta_{ki}| = B\left(|\lambda| \rho + \delta \right), \tag{1.10}$$

где $B = \dfrac{\sum_{k=0}^{n} |\Delta_{ki}|}{|\Delta(\lambda)|}$.

Отметим, что значения $\Delta(\lambda)$ и Δ_{ki} могут быть найдены в процессе вычислений и поэтому B является вычислимой константой в отличие от ρ, способ нахождения которой еще предстоит указать.

Рассмотрим остаточный член $\rho(x)$ более подробно.

$$\begin{aligned}
\rho(x) &= \int_a^b K(x,s) u(s) ds - \sum_{k=0}^{n} A_k K(x,x_k) u(x_k) = \left[u(x) = f(x) + \lambda \int_a^b K(x,s) u(s) ds \right] = \\
&= \int_a^b K(x,s) \left[f(s) + \lambda \int_a^b K(s,t) u(t) dt \right] ds - \sum_{k=0}^{n} A_k K(x,x_k) \left[f(x_k) + \lambda \int_a^b K(x_k,t) u(t) dt \right] = \\
&= \int_a^b K(x,s) f(s) ds - \sum_{k=0}^{n} K(x,x_k) f(x_k) + \lambda \int_a^b \left[\int_a^b K(x,s) K(s,t) ds - \sum_{k=0}^{n} A_k K(x,x_k) K(x_k,t) \right] u(t) dt = \\
&= \rho_f(x) + \lambda \int_a^b \rho_K(x,t) u(t) dt,
\end{aligned} \tag{1.11}$$

где

$$\rho_f(x) = \int_a^b K(x,s) f(s) ds - \sum_{k=0}^{n} A_k K(x,x_k) f(x_k)$$

$$\rho_K(x,t) = \int_a^b K(x,s)K(s,t)ds - \sum_{k=0}^{n} A_k K(x,x_k)K(x_k,t).$$

Так как $\rho_f(x)$ и $\rho_K(x,t)$ могут быть вычислены, поскольку $f(x)$ и $K(x,s)$ известны, то из (1.11) получим:

$$|\rho(x_i)| \le \rho \le \max_{x\in[a;b]}|\rho(x)| \le \rho_f + |\lambda|(b-a)\rho_K H, \qquad (1.12)$$

где $H = \max_{x\in[a;\,b]}|u(x)|$.

Подставляя (1.12) в (1.10), найдем:

$$|\varepsilon_i| \le B\left[|\lambda|\left(\rho_f + |\lambda|(b-a)\rho_K \mathrm{H}\right) + \delta\right],\ i = \overline{0,n}. \qquad (1.13)$$

Оценка (1.13) позволяет во многих случаях сделать заключение о сходимости вычислительного процесса. Однако ее недостатком является наличие величины H – максимума модуля неизвестной функции.

Оценим H через вычислимые величины. Определив приближенное решение во всех точках отрезка $[a;\,b]$ по формуле (1.5), для погрешности $\varepsilon(x)$ получим:

$$\varepsilon(x) = u(x) - y(x) = \lambda\left[\sum_{k=0}^{n} A_k K(x,x_k)\varepsilon_k + \rho(x)\right].$$

Отсюда, используя (1.12), (1.13), будем иметь (в предположении, что использованная квадратурная формула имеет положительные коэффициенты):

$$|\varepsilon(x)| \le |\lambda|\left[\left|\sum_{k=0}^{n} A_k K(x,x_k)\varepsilon_k\right| + |\rho(x)|\right] \le |\lambda|\left[\max_{0\le k\le n}|\varepsilon_k| \cdot \max_{(x,s)}|K(x,s)| \cdot \left|\sum_{k=0}^{n} A_k\right| + |\rho(x)|\right] \le$$

$$\le |\lambda|\left\{(b-a)MB\left[|\lambda|\left(\rho_f + |\lambda|(b-a)\rho_K H\right) + \delta\right] + \rho_f + |\lambda|(b-a)\rho_K H\right\}.$$

Здесь использовано обозначение $M = \max_{(x,s)}|K(x,s)|$. Пусть теперь $\widetilde{H} = \max_{x\in[a;b]}|y(x)|$. Эта величина вычислима. Так как $u(x) = y(x) + \varepsilon(x)$, то отсюда $H \le \widetilde{H} + \max_{x\in[a;b]}|\varepsilon(x)|$, и, применяя оценку для $|\varepsilon(x)|$, имеем:

$$H \le \widetilde{H} + |\lambda|\left\{(b-a)MB\left[|\lambda|\left(\rho_f + |\lambda|(b-a)\rho_K H\right) + \delta\right] + \rho_f + |\lambda|(b-a)\rho_K H\right\}$$

или

$$H\left[1 - |\lambda|^3(b-a)^2 MB\rho_K - |\lambda|^2(b-a)\rho_K\right] \le \widetilde{H} + |\lambda|\left[\rho_f + (b-a)MB\left(|\lambda|\rho_f + \delta\right)\right].$$

Отсюда при выполнении условия

$$1 - |\lambda|^2(b-a)\rho_K\left[|\lambda|(b-a)MB + 1\right] > 0$$

вытекает неравенство для H:

$$H \le \frac{\widetilde{H} + |\lambda|\left[\rho_f + (b-a)MB\left(|\lambda|\rho_f + \delta\right)\right]}{1 - |\lambda|^2 (b-a)\rho_K \left[|\lambda|(b-a)MB + 1\right]}. \qquad (1.14)$$

Таким образом, совместно оценки (1.13), (1.14) являются вычислимыми.

Учитывая приведенные оценки, можно сформулировать следующие соображения, касающиеся выбора квадратурной формулы для замены интеграла в (1.1): естественно выбирать такую квадратурную формулу, чтобы ее остаток $\rho(x)$ был по возможности малым. Этого можно достичь двумя путями:

1. за счет увеличения числа узлов;
2. за счет повышения алгебраической степени точности квадратурной формулы.

При этом следует учитывать, что увеличение числа узлов ведет к увеличению объема работы при решении системы (1.4), увеличение же алгебраической степени точности квадратурной формулы даст эффект лишь тогда, когда интегрируемые функции обладают достаточными свойствами гладкости (в первую очередь это касается ядра $K(x,s)$).

Заметим также, что приведенные выше оценки погрешности, хоть и являются вычислимыми, однако требуют очень серьезной аналитической работы. Поэтому практически представление о погрешности можно составить, используя расчеты на вложенных сетках (аналог правила Рунге).

§ 2. Метод замены ядра на вырожденное

Определение. Ядро $K(x,s)$ называется вырожденным, если оно может быть представлено в виде

$$K(x,s) = \sum_{i=0}^{n} \alpha_i(x)\beta_i(s). \qquad (2.1)$$

Системы $\alpha_i(x)$ и $\beta_i(s)$ $\left(i = \overline{0,n}\right)$ в (2.1) естественно считать линейно независимыми, так как в противном случае число слагаемых в (2.1) можно было бы уменьшить.

Примеры:

1. $K(x,s) = \mathrm{e}^{x+s} = \mathrm{e}^x \cdot \mathrm{e}^s = \alpha_0(x)\cdot\beta_0(s)$;
2. $K(x,s) = \sin(x+s) = \sin x \cdot \cos s + \cos x \cdot \sin s = \alpha_0(x)\cdot\beta_0(s) + \alpha_1(x)\cdot\beta_0(s)$.

Для вырожденных ядер уравнение Фредгольма второго рода (1.1) решается в аналитическом виде за конечное число действий.

Действительно, перепишем (1.1) в виде

$$u(x) = \lambda\int_a^b K(x,s)u(s)ds + f(x) = \lambda\int_a^b\left[\sum_{i=0}^{n}\alpha_i(x)\beta_i(s)\right]u(s)ds + f(x) =$$

$$= \lambda\sum_{i=0}^{n}\alpha_i(x)\int_a^b \beta_i(s)u(s)ds + f(x) = f(x) + \lambda\sum_{i=0}^{n} C_i\alpha_i(x),$$

т.е. фактически мы нашли вид точного решения

$$u(x) = f(x) + \lambda \sum_{i=0}^{n} C_i \alpha_i(x), \tag{2.2}$$

где

$$C_i = \int_a^b \beta_i(s) u(s) ds, \quad i = \overline{0,n}. \tag{2.3}$$

Чтобы найти C_i, подставим (2.2) в (2.3):

$$C_i = \int_a^b \beta_i(s) \left[f(s) + \lambda \sum_{j=0}^{n} C_j \alpha_j(s) \right] ds, \quad i = \overline{0,n}$$

или

$$C_i - \lambda \sum_{j=0}^{n} C_j a_{ij} = b_i, \ i = \overline{0,n}, \tag{2.4}$$

где

$$b_i = \int_a^b \beta_i(s) f(s) ds; \quad a_{ij} = \int_a^b \beta_i(s) \alpha_j(s) ds, \ i = \overline{0,n}; \ j = \overline{0,n}. \tag{2.5}$$

Таким образом, для определения коэффициентов C_i формулы (2.2) получаем систему линейных алгебраических уравнений. Если определитель ее

$$\Delta(\lambda) = \begin{vmatrix} 1-\lambda a_{00} & -\lambda a_{01} & \dots & -\lambda a_{0n} \\ -\lambda a_{10} & 1-\lambda a_{11} & \dots & -\lambda a_{1n} \\ & & \dots & \\ -\lambda a_{n0} & -\lambda a_{n1} & \dots & 1-\lambda a_{nn} \end{vmatrix}$$

отличен от нуля, то мы найдем единственным образом набор констант C_i и, следовательно, построим точное решение $u(x)$. Случай $\Delta(\lambda) = 0$ соответствует собственным значениям.

Изложенное выше позволяет указать метод приближенного решения интегрального уравнения (1.1), основной идеей которого является замена ядра исходного интегрального уравнения $K(x,s)$ близким к нему вырожденным ядром $\widetilde{K}(x,s)$ и последующее решение этого уравнения с вырожденным ядром изложенным выше способом.

Укажем несколько способов такой замены:

1. Разложение ядра в ряд Тейлора:
   а) если ядро $K(x,s)$ обладает достаточной гладкостью по переменной x на отрезке $[a; b]$, то в качестве вырожденного ядра можно взять соответствующей длины отрезок ряда Тейлора по x:

$$\widetilde{K}(x,s) = \sum_{i=0}^{n} \frac{(x-x_0)^i}{i!} \frac{\partial^i K(x_0,s)}{\partial x^i}, \tag{2.6}$$

   где x_0 – некоторая точка из отрезка $[a; b]$ (ее выбор может быть подчинен, например, требованию минимизации остатка ряда).

   Очевидно, в данном случае ядро имеет вид (2.1), в котором

$$\alpha_i(x) = (x - x_0)^i, \quad \beta_i(s) = \frac{1}{i!}\frac{\partial^i K(x_0, s)}{\partial x^i}.$$

б) если ядро $K(x,s)$ достаточно гладкое по переменной s на $[a;b]$, то аналогично можем применить разложение в ряд Тейлора по переменной s:

$$\widetilde{K}(x,s) = \sum_{i=0}^{m} \frac{(s - s_0)^i}{i!}\frac{\partial^i K(x, s_0)}{\partial s^i}. \tag{2.7}$$

При таком способе замены

$$\alpha_i(x) = \frac{1}{i!}\frac{\partial^i K(x, s_0)}{\partial s^i}, \quad \beta_i(s) = (s - s_0)^i.$$

в) для построения вырожденного ядра можно также использовать конечный отрезок двойного ряда Тейлора:

$$\widetilde{K}(x,s) = \sum_{i=0}^{n} \frac{1}{i!}\left[(x - x_0)\frac{\partial}{\partial x} + (s - s_0)\frac{\partial}{\partial s}\right]^i K(x_0, s_0), \quad (x_0,\ s_0 \in [a;b]). \tag{2.8}$$

В последнем случае, очевидно, нам всегда гарантировано точное вычисление интегралов во второй из формул (2.5).

2. Использование ортогональных разложений.

Рассмотрим этот прием на примере применения ряда Фурье. Известно, что непрерывная на отрезке $[-l;l]$ функция допускает разложение в ряд Фурье (например, по косинусам, если она четна). Поэтому можно положить (в предположении, что ядро $K(x,s)$ непрерывно по x и четно по этой же переменной)

$$\widetilde{K}(x,s) = \frac{1}{2}a_0(s) + \sum_{i=1}^{n} a_i(s)\cos\frac{i\pi x}{l}, \tag{2.9}$$

где

$$a_i(s) = \frac{2}{l}\int_0^l K(x,s)\cos\frac{i\pi x}{l}\,dx, \quad i = \overline{0,n}.$$

При этом следует иметь в виду, что исходный отрезок $[a;b]$ линейной заменой может быть превращен в нужный.

Аналогичные формулы можно записать, если использовать отрезки рядов Фурье по переменной s, либо по обеим переменным.

3. Интерполяционные способы замены ядра.

Рассмотрим для примера применение алгебраического интерполирования по значениям функции. Выбрав на отрезке $[a;b]$ $(n+1)$ точек $a \le x_0 < x_1 < \ldots < x_n \le b$ (узлов интерполирования), можем записать:

$$\widetilde{K}(x,s) = \sum_{i=0}^{n} \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)} K(x_i, s). \tag{2.10}$$

Аналогично может быть использована замена ядра интерполяционным многочленом по переменной s:

$$\widetilde{K}(x,s)=\sum_{j=0}^{m}\frac{\omega_{m+1}(s)}{(s-s_j)\omega'_{m+1}(s_j)}K(x,s_j). \qquad (2.11)$$

Иногда также целесообразным бывает использование интерполирования по обеим переменным. Так, например, использование повторного интерполирования приводит к формуле

$$\widetilde{K}(x,s)=\sum_{i=0}^{n}\sum_{j=0}^{m}\frac{\omega_{n+1}(x)\omega_{m+1}(s)}{(x-x_i)(s-s_j)\omega'_{n+1}(x_i)\omega'_{m+1}(s_j)}K(x_i,s_j). \qquad (2.12)$$

При этом, очевидно, вместо представлений Лагранжа могут быть использованы и другие.

4. Способ Бэтмена (предложен в 1922 году).

Аппроксимирующее ядро $\widetilde{K}(x,s)$ предлагается определять с помощью равенства

$$\begin{vmatrix} \widetilde{K}(x,s) & K(x,s_0) & \dots & K(x,s_n) \\ K(x_0,s) & K(x_0,s_0) & \dots & K(x_0,s_n) \\ & & \dots & \\ K(x_n,s_0) & K(x_n,s_0) & \dots & K(x_n,s_n) \end{vmatrix}=0,$$

где $x_0,\dots,x_n,\ s_0,\dots,s_n$ – некоторые точки из отрезка $[a;b]$. Представляя элементы первого столбца записанного определителя в виде

$$\widetilde{K}(x,s)+0,\ 0+K(x_0,s),\dots,0+K(x_n,s)$$

и разлагая определитель в сумму двух определителей, получим:

$$\widetilde{K}(x,s)=-\frac{1}{\Delta}\begin{vmatrix} 0 & K(x,s_0) & \dots & K(x,s_n) \\ K(x_0,s) & K(x_0,s_0) & \dots & K(x_0,s_n) \\ & & \dots & \\ K(x_n,s_0) & K(x_n,s_0) & \dots & K(x_n,s_n) \end{vmatrix}, \qquad (2.13)$$

где

$$\Delta=\begin{vmatrix} K(x_0,s_0) & \dots & K(x_0,s_n) \\ & \dots & \\ K(x_n,s_0) & \dots & K(x_n,s_n) \end{vmatrix}. \qquad (2.14)$$

Упражнение. Показать, что в вырожденное ядро, определяемое формулами (2.13), (2.14), удовлетворяет условиям $\widetilde{K}(x_i,s)=K(x_i,s);\ \ \widetilde{K}(x,s_j)=K(x,s_j)$.

§ 3. Метод последовательных приближений

Вновь рассмотрим интегральное уравнение Фредгольма второго рода (1.1). Будем искать его решение в виде степенного ряда

$$u(x)=\sum_{i=0}^{\infty}\lambda^i\varphi_i(x), \tag{3.1}$$

где λ – числовой параметр из уравнения (1.1), а функции $\varphi_i(x)$ подлежат определению.

Подставим ряд (3.1) в исходное интегральное уравнение (1.1).

$$\sum_{i=0}^{\infty}\lambda^i\varphi_i(x)-\lambda\int_a^b K(x,s)\sum_{i=0}^{\infty}\lambda^i\varphi_i(s)ds=f(x).$$

Меняя порядок суммирования и интегрирования (в предположении, что ряд (3.1) сходится) и приравнивая коэффициенты при одинаковых степенях λ, получим рекуррентные соотношения, позволяющие последовательно находить (быть может, приближенно) функциональные коэффициенты ряда (3.1):

$$\begin{cases}\varphi_0(x)=f(x),\\ \varphi_i(x)=\int_a^b K(x,s)\varphi_{i-1}(s)ds, \quad i=1,\,2,\ldots\end{cases} \tag{3.2}$$

Таким образом, алгоритм построения последовательности приближений определен. Исследуем его сходимость.

Пусть в области $R=[a;b]\times[a;b]$ выполняется неравенство $|K(x,s)|\le M$, и на отрезке $[a;b]$ – неравенство $|f(x)|\le N$.

Тогда из формул (3.2) последовательно получим:

$$|\varphi_0(x)|=|f(x)|\le N,$$

$$|\varphi_1(x)|=\left|\int_a^b K(x,s)\varphi_0(s)ds\right|\le\int_a^b|K(x,s)||\varphi_0(s)|ds\le NM(b-a),$$

$$|\varphi_2(x)|=\left|\int_a^b K(x,s)\varphi_1(s)ds\right|\le\int_a^b|K(x,s)||\varphi_1(s)|ds\le NM^2(b-a)^2,$$

$$\ldots\ldots$$

$$|\varphi_i(x)|\le NM^i(b-a)^i,\ i=0,\,1,\,2,\ldots$$

Учитывая полученные оценки, видим, что ряд (3.1) мажорируется числовым рядом $N\sum_{i=0}^{\infty}\left(|\lambda|M(b-a)\right)^i$, представляющим собой геометрическую прогрессию, и, следовательно, сходящимся при выполнении условия

$$|\lambda|M(b-a)<1. \tag{3.3}$$

Таким образом, если параметры исходного интегрального уравнения будут удовлетворять условию (3.3), то ряд (3.1) равномерно сходится на отрезке $[a; b]$. Тогда в качестве приближенного решения можно взять

$$y(x) = y_n(x) = \sum_{i=0}^{n} \lambda^i \varphi_i(x). \tag{3.4}$$

Оценим погрешность такого решения (в предположении, что все интегралы в (3.2) вычисляются точно). Имеем:

$$|\varepsilon_n(x)| = |u(x) - y_n(x)| = \left| \sum_{i=n+1}^{\infty} \lambda^i \varphi_i(x) \right| \le N|\lambda|^{n+1} M^{n+1} (b-a)^{n+1} (1 + |\lambda| M(b-a) + \cdots) =$$

$$= N[|\lambda| M(b-a)]^{n+1} \frac{1}{1 - |\lambda| M(b-a)}, \quad x \in [a; b]. \tag{3.5}$$

Из этой оценки следует равномерная сходимость $\varepsilon_n(x)$ к нулю. Отсюда видно также, что $y_n(x) \underset{n\to\infty}{\to} u(x)$ по крайней мере со скоростью геометрической прогрессии.

Заметим, однако, что все нужные интегралы в (3.2), как правило, вычисляются приближенно, поэтому оценка истинной погрешности будет несколько отличаться от полученной.

Следует также иметь в виду, что метод последовательных приближений может употребляться и в другой форме, несколько более удобной с точки зрения машинной реализации. Действительно, перепишем (1.1) в виде

$$u(x) = \lambda \int_a^b K(x, s) u(s) ds + f(x).$$

Получим (см. также «Метод простой итерации») вид, удобный для итерации. Выбирая в качестве начального приближения произвольную функцию $y_0(x)$, построим последовательность приближений

$$y_{n+1}(x) = \lambda \int_a^b K(x, s) y_n(s) ds + f(x), \quad n = 0, 1, \ldots, \tag{3.6}$$

которая при $y_0(x) \equiv 0$ будет полностью совпадать с (3.2), (3.4).

В то же время, запись метода последовательных приближений в форме (3.6) позволяет трактовать условие сходимости (3.5) как условие сжимаемости отображения $\varphi(u) = \lambda \int_a^b K(x, s) u(s) ds$. С другой стороны, процедура исследования сходимости метода последовательных приближений в форме (3.2), (3.4) говорит о том, что условие (3.5) является достаточным.

§ 4. Методы решения интегральных уравнений Вольтерра второго рода

4.1. Метод механических квадратур

Как мы уже отмечали, чисто формально уравнение Вольтерра можно считать частным случаем интегрального уравнения Фредгольма, у которого $K(x,s) \equiv 0$ при $s > x$. Поэтому алгоритм метода механических квадратур, рассмотренный нами выше, может рассматриваться и как алгоритм решения интегрального уравнения Вольтерра второго рода. Однако на таком пути мы встретимся с теоретическими трудностями при оценке погрешности, так как при таком подходе ядро $\widetilde{K}(x,s)$ оказывается разрывной функцией и для величины ρ_K не может быть получено сколько-нибудь удовлетворительных оценок.

Поэтому повторим вывод алгоритма, несколько его видоизменив.

Итак, рассмотрим интегральное уравнение Вольтерра второго рода

$$u(x) - \lambda \int_a^x K(x,s)u(s)ds = f(x),\ x \in [a;\, b]. \tag{4.1}$$

Выберем на отрезке $[a;\, b]$ $(n+1)$ точек $a \le x_0 < x_1 < \ldots < x_n \le b$ и рассмотрим уравнение (4.1) в этих точках:

$$u(x_i) - \lambda \int_a^{x_i} K(x_i,s)u(s)ds = f(x_i),\ i = \overline{0,n}. \tag{4.2}$$

Интеграл в (4.2) представляет собой интеграл с постоянными (!) пределами. Заменим его квадратурной суммой, использующей значения подынтегральной функции в точках $x_0,\, x_1,\, \ldots,\, x_i$:

$$u(x_i) - \lambda \sum_{k=0}^{i} A_K^{(i)} K(x_i, x_k)u(x_k) = f(x_i) + \lambda \rho^{(i)}(x_i). \tag{4.3}$$

Здесь $A_k^{(i)}$ – коэффициенты выбранной квадратурной формулы (принципиально для каждого значения i возможно использование своей квадратурной формулы), а $\rho^{(i)}(x)$ – ее остаток.

Отбрасывая в (4.3) остаточный член, получим систему линейных алгебраических уравнений для определения приближенных значений решения интегрального уравнения (4.1) в узлах выбранной сетки (используем те же обозначения, что и в § 1 предыдущей главы):

$$y_i - \lambda \sum_{k=0}^{i} K_{ik} y_k = f_i,\ i = \overline{0,n},$$

или в развернутом виде

$$\begin{cases} \left(1 - \lambda A_0^{(0)} K_{00}\right) y_0 = f_0, \\ -\lambda A_0^{(1)} K_{10} y_0 + \left(1 - \lambda A_1^{(1)} K_{11}\right) y_1 = f_1, \\ \qquad \ldots \\ -\lambda A_0^{(n)} K_{n0} y_0 - \cdots - \lambda A_{n-1}^{(n)} K_{nn-1} y_{n-1} + \left(1 - \lambda A_n^{(n)} K_{nn}\right) y_n = f_n. \end{cases} \tag{1.4}$$

Матрица данной системы является нижней треугольной. Поэтому решение системы (1.4) по сути представляет собой обратный ход метода Гаусса.

4.2. Метод последовательных приближений

Аналогично описанному выше (см. § 3) решение уравнения (4.1) может быть получено в виде ряда

$$u(x)=\sum_{i=0}^{\infty}\lambda^i\varphi_i(x), \tag{1.5}$$

где $\varphi_i(x)$ на сей раз определяются по формулам

$$\begin{cases}\varphi_0(x)=f(x),\\ \varphi_i(x)=\int\limits_a^x K(x,s)\varphi_{i-1}(s)ds, \quad i=1,\,2,\ldots\end{cases} \tag{1.6}$$

Таким образом, до сих пор все фактически совпадает с алгоритмом метода последовательных приближений для интегральных уравнений Фредгольма второго рода.

Исследуем сходимость полученного алгоритма, используя для этих целей тот же прием построения мажоранты, что и выше.

Если, как и ранее, предположить, что в области $R=[a;b]\times[a;b]$ выполняется неравенство $|K(x,s)|\le M$, и на отрезке $[a;b]$ – неравенство $|f(x)|\le N$, то

$$|\varphi_0(x)|=|f(x)|\le N,\ \ a\le x\le b,$$

$$|\varphi_1(x)|=\left|\int\limits_a^x K(x,s)\varphi_0(s)ds\right|\le\int\limits_a^x|K(x,s)||\varphi_0(s)|ds\le NM(x-a),\ \ a\le x\le b,$$

$$|\varphi_2(x)|=\left|\int\limits_a^x K(x,s)\varphi_1(s)ds\right|\le\int\limits_a^x|K(x,s)||\varphi_1(s)|ds\le NM^2\int\limits_a^x(s-a)ds\le NM^2\frac{(x-a)^2}{2!},\ a\le x\le b,$$

……

$$|\varphi_i(x)|\le NM^i\frac{(x-a)^i}{i!},\ \ a\le x\le b,\ \ i=0,\,1,\,2,\ldots$$

Таким образом, ряд (1.5) в случае интегрального уравнения Вольтерра второго рода будет мажорироваться степенным рядом $N\sum\limits_{i=0}^{\infty}\frac{\left[|\lambda|M(x-a)\right]^i}{i!}$, который, как известно, сходится при любом x и λ (к функции $N\mathrm{e}^{|\lambda|M(x-a)}$), т.е. в отличие от интегральных уравнений Фредгольма второго рода сходимость метода последовательных приближений для уравнения (4.1) не накладывает ограничений на количественные характеристики параметров исходной задачи.

Для погрешности $\varepsilon_n(x)$ приближенного решения

$$y_n(x)=\sum_{i=0}^{n}\lambda^i\varphi_i(x)$$

получим:

$$|\varepsilon_n(x)| = \left|\sum_{i=n+1}^{\infty}\lambda^i\varphi_i(x)\right| \le N\frac{|\lambda|^{n+1}M^{n+1}(x-a)^{n+1}}{(n+1)!}\left(1+\frac{|\lambda|M(x-a)}{n+2}+\cdots\right) \le \left[q=|\lambda|M(b-a)\right] \le$$

$$\le N\frac{q^{n+1}}{(n+1)!}\left(1+\frac{q}{n+2}+\frac{q^2}{(n+2)(n+3)}+\ldots\right) < N\frac{q^{n+1}}{(n+1)!}\left[1+\frac{q}{n+2}+\frac{q^2}{(n+2)^2}+\ldots\right] =$$

$$= N\frac{q^{n+1}}{(n+1)!}\cdot\frac{1}{1-\dfrac{q}{n+2}},\ a\le x\le b,\ n>q-2.$$

ГЛАВА IX

Проекционные методы решения интегральных уравнений

Очень важной и динамично развивающейся в последнее время группой методов, предназначенных для решения функциональных уравнений (в том числе и интегральных) являются так называемые проекционные методы, основной идеей которых является поиск приближенных решений в подпространствах более простой по сравнению с исходным пространством, в котором поставлена исходная задача, структуры. Очень часто такими подпространствами являются линейные оболочки, натянутые на известные системы координатных функций и, таким образом, для определения приближенного решения задачи оказывается достаточным подобрать коэффициенты соответствующей линейной комбинации. Идея, на основе которой строится замыкающая система уравнений для определения коэффициентов, и отличает проекционные методы друг от друга.

§ 1. Метод моментов и метод Галеркина

Зададим две системы функций:

I. $\varphi_0(x),\ \varphi_1(x),\ldots$.

Эта система составляет базис подпространств, в которые проектируется решение исходного интегрального уравнения. Поэтому требования, предъявляемые к ней, должны быть такими:

1. При любом значении i функции $\varphi_i(x)$ непрерывны;
2. При любом конечном значении n система $\{\varphi_i(x)\}_{i=1}^{n}$ является линейно независимой;
3. Система $\{\varphi_i(x)\}$ обладает свойством C – полноты на множестве непрерывных функций (это означает следующее: для любой функции $F(x)\in C[a;b]$ и любого $\varepsilon>0$ существует n и набор коэффициентов $c_0, c_1,..., c_n$ такие, что $\left|F(x)-\sum_{i=0}^{n}c_i\varphi_i(x)\right|<\varepsilon$ для всех $x\in[a;b]$).

II. $\psi_0(x), \psi_1(x),\ldots$.

Эту систему будем использовать для возможно лучшего в каком-то смысле выполнения исходного уравнения на приближенном решении. Будем предполагать все функции этой системы непрерывными, линейно независимыми, а саму систему – замкнутой на множестве $C[a;b]$. Применительно к нашим целям это означает, что если

$\int_a^b f(x)\psi_i(x)dx = 0,\ i = 0,1,...$, то отсюда с необходимостью следует, что $f(x) \equiv 0$ (используя терминологию скалярного произведения, это можно переписать следующим образом: $(f,\psi_i) = 0,\ i = 0,1,... \Leftrightarrow f \equiv 0$).

Составим линейную комбинацию

$$u_n(x) = f(x) + \sum_{i=0}^{n} c_i \varphi_i(x). \tag{1.1}$$

Она содержит $(n+1)$ произвольных коэффициентов. Опишем требования, на основании которых можно осуществить их выбор.

Переписав исходную задачу в виде

$$Lu \equiv u(x) - \lambda \int_a^b K(x,s)u(s)ds = f(x), \tag{1.2}$$

в силу замкнутости системы $\{\psi_i(x)\}$ получим требование: чтобы $u_n(x)$ была решением уравнения (1.2), т.е. чтобы имело место тождество $Lu_n - f \equiv 0$, необходимо и достаточно выполнение бесконечного множества равенств

$$\left(Lu_n - f, \psi_j\right) = 0,\ j = 0,1,2,...$$

Однако в нашем распоряжении имеется лишь $(n+1)$ коэффициентов c_i, выбором которых мы можем распоряжаться. Следовательно, мы имеем возможность удовлетворить только первым $(n+1)$ из выписанных условий:

$$\left(Lu_n - f, \psi_j\right) = 0,\ j = 0,1,...,n. \tag{1.3}$$

Условия (1.3) дают систему линейных алгебраических уравнений для определения коэффициентов c_i. Запишем ее более подробно.

Так как оператор L линейный, то $Lu_n = Lf + \sum_{i=0}^{n} c_i L\varphi_i$. Поэтому из (1.3) получим:

$$\left(Lf + \sum_{i=0}^{n} c_i L\varphi_i - f, \psi_j\right) = 0,\ j = 0,1,...,n$$

или (учитывая также и линейность скалярного произведения)

$$\sum_{i=0}^{n} c_i \left(L\varphi_i, \psi_j\right) = \left(f - Lf, \psi_j\right),\ j = 0,1,...,n.$$

Таким образом, система для нахождения коэффициентов линейной комбинации (1.1), определяющей приближенное решение, имеет вид

$$\sum_{i=0}^{n} a_{ji} c_i = b_j,\ j = 0,1,...,n, \tag{1.4}$$

где

$$a_{ji} = (L\varphi_i, \psi_j) = \int_a^b \left[\varphi_i(x) - \lambda \int_a^b K(x,s)\varphi_i(s)ds \right] \psi_j(x)dx =$$

$$= \int_a^b \varphi_i(x)\psi_j(x)dx - \lambda \int_a^b \int_a^b K(x,s)\varphi_i(s)\psi_j(x)dsdx,$$

(1.5)

$$b_j = (f - Lf, \psi_j) = \int_a^b \left[f(x) - f(x) + \lambda \int_a^b K(x,s)f(s)ds \right] \psi_j(x)dx =$$

$$= \lambda \int_a^b \int_a^b K(x,s)f(s)\psi_j(x)dsdx.$$

При сформулированных выше требованиях, предъявляемых к системам функций $\{\varphi_i(x)\}$ и $\{\psi_i(x)\}$ система (1.4), (1.5) будет иметь единственное решение. Найдя его тем или иным способом, построим приближенное решение $u_n(x)$.

Если системы $\{\varphi_i(x)\}$ и $\{\psi_i(x)\}$ ***совпадают***, то получаем ***алгоритм метода Галеркина***. В дальнейшем будем говорить именно о нем.

1.1. Связь метода Галеркина с заменой ядра вырожденным

Покажем, что применение метода Галеркина равносильно замене ядра $K(x,s)$ вырожденным ядром, строящимся некоторым специальным образом. Действительно, предполагая ортонормированность системы $\{\varphi_i(x)\}$ (если это не так, то всегда можно применить процедуру ортогонализации), разложим ядро $K(x,s)$ как функцию переменной x в ряд Фурье по этой системе и за $\tilde{K}(x,s)$ примем n-ю частичную сумму этого ряда. Получим:

$$\tilde{K}(x,s) = \sum_{i=0}^{n} \beta_i(s)\varphi_i(x),$$

где

$$\beta_i(s) = \int_a^b K(x,s)\varphi_i(x)dx.$$

Если теперь для уравнения

$$u(x) - \lambda \int_a^b \tilde{K}(x,s)u(s)ds = f(x)$$

записать (в соответствии с теорией уравнений с вырожденным ядром) точное решение

$$u(x) = f(x) + \lambda \sum_{i=0}^{n} c_i\varphi_i(x),$$

то для определения коэффициентов c_i имеем систему (см. формулы (2.4), (2.5) главы **VIII**)

$$c_i - \lambda \sum_{j=0}^{n} a_{ij} c_j = b_i, \quad i = 0,1,...,n,$$

где

$$b_i = \int_a^b f(s)\beta_i(s)ds = \int_a^b \int_a^b K(x,s) f(s)\varphi_i(x)dsdx,$$

$$a_{ij} = \int_a^b \beta_i(s)\varphi_j(s)ds = \int_a^b \int_a^b K(x,s)\varphi_i(s)\varphi_j(x)dsdx,$$

совпадающую с (1.4), (1.5) при $\psi_i(x) \equiv \varphi_i(x), \ i = 0,1,...$.

Замечание. Возможна организация работы по описанным выше методам с несколько иным представлением приближенного решения, немного отличающимся от (1.1):

$$u_n(x) = \sum_{i=0}^{n} c_i \varphi_i(x). \qquad (1.1')$$

В этом случае система для определения коэффициентов c_i также будет иметь вид (1.4) с той лишь разницей, что для коэффициентов b_j необходимо использовать формулу

$$b_j = (f, \psi_j) = \int_a^b f(x)\psi_j(x)dx. \qquad (1.5')$$

§ 2. Другие проекционные методы

2.1. Метод наименьших квадратов

Рассмотрим вновь интегральное уравнение (1.2) предыдущего параграфа. Легко видеть, что решение этой задачи эквивалентно задаче нахождения минимума функционала

$$J(u) = (Lu - f, Lu - f), \qquad (2.1)$$

(где, по-прежнему, $(f,g) = \int_a^b f(x)g(x)dx$ обозначает скалярное произведение) и, следовательно, может быть заменено последней, причем, в соответствии с общей идеей проекционных методов, минимум будем искать в подпространствах конечной размерности.

Задавая систему функций $\{\varphi_i(x)\}$, описанную в предыдущем параграфе (с такими же свойствами), приближенное решение будем искать в виде (1.1′):

$$u_n(x) = \sum_{i=0}^{n} c_i \varphi_i(x). \qquad (2.2)$$

Функционал (2.1) на приближенном решении примет вид

$$J(u_n) = \left(\sum_{i=0}^{n} c_i L\varphi_i - f, \sum_{i=0}^{n} c_i L\varphi_i - f \right)$$

и будет являться функцией переменных $c_0, c_1, ..., c_n$ (коэффициентов комбинации (2.2)).

Записывая необходимое условие минимума первого порядка, получим:

$$\frac{\partial J(u_n)}{\partial c_j} = 2\left(\sum_{i=0}^{n} c_i L\varphi_i - f, L\varphi_j \right) = 0, \ \ j = 0,1,...,n.$$

Таким образом, имеем систему линейных алгебраических уравнений для определения коэффициентов $c_0, c_1, ..., c_n$:

$$\sum_{i=0}^{n} a_{ji} c_i = b_j, \ \ j = 0,1,...,n, \tag{2.3}$$

где

$$a_{ji} = \left(L\varphi_i, L\varphi_j \right) = \int_a^b \left[\varphi_i(x) - \lambda \int_a^b K(x,s)\varphi_i(s)ds \right] \left[\varphi_j(x) - \lambda \int_a^b K(x,s)\varphi_j(s)ds \right] dx,$$

$$b_j = \left(f, L\varphi_j \right) = \int_a^b f(x) \left[\varphi_j(x) - \lambda \int_a^b K(x,s)\varphi_j(s)ds \right] dx. \tag{2.4}$$

2.2. Метод коллокации

Как и выше, решаем уравнение $Lu(x) = f(x)$. В соответствии с общей идеологией проекционных методов приближенное решение будем искать в виде (2.2) с использованием введенной ранее системы функций $\{\varphi_i(x)\}$.

По сути, конечным «продуктом» метода является опять-таки система линейных алгебраических уравнений. Отличия состоят лишь в способе ее получения (и, как следствие – в виде элементов матрицы и свободных членов). Здесь для этих целей используется идея обращения в нуль невязки уравнения (1.2) для приближенного решения на выбранном множестве точек. Таким образом, требуется выполнение соотношений

$$Lu_n(x_j) = f(x_j), \ \ j = 0,1,...,n \tag{2.5}$$

или

$$\sum_{i=0}^{n} a_{ji} c_i = f(x_j), \ \ j = 0,1,...,n, \tag{2.6}$$

где

$$a_{ji} = L\varphi_i(x_j) = \varphi_i(x_j) - \lambda \int_a^b K(x_j, s)\varphi_i(s)ds. \tag{2.7}$$

Чтобы система (2.7) имела единственное решение, необходимо потребовать отличия от нуля ее определителя:

$$\Delta = \begin{vmatrix} L\varphi_0(x_0) & \dots & L\varphi_n(x_0) \\ \dots & \dots & \dots \\ L\varphi_0(x_n) & \dots & L\varphi_n(x_n) \end{vmatrix} \neq 0.$$

Как мы помним, в этом случае теоретически достаточно выполнения требования (которое, однако, на практике совсем не просто выполнить), чтобы система $\{L\varphi_i(x)\}$ являлась системой функций Чебышева на отрезке $[a;b]$. Это можно считать дополнительными ограничениями, накладываемыми на систему $\{\varphi_i(x)\}$.

В качестве общего замечания отметим следующее:

1. Для решения интегральных уравнений можно применять и другие идеи (например, использовать для этих целей сплайн-приближения);
2. Практически все изложенные методы (за исключением, разве что, метода замены ядра на вырожденное, в котором явно используется линейность задачи) можно применять и для решения других типов интегральных уравнений (в том числе – нелинейных) с соответствующими изменениями в алгоритме.

ГЛАВА IX

Решение интегральных уравнений первого рода

В предыдущих главах данного раздела мы ввели понятие интегрального уравнения, а также привели классификацию интегральных уравнений по типам и основные способы решения уравнений Фредгольма и Вольтера второго рода, оставляя за скобками (как правило) интегральные уравнения первого рода. Основной причиной этого является тот факт, что уравнения первого рода по сути своей относятся к классу некорректных задач. Учитывая данное замечание, рассмотрим сейчас несколько подробнее некоторые подходы к решению таких задач.

§ 1. Основные определения и примеры

Большинство некорректных задач может быть приведено к операторному уравнению первого рода, имеющему вид

$$Au = f,\ \ u \in U,\ \ f \in F, \tag{1.1}$$

в котором по заданному оператору A (не обязательно линейному), действующему из пространства U в пространство F, и по заданному элементу $f \in F$ требуется определить решение $u \in U$.

В частном случае имеем

1^0. Интегральное уравнение Фредгольма первого рода:

$$\int_a^b K(x,s)u(s)ds = f(x),\ \ c \le x \le d; \tag{1.2}$$

2^0. Интегральное уравнение Вольтера первого рода

$$\int_a^x K(x,s)u(s)ds = f(x),\ \ c \le x \le d; \tag{1.3}$$

Определение. Задача отыскания элемента $u \in U$ из уравнения (1.1) называется корректной (или корректно поставленной), если при любой фиксированной правой части $f = f_0 \in F$ ее решение

а) существует в пространстве U;
б) единственно в пространстве U;
в) устойчиво в пространстве U, т.е. непрерывно зависит от $f \in F$.

Данное определение называют также определением корректности по Адамару.

Если хотя бы одно из условий а) – в) не выполняется, то задачу называют некорректной (или некорректно поставленной).

Таким образом, корректность задачи связана с наличием обратного оператора A^{-1}, определенного и непрерывного на всем пространстве F.

Простейшим примером некорректно поставленной задачи могут служить системы линейных алгебраических уравнений (о чем, впрочем, мы говорили при рассмотрении соответствующей темы). Запишем соответствующую задачу в матричном виде

$$Au = f.$$

Теоретически возможны следующие ситуации:
1) A – квадратная матрица, причем $\det A = 0$. Тогда решений либо не существует вовсе, либо их будет бесконечно много. Таким образом, нарушаются либо условие а), либо условие б);
2) A – квадратная матрица, но $\det A \neq 0$. Тогда система имеет единственное решение, но при нарушении условия в) имеем плохо обусловленную задачу, сложность решения которой отмечалась ранее;
3) A – прямоугольная матрица. Тогда, как это следует из общей теории линейных систем, также нарушаются условия а) и б) из определения корректности.

Покажем сейчас, что интегральное уравнение Фредгольма первого рода (1.2) также является некорректной задачей.

Рассмотрим более простой для исследования частный случай. Пусть ядро $K(x,s)$ вещественно и симметрично, т.е. $K(x,s) = K(s,x)$. Предположим также, что $K(x,s)$ и $f(x)$ непрерывны. Тогда, как известно (см., например, книгу: А.Б. Антоневич, Я.В. Радыно. «Функциональный анализ и интегральные уравнения»), существует полная ортонормированная система собственных функций $\varphi_i(x)$ оператора A:

$$A\varphi_i(x) = \int_a^b K(x,s)\varphi_i(s)ds = \lambda_i\varphi_i(x), \quad i = 0,1,...$$

$$(\varphi_i, \varphi_j) = \int_a^b \varphi_i(s)\varphi_j(s)ds = \delta_i^j.$$

При этом ядро $K(x,s)$ раскладывается в сходящийся ряд по собственным функциям (ряд Фурье)

$$K(x,s) = \sum_{i=0}^{\infty} \lambda_i\varphi_i(x)\varphi_i(s),$$

где сходимость ряда в правой части понимается в L_2-норме:

$$\|K(x,s)\| = \sqrt{\int_a^b\int_a^b |K(x,s)|^2 dxds}.$$

Отсюда, в частности, следует, что $\|K\|^2 = \sum_{i=0}^{\infty} |\lambda_i|^2$ и, следовательно, $\lambda_i \underset{i\to\infty}{\to} 0$.

Рассмотрим случай, когда $\lambda_i \neq 0$ при $0 \le i \le N$ и все $\lambda_i = 0$ при $i > N$. Тогда ядро имеет вид

$$K(x,s) = \sum_{i=0}^{N} \lambda_i \varphi_i(x) \varphi_i(s),$$

т.е. является вырожденным и уравнение (1.2) может быть переписано в виде

$$\int_a^b K(x,s)u(s)ds = \sum_{i=0}^{N} \lambda_i \int_a^b \varphi_i(x)\varphi_i(s)u(s)ds = \sum_{i=0}^{N} \lambda_i (\varphi_i, u)\varphi_i(x) = f(x).$$

Отсюда следует, что задача может иметь решение только в том случае, когда $f(x)$ является линейной комбинацией функций $\varphi_0(x),...,\varphi_N(x)$, т.е. записывается в виде

$$f(x) = \sum_{i=0}^{N} f_i \varphi_i(x).$$

Легко видеть, что решением в этом случае является функция

$$u(x) = u_0(x) = \sum_{i=0}^{N} \frac{f_i}{\lambda_i} \varphi_i(x).$$

Действительно, если искать решение $u(x)$ в виде (в полном согласии с теорией уравнений с вырожденным ядром)

$$u(x) = \sum_{i=0}^{N} C_i \varphi_i(x),$$

то

$$(u, \varphi_k) = C_k$$

и тогда

$$\int_a^b K(x,s)u(s)ds = \sum_{i=0}^{N} \lambda_i (\varphi_i, u)\varphi_i(x) = \sum_{i=0}^{N} C_i \lambda_i \varphi_i(x) = f(x) = \sum_{i=0}^{N} f_i \varphi_i(x),$$

откуда $C_i = \frac{f_i}{\lambda_i}$.

В то же время, любая функция $u(x)$, представимая в виде

$$u(x) = u_0(x) + \sum_{i=N+1}^{\infty} C_i \varphi_i(x),$$

где $\sum_{i=N+1}^{\infty} |C_i|^2 < +\infty$, также будет решением уравнения (1.2) (***проверьте*!**).

Таким образом, в рассматриваемом случае задача (1.2) может не иметь решения; в случае же, когда это решение существует, оно не единственно.

Совершенно аналогично можно рассмотреть случай, когда все собственные значения λ_i отличны от нуля.

Решение в этом случае представимо в виде ряда

$$u(x) = \sum_{i=0}^{\infty} \frac{f_i}{\lambda_i} \varphi_i(x),$$

где, опять-таки, f_i – коэффициенты разложения $f(x)$ в ряд по собственным функциям $\varphi_i(x)$. Если $f(x) \in L_2[a;b]$, то такой ряд будет сходиться в норме пространства L_2, т.е. решение в этом случае будет существовать и окажется единственным (два решения могут не совпадать на множестве меры нуль).

Но в то же время, если правые части уравнения (1.2) $f^n(x),\ n = 0, 1, 2,...$ сходятся к некоторой функции $f(x)$ в пространстве $L_2[a;b]$, т.е.

$$\| f^n - f \|^2 = \sum_{i=0}^{\infty} \left(f_i^n - f_i\right)^2 \xrightarrow[n\to\infty]{} 0,$$

то норма разности соответствующих решений уравнения (1.2), выражаемая равенством

$$\| u^n - u \|^2 = \sum_{i=0}^{\infty} \frac{\left(f_i^n - f_i\right)^2}{\lambda_i^2},$$

не только не обязана стремиться к нулю, но и может быть бесконечно большой. В этом легко убедиться, положив $f^n(x) = f(x) + \sqrt{|\lambda_n|}\,\varphi_n(x)$. Тогда

$$\| f^n - f \|^2 = \| \sqrt{|\lambda_n|}\,\varphi_n \|^2 = |\lambda_n| \xrightarrow[n\to\infty]{} 0.$$

В то же время

$$\| u^n - u \|^2 = \left\| u + \frac{\sqrt{|\lambda_n|}}{|\lambda_n|}\varphi_n - u \right\|^2 = \frac{1}{|\lambda_n|} \underset{n\to\infty}{\to} \infty.$$

Следовательно, устойчивость решений отсутствует.

Таким образом, в рассмотренных частных случаях возможны нарушения всех трех условий корректности.

Отметим также, что для уравнения (1.3) справедливы аналогичные результаты. При этом, в частности, простейшее интегральное уравнение Вольтера первого рода

$$\int_a^x u(s)ds = f(x) - f(a) \qquad (1.4)$$

эквивалентно задаче нахождения производной, поскольку решением (1.4) (в случае если $f(x)$ дифференцируема (!)) является функция $u(x) = f'(x)$. О некорректности последней мы уже упоминали ранее.

§ 2. Метод регуляризации решения некорректных задач

Как следует из изложенного выше, непосредственно решать некорректно поставленные задачи при неточно заданной правой части бессмысленно. Если $\bar{f}(x)$ задана с погрешностью $\delta f(x)$, то соответствующее решение $u_\delta(x)$ или не существует, или отличается от искомого решения $\bar{u}(x)$ на величину $\delta u(x)$, которая может быть большой.

Даже если $f(x)$ задана точно, но отыскание решения выполняется численными методами, то неизбежно вносится погрешность метода и округления. Это снова приводит к большой погрешности решения $\delta u(x)$.

Однако никто не обязывает нас непосредственно решать исходную задачу

$$Au = f \tag{2.1}$$

с возмущенной правой частью. Всегда можно попытаться заменить эту задачу некоторой «близкой» задачей, решение которой будет «близко» к $u(x)$. Символически запишем измененную задачу в виде

$$A_\alpha u_\alpha = f\,, \tag{2.2}$$

где $\alpha > 0$ – некоторый параметр (параметр регуляризации), а ее решение будем обозначать $u_\alpha(x)$.

Определение. Оператор A_α называют регуляризирующим, если:

а) задача (2.2) является корректно поставленной в классе правых частей F при любом $\alpha > 0$;

б) существуют такие функции $\alpha(\delta)$ и $\delta(\varepsilon)$, что если $\left\| f - \bar{f} \right\|_F \le \delta(\varepsilon)$, то $\left\| u_{\alpha(\delta)} - \bar{u} \right\|_U \le \varepsilon$.

Таким образом, если найден регуляризирующий оператор A_α, то задача (2.2) имеет решения при любых $f \in F$, в том числе и отличающихся от $\bar{f}$ на любого вида погрешность δf; эта задача устойчива, так что ее можно решать обычными численными методами. При правильно подобранном параметре регуляризации α ее решение $u_\alpha(x)$ достаточно мало отличается от нужного нам решения $\bar{u}(x)$ исходной задачи (2.1).

2.1. Вариационный метод регуляризации

Рассмотрим уравнение Фредгольма первого рода

$$\int_a^b K(x,s)u(s)ds = f(x),\ c \le x \le d\,. \tag{2.3}$$

Будем считать, что ядро его непрерывно и таково, что в случае $f(x) \equiv 0$ имеет только тривиальное решение $u(x) \equiv 0$. Тогда при любой правой части $f(x) \in F$ решение либо единственное, либо не существует. Тем самым интегральный оператор

$$A(x, u(s)) = \int_a^b K(x,s)u(s)ds \tag{2.4}$$

отображает U в F взаимно однозначно.

Исходную задачу (2.3) можно заменить эквивалентной вариационной задачей

$$\int_c^d [A(x,u(s))-f(x)]^2\,dx \to \min\,. \qquad (2.5)$$

Рассмотрим измененную задачу

$$M(\alpha, f(x),u(s)) = \int_c^d [A(x,u(s))-f(x)]^2\,dx + \alpha\,\Omega(u(s)) \to \min\,, \qquad (2.6)$$

где $\Omega(u(s))$ – так называемый ***тихоновский стабилизатор***. Чаще всего в качестве $\Omega(u)$ берут функционал

$$\Omega(u) = \Omega_n(u) = \|u\|^2_{W_2^n} = \int_a^b \left[\sum_{k=0}^{n} p_k(s)\left(\frac{d^k u(s)}{ds^k}\right)^2\right] ds$$

при некотором значении n (здесь W_2^n – пространство Соболева, а все весовые функции $p_k(s)$ непрерывны и неотрицательны).

***Теорема* 1.** Задача (2.6) имеет решение $u_\alpha(x)$ при любых $f(x)\in F$ и $\alpha>0$.

Доказательство.

При $\alpha>0$ функционал $M(\alpha,f,u)$ ограничен снизу. Поэтому при данных α и $f(x)$ он имеет точную нижнюю грань: $\overline{M} = \inf\limits_{u\in U} M(\alpha,f,u)$. Выберем некоторую минимизирующую последовательность $u_i(s)$ так, что $\lim\limits_{i\to\infty} M_i = \overline{M}$, где $M_i = M(\alpha,f,u_i)$. Упорядочим ее так, чтобы M_i не возрастали. Тогда

$$\alpha\,\Omega(u_i) \le M_i \le M_0 = \text{const}$$

или

$$\Omega(u_i) \le \frac{1}{\alpha} M_0\,.$$

Таким образом, последовательность $u_i(s)$ принадлежит множеству таких $u(s)$, для которых $\Omega(u)\le \text{const}$. А это множество, как известно, является компактом в U. Поэтому из последовательности $u_i(s)$ можно выделить подпоследовательность $u_{i_k}(s)$, сходящуюся по норме к некоторой $u_\alpha(s)\in U$. В силу непрерывности функционал $M(\alpha,f,u)$ на этой функции достигает своей точной нижней грани. Тем самым $u_\alpha(s)\in U$ есть решение задачи (2.6). ☒

***Теорема* 2.** Алгоритм (2.6) является регуляризирующим для задачи (2.5).

Доказательство.

Пусть $\bar{u}(s)$ – решение задачи (2.5) с правой частью $\bar{f}(x)$, $\tilde{u}_\alpha(s)$ – решение задачи (2.6) с приближенной правой частью $\tilde{f}(x)$ и $f_\alpha(x) = A(x,\tilde{u}_\alpha(s))$.

Поскольку функционал $M\left(\alpha,\tilde{f},u\right)$ достигает минимума на элементе $\tilde{u}_\alpha$, то справедливо неравенство

$$M\left(\alpha,\tilde{f},\tilde{u}_\alpha\right)\le M\left(\alpha,\tilde{f},\bar{u}\right).$$

Отсюда получаем:

$$\alpha\,\Omega(\tilde{u}_\alpha)\le M\left(\alpha,\tilde{f},\tilde{u}_\alpha\right)\le M\left(\alpha,\tilde{f},\bar{u}\right)=\int_c^d\left[A(x,\bar{u})-\tilde{f}(x)\right]^2dx+\alpha\,\Omega(\bar{u})=$$

$$=\int_c^d\left[\bar{f}(x)-\tilde{f}(x)\right]^2dx+\alpha\,\Omega(\bar{u})=\left\|\bar{f}-\tilde{f}\right\|_{L_2}^2+\alpha\,\Omega(\bar{u}). \tag{2.7}$$

Пусть приближенные правые части удовлетворяют условию

$$\left\|\bar{f}-\tilde{f}\right\|_{L_2}\le C\sqrt{\alpha}\,, \tag{2.8}$$

где C – некоторая константа.
Тогда из неравенства (2.7) следует, что

$$\Omega(\tilde{u}_\alpha)\le C^2+\Omega(\bar{u})=\text{const}\,. \tag{2.9}$$

Значит, $\tilde{u}_\alpha$ принадлежит компактному множеству U_0 функций из U (заметим, что $\bar{u}$ также принадлежит U_0).

Множество F_0 функций $f_\alpha(x)$ есть образ множества U_0 при отображении A. По предположению оператор A непрерывен и таков, что обратное отображение единственно. Поэтому обратное отображение F_0 в компактное множество U_0 при помощи нерегуляризованного оператора A^{-1} будет непрерывным в норме пространства U. Следовательно, по заданному $\varepsilon>0$ всегда найдется $\beta(\varepsilon)$ такое, что из условия $\left\|f_\alpha-\bar{f}\right\|\le\beta(\varepsilon)$ следует выполнение неравенства $\left\|\tilde{u}_\alpha-\bar{u}\right\|\le\varepsilon$.

Заметим, что

$$\left\|f_\alpha-\tilde{f}\right\|^2=\int_c^d\left[f_\alpha(x)-\tilde{f}(x)\right]^2dx=\int_c^d\left[A(x,\tilde{u}_\alpha)-\tilde{f}(x)\right]^2dx\le M\left(\alpha,\tilde{f},\tilde{u}_\alpha\right)\le\alpha\left(C^2+\Omega(\bar{u})\right).$$

Отсюда с учетом условия (2.8) следует:

$$\left\|f_\alpha-\bar{f}\right\|\le\left\|f_\alpha-\tilde{f}\right\|+\left\|\tilde{f}-\bar{f}\right\|\le\sqrt{\alpha}\left(C+\sqrt{C^2+\Omega(\bar{u})}\right). \tag{2.10}$$

Выберем α так, чтобы

$$\alpha\le\alpha_0(\varepsilon)\equiv\left(\frac{\beta(\varepsilon)}{C+\sqrt{C^2+\Omega(\bar{u})}}\right)^2. \tag{2.11}$$

Тогда правая часть неравенства (2.10) не будет превосходить $\beta(\varepsilon)$, откуда следует, что $\left\|\tilde{u}_\alpha-\bar{u}\right\|\le\varepsilon$.

Таким образом, по заданному ε нашлись $\alpha_0(\varepsilon)$ и $\delta(\alpha)=C\sqrt{\alpha}$ такие, что выполнение неравенств $\alpha\le\alpha_0(\varepsilon)$ и $\left\|\tilde{f}-\bar{f}\right\|\le\delta(\alpha)$ влечет выполнение неравенства $\left\|\tilde{u}_\alpha-\bar{u}\right\|\le\varepsilon$.

⊠

Следствие. Задача (2.6) корректно поставлена.

Доказательство немедленно следует из того, что заключительная строка доказательства теоремы, по сути, означает непрерывную зависимость решений от правых частей задачи.

Замечание. Сходимость в пространстве W_2^n означает, что n-я производная от сходится среднеквадратично, а сама функция и все остальные (до порядка $(n-1)$ включительно) – равномерно. Таким образом, использование стабилизатора $\Omega_n(u)$ обеспечивает слабую регуляризацию при $n=0$, сильную при $n=1$ и $(n-1)$-го порядка гладкости при $n>1$.

2.1.1. Выбор параметров регуляризации

В ряде прикладных задач известно, что правые части имеют характерную погрешность $\left\|\widetilde{f}-\bar{f}\right\|$ порядка некоторой заданной величины δ. Если при этом выбрать α настолько малым, что нарушится критерий (2.8), то устойчивость расчетов станет недостаточной, так что регуляризованное решение $\widetilde{u}_\alpha$ будет заметно «разболтанным». Если же α настолько велико, что не соблюден критерий (2.11), то регуляризованное решение $\widetilde{u}_\alpha$ будет чрезмерно сглажено, что также нежелательно.

Вдобавок непосредственно проверить выполнение критериев (2.8), (2.11) не удается, поскольку $\beta(\varepsilon)$ неизвестно. Поэтому оптимальный выбор параметра α является сложной задачей.

Обычно на практике проводят расчеты с несколькими значениями параметра α, составляющими геометрическую прогрессию (например, 10^{-1}, 10^{-2},...) (или по какому-либо другому закону), из полученных результатов выбирают наилучший либо визуально, либо по какому-либо критерию правдоподобия.

Такое поведение характерно для некорректных задач. Например, приближенное вычисление производной на сетке с шагом h с помощью численного дифференцирования приводит к следующему: характерной погрешностью метода является величина вида Ch^m, а полная (включая погрешность входных данных) – $E=Ch^m+\dfrac{\delta}{h}$. Следовательно, оптимальным значением величины шага является значение $h=h_{\text{опт}}=\sqrt[m+1]{\dfrac{\delta}{Cm}}$.

Выбор n**.** Аналогично предыдущим рассуждениям можно отметить, что при чрезмерно больших значениях n регуляризованное решение сильно сглаживается. Значение $n=0$ обеспечивает лишь среднеквадратичную сходимость $\widetilde{u}_\alpha$ к $\bar{u}$. Поэтому наиболее часто используют значение $n=1$.

2.2. Уравнение Эйлера

Учитывая явный вид операторов A и Ω, перепишем задачу (2.6) следующим образом:

$$\alpha\sum_{k=0}^{n}\int_a^b p_k(s)\left[u^{(k)}(s)\right]^2 ds+\int_c^d\left[\int_a^b K(x,s)u(s)ds-f(x)\right]^2 dx\to\min. \qquad (2.12)$$

Из теории вариационного исчисления известно, что функция $\widetilde{u}(x)$, доставляющая решение задачи (2.12), удовлетворяет ***уравнению Эйлера,*** смысл которого аналогичен необходимому условию минимума первого порядка для функций: первая вариация равна нулю. Полагая $u\sim u_1:=u+\delta u$, обнуляем те слагаемые, которые содержат первые степени δ:

$$0 = \alpha \sum_{k=0}^{n} \int_a^b p_k(s) u^{(k)}(s) \delta u^{(k)}(s) ds + \int_c^d \left[\int_a^b K(x,\eta) u(\eta) d\eta - f(x) \right] \int_a^b K(x,s) \delta u(s) ds dx . \quad (2.13)$$

Интегралы, стоящие под знаком суммы, будем вычислять последовательным интегрированием по частям:

$$\int_a^b p_k(s) u^{(k)}(s) \delta u^{(k)}(s) ds = \delta u^{(k-1)}(s) p_k(s) u^{(k)}(s) \Big|_a^b - \int_a^b \delta u^{(k-1)}(s) \frac{d}{ds} \left[p_k(s) u^{(k)}(s) \right] ds =$$

$$= \sum_{j=0}^{k-1} (-1)^j \delta u^{(k-1-j)}(s) \frac{d^j}{ds^j} \left[p_k(s) u^{(k)}(s) \right] \Big|_a^b + (-1)^k \int_a^b \delta u(s) \frac{d^k}{ds^k} \left[p_k(s) u^{(k)}(s) \right] ds.$$

Подставляя это выражение в уравнение вариации (2.13), получим:

$$\alpha \sum_{k=0}^{n} \sum_{j=0}^{k-1} (-1)^j \delta u^{(k-1-j)}(s) \frac{d^j}{ds^j} \left[p_k(s) u^{(k)}(s) \right] \Big|_a^b + \alpha \sum_{k=0}^{n} (-1)^k \int_a^b \delta u(s) \frac{d^k}{ds^k} \left[p_k(s) u^{(k)}(s) \right] ds +$$

$$+ \int_c^d \left[\int_a^b K(x,\eta) u(\eta) d\eta - f(x) \right] \int_a^b K(x,s) \delta u(s) ds dx = 0.$$

Теперь поменяем порядок суммирования (что равносильно собиранию коэффициентов при $\delta u^{(k)}$):

$$\alpha \sum_{j=1}^{n} \delta u^{(j-1)}(s) \sum_{k=j}^{n} (-1)^{k-j} \frac{d^{k-j}}{ds^{k-j}} \left[p_k(s) u^{(k)}(s) \right] \Big|_a^b + \alpha \sum_{k=0}^{n} (-1)^k \int_a^b \delta u(s) \frac{d^k}{ds^k} \left[p_k(s) u^{(k)}(s) \right] ds +$$

$$+ \int_c^d \left[\int_a^b K(x,\eta) u(\eta) d\eta \right] \left[\int_a^b K(x,s) \delta u(s) ds \right] dx = \int_c^d f(x) \int_a^b K(x,s) \delta u(s) ds dx .$$

Введем обозначение $q_j(u) = \sum_{k=j}^{n} (-1)^{k-j} \frac{d^{k-j}}{ds^{k-j}} \left[p_k(s) u^{(k)}(s) \right]$.

Тогда, полагая

$$q_j(u(a)) = q_j(u(b)) = 0, \quad j = \overline{1,n} \quad (2.14)$$

и приравнивая коэффициенты при $\delta u(s)$ под знаками интеграла справа и слева (эта процедура эквивалентна выбору в качестве вариации δ - функции), получим:

$$\alpha \sum_{k=0}^{n} (-1)^k \frac{d^k}{ds^k} \left[p_k(s) u^{(k)}(s) \right] + \int_a^b \left[\int_c^d K(x,\eta) K(x,s) dx \right] u(\eta) d\eta = \int_c^d K(x,s) f(x) dx .$$

Вводя обозначения

$$Q(s,\eta) = \int_c^d K(x,\eta) K(x,s) dx, \quad \Phi(s) = \int_c^d K(x,s) f(x) dx , \quad (2.15)$$

последнее уравнение перепишем в виде

$$\alpha\sum_{k=0}^{n}(-1)^k\frac{d^k}{ds^k}\left[p_k(s)u^{(k)}(s)\right]+\int_a^b Q(s,\eta)u(\eta)d\eta=\Phi(s). \qquad (2.16)$$

(2.14) – (2.16) и представляет собой искомое уравнение Эйлера и является интегро-дифференциальным уравнением, ядро $Q(x,s)$ которого определено на квадрате $[a;b]\times[a;b]$, симметрично и непрерывно, а правая часть $\Phi(s)$ непрерывна.

В частном случае слабой регуляризации (при $n=0$) (2.14) – (2.16) превращается в

$$\alpha u(s)+\int_a^b Q(s,\eta)u(\eta)d\eta=\Phi(s),\ \ a\le s\le b, \qquad (2.17)$$

т.е. в этом случае регуляризованная задача представляет собой обычное интегральное уравнение Фредгольма второго рода, способы решения которого мы изучали выше.

2.3. Замечание о решении плохо обусловленных линейных систем

Описанным выше способом можно решать и системы линейных алгебраических уравнений. Рассмотрим вкратце эту ситуацию.

Пусть задана линейная система

$$Au=f,$$

где u и f – конечномерные векторы.

Выбирая в стабилизирующем функционале $\Omega_n(u)$ $n=0$, по аналогии с изложенным выше, запишем задачу минимизации вариационного функционала

$$M(\alpha,f,u)=\|Au-f\|^2+\alpha\|u\|^2\to\min\ \text{(здесь } \|a\|^2=(a,a)). \qquad (2.18)$$

Формально $n=0$ соответствует слабой регуляризации, но в конечномерном пространстве все нормы эквивалентны. Поэтому сходимость регуляризованного решения (при $\alpha\to 0$) будет равномерной.

Поскольку (2.18) является квадратичной формой относительно u, то нахождение минимума последней сводится к решению линейной алгебраической системы

$$\left(A^TA+\alpha E\right)u=A^Tf.$$

Благодаря слагаемому αE эта система хорошо обусловлена (по крайней мере, при не слишком малых $\alpha>0$), Поэтому ее можно решать с помощью стандартных алгоритмов.

Описанный алгоритм может бать применен также к решению систем с вырожденной матрицей A.

РАЗДЕЛ V

Методы решения задачи Коши для обыкновенных дифференциальных уравнений

С помощью обыкновенных дифференциальных уравнений можно описывать движение системы взаимодействующих материальных точек (динамика системы материальных точек), концентрации реагирующих веществ в химических превращениях (задачи химической кинетики), задачи теории электрических цепей, сопротивления материалов и т.п. Учитывая, что в большинстве своем – это задачи динамики (т.е. изменения состояния некоторой системы с течением времени), независимую переменную будем обозначать буквой t.

Конкретная прикладная задача может приводить к дифференциальному уравнению любого порядка или к системе таких уравнений (опять-таки, любого порядка). При этом каждое из уравнений может иметь различный вид (в том числе и задаваться неявно). Мы в нашем курсе будем, однако, основное внимание уделять некоторому частному, но достаточно важному и широко распространенному классу задач, а именно: мы будем рассматривать уравнения, разрешенные относительно старшей производной.

Так, например, уравнение n-го порядка, разрешенное относительно старшей производной, выглядит следующим образом:

$$u^{(n)}(t) = f\left(t, u, u', \ldots, u^{(n-1)}\right), \tag{1}$$

где f – некоторая заданная функция от $(n+1)$ аргументов.

В то же время, хорошо известно, что задачу (1) с помощью замены $u^{(k)}(t) = u_{k+1}(t)$ можно свести к эквивалентной системе обыкновенных дифференциальных уравнений первого порядка

$$\begin{cases} u_k'(t) = u_{k+1}(t), \ k = 1, \ldots, n-1, \\ u_n'(t) = f(t, u_1, \ldots, u_n). \end{cases}$$

Аналогично произвольную систему обыкновенных дифференциальных уравнений любого порядка можно заменить некоторой эквивалентной системой уравнений первого порядка (естественно, с увеличением ее размерности по сравнению с исходной).

Учитывая сказанное, как правило, рассматривают системы уравнений первого порядка

$$u_k'(t) = f_k(t, u_1, \ldots, u_n), \ k = \overline{1, n},$$

которые в векторной форме имеют вид

$$u'(t) = f(t, u), \tag{2}$$

(здесь $u = (u_1, \ldots, u_n)^T$, $f = (f_1, \ldots, f_n)^T$).

Известно, что система имеет множество решений, которое в общем случае зависит от n параметров $C = (C_1, \ldots, C_n)^T$: $u(t) = u(t; C)$.

Для определения значений этих параметров, т.е. для выделения конкретного решения необходимо задать n дополнительных ограничений на функции $u_k(t)$. Эти ограни-

чения, вообще говоря, могут иметь самый разнообразный характер, хотя достаточно часто в качестве таковых используются значения указанных функций (или линейных комбинаций от них) в определенных точках промежутка, на котором поставлена задача.

Различают три основных типа задач для обыкновенных дифференциальных уравнений. Это:

1. Задачи Коши (или начальные задачи);
2. краевые задачи (или задачи с граничными условиями);
3. задачи на собственные значения (задачи Штурма-Лиувилля).

Задача Коши имеет дополнительные условия вида

$$u_k(t_0) = u_k^0, \ k = \overline{1,n}, \tag{3}$$

т.е. заданы значения всех функций в одной и той же точке $t = t_0$.

§ 0. Классификация методов решения задачи Коши

Для простоты изложения основных идей вычислительных методов решения задачи Коши в дальнейшем будем рассматривать (если это не оговорено особо) случай одного обыкновенного дифференциального уравнения первого порядка. Как правило, эти идеи, равно как и полученные с их помощью алгоритмы, легко переносятся на случай систем вида (2) (а следовательно, и на случай уравнений высших порядков).

Итак, пусть на отрезке $t_0 \le t \le T$ требуется найти решение $u(t)$ дифференциального уравнения

$$u'(t) = f(t,u), \tag{0.1}$$

удовлетворяющее начальному условию

$$u(t_0) = u_0. \tag{0.2}$$

Условия существования и единственности решения поставленной задачи Коши будем считать выполненными. Будем также предполагать, что функция $f(t,u)$ обладает необходимой по ходу изложения гладкостью.

Существующие приближенные алгоритмы можно (с определенной долей условности) разделить на

1. ***аналитические*** (когда решение получается в виде аналитически заданной функции);
2. ***численные*** (когда решение находят в виде таблицы значений в узлах заданной, либо параллельно построенной сетки).

Простейшим из аналитических методов является метод последовательных приближений или метод Пикара. Этот метод позволяет получать в аналитическом виде последовательность приближений $u_m(t), \ m = 0,1,\ldots$ к решению $u(t)$ по следующему правилу:

$$u_m(t) = u_0 + \int_{t_0}^{t} f(x, u_{m-1}(x))dx, \ t_0 \le t \le T, \ \ m = 1, 2, \ldots,$$

$$u_0(t) \equiv u_0.$$

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

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

$$u_m(t) = \sum_{i=0}^{m} \frac{(t-t_0)^i}{i!} u^{(i)}(t_0), \quad t_0 \le t \le T, \tag{0.3}$$

где

$$u^{(0)}(t_0) = u(t_0) = u_0, \quad u^{(1)}(t_0) = u'(t_0) = f(t_0, u_0),$$

а значения $u^{(i)}(t_0)$, $i = 2, 3, \ldots, m$ находят по формулам, полученным последовательным дифференцированием уравнения (0.1):

$$\begin{aligned} u^{(2)}(t_0) &= u''(t_0) = f_t(t_0, u_0) + f_u(t_0, u_0) \cdot f(t_0, u_0), \\ u^{(3)}(t_0) &= u'''(t_0) = f_{tt}(t_0, u_0) + 2 f_{tu}(t_0, u_0) \cdot f(t_0, u_0) + f_{uu}(t_0, u_0) \cdot f^2(t_0, u_0) + \\ &\quad + f_u(t_0, u_0)\left(f_t(t_0, u_0) + f_u(t_0, u_0) \cdot f(t_0, u_0)\right), \\ &\quad \ldots\ldots \end{aligned} \tag{0.4}$$

Для значений t, близких к t_0, метод рядов (0.3) – (0.4) при достаточно большом m дает хорошее приближение к точному решению $u(t)$ задачи (0.1) – (0.2). Однако с увеличением расстояния $t - t_0$ погрешность приближенного равенства $u(t) \approx u_m(t)$, вообще говоря, возрастает по абсолютной величине, и правило (0.3) становится вовсе непригодным, когда t выходит за пределы области сходимости соответствующего ряда.

Более предпочтительными в таких случаях будут численные методы решения задачи Коши, позволяющие в узлах сетки $t_0 < t_1 < t_2 < \ldots < t_N = T$ последовательно находить значения $y_j \approx u(t_j)$, $j = 1, 2, \ldots, N$ приближенного решения.

Большинство численных методов решения рассматриваемой задачи Коши можно записать в виде

$$y_{j+1} = F(y_{j-q}, y_{j-q+1}, \ldots, y_j, y_{j+1}, \ldots, y_{j+s}), \tag{0.5}$$

где F – некоторая известная функция указанных аргументов, определяемая способом построения метода и зависящая от вида уравнения (0.1) и избранной сетки узлов.

При $q = 0$ и $s \in \{0, 1\}$ такие вычислительные правила называют одношаговыми, а при $q \ge 1$ или $s > 1$ – многошаговыми.

Как одношаговые, так и многошаговые методы вида (0.5) называют явными в случае $s = 0$ и неявными при $s \ge 1$. При $s > 1$ многошаговые правила часто называют методами с забеганием вперед.

Если правило (0.5) является одношаговым, то вычисления по нему можно начинать со значения $j = 0$ и проводить до значения $j = N - 1$ включительно. В случае же многошаговых методов указанного вида, вообще говоря, нарушается однородность вычислительного процесса, и для нахождения первых q значений $y_1, \ldots, y_q$ и последних $s - 1$ значений требуется применение специальных (отличных от базового) правил. В этом смысле одношаговые правила оказываются предпочтительнее. Удобнее пользоваться ими и в том случае, когда шаг сетки $\tau_j = t_{j+1} - t_j$ не является постоянным для всех значений j (например, когда величина τ выбирается компьютером автоматически по результатам вычислений).

Лучше работают одношаговые методы и в областях резкого изменения функций. В то же время, в экономичности решения доступных им задач, как правило, выигрывают многошаговые методы.

Прежде чем перейти к рассмотрению конкретных вычислительных правил, остановимся на некоторых понятиях, используемых в теории и практике решения дифференциальных уравнений.

1. Невязку численного метода (0.5) на точном решении задачи (0.1)

$$r(t_j, \tau) = u(t_{j+1}) - F(u(t_{j-q}), \ldots, u(t_j), u(t_{j+1}), \ldots, u(t_{j+s})) \qquad (0.6)$$

будем называть локальной погрешностью метода (0.5).

2. Величину

$$\psi(t_j, \tau) = \frac{u(t_{j+1}) - u(t_j)}{\tau} - \frac{F(u(t_{j-q}), \ldots, u(t_j), u(t_{j+1}), \ldots, u(t_{j+s})) - u(t_j)}{\tau} \equiv \frac{r(t_j, \tau)}{\tau}$$

будем называть погрешностью аппроксимации дифференциальной задачи (0.1) разностной задачей (0.5).

Как правило, величину $\psi(t_j, \tau)$ (или $r(t_j, \tau)$) стараются разложить в ряд Тейлора по степеням шага сетки τ.

Если при этом $\psi(t_j, \tau) = O(\tau^p)$, $p \geq 1$, то метод (0.5) называют методом p-го порядка точности.

ГЛАВА X

Одношаговые методы решения задачи Коши

Дополнительная литература:

1. Хайрер Э., Нерсетт С., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990.

§ 1. Пошаговый вариант метода рядов

Прежде всего еще раз отметим, что основная отличительная черта численных методов – это тот факт, что они представляют собой алгоритмы вычисления приближенных (а иногда - точных) значений искомого решения $u(t)$ на некоторой выбранной сетке значений аргумента t_j. Они не позволяют найти общее решение системы, а дают какое-либо

частное. При этом одношаговые методы для нахождения решения в очередном узле сетки t_{j+1} используют информацию о задаче только из отрезка $[t_j; t_{j+1}]$.

Итак, будем считать, что процесс решения задачи (0.1), (0.2) доведен до некоторой точки t_j ($0 \le j < N$) (таким образом, нам известны значения y_k, $k = 0,1, \ldots, j$)

Построим сейчас простейший вычислительный алгоритм для нахождения решения в точке $t_{j+1} = t_j + \tau_j$ сетки. Поскольку при построении одношаговых методов используется информация о решаемой задаче лишь в пределах одного шага интегрирования, то можно без ущерба для понимания не писать индекс j, обозначающий номер шага процесса.

Для решения поставленной задачи воспользуемся формулой (0.3), положив в ней вместо t_0 t_j, а вместо $t - t_{j+1}$. В результате получим:

$$y_{j+1} = \sum_{i=0}^{m} \frac{\tau^i}{i!} y_j^{(i)}, \tag{1.1}$$

где производные $y_j^{(i)}$ вычисляются, как и ранее, по формулам (0.4), в которых вместо t_0 будет стоять t_j, а вместо $u - y$.

Очевидно, решение в (1.1) разложено по последовательным главным частям (естественно, при достаточно малых τ) и по величине поправки (т.е. очередного слагаемого в сумме) можно судить о том, с какой локальной погрешностью получено интересующее нас значение.

Конкретными примерами методов типа (1.1) могут служить:

1. $m = 1$: - метод первого порядка

$$y_{j+1} = y_j + \tau f(t_j, y_j) \tag{1.2}$$

(другие названия метода (1.2) – явный метод Эйлера, метод ломаных);

2. $m = 2$: - метод второго порядка

$$y_{j+1} = y_j + \tau f(t_j, y_j) + \frac{\tau^2}{2}\left(f_t(t_j, y_j) + f(t_j, y_j) f_u(t_j, y_j)\right). \tag{1.3}$$

Поменяв точку, в окрестности которой строится разложение, с t_0 на t_j, мы в какой-то степени ослабили первый из недостатков изложенного в § 0 метода рядов, связанный со сходимостью. В то же время, второй недостаток, связанный с необходимостью нахождения большого числа различных функций $\left(\frac{m(m+1)}{2}\right)$ здесь не только не исчезает, а наоборот, усиливается, поскольку делать это теперь нужно на каждом шаге. Поэтому при $m > 1$ указанный метод применяется редко.

§ 2. Способ Рунге-Кутта построения одношаговых методов

Проинтегрировав уравнение (0.1) по отрезку $[t_j; t_{j+1}]$, получим равенство

$$u(t_j + \tau) = u(t_j) + \int_{t_j}^{t_{j+1}} f(x, u(x))dx, \tag{2.1}$$

которое связывает значения решения рассматриваемого уравнения в двух соседних узлах сетки. Указав эффективный способ вычисления интеграла в (2.1), мы получим одно из приближенных правил численного интегрирования уравнения (0.1). В силу требования одношаговости конструируемых правил при нахождении значения указанного интеграла мы можем использовать информацию о функции $f(t,u(t))$ лишь из отрезка $\lfloor t_j;\, t_j+\tau \rfloor$. По постановке задачи значение этой функции в точке t_j нам известно. Поэтому для вычисления интеграла можно применить, например, квадратурную формулу левых прямоугольников

$$\int_a^b \varphi(x)dx \approx (b-a)\varphi(a).$$

В результате получим метод (первого порядка точности),

$$y_{j+1} = y_j + \tau f(t_j, y_j),$$

полностью совпадающий с (1.2).

Аналогично, применив для вычисления интеграла в) (2.1 формулу правых прямоугольников, получим неявный метод Эйлера

$$y_{j+1} = y_j + \tau f(t_{j+1}, y_{j+1}), \tag{2.2}$$

который также является методом первого порядка.

Воспользовавшись для замены интеграла в (2.1) формулой трапеций, получим одношаговый метод численного интегрирования (неявный метод трапеций)

$$y_{j+1} = y_j + \frac{\tau}{2}\left(f(t_j, y_j) + f(t_{j+1}, y_{j+1})\right), \tag{2.3}$$

являющийся методом второго порядка.

Заметим, что методы (2.2) и (2.3) (как и любой другой неявный метод) требуют для определения искомого решения y_{j+1} на каждом шаге решать нелинейное уравнение, что далеко не всегда просто.

Встает вопрос: как построить явные одношаговые методы, более точные, чем (1.2), и не использующие производных от функции f.

В способе Рунге-Кутта предлагается использовать следующий специальный прием приближенного вычисления интеграла в (2.1).

Введем обозначение $\Delta u = u(t_j+\tau) - u(t_j)$ и сделаем в (2.1) замену переменной интегрирования по формуле $x = t_j + \alpha\tau$. Тогда (2.1) можно переписать в виде

$$\Delta u = \tau \int_0^1 f(t_j + \alpha\tau, u(t_j + \alpha\tau))d\alpha\,. \tag{2.4}$$

Зададимся тремя наборами параметров:

$c_1, c_2, \ldots, c_s$ (вектор $c = (c_1, c_2, \ldots, c_s)^T$),

$a_{11}, a_{12}, \ldots, a_{1s}$

$\ldots$ (матрица $A_{s\times s}$),

$a_{s1}, a_{s2}, \ldots, a_{ss}$

$b_1, b_2, \ldots, b_s$ (вектор $b = (b_1, b_2, \ldots, b_s)^T$).

При помощи первых двух наборов построим величины

$$k_1 = f\left(t_j + c_1\tau, y_j + \tau\sum_{l=1}^{s} a_{1l}k_l\right),$$

$$k_2 = f\left(t_j + c_2\tau, y_j + \tau\sum_{l=1}^{s} a_{2l}k_l\right), \qquad (2.5)$$

$$\ldots\ldots$$

$$k_s = f\left(t_j + c_s\tau, y_j + \tau\sum_{l=1}^{s} a_{sl}k_l\right).$$

Каждая из величин $k_i = f\left(t_j + c_i\tau, y_j + \tau\sum_{l=1}^{s} a_{il}k_l\right)$, вообще говоря, не равна значению $f\left(t_j + c_i\tau, u\left(t_j + c_i\tau\right)\right)$, однако при соответствующем выборе параметров их можно надеяться сделать близкими. А это, в свою очередь, дает основание надеяться при помощи параметров b_i составить такую линейную комбинацию величин k_i $\left(i = \overline{1,s}\right)$, которая будет являться аналогом квадратурной суммы и позволит вычислить приближенное значение приращения:

$$\Delta y = \tau\sum_{i=1}^{s} b_i k_i$$

или

$$y_{j+1} = y_j + \tau\sum_{i=1}^{s} b_i k_i\ . \qquad (2.6)$$

(2.5), (2.6) в литературе носит название s-стадийного метода Рунге-Кутта.

Иногда его записывают в сокращенном виде

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \qquad (2.7)$$

Последний носит название таблицы Батчера.

При этом, если $a_{in} = 0$ при $n \geq i$ для всех i (и $c_1 = 0$), то вектор k_i может быть вычислен явным образом по значениям $k_1, \ldots, k_{i-1}$. Поэтому такие методы называют ***явными*** методами Рунге-Кутта.

Если $a_{in} = 0$ при $n > i$ и хотя бы при одном значении i $a_{ii} \neq 0$, то получающиеся методы носят название ***диагонально неявных***, и, наконец, если среди элементов матрицы A имеются отличные от нуля и выше ее главной диагонали, то такие методы называют просто ***неявными***.

Мы в ближайшее время будем заниматься ***явными*** методами Рунге-Кутта.

Итак, построить метод Рунге-Кутта – значит указать конкретную таблицу Батчера (или, что то же, - конкретные наборы параметров c, A и b).

Естественно выбирать их таким образом, чтобы конструируемый метод (2.6), (2.5) (или (2.7)) имел по возможности более высокий порядок точности. В предположении, что правая часть уравнения (0.1) (функция f) является достаточно гладкой, можно записать разложение локальной погрешности формулы (2.6) $r(t_j,\tau)$ в ряд Тейлора с остаточным членом в форме Лагранжа:

$$r(t_j,\tau)=\Delta u-\tau\sum_{i=1}^{s}b_ik_i=\sum_{l=0}^{k}\frac{\tau^l}{l!}r^{(l)}(t_j,0)+\frac{\tau^{k+1}}{(k+1)!}r^{(k+1)}(t_j,\theta\tau),\quad 0<\theta<1.$$

Если теперь подобрать параметры c, A и b так, чтобы выполнялись условия

$$r^{(l)}(t_j,0)=0,\quad l=0,1,\ldots,k, \tag{2.8}$$

то локальная погрешность метода примет вид

$$r(t_j,\tau)=\frac{\tau^{k+1}}{(k+1)!}r^{(k+1)}(t_j,\theta\tau), \tag{2.9}$$

т.е. метод будет методом k-го порядка точности.

Практически при построении методов может оказаться более целесообразной следующая схема действий: составляют разложение по степеням τ величины

$$u(t_j+\tau)=u(t_j)+\tau u'(t_j)+\frac{\tau^2}{2}u''(t_j)+\cdots=u+\tau f+\frac{\tau^2}{2}(f_t+f_uf)+\cdots.$$

Аналогичное разложение составляется для правой части формулы (2.6) (на точном решении задачи (0.1)), т.е. для комбинации

$$u(t_j)+\tau\sum_{i=1}^{s}b_ik_i=u(t_j)+\tau\sum_{i=1}^{s}b_if\left(t_j+c_i\tau,u(t_j)+\tau\left(a_{i1}k_1+\cdots+a_{ii-1}k_{i-1}\right)\right).$$

После этого требуют, чтобы полученные разложения совпадали до членов с возможно более высокими степенями τ для произвольной функции f.

При произвольных s и k систему уравнений для определения параметров c, A и b записать достаточно сложно и мы этим займемся немного позже. А пока рассмотрим применение указанного подхода к построению конкретных примеров методов невысоких порядков точности.

2.1. Методы первого порядка точности

Зададимся минимальным значением $s=1$, что равнозначно введению лишь одного параметра b_1. Равенство (2.6) в этом случае будет иметь вид

$$y_{j+1}=y_j+\tau b_1k_1=y_j+\tau f(t_j,y_j),$$

а погрешность $r(t_j,\tau)$ примет вид

$$r(t_j,\tau)=u(t_j+\tau)-u(t_j)-\tau b_1 f(t_j,u(t_j)).$$

Учитывая простоту конструкции, непосредственно вычислим производные от локальной погрешности:

$$r'(t_j,\tau)=u'(t_j+\tau)-b_1 f(t_j,u(t_j)),$$

$$r''(t_j,\tau)=u''(t_j+\tau).$$

Так как $r''(t_j,\tau)$ не зависит от b_1, то уже при $l=2$ условие (2.8) в случае произвольной функции f удовлетворено быть не может. Поэтому $k=1$ и система (2.8) принимает вид

$$r'(t_j,0)=u'(t_j)-b_1 f(t_j,u(t_j))=(1-b_1)f(t_j,u(t_j))=0\,.$$

Отсюда следует, что $b_1=1$. В этом случае получим хорошо уже знакомый нам явный метод Эйлера, который в компактной записи (2.7) имеет вид

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Его локальная погрешность, согласно (2.9), будет иметь вид

$$r(t_j,\tau)=\frac{\tau^2}{2}r''(t_j,\theta\tau)=\frac{\tau^2}{2}u''(t_j+\theta\tau).$$

2.2. Методы второго порядка точности

Положим $s=2$ (при $s=1$, как мы видели, явных методов порядка выше первого получить нельзя). Тогда

$$y_{j+1}=y_j+\tau(b_1k_1+b_2k_2)\,.$$

Выполняя указанные выше разложения правой и левой частей последней формулы на точном решении в ряды по степеням τ, имеем:

$$u(t_j+\tau)=u(t_j)+\tau u'(t_j)+\frac{\tau^2}{2}u''(t_j)+\frac{\tau^3}{6}u'''(t_j)+O(\tau^4)=$$

$$=u+\tau f+\frac{\tau^2}{2}(f_t+f_u f)+\frac{\tau^3}{6}\left(f_{tt}+2f_{tu}f+f_{uu}f^2+f_u(f_t+f_u f)\right)+O(\tau^4). \qquad (*)$$

$$u(t_j)+\tau(b_1k_1+b_2k_2)=u(t_j)+\tau b_1 f(t_j,u(t_j))+\tau b_2 f(t_j+c_2\tau,u(t_j)+\tau a_{21}f(t_j,u(t_j)))=$$

$$=u+\tau b_1 f+\tau b_2\left[f+c_2\tau f_t+\tau a_{21}f_u f+\frac{c_2^2\tau^2}{2}f_{tt}+c_2a_{21}\tau^2 f_{tu}f+\frac{a_{21}^2\tau^2}{2}f_{uu}f^2\right]+O(\tau^4)=\ (**)$$

$$=\tau(b_1+b_2)f+\tau^2 b_2(c_2f_t+a_{21}f_u f)+\frac{\tau^3 b_2}{2}(c_2^2f_{tt}+2c_2a_{21}f_{tu}f+a_{21}^2f_{uu}f^2)+O(\tau^4).$$

Приравняем в правых частях разложений (*) и (**) коэффициенты при одинаковых степенях τ и сомножителях, зависящих от f, одинакового вида.

Тем самым на выбор четырех параметров c_2, a_{21}, b_1 и b_2 будут наложены три условия:

$$\begin{cases} b_1+b_2=1, \\ 2b_2c_2=1, \\ 2b_2a_{21}=1. \end{cases} \qquad (2.10)$$

Непосредственно из разложений следует, что в случае $s=2$ для произвольных f нельзя добиться совпадения всех членов с множителем τ^3 за счет выбора введенных параметров. Поэтому при $s=2$ максимальный порядок точности правил типа Рунге-Кутта равен 2. С другой стороны, система (2.10), очевидно, имеет бесчисленное множество решений, например, такое: $c_2=a_{21}=\alpha;\ b_2=\frac{1}{2\alpha};\ b_1=1-\frac{1}{2\alpha}$, где в качестве α может быть взято, вообще говоря, любое отличное от нуля число (хотя требование одношаговости, более естественными будут $\alpha\in(0;1]$). Таким образом, существует однопараметрическое семейство методов второго порядка точности, задаваемое таблицей Батчера

0	0	0
α	α	0
	$1-\frac{1}{2\alpha}$	$\frac{1}{2\alpha}$

или

$$\begin{cases} y_{j+1}=y_j+\frac{\tau}{2\alpha}((2\alpha-1)k_1+k_2), \\ k_1=f(t_j,y_j), \\ k_2=f(t_j+\alpha\tau,y_j+\tau\alpha k_1). \end{cases} \qquad (2.11)$$

Наиболее употребительными из формул (2.11) являются их частные случаи при $\alpha=\frac{1}{2}$, т.е.

0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0
	0	1

или

$$\begin{cases} y_{j+1}=y_j+\tau k_2, \\ k_1=f(t_j,y_j), \\ k_2=f\left(t_j+\frac{\tau}{2},y_j+\frac{\tau}{2}k_1\right). \end{cases} \qquad (2.12)$$

(аналог формулы средних прямоугольников (!)), и $\alpha=1$:

0	0	0
1	1	0
	$\frac{1}{2}$	$\frac{1}{2}$

или

$$\begin{cases} y_{j+1} = y_j + \dfrac{\tau}{2}(k_1 + k_2), \\ k_1 = f(t_j, y_j), \\ k_2 = f(t_j + \tau, y_j + \tau k_1). \end{cases} \qquad (2.13)$$

(аналог формулы трапеций (!)).

Локальная погрешность любого из методов типа (2.11), как следует из разложений (*) и (**), может быть представлена в виде

$$r(t_j, \tau) = \frac{\tau^3}{6}\left[f_{tt}(1 - 3c_2^2 b_2) + 2ff_{tu}(1 - 3c_2 a_{21} b_2) + f^2 f_{uu}(1 - 3a_{21}^2 b_2) + f_u(f_t + f_u f)\right] + O(\tau^4). \qquad (2.14)$$

Исходя из этого, свободный параметр α иногда выбирают таким образом, чтобы в этом представлении обратилась в нуль хотя бы часть слагаемых. Так как $c_2 = a_{21} = \alpha$, $b_2 = \dfrac{1}{2\alpha}$, то в этом случае получим:

$$1 - 3\alpha^2 \cdot \frac{1}{2\alpha} = 0,$$

откуда $\alpha = \dfrac{2}{3}$. При таком выборе α (2.14) существенно упростится:

$$r(t_j, \tau) = \frac{\tau^3}{6} f_u(f_t + f_u f) + O(\tau^4),$$

а соответствующий метод Рунге-Кутта будет иметь вид

0	0	0
$\frac{2}{3}$	$\frac{2}{3}$	0
	$\frac{1}{4}$	$\frac{3}{4}$

или

$$\begin{cases} y_{j+1} = y_j + \dfrac{\tau}{4}(k_1 + 3k_2), \\ k_1 = f(t_j, y_j), \\ k_2 = f\left(t_j + \dfrac{2\tau}{3}, y_j + \dfrac{2\tau}{3} k_1\right). \end{cases} \qquad (2.15)$$

2.3. Условия порядка для явных методов Рунге-Кутта

Изучим сейчас общую структуру условий, определяющих порядок метода. Для упрощения вывода преобразуем уравнение (0.1) к автономной форме путем добавления t к зависимым переменным:

$$\begin{pmatrix} t \\ u \end{pmatrix}' = \begin{pmatrix} 1 \\ f(t,u) \end{pmatrix}. \qquad (0.1')$$

Таким образом, вместо одного уравнения (в скалярном случае) мы имеем уже систему. Поэтому в дальнейшем (в этом пункте) будем иметь дело с системами, при этом компоненты векторов мы будем обозначать верхними заглавными буквами. Тогда автономную систему общего вида можно записать так:

$$\left(u^J\right)' = f^J\left(u^1,...,u^n\right),\ J=1,...,n \tag{2.16}$$

Зафиксируем также в схеме (2.6) узел сетки (т.е. будем полагать $t_j = t_0$) и сделаем запись формул (2.6) более симметричной, перейдя от функций $k_i = f\left(g_i\right)$ к их аргументам:

$$\begin{cases} g_i^J = y_0^J + \sum\limits_{j=1}^{i-1} a_{ij}\tau f^J\left(g_j^1,\ldots,g_j^n\right),\ i=1,\ldots s, \\ y_1^J = y_0^J + \sum\limits_{j=1}^{s} b_j\tau f^J\left(g_j^1,\ldots,g_j^n\right). \end{cases} \tag{2.17}$$

В частности, если система (2.16) получается из (0.1′), то (2.17) при $J=1$ дает

$$g_i^1 = y_0^1 + \sum_{j=1}^{i-1} a_{ij}\tau\ . \tag{*}$$

Обычно в явных методах Рунге-Кутта параметры c_i таблицы Батчера удовлетворяют условию (и мы это видели в частных случаях)

$$c_i = \sum_j a_{ij}\ . \tag{2.18}$$

С учетом последнего условия равенство (*) примет вид

$$g_i^1 = y_0^1 + \sum_{j=1}^{i-1} a_{ij}\tau = t_0 + c_i\tau\ ,$$

т.е. будет соответствовать «штатному» виду формул (2.6) для неавтономной системы.

Таким образом, если выполнено условие (2.18), то для вывода условий порядка достаточно рассмотреть автономную систему (2.16).

Как мы уже отмечали ранее, для получения условий порядка нужно сравнивать ряды Тейлора для u_1^J и выражения, стоящего в правой части формул (2.17) (естественно, при подстановке туда точного решения u вместо y). Для этой цели вычислим сначала значения производных u_1^J и g_i^J по τ при $\tau = 0$. Ввиду внешнего сходства обеих формул (2.17) достаточно проделать это для g_i^J. В правые части этих формул входят выражения вида $\tau\varphi(\tau)$ и мы воспользуемся формулой Лейбница

$$\left(\tau\varphi(\tau)\right)^{(q)}\Big|_{\tau=0} = q\cdot\left(\varphi(\tau)\right)^{(q-1)}\Big|_{\tau=0} \tag{2.19}$$

Имеем:

$q=0$: $$\left(g_i^J\right)^{(0)}\Big|_{\tau=0} = u_0^J, \tag{П.0}$$

$q=1$:

$$\left(g_i^J\right)^{(1)}\Big|_{\tau=0}=\sum_j a_{ij}f^J, \tag{П.1}$$

$q=2$:

Так как

$$\left(f^J\left(g_j\right)\right)^{(1)}=\sum_K f_K^J\left(g_j\right)\cdot\left(g_j^K\right)^{(1)}, \tag{Ф.1}$$

где $f_K^J=\dfrac{\partial f^J}{\partial u^K}$ (формула производной сложной функции), то

$$\left(g_i^J\right)^{(2)}\Big|_{\tau=0}=\left[\varphi(\tau)=\sum_j a_{ij}f^J\left(g_j\right)\right]=2\sum_j a_{ij}\left(f^J\left(g_j\right)\right)^{(1)}\Big|_{\tau=0}=2\sum_j a_{ij}\sum_K f_K^J\left(g_j\right)\left(g_j^K\right)^{(1)}\Big|_{\tau=0}=$$

$$=2\sum_j a_{ij}\sum_K f_K^J\sum_k a_{jk}f^K=2\sum_{j,k}a_{ij}a_{jk}\sum_K f_K^J f^K. \tag{П.2}$$

Аналогично

$$\left(g_i^j\right)^{(3)}\Big|_{\tau=0}=3\sum_j a_{ij}\left(f^J\left(g_j\right)\right)^{(2)}\Big|_{\tau=0}. \tag{2.20}$$

Продифференцировав (Ф.1), получим:

$$\left(f^J\left(g_j\right)\right)^{(2)}=\left(\sum_K f_K^J\left(g_j\right)\left(g_j^K\right)^{(1)}\right)^{(1)}=\sum_{K,L}f_{KL}^J\left(g_j\right)\left(g_j^L\right)^{(1)}\left(g_j^K\right)^{(1)}+\sum_K f_K^J\left(g_j\right)\left(g_j^K\right)^{(2)} \tag{Ф.2}$$

Подставляя это выражение в (2.20), будем иметь:

$$\left(g_i^j\right)^{(3)}\Big|_{\tau=0}=3\sum_j a_{ij}\left[\sum_{K,L}f_{KL}^J\left(g_j\right)\left(g_j^L\right)^{(1)}\left(g_j^K\right)^{(1)}+\sum_K f_K^J\left(g_j\right)\left(g_j^K\right)^{(2)}\right]\Bigg|_{\tau=0}=$$

$$=3\sum_{j,k,l}a_{ij}a_{jk}a_{jl}\sum_{K,L}f_{KL}^J f^K f^L+6\sum_{j,k,l}a_{ij}a_{jk}a_{kl}\sum_{K,L}f_K^J f_L^K f^L \tag{П.3}$$

Для правой части последней из формул (2.17) ввиду симметрии будут справедливы те же формулы (П.1) – (П.3), если в них заменить a_{ij} на b_j.

Найдем теперь производные точного решения:

$$\left(u^J\right)^{(1)}=f^J(u) \tag{Т.1}$$

$$\left(u^J\right)^{(2)}=\sum_K f_K^J(u)\cdot\left(u^K\right)^{(1)}=\sum_K f_K^J f^K \tag{Т.2}$$

$$\left(u^J\right)^{(3)}=\left(\sum_K f_K^J f^K\right)^{(1)}=\sum_K\sum_L\left(f_{KL}^J f^L f^K+f_K^J f_L^K f^L\right)=$$

$$=\sum_{K,L}f_{KL}^J f^K f^L+\sum_{K,L}f_K^J f_L^K f^L \tag{Т.3}$$

Учитывая полученные формулы, легко выписать условия, при которых метод Рунге-Кутта имеет третий порядок:

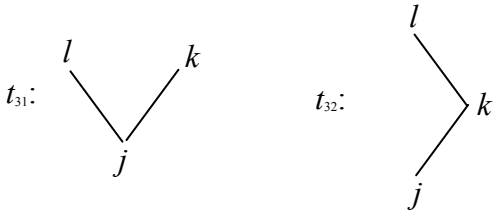
$$\begin{cases} \sum\limits_j b_j = 1, \\ 2\sum\limits_{j,k} b_j a_{jk} = 1, \\ 3\sum\limits_{j,k,l} b_j a_{jk} a_{jl} = 1, \\ 6\sum\limits_{j,k,l} b_j a_{jk} a_{kl} = 1. \end{cases} \tag{2.21}$$

Если к ним теперь добавить условия (2.18), то мы получим условия порядка методов Рунге-Кутта в том виде, в котором они встречаются в литературе:

$$\begin{cases} \sum\limits_j b_j = 1, \\ \sum\limits_j b_j c_j = \frac{1}{2}, \\ \sum\limits_j b_j c_j^2 = \frac{1}{3}, \\ \sum\limits_{j,k} b_j a_{jk} c_k = \frac{1}{6}, \\ \sum\limits_j a_{ij} = c_i . \end{cases} \tag{2.22}$$

2.3.1. Деревья и элементарные дифференциалы

Проведенная выше унификация работы позволила относительно несложно получить при произвольном числе стадий условия третьего порядка. Тем не менее, получающиеся формулы достаточно громоздки. Поэтому перейдем к графическому представлению рассмотренных выше операций дифференцирования. Для этого заметим, что в каждом члене формулы (П.3) индексы j, k, l связаны в пары в качестве нижних индексов в коэффициентах a_{jk}, a_{jl}, a_{kl}, и точно таким же образом индексы J, K, L связаны попарно в качестве верхних и нижних индексов в выражениях f_{KL}^J, f_K^J, f_L^K. Графически эти связи для первого и второго членов соответственно можно представить следующим образом:



Назовем эти объекты «помеченными деревьями», потому что это связные графы (деревья), вершины которых помечены индексами суммирования. Их можно представить и как ***отображения*** $l \to j, k \to j$ и $l \to k,\ k \to j$. Эти отображения каждой вершине графа (кроме одной) ставят в соответствие другую вершину, связанную с ней направленным вниз ребром.

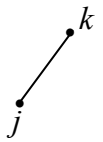
***Определение* 1**. Пусть A – упорядоченное множество индексов: $A=\{j<k<l<m<\cdots\}$ и A_q – его подмножество, состоящее из первых q индексов. Назовем (корневым) помеченным деревом порядка q $(q\geq 1)$ отображение $t: A_q \setminus \{j\} \to A_q$ такое, что $t(z)<z$ для всех $z \in A_q \setminus \{j\}$.

Множество всех помеченных деревьев порядка q обозначим LT_q. z будем называть сыном $t(z)$, а $t(z)$ – отцом z (t – отображение сыновей на отцов). Вершина j (праотец всей династии) называется корнем дерева t. Порядок q помеченного дерева равен числу всех его вершин. Будем обозначать его $\rho(t)$.

Примеры:

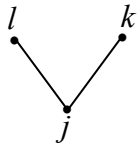
$q=1 \Rightarrow A_1=\{j\},\ A_1 \setminus \{j\}=\varnothing$; граф выглядит так:

$q=2 \Rightarrow A_2=\{j,k\},\ A_2 \setminus \{j\}=\{k\} \Rightarrow t(k)=j$; граф:

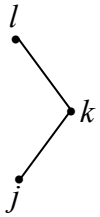


$q=3 \Rightarrow A_3=\{j,k,l\},\ A_3 \setminus \{k,l\} \Rightarrow$

а) $t(k)=j,\ t(l)=j$; граф:



б) $t(k)=j,\ t(l)=k$; граф:



Упражнение. Найти множество LT_4.

***Определение* 2.** Назовем выражение

$$F^J(t)(u)=\sum_{K,L,\ldots} f^J_{K\ldots}(u) f^K_{\ldots}(u) f^L_{\ldots}(u)\ldots \tag{2.23}$$

элементарным дифференциалом, соответствующим помеченному дереву $t \in LT_q$.

Здесь суммирование производится по $q-1$ индексам $K, L, \ldots$ (которые соответствуют множеству $A_q \setminus \{j\}$) и каждое слагаемое является произведением q символов f, верхние индексы которых пробегают все множество A_q вершин дерева t, а нижние индексы – множество соответствующих сыновей.

Если множество A_q записать в виде $A_q=\{j_1<j_2<\cdots<j_q\}$, то определение для $F(t)$ можно записать следующим образом:

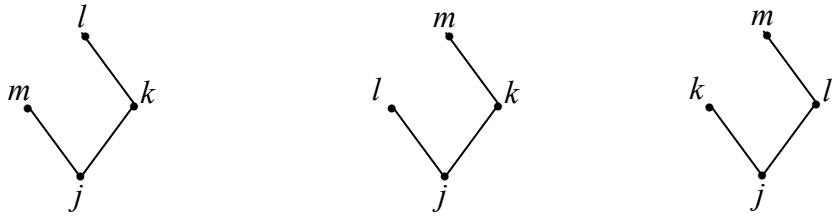
$$F^{J_1}(t)(u)=\sum_{J_2,\ldots,J_q} \prod_{i=1}^{q} f^{J_i}_{t^{-1}(J_i)}(u), \tag{2.24}$$

поскольку сыновья индекса составляют его полный прообраз при отображении t.

Примерами элементарных дифференциалов являются

$$F^J(t_{31})(u)=\sum_{K,L} f^J_{KL} f^K f^L, \quad F^J(t_{32})(u)=\sum_{K,L} f^J_K f^K_L f^L .$$

Теперь обратим внимание на то, что, например, следующие три помеченных дерева из LT_4



имеют одинаковую топологическую структуру; более того, их элементарные дифференциалы $\sum_{K,L,M} f^J_{KM} f^K_L f^M f^L$, $\sum_{K,L,M} f^J_{KL} f^K_M f^L f^M$ и $\sum_{K,L,M} f^J_{KL} f^K f^L_M f^M$ совпадают, поскольку они отличаются лишь обозначениями индексов суммирования.

***Определение* 3.** Два помеченных дерева t и u назовем эквивалентными, если они имеют одинаковый порядок q и существует подстановка $\sigma : A_q \to A_q$ такая, что $\sigma(j) = j$ и $t\sigma = \sigma u$ на множестве $A_q \setminus \{j\}$.

Таким образом, мы определили на множестве LT_q отношение эквивалентности.

Определение* 4.** Класс эквивалентности помеченных деревьев порядка q назовем (корневым***) ***деревом порядка*** q.

Множество всех деревьев порядка q обозначим T_q. Порядок дерева определим как порядок его представителя из LT_q и тоже обозначим $\rho(t)$. Через $\alpha(t)$ будем обозначать число элементов в классе эквивалентности t, т.е. фактически число возможных монотонных индексаций t.

Геометрически дерево отличается от помеченного дерева тем, что у него опущены все индексы. Часто бывает целесообразно включать в рассмотрение пустое дерево $\varnothing$ как единственное дерево порядка 0. Единственное дерево порядка 1 обозначают τ.

2.3.2. Разложение Тейлора для точного решения

Теперь, используя описанные выше объекты, мы можем для q-й производной точного решения сформулировать результат в общем виде.

***Теорема* 1.** Для точного решения системы (2.16) справедлива формула

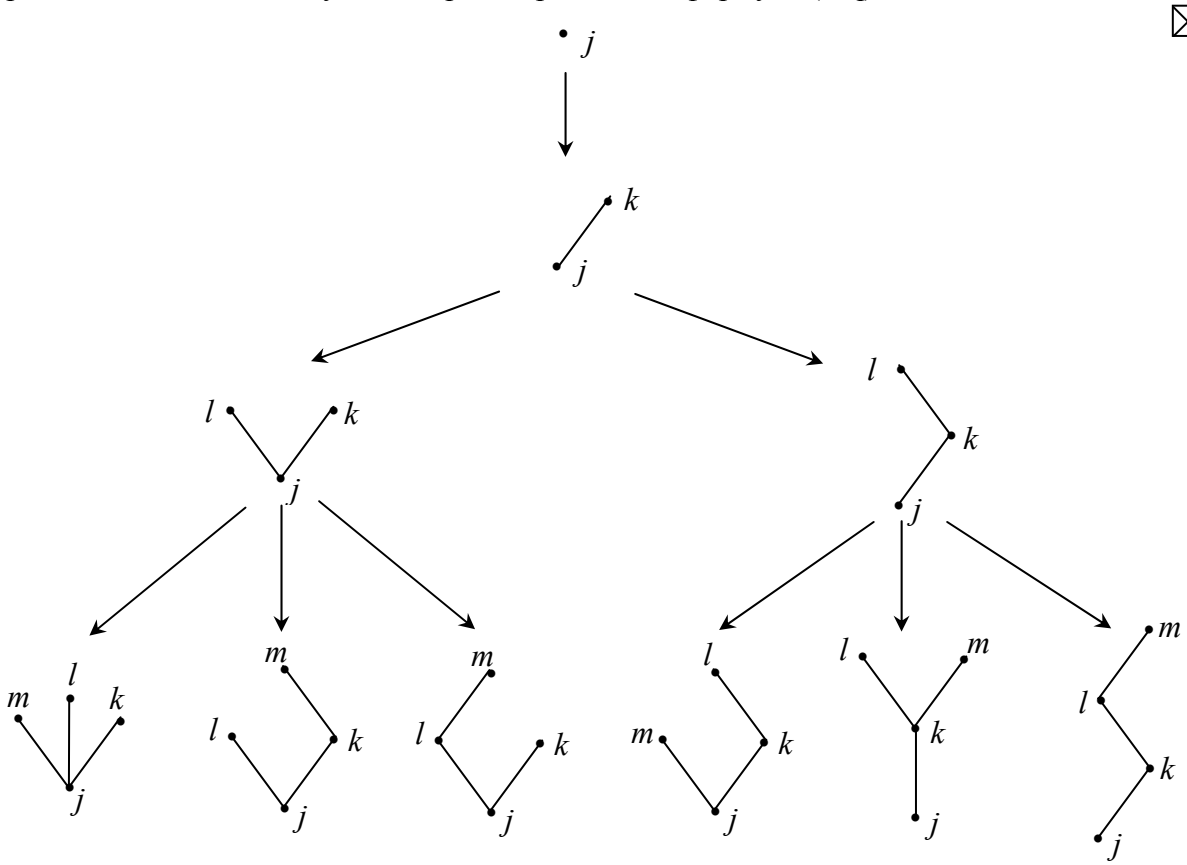
$$\left(u^J\right)^{(q)}\Big|_{t=t_0} = \sum_{t\in LT_q} F^J(t)(u_0) = \sum_{t\in T_q} \alpha(t) F^J(t)(u_0). \qquad \text{(T.q)}$$

Доказательство.

В соответствии с методом математической индукции получаем: теорема справедлива при $q = 1, 2, 3$ (см. формулы (T.1) – (T.3)). Индуктивный переход осуществляется с помощью следующих соображений: чтобы вычислить производную порядка $(q+1)$, необходимо продифференцировать формулу (T.q), в которой $(q-1)!$ членов соответствуют всем помеченным деревьям порядка q и каждый из которых содержит q сомножителей вида $f^{\cdots}_{\cdots}$, отвечающих q вершинам этих деревьев. Дифференцирование этих членов по правилу Лейбница и подстановка правой части уравнения (2.16) вместо производных u' геометрически интерпретируется как добавление к каждой вершине нового ребра с вершиной, помеченной новым индексом суммирования. Очевидно, в этом процессе появляются все помеченные деревья порядка $q+1$ и каждое из них только один раз (поскольку каждое из

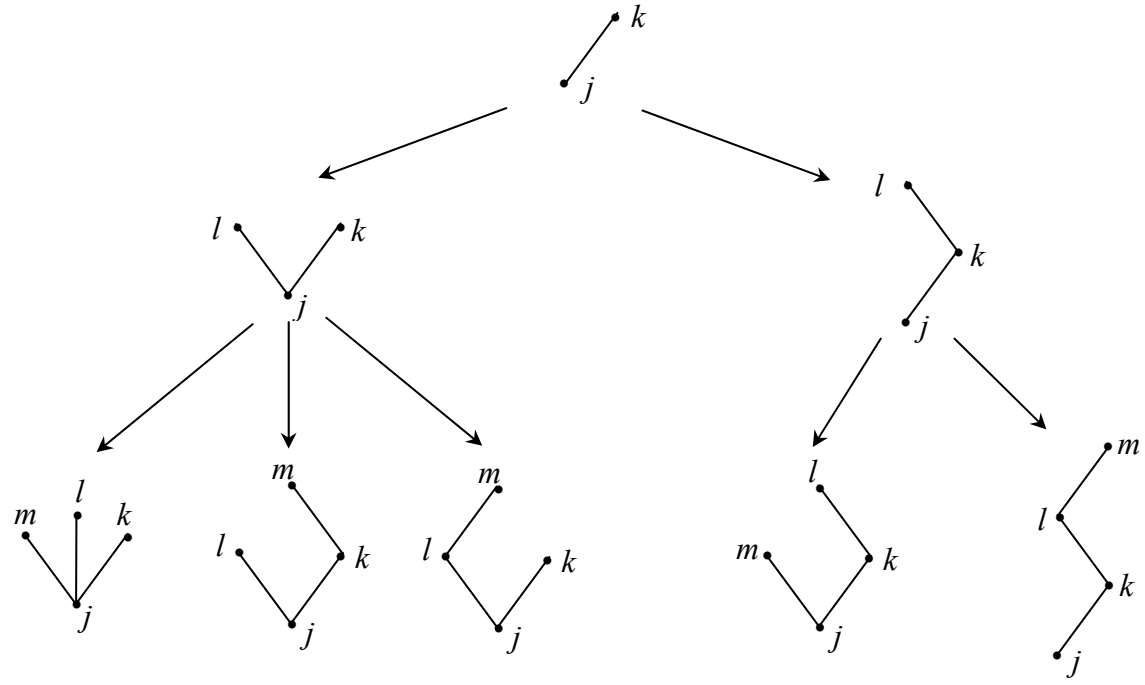
$(q-1)!$ деревьев порождает q новых деревьев, то их общее количество будет равно $q!$). (Поясняющую картинку см. ниже).

Если после этого сгруппировать все члены с одинаковыми элементарными дифференциалами, то мы получим второе выражение в формуле (T.q).



2.3.3. Формула Фаа ди Бруно

Займемся теперь вычислением q-й производной численного решения u_1 и вели-



чин g_j (т.е. правой части формул (2.17) на точном решении).Для этого сначала необходимо получить обобщение формул (Ф.1) и (Ф.2) на случай q-й производной от суперпозиции двух функций. Проведем соответствующий вывод, используя поясняющую картинку.

Формула (Ф.2) состоит из двух членов, причем первый член содержит три множителя, а второй – только два, так что в его гра́фе индекс l – «пустой». В формуле он отсутствует, его назначение – показать, что нужно взять вторую производную. Следовательно, дифференцируя выражение (Ф.2), мы получим *пять* членов:

$$\left(f^J\left(g_j\right)\right)^{(3)} = \sum_{K,L,M} f^J_{KLM}\left(g_j\right)\left(g^K_j\right)^{(1)}\left(g^L_j\right)^{(1)}\left(g^M_j\right)^{(1)} + \sum_{K,L} f^J_{KL}\left(g_j\right)\left(g^K_j\right)^{(2)}\left(g^L_j\right)^{(1)} +$$

$$+ \sum_{K,L} f^J_{KL}\left(g_j\right)\left(g^K_j\right)^{(1)}\left(g^L_j\right)^{(2)} + \sum_{K,M} f^J_{KM}\left(g_j\right)\left(g^K_j\right)^{(2)}\left(g^M_j\right)^{(1)} + \sum_{K} f^J_{K}\left(g_j\right)\left(g^K_j\right)^{(3)} \qquad (Ф.3)$$

Соответствующие деревья изображены в третьем ряду на рисунке. При каждом дифференцировании мы производим следующие операции:

1) дифференцируем первый множитель $f^J_{K\ldots}$, т.е. добавляем к дереву новую ветвь, растущую от корня j;
2) увеличиваем на единицу порядок производной каждого из множителей g, что графически представляется удлинением соответствующей ветви.

Каждый раз мы добавляем новый индекс. Все полученные при этом деревья имеют «специальный» вид: разветвления у них возможны только у корня.

***Определение* 5.** Обозначим LS_q множество *специальных помеченных деревьев* порядка q, т.е. таких помеченных деревьев, у которых разветвления встречаются только у корня.

Лемма (формула Фаа ди Бруно). При $q \ge 1$ справедлива формула

$$\left(f^J\left(g\right)\right)^{(q-1)} = \sum_{u \in LS_q} \sum_{K_1,\ldots,K_m} f^J_{K_1\ldots K_m}\left(g\right)\left(g^{K_1}\right)^{(\delta_1)}\ldots\left(g^{K_m}\right)^{(\delta_m)}. \qquad (Ф.q\text{-}1)$$

Для каждого $u \in LS_q$ здесь m – число выходящих из корня ветвей, $\delta_1,\ldots,\delta_m$ – количество вершин на этих ветвях, так что $q = 1 + \delta_1 + \cdots + \delta_m$.

Доказательство данного утверждения проводится с применением метода математической индукции по аналогии с доказательством теоремы 1.

Замечание. В формуле (Ф.q-1) отсутствуют коэффициенты, указывающие кратности членов, так как мы пользуемся помеченными деревьями.

2.3.4. Производные численного решения

Пусть t – помеченное дерево с корнем j. Введем величину

$$\Phi_j(t) = \sum_{k,l,\ldots} a_{jk}a_{\ldots}\ldots, \qquad (2.25)$$

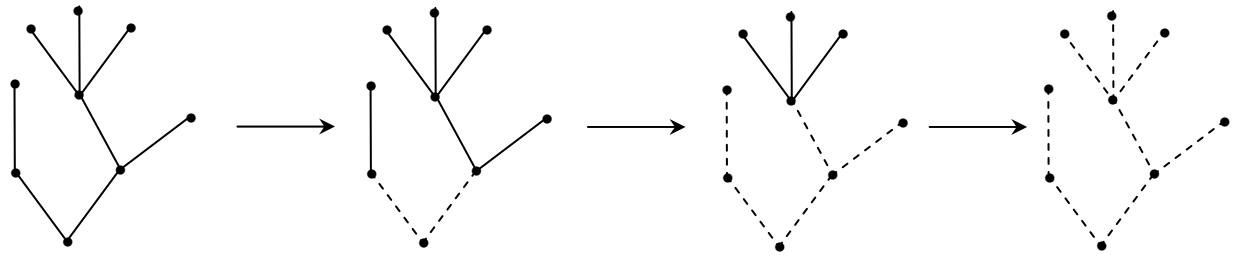
где суммирование производится по $q-1$ индексам k, l, … (как в *Определении* 2), каждое слагаемое является произведением $q-1$ коэффициентов a, у каждого из которых индексы составляют одну из пар отец-сын дерева t.

Если множество A_q записано в виде $A_q = \{j_1 < j_2 < \ldots < j_q\}$, то (2.25) примет вид

$$\Phi_{j_1}(t) = \sum_{j_2,\ldots,j_q} a_{t(j_2)j_2} \ldots a_{t(j_q)j_q} \,. \tag{2.26}$$

Для каждого дерева $t \in LT_q$ определим величину $\gamma(t)$ как произведение $\rho(t)$ на порядки всех деревьев, которые получаются из t, если последовательно удалять их корни (вместе с инцидентными им ребрами).

Пример:



$$\gamma(t) = 9 \times 2 \cdot 6 \times 1 \cdot 4 \cdot 1 \times 1 \cdot 1 \cdot 1 = 9 \cdot 12 \cdot 4 = 432$$

***Теорема* 2.** Для производных g_i^J имеет место равенство

$$\left(g_i^J\right)^{(q)}\Big|_{\tau=0} = \sum_{t \in LT_q} \gamma(t) \sum_j a_{ij} \Phi_j(t) F^J(t)(u_0) \tag{П.q}$$

Для производных определяемого методом (2.17) численного решения справедлива формула

$$\left(u_1^J\right)^{(q)}\Big|_{\tau=0} = \sum_{t \in LT_q} \gamma(t) \sum_j b_j \Phi_j(t) F^J(t)(u_0) = \sum_{t \in T_q} \alpha(t)\gamma(t) \sum_j b_j \Phi_j(t) F^J(t)(u_0) \,. \tag{2.27}$$

Доказательство.

Как уже отмечалось ранее, достаточно доказать первое равенство. Докажем его индукцией по q, следуя тем же путем, которым были получены ранее равенства (П.1) – (П.3). Применив к (2.17) формулу Лейбница, получим:

$$\left(g_i^J\right)^{(q)}\Big|_{\tau=0} = q \sum_j a_{ij} \left(f^J(g_i)\right)^{(q-1)}\Big|_{u=u_0} \,.$$

Далее воспользуемся формулой Фаа ди Бруно, используя при этом индуктивное предположение о справедливости формул (П.1) – (П.q-1) (т.е. подставляя соответствующие выражения в (Ф.q-1) вместо производных $\left(g_j^{K_s}\right)^{(\delta_s)}$) и учитывая, что всегда $\delta_s < q$). В итоге будем иметь:

$$\left(g_i^J\right)^{(q)}\Big|_{\tau=0} =$$
$$= q \sum_{u \in LS_q} \sum_{t_1 \in LT_{\delta_1}} \ldots \sum_{t_m \in LT_{\delta_m}} \gamma(t_1)\ldots\gamma(t_m) \sum_j a_{ij} \sum_{k_1} a_{jk_1} \Phi_{k_1}(t_1) \ldots \sum_{k_m} a_{jk_m} \Phi_{k_m}(t_m) \sum_{K_1,\ldots,K_m} f^J_{K_1\ldots K_m} \cdot F^{K_1}(t_1)(u_0)\ldots F^{K_m}(t_m)(u_0)$$

Остается уяснить, что каждой совокупности деревьев $\{u, t_1, \ldots, t_m\}$, $u \in LS_q$, $t_s \in LT_{\delta_s}$, $s = 1, \ldots, m$, соответствует определенное помеченное дерево $t \in LT_q$ такое, что

$$\gamma(t) = q \cdot \gamma(t_1) \cdot \ldots \cdot \gamma(t_m),$$

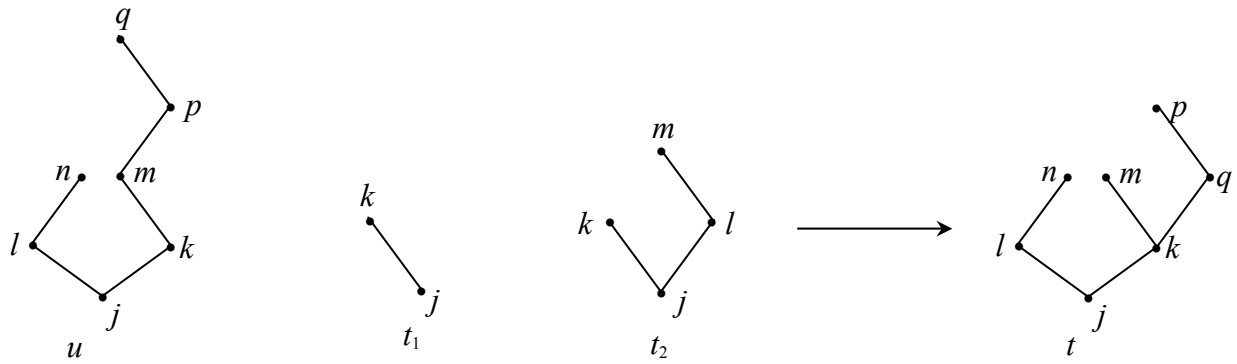
$$F^J(t)(u) = \sum_{K_1, \ldots, K_m} f^J_{K_1 \ldots K_m}(u) F^{K_1}(t_1)(u) \ldots F^{K_m}(t_m)(u),$$

$$\Phi_j(t) = \sum_{k_1, \ldots, k_m} a_{jk_1} \ldots a_{jk_m} \Phi_{k_1}(t_1) \ldots \Phi_{k_m}(t_m).$$

Это помеченное дерево t получается заменой ветвей дерева u деревьями $t_1, \ldots, t_m$, и соответствующие индексы переносятся с u на t в естественном порядке, т.е. сохраняется их упорядоченность. Таким способом оказываются построенными *все* деревья $t \in LT_q$, причем каждое возникает только один раз. В итоге, учитывая последние четыре формулы, получаем требуемое.

⊠

В качестве пояснения сказанного при доказательстве теоремы 2 приведем пример взаимно однозначного соответствия $\{u, t_1, \ldots, t_m\} \to t$.



Замечание. Описанное построение дерева t может быть использовано для индуктивного определения деревьев. Если обозначить $t = [t_1, \ldots, t_m]$ дерево, после удаления корня которого (вместе с инцидентными ему ребрами) останутся деревья $t_1, \ldots, t_m$, то в конечном счете все деревья могут быть выражены через τ. Например, $t_{21} = [\tau]$, $t_{31} = [\tau, \tau]$, $t_{32} = [[\tau]]$ и т.п.

Теперь, сопоставляя теоремы 1 и 2, окончательно получим результат:

***Теорема* 3.** Чтобы явный метод Рунге-Кутта (2.17) имел порядок p, необходимо и достаточно выполнение равенств

$$\sum_j b_j \Phi_j(t) = \frac{1}{\gamma(t)} \tag{2.28}$$

для всех деревьев порядка меньшего либо равного p.

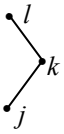
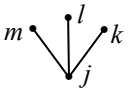
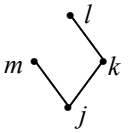
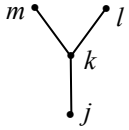
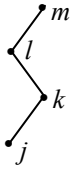
В заключение данного пункта приведем некоторую справочную информацию. Ниже, в таблице 1, указано число деревьев порядков от 1 до 10.

Таблица 1.

q	1	2	3	4	5	6	7	8	9	10
$card(T_q)$	1	1	2	4	9	20	48	115	286	719

Далее, в таблице 2, приведены все характеристики деревьев до четвертого порядка включительно, а также вид связанных с ними величин.

Таблица 2.

q	t	граф	$\gamma(t)$	$\alpha(t)$	$F^J(t)(u)$	$\Phi_j(t)$
0	Ø	Ø	1	1	u^J	
1	τ		1	1	f^J	1
2	t_{21}		2	1	$\sum_K f_K^J f^K$	$\sum_k a_{jk}$
3	t_{31}		3	1	$\sum_{K,L} f_{KL}^J f^K f^L$	$\sum_{k,l} a_{jk} a_{jl}$
	t_{32}		6	1	$\sum_{K,L} f_K^J f_L^K f^L$	$\sum_{k,l} a_{jk} a_{kl}$
4	t_{41}		4	1	$\sum_{K,L,M} f_{KLM}^J f^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{jl} a_{jm}$
	t_{42}		8	3	$\sum_{K,L,M} f_{KM}^J f_L^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{jm} a_{kl}$
	t_{43}		12	1	$\sum_{K,L,M} f_K^J f_{LM}^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{km}$
	t_{44}		24	1	$\sum_{K,L,M} f_K^J f_L^K f_M^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{lm}$

Наконец, в таблице 3 указано количество условий порядка при различных значениях числа p.

Таблица 3

p	1	2	3	4	5	6	7	8	9	10
Число условий	1	2	4	8	17	37	85	200	486	1205

2.4. Методы третьего и четвертого порядка точности

Пользуясь теоремой 3 (и технически слегка подглядывая в таблицу 2), мы теперь достаточно легко можем выписать условия порядка для случаев $p=3$ и $p=4$.

Имеем:

$p = 3$

$$\begin{cases} \sum\limits_j b_j = 1, \\ \sum\limits_{j,k} b_j a_{jk} = \dfrac{1}{2}, \\ \sum\limits_{j,k,l} b_j a_{jk} a_{jl} = \dfrac{1}{3}, \\ \sum\limits_{j,k,l} b_j a_{jk} a_{kl} = \dfrac{1}{6}. \end{cases} \tag{2.29}$$

Последнюю систему, учитывая условие (2.18), можно переписать в виде

$$\begin{cases} \sum\limits_j b_j = 1, \\ \sum\limits_j b_j c_j = \dfrac{1}{2}, \\ \sum\limits_j b_j c_j^2 = \dfrac{1}{3}, \\ \sum\limits_{j,k} b_j a_{jk} c_k = \dfrac{1}{6}. \end{cases} \tag{2.30}$$

Отсюда, положив, например, $s = 3$, получаем:

$$\begin{cases} b_1 + b_2 + b_3 = 1, \\ b_2 c_2 + b_3 c_3 = \dfrac{1}{2}, \\ b_2 c_2^2 + b_3 c_3^3 = \dfrac{1}{3}, \\ b_3 a_{32} c_2 = \dfrac{1}{6}, \\ c_2 = a_{21}, \\ c_3 = a_{31} + a_{32}. \end{cases}$$

Одним из алгоритмов третьего порядка точности, получающихся в результате решения последней системы, является, например, такой:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \dfrac{1}{3} & \dfrac{1}{3} & 0 & 0 \\ \dfrac{2}{3} & 0 & \dfrac{2}{3} & 0 \\ \hline & \dfrac{1}{4} & 0 & \dfrac{3}{4} \end{array}$$

$p = 4$:

Соответствующая система уравнений будет иметь вид:

$$\begin{cases} \sum\limits_{j} b_j = 1, \\ \sum\limits_{j,k} b_j a_{jk} = \frac{1}{2}, \\ \sum\limits_{j,k,l} b_j a_{jk} a_{jl} = \frac{1}{3}, \\ \sum\limits_{j,k,l} b_j a_{jk} a_{kl} = \frac{1}{6}, \\ \sum\limits_{j,k,l,m} b_j a_{jk} a_{jl} a_{jm} = \frac{1}{4}, \\ \sum\limits_{j,k,l,m} b_j a_{jk} a_{jm} a_{kl} = \frac{1}{8}, \\ \sum\limits_{j,k,l,m} b_j a_{jk} a_{kl} a_{km} = \frac{1}{12}, \\ \sum\limits_{j,k,l,m} b_j a_{jk} a_{kl} a_{lm} = \frac{1}{24} \end{cases} \quad \text{или, с учетом (2.18),} \quad \begin{cases} \sum\limits_{j} b_j = 1, \\ \sum\limits_{j} b_j c_j = \frac{1}{2}, \\ \sum\limits_{j} b_j c_j^2 = \frac{1}{3}, \\ \sum\limits_{j,k} b_j a_{jk} c_k = \frac{1}{6}, \\ \sum\limits_{j} b_j c_j^3 = \frac{1}{4}, \\ \sum\limits_{j,k} b_j c_j a_{jk} c_k = \frac{1}{8}, \\ \sum\limits_{j,k} b_j a_{jk} c_k^2 = \frac{1}{12}, \\ \sum\limits_{j,k,l} b_j a_{jk} a_{kl} c_l = \frac{1}{24}. \end{cases} \tag{2.31}$$

Отсюда при $s = 4$ имеем систему

$$\begin{cases} b_1 + b_2 + b_3 + b_4 = 1, \\ b_2 c_2 + b_3 c_3 + b_4 c_4 = \frac{1}{2}, \\ b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = \frac{1}{3}, \\ b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^2 = \frac{1}{4}, \\ \\ b_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{6}, \\ b_3 a_{32} c_2 c_3 + b_4 c_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{8}, \\ b_3 a_{32} c_2^2 + b_4 (a_{42} c_2^2 + a_{43} c_3^2) = \frac{1}{12}, \\ b_4 a_{43} a_{32} c_2 = \frac{1}{24}, \\ c_2 = a_{21}, \\ c_3 = a_{31} + a_{32}, \\ c_4 = a_{41} + a_{42} + a_{43}, \end{cases}$$

одним из решений которой является метод, записанный ниже.

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Именно этот метод в технической литературе и называют методом Рунге-Кутта.

Упражнения. 1. Построить другие примеры методов третьего и четвертого порядков;

2. Выписать условия порядка для $p=5$;

3. Найти все деревья порядка 6.

§ 3. Способ последовательного повышения порядка точности построения одношаговых методов

Как и в способе Рунге-Кутта, при построении методов численного решения задачи Коши (0.1), (0.2) будем исходить из интегрального соотношения (2.4), которое сейчас перепишем в виде

$$u(t_j+\tau)=u(t_j)+\tau\int_0^1 z_j(\alpha)d\alpha\,, \tag{3.1}$$

где $z_j(\alpha)=f(t_j+\alpha\tau,u(t_j+\alpha\tau))$.

Заменим интеграл в (3.1) квадратурной суммой $\sum_{i=0}^{q}A_i z_j(\alpha_i)$. Тогда будем иметь:

$$u(t_j+\tau)\approx u(t_j)+\tau\sum_{i=0}^{q}A_i z_j(\alpha_i)=u(t_j)+\tau\sum_{i=0}^{q}A_i f(t_j+\alpha_i\tau,u(t_j+\alpha_i\tau)). \tag{3.2}$$

Выбор параметров A_i, α_i, $i=0,1,\ldots,q,$ в этом приближенном равенстве будем осуществлять, например, исходя из требования, чтобы квадратурная формула

$$\int_0^1 z_j(\alpha)d\alpha\approx\sum_{i=0}^{q}A_i z_j(\alpha_i) \tag{3.3}$$

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

$$\begin{cases}\sum\limits_{i=0}^{q}A_i=1,\\ \sum\limits_{i=0}^{q}A_i\alpha_i^{\,j}=\dfrac{1}{j+1},\ j=1,\ldots,k-1.\end{cases} \tag{3.4}$$

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

$$u(t_j+\tau)\approx u(t_j)+\sum_{i=0}^{q}A_i f(t_j+\alpha_i\tau,u(t_j+\alpha_i\tau))=u(t_j)+\sum_{i=0}^{q}A_i u'(t_j+\alpha_i\tau)$$

совпадали до членов с τ^k включительно. Тогда, очевидно, локальная погрешность формулы (3.2) будет иметь вид

$$r(t_j,\tau)=\tau^{k+1}u^{(k+1)}(t_j)\left[\frac{1}{(k+1)!}-\frac{1}{k!}\sum_{i=0}^{q}A_i\alpha_i^k\right]+O(\tau^{k+2}). \qquad (3.5)$$

Так как весовая функция в случае интеграла $\int_0^1 z_j(\alpha)d\alpha$ равна 1, то квадратурная формула вида (3.3), имеющая алгебраическую степень точности, равную $2q+1$ (формула наивысшей алгебраической степени точности), может быть построена и притом единственным образом для любого значения $q\ge 0$. Поэтому при $k=2q+2$ система (3.4) имеет единственное решение, при этом $0<A_i\le 1,\ 0<\alpha_i<1,\ i=\overline{0,q}$. Следовательно, при $1\le k\le 2q+2$ у этой системы существует хотя бы одно решение. Таким образом, приближенное равенство (3.2) может быть построено (т.е. вопрос о разрешимости системы здесь целиком решается на базе теории квадратурных формул).

Если бы в (3.2) все значения $u(t_j+\alpha_i\tau),\ i=\overline{0,q}$, были известны точно, то это приближенное равенство позволяло бы найти искомое значение $u(t_j+\tau)$ соответствующего решения задачи Коши по известному значению $u(t_j)$ этого решения с локальной ошибкой порядка τ^{k+1}.

Хотя точными значениями $u(t_j+\alpha_i\tau)$ мы не располагаем, но, подобно (3.2), заменив там τ на $\alpha_i\tau$, нетрудно указать правила для их приближенного вычисления через значения $u(t_j+\alpha_i\beta_{in}\tau)$, для нахождения которых, в свою очередь, можно построить подобные же рекурсивные формулы. При этом следует иметь в виду, что наличие множителя τ перед суммой в формуле (3.2) позволяет находить значения $u(t_j+\alpha_i\tau)$ с локальной ошибкой порядка τ^k, значения $u(t_j+\alpha_i\beta_{in}\tau)$ – с ошибкой порядка τ^{k-1}, и т.д., понижая на каждом шаге рекурсии требования к порядку точности на единицу. Параметры соответствующих приближенных равенств должны удовлетворять системе уравнений типа (3.4), в которой с понижением требований к точности на порядок следует уменьшать на единицу и количество уравнений (отбрасывая при этом последнее из них). При этом часто бывает целесообразным уменьшать и число q, определяющее количество подлежащих выбору параметров.

Следуя такой схеме действий, придем, наконец, к приближенным равенствам

$$u(t_j+\alpha_i\beta_{in}\ldots\gamma_{in\ldots l}\tau)\approx u(t_j)+\alpha_i\beta_{in}\ldots\gamma_{in\ldots l}\tau f(t_j,u(t_j)), \qquad (3.6)$$

на которых процесс замыкается. Погрешность таких равенств будет, очевидно, величиной порядка τ^2. Они получаются из равенств типа (3.2) в случае, когда квадратурная формула (3.3) является простейшей формулой левых прямоугольников.

Приведем сейчас примеры методов, полученных описанным выше способом, условившись предварительно о следующих обозначениях:

$$\overset{[k]}{y}_{j+\alpha}=u(t_j+\alpha\tau)+O(\tau^k),\ \ \overset{[k]}{f}_{j+\alpha}=f\left(t_j+\alpha\tau,\overset{[k]}{y}_{j+\alpha}\right).$$

а) *Методы первого порядка точности*.

Система (3.4) в этом случае вырождается в единственное уравнение

$$\sum_{i=0}^{q} A_i = 1. \tag{3.7}$$

Параметры α_i, $i = \overline{0,q}$, могут принимать, вообще говоря, любые фиксированные значения. Однако для случая одношаговых методов выбор этих параметров должен быть подчинен ограничению $0 \le \alpha_i \le 1$. Положив в (3.7), например, $q = 0$, найдем: $A_0 = 1$. Задавая теперь $\alpha_0 = 0$, получим формулу *явного метода Эйлера*:

$$y^{[2]}_{j+1} = y^{[2]}_j + \tau f^{[2]}_j . \tag{3.8}$$

б) *Методы второго порядка точности*.

В этом случае требование (3.7) нужно дополнить условием

$$\sum_{i=0}^{q} A_i \alpha_i = \frac{1}{2}. \tag{3.9}$$

При $q = 0$ система (3.7), (3.9) имеет единственное решение $A_0 = 1,\ \alpha_0 = \frac{1}{2}$, что приводит к следующему вычислительному правилу:

$$\begin{cases} y^{[2]}_{j+\frac{1}{2}} = y^{[3]}_j + \dfrac{\tau}{2} f^{[3]}_j, \\ y^{[3]}_{j+1} = y^{[3]}_j + \tau f^{[2]}_{j+\frac{1}{2}}. \end{cases} \tag{3.10}$$

При $q = 1$ система (3.7), (3.9) примет вид

$$\begin{cases} A_0 + A_1 = 1, \\ A_0\alpha_0 + A_1\alpha_1 = \dfrac{1}{2}. \end{cases}$$

Выбрав, например, $\alpha_0 = 0,\ \alpha_1 = 1$, найдем: $A_0 = A_1 = \frac{1}{2}$ и получим ***неявный метод трапеций***

$$y^{[3]}_{j+1} = y^{[3]}_j + \frac{\tau}{2}\left(f^{[3]}_j + f^{[3]}_{j+1} \right). \tag{3.11}$$

Используя формулу типа (3.6) (или, что то же самое, явный метод Эйлера), (3.11) можно преобразовать в ***явный метод трапеций***:

$$\begin{cases} y^{[2]}_{j+1} = y^{[3]}_j + \tau f^{[3]}_j, \\ y^{[3]}_{j+1} = y^{[3]}_j + \dfrac{\tau}{2}\left(f^{[3]}_j + f^{[2]}_{j+1} \right). \end{cases} \tag{3.12}$$

Заметим, что наряду с (3.12) можно записать и более экономичный вариант явного метода трапеций, требующий, в отличие от (3.12), вычисления всего одного значения правой части исходной задачи на каждый узел сетки, кроме первого:

$$\begin{cases} {}^{[2]}y_{j+1} = {}^{[3]}y_j + \tau\, {}^{[2]}f_j, \\ {}^{[3]}y_{j+1} = {}^{[3]}y_j + \dfrac{\tau}{2}\left({}^{[2]}f_j + {}^{[2]}f_{j+1} \right). \end{cases} \tag{3.13}$$

в) *Методы третьего порядка точности*.

К уравнениям (3.7), (3.9) добавляется еще одно:

$$\sum_{i=0}^{q} A_i \alpha_i^2 = \frac{1}{3}. \tag{3.14}$$

Положив $q = 1$ (при $q = 0$ система, очевидно, несовместна), получим:

$$\begin{cases} A_0 + A_1 = 1, \\ A_0\alpha_0 + A_1\alpha_1 = \dfrac{1}{2}, \\ A_0\alpha_0^2 + A_0\alpha_1^2 = \dfrac{1}{3}. \end{cases}$$

Отсюда, положив $\alpha_0 = 0$, находим: $\alpha_1 = \dfrac{2}{3}$, $A_1 = \dfrac{3}{4}$, $A_0 = \dfrac{1}{4}$. Используя для вычисления ${}^{[3]}y_{j+\frac{2}{3}}$ формулы (3.10) (с заменой τ на $\dfrac{2}{3}\tau$), окончательно будем иметь:

$$\begin{cases} {}^{[2]}y_{j+\frac{1}{3}} = {}^{[4]}y_j + \dfrac{\tau}{3}\, {}^{[4]}f_j, \\ {}^{[3]}y_{j+\frac{2}{3}} = {}^{[4]}y_j + \dfrac{2\tau}{3}\, {}^{[2]}f_{j+\frac{1}{3}}, \\ {}^{[4]}y_{j+1} = {}^{[4]}y_j + \dfrac{\tau}{4}\left({}^{[4]}f_j + 3\, {}^{[3]}f_{j+\frac{2}{3}} \right). \end{cases} \tag{3.15}$$

При $q = 2$ можно построить, например, такой вычислительный алгоритм третьего порядка точности, базирующийся на квадратурной формуле Симпсона:

$$\begin{cases} {}^{[2]}y_{j+\frac{1}{4}} = {}^{[4]}y_j + \dfrac{\tau}{4}\, {}^{[i]}f_j, \\ {}^{[3]}y_{j+\frac{1}{2}} = {}^{[4]}y_j + \dfrac{\tau}{2}\, {}^{[2]}f_{j+\frac{1}{4}}, \\ {}^{[3]}y_{j+1} = {}^{[4]}y_j + \tau\, {}^{[3]}f_{j+\frac{1}{2}}, \\ {}^{[4]}y_{j+1} = {}^{[4]}y_j + \dfrac{\tau}{6}\left({}^{[i]}f_j + 4\, {}^{[3]}f_{j+\frac{1}{2}} + {}^{[3]}f_{j+1} \right). \end{cases} \tag{3.16}$$

Здесь i может принимать значения 3 или 4. При $i = 4$ построенное правило на один узел сетки требует четырехкратного обращения к блоку нахождения значений правой части исходного уравнения. В случае же $i = 3$ точность результата, вообще говоря, несколько по-

нижается (при сохранении порядка), однако в основном счете число обращений к блоку вычисления значений функции $f(t,u)$ сокращается до трех на узел сетки.

Отметим также, что построенный численный метод, как и правила (3.12) и (3.13), имеют предсказывающе-исправляющий характер. Приближенное значение величины $u(t_{j+1})$, найденное с локальной погрешностью порядка τ^3, уточняется затем по формуле, локальная погрешность которой имеет четвертый порядок. Сравнение значений $\overset{[3]}{y}_{j+1}$ и $\overset{[4]}{y}_{j+1}$ дает практическую возможность по ходу вычислений без дополнительных вычислительных затрат составить представление о локальной точности полученного приближения к $u(t_{j+1})$. Такое сравнение, в частности, может быть положено в основу правила автоматического выбора шага интегрирования.

На примере приведенных вычислительных правил легко видеть, что описанный способ построения методов численного интегрирования дифференциальных уравнений удовлетворяет принципу модульности, когда сложные вычислительные правила компонуются на основе более простых типовых расчетных формул. Так, например, в случае метода (3.16) мы имеем цепочку явный метод Эйлера – формула средних прямоугольников – формула Симпсона.

Упражнение. Построить примеры методов четвертого порядка точности.

§ 4. Практический контроль погрешности приближенного решения

В ходе расчетов всегда желательно иметь представление о том, сколь далеко полученное приближенное решение от истинного решения исходной задачи и в соответствии с величиной оценки погрешности выбирать шаг численного интегрирования. Большинство из известных методик, применяемых для этих целей, оценивают главный член погрешности метода.

4.1. Правило Рунге

Как мы видели, главный член погрешности аппроксимации метода k-го порядка имеет вид $\tau^k \rho(t)$, т.е.

$$u(t_j+\tau) = y(t_j+\tau) + \tau^k \rho(t_j) + O(\tau^{k+p}).$$

Тогда, проведя расчеты из одной и той же точки t_j с двумя различными шагами τ_1 и τ_2, получим:

$$\begin{cases} u(t_j+\tau) \approx y_{\tau_1}(t_j+\tau) + \tau_1^k \rho(t_j), \\ u(t_j+\tau) \approx y_{\tau_2}(t_j+\tau) + \tau_2^k \rho(t_j), \end{cases}$$

откуда

$$\rho(t_j) \approx \frac{y_{\tau_1} - y_{\tau_2}}{\tau_2^k - \tau_1^k}. \qquad (4.1)$$

Эта формула дает возможность после проведения вычислений до точки $t_j+\tau$ с шагами τ_1 и τ_2 получить приближенные значения величин погрешностей для каждого из при-

ближенных значений решения y_{τ_1} и y_{τ_2}. Кроме того, при заданной границе ε допустимой погрешности на основании приближенного равенства

$$\varepsilon \approx \left|\rho\left(t_j\right)\right|\tau_\varepsilon^k$$

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

$$\tau_\varepsilon = \sqrt[k]{\varepsilon\left|\frac{\tau_2^k - \tau_1^k}{y_{\tau_1} - y_{\tau_2}}\right|}. \tag{4.2}$$

В практике вычислений достаточно часто в качестве τ_1 и τ_2 выбирают τ и $\frac{\tau}{2}$ соответственно (прием, предложенный еще Рунге). В этом случае формулы (4.1) и (4.2) принимают вид

$$\rho\left(t_j\right) \approx \frac{y_{\frac{\tau}{2}} - y_\tau}{\tau^k\left(1 - \frac{1}{2^k}\right)},$$

$$\tau_\varepsilon = \frac{\tau}{2}\sqrt[k]{\frac{\left(2^k - 1\right)\varepsilon}{\left|y_{\frac{\tau}{2}} - y_\tau\right|}},$$

а сам способ оценки носит название ***двойного пересчета***.

Замечание. Значение решения в очередной точке сетки только тогда следует считать найденным (вместе с координатой узла сетки), когда значение $\rho\left(t_j\right)$, найденное по формуле (4.1) (или ее аналогу) будет в пределах допустимой погрешности. И только после этого следует приступать к очередному шагу.

4.2. Использование вложенных методов

Как мы уже отмечали в § 3, способ последовательного повышения порядка точности часто приводит к тому, что приближенное решение получается в одной и той же точке, но с разным порядком погрешности относительно шага сетки. Такими, например, являются явный метод трапеций (3.12) и метод третьего порядка точности (3.16).

Тогда разность между этими величинами дает возможность вычислить главный член локальной погрешности метода более низкого порядка (или, если разделить ее на τ, то слово «локальной» можно опустить). Далее действия должны быть такими же, как и в описанном выше подходе Рунге.

4.2.1. Вложенные методы Рунге-Кутта

Методы Рунге-Кутта, разобранные нами в § 2, в изложенной там схеме не предоставляют возможности оценки погрешности, аналогичной изложенной выше. С другой стороны, принципиальная возможность использовать идею вложенности существует.

В общем случае нам нужно найти такую таблицу Батчера

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
c_2 & a_{21} & & & & & \\
c_3 & a_{31} & a_{32} & & & & \\
\vdots & & \cdots & & & & \\
c_s & a_{s1} & a_{s2} & \ldots & a_{ss-1} & & \\
\hline
 & b_1 & b_2 & \ldots & b_{s-1} & b_s & \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \ldots & \hat{b}_{s-1} & \hat{b}_s &
\end{array}
\tag{4.3}
$$

чтобы величина

$$y_{j+1} = y_j + \tau(b_1 k_1 + \cdots + b_s k_s)$$

имела порядок p, а величина

$$y_{j+1} = y_j + \tau\left(\hat{b}_1 k_1 + \cdots + \hat{b}_s k_s\right)$$

порядок q (обычно $q = p-1$ или $q = p+1$).

Поясним эту идею на конкретном примере, построив вложенные формулы порядков 2 и 3.

Тогда таблица (4.3) будет иметь вид (полагаем $s = 3$, так как при $s = 2$ явных методов Рунге-Кутта третьего порядка построить нельзя)

$$
\begin{array}{c|ccc}
0 & & & \\
c_2 & a_{21} & & \\
c_3 & a_{31} & a_{32} & \\
\hline
 & b_1 & b_2 & b_3 \\
\hline
 & \hat{b}_1 & \hat{b}_2 & \hat{b}_3
\end{array}
$$

Параметры этой таблицы должны удовлетворять условиям

$$
\begin{cases} b_1 + b_2 + b_3 = 1, \\ b_2 c_2 + b_3 c_3 = \dfrac{1}{2} \end{cases}
\quad \text{(второй порядок)} \quad \text{и} \quad
\begin{cases} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 = 1, \\ \hat{b}_2 c_2 + \hat{b}_3 c_3 = \dfrac{1}{2}, \\ \hat{b}_2 c_2^2 + \hat{b}_3 c_3^2 = \dfrac{1}{3}, \\ \hat{b}_3 a_{32} c_2 = \dfrac{1}{6} \end{cases}
\quad \text{(третий порядок).}
$$

Выбрав $c_2 = 1$ и $b_3 = 0$, из первых двух уравнений получим: $b_2 = b_1 = \dfrac{1}{2}$. Осталось четыре уравнения с пятью неизвестными. Если положить $c_3 = \dfrac{1}{2}$, то $\hat{b}_1 = \dfrac{1}{6}$, $\hat{b}_2 = \dfrac{1}{6}$, $\hat{b}_3 = \dfrac{4}{6}$

и $a_{32} = \frac{1}{4}$. Таким образом, получившийся метод имеет вид (его аббревиатура в литературе – RKF2(3) по первым буквам фамилий: Рунге-Кутта-Фельберга))

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{4}{6} \end{array} \qquad (4.4)$$

Автором данного алгоритма – Фельбергом – были построены и другие варианты вложенных формул различных порядков точности.

Упражнение. Построить другие примеры вложенных методов типа Рунге-Кутта.

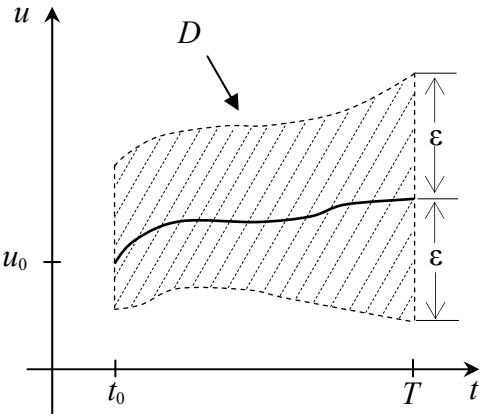
§ 5. Сходимость одношаговых методов решения задачи Коши

Исследуем сейчас вопрос о сходимости одношаговых методов решения задачи Коши, частные случаи построения которых мы рассмотрели выше.

Предположим, что исходное дифференциальное уравнение (0.1) имеет единственное решение на отрезке $t_0 \le t \le T$ не только при начальных данных (0.2), но и в случае любых начальных данных вида $u(\xi) = \eta$, где

$$(\xi,\eta) \in D = \{(\xi,\eta) : t_0 \le \xi \le T;\ u(\xi,T,u(T,t_0,u_0)-\varepsilon) \le \eta \le u(\xi,T,u(T,t_0,u_0)+\varepsilon)\},$$

а через $u(t,\xi,\eta)$ обозначено значение в точке t решения $u(t)$ уравнения (0.1) при начальных данных $u(\xi) = \eta$. Величина ε при этом может быть истолкована как верхняя граница допустимой погрешности в нахождении решения задачи (0.1), (0.2) в точке $t = T$ (см. рис.).



Вначале докажем вспомогательное утверждение, с помощью которого можно находить производные от решения по начальным данным.

Лемма. Имеет место соотношение

$$\frac{\partial u(t,\xi,\eta)}{\partial \eta} = \exp \int_{\xi}^{t} \frac{\partial f(x,u(x,\xi,\eta))}{\partial u}\, dx\,. \qquad (5.1)$$

Доказательство.

Так как $u(t,\xi,\eta)$ удовлетворяет исходному уравнению (0.1), то $\frac{\partial u(t,\xi,\eta)}{\partial t} = f(t,u(t,\xi,\eta))$.

Продифференцируем последнее соотношение по переменной η:

$$\frac{\partial}{\partial \eta}\frac{\partial u(t,\xi,\eta)}{\partial t} = \frac{\partial f(t,u(t,\xi,\eta))}{\partial u} \cdot \frac{\partial u(t,\xi,\eta)}{\partial \eta}$$

или, меняя порядок дифференцирования слева и вводя обозначение $\frac{\partial u(t,\xi,\eta)}{\partial \eta} = z$:

$$\frac{\partial}{\partial t} z(t) = \frac{\partial f(t,u(t,\xi,\eta))}{\partial \eta} \cdot z(t). \tag{5.2}$$

Так как $z(\xi) = \frac{\partial}{\partial \eta} u(\xi,\xi,\eta) = \frac{\partial}{\partial \eta} \eta = 1$, то из (5.2) получим:

$$z(t) = \frac{\partial u(t,\xi,\eta)}{\partial t} = \exp \int_{\xi}^{t} \frac{\partial f(x,u(x,\xi,\eta))}{\partial u} dx.$$

⊠

Пусть $M > 0$ таково, что

$$\left| \frac{\partial u(t,\xi,\eta)}{\partial \eta} \right| \le M \text{ для всех } \xi, \eta \in D, \ \xi \le t. \tag{*}$$

Будем считать, что $u(t,\xi,\eta)$ имеет все необходимые по ходу изложения производные по t, порядок которых определяется конструкцией избранного метода

$$y_{j+1} = F(y_j), \ j = 0,1,...,N-1. \tag{5.3}$$

Здесь, как и ранее, y_i – приближенное значение для $u(t_i,t_0,u_0)$, полученное при условии точного выполнения всех предусмотренных в (5.3) операций.

В действительности же мы вместо (5.3) имеем соотношения вида

$$\widetilde{y}_{j+1} = F(\widetilde{y}_j) - \delta_{j+1}, \ j = 0,1,...,N-1. \tag{5.4}$$

Разность $\varepsilon_0 = y_0 - \widetilde{y}_0$ будем называть ***погрешностью начального условия***, а величину $-\delta_{j+1}$ – ***погрешностью округления*** на $(j+1)$-м шаге вычислительного процесса. При отсутствии ошибок начальных данных и округлений величину

$$\varepsilon_j = u(t_j,t_0,u_0) - y_j = u(t_j,t_0,u_0) - u(t_j,t_j,y_j)$$

обычно называют ***погрешностью метода***.

На практике интерес представляет величина погрешности приближенного решения

$$\widetilde{\varepsilon}_j = u(t_j,t_0,u_0) - \widetilde{y}_j = u(t_j,t_0,u_0) - u(t_j,t_j,\widetilde{y}_j). \tag{5.5}$$

Погрешность формулы (5.3) на каждом шаге реального вычислительного процесса можно определить посредством равенства

$$u(t_{j+1},t_j,\widetilde{y}_j) = F(u(t_j,t_j,\widetilde{y}_j)) + r_{j+1} = F(\widetilde{y}_j) + r_{j+1}. \tag{5.6}$$

Здесь r_{j+1} – фактически – локальная погрешность метода. Тогда, вычитая из (5.6) (5.4), получим:

$$u(t_{j+1},t_j,\widetilde{y}_j) - \widetilde{y}_{j+1} = r_{j+1} + \delta_{j+1}. \tag{5.7}$$

(5.7) дает возможность оценить реально допускаемое отклонение на каждом шаге вычислительного процесса.

Теперь вернемся к соотношению (5.5), дающему глобальную погрешность:

$$\widetilde{\varepsilon}_j = u(t_j, t_0, y_0) - u(t_j, t_j, \widetilde{y}_j) = u(t_j, t_0, y_0) - u(t_j, t_0, \widetilde{y}_0) + \sum_{i=1}^{j} \left[u(t_j, t_{i-1}, \widetilde{y}_{i-1}) - u(t_j, t_i, \widetilde{y}_i) \right].$$

Отсюда, поскольку $u(t_j, t_{i-1}, \widetilde{y}_{i-1}) = u(t_j, t_i, u(t_i, t_{i-1}, \widetilde{y}_{i-1}))$, то

$$\widetilde{\varepsilon}_j = u(t_j, t_0, y_0) - u(t_j, t_j, \widetilde{y}_0) + \sum_{i=1}^{j} \left[u(t_j, t_i, u(t_i, t_{i-1}, \widetilde{y}_{i-1})) - u(t_j, t_i, \widetilde{y}_i) \right].$$

Последнее соотношение, используя формулу Лагранжа о конечном приращении, перепишем в виде

$$\widetilde{\varepsilon}_j = (y_0 - \widetilde{y}_0) \frac{\partial u(t_j, t_0, \widetilde{\eta}_0)}{\partial \eta} + \sum_{i=1}^{j} \left[u(t_i, t_{i-1}, \widetilde{y}_{i-1}) - \widetilde{y}_i \right] \frac{\partial u(t_j, t_i, \widetilde{\eta}_i)}{\partial \eta} =$$

$$= \widetilde{\varepsilon}_0 \frac{\partial u(t_j, t_0, \widetilde{\eta}_0)}{\partial \eta} + \sum_{i=1}^{j} (r_i + \delta_i) \frac{\partial u(t_j, t_i, \widetilde{\eta}_i)}{\partial \eta}. \qquad (5.8)$$

Таким образом, отсюда, учитывая соотношение (*), получаем оценку

$$|\widetilde{\varepsilon}_j| \le \left(\varepsilon_0 + \sum_{i=1}^{j} (|r_i| + |\delta_i|) \right) M \le (\varepsilon_0 + j(r+\delta)) M \le (\varepsilon_0 + N(r+\delta)) M = \left(\varepsilon_0 + \frac{(r+\delta)(T-t_0)}{\tau} \right) M,$$

$$j = 1, 2, ..., N.$$

На основании этой оценки можно утверждать, что если $\varepsilon_0 \to 0$, $\dfrac{\delta}{\tau} \to 0$ и $\dfrac{r}{\tau} \to 0$ при $\tau \to 0$, то в любой точке отрезка $[t_0; T]$ приближенное решение задачи Коши, полученное с помощью одношагового метода (5.3), будет сходиться к точному решению этой задачи.

В частности, когда $\widetilde{\varepsilon}_0 = 0$, $\delta_i = 0$ имеем условие сходимости в виде $\dfrac{r}{\tau} \underset{\tau \to 0}{\to} 0$. При этом обычно для локальной погрешности справедливо представление $r = O(\tau^{p+1})$ для некоторого $p > 0$. Поэтому данное условие автоматически выполняется, причем скорость сходимости будет иметь p-й порядок.

Отметим, что соотношение (5.8) с учетом (5.1) позволяет провести более подробный анализ зависимости погрешности от свойств исходного дифференциального уравнения. Так, например, если $\dfrac{\partial f}{\partial u} > 0$, то $\dfrac{\partial u}{\partial \eta} > 1$ и влияние погрешностей начальных данных и округления растет, а в случае $\dfrac{\partial f}{\partial u} < 0$ – ослабевает.

ГЛАВА XI

Многошаговые методы решения задачи Коши

Рассмотрим сейчас другой класс численных методов решения задачи Коши, отказавшись от условия одношаговости.

Вновь воспользуемся интегральным соотношением типа (2.1) предыдущей главы, которое сейчас будет удобно переписать в виде

$$u(t_{j+1}) = u(t_j) + \int_{t_j}^{t_{j+1}} u'(x)dx\,. \tag{1.1}$$

Предполагая известными (уже найденными) решения $y_{j-k},...,y_j$, заменим функцию $u'(x)$, стоящую под знаком интеграла, некоторым ее интерполяционным приближением по выписанным выше значениям:

$$u'(x) \approx \varphi(x, y_{j-k},..., y_j), \tag{1.2}$$

где φ – некоторая известная функция.

Очевидно, после такой замены интеграл может быть непосредственно вычислен.

В результате получим некоторый явный многошаговый метод решения задачи Коши.

§ 1. Методы Адамса

1.1. Экстраполяционные методы Адамса

Чаще всего в (1.2) используется многочленное приближение. Ранее мы рассматривали наиболее часто используемые представления соответствующих интерполяционных многочленов. Одним из них является, например, интерполяционный многочлен в форме Лагранжа:

$$u'(x) \approx L_k(x) = \sum_{i=0}^{k} \frac{\omega_{k+1}(x)}{(x - t_{j-i})\omega'_{k+1}(t_{j-i})} y'(t_{j-i}) = \sum_{i=0}^{k} \frac{\omega_{k+1}(x)}{(x - t_{j-i})\omega'_{k+1}(t_{j-i})} f(t_{j-i}, y_{j-i}). \tag{1.3}$$

Подставляя это выражение в (1.1), получим численный метод вида

$$y_{j+1} = y_j + \sum_{i=0}^{k} \beta_i f(t_{j-i}, y_{j-i}), \tag{1.4}$$

где

$$\beta_i = \int_{t_j}^{t_{j+1}} \frac{\omega_{k+1}(x)}{(x - t_{j-i})\omega'_{k+1}(t_{j-i})} dx, \quad i = \overline{0,k}\,. \tag{1.5}$$

(1.4), (1.5) – $(k+1)$-шаговый экстраполяционный метод Адамса на неравномерной сетке.

Очевидно, с ростом k растет как порядок точности рассматриваемых методов, так и их сложность. В частном случае, при $k = 1$, получим:

$$u'(x) \approx L_1(x) = \frac{x - t_{j-1}}{t_j - t_{j-1}} f_j + \frac{x - t_j}{t_{j-1} - t_j} f_{j-1}$$

или, используя обозначение $t_j - t_{j-1} = \tau_{j-1}$

$$u'(x) \approx L_1(x) = \frac{x - t_{j-1}}{\tau_{j-1}} f_j - \frac{x - t_j}{\tau_{j-1}} f_{j-1}.$$

Отсюда

$$y_{j+1} = y_j + \int_{t_j}^{t_{j+1}} \left(\frac{x - t_{j-1}}{\tau_{j-1}} f_j - \frac{x - t_j}{\tau_{j-1}} f_{j-1} \right) dx =$$

$$= y_j + \frac{1}{2\tau_{j-1}} \left\{ \left[(t_{j+1} - t_{j-1})^2 - (t_j - t_{j-1})^2 \right] f_j - (t_{j+1} - t_j)^2 f_{j-1} \right\} = y_j + \frac{(\tau_j + \tau_{j-1})^2 - \tau_{j-1}^2}{2\tau_{j-1}} f_j - \frac{\tau_j^2}{2\tau_{j-1}} f_{j-1}.$$

Упражнение. Рассмотреть случай $k = 2$.

1.1.1. Экстраполяционные методы Адамса на равномерной сетке

Если сетка узлов равномерна, т.е. когда для всех значений j $\tau_j = \text{const} = \tau$, то предлагаемые методы имеют значительно более простой вид. Дадим их подробное описание.

Делая вновь, как и в способе Рунге-Кутта, замену переменных в интегральном тождестве (1.1), перепишем его в виде

$$u(t_{j+1}) = u(t_j) + \tau \int_0^1 u'(t_j + \alpha\tau) d\alpha. \qquad (1.1')$$

Тогда описанная выше процедура приведет к методу

$$y_{j+1} = y_j + \tau \sum_{i=0}^{k} \beta_i f_{j-i}, \qquad (1.6)$$

где

$$\beta_i = \frac{(-1)^i}{i! \cdot (k-i)!} \int_0^1 \frac{\alpha(\alpha+1)\ldots(\alpha+k)}{\alpha+i} d\alpha, \quad i = \overline{0,k}. \qquad (1.7)$$

В частности, имеем:

1) при $k = 0$:

$$\beta_0 = \int_0^1 d\alpha = 1$$

и (1.6) примет вид

$$y_{j+1} = y_j + \tau f_j.$$

2) при $k = 1$:

$$\beta_0 = \int_0^1 (\alpha+1)d\alpha = \frac{3}{2};\ \ \beta_1 = -\int_0^1 \alpha d\alpha = -\frac{1}{2}$$

и

$$y_{j+1} = y_j + \frac{\tau}{2}\left(3f_j - f_{j-1}\right).$$

3) при $k = 2$:

$$\beta_0 = \frac{1}{2}\int_0^1 (\alpha+1)(\alpha+2)d\alpha = \frac{23}{12};\ \ \beta_1 = -\int_0^1 \alpha(\alpha+2)d\alpha = -\frac{4}{3};\ \ \beta_2 = \frac{1}{2}\int_0^1 \alpha(\alpha+1)d\alpha = \frac{5}{12}$$

и

$$y_{j+1} = y_j + \frac{\tau}{12}\left(23f_j - 16f_{j-1} + 5f_{j-2}\right).$$

Заметим, что более известно ***другое представление*** экстраполяционных методов Адамса, базирующееся на представлении интерполяционного многочлена в форме Ньютона. Поскольку интерполирование ведется по узлам $t_j, t_{j-1}, ..., t_{j-k}$, то естественно применять интерполяционную формулу Ньютона для конца таблицы:

$$u'\left(t_j + \alpha\tau\right) \approx u'\left(t_j\right) + \frac{\alpha}{1!}\Delta u'\left(t_{j-1}\right) + \frac{\alpha(\alpha+1)}{2!}\Delta^2 u'\left(t_{j-2}\right) + \cdots + \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{k!}\Delta^k u'\left(t_{j-k}\right), \quad (1.8)$$

где остаток интерполирования имеет вид

$$r_k\left(t_j + \alpha\tau\right) = \tau^{k+1}\frac{\alpha(\alpha+1)\ldots(\alpha+k)}{(k+1)!}u^{(k+2)}(\xi),\ \ t_{j-k} \le \xi \le t_{j+1}.$$

Подставляя вместо u' ее представление в (1.1′) и выполняя интегрирование, получим экстраполяционный метод Адамса в виде

$$y_{j+1} = y_j + \tau\sum_{i=0}^{k} C_i \Delta^i f_{j-i}, \quad (1.9)$$

где

$$C_i = \frac{1}{i!}\int_0^1 \alpha(\alpha+1)\ldots(\alpha+i-1)d\alpha. \quad (1.10)$$

Очевидно, локальная погрешность метода (1.9), (1.10), учитывая (1.8), будет иметь вид

$$r_k\left(t_j, \tau\right) = \tau^{k+2}\int_0^1 \frac{\alpha(\alpha+1)\ldots(\alpha+k)}{(k+1)!}u^{(k+2)}(\xi)d\alpha = C_{k+1}\tau^{k+2}u^{(k+2)}(\xi'), \quad (1.11)$$

т.е. метод (1.9), (1.10) является методом $(k+1)$-го порядка точности. Из (1.10) следует, что

$$C_0 = \int_0^1 d\alpha = 1;\ \ C_1 = \frac{1}{1!}\int_0^1 \alpha d\alpha = \frac{1}{2};\ \ C_2 = \frac{1}{2!}\int_0^1 \alpha(\alpha+1)d\alpha = \frac{5}{12};\ldots.$$

Таким образом, (1.9) может быть переписан в виде

$$y_{j+1} = y_j + \tau\left(f_j + \frac{1}{2}\Delta f_{j-1} + \frac{5}{12}\Delta^2 f_{j-2} + \cdots + C_k \Delta^k f_{j-k} \right).$$

При этом приближенное решение вновь (как и в пошаговом варианте метода рядов) оказывается разложенным по последовательным главным частям.

Как уже отмечалось ранее, многошаговые методы характеризуются известной неоднородностью вычислительного процесса. Первые k значений y_j (начало таблицы) должны быть вычислены каким-либо другим способом (например, с помощью одношаговых методов).

1.2. Интерполяционные методы Адамса

Все изложенное в предыдущем пункте можно было бы повторить при одном лишь отличии: интерполирование функции $u'(x)$ проводить не по узлам $t_j, t_{j-1}, ..., t_{j-k}$, а по узлам $t_{j+1}, t_j, ..., t_{j-k}$.

Мы, однако, вновь более подробно остановимся на случае равномерной сетки узлов, так как тогда расчетные формулы будут иметь наиболее простой вид.

Сделаем в (1.1) замену переменной $x = t_{j+1} + \alpha\tau$. Получим:

$$u(t_{j+1}) = u(t_j) + \tau \int_{-1}^{0} u'(t_{j+1} + \alpha\tau) d\alpha. \qquad (1.1'')$$

Вновь проинтерполируем подынтегральную функцию по формуле Ньютона для конца таблицы:

$$u'(t_{j+1} + \alpha\tau) \approx u'(t_{j+1}) + \frac{\alpha}{1!}\Delta u'(t_j) + \frac{\alpha(\alpha+1)}{2!}\Delta^2 u'(t_{j-1}) + \cdots + \frac{\alpha(\alpha+1)\ldots(\alpha+k)}{k!}\Delta^{k+1} u'(t_{j-k}).$$

Остаток интерполирования имеет вид

$$\rho_{k+1}(t_{j+1} + \alpha\tau) = \tau^{k+2} \frac{\alpha(\alpha+1)\ldots(\alpha+k+1)}{(k+2)!} u^{(k+3)}(\xi), \quad t_{j-k} \le \xi \le t_{j+1}.$$

Выполнив интегрирование, получим интерполяционный метод Адамса вида

$$y_{j+1} = y_j + \tau \sum_{i=0}^{k+1} C_i^* \Delta^i f_{j+1-i}, \qquad (1.12)$$

где

$$C_i^* = \frac{1}{i!} \int_{-1}^{0} \alpha(\alpha+1)\ldots(\alpha+i-1) d\alpha, \qquad (1.13)$$

а для локальной погрешности метода справедливо представление

$$r_k(t_j, \tau) = \tau^{k+3} \int_0^1 \frac{\alpha(\alpha+1)\ldots(\alpha+k+1)}{(k+2)!} u^{(k+3)}(\xi) d\alpha = C_{k+2}^* \tau^{k+3} u^{(k+3)}(\xi'), \quad t_{j-k} \le \xi' \le t_{j+1}, \qquad (1.14)$$

В частности, из (1.13) следует, что

$$C_0^* = \int_{-1}^{0} d\alpha = 1;\ \ C_1^* = \frac{1}{1!}\int_{-1}^{0} \alpha d\alpha = -\frac{1}{2};\ \ C_2^* = \frac{1}{2!}\int_{-1}^{0} \alpha(\alpha+1) d\alpha = -\frac{1}{12};\ldots,$$

т.е. (1.12) можно переписать в виде

$$y_{j+1} = y_j + \tau\left(f_{j+1} - \frac{1}{2}\Delta f_j - \frac{1}{12}\Delta^2 f_{j-1} - \cdots + C_{k+1}^*\Delta^{k+1} f_{j-k} \right). \qquad (1.12')$$

Замечание. Ранее мы получили два представления для экстраполяционных методов Адамса (через конечные разности и значения функции). Здесь также легко получить второе (через значения функции) представление. Для этого достаточно либо просто заменить в (1.12′) конечные разности их выражениями через значения функции, либо, как и выше, интерполяционный многочлен брать в форме Лагранжа.

В любом варианте получим такое семейство методов:

$y_{j+1} = y_j + \tau f_{j+1}$ – метод первого порядка (неявный метод Эйлера);

$y_{j+1} = y_j + \frac{\tau}{2}\left(f_{j+1} + f_j\right)$ – метод второго порядка (неявный метод трапеций);

$y_{j+1} = y_j + \frac{\tau}{12}\left(5f_{j+1} + 8f_j - f_{j-1}\right)$ – метод третьего порядка;

……

Заметим также, что в компьютерной реализации последние представления используются чаще, нежели представления через конечные разности.

§ 2. Общие линейные многошаговые методы, не использующие старших производных

Рассмотренные выше семейства экстраполяционных и интерполяционных методов Адамса являются частными случаями более общего семейства методов вида

$$\sum_{i=-1}^{k} a_i y_{j-i} = \tau \sum_{i=-1}^{k} b_i f\left(t_{j-i}, y_{j-i}\right), \qquad (2.1)$$

которые носят название линейных многошаговых методов (ЛММ) без старших производных.

Очевидно, для методов Адамса $a_{-1} = 1,\ a_0 = -1,\ a_1 = \ldots = a_k = 0$.

Получим сейчас общий вид условий, которым должны удовлетворять коэффициенты a_i и b_i метода (2.1). Для этих целей воспользуемся той же идеей, что и при рассмотрении методов Рунге-Кутта: порядок метода должен быть максимальным. Запишем выражение для локальной погрешности метода (2.1):

$$r\left(t_j, \tau\right) = \sum_{i=-1}^{k}\left[a_i u\left(t_j - i\tau\right) - \tau b_i f\left(t_j - i\tau, u\left(t_j - i\tau\right)\right)\right]. \qquad (2.2)$$

Поскольку

$$u(t_j - i\tau) = u_j + \frac{(-i)\tau}{1!}u'_j + \frac{(-i)^2\tau^2}{2!}u''_j + \cdots,$$

$$f(t_j - i\tau, u(t_j - i\tau)) = u'(t_j - i\tau) = u'_j + \frac{(-i)\tau}{1!}u''_j + \frac{(-i)^2\tau^2}{2!}u'''_j + \cdots,$$

то, подставив эти разложения в (2.2), будем иметь:

$$r(t_j, \tau) = \left(\sum_{i=-1}^{k} a_i\right)u_j - \frac{\tau}{1!}\sum_{i=-1}^{k}(ia_i + b_i)u'_j + \frac{\tau^2}{2!}\sum_{i=-1}^{k}(i^2 a_i + 2ib_i)u''_j + \cdots + \frac{(-\tau)^l}{l!}\sum_{i=-1}^{k} i^{l-1}(ia_i + lb_i)u_j^{(l)} + \cdots.$$

Отсюда видно, что метод (2.1) будет иметь порядок точности p, если выполнены условия

$$\begin{cases} \sum_{i=-1}^{k} a_i = 0, \\ \sum_{i=-1}^{k}(ia_i + lb_i)i^{l-1} = 0, \quad l = 1, 2, ..., p. \end{cases} \tag{2.3}$$

Используя условия порядка (2.3), можно на основе конструкции (2.1) получить большинство из используемых ныне в вычислительной практике линейных многошаговых методов.

***Упражнение* 1.** Получить условия порядка, аналогичные (2.3), для экстраполяционных и интерполяционных методов Адамса.

***Замечание* 1.** Аналогичную (2.1) конструкцию можно получить, если отказаться от «запрета» на использование старших производных от решения. Она будет иметь вид

$$\sum_{i=-1}^{k} a_i y_{j-i} = \sum_{l=0}^{p} \tau^{l+1} \sum_{i=-1}^{k} b_{il} f^{(l)}(t_{j-i}, y_{j-i}). \tag{2.4}$$

В литературе эта конструкция носит название обобщенных линейных многошаговых методов.

***Упражнение* 2.** Получить условия порядка для обобщенных ЛММ.

***Замечание* 2.** Чаще оказывается целесообразным часть параметров метода отдать не на достижение максимального порядка аппроксимации, а на то, чтобы добиться выполнения некоторых других важных свойств (например, расширения области устойчивости).

§ 3. Понятие устойчивости численных методов решения задачи Коши

Вновь, если не оговорено особо, будем предполагать, что рассматривается случай одного обыкновенного дифференциального уравнения первого порядка.

Ранее мы рассматривали достаточно общие определения, касающиеся таких понятий как «корректность», «устойчивость», «сходимость» численных методов. Естественно, такие общие определения очень часто нуждаются в конкретизации, для того чтобы их можно было реально применить к исследованию конкретных свойств того или иного алгоритма.

С этих позиций обсудим сейчас термин «устойчивость» применительно к методам решения задачи Коши.

Прежде всего, заметим, что все численные методы решения задачи Коши представляют собой ***разностные уравнения*** различных порядков: одношаговые методы – первого порядка, многошаговые (k-шаговые) – k-го порядка.

Далее, вполне очевидным представляется тот факт, что численный метод должен более или менее адекватно отражать истинные свойства решения исследуемой (решаемой) дифференциальной задачи и причем, желательно, в достаточно широком диапазоне правых частей уравнения (т.е. функций $f(t,u)$).

В частности, положив $f(t,u)\equiv 0$, получим следующую запись линейного многошагового метода (2.1):

$$a_{-1}y_{j+1}+a_0y_j+\cdots+a_ky_{j-k}=0,$$

т.е. будем иметь линейное разностное уравнение порядка $(k+1)$ с постоянными коэффициентами. Его общее решение может быть записано в виде

$$y_j=\sum_{i=-1}^{k}C_iq_i^j,$$

где C_i – коэффициенты, определяемые из дополнительных (например, начальных) условий, а q_i, $i=\overline{0,k}$ – простые вещественные корни характеристического уравнения (в других случаях все организуется аналогично дифференциальному случаю)

$$a_{-1}q^{k+1}+a_0q^k+\cdots+a_k=0. \qquad (3.1)$$

Отсюда следует, что если $|q_i|>1$ хотя бы для некоторого значения i, то $|y_j|\underset{j\to\infty}{\to}\infty$, в то время как решение исходного уравнения $u'=f(t,u)$ при $f(t,u)\equiv 0$, очевидно, представляет собой константу.

Аналогичная ситуация имеет место и в случае, если $|q_{i_0}|=1$, но кратность этого корня выше единицы, так как тогда в конструкции общего решения перед сомножителем $q_{i_0}^j$ появляется многочленный по j коэффициент, учитывающий кратность.

Определение* 1.** Будем говорить, что численный метод удовлетворяет ***условию корней, если все корни $q_0,...,q_k$ характеристического уравнения (3.1) лежат внутри или на границе единичного круга комплексной плоскости, причем на границе круга нет кратных корней.

Очевидно, метод, не удовлетворяющий условию корней (или корневому условию), для вычислений не пригоден. Это следует иметь в виду при построении различных алгоритмов. Заметим, что рассмотренные нами ранее классы методов (как Адамса, так и одношаговые типа Рунге-Кутта и последовательного повышения порядка точности) корневому условию удовлетворяют, поскольку для них характеристическое уравнение (3.1) будет иметь вид

$$a_{-1}q+a_0=0$$

при $a_{-1}=1,\ a_0=-1$, т.е.

$$q-1=0$$

и имеет единственный корень $q=1$.

Однако не все численные методы, удовлетворяющие корневому условию, всегда пригодны для расчетов. Понятно, что при произвольной правой части дифференциального

уравнения (функции f) вряд ли возможно получить сколько-нибудь эффективных оценок по этому поводу. Поэтому в отличие от такого свойства численных методов как аппроксимация, о которой мы говорили выше, и которая может быть исследована принципиально при любой правой части, другие свойства приближенных методов, характеризующие ***качественное поведение*** приближенного решения, могут быть исследованы только на задачах определенного вида – ***модельных*** уравнениях.

В частности, при исследовании свойства устойчивости в качестве такого модельного уравнения чаще всего рассматривают уравнение

$$u'(t) = \lambda u(t), \tag{3.2}$$

где λ – произвольное комплексное число с отрицательной вещественной частью $(\operatorname{Re}\lambda < 0)$. Отчасти выбор в качестве модели уравнения (3.2) объясняется тем, что оно представляет собой линейное (однородное) приближение к задаче общего вида (вспомним, что исследование устойчивости по Ляпунову для дифференциальных уравнений проводится на линейном приближении).

Отметим, что поскольку точное решение уравнения (3.2) имеет вид $u(t) = C\,\mathrm{e}^{\lambda t}$, то при указанных значениях λ задача (3.2) устойчива, причем $u(t) \underset{t\to\infty}{\to} 0$ и в частности, при любом значении шага τ выполняется условие

$$|u(t+\tau)| \le |u(t)|,$$

т.е. модуль решения монотонно не возрастает. Естественно было бы потребовать качественно такого же поведения и от приближенного решения, доставляемого тем или иным численным методом.

Легко видеть, что применение линейного многошагового метода (2.1) к решению уравнения (3.2) приводит к разностному уравнению вида

$$\sum_{i=-1}^{k} (a_i - z b_i) y_{j-i} = 0, \tag{4.3}$$

где $z = \tau\lambda$.

Аналогичные разностные уравнения (но только ***первого*** порядка) получаются, если к решению уравнения (3.2) применить соответствующий метод Рунге-Кутта или последовательного повышения порядка точности.

Определение* 2.** Численный метод решения задачи Коши будем называть ***устойчивым при некотором значении z, если при данном значении z устойчиво соответствующее ему разностное уравнение (типа (3.3)), получающееся вследствие применения исследуемого метода к решению модельного уравнения (3.2).

Очевидно, для того чтобы метод был устойчивым, достаточно, чтобы все корни соответствующего характеристического уравнения по модулю не превосходили единицы.

Определение* 3. *Областью устойчивости численного метода будем называть множество всех точек z комплексной плоскости, для которых данный метод устойчив.

Определение* 4. *Интервалом устойчивости численного метода будем называть пересечение области устойчивости с вещественной осью координат.

Рассмотрим некоторые примеры.

1^0. Явный метод Эйлера:

$$y_{j+1} = y_j + \tau f_j.$$

Применяя его к уравнению (3.2), будем иметь:

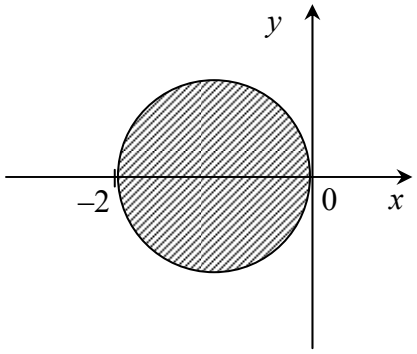
$$y_{j+1} = (1+z)y_j .$$

Единственный корень соответствующего характеристического уравнения здесь $q = 1+z$. Поэтому условие устойчивости примет вид

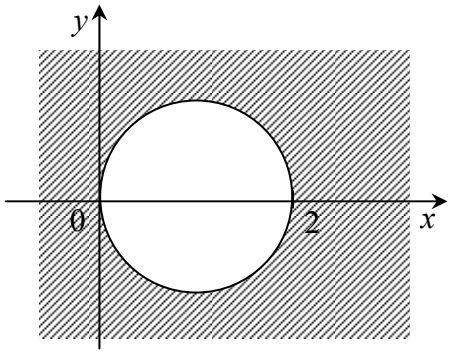
$$|1+z| \le 1 . \qquad (3.4)$$

Если λ – комплексное число, то, записав z в виде $z = x+iy$, перепишем (3.4) в виде

$$|1+x+iy| \le 1 \qquad \text{или} \qquad (1+x)^2 + y^2 \le 1 .$$



Очевидно, последнее неравенство, описывающее область устойчивости явного метода Эйлера, задает в комплексной плоскости круг единичного радиуса с центром в точке $(-1;0)$. Пересечением данной области с вещественной осью будет отрезок $[-2;0]$, который и будет интервалом устойчивости явного метода Эйлера. (Заметим, что интервал устойчивости на самом деле искать гораздо проще, чем область, поскольку для этого достаточно просто решить неравенство (3.4) над полем вещественных чисел).



2^0. Неявный метод Эйлера:

$$y_{j+1} = y_j + \tau f_{j+1} ,$$

откуда

$$(1-z)y_{j+1} = y_j \qquad \text{или} \qquad y_{j+1} = \frac{1}{1-z} y_j .$$

Тогда $q = \dfrac{1}{1-z}$ и условие устойчивости примет вид

$$\left|\frac{1}{1-z}\right| \le 1 \qquad \text{или} \qquad |1-z| \ge 1 .$$

Таким образом, областью устойчивости неявного метода Эйлера является внешность круга единичного радиуса с центром в точке $(1;0)$. Интервалом устойчивости является вся числовая прямая, кроме промежутка $(0;2)$. Следовательно, при $\operatorname{Re}\lambda < 0$ шаг численного интегрирования τ может быть любым (в то время как для явного Эйлера от ограничен сверху величиной $\dfrac{2}{|\lambda|}$).

Определение* 5.** Численный метод будем называть ***A-устойчивым, если его область устойчивости содержит всю левую полуплоскость $\operatorname{Re} z < 0$.

Сущность данного определения состоит в том, что A-устойчивый метод является абсолютно (т.е. при любых $\tau > 0$) устойчивым, если устойчиво решение исходного дифференциального уравнения. Отметим, что неявный метод Эйлера является A-устойчивым, а явный – нет. Заметим, однако, что неявный метод Эйлера является и в той области, где исходная задача является неустойчивой (!).

Справедлива следующая

Теорема. Не существует явных линейных А-устойчивых методов.

Это означает, что при использовании таких методов всегда будут иметь место ограничения на выбор допустимой величины шага τ, подобные ограничению, характерному для явного метода Эйлера. В то же время среди неявных линейных методов А-устойчивые, как мы видели, существуют.

Подводя итоги, можем сформулировать следующую процедуру исследования устойчивости численных методов решения задачи Коши:

1^0. Применяя исследуемый метод к решению уравнения (3.2), получаем разностное уравнение, которому удовлетворяет приближенное решение;

2^0. Записываем соответствующее характеристическое уравнение;

3^0. Находим корни характеристического уравнения $(q_i,\ i=1,...,k)$;

4^0. Решение системы неравенств $|q_i|\le 1,\ i=1,...,k,$ дает искомую область устойчивости.

Заметим, однако, что практическая реализация изложенного алгоритма может натолкнуться на значительные технические трудности (особенно это касается случаев комплексных λ для многостадийных методов, а также методов многошаговых). Поэтому практически для построения областей устойчивости используют прием, который носит название ***метод множества точек границы*** и состоит в следующем: точка z комплексной плоскости будет принадлежать границе области устойчивости, если при данном значении z выполняется равенство $\max\limits_i|q_i|=1\overset{def}{=}|q^*|$ или $q^*=\mathrm{e}^{i\varphi}$, $\varphi\in[0;2\pi)$ (здесь i – мнимая единица). Решая записанное уравнение относительно z (возможно, при фиксированных значениях φ из указанного промежутка), мы получаем множество точек, составляющих границу области устойчивости. Далее остается определить (например, путем подстановки), по какую сторону границы находится сама область.

Приведем примеры исследования устойчивости:

1^0. Метод последовательного повышения порядка точности второго порядка (см. (3.10) предыдущей главы)

$$\begin{cases} y_{j+\frac12}=y_j+\dfrac{\tau}{2}f_j,\\ y_{j+1}=y_j+\tau f_{j+\frac12}. \end{cases}$$

Применяя данный метод к решению уравнения (3.2), получим:

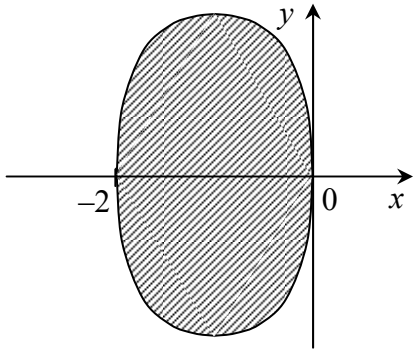
$$y_{j+1}=\left(1+z+\frac{z^2}{2}\right)y_j.$$

Отсюда

$$q=1+z+\frac{z^2}{2}=e^{i\varphi}$$

и

$$z=z(\varphi)=-1\pm\sqrt{2e^{i\varphi}-1}.$$



Кривая $z(\varphi)$ и есть граница области устойчивости, а сама область есть внутренность данной кривой (изображена на рисунке).

***Упражнение* 1.** Построить области устойчивости методов Рунге-Кутта третьего и четвертого порядков точности.

2^0. Экстраполяционный метод Адамса второго порядка (см. (1.6), (1.7) данной главы)

$$y_{j+1} = y_j + \frac{\tau}{2}\left(3f_j - f_{j-1}\right).$$

Разностное уравнение имеет вид

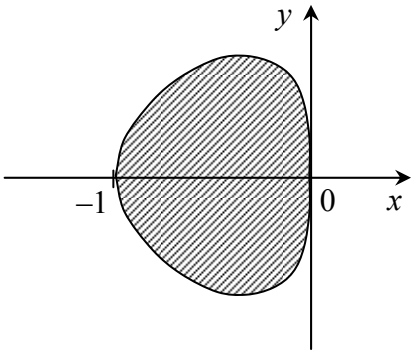
$$y_{j+1} - \left(1 + \frac{3}{2}z\right)y_j + \frac{z}{2}y_{j-1} = 0.$$

Тогда характеристическое уравнение будет таким:

$$q^2 - \left(1 + \frac{3}{2}z\right)q + \frac{z}{2} = 0.$$

Подставим сюда $q = \mathrm{e}^{i\varphi}$ и найдем границу области устойчивости:

$$z(\varphi) = 2\frac{\mathrm{e}^{2i\varphi} - \mathrm{e}^{i\varphi}}{3\mathrm{e}^{i\varphi} - 1}. \qquad (*)$$



Область устойчивости, являющаяся внутренностью данной кривой, изображена на рисунке.

***Замечание* 1.** Для определения интервала устойчивости вдоль вещественной оси для линейных многошаговых методов в уравнение типа (*) достаточно подставить $\varphi = \pi$. Так, для указанного выше метода получим левую границу интервала

$$z(\pi) = 2 \cdot \frac{1+1}{3 \cdot (-1) - 1} = -1,$$

т.е. экстраполяционный метод Адамса второго порядка устойчив на отрезке $z \in [-1;0]$.

***Упражнение* 2.** Исследовать устойчивость всех методов Адамса, рассмотренных выше.

***Замечание* 2.** В случае систем обыкновенных дифференциальных уравнений можно показать, что соответствующие условия устойчивости будут иметь такой же вид, как и в случае одного уравнения, но с заменой параметра λ на максимальное по модулю собственное значение матрицы Якоби системы.

ГЛАВА XII

Жесткие задачи и методы их решения

§ 1. Явление жесткости

Задачи, называемые жесткими, весьма разнообразны, и дать математически строгое определение жесткости непросто. Поэтому в литературе можно встретить различные определения жесткости, отличающиеся степенью строгости. Сущность же явления жесткости состоит в том, что решение, которое необходимо вычислить, меняется медленно, однако в любой его окрестности существуют быстро затухающие возмущения. Характерное время затухания их называют пограничным слоем. Наличие таких возмущений затрудняет получение медленно меняющегося решения численным способом. При этом жесткими могут как скалярные дифференциальные уравнения, так и, что встречается осо-

бенно часто, системы обыкновенных дифференциальных уравнений. Приведем вначале некоторые примеры.

1^0. Скалярное уравнение

$$\begin{cases} u'(t) = \lambda u(t) + F'(t) - \lambda F(t), \ \ t > 0, \ \lambda << 0, \\ u(0) = u_0, \end{cases} \tag{1.1}$$

где $F(t)$ – медленно меняющаяся функция, зависящая только от t (например, $F(t) = \operatorname{th} t$). Решение задачи (1.1) имеет вид

$$u(t) = F(t) + \mathrm{e}^{\lambda t}[u_0 - F(0)].$$

Так как $\lambda << 0$, то ясно, что уже после очень небольшого отрезка времени второе слагаемое в решении практически отсутствует и, таким образом, на большей части отрезка интегрирования преобладает медленно меняющаяся функция, которая принципиально может быть достаточно хорошо приближенно описана на сетке с крупным шагом τ. В то же время, если применить к решению задачи (1.1) явный метод Эйлера (или любой другой явный метод типа Рунге-Кутта), то легко видеть, что допустимый шаг интегрирования, как и в случае модельного уравнения (3.2) предыдущей главы, будет определяться величиной $\tau \le -\dfrac{2}{\lambda}$ (или $\tau \le -\dfrac{a}{\lambda}$), т.е. будет очень малым (***показать!***).

2^0. Рассмотрим теперь систему из двух независимых уравнений

$$\begin{cases} u_1'(t) = -\lambda_1 u_1(t), \\ u_2'(t) = -\lambda_2 u_2(t), \ \ t > 0, \ \ \lambda_2 >> \lambda_1 > 0. \end{cases} \tag{1.2}$$

Эта система имеет решение $u(t) = (u_1(t), u_2(t))^T = \left(u_1^0 \mathrm{e}^{-\lambda_1 t}, u_2^0 \mathrm{e}^{-\lambda_2 t}\right)^T$. При выписанных условиях на λ_1 и λ_2, очевидно, компонента $u_2(t)$ решения затухает гораздо быстрее, чем $u_1(t)$ и, начиная с некоторого момента t поведение вектора $u(t)$ почти полностью определяется компонентой $u_1(t)$. Однако при решении системы (1.2) численным методом величина шага интегрирования, как правило, определяется компонентой $u_2(t)$, не существенной с точки зрения поведения решения системы. Например, используя тот же явный метод Эйлера, мы из первого уравнения имеем ограничение на шаг $\tau \le \dfrac{2}{\lambda_1}$, а из второго – $\tau \le \dfrac{2}{\lambda_2}$ и, таким образом, ясно, что для решения системы (1.2) как цельного математического объекта шаг τ ограничен величиной $\dfrac{2}{\lambda_2}$.

Такая же ситуация типична и при решении любой системы обыкновенных дифференциальных уравнений вида

$$u'(t) = Au(t), \tag{1.3}$$

если матрица этой системы имеет большой разброс собственных значений.

Определение. Система обыкновенных дифференциальных уравнений (1.3) с постоянной $(n \times n)$-матрицей A называется жесткой, если:

1) $\operatorname{Re}\lambda_k < 0,\ k = \overline{1,n}$ (т.е. задача устойчива);

2) отношение $S = \dfrac{\max\limits_{1\le k\le n}|\operatorname{Re}\lambda_k|}{\min\limits_{1\le k\le n}|\operatorname{Re}\lambda_k|}$ велико (например, $S > 10$);

3) промежуток интегрирования велик по сравнению с длиной погранслоя.

Число S иногда называют коэффициентом жесткости системы (1.3).

Если в (1.3) матрица A будет зависеть от t, то, очевидно, и $S = S(t)$, т.е. коэффициент жесткости может меняться с течением времени.

Поскольку система нелинейных обыкновенных дифференциальных уравнений вида $u'(t) = f(t,u(t))$ может быть в окрестности некоторого известного решения $v(t)$ заменена линейной системой

$$u'(t) = f_u(t, v+\theta(u-v))u + b(t),$$

где f_u – матрица Якоби системы, а $b(t) = f(t,v) - f_u(t, v+\theta(u-v))v$, то понятие жесткости для нелинейных систем может быть определено аналогично. Заметим, однако, что за пределами класса систем линейных обыкновенных дифференциальных уравнений с постоянной матрицей полагаться на спектр как на источник надежной информации о распространении погрешности уже нельзя (это показывают известные из литературы примеры (см., например, Деккер, Вервер: Устойчивость методов Рунге-Кутта для жестких нелинейных дифференциальных уравнений)).

§ 2. Методы, применяемые для решения жестких систем

Учитывая все сказанное выше, можно сделать вывод, что для решения жестких задач наиболее пригодны те численные методы, которые требуют наиболее слабых ограничений на величину шага численного интегрирования из соображений устойчивости. В настоящее время наиболее часто для этих целей используют либо неявные методы, среди которых, как мы видели, встречаются А-устойчивые (правда, при этом не следует думать, что все неявные методы будут в этом смысле хорошими), либо методы, специально сконструированные для решения задач конкретного вида.

2.1. Неявные методы Рунге-Кутта

Важным классом одношаговых методов, применяемых для решения жестких задач, являются неявными методами неявные методы Рунге-Кутта, вид которых и технология построения укладываются в общую схему, изложенную ранее (глава X, § 2). Пользуясь полученными там же условиями порядка, рассмотрим подробнее некоторые частные случаи.

1^0. Одностадийные методы.

Эти методы имеют вид

$$\begin{array}{c|c} c_1 & a_{11} \\ \hline & b_1 \end{array}$$

В отличие от аналогичных явных методов Рунге-Кутта, они зависят от трех параметров. Поэтому здесь возможно построение методов не только первого порядка.

1) методы первого порядка.

При выводе условий порядка (формула (2.28) главы X) мы фактически не пользовались свойством явности. И действительно, эти условия оказываются универсальными и в нашем случае имеют вид

$$\begin{cases} b_1 = 1, \\ c_1 = a_{11}. \end{cases} \tag{2.1}$$

Таким образом, имеем два уравнения с тремя неизвестными. Полагая $c_1 = a_{11} = \alpha$, получим однопараметрическое семейство методов первого порядка

$$\begin{array}{c|c} \alpha & \alpha \\ \hline & 1 \end{array}$$

или в развернутом виде

$$\begin{cases} y_{j+1} = y_j + \tau k_1, \\ k_1 = f\left(t_j + \alpha\tau, y_j + \alpha\tau k_1\right). \end{cases} \tag{2.2}$$

При выработке рекомендаций по выбору параметра вспомним, что основной целью при конструировании неявных методов является, по сути дела, расширение области устойчивости (вплоть до А-устойчивости). Поэтому проведем исследование устойчивости построенного семейства. Полагая $f = \lambda y$, имеем:

$$k_1 = \lambda\left(y_j + \alpha\tau k_1\right),$$

откуда

$$k_1 = \frac{\lambda}{1-\alpha z} y_j$$

и, следовательно,

$$y_{j+1} = y_j + \frac{z}{1-\alpha z} y_j,$$

т.е.

$$y_{j+1} = \frac{1+(1-\alpha)z}{1-\alpha z} y_j.$$

Неравенство

$$\left|\frac{1+(1-\alpha)z}{1-\alpha z}\right| \le 1$$

равносильно неравенству

$$|1+(1-\alpha)z|^2 \le |1-\alpha z|^2$$

или

$$(1+(1-\alpha)x)^2 + (1-\alpha)^2 y^2 \le (1-\alpha x)^2 + \alpha^2 y^2,$$

где x и y – соответственно вещественная и мнимая части комплексного числа $z = \tau\lambda$. Приводя в последнем неравенстве подобные, перепишем его в виде

$$2x + (1-2\alpha)x^2 + (1-2\alpha)y^2 \le 0. \tag{2.3}$$

Теперь остается рассмотреть три случая:

а) $\alpha = \frac{1}{2}$. В этом случае (2.3) превращается в неравенство $x \le 0$, что равносильно А-устойчивости (2.2), так как область устойчивости в точности совпадает с левой полуплоскостью;

б) $\alpha > \frac{1}{2}$. В этом случае (2.3) может быть переписано в виде

$$\left(x - \frac{1}{2\alpha - 1}\right)^2 + y^2 \geq \left(\frac{1}{2\alpha - 1}\right)^2.$$

Последнее же означает, что областью устойчивости является внешность круга радиуса $\frac{1}{2\alpha - 1}$ с центром в точке $\left(\frac{1}{2\alpha - 1}, 0\right)$ и, как легко видеть, эта область целиком содержит всю левую полуплоскость, т.е. для всех рассматриваемых α метод (2.2) также является А-устойчивым;

в) $\alpha < \frac{1}{2}$. В этом случае (2.3) следует переписать в виде

$$\left(x + \frac{1}{1 - 2\alpha}\right)^2 + y^2 \leq \left(\frac{1}{1 - 2\alpha}\right)^2.$$

Следовательно, областью устойчивости является внутренность круга радиуса $\frac{1}{1 - 2\alpha}$ с центром в точке $\left(-\frac{1}{1 - 2\alpha}, 0\right)$, т.е. (2.2) в этом случае является условно устойчивым.

Таким образом, проведенный анализ показывает, что выбор параметра α в (2.2) должен быть подчинен условию $\alpha \geq \frac{1}{2}$.

2) методы второго порядка

В этом случае к уравнениям (2.1), рассмотренным выше, добавляется еще одно уравнение:

$$b_1 c_1 = \frac{1}{2}.$$

Получившаяся система из трех уравнений с тремя неизвестными, как легко видеть, имеет единственное решение: $b_1 = 1;\ c_1 = a_{11} = \frac{1}{2}$. Таким образом, единственный одностадийный метод второго порядка имеет вид

$\frac{1}{2}$	$\frac{1}{2}$
	1

или в развернутом виде

$$\begin{cases} y_{j+1} = y_j + \tau k_1, \\ k_1 = f\left(t_j + \frac{\tau}{2}, y_j + \frac{\tau}{2} k_1\right). \end{cases} \tag{2.4}$$

Очевидно, полученный метод является частным случаем семейства (2.2), соответствующим случаю $\alpha = \frac{1}{2}$, и потому является А-устойчивым.

2^0. Двухстадийные методы.

В соответствии с общей схемой эти методы в форме Батчера имеют вид

c_1	a_{11}	a_{12}
c_2	a_{21}	a_{22}
	b_1	b_2

и, таким образом, зависят от восьми произвольных параметров. Учитывая это, исследование начнем со случая

а) методы второго порядка.

Условия порядка в этом случае будут иметь вид

$$\begin{cases} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \dfrac{1}{2}, \\ c_1 = a_{11} + a_{12}, \\ c_2 = a_{21} + a_{22}. \end{cases} \tag{2.5}$$

В этом варианте свободных параметров все равно достаточно много (четыре), и мы, не задаваясь целью провести полное исследование, распорядимся их выбором с целью упрощения реализации получаемых методов. Для этого зададим верхнюю строку таблицы Батчера нулевой (т.е. положим $a_{11} = a_{12} = c_1 = \ = 0$). Кроме того, положим $c_2 = \alpha$, $a_{22} = \beta$. Тогда оставшиеся неизвестные однозначно определятся из системы (2.5): $a_{21} = \alpha - \beta,\ b_2 = \dfrac{1}{2\alpha},\ b_1 = \dfrac{2\alpha - 1}{2\alpha}$. Таким образом, получаем двухпараметрическое семейство методов второго порядка точности

0	0	0
α	$\alpha - \beta$	β
	$\dfrac{2\alpha - 1}{2\alpha}$	$\dfrac{1}{2\alpha}$

или в развернутой форме

$$\begin{cases} y_{j+1} = y_j + \dfrac{\tau}{2\alpha}[(2\alpha - 1)k_1 + k_2], \\ k_1 = f(t_j, y_j), \\ k_2 = f(t_j + \alpha\tau, y_j + \tau[(\alpha - \beta)k_1 + \beta k_2]). \end{cases} \tag{2.6}$$

Заметим, что построенный метод является своеобразным аналогом семейства явных двухстадийных методов Рунге-Кутта второго порядка (см. (2.11) главы X).

Вновь анализ возможного выбора параметров проведем на основе требования устойчивости. Применяя (2.6) к модельному уравнению, будем иметь:
$k_1 = \lambda y_j$. Тогда для определения k_2 получим уравнение

$$k_2 = \lambda\left[y_j + \tau(\alpha - \beta)\lambda y_j + \tau\beta k_2\right],$$

откуда

$$k_2 = \lambda y_j \cdot \frac{1 + (\alpha - \beta)z}{1 - \beta z}.$$

Следовательно,

$$y_{j+1} = y_j + \frac{\tau}{2\alpha}\left[(2\alpha - 1)\lambda y_j + \lambda y_j \cdot \frac{1 + (\alpha - \beta)z}{1 - \beta z}\right] = y_j\left[1 + \frac{z}{2\alpha}\left(2\alpha - 1 + \frac{1 + (\alpha - \beta)z}{1 - \beta z}\right)\right] =$$

$$= y_j\left[1+\frac{z}{2\alpha}\cdot\frac{2\alpha-1-\beta(2\alpha-1)z+1+(\alpha-\beta)z}{1-\beta z}\right] = y_j\left[1+\frac{z}{2\alpha}\cdot\frac{2\alpha+\alpha z(1-2\beta)}{1-\beta z}\right] =$$

$$= y_j\left[1+\frac{z+z^2\left(\frac{1}{2}-\beta\right)}{1-\beta z}\right] = y_j\cdot\frac{1+(1-\beta)z+\left(\frac{1}{2}-\beta\right)z^2}{1-\beta z}.$$

Отсюда видим, что требование А-устойчивости может быть выполнено лишь при значении параметра β, равном $\frac{1}{2}$ (так как в противном случае степень числителя множителя перехода будет выше степени знаменателя, и, следовательно, при больших по модулю значениях z множитель перехода будет заведомо больше единицы по модулю, при $\beta=\frac{1}{2}$ мы получим исследованный выше случай). Таким образом, в (2.6) параметр β следует полагать равным $\frac{1}{2}$. Отметим в этом варианте два частных значения α:

1) $\alpha = 1$. В этом случае получаем метод,

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

являющийся аналогом неявного метода трапеций, о котором мы упоминали ранее.

2) $\alpha = \frac{1}{2}$. Получающийся в этом случае метод

0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$
	0	1

по сути, является методом (2.4) (т.е. одностадийным (!)) (аналог неявной формулы средних прямоугольников).

б) Методы третьего порядка.

Условия порядка имеют вид

$$\begin{cases} b_1+b_2=1, \\ b_1c_1+b_2c_2=\frac{1}{2}, \\ b_1c_1^2+b_2c_2^2=\frac{1}{3}, \\ b_1(a_{11}c_1+a_{12}c_2)+b_2(a_{21}c_1+a_{22}c_2)=\frac{1}{6}, \\ c_1=a_{11}+a_{12}, \\ c_2=a_{21}+a_{22}. \end{cases}$$

Таким образом, здесь шесть уравнений и восемь неизвестных. Исходя из соображений простоты дальнейшей реализации метода, положим $a_{12} = 0,\ a_{11} = a_{22}$. Тогда, с учетом пятого уравнения, систему можно переписать в виде

$$\begin{cases} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \dfrac{1}{2}, \\ b_1 c_1^2 + b_2 c_2^2 = \dfrac{1}{3}, \\ b_1 c_1^2 + b_2 (a_{21} c_1 + c_1 c_2) = \dfrac{1}{6}, \\ c_2 = a_{21} + c_1. \end{cases}$$

Подставляя в четвертое уравнение вместо a_{21} его выражение через c_1 и c_2, получим систему из четырех уравнений с четырьмя неизвестными:

$$\begin{cases} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \dfrac{1}{2}, \\ b_1 c_1^2 + b_2 c_2^2 = \dfrac{1}{3}, \\ b_1 c_1^2 + b_2 \left(2 c_1 c_2 - c_1^2\right) = \dfrac{1}{6}. \end{cases}$$

Вычитая из второго уравнения последней системы первое, умноженное на c_1, из третьего – второе, умноженное на c_1, и из четвертого – третье, перепишем систему в виде

$$\begin{cases} b_1 + b_2 = 1, \\ b_2 (c_2 - c_1) = \dfrac{1}{2} - c_1, \\ b_2 c_2 (c_2 - c_1) = \dfrac{1}{3} - \dfrac{1}{2} c_1, \\ b_2 (c_2 - c_1)^2 = \dfrac{1}{6}. \end{cases} \qquad (*)$$

Разделив третье уравнение полученной системы на второе, получим:

$$c_2 = \frac{\dfrac{1}{3} - \dfrac{1}{2} c_1}{\dfrac{1}{2} - c_1}. \qquad (2.7)$$

Проделав аналогичную операцию с четвертым и вторым уравнениями, найдем:

$$c_2 - c_1 = \frac{\dfrac{1}{6}}{\dfrac{1}{2} - c_1}$$

или

$$c_2 = \frac{\frac{1}{2}c_1 - c_1^2 + \frac{1}{6}}{\frac{1}{2} - c_1}. \qquad (2.8)$$

Приравнивая теперь правые части формул (2.7) и (2.8), получим уравнение

$$\frac{1}{3} - \frac{1}{2}c_1 = \frac{1}{2}c_1 - c_1^2 + \frac{1}{6}$$

или

$$c_1^2 - c_1 + \frac{1}{6} = 0.$$

Отсюда

$$c_1 = \frac{3 \pm \sqrt{3}}{6}.$$

Для определенности положим $c_1 = \frac{3-\sqrt{3}}{6}$. Тогда из (2.7) следует, что $c_2 = \frac{3+\sqrt{3}}{6}$. Следовательно (четвертое из уравнений системы (*)), $b_2 = \frac{1}{2}$. Наконец, $b_1 = \frac{1}{2}$ и $a_{21} = c_2 - c_1 = \frac{\sqrt{3}}{3}$.

Таким образом, искомый метод третьего порядка будет иметь вид

$\frac{3-\sqrt{3}}{6}$	$\frac{3-\sqrt{3}}{6}$	0
$\frac{3+\sqrt{3}}{6}$	$\frac{\sqrt{3}}{3}$	$\frac{3-\sqrt{3}}{6}$
	$\frac{1}{2}$	$\frac{1}{2}$

или в развернутой форме

$$\begin{cases} y_{j+1} = y_j + \dfrac{\tau}{2}[k_1 + k_2], \\ k_1 = f\left(t_j + \dfrac{3-\sqrt{3}}{6}\tau, y_j + \dfrac{3-\sqrt{3}}{6}\tau k_1\right), \\ k_2 = f\left(t_j + \dfrac{3+\sqrt{3}}{6}\tau, y_j + \tau\left[\dfrac{\sqrt{3}}{3}k_1 + \dfrac{3-\sqrt{3}}{6}k_2\right]\right). \end{cases} \qquad (2.9)$$

(2.9) – пример ***диагонально неявного*** метода.

***Упражнение* 1.** Исследовать устойчивость метода (2.9).

***Упражнение* 2.** Построить и исследовать двухстадийный метод четвертого порядка точности.

2.2. Многошаговые методы, основанные на формулах дифференцирования назад (ФДН-методы или BDF-методы)

Положим в конструкции (2.1) общих линейных многошаговых методов (§ 2 предыдущей главы) $b_0 = b_1 = \ldots = b_k = 0$. Тогда получим конструкцию следующего вида:

$$\sum_{i=-1}^{k} a_i y_{n-i} = \tau f\left(t_{j+1}, y_{j+1}\right). \tag{2.10}$$

Именно эта конструкция в литературе и носит название формул дифференцирования назад. Смысл ее состоит в том, чтобы не использовать в алгоритме вычисление производных от решения (или правых частей дифференциальной задачи) в тех точках сетки, значение в которых уже известно, причем приближенно. В этом случае можно надеяться на достижение более сильных свойств устойчивости, чем, скажем, для рассматривавшихся нами ранее (***тоже неявных*** (**!**)) интерполяционных методов Адамса.

Условия порядка (2.3) (см. тот же § 2 предыдущей главы) для (2.10) примут вид

$$\begin{cases} \sum\limits_{i=-1}^{k} a_i = 0, \\ \sum\limits_{i=-1}^{k} i^l a_i = (-1)^l l, \;\; l = 1, 2, \ldots, p. \end{cases} \tag{2.11}$$

Построим конкретные примеры таких методов:

1^0. При $p = 1$ имеем два уравнения. Поэтому достаточно положить $k = 0$. Тогда (2.11) примет вид

$$\begin{cases} a_{-1} + a_0 = 0, \\ -a_{-1} = -1, \end{cases}$$

откуда $a_{-1} = 1,\ a_0 = -1$. Соответствующий метод дифференцирования назад первого порядка будет иметь выглядеть следующим образом:

$$y_{j+1} - y_j = \tau f\left(t_{j+1}, y_{j+1}\right).$$

Легко узнать в нем изучавшийся нами ранее неявный метод Эйлера (который, как мы помним, является А-устойчивым).

2^0. Пусть теперь $p = 2$. В этом случае уравнений будет три и k достаточно положить равным 2. Тогда (2.11) примет вид

$$\begin{cases} a_{-1} + a_0 + a_1 = 0, \\ -a_{-1} + a_1 = -1, \\ a_{-1} + a_1 = 2 \end{cases}$$

и имеет решение $a_1 = \dfrac{1}{2},\ a_{-1} = \dfrac{3}{2},\ a_0 = -2$, что приводит к методу дифференцирования назад второго порядка

$$\frac{3}{2} y_{j+1} - 2 y_j + \frac{1}{2} y_{j-1} = \tau f\left(t_{j+1}, y_{j+1}\right). \tag{2.12}$$

***Упражнение* 1.** Показать, что метод (2.12) является А-устойчивым.

***Упражнение* 2.** Построить методы дифференцирования назад третьего и четвертого порядков и исследовать их на устойчивость.

Результаты выполнения упражнений должны помочь нам убедиться в том, что заявленные выше улучшенные свойства устойчивости формул дифференцирования назад по сравнению с соответствующими интерполяционными методами Адамса действительно имеют место, что позволяет использовать их для решения жестких задач.

2.3. Реализация неявных методов

Непосредственно по виду всех рассмотренных нами выше неявных методов можно сделать вывод, что для того, чтобы превратить их в алгоритмы для вычисления приближенного решения задачи Коши в очередной точке сеточной области, необходимо «приложить» к каждому из них некоторый алгоритм решения соответствующего уравнения или системы. При этом неявные многошаговые методы представляют собой, как правило (мы других типов не рассматривали), одно уравнение, если исходная задача – это задача Коши для одного уравнения, и систему уравнений, если исходная задача – это задача Коши для системы уравнений. В то же время неявные одношаговые методы (типа Рунге-Кутта) практически всегда представляют собой системы уравнений. Способы решения таких задач мы рассматривали ранее. Формально любой из них может быть применен и в нашем случае. Однако не все получающиеся при этом алгоритмы будут одинаково успешными. Поясним ситуацию (а заодно и продемонстрируем, как выглядят соответствующие алгоритмы) на примере неявного метода Эйлера, применяемого к решению одного нелинейного уравнения. Как мы помним, формула его имеет вид

$$y_{j+1} = y_j + \tau f\left(t_{j+1}, y_{j+1}\right). \tag{2.13}$$

а) Метод итерации.

Простейшим из методов, применяемых для решения нелинейных уравнений, является метод итераций. Технически в нашем случае его применять особенно удобно, ибо формула (2.13) фактически дает канонический вид нелинейного уравнения, готовый к применению метода итераций. Следовательно, для нахождения приближенного значения y_{j+1} можно записать итерационный процесс

$$\overset{k+1}{y}_{j+1} = y_j + \tau f\left(t_{j+1}, \overset{k}{y}_{j+1}\right),\ k = 0,1,\ldots. \tag{2.14}$$

В качестве начального приближения $\overset{0}{y}_{j+1}$ в простейшем варианте может быть взято значение решения в предыдущем узле сетки, т.е. $\overset{0}{y}_{j+1} = y_j$. Иногда для предсказания начального приближения используют некоторый явный метод решения задачи Коши (например, в нашем случае – явный метод Эйлера, вполне согласованный с (2.13) по порядку, т.е. $\overset{0}{y}_{j+1} = y_j + \tau f\left(t_j, y_j\right)$).

Итерации продолжают до достижения сходимости (т.е., например, до выполнения неравенства $\left|\overset{k+1}{y}_{j+1} - \overset{k}{y}_{j+1}\right| \le \varepsilon$, где ε – заданная величина погрешности, которая должна быть, очевидно, согласована с пользовательским требованием к точности нахождения решения исходной задачи Коши (по крайней мере, быть не больше последней)). В качестве

критерия для остановки итерационного процесса можно использовать и соответствующий аналог относительной погрешности.

Вспомним, однако, что метод итерации имеет ограничения на область сходимости. В частности, достаточное условие сходимости может иметь вид $\left|\frac{\partial\varphi(y_{j+1})}{\partial y}\right|<1$, где $\varphi(y)$ – функция, стоящая в правой части канонического представления. В нашем случае это ограничение примет вид $\tau\left|\frac{\partial f(t_{j+1},y_{j+1})}{\partial y}\right|<1$ и дает ограничение на допустимую величину шага сетки, аналогичную соответствующему ограничению, накладываемому явным методом Эйлера.

Таким образом, метод итераций уничтожает одно из главных достоинств неявных методов – их возможную А-устойчивость, – и поэтому, несмотря на простоту, не может быть рекомендован к широкому использованию (по крайней мере, при решении жестких задач).

б) Метод Ньютона.

Переписав (2.13) в виде $g(y_{j+1})=0$, где $g(y_{j+1})=y_{j+1}-y_j-\tau f(t_{j+1},y_{j+1})$, можем записать для нахождения неизвестной величины y_{j+1} итерационный метод Ньютона:

$$\overset{k+1}{y}_{j+1}=\overset{k}{y}_j-\frac{\overset{k}{y}_{j+1}-y_j-\tau f\left(t_{j+1},\overset{k}{y}_{j+1}\right)}{1-\tau\dfrac{\partial f\left(t_{j+1},\overset{k}{y}_{j+1}\right)}{\partial y}},\ k=0,1,\ldots. \qquad (2.15)$$

Выбор начального приближения может быть осуществлен аналогично рассмотренному выше для метода итераций. При этом, однако, необходимо иметь в виду, что в случае жестких задач использование явных алгоритмов для предсказания начального приближения следует осуществлять с известной долей осторожности. Контроль сходимости также осуществляется аналогично.

Напомним, что сходимость метода Ньютона практически регламентируется только качеством выбора начального приближения и поэтому это – один из лучших способов реализации неявных численных методов.

Упражнение. Записать алгоритм реализации методов (2.4) и (2.9) с помощью методов: а) Ньютона; б) секущих в случае решения задачи Коши для а) одного уравнения; б) системы уравнений.

Замечание. Ограничившись одной итерацией метода Ньютона (2.15) при выборе $\overset{0}{y}_{j+1}=y_j$, получим простейший пример ***нелинейного*** явного численного метода решения задачи Коши:

$$y_{j+1}=y_j+\frac{\tau f(t_{j+1},y_j)}{1-\tau\dfrac{\partial f(t_{j+1},y_j)}{\partial y}}. \qquad (2.16)$$

Несложно проверить, что в этом случае получается ***явный А-устойчивый*** метод.

РАЗДЕЛ VI

Методы решения граничных задач для обыкновенных дифференциальных уравнений

Граничные или многоточечные задачи представляют собой более сложный тип задач по сравнению с задачами Коши. Здесь и сам характер постановки задач более общий. Задаются, как правило, не значения искомой функции или ее производных, а лишь связи между этими значениями. При решении подобных задач более сложен как сам процесс поиска решения, так и вопросы существования и единственности решения.

Будем считать, что на отрезке $[a;b]$ задано обыкновенное дифференциальное уравнение n-го порядка (можно рассматривать систему из n уравнений)

$$u^{(n)} = f\left(x, u, u', \ldots, u^{(n-1)}\right),$$

выбраны k различных точек $x_1 < x_2 < \cdots < x_k$ и заданы некоторые n связей (условий)

$$v_j\left[u(x_1), u'(x_1), \ldots, u^{(n-1)}(x_1), \ldots, u(x_k), u'(x_k), \ldots, u^{(n-1)}(x_k)\right] = 0, \;\; j = 1, 2, ..., n\,.$$

Тогда говорят, что для нашего уравнения поставлена многоточечная (k-точечная) задача.

В частном случае, когда $k = 1$, а функции v_j имеют тривиальный вид, мы получаем задачу Коши.

Сейчас (и, как правило, далее) рассмотрим случай $k = 2$ и $x_1 = a,\; x_2 = b$. В этом случае задача называется ***граничной*** (или краевой).

Если исходное дифференциальное уравнение или хотя бы одно из дополнительных условий нелинейны, то мы имеем нелинейную граничную задачу. В противном случае задача называется линейной. В достаточно общем виде линейную граничную задачу можно записать следующим образом:

$$\begin{cases} Lu \equiv u^{(n)}(x) + p_1(x)u^{(n-1)}(x) + \ldots + p_{n-1}(x)u'(x) + p_n(x)u(x) = f(x), & (1) \\ l_j(u) \equiv \sum\limits_{i=0}^{n-1}\left[\alpha_{ij}u^{(i)}(a) + \beta_{ij}u^{(i)}(b)\right] = A_j\,, \quad j = 1, 2, ..., n. & (2) \end{cases}$$

ГЛАВА XIII

Методы, основанные на сведении к решению задач Коши

§ 1. Метод редукции граничных задач к задачам Коши

Будем иметь в виду линейную граничную задачу (1), (2). Сейчас нас будет интересовать процедура сведения решения такой задачи к решению задач с начальными условиями. Предлагаемый ниже алгоритм существенно использует общий вид решения дифференциального уравнения (1).

Как известно из теории дифференциальных уравнений, общее решение линейного уравнения (1) может быть записано в виде

$$u(x) = u_0(x) + \sum_{i=1}^{n} C_i u_i(x)\,, \qquad (1.1)$$

где $u_0(x)$ – некоторое частное решение неоднородного уравнения (1), т.е.

$$Lu_0(x) = f(x),$$

а $u_i(x),\ i = \overline{1,n}$ – решения соответствующего однородного уравнения, образующие линейно независимую систему:

$$Lu_i(x) = 0,\ i = \overline{1,n}.$$

C_i – произвольные постоянные, конкретные значения которых определяются дополнительными условиями (2).

Если бы нам были известны функции $u_0(x)$ и $u_i(x),\ i = \overline{1,n}$, то те значения произвольных постоянных C_i, которые соответствуют искомому решению, были бы легко найдены из условий (2), а именно:

$$l_j(u_0) + \sum_{i=1}^{n} C_i \cdot l_j(u_i) = A_j,\ j = 1,2,...,n, \tag{1.2}$$

и, таким образом, имеем для определения названных констант систему из n линейных алгебраических уравнений с n неизвестными.

Если определитель матрицы системы (1.2) отличен от нуля, то она будет иметь единственное решение и, следовательно, мы получим единственный набор C_i. В противном же случае будем иметь либо бесконечное множество решений, либо неразрешимость в зависимости от ранга расширенной матрицы граничных условий.

Весь вопрос решения граничной задачи, таким образом, сводится к тому, как найти частное решение неоднородного уравнения $u_0(x)$ и систему $u_i(x)$. Поскольку речь идет о численном решении соответствующей задачи и мы на данный момент для дифференциальных уравнений знакомы лишь с методами решения задач Коши, то естественно в качестве искомых функций взять решения некоторых задач Коши.

Учитывая, что $u_0(x)$ – ***произвольное*** частное решение неоднородного уравнения, задачу Коши для его определения можно взять, например, в виде (использовав ***простейшие*** начальные условия)

$$\begin{cases} Lu_0(x) = f(x), \\ u_0(a) = 0, \\ \dots \\ u_0^{(n-1)}(a) = 0. \end{cases} \tag{1.3}$$

Аналогичным образом, поскольку единственным ограничением для системы $u_i(x)$ является ее линейная независимость, то соответствующий набор начальных условий для их определения должен быть подчинен условию: определитель Вронского искомой системы в точке a должен быть отличен от нуля. Поэтому простейший тип таких начальных условий приводит к задачам вида

$$\begin{cases} Lu_i(x) = 0, \\ u_i^{(j)}(a) = \delta_i^j,\ j = \overline{0,n-1};\ i = \overline{1,n}. \end{cases} \tag{1.4}$$

В этом случае определитель Вронского есть определитель единичной матрицы, что обеспечивает выполнение сформулированных выше требований.

Таким образом, мы свели решение граничной задачи к решению $(n+1)$ задач Коши (1.3), (1.4) и системы линейных алгебраических уравнений (1.2) для определения постоянных интегрирования C_i.

Очевидно, что граничные условия могут быть и нелинейными, что, в конечном итоге, приведет к необходимости решать вместо линейной системы (1.2) соответствующую систему нелинейных уравнений.

***Замечание* 1.** Все указанные выше задачи Коши могут быть решены численно с использованием изучавшихся нами ранее методов. При этом сетки, на которых отыскиваются решения задач (1.3), (1.4), должны иметь непустое пересечение узловых точек (в оптимуме – совпадать), ибо только на этом пересечении и может быть построена соответствующая линейная комбинация (1.1). Это накладывает некоторые ограничения на процедуры поиска решений задач Коши с апостериорной оценкой погрешности.

***Замечание* 2.** Описанный выше алгоритм может быть применен к поиску решения не только задачи (1), (2), но и любой другой, общее решение которой имеет вид (1.1) (например, граничной задачи для системы линейных обыкновенных дифференциальных уравнений).

§ 2. Метод стрельбы (пристрелки) для линейных граничных задач

Заметим, что в изложенном выше алгоритме метода редукции фигурируют слова «произвольный» по отношению к построению частного решения неоднородного уравнения и линейно независимой системы решений однородного уравнения. Если отказаться от этой произвольности в пользу того, чтобы некоторые комбинации найденных частных решений удовлетворяли части граничных условий, то это может позволить сократить общее количество решаемых задач Коши.

Изучим вначале применение этой идеи на частном примере задачи

$$\begin{cases} Lu(x) \equiv u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \;\; a \le x \le b, \\ u(a) = A, \;\; u(b) = B. \end{cases} \tag{2.1}$$

Учитывая простоту граничных условий, можно предложить следующее: рассмотрим частное решение $u_0(x)$ неоднородного уравнения, которое удовлетворяет левому граничному условию. Значение $u_0'(a)$ при этом может задаваться произвольно. Таким образом, функция $u_0(x)$ находится как решение задачи

$$\begin{cases} Lu_0(x) = f(x), \\ u_0(a) = A, \\ u_0'(a) = \eta_0, \;\; \eta_0 - \text{любое}. \end{cases} \tag{2.2}$$

Рассмотрим также частное решение однородного уравнения $u_1(x)$, удовлетворяющее условиям

$$\begin{cases} Lu_1(x) = 0, \\ u_1(a) = 0, \\ u_1'(a) = \eta_1, \;\; \eta_1 \ne 0 - \text{любое}. \end{cases} \tag{2.3}$$

Тогда составленная из функций $u_0(x)$ и $u_1(x)$ линейная комбинация

$$u(x) = u_0(x) + Cu_1(x) \tag{2.4}$$

заведомо удовлетворяет дифференциальному уравнению (2.1) и левому граничному условию при любом значении произвольной постоянной C. Выбором же последней следует распорядиться таким образом, чтобы удовлетворить второму граничному условию. Исходя из этого, получится уравнение

$$u_0(b) + Cu_1(b) = B\,,$$

которое в случае $u_1(b) \neq 0$ имеет решение

$$C = \frac{B - u_0(b)}{u_1(b)}$$

и, таким образом, решение задачи (2.1) может быть вычислено по формуле (2.4).

По сравнению с классическим вариантом метода редукции здесь необходимо решать всего две задачи Коши.

Перейдем теперь к рассмотрению общего случая краевой задачи для системы линейных обыкновенных дифференциальных уравнений первого порядка. Итак, пусть имеем краевую задачу

$$\begin{cases} u'(x) = A(x)u(x) + f(x), \;\; a < x < b \\ Bu(a) = c, \\ Du(b) = d, \end{cases} \tag{2.5}$$

где u, f, c, d – векторы размерностей соответственно n, n, $n-r$, r, а A, B, D – матрицы размерностей $n \times n$, $(n-r) \times n$, $r \times n$. В дальнейшем будем предполагать, что ранг матрицы B равен $n-r$, а ранг матрицы D равен r. Рассмотрим линейную систему алгебраических уравнений

$$Bu = c\,, \tag{2.6}$$

задающую граничное условие на левом конце отрезка интегрирования. Так как по предположению ранг матрицы B равен $n-r$, то общее решение системы (2.6) может быть записано в виде

$$u_0 + \sum_{i=1}^{r} C_i u_i\,,$$

где u_0 – произвольное решение неоднородной системы (2.6), а $u_1, u_2, ..., u_r$ – произвольная система из r линейно независимых решений однородной системы $Bu = 0$. Пусть $u_1, u_2, ..., u_r, u_0$ – какой-либо набор таких векторов. Тогда при помощи численного интегрирования найдем частное решение неоднородной системы

$$\begin{cases} u_0'(x) = A(x)u_0(x) + f(x), \\ u_0(a) = u_0 \end{cases} \tag{2.7}$$

и решения однородных систем

$$\begin{cases} u_j'(x) = A(x)u_j(x), \\ u_j(a) = u_j, \qquad\qquad j = \overline{1,r}. \end{cases} \tag{2.8}$$

Теперь заметим, что всякая функция вида

$$u(x) = u_0(x) + \sum_{j=1}^{r} C_j u_j(x) \tag{2.9}$$

при любых значениях произвольных постоянных C_j удовлетворяет исходной системе (2.5) и левому граничному условию. Таким образом, остается определить эти постоянные, удовлетворив правому граничному условию:

$$D\left(u_0(b) + \sum_{j=1}^{r} C_j u_j(b) \right) = d\,. \tag{2.10}$$

(2.10) представляет собой систему из r линейных алгебраических уравнений с r неизвестными. Матрица D этой системы невырождена (в соответствии с предположением ее ранг равен r), поэтому задача имеет единственное решение.

Окончательно алгоритм решения граничной задачи (2.5) может выглядеть следующим образом:

1. Находим частное решение линейной неоднородной системы алгебраических уравнений (2.6), соответствующей левому граничному условию, и r линейно независимых частных решений соответствующей (2.6) линейной однородной системы;
2. Решаем задачи Коши (2.7), (2.8) (в количестве $(r+1)$ штук), начальными условиями которых являются найденные на первом этапе частные решения системы (2.6) и соответствующей ей однородной;
3. Решаем систему линейных алгебраических уравнений (2.10) относительно произвольных постоянных C_j;
4. По формуле (2.9) определяем решение исходной граничной задачи.

§ 3. Метод дифференциальной прогонки

Другим алгоритмом, применяемым для решения линейных граничных задач, является метод дифференциальной прогонки. Технику применения одной из простейших разновидностей этого алгоритма рассмотрим на примере задачи

$$\begin{cases} Lu(x) \equiv u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \;\; a \le x \le b, \\ \alpha_0 u(a) + \alpha_1 u'(a) = A, \\ \beta_0 u(b) + \beta_1 u'(b) = B. \end{cases} \tag{3.1}$$

Существенным моментом здесь (как, впрочем, и в задачах, решавшихся в предыдущем параграфе) является то, что граничные условия разделены по концам.

Как мы уже отмечали ранее, общее решение уравнения (3.1) имеет вид

$$u(x) = u_0(x) + C_1 u_1(x) + C_2 u_2(x)\,.$$

Выделим теперь то подмножество решений, которое удовлетворяет одному из граничных условий (для определенности – левому). В итоге получим однопараметриче-

ское семейство решений, которое можно рассматривать как общее решение некоторого линейного дифференциального уравнения первого порядка

$$u'(x)+P(x)u(x)=F(x). \tag{3.2}$$

Мы не знаем коэффициентов этого уравнения, но и $P(x)$, и $F(x)$ можно найти. Для этого следует учесть, что любое решение уравнения (3.2) есть решение исходного уравнения второго порядка (3.1) и, кроме того, удовлетворяет левому граничному условию из (3.1). Удовлетворим сначала первому требованию. Так как из (3.2) следует, что

$$u'(x)=F(x)-P(x)u(x),$$

то

$$u''(x)=F'(x)-P'(x)u(x)-P(x)u'(x)=F'(x)-P'(x)u(x)-P(x)[F(x)-P(x)u(x)].$$

Подставив эти выражения в исходное уравнение (3.1), будем иметь:

$$F'(x)-P'(x)u(x)-P(x)[F(x)-P(x)u(x)]+p(x)[F(x)-P(x)u(x)]+q(x)u(x)=f(x).$$

Поскольку мы хотим выполнения этого равенства при любых $u(x)$, то, приравнивая коэффициенты при $u(x)$ в левой и правой части последнего равенства, а также – отдельно – свободные члены, получим:

$$\begin{cases} P'(x)+[p(x)-P(x)]P(x)=q(x), \\ F'(x)+[p(x)-P(x)]F(x)=f(x). \end{cases} \tag{3.3}$$

Таким образом, для определения коэффициентов уравнения (3.2) получим систему обыкновенных дифференциальных уравнений первого порядка (3.3).

Чтобы выделить конкретные $P(x)$ и $F(x)$, удовлетворим левому граничному условию из (3.1): для любой функции $u(x)$ должно выполняться равенство

$$\alpha_0 u(a)+\alpha_1[F(a)-P(a)u(a)]=A.$$

Как и выше, приравнивая коэффициенты при $u(a)$ и свободные члены, найдем:

$$\begin{cases} P(a)=\dfrac{\alpha_0}{\alpha_1}, \\ F(a)=\dfrac{A}{\alpha_1}. \end{cases} \tag{3.4}$$

Отсюда следует, что при $\alpha_1 \neq 0$ (это – одно из условий применимости рассматриваемого варианта метода прогонки) мы для определения функций $P(x)$ и $F(x)$ получаем задачу Коши (3.3), (3.4). Решив данную задачу, мы построим уравнение (3.2). Теперь можно будет потребовать, чтобы его решение удовлетворяло второму из граничных условий задачи (3.1). На основании этого требования получим систему линейных алгебраических уравнений

$$\begin{cases} \beta_0 u(b)+\beta_1 u'(b)=B, \\ P(b)u(b)+u'(b)=F(b). \end{cases} \tag{3.5}$$

Если $\Delta = \begin{vmatrix} \beta_0 & \beta_1 \\ P(b) & 1 \end{vmatrix} \neq 0$, то мы отсюда единственным образом найдем значения $u(b)$ и $u'(b)$:

$$\begin{cases} u(b) = \dfrac{\begin{vmatrix} B & \beta_1 \\ F(b) & 1 \end{vmatrix}}{\Delta}, \\ u'(b) = \dfrac{\begin{vmatrix} \beta_0 & B \\ P(b) & F(b) \end{vmatrix}}{\Delta}. \end{cases} \tag{3.6}$$

После этого, решая уравнение (3.2) с $u(b)$ из (3.6) в качестве начального (точнее, конечного) условия, найдем решение задачи (3.1) (заметим, что последняя задача Коши решается в противоположном направлении (от правого конца отрезка к левому), что предполагает использование в формулах численных методов, применяемых для этих целей, отрицательного значения шага).

Замечание* 1.** Описанный вариант метода дифференциальной прогонки носит название метода ***левой прогонки, поскольку мы строили уравнение (3.2) удовлетворяющим левому граничному условию. Если же $\alpha_1 = 0$, то нам не удастся воспользоваться формулами (3.4) для вычисления начальных условий для функций $P(x)$ и $F(x)$. В этом случае можно (если $\beta_1 \neq 0$) построить формулы типа (3.4), получить их из соображений удовлетворения правому граничному условию. Таким образом, внося необходимые коррективы в дальнейшие рассуждения, придем к методу ***правой прогонки***. Если же $\alpha_1 = \beta_1 = 0$, то метод дифференциальной прогонки для данной задачи не применим.

***Замечание* 2.** На аналогичной идеологии могут быть построены варианты метода дифференциальной прогонки и для решения линейных граничных задач более общего вида (например, типа (2.5) из предыдущего параграфа).

§ 4. Метод стрельбы для нелинейных граничных задач

Вначале рассмотрим нелинейную граничную задачу для системы обыкновенных дифференциальных уравнений первого порядка:

$$\begin{cases} u'(x) = f(x, u(x)), \\ B(u(a)) = 0, \\ D(u(b)) = 0. \end{cases} \tag{4.1}$$

Здесь $u = (u_1, u_2, ..., u_n)^T$, $B = (b_1, b_2, ..., b_{n-r})^T$, $D = (d_1, d_2, ..., d_r)^T$, причем $u(x)$ – искомая вектор-функция, а B и D – заданные нелинейные вектор-функции указанного количества аргументов. Основой для создания алгоритма решения задачи (4.1) путем сведения к решению задачи Коши может служить очень простой факт: решение задачи зависит от начальных данных. Поэтому, если добавить, например, недостающие на левом конце отрезка интегрирования r уравнений связи вида $g_i(u(a)) = \eta_i$, $i = \overline{1, r}$, (здесь $g_i(u(a))$ –заданные функции векторного аргумента, а η_i – заданные числовые параметры) таким образом, чтобы система нелинейных уравнений

$$\begin{cases} b_i(u(a)) = 0, \ i = \overline{1, n-r}, \\ g_j(u(a)) = \eta_j, \ j = \overline{1, r} \end{cases} \tag{4.2}$$

позволяла однозначно определить вектор начальных данных $u(a)$ как функцию параметров η_j: $u(a)=\omega(\eta),\ \eta=(\eta_1,...,\eta_r)$, то решение системы уравнений (4.1) с найденными начальными условиями также будет функцией вектора параметров η_i: $u(x)=u(x,\eta)$. Эта функция, очевидно, удовлетворяет левым граничным условиям. Поэтому вектор параметров η может быть определен путем подстановки в правое граничное условие:

$$\psi(\eta)\overset{def}{=}D(u(b,\eta))=0\,. \tag{4.3}$$

Таким образом, формально задача свелась к решению нелинейной системы уравнений относительно вектора параметров.

Основная проблема, возникающая при решении системы (4.3), состоит в том, что мы не знаем аналитического вида функциональной зависимости $\psi(\eta)$, но в то же время имеем техническую возможность при фиксированном значении вектора параметров η вычислять значение функции ψ. Для этого необходимо решить систему (4.2) (этим самым мы задаем начальные условия для исходной системы (4.1)), затем с найденными начальными условиями решить (численно) задачу Коши и найденное решение подставить в правое граничное условие. Это и будет искомое значение функции ψ. Учитывая сказанное, в качестве метода для решения системы (4.3) следует выбирать, с одной стороны, достаточно быстро сходящийся, а с другой – тот, алгоритм которого «в состоянии» обходиться ьез вычисления производных от функции ψ (последняя задача не относится к разряду неразрешимых, но является очень трудоемкой).

Опишем алгоритм (метода стрельбы) более подробно на примере системы вида (4.1) при $n=2,\ r=1$, т.е.

$$\begin{cases} u'(x)=f(x,u(x),v(x)),\\ v'(x)=g(x,u(x),v(x)),\\ \varphi(u(a),v(a))=0,\\ \psi(u(b),v(b))=0 \end{cases} \tag{4.4}$$

В соответствии с изложенной выше общей схемой выберем произвольно значение $u(a)=\eta$, рассмотрим левое граничное условие как алгебраическое уравнение

$$\varphi(\eta,v(a))=0$$

и определим удовлетворяющее ему значение $v(a)=\xi(\eta)$. Возьмем значения $u(a)=\eta$ и $v(a)=\xi(\eta)$ в качестве начальных условий задачи Коши для системы (4.4) и проинтегрируем полученную задачу любым подходящим численным методом. При этом получим решение $u(x;\eta),\ v(x;\eta)$, зависящее от η как от параметра.

Значение $\xi(\eta)$ выбрано так, что найденное решение задачи Коши удовлетворяет левому граничному условию задачи (4.4). Однако второму граничному условию это решение, вообще говоря, не удовлетворяет: при его подстановке левая часть правого граничного условия, рассматриваемая как функция параметра η

$$\overline{\psi}(\eta)=\psi(u(b;\eta),v(b,\eta))\,, \tag{4.5}$$

не обратится в нуль. Таким образом, необходимо каким-либо образом менять числовые значения параметра η, пока не подберем такое значение, для которого $\overline{\psi}(\eta)\approx 0$ с требуе-

мой точностью, т.е., как и было отмечено в общей схеме, решение граничной задачи (4.4) в конечном итоге сводится к нахождению корня алгебраического уравнения

$$\overline{\psi}(\eta) = 0 . \tag{4.6}$$

Простейшим из методов его решения, которые целесообразно применять в данном случае, является метод дихотомии. При его реализации делают пробные «выстрелы» – расчеты с наугад (если нет каких-либо специальных соображений) выбранными значениями η_i до тех пор, пока среди величин $\overline{\psi}(\eta_i)$ не окажутся разные по знаку. Пара таких значений η_i, η_{i+1} образует «вилку». С математической точки зрения мы таким образом отделим некоторый корень уравнения (4.6)). Далее, последовательно деля отрезок $[\eta_i;\eta_{i+1}]$ пополам до получения нужной точности, производим «пристрелку» (уточнение) параметра η. Благодаря этому процессу весь метод получил название метода стрельбы.

Однако нахождение каждого нового значения функции $\overline{\psi}(\eta)$ требует численного интегрирования системы (4.4), т.е. достаточно трудоемко. Поэтому корень уравнения (4.6) желательно находить с помощью метода, обладающего более быстрой сходимостью, нежели дихотомия. Часто таковым является разностный аналог метода Ньютона – метод секущих. В этом случае первые два расчета делают с наудачу выбранными значениями η_0, η_1, а следующие вычисляют по формуле

$$\eta_{i+1} = \eta_i - \frac{(\eta_i - \eta_{i-1})\overline{\psi}(\eta_i)}{\overline{\psi}(\eta_i) - \overline{\psi}(\eta_{i-1})},\ i = 1,2,... \tag{4.7}$$

Сходимость итерационного процесса (4.7), очевидно, зависит от выбора η_0 и η_1.

Замечание. Очевидно, при решении общей задачи (4.1) приведенные здесь соображения становятся значительно сложнее с точки зрения технической их реализации.

ГЛАВА XIV

Проекционные методы решения граничных задач

С идеологией данной группы методов решения операторных уравнений мы познакомились при изучении методов решения интегральных уравнений. Напомним, что эти методы называют проекционными по той причине, что первоначальное пространство, в котором поставлена исходная задача, проектируют в некоторое подпространство более простой структуры, где и отыскивают приближение. Часто элементами подпространства являются известные координатные функции, и остается только подобрать коэффициенты линейной комбинации. Приближенное решение здесь получается в аналитическом виде.

§ 1. Вариационные методы решения граничных задач

1.1. Вариационная задача для операторных уравнений

Пусть H – некоторое вещественное гильбертово пространство и A – линейный оператор, определенный на множестве H_A, всюду плотном в H.

Рассмотрим операторное уравнение

$$Au = f , \tag{1.1}$$

где $f \in H$ – заданный элемент, а $u \in H_A$ – искомый.

Оператор A будем предполагать положительным и самосопряженным. Тогда уравнение (1.1) не может иметь более одного решения.

Действительно, пусть имеются два решения $u_1 \neq u_2$. Вычитая равенства $Au_1 = f$ и $Au_2 = f$, получим: $A(u_1 - u_2) = 0$ и, следовательно, $(A(u_1 - u_2), u_1 - u_2) = 0$. Поскольку по условию $A > 0$, то отсюда с необходимостью следует, что $u_1 - u_2 = 0$, т.е. $u_1 = u_2$, что противоречит предположению.

Таким образом, если задача (1.1) при сделанных относительно оператора A предположениях имеет решение, то это решение будет единственным.

***Теорема* 1.** Если уравнение (1.1) имеет некоторое решение u_1, то это решение доставляет минимум функционалу

$$J(u) = (Au, u) - 2(f, u). \tag{1.2}$$

Доказательство.

Возьмем произвольный элемент $v \in H_A$ и положим $v = u_1 + t$. Тогда

$$J(v) = (Av, v) - 2(f, v) = (A(u_1 + t), u_1 + t) - 2(f, u_1 + t) =$$

$$= J(u_1) + 2(Au_1 - f, t) + (At, t) = J(u_1) + (At, t) \geq J(u_1).$$

Справедливо и обратное утверждение. ⊠

***Теорема* 2.** Если найдется такой элемент $u_1 \in H_A$, который доставляет минимум функционалу $J(u)$, то этот элемент будет решением уравнения (1.1).

Доказательство.

Возьмем произвольный элемент $v \in H_A$ и произвольное вещественное число λ. Тогда $u_1 + \lambda v \in H_A$ и по условию $J(u_1 + \lambda v) \geq J(u_1)$. С другой стороны

$$J(u_1 + \lambda v) = (A(u_1 + \lambda v), u_1 + \lambda v) - 2(f, u_1 + \lambda v) = J(u_1) + 2\lambda(Au_1 - f, v) + \lambda^2 (Av, v).$$

Поэтому для всех λ выполняется неравенство

$$\lambda^2 (Av, v) + 2\lambda(Au_1 - f, v) \geq 0.$$

Последнее же неравенство справедливо для любых действительных λ только в том случае, когда $(Au_1 - f, v) = 0$.

Так как H_A всюду плотно в H, то отсюда следует, что $Au_1 - f = 0$. ⊠

1.2. Граничная задача для обыкновенного дифференциального уравнения второго порядка как операторное уравнение

Вернемся сейчас к граничной задаче для обыкновенного дифференциального уравнения второго порядка (с краевыми условиями первого либо третьего рода) и выяс-

ним, какой вид она должна иметь для того, чтобы непосредственно из общей теории следовала ее равносильность некоторой вариационной задаче типа (1.2).

Как уже отмечалось в общей теории, оператор A должен быть самосопряженным и положительным.

Пусть исходное дифференциальное уравнение имеет вид

$$Lu(x) \equiv p_0(x)u''(x) + p_1(x)u'(x) + p_2(x)u(x) = f(x),\ x \in [a;b].$$

Тогда условие самосопряженности оператора L имеет вид

$$(Lu, v) = (u, Lv)$$

или

$$\int_a^b [(p_0u'' + p_1u' + p_2u)v - (p_0v'' + p_1v' + p_2v)u]dx \equiv 0,$$

причем это соотношение выполняется для любых допустимых функций u и v.

Выполняя в слагаемых, содержащих вторые производные, интегрирование по частям, имеем:

$$(Lu, v) - (u, Lv) = p_0(u'v - uv')\Big|_a^b + \int_a^b \Big[-u'(p_0v)' + v'(p_0u)' + p_1(u'v - uv')\Big]dx =$$

$$= p_0(u'v - uv')\Big|_a^b + \int_a^b (p_0' - p_1)(uv' - u'v)dx = 0.$$

Отсюда, учитывая, что $p_0(x) \neq 0$, а также то, что $u(x)$ и $v(x)$ – произвольные допустимые функции, получим:

$$\begin{cases} (u'v - uv')(b) = (u'v - uv')(a) = 0, & (1.3) \\ p_0'(x) - p_1(x) = 0. & (1.4) \end{cases}$$

Рассмотрим теперь по отдельности случаи граничных условий первого и третьего рода (условия второго рода – частный случай условий третьего рода).

Итак, пусть вначале граничные условия имеют вид

$$\begin{cases} u(a) = A, \\ u(b) = B. \end{cases} \qquad (*)$$

Тогда из условия (1.3) следует, что допустимые функции должны обращаться на границе в нуль, т.е. $A = B = 0$. Кроме того, условие (1.4) означает, что

$$p_0(x)u''(x) + p_1(x)u'(x) = (p_0(x)u'(x))'.$$

Таким образом, первая краевая задача, исходя из условия самосопряженности дифференциального оператора, будет иметь вид

$$\begin{cases} Lu(x) \equiv (p_0(x)u'(x))' + p_2(x)u(x) = f(x), \ a \le x \le b, \\ u(a) = u(b) = 0. \end{cases}$$

Выясним теперь, какие условия на коэффициенты $p_0(x)$ и $p_2(x)$ накладывает требование положительности оператора L. Поскольку для всех отличных от тождественного нуля функций $u(x)$ должно выполняться неравенство $(Lu,u) > 0$, то имеем:

$$(Lu,u) = \int_a^b \left[(p_0u')' + p_2u\right]u dx = p_0u' \cdot u\Big|_a^b + \int_a^b \left[-p_0 \cdot (u')^2 + p_2u^2\right]dx =$$

$$= \int_a^b \left[-p_0(x) \cdot (u'(x))^2 + p_2(x)u(x)^2\right]dx > 0.$$

Отсюда следует, что (поскольку $u(x)$ – произвольная допустимая функция)

$$\begin{cases} p_0(x) \le c_0 < 0, \\ \mathrm{p}_2(x) \ge 0 \end{cases} \quad \text{для всех } x \in [a;b].$$

Учитывая полученные результаты, в дальнейшем первую краевую задачу с самосопряженным положительным оператором будем записывать в виде

$$\begin{cases} Lu(x) \equiv -(p(x)u'(x))' + q(x)u(x) = -f(x), \ a \le x \le b, \\ u(a) = u(b) = 0. \end{cases} \tag{1.5}$$

где $p(x) \ge p_0 > 0$, $p(x) \in C^1[a;b]$ и $q(x) \in C[a;b]$, $q(x) \ge 0$, $f(x) \in C[a;b]$.

Заметим, что в этом случае функционал (1.2) примет вид

$$J(u) = (Lu,u) - 2(f,u) = \int_a^b \left[-(p(x)u'(x))' + q(x)u(x) + 2f(x)\right]u(x)dx =$$

$$= \int_a^b \left[p(x)(u'(x))^2 + q(x)u^2(x) + 2f(x)u(x)\right]dx, \tag{1.6}$$

т.е. задача (1.5) равносильна задаче минимизации функционала (1.6).

Случай неоднородных граничных условий (*) может быть сведен к рассмотренному. Действительно, зафиксируем некоторую функцию $v(x)$, удовлетворяющую условиям (*), и представим решение исходной задачи с неоднородными условиями в виде $u(x) = u_1(x) + v(x)$. Тогда функция $u_1(x)$ удовлетворяет нулевым граничным условиям и уравнению

$$Lu_1(x) = Lu(x) - Lv(x) = -f(x) - Lv(x) = -f(x) + (p(x)v'(x))' - q(x)v(x).$$

Решение этой задачи равносильно минимизации функционала (1.6)

$$J(u_1)=\int_a^b\left[p(x)(u_1'(x))^2+q(x)u_1^2(x)+2\Big(f(x)-(p(x)v'(x))'+q(x)v(x)\Big)u_1(x)\right]dx=$$

$$=\int_a^b\left[p(u_1')^2+qu_1^2+2fu_1-2(pv')'u_1+2qvu_1\right]dx=\begin{bmatrix}\text{интегрируем слагаемое}\\(pv')'u_1\ \text{ по частям}\end{bmatrix}=$$

$$=-2pv'u_1\Big|_a^b+\int_a^b\left[p(u_1')^2+qu_1^2+2fu_1+2pv'u_1'+2qvu_1\right]dx=$$

$$=\int_a^b\left[p(u_1'+v')^2+q(u_1+v)^2+2f(u_1+v)\right]dx-\int_a^b\left[p(v')^2+qv^2+2fv\right]dx=J(u)-J(v).$$

Так как $v(x)$ – фиксированная функция, то минимизация функционала $J(u_1)$ равносильна минимизации функционала $J(u)$

Рассмотрим теперь граничные условия третьего рода

$$\begin{cases}\alpha_0u(a)+\alpha_1u'(a)=A,\\ \beta_0u(b)+\beta_1u'(b)=B\end{cases}\tag{1.7}$$

и выясним, когда выполняется условие (1.3). Так как из (1.7) следует, что ($\alpha_1\neq 0,\ \beta_1\neq 0$)

$$u'(a)=\frac{A-\alpha_0u(a)}{\alpha_1},\ \ u'(b)=\frac{B-\beta_0u(b)}{\beta_1}\tag{1.8}$$

(заметим, что точно таким же граничным условиям удовлетворяет и функция $v(x)$), то (1.3) преобразуется к виду

$$\frac{B-\beta_0u(b)}{\beta_1}v(b)-u(b)\frac{B-\beta_0v(b)}{\beta_1}=\frac{A-\alpha_0u(a)}{\alpha_1}v(a)-u(a)\frac{A-\alpha_0v(a)}{\alpha_1}\Rightarrow$$

$$\Rightarrow\ \frac{B}{\beta_1}(v(b)-u(b))=\frac{A}{\alpha_1}(v(a)-u(a))\equiv 0,$$

откуда, учитывая произвольность допустимых функций, непосредственно получаем: $A=B=0$, т.е. граничные условия (1.7) должны быть однородного типа (с нулевыми правыми частями). При этом (1.8) перепишутся в виде

$$u'(a)=-\frac{\alpha_0}{\alpha_1}u(a),\ \ u'(b)=-\frac{\beta_0}{\beta_1}u(b).\tag{1.9}$$

Рассматривая дифференциальный оператор в виде (1.5), выясним условия положительной определенности (помимо $p(x)\geq p_0>0$ и $q(x)\geq 0$):

$$(Lu,u)=\int_a^b\Big[-(p(x)u'(x))'+q(x)u(x)\Big]u(x)dx=-p(x)u'(x)u(x)\Big|_a^b+\int_a^b\left[p(u')^2+qu^2\right]dx.$$

Второе слагаемое при отмеченных условиях положительно, а первое с учетом (1.9) примет вид

$$p(b)\cdot\frac{\beta_0}{\beta_1}\cdot u^2(b)-p(a)\cdot\frac{\alpha_0}{\alpha_1}\cdot u^2(a).$$

Отсюда непосредственно следует:

$$\frac{\beta_0}{\beta_1}\ge 0,\ \ \frac{\alpha_0}{\alpha_1}\le 0.$$

Таким образом, при выполнении найденных условий задача

$$\begin{cases} Lu(x)\equiv -(p(x)u'(x))'+q(x)u(x)=-f(x),\ a\le x\le b,\\ \alpha_0 u(a)+\alpha_1 u'(a)=0,\\ \beta_0 u(b)+\beta_1 u'(b)=0 \end{cases} \tag{1.10}$$

равносильна задаче минимизации функционала $J(u)$, который теперь примет вид

$$J(u)=(Lu,u)-2(f,u)=p(b)\frac{\beta_0}{\beta_1}u^2(b)-p(a)\frac{\alpha_0}{\alpha_1}u^2(a)+ \\ +\int_a^b\left[p(x)(u'(x))^2+q(x)u^2(x)+2f(x)u(x)\right]dx. \tag{1.11}$$

Упражнение. Исследовать случай A и B, отличных от нуля (и подправить (!) функционал (1.11)).

1.3. Метод Ритца нахождения минимума функционала

После того как для заданной граничной задачи построен функционал $J(u)$, минимизация которого эквивалентна отысканию решения исходной задачи, встает вопрос о том, каким образом элемент, доставляющий минимум, может быть найден.

Такая задача (минимизации функционала) имеет смысл и самостоятельно, без привязки к дифференциальным уравнениям. Поэтому основные понятия и идеи рассмотрим на примере задачи минимизации функционала несколько более общего вида

$$J(u)=\int_a^b F(x,u,u')dx \tag{1.12}$$

на множестве функций $u(x)\in C^1[a;b]$, удовлетворяющих граничным условиям (*).

Предположим, что множество значений функционала $J(u)$ ограничено снизу и существует такая допустимая функция $u^*(x)$, что

$$J(u^*)=\min_u J(u)=m.$$

Тогда, если существует последовательность допустимых функций $u_n(x),\ n=0,1,...$, для которой соответствующая ей последовательность функционалов $J(u_n)$ сходится к минимуму m, т.е.

$$m_n = J(u_n) \underset{n\to\infty}{\to} m = J(u^*),$$

то такая последовательность называется ***минимизирующей***.

Из сходимости последовательности функционалов, вообще говоря, не всегда следует сходимость последовательности их аргументов, т.е. из того, что последовательность $\{u_n(x)\}$ – минимизирующая, еще не следует, что она сходится к $u^*(x)$. Эта сходимость будет иметь место только при определенных условиях, которым должен быть подчинен способ построения минимизирующей последовательности.

В качестве n-го приближения к $u^*(x)$ будем брать n-й член некоторой последовательности $\{u_n(x)\}$. Способ, который мы рассмотрим далее, принадлежит в В. Ритцу и был предложен им в 1908 г. Его основная идея состоит в замене задачи нахождения минимума функционала более простой задачей поиска минимума функции.

Рассмотрим семейство функций

$$u_n(x) = \varphi(x, a_1, a_2, ..., a_n),\ n=1,2,... , \tag{1.13}$$

где φ – некоторая заданная функция, а $a_1,\ a_2, ..., a_n$ – числовые параметры. Будем считать, что при любых конечных значениях параметров a_i каждая функция этого семейства удовлетворяет условиям:

1) $\varphi(x, a_1, a_2, ..., a_n) \in C^1[a;b]$;

2) $\varphi(a, a_1, a_2, ..., a_n) = A;\ \ \varphi(b, a_1, a_2, ..., a_n) = B$.

Выписанные условия означают, что все функции рассматриваемого семейства являются допустимыми.

Легко видеть, что значение функционала (1.12) на функции $u_n(x)$ представляет собой некоторую функцию, аргументами которой являются числовые параметры $a_1,\ a_2, ..., a_n$, т.е.

$$J(u_n) = \int_a^b F(x, u_n, u_n')dx = \Phi(a_1, a_2, ..., a_n).$$

Таким образом, задача об отыскании минимума функционала $J(u_n)$ по всевозможным функциям $u_n(x)$ свелась к задаче отыскания минимума функции $\Phi(a_1, a_2, ..., a_n)$. При этом

$$J(u_n^*) = \min_{a_1,...,a_n} \Phi(a_1, a_2, ..., a_n) = m_n = \Phi(a_1^*, a_2^*, ..., a_n^*).$$

Записав необходимые условия минимума первого порядка, получим следующую систему уравнений для определения параметров $a_1^*, a_2^*, ..., a_n^*$:

$$\frac{\partial \Phi(a_1^*, a_2^*, ..., a_n^*)}{\partial a_i} = 0,\ \ i = 1,2,...,n\ . \tag{1.14}$$

Если система (1.14) (в общем случае нелинейная) имеет решение, то найденный набор a_i^* мы и возьмем в качестве искомого. Тем самым будет построена последовательность Ритца. При этом, очевидно, $J(u_n) \geq J(u_n^*)$.

Если при этом взятое нами семейство функций $\{u_n(x)\}$ вида (1.13) будет достаточно широким и будет хорошо отражать свойства класса допустимых функций, то можно надеяться, что построенная последовательность будет минимизирующей, т.е.

$$\lim_{n\to\infty} J(u_n^*) = m = J(u^*)$$

Имеет место

***Теорема* 3.** Если функция $F(x,u,u')$ непрерывна в области $a \leq x \leq b; -\infty < u,u' < +\infty$, а семейство функций (1.13) расширяется с увеличением n и обладает свойством C^1-полноты, то построенная по Ритцу последовательность $\{u_n^*(x)\}$ – минимизирующая.

Доказательство.

Пусть $u^*(x)$ – функция, доставляющая минимум функционалу $J(u)$ на классе G допустимых функций. Так как $u^*(x) \in G$ и система функций $\{u_n(x)\}$ обладает свойством C^1-полноты, то для любого $\delta > 0$ найдутся такие n и $\widetilde{a}_1,...,\widetilde{a}_n$, что для всех x из отрезка $[a;b]$ будут выполнены неравенства

$$|u^*(x) - \widetilde{u}_n(x)| \leq \delta, \ |(u^*)'(x) - \widetilde{u}_n'(x)| \leq \delta .$$

Учитывая непрерывность функции $F(x,u,u')$, можно для любого значения $\varepsilon > 0$ выбрать такое δ, что будут иметь место неравенства

$$0 \leq J(\widetilde{u}_n) - J(u^*) \leq \varepsilon . \qquad (*)$$

В то же время для последовательности Ритца $\{u_n^*(x)\}$ выполняются неравенства

$$J(u^*) \leq J(u_n^*) \leq J(\widetilde{u}_n) .$$

Поэтому из неравенств (*) следует, что

$$0 \leq J(u_n^*) - J(u^*) \leq \varepsilon ,$$

а отсюда в силу произвольности ε имеем:

$$\lim_{n\to\infty} J(u_n^*) = J(u^*) = m .$$

⊠

1.4. Сходимость минимизирующей последовательности к минимизирующей функции

Вновь возвратимся к нашей конкретной граничной задаче

$$\begin{cases} Lu(x) \equiv -(p(x)u'(x))' + q(x)u(x) = -f(x), \\ u(a) = A, \\ u(b) = B. \end{cases} \tag{1.15}$$

Справедлива

***Теорема* 4.** Если выполняются условия:

1) $p(x) \geq p_0 > 0$ и $p(x) \in C^1[a;b]$;

2) $q(x) \geq 0$ и $q(x), f(x) \in C[a;b]$;

3) последовательность функций $\{u_n(x)\}$ является минимизирующей для вариационной задачи (1.6),

то эта последовательность функций будет равномерно сходящейся на отрезке $[a;b]$ к $u^{*(x)}$ – решению граничной задачи (1.15).

Доказательство.

Рассмотрим функцию $\varepsilon_n(x) = u^*(x) - u_n(x)$, где $u_n(x)$ – n-й элемент произвольной минимизирующей последовательности. Так как $\varepsilon_n(a) = 0$, то

$$\varepsilon_n(x) = \int_a^x \varepsilon_n'(t)dt \ .$$

Отсюда, используя неравенство Коши-Буняковского, получим:

$$|\varepsilon_n(x)| = \left|\int_a^x \varepsilon_n'(t)dt\right| \leq \left|\int_a^x 1^2\,dt\right|^{\frac{1}{2}} \cdot \left|\int_a^x (\varepsilon_n'(t))^2\,dt\right|^{\frac{1}{2}} \leq \sqrt{b-a}\left(\int_a^b (\varepsilon_n'(t))^2\,dt\right)^{\frac{1}{2}} \leq$$

$$\leq \sqrt{b-a}\left(\int_a^b \frac{p(t)}{\min\limits_{x\in[a;b]} p(x)}(\varepsilon_n'(t))^2\,dt\right)^{\frac{1}{2}} \leq \sqrt{\frac{b-a}{\min\limits_{x\in[a;b]} p(x)}}\left\{\int_a^b \left[p(t)(\varepsilon_n'(t))^2 + q(t)\varepsilon_n^2(t)\right]dt\right\}^{\frac{1}{2}} =$$

$$= \sqrt{\frac{b-a}{p_0}}\sqrt{J_n(u) - J(u^*)}.$$

Переход к разности функционалов здесь осуществлен по следующей схеме:

$$\int_a^b \left[p(t)(\varepsilon_n'(t)) + q(t)\varepsilon_n^2(t)\right]dt = (L\varepsilon_n, \varepsilon_n) = (L(u^* - u_n), u^* - u_n) = (Lu^*, u^*) - 2(Lu^*, u_n) + (Lu_n, u_n) =$$

$$= (Lu_n, u_n) + 2(f, u_n) - (Lu^*, u^*) + 2(Lu^*, u^*) = J(u_n) - J(u^*).$$

Так как $\{u_n(x)\}$ – минимизирующая последовательность, то $J(u_n) \underset{n\to\infty}{\to} J(u^*)$, причем независимо от x. Поэтому $\varepsilon_n(x) \underset{n\to\infty}{\to} 0$ или $u_n(x) \underset{n\to\infty}{\to} u^*(x)$. ⊠

1.5. Построение минимизирующей последовательности по Ритцу для граничной задачи (1.15)

На практике часто с целью упрощения системы (1.14) функции (1.13) берут в виде обобщенных полиномов. Тогда построение минимизирующей по методу Ритца может выглядеть следующим образом. Сначала выбирается последовательность координатных функций $\{\varphi_k(x)\},\ k=0,1,2,...$, удовлетворяющих условиям:

1) $\varphi_k(x)\in C^1[a;b],\ k=0,1,2,...$;

2) для функции $\varphi_0(x)$ должны выполняться граничные условия, т.е. $\varphi_0(a)=A,\ \varphi_0(b)=B$; другие функции $\varphi_k(x)$ должны удовлетворять однородным граничным условиям такого же типа, т.е. в нашем случае $\varphi_k(a)=\varphi_k(b)=0,\ k=1,2,...$;

3) при любом конечном n функции $\varphi_1(x),...,\varphi_n(x)$ линейно независимы;

4) образованное по $\{\varphi_k(x)\}$ семейство $\{u_n(x)\}$, где

$$u_n(x)=\varphi_0(x)+\sum_{k=1}^{n}a_k\varphi_k(x),$$

обладает свойством C^1-полноты.

Если систему функций $\{\varphi_k(x)\}$ выбрать указанным образом, то при любом выборе параметров $a_1,...,a_n$ функция $u_n(x)$ будет непрерывно дифференцируемой и удовлетворять граничным условиям, т.е. – допустимой.

Заметим также, что при таком выборе $u_n(x)$ функционал $J(u_n)$ будет квадратичным и поэтому вместо системы (1.14) для определения параметров $a_1,...,a_n$ часто рассматривают другую систему:

$$\frac{1}{2}\frac{\partial J(u_n)}{\partial a_i}=0,\ i=1,...,n. \qquad (1.14')$$

Распишем эту систему более подробно:

$$\frac{1}{2}\frac{\partial J(u_n)}{\partial a_i}=\int_a^b(pu_n'\varphi_i'+qu_n\varphi_i+f\varphi_i)dx=0,\ i=\overline{1,n} \qquad (*)$$

или

$$\int_a^b\left[p\left(\varphi_0'+\sum_{j=1}^{n}a_j\varphi_j'\right)\varphi_i'+q\left(\varphi_0+\sum_{j=1}^{n}a_j\varphi_j\right)\varphi_i+f\varphi_i\right]dx=0,\ i=\overline{1,n}.$$

Поменяв местами порядок суммирования и интегрирования и собирая коэффициенты при α_j, получим:

$$\sum_{j=1}^{n}a_j\int_a^b(p\varphi_i'\varphi_j'+q\varphi_i\varphi_j)dx+\int_a^b(p\varphi_0'\varphi_i'+q\varphi_0\varphi_i+f\varphi_i)dx=0,\ i=\overline{1,n}$$

или, если ввести обозначения

$$\alpha_{ij} = \int_a^b \left(p\varphi_i'\varphi_j' + q\varphi_i\varphi_j\right)dx,\ \beta_i = \int_a^b \left(p\varphi_0'\varphi_i' + q\varphi_0\varphi_j + f\varphi_i\right)dx,\ i = \overline{1,n};\ j = \overline{1,n} \qquad (1.16)$$

то система (1.14′) примет вид

$$\sum_{j=1}^{n} \alpha_{ij} a_j + \beta_i = 0,\ i = \overline{1,n}. \qquad (1.17)$$

Посмотрим, когда она разрешима и определенна. Рассмотрим соответствующую однородную систему:

$$\sum_{j=1}^{n} \alpha_{ij} a_j = 0,\ i = \overline{1,n}.$$

Учитывая (*), запишем ее в виде

$$\int_a^b (pz_n'\varphi_i' + qz_n\varphi_i)dx = 0,\ i = \overline{1,n},$$

где

$$z_n(x) = \sum_{j=1}^{n} a_j\varphi_j(x).$$

Умножим теперь i-е уравнение системы на a_i и просуммируем все уравнения по i от единицы до n. В итоге получим:

$$\int_a^b \left(p(z_n')^2 + qz_n^2\right)dx = 0.$$

Так как интеграл равен нулю и подынтегральная функция неотрицательна, то отсюда имеем:

$$p(z_n')^2 + qz_n^2 \equiv 0,$$

а следовательно, поскольку каждое слагаемое неотрицательно,

$$p(z_n')^2 \equiv 0 \text{ и } qz_n^2 \equiv 0.$$

Далее, в силу того что $p(x) > 0$, то $z_n'(x) \equiv 0$, т.е. $z_n(x) = \text{const}$ или, поскольку $z_n(a) = = z_n(b) = 0$, то $z_n(x) \equiv 0$.

Таким образом, мы получили, что

$$\sum_{j=1}^{n} a_j\varphi_j(x) = z_n(x) \equiv 0.$$

Отсюда, в силу линейной независимости системы $\{\varphi_i(x)\}$ следует, что $a_1 = a_2 = = \ldots = a_n = 0$, т.е. однородная система имеет лишь тривиальное решение, а поэтому соот-

ветствующая ей неоднородная система (1.16), (1.17) разрешима, причем единственным образом.

Легко видеть, что последовательность $\{u_n(x)\}$ будет минимизирующей, поскольку все условия *Теоремы* 3 при таком выборе $\varphi_i(x)$ будут выполнены.

В качестве системы функций $\{\varphi_i(x)\}$ могут быть взяты следующие системы:

1) $\varphi_i(x)=(x-a)^i(b-x)$ или $\varphi_i(x)=(x-a)(b-x)^i$, $i=\overline{1,n}$; (1.18)

2) $\varphi_k(x)=\sin k\pi\dfrac{x-a}{b-a}$, $k=\overline{1,n}$. (1.19)

Упражнения.

1) Проверить выполнение свойств 1)-4) для функций (1.18)-(1.19);
2) Построить системы, аналогичные (1.18), (1.19), для решения третьей краевой задачи.

§ 2. Метод моментов и метод Галеркина решения граничных задач

Пусть при $x\in[a;b]$ задано дифференциальное уравнение

$$F(x,u,u',u'')=0 \tag{2.1}$$

с граничными условиями

$$\begin{cases}u(a)=A,\\ u(b)=B.\end{cases} \tag{2.2}$$

Будем считать, что граничная задача (2.1), (2.2) имеет на отрезке $[a;b]$ единственное решение, принадлежащее классу $C^2[a;b]$.

Следуя общей идее метода моментов (см. Глава VIII, § 1), рассмотрим две системы функций:

I. Система функций $\{\psi_k(x)\}$, $k=1,...,\infty$, подчиненная условиям

1) $\psi_k(x)\in C[a;b]$, $k=1,2,...$;

2) функции $\psi_k(x)$ образуют замкнутую систему, т.е. из того что

$$\int_a^b f(x)\psi_k(x)dx=0,\ k=1,2,...$$

должно с необходимостью следовать, что $f(x)\equiv 0$, если $f(x)\in C[a;b]$.

II. Система функций $\{\varphi_k(x)\}$, $k=0,1,...$, удовлетворяющая условиям:

1) $\varphi_k(x)\in C^2[a;b]$, $k=0,1,...$;

2) при любом конечном n функции $\varphi_1(x),...,\varphi_n(x)$ линейно независимы на $[a;b]$;

3) функция $\varphi_0(x)$ удовлетворяет граничным условиям (2.2) а $\varphi_1(x),...,\varphi_n(x)$ удовлетворяют требованию $\varphi_k(a)=\varphi_k(b)=0$, $k=1,2,...$;

4) функции $\varphi_k(x)$ образуют в классе $C^2[a;b]$ полную систему.

Перейдем к построению приближенного решения граничной задачи (2.1), (2.2) по методу моментов. Функцию F в (2.1) будем считать непрерывной по всем аргументам в области $\{a\le x\le b,\ -\infty<u,u',u''<\infty\}$.

Составим линейную комбинацию

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^{n} a_k \varphi_k(x), \tag{2.3}$$

где a_k – некоторые параметры. В силу выбора функций $\varphi_k(x)$ $u_n(x)$ удовлетворяет граничным условиям (2.2) при любых значениях параметров a_k. Обсудим сейчас проблему выбора последних.

Подставим приближенное решение (2.3) в исходное дифференциальное уравнение (2.1):

$$F(x, u_n, u_n', u_n'') = \delta(x, a_1, a_2, ..., a_n) = \delta_n(x).$$

Здесь функция $\delta_n(x)$ играет роль невязки на приближенном решении. При любом n $\delta_n(x)$ непрерывны, так как непрерывна функция F. Очевидно, выбор параметров $a_1, ..., a_n$ следует осуществлять таким образом, чтобы величина $\delta_n(x)$ была близка к нулю.

Ясно, что если бы удалось удовлетворить бесконечное число требований

$$\int_a^b \delta_n(x)\psi_i(x)dx = 0, \ i = 1, 2, ..., \tag{2.4}$$

то отсюда в силу замкнутости системы функций $\{\psi_k(x)\}$ следовало бы, что $\delta_n(x) = 0$, т.е. $u_n(x)$ – точное решение задачи (2.1), (2.2). Однако за счет выбора конечного числа параметров бесконечному числу требований практически невозможно. Поэтому было бы разумно параметры $a_1, ..., a_n$ выбирать, удовлетворяя первым из n требований системы (2.4), т.е.

$$\int_a^b \delta_n(x)\psi_i(x)dx = 0, \ i = 1, 2, ..., n, \tag{2.4′}$$

В этом случае можно надеяться, что $\delta_n(x)$ будет близко к нулю и при выполнении некоторых дополнительных условий $\delta_n(x) \underset{n\to\infty}{\to} 0$. Можно также ожидать, что $u_n(x)$, определяемое по формуле (2.3), также будет близко к $u(x)$.

Таким образом, в методе моментов приближенное решение ищется в виде (2.3), причем параметры $a_1, ..., a_n$ определяются из системы уравнений (2.4′).

Если уравнение (2.1) будет линейным, т.е.

$$F(x, u, u', u'') \equiv Lu(x) - f(x) = 0$$

где

$$Lu(x) \equiv u''(x) + p(x)u(x) + q(x)u(x), \tag{2.5}$$

то запись системы (2.4′) упростится. В этом случае мы получим систему

$$\sum_{k=1}^{n} c_{ki} a_k - D_i = 0, \ i = \overline{1, n}, \tag{2.6}$$

где

$$c_{ki} = \int_a^b L\varphi_k(x)\psi_i(x)dx, \quad D_i = \int_a^b [f(x) - L\varphi_0(x)]dx\,. \tag{2.7}$$

В частном случае, когда системы функций $\{\psi_i(x)\}$ и $\{\varphi_i(x)\}$ совпадают (при этом все сформулированные к ним ранее требования остаются в силе), мы получим (как это уже отмечалось в Главе VIII) алгоритм метода Галеркина.

Можно также показать, что в случае, когда уравнение (2.1) является уравнением Эйлера, т.е.

$$F(x,u,u',u'') \equiv \frac{d}{dx}\frac{\partial f(x,u,u')}{\partial u'} - \frac{\partial f(x,u,u')}{\partial u} = 0\,,$$

системы уравнений для определения параметров по методу Галеркина и по методу Ритца будут совпадать, т.е. при одинаковом n два этих метода дают одинаковое приближенное решение, определяемое по формуле (2.3).

§ 3. Метод наименьших квадратов решения граничных задач

Этот метод, как и рассмотренные выше, применим к любым функциональным уравнениям. В частности, мы уже применяли его к решению интегральных уравнений. Там же мы сформулировали основные его черты как представителя семейства проекционных методов и отличия от других представителей данного семейства. Рассмотрим сейчас его применение к решению дифференциальных уравнений на примере задачи (2.1), (2.2). Предположения относительно функции F оставим прежними. Приближенное решение, как и ранее, будем искать в виде (2.3), т.е.

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^{n} a_k\varphi_k(x)\,, \tag{3.1}$$

причем $\varphi_k(x)$ удовлетворяют тем же четырем требованиям, что и в методе моментов.

Вновь составим невязку

$$\delta_n(x) = \delta(x,a_1,...,a_n) = F(x,u_n,u_n',u_n'')\,.$$

Однако на сей раз выбор параметров $a_1,...,a_n$ подчиним другому требованию: будем добиваться среднеквадратичной малости невязки $\delta_n(x)$ на отрезке $[a;b]$ (или, что то же самое, квадрата L_2-нормы невязки). Минимизация ее эквивалентна минимизации функционала

$$J(u_n) = \int_a^b F^2(x,u_n,u_n',u_n'')dx = \Phi(a_1,...,a_n)\,.$$

Записывая необходимые условия минимума первого порядка, получим систему уравнений

$$\frac{\partial\Phi(a_1,...,a_n)}{\partial a_i} = 0, \quad i = \overline{1,n}\,, \tag{3.2}$$

из которой находим (если это возможно) значения параметров $a_1,...,a_n$. Исследуем подробнее линейный случай, т.е. случай, когда уравнение (2.1) имеет вид (2.5). Тогда система (3.2) может быть записана в виде

$$\sum_{j=1}^{n} c_{ij} a_j - b_i = 0,\ i = \overline{1,n}, \tag{3.3}$$

где

$$c_{ij} = \int_a^b L\varphi_i(x) L\varphi_j(x)dx,\ i,j = \overline{1,n},$$

$$b_i = \int_a^b (f(x) - L\varphi_0(x))\varphi_i(x)dx,\ i = \overline{1,n}. \tag{3.4}$$

Разрешимость системы (3.3), (3.4) зависит не только от свойств системы функций $\{\varphi_i(x)\}$, но также и от природы рассматриваемой граничной задачи, в частности, от того, имеет ли однородная граничная задача

$$\begin{cases} Lu(x) = 0, \\ u(a) = u(b) = 0 \end{cases}$$

только нулевое решение.

§ 4. Метод коллокации решения граничных задач

Вновь рассмотрим граничную задачу (2.1), (2.2), систему функций $\{\varphi_k(x)\}$, удовлетворяющую прежним требованиям, а приближенное решение ищем в виде

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^{n} a_k \varphi_k(x).$$

В соответствии с идеологией метода коллокации, изложенной ранее на примере применения метода для решения интегральных уравнений, потребуем, чтобы невязка $\delta_n(x)$ была мала в следующем смысле: чтобы в некоторых заданных точках отрезка $[a;b]$ $x_1,...,x_n$ эта невязка обращалась в нуль:

$$\delta_n(x_i) = 0,\ i = \overline{1,n}. \tag{4.1}$$

В итоге получаем систему (в общем случае нелинейных) уравнений для определения параметров $a_1,...,a_n$ приближенного решения.

Если исходная задача имеет вид (2.5) (т.е. становится линейной), то и система (4.1) также станет линейной:

$$\sum_{i=1}^{n} a_i L\varphi_i(x_j) = f(x_j) - L\varphi_0(x_j),\ j = \overline{1,n}. \tag{4.2}$$

Для разрешимости последней необходимо выполнение условия

$$\begin{vmatrix} L\varphi_1(x_1) & \cdots & L\varphi_n(x_1) \\ & \cdots & \\ L\varphi_1(x_n) & \cdots & L\varphi_n(x_n) \end{vmatrix} \neq 0 . \qquad (4.3)$$

Требование (4.2), как легко видеть, равносильно тому, чтобы система функций $\{L\varphi_i(x)\}$ была системой функций Чебышева на отрезке $[a;b]$.

Поскольку метод коллокации можно рассматривать и как решение задачи об интерполировании функции $f(x)$ обобщенным многочленом, построенным по системе функций $\{L\varphi_i(x)\}$ на заданном множестве узлов $x_1,...,x_n$, то задача выбора последних также имеет немаловажное значение. В то же время, как мы помним, существуют большие проблемы со сходимостью интерполяционных процессов (см. соответствующий материал по теории интерполирования). Поэтому в целом, несмотря на простоту системы (4.3), метод коллокации в изложенном виде применяется сравнительно редко.

Замечание. Все изложенные выше алгоритмы проекционно-вариационного типа можно рассматривать с точки зрения теории приближения функций с той лишь разницей, что вместо совпадения на множестве точек (как в методе коллокации) рассматриваются интегральные аналоги этих условий. Так, например, метод наименьших квадратов – это, фактически построение наилучшего среднеквадратичного приближения к функции $f(x)$ и т.п.).

РАЗДЕЛ VII

Сеточные методы решения граничных задач

Дополнительная литература:

1. Самарский А.А. Теория разностных схем. – М.: Наука. 1977.
2. Самарский А.А., Гулин А.В. Устойчивость разностных схем. – М.: Наука. 1973.
3. Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. – М.: Наука. 1978.
4. Самарский А.А., Андреев В.Б. Разностные методы решения эллиптических уравнений. – М.: Наука. 1976.

Для решения граничных задач (причем не только для обыкновенных дифференциальных уравнений, но и для уравнений с частными производными) помимо изученных нами ранее алгоритмов достаточно широко и уже давно применяется еще один подход, обладающий известной универсальностью: подход, позволяющий достаточно несложным образом свести решение указанной выше задачи к решению системы уравнений, неизвестными которой, как правило, являются значения приближенного решения на заданном каким-либо способом множестве точек (узлов). Получающиеся алгоритмы называются сеточными, а метод их получения – методом сеток (или конечных разностей). Раздел численных методов, посвященный теории метода сеток, носит название теории разностных схем. Далее мы познакомимся с ее основными моментами.

ГЛАВА XV

Основные понятия теории разностных схем

§ 1. Сетки и сеточные функции

Система алгебраических уравнений, заменяющая исходную дифференциальную задачу и зависящая от шага замены как от параметра, обычно и называется разностной схемой.

Для того чтобы написать разностную схему, приближенно описывающую рассматриваемую дифференциальную задачу, необходимо совершить следующие два шага:

1. Заменить область непрерывного изменения аргумента областью дискретного его изменения;
2. Заменить дифференциальные операторы некоторыми разностными операторами, а также сформулировать аналогичным образом разностные аналоги краевых условий.

Остановимся на этих вопросах подробнее.

При численном решении той или иной математической задачи (связанной с решением функциональных уравнений) мы, очевидно, не можем воспроизводить решение для всех значений аргумента, изменяющегося внутри некоторой области евклидова пространства. Естественно поэтому выбрать в этой области некоторое конечное подмножество точек и приближенное решение искать только в этих точках. Такое множество точек мы в дальнейшем будем называть ***сеткой*.** Отдельные точки этого множества будем называть ***узлами*.** Функцию, определенную в узлах сетки, будем называть ***сеточной*** функций.

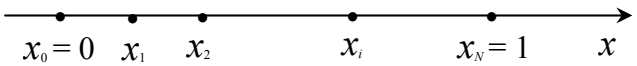
Таким образом, мы заменили область непрерывного изменения аргумента сеткой, т.е. областью дискретного изменения аргумента. Иными словами, мы осуществили ***аппроксимацию*** пространства решений исходной дифференциальной задачи пространством сеточных функций.

Свойства приближенного (разностного) решения, и в частности, его близость к точному решению, зависят от выбора сетки.

Рассмотрим сейчас примеры наиболее часто используемых типов сеточных областей.

1^0. Равномерная сетка на отрезке.

Рассмотрим стандартный отрезок $[0;1]$ и разобьем его на заданное число N равных частей.



Расстояние между соседними узлами $x_i - x_{i-1} = h = \dfrac{1}{N}$ назовем шагом сетки, а точки деления $x_i = ih$ примем в качестве узлов сетки. Множество всех узлов x_i и составляет равномерную сетку на отрезке $[0;1]$, которую в дальнейшем будем обозначать ω_h:

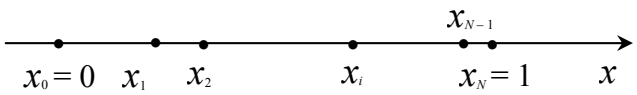
$$\omega_h = \left\{ x_i = ih;\ i = 1,2,...,N-1;\ h = \frac{1}{N} \right\}.$$

В это множество можно включать и граничные узлы. Обозначение такой сетки – $\overline{\omega}_h$:

$$\overline{\omega}_h = \left\{ x_i = ih;\ i = 0,1,2,...,N;\ h = \frac{1}{N} \right\}.$$

На отрезке $[0;1]$ вместо функции непрерывного аргумента $u(x)$ будем рассматривать функцию дискретного аргумента $y_h(x_i)$. Значения этой функции вычисляются только в узлах x_i, а сама функция зависит от шага сетки h как от параметра.

2^0. Неравномерная сетка на отрезке.



Вновь рассмотрим отрезок $[0;1]$. Вводя произвольные точки $0 < x_1 < x_2 < \ldots < x_{N-1} < 1$, разобьем его на N частей. Множество узлов

$$\hat{\overline{\omega}}_h := \{x_i,\ i = 0,...,N;\ x_0 = 0,\ x_N = 1\}$$

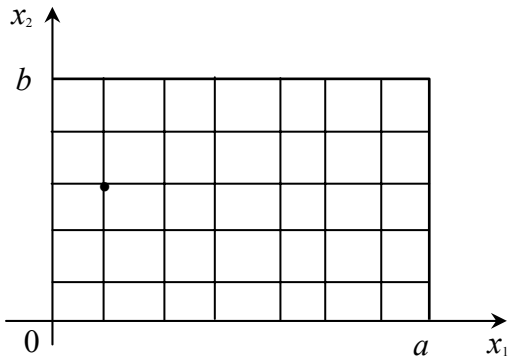
образует неравномерную сетку на отрезке $[0;1]$. Расстояние между соседними узлами – шаг сетки – равно $h_i = x_i - x_{i-1}$ и зависит от номера i узла, т.е. является сеточной функцией. Шаги сетки $\hat{\overline{\omega}}_h$ удовлетворяют условию нормировки

$$\sum_{i=1}^{N} h_i = 1.$$

3^0. Сетки в двумерном случае.

а) Сетка в прямоугольнике

Пусть исходная область G является прямоугольником с основанием a и высотой b, основание которого лежит на оси Ox_1, а одна из боковых сторон – на оси Ox_2, т.е.



$$\overline{G} = \{(x_1, x_2) : 0 \le x_1 \le a;\ 0 \le x_2 \le b\}.$$

Разобьем отрезки $[0;a]$ и $[0;b]$ на N_1 и N_2 частей соответственно. Пусть точки деления на оси Ox_1 имеют координаты x_{1,i_1}, а на оси Ox_2 – x_{2,i_2}, причем

$$0 = x_{1,0} < x_{1,1} < \cdots < x_{1,N_1-1} < x_{1,N_1} = a,$$

$$0 = x_{2,0} < x_{2,1} < \cdots < x_{2,N_2-1} < x_{2,N_2} = b.$$

Через точки деления проведем два семейства прямых

$$x_1 = x_{1,i_1},\ i_1 = 0,1,...,N_1;$$

$$x_2 = x_{2,i_2},\ i_2 = 0,1,...,N_2,$$

параллельных соответствующим координатным осям.

В качестве узлов сетки возьмем точки пересечения этих прямых. Общее число узлов равно $(N_1+1)\cdot(N_2+1)$ и все они принадлежат прямоугольнику $\overline{G}$. Распределение узлов характеризуется векторным параметром $h = \{h_{1,1},...,h_{1,N_1};\ h_{2,1},...,h_{2,N_2}\}$, составленным из шагов по каждому направлению:

$$h_{1,i_1} = x_{1,i_1} - x_{1,i_1-1},\ i_1 = 1,...,N_1;$$

$$h_{2,i_2} = x_{2,i_2} - x_{2,i_2-1},\ i_2 = 1,...,N_2.$$

Если все шаги сетки как по направлению x_1, так и по направлению x_2, равны между собой, т.е. $h_{1,1} = h_{1,2} = \cdots h_{1,N_1} =: h_1 = \dfrac{a}{N_1}$, $h_{2,1} = h_{2,2} = \cdots h_{2,N_2} =: h_2 = \dfrac{b}{N_2}$, то сетка называется ***равномерной*** и обозначается

$$\overline{\omega}_h = \overline{\omega}_{h_1 h_2} = \left\{ \left(x_{1,i_1}, x_{2,i_2}\right) : x_{1,i_1} = i_1 h_1;\ x_{2,i_2} = i_2 h_2;\ h_1 = \frac{a}{N_1},\ h_2 = \frac{b}{N_2};\ i_1 = \overline{0,N_1},\ i_2 = \overline{0,N_2} \right\}.$$

В противном случае сетка называется ***неравномерной*** и обозначается

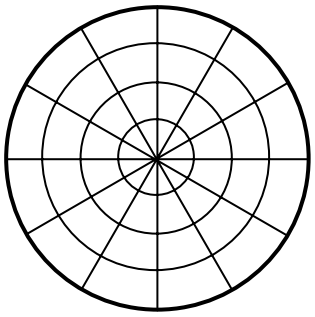
$$\hat{\overline{\omega}}_h = \hat{\overline{\omega}}_{h_1 h_2} = \hat{\overline{\omega}}_{h_1} \times \hat{\overline{\omega}}_{h_2} = \left\{ \left(x_{1,i_1}, x_{2,i_2}\right),\ i_1 = \overline{0,N_1},\ i_2 = \overline{0,N_2};\ x_{1,0} = 0,\ x_{1,N_1} = a;\ x_{2,0} = 0,\ x_{2,N_2} = b \right\}.$$

Если $h_1 = h_2$, то сетку называют ***квадратной***.

Неравномерные сетки бывают эффективны в случае сильно неоднородного решения. Узлы сетки в этом случае сгущают в зоне сильно изменяющегося решения за счет уменьшения их плотности на участках, где решение меняется слабо, не увеличивая их общего количества.

б) Сетка в криволинейной ортогональной системе координат

В качестве примера рассмотрим область $\overline{G}$, имеющую вид круга радиуса R. Двумерную задачу в такой области удобно формулировать в полярных координатах (r,φ), поместив полюс в центр круга. Тогда $\overline{G} = \{(r,\varphi) : 0 \le r \le R,\ 0 \le \varphi < 2\pi\}$.



Проведем два семейства кривых, параллельных координатным линиям:

$$r = r_i,\ i = \overline{1, N_1};$$

$$\varphi = \varphi_j,\ j = \overline{1, N_2 - 1}.$$

Кривые первого семейства являются концентрическими окружностями, а второе семейство образуют лучи, исходящие из полюса. Узлами рассматриваемой сетки будут точки пересечения данных линий, т.е. множество точек $\overline{\omega}_{r\varphi} = \left\{(r_i, \varphi_j),\ i = \overline{0, N_1},\ j = \overline{0, N_2 - 1};\ r_0 = 0,\ r_{N_1} = R;\ \varphi_0 = 0 = \varphi_{N_2}\right\}$.

в) Пространственно-временна́я сетка в прямоугольнике

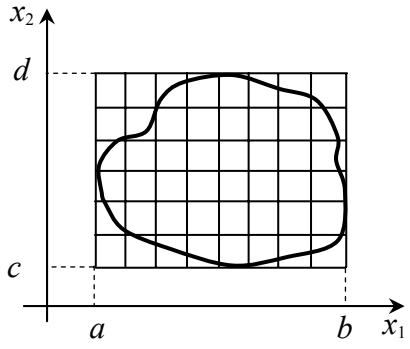
Сетка строится аналогично случаю б), но в других обозначениях: рассматривается область $\overline{G} = \{(x,t) : 0 \le x \le a,\ 0 \le t \le T\}$ и равномерная сетка на ней имеет вид

$$\overline{\omega}_{h\tau} = \left\{(x_i, t_j) : x_i = ih,\ t_j = j\tau;\ h = \frac{a}{N_1},\ \tau = \frac{T}{N_2};\ i = \overline{0, N_1},\ j = \overline{0, N_2}\right\},$$

а неравномерная –

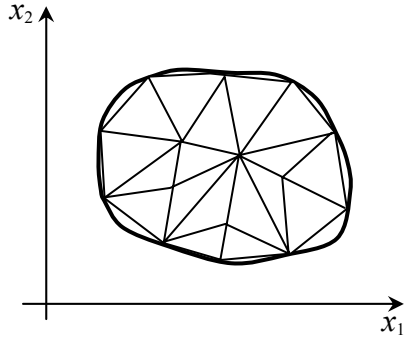
$$\hat{\overline{\omega}}_{h\tau} = \left\{(x_i, t_j) : x_i = x_{i-1} + h_i,\ t_j = t_{j-1} + \tau_j;\ i = \overline{0, N_1},\ j = \overline{0, N_2};\ x_0 = 0,\ x_{N_1} = a;\ t_0 = 0,\ t_{N_2} = T\right\}.$$

г) прямоугольная сетка в области сложной формы



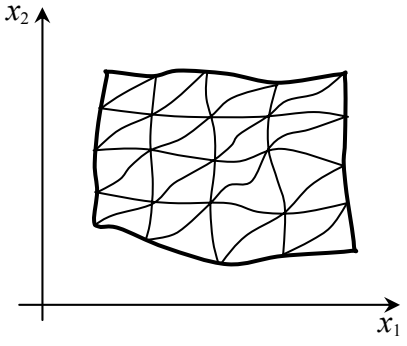
Пусть в плоскости Ox_1x_2 задана область сложной формы G с границей Γ. Заключим область $\overline{G}$ в прямоугольник $\overline{\Pi} = [a;b] \times [c;d]$ (этот прямоугольник может быть как в некотором смысле минимальным, когда на каждой стороне прямоугольника существуют точки области $\overline{G}$, принадлежащие этой стороне, так и любым прямоугольником, содержащим в себе описанный минимальный). После этого зададим на данном прямоугольнике сетку $\overline{\omega}_h$. Те узлы сетки $\overline{\omega}_h$, которые принадлежат области $\overline{G}$, а также точки пересечения прямых, образующих сетку $\overline{\omega}_h$ с границей Γ, и составляют искомую прямоугольную сетку $\overline{\Omega}_h$ в области сложной формы. Несмотря на то, что исходная сетка $\overline{\omega}_h$ равномерна по каждому направлению, построенная сетка $\overline{\Omega}_h$ таковой может не оказаться. Легко видеть, что равномерность может быть нарушена вблизи границы.

д) треугольная сетка в области сложной формы



Вновь рассмотрим область $\overline{G}$ сложной формы, описанную выше. Выберем на границе Γ области множество точек, являющихся узлами ломаной. Эта ломаная будет границей некоторого многоугольника Π, которым мы приближенно заменим исходную область $\overline{G}$. Многоугольник Π покроем множеством треугольником, каждая пара которых либо вовсе не имеет общих точек, либо имеет общую вершину, либо имеет общую сторону. Таким образом, получим треугольную сетку в области

сложной формы. Все узлы (вершины треугольников) здесь можно занумеровать одним индексом: $\xi_p = (x_{1,p}, x_{2,p}),\ p = \overline{0,N}$. Эти узлы выбираются внутри и на границе области $\overline{G}$, вообще говоря, произвольно, исходя из структуры решения и требований точности.



Если же область $\overline{G}$ представляет собой некоторый криволинейный четырехугольник, то в этом случае узлы удобно нумеровать двумя индексами. Сначала выбираем узлы на границе – равное количество на противоположных сторонах криволинейного четырехугольника. Противоположные точки соединяем двумя семействами попарно непересекающихся кривых, включая границы (сравнить: процесс построения параметрических сплайнов от двух переменных). Точки пересечения этих семейств образуют узлы: $\xi_{i_1 i_2} = (x_{1,i_1}, x_{2,i_2}),\ i_1 = \overline{0,N_1};\ i_2 = \overline{0,N_2}$. Элементами сетки являются криволинейные треугольники.

Аналогичным образом строятся сетки и в задачах с количеством независимых переменных $n > 2$.

Итак, область $\overline{G}$ изменения аргумента x мы заменяем сеткой $\overline{\omega}_h$, т.е. конечным множеством точек $x_i \in \overline{G}$. Вместо функций $u(x)$ непрерывного аргумента $x \in \overline{G}$ будем рассматривать сеточные функции $y(x_i)$, т.е. функции точки x_i, являющейся узлом сетки $\overline{\omega}_h$. Сеточную функцию можно представить в виде вектора. Если перенумеровать все узлы в некотором порядке $(x_1, x_2, ..., x_N)$, то значения сеточной функции в этих узлах можно рассматривать как компоненты вектора $Y = (y_1, y_2, ..., y_N)$. Если область $\overline{G}$, в которой построена сетка, конечна, то размерность N вектора Y также конечна. В случае неограниченной области G сетка состоит из бесконечного числа узлов и, следовательно, размерность вектора Y также бесконечна.

Обычно рассматриваются множества сеток $\{\omega_h\}$, зависящие от шага h как от параметра. Поэтому и сеточные функции $y_h(x)$ зависят от параметра h (или от числа узлов N в случае равномерной сетки). Если сетка ω_h неравномерна, то под h следует понимать вектор $h = (h_1, h_2, ..., h_N)$. Это же замечание относится и к случаю, когда область G многомерна, т.е. $x = (x_1, x_2, ..., x_p)$; тогда $h = (h_1, h_2, ..., h_p)$, если сетка ω_h равномерна по каждому из аргументов $x_1, x_2, ... x_p$.

Функции $u(x)$ непрерывного аргумента $x \in G$ являются элементами некоторого функционального пространства H_0. Множество сеточных функций $y_h(x)$ образует пространство H_h. Таким образом, используя метод конечных разностей, мы заменяем пространство H_0 функций $u(x)$ непрерывного аргумента пространством H_h сеточных функций $y_h(x)$.

Рассматривая множество сеток $\{\omega_h\}$, мы получаем множество пространств сеточных функций $\{H_h\}$, зависящих от параметра h. В линейном пространстве H_h вводится норма $\|\cdot\|_h$, являющаяся сеточным аналогом нормы $\|\cdot\|_0$ в исходном пространстве H_0.

Укажем простейшие типы норм в пространстве H_h сеточных функций для случая сетки $\overline{\omega}_h = \left\{x_i = ih,\ i = \overline{0,N},\ h = \frac{1}{N}\right\}$ на отрезке $[0;1]$ (индекс h у y_h тогда будем опускать):

1^0. Сеточный аналог нормы в пространстве C ($H_0 = C$):

$$\| y \|_C = \max_{x \in \omega_h} |y(x)| \quad \text{или} \quad \| y \|_C = \max_{0 \le i \le N} |y_i| ;$$

2^0. Сеточные аналоги нормы в пространстве L_2 ($H_0 = L_2$):

$$\| y \| = \left(\sum_{i=1}^{N-1} h y_i^2 \right)^{\frac{1}{2}} \quad \text{или} \quad \| y] | = \left(\sum_{i=1}^{N} h y_i^2 \right)^{\frac{1}{2}}.$$

Пусть $u(x)$ – решение исходной непрерывной задачи, $u \in H_0$, y_h – решение приближенной (разностной) задачи, $y_h \in H_h$. Основной интерес для теории приближенных методов представляет оценка близости y_h к u. Однако y_h и u являются элементами различных функциональных пространств. Поэтому для изучения вопроса о близости y_h и u принципиально имеются две возможности:

1) Сеточная функция y_h, заданная в узлах сетки $\omega_h(G)$, доопределяется (например, с помощью интерполяции) во всех остальных точках области G. В результате получаем функцию $\widetilde{y}(x,h)$ непрерывного аргумента $x \in G$. Разность $\widetilde{y}(x,h) - u(x)$ принадлежит пространству H_0, а близость y_h к u характеризуется числом $\| \widetilde{y}(x,h) - u(x) \|_0$ (см., например, оценку погрешности метода механических квадратур);
2) Пространство H_0 отображается на пространство H_h. Каждой функции $u(x) \in H_0$ ставится в соответствие сеточная функция $u_h(x)$, $x \in \omega_h$ так что $u_h = P_h u \in H_h$, где P_h – линейный оператор из H_0 в H_h. Это соответствие можно осуществить поразному, выбирая различные операторы P_h. Если, например, $u(x)$ – непрерывная функция, то можно положить (и чаще всего делается именно так) $u_h(x) = u(x)$, $x \in \omega_h$. Иногда определяют $u_h(x)$, $x \in \omega_h$ как интегральное среднее значение $u(x)$ по некоторой окрестности (например, диаметра $O(h)$) данного узла x_i. Имея сеточную функцию u_h, образуем разность $y_h - u_h \in H_h$. Близость y_h к u будет, следовательно, характеризоваться числом $\| y_h - u_h \|_h$. При этом естественно требовать, чтобы норма $\| \cdot \|_h$ аппроксимировала норму $\| \cdot \|_0$

$$\lim_{h \to 0} \| u_h \|_h = \| u \|_0 \quad \text{для всех } u \in H_0 .$$

Это условие будем называть условием согласованности норм в H_h и H_0.

При изучении вопроса о близости y_h к u чаще всего используется второй подход.

§ 2. Разностная аппроксимация дифференциальных операторов

После того как область G заменена сеточной областью ω_h, можно переходить к следующему этапу: замене дифференциального оператора его разностным аналогом. Такую замену обычно называют аппроксимацией дифференциального оператора разностным оператором.

Рассмотрим этот вопрос несколько подробнее. Итак, пусть задан линейный дифференциальный оператор L, действующий на функцию $u = u(x)$. Для того чтобы аппрок-

симировать его в любой точке сетки ω_h разностным оператором L_h, действующим на сеточную функцию u_h, необходимо вначале указать (выбрать) ***шаблон***, т.е. множество узлов $Ш(x)$ сетки, которое будет непосредственно использоваться при аппроксимации оператора L оператором L_h в точке $x \in \omega_h$. Обычно выбор шаблона $Ш(x)$ зависит от порядка производных, входящих в оператор L, а также от некоторых других моментов.

Сама же аппроксимация может быть осуществлена двумя способами:

а) методом неопределенных коэффициентов;

б) методом численного дифференцирования.

Прежде чем рассматривать эти способы, напомним, что разность

$$\psi(x) = L_h u(x) - Lu(x), \ x \in \omega_h$$

называется ***погрешностью аппроксимации*** дифференциального оператора L разностным оператором L_h в точке $x \in \omega_h$.

Кроме того, будем говорить, что L_h аппроксимирует дифференциальный оператор L с порядком $m > 0$ в точке $x \in \omega_h$, если $\psi(x) = O\left(|h|^m\right)$.

Рассмотрим теперь подробнее способы построения разностных операторов.

1^0. Способ неопределенных коэффициентов.

Выбрав шаблон $Ш(x)$, разностную аппроксимацию $L_h u(x)$ будем искать в виде линейной комбинации значений функции u в точках этого шаблона:

$$L_h u(x) = \sum_{\xi \in Ш(x)} A_h(x,\xi) u(\xi), \tag{2.1}$$

где $A_h(x,\xi)$ – неизвестные коэффициенты, выбор которых осуществляется таким образом, чтобы погрешность аппроксимации $\psi(x)$ имела в точке x заданный (чаще всего – максимально возможный в данной ситуации) порядок. Практически это осуществляется путем разложения (при естественном предположении законности этой операции) погрешности аппроксимации $\psi(x)$ в ряд Тейлора

$$\psi(x) = \sum_{\xi \in Ш(x)} A_h(x,\xi) u(\xi) - Lu(x) = \sum_{|j| \geq 0} B_h^{(j)}(x) u^{(j)}(x)$$

и приравниванием к нулю заданного (максимального) количества первых членов разложения. После этого, решив получившуюся систему линейных алгебраических уравнений, найдем коэффициенты $A_h(x,\xi)$ и по формуле (2.1) запишем искомый разностный оператор.

Несложно заметить, что для аппроксимации дифференциального оператора, содержащего производную порядка k по некоторой независимой переменной, необходимо использовать шаблон, содержащий не менее $(k+1)$ точек вдоль координатного направления по данной независимой переменной, поскольку при разложении функции $u(x)$ в ряд Тейлора по данной переменной производная k-го порядка при фиксированных других будет находиться на $(k+1)$-м месте.

Приведем примеры построения простейших разностных аппроксимаций

Пример 1. $Lu(x) = \dfrac{du(x)}{dx} = u'(x)$.

Необходимый для аппроксимации шаблон должен содержать, по крайней мере, две точки, одна из которых – точка аппроксимации x.

а) Выберем в качестве шаблона $Ш(x)$ узлы x и $x+h$, т.е. $Ш(x) = \{x, x+h\}$. Тогда, согласно (2.1), разностный оператор L_h будем искать в виде

$$L_h u(x) = a_0 u(x) + a_1 u(x+h).$$

Запишем погрешность аппроксимации $\psi(x)$

$$\psi(x) = a_0 u(x) + a_1 u(x+h) - u'(x)$$

и разложим ее в ряд Тейлора в окрестности точки x:

$$\psi(x) = a_0 u(x) + a_1 \left[u(x) + h u'(x) + \frac{h^2}{2} u''(x) + \cdots \right] - u'(x) =$$

$$= (a_0 + a_1) u(x) + (h a_1 - 1) u'(x) + \frac{h^2}{2} a_1 u''(x) + \cdots.$$

Приравнивая к нулю первые коэффициенты разложения, получим:

$$\begin{cases} a_0 + a_1 = 0, \\ h a_1 - 1 = 0, \end{cases}$$

откуда $a_1 = \frac{1}{h},\ a_0 = -\frac{1}{h}$.

Таким образом,

$$L_h u := u_x = \frac{u(x+h) - u(x)}{h}. \qquad (2.2)$$

Аппроксимация (2.2) носит название ***правой разностной производной***. При этом

$$\psi(x) = \frac{h^2}{2} a_1 u''(x) + \cdots = \frac{h^2}{2} \cdot \frac{1}{h} u''(x) + \cdots = \frac{h}{2} u''(x) + \cdots = O(h),$$

т.е. правая разностная производная аппроксимирует исходный дифференциальный оператор первой производной с первым порядком.

б) Выбрав в качестве шаблона $Ш(x)$ множество узлов $Ш(x) = \{x-h,\ x\}$, точно так же легко получить следующую аппроксимацию:

$$L_h u := u_{\bar{x}} = \frac{u(x) - u(x-h)}{h}. \qquad (2.3)$$

(2.3) – ***левая разностная производная***, причем

$$\psi(x) = u_{\bar{x}}(x) - u'(x) = -\frac{h}{2} u''(x) + O(h^2) = O(h).$$

в) На шаблоне из трех точек $(Ш(x) = \{x-h,\ x,\ x+h\})$ можно получить (если не требовать максимального порядка аппроксимации) однопараметрическое семейство разностных операторов

$$L_h^{(\sigma)} u = \sigma u_x + (1-\sigma) u_{\bar{x}}, \qquad (2.4)$$

где σ – любое вещественное число. При этом

$$\psi(x) = L_h^{(\sigma)}u(x) - u(x) = (2\sigma - 1)\frac{h}{2}u''(x) + O(h^2).$$

Отсюда следует, что при любом $\sigma \neq \frac{1}{2}$ разностный оператор $L_h^{(\sigma)}u$ имеет первый порядок аппроксимации. В то же время при $\sigma = \frac{1}{2}$, как легко видеть, погрешность аппроксимации становится величиной второго порядка. Сам разностный оператор в этом случае принимает вид

$$L_h^{(0.5)}u = \frac{u_x + u_{\bar{x}}}{2} := u_{\mathring{x}} = \frac{u(x+h) - u(x-h)}{2h} \tag{2.5}$$

и называется ***центральной разностной производной***, а его погрешность –

$$\psi(x) = u_{\mathring{x}}(x) - u'(x) = \frac{h^2}{6}u'''(x) + \cdots = O(h^2).$$

2^0. Способ численного дифференцирования (см. также § 7 главы IV).

Формальная схема данного способа выглядит следующим образом: выбрав шаблон $Ш(x)$, заменяем на этом шаблоне функцию $u(x)$ интерполяционным многочленом (или в общем случае интерполяционной функцией заданного вида):

$$u(x) = P(x) + r(x).$$

После этого применим к последнему равенству дифференциальный оператор L, аппроксимацию которого мы ищем:

$$Lu(x) = LP(x) + Lr(x). \tag{2.6}$$

Заметим, что равенство (2.6) справедливо для всех значений x, а не только в узлах сетки. Поэтому, рассмотрев его в интересующем нас узле $x \in \omega_h$, получим:

$$\begin{aligned} L_h u(x) &= LP(x), \\ \psi(x) &= -Lr(x). \end{aligned} \tag{2.7}$$

Получим описанным способом аппроксимации примера 1.

а) Здесь $Ш(x_i) = \{x_i, x_i + h\}$. Тогда

$$P(x) = P_1(x) = u(x_i) + (x - x_i)u(x_i, x_i + h);$$

$$P_1'(x) = u(x_i, x_i + h) = \frac{u(x_i + h) - u(x_i)}{h} = u_{x,i}.$$

Так как

$$r(x)=\frac{\omega_2(x)u''(\xi)}{2!},$$

то

$$\psi(x_i)=-r'(x_i)=-\frac{1}{2!}\left[\omega_2'(x)u''(\xi)+\omega_2(x)\frac{d}{dx}u''(\xi)\right]\Big|_{x=x_i}=-\frac{1}{2!}\omega_2'(x_i)u''(\xi)=\frac{h}{2}u''(\xi),$$

т.е. получили результаты, полностью согласующиеся с предыдущими.

Упражнение. Используя способ численного дифференцирования, получить аппроксимации (2.4), (2.5).

Пример 2. $Lu(x)=u''(x)=\dfrac{d^2u(x)}{dx^2}$.

Здесь уже минимально необходимое количество узлов шаблона равно трем. Выбрав в качестве такового $Ш(x)=\{x-h,x,x+h\}$, любым из описанных выше способов построим разностный оператор

$$L_hu(x):=u_{\bar{x}x}=\frac{u(x+h)-2u(x)+u(x-h)}{h^2},\qquad(2.8)$$

который носит название ***второй разностной производной***. При этом погрешность аппроксимации имеет вид

$$\psi(x)=\frac{h^2}{12}u^{\mathrm{IV}}(x)+O\left(h^4\right)=O\left(h^2\right),\qquad(2.9)$$

т.е., является величиной второго порядка, а не первого, как следовало бы ожидать. Этот факт объясняется совпадением точки аппроксимации с центром симметрии шаблона.

Пример 3. $Lu(x)=u^{\mathrm{IV}}(x)=\dfrac{d^4u(x)}{dx^4}$.

Выбрав минимально необходимый шаблон из пяти точек вида (вновь выбираем множество узлов максимально симметричным) $Ш(x)=\{x-2h,x-h,x,x+h,x+2h\}$, найдем следующий разностный оператор:

$$L_hu(x):=u_{\bar{x}\bar{x}\bar{x}\bar{x}}=\frac{u(x+2h)-4u(x+h)+6u(x)-4u(x-h)+u(x-2h)}{h^4}.\qquad(2.10)$$

Полученный разностный оператор называется ***четвертой разностной*** производной. Проводя соответствующие разложения в ряд Тейлора, можем записать

$$u_{\bar{x}\bar{x}\bar{x}\bar{x}}=u^{\mathrm{IV}}(x)+\frac{h^2}{6}u^{\mathrm{VI}}(x)+O\left(h^4\right),$$

откуда

$$\psi(x)=\frac{h^2}{6}u^{\mathrm{VI}}(x)+O\left(h^4\right)=O\left(h^2\right),$$

т.е. четвертая разностная производная также имеет второй порядок аппроксимации.

Замечание. Символы, используемые для обозначения второй, четвертой и т.д. разностных производных, в действительности не просто являются обозначениями, но и

предписывают применение в указанном порядке и указанное число раз операторов правой и левой разностных производных. В самом деле, например,

$$u_{\bar{x}x}(x) = \left(u_{\bar{x}}(x)\right)_x = \frac{u_{\bar{x}}(x+h) - u_{\bar{x}}(x)}{h} = \frac{\dfrac{u(x+h)-u(x)}{h} - \dfrac{u(x)-u(x-h)}{h}}{h} =$$

$$= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}.$$

Разложение погрешности аппроксимации в ряд по степеням h в принципе можно использовать для повышения порядка аппроксимации. Действительно, из (2.9) имеем:

$$u_{\bar{x}x} - u'' = \frac{h^2}{12}u^{IV} + O\left(h^4\right) = [\text{используем формулу (2.10)}] =$$

$$= \frac{h^2}{12}\left[u_{\bar{x}x\bar{x}x} + O\left(h^2\right)\right] + O\left(h^4\right) = \frac{h^2}{12}u_{\bar{x}x\bar{x}x} + O\left(h^4\right).$$

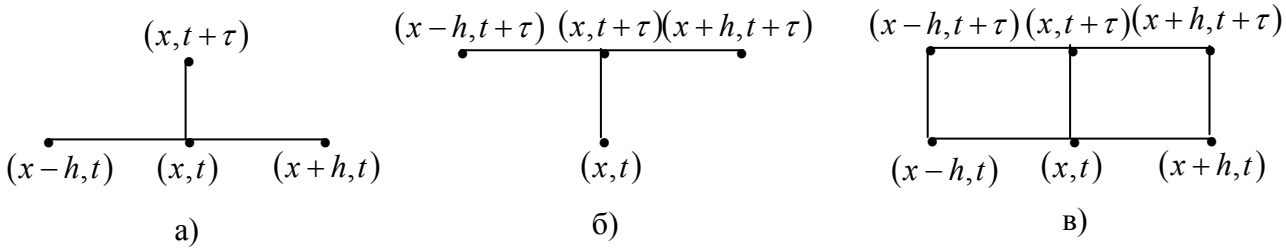
Отсюда следует, что оператор $L_h u = u_{\bar{x}x} - \dfrac{h^2}{12}u_{\bar{x}x\bar{x}x}$, определенный на пятиточечном шаблоне $Ш(x) = \{x-2h, x-h, x, x+h, x+2h\}$, аппроксимирует оператор $Lu = u''$ с четвертым порядком.

Принципиально такой процесс повышения порядка аппроксимации можно продолжить и дальше и получить любой порядок аппроксимации в классе достаточно гладких функций. При этом количество узлов шаблона, естественно, возрастает. Однако указанный прием повышения порядка аппроксимации не всегда можно рекомендовать для практического применения, так как качество получающихся при этом разностных операторов ухудшается (увеличивается объем работы, могут возникнуть проблемы с устойчивостью и т.п.).

Пример 4. $Lu = \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2}$, $u = u(x,t)$.

Прежде чем приступить к построению соответствующих разностных операторов, заметим, что, зная аппроксимации типа (2.2), (2.3) для первой производной и (2.8) для второй, можно записать разностные аппроксимации большинства дифференциальных операторов, использующихся в приложениях (по крайней мере, простейшие).

Сконструируем теперь шаблон для аппроксимации интересующего нас дифференциального оператора Lu. Ранее мы видели, что для аппроксимации оператора первой производной необходимо, как минимум, две точки, а второй – три. Поэтому наш шаблон в простейшем случае может иметь один из следующих видов:



а) б) в)

Используя шаблон а), можем, очевидно, записать такой разностный оператор:

$$L_{h\tau}^{(0)}u = \frac{u(x,t+\tau)-u(x,t)}{\tau} - \frac{u(x+h,t)-2u(x,t)+u(x-h,t)}{h^2}. \qquad (2.11)$$

Для сокращения записи в дальнейшем будем использовать следующие обозначения:

$$u = u(x,t);\ \hat{u} = u(x,t+\tau);\ \check{u} = u(x,t-\tau).$$

В этих обозначениях формула (2.11) может быть переписана в виде

$$L_{h\tau}^{(0)}u = u_t - u_{\bar{x}x}. \qquad (2.12)$$

Используя шаблон б), аналогично можем записать

$$L_{h\tau}^{(1)}u = u_t - \hat{u}_{\bar{x}x}. \qquad (2.13)$$

Взяв линейную комбинацию операторов (2.12) и (2.13), получим однопараметрическое семейство разностных операторов

$$L_{h\tau}^{(\sigma)}u = u_t - \left(\sigma\hat{u}_{\bar{x}x} + (1-\sigma)u_{\bar{x}x}\right), \qquad (2.14)$$

определенных при $\sigma \neq 0$ и $\sigma \neq 1$ на шеститочечном шаблоне в) (случай $\sigma = 0$ дает разностный оператор (2.12), а $\sigma = 1$ – (2.13)).

Для оценки порядка разностной аппроксимации воспользуемся формулами

$$u_t = \frac{\partial u(x,t)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x,t)}{\partial t^2} + O(\tau^2) = \frac{\partial u(x,t+\tau)}{\partial t} - \frac{\tau}{2}\frac{\partial^2 u(x,t+\tau)}{\partial t^2} + O(\tau^2) = \frac{\partial u\left(x,t+\frac{\tau}{2}\right)}{\partial t} + O(\tau^2),$$

$$u_{\bar{x}x} = \frac{\partial^2 u(x,t)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(x,t)}{\partial x^4} + O(h^4) = \frac{\partial^2 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2} - \frac{\tau}{2}\frac{\partial^3 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2),$$

$$\hat{u}_{\bar{x}x} = \frac{\partial^2 u(x,t+\tau)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(x,t+\tau)}{\partial x^4} + O(h^4) = \frac{\partial^2 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2} + \frac{\tau}{2}\frac{\partial^3 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2).$$

Подставляя эти разложения в (2.12), (2.13) и (2.14), получим:

$$L_{h\tau}^{(0)}u = \frac{\partial u(x,t)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x,t)}{\partial t^2} + O(\tau^2) - \frac{\partial^2 u(x,t)}{\partial x^2} - \frac{h^2}{12}\frac{\partial^4 u(x,t)}{\partial x^4} + O(h^4) = Lu(x,t) + O(\tau + h^2);$$

$$L_{h\tau}^{(1)}u = \frac{\partial u(x,t+\tau)}{\partial t} - \frac{\tau}{2}\frac{\partial^2 u(x,t+\tau)}{\partial t^2} + O(\tau^2) - \frac{\partial^2 u(x,t+\tau)}{\partial x^2} - \frac{h^2}{12}\frac{\partial^4 u(x,t+\tau)}{\partial x^4} + O(h^4) =$$

$$= Lu(x,t+\tau) + O(\tau + h^2);$$

$$L_{h\tau}^{(\sigma)} = \frac{\partial u\left(x,t+\frac{\tau}{2}\right)}{\partial t} + O\left(\tau^2\right) - \sigma\left(\frac{\partial^2 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2} + \frac{\tau}{2}\frac{\partial^3 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2 \partial t} + O\left(\tau^2+h^2\right)\right) -$$

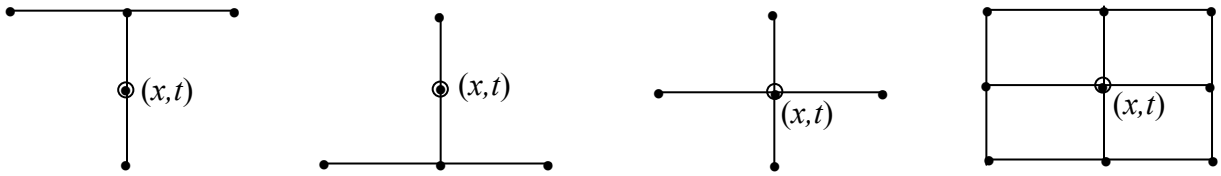
$$-(1-\sigma)\left(\frac{\partial^2 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2} - \frac{\tau}{2}\frac{\partial^3 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2 \partial t} + O\left(\tau^2+h^2\right)\right) = \frac{\partial u\left(x,t+\frac{\tau}{2}\right)}{\partial t} - \frac{\partial^2 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2} +$$

$$+(1-2\sigma)\frac{\tau}{2}\frac{\partial^3 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2 \partial t} + O\left(\tau^2+h^2\right) = Lu\left(x,t+\frac{\tau}{2}\right) + (1-2\sigma)\frac{\tau}{2}\frac{\partial^3 u\left(x,t+\frac{\tau}{2}\right)}{\partial x^2 \partial t} + O\left(\tau^2+h^2\right).$$

Таким образом, оператор $L_{h\tau}^{(\sigma)}$ аппроксимирует оператор L со вторым порядком по h при любом значении параметра σ, с первым порядком по τ при $\sigma \neq 0.5$ (в том числе при $\sigma = 0$ и при $\sigma = 1$) и со вторым порядком по τ при $\sigma = 0.5$.

Пример 5. $Lu = \dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2}$.

В этом случае при конструировании шаблона необходимо учесть, что в операторе L присутствует вторая производная по t. Таким образом, минимально возможными будут следующие конфигурации шаблона:



а) б) в) г)

Соответствующие разностные операторы будут иметь вид:

$$L_{h\tau}^{(1,0)}u = u_{\bar{t}t} - \hat{u}_{\bar{x}x} \quad \left(\text{на шаблоне а), здесь } u_{\bar{t}t} = \frac{\hat{u} - 2u + \breve{u}}{\tau^2}\right); \qquad (2.15)$$

$$L_{h\tau}^{(0,0)}u = u_{\bar{t}t} - u_{\bar{x}x} \quad (\text{на шаблоне в})); \qquad (2.16)$$

$$L_{h\tau}^{(0,1)}u = u_{\bar{t}t} - \breve{u}_{\bar{x}x} \quad (\text{на шаблоне б})); \qquad (2.17)$$

На девятиточечном шаблоне г) можно записать (по аналогии с примером 4, комбинируя разностные операторы (2.15) – (2.17)) двухпараметрическое семейство разностных операторов

$$L_{h\tau}^{(\sigma_1,\sigma_2)}u = u_{\bar{t}t} - \left(\sigma_1\hat{u}_{\bar{x}x} + (1-\sigma_1-\sigma_2)u_{\bar{x}x} + \sigma_2\breve{u}_{\bar{x}x}\right), \qquad (2.18)$$

частными случаями которого являются операторы (2.15) – (2.17) (соответствующие значения σ_1 и σ_2 использованы в их обозначениях).

С помощью разложений, аналогичных использованным в примере 4, несложно показать, что оператор (2.16) имеет погрешность аппроксимации $O\left(\tau^2+h^2\right)$. Этот же порядок имеет и оператор (2.18) при $\sigma_1=\sigma_2=\sigma$, где σ – любое число.

Упражнение. Исследовать погрешность аппроксимации разностных операторов (2.15), (2.17) и (2.18) (при $\sigma_1\neq\sigma_2$).

Следует заметить, что параметры σ_1 и σ_2, так же, как и параметр σ в операторе (2.14), управляют не только порядком аппроксимации, но и устойчивостью соответствующей разностной схемы.

Пример 6. $Lu=u''$.

Вновь вернемся к дифференциальному оператору примера 2, но на сей раз рассмотрим сетку $\hat{\bar{\omega}}_h$ и, следовательно, для построения разностной аппроксимации зададим ***нерегулярный*** трехточечный шаблон $Ш(x)=\{x-h_-,x,x+h_+\}$. Наряду с использованными нами ранее обозначениями для правой и левой разностных производных, которые на данном шаблоне запишутся следующим образом:

$$u_{\bar{x}}=\frac{u(x)-u\left(x-h_-\right)}{h_-};\quad u_x=\frac{u\left(x+h_+\right)-u(x)}{h_+}$$

введем еще и такие:

$$\hbar=\frac{h_++h_-}{2}\quad\text{и}\quad u_{\hat{x}}=\frac{u\left(x+h_+\right)-u(x)}{\hbar}.$$

Тогда для разностной аппроксимации рассматриваемого дифференциального оператора любым из описанных выше способов может быть получено (***проделать***!) следующее выражение

$$L_hu=\frac{1}{\hbar}\left[\frac{u\left(x+h_+\right)-u(x)}{h_+}-\frac{u(x)-u\left(x-h_-\right)}{h_-}\right]=\frac{u_x-u_{\bar{x}}}{\hbar}=:u_{\bar{x}\hat{x}}.\qquad(2.19)$$

Заметим, что при $h_+=h_-=h$ оператор (2.19) совпадает с (2.8).

Так как

$$u_x=u'(x)+\frac{h_+}{2}u''(x)+\frac{h_+^2}{6}u'''(x)+O\left(h_+^3\right),$$

$$u_{\bar{x}}=u'(x)-\frac{h_-}{2}u''(x)+\frac{h_-^2}{6}u'''(x)+O\left(h_-^3\right),$$

то

$$L_hu=u''(x)+\frac{h_+^2-h_-^2}{6\hbar}u'''(x)+O\left(\hbar^2\right)=u''(x)+\frac{h_+-h_-}{3}u'''(x)+O\left(\hbar^2\right).$$

Таким образом, разностный оператор (2.19) при $h_+\neq h_-$ имеет первый порядок аппроксимации.

§ 3. Погрешность аппроксимации на сетке

До сих пор мы рассматривали локальную разностную аппроксимацию (аппроксимацию в точке). Обычно же требуется оценка порядка разностной аппроксимации на всей сетке (заметим: эти порядки ***могут не совпадать***).

Пусть ω_h – сетка в некоторой области G p-мерного пространства, H_h – линейное пространство сеточных функций, заданных на ω_h, H_0 – пространство гладких функций $u(x)$, $\|\cdot\|_0$ – норма в H_0, $\|\cdot\|_h$ – норма в H_h. Как и ранее, будем предполагать, что:

1) существует оператор проектирования $P_h : P_h u = u_h \in H_h$ для всех функций $u \in H_0$;
2) нормы $\|\cdot\|_h$ и $\|\cdot\|_0$ согласованы, т.е. $\lim\limits_{|h|\to 0} \|u_h\|_h = \|u\|_0$.

Рассмотрим некоторый оператор $L : H_0 \to H_0$ и оператор $L_h : H_h \to H_h$. Назовем ***погрешностью аппроксимации*** дифференциального оператора L разностным оператором L_h сеточную функцию

$$\psi_h = L_h u_h - (Lu)_h ,$$

где $u_h = P_h u$, $(Lu)_h = P_h(Lu)$, а u – произвольный элемент из H_0.

Если $\|\psi_h\|_h \underset{|h|\to 0}{\to} 0$, то будем говорить, что разностный оператор L_h аппроксимирует дифференциальный оператор L на сетке ω_h.

Если

$$\|\psi_h\|_h = \|L_h u_h - (Lu)_h\|_h = O\left(|h|^m\right), \tag{3.1}$$

или, что то же самое,

$$\|L_h u_h - (Lu)_h\|_h \le M|h|^m ,$$

где M – не зависящая от $|h|$ константа, а $m > 0$

то будем говорить, что разностный оператор L_h аппроксимирует дифференциальный оператор L на сетке ω_h с порядком m.

Под символом $|h|$ здесь понимается следующее:

а) если $h = (h_1,...,h_p)$ (в случае p-мерного пространства), то, например, $|h| = \sqrt{h_1^2 + \cdots + h_p^2}$.
При этом может оказаться, что аппроксимации по каждому из h_α, $\alpha = 1,...,p$ различны по порядку. Тогда вместо (3.1) можно записать неравенство

$$\|L_h u_h - (Lu)_h\|_h \le M \sum_{\alpha=1}^{p} h_\alpha^{m_\alpha} ,$$

где $m_\alpha > 0$ – порядок аппроксимации по α-й компоненте.

Если теперь положить $m = \min\{m_1, m_2,..., m_p\}$, то получим оценку (3.1).

б) если сетка ω_h – одномерная и неравномерная, т.е. $h = (h_1,...,h_N)$, где N – число узлов, то, например, $|h| = \max\limits_{1\le i\le N} h_i$, или так же, как в предыдущем случае.

Рассмотрим примеры.

Пример 1. Разностная аппроксимация на неравномерной сетке (пример 6 из предыдущего параграфа).

Здесь

$$Lu = \frac{d^2u}{dx^2};\ \ u \in H_0 = C^4[0;1];\ \ \hat{\overline{\omega}}_h = \left\{x_i,\, i = \overline{0,N};\ x_0 = 0,\ x_N = 1\right\};\ \ (L_h u)_i = u_{\overline{x}\hat{x},i},$$

причем

$$\psi_i = \frac{h_{i+1} - h_i}{3} u_i'' + O\left(\hbar_i^2\right),\ \ i = 1, 2, ..., N-1.$$

Отсюда видно, что оператор $L_h u$ имеет в сеточной норме C первый порядок аппроксимации:

$$\|\psi_h\|_C = \max_{1 \le i \le N-1} |\psi_i| = O(h),\ \ h = \max_{1 \le i \le N} h_i.$$

В сеточной L_2-норме также получим первый порядок аппроксимации:

$$\|\psi_h\|_{L_2} = \left(\sum_{i=1}^{N-1} \hbar_i \psi_i^2\right)^{\frac{1}{2}} = O(h),\ \ h = \max_{1 \le i \le N} h_i.$$

Однако в норме

$$\|\psi_h\|_{(-1)} = \left[\sum_{i=1}^{N-1} h_i \left(\sum_{k=1}^{i} \hbar_k \psi_k\right)^2\right]^{\frac{1}{2}} \tag{3.2}$$

ψ_h имеет второй порядок, т.е. $\|\psi_h\|_{(-1)} = O\left(h^2\right)$, где $h = \max\limits_{1 \le i \le N} h_i$. Действительно,

$$\psi_i = \frac{h_{i+1}^2 - h_i^2}{6\hbar_i} u_i''' + O\left(\hbar_i^2\right)$$

и, так как $u_i''' = u_{i+1}''' + O\left(h_{i+1}\right)$, то

$$\psi_i = \frac{h_{i+1}^2 u_{i+1}''' - h_i^2 u_i'''}{6\hbar_i} + \psi_i^* = \mathring{\psi}_i + \psi_i^*,$$

где $\psi_i^* = O\left(h^2\right)$ в любой норме. Главный же член разложения имеет так называемый ***дивергентный*** вид. Поэтому

$$S_i = \sum_{k=1}^{i} \hbar_k \mathring{\psi}_k = \frac{1}{6} \sum_{k=1}^{i} \left(h_{k+1}^2 u_{k+1}''' - h_k^2 u_k'''\right) = \frac{h_{i+1}^2 u_{i+1}''' - h_i^2 u_i'''}{6},$$

т.е. $|S| \le Mh^2$.

Следовательно,

$$\left\|\mathring{\psi}_h\right\|_{(-1)} = \left(\sum_{i=1}^{N-1} h_i S_i^2\right)^{\frac{1}{2}} = O\left(h^2\right),$$

а так как

$$\|\psi_h\|_{(-1)} \le \left\|\mathring{\psi}_h\right\|_{(-1)} + \left\|\psi_h^*\right\|_{(-1)} = O\left(h^2\right),$$

то $\|\psi_h\|_{(-1)} = O(h^2)$.

Этот пример показывает, что исследование локальной аппроксимации может быть недостаточным для суждения о качестве разностного оператора. Выбор же подходящей нормы всякий раз должен быть предметом изучения, поскольку связан со структурой исходного дифференциального оператора.

Если ищется решение нестационарного уравнения (например, теплопроводности или колебаний), то переменная t выделяется по физическому смыслу (время). Поэтому говорят о порядке аппроксимации отдельно по временно́й переменной t и отдельно по пространственной (пространственным) переменной x. При этом чаще всего используют нормы

$$\| y \|_{h\tau} = \max_{t\in\omega_\tau} \| y(t) \|_h \tag{3.3}$$

или

$$\| y \|_{h\tau} = \left[\sum_{t\in\omega_\tau} \tau \| y(t) \|_h^2 \right]^{\frac{1}{2}}. \tag{3.4}$$

Пример 2. $Lu = \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2}$, $L_{h\tau}u = u_t - u_{\bar{x}x}$.

Тогда в случае, если $u(x,t)$ имеет четыре непрерывных производных по x и две – по t, то (см. Пример 4 предыдущего параграфа)

$$\psi_{h\tau}(x,t) = L_{h\tau}u - (Lu)_{h\tau} = O(\tau + h^2).$$

Отсюда следует, что $L_{h\tau}$ аппроксимирует исходный оператор L со вторым порядком по x и первым по t в любой из норм, рассмотренных выше.

§ 4. Постановка разностной задачи

До сих пор мы занимались приближенной заменой дифференциальных операторов разностными. Однако дифференциальные задачи в целом помимо собственно дифференциальных уравнений включают в себя еще и дополнительные условия, которые и обеспечивают выделение из всей совокупности возможных решений единственного. Поэтому при формулировке разностной задачи, помимо аппроксимации дифференциального уравнения, необходимо описывать в разностном виде еще и эти дополнительные условия. Совокупность разностных уравнений, аппроксимирующих дифференциальное уравнение и дополнительные условия, и называют ***разностной схемой***.

При этом используется две формы записи разностных схем (безындексная и индексная), которыми необходимо научиться пользоваться. Вновь обратимся к примерам.

Пример 1. Задача Коши для обыкновенного дифференциального уравнения первого порядка:

$$\begin{cases} u'(t) = f(t), \ t > 0, \\ u(0) = u_0. \end{cases} \tag{4.1}$$

Выберем равномерную сетку $\omega_\tau = \{t_j = j\tau; \ \tau > 0; \ j = 0,1,...\}$. Тогда в соответствие дифференциальной задаче можно поставить разностную схему, которая в безындексной форме имеет вид

$$\begin{cases} y_t = \varphi, \\ y(0) = u_0 \end{cases} \tag{4.2}$$

а в индексной

$$\begin{cases} \dfrac{y^{j+1} - y^j}{\tau} = \varphi^j, \ j = 0,1,... \\ y^0 = u_0 . \end{cases} \tag{4.2′}$$

При этом правую часть φ можно задавать различными способами, лишь бы выполнялось условие $\varphi - f = O(\tau)$, например,

$$\varphi(t) = f(t) \quad \text{или} \quad \varphi(t) = \frac{f(t) + f(t+\tau)}{2} \ \text{ при } t \in \omega_\tau ,$$

что в индексной форме выглядит соответственно

$$\varphi^j = f(t_j) \quad \text{или} \quad \varphi^j = \frac{f(t_j) + f(t_{j+1})}{2} \ \text{ при } j = 0,1,... .$$

Для нахождения решения, как это следует из (4.2′), получаем рекуррентную формулу

$$\begin{cases} y^{j+1} = y^j + \tau\varphi^j, \ \ j = 0,1,... \\ y^0 = u_0 . \end{cases} \tag{4.3}$$

Пример 2. Задача Коши для системы обыкновенных дифференциальных уравнений первого порядка.

$$\begin{cases} \dfrac{du(t)}{dt} + Au(t) = 0, \ t > 0, \\ u(0) = u_0 , \end{cases} \tag{4.4}$$

где A – $n \times n$-матрица, $u = (u_1,...,u_n)^T$.

Выберем сетку так же, как и в Примере 1. Тогда разностная схема может иметь вид (явная схема Эйлера)

$$\begin{cases} y_t + Ay = 0, \\ y(0) = u_0 \end{cases} \quad \text{или} \quad \begin{cases} \dfrac{y^{j+1} - y^j}{\tau} + Ay^j = 0, \ j = 0,1,... \\ y^0 = u_0 \end{cases} \tag{4.5}$$

(здесь $y^j = (y_1^j,...,y_n^j)^T$).

Решения данной разностной задачи могут быть найдены по рекуррентным формулам типа (4.3):

$$\begin{cases} y^{j+1} = y^j - \tau Ay^j, \ j = 0,1,... \\ y^0 = u_0 . \end{cases}$$

Пример 3. Краевая задача для обыкновенного дифференциального уравнения второго порядка:

$$\begin{cases} u''(x) = -f(x), \ 0 < x < 1, \\ u(0) = \mu_0, \\ u(1) = \mu_1 . \end{cases} \tag{4.6}$$

Вновь выберем равномерную сетку $\overline{\omega}_h$. Тогда разностная схема может иметь вид

$$\begin{cases} y_{\bar{x}x} = -\varphi, \ x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1 \end{cases} \quad \text{или} \quad \begin{cases} \dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = -\varphi_i, \ i = 1,...,N-1, \\ y_0 = \mu_0, \\ y_N = \mu_1. \end{cases} \tag{4.7}$$

Сеточная функция φ, как и в Примере 1, может быть выбрана лишь исходя из условия $\varphi - f = O(h^2)$. Решение задачи (4.7) может быть найдено (напомним: это – система линейных алгебраических уравнений (!) с трехдиагональной матрицей) с помощью метода разностной прогонки.

Пример 4. Первая краевая задача для уравнения теплопроводности:

$$\begin{cases} Lu = \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial t^2} = f(x,t), \ 0 < x < 1, \ 0 < t \le T, \\ u(x,0) = u_0(x), \\ u(0,t) = \mu_0(t), \\ u(1,t) = \mu_1(t). \end{cases} \tag{4.8}$$

Выбрав равномерную сетку $\overline{\omega}_{h\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$ и простейший четырехточечный шаблон (см. Пример 4а из § 2), получим разностную схему

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi, \ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x), \ x \in \overline{\omega}_h, \\ y(0,t) = \mu_0(t), \ t \in \overline{\omega}_\tau, \\ y(1,t) = \mu_1(t), \ t \in \overline{\omega}_\tau \end{cases} \tag{4.9}$$

или в индексной форме

$$\begin{cases} \dfrac{y_i^{j+1} - y_i^j}{\tau} = \dfrac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j, \ i = \overline{1, N_1 - 1}; \ j = \overline{0, N_2 - 1}; \\ y_i^0 = u_0(x_i), \ i = \overline{0, N_1}, \\ y_0^j = \mu_0(t_j), \ j = \overline{0, N_2}, \\ y_{N_1}^j = \mu_1(t_j), \ j = \overline{0, N_2}. \end{cases}$$

Функция φ здесь снова выбирается из условия, аналогичного предыдущим:
$\varphi - f = O(\tau + h^2)$, например, $\varphi_i^j = f(x_i, t_j)$ или $\varphi_i^j = f\left(x_i, t_{j+\frac{1}{2}}\right)$ и т.п.

(4.9) – пример ***явной*** разностной схемы: значения решения на ***верхнем*** временно́м слое y^{j+1} определяются через значения решения на предыдущем слое по явным рекуррентным формулам:

$$y_i^{j+1} = y_i^j + \tau\left(y_{\bar{x}x,i}^j + \varphi_i^j\right), \ i = \overline{1, N_1 - 1}; \ j = \overline{0, N_2 - 1}.$$

Значения y_i^0, y_0^j и $y_{N_1}^j$ при этом известны из начального и граничных условий.

Аналогичным образом, выбрав для аппроксимации дифференциального оператора L в (4.8) шаблон примера 4а из § 2, можем записать ***неявную*** разностную схему:

$$\begin{cases} y_t = \hat{y}_{\bar{x}x} + \varphi, \ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x), \ x \in \overline{\omega}_h, \\ y(0,t) = \mu_0(t), \ t \in \overline{\omega}_\tau, \\ y(1,t) = \mu_1(t), \ t \in \overline{\omega}_\tau \end{cases} \tag{4.9}$$

или в индексной форме

$$\begin{cases} \dfrac{y_i^{j+1} - y_i^j}{\tau} = \dfrac{y_{i+1}^{j+1} - 2y_i^{j+1} + y_{i-1}^{j+1}}{h^2} + \varphi_i^j, \ i = \overline{1, N_1 - 1}; \ j = \overline{0, N_2 - 1}; \\ y_i^0 = u_0(x_i), \ i = \overline{0, N_1}, \\ y_0^j = \mu_0(t_j), \ j = \overline{0, N_2}, \\ y_{N_1}^j = \mu_1(t_j), \ j = \overline{0, N_2}. \end{cases}$$

Значения решения на верхнем временно́м слое y^{j+1} вновь определяются через значения решения на предыдущем слое, но на сей раз путем решения соответствующей системы линейных алгебраических уравнений с трехдиагональной матрицей (например, как в примере 3, с помощью метода разностной прогонки).

§ 5. Сходимость и точность разностных схем

Итак, пусть дифференциальной задаче

$$\begin{cases} Lu = f(x), \ x \in G, & (5.1) \\ lu = \mu(x), \ x \in \Gamma, & (5.2) \end{cases}$$

на сетке $\omega_h + \gamma_h$ поставлена в соответствие разностная схема

$$\begin{cases} L_h y_h = \varphi_h, \ x \in \omega_h, & (5.3) \\ l_h y_h = \chi_h(x), \ x \in \gamma_h. & (5.4) \end{cases}$$

Основной целью всякого приближенного метода является получение решения исходной непрерывной задачи с заданной точностью $\varepsilon > 0$ за конечное число действий $N(\varepsilon)$. Чтобы выяснить принципиальную возможность приближения решения u задачи (5.1)-(5.2) решением y_h задачи (5.3)-(5.4), сравним y_h и $u(x)$. Это сравнение, как обычно, будем проводить в пространстве H_h.

Пусть u_h – значение функции $u(x)$ на сетке $\overline{\omega}_h$. Рассмотрим погрешность разностной схемы (5.3)-(5.4): $z_h = y_h - u_h$ и выпишем задачу для ее определения. Подставляя $y_h = z_h + u_h$ в разностные уравнения (5.3) и (5.4), получим:

$$\begin{cases} L_h z_h + L_h u_h = \varphi_h, \\ l_h z_h + l_h u_h = \chi_h \end{cases}$$

или

$$\begin{cases} L_h z_h = \varphi_h - L_h u_h := \psi_h, \ x \in \omega_h, \\ l_h z_h = \chi_h - l_h u_h := \nu_h, \ x \in \gamma_h. \end{cases} \tag{5.5}$$

Правые части задачи (5.5) (ψ_h и ν_h) называются соответственно погрешностью аппроксимации уравнения (5.1) разностным уравнением (5.3) и граничных условий (5.2) – разностными граничными условиями (5.4) на решении дифференциальной задачи (5.1)-(5.2).

Для оценки погрешности схемы z_h и погрешности аппроксимации ψ_h и ν_h введем на множестве сеточных функций нормы $\|\cdot\|_{(1_h)}$, $\|\cdot\|_{(2_h)}$ и $\|\cdot\|_{(3_h)}$ соответственно.

Будем говорить, что решение разностной задачи (5.3)-(5.4) ***сходится*** к решению задачи (5.1)-(5.2) (или, что то же самое: схема (5.3)-(5.4) ***сходится***), если

$$\| z_h \|_{(1_h)} = \| y_h - u_h \|_{(1_h)} \underset{|h|\to 0}{\to} 0 .$$

Разностная схема ***сходится со скоростью*** $O\left(|h|^n\right)$ или ***имеет*** n***-й порядок точности***, если при достаточно малом $|h| \le h_0$ выполняется неравенство

$$\| z_h \|_{(1_h)} = \| y_h - u_h \|_{(1_h)} \le M |h|^n ,$$

где M – константа, не зависящая от h и $n > 0$.

Говорят также, что разностная схема (5.3)-(5.4) ***обладает*** n***-м порядком аппроксимации***, если

$$\| \psi_h \|_{(2_h)} = O\left(|h|^n\right), \ \| \nu_h \|_{(3_h)} = O\left(|h|^n\right).$$

Обозначая f_h и $(Lu)_h$ значения $f(x)$ и $Lu(x)$ на сетке ω_h и учитывая, что $(f - Lu)_h = 0$, запишем ψ_h в виде

$$\psi_h = \varphi_h - L_h u_h = (u_h - L_h u_h) - (f_h - (Lu)_h) = (\varphi_h - f_h) + ((Lu)_h - L_h u_h) = \psi_h^{(1)} + \psi_h^{(2)} . \quad (5.6)$$

Таким образом, погрешность аппроксимации схемы ψ_h складывается из погрешности аппроксимации $\psi_h^{(1)} = \varphi_h - f_h$ правой части и погрешности аппроксимации $\psi_h^{(2)} = (Lu)_h - L_h u_h$ дифференциального оператора.

Так как ψ_h есть погрешность аппроксимации в классе решений дифференциального уравнения, то условие $\| \psi_h \|_{(2_h)} = O\left(|h|^n\right)$ может быть выполнено, если $\psi_h^{(1)}$ и $\psi_h^{(2)}$ не имеют по отдельности n-го порядка.

Возникает вопрос: как зависит порядок точности схемы от порядка аппроксимации на решении? Прежде чем дать на него ответ, напомним понятие корректной постановки задачи применительно к разностным схемам:

1) решение y_h разностной задачи существует и единственно для всех входных данных φ_h из некоторого допустимого семейства;
2) решение y_h непрерывно зависит от φ_h, причем эта зависимость равномерна по h.

Второе условие корректности означает, что существует константа $M > 0$, не зависящая от h и такая, что при достаточно малом $|h| \le h_0$ выполняется неравенство

$$\| y_h - \tilde{y}_h \|_{(1_h)} \le M \left(\| \varphi_h - \tilde{\varphi}_h \|_{(2_h)} + \| \chi_h - \tilde{\chi}_h \|_{(3_h)} \right), \quad (5.7)$$

где $\tilde{y}_h$ – решение задачи с правой частью $\tilde{\varphi}_h$.

Второе свойство (непрерывной зависимости решения разностной задачи от входных данных), выражаемое неравенством (5.6), называется ***устойчивостью*** (по входным данным).

Теперь ответ на поставленный выше вопрос дает следующая

Теорема (Лакса). Если линейная разностная схема устойчива и аппроксимирует исходную дифференциальную задачу, то она сходится, причем порядок точности схемы определяется ее порядком аппроксимации.

Доказательство.

Если оператор $\widetilde{L}_h = (L_h, l_h)$ линеен и разностная схема (5.3)-(5.4) корректна, то на основании соотношения (5.7) можно записать:

$$\|z_h\|_{(1_h)} \le M\|\widetilde{\psi}_h\|_{(2_h)} \quad \text{или} \quad \|z_h\|_{(1_h)} \le M\big(\|\psi_h\|_{(2_h)} + \|\nu_h\|_{(3_h)}\big). \tag{5.8}$$

Так как разностная схема аппроксимирует исходную дифференциальную задачу, то отсюда непосредственно получаем утверждение теоремы. ⊠

§ 6. Повышение порядка аппроксимации разностных схем

Как уже отмечалось выше, скорость сходимости разностной схемы, если последняя устойчива, совпадает с порядком ее аппроксимации на решении исходной дифференциальной задачи, причем, как это следует из формулы (5.6), порядок аппроксимации может быть более высоким, чем порядок аппроксимации дифференциального оператора разностным. Этот факт может быть использован для повышения порядка аппроксимации разностной схемы без увеличения геометрических размеров шаблона. Рассмотрим этот прием на примерах.

Пример 1. Рассмотрим задачу примера 1 из § 4:

$$\begin{cases} u'(t) = f(t), \\ u(0) = u_0. \end{cases} \tag{6.1}$$

В том же параграфе мы записали для нее разностную схему

$$\begin{cases} y_t = \varphi, \ t \in \omega_\tau \\ y(0) = u_0. \end{cases} \tag{6.2}$$

Найдем невязку разностного уравнения на решении $u(t)$ уравнения (6.1):

$$\psi_\tau(t) = u_t(t) - \varphi(t).$$

Так как

$$u(t+\tau) = u(t) + \tau u'(t) + \frac{\tau^2}{2}u''(t) + \frac{\tau^3}{6}u'''(t) + O(\tau^4)$$

и

$$u'(t) = f(t), \ u''(t) = f'(t), \ u'''(t) = f''(t), \dots,$$

то

$$\psi_\tau(t) = u'(t) + \frac{\tau}{2}u''(t) + \frac{\tau^2}{6}u'''(t) - \varphi(t) + O(\tau^3).$$

Отсюда видим, что:

1) выбирая $\varphi(t)=f(t)+O(\tau)$ (например, $\varphi(t)=f(t)$ (в индексной форме $\varphi^j=f^j$)), получим разностную схему первого порядка аппроксимации;

2) если же выбрать $\varphi(t)=f(t)+\frac{\tau}{2}f'(t)+O(\tau^2)$ (например, $\varphi(t)=f\left(t+\frac{\tau}{2}\right)$ (в индексной форме $\varphi^j=f^{j+\frac{1}{2}}$), или $\varphi(t)=f(t)+\frac{\tau}{2}f_t(t)$ (в индексной форме $\varphi^j=f^j+\frac{\tau}{2}f_t^j$), или $\varphi(t)=\frac{f(t)+f(t+\tau)}{2}$ (в индексной форме $\varphi^j=\frac{f^j+f^{j+1}}{2}$) и т.п.), то разностная схема становится схемой второго порядка;

3) аналогично, выбрав $\varphi(t)=f(t)+\frac{\tau}{2}f'(t)+\frac{\tau^2}{6}f''(t)+O(\tau^3)$ (например, $\varphi(t)=f(t)+\frac{\tau}{2}f_{\mathring{t}}(t)+\frac{\tau^2}{6}f_{\bar{t}t}(t)$ (в индексной форме $\varphi^j=f^j+\frac{\tau}{2}f_{\mathring{t}}^j+\frac{\tau^2}{6}f_{\bar{t}t}^j$) или $\varphi(t)=\frac{5f(t+\tau)+8f(t)-f(t-\tau)}{12}$ (в индексной форме $\varphi^j=\frac{5f^{j+1}+8f^j-f^{j-1}}{12}$)), то получим схему третьего порядка.

Пример 2. Пусть дифференциальная задача имеет вид

$$\begin{cases} u''(x)-qu(x)=-f(x),\ 0<x<1,\ q=const, \\ u(0)=\mu_0, \\ u(1)=\mu_1. \end{cases} \tag{6.3}$$

На равномерной сетке $\overline{\omega}_h$ запишем для нее трехточечную разностную схему

$$\begin{cases} y_{\bar{x}x}-dy=-\varphi,\ x\in\omega_h, \\ y(0)=\mu_0, \\ y(1)=\mu_1. \end{cases} \tag{6.4}$$

Невязка разностных граничных условий на решении задачи (6.3) здесь, очевидно, равна нулю (ибо последние не содержат производных). Поэтому рассмотрим невязку разностного уравнения (точкой, в которой это делается, мы будем считать произвольный узел сетки ω_h):

$$\psi_h=u_{\bar{x}x}-du+\varphi.$$

Так как

$$u_{\bar{x}x}=u''+\frac{h^2}{12}u^{IV}+O(h^4)$$

а

$$u''=qu-f,$$

то

$$\psi_h=u''+\frac{h^2}{12}u^{IV}-du+\varphi+O(h^4)=qu-f+\frac{h^2}{12}u^{IV}-du+\varphi+O(h^4)=$$

$$=(q-d)u+(\varphi-f)+\frac{h^2}{12}u^{IV}+O(h^4).$$

Отсюда видим, что $\psi_h = O(h^2)$ при $d = q + O(h^2),\ \varphi = f + O(h^2)$.

Продифференцировав дважды равенство $u'' = qu - f$, будем иметь:

$$u^{IV} = qu'' - f'' = q(qu - f) - f'' .$$

Тогда невязку можно переписать в виде

$$\psi_h = (q-d)u + (\varphi - f) + \frac{h^2}{12} q(qu - f) - \frac{h^2}{12} f'' + O(h^4) =$$

$$= \left[\varphi - f - \frac{h^2}{12}(f'' + qf)\right] + \left[\left(q + \frac{h^2}{12} q^2\right) - d\right] u + O(h^4).$$

Таким образом, если положить

$$d = q + \frac{h^2}{12} q^2 + O(h^4);\ \ \varphi = f + \frac{h^2}{12}(qf + f'') + O(h^4) = f + \frac{h^2}{12}(qf + f_{\bar{x}x}) + O(h^4),$$

то получим трехточечную разностную схему повышенного (четвертого) порядка аппроксимации на решении исходного дифференциального уравнения (6.3).

Пример 3. Третья краевая задача для обыкновенного дифференциального уравнения второго порядка

$$\begin{cases} u''(x) - qu(x) = -f(x),\ \ 0 < x < 1,\ \ q = const, \\ u'(0) = \sigma_0 u(0) - \mu_0, \\ u(1) = \mu_1. \end{cases} \tag{6.5}$$

Вновь выбрав равномерную сетку $\overline{\omega}_h$, разностное уравнение запишем в виде

$$y_{\bar{x}x} - dy = -\varphi, \tag{6.6}$$

где $\varphi = f + O(h^2),\ d = q + O(h^2)$.

Краевое условие при $x = 1$ удовлетворяется точно:

$$y(1) = \mu_1 . \tag{6.7}$$

Производную $u'(0)$ заменим ***правой*** разностной производной $y_x(0)$ (использование для этих целей левой разностной производной приведет к тому, что одна из используемых точек шаблона (конкретно $x = -h$) выйдет за пределы области, в которой определена задача). Тогда краевое условие при $x = 0$ запишется в виде

$$y_x(0) = \sigma_0\, y(0) - \mu_0 \qquad (\text{или} \quad l_h y = \mu_0), \tag{6.8}$$

причем разностный оператор l_h определен на двухточечном шаблоне $\{0, h\}$, что имеет важное значение при реализации разностной схемы (6.6) – (6.8), поскольку для стандарт-

ного варианта метода разностной прогонки граничные условия должны быть не более чем двухточечными.

Выше (см. пример 2) мы видели, что указанный выбор сеточных функций φ и d приводит к тому, что погрешность аппроксимации разностного уравнения (6.6) $\psi_h = O(h^2)$. С другой стороны, так как

$$u(h) = u(0) + hu'(0) + \frac{h^2}{2}u''(0) + O(h^3),$$

то

$$\nu_h(0) = u_x(0) - \sigma_0 u(0) + \mu_0 = u'(0) + \frac{h}{2}u''(0) + O(h^2) - \sigma_0 u(0) + \mu_0 =$$

$$= [u'(0) - (\sigma_0 u(0) - \mu_0)] + \frac{h}{2}u''(0) + O(h^2) = \frac{h}{2}u''(0) + O(h^2),$$

т.е. $\nu_h(0) = O(h)$.

Таким образом, согласно общей схеме рассуждений, порядок аппроксимации разностной схемы (6.6) – (6.8) равен единице.

Подправим разностное граничное условие (6.8) таким образом, чтобы получить второй порядок аппроксимации. Для этого, введя вместо фиксированных коэффициентов σ_0 и μ_0 из граничного условия исходной задачи подлежащие выбору сеточные коэффициенты (параметры) $\bar{\sigma}_0$ и $\bar{\mu}_0$, будем искать его в виде

$$y_x(0) = \bar{\sigma}_0 y(0) - \bar{\mu}_0. \qquad (6.8')$$

Проделав выкладки, аналогичные проведенным выше, получим:

$$\nu_h(0) = u'(0) - (\bar{\sigma}_0 u(0) - \bar{\mu}_0) + \frac{h}{2}u''(0) + O(h^2).$$

Из исходного дифференциального уравнения (6.5) (предполагая его выполняющимся и при $x = 0$) найдем: $u''(0) = qu(0) - f(0)$, а из первого граничного условия – $u'(0) = \sigma_0 u(0) - \mu_0$. Тогда

$$\nu_h(0) = \sigma_0 u(0) - \mu_0 - (\bar{\sigma}_0 u(0) - \bar{\mu}_0) + \frac{h}{2}(qu(0) - f(0)) + O(h^2) =$$

$$= \left[\sigma_0 - \left(\bar{\sigma}_0 - \frac{h}{2}q\right)\right]u(0) + \left[\bar{\mu}_0 - \left(\mu_0 + \frac{h}{2}f(0)\right)\right] + O(h^2) =$$

$$= \left[\bar{\mu}_0 - \left(\mu_0 + \frac{h}{2}f(0)\right)\right] - \left[\bar{\sigma}_0 - \left(\sigma_0 + \frac{h}{2}q\right)\right]u(0) + O(h^2).$$

Отсюда следует, что, выбрав

$$\bar{\sigma}_0 = \sigma + \frac{h}{2}q, \quad \bar{\mu}_0 = \mu_0 + \frac{h}{2}f(0), \qquad (6.9)$$

получим разностное граничное условие (6.8′) (на том же двухточечном шаблоне) второго порядка аппроксимации.

Пример 4. Третья краевая задача для уравнения теплопроводности

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + f(x,t),\ 0 < x < 1,\ 0 < t \le T, \\ u(x,0) = u_0(x),\ 0 \le x \le 1, \\ \dfrac{\partial u(0,t)}{\partial x} = \sigma_0 u(0,t) - \mu_0(t),\ 0 \le t \le T, \\ u(1,t) = \mu_1(t),\ \ 0 \le t \le T. \end{cases} \tag{6.10}$$

На сетке $\overline{\omega}_{h\tau}$, используя простейший четырехточечный шаблон (пример 4а из § 2), напишем явную разностную схему

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi,\ (x,t) \in \omega_{h\tau} \\ y(x,0) = u_0(x),\ x \in \overline{\omega}_h, \\ y(1,t) = \mu_1(t),\ t \in \overline{\omega}_\tau. \end{cases} \tag{6.11}$$

Эта разностная схема при условии $\varphi(x,t) = f(x,t) + O(\tau + h^2)$ аппроксимирует дифференциальную задачу (6.10) (кроме первого граничного условия) с погрешностью $O(\tau + h^2)$. Однако, заменяя в условии при $x = 0$ производную $\dfrac{\partial u}{\partial x}$ правой разностной производной (как и в предыдущем примере), получим разностное граничное условие

$$y_x(0,t) = \sigma_0 y(0,t) - \mu_0(t),\ t \in \overline{\omega}_\tau, \tag{6.12}$$

имеющее, как легко видеть, лишь первый порядок аппроксимации по переменной x.

Получим сейчас разностное граничное условие с порядком аппроксимации $O(\tau + h^2)$, не увеличивая числа точек шаблона по x. Как и в примере 3 будем искать требуемую аппроксимацию в виде

$$y_x(0,t) = \overline{\sigma}_0 y(0,t) - \overline{\mu}_0(t),\ t \in \overline{\omega}_\tau, \tag{6.13}$$

где $\overline{\sigma}_0$ и $\overline{\mu}_0(t)$ – сеточные функции, подлежащие определению. Исследуем погрешность аппроксимации:

$$\nu(0,t) = u_x(0,t) - \overline{\sigma}_0 u(0,t) + \overline{\mu}_0(t) = \frac{\partial u(0,t)}{\partial x} + \frac{h}{2}\frac{\partial^2 u(0,t)}{\partial x^2} + O(h^2) - \overline{\sigma}_0 u(0,t) + \overline{\mu}_0(t) =$$

$$= \sigma_0 u(0,t) - \mu_0(t) + \frac{h}{2}\frac{\partial^2 u(0,t)}{\partial x^2} + O(h^2) - \overline{\sigma}_0 u(0,t) + \overline{\mu}_0(t) = \left[\frac{\partial^2 u(0,t)}{\partial x^2} = \frac{\partial u(0,t)}{\partial t} - f(0,t) \right] =$$

$$= (\sigma_0 - \overline{\sigma}_0) u(0,t) - \mu_0(t) + \overline{\mu}_0(t) + \frac{h}{2}\frac{\partial u(0,t)}{\partial t} - \frac{h}{2} f(0,t) + O(h^2) = (\sigma_0 - \overline{\sigma}_0) u(0,t) +$$

$$+ \left(\overline{\mu}_0(t) - \mu_0(t) + \frac{h}{2} u_{\bar{t}}(0,t) - \frac{h}{2} f(0,t) \right) + O(\tau^2 + h^2).$$

При проведении записанных выкладок использованы как исходное дифференциальное уравнение (в предположении, что оно выполняется при $x = 0$), так и граничное условие на левом конце отрезка изменения переменной x. При этом был применен прием изменения направления дифференцирования, который позволил вторую производную по x заменить первой производной по t (последнюю затем заменяем разностным аналогом).

Анализируя полученное выражение, приходим к выводу, что при

$$\overline{\sigma}_0 = \sigma,\ \overline{\mu}_0(t) = \mu_0(t) + \frac{h}{2} f(0,t) - \frac{h}{2} y_{\bar{t}}(0,t) \tag{6.14}$$

разностное граничное условие имеет погрешность аппроксимации порядка $O\left(\tau^2 + h^2\right)$.

Таким образом, в безындексной форме явная разностная схема, аппроксимирующая задачу (6.10) с погрешностью порядка $O\left(\tau + h^2\right)$, имеет вид

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi,\ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x),\ x \in \overline{\omega}_h, \\ y_x(0,t) = \sigma_0 y(0,t) - \mu_0(t) - \dfrac{h}{2} f(0,t) + \dfrac{h}{2} y_{\bar{t}}(0,t),\ t \in \omega_\tau, \\ y(1,t) = \mu_1(t),\ t \in \omega_\tau. \end{cases} \tag{6.15}$$

Распишем ее, как и в § 4, в индексной форме и укажем способ реализации:

$$\begin{cases} \dfrac{y_i^{j+1} - y_i^j}{\tau} = \dfrac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j,\ i = \overline{1, N_1 - 1};\ j = \overline{0, N_2 - 1}, \\ y_i^0 = u_0(x_i),\ i = \overline{0, N_1}, \\ \dfrac{y_1^{j+1} - y_0^{j+1}}{h} - \dfrac{h}{2} \dfrac{y_0^{j+1} - y_0^j}{\tau} = \sigma_0 y_0^{j+1} - \mu_0\left(t_{j+1}\right) - \dfrac{h}{2} f\left(0, t_{j+1}\right),\ j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1\left(t_{j+1}\right),\ j = \overline{0, N_2 - 1}. \end{cases}$$

Таким образом, решение на слое t_{j+1} может быть определено по рекуррентным формулам

$$\begin{cases} y_i^{j+1} = y_i^j + \dfrac{\tau}{h^2}\left(y_{i+1}^j - 2y_i^j + y_{i-1}^j\right) + \tau\varphi_i^j,\ i = \overline{1, N_1 - 1}, \\ y_0^{j+1} = \dfrac{1}{\dfrac{1}{h} + \dfrac{h}{2\tau} + \sigma_0}\left(\dfrac{1}{h} y_1^{j+1} + \dfrac{h}{2\tau} y_0^j + \mu_0\left(t_{j+1}\right) + \dfrac{h}{2} f\left(0, t_{j+1}\right)\right),\ j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1\left(t_{j+1}\right),\ j = \overline{0, N_2 - 1}. \end{cases}$$

Начальное значение решения (при $t = 0$) определяется из начального условия:

$$y_i^0 = u_0(x_i),\ i = \overline{0, N_1}.$$

Аналогичным образом может быть построена неявная разностная схема с погрешностью порядка $O\left(\tau + h^2\right)$.

Упражнение. Построить указанную схему.

Пример 5. Первая краевая задача для уравнения колебаний струны

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = \dfrac{\partial^2 u}{\partial x^2} + f(x,t),\ 0 < x < 1,\ 0 < t \le T, \\ u(x,0) = u_0(x),\ 0 \le x \le 1, \\ \dfrac{\partial u(x,0)}{\partial t} = u_1(x),\ 0 \le x \le 1, \\ u(0,t) = \mu_0(t),\ 0 \le t \le T, \\ u(1,t) = \mu_1(t),\ 0 \le t \le T. \end{cases} \tag{6.16}$$

На сетке $\omega_{h\tau}$, используя шаблон «крест», запишем явную разностную схему:

$$\begin{cases} y_{\bar{t}t} = y_{\bar{x}x} + f,\ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x),\ x \in \overline{\omega}_h, \\ y_t(x,0) = u_1(x),\ x \in \overline{\omega}_h, \\ y(0,t) = \mu_0(t),\ t \in \omega_\tau, \\ y(1,t) = \mu_1(t),\ t \in \omega_\tau. \end{cases} \tag{6.17}$$

Разностное уравнение имеет погрешность аппроксимации $O(\tau^2 + h^2)$, первое начальное условие и оба граничных аппроксимируются точно. Второе начальное условие имеет погрешность аппроксимации

$$\nu_1(x,0) = u_t(x,0) - u_1(x) = \frac{\partial u(x,0)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x,0)}{\partial t^2} + O(\tau^2) - u_1(x) = \frac{\tau}{2}\frac{\partial^2 u(x,0)}{\partial t^2} + O(\tau^2) = O(\tau).$$

Чтобы обеспечить разностной схеме второй порядок аппроксимации по времени, поднимем порядок аппроксимации второго начального условия, не расширяя шаблона. Как и в предыдущих примерах, будем искать новое разностное условие в виде

$$y_t(x,0) = \overline{u}_1(x),$$

где $\overline{u}_1(x)$ – неизвестная пока сеточная функция.

Имеем:

$$\nu_1(x,0) = u_t(x,0) - \overline{u}_1(x) = \frac{\partial u(x,0)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x,0)}{\partial t^2} + O(\tau^2) - \overline{u}_1(x) =$$

$$= \left[\frac{\partial^2 u(x,0)}{\partial t^2} = \frac{\partial^2 u(x,0)}{\partial x^2} + f(x,0) = u_0''(x) + f(x,0) \right] = u_1(x) + \frac{\tau}{2}(u_0''(x) + f(x,0)) +$$

$$+ O(\tau^2) - \overline{u}_1(x).$$

Отсюда следует, что в качестве $\overline{u}_1(x)$ можно взять, например,

$$\overline{u}_1(x) = u_1(x) + \frac{\tau}{2}(u_0''(x) + f(x,0))$$

(здесь можно использовать вместо $u_0''(x)$ (это – ***известная*** функция) и ее разностный аналог).

Таким образом, разностная схема второго порядка аппроксимации имеет вид:
в безындексной форме:

$$\begin{cases} y_{\bar{t}t} = y_{\bar{x}x} + f,\ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x),\ x \in \overline{\omega}_h, \\ y_t(x,0) = u_1(x) + \dfrac{\tau}{2}\left(u_0''(x) + f(x,0)\right),\ x \in \overline{\omega}_h, \\ y(0,t) = \mu_0(t),\ t \in \omega_\tau, \\ y(1,t) = \mu_1(t),\ t \in \omega_\tau, \end{cases} \tag{6.18}$$

в индексной форме:

$$\begin{cases} \dfrac{y_i^{j+1} - 2y_i^j + y_i^{j-1}}{\tau^2} = \dfrac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + f_i^j,\ i = \overline{1, N_1 - 1};\ j = \overline{1, N_2 - 1}, \\ y_i^0 = u_0(x_i),\ i = \overline{0, N_1}, \\ \dfrac{y_i^1 - y_i^0}{\tau} = u_1(x_i) + \dfrac{\tau}{2}\left(u_0''(x_i) + f_i^0\right),\ i = \overline{0, N_1}, \\ y_0^{j+1} = \mu_0(t_{j+1}),\ j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1(t_{j+1}),\ j = \overline{0, N_2 - 1}. \end{cases}$$

Решение соответствующей разностной задачи может быть найдено по формулам

$$\begin{cases} y_i^0 = u_0(x_i),\ i = \overline{0, N_1}, \\ y_i^1 = y_i^0 + \tau\left[u_1(x_i) + \dfrac{\tau}{2}\left(u_0''(x_i) + f_i^0\right)\right],\ i = \overline{0, N_1}, \\ y_i^{j+1} = 2y_i^j - y_i^{j-1} + \dfrac{\tau^2}{h^2}\left(y_{i+1}^j - 2y_i^j + y_{i-1}^j\right) + \tau^2 f_i^j,\ i = \overline{1, N_1 - 1}, \\ y_0^{j+1} = \mu_0(t_{j+1}),\ j = \overline{1, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1(t_{j+1}),\ j = \overline{1, N_2 - 1}. \end{cases}$$

§ 7. Математический аппарат теории разностных схем

Как уже отмечалось ранее, основной задачей теории разностных схем является получение априорных оценок решения разностной задачи через ее входные данные. Для этих целей нам в дальнейшем потребуются различные формулы преобразования разностных выражений, а также знание общей методики работы с разностными выражениями.

7.1. Некоторые разностные формулы

Получим сейчас простейшие формулы, проводя аналогию с соответствующими формулами дифференциального исчисления.

1^0. ***Формулы разностного дифференцирования произведения***.

Как известно, в дифференциальном исчислении существует формула дифференцирования произведения

$$(uv)' = u'v + uv' .$$

Для сеточных функций ранее мы ввели (на двухточечном шаблоне) два типа разностных производных – правые и левые. Соответственно этому имеется и две формулы разностного дифференцирования произведения:

$$(uv)_x = u_x v + u^{(+1)} v_x = u_x v^{(+1)} + u v_x , \tag{7.1}$$

$$(uv)_{\bar{x}} = u_{\bar{x}} v + u^{(-1)} v_{\bar{x}} = u_{\bar{x}} v^{(-1)} + u v_{\bar{x}} , \tag{7.2}$$

где $f^{(\pm 1)} = f(x \pm h)$.

Докажем, например, первое из этих равенств. Записывая его в индексной форме, получим:

$$\frac{u_{i+1}v_{i+1} - u_i v_i}{h} = \frac{u_{i+1} - u_i}{h} v_i + u_{i+1} \frac{v_{i+1} - v_i}{h} .$$

Справедливость последнего равенства очевидна.

2^0. *Формулы суммирования по частям*.

Эти формулы являются разностными аналогами формулы интегрирования по частям

$$\int_0^1 uv'dx = uv\Big|_0^1 - \int_0^1 u'vdx .$$

Для сеточных функций, как и в предыдущем случае, имеются формулы двух типов:

$$(u, v_x) = u_N v_N - u_0 v_1 - (u_{\bar{x}}, v], \tag{7.3}$$

$$(u, v_{\bar{x}}) = u_N v_{N-1} - u_0 v_0 - [u_x, v), \tag{7.4}$$

где

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

Докажем, например, формулу (7.3):

$$(u, v_x) = \sum_{i=1}^{N-1} (uv_x)_i h = \left[uv_x = (uv)_x - u_x v^{(+1)}\right] = \sum_{i=1}^{N-1} (uv)_{x,i} h - \sum_{i=1}^{N-1} u_{x,i} v_{i+1} h =$$

$$= u_N v_N - u_1 v_1 - \sum_{i=1}^{N-1} u_{\bar{x},i+1} v_{i+1} h = u_N v_N - u_1 v_1 - \sum_{i=1}^{N} u_{\bar{x},i+1} v_{i+1} h + u_{\bar{x},1} v_1 h =$$

$$= u_N v_N - u_1 v_1 + u_1 v_1 - u_0 v_1 - \sum_{i=1}^{N} u_{\bar{x},i+1} v_{i+1} h = u_N v_N - u_0 v_1 - (u_{\bar{x}}, v],$$

что и требовалось доказать.

3^0. *Первая разностная формула Грина*.

Равенство

$$\int_0^1 u(kv')'\,dx = -\int_0^1 ku'v'dx + kuv'\Big|_0^1$$

в дифференциальном исчислении обычно называют первой формулой Грина.

Для сеточных функций аналог первой формулы Грина можно получить, пользуясь формулами суммирования по частям. Подставляя в (7.3) $u = z,\ \ v = ay_{\bar{x}}$, получим

$$\left(z,(ay_{\bar{x}})_x\right) = -(ay_{\bar{x}}, z_{\bar{x}}] + a_N y_{\bar{x},N} z_N - a_1 y_{x,0} z_0 . \qquad (7.5)$$

(7.5) – первая разностная формула Грина.

Отметим некоторые частные случаи, имеющие более простой вид и часто используемые на практике.

Если $z_0 = z_N = 0$, то первая разностная формула Грина имеет вид

$$(z,(ay_{\bar{x}})) = -(ay_{\bar{x}}, z_{\bar{x}}] \qquad \text{или} \qquad (z,\Lambda y) = -(ay_{\bar{x}}, z_{\bar{x}}],\ \ \Lambda y = (ay_{\bar{x}})_x ; \quad (7.5')$$

при $z = y$ отсюда получаем:

$$(\Lambda y, y) = -\left(a,(y_{\bar{x}})^2\right], \qquad (7.5'')$$

что на практике может быть использовано для исследования знакопостоянства разностного оператора Λ (при знакопостоянстве сеточной функции a).

4^0. *Вторая разностная формула Грина*.

В интегральном исчислении вторая формула Грина имеет вид

$$\int_0^1 u(kv')'\,dx - \int_0^1 v(ku')'\,dx = k(uv' - vu')\Big|_0^1 .$$

Чтобы получить ее разностный аналог, запишем на основании (7.5) соотношение

$$\left(y,(az_{\bar{x}})_x\right) = -(az_{\bar{x}}, y_{\bar{x}}] + a_N z_{\bar{x},N} y_N - a_1 z_{x,0} y_0 . \qquad (7.6)$$

Теперь, вычитая из (7.5) (7.6), будем иметь:

$$\left(z,(ay_{\bar{x}})_x\right) - \left(y,(az_{\bar{x}})_x\right) = a_N (zy_{\bar{x}} - yz_{\bar{x}})_N - a_1 (zy_x - yz_x)_0 . \qquad (7.7)$$

(7.7) – вторая разностная формула Грина. Из нее, в частности, в случае, когда сеточные функции y и z обращаются в ноль при $x = 0$ и $x = 1$, непосредственно следует равенство

$$(z,\Lambda y) = (y,\Lambda z),$$

которое означает самосопряженность введенного выше разностного оператора Λ.

5^0. *Неравенство Коши – Буняковского и ε-неравенство*.

Напомним здесь известные из курса анализа неравенства. Одно из них – неравенство Коши – Буняковского – имеет вид

$$|(u,v)| \le \|u\| \cdot \|v\|,$$

где (u,v) – скалярное произведение в некотором линейном пространстве (в том числе, и в пространстве сеточных функций), а $\|u\| = \sqrt{(u,u)}$. В нашем случае под скалярным произведением будем понимать любое из введенных ранее скалярных произведений в пространстве сеточных функций. Второе из неравенств – ε-неравенство – имеет вид

$$|ab| \le \varepsilon a^2 + \frac{b^2}{4\varepsilon},$$

где ε – любое положительное число. Из него, в частности, получаем неравенство

$$|(u,v)| \le \|u\| \cdot \|v\| \le \varepsilon \|u\|^2 + \frac{1}{4\varepsilon}\|v\|^2.$$

Упражнение. Доказать все недоказанные формулы.

7.2. Отыскание собственных функций и собственных значений на примере простейшей разностной задачи

Применение известного из курса уравнений в частных производных метода разделения переменных в теории разностных схем приводит к появлению разностных задач на собственные значения.

Рассмотрим сейчас задачу об отыскании собственных значений для простейшего разностного оператора.

Предварительно напомним основные факты, связанные с простейшей задачей об отыскании собственных функций и собственных значений для дифференциального оператора второй производной. В математической формулировке задача выглядит следующим образом: найти, при каких значениях числового параметра λ существуют нетривиальные решения краевой задачи

$$\begin{cases} u''(x) + \lambda u(x) = 0, \;\; 0 < x < l, \\ u(0) = u(l) = 0 \end{cases} \tag{7.8}$$

и указать эти решения.

Известно следующее:

1. Нетривиальные решения задачи (7.8) – собственные функции $u_k(x)$ – и отвечающие им собственные значения λ_k – выражаются следующим образом:

$$u_k(x) = \sqrt{\frac{2}{l}} \sin\frac{k\pi x}{l}, \;\; \lambda_k = \frac{k^2\pi^2}{l^2}, \;\; k = 1,2,\ldots; \tag{7.9}$$

2. Собственные функции $u_k(x)$ образуют ортонормированную систему:

$$\int_0^1 u_k(x) u_m(x) dx = \delta_k^m;$$

3. Если $f(x)$ дважды непрерывно дифференцируема и удовлетворяет однородным краевым условиям $(f(0) = f(l) = 0)$, то она представима в виде равномерно сходящегося ряда:

$$f(x)=\sum_{k=1}^{\infty} f_k u_k(x),$$

где

$$f_k=\int_0^l f(x)u_k(x)dx\,,$$

причем

$$\|f\|^2=\int_0^l f^2(x)dx=\sum_{k=1}^{\infty} f_k^2\,.$$

Упражнение. Доказать соответствующие соотношения.

Поставим в соответствие дифференциальной задаче (7.8) на равномерной сетке $\overline{\omega}_h[0;l]$ разностную задачу

$$\begin{cases} y_{\bar{x}x}+\lambda y=0,\ x\in\omega_h\,, \\ y(0)=y(l)=0 \end{cases} \tag{7.10}$$

об отыскании нетривиальных решений – собственных функций $y_k(x)$ и соответствующих им собственных значений.

Перейдем в (7.10) к индексной форме записи:

$$y_{i+1}-2\left(1-\frac{h^2\lambda}{2}\right)y_i+y_{i-1}=0,\ i=\overline{1,N-1}\,. \tag{7.11}$$

(7.11) представляет собой разностное уравнение второго порядка с постоянными коэффициентами. Заметим, что при наличии у соответствующего ему характеристического уравнения

$$r^2-2\left(1-\frac{h^2\lambda}{2}\right)r+1=0 \tag{7.12}$$

вещественных корней построить нетривиальное решение задачи (7.10) (как и в случае исходной дифференциальной задачи (7.8) (!)) не удается. Поэтому необходимым условием разрешимости задачи (7.10) является отрицательность дискриминанта квадратного уравнения (7.12), которая, как легко видеть, имеет место при $\lambda\in\left(0;\frac{4}{h^2}\right)$. Поскольку при этом также $r_1r_2=1$, то решение уравнения (7.11) будем искать (что непосредственно следует из общей теории разностных уравнений) в виде $y=\sin\alpha x$, где постоянная α подлежит определению. Поскольку в этом случае

$$y_{i+1}+y_{i-1}=[x_i=x]=\sin\alpha(x+h)+\sin\alpha(x-h)=2\sin\alpha x\cos\alpha h\,,$$

то из (7.11) следует:

$$2\sin\alpha x\cos\alpha h=2\left(1-\frac{h^2\lambda}{2}\right)\sin\alpha x\,.$$

Так как мы ищем нетривиальное решение, т.е. $\sin\alpha x$ отлично от тождественного нуля, то отсюда следует

$$1-\frac{h^2\lambda}{2}=\cos\alpha h,$$

т.е.

$$\lambda=\frac{2}{h^2}(1-\cos\alpha h)=\frac{4}{h^2}\sin^2\frac{\alpha h}{2}.$$

Значение параметра α выберем так, чтобы функция $y=\sin\alpha x$ удовлетворяла граничным условиям задачи (7.10): $y(0)=y(l)=0$. При $x=0$ $\sin\alpha x=0$ при любых значениях α, а при $x=l$ имеем:

$$\sin\alpha l=0,$$

откуда

$$\alpha l=k\pi,\ k=1,...,N-1.$$

Тогда

$$\alpha=\frac{k\pi}{l}=\alpha_k,\ k=1,...,N-1.$$

Таким образом, мы получили собственные функции и собственные значения задачи (7.10). Перечислим их свойства:

1. Множество собственных функций и собственных значений задачи (7.10) имеет вид

$$y^{(k)}(x)=\sin\frac{k\pi x}{l},\ \ \lambda_k=\frac{4}{h^2}\sin^2\frac{k\pi h}{2l},\ \ k=1,...,N-1. \qquad (7.13)$$

2. Собственные значения λ_k перенумерованы в порядке возрастания, причем

$$0<\lambda_1=\frac{4}{h^2}\sin^2\frac{\pi h}{2l}<\lambda_2<\cdots<\lambda_{N-1}=\frac{4}{h^2}\sin^2\frac{\pi(N-1)h}{2l}=\frac{4}{h^2}\cos^2\frac{\pi h}{2l}<\frac{4}{h^2}. \qquad (7.14)$$

Отсюда, в частности, следует, что все собственные значения задачи (7.10) положительны.

3. Собственные функции задачи (7.10) $y^{(k)}(x)$, $y^{(m)}(x)$, отвечающие различным собственным значениям, ортогональны:

$$\left(y^{(k)},y^{(m)}\right)=0 \quad \text{при } k\neq m.$$

Для доказательства этого факта воспользуемся второй разностной формулой Грина, записанной для однородных краевых условий:

$$0=\left(y^{(k)}_{\bar{x}x},y^{(m)}\right)-\left(y^{(k)},y^{(m)}_{\bar{x}x}\right)=(\lambda_k-\lambda_m)\left(y^{(k)},y^{(m)}\right)$$

Так как $\lambda_k\neq\lambda_m$, то отсюда следует: $\left(y^{(k)},y^{(m)}\right)=0$.

4. $\left\|y^{(k)}\right\|=\sqrt{\frac{l}{2}}$.

Действительно,

$$\left\|y^{(k)}\right\|^2=\left(y^{(k)},y^{(k)}\right)=\sum_{s=1}^{N-1}\left(y^{(k)}(x_s)\right)^2h=\sum_{k=1}^{N-1}h\sin^2\frac{k\pi x_s}{l}=\sum_{k=1}^{N-1}\frac{h}{2}\left(1-\cos\frac{2k\pi x_s}{l}\right). \qquad (7.15)$$

Вычислим сумму косинусов: пусть $q_k = \exp\left(i\frac{2k\pi h}{l}\right)$. Тогда

$$q_k^s = \exp\left(i\frac{2k\pi hs}{l}\right) = \exp\left(i\frac{2k\pi x_s}{l}\right);\quad q_k^N = \exp\left(i\frac{2k\pi x_N}{l}\right) = \exp(2k\pi i) = 1$$

и

$$\sum_{s=1}^{N-1} h\cos\frac{2k\pi x_s}{l} = \mathrm{Re}\sum_{s=1}^{N-1} hq_k^s = \mathrm{Re}\left(h\frac{q_k^N - q_k}{q_k - 1}\right) = \mathrm{Re}\left(h\frac{1-q_k}{q_k-1}\right) = -h\,.$$

Тогда из (7.15) получим:

$$\left\|y^{(k)}\right\|^2 = \frac{h}{2}(N-1+1) = \frac{Nh}{2} = \frac{l}{2},$$

что и требовалось установить.
Следовательно, набор сеточных функций

$$\mu^{(k)}(x) = \sqrt{\frac{2}{l}}y^{(k)}(x) = \sqrt{\frac{2}{l}}\sin\frac{k\pi x}{l},\quad k = \overline{1, N-1} \tag{7.16}$$

образует ортонормированную систему.

5. Пусть на сетке $\overline{\omega}_h$ задана функция $f(x)$, причем $f(0) = f(l) = 0$. Тогда она представима в виде

$$f(x) = \sum_{k=1}^{N-1} f_k\mu^{(k)}(x), \tag{7.17}$$

где $f_k = \left(f(x), \mu^{(k)}(x)\right)$, причем справедливо равенство

$$\|f\|^2 = \sum_{k=1}^{N-1} f_k^2\,. \tag{7.18}$$

Действительно,

$$\|f\|^2 = (f,f) = \left(\sum_{k=1}^{N-1} f_k\mu^{(k)}, \sum_{k=1}^{N-1} f_k\mu^{(k)}\right) = \sum_{k=1}^{N-1} f_k^2\,,$$

так как $\left(\mu^{(k)}, \mu^{(m)}\right) = \delta_k^m$.

7.3. Разностные аналоги теорем вложения

При оценке различных свойств разностных схем часто используются неравенства, связывающие нормы в различных функциональных пространствах, соответствующие простейшим теоремам вложения Соболева.

***Лемма* 1.** Для всякой сеточной функции $y(x)$, заданной на сетке $\overline{\omega}_h$ и обращающейся в нуль при $x = 0$ и $x = 1$, справедливо неравенство

$$\|y\|_C \le \frac{1}{2}\|y_{\bar{x}}\|. \tag{7.19}$$

Доказательство.

На сетке $\overline{\omega}_h$ справедливо тождество

$$y^2(x) = (1-x)y^2(x) + xy^2(x). \tag{7.20}$$

Так как $y(0) = y(1) = 0$, то

$$y^2(x) = \left(\sum_{x'=h}^{x} y_{\bar{x}}(x')h\right)^2 \qquad \text{или} \qquad y^2(x) = \left(\sum_{x'=x+h}^{1} y_{\bar{x}}(x')h\right)^2.$$

Подставляя эти выражения в (7.20), получим:

$$y^2(x) = (1-x)\left(\sum_{x'=h}^{x} y_{\bar{x}}(x')h\right)^2 + x\left(\sum_{x'=x+h}^{1} y_{\bar{x}}(x')h\right)^2.$$

Теперь, используя неравенство Коши-Буняковского, отсюда получим:

$$y^2(x) \le (1-x)\sum_{x'=h}^{x} h \cdot \sum_{x'=h}^{x} h y_{\bar{x}}^2(x') + x\sum_{x'=x+h}^{1} h \cdot \sum_{x'=x+h}^{1} h y_{\bar{x}}^2(x') = x(1-x)\sum_{x'=h}^{1} h y_{\bar{x}}^2(x') = x(1-x)\|y_{\bar{x}}\|^2,$$

а так как $x(1-x) \le \frac{1}{4}$, то $y^2(x) \le \frac{1}{4}\|y_{\bar{x}}\|^2$ или $\|y\|_C \le \frac{1}{2}\|y_{\bar{x}}\|$. ⊠

***Замечания*:**

1. Для $\overline{\omega}_h(0;l)$ неравенство (7.19) следует переписать в виде

$$\|y\|_C \le \frac{\sqrt{l}}{2}\|y_{\bar{x}}\|. \tag{7.19′}$$

2. Если $y(0)\cdot y(l) \ne 0$, то (7.19′), вообще говоря, неверно.
3. Для случая неравномерной сетки (7.19) остается в силе.

Упражнение. Доказать 1. – 3.

Лемма 2. Для всякой функции $y(x)$, заданной на сетке $\overline{\omega}_h(0;l)$ и обращающейся в нуль при $x = 0$ и $x = l$, справедливы оценки

$$\frac{h^2}{4}\|y_{\bar{x}}\|^2 \le \|y\|^2 \le \frac{l^2}{8}\|y_{\bar{x}}\|^2. \tag{7.21}$$

Доказательство.

Разложим $y(x)$ по собственным функциям задачи (7.10):

$$y(x) = \sum_{k=1}^{N-1} C_k \mu^{(k)}(x),$$

причем

$$C_k = \left(y(x), \mu^{(k)}(x)\right), \ \|y\|^2 = \sum_{k=1}^{N-1} C_k^2.$$

В силу первой разностной формулы Грина

$$\|y_{\bar{x}}]\|^2 = (-y_{\bar{x}x}, y),$$

а так как $\mu_{\bar{x}x}^{(k)} = -\lambda_k \mu^{(k)}$, то

$$\|y_{\bar{x}}]\|^2 = \left(\sum_{k=1}^{N-1} C_k \lambda_k \mu^{(k)}, \sum_{k=1}^{N-1} C_k \mu^{(k)}\right) = \sum_{k=1}^{N-1} C_k^2 \lambda_k \ .$$

Отсюда получаем:

$$\lambda_1 \|y\|^2 \le \|y_{\bar{x}}]\|^2 \le \lambda_{N-1} \|y\|^2,$$

где

$$\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l}, \ \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l}.$$

Оценим λ_1 снизу. Пусть $\alpha = \dfrac{\pi h}{2l}$. Тогда

$$\lambda_1 = \frac{\pi^2}{l^2} \left(\frac{\sin \alpha}{\alpha}\right)^2.$$

Так как $h \le \dfrac{l}{2}$, то $\alpha \in \left(0; \dfrac{\pi}{4}\right]$ и, следовательно, учитывая, что функция $f(\alpha) = \dfrac{\sin \alpha}{\alpha}$ монотонно убывает на данном промежутке, получаем:

$$\min_{\alpha \in \left(0; \frac{\pi}{4}\right]} \frac{\sin \alpha}{\alpha} = f\left(\frac{\pi}{4}\right) = \frac{2\sqrt{2}}{\pi}.$$

Поэтому

$$\lambda_1 \ge \frac{\pi^2}{l^2} \cdot \left(\frac{2\sqrt{2}}{\pi}\right)^2 = \frac{8}{l^2}.$$

С другой стороны, $\lambda_{N-1} < \dfrac{4}{h^2}$. Таким образом, получаем оценку

$$\frac{8}{l^2} \|y\|^2 \le \|y_{\bar{x}}]\|^2 \le \frac{4}{h^2} \|y\|^2,$$

откуда непосредственно следует доказываемое неравенство (7.21). ⊠

7.4. Метод энергетических неравенств

Одним из общих и весьма эффективных способов получения априорных оценок является ***метод энергетических неравенств***. Приведем пример использования данного метода для получения априорных оценок применительно к разностным задачам.

Пусть имеем модельную задачу

$$\begin{cases} u''(x) + f(x) = 0, \ 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{7.22}$$

Введем на отрезке $[0;1]$ равномерную сетку $\overline{\omega}_h$ и заменим (7.22) разностной схемой

$$\begin{cases} y_{\bar{x}x} + f(x) = 0, \ x \in \omega_h, \\ y(0) = y(1) = 0. \end{cases} \tag{7.23}$$

Умножив разностное уравнение (7.23) скалярно на искомую функцию y, получим:

$$(y_{\bar{x}x}, y) + (f, y) = 0.$$

Применив к первому слагаемому первую разностную формулу Грина, перепишем полученное равенство в виде

$$\|y_{\bar{x}}]\|^2 = (f, y). \tag{7.24}$$

Согласно неравенству Коши – Буняковского $|(f, y)| \le \|f\| \cdot \|y\|$, а в силу Леммы 2 имеем оценку $\|y\|^2 \le \frac{l^2}{8}\|y_{\bar{x}}]\|^2$ или $\|y\| \le \frac{l}{2\sqrt{2}}\|y_{\bar{x}}]\|$. Поэтому для скалярного произведения получим оценку сверху вида

$$|(f, y)| \le \frac{1}{2\sqrt{2}}\|f\| \cdot \|y_{\bar{x}}]\|.$$

Таким образом, используя Лемму 1, из (7.24) последовательно будем иметь:

$$\|y_{\bar{x}}]\|^2 = (f, y) \le \frac{1}{2\sqrt{2}}\|f\| \cdot \|y_{\bar{x}}]\|,$$

$$2\|y\|_C \le \|y_{\bar{x}}]\| \le \frac{1}{2\sqrt{2}}\|f\|$$

и, наконец,

$$\|y\|_C \le \frac{1}{4\sqrt{2}}\|f\|. \tag{7.25}$$

(7.25) – априорная оценка решения разностной задачи (7.23) через входные данные.

Покажем, как ее можно использовать для оценки скорости сходимости разностной схемы (7.23). Запишем уравнение для погрешности. Если $z = y - u$, то для z имеем задачу

$$\begin{cases} z_{\bar{x}x} + \psi(x) = 0, \ x \in \omega_h, \\ z(0) = z(1) = 0, \end{cases} \tag{7.26}$$

где $\psi(x)$ – погрешность аппроксимации разностной схемы на решении задачи (7.22). Согласно (7.25) можем записать:

$$\|z\|_C \frac{1}{4\sqrt{2}}\|\psi\|,$$

а так как $\psi = O(h^2)$, то, следовательно, $\|z\|_C = \|y - u\|_C \le Mh^2$, т.е. решение разностной задачи (7.23) сходится к решению дифференциальной задачи (7.22) со скоростью $O(h^2)$.

В заключение отметим, что метод энергетических неравенств является достаточно универсальным.

ГЛАВА XVI

Способы построения разностных схем

§ 1. Требования, предъявляемые к разностным схемам

Выше мы приводили примеры разностных аппроксимаций для дифференциальных операторов различных порядков, а также разностных схем для дифференциальных уравнений первого и второго порядка, в том числе с граничными условиями, содержащими производные от искомого решения.

При этом в случае дифференциальных уравнений с переменными коэффициентами задача построения разностной схемы может существенно усложниться. На заданном шаблоне мы можем построить бесчисленное множество разностных схем, эквивалентных по порядку аппроксимации. Так, например, для дифференциального уравнения

$$\frac{d^2u}{dx^2} - q(x)u = -f(x)$$

на трехточечном шаблоне можно построить однопараметрическое семейство разностных аппроксимаций

$$L_h y_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - d_i y_i = -f_i, \quad d_i = \alpha q_{i-1} + (1-2\alpha)q_i + \alpha q_{i+1},$$

имеющих при любом вещественном значении параметра α второй порядок аппроксимации. Аналогичную конструкцию можно использовать и для аппроксимации функции $f(x)$. Кроме того, к коэффициенту d_i (равно как и к f_i) можно без нарушения порядка аппроксимации добавлять слагаемые вида βh^2, где β – произвольное не зависящее от шага h число.

Таким образом, возникает задача выбора разностных схем из множества допустимых схем, заданных на некотором шаблоне и имеющих один и тот же порядок аппроксимации. Для этого необходимо сформулировать требования, которые следует предъявлять к разностным схемам. Интуитивно понятно, что любой приближенный метод должен давать возможность найти численное решение с заданной точностью ε за конечное число действий $Q(\varepsilon)$. Естественно поэтому стремиться минимизировать величину $Q(\varepsilon)$, т.е. найти ***оптимальный*** метод.

При фиксированном методе решения системы объем вычислений зависит от ее порядка. Он тем меньше, чем крупнее шаг сетки. Однако уменьшение числа узлов сетки приводит к уменьшению точности схемы. Поэтому желательно иметь схему с возможно более высоким порядком точности (который зависит от гладкости коэффициентов дифференциального уравнения, начальных и граничных условий). Практически это означает, что надо искать схемы с минимальным шаблоном, имеющие максимально возможный на этом шаблоне порядок аппроксимации.

Итак, количественные требования к семейству разностных схем могут выглядеть следующим образом:

1) определенный порядок аппроксимации;
2) максимальный порядок точности на всем классе решаемых задач;
3) экономичность, т.е. минимум операций при машинной реализации сеточных уравнений.

Необходимо выделить также следующие качественные характеристики разностных схем:

1) схема должна быть однородной, т.е. сеточные уравнения для любой задачи из рассматриваемого класса K и любой сетки в любом узле должны записываться единообразно, по одному и тому же закону;
2) система разностных уравнений должна быть разрешимой на любой допустимой сетке и для любой задачи из рассматриваемого класса K;
3) схема должна быть сходящейся для любой задачи из рассматриваемого класса K.

1.1. Однородные разностные схемы

Однородность разностной схемы означает, что все ее коэффициенты являются функционалами коэффициентов дифференциальной задачи, зависящими от шага сетки h как от параметра и не зависящими от узла сетки и от выбора коэффициентов задачи.

Формальная схема может выглядеть следующим образом. Пусть заданы:

а) целочисленный шаблон $Ш = \{-m_1, -m_1+1, ..., -1, 0, 1, ..., m_2\}$, где $m_1 > 0$ и $m_2 > 0$ – целые числа, на котором определяется сеточная функция $\bar{y}(j),\ j \in Ш$;

б) шаблон $\Sigma = \{-m_1 \le s \le m_2\}$, на котором определена вектор-функция $\bar{k}(s)$ коэффициентов исходной дифференциальной задачи (концы шаблонов $Ш$ и Σ могут и не совпадать).

Обозначим через $A_j^h\left(\bar{k}(s)\right),\ F^h\left(\bar{k}(s)\right),\ j \in Ш,\ s \in \Sigma$, шаблонные функционалы. Рассматривается функционал

$$\Phi^h(\bar{y}(j)) = \sum_{j=-m_1}^{m_2} A_j^h\left(\bar{k}(s)\right)\bar{y}(j) + F^h\left(\bar{k}(s)\right)$$

и от него осуществляется переход к однородной схеме следующим образом: полагая $\bar{y}(j) = y^h(x_i + jh),\ \bar{k}(s) = k(x_i + sh)$ и пользуясь выражением для Φ^h, получаем однородную разностную схему

$$\left(L_h y^h + F^h\right)_i = \sum_{i=-m_1}^{m_2} A_j^h\left(k(x_i + sh)\right) y^h(x_i + jh) + F^h\left(k(x_i + sh)\right) = 0,$$

где $y^h(x_i)$ – сеточная функция, $k(x)$ – вектор-функция непрерывного аргумента.

Семейство однородных схем задано, если заданы шаблонные функционалы $A_j^h\left(\bar{k}(s)\right)$ и $F^h\left(\bar{k}(s)\right),\ j = \overline{-m_1, m_2}$. Произвол в их выборе должен быть ограничен требованиями разрешимости, аппроксимации определенного порядка, экономичности.

Проиллюстрируем понятие однородности на примере трехточечных схем для задачи

$$\begin{cases} \dfrac{d}{dx}\left(k(x)\dfrac{du(x)}{dx}\right) - q(x)u(x) = -f(x),\ \ 0 < x < 1, \\ u(0) = \mu_0,\ u(1) = \mu_1,\ \ k(x) \ge c > 0,\ q(x) \ge 0. \end{cases} \tag{1.1}$$

На сетке $\bar{\omega}_h$ рассмотрим трехточечный шаблон $\{x_{i-1}, x_i, x_{i+1}\}$, так что $Ш = \{-1, 0, 1\}$ и $m_1 = m_2 = 1$. Пусть коэффициентный шаблон имеет вид $\Sigma = \{-1 \le s \le 1\}$ и $A^h\left(\bar{k}(s)\right)$, $B^h\left(\bar{k}(s)\right),\ F^h\left(\bar{k}(s)\right)$ – шаблонные функционалы. Тогда однородная разностная схема будет иметь вид

$$\begin{cases} \dfrac{1}{h}\left(b_i \dfrac{y_{i+1}-y_i}{h} - a_i \dfrac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i, \ i = 1,\ldots,N-1, \\ y_0 = \mu_0, \ y_N = \mu_1, \end{cases} \tag{1.2}$$

причем коэффициенты ее вычисляются во всех узлах $x_i \in \omega_h$ и для любых $k(x), q(x), f(x)$ одинаково:

$$a_i = A^h(k(x_i + sh)), \ b_i = B^h(k(x_i + sh)), \ d_i = F^h(q(x_i + sh)), \ \varphi_i = F^h(f(x_i + sh)).$$

Для простоты здесь каждый из коэффициентов разностного уравнения зависит только от соответствующего коэффициента дифференциального уравнения (причем для d_i и φ_i эта зависимость одинакова). В общем случае A^h, B^h и F^h – нелинейные функционалы, но мы далее будем предполагать их линейными и не зависящими от h.

Найдем сейчас условия, при которых разностная схема (1.2) имеет второй порядок аппроксимации.

Так как

$$\frac{u_{i+1}-u_i}{h} = u_i' + \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + O(h^3),$$

$$\frac{u_i - u_{i-1}}{h} = u_i' - \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + O(h^3),$$

и исходное дифференциальное уравнение, раскрыв скобки в его левой части, можно переписать в виде $ku'' + k'u' - qu + f = 0$, то

$$\psi_i = \frac{1}{h}\left(b_i \frac{u_{i+1}-u_i}{h} - a_i \frac{u_i - u_{i-1}}{h} \right) - d_i u_i + \varphi_i - (k_i u_i'' + k_i' u_i' - q_i u_i + f_i) =$$

$$= \left(\frac{b_i + a_i}{2} - k_i \right) u_i'' + \left(\frac{b_i - a_i}{h} - k_i' \right) u_i' - (d_i - q_i) u_i + \varphi_i - f_i + O(h^2).$$

Таким образом, схема (1.2) будет иметь второй порядок аппроксимации, если

$$\begin{cases} \dfrac{b_i + a_i}{2} = k_i + O(h^2); \\ \dfrac{b_i - a_i}{h} = k_i' + O(h^2); \\ d_i = q_i + O(h^2); \\ \varphi_i = f_i + O(h^2) \end{cases} \tag{1.3}$$

Воспользовавшись разложениями

$$k(x+sh)=k(x)+shk'(x)+\frac{s^2h^2}{2}k''(x)+O(h^3),$$

$$f(x+sh)=f(x)+shf'(x)+O(h^2);$$

$$q(x+sh)=q(x)+shq'(x)+O(h^2),$$

получим:

$$a_i=A(k(x_i+sh))=A\left[k_i+shk_i'+\frac{s^2h^2}{2}k_i''+O(h^3)\right]=$$

$$=A(1)k_i+hk_i'A(s)+\frac{h^2}{2}k_i''A(s^2)+O(h^3);$$

$$b_i=B(k(x_i+sh))=B\left[k_i+shk_i'+\frac{s^2h^2}{2}k_i''+O(h^3)\right]=$$

$$=B(1)k_i+hk_i'B(s)+\frac{h^2}{2}k_i''B(s^2)+O(h^3);$$

$$d_i=F(q(x_i+sh))=F\left[q_i+shq_i'+O(h^2)\right]=F(1)q_i+hq_i'F(s)+O(h^2);$$

$$\varphi_i=F(f(x_i+sh))=F\left[f_i+shf_i'+O(h^2)\right]=F(1)f_i+hf_i'F(s)+O(h^2).$$

Тогда условия (1.3) перепишутся в виде

$$\begin{cases}\dfrac{A(1)+B(1)}{2}k_i+hk_i'\dfrac{A(s)+B(s)}{2}+O(h^2)=k_i+O(h^2);\\ \dfrac{B(1)-B(1)}{h}k_i+hk_i'\dfrac{B(s)-A(s)}{h}+\dfrac{h^2}{2}k_i''\dfrac{B(s^2)-A(s^2)}{h}+O(h^2)=k_i'+O(h^2);\\ F(1)q_i+hq_i'F(s)+O(h^2)=q_i+O(h^2);\\ F(1)f_i+hf_i'F(s)+O(h^2)=f_i+O(h^2).\end{cases}$$

Отсюда, приравнивая коэффициенты при одинаковых степенях h, получаем:

$$\begin{cases}\dfrac{A(1)+B(1)}{2}=1;\\ \dfrac{B(1)-A(1)}{h}=0;\\ \dfrac{A(s)+B(s)}{2}=0;\\ B(s)-A(s)=1;\\ F(1)=1;\\ F(s)=0;\\ \dfrac{B(s^2)-A(s^2)}{2}=0;\end{cases}\qquad (*)$$

Решая данную систему, находим:

$$\begin{cases} A(1) = B(1) = F(1) = 1; \\ B(s) = \dfrac{1}{2}; \\ A(s) = -\dfrac{1}{2}; \\ F(s) = 0; \\ B(s^2) = A(s^2) \end{cases} \tag{1.4}$$

Требование разрешимости системы разностных уравнений (1.2) будет выполнено, если $a_i > 0$, $b_i > 0$, $a_i + b_i + h^2 d_i \geq a_i + b_i$ (это ведь не что иное как условие применимости метода разностной прогонки). Для выполнения этих условий достаточно потребовать, чтобы функционалы A, B и F были положительны. Экономичность разностной схемы гарантируется алгоритмом разностной прогонки.

В простейшем случае A, B и F представляют собой линейные комбинации значений функций $\bar{k}(s)$ и $\bar{f}(s)$ в конечном числе точек на шаблоне Σ, например

$$\begin{cases} A\left(\bar{k}(s)\right) = \alpha_{-1}\bar{k}(-1) + \alpha_0\bar{k}(0) + \alpha_1\bar{k}(1), \\ B\left(\bar{k}(s)\right) = \beta_{-1}\bar{k}(-1) + \beta_0\bar{k}(0) + \beta_1\bar{k}(1), \\ F\left(\bar{k}(s)\right) = \gamma_{-1}\bar{k}(-1) + \gamma_0\bar{k}(0) + \gamma_1\bar{k}(1), \end{cases}$$

так что

$$\begin{cases} a_i = \alpha_{-1}k_{i-1} + \alpha_0 k_i + \alpha_1 k_{i+1}, \\ b_i = \beta_{-1}k_{i-1} + \beta_0 k_i + \beta_1 k_{i+1}, \\ d_i = \gamma_{-1}q_{i-1} + \gamma_0 q_i + \gamma_1 q_{i+1}, \\ \varphi_i = \gamma_{-1}f_{i-1} + \gamma_0 f_i + \gamma_1 f_{i+1}. \end{cases}$$

В этом случае условия второго порядка аппроксимации (1.4) могут быть переписаны в виде системы соотношений, связывающих коэффициенты α_i, β_i, γ_i:

$$\begin{cases} \alpha_{-1} + \alpha_0 + \alpha_1 = 1, \\ \beta_{-1} + \beta_0 + \beta_1 = 1, \\ \gamma_{-1} + \gamma_0 + \gamma_1 = 1, \\ -\gamma_{-1} + \gamma_1 = 0, \\ \\ -\beta_{-1} + \beta_1 = \dfrac{1}{2}, \\ -\alpha_{-1} + \alpha_1 = -\dfrac{1}{2}, \\ \beta_{-1} + \beta_1 = \alpha_{-1} + \alpha_1. \end{cases}$$

Таким образом, при обсуждаемом способе задания шаблонных функционалов существует двухпараметрическое семейство трехточечных однородных разностных схем второго порядка.

Замечание. Чтобы разностная схема (1.3) имела ***первый*** порядок аппроксимации (т.е. вообще аппроксимировала задачу) в рассмотренной выше системе (*) достаточно оставить только четыре уравнения

$$\begin{cases} \dfrac{A(1)+B(1)}{2}=1, \\ \dfrac{B(1)-A(1)}{h}=0, \\ B(s)-A(s)=1, \\ F(1)=1, \end{cases}$$

откуда следует:

$$\begin{cases} A(1)=B(1)=F(1)=1, \\ B(s)-A(s)=1. \end{cases} \tag{1.5}$$

1.2. Консервативные разностные схемы

Помимо формальных требований разрешимости, аппроксимации, экономичности необходимо обеспечить также сходимость. Вообще говоря, это следует из аппроксимации и устойчивости (*теорема Лакса*). Однако в реальном вычислительном процессе шаг сетки не должен быть слишком малым.

Чтобы получить хорошее приближение на реальных сетках, необходимо, как показывает практика, пользоваться схемами, хорошо отражающими основные свойства дифференциальных уравнений.

Уравнения же математической физики выражают, как правило, законы сохранения в дифференциальной форме. Так, например, уравнение

$$\frac{d}{dx}\left(k(x)\frac{du}{dx}\right)=-f(x),\ 0<x<1 \tag{1.6}$$

можно трактовать как уравнение стационарного распределения температуры $u(x)$ в стержне $0<x<1$ с коэффициентом теплопроводности $k(x)$. Интегрируя это уравнение по переменной x от $x^{(1)}$ до $x^{(2)}$, получим закон сохранения тепла на отрезке $x^{(1)}\le x\le x^{(2)}$:

$$W\left(x^{(2)}\right)-W\left(x^{(1)}\right)=\int_{x^{(1)}}^{x^{(2)}} f(x)dx,\quad W(x)=-k(x)\frac{du}{dx}. \tag{1.7}$$

Слева в равенстве (1.7) стоит разность тепловых потоков на концах отрезка, справа – количество выделившегося (поглотившегося) тепла.

Разностные схемы, которые выражают законы сохранения на сетке, называют ***консервативными*** разностными схемами.

Поясним смысл консервативности на примере разностной схемы (1.2) для задачи (1.6). Так как $q(x)=0$, то (1.2) имеет вид

$$\frac{1}{h}\left(b_i\frac{y_{i+1}-y_i}{h}-a_i\frac{y_i-y_{i-1}}{h}\right)=-\varphi_i$$

или

$$\frac{1}{h}\left(a_{i+1}\frac{y_{i+1}-y_i}{h}-a_i\frac{y_i-y_{i-1}}{h}\right)=-\varphi_i-\frac{b_i-a_{i+1}}{h}\cdot\frac{y_{i+1}-y_i}{h}.$$

Просуммировав по сетке от i_1 до i_2 $\left(x^{(1)}=i_1h;\ x^{(2)}=i_2h\right)$, получим:

$$W^h_{i_2+1}-W^h_{i_1}=\sum_{i=i_1}^{i_2}h\varphi_i+\sum_{i=i_1}^{i_2}(b_i-a_{i+1})\frac{y_{i+1}-y_i}{h}, \tag{1.8}$$

где $W^h_i=-a_i\frac{y_i-y_{i-1}}{h}$.

(1.8) – сеточный аналог закона сохранения (1.7).

В правую часть равенства (1.7) входит величина дисбаланса

$$D=\sum_{i=i_1}^{i_2}(b_i-a_{i+1})\frac{y_{i+1}-y_i}{h},$$

которая обратится в нуль для любых сеточных функций y_i только при условии

$$b_i=a_{i+1}. \tag{1.9}$$

Условие (1.9) – необходимое и достаточное условие консервативности разностной схемы (в данном случае при условии $y_0=y_N=0$ оно совпадает с условием самосопряженности оператора Λy).

Консервативность является важным свойством. Покажем, что консервативность является необходимым условием сходимости разностной схемы в случае простейшей задачи

$$\begin{cases}\dfrac{d}{dx}\left(k(x)\dfrac{du}{dx}\right)=0,\ 0<x<1,\\ u(0)=1,\ u(1)=0\end{cases} \tag{1.10}$$

в классе кусочно-постоянных коэффициентов:

$$k(x)=\begin{cases}k_1,\ \text{если}\ 0<x<\xi,\\ k_2,\ \text{если}\ \xi<x<1,\end{cases} \tag{1.11}$$

где ξ – иррациональное число: $\xi=x_n+\theta h,\ 0<\theta<1$.

Как известно, точное решение такой задачи в точках разрыва коэффициентов удовлетворяет условиям сопряжения (непрерывности температуры и теплового потока):

$$\begin{cases}[u]=u(\xi+0)-u(\xi-0)=0,\\ [ku']=k_2u'(\xi+0)-k_1u'(\xi-0)=0.\end{cases}$$

Исследование начнем с нахождения аналитического решения поставленной задачи.

Поскольку на промежутке $[0;\xi)$ уравнение имеет вид $u''(x)=0$, то $u(x)=C_0+C_1x$, $x\in[0;\xi)$, а так как при этом $u(0)=1$, то $C_0=1$. Таким образом, $u(x)=1-\gamma_0x,\ x\in[0;\xi)$.

Аналогично при $x\in(\xi;1]$ также $u''(x)=0$, откуда $u(x)=C_2+C_3x,\ x\in(\xi;1]$ и в силу граничного условия на правом конце отрезка $C_2+C_3=0$, т.е. $u(x)=\delta_0(1-x),\ x\in(\xi;1]$.

Таким образом,

$$u(x)=\begin{cases}1-\gamma_0 x,\ 0\le x<\xi,\\ \delta_0(1-x),\ \xi<x\le 1.\end{cases}$$

В точке $x=\xi$ функцию $u(x)$ доопределим ее предельным значением, которое, в силу первого из условий сопряжения, существует. При этом само условие примет вид

$$1-\gamma_0\xi=\delta_0(1-\xi).$$

Аналогично второе условие сопряжения, учитывая значения коэффициента теплопроводности, может быть переписано в виде

$$-k_1\gamma_0=-k_2\delta_0.$$

Следовательно, получаем систему из двух уравнений для определения параметров γ_0 и δ_0

$$\begin{cases}1-\gamma_0\xi=\delta_0(1-\xi),\\ k_1\gamma_0=k_2\delta_0.\end{cases}$$

Из второго уравнения этой системы получаем: $\delta_0=\dfrac{k_1}{k_2}\gamma_0$. Подставляя это выражение в первое уравнение, находим:

$$\gamma_0\left[\frac{k_1}{k_2}(1-\xi)+\xi\right]=1.$$

Отсюда

$$\gamma_0=\frac{1}{\Delta_0},$$

где

$$\Delta_0=\frac{k_1}{k_2}(1-\xi)+\xi.$$

Применим теперь для решения задачи (1.10), (1.11) однородную разностную схему вида (1.2):

$$\begin{cases}\dfrac{1}{h}\left(b_i\dfrac{y_{i+1}-y_i}{h}-a_i\dfrac{y_i-y_{i-1}}{h}\right)=0,\ i=\overline{1,N-1},\\ y_0=1,\\ y_N=0.\end{cases}\qquad(1.12)$$

Здесь, как и выше, $b_i=B(k(x_i+sh)),\ a_i=A(k(x_i+sh))$.

Так как рассматриваемая разностная схема обладает аппроксимацией (по крайней мере, первого порядка), то должны выполняться условия (1.5). Следовательно, на всех отрезках длины $2h$, не содержащих точки разрыва коэффициента теплопроводности, коэффициенты схемы a_i, b_i будут постоянны ($A(k_1)=k_1\cdot A(1)=k_1$ и т.п.), т.е. справедливы равенства

$$\begin{cases}a_i=b_i=k_1,\ 0<i<n,\\ a_i=b_i=k_2,\ n+1<i<N.\end{cases}$$

Таким образом, рассматриваемое разностное уравнение при $i=\overline{1,n-1}$ и $i=\overline{n+2,N-1}$ может быть переписано в виде

$$y_{i+1}-2y_i+y_{i-1}=0$$

и его решение, учитывающее граничные условия по аналогии с точным может быть определено в виде

$$y_i=\begin{cases}1-\gamma x_i,\ 0\le i\le n,\\ \delta(1-x_i),\ n+1\le i\le N.\end{cases} \tag{1.13}$$

Запишем теперь разностное уравнение (1.12) при двух оставшихся значениях индекса i: при $i=n$ и $i=n+1$:

$$\begin{cases}b_n(y_{n+1}-y_n)-a_n(y_n-y_{n-1})=0,\\ b_{n+1}(y_{n+2}-y_{n+1})-a_{n+1}(y_{n+1}-y_n)=0.\end{cases} \tag{1.14}$$

Умножая первое из этих уравнений на a_{n+1}, а второе – на b_n и складывая, получим:

$$b_{n+1}b_n(y_{n+2}-y_{n+1})=a_na_{n+1}(y_n-y_{n-1}). \tag{1.15}$$

Поскольку из (1.13) следует, что

$$\begin{aligned}y_n-y_{n-1}&=(1-\gamma x_n)-(1-\gamma x_{n-1})=-\gamma(x_n-x_{n-1})=-\gamma h,\\ y_{n+2}-y_{n+1}&=\delta(1-x_{n+2})-\delta(1-x_{n+1})=-\delta(x_{n+2}-x_{n+1})=-\delta h,\end{aligned} \tag{1.16}$$

то из (1.15) находим:

$$b_{n+1}b_n(-\delta h)=a_na_{n+1}(-\gamma h),$$

т.е.

$$\delta=\frac{a_na_{n+1}}{b_nb_{n+1}}\gamma. \tag{1.17}$$

Подставляя найденную связь, например, в первое из уравнений системы (1.14) и учитывая, что

$$y_{n+1}-y_n=\delta(1-x_{n+1})-(1-\gamma x_n),$$

будем иметь:

$$b_n\left(\gamma\frac{a_na_{n+1}}{b_nb_{n+1}}(1-x_{n+1})-1+\gamma x_n\right)+a_n\gamma h=0,$$

или

$$\gamma\left[\frac{a_na_{n+1}}{b_nb_{n+1}}(1-x_{n+1})+x_n+\frac{a_n}{b_n}h\right]=1,$$

т.е.

$$\gamma=\frac{1}{\dfrac{a_n}{b_n}h+x_n+\dfrac{a_na_{n+1}}{b_nb_{n+1}}(1-x_{n+1})}. \tag{1.18}$$

Таким образом, разностное решение в узлах сетки определяется однозначно формулами (1.13), (1.17), (1.18). Распространим его на весь отрезок $[0;1]$ путем линейной интерполяции. Применим для этих целей, например, интерполяционный многочлен Ньютона:

$$\widetilde{y}(x;h) = y_i + \frac{x - x_i}{h}(y_{i+1} - y_i),\ x \in [x_i; x_{i+1}].$$

Тогда, используя формулы (1.13) и (1.16), при всех $i \le n-1$ будем иметь

$$\widetilde{y}(x;h) = 1 - \gamma x_i + \frac{x - x_i}{h}(-\gamma h) = 1 - \gamma x,$$

а при $i \ge n+1$ –

$$\widetilde{y}(x;h) = \delta(1 - x_i) + \frac{x - x_i}{h}(-\delta h) = \delta(1 - x).$$

На отрезке $[x_n; x_{n+1}]$ доопределение выполним естественным образом: слева от точки ξ – как для $i \le n-1$, а справа – как для $i \ge n+1$. Таким образом,

$$\widetilde{y}(x;h) = \begin{cases} 1 - \gamma x,\ 0 \le x \le \xi, \\ \delta(1 - x),\ \xi \le x \le 1. \end{cases}$$

Сходимость найденного приближенного решения к точному решению $u(x)$ задачи, т.е. выполнение соотношения $\widetilde{\gamma}(x;h) \underset{h \to 0}{\to} u(x)$, равносильна, учитывая вид найденного выше точного решения, выполнению соотношений

$$\gamma \underset{h \to 0}{\to} \gamma_0,\ \ \delta \underset{h \to 0}{\to} \delta_0.$$

Последние же соотношения с учетом вида всех входящих в них констант, выполняются при условии

$$\frac{a_n a_{n+1}}{b_n b_{n+1}} \underset{h \to 0}{\to} \frac{k_1}{k_2}$$

или

$$R_n = \frac{b_n b_{n+1}}{k_2} - \frac{a_n a_{n+1}}{k_1} \underset{h \to 0}{\to} 0. \qquad (1.19)$$

Чтобы проверить полученное условие сходимости, зададим конкретный вид шаблонных функционалов $A(k(x_i + sh))$ и $B(k(x_i + sh))$, ибо входящие в него коэффициенты разностного уравнения только лишь приведенными выше условиями аппроксимации не определяются. Пусть, например, это будут рассмотренные ранее линейные комбинации значений функции в трех соседних узлах сетки, т.е.

$$\begin{cases} a_i = \alpha_{-1} k_{i-1} + \alpha_0 k_i + \alpha_1 k_{i+1}, \\ b_i = \beta_{-1} k_{i-1} + \beta_0 k_i + \beta_1 k_{i+1}. \end{cases}$$

Тогда обсуждавшиеся выше условия аппроксимации (1.5) примут вид

$$\begin{cases} \alpha_{-1} + \alpha_0 + \alpha_1 = 1, \\ \beta_{-1} + \beta_0 + \beta_1 = 1, \\ \beta_1 - \beta_{-1} = 1 + \alpha_1 - \alpha_{-1}. \end{cases} \qquad (1.20)$$

При этом из соображений разрешимости, как отмечалось ранее, все коэффициенты α_i и β_i должны быть неотрицательными.

При таком выборе функционалов получим:

$$\begin{cases} a_i = k_1,\ i = \overline{1, n-1}, \\ a_i = k_2,\ i = \overline{n+2, N-1}, \\ a_n = \alpha_{-1}k_1 + \alpha_0 k_1 + \alpha_1 k_2 = (\alpha_{-1} + \alpha_0)k_1 + \alpha_1 k_2 = (1-\alpha_1)k_1 + \alpha_1 k_2, \\ a_{n+1} = \alpha_{-1}k_1 + \alpha_0 k_2 + \alpha_1 k_2 = \alpha_{-1}k_1 + (\alpha_0 + \alpha_1)k_2 = \alpha_{-1}k_1 + (1-\alpha_{-1})k_2, \end{cases}$$

$$\begin{cases} b_i = k_1,\ i = \overline{1, n-1}, \\ b_i = k_2,\ i = \overline{n+2, N-1}, \\ b_n = \beta_{-1}k_1 + \beta_0 k_1 + \beta_1 k_2 = (\beta_{-1} + \beta_0)k_1 + \beta_1 k_2 = (1-\beta_1)k_1 + \beta_1 k_2, \\ b_{n+1} = \beta_{-1}k_1 + \beta_0 k_2 + \beta_1 k_2 = \beta_{-1}k_1 + (\beta_0 + \beta_1)k_2 = \beta_{-1}k_1 + (1-\beta_{-1})k_2. \end{cases}$$

Отсюда видим, что поскольку коэффициенты $a_n, a_{n+1}, b_n, b_{n+1}$ не зависят от шага h, то полученное выше условие сходимости (1.19) равносильно равенству

$$R_n = \frac{b_n b_{n+1}}{k_2} - \frac{a_n a_{n+1}}{k_1} = 0.$$

Подставляя сюда полученные выражения коэффициентов $a_n, a_{n+1}, b_n, b_{n+1}$, будем иметь:

$$\frac{1}{k_2}[(1-\beta_1)k_1 + \beta_1 k_2][\beta_{-1}k_1 + (1-\beta_{-1})k_2] - \frac{1}{k_1}[(1-\alpha_1)k_1 + \alpha_1 k_2][\alpha_{-1}k_1 + (1-\alpha_{-1})k_2] =$$

$$= \left[\text{введем обозначение } t = \frac{k_1}{k_2}\right] =$$

$$= \frac{k_2^2}{k_2}[(1-\beta_1)t + \beta_1][\beta_{-1}t + (1-\beta_{-1})] - \frac{k_2^2}{k_1}[(1-\alpha_1)t + \alpha_1][\alpha_{-1}t + (1-\alpha_{-1})] =$$

$$= \frac{k_2^2}{k_1}\{t[(1-\beta_1)t + \beta_1][\beta_{-1}t + (1-\beta_{-1})] - [(1-\alpha_1)t + \alpha_1][\alpha_{-1}t + (1-\alpha_{-1})]\} = 0.$$

Раскрывая скобки и отбрасывая отличный от нуля коэффициент $\frac{k_2^2}{k_1}$, рассмотрим это равенство как многочлен по переменной t:

$$\beta_{-1}(1-\beta_1)t^3 + [\beta_1\beta_{-1} + (1-\beta_1)(1-\beta_{-1}) - \alpha_{-1}(1\alpha_1)]t^2 +$$

$$+ [\beta_1(1-\beta_{-1}) - (1-\alpha_1)(1-\alpha_{-1}) - \alpha_1\alpha_{-1}]t - \alpha_1(1-\alpha_{-1}) = 0.$$

Требуя тождественного по t равенства нулю, приравниваем нулю коэффициенты при всех степенях t. В итоге получим систему уравнений, связывающую параметры α и β:

$$\begin{cases} \beta_{-1}(1-\beta_1)=0, \\ \alpha_1(1-\alpha_{-1})=0, \\ \beta_1\beta_{-1}+(1-\beta_1)(1-\beta_{-1})-\alpha_{-1}(1-\alpha_1)=0, \\ \beta_1(1-\beta_{-1})-(1-\alpha_1)(1-\alpha_{-1})-\alpha_1\alpha_{-1}=0. \end{cases} \tag{1.21}$$

Исследуем решения этой системы. Из первого уравнения имеем: либо $\beta_{-1}=0$, либо $\beta_1=1$. В первом случае вновь возможны две версии:

а) $\alpha_{-1}=1$. Тогда оставшиеся уравнения системы перепишутся в виде

$$\begin{cases} 1-\beta_1-1+\alpha_1=0, \\ \beta_1-\alpha_1=0, \end{cases}$$

откуда $\beta_1=\alpha_1$;

б) $\alpha_1=0$. Аналогично предыдущему случаю имеем: оставшиеся уравнения отличаются только знаком и, следовательно, $\beta_1=1-\alpha_{-1}$;

Во втором случае также имеем две версии:

в) $\alpha_{-1}=1$. Тогда оставшиеся уравнения системы перепишутся в виде

$$\begin{cases} \beta_{-1}-1+\alpha_1=0, \\ 1-\beta_{-1}-\alpha_1=0, \end{cases}$$

откуда $\beta_{-1}=1-\alpha_1$;

г) $\alpha_1=0$. Аналогично предыдущему случаю имеем: оставшиеся уравнения отличаются только знаком и, следовательно, $\beta_{-1}=\alpha_{-1}$.

Вспомним теперь, что помимо системы (1.21) коэффициенты α_i и β_i должны также удовлетворять системе условий порядка (1.20), а также условию положительности соответствующих функционалов.

В случае а) система (1.20) примет вид

$$\begin{cases} 1+\alpha_0+\alpha_1=1, \\ 0+\beta_0+\beta_1=1, \\ \beta_1-0=1+\alpha_1-1, \end{cases}$$

откуда, учитывая, что $\alpha_i\ge 0$, имеем: $\alpha_0=\alpha_1=0$, $\beta_1=0$, $\beta_0=1$. Таким образом, в этом варианте

$$a_i=k_{i-1},\ b_i=k_i=a_{i+1}.$$

В случае б) аналогично имеем:

$$\begin{cases} \alpha_{-1}+\alpha_0+0=1, \\ 0+\beta_0+\beta_1=1, \\ \beta_1-0=1+0-\alpha_{-1}, \end{cases}$$

откуда $\alpha_0=1-\alpha_{-1}$, $\beta_0=\alpha_{-1}$, $\beta_1=1-\alpha_{-1}$, т.е.

$$a_i=\alpha_{-1}k_{i-1}+(1-\alpha_{-1})k_i,\ b_i=\alpha_{-1}k_i+(1-\alpha_{-1})k_{i+1}=a_{i+1}.$$

В случае в):

$$\begin{cases}\alpha_{-1}+\alpha_0+0=1,\\ \beta_{-1}+\beta_0+1=1,\\ 1-\beta_{-1}=1+0-\alpha_{-1},\end{cases}$$

откуда $\alpha_{-1}=0,\ \alpha_0=1,\ \beta_{-1}=\beta_0=0$, т.е.

$$a_i=k_i,\ b_i=k_{i+1}=a_{i+1}.$$

Наконец, в случае г):

$$\begin{cases}1+\alpha_0+\alpha_1=1,\\ \beta_{-1}+\beta_0+1=1,\\ 1-\beta_{-1}=1+\alpha_1-1,\end{cases}$$

откуда $\beta_0=\beta_{-1}=0,\ \alpha_1=1,\ \alpha_0=-1$. Полученное решение противоречит требованию положительности функционала $A(k(x_i+sh))$.

Окончательно получаем, что во всех случаях коэффициенты разностной схемы удовлетворяют условию консервативности (1.9).

В то же время простейшая разностная схема ***второго*** порядка, полученная путем раскрытия скобок в уравнении (1.10) с последующей заменой производных разностными отношениями, условию сходимости (1.19) не удовлетворяет.

Действительно, переходя к уравнению

$$ku''+k'u'=0$$

и выполняя замену производных, получим разностную схему

$$k_i\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}+\frac{k_{i+1}-k_{i-1}}{2h}\cdot\frac{y_{i+1}-y_{i-1}}{2h}=0 \qquad (1.22)$$

или

$$\frac{1}{h}\left(b_i\frac{y_{i+1}-y_i}{h}-a_i\frac{y_i-y_{i-1}}{h}\right)=0,$$

где

$$a_i=k_i-\frac{k_{i+1}-k_{i-1}}{4},\ b_i=k_i+\frac{k_{i+1}-k_{i-1}}{4}.$$

Отсюда, во-первых, следует, что такие коэффициенты не всегда будут положительны. В самом деле, так как $a_n=k_1-\frac{k_2-k_1}{4}=\frac{5k_1-k_2}{4}$, то отсюда следует ограничение $\frac{k_1}{k_2}>\frac{1}{5}$. С другой стороны, $b_{n+1}=k_2+\frac{k_2-k_1}{4}=\frac{5k_2-k_1}{4}$, откуда $\frac{k_1}{k_2}<5$. Таким образом, разрешимость разностной схемы (1.22) гарантирована только при $\frac{k_1}{k_2}\in\left(\frac{1}{5};5\right)$.

Проверим теперь выполнение условия сходимости (1.19):

$$R_n=\frac{b_nb_{n+1}}{k_2}-\frac{a_na_{n+1}}{k_1}=\frac{(3k_1+k_2)(5k_2-k_1)}{16k_2}-\frac{(5k_1-k_2)(3k_2+k_1)}{16k_1}=\frac{3(k_2-k_1)^3}{16k_1k_2}=0.$$

Последнее же равенство возможно только лишь в случае $k_1 = k_2$.

Таким образом, консервативность разностной схемы является необходимым условием ее сходимости в классе кусочно-непрерывных коэффициентов.

Замечание. Можно показать, что это – также и ***достаточное*** условие.

§ 2. Интегро-интерполяционный метод (метод баланса) построения разностных схем

Рассмотрим сейчас один из способов построения разностных схем для дифференциального уравнения с переменными коэффициентами, который позволяет автоматически удовлетворить требованию консервативности при наличии у исходного дифференциального оператора требования самосопряженности. Ранее мы выяснили, что соответствующее дифференциальное уравнение второго порядка должно иметь вид

$$Lu(x) \equiv (k(x)u'(x))' - q(x)u(x) = -f(x),\ 0 < x < 1. \tag{2.1}$$

Будем рассматривать уравнение (2.1) как уравнение стационарного распределения тепла в стержне. Для него справедлив закон сохранения тепла (уравнение баланса), который на отрезке $\left[x^{(1)}; x^{(2)}\right]$ имеет вид

$$W\left(x^{(1)}\right) - W\left(x^{(2)}\right) - \int_{x^{(1)}}^{x^{(2)}} q(x)u(x)dx + \int_{x^{(1)}}^{x^{(2)}} f(x)dx = 0. \tag{2.2}$$

Это уравнение, очевидно, может быть получено путем интегрирования исходного дифференциального уравнения (2.1) по указанному отрезку. Здесь $W(x) = -k(x)\dfrac{du(x)}{dx}$ – тепловой поток, $k(x) > 0$ – коэффициент теплопроводности, $u(x)$ – температура.

Воспользуемся уравнением (2.2) для написания разностной схемы, аппроксимирующей дифференциальное уравнение (2.1). Пусть на отрезке $[0;1]$ задана сетка $\overline{\omega_h}$ и $x_{i-0.5} = x_i - 0.5h$, $x_{i+0.5} = x_i + 0.5h$. Запишем уравнение баланса (2.2) для отрезка $\left[x_{i-0.5}; x_{i+0.5}\right]$:

$$W_{i-0.5} - W_{i+0.5} - \int_{x_{i-0.5}}^{x_{i+0.5}} q(x)u(x)dx + \int_{x_{i-0.5}}^{x_{i+0.5}} f(x)dx = 0. \tag{2.3}$$

Чтобы построить разностную схему, аппроксимируем тепловой поток и первый из интегралов в (2.3). Заменим функцию $u(x)$ на отрезке $\left[x_{i-0.5}; x_{i+0.5}\right]$ интерполяционным многочленом нулевой степени: $u(x) \approx P_0(x) = u(x_i) =: u_i$, $x \in [x_{i-0.5}; x_{i+0.5}]$. Тогда

$$\int_{x_i-0.5}^{x_{i+0.5}} q(x)u(x)dx \approx hd_iu_i,\ \text{где}\ d_i = \frac{1}{h}\int_{x_{i-0.5}}^{x_{i+0.5}} q(x)u(x)dx,\ i = \overline{1, N-1}. \tag{2.4}$$

Теперь займемся тепловым потоком. Выразив производную от температуры:

$$\frac{du(x)}{dx} = -\frac{W(x)}{k(x)},$$

проинтегрируем последнее равенство по отрезку $[x_{i-1}; x_i]$:

$$u_{i-1} - u_i = \int_{x_{i-1}}^{x_i} \frac{W(x)}{k(x)} dx \,. \tag{2.5}$$

Так как в интегральное соотношение (2.3) тепловой поток входит в полуцелых точках, то, по аналогии с проделанной выше операцией по интерполированию температуры проинтерполируем $W(x)$ с помощью многочлена нулевой степени: $W(x) \approx P_0^*(x) = W_{i-0.5}$ при $x \in [x_{i-1}; x_i]$. Тогда из (2.5) получим:

$$u_i - u_{i-1} \approx -W_{i-0.5} \cdot \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}$$

или

$$W_{i-0.5} \approx -a_i \frac{u_i - u_{i-1}}{h} = -a_i u_{\bar{x},i}, \tag{2.6}$$

где

$$a_i = \left[\frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right]^{-1}, \; i = \overline{1, N} \,. \tag{2.7}$$

Подставляя в (2.3) выражения (2.4) и (2.6), получим:

$$a_{i+1} u_{\bar{x},i+1} - a_i u_{\bar{x},i} - h d_i u_i + h \varphi_i \approx 0 \,,$$

где

$$\varphi_i = \frac{1}{h} \int_{x_{i-0.5}}^{x_{i+0.5}} f(x) dx \,. \tag{2.8}$$

Разделив полученное приближенное равенство на h и переходя к точному равенству для приближенных значений решения в узлах сетки, получим разностную схему:

$$\frac{1}{h} \left(a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i \,, \; i = \overline{1, N-1}$$

или

$$\left(a y_{\bar{x}} \right)_x - d y = -\varphi, \; x \in \omega_h \,, \tag{2.9}$$

коэффициенты которой вычисляются по формулам (2.4), (2.7), (2.8).

Займемся теперь аппроксимацией граничных условий третьего рода для уравнения (2.1). Пусть это условие задано в точке $x = 0$ и имеет вид

$$u'(0) = \widetilde{\kappa}_0 u(0) - \widetilde{g}_0 \,.$$

По предположению коэффициент $k(x)$ в (2.1) больше нуля. Поэтому перепишем данное условие в несколько более удобном виде

$$k(0) u'(0) = \kappa_0 u(0) - g_0 \,. \tag{2.10}$$

Запись граничного условия в виде (2.10) более естественна, так как теперь в левой его части с точностью до знака стоит величина теплового потока. Для аппроксимации полученного условия запишем уравнение баланса (2.2) на отрезке $\left[0; \frac{h}{2} \right]$:

$$W_0 - W_{0.5} - \int_0^{\frac{h}{2}} q(x)u(x)dx + \int_0^{\frac{h}{2}} f(x)dx = 0\,. \qquad (2.11)$$

Теперь из (2.10) имеем: $W_0 = -k(0)u'(0) = -\kappa_0 u(0) + g_0$. Для аппроксимации $W_{0.5}$ воспользуемся формулой (2.6) при $i = 1$, т.е. $W_{0.5} \approx -a_1 u_{\bar{x},1} = -a_1 u_{x,0}$, а также проинтерполируем функцию $u(x)$, стоящую под знаком первого интеграла, с помощью многочлена нулевой степени: $u(x) \approx P_0(x) = u_0$ при $x \in \left[0; \frac{h}{2}\right]$. С учетом сказанного получим разностную аппроксимацию граничного условия (2.10):

$$a_1 y_{x,0} = \left(\kappa_0 + \frac{h}{2} d_0\right) y_0 - \left(g_0 + \frac{h}{2}\varphi_0\right), \qquad (2.12)$$

где

$$d_0 = \frac{2}{h}\int_0^{\frac{h}{2}} q(x)dx\,,\ \varphi_0 = \frac{2}{h}\int_0^{\frac{h}{2}} f(x)dx\,. \qquad (2.13)$$

Аналогичным образом, записав граничное условие на правом конце отрезка в виде

$$-k(1)u'(1) = \kappa_1 u(1) - g_1 \qquad (2.14)$$

и используя уравнение баланса (2.2), записанное для отрезка $\left[1 - \frac{h}{2}; 1\right]$, получим его разностную аппроксимацию в виде

$$-a_N y_{\bar{x},N} = \left(\kappa_1 + \frac{h}{2} d_N\right) y_N - \left(g_1 + \frac{h}{2}\varphi_N\right), \qquad (2.15)$$

где

$$d_N = \frac{2}{h}\int_{1-\frac{h}{2}}^{1} q(x)dx\,,\ \varphi_N = \frac{2}{h}\int_{1-\frac{h}{2}}^{1} f(x)dx\,. \qquad (2.16)$$

Таким образом, дифференциальная задача (2.1), (2.10), (2.14) аппроксимирована разностной схемой (2.9), (2.12), (2.15).

2.1. Аппроксимация и сходимость построенной разностной схемы

Для простоты изложения рассмотрим случай, когда коэффициенты исходной дифференциальной задачи обладают достаточной гладкостью.

Определимся с видом шаблонных функционалов. Легко видеть, что

$$\frac{1}{a_i} = A\left(\frac{1}{k(x_i + sh)}\right) = \frac{1}{h}\int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} = \int_{-1}^{0} \frac{ds}{k(x_i + sh)},\ \frac{1}{a_{i+1}} = B\left(\frac{1}{k(x_i + sh)}\right),$$

$$d_i = F(q(x_i + sh)) = \int_{-0.5}^{0.5} q(x_i + sh)ds\,,\ \varphi_i = F(f(x_i + sh)) = \int_{-0.5}^{0.5} f(x_i + sh)ds\,,\ i = \overline{1, N-1}.$$

Поэтому

$$A(1)=\int_{-1}^{0} ds=1,\ A(s)=\int_{-1}^{0} sds=-\frac{1}{2},\ B(s)=A(s+1)=\frac{1}{2},\ B(s^2)-A(s^2)=\int_{0}^{1} s^2ds-\int_{-1}^{0} s^2ds=0,$$

$$F(1)=\int_{-0.5}^{0.5} ds=1,\ F(s)=\int_{-0.5}^{0.5} sds=0.$$

Отсюда, учитывая условия (1.4), делаем вывод, что разностное уравнение (2.9) имеет второй порядок аппроксимации.

Для левого граничного условия непосредственно получаем:

$$\nu(0)=a_1u_{x,0}-\left(\kappa_0+\frac{h}{2}d_0\right)u(0)+\left(g_0+\frac{h}{2}\varphi_0\right),$$

а поскольку

$$a_1u_{x,0}=\left[k\left(\frac{h}{2}\right)+O(h^2)\right]\cdot\frac{u(h)-u(0)}{h}=\left[k\left(\frac{h}{2}\right)+O(h^2)\right]\cdot\left[u'\left(\frac{h}{2}\right)+O(h^2)\right]=$$

$$=-W\left(\frac{h}{2}\right)+O(h^2)=-\left[W(0)+\frac{h}{2}W'(0)+O(h^2)\right]=k(0)u'(0)+\frac{h}{2}(ku')'(0)+O(h^2),$$

то, используя уравнение (2.1), имеем:

$$\nu(0)=\kappa_0u(0)-g_0+\frac{h}{2}(q(0)u(0)-f(0))-\left(\kappa_0+\frac{h}{2}d_0\right)u(0)+g_0+\frac{h}{2}\varphi_0=$$

$$=\frac{h}{2}(q(0)-d_0)u(0)+\frac{h}{2}(\varphi_0-f(0))+O(h^2).$$

Так как

$$q(0)-d_0=q(0)-\frac{2}{h}\int_0^{\frac{h}{2}} q(x)dx=q(0)-\frac{2}{h}\left[\frac{h}{2}q(0)+O(h^2)\right]=O(h)$$

и аналогично $\varphi_0-f(0)=O(h)$, то $\nu(0)=O(h^2)$, т.е. левое разностное граничное условие аппроксимирует условие (2.10) со вторым порядком.

Аналогичным образом показывается, что правое граничное условие также имеет второй порядок.

Таким образом, разностная схема, построенная с помощью рассмотренного варианта метода баланса, в классе достаточно гладких коэффициентов обладает вторым порядком аппроксимации.

Исследуем теперь вопрос о сходимости данной разностной схемы (сделав, однако, предположение о том, что на правом конце отрезка задано условие первого рода, так как это несколько упрощает выкладки). Тогда задача для погрешности $z_h=y_h-u_h$ будет выглядеть следующим образом:

$$\begin{cases} \left(az_{\bar{x}}\right)_x - dz = -\psi, \ x \in \omega_h, \\ a_1 z_{x,0} = \widetilde{\kappa}_0 z_0 - \nu(0), \\ z_N = 0. \end{cases}$$

Умножим разностное уравнение скалярно на сеточную функцию z:

$$\left(\left(az_{\bar{x}}\right)_x, z\right) - \left(d, z^2\right) = -(\psi, z).$$

Применяя первую разностную формулу Грина и учитывая условие $z_N = 0$, получим:

$$-\left(a, z_{\bar{x}}^2\right] - a_1 z_{x,0} z_0 - \left(d, z^2\right) = -(\psi, z)$$

или, так как $a_1 z_{x,0} = \widetilde{\kappa}_0 z_0 - \nu(0)$,

$$-\left(a, z_{\bar{x}}^2\right] - \widetilde{\kappa}_0 z_0^2 - \left(d, z^2\right) = -(\psi, z) - \nu(0) z_0 .$$

Умножая на -1, перепишем последнее соотношение в виде

$$\left(a, z_{\bar{x}}^2\right] + \widetilde{\kappa}_0 z_0^2 + \left(d, z^2\right) = (\psi, z) + \nu(0) z_0 . \qquad (2.17)$$

По предположению $k(x) \ge c_1 > 0$, $q(x) \ge 0$, $\kappa_0 \ge 0$. Поэтому

$$\left(a, z_{\bar{x}}^2\right] \ge c_1 \left| |z_{\bar{x}}] \right|^2, \ \left(d, z^2\right) \ge 0, \ \widetilde{\kappa}_0 z_0^2 \ge 0 .$$

Следовательно, из (2.17) получаем:

$$c_1 \left\| z_{\bar{x}}] \right\|^2 \le |(\psi, z)| + |\nu(0)||z_0| .$$

Оценим сверху правую часть этого неравенства:

$$|(\psi, z)| + |\nu(0)||z_0| \le \sum_{i=1}^{N-1} h|\psi_i||z_i| + |\nu(0)||z_0| \le \|z\|_C \left(\|\psi\|_C \sum_{i=1}^{N-1} h + |\nu(0)| \right) = \|z\|_C \left(\|\psi\|_C + |\nu(0)| \right).$$

С другой стороны, для сеточной функции z, обращающейся в нуль при $x = l$ (т.е. $z_N = 0$) справедлив аналог теоремы вложения $\|z\|_C \le \sqrt{l} \left\| z_{\bar{x}}] \right\|$ (***доказать*!**). Поэтому имеем:

$$c_1 \left\| z_{\bar{x}}] \right\|^2 \ge \frac{c_1}{l} \|z\|_C^2 = c_1 \|z\|_C^2$$

и, следовательно,

$$c_1 \|z\|_C^2 \le \|z\|_C \left(\|\psi\|_C + |\nu(0)| \right),$$

откуда

$$\|z\|_C \le \frac{1}{c_1} \left(\|\psi\|_C + |\nu(0)| \right).$$

Из полученного неравенства следует сходимость данной разностной схемы в норме C со вторым порядком в классе гладких коэффициентов.

Заметим, что в данном доказательстве важны, по сути, только формулы, задающие второй порядок аппроксимации. Поэтому для практических целей удобно иметь возможно более простые формулы для нахождения сеточных функций a, d, φ, использующие значения коэффициентов исходного уравнения $k(x), q(x), f(x)$ в отдельных точках. Обычно используют шаблон из одной или двух точек, полагая, например,

$$a_i = k_{i-0.5} = k\left(x_i - \frac{h}{2}\right) \quad \left(\text{здесь } A\left(\bar{k}(s)\right) = \bar{k}(-0.5)\right),$$

$$d_i = q_i, \ \varphi_i = f_i \quad \left(F\left(\bar{f}(s)\right) = \bar{f}(0)\right)$$

или

$$a_i = \frac{k_i + k_{i-1}}{2} \quad \left(A\left(\bar{k}(s)\right) = \frac{1}{2}\left(\bar{k}(-1) + \bar{k}(0)\right)\right),$$

или

$$a_i = \frac{2k_i k_{i-1}}{k_i + k_{i-1}} \quad \left(\frac{1}{A\left(\bar{k}(s)\right)} = \frac{1}{2}\left(\frac{1}{\bar{k}(-1)} + \frac{1}{\bar{k}(0)}\right)\right).$$

Фактически речь идет о замене интегралов в формулах (2.4), (2.7), (2.8) квадратурными формулами, обладающими необходимой точностью (в том числе – нелинейными).

Замечание. Все полученные при этом разностные схемы будут сходиться (но не в норме C (!)) также и в классе кусочно-гладких коэффициентов, однако скорость сходимости станет равной единице, и только исходная разностная схема с интегральным представлением коэффициентов сохраняет второй порядок. Поэтому ее называют ***наилучшей*** консервативной однородной разностной схемой.

§ 3. Вариационно-проекционные подходы к построению разностных схем

3.1. Метод Ритца построения разностных схем

Рассмотренный выше способ построения разностных схем автоматически приводит к консервативным разностным схемам, если оператор исходной дифференциальной задачи является самосопряженным. Изложим сейчас еще один способ, позволяющий добиться такого же эффекта.

Ранее мы рассматривали классический вариант метода Ритца решения граничных задач. Напомним, что решение всякого операторного уравнения с самосопряженным положительным оператором может быть сведено к эквивалентной задаче отыскания функции, доставляющей минимум некоторому функционалу (функционалу Ритца). Так, например, нахождение решения краевой задачи

$$\begin{cases} \left(k(x)u'(x)\right)' - q(x)u(x) = -f(x), \ 0 < x < 1, \ k(x) \ge c_1 > 0, \ q(x) \ge 0, \\ k(0)u'(0) = \kappa_0 u(0) - g_0, \ \kappa_0 \ge 0, \\ -k(1)u'(1) = \kappa_1 u(1) - g_1, \ \kappa_1 \ge 0, \end{cases} \tag{3.1}$$

эквивалентно задаче отыскания функции $u(x)$, доставляющей минимум функционалу

$$J(u) = \frac{1}{2}[u,u] - \int_0^1 f(x)u(x)dx - g_0 u(0) - g_1 u(1), \tag{3.2}$$

где

$$[u,v]=\int_0^1 [k(x)u'(x)v'(x)+q(x)u(x)v(x)]dx+\kappa_0 u(0)v(0)+\kappa_1 u(1)v(1), \qquad (3.3)$$

для которого уравнение (3.1) есть уравнение Эйлера.

Известно, что если входные параметры задачи удовлетворяют указанным в (3.1) условиям, то минимум функционала (3.2) существует и соответствующий элемент принадлежит пространству $W_2^1[0;1]$. В соответствии с идеей Ритца построим последовательность конечномерных подпространств $V_n \in W_2^1$ и вместо того, чтобы искать минимум на W_2^1, будем искать его на V_n. Пусть размерность подпространства V_n равна n и $\eta_i^{(n)}$, $i=\overline{0,n-1}$ – базис этого подпространства, т.е. любой элемент u_n этого подпространства представим в виде

$$u_n=\sum_{i=0}^{n-1} a_i \eta_i^{(n)}. \qquad (3.4)$$

Подставляя записанное представление для u_n вместо u в функционал $I(u)$, получим функцию n переменных $a_0, a_1, ..., a_{n-1}$. Так как мы желаем получить минимум этой функции, то числа a_i должны удовлетворять системе уравнений

$$\frac{\partial J(u_n)}{\partial a_i}=0,\ i=\overline{0,n-1}. \qquad (3.5)$$

Решив эту систему, мы получим определенные значения параметров $a_0, a_1, ..., a_{n-1}$, дающие $I(u_n)$ абсолютный минимум, а затем по формуле (3.4) получим требуемое приближенное решение. Найдем вид системы (3.4), исходя из конкретного функционала (3.2):

$$J(u_n)=\frac{1}{2}\left[\sum_{i=0}^{n-1} a_i\eta_i^{(n)}, \sum_{i=0}^{n-1} a_i\eta_i^{(n)}\right]-\int_0^1 f(x)\sum_{i=0}^{n-1} a_i\eta_i^{(n)}(x)dx-g_0\sum_{i=0}^{n-1} a_i\eta_i^{(n)}(0)-g_1\sum_{i=0}^{n-1} a_i\eta_i^{(n)}(1).$$

Тогда

$$\frac{\partial J(u_n)}{\partial a_j}=\sum_{i=0}^{n-1} a_i\left[\eta_i^{(n)},\eta_j^{(n)}\right]-\int_0^1 f(x)\eta_j^{(n)}(x)dx-g_0\eta_j^{(n)}(0)-g_1\eta_j^{(n)}(1)=0$$

или

$$\sum_{j=0}^{n-1}\alpha_{ij}a_j=\beta_i,\ i=\overline{0,n-1}, \qquad (3.6)$$

где

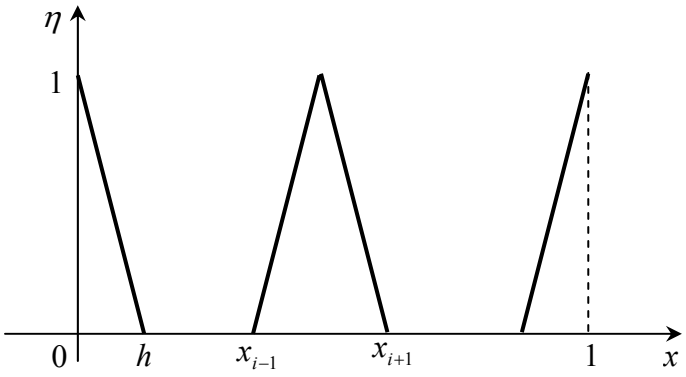
$$\alpha_{ij}=\left[\eta_i^{(n)},\eta_j^{(n)}\right],\ \beta_i=\int_0^1 f(x)\eta_i^{(n)}(x)dx+g_0\eta_i^{(n)}(0)+g_1\eta_i^{(n)}(1). \qquad (3.7)$$

Поскольку наша цель – построение разностной схемы с помощью метода Ритца, то достичь этой цели можно с помощью специального подбора координатных функций.

Пусть на отрезке $[0;1]$ задана сетка $\overline{\omega}_h$ и $N=n-1$. Тогда система (3.6) будет иметь вид трехточечной разностной схемы, если матрица этой системы будет трехдиагональной, т.е. если коэффициенты α_{ij} будут равны нулю при $|i-j|>1$. При этих условиях

система (3.6) будет классической разностной схемой, если в качестве параметров в разложении (3.4) будут выбраны значения функции u_n в узлах сетки $\overline{\omega}_h$.

Матрица системы (3.6) будет трехдиагональной, если базисные функции подпространства V_n при $|i-j|\geq 2$ будут ортогональны в смысле скалярного произведения (3.3). Это условие, очевидно, будет выполнено, если в качестве $\eta_i^{(N+1)}(x)$ взять функции, которые отличны от нуля только при $|x-x_i|\leq h$, $x\in[0;1]$ (такие функции называются ***функциями с конечным носителем*** или ***финитными***). Так как значения u_{N+1} в узле являются коэффициентами разложения, то должны выполняться условия $\eta_i^{(N+1)}(x_i)=1$. Простейшими функциями указанного вида, принадлежащими W_2^1, являются функции



$$\eta_i^{(N+1)}(x)=\begin{cases}0,\ \text{если}\ x\overline{\in}[x_{i-1};x_{i+1}],\\ \dfrac{x-x_{i-1}}{h},\ \text{если}\ x\in[x_{i-1};x_i],\ i=\overline{1,N-1},\\ \dfrac{x_{i+1}-x}{h},\ \text{если}\ x\in[x_i;x_{i+1}],\end{cases}$$

$$\eta_0^{(N+1)}(x)=\begin{cases}\dfrac{h-x}{h},\ \text{если}\ x\in[0;h],\\ 0,\ \text{если}\ x\overline{\in}[0;h],\end{cases}\qquad(3.8)$$

$$\eta_N^{(N+1)}(x)=\begin{cases}\dfrac{x-1+h}{h},\ \text{если}\ x\in[1-h;1],\\ 0,\ \text{если}\ x\overline{\in}[1-h;1].\end{cases}$$

Если координатные функции $\eta_i^{(N+1)}(x)$ выбрать по формулам (3.8), то система (3.6) примет вид

$$\begin{cases}\alpha_{ii-1}y_{i-1}+\alpha_{ii}y_i+\alpha_{ii+1}y_{i+1}=\beta_i,\ i=\overline{1,N-1},\\ \alpha_{00}y_0+\alpha_{01}y_1=\beta_0,\\ \alpha_{NN-1}y_{N-1}+\alpha_{NN}y_N=\beta_N.\end{cases}\qquad(3.9)$$

Пользуясь формулами (3.7), (3.8), вычислим коэффициенты α_{ii} и β_i:

$$\alpha_{ii}=\frac{1}{h^2}\left[\int_{x_{i-1}}^{x_{i+1}}k(x)dx+\int_{x_{i-1}}^{x_i}q(x)(x-x_{i-1})^2dx+\int_{x_i}^{x_{i+1}}q(x)(x_{i+1}-x)^2dx\right],\ i=\overline{1,N-1},$$

$$\alpha_{00}=\frac{1}{h^2}\left[\int_0^h k(x)dx+\int_0^h q(x)(x-h)^2dx\right]+\kappa_0,$$

$$\alpha_{NN} = \frac{1}{h^2}\left[\int_{1-h}^{1} k(x)dx + \int_{1-h}^{1} q(x)(x-1+h)^2 dx\right] + \kappa_1,$$

$$\alpha_{ii+1} = \alpha_{i+1i} = \frac{1}{h^2}\left[-\int_{x_i}^{x_{i+1}} k(x)dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)(x-x_i)dx\right],\ i=\overline{1,N-1},$$

$$\beta_i = \frac{1}{h}\left[\int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1})dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1}-x)dx\right],\ i=\overline{1,N-1},$$

$$\beta_0 = \frac{1}{h}\int_0^h f(x)(h-x)dx + g_0,$$

$$\beta_N = \frac{1}{h}\int_{1-h}^{1} f(x)(x-1+h)dx + g_1.$$

Полученную систему уравнений (3.9) легко записать в стандартном для однородных консервативных схем виде (2.9), (2.12), (2.15). Действительно, разностное уравнение (2.9) в развернутом виде выглядит следующим образом:

$$\frac{a_i}{h^2}y_{i-1} - \left(\frac{a_i + a_{i+1}}{h^2} + d_i\right)y_i + \frac{a_{i+1}}{h^2}y_{i+1} = -\varphi_i.$$

Чтобы получить такой вид из (3.9), очевидно, необходимо положить

$$a_i = -h\alpha_{ii-1},\ \varphi_i = \frac{1}{h}\beta_i,\ d_i = \frac{1}{h}\left(\alpha_{ii-1} + \alpha_{ii} + \alpha_{ii+1}\right).$$

Аналогично разбираемся и с граничными условиями. Поэтому при

$$\begin{cases} a_i = \dfrac{1}{h}\left[\displaystyle\int_{x_{i-1}}^{x_i} k(x)dx - \int_{x_{i-1}}^{x_i} q(x)(x_i-x)(x-x_{i-1})dx\right],\ i=\overline{1,N}, \\ d_i = \dfrac{1}{h^2}\left[\displaystyle\int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1})dx - \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)dx\right],\ i=\overline{1,N-1}, \\ d_0 = \dfrac{2}{h^2}\displaystyle\int_0^h q(x)(h-x)dx;\ \ d_N = \frac{2}{h^2}\int_{1-h}^{1} q(x)(x-1+h)dx, \\ \varphi_i = \dfrac{1}{h^2}\left[\displaystyle\int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1})dx - \int_{x_i}^{x_{i+1}} f(x)(x_{i+1}-x)dx\right],\ i=\overline{1,N-1}, \\ \varphi_0 = \dfrac{2}{h^2}\displaystyle\int_0^h f(x)(h-x)dx;\ \ \varphi_N = \frac{2}{h^2}\int_{1-h}^{1} f(x)(x-1+h)dx \end{cases} \qquad (3.10)$$

система (3.9) превратится в стандартную разностную схему

$$\begin{cases} \left(a y_{\bar{x}}\right)_x - dy = -\varphi, \ x \in \omega_h, \\ a_1 y_{x,0} = \left(\kappa_0 + \dfrac{h}{2} d_0\right) y_0 - \left(g_0 + \dfrac{h}{2}\varphi_0\right), \\ -a_N y_{\bar{x},N} = \left(\kappa_1 + \dfrac{h}{2} d_N\right) y_N - \left(g_1 + \dfrac{h}{2}\varphi_N\right). \end{cases} \tag{3.11}$$

Упражнение. Определить порядок аппроксимации разностной схемы (3.10), (3.11).

3.2. Метод аппроксимации квадратичного функционала

В предыдущем пункте мы использовали эквивалентность задачи (3.1) задаче об отыскании минимума функционала (3.2), (3.3), который в «сборном» виде будет выглядеть следующим образом:

$$J(u) = \int_0^1 \left[k(x)(u'(x))^2 + q(x)u^2(x)\right]dx - 2\int_0^1 f(x)u(x)dx + \kappa_0 u^2(0) + \kappa_1 u^2(1) - 2g_0 u(0) - 2g_1 u(1), \tag{3.12}$$

для построения разностной схемы методом Ритца. При этом в качестве стандартной идеи использовалась идея аппроксимации пространства, в котором отыскивался минимум функционала, подпространствами конечной размерности.

Сейчас же мы поступим по-другому: аппроксимируем не пространство, а сам функционал (3.12). Для этого заменим на сетке $\overline{\omega}_h$ интегралы, входящие в выражение (3.12), квадратурными формулами, предварительно переписав его в виде

$$J(u) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} k(x)(u'(x))^2 dx + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \left(q(x)u^2(x) - 2f(x)u(x)\right)dx + \kappa_0 u^2(0) + \kappa_1 u^2(1) - 2g_0 u(0) - 2g_1 u(1).$$

После этого аппроксимируем интегралы следующим образом:

$$\int_{x_{i-1}}^{x_i} k(x)(u'(x))^2 dx \approx a_i \left(u_{\bar{x},i}\right)^2 \cdot h, \ -\text{аналог формулы средних прямоугольников}$$

$$\int_{x_{i-1}}^{x_i} \left[q(x)u^2(x) - 2f(x)u(x)\right]dx \approx \frac{h}{2}\left[\left(q_i u_i^2 - 2f_i u_i\right) + \left(q_{i-1}u_{i-1}^2 - 2f_{i-1}u_{i-1}\right)\right] - \text{формула трапеций}.$$

Здесь a_i – некоторый функционал, зависящий от коэффициента $k(x)$ на отрезке $[x_{i-1}; x_i]$, например, $a_i = \dfrac{1}{h}\int_{x_{i-1}}^{x_i} k(x)dx$ или $a_i = k_{i-0.5}$ и т.п.

Тогда вместо $J(u)$ получим функционал

$$J_h(y)=\sum_{i=1}^{N} h a_i y_{\bar{x},i}^2+\sum_{i=1}^{N-1} h\left(q_i y_i^2-2f_i y_i\right)+\frac{h}{2}\left[q_0 y_0^2+q_N y_N^2-2f_0 y_0-2f_N y_N\right]+$$

$$+\kappa_0 y_0^2+\kappa_1 y_N^2-2g_0 y_0-2g_1 y_N$$

или

$$J_h(y)=\sum_{i=1}^{N} h a_i y_{\bar{x},i}^2+\sum_{i=1}^{N-1} h\left(q_i y_i^2-2f_i y_i\right)+\left(\kappa_0+\frac{h}{2}q_0\right)y_0^2+\left(\kappa_1+\frac{h}{2}q_N\right)y_N^2-$$

$$-2\left(g_0+\frac{h}{2}f_0\right)y_0-2\left(g_1+\frac{h}{2}f_N\right)y_N.$$

В результате имеем: $J_h(y)$ есть функция $(N+1)$ переменных y_i, и для того чтобы найти уравнения, определяющие точку ее минимума, необходимо приравнять нулю первые производные этой функции по переменным y_i:

$$\frac{\partial J_h(y)}{\partial y_i}=0,\ i=\overline{0,N}.$$

Тогда при $i=\overline{1,N-1}$ получим следующие уравнения:

$$2ha_{i+1}y_{\bar{x},i+1}\cdot\left(-\frac{1}{h}\right)+2ha_i y_{\bar{x},i}\cdot\frac{1}{h}+2hq_i y_i-2hf_i=0,\ i=\overline{1,N-1}.$$

Аналогично при $i=0$ будем иметь

$$2ha_1 y_{\bar{x},1}\cdot\left(-\frac{1}{h}\right)+2\left(\kappa_0+\frac{h}{2}q_0\right)y_0-2\left(g_0+\frac{h}{2}f_0\right),$$

а при $i=N$ –

$$2ha_N y_{\bar{x},N}\cdot\frac{1}{h}+2\left(\kappa_1+\frac{h}{2}q_N\right)y_N-2\left(g_1+\frac{h}{2}f_N\right).$$

От полученных соотношений очевиден переход к трехточечной разностной схеме

$$\begin{cases}\left(ay_{\bar{x}}\right)_x-qy=-f,\ x\in\omega_h,\\ a_1 y_{x,0}=\left(\kappa_0+\frac{h}{2}q_0\right)y_0-\left(g_0+\frac{h}{2}f_0\right),\\ -a_N y_{\bar{x},N}=\left(\kappa_1+\frac{h}{2}q_N\right)y_N-\left(g_1+\frac{h}{2}f_N\right).\end{cases}\qquad(3.13)$$

Упражнение. Показать, что разностная схема (3.13) при надлежащем выборе a_i имеет второй порядок аппроксимации.

Замечание. Описанную процедуру можно организовать и таким образом, чтобы, аналогично (3.11), ее коэффициенты были представимы в некоторой интегральной форме.

3.3. Метод Галеркина построения разностных схем

Очень близким к изложенному выше методу Ритца, но имеющим несколько более широкую область применимости, является другой проекционный метод – метод Галеркина (Бубнова – Галеркина). Этот метод применим, в частности, и тогда, когда задача не является самосопряженной. Рассмотрим технику его использования для построения разностных схем на примере задачи

$$\begin{cases} \dfrac{d}{dx}\left(k(x)\dfrac{du(x)}{dx}\right)+r(x)\dfrac{du(x)}{dx}-q(x)u(x)=-f(x),\ 0<x<1, \\ k(0)\dfrac{du(0)}{dx}=\kappa_0 u(0)-g_0, \\ -k(1)\dfrac{du(1)}{dx}=\kappa_1 u(1)-g_1. \end{cases} \tag{3.14}$$

В соответствии с общей идеей метода Галеркина коэффициенты a_i приближенного решения

$$u_n=\sum_{i=0}^{n-1} a_i \eta_i^{(n)}$$

находятся из условия ортогональности невязки Lu_n-f ко всем базисным функциям $\eta_i^{(n)}$. В случае задачи (3.14) равенство $(Lu-f,v)=0$ принимает вид (первое слагаемое под знаком интеграла, содержащее вторые производные, интегрируем по частям и пользуемся граничными условиями при вычислении двойной подстановки)

$$\int_0^1 [k(x)u'(x)v'(x)-r(x)u'(x)v(x)+q(x)u(x)v(x)-f(x)v(x)]dx+$$

$$+\kappa_0 u(0)v(0)+\kappa_1 u(1)v(1)-g_0 v(0)-g_1 v(1)=0. \tag{3.15}$$

Выберем такое же подпространство координатных функций, как и в методе Ритца: оно определяется формулами (3.8). Тогда приближенное решение $u_{N+1}(x)$ примет вид

$$u_{N+1}(x)=\sum_{j=0}^{N} y_j \eta_j^{(N+1)}(x), \tag{3.16}$$

где y_j – приближенные решения задачи (3.14) в узлах сетки $\overline{\omega}_h$. Подставляя (3.16) в (3.15) и выбирая в качестве поверочной функции $v(x)$ $\eta_i^{(N+1)}(x)$, $i=\overline{0,N}$, имеем:

$$\sum_{j=0}^{N}\left\{\int_0^1\left[k(x)y_j\frac{d}{dx}\eta_j^{(N+1)}(x)\frac{d}{dx}\eta_i^{(N+1)}(x)-r(x)y_j\frac{d}{dx}\eta_j^{(N+1)}(x)\eta_i^{(N+1)}(x)+q(x)y_j\eta_j^{(N+1)}(x)\eta_i^{(N+1)}(x)-\right.\right.$$

$$\left.-f(x)\frac{d}{dx}\eta_i^{(N+1)}(x)\right]dx+\kappa_0 y_j\eta_j^{(N+1)}(0)\eta_i^{(N+1)}(0)+\kappa_1 y_j\eta_j^{(N+1)}(1)\eta_i^{(N+1)}(1)-$$

$$\left.-g_0\eta_i^{(N+1)}(0)-g_1\eta_i^{(N+1)}(1)\right\}=0.$$

Учитывая вид координатных функций, отсюда получаем систему уравнений

$$\begin{cases} \alpha_{ii-1}y_{i-1} + \alpha_{ii}y_i + \alpha_{ii+1}y_{i+1} = \beta_i, \ i = \overline{1, N-1}, \\ \alpha_{00}y_0 + \alpha_{01}y_1 = \beta_0, \\ \alpha_{NN-1}y_{N-1} + \alpha_{NN}y_N = \beta_N, \end{cases} \tag{3.17}$$

где

$$\alpha_{ii-1} = \frac{1}{h^2}\left[-\int_{x_{i-1}}^{x_i} k(x)dx + \int_{x_{i-1}}^{x_i} r(x)(x - x_{i-1})dx + \int_{x_{i-1}}^{x_i} q(x)(x_i - x)(x - x_{i-1})dx \right],$$

$$\alpha_{ii} = \frac{1}{h^2}\left[\int_{x_{i-1}}^{x_{i+1}} k(x)dx - \int_{x_{i-1}}^{x_i} r(x)(x - x_{i-1})dx + \int_{x_i}^{x_{i+1}} r(x)(x_{i+1} - x)dx + \int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})^2 dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1} - x)^2 dx \right],$$

$$\alpha_{ii+1} = \frac{1}{h^2}\left[-\int_{x_i}^{x_{i+1}} k(x)dx - \int_{x_i}^{x_{i+1}} r(x)(x_{i+1} - x)dx + \int_{x_i}^{x_{i+1}} q(x)(x - x_i)(x_{i+1} - x)dx \right], \ i = \overline{1, N-1},$$

$$\alpha_{00} = \frac{1}{h^2}\left[\int_0^h k(x)dx + \int_0^h r(x)(h - x)dx + \int_0^h q(x)(h - x)^2 dx \right] + \kappa_0,$$

$$\alpha_{01} = \frac{1}{h^2}\left[-\int_0^h k(x)dx - \int_0^h r(x)(h - x)dx + \int_0^h q(x)x(h - x)dx \right],$$

$$\alpha_{NN-1} = \frac{1}{h^2}\left[-\int_{1-h}^1 k(x)dx + \int_{1-h}^1 r(x)(x - 1 + h)dx + \int_{1-h}^1 q(x)(1 - x)(x - 1 + h)dx \right],$$

$$\alpha_{NN} = \frac{1}{h^2}\left[\int_{1-h}^1 k(x)dx - \int_{1-h}^1 r(x)(x - 1 + h)dx + \int_{1-h}^1 q(x)(x - 1 + h)^2 dx \right] + \kappa_1,$$

$$\beta_i = \frac{1}{h}\left[\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x)dx \right], \ i = \overline{1, N-1},$$

$$\beta_0 = \frac{1}{h}\int_0^h f(x)(h - x)dx + g_0,$$

$$\beta_N = \frac{1}{h}\int_{1-h}^1 f(x)(x - 1 + h)dx + g_1.$$

Точно так же, как мы это поделали в методе Ритца, систему (3.17) можно привести к стандартному безындексному виду

$$\begin{cases} \left(a y_{\bar{x}}\right)_x + b^+ y_x + b^- y_{\bar{x}} - dy = -\varphi, \ x \in \omega_h \\ \left(a_1 + b_0^+ h\right) y_{x,0} = \left(\kappa_0 + \dfrac{h}{2} d_0\right) y_0 - \left(g_0 + \dfrac{h}{2}\varphi_0\right), \\ -\left(a_N - b_N^+ h\right) y_{\bar{x},N} = \left(\kappa_1 + \dfrac{h}{2} d_N\right) y_N - \left(g_1 + \dfrac{h}{2}\varphi_N\right), \end{cases} \tag{3.18}$$

где

$$\begin{cases} a_i = \dfrac{1}{h}\left[\int\limits_{x_{i-1}}^{x_i} k(x)dx - \int\limits_{x_{i-1}}^{x_i} q(x)(x_i - x)(x - x_{i-1})dx\right], \ i = \overline{1,N}, \\ b_i^+ = \dfrac{1}{h^2}\int\limits_{x_i}^{x_{i+1}} r(x)(x_{i+1} - x)dx, \ i = \overline{0,N-1}, \\ b_i^- = \dfrac{1}{h^2}\int\limits_{x_{i-1}}^{x_{i+1}} r(x)(x - x_{i-1})dx, \ i = \overline{1,N}, \\ d_i = \dfrac{1}{h^2}\left[\int\limits_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})dx - \int\limits_{x_i}^{x_{i+1}} q(x)(x_{i+1} - x)dx\right], \ i = \overline{1,N-1}, \\ d_0 = \dfrac{2}{h^2}\int\limits_0^h q(x)(h - x)dx; \ \ d_N = \dfrac{2}{h^2}\int\limits_{1-h}^1 q(x)(x - 1 + h)dx, \\ \varphi_i = \dfrac{1}{h^2}\left[\int\limits_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})dx - \int\limits_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x)dx\right], \ i = \overline{1,N-1}, \\ \varphi_0 = \dfrac{2}{h^2}\int\limits_0^h f(x)(h - x)dx; \ \ \varphi_N = \dfrac{2}{h^2}\int\limits_{1-h}^1 f(x)(x - 1 + h)dx. \end{cases}$$

Упражнение. Показать, что разностная схема (3.18) имеет второй порядок аппроксимации.

3.4. Метод аппроксимации интегрального тождества

Этот метод находится в таком же отношении к методу Галеркина, как метод аппроксимации квадратичного функционала к методу Ритца.

Для построения разностной схемы на сетке $\overline{\omega}_h$ аппроксимируем интегральное тождество (3.15) сумматорным тождеством для сеточных функций (поэтому метод также называют методом сумматорных тождеств), используя технику, изложенную в п. 3.2: переписав тождество в виде

$$J(u,v) = \sum_{i=1}^{N}\left\{\int\limits_{x_{i-1}}^{x_i} k(x)u'(x)v'(x)dx + \int\limits_{x_{i-1}}^{x_i}\left[q(x)u(x)v(x) - f(x)v(x) - r(x)u'(x)v(x)\right]dx + \right.$$

$$\left. + \kappa_0 u(0)v(0) + \kappa_1 u(1)v(1) - g_0 v(0) - g_1 v(1)\right\} = 0,$$

заменим первый интеграл под знаком суммы некоторым аналогом квадратурной формулы средних прямоугольников (при этом производные аппроксимируем левыми разностными),

а второй – квадратурной формулой трапеций (производную заменяем центральной разностной):

$$\int_{x_{i-1}}^{x_i} k(x)u'(x)v'(x)dx \approx ha_i y_{\bar{x},i} v_{\bar{x},i},$$

$$\int_{x_{i-1}}^{x_i} [q(x)u(x)v(x) - f(x)v(x) - r(x)u'(x)v(x)]dx \approx \frac{h}{2}\left[q_{i-1}y_{i-1}v_{i-1} - f_{i-1}v_{i-1} + q_i y_i v_i - f_i v_i - r_i y_{\overset{\circ}{x},i} v_i \right].$$

В результате получим сумматорное тождество

$$J_h(y_h, v_h) = \sum_{i=1}^{N} ha_i y_{\bar{x},i} v_{\bar{x},i} + \sum_{i=1}^{N-1} h\left[q_i y_i v_i - f_i v_i - r_i y_{\overset{\circ}{x},i} v_i \right] + \frac{h}{2} q_0 y_0 v_0 - \frac{h}{2} f_0 v_0 - \frac{h}{2} r_0 y_{\bar{x},1} v_0 +$$

$$+ \frac{h}{2} q_N y_N v_N - \frac{h}{2} f_N v_N - \frac{h}{2} r_N y_{\bar{x},N} v_N + \kappa_0 y_0 v_0 + \kappa_1 y_N v_N - g_0 v_0 - g_1 v_N =$$

$$= \sum_{i=1}^{N} ha_i y_{\bar{x},i} v_{\bar{x},i} + \sum_{i=1}^{N-1} h\left[q_i y_i v_i - f_i v_i - r_i y_{\overset{\circ}{x},i} v_i \right] + \left(\kappa_0 + \frac{h}{2} q_0 \right) y_0 v_0 + \left(\kappa_1 + \frac{h}{2} q_N \right) y_N v_N -$$

$$- \frac{h}{2} r_0 y_{x,0} v_0 - \frac{h}{2} r_N y_{\bar{x},N} v_N - \left(g_0 + \frac{h}{2} f_0 \right) v_0 - \left(g_1 + \frac{h}{2} f_N \right) v_N = 0.$$

В этом тождестве v – произвольная сеточная функция. Выбирая ее равной единице в одном из узлов сетки и равной нулю в остальных, получим уравнение в той точке, где v отлична от нуля. Перебирая таким образом все узлы, получим разностную схему

$$\begin{cases} \left(a y_{\bar{x}}\right)_x + r(x) y_{\overset{\circ}{x}} - qy = -f(x), \ x \in \omega_h \\ \left(a_1 + \frac{h}{2} r(0) \right) y_{x,0} = \left(\kappa_0 + \frac{h}{2} q(0) \right) y_0 - \left(g_0 + \frac{h}{2} f(0) \right), \\ -\left(a_N - \frac{h}{2} r(1) \right) y_{\bar{x},N} = \left(\kappa_1 + \frac{h}{2} q(1) \right) y_N - \left(g_1 + \frac{h}{2} f(1) \right), \end{cases} \tag{3.19}$$

имеющую второй порядок аппроксимации.

Замечание* 1**. Проекционные подходы к построению разностных схем, разобранные нами в данном параграфе, в современной литературе достаточно часто относят к ***методам конечных элементов.

***Замечание* 2.** Все рассмотренные нами выше способы построения разностных схем могут быть распространены как на случай неравномерной сетки на отрезке, так и на случай функций многих независимых переменных.

ГЛАВА XVII

Методы исследования устойчивости разностных схем

§ 1. Принцип максимума

Ранее мы отмечали важность такого свойства разностных схем как устойчивость. Ее исследование, как правило, состоит в получении априорных оценок решения разностной задачи через ее входные данные.

Для оценок в равномерной метрике разностных эллиптических и параболических уравнений, а также разностных уравнений переноса, применяется ***принцип максимума***. Он позволяет получить равномерные оценки решения через правую часть уравнения, граничные и начальные данные. Опишем его подробнее.

Пусть Ω – некоторое конечное множество точек $x=(x_1,...,x_p)$ p-мерного евклидова пространства (сетка). Пусть также в каждой точке $x\in\Omega$ задан шаблон $Ш(x)\subset\Omega$. Через $Ш'(x)$, как и ранее, обозначим окрестность точки x, т.е. $Ш'(x)=Ш(x)\setminus\{x\}$.

Рассмотрим уравнение

$$Sy(x)=F(x),\ x\in\Omega, \tag{1.1}$$

где $y(x)$ – искомая функция, $F(x)$ – заданная сеточная функция, а S – линейный оператор, определяемый формулой

$$Sv(x)=A(x)v(x)-\sum_{\xi\in Ш'(x)}B(x,\xi)v(\xi), \tag{1.2}$$

коэффициенты которого $A(x)$ и $B(x,\xi)$ – заданные сеточные функции x и ξ. Будем далее предполагать, что они удовлетворяют условиям

$$\begin{aligned}&1)\ A(x)>0,\ B(x,\xi)>0 \text{ для всех } x\in\Omega,\ \xi\in Ш'(x);\\ &2)\ D(x)\equiv A(x)-\sum_{\xi\in Ш'(x)}B(x,\xi)\geq 0.\end{aligned} \tag{1.3}$$

Пусть x – произвольный узел сетки Ω. Тогда возможны два случая:

а) $Ш'(x)=\varnothing$;

б) $Ш'(x)$ содержит хотя бы один узел $\xi\in\Omega$.

Если имеет место первый случай, т.е. $Ш'(\bar{x})=\varnothing$, то уравнение(1.1) при $x=\bar{x}$ имеет вид

$$A(\bar{x})y(\bar{x})=F(\bar{x})$$

или

$$y(\bar{x})=g(\bar{x}).$$

Такую точку будем называть граничным узлом (и писать $\bar{x}\in\gamma$), а остальные узлы, окрестность которых состоит, по крайней мере, из одной точки, – внутренними (обозначаем: множество ω). Согласно сказанному $\omega\cup\gamma=\Omega$. Отметим, что с такой точки зрения в случае краевых условий второго или третьего рода для эллиптических уравнений граничных узлов ***нет***.

Будем предполагать также, что сетка Ω – связная, т.е. для любых двух узлов $\bar{x}$, $\bar{\bar{x}}$, не являющихся одновременно граничными (например, для определенности, $\bar{x}\in\omega$)

можно указать такую последовательность узлов $x_1, x_2, ..., x_m$, что каждый последующий узел принадлежит окрестности предыдущего, т.е.

$$x_1 \in Ш'(\bar{x}),\ x_2 \in Ш'(x_1), ..., x_{m+1} \in Ш'(x_m),\ \bar{\bar{x}} \in Ш'(x_{m+1}). \qquad (1.4)$$

***Теорема* 1.** (Принцип максимума) Пусть $y(x)$ – отличная от тождественной постоянной сеточная функция, определенная на связной сетке Ω, и пусть на ω выполняются условия (1.3). Тогда из условия $Sy(x) \le 0$ $(Sy(x) \ge 0)$ на ω следует, что $y(x)$ не может принимать наибольшего положительного (наименьшего отрицательного) значения во внутренних узлах сетки Ω.

Доказательство.

Пусть, для определенности, $Sy(x) \le 0$ и существует узел $\bar{x} \in \omega$, в котором

$$y(\bar{x}) = \max_{x \in \omega} y(x) > 0.$$

Тогда в этом узле

$$Sy(\bar{x}) \equiv A(\bar{x})y(\bar{x}) - \sum_{\xi \in Ш'(\bar{x})} B(\bar{x}, \xi)y(\xi) = D(\bar{x})y(\bar{x}) + \sum_{\xi \in Ш'(\bar{x})} B(\bar{x}, \xi)(y(\bar{x}) - y(\xi)) \ge 0.$$

Так как по условию теоремы $Sy(\bar{x}) \le 0$, то отсюда следует, что $Sy(\bar{x}) = 0$, а значит, в силу неравенств $B(\bar{x}, \xi) > 0$, $y(\bar{x}) \ge y(\xi)$, $y(\bar{x}) > 0$ и $D(\bar{x}) \ge 0$ имеем: $D(\bar{x}) = 0$ и $y(\bar{x}) = y(\xi)$ для всех $\xi \in Ш'(\bar{x})$.

Поскольку сеточная функция $y(x)$ отлична от тождественной константы при $x \in \omega$, то существует узел $\bar{\bar{x}} \in \omega$ такой что $y(\bar{\bar{x}}) < y(\bar{x})$. В силу связности сетки Ω можно указать последовательность узлов $x_1, ..., x_m$, удовлетворяющую условиям (1.4). Тогда $y(x_1) = y(\bar{x})$. Повторяя рассуждения, приведенные выше, получим:

$$y(x_1) = y(x_2) = \ldots = y(x_m) = y(\bar{x}).$$

Следовательно, для точки x_m получим неравенство

$$Sy(x_m) = D(x_m)y(x_m) + \sum_{\xi \in Ш'(x_m)} B(x_m, \xi)(y(x_m) - y(\xi)) \ge B(x_m, \bar{\bar{x}})(y(x_m) - y(\bar{\bar{x}})) > 0,$$

которое противоречит условию теоремы. ⊠

Замечание. Возможна несколько более общая формулировка доказанной теоремы: не на всей сетке ω, а на некотором связном подмножестве $\Omega' \subset \Omega$.

***Следствие* 1.** Пусть $Sy(x) \le 0$ $(Sy(x) \ge 0)$ на связной сетке Ω и существует, по крайней мере, один узел $x_0 \in \Omega$, для которого

$$D(x_0) > 0. \qquad (1.5)$$

Тогда $y(x) \le 0$ $(y(x) \ge 0)$ на сетке Ω.

Доказательство.

Пусть $Sy(x) \le 0$. Если $y(x) \equiv \text{const}$ на Ω, то

$$Sy(x_0) = D(x_0)y(x_0) + \sum_{\xi \in Ш'(x_0)} B(x_0, \xi)(y(x_0) - y(\xi)) = D(x_0)y(x_0) \le 0.$$

Поэтому $y(x) \equiv y(x_0) \le 0$.

Если же $y(x)$ не является тождественно постоянной, то $y(x) \le 0$ на основании принципа максимума (в соответствии с последним наибольшее положительное значение функция при указанных условиях может принимать только на границе, а там, поскольку $D(\xi) = 1$ для всех $\xi \in \gamma$, $y(\xi) \le 0$). ⊠

***Следствие* 2.** Пусть оператор S удовлетворяет на сетке Ω условиям (1.3), (1.5). Тогда задача (1.1), (1.2) имеет единственное решение.

Доказательство.

Убедимся, что однородная задача, соответствующая (1.1), имеет лишь тривиальное решение. Поскольку $Sy(x) = 0 \le 0$, то согласно *Следствию* 1 $y(x) \le 0$ для всех $x \in \Omega$. С другой стороны, так как $Sy(x) = 0 \ge 0$, то по тому же *Следствию* 1 $y(x) \ge 0$ для всех $x \in \Omega$. Отсюда $y(x) \equiv 0$ для всех $x \in \Omega$. ⊠

***Теорема* 2** (теорема сравнения). Пусть $y(x)$ – решение задачи (1.1) – (1.3), (1.5), а $\bar{y}(x)$ – решение той же задачи с правой частью $\bar{F}(x)$. Тогда из условия $|F(x)| \le \bar{F}(x)$ следует, что $|y(x)| \le \bar{y}(x)$ на Ω.

Доказательство.

Сложим и вычтем уравнения $Sy(x) = F(x)$ и $S\bar{y}(x) = \bar{F}(x)$. В силу линейности оператора S получим:

$$\begin{cases} S(\bar{y} + y) = \bar{F}(x) + F(x) \ge 0, \\ S(\bar{y} - y) = \bar{F}(x) - F(x) \ge 0. \end{cases}$$

Из первого из этих неравенств с помощью *Следствия* 1 получаем, что $\bar{y}(x) + y(x) \ge 0$, а из второго – $\bar{y}(x) - y(x) \ge 0$. Объединяя эти неравенства, получаем: $|y(x)| \le \bar{y}(x)$. ⊠

Таким образом, решение задачи (1.1), (1.2) можно оценить с помощью ***мажорантной функции*** $\bar{y}(x)$, которая удовлетворяет уравнению $S\bar{y}(x) = \bar{F}(x)$ с правой частью $\bar{F}(x) \ge |F(x)|$ (например, $\bar{F}(x) = \|F(x)\|_C$ или $\bar{F}(x) = |F(x)|$ и т.п.).

Теорема сравнения позволяет сразу получить оценку решения первой краевой задачи в случае однородного уравнения. Имеет место

***Следствие* 3.** Для решения задачи $\begin{cases} Sy(x) = 0, \ x \in \omega, \\ y(x) = \mu(x), \ x \in \gamma \end{cases}$ имеет место априорная оценка

$$\max_{x \in \omega} |y(x)| \le \max_{x \in \gamma} |\mu(x)| \qquad \text{или} \qquad \|y\|_{\bar{C}} \le \|\mu\|_{C_\gamma}. \tag{1.6}$$

Доказательство.

Пусть $\bar{y}(x)$ – решение задачи $\begin{cases} S\bar{y}(x) = 0, \ x \in \omega, \\ \bar{y}(x) = |\mu(x)|, \ x \in \gamma. \end{cases}$

Тогда на основании *Теоремы* 2 $|y(x)| \le \bar{y}(x)$. Поэтому если $\bar{y}(x) \equiv \text{const}$ на Ω, то $\max\limits_{x\in\Omega} \bar{y}(x) = \max\limits_{x\in\gamma} |\mu(x)|$ и, следовательно, неравенство (1.6) справедливо. Если же $\bar{y}(x)$ отлично от тождественной константы, то в силу *Теоремы* 1 максимальное значение этой функции может достигаться только на границе, т.е. снова $\|\bar{y}(x)\|_{\bar{C}} \le \|\mu(x)\|_{C_\gamma}$. ☒

С помощью доказанных утверждений можно получить оценки и для решения неоднородной задачи.

***Теорема* 3.** Если $D(x) > 0$ для всех $x \in \Omega$, то для решения задачи (1.1) – (1.3) верна априорная оценка

$$\max_{x\in\Omega} |y(x)| \le \max_{x\in\Omega} \frac{|F(x)|}{D(x)} \qquad \text{или} \qquad \|y\|_{\bar{C}} \le \left\|\frac{F}{D}\right\|_{\bar{C}}. \tag{1.7}$$

Доказательство.

Пусть $\bar{y}(x)$ – решение задачи $S\bar{y}(x) = |F(x)|$, $x \in \Omega$. Тогда по *Теореме* 2 $|y(x)| \le \bar{y}(x)$. Функция $\bar{y}(x)$ достигает наибольшего значения в некотором узле $\bar{x}$: $\bar{y}(\bar{x}) = \max\limits_{x\in\Omega} \bar{y}(x) > 0$. Тогда

$$Sy(\bar{x}) = D(\bar{x})\bar{y}(\bar{x}) + \sum_{\xi\in Ш'(\bar{x})} B(\bar{x},\xi)(\bar{y}(\bar{x}) - \bar{y}(\xi)) = |F(\bar{x})|$$

и так как $\bar{y}(x) \ge \bar{y}(\xi)$, то $D(\bar{x})\bar{y}(\bar{x}) \le |F(\bar{x})|$. Отсюда $\bar{y}(x) \le \dfrac{|F(\bar{x})|}{D(\bar{x})}$ и поэтому

$$\|y\|_{\bar{C}} \le \bar{y}(\bar{x}) = \max_{x\in\Omega} \bar{y}(x) = \|\bar{y}\|_{\bar{C}} \le \frac{|F(\bar{x})|}{D(\bar{x})} \le \max_{x\in\Omega} \left|\frac{F(x)}{D(x)}\right| = \left\|\frac{F}{D}\right\|_{\bar{C}}.$$

☒

Аналогично доказывается

***Теорема* 4.** Пусть сетка Ω разбита на два непересекающихся непустых подмножества Ω' и Ω'', причем Ω' – связное. Если $F(x) \equiv 0$ на Ω', а на Ω'' $F(x)$ отлична от тождественного нуля и $D(x) > 0$, то для решения задачи (1.1) – (1.3) справедлива оценка

$$\|y\|_{\bar{C}} = \max_{x\in\Omega} |y(x)| \le \max_{x\in\Omega'} \frac{|F(x)|}{D(x)}. \tag{1.8}$$

Упражнение. Доказать *Теорему* 4.

1.1. Примеры исследования устойчивости с помощью принципа максимума

Фактически исследование состоит в том, чтобы:

1) привести задачу к виду (1.1) – (1.2);

2) проверить условия (1.3), (1.5).

При решении первой части задачи в качестве ориентировки следует помнить, что коэффициент $A(x)$ должен быть ***диагональным*** элементом матрицы при записи задачи в матрично-векторной форме.

Пример 1. Задача Коши для уравнения переноса

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0, \ t > 0, \ -\infty < x < +\infty, \\ u(x,0) = u_0(x), \ a = \text{const} > 0. \end{cases}$$

Зададим сетку $\omega_{h\tau} = \omega_h \times \omega_\tau$ и на ней запишем простейшую разностную схему, аппроксимирующую поставленную задачу

$$\begin{cases} y_t + a y_{\bar{x}} = 0, \\ y(x,0) = u_0(x) \end{cases}$$

или в индексной форме

$$\begin{cases} \dfrac{y_k^{j+1} - y_k^j}{\tau} + a \dfrac{y_k^j - y_{k-1}^j}{h} = 0, \ j = 0,1,... \\ y_k^0 = u_0(x_k). \end{cases}$$

Решение данной разностной задачи, очевидно, должно находиться послойно. Единственное же значение на $(j+1)$-м временно́м слое, подлежащее определению – y_k^{j+1}. Поэтому коэффициент при нем – искомый диагональный элемент, т.е. в канонической записи (1.2) $x = (x_k, t_{j+1})$. Поэтому схему перепишем в виде

$$\frac{1}{\tau} y_k^{j+1} = \left(\frac{1}{\tau} - \frac{a}{h} \right) y_k^j + \frac{a}{h} y_{k-1}^j.$$

Поэтому

$$A(x) = \frac{1}{\tau}, \ B_1 = \frac{1}{\tau} - \frac{a}{h}, \ B_2 = \frac{a}{h}, \ D(x) = A(x) - (B_1 + B_2) = \frac{1}{\tau} - \left(\frac{1}{\tau} - \frac{a}{h} + \frac{a}{h} \right) = 0.$$

Следовательно, условия (1.3) примут вид

$$\begin{cases} A(x) = \dfrac{1}{\tau} > 0, \\ B_1 = \dfrac{1}{\tau} - \dfrac{a}{h} \geq 0, \\ B_2 = \dfrac{a}{h} \geq 0, \\ D(x) = 0 \geq 0. \end{cases}$$

Заметим, что для коэффициентов $B(x,\xi)$ мы используем в условиях нестрогие неравенства, поскольку равенство нулю того или иного коэффициента, по сути, означает отсутствие в шаблоне соответствующего узла. В полученной системе первое, третье и четвертое неравенства (учитывая знак коэффициента a) выполняются автоматически, а второе приводит к ограничению, связывающему допустимые шаги сетки: $\dfrac{a\tau}{h} \leq 1$ (в литературе его называют ***условием Куранта***). При выполнении найденного условия легко (на основании *Следствия* 3) получить оценку разностного решения: $\|y\|_{\bar{C}} \leq \|u_0\|_{C_\gamma}$.

Пример 2. Первая краевая задача для уравнения теплопроводности

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + f(x,t), \ 0 < x < 1, \ t > 0, \\ u(x,0) = u_0(x), \ 0 \leq x \leq 1, \\ u(0,t) = \mu_0(t), \ t \geq 0, \\ u(1,t) = \mu_1(t), \ t \geq 0. \end{cases}$$

На сетке $\overline{\omega}_{h\tau}$ запишем ***явную*** разностную схему:

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi,\ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x),\ x \in \overline{\omega}_h, \\ y(0,t) = \mu_0(t),\ t \in \omega_\tau, \\ y(1,t) = \mu_1(t),\ \omega_\tau. \end{cases}$$

Расписав разностное уравнение в индексной форме, получим:

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j .$$

Отсюда, учитывая послойный принцип реализации и единственный узел сетки на верхнем временно́м слое, находим: $x = (x_i, t_{j+1})$. Поэтому, умножив уравнение в индексной форме на τ, перепишем его в виде

$$y_j^{j+1} = \left(1 - \frac{2\tau}{h^2}\right) y_i^j + \frac{\tau}{h^2}\left(y_{i+1}^j + y_{i-1}^j\right) + \tau\varphi_i^j .$$

Таким образом, $A(x) = 1 > 0,\ B_2 = B_3 = \frac{\tau}{h^2} > 0$. Неравенство $B_1 = 1 - \frac{2\tau}{h^2} \geq 0$ приводит к ограничению $\tau \leq \frac{h^2}{2}$ (***условие Куранта*** для уравнения теплопроводности), а последнее из условий

$$D(x) = 1 - \left(1 - \frac{2\tau}{h^2} + \frac{\tau}{h^2} + \frac{\tau}{h^2}\right) \equiv 0$$

очевидным образом выполняется. Следовательно, в равномерной метрике явная разностная схема для уравнения теплопроводности устойчива при выполнении условия $\tau \leq \frac{h^2}{2}$.

1.2. Монотонные разностные схемы для обыкновенных дифференциальных уравнений второго порядка

Используя полученные в начале параграфа результаты, несложно исследовать и конкретные разностные схемы в случае граничных задач для обыкновенного дифференциального уравнения второго порядка.

Пусть, например, исходная дифференциальная задача имеет вид

$$\begin{cases} \frac{d}{dx}\left(k(x)\frac{du(x)}{dx}\right) - q(x)u(x) = -f(x),\ 0 < x < 1, \\ k(x) \geq k_0 > 0,\ q(x) \geq 0, \\ u(0) = \mu_0, \\ u(1) = \mu_1. \end{cases} \tag{1.9}$$

Запишем для нее однородную консервативную разностную схему

$$\begin{cases} \left(a y_{\bar{x}}\right)_x - dy = -\varphi, \ x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1. \end{cases} \tag{1.10}$$

Расписывая ее в индексной форме, имеем (уравнения, описывающие граничные условия, в данной задаче тривиальны):

$$\frac{1}{h}\left(a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i .$$

Исходя из принципа «точка x должна соответствовать диагональному элементу», получаем: $x = x_i$. Поэтому, собрав подобные, перепишем наше разностное уравнение в виде

$$\left(\frac{1}{h^2}(a_{i+1} + a_i) + d_i \right) y_i - \left(\frac{1}{h^2} a_{i+1} y_{i+1} + \frac{1}{h^2} a_i y_{i-1} \right) = \varphi_i ,$$

откуда

$$A(x) = \frac{1}{h^2}(a_{i+1} + a_i) + d_i , \ B_1 = \frac{1}{h^2} a_{i+1} , \ B_2 = \frac{1}{h^2} a_i , \ D(x) = d_i .$$

Очевидно, если функционалы, с помощью которых вычисляются коэффициенты разностной схемы, сохраняют свойства коэффициентов исходной дифференциальной задачи (в частности, положительность), то записанная разностная схема будет удовлетворять условиям принципа максимума при всех значениях параметра h. Такие разностные схемы называют ***монотонными***.

Не следует думать, что монотонность – «врожденное» свойство разностных схем. Чтобы убедиться в обратном, рассмотрим задачу более общего, нежели (1.9), вида

$$\begin{cases} Lu(x) \equiv \dfrac{d}{dx}\left(k(x) \dfrac{du(x)}{dx} \right) + r(x) \dfrac{du(x)}{dx} - q(x) u(x) = -f(x), \ 0 < x < 1, \\ k(x) \ge k_0 > 0, \ q(x) \ge 0, \ |r(x) \le C|, \\ u(0) = \mu_0, \\ u(1) = \mu_1. \end{cases} \tag{1.11}$$

Для этой задачи легко записать разностную схему второго порядка, заменив производную $\dfrac{du}{dx}$ в слагаемом $r(x)\dfrac{du}{dx}$ центральной разностной производной:

$$\begin{cases} \left(a y_{\bar{x}}\right)_x + r y_{\mathring{x}} - qy = -f, \ x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1. \end{cases} \tag{1.12}$$

Проделывая выкладки, аналогичные приведенным выше, получаем:

$$\left(\frac{a_{i+1} + a_i}{h^2} + q_i \right) y_i - \left(\frac{a_{i+1}}{h^2} + \frac{r_i}{2h} \right) y_{i+1} - \left(\frac{a_i}{h^2} - \frac{r_i}{2h} \right) y_{i-1} = f_i ,$$

т.е.

$$A(x)=\frac{a_{i+1}+a_i}{h^2}+q_i,\ \ B_1=\frac{a_{i+1}}{h^2}+\frac{r_i}{2h},\ \ B_2=\frac{a_i}{h^2}-\frac{r_i}{2h},\ \ D(x)=q_i.$$

Отсюда видим, что сеточные коэффициенты $A(x)$ и $D(x)$ удовлетворяют условиям принципа максимума при всех h, в то время как неотрицательность коэффициентов B_1 и B_2 приводит к ограничению на шаг сетки вида $h\le\frac{2k(x)}{|r(x)|}$, которое становится достаточно обременительным, если $|r(x)|>>1$.

В то же время, если воспользоваться для аппроксимации $\frac{du}{dx}$ в слагаемом $r(x)\frac{du}{dx}$ односторонними производными (правой при $r(x)\ge 0$ и левой при $r(x)\le 0$: так называемая ***аппроксимация против потока***), то полученная разностная схема

$$\begin{cases}\left(ay_{\bar{x}}\right)_x+r^+y_x+r^-y_{\bar{x}}-qy=-f,\ x\in\omega_h,\\ y(0)=\mu_0,\\ y(1)=\mu_1\end{cases}\tag{1.13}$$

будет монотонной (***проверить!***), но ее порядок равен единице.

Построим монотонную схему второго порядка точности, содержащую односторонние производные, учитывающие знак коэффициента $r(x)$. Для этого, как оказывается, достаточно написать монотонную схему с односторонними производными типа (1.13) для уравнения с возмущенными коэффициентами

$$\widetilde{L}u(x)\equiv\kappa\frac{d}{dx}\left(k(x)\frac{du(x)}{dx}\right)+r(x)\frac{du(x)}{dx}-q(x)u(x)=-f(x),\tag{1.14}$$

где

$$\kappa=\frac{1}{1+R},\ \ R=\frac{h|r|}{2k}\ -\text{разностное число Рейнольдса.}$$

Аппроксимируем слагаемое $r(x)\frac{du(x)}{dx}$ выражением

$$(ru')_i=\left(\frac{r}{k}(ku')\right)_i\sim b_i^+a_{i+1}u_{x,i}+b_i^-a_iu_{\bar{x},i},$$

где $b_i^\pm=F\left(\widetilde{r}^{\pm}(x_i+sh)\right)$, $\widetilde{r}^{\pm}=\frac{r^\pm}{k}$, а F – шаблонный функционал, используемый для вычисления коэффициентов d и φ разностной схемы (например, можно просто положить $b_i^+=\frac{r_i^+}{k_i}=\frac{r_i+|r_i|}{2k_i}$, $b_i^-=\frac{r_i^-}{k_i}=\frac{r_i-|r_i|}{2k_i}$).

В результате получаем однородную разностную схему

$$\begin{cases}\kappa\left(ay_{\bar{x}}\right)_x+b^+a^{(+1)}y_x+b^-ay_{\bar{x}}-dy=-\varphi,\ x\in\omega_h,\\ a^{(+1)}=a(x+h),\ \kappa=\frac{1}{1+R},\ R=\frac{h|r|}{2k},\\ y(0)=\mu_0,\\ y(1)=\mu_1\end{cases}\tag{1.15}$$

Приводя (1.15) к каноническому виде по изложенной ранее схеме, имеем:

$$B_1 = \frac{\kappa a_{i+1}}{h^2} + \frac{b_i^+ a_{i+1}}{h} > 0,\ B_2 = \frac{\kappa a_i}{h^2} - \frac{b_i^- a_i}{h} > 0,\ A(x) = B_1 + B_2 + d_i > 0,\ D(x) = d_i \geq 0.$$

Таким образом, разностная схема (1.15) является монотонной.

Погрешность аппроксимации этой схемы

$$\psi = \kappa\left(a u_{\bar{x}}\right)_x + b^+ a^{(+1)} u_x + b^- a u_{\bar{x}} - du + \varphi - \left(Lu + f\right)$$

представим в виде суммы

$$\psi = \psi^{(1)} + \psi^{(2)},$$

$$\psi^{(1)} = \left[\left(a u_{\bar{x}}\right)_x - du + \varphi\right] - \left[\left(ku'\right)' - qu + f\right],$$

$$\psi^{(2)} = \left[(\kappa - 1)\left(a u_{\bar{x}}\right)_x + b^+ a^{(+1)} u_x + b^- a u_{\bar{x}}\right] - ru'.$$

Как мы помним, для достаточно гладких функций $\psi^{(1)} = O\left(h^2\right)$. В то же время,

$$b^+ = \tilde{r}^+ + O\left(h^2\right),\ b^- = \tilde{r}^- + O\left(h^2\right),\ k\tilde{r}^{\pm} = r^{\pm},\ r^+ + r^- = r,\ r^+ - r^- = |r|,$$

$$a u_{\bar{x}} = ku' - \frac{h}{2}\left(ku'\right)' + O\left(h^2\right),\ a^{(+1)} u_x = ku' + \frac{h}{2}\left(ku'\right)' + O\left(h^2\right),$$

$$\left(a u_{\bar{x}}\right)_x = \left(ku'\right)' + O\left(h^2\right).$$

Поэтому

$$b^+ a^{(+1)} u_x + b^- a u_{\bar{x}} = \left[\tilde{r}^+ + O\left(h^2\right)\right]\cdot\left[ku' + \frac{h}{2}\left(ku'\right)' + O\left(h^2\right)\right] + \left[\tilde{r}^- + O\left(h^2\right)\right]\cdot\left[ku' - \frac{h}{2}\left(ku'\right)' + O\left(h^2\right)\right] =$$

$$= \left(\tilde{r}^+ + \tilde{r}^-\right)ku' + \frac{h}{2}\left(ku'\right)'\left(\tilde{r}^+ - \tilde{r}^-\right) + O\left(h^2\right) = ru' + \frac{h}{2}\left(ku'\right)'\cdot\frac{|r|}{k} + O\left(h^2\right).$$

Следовательно,

$$\psi^{(2)} = \left[\kappa - 1 = \frac{1}{1+R} - 1 = -\frac{R}{1+R}\right] = -\frac{R}{1+R}\left(ku'\right)' + ru' + \frac{h}{2}\left(ku'\right)'\cdot\frac{|r|}{k} + O\left(h^2\right) - ru' =$$

$$= \left(ku'\right)'\cdot\left(R - \frac{R}{1+R}\right) + O\left(h^2\right) = \left(ku'\right)'\cdot\frac{R^2}{1+R} + O\left(h^2\right) = O\left(h^2\right),$$

так как $R = \frac{h|r|}{2k} = O(h)$.

Таким образом, построенная разностная схема (1.15) имеет второй порядок и является монотонной. Ее целесообразно использовать в случае быстро меняющейся функции $r(x)$.

Замечание. Монотонную разностную схему второго порядка несложно написать, если от уравнения (1.11) перейти к уравнению

$$\frac{d}{dx}\left(\mu(x)k(x)\frac{du(x)}{dx}\right)-\mu(x)q(x)u(x)=-\mu(x)f(x),$$

где $\mu(x)=\exp\left(\int\frac{r(x)}{k(x)}dx\right)$, т.е. преобразовав его к самосопряженному виду.

Упражнение. Построить соответствующую разностную схему.

§ 2. Метод разделения переменных

Этот метод применяется для строгого обоснования многих линейных разностных схем и нестрогого исследования большинства нелинейных задач. Теоретические основы метода и его практическое применение традиционно изучают в курсе «Уравнений математической физики».

Технически поиск частного решения уравнения в виде произведения функций, каждая из которых зависит только от одной независимой переменной, приводит к необходимости решать задачу на собственные значения для некоторого дифференциального оператора. И, таким образом, решение задачи получается в виде ряда по собственным функциям данных операторов. При этом, конечно же, хорошо, если данная система оказывается ортогональной (или, более того, ортонормированной).

В разностном варианте технически и теоретически все остается таким же: ищем частное решение в виде произведения функций, каждая из которых зависит от одной (своей) независимой переменной, переходим к задаче на собственные значения. После ее решения можно делать некоторые заключения об исследуемых свойствах решения разностной задачи (в частности, об устойчивости).

Конечно, вместо поиска конкретных систем собственных функций и собственных значений для каждого разностного оператора можно пользоваться и некоей универсальной системой сеточных функций, которая являлась бы ортогональной системой собственных функций любого оператора разностного дифференцирования на равномерной сетке (по аналогии с тем, как мы проводим разложение в ряд Фурье на всей числовой прямой функций, периодических с периодом l, или периодически продолжая их на всю числовую прямую).

В этом случае также можно сеточную функцию y_h, определенную на сетке $\overline{\omega}_h=\left\{x_k=kh,\ k=0,1,...,N;\ h=\frac{l}{N}\right\}$ доопределить на всей числовой прямой с координатами $x_k=kh$, $k=0,\pm1,\pm2,...$ так, чтобы получилась l-периодическая сеточная функция. Множество таких функций обозначим M_h.

В пространстве M_h введем скалярное произведение по формуле

$$(v_h,y_h)=\sum_{s=0}^{N-1}hv_h(x_s)\overline{y}_h(x_s).$$

Тогда примером системы линейно независимых l-периодических ортогональных в M_h функций являются функции

$$\mu_k(x)=\exp\left(ik\frac{2\pi}{l}x\right),\ k=0,\pm1,\pm2,...,\pm\frac{N-1}{2}\qquad(\text{или } k=0,1,...,N-1).$$

Действительно,

$$(\mu_k(x),\mu_m(x))=\sum_{s=0}^{N-1}h\exp\left(ik\frac{2\pi}{l}x_s\right)\cdot\exp\left(-im\frac{2\pi}{l}x_s\right)=\sum_{s=0}^{N-1}h\exp\left(i(k-m)\frac{2\pi}{l}sh\right)=Nh\delta_k^m=l\delta_k^m.$$

Следовательно, любую функцию из M_h можно разложить в «сумму Фурье»

$$y_h(x)=\sum_{k=0}^{N-1}a_k\mu_k(x).$$

Заметим также, что $\mu_k(x)$ являются собственными функциями операторов правой и левой разностных производных. Действительно,

$$(\mu_k(x))_x=\frac{\mu_k(x+h)-\mu_k(x)}{h}=\frac{\exp\left(ik\frac{2\pi}{l}x\right)\cdot\exp\left(ik\frac{2\pi}{l}h\right)-\exp\left(ik\frac{2\pi}{l}x\right)}{h}=$$

$$=\mu_k(x)\cdot\frac{\exp\left(ik\frac{2\pi}{l}h\right)-1}{h}.$$

Видим отсюда, что собственным значением оператора правой разностной производной, соответствующим собственной функции $\mu_k(x)$, является $\lambda_k=\dfrac{\exp\left(ik\frac{2\pi}{l}h\right)-1}{h}$. Аналогичный результат имеет место для левой разностной производной. Следовательно, функции $\mu_k(x)$ образуют полную систему собственных функций для любого разностного оператора L_h вида

$$L_hy_h=\sum_{s,q}a_{sq}D^s\overline{D}^qy_h,$$

где

$$D^sy_h=y_{\underbrace{xx\ldots x}_{s}},\ \overline{D}^qy_h=y_{\underbrace{\bar{x}\bar{x}\ldots\bar{x}}_{q}}.$$

Это позволяет изучать вопросы исследования устойчивости разностных схем с использованием данной системы функций.

Рассмотрим применение к линейным двухслойным разностным схемам, записываемым в каноническом виде

$$By_t+Ay=\varphi,\qquad(2.1)$$

где B и A – некоторые разностные операторы, действующие по пространственной переменной x.

При фиксированной правой части погрешность z приближенного решения удовлетворяет однородному уравнению

$$B\hat{z}=(B-\tau A)z.\qquad(2.2)$$

Будем, в соответствии с изложенным выше, искать частное решение в виде

$$z(x_s,t_j)=q_k^j\exp\left(ik\frac{2\pi}{l}x_s\right). \tag{2.3}$$

При этом, очевидно, $\hat{z}=q_k z$, так что q_k есть множитель роста k-й гармоники при переходе с одного временно́го слоя на другой. Подставляя (2.3) в (2.2), получим уравнение для определения q_k:

$$q_k^{j+1}\lambda_k(B)\exp\left(ik\frac{2\pi}{l}x_s\right)=q_k^j\left[\lambda_k(B)-\tau\lambda_k(A)\right]\exp\left(ik\frac{2\pi}{l}x_s\right).$$

Отсюда

$$q_k=1-\tau\frac{\lambda_k(A)}{\lambda_k(B)}$$

(при этом, естественно, исходная разностная схема должна быть разностной схемой с постоянными коэффициентами).

Теперь остается оценить «степень роста». Имеет место

Теорема (признак устойчивости). Разностная схема (2.1) с постоянными коэффициентами устойчива по начальным данным, если для всех k выполняется неравенство

$$|q_k|\le 1+C\tau,\ C\ -\text{константа}. \tag{2.4}$$

Доказательство.

Разложим произвольную ошибку начальных данных по системе $\mu_k(x)$:

$$z(x_s,t_0)=\sum_{k=0}^{N-1}a_k\exp\left(ik\frac{2\pi}{l}x_s\right).$$

Тогда, поскольку разностная схема (2.1) линейна, то

$$z(x_s,t_j)=\sum_{k=0}^{N-1}a_k q_k^j\exp\left(ik\frac{2\pi}{l}x_s\right).$$

Следовательно, учитывая ортогональность системы $\{\mu_k(x)\}$, имеем:

$$\left\|z^j\right\|^2=\left(z^j,z^j\right)=l\cdot\sum_{k=0}^{N-1}\left|q_k^j\right|^2\cdot|a_k|^2\le l\cdot\max_{0\le k\le N-1}|q_k|^{2j}\sum_{k=0}^{N-1}|a_k|^2=\max_{0\le k\le N-1}|q_k|^{2j}\cdot\left\|z^0\right\|^2\le$$

$$\le(1+C\tau)^{2j}\left\|z^0\right\|^2\le\exp(2Cj\tau)\left\|z^0\right\|^2\le\exp(2CT)\left\|z^0\right\|^2.$$

⊠

***Замечание* 1.** Фактически константа C не должна быть слишком большой, поэтому при проверке сформулированного признака обычно полагают $C=0$.

Следствие. Если хотя бы для одного k величину $|q_k|$ нельзя мажорировать величиной $1+C\tau$, то схема неустойчива.

***Замечание* 2.** Практическое использование метода разделения переменных обычно состоит в следующем:

1) полагают $y_k^j = q^j \, \mathrm{e}^{ik\varphi}$, где $\varphi \in [0; 2\pi)$ (по сути, используется обозначение $\varphi = \dfrac{2\pi}{l} x_s$);
2) подставляя это выражение в исследуемую разностную схему, находят q;
3) проверяют условие $|q| \le 1$.

Пример. Исследуем описанным способом разностную схему для уравнения переноса, изученную нами в предыдущем параграфе с помощью принципа максимума:

$$y_t + a y_{\bar{x}} = 0 .$$

Расписав разностное уравнение в индексной форме, имеем:

$$\frac{y_k^{j+1} - y_k^j}{\tau} + a \frac{y_k^j - y_{k-1}^j}{h} = 0 . \tag{2.5}$$

Пусть теперь $y_k^j = q^j \, \mathrm{e}^{ik\varphi}$. Подставляя это выражение в (2.5), получаем:

$$\frac{q^{j+1} \, \mathrm{e}^{ik\varphi} - q^j \, \mathrm{e}^{ik\varphi}}{\tau} + a \frac{q^j \, \mathrm{e}^{ik\varphi} - q^j \, \mathrm{e}^{i(k-1)\varphi}}{h} = 0 .$$

Отсюда, сократив на $q^j \, \mathrm{e}^{ik\varphi}$, находим:

$$q = 1 - \frac{a\tau}{h} + \frac{a\tau}{h} \mathrm{e}^{-i\varphi}$$

или (полагая $\gamma = \dfrac{a\tau}{h}$)

$$q = 1 - \gamma + \gamma \, \mathrm{e}^{-i\varphi} = 1 - \gamma\left(1 - \mathrm{e}^{-i\varphi}\right).$$

Следовательно,

$$|q|^2 = \left(1 - \gamma + \gamma \cos\varphi\right)^2 + \gamma^2 \sin^2 \varphi = 1 - 2\gamma + 2\gamma \cos\varphi - 2\gamma^2 \cos\varphi + 2\gamma^2 .$$

Поэтому неравенство $|q|^2 \le 1$ может быть переписано в виде

$$-\gamma\left(1 - \cos\varphi\right) + \gamma^2\left(1 - \cos\varphi\right) \le 0$$

или

$$\gamma\left(1 - \cos\varphi\right)\left(1 - \gamma\right) \ge 0 ,$$

откуда, учитывая положительность коэффициента a, получаем ограничение $\gamma \le 1$, которое совпадает с условием Куранта, полученным нами ранее.

Замечание. Полученное совпадение, вообще говоря, случайно, поскольку речь идет об устойчивости в ***различных*** нормах.

§ 3. Метод энергетических неравенств

Ранее мы приводили простейший пример использования метода для получения априорных оценок. Сейчас используем его для получения критерия устойчивости двухслойных разностных схем, записанных в канонической форме

$$By_t + Ay = \varphi\,, \tag{3.1}$$

Для упрощения выкладок детальное изложение проведем в предположении, что:

1) операторы A и B не зависят от t (постоянны);
2) $B > 0$ (положителен);
3) $A = A^* > 0$ (положителен и самосопряжен).

Умножив уравнение (3.1) скалярно на сеточную функцию $2\tau\, y_t$, получим:

$$2\tau\left(By_t, y_t\right) + 2\tau\left(Ay, y_t\right) = 2\tau\left(\varphi, y_t\right). \tag{3.2}$$

Так как

$$y = \frac{\hat{y}+y}{2} - \frac{\hat{y}-y}{2} = \frac{1}{2}\left(\hat{y}+y\right) - \frac{\tau}{2}y_t\,,$$

то (3.2) перепишем в виде

$$2\tau\left(\left(B - \frac{\tau}{2}A\right)y_t, y_t\right) + \left(A\left(\hat{y}+y\right), \hat{y}-y\right) = 2\tau\left(\varphi, y_t\right).$$

Поскольку

$$\left(A\left(\hat{y}+y\right), \hat{y}-y\right) = \left(A\hat{y}, \hat{y}\right) - \left(Ay, y\right),$$

то отсюда имеем:

$$2\tau\left(\left(B - \frac{\tau}{2}A\right)y_t, y_t\right) + \left(A\hat{y}, \hat{y}\right) = \left(Ay, y\right) + 2\tau\left(\varphi, y_t\right). \tag{3.3}$$

(3.3) – энергетическое тождество для разностной схемы (3.1).

Теорема. Условие

$$B \geq \frac{\tau}{2}A \tag{3.4}$$

необходимо и достаточно для устойчивости в H_A по начальным данным разностной схемы (3.1), т.е. для выполнения неравенства

$$\left\|y^j\right\|_A \leq \left\|y^0\right\|_A,\ \ j = 1,2,\ldots. \tag{3.5}$$

(здесь $\|y\|_A = \sqrt{\left(Ay, y\right)}$).

Доказательство.

Достаточность.

Пусть выполнено условие (3.4). Из энергетического тождества (3.3) при $\varphi = 0$ (мы исследуем устойчивость по начальным данным) следует

$$2\tau\left(\left(B - \frac{\tau}{2}A\right)y_t, y_t\right) + \left(A\hat{y}, \hat{y}\right) = \left(Ay, y\right).$$

Отсюда в силу (3.4) имеем:

$$\left(A\hat{y},\hat{y}\right)\le\left(Ay,y\right)$$

или

$$\left\|\hat{y}\right\|_A^2\le\left\|y\right\|_A^2,$$

а тогда

$$\left\|y^{j+1}\right\|_A\le\left\|y^j\right\|_A\le\cdots\le\left\|y^0\right\|_A.$$

Необходимость.

Пусть разностная схема (3.1) устойчива по начальным данным и выполнено неравенство (3.5). Докажем, что отсюда следует операторное неравенство (3.4), т.е.

$$\left(Bv,v\right)\ge\frac{\tau}{2}\left(Av,v\right)\text{ для всех }v\in H.$$

Запишем энергетическое тождество при $t=0$:

$$2\tau\left(\left(B-\frac{\tau}{2}A\right)y_t(0),y_t(0)\right)+\left(Ay^1,y^1\right)=\left(Ay^0,y^0\right).$$

В силу неравенства (3.5) это тождество может быть выполнено только при

$$2\tau\left(\left(B-\frac{\tau}{2}A\right)y_t(0),y_t(0)\right)=\left(Ay^0,y^0\right)-\left(Ay^1,y^1\right)\ge 0,$$

т.е.

$$2\tau\left(\left(B-\frac{\tau}{2}A\right)y_t(0),y_t(0)\right)\ge 0. \qquad (*)$$

Так как $y^0\in H$ – произвольный, то и элемент $v=y_t(0)=B^{-1}Ay^0\in H$ также произволен. В самом деле, задавая любой элемент $v=y_t(0)\in H$, находим $y^0=-A^{-1}Bv\in H$, так как оператор A^{-1} существует. Таким образом, неравенство (*) выполнено при любых $v=y_t(0)\in H$, т.е. имеет место операторное неравенство (3.4).

⊠

***Замечание* 1.** Условие (3.4) достаточно для устойчивости схемы (3.1), если $B=B(t)$ – несамосопряженный положительный оператор.

***Замечание* 2.** Если исходное семейство разностных схем другое (т.е. операторы A и B удовлетворяют некоторым условиям, отличным от сформулированных выше) (например, оба положительные, самосопряженные и постоянные и т.п.), то аналогичным путем могут быть установлены и другие условия (чаще всего – достаточные) устойчивости (например, в норме $\|\cdot\|_B$)

Пример. Вновь обратимся к явной разностной схеме для уравнения теплопроводности с нулевыми граничными условиями и нулевой правой частью:

$$\begin{cases}\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2},\ 0<x<1,\ t>0,\\ u(x,0)=u_0(x),\ 0\le x\le 1,\\ u(0,t)=u(1,t)=0.\end{cases}$$

Соответствующая разностная схема на сетке $\overline{\omega}_{h\tau}$ имеет вид

$$\begin{cases} y_t = y_{\bar{x}x},\ (x,t) \in \omega_{h\tau}, \\ y(x,0) = u_0(x),\ x \in \overline{\omega}_h, \\ y(0,t) = y(1,t) = 0,\ t \in \omega_\tau. \end{cases}$$

Приведем ее к виду (3.1):

$$y_t - y_{\bar{x}x} = 0$$

или

$$By_t + Ay = 0,$$

где

$$B = E,\ A = -\Lambda.$$

Условие (3.4) теперь примет вид

$$E - \frac{\tau}{2} A \geq 0.$$

Так как $A \leq \|A\| E$, то неравенство можно усилить:

$$E - \frac{\tau}{2} A \geq E - \frac{\tau}{2} \|A\| E = \left(1 - \frac{\tau}{2} \|A\|\right) E \geq 0.$$

Отсюда следует:

$$1 - \frac{\tau}{2} \|A\| \geq 0$$

или

$$\tau \leq \frac{2}{\|A\|}.$$

Так как оператор A – самосопряженный, то $\|A\| = \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} < \frac{4}{h^2}$. Поэтому полученное условие устойчивости можно записать в более удобном виде $\tau \leq \frac{h^2}{2}$.

Замечание. В случае самосопряженных операторов B и A вместо операторного неравенства (3.4) можно перейти к системе числовых неравенств вида

$$\lambda_k(B) \geq \frac{\tau}{2} \lambda_k(A).$$

Последнее вполне аналогично условию спектральной устойчивости, полученному нами выше для метода разделения переменных.