САНКТ–ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

КАФЕДРА ВЫЧИСЛИТЕЛЬНОЙ ФИЗИКИ

В.А.БУСЛОВ , С.Л.ЯКОВЛЕВ

ЧИСЛЕННЫЕ МЕТОДЫ I.
ИССЛЕДОВАНИЕ ФУНКЦИЙ

КУРС ЛЕКЦИЙ

САНКТ–ПЕТЕРБУРГ

2001

Утверждено на заседании кафедры
вычислительной физики
печатается по решению методической комиссии
физического факультета СПбГУ

А В Т О Р Ы : В.А.БУСЛОВ, С.Л.ЯКОВЛЕВ

Р Е Ц Е Н З Е Н Т : докт. физ.-мат. наук С.Ю.СЛАВНОВ

Курс лекций состоит из двух частей. Настоящая первая часть посвящена численным аппроксимациям функций и, связанным с этим вопросам дифференцирования и интегрирования, вторая — решению уравнений, в том числе и дифференциальным. Издание представляет собой изложение вводных лекций по численным методам, читавшихся на протяжении ряда лет авторами в первом семестре II курса физического факультета СПбГУ. С этим связано ограничение материала вошедшего в учебник, поскольку ко второму курсу студенты еще не обладают достаточной математической подготовкой, необходимой для реализации многих численных методов. В частности, не освещены вопросы численного решения дифференциальных уравнений в частных производных, некорректных задач и ряда других, относящихся к численным методам, преподаваемым на IV курсе физического факультета. Тем не менее некоторые вопросы вводного курса численных методов требуют предварительных знаний, выходящих за рамки об'ема математических сведений, получаемых студентами на I-м и даже II-м курсе, поэтому авторы сочли как необходимым так и возможным, включить в соответствующих местах базовые сведения из функционального анализа и математической физики, чтобы сделать изложение материала в разумных пределах независимым от априорных знаний читателя.

В пособии принята нумерация формул по главам. Приведенная библиография частично представляет собой источник справочного материала, но, в основном, рассчитана на дальнейшее изучение численных методов.

Авторы рады возможности выразить свою благодарность нашему коллеге С.Ю.Славянову, прочитавшему рукопись и сделавшему ряд ценных замечаний, и признательны Т.В.Фроловой за помощь в наборе текста.

Глава 1

Введение. Пространства с метрикой

В численных методах математическая задача решается как правило приближенно. Получаемое приближение в том или ином смысле должно быть "близко расположенным"к истинному решению, поэтому понятию близости необходимо придать четкий математический смысл, чтобы иметь критерий сравнения и возможность утверждать, что такое-то приближение есть "хорошее"приближение, а такое-то — нет. Все об'екты, которые изучаются в численных методах, принадлежат некоторым пространствам (пространствам функций, векторным пространствам) с различными свойствами. Общим для всех этих пространств является понятие расстояния, которое и является мерой близости элементов. Поэтому естественно начать изучение численных методов с наиболее общего пространства, для любой пары элементов которого, определено расстояние. Таковым является *метрическое пространство.*

Определение. Пусть M — некоторое множество и ρ — бинарная функция называемая *метрикой*, $\rho : M \times M \to \mathbf{R}_+ = [0, \infty)$, такая что для любых элементов множества M выполнено

$1)\rho(x, y) = 0 \Leftrightarrow x = y;$

$2)\rho(x, y) = \rho(y, x);$

$3)\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ — неравенство треугольника,

тогда пара (M, ρ) называется *метрическим пространством.*

Заметим, что если на одном и том же множестве M определить другую бинарную функцию с указанными свойствами, то мы будем иметь, соответственно, и другое метрическое пространство.

Структура метрического пространства позволяет говорить о сходимости последовательностей элементов данного пространства. Именно, последовательность x_i называется сходящейся (по метрике) к элементу x если

$$\lim_{i \to \infty} \rho(x_i, x) = 0 \ .$$

На практике же, имея дело с последовательностями приближений x_i к точному решению x поставленной задачи, проверять выполнение этого условия зачастую не представляется возможным, поскольку само это решение, как правило, неизвестно, и есть лишь возможность сравнивать приближения x_i между собой, то есть выяснять является ли данная последовательность фундаментальной (последовательностью Коши). Напомним, что последовательность x_i называется фундаментальной, если $\forall \ \epsilon > 0 \ \ \exists N \ \ \forall i, k > N \ \ \rho(x_i, x_k) < \epsilon$. Метрическое пространство, в котором любая фундаментальная последовательность имеет предел, принадлежащий этому же пространству, называется *полным.* Как известно, любое неполное метрическое про-

странство (M,ρ) можно пополнить [15] единственным образом с сохранением расстояния, если понимать под элементами пополнения (M^*,ρ^*) классы эквивалентных друг другу фундаментальных последовательностей (последовательности $\{x_n\}$ и $\{y_n\}$ называются эквивалентными, если $\rho(x_n,y_n)\to 0$ при $n\to\infty$), а в качестве новой метрики ρ^* принять следующую: $\rho^*(x^*,y^*)=\lim\limits_{k\to\infty}\rho(x_k,y_k)$, где элементы x_k, y_k являются k-ми элементами из произвольных представителей $\{x_n\}$ и $\{y_n\}$ классов эквивалентных последовательностей, отвечающих элементам x^* и y^* соответственно. При этом элементы пространства (M,ρ) всюду плотны в (M^*,ρ^*).

Простым примером неполного метрического пространства служит, например, отрезок $(0,1]$ с метрикой $\rho(x,y)=|x-y|$. Это пространство неполное, так как последовательность $1/n$, очевидно, является фундаментальной, но ее пределом является 0, и он не принадлежит отрезку (0,1]. Пополнением является замкнутый отрезок [0,1].

Естественно, что полнота пространства, является существенным обстоятельством для численных методов, поскольку последовательность приближений, принадлежащих одному классу объектов, может иметь предел этому классу не принадлежащий и, следовательно, не обладать требуемыми свойствами.

Как правило в численных методах задача поиска решения x методом последовательных приближений может быть сформулирована в виде задачи о нахождении неподвижной точки некоторого сжимающего отображения A :

$$Ax=x\ . \tag{1}$$

Напомним, что точка x называется *неподвижной точкой* отображения A заданного в метрическом пространстве, если выполнено (1). Само же отображение A называется *сжимающим* или сжатием, если существует такое число $0<\alpha<1$, что для любых элементов x,y метрического пространства

$$\rho(Ax,Ay)\le\alpha\rho(x,y)\ . \tag{2}$$

Заметим, что всякое сжимающее отображение непрерывно, поскольку если $x_n\to x$, то из (2) следует, что $Ax_n\to Ax$.

Важным свойством сжимающего отображения является существование неподвижной точки. Именно, справедлива

Теорема (принцип сжимающих отображений). *Всякое сжимающее отображение, определенное в полном метрическом пространстве имеет одну и только одну неподвижную точку.*

Доказательство. Пусть x_0 — произвольная точка метрического пространства. Определим последовательность $x_n=Ax_{n-1}$. Покажем, что она фундаментальная. Будем считать для определенности, что $m\ge n$, тогда

$$\rho(x_n,x_m)=\rho(A^nx_0,A^mx_0)\le\alpha^n\rho(x_0,x_{m-n})\le$$

$$\le\alpha^n\{\rho(x_0,x_1)+\rho(x_1,x_2)+\ldots+\rho(x_{m-n-1},x_{m-n})\}\le$$

$$=\alpha^n\rho(x_0,x_1)\{1+\alpha+\alpha^2+\ldots+\alpha^{m-n-1}\}=\alpha^n\rho(x_0,x_1)\frac{1-\alpha^{m-n}}{1-\alpha}\le$$

$$\le\alpha^n\rho(x_0,x_1)\frac{1}{1-\alpha}\underset{n\to\infty}{\longrightarrow}0\ .$$

Таким образом последовательность x_n фундаментальная и, следовательно, в силу полноты пространства имеет предел. Обозначим его через x. Убедимся, что x является неподвижной точкой. Действительно, из

непрерывности отображения A

$$Ax = A \lim_{n\to\infty} x_n = \lim_{n\to\infty} Ax_n = \lim_{n\to\infty} x_{n+1} = x \ .$$

Осталось удостовериться, что неподвижная точка является единственной. Пусть

$$Ax = x \ , \qquad Ay = y \ ,$$

тогда из (2)

$$\rho(x,y) = \rho(Ax,Ay) \le \alpha\rho(x,y) \ ,$$

откуда $\rho(x,y) = 0$, что в силу определения метрического пространства означает, что $x = y$. В дальнейшем мы будем неоднократно пользоваться принципом сжимающих отображений.

Определим также понятие *порядка сходимости*. Пусть последовательность x_n сходится: $\lim\limits_{n\to\infty} x_n = x$ и

$$d = \lim_{n\to\infty} \frac{\ln\rho(x_{n+1},x)}{\ln\rho(x_n,x)} \ . \tag{3}$$

Если существует конечный предел (3), то он и называется порядком сходимости. Выражение (3) можно записать и в другой форме. Именно:

$$\lim_{n\to\infty} \frac{\rho(x_{n+1},x)}{\rho^d(x_n,x)} = C \ , \tag{4}$$

где C некоторая отличная от 0 и не равная бесконечности константа. Из этого выражения видно, что чем выше порядок сходимости d, тем быстрее последовательность x_n сходится к пределу.

Метрическое пространство является слишком общим понятием. Как правило, об'екты, с которыми приходится иметь дело, обладают свойствами не только метрического пространства, но и рядом дополнительных свойств. Напомним предварительно некоторые определения из абстрактной алгебры [20].

Определение. Класс G об'ектов (элементов) a, b, c, ... называется *группой*, если определена бинарная операция, которая каждой паре элементов a, b ставит в соответствие некоторый об'ект (результат операции) $a \cdot b$ так, что:

1) $a \cdot b$ является элементом класса (замкнутость относительно операции $\cdot$);

2) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ (ассоциативность);

3) G содержит (левую) *единицу* e такую, что для любого a из G, $e \cdot a = a$;

4) для любого элемента $a \in G$ в G существует (левый) *обратный* элемент a^{-1} такой, что $a^{-1} \cdot a = e$.

Операцию $\cdot$, определяющую группу, называют (абстрактным) *умножением* и часто опускают при записи: $ab = a \cdot b$. Группа *коммутативна* или *абелева*, если любые ее элементы перестановочны. Определяющую операцию в коммутативной группе часто называют (абстрактным) *сложением*, обозначая ее $+$, а единичный элемент называют *нуль* и обозначают 0. Обратный элемент к a записывают как $-a$; при этом пишут $a + (-b) = a - b$.

Нетрудно убедиться, что каждая группа имеет единственную левую и правую единицы, и эти единицы равны, равно как каждый элемент имеет единственный левый и правый обратные и эти элементы равны. Отсюда следуют законы сокращения ($ab = ac \Rightarrow b = c$; $ca = cb \Rightarrow b = c$) и существование единственного решения x уравнения $ax = b$ (или $xa = b$), т.е. однозначно определено "деление".

Определение. Класс R об'ектов (элементов) a, b, c, ... называется *кольцом*, если определены уже две бинарные операции, обычно называемые сложением и умножением, такие, что:

1) R есть абелева группа по сложению;

2) $ab \in R$ (замкнутость по отношению к умножению);

3) $a(bc) = (ab)c$ (ассоциативность умножения);

4) $a(b + c) = ab + ac$, $(a + b)c = ac + bc$ (дистрибутивные законы).

Заметим, что $a \cdot 0 = 0 \cdot a = 0$. Два элемента $a \neq 0$ и $b \neq 0$, для которых $ab = 0$, называются соответственно *левым* и *правым делителями нуля*. Например, непрерывные функции на конечном интервале образуют коммутативное кольцо, содержащее делители нуля. В кольце без делителей нуля из $ab = 0$ следует, что либо $a = 0$ либо $b = 0$ и действуют законы сокращения. Если кольцо R содержит и левую и правую единицы, то они единственны и совпадают, а R называется *кольцом с единицей*. Аналогично, если элемент обладает и левым и правым обратным, то они также единственны и совпадают.

Определение. *Тело* B – кольцо с единицей, в котором для каждого ненулевого элемента существует мультипликативный обратный (т.е. $B \setminus \{0\}$ – группа по умножению).

Коммутативное тело называется полем.

Определение. Непустое множество L называется *линейным* или *векторным пространством* над полем $\mathbf{F}$, если оно удовлетворяет следующим условиям:

1. L — коммутативная группа по (векторному) сложению;

2. Для любого элемента α поля $\mathbf{F}$ и любого $x \in L$ определен элемент $\alpha x \in L$ (замкнутость по отношению к умножению на элемент поля (скаляр)), причем

а) $\alpha(\beta x) = (\alpha\beta)x$ (ассоциативный закон для умножения на скаляр);

б) $1 \cdot x = x$;

в) $(\alpha + \beta)x = \alpha x + \beta x$, $\alpha(x + y) = \alpha x + \alpha y$ (дистрибутивные законы).

Если в качестве поля $\mathbf{F}$ выступает поле действительных чисел $\mathbf{R}$ или поле комплексных чисел $\mathbf{C}$, то различают, соответственно, действительные (вещественные) и комплексные линейные пространства.

Всякую функцию f заданную на линейном пространстве L со значениями в поле $\mathbf{F}$ ($f : L \to \mathbf{F}$) мы будем называть *функционалом*. Функционал f называется линейным, если для любых $x, y \in L$ и $\alpha, \in \mathbf{F}$ выполнено:

1) $f(x + y) = f(x) + f(y)$ (аддитивность);

2) $f(\alpha x) = \alpha f(x)$ (однородность).

Функционал f называется непрерывным в точке x, если для любой последовательности x_n из $x_n \to x$ следует, что $f(x_n) \to f(x)$. Нетрудно убедиться, что если линейный функционал непрерывен в одной точке x, то он непрерывен и во всем L. Действительно, пусть $y_n \to y$, тогда $x + y_n - y \to x$ и из непрерывности функционала в точке x следует, что $f(x + y_n - y) \to f(x)$, откуда по линейности функционала заключаем, что $f(y_n) \to f(y)$. Обычно непрерывность линейного функционала проверяют в нуле (то есть, если для любой последовательности $x_n \to 0 \quad \Rightarrow f(x_n) \to 0$, то линейный функционал непрерывен).

Определение. Функционал f, заданный на линейном пространстве L со значениями в $\mathbf{R}_+ = [0, \infty)$, называется *нормой*, если

1) $f(x) = 0 \Leftrightarrow x = 0$;

2) $f(\alpha x) = |\alpha| f(x)$;

3) $f(x + y) \leq f(x) + f(y)$.

Линейное пространство, на котором задана некоторая норма, называется *нормированным пространством*. Норму элемента x принято обозначать $\|x\|$.

Всякая норма порождает в L и метрику

$$\rho(x,y) = ||x-y|| \ ,$$

то есть превращает нормированное пространство в метрическое. Обратное неверно. Нормированное пространство, полное по метрике порожденной нормой, называется *Банаховым* пространством.

Примеры нормированных пространств

1. Примером нормированного пространства может служить пространство функций $L^p_{[a,b]}$, $1 \le p < \infty$:

$$f \in L^p_{[a,b]} \Leftrightarrow ||f||_p \stackrel{\text{def}}{=} \Big(\int\limits_a^b |f(t)|^p dt \Big)^{\frac{1}{p}} < \infty \ . \tag{5}$$

2. Пространство непрерывных функций $C_{[a,b]}$ с нормой

$$||f|| = \max_{a \le t \le b} |f(t)| \ . \tag{6}$$

Пространство непрерывных функций полно по метрике, порожденной этой нормой.

Еще более содержательным об'ектом являются евклидовы пространства — пространства со скалярным произведением. В действительном линейном пространстве L *скалярное произведение* определяется как бинарная функция на L (в дальнейшем обозначаемая $\langle \cdot, \cdot \rangle$) со значениями в $\mathbf{R}$, удовлетворяющая следующим условиям:

1) $\langle x, y \rangle = \langle y, x \rangle$;

2) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$;

3) $\langle x, x \rangle \ge 0$, причем $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.

Действительное линейное пространство с фиксированным в нем скалярным произведением называется действительным *евклидовом пространством*.

Скалярное произведение в комплексном линейном пространстве L — это бинарная функция $\langle \cdot, \cdot \rangle$, определенная для любой пары элементов $x, y \in L$, со значениями в $\mathbf{C}$, удовлетворяющая следующим условиям:

1) $\langle x, y \rangle = \overline{\langle y, x \rangle}$;

2) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$;

3) $\langle x, x \rangle \ge 0$, причем $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.

Условия 2) и 3) совпадают для комплексных и действительных евклидовых пространств совпадают, различие лишь в условии 1).

Наконец, евклидово пространство называется *гильбертовым*, если оно сепарабельно (т.е. в нем существует счетный базис (напомним, множество называется счетным, если между ним и множеством натуральных чисел можно установить взаимно однозначное соответствие)) и полно по метрике, порожденной скалярным произведением. Пространство L^2 является гильбертовым [15].

Глава 2

Аппроксимации функций

Термин *аппроксимация* означает приближение. Функция f является аппроксимацией функции g, если она в том или ином смысле близка к g (скажем, по той или иной норме). В ситуации, когда функция f ищется так, чтобы она совпадала g в конечном наборе точек, то ее, равно как и сам процесс поиска, называют *интерполяцией*. При этом, если интерес представляют приближенные значения функции g (т.е. значения функции f), находящиеся вне отрезка с заданным набором точек (это касается лишь вещественных функций, разумеется), то наряду с термином интерполяция употребляется также термин *экстраполяция*.

2.1 Интерполяция

2.1.1 Задача интерполяции

Пусть задана таблица чисел $\{x_i, f_i\}_{i=0}^N$, $i = 0, 1, ..., N$; $x_0 < x_1 < ... < x_N$.

Определение. Всякая функция $f(x)$ такая, что $f(x_i) = f_i$; $i = 0, 1, ..., N$ называется *интерполирующей (интерполяцией)* для таблицы $\{x_i, f_i\}_{i=0}^N$.

Задача интерполяции состоит в отыскании (построении) интерполирующей функции (т.е. принимающей в заданных узлах интерполяции x_i заданные значения f_i) и принадлежащей заданному классу функций. Разумеется задача интерполяции может иметь или не иметь решение (и при том не единственное), все зависит от "заданного класса функций". Необходимо выяснить условия, при которых задача интерполяции была бы корректно поставлена. Один из способов интерполяции состоит в том, что интерполирующая функция ищется в виде линейной комбинации некоторых конкретных функций. Такая интерполяция называется линейной. Только линейные интерполяции мы и будем рассматривать в дальнейшем.

2.1.2 Чебышевские системы функций

Пусть $\{\varphi_i(x)\}_{i=0}^N$ — некоторый набор функций на $[a, b]$, скажем $\varphi_i(x) \in C_{[a,b]}$. Рассмотрим линейную оболочку $\mathbf{H} = \bigvee_{i=0}^{N} \varphi_i(x)$, она по определению состоит из функций представимых в виде $\sum_{i=0}^{N} a_i\varphi_i(x)$, где a_i — некоторые числа. Будем искать решение задачи интерполяции в классе функций, принадлежащих $\mathbf{H}$. Можно считать, что $\varphi_i(x)$ — линейно независимые функции (в противном случае, если задача интерполяции разрешима, то ее решение заведомо не единственно). Однако одного этого ограничения для

однозначной разрешимости недостаточно.

Примеры. Пусть задана таблица

x_i	0	1
f_i	0	1

.

1. Возьмем в качестве $\mathbf{H}$ оболочку синусоидальных функций $\mathbf{H} = \bigvee_{k=0}^{N} \sin(\pi k x)$. Всякая функция из $\mathbf{H}$ представляется в виде $f(x) = \sum_{k=0}^{N} a_k \sin \pi k x$ и, следовательно, для нее $f(0) = 0$, $f(1) = 0$ и она не удовлетворяет второму условию таблицы: $f(1) = 0 \neq 1$, таким образом решений нет.

2. Пусть теперь $\mathbf{H}$ — линейная оболочка степенных функций $\mathbf{H} = \bigvee_{k=0}^{N} x^k$ и $N \geq 2$, скажем $N = 2$. Тогда

$$f(x) = a_0 + a_1 x + a_2 x^2 \ .$$

Поскольку $f(0) = 0$, то $a_0 = 0$. Далее, $f(1) = 1$ и значит $a_1 + a_2 = 1$, и решений бесконечно много, лишь бы выполнялось последнее условие .

Нетрудно видеть, что во втором примере для существования и единственности решения можно использовать в качестве $\mathbf{H}$ функции вида $a_0 + a_k x^k$. То есть, для единственности решения задачи интерполяции естественно использовать следующее ограничение: число узлов должно равняться размерности интерполирующего пространства. Однако, как показывает первый пример, для разрешимости задачи интерполяции и этого ограничения недостаточно.

Выясним условия, при которых задача интерполяции разрешима однозначно. Задача линейной интерполяции выглядит следующим образом: пусть $f \in \mathbf{H}$, где H — линейная оболочка некоторых функций $\varphi_i(x)$, $i = 0, 1, 2, \cdots, N$, необходимо удовлетворить системе равенств

$$f(x_i) = f_i = \sum_{k=0}^{N} a_k \varphi_k(x_i) \ . \tag{1}$$

То есть, требуется найти набор чисел $\{a_k\}_{k=0}^{N}$ так, чтобы функция $f(x)$ удовлетворяла заданной таблице $\{x_i, f_i\}$. Слово "линейная"в формулировке означает, что функции φ_i входят в (1) линейным образом (или, что то же самое, функция f принадлежит линейной оболочке функций φ_i).

Обозначим матрицу $\{\varphi_k(x_i)\}$ через Φ. Пусть $\mathbf{f}$ — вектор с компонентами f_i , и $\mathbf{a} = (a_0, a_1, \ldots, a_N)^T$, тогда система (1) эквивалентна задаче

$$\Phi^T \mathbf{a} = \mathbf{f} \ .$$

Если $\det \Phi \neq 0$, то эта система разрешима единственным образом.

Определение. Система функций $\{\varphi_i\}$, для которой $\det \Phi \neq 0$, называется *чебышевской*.

Заметим, что любая чебышевская система функций автоматически линейно независима. Важным примером чебышевских систем являются многочлены.

2.1.3 Интерполяция многочленами

Особое место многочленов

Выделенность многочленов (полиномов) обусловлена целым рядом обстоятельств.

1) Полиномы $p_n(x)$ легко вычислять.

2) Множество полиномов плотно в пространстве непрерывных функций $C_{[a,b]}$, в силу теоремы Вейерштрасса, формулировку которой мы приведем.

Теорема Вейерштрасса. *Для любой функции* $f \in C_{[a,b]}$ *и для любого* $\varepsilon > 0$ *существует такое* n *и такой полином* $p_n(x)$, $\deg p_n(x) = n$, *что* $||f - p_n||_{C_{[a,b]}} < \varepsilon$.

3) Полиномы являются чебышевской системой для любой системы несовпадающих узлов.

В самом деле, пусть $\{\varphi_i(x)\}_{i=0}^N = \{1\,,\, x\,,\, x^2\,,\, \ldots\,,\, x^N\}$, тогда $\det\Phi$ совпадает с определителем Вандермонда

$$\Delta(x_0, x_1, \ldots, x_N) = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^N \\ 1 & x_1 & x_1^2 & \ldots & x_1^N \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \ldots & x_N^N \end{vmatrix} = \prod_{N \geq k \geq m \geq 0} (x_k - x_m)\ ,$$

который, очевидно, не равен нулю если $x_k \neq x_m$ при $k \neq m$. Убедимся в справедливости представления определителя Вандермонда по индукции [16]. Действительно, пусть для индекса равного $N-1$ последняя формула верна. Вычтем в определителе $\Delta(x_0, x_1, \ldots, x_N)$ из каждого столбца предшествующий, умноженный на x_0, тогда

$$\Delta(x_0, \ldots, x_N) = \begin{vmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & x_1 - x_0 & x_1^2 - x_1 x_0 & \ldots & x_1^N - x_1^{N-1} x_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N - x_0 & x_N^2 - x_N x_0 & \ldots & x_N^N - x_N^{N-1} x_0 \end{vmatrix} =$$

$$= (x_1 - x_0)(x_2 - x_0)\ldots(x_N - x_0) \begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{N-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \ldots & x_N^{N-1} \end{vmatrix} =$$

$$= (x_1 - x_0)(x_2 - x_0)\ldots(x_N - x_0) \prod_{N \geq k \geq m \geq 1} (x_k - x_m) = \prod_{N \geq k \geq m \geq 0} (x_k - x_m).$$

Таким образом, задача интерполяции для таблицы $\{x_i, f_i\}_{i=0}^N$ разрешима единственным образом в линейной оболочке степенных функций $\mathbf{H} = \bigvee\limits_{k=0}^{N} x^k$. Возникает вопрос: как строить интерполяционный полином $p_N(x)$, ведь есть свобода выбора базиса в $\mathbf{H}$ или, что то же самое, свобода формы записи. Брать в качестве $\varphi_k(x)$ собственно степени x^k зачастую оказывается неудобным. В частности, например, на отрезке $[a,b] = [0,1]$ степенные функции высоких порядков ведут себя весьма схожим образом: степени x^i и x^j "почти линейно зависимы"(они очень похожи друг на друга) и при этом получается почти вырожденная матрица Ф. Задача нахождения коэффициентов a_k при степенях x оказывается плохо обусловленной. Небольшое варьирование входных данных (значений f_i) приводит к значительным изменениям величин a_k. Если же в $\mathbf{H} = \bigvee\limits_{i=0}^{N} x^i$ выбрать другой базис, то это будет отвечать тому, что вместо определителя Вандермонда ($\det\Phi$) и самой матрицы Ф, необходимой для отыскания коэффициентов a_k в задаче

$$f(x_k) = a_0 + a_1 x_k + a_2 x_k^2 + \ldots + a_N x_k^N\ ,\quad k = 0\ ,\ 1\ ,\ \ldots\ ,\ N\ ,$$

мы будем иметь некоторую другую задачу

$$f(x_k) = b_0 p_0(x_k) + b_1 p_1(x_k) + \ldots + b_N p_N(x_k)\ ,\quad k = 0\,,\, 1\,,\, \ldots\,,\, N\ , \tag{2}$$

где $p_k \in \mathbf{H}$. Коэффициенты b_k определяются из равенства

$$f(x) = \sum_{i=0}^{N} b_i p_i(x) = \sum_{j=0}^{N} a_j x^k\ ,$$

при этом $p_i(x) = \sum C_{ik} x^k$, то есть

$$f(x) = \sum_{i=0}^{N} b_i \sum_{k=0}^{N} C_{ik} x^k = \sum_{k=0}^{N} \sum_{i=0}^{N} \left(b_i C_{ik} \right) x^k = \sum_{k=0}^{N} a_k x^k \ ,$$

или в матричной форме

$$C^T \mathbf{b} = \mathbf{a} \ .$$

Таким образом, если $\det C \neq 0$, то новая задача (2) так же разрешима единственным образом. Невырожденность C эквивалентна тому, что $\{p_k(x)\}_{k=0}^N$ образуют базис в $\mathbf{H}$ (следствие линейной алгебры). В частности, если полиномы $\{p_k(x)\}_{k=0}^N \subset \mathbf{H}$ таковы, что $\deg\, p_k = k$, то они автоматически линейно независимы и образуют базис в $\mathbf{H}$ и, следовательно, задача интерполяции разрешима единственным образом.

Интерполяционный полином в форме Лагранжа

Один из возможных подходов к решению задачи интерполяции многочленами, состоит в том, чтобы матрица Φ имела по возможности простой вид. Именно, рассмотрим задачу интерполяции : пусть дана интерполяционная таблица $\{x_k, f_k\}_{k=0}^N$. Требуется найти полином $p_N(x)$ степени N удовлетворяющий этой таблице.

Введем базис в $\mathbf{H} = \bigvee_{i=0}^{N} x^i$, в котором матрица Φ представляет собой единичную, обозначим его $\{\mathcal{L}_k(x)\}_{k=0}^N$, то есть

$$\mathcal{L}_k(x_i) = \delta_{ki} \, ; \, \Phi = I \ .$$

Отсюда $p_N(x) = \sum_{k=0}^{N} a_k \mathcal{L}_k(x)$, и

$$f_i = p_N(x_i) = \sum_{k=0}^{N} a_k \mathcal{L}_k(x_i) = a_i \ \ ,$$

или

$$p(x) = \sum_{k=0}^{N} f_k \mathcal{L}_k(x) \ .$$

Как построить лагранжевы полиномы $\mathcal{L}_k(x)$? Поскольку $\mathcal{L}_k(x_i) = 0$ при $i \neq k$, то такой полином $\mathcal{L}_k(x)$ имеет N корней и следовательно является полиномом степени N . Таким образом $\mathcal{L}_k(x) = C_k \prod_{i \neq k} (x - x_i)$, причем $\mathcal{L}_k(x_k) = 1$, поэтому $C_k = \prod_{i \neq k} \frac{1}{(x_k - x_i)}$, следовательно

$$\mathcal{L}_j(x) = \prod_{k \neq j} \frac{x - x_k}{x_j - x_k} \ .$$

Окончательно, решение задачи интерполяции принимает вид

$$p(x) = \sum_{j=0}^{N} f_j \prod_{k \neq j} \frac{x - x_k}{x_j - x_k} \ .$$

Интерполяционный полином в форме Ньютона

Интерполяционный полином степени N , проходящий через заданные $(N+1)$ точку $\{x_i, f_i\}_{i=0}^N$ единственнен. Однако запись его в форме Лагранжа может для некоторых задач оказаться неудобной. Это связано с тем обстоятельством, что все Лагранжевы полиномы $\mathcal{L}_k(x)$ имеют одну и ту же степень — N . В частности, если к интерполяционной сетке $\{x_i, f_i\}$ добавлять новые точки, то нельзя воспользоваться ранее построенными лагранжевыми полиномами, и приходится для более высоких степеней их строить заново.

Будем решать задачу интерполяции выбрав в $\mathbf{H}$ новый базис $\{\mathcal{N}_k(x)\}_{k=0}^N$:

$$\mathcal{N}_0(x) = 1 \ ; \ \mathcal{N}_k(x) = \prod_{i<k}(x - x_i) \ , \ k = 1 \, , \, \dots \, , \, N \ .$$

В том, что это действительно базис в $\mathbf{H}$ легко убедиться, поскольку $\deg \mathcal{N}_k(x) = k$, и тем самым ньютоновы многочлены $\mathcal{N}_i$ линейно независимы. Итак, будем искать интерполяционный полином $p(x)$ в виде

$$p(x) = \sum_{k=0}^{N} a_k \mathcal{N}_k(x) \ .$$

Такое представление решения и является записью интерполяционного полинома в форме Ньютона. Заметим, что $\mathcal{N}_k(x_j) = 0$ при $j < k$. Сами коэффициенты a_k находим из системы: $p(x_j) = f_j \ , \ j = 0 \, , \, \dots \, , \, N$ или

$$\begin{cases} a_0 = f_0 \ , \\ a_0 + a_1(x_1 - x_0) = f_1 \ , \\ \dots \\ \sum\limits_{k=0}^{m} a_k \mathcal{N}_k(x_l) = f_m \ , \\ \dots \\ \sum\limits_{k=0}^{N} a_k \mathcal{N}_k(x_N) = f_N \ . \end{cases}$$

Это треугольная система. Из первого уравнения определяется a_0 , затем, зная a_0 , из второго уравнения определяем a_1 , и так далее.

Можно решить ту же задачу и более "элегантно". Введем так называемые *разделенные разности*. Разделенные разности 0-го порядка — это просто значения функции $f_i = f(x_i)$. Разделенные разности более высоких порядков определяются рекуррентно:

1 порядка $f_{ij} = f(x_i, x_j) = \frac{f_i - f_j}{x_i - x_j}$;

2 порядка $f_{ijk} = f(x_i, x_j, x_k) = \frac{f_{ij} - f_{jk}}{x_i - x_k}$;

. . . . . .

k порядка $f_{\beta_0, \beta_1 \dots \beta_k} = \frac{f_{\beta_0 \beta_1 \dots \beta_{k-1}} - f_{\beta_1 \beta_2 \dots \beta_k}}{x_0 - x_k}$.

Нетрудно видеть, что разделенные разности имеют размерность соответствующих производных. Решение задачи интерполяции дается следующим выражением

$$p(x) = \sum_{k=0}^{N} f_{012 \, \dots \, k} \mathcal{N}_k(x) \ .$$

Чтобы убедиться в справедливости этого представления, рассмотрим разделенные разности интерполяционного полинома $p(x)$, в которых в качестве первого из аргументов выступает сама переменная x , а остальными являются точки интерполяции. Степень этого полинома равна N . Разность $p(x) - p(x_0) = p(x) - f_0$ обращается в ноль в точке x_0 и, следовательно, делится на $x - x_0$. Таким образом разделенная разность $p_{x0} = \frac{p(x) - p(x_0)}{x - x_0}$, рассматриваемая как функция x , является полиномом степени $N - 1$. Аналогично,

вторая разделенная разность $p_{x01} = \frac{p_{x0}-p_{01}}{x-x_1}$ есть полином по x степени $N-2$, разделенная разность N-го порядка $p_{x012\ldots N-1}$ уже не зависит от x и является константой, и разности более высокого порядка тождественно равны нулю. Таким образом,

$$p(x) = p_0 + (x-x_0)p_{x0} = p_0 + (x-x_0)[p_{01} + (x-x_1)p_{x01}] =$$

$$p_0 + (x-x_0)p_{01} + (x-x_0)(x-x_1)[p_{012} + (x-x_3)p_{x012}] = \ldots =$$
$$= \sum_{k=0}^{N} p_{012\ldots k} \prod_{i=0}^{k-1}(x-x_i) \ .$$

Осталось заметить, что поскольку в узлах интерполяции x_i значения интерполяционного полинома равны табличным значениям f_i, то и $f_{01\ldots k} = p_{01\ldots k}$.

2.1.4 Погрешность интерполяции

Пусть $f(x)$ — некоторая функция и пусть $\{x_i, f_i\}_{i=0}^N$ — интерполяционная таблица, которой эта функция удовлетворяет (то есть $f(x_i) = f_i$). По этой же интерполяционной сетке можно построить и интерполяционный полином $p_N(x)$. Возникает естественный вопрос: насколько различаются между собой функция $f(x)$ и полином $p_N(x)$, удовлетворяющие одной и той же таблице? Если никаких свойств гладкости не потребовать от функции f, то и сказать ничего определенного нельзя. Однако, при достаточной гладкости функции f можно оценить разность $f(x) - p_N(x)$, именно, справедлива

Теорема. *Пусть $f \in C^{N+1}[a,b]$ и p_N — интерполяционный полином, удовлетворяющие одной и той же сетке значений $\{x_i, f_i\}_{i=0}^N$, тогда для любой точки $x \in [a,b]$ существует такая точка $\xi(x)$, что*

$$f(x) - p_N(x) = \frac{f^{N+1}(\xi(x))}{(N+1)!}\mathcal{N}_{N+1}(x) \ ,$$

где $\mathcal{N}_{N+1}(x) = (x-x_0)(x-x_1)\ldots(x-x_N)$.

Доказательство. Представим погрешность в виде

$$f(x) - p_N(x) = \mathcal{N}_{N+1}(x)r(x) \ .$$

Такое представление естественно, поскольку и разность $f - p_N$ и $\mathcal{N}_{N+1}$ в точках x_i, $i = 0, 1, \ldots, N$ обращаются в ноль:

$$[(x) - p_N(x)]|_{x=x_i} = 0 \quad , \quad i = 0\,,\, 1\,,\, \ldots\,,\, N \ .$$

При этом $r(x) \in C_{[a,b]}$. Введем также вспомогательную функцию

$$q(\xi) = f(\xi) - p_N(\xi) - \mathcal{N}_{N+1}(\xi)r(x) \ .$$

Здесь x — параметр, $\xi \in [a,b]$. Очевидно, что $q(\xi) = 0$ в точках $\xi = x_0\,,\, x_1\,,\, \ldots\,,\, x_N\,,\, x$. Далее, если $f \in C^{N+1}$ то и $q \in C^{N+1}$. Напомним, что для функции, принадлежащей C^1, между двумя корнями имеется по крайней мере один нуль производной. Следовательно, между крайними из $N+2$ нулями функции $q(\xi)$ лежит хотя бы один нуль $(N+1)$-ой производной. Выпишем эту производную:

$$q^{N+1}(\xi) = f^{N+1}(\xi) - (N+1)!r(x) \ .$$

Пусть она обращается в нуль в точке $\xi(x)$: $q(\xi(x)) = 0$ и, следовательно, в этой точке $\xi(x)$ выполнено

$$r(x) = \frac{f^{N+1}(\xi(x))}{(N+1)!} \ ,$$

откуда утверждение теоремы следует непосредственно.

Замечание. Приведенным рассуждением о корнях вспомогательной функции q можно воспользоваться только, если $x \neq x_i$, так как при $x = x_i$ функция $q(x)$ имеет лишь $N+1$ корень. Однако при $x = x_i$ условия теоремы выполнены автоматически, поскольку $f(x_i) = p_N(x_i)$ и $\mathcal{N}_{N+1}(x_i) = 0$.

Из условия теоремы следует априорная оценка

$$|f(x) - p_N(x)| \leq \max_{\xi \in [a,b]} \left[\frac{|f^{N+1}(\xi)|}{(N+1)!} |\mathcal{N}_{N+1}(x)| \right] \leq \frac{||f^{N+1}||_C}{(N+1)!} |\mathcal{N}_{N+1}(x)| \ .$$

Пример. Оценить погрешность функции $y = \sqrt{x}$ на промежутке [100,144] с узлами 100, 121, 144 с помощью интерполяционного полинома второй степени (в форме Лагранжа или Ньютона — это все равно, поскольку это один и тот же полином, разница может возникнуть только, если вычисление коэффициентов происходит не точно, а с некоторой погрешностью).

Решение. Для того, чтобы оценить погрешность вовсе нет необходимости строить сам интерполяционный полином, достаточно воспользоваться полученной оценкой. Итак $N = 2$, $y' = \frac{1}{2}x^{-\frac{1}{2}}$, $y'' = -\frac{1}{4}x^{-\frac{3}{2}}$, $y''' = \frac{3}{8}x^{-\frac{5}{2}}$, следовательно
$\max|y'''| \leq \frac{3}{8}(100)^{-\frac{5}{2}} = \frac{3}{8}10^{-5}$, откуда

$$|p_N(x) - y(x)| \leq \frac{3}{8}\, 10^{-5}\, \frac{1}{3!} \max|(x-100)(x-121)(x-144)| < 3 \cdot 10^{-3} \ .$$

Таким образом, даже не считая сам интерполяционный полином $p_N(x)$ мы оценили погрешность.

2.1.5 Оценка $\mathcal{N}_{N+1}(x)$.

При произвольном расположении узлов оценить модуль $\mathcal{N}_{N+1}$ довольно сложно. Для равномерной сетки ситуация выглядит проще. Проведем грубую оценку. Пусть $x \in [x_{k-1}, x_k]$, тогда

$$|x_0 - x| \leq kh \, , \, |x_1 - x| \leq (k-1)h \, , \, \ldots \, , \, |x_{k-1} - x| \leq h \ ,$$

$$|x_k - x| \leq h \, , \, |x_{k+1} - x| \leq 2h \, , \, \ldots \, , \, |x_N - x| \leq (N-k+1)h \ ,$$

откуда $|\mathcal{N}_{N+1}| \leq (N-k+1)!k!h^{N+1}$, и

$$||f - p_n||_C \leq ||f^{N+1}||_C \underbrace{\frac{k!(N+1-k)!}{(N+1)!}}_{1/C_{N+1}^k} h^{N+1} \ ,$$

то есть $|f - p_N| = O(h^{N+1})$. В этой ситуации говорят, что интерполяционный многочлен $p_N(x)$ имеет погрешность $O(h^{N+1})$.

Замечание. Можно подобрать узлы так, чтобы величина $\max|\mathcal{N}_{N+1}(x)|$ была меньше, чем у любого другого полинома той же степени с единичным старшим коэффициентом (такие полиномы наименее отклоняющиеся от нуля — многочлены Чебышева). Узлы расположены редко в середине и сгущаются у концов промежутка.

2.1.6 Сходимость интерполяции. Примеры

Хотя теорема Вейерштрасса и утверждает полноту полиномов, однако она ничего не говорит относительно того, как строить такие полиномы p_n . Возникают вопросы:

1. Как выбирать интерполяционную таблицу $\{x_i, f_i\}$?

2. Сходится ли в том или ином смысле последовательность аппроксимационных полиномов к интерполируемой функции?

Для увеличения точности можно использовать следующие методы построения полинома:

1. Уменьшение шага сетки, при постоянной степени N интерполяционного полинома. В этой ситуации интерполяционный полином хорошо описывает поведение функции $f \in C^{N+1}$ лишь на небольшом промежутке (длины hN).

2. Разумное размещение узлов. Обычно это означает выбор в качестве узлов корней многочленов Чебышева.

3. Увеличение числа узлов и, тем самым, увеличение степени интерполяционного полинома.

Известно, что если $y(x)$ — целая функция (т.е. разлагается в степенной ряд с бесконечным радиусом сходимости на комплексной плоскости), то при произвольном расположении узлов на любом промежутке $[a, b]$, $p_N(x) \to y(x)$ равномерно (т.е. по норме пространства непрерывных функций) при $N \to \infty$. Однако если функция бесконечно дифференцируема лишь в вещественном смысле: $f \in C^{\infty}_{(-\infty,\infty)}$, то это уже не гарантирует сходимости последовательности интерполяционных полиномов к функции f при увеличении числа узлов.

Пример.

$$f(x) = \begin{cases} 0 & x \le 0 \ , \\ e^{-\frac{1}{x}} & 0 < x \ . \end{cases}$$

Построим последовательность интерполяционных полиномов по точкам отрицательной полуоси. Все они тождественно равны нулю $p_n(x) \equiv 0$ и сходимости к функции f нет. Правда узлы мы выбрали весьма неэффективно.

При равномерном расположении узлов также не всегда удается добиться сходимости. Причина здесь заключается в том, что в оценку интерполяции входит производная от интерполируемой функции. В случае если f не обладает достаточной гладкостью, то и оценка теряет смысл.

Пример Бернштейна.

$$y(x) = |x| \ , \quad x \in [-1, 1] \ .$$

Бернштейн показал, что для равномерной сетки значения $p_N(x)$ между узлами интерполяции неограниченно возрастают при $N \to \infty$ в окрестности точек -1, 1. Заметим, что функция $|x|$ недифференцируема в нуле, но в окрестности нуля интерполяционные полиномы высокой степени достаточно хорошо передают поведение функции модуль.

В известном смысле общего метода построения последовательности интерполяционных полиномов нет. И основанием, чтобы утверждать столь "пренеприятнейшее известие", является

Теорема Фабера (NO GO Theorem). *Пусть x_i^j — произвольный интерполяционный массив на $[a, b]$:*

$$
\begin{array}{cccccc}
x_0^0 & & & & & \\
x_0^1 & x_1^1 & & & & \\
x_0^2 & x_1^2 & x_2^2 & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
x_0^n & x_1^n & x_2^n & \dots & x_n^n & \\
\dots & \dots & \dots & \dots & \dots & \dots
\end{array} .
$$

Тогда существует такая функция $g \in C_{[a,b]}$ *и такая точка* $x_* \in [a,b]$ *, что последовательность интерполяционных полиномов, построенная по строкам этого массива и совпадающая в них с* g*, не стремится в точке* x_* *к* $g(x_*)$.

Таким образом, равномерной сходимости, вообще говоря, добиться не удается. Как "преодолеть"теорему Фабера? Необходимо отказаться от поточечной сходимости и заменить ее на сходимость в среднем. Именно, верно следующее утверждение.

Пусть $P_n(x)$ — система многочленов ортогональных с весом $\rho \in C_{[a,b]}$ на промежутке $[a,b]$:

$$
\int_a^b P_n(x)P_m(x)\rho dx = \delta_{nm} , \quad \rho(x) > 0 .
$$

Пусть $x_m^{(n)}$ суть корни P_n. Все они вещественные и простые и принадлежат промежутку (a,b) (см. гл. "Численное интегрирование"). Возьмем корни $x_m^{(N+1)}$ ортогонального полинома P_{N+1} в качестве узлов интерполяции, и по ним построим полином p_N N-ой степени проходящий через $N+1$ точку: $f(x_m^{(N+1)}) = p_N(x_m^{(N+1)}) , m = 0 , \dots , N$. Тогда для любой непрерывной на $[a,b]$ функции f

$$
\int_a^b [f(x) - p_N(x)]^2 \rho(x) dx \underset{N\to\infty}{\longrightarrow} 0 .
$$

Выбирая ту или иную весовую функцию, получаем различные ортогональные полиномы. Наиболее употребительными ортогональными полиномами являются полиномы Якоби, Лежандра, Чебышева, Лагерра, Эрмита.

2.1.7 Сплайны

Как мы видели, увеличение степени интерполирующего полинома далеко не всегда приводит к желаемому результату. Зачастую более эффективным способом интерполяции на сетке $\{x_i, f_i\}_{i=0}^N$ оказывается использование сплайнов. Дадим соответствующие определения.

Пусть $x_0 < x_1 < x_2 \cdots < x_N$ — некоторые числа. Рассмотрим кусочно полиномиальную функцию S_n^ν ($\nu \le n$, ν, n натуральные), заданную на промежутке $[x_0, x_N]$ такую, что на каждом промежутке $[x_{i-1}, x_i]$ она представляет собой некоторый полином p_n^i степени n :

$$
S_n^\nu(x) = p_n^i(x), \quad x \in [x_{i-1}, x_i], \quad i = 1, 2, \cdots N ,
$$

и, при этом, рассматриваемая на всем промежутке $[x_0, x_N]$ функция S_n^ν имеет $n - \nu$ непрерывных производных, то есть $S_n^\nu \in C^{n-\nu}_{[x_0,x_N]}$, и, следовательно, для полиномов p_n^i во всех внутренних точках $x_1, x_2, \cdots x_{N-1}$ промежутка $[x_0, x_N]$ выполнено

$$
\frac{d^k}{dx^k} p_n^i|_{x=(x_i)} = \frac{d^k}{dx^k} p_n^{i+1}|_{x=(x_i)} , \quad i = 1, 2, \cdots, N-1 , \quad k = 0, 1, 2, \cdots n - \nu .
$$

Определение. Функция $S_n^\nu(x)$ называется *сплайном порядка n (степени n) дефекта ν*, . Точки x_i называются *узлами сплайна*.

Очевидно, что дефект сплайна ν равен его порядку n "минус" число его непрерывных производных.

Существует ли хотя бы один сплайн? Разумеется, да. В частности полином n-ой степени есть одновременно сплайн S_n^ν для всех ν от 0 до n . В качестве примера сплайна также отметим, что S_1^1 представляет собой ломанную, точками излома которой являются узлы сплайна.

Всякий сплайн S_n^ν , до тех пор пока мы от него ничего не потребовали кроме как являться кусочно полиномиальной функцией, обладает некоторым числом свободных параметров, которыми мы можем распоряжаться по своему усмотрению. Чему равно это число? У нас имеется N промежутков $\Delta_k = [x_{k-1}, x_k]$, $k = 1, 2, \cdots, N$. На каждом из этих промежутков сплайн S_n^ν должен представлять собой некоторый полином n-степени $p_n^k = \sum_{j=0}^{n} a_j^{(k)} x^j$, который имеет $n+1$ свободный параметр. Таким образом, общее число свободных параметров равно $N(n+1)$. Однако, одновременно с этим, на сам сплайн наложено некоторое количество условий гладкости во внутренних узлах сплайна $x_1, x_2, \cdots, x_{N-1}$. Сплайн должен быть непрерывен и непрерывными должны быть $n-\nu$ его производных в $N-1$ точке $\{x_i\}_{i=1}^{N-1}$, то есть из общего числа параметров $N(n+1)$ мы должны вычесть число условий гладкости, равное $(N-1)(n-\nu+1)$. В итоге, число F действительно свободных параметров равно

$$F = \nu(N-1) + n + 1 \ .$$

Пусть теперь нам задана некоторая интерполяционная таблица $\{x'_i, y_i\}_{i=0}^{M}$ (точки x'_i — это узлы интерполяции, и они вовсе не обязаны совпадать с узлами сплайна x_i), и мы хотим найти сплайн S_n^ν , который бы этой таблице удовлетворял: $S_n^\nu(x'_i) = y_i$, $i = 0\,,\, 1\,,\, \ldots\,,\, M$. Существует ли решение такой задачи интерполяции и единственно ли оно? Если число свободных параметров сплайна $F = \nu(N-1) + n + 1$ совпадает с числом $M+1$ условий интерполяции (условия равенства конкретным значениям f_i в $M+1$ узле интерполяции), то можно надеяться, что ответ положительный, хотя это и не всегда так, и ответ, в частности, зависит от взаимного расположения узлов интерполяции и узлов сплайна. Так, например, если между двумя соседними узлами x_{i-1} и x_i сплайна S_n^ν находится более чем $n+1$ узел интерполяции, то задача некорректна, поскольку от полинома n-ой степени неестественно требовать, чтобы он проходил более чем через $n+1$ заданную точку $\{x'_i, f_i\}$. Кроме того, на практике два условия обычно резервируют под граничные значения сплайна или его производных. Скажем интерполируемая функция известна нам в некотором количестве точек и при этом удовлетворяет некоторому дифференциальному уравнению и граничным условиям. Можно в принципе потребовать чтобы этим же условиям удовлетворял и интерполирующий её сплайн (хотя никакому уравнению он, разумеется, не удовлетворяет). Для некоторых конкретных сплайнов (например, кубический сплайн S_3^1) есть более естественные соображения, по которым имеет смысл два условия использовать как граничные. Чуть позже мы этого коснемся.

Параболический сплайн S_2^1

Для параболического сплайна число свободных параметров $F = \nu(N-1)+n+1 = N+2$ и его используют, чтобы удовлетворить интерполяционной таблице $\{x'_i, f_i\}$ с N узлами интерполяции, оставляя два параметра под граничные условия, которые задаются в крайних узлах сплайна x_0 и x_N. Узлы интерполяции

x'_i располагают между соседними узлами сплайна x_i и x_{i+1}:

$$x'_i \in (x_i, x_{i+1}) \ . \tag{3}$$

Справедлива теорема (приводимая нами без доказательства) которая утверждает, что при выполнении условия (3) задача интерполяции параболическим сплайном корректна, т.е. для любой интерполяционной таблицы $\{x'_i, f_i\}_{i=0}^N$ интерполирующий её сплайн существует и единственнен при граничных условиях вида

$$\alpha_1 S(a) + \beta_1 S'(a) = \gamma_1 \ , \quad \alpha_2 S(b) + \beta_2 S'(b) = \gamma_2 \ ,$$

$$\alpha_i^2 + \beta_i^2 \neq 0 \ , \quad i = 1, 2, \quad a = x_0, \quad b = x_N \ .$$

Для параболического сплайна узлы интерполяции и узлы сплайна не совпадают, и это обстоятельство можно использовать для повышения точности интерполяции функции $f(x)$. Именно, поскольку парабола не имеет точек перегиба, то естественно точки перегиба интерполируемой функции $f(x)$ выбирать в качестве узлов сплайна, а точки локальных экстремумов $f(x)$ — в качестве узлов интерполяции.

Задача интерполяции кубическим сплайном $S_3^1(x)$

Пусть нам задана интерполяционная таблица $\{x_i, y_i\}_0^N$ и требуется найти кубический сплайн $S_3^1(x)$, узлы которого совпадают с узлами интерполяции, и который бы этой таблице удовлетворял: $S_3^1(x_i) = y_i$, $i = 0\,, 1\,, \ldots\,, N$. Для решения этой задачи прежде всего определим число свободных параметров и количество условий, которым необходимо удовлетворить.

Для сплайна S_n^ν число свободных параметров F равно

$$F = \nu(N-1) + n + 1 = 1(N-1) + 3 + 1 = N + 3 \ .$$

При этом необходимо удовлетворить $(N+1)$-му условию равенства сплайна значениям интерполяционной таблицы. Два оставшихся свободных параметра используют под граничные значения. Перечислим наиболее употребительные граничные условия для кубического сплайна.

1. $S_3^{1\prime\prime}(x_0) = S_3^{1\prime\prime}(x_N) = 0$ — естественный (натуральный) сплайн;
2. $S_3^{1\prime\prime}(x_0) = A\,,\ S_3^{1\prime\prime}(x_N) = B$;
3. периодический сплайн $S^{(\rho)}(a) = S^{(\rho)}(b)\,,\ \rho = 0\,, 1\,, 2\ .$

Вопрос. Почему в случае периодического сплайна указано 3 условия, а не 2 ... или их все таки 2?

Свойство минимальной кривизны

Выделенность естественного сплайна обусловлена тем, что он имеет минимальную среднюю кривизну среди всех функций, удовлетворяющих заданной интерполяционной таблице. Именно, справедлива

Теорема (Холидей). *Пусть* $\Delta = [a, b]$, $a = x_0 < x_1 < \ldots < x_N = b$ *и* $\{y_i\}_{i=0}^N$ *— некоторые числа. Среди всех дважды непрерывно дифференцируемых функций* F *, таких что* $f(x_i) = y_i$, $f''(a) = f''(b) = 0$ *на естественном сплайне* S_3^1 *достигается минимум функционала* $\mathcal{F}(f) = \int\limits_a^b (f''(x))^2 dx$.

Доказательство. Пусть $f \in C^2_{[a,b]}$ и удовлетворяет таблице $\{x_i, y_i\}_{i=0}^N$. Обозначим для удобства $S(x) = S_3^1(x)$. Рассмотрим разность

$$\mathcal{F}(f) - \mathcal{F}(S) = \int_a^b \{(f'')^2 - (S'')^2\} dx =$$

$$= \int_a^b (f'' - S'')^2 + \int_a^b (2f''S'' - 2S''^2) dx = \mathcal{I}_1 + 2\mathcal{I}_2 \ .$$

Очевидно $\mathcal{I}_1 \geq 0$. Рассмотрим второй интеграл:

$$\mathcal{I}_2 = \int_a^b (f'' - S'')S'' dx = \sum_{i=1}^N \int_{x_{i-1}}^{x_i} (f'' - S'')S'' dx \ .$$

Возьмем его по частям:

$$\sum_{i=1}^N (f' - S')S''\Big|_{x_{i-1}}^{x_i} - \sum_{i=1}^N \int_{x_{i-1}}^{x_i} (f' - S')S''' dx \ .$$

Первая сумма равна 0 из граничных условий, поэтому

$$\mathcal{I}_2 = -\sum_{i=1}^N S_i''' \int_{x_{i-1}}^{x_i} (f' - S') dx = \sum_{i=1}^N S_i''' (f - S)\Big|_{x_{i-1}}^{x_i} = 0 \ ,$$

поскольку значение третьей производной на i-ом промежутке постоянно, а f и S интерполируют таблицу. Таким образом, $\mathcal{F}(f) - \mathcal{F}(S) \geq 0$, что и требовалось доказать.

Существование и единственность кубического сплайна

Пока мы знаем, что для сплайна $S_3^1(x)$ с заданными граничными условиями количество неизвестных, которые необходимо определить (т.е. свободных параметров) совпадает с количеством уравнений. Таким образом в принципе сплайн $S_3^1(x)$ может существовать. Однако совпадение количества уравнений и количества неизвестных не гарантирует ни существования ни единственности решения. Рассмотрим подробно этот вопрос.

Возьмем вторую производную от сплайна S_3^1, который для сокращения записи будем обозначать просто $S(x)$. $S''(x)$ — это ломаная, или кусочно-линейная непрерывная функция. Обозначим через M_i значения второй производной от сплайна в точках x_i: $M_i = S''(x_i)$, $i = 0\,,\, 1\,,\, \ldots\,,\, N$. Поскольку на любом промежутке $[x_{i-1}, x_i]$ $S''(x)$ является линейной функцией, проходящей через точки (x_{i-1}, M_{i-1}) и (x_i, M_i), то она очевидно имеет вид:

$$S''(x) = M_i \frac{x - x_{i-1}}{x_i - x_{i-1}} + M_{i-1} \frac{x_i - x}{x_i - x_{i-1}}, \quad x \in [x_{i-1}, x_i] \ .$$

Обозначим $x_i - x_{i-1} = h_i$, $i = 1\,,\, \ldots\,,\, N$, тогда

$$S''(x) = \frac{M_i}{h_i}(x - x_{i-1}) + \frac{M_{i-1}}{h_i}(x_i - x) \ .$$

Интегрируя по промежутку $[x_{i-1}, x_i]$, получаем

$$S'(x) = \frac{M_i}{2h_i}(x - x_{i-1})^2 - \frac{M_{i-1}}{2h_i}(x_i - x)^2 + d_i \ ,$$

и интегрируя еще раз, представим сплайн в виде

$$S(x) = \frac{M_i}{6h_i}(x - x_{i-1})^3 + \frac{M_{i-1}}{6h_i}(x_i - x)^3 + \frac{d_i}{2}(x - x_{i-1}) - \frac{d_i}{2}(x_i - x) + c_i \ .$$

Константы d_i и c_i пока неизвестны. Чтобы их найти воспользуемся собственно уравнениями интерполяции $S(x_{i-1}) = y_{i-1}$ и $S(x_i) = y_i$, которые в нашем случае принимают вид

$$\frac{M_{i-1}}{6}h_i^2 - \frac{d_i}{2}h_i + c_i = y_{i-1} \quad , \quad \frac{M_i}{6}h_i^2 + \frac{d_i}{2}h_i + c_i = y_i \ .$$

Складывая и вычитая эти уравнения получаем

$$c_i = \frac{y_i + y_{i-1}}{2} - \frac{M_i + M_{i-1}}{12}h_i^2 \quad , \quad d_i = \frac{y_i - y_{i-1}}{h_i} - \frac{M_i - M_{i-1}}{6}h_i \ .$$

Таким образом c_i и d_i можно определить через величины M_i , если бы последние были известны. Чтобы их определить, используем непрерывность $S'(x)$ во внутренних узлах $x_1 , \dots , x_{N-1}$. Эти условия записываются в виде

$$\frac{M_i h_i}{2} + d_i = -\frac{M_i h_{i+1}}{2} + d_{i+1} \ .$$

Подставим сюда выражение для величин d_i:

$$\frac{M_i}{2}h_i \ + \ \frac{y_i - y_{i-1}}{h_i} \ - \ \frac{M_i - M_{i-1}}{6}h_i =$$

$$= -\frac{M_i}{2}h_{i+1} \ + \ \frac{y_{i+1} - y_i}{h_{i+1}} \ - \ \frac{M_{i+1} - M_i}{6}h_{i+1} \ ,$$

то есть

$$\frac{h_i}{h_i + h_{i+1}}M_{i-1} + 2M_i + \frac{h_{i+1}}{h_i + h_{i+1}}M_{i+1} = \frac{6}{h_i + h_{i+1}}\left(\frac{y_{i+1} - y_i}{h_{i+1}} \ - \ \frac{y_i - y_{i-1}}{h_i}\right) \ ,$$

$i = 1 , 2 , \dots , N-1$. Это система из $N-1$ уравнения с $(N+1)$-ой неизвестной величиной $M_0 , M_1 , \dots , M_N$. Ее необходимо дополнить двумя уравнениями исходя из граничных условий . Возьмем, скажем, однородные граничные условия

$$\begin{cases} M_0 = 0, \\ M_N = 0. \end{cases}$$

Матрица, соответствующая системе полученных уравнений, трехдиагональна, и при этом является матрицей с диагональным преобладанием. Напомним, что квадратная матрица D (вообще говоря комплексная) называется матрицей с *диагональным преобладанием* если для элементов d_{ij} любой строки выполнено

$$|d_{ii}| > \sum_{j \neq i}^{N} |d_{ij}| \ .$$

Приведем без доказательства следующее утверждение.

Теорема (Гершгорин). *Собственные значения квадратной матрицы* $D = \{d_{ij}\}_{i,j=1}^{N}$ *лежат в объединении кругов*

$$|z - d_{ii}| \leq \sum_{j \neq i} |d_{ij}| \ .$$

Прямым следствием теоремы Гершгорина является невырожденность матриц с доминирующей главной диагональю, поскольку для таких матриц указанное объединение кругов не содержит точку $z = 0$ и, следовательно, матрица не имеет нулевого собственного значения, а значит и невырождена. Таким образом система уравнений для определения величин M_i однозначно разрешима, тем самым существование и единственность кубического сплайна можно считать доказанными. Саму же возникшую трехдиагональную систему удобно решать методом прогонки, который рассматривается в соответствующей главе. Там же показано (независимо от теоремы Гершгорина), что для матриц с диагональным преобладанием метод прогонки заведомо разрешим.

Базис в пространстве сплайнов с однородными граничными условиями

Для всякой интерполяционной таблицы

$$x_0\,,\,x_1\,,\,\ldots\,,\,x_N$$

$$y_0\,,\,y_1\,,\,\ldots\,,\,y_N$$

существует единственный кубический сплайн $S_3^1(x)=S(x)$, который ей удовлетворяет

$$S(x_i)=y_i\,,\quad ß=0\,,\,1\,,\,\ldots\,,\,N\ ,$$

и удовлетворяет однородным граничным условиям (так называемый *натуральный* или *естественный* сплайн)

$$S''(x_0)=S''(x_N)=0\ .$$

Варьируя величины y_i (считая, что узлы x_i фиксированы) мы получим пространство $M(x_0\,,\,x_1\,,\,\ldots\,,\,x_N)$ интерполяционных естественных сплайнов размерности $\dim=N+1$ с узлами $\{x_i\}_0^N$. В самом деле, введем в качестве базиса в M следующие сплайны: $\{S_k(x)\}_{k=0}^N:\ S_k(x_i)=\delta_{ik}\ \ i=1\,,\,\ldots\,,\,N$. Каждый такой сплайн $S_k(x)$ существует и единственнен. Рассмотрим комбинацию

$$S(x,\{\alpha_i\}_{i=0}^N)\equiv\sum_{k=0}^N\alpha_kS_k(x)\ .$$

Предположим, что $S(x,\{\alpha\})\equiv 0$, тогда

$$S(x_i,\{\alpha\})=S_i(x_i)\alpha_i=0$$

и, следовательно, все $\alpha_i=0$, т.е. сплайны $S_i(x)$ линейно независимые, и они покрывают все M. Следовательно любой кубический сплайн $S\in M(x_0\,,\,\ldots\,,\,x_N)$ единственным образом представим в виде

$$S(x,\{y\})=\sum_{k=0}^N y_kS_k(x)\ .$$

А как быть в случае сплайнов с ненулевыми граничными условиями, скажем с условиями $S''(x_0)=A$, $S''(x_N)=B$? Множество таких сплайнов уже не образует пространство, поскольку не является линейным (сумма таких сплайнов не удовлетворяет граничным условиям). Чтобы описать эту ситуацию построим сплайн специального вида $S^{(0,N)}(x)$ такой, что

1) $S^{(0,N)}(x_i)=0\ ,\ \ i=0,1,2,\ldots,N$;

2) $S^{(0,N)}(x)$ удовлетворяет заданным неоднородным граничным условиям.

Такой сплайн существует и единственнен по доказанному нами утверждении о существовании и единственности кубического сплайна. Произвольный сплайн с узлами $\{x_i\}_0^N$ и с неоднородными граничными условиями представим единственным образом в виде

$$S(x)=S^{(0,N)}(x)+\sum_{k=0}^N y_kS_k(x)\ ,$$

где S_k ранее построенные базисные сплайны пространства естественных сплайнов. Таким образом кубические сплайны описаны полностью.

2.2 Аппроксимации Паде

2.2.1 "Наивный"подход

Приближение функции с помощью интерполяционного полинома или с помощью сплайна основано на использовании значений интерполируемой функции в некотором количестве точек, и как сплайн так и интерполяционный полином должны в этих точках иметь значения, совпадающие с соответствующими значениями интерполируемой функции. Можно, однако, рассматривать приближения не связанные жестко со значениями функции в наборе точек. В частности отрезок ряда Тейлора-Маклорена $\sum\limits_{k=0}^{N} \frac{f^{(k)}(x-x_0)}{k!}(x-x_0)^k$ может достаточно хорошо приближать функцию в окрестности точки разложения x_0 (с точностью $o((x-x_0)^N)$) и, при этом, не быть связанным с интерполяционной таблицей (он определяется лишь значениями производных в одной единственной точке x_0). Отрезок ряда Тейлора-Маклорена представляет собой один из способов *аппроксимации*. Более общими аппроксимациями являются *аппроксимации Паде* [6].

Пусть функция f (вообще говоря комплекснозначная) задана своим рядом Тейлора. Для удобства будем считать, что точкой разложения является нуль

$$f(z) = \sum_{i=0}^{\infty} c_i z^i \ . \tag{4}$$

На этот ряд можно смотреть и как на формальный (т.е. возможно и не сходящийся нигде ни к какой функции).

Дадим сначала предварительное определение, а точное — несколько позже.

"Наивное"определение. Назовем $[L/M]_f$-*аппроксимацией Паде* отношение двух полиномов

$$[L/M]_f = \frac{\sum\limits_{i=0}^{L} p_i z^i}{\sum\limits_{i=0}^{M} q_i z^i} \quad , \quad q_0 = 1 \ , \tag{5}$$

разложение в ряд Тейлора которого, совпадает с первыми коэффициентами ряда f настолько, насколько это возможно.

Всего мы имеем $L+M+1$ свободных параметров (т.к. $q_0 = 1$), следовательно, можно надеяться на то, что их выбором удастся добиться выполнения следующего равенства

$$\sum_{i=0}^{\infty} c_i z^i = \frac{\sum\limits_{i=0}^{L} p_i z^i}{\sum\limits_{i=0}^{M} q_i z^i} + O(z^{L+M+1}) \ . \tag{6}$$

Пример. Пусть

$$f(z) = 1 - \frac{1}{2}z + \frac{1}{3}z^2 - \frac{1}{4}z^3 + \cdots \ .$$

Легко видеть, что

$$[1/0]_f = 1 - \frac{1}{2}z = f(z) + O(z^2) \ , \quad [0/1]_f = \frac{1}{1+\frac{1}{2}z} = f(z) + O(z^2) \ ,$$

$$[1/1]_f = \frac{1+az}{1+bz} = (1+az)(1-bz+b^2z^2-\cdots) = 1 + (a-b)z + (b^2-ab)z^2 + \cdots \ ,$$

откуда $a-b = -1/2$, $b(b-a) = 1/3$, то есть $a = 1/6$, $b = 2/3$, и

$$[1/1]_f = \frac{1+\frac{1}{6}z}{1+\frac{2}{3}z} = f(z) + O(z^3) \ .$$

Домножим равенство (6) на полином Q.

$$\left(\sum_{i=0}^{M} q_i z^i\right)\left(\sum_{i=0}^{\infty} c_i z^i\right) = \sum_{i=0}^{L} p_i z^i + O(z^{L+M+1}) \ . \tag{7}$$

Заметим, что коэффициенты при степенях z^{L+1}, z^{L+2}, $\cdots$, z^{L+M} равны нулю. Сосчитаем коэффициенты при этих степенях, положив $c_j = 0$ при $j < 0$.

$$\begin{aligned}
z^{L+1} \ &: \quad c_{L+1}q_0 + c_L q_1 + \cdots + c_{L-M+2}q_{M-1} + c_{L-M+1}q_M = 0 \ , \\
z^{L+2} \ &: \quad c_{L+2}q_0 + c_{L+1}q_1 + \cdots + c_{L-M+3}q_{M-1} + c_{L-M+2}q_M = 0 \ , \\
&- \\
z^{L+M} \ &: \quad c_{L+M}q_0 + c_{L+M-1}q_1 + \cdots + c_{L+1}q_{M-1} + c_L q_M = 0 \ .
\end{aligned} \tag{8}$$

Поскольку q_0 мы положили равным 1, то имеем M уравнений на M неизвестных коэффициентов q_i, которые удобно записать в матричной форме:

$$\begin{pmatrix}
c_L & c_{L-1} & \cdots & c_{L-M+2} & c_{L-M+1} \\
c_{L+1} & c_L & \cdots & c_{L-M+3} & c_{L-M+2} \\
\cdots & \cdots & \ddots & \cdots & \cdots \\
c_{L+M-1} & c_{L+M-2} & \cdots & c_{L+1} & c_L
\end{pmatrix}
\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_M \end{pmatrix}
= -
\begin{pmatrix} c_{L+1} \\ c_{L+2} \\ \vdots \\ c_{L+M} \end{pmatrix} .$$

Если определитель матрицы, фигурирующей в этой линейной системе, не обращается в нуль, то из неё можно определить коэффициенты q_i . Найдя их и приравнивая в (7) коэффициенты при $1, z, z^2, \cdots, z^L$, находим и коэффициенты полинома P :

$$\begin{gathered}
p_0 = c_0, \\
p_1 = c_1 + q_1 c_0, \\
p_2 = c_2 + q_1 c_1 + q_2 c_0, \\
- - - - - - - - - - - -, \\
p_L = c_L + \sum_{i=1}^{\min(L,M)} q_i c_{L-i} \ .
\end{gathered}$$

Последние две системы уравнений называются уравнениями Паде.

Замечание 1. Если указанная система разрешима, то тейлоровское разложение f совпадает с $[L/M]_f$ с точностью до $O(z^{L+M+1})$.

Замечание 2. Коэффициенты c_i могут быть такими, что степенной ряд (4) везде расходится (радиус сходимости равен нулю) и является формальным. Однако при этом, скажем, диагональные ($L = M$) аппроксимации Паде могут сходится при $M \to \infty$ к некоторой функции F . На этом основывается идея о том, что можно с помощью аппроксимаций Паде построить аналог аналитического продолжения. То есть, функция f может быть задана в некоторой области, а ее аппроксимации Паде при этом сходятся в более широкой области.

Замечание 3. Отрезок ряда Тейлора хорошо аппроксимирует функцию лишь в окрестности точки разложения, тогда как аппроксимация Паде зачастую хорошо приближает функцию в значительно более широкой области.

Пример. Пусть

$$f(z) = \sqrt{\frac{1+\frac{1}{2}z}{1+2z}}\ .$$

Отрезок ряда Тейлора функции f из трех членов представляет собой параболу и на вещественной оси при $z = x \to \infty$ стремится к бесконечности, тогда как сама функция $f(z)$ остается при этом ограниченной. [1/1]-аппроксимация Паде имеет погрешность нигде не превышающую 8 процентов (в том числе и на бесконечности).

2.2.2 Детерминантное Представление полиномов Паде

Заметим, что система (8) для определения величин q_i позволяет предъявить некоторый многочлен $\tilde{Q}^{[M/L]_f}(z)$, коэффициенты которого этой системе удовлетворяют:

$$\tilde{Q}^{[M/L]_f}(z) = \begin{vmatrix} c_{L+1} & c_L & \cdots & c_{L-M+2} & c_{L-M+1} \\ c_{L+2} & c_{L+1} & \cdots & c_{L-M+3} & c_{L-M+2} \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ c_{L+M} & c_{L+M-1} & \cdots & c_{L+1} & c_L \\ 1 & z & \cdots & z^{M-1} & z^M \end{vmatrix}.$$

Определим теперь соответствующий многочлен $\tilde{P}^{[L/M]}$ из соотношения

$$\tilde{Q}^{[L/M]_f}(z)\sum_{i=0}^{\infty} c_i z^i - \tilde{P}^{[L/M]_f}(z) = O(z^{L+M+1})\ . \tag{9}$$

Имеем

$$\tilde{Q}^{[L/M]_f}(z)\sum_{i=0}^{\infty} c_i z^i = \begin{vmatrix} c_{L+1} & c_L & \cdots & c_{L-M+2} & c_{L-M+1} \\ c_{L+2} & c_{L+1} & \cdots & c_{L-M+3} & c_{L-M+2} \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ c_{L+M} & c_{L+M-1} & \cdots & c_{L+1} & c_L \\ \sum\limits_{i=0}^{\infty} c_i z^i & \sum\limits_{i=0}^{\infty} c_i z^{i+1} & \cdots & \sum\limits_{i=0}^{\infty} c_i z^{M+i-1} & \sum\limits_{i=0}^{\infty} c_i z^{M+i} \end{vmatrix}.$$

Домножим первую строку на z^{L+1} и вычтем ее из последней строки. Вторую строку домножим на z^{L+2} и также вычтем из последней строки и т.д. M-ую строку домножим на z^{L+M} и вычтем из последней строки. В результате в каждой сумме в последней строке будут отсутствовать члены со степенями z равными $L+1, L+2, \cdots, L+M$. Если теперь выделить из последнего определителя все члены до степени z^L включительно, то он представляется в виде

$$\begin{vmatrix} c_{L+1} & c_L & \cdots & c_{L-M+2} & c_{L-M+1} \\ c_{L+2} & c_{L+1} & \cdots & c_{L-M+3} & c_{L-M+2} \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ c_{L+M} & c_{L+M-1} & \cdots & c_{L+1} & c_L \\ \sum\limits_{i=0}^{L} c_i z^i & \sum\limits_{i=0}^{L-1} c_i z^{i+1} & \cdots & \sum\limits_{i=0}^{L-M+1} c_i z^{M+i-1} & \sum\limits_{i=0}^{L-M} c_i z^{M+i} \end{vmatrix} +$$

$$+O(z^{M+L+1}) = \tilde{P}^{[L/M]_f}(z) + O(z^{M+L+1})\ .$$

Итак доказано представление (9). Иначе говоря доказана

Теорема. *Для любого ряда* $\sum\limits_{i=0}^{\infty} c_i z^i$ *существуют такие полиномы* P *и* Q *степеней не выше* L *и* M *соответственно, что*

$$Q(z)\sum_{i=0}^{\infty} c_i z^i - P(z) = O(z^{L+M+1}) \ . \tag{10}$$

Заметим, что при доказательстве возможности представления (7) мы нигде не пользовались тем вырождена или не вырождена матрица составленная из коэффициентов c_i.

Определение. Определитель

$$\tilde{Q}^{[L/M]}(0) = \begin{vmatrix} c_L & c_{L-1} & \cdots & c_{L-M+2} & c_{L-M+1} \\ c_{L+1} & c_L & \cdots & c_{L-M+3} & c_{L-M+2} \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ c_{L+M-1} & c_{L+M-2} & \cdots & c_{L+1} & c_L \end{vmatrix}$$

называется определителем Ханкеля.

Отметим, что из наших рассуждений следует справедливость следующей теоремы.

Теорема. *Если* $\tilde{Q}^{[L/M]}(0) \neq 0$, *то существуют единственные (с точностью до множителя) многочлены* $P(z)$ *и* $Q(z)$ *степеней не выше* L *и* M *соответственно, такие что*

$$\sum_{i=0}^{\infty} c_i z^i - \frac{P(z)}{Q(z)} = O(z^{M+L+1}) \ .$$

Пример (недостатки наивного подхода). Построим $[1/1]$-аппроксимацию для $f(z) = 1 + z^2$. Требуется добиться равенства

$$\frac{p_0 + p_1 z}{q_0 + q_1 z} = 1 + z^2 + O(z^3) \ ,$$

откуда

$$p_0 + p_1 z = q_0 + q_1 z + q_0 z^2 + O(z^3) \ ,$$

и

$$q_0 = p_0, \quad p_1 = q_1 \ ,$$

то есть $[1/1] = 1$ и поставленная задача решений не имеет. Обратимся теперь к детерминантным формулам.

$$\tilde{Q}^{[1/1]} = \begin{vmatrix} c_2 & c_1 \\ 1 & z \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & z \end{vmatrix} = z \ ,$$

$$\tilde{P}^{[1/1]} = \begin{vmatrix} c_2 & c_1 \\ c_0 + c_1 z & c_0 z \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & z \end{vmatrix} = z \ .$$

Убедимся, что равенство (7), тем не менее, имеет место:

$$z(1 + z^2) - z = z^3 = O(z^3) \ .$$

Дадим теперь строгое определение аппроксимаций Паде.

Определение. Пусть P и Q полиномы степеней не выше L и M соответственно, $Q(0) \neq 0$ и

$$\frac{P}{Q} - f(z) = O(z^{L+M+1}) \ ,$$

тогда отношение P/Q называется $[L/M]$-аппроксимацией Паде.

Отметим некоторые легко проверяемые свойства аппроксимаций Паде.

Теорема. *Пусть $g = f^{-1}$ и $f(0) \neq 0$, тогда $[M/L]_g = [L/M]_f$, при условии, что хотя бы одна из этих аппроксимаций существует.*

Доказательство. Пусть, скажем, существует аппроксимация $[L/M]_f$, тогда $[L/M]_f(z) = \frac{P_L(z)}{Q_M(z)}$ и $P_L(0) \neq 0$, поскольку $[L/M]_f(0) = f(0) \neq 0$ и, следовательно

$$g(z) - \frac{Q_M(z)}{P_L(z)} = \frac{P_L(z) - f(z)Q_M(z)}{f(z)P_L(z)} = O(z^{L+M+1}) \ ,$$

что и требовалось доказать.

Теорема (инвариантность диагональных аппроксимаций при дробно-линейных преобразованиях сохраняющих начало координат). *Пусть $w = \frac{az}{1+bz}$. Положим $g(w) = f(z)$, тогда $[M/M]_g(w) = [M/M]_f(z)$ при условии, что хотя бы одна из этих аппроксимаций существует.*

Доказательство. Пусть существует аппроксимация

$$[M/M]_g(w) = \frac{\sum\limits^{M} a_k w^k}{\sum\limits^{M} b_k w^k} = g(w) + O(z^{2M+1}) \ .$$

Введем полиномы A_M и B_M по z степени не выше M:

$$A(z) = (1+bz)^M \sum^{M} a_k \left(\frac{az}{1+bz}\right)^k \ , \quad B(z) = (1+bz)^M \sum^{M} b_k \left(\frac{az}{1+bz}\right)^k \ ,$$

тогда

$$\frac{A_M(z)}{B_M(z)} = f(z) + O(z^{2M+1}) \ ,$$

поскольку 0 переходит в 0.

Теорема(инвариантность диагональных аппроксимаций относительно дробно-линейных функций). *Пусть $g(z) = \frac{a+bf(z)}{c+df(z)}$ и $c + df(0) \neq 0$, тогда*

$$[M/M]_g(z) = \frac{a + b[M/M]_f}{c + d[M/M]_f} \ ,$$

если $[M/M]_f$ существует.

Доказательство.

$$\frac{a + b[M/M]_f(z)}{c + d[M/M]_f(z)} = \frac{P_M(z)}{Q_M(z)} \ ,$$

где P_M и Q_M полиномы степени не выше M, причем $Q_M(0) \neq 0$. Следовательно,

$$\frac{P_M(z)}{Q_M(z)} - g(z) = \frac{(bc-ad)\left\{[M/M]_f(z) - f(z)\right\}}{\{c + d[M/M]_f(z)\}(c + df(z))} = O(z^{2M+1}) \ .$$

2.2.3 Аппроксимации Паде в бесконечно удаленной точке

Пусть

$$f(z) = \sum_{k=0}^{\infty} \frac{f_k}{z^{k+1}}$$

— формальный ряд по обратным степеням z. Поставим следующую задачу. Пусть N — натуральное. Требуется найти многочлен $Q_N \neq 0$, $\deg Q_N \leq N$, такой что

$$Q_N(z)f(z) - P_N(z) = \frac{c}{z^{N+1}} + \dots \quad , \tag{11}$$

где $P_N(z)$ — полиномиальная часть ряда $Q_N(z)f(z)$.

Решение этой задачи существует и $\deg P_N \leq N$. Если пара (P_N, Q_N) не единственна (не только с точностью до множителя), то тем не менее отношение P_N/Q_N определяет одну и ту же рациональную функцию для любой пары Паде. Действительно, пусть

$$Q'_N(z)f(z) - P'_N(z) = \frac{c'}{z^{N+1}} + \dots \ , \quad Q''_N(z)f(z) - P''_N(z) = \frac{c''}{z^{N+1}} + \dots \ ,$$

тогда домножив первое равенство на $Q''_N(z)$, а второе на $Q'_N(z)$ и вычтя второе из первого, получим

$$Q'_N(z)P''_N(z) - Q''_N(z)P'_N(z) = \frac{c}{z} + \dots \ ,$$

и, поскольку, в левой части равенства стоит многочлен, а в правой разложение идет лишь по отрицательным степеням z, получаем что $Q'_N(z)P''_N(z) - Q''_N(z)P'_N(z) = 0$.

Отношение $\pi_N(z) = Q_N/P_N$ называется N-ой диагональной аппроксимацией Паде ряда f. Ясно, что $\pi_N(f(1/z), z) = \pi_N(f(z), 1/z)$.

Если для любой N-ой пары Паде $\deg Q_N = N$, то индекс N называют *нормальным* (для ряда f). Множество нормальных индексов обозначим $\Lambda(f)$. Установим детерминантный критерий нормальности. Пусть $H_0 = 1$ и

$$H_N = \begin{vmatrix} f_0 & f_1 & \dots & f_{N-1} \\ f_1 & f_2 & \dots & f_N \\ \vdots & \vdots & \ddots & \vdots \\ f_{N-1} & f_N & \dots & f_{2N-2} \end{vmatrix}$$

— определители Ханкеля, построенные по ряду $f(z)$.

Утверждение. $N \in \Lambda \ \Leftrightarrow \ H_N \neq 0$.

Доказательство. Индекс $N = 0$ всегда нормален ($H_N = 1$). При $N > 0$ запишем в явном виде систему линейных уравнений для определения коэффициентов q_k многочлена Q_N . Пусть $Q_N(z) = \sum\limits_{k=0}^{N} q_k z^k$, тогда условия равенства нулю коэффициентов при степенях $(1/z)^n$, $n = 1, 2, \dots, N$ принимают вид

$$\begin{gathered} f_0 q_0 + f_1 q_1 + \dots + f_N q_N = 0 \ , \\ f_1 q_0 + f_2 q_1 + \dots + f_{N+1} q_N = 0 \ , \\ \dots\dots\dots \ , \\ f_{N-1} q_0 + f_N q_1 + \dots + f_{2N-1} q_N = 0 \ . \end{gathered} \tag{12}$$

Если $N \notin \Lambda$, то существует ненулевое решение выписанной системы с $q_N = 0$ и, следовательно, $H_N = 0$. Пусть теперь $N \in \Lambda$. Тогда по определению нормального индекса система (12) с $q_N = 0$ имеет лишь тривиальное решение, поэтому $H_N \neq 0$.

Отметим некоторые легко проверяемые свойства нормальных индексов. Если $N \in \Lambda$, то N-ая пара Паде единственна (с точностью до умножения на отличное от нуля число), многочлены P_N и Q_N при этом взаимно просты и $\deg \pi_N = N$.

Следующее утверждение полностью описывает структуру последовательности диагональных аппроксимаций Паде — эта последовательность оказывается состоит только из аппроксимаций, отвечающих нормальным индексам.

Утверждение. Пусть $N \in \Lambda$, J — целое, $J > N$ и $(N, J] \cap \Lambda = \emptyset$, тогда $\pi_J = \pi_N$.

Доказательство. Запишем π_J в виде несократимой дроби: $\pi_J = P/Q$. Пусть $\deg \pi_J = r$. Поскольку $J \notin \Lambda$, то $r < J$. Покажем, что индекс r нормален. Пара (P, Q) — r-ая пара Паде и $\pi_r = \pi_J$ (по построению), $\deg \pi_r = r$, $r \in \Lambda$. Ясно, что $r \leq N$ (поскольку индекс r нормален), причем (P, Q) — также и N-ая пара, и, следовательно, $\pi_N = \pi_J$.

Заметим, что знаменатель Паде $Q_N(z)$ можно записать в виде определителя

$$Q_N(z) = \begin{vmatrix} f_0 & f_1 & \dots & f_N \\ f_1 & f_2 & \dots & f_{N+1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N-1} & f_N & \dots & f_{2N-1} \\ 1 & z & \dots & z^N \end{vmatrix} .$$

При этом если существует такая область $F \subseteq R$ и конечная мера μ (напомним, что мера есть счетно аддитивная неотрицательная функция множества, конечность меры означает что ее значение на всем множестве F, где она определена, конечно: $\mu(F) < \infty$) , что величины f_k представляют собой ее моменты, то есть $f_k = \int\limits_F x^k d\mu(x)$, $k = 0, 1 \dots$, то многочлены Q_N являются ортогональными с мерой μ:

$$\int\limits_F Q_N(x) Q_M(x) d\mu(x) = 0 \ , \quad N \neq M \ .$$

Подробнее об ортогональных полиномах см. в главе "Численное интегрирование". Сама задача о нахождении меры μ по заданной последовательности чисел f_k называется проблемой моментов. В зависимости от области интегрирования F выделяют 3 классических случая:

1) $F = R$ — проблема моментов Гамбургера;

2) $F = [0, \infty)$ — проблема моментов Стильтьеса;

3) $F = [0, 1]$ — проблема моментов Хаусдорфа.

Отметим в заключение, что если числа $\{f_k\}_0^\infty$ являются моментами некоторой меры то все определители Ханкеля H_n больше нуля. Если же последовательность $\{f_k\}_0^\infty$ такова, что все $H_n > 0$, то проблема моментов Гамбургера разрешима.

Глава 3

Численное дифференцирование

Естественным способом приближенного дифференцирования является дифференцирование не самой функции, а интерполяционного полинома или сплайна построенного по ее табличным значениям. Можно также дифференцировать аппроксимации Паде и вообще произвольные аппроксимации функции, производные от которой мы хотим определить. Вопрос лишь в том, каковы затраты и какова точность такого дифференцирования.

3.1 Дифференцирование интерполяционного полинома

Самым простым способом приближенного дифференцирования функции является дифференцирование интерполяционного полинома, построенного по некоторой сетке ее значений, который удобно представить в форме Ньютона

$$p(x) = \sum_{k=0}^{N} f_{012\,\ldots\,k}\mathcal{N}_k(x) = \sum_{k=0}^{N} f_{012\,\ldots\,k}\prod_{i=0}^{k-1}(x-x_i)\ .$$

Его n-ая производная имеет вид

$$p^{(n)}(x) = n!\Big\{f_{01\ldots n} + \big[\sum_{i=0}^{n}(x-x_i)\big]f_{01\ldots n+1} + \\ + \big[\sum_{j>i\geq 0}^{j=n+1}(x-x_i)(x-x_j)\big]f_{01\ldots n+2} + \ldots\Big\}\ . \tag{1}$$

Поскольку погрешность интерполяционного полинома есть

$$f(x) - p(x) = \frac{f^{N+1}(\xi(x))}{(N+1)!}\prod_{i=0}^{N}(x-x_i)\ ,$$

то погрешность дифференцирования оценивается выражением

$$f^{(n)}(x) - p^{(n)}(x) \lesssim const\frac{||f^{(N+1)}||_C}{(N+1-n)!}\max_i[x-x_i]^{N+1-n}\quad ,$$

то есть, каждое дифференцирование на один порядок снижает точность. Если же в (1) оставить только первый член, то поскольку порядок погрешности определяется первым отброшенным членом $\sum_{i=0}^{n}(x-x_i)f_{01\ldots n+1}$, получаем следующее приближенное выражение для производных

$$f^{(n)}(x) \simeq n!\Big[f_{01\ldots n} + O\Big(\sum_{i=0}^{n}(x-x_i)\Big)\Big]\quad . \tag{2}$$

Пусть $h = \max\limits_i h_i$, где $h_i = x_{i+1} - x_i$. Из (2) видно, что разделенная разность n-го порядка, домноженная на $n!$, аппроксимирует производную n-го порядка с точностью $O(h)$. Действительно, если $x \in [x_0, x_n]$, то максимум модуля суммы $\sum\limits_{i=0}^{n}(x - x_i)$ достигается в одной из точек $x = x_0$ или $x = x_n$ и не превосходит величины $h + 2h + \ldots nh = \frac{n(n+1)}{2}h$.

Поскольку разделенная разность $f_{01\ldots n}$ не содержит самой переменной x , то возникает вопрос: производную в какой точке она аппроксимирует точнее всего. Эта точка определяется условием равенства нулю первого отброшенного члена, то есть условием $\sum\limits_{i=0}^{n}(x - x_i) = 0$, откуда $x_* = \sum\limits_{i=0}^{n} x_i/(n+1)$, и представляет собой положение центра масс точек $x_0, x_1, \ldots, x_n$. В этой точке порядок точности приближенной производной n-го порядка на единицу выше и равен $O(h^2)$. Если же в (1) оставить два первых члена, то порядок точности такой формулы приближенного дифференцирования будет $O(h^2)$. При этом в двух точках, определяемых как корни квадратного уравнения $\sum\limits_{j>i\geq 0}^{j=n+1}(x-x_i)(x-x_j) = 0$, точность будет на порядок выше. Аналогично, k-членная формула имеет порядок точности $O(h^k)$, точки повышенной точности есть корни уравнения k-го порядка. Решать такое уравнение вряд ли целесообразно, однако, как нетрудно заметить, если точка x такова, что узлы $x_0, x_1, \ldots x_{n+k-1}$ расположены относительно неё симметрично и k нечетно, то эта точка является точкой повышенной точности. Разумеется для произвольной сетки, такое условие, как правило, не реализуется. Однако такая точка заведомо существует (при произвольном k), если шаг сетки постоянный и находится она посередине между крайними узлами: $x_* = (x_0 + x_{n+k-1})/2$.

3.2 Конечные разности

Естественным способом описания приближенного дифференцирования наряду с разделенными разностями является использование конечных разностей.

Пусть $f(x) \in C$, обозначим через $\Delta x = h$ приращение аргумента.

Определение. Выражение $\Delta_h f = \Delta f(x) = f(x+h) - f(x)$ называется *первой разностью* (или *конечной разностью первого порядка*) шага h функции $f(x)$.

Конечные разности высших порядков определим естественными рекуррентными соотношениями. Именно, положим $\Delta^n_{h_1 h_2 \ldots h_n} f = \Delta_{h_n}(\Delta^{n-1}_{h_1 h_2 \ldots h_{n-1}} f)$ — конечная разность n-го порядка. Это определение не зависит от последовательности применения сдвигов h_i :

$\Delta^n_{h_1 h_2 \ldots h_n} f = \Delta^n_{h_{i_1} h_{i_2} \ldots h_{i_n}} f$, где $\{i_1, i_2 \ldots i_n\}$ — произвольная перестановка индексов $1, 2, \ldots n$.

Отметим связь конечных разностей с многочленами. Пусть $p(x) = a_N x^N + a_{N-1}x^{N-1} + \ldots + a_0$ — полином степени N, тогда

а)$\Delta^N_{h_1 h_2 \ldots h_N} p(x) = N! a_N h_1 h_2 \ldots h_N = const$;

б)$\Delta^{N+1}_{h_1 h_2 \ldots h_{N+1}} p(x) = 0$.

Доказательство. Для достаточно убедиться в том, что

$$\Delta^k_{h_1 h_2 \ldots h_k} x^N = N(N-1)\ldots(N-k+1) h_1 h_2 \ldots h_k x^{N-k} \quad , \quad k \leq N \quad .$$

Действительно,

$$\Delta_{h_1} x^N = (x + h_1)^N - x^N = \sum_{k=0}^{N} \overset{k}{N} h^k x^{N-k} - x^N = N h_1 x^{N-1} + \ldots \quad .$$

Применим теперь Δ_{h_2} к $\Delta_{h_1}x^N$:

$$\Delta^2_{h_1h_2}x^N = \Delta_{h_2}(\Delta_{h_1}x^N) = N(N-1)h_1h_2x^{N-2} + \dots \ ,$$

и так далее.

Заметим, что эти свойства аналогичны соответствующим свойствам разделенных разностей: $p_{01\dots N} = const$, $p_{01\dots N+1} = 0$. Указанное сходство разделенных и конечных разностей не ограничивается этим. Пусть шаг h постоянный, обозначим $\Delta^k = \Delta^k_{\underbrace{hh\dots h}_{k}}$, тогда

$$k!f_{01\dots k} = \frac{\Delta^k f_0}{h^k} \ ,$$

где $\Delta^k f_0 = \Delta^k f(x)|_{x=x_0}$. Действительно $f_{01} = \frac{(f_1-f_0)}{(x_1-x_0)} = \frac{\Delta f_0}{h}$. Далее поступим по индукции. Пусть при индексе равном $k-1$ равенство имеет место, тогда

$$f_{01\dots k} = \frac{f_{12\dots k} - f_{01\dots k-1}}{x_k - x_0} = \frac{\frac{1}{(k-1)!h^{k-1}}(\Delta^{k-1}f_1 - \Delta^{k-1}f_0)}{kh} = \frac{\Delta^k f_0}{k!h^k} \ .$$

Заметим, что напрашивающееся обобщение для неравномерной сетки (непостоянного шага), а именно равенство величины $k!f_{01\dots k}$ отношению $\frac{\Delta^k_{h_1h_2\dots k_k}f_0}{h_1h_2\dots h_k}$, очевидно, не имеет места. Предоставляем читателю убедиться в этом самостоятельно (без всяких вычислений!).

Итак введен оператор Δ действующий на функцию $f(x)$ по правилу $\Delta f(x) \equiv f(x+h) - f(x)$. Отметим дальнейшие свойства конечных разностей:

1) Линейность: $\Delta(\alpha f + \beta g) = \alpha\Delta f + \beta\Delta g$;

2) $\Delta^k(\Delta^l f) = \Delta^{k+l}f = \Delta^l(\Delta^k f)$;

3) Связь с производной: $\frac{d}{dx} = \frac{1}{\Delta x}ln(1+\Delta)$.

Последнее равенство формальное и понимать его нужно в следующем смысле

$$\Delta f = \exp\{h\frac{d}{dx}\}f - f \ ,$$

где подразумевается, что f — аналитическая, т.е., в частности, раскладывается в ряд Тейлора и совпадает с ним в некотором круге на комплексной плоскости

$$f(x+h) = \sum_{n=0}^{\infty}\frac{1}{n!}\left(h\frac{d}{dx}\right)^n f(x) = \exp\{h\frac{d}{dx}\}f(x) \ .$$

Таким образом оператор дифференцирования можно с любой степенью точности аппроксимировать конечными разностями:

$$\frac{d}{dx} = \frac{ln(1+\Delta)}{h} = \frac{1}{h}\left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} + \dots + \frac{(-1)^{n+1}\Delta^n}{n} + \dots\right) \ . \tag{3}$$

Обрезая это выражение на той или иной степени Δ , можно получить выражение для производной в точке x с любой степенью точности. Из этой формулы, в частности, приближенно получается $\frac{df}{dx} \simeq \frac{\Delta f}{h} = \frac{f(x+h)-f(x)}{h}$, а оставляя два члена разложения, получаем

$$\frac{df}{dx} \simeq \frac{1}{h}\left(\Delta - \frac{\Delta^2}{2}\right)f = \frac{1}{h}\left(2f(x+h) - \frac{f(x+2h)}{2} - \frac{3}{2}f(x)\right) \ .$$

Выражение для производных высших порядков получаем из (3), скажем вторая производная имеет следующее представление

$$\frac{d}{dx}f = \frac{1}{h^2}ln(1+\Delta)ln(1+\Delta)f \ .$$

4) Выражение последовательных значений функции через конечные разности: $f(x+kh) = \sum\limits_{s=0}^{k} C_k^s \Delta^s f(x)$.

Доказательство: Действительно

$$f(x+h) = f(x) + \Delta f(x) = (1+\Delta) f(x) \ ,$$

$$f(x+2h) = (1+\Delta) f(x+h) = (1+\Delta)^2 f(x) \ ,$$

$$\dots \ ,$$

$$f(x+kh) = (1+\Delta)^k f(x) \ ,$$

и, раскладывая по биному $(1+\Delta)^k = \sum\limits_{s=0}^{k} C_k^s \Delta^s$, где $C_k^s = \frac{k(k-1)\dots(k-s+1)}{s!} = \frac{k!}{(k-s)!s!}$, получаем искомое выражение.

5) Выражение конечных разностей через значения функции: $\Delta^k f(x) = \sum\limits_{s=0}^{k} C_k^s (-1)^s f(x + (k-s)h)$.

Доказательство. Представим $\Delta = (1+\Delta) - 1$, тогда

$$\Delta^k f(x) = [(1+\Delta) - 1]^k f(x) = \sum_{s=0}^{k} C_k^s (1+\Delta)^{k-s} (-1)^s f(x) =$$

$$= \sum_{s=0}^{k} C_k^s (-1)^s f(x + (k-s)h),$$

или расписывая подробно:

$$\Delta^k f(x) = f(x+kh) - C_k^1 f(x + (k-1)h) + C_k^2 f(x + (k-2)h) +$$

$$+ \dots + (-1)^k f(x).$$

6) Формула конечных приращений Лагранжа:

$$\Delta^k f(x) = (\Delta x)^k f^{(k)}(x + \Theta k \Delta x) \ ,$$

где $0 < \Theta < 1$ и $f \in C^k$.

Доказательство. Доказательство мы будем проводить по индукции. База индукции $\Delta f = \Delta x f'(x + \Theta \Delta x)$ имеет место в силу теоремы Лагранжа о среднем значении производной (напомним, что для дифференцируемой на отрезке $[x, x+\Delta x]$ функции теорема Лагранжа утверждает, что на этом же промежутке найдется точка ξ, такая что $\frac{\Delta f}{\Delta x} = \frac{f(x+h) - f(x)}{\Delta x} = f'(\xi)$, где $\xi \in [x, x + \Delta x]$). Далее пусть при индексе равном k формула справедлива:

$$\Delta^k f(x) = (\Delta x)^k f^{(k)}(x + \Theta k \Delta x) \ \ .$$

Тогда

$$\Delta^{k+1} f(x) = \Delta(\Delta^k f) = \Delta[f^{(k)}(x + k\Theta\Delta x)] \Delta^k x =$$

$$= \Delta^k x [f^{(k)}(x + \Delta x + k\Theta\Delta x) - f^{(k)}(x + k\Theta\Delta x)] \ .$$

Продолжим это равенство используя теорему Лагранжа

$$= (\Delta x)^{k+1} f^{(k+1)}(x + k\Theta\Delta x + \Theta'\Delta x) = (\Delta x)^{k+1} f^{(k+1)}(x + (k\Theta + \Theta')\Delta x) \ .$$

Здесь $\Theta' < 1$ (равно как и Θ). Введем $\Theta'' = \frac{k\Theta+\Theta'}{k+1}$, тогда последняя формула переписывается в виде

$$(\Delta x)^{k+1} f^{(k+1)}(x + (k+1)\Theta''\Delta x) \ .$$

При этом, как нетрудно убедиться $\Theta'' < 1$, таким образом формула конечных приращений доказана.

Следствие свойства 6). $f^{(k)}(x) = \frac{\Delta^k f}{(\Delta x)^k} + o(1)$.

Действительно, $\frac{\Delta^k f}{(\Delta x)^k} = f^{(k)}(x + \Theta k \Delta x)$, откуда устремляя $\Delta x \to 0$, получаем $f^{(k)}(x) = \lim\limits_{\Delta x \to 0} \frac{\Delta^k f}{(\Delta x)^k}$.

3.2.1 Оператор Δ и обобщенная степень

Определение. Обобщенной степенью числа x называется выражение

$$x^{[n]} \equiv x(x-h)(x-2h)\dots(x-(n-1)h) \ , \quad x^{[0]} \equiv 1.$$

Заметим, что если $h = 0$, то $x^{[n]} = x^n$.

Свойство. $\Delta^k x^{[n]} = n(n-1)\dots(n-(k-1))h^k x^{[n-k]}$.

Доказательство.

$$\Delta x^{[n]} = (x+h)^{[n]} - x^{[n]} =$$

$$= (x+h)x(x-h)\dots(x-(n-2)h) - x(x-h)\dots(x-(n-1)h) =$$

$$= x(x-h)\dots(x-(n-2)h)[x+h-(x-(n-1)h)] = nhx^{[n-1]} \ ,$$

применяя Δ еще раз, получаем

$$\Delta^2 x^{[n]} = \Delta(\Delta x^{[n]}) = \Delta(nhx^{[n-1]}) = nh(n-1)hx^{[n-2]} =$$

$$= n(n-1)h^2 x^{[n-2]} \ ,$$

и так далее.

Таким образом действие оператора Δ на обобщенную степень аналогично дифференцированию обычных степеней:

$$d^k x^n = n(n-1)\dots(n-(k-1))x^{n-k}(dx)^k \ .$$

3.2.2 Интерполяционный многочлен Ньютона для равноотстоящих узлов

Пусть в точках $x_0 \, , x_1 \, , \dots \, , x_N \ : \ x_i = x_0 + ih$, заданы значения $f_0 \, , f_1 \, , \dots \, , f_N$. Решим задачу интерполяции, то есть построим полином

$$p(x) \ : \ p(x_i) = f_i \, , \quad i = 0 \, , 1 \, , \dots \, , N \ , \quad \deg p(x) = N \ . \tag{4}$$

Интерполяционный полином, удовлетворяющий табличным значениям $\{x_i, f_i\}_{i=0}^N$, в форме Ньютона имеет вид

$$p(x) = \sum_{k=0}^{N} f_{012\dots k} \mathcal{N}_k(x) \ .$$

Для постоянного шага h выполнено: $k! f_{01\dots k} = \frac{\Delta^k f_0}{h^k}$, при этом $\mathcal{N}_k(x) = \prod\limits_{i=0}^{k-1}(x - x_i) = (x-x_0)^{[k]}$, таким образом решение задачи интерполяции принимает вид

$$p(x) = f_0 + \frac{\Delta f_0}{h}(x-x_0)^{[1]} + \frac{1}{2!}\frac{\Delta^2 f_0}{h^2}(x-x_0)^{[2]} + \dots + \frac{1}{N!}\frac{\Delta^N f_0}{h^N}(x-x_0)^{[N]} \ .$$

Заметим, что сами условия (4) можно также переписать в виде: $\Delta^k p(x)|_{x=x_0} = \Delta^k f_0$. Действительно, из свойства 5) конечных разностей

$$\Delta^k p(x_0) = p(x_0 + kh) - C_k^1 p(x_0 + (k-1)h) + \ldots + (-1)^k p(x_0) =$$

$$= f_k - C_k^1 f_{k-1} + \ \ldots \ + f_0 = \Delta^k f_0 \ .$$

Проверим, что построенный полином $p(x)$ действительно удовлетворяет условиям интерполяции:

1) $p(x_0) = f_0$, что следует из формы записи полинома;

2) $p(x_k) = p_0 + \frac{\Delta p_0}{h}(x_k - x_0)^{[1]} + \ldots + \frac{\Delta^k p_0}{k! h^k}(x_k - x_0)^{[k]} + 0$.

Поскольку $x_k - x_0 = kh$, то

$$(x_k - x_0)^{[m]} = kh(kh-h)\ldots(kh-(m-1)h) = h^m k(k-1)\ldots(k-(m-1)) \ ,$$

и, следовательно,

$$p(x_k) = f_0 + \frac{\Delta f_0}{h}kh + \frac{\Delta^2 f_0}{2!h^2}h^2 k(k-1) + \ldots + \frac{\Delta^k f_0}{k!h^k}h^k k(k-1)\ldots 1 =$$

$$= f_0 + \Delta f_0 k + \frac{\Delta^2 f_0}{2!}k(k-1) + \ldots + \frac{\Delta^k f_0}{k!h^k}k(k-1)\ldots 1 =$$

$$= \sum_{m=0}^{k} C_k^m \Delta^m f_0 = (1+\Delta)^k f_0 = f_k$$

по свойству конечных разностей.

Замечание. Если $h \to 0$, то полином $p(x)$ стремится к отрезку ряда Тейлора функции f , так как в этом случае $\frac{\Delta^m f_0}{(\Delta x)^m} \to f^{(m)}(x_0)$, $(x - x_0)^{[m]} \to (x - x_0)^m$ и

$$p(x) \to f_0 + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^{(N)}(x_0)}{N!}(x - x_0)^N =$$

$$= \sum_{k=0}^{N} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \ .$$

Можно интерполяционный полином записать также в следующей форме:

$$p(x) = f_0 + q\Delta f_0 + \frac{q(q-1)}{2!}\Delta^2 f_0 + \ldots + \frac{q(q-1)\ldots(q-N+1)}{N!}\Delta^N f_0,$$

где $q = \frac{x - x_0}{h}$. Действительно

$$\frac{(x - x_0)^{[m]}}{h^m} = \frac{(x - x_0)(x - x_0 - h)\ldots(x - x_0 - (m-1)h)}{h \ h \ \ldots \ h} =$$

$$= q(q-1)\ldots(q-m+1) \ .$$

Глава 4

Численное интегрирование

4.1 Наводящие соображения

При приближенном вычислении интегралов вида

$$I = \int_a^b f(x)\rho(x)dx \ ,$$

где f — интегрируемая функция, ρ — *вес* или *весовая функция* со свойствами

1) $\rho \in C_{(a,b)}$;

2) ρ интегрируемая на $[a,b]$;

3) $\rho > 0$,

естественно использовать следующий прием.

Проинтерполируем интегрируемую функцию f с помощью чебышевской системы функций $\{\varphi_i\}_{i=0}^N$ по её значениям $f_i = f(x_i)$ в некоторых узлах $\{x_i\}_{i=0}^N$ промежутка $[a,b]$. Тогда функцию f можно представить в виде

$$f(x) = \sum_{i=0}^{N} \alpha_i \varphi_i(x) + r_N(x) \ , \tag{1}$$

где $r_N(x)$ соответствующая невязка, а коэффициенты α_i линейно выражаются через значения f_j (см. раздел "Интерполяция"): $\alpha_i = \sum\limits_{j=0}^{N} [\Phi^T]_{ij}^{-1} f_j$, где Φ невырожденная матрица с элементами $\varphi_i(x_j)$. Для удобства будем считать, что базис в линейной оболочке $\bigvee\limits_{i=0}^{N} \varphi_i$ выбран таким, что матрица Φ единичная, то есть $\alpha_i = f_i$, тогда

$$I = \sum_{i=0}^{N} f(x_i)\lambda_i + R_N(f,\rho) \ , \tag{2}$$

где введены обозначения

$$\lambda_i = \int_a^b \varphi_i(x)\rho(x)dx \ , \quad R_N(f,\rho) = \int_a^b r_i(x)\rho(x)dx \ , \tag{3}$$

Если в (2) отбросить погрешность $R_N(f,\rho)$, то оставшееся выражение

$$I = \int_a^b f(x)\rho(x)dx \approx \sum_{i=0}^{N} \lambda_i f(x_i) \tag{4}$$

называется *квадратурной* формулой. Естественно, что для того чтобы можно было использовать квадратурную формулу необходимо выбрать чебышевскую систему таким образом, чтобы *веса* λ_i квадратурной формулы могли быть сосчитаны явно. Обычно в качестве такой системы используются полиномы.

Замечание. В принципе любая запись вида (4) (с произвольными весами λ_i) является квадратурной формулой, однако ценность такой формулы может и вовсе отсутствовать, если числа (веса) λ_i выбраны неразумно. Кроме того, дополнительной точности можно добиться за счет эффективного расположения *узлов* x_i квадратурной формулы.

Возникает естественный вопрос: А что является мерой точности квадратурной формулы, ведь при интегрировании различных функций f погрешность $R_N(f,\rho)$ может быть существенно разной? В связи с этим естественно выделить некоторый класс функций, на котором и проверяется величина погрешности. Если в качестве такого класса используются полиномы, то говорят об алгебраической точности квадратурной формулы.

Определение. *Алгебраической степенью точности* квадратурной формулы называется максимальное число M такое, что при интегрировании любых полиномов степени не превосходящей M приближенное равенство (4) превращается в тождество (т.е. невязка (погрешность $R_N(f,\rho)$) квадратурной формулы равна нулю если $f = p_k$ полином степени $k \le M$).

Заметим, что если в качестве чебышевской системы использовать полиномы, то при условии, что веса λ_i сосчитаны точно (по формуле (3)), квадратурная формула (4) имеет алгебраическую степень точности M не ниже N , поскольку для полиномов степени до N невязка $r_N(x)$ тождественно равна нулю, так как в этом случае интерполяционный полином просто совпадает с f.

4.2 Квадратурные формулы Ньютона-Котеса

Пусть вес $\rho \equiv 1$, $x_0 = a$, $x_N = b$. Используем в качестве чебышевской системы $\varphi_i(x)$ полиномы Лагранжа:

$$\varphi_i(x) \equiv \mathcal{L}_N^{(i)}(x) \quad , \quad \mathcal{L}_N^{(i)}(x) = \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)} \ ,$$

тогда $f(x) = \sum\limits_{i=0}^{N} \mathcal{L}_N^{(i)}(x) f(x_i) + r_N(x)$ и

$$I = \sum_{i=0}^{N} \lambda_i f(x_i) + \int_a^b r_N(x) dx \ ,$$

где

$$\lambda_i = \int_a^b \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)} dx \ , \tag{5}$$

т.е. веса квадратурной формулы могут быть сосчитаны явно. Сама же полученная формула приближенного интегрирования называется *квадратурной формулой Ньютона-Котеса*.

4.2.1 Случай равноотстоящих узлов

Получим выражения для весов в случае равноотстоящих узлов. В этой ситуации $x_k = x_0 + kh$, $k = 0, 1, \ldots, N$, и

$$\prod_{j \neq i}(x - x_j) = (x - x_0)^{[i]}(x - x_{i+1})^{[N-i]} ,$$

а также

$$\prod_{j \neq i}(x_i - x_j) = \underbrace{ih(i-1)h \ldots h}_{i \ p}\underbrace{(-h) \ldots (-1)(N-i)h}_{(N-i) \ p} =$$

$$= i!(N-i)!(-1)^{N-i}h^N .$$

Таким образом

$$\lambda_i = \frac{(-1)^{N-i}}{i!(N-i)!h^N}\int_a^b \frac{\prod_{j=0}^{N}(x - x_0 - jh)}{(x - x_0 - ih)}dx .$$

Положим $\frac{x-x_0}{h} = q$, $a = x_0$, $b = x_N$ и заметим, что $\frac{h}{b-a} = \frac{1}{N}$, тогда

$$\lambda_i = \frac{(-1)^{N-i}}{i!(N-i)!}h\int_0^N dq\frac{\prod_{j=0}^{N}(q-j)}{q-i} .$$

Окончательно, обычно вводят несколько другие коэффициенты, называемые *коэффициентами Котеса*: $H_i = \frac{1}{b-a}\lambda_i$, при этом квадратурная формула принимает вид

$$\int_a^b f(x)dx = (b-a)\sum_{i=0}^{N} H_i f(x_i) + R(f) .$$

Свойства коэффициентов Котеса H_i:
1) $\sum\limits_{i=0}^{N} H_i = 1$;
2) $H_i = H_{N-i}$.

Доказательство.

1) Поскольку квадратурная формула Ньютона-Котеса точна для полиномов степени не превосходящей N , то в частности если взять в качестве функции f функцию тождественно равную 1, то

$$\int_a^b dx = (b-a)\sum_{i=0}^{N} H_i f(x_i) = (b-a)\sum_{i=0}^{N} H_i = (b-a) ,$$

откуда свойство 1) следует непосредственно.

2) Коэффициент H_i равен

$$H_i = \frac{1}{N}\frac{(-1)^{N-i}}{i!(N-i)!}\int_0^N dq\frac{\prod_{j=0}^{N}(q-j)}{q-i} ,$$

при этом

$$H_{N-i} = \frac{1}{N}\frac{(-1)^{N-N+i}}{(N-i)!(N-N+i)!}\int_0^N dq\frac{\prod_{j=0}^{N}(q-j)}{q-N+i} =$$

$$= \frac{(-1)^i}{N(N-i)!i!}\int_0^N dq\frac{\prod_{j=0}^{N}(q-j)}{q-N+i} .$$

Произведем замену переменной $q - N = -p$, $dq = dp$, тогда

$$\int_0^N dq \frac{\prod_{j=0}^{N}(q-j)}{q-N+i} = \int_0^N dp \frac{\prod_{j=0}^{N}(N-p-j)}{-(p-i)} = (-1)^N \int_0^N dp \frac{\prod_{j=0}^{N}(p-j)}{p-i} \ ,$$

откуда $H_{N-i} = H_i$.

4.2.2 Оценка погрешности квадратурных формул Ньютона-Котеса

Для погрешности интерполирования $r(x)$ функции $f(x)$ интерполяционным полиномом $p(x)$ у нас было получено выражение

$$r(x) \equiv f(x) - p_N(x) = \frac{f^{(N+1)}(\eta)}{(N+1)!}\mathcal{N}_{N+1}(x) \ ,$$

где точка η зависит от x : $\eta = \eta(x)$ и $\mathcal{N}_{N+1}(x) = \prod_{i=0}^{N}(x - x_i)$. Таким образом

$$R_N(f,1) = \int_a^b r_n(x)dx = \int_a^b \frac{f^{(N+1)}(\eta)}{(N+1)!}\mathcal{N}_{N+1}(x)dx \ ,$$

и

$$|R_N(f,1)| \le \frac{||f^{(N+1)}||_{C[a,b]}}{(N+1)!} \int_a^b \mathcal{N}_{N+1}(x)dx \ .$$

В частности, если $f(x)$ — это полином степени $\deg f \le N$ то $R_N(f,1) = 0$, то есть действительно квадратурная формула Ньютона-Котеса с $(N+1)$ узлом точна для полиномов степени не превосходящей N.

4.3 Формулы Гаусса-Кристофеля

4.3.1 Пределы алгебраической степени точности

Выясним какой может быть алгебраическая степень точности M квадратурной формулы с L узлами $x_1, x_2, \ldots, x_L$:

$$\int_a^b f(x)\rho(x)dx \approx \sum_{k=1}^{L} \lambda_k f(x_k) \ . \tag{6}$$

Частичный ответ на этот вопрос дает

Лемма.

а) *для любой квадратурной формулы* $M \le 2L - 1$;

б) *для любой данной системы узлов* $\{x_i\}_{i=1}^{L}$ *существуют такие* λ_k, *что алгебраическая степень точности* $M \ge L - 1$.

Доказательство.

а) Сначала приведем нестрогое рассуждение. Подсчитаем число свободных параметров квадратурной формулы. Оно равно $2L$ (L весов λ_i и L узлов x_i). Полином же степени M содержит $M + 1$ параметр. Приравняем эти величины: $M + 1 = 2L$, то есть M не может превосходить $2L - 1$.

Строгое же доказательство состоит в том, что мы просто предложим полином степени $2L$, для которого (6) не может быть тождеством. Действительно пусть $f(x) = [\prod_{i=1}^{L}(x - x_i)]^2$, тогда $f(x) \ge 0$ и поскольку вес

$\rho(x)$ неотрицателен и не равен тождественно нулю, то $\int\limits_a^b f(x)\rho(x)dx > 0$, с другой стороны $\sum\limits_{k=1}^{L} \lambda_k f(x_k) = 0$, поскольку $f(x_k) = 0$.

б) Введем моменты

$$c_l = \int\limits_a^b x^l \rho(x)dx \ .$$

Если (6) — строгое равенство для полиномов степени до M, то должно быть выполнено:

$$\int\limits_a^b x^l \rho(x)dx = c_l = \sum_{k=1}^{L} \lambda_k x_k^l \ , \ l = 0\,,\, 1\,,\, \ldots\,,\, M$$

Заметим, что это система из $M+1$ линейного уравнения на L чисел λ_k и она становится однозначно разрешимой при $M = L-1$, поскольку определитель этой системы — определитель Вандермонда и, следовательно, отличен от нуля, поэтому веса λ_k существуют и единственны. Отметим также, что явное выражение для весов имеет вид

$$\lambda_k = \int\limits_a^b \prod_{j\neq k} \frac{(x - x_j)}{(x_k - x_j)} \rho(x)dx \ , \tag{7}$$

что естественно совпадает с (5) при $\rho(x) \equiv 1$.

Итак, алгебраическая степень точности не может превышать величину $2L-1$, а может ли она равняться этому числу? — Да, может!

Определение. Квадратурные формулы наивысшей алгебраической степени точности ($M = 2L-1$) называются *квадратурными формулами Гаусса-Кристофеля.*

Займемся построением формул Гаусса-Кристофеля. Если узлы уже известны, то веса можно λ_k определить используя определитель Вандермонда (и получить выражение (7)), но это гарантирует алгебраическую степень точности лишь до значения $M = L-1$. Значит вопрос заключается в "разумном"расположении узлов x_k. Для решения этой задачи нам потребуются некоторые сведения об ортогональных полиномах (корни которых и являются узлами квадратурных формул Гаусса-Кристофеля).

4.3.2 Ортогональные полиномы

Теорема. *Пусть задана весовая функция ρ со свойствами 1)-3), тогда в $L_{2,\rho}$ существует и единственна полная система ортогональных полиномов $P_n(x)$:*

$$\langle P_n, P_m \rangle_{L_{2,\rho}} = \int\limits_a^b P_n(x)P_m(x)\rho(x)dx = \delta_{nm}||P_n||^2_{L_{2,\rho}} \ ,$$

такая что $\deg P_n = n$.

Напомним,что система векторов $\{\varphi_i\}$ нормированного пространства E, называется полной если наименьшее замкнутое (т.е. содержащее все свои предельные точки) подпространство, содержащее $\{\varphi_k\}$, есть все E. В конечномерном нормированном пространстве всякое подпространство автоматически замкнуто. бесконечномерном случае это не так. Например, в пространстве непрерывных функций $C_{[a,b]}$ (со своей нормой: $||f|| = \max\limits_{x\in[a,b]} |f(x)|$) полиномы образуют подпространство, но не замкнутое. Однако, в силу теоремы Вейерштрасса, система функций $\{x^k\}_{k=0}^{\infty}$ является полной в $C_{[a,b]}$.

Доказательство. Докажем существование и единственность без проверки полноты. Предъявим эти полиномы с точностью до множителя явно:

$$P_n(x) = A_n \begin{vmatrix} c_0 & c_1 & \dots & c_n \\ c_1 & c_2 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ 1 & x & \dots & x^n \end{vmatrix} .$$

Здесь, A_n – нормировочные константы. Для проверки существования, необходимо убедиться, что $P_n \perp x^m$, $m < n$. Действительно

$$\int_a^b x^m P_n(x)\rho(x)dx = A_n \int_a^b x^m \begin{vmatrix} c_0 & c_1 & \dots & c_n \\ c_1 & c_2 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ 1 & x & \dots & x^n \end{vmatrix} \rho(x)dx =$$

$$= A_n \begin{vmatrix} c_0 & c_1 & \dots & c_n \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ c_m & c_{m+1} & \dots & c_{m+n} \end{vmatrix} = 0 ,$$

если $m \le n-1$ (определитель с двумя одинаковыми строками). Таким образом ортогональные полиномы существуют.

Поскольку степени x^m линейно независимы, то ортогональные полиномы можно также построить и стандартной процедурой ортогонализации (Гильберта-Шмидта):

$$P_0 = \frac{1}{||1||_{L_{2,\rho}}} , \; P_1 = \frac{x - \langle x, 1\rangle_{L_{2,\rho}} 1}{||x - \langle x, 1\rangle_{L_{2,\rho}} 1||_{L_{2,\rho}}} , \dots ,$$

$$P_l = \frac{x^l - \sum_{k=1}^{l-1} \langle x^l, P_k\rangle_{L_{2,\rho}} P_k}{||x^l - \sum_{k=1}^{l-1} \langle x^l, P_k\rangle_{L_{2,\rho}} P_k||_{L_{2,\rho}}} .$$

Проверим теперь единственность. Пусть существует другой полином G_k степени k, такой что $G_k \perp P_i$, $i = 1, \dots, k-1$. Разложим его по системе P_k: $G_k = \sum_{i=0}^{k} c_i P_i$. Домножим это равенство на P_l и проинтегрируем с весом ρ по отрезку $[a, b]$ (т.е. рассмотрим скалярное произведение), тогда $\langle g_k, P_l\rangle = 0 = c_l$ при $l < k$ и, следовательно $g_k = c_k P_k$.

Вопрос: А где мы используем свойства ρ?

Дело в том, что если ρ удовлетворяет свойствам 1)-3), то форма

$$\langle f, g\rangle_{L_{2,\rho}} = \int_a^b f(x)\bar{g}(x)\rho(x)dx$$

действительно определяет скалярное произведение.

4.3.3 Свойства ортогональных полиномов

Путь задана система ортогональных с весом ρ полиномов $P_n(x)$. Справедлива

Теорема. *Все корни* $P_n(x)$ *вещественные, простые и принадлежат отрезку* (a,b) .

Доказательство. Пусть $P_n(x)$ имеет k вещественных корней x_i на отрезке (a,b) нечетной кратности. Положим

$$q_k(x) = \begin{cases} 1, & k = 0 \\ \prod\limits_{j=1}^{k}(x - x_j), & k > 0 \end{cases} ,$$

где корни $x_i \in (a,b)$ взятые без учета кратности, т.е. входят в произведение только один раз. Тогда произведение $P_n(x)q_k(x)$ не меняет знак на промежутке (a,b) , и, следовательно,

$$\int\limits_a^b P_n(x)q_k(x)\rho(x)dx \neq 0 .$$

Однако при $k < n$ интеграл должен равняться 0 в силу ортогональности P_n полиномам меньшей степени. Таким образом $k = n$.

Теорема. *Если алгебраическая степень точности квадратурной формулы с* L *узлами* x_k *равна* $2L-1$, *то узлы* x_k *— суть корни ортогонального полинома* $P_L(x)$.

Доказательство. Пусть $\mathcal{N}_L(x) = \prod\limits_{i=1}^{L}(x - x_i)$, где x_i — узлы квадратурной формулы и пусть её алгебраическая степень точности равна $2L - 1$. Рассмотрим функцию $f(x) = \mathcal{N}_L(x)P_m(x)$, где $m \leq L - 1$, являющуюся полиномом степени не превосходящей $2L - 1$. Для такой функции квадратурная формула точна по условию, и, следовательно,

$$\int\limits_a^b f(x)\rho(x)dx = \sum_{k=1}^{L} f(x_k)\lambda_k = 0 ,$$

то есть $\int\limits_a^b \mathcal{N}_L(x)P_m(x)dx = 0$ и значит $\mathcal{N}_L \perp P_m$. Таким образом $\mathcal{N}_L$ является ортогональным полиномом и в силу единственности с точностью до множителя совпадает с P_L .

Пусть теперь корни x_i ортогонального полинома $P_L(x)$ являются узлами квадратурной формулы. Покажем, что алгебраическая степень точности квадратурной формулы может равняться $2L - 1$. Проаппроксимируем функцию $f(x)$ полиномом $g(x)$ степени $L - 1$ по ее значениям в точках x_i :

$$g(x) = \sum_{i=1}^{L} f(x_i)\mathcal{L}_i(x) \quad , \quad \mathcal{L}_i(x) = \prod_{j\neq i}^{L} \frac{(x - x_j)}{(x_i - x_j)} .$$

Пусть $I = \int\limits_a^b f(x)\rho(x)dx$, $J = \int\limits_a^b g(x)\rho(x)dx$. Тогда $I = J$, если f — полином степени до $L - 1$ включительно, поскольку в этом случае $f = g$. Но если f | полином степени до $2L - 1$, то разность полиномов f и g также полином степени не превосходящей $2L - 1$, причем $(f - g)|_{x=x_j} = 0$, и, следовательно, справедливо представление $f - g = P_L q_{L-1}$, где q_{L-1} — некоторый полином степени до $L - 1$. Тогда

$$I - J = \int\limits_a^b \rho(x)[f(x) - g(x)]dx = \int\limits_a^b \rho(x)P_L(x)q_{L-1}(x)dx = 0 ,$$

то есть алгебраическая степень точности квадратурной формулы равна $2L-1$, если её узлы — корни ортогонального полинома. Веса при этом равны

$$\lambda_k = \int\limits_a^b \mathcal{L}_k(x)\rho(x)dx = \int\limits_a^b \prod_{i\neq k}^{L} \frac{(x-x_i)}{(x_k-x_i)}\rho(x)dx \ .$$

Отметим, что корни соседних ортогональных полиномов P_L и P_{L-1} различны (на самом деле между двумя последовательными корнями x_i и x_{i+1} полинома P_L лежит ровно один корень $\tilde{x}_i$ полинома P_{L-1}). Действительно, пусть $f_k(x) = \frac{P_L(x)P_{L-1}(x)}{(x-x_k)}$, тогда $\deg f_k = 2L-1$ и формула Гаусса-Кристофеля с узлами x_i , являющимися корнями полинома P_L , для такой функции точна. Она, как легко увидеть принимает вид

$$\lambda_k P'_L(x_k)P_{L-1}(x_k) = \int\limits_a^b \frac{P_L(x)P_{L-1}(x)}{(x-x_k)}\rho(x)dx \ .$$

Пусть a_i — старшие коэффициенты ортогональных полиномов P_i , тогда

$$P_L(x) = a_L\prod(x-x_i) \ , \quad P_{L-1}(x) = a_{L-1}\prod(x-\tilde{x}_i) \ ,$$

и справедливо представление

$$\frac{P_L(x)}{x-x_k} = \frac{a_L}{a_{L-1}}P_{L-1}(x) + q_{L-2}(x) \ ,$$

где $q_{L-2}(x)$ — некоторый полином степени не выше $L-2$. Таким образом с учетом ортогональности

$$\lambda_k = \frac{1}{P'_L(x_k)P_{L-1}(x_k)}\int\limits_a^b \frac{a_L}{a_{L-1}}P^2_{L-1}(x)\rho(x)dx = \frac{a_L||P_{L-1}||^2_{L_{2,\rho}}}{a_{L-1}P'_L(x_k)P_{L-1}(x_k)} \ ,$$

но так как $\lambda_k \neq \infty$ (для весов уже получено явное выражение (7), да и кроме того, зная узлы, веса можно однозначно определить через определитель Вандермонда), то $P_{L-1}(x_k) \neq 0$, и значит ни один из корней полинома P_L не может являться корнем полинома P_{L-1} . Попутно мы нашли и другое выражение для весов λ_k .

Свойства весов

1) $\lambda_k > 0$.

Доказательство. Пусть $f_k(x) = \left[\frac{\mathcal{N}_L(x)}{x-x_k}\right]^2$. Это полином степени $2L-2$, равный 0 во всех узлах, кроме $x = x_k$, для него формула Гаусса-Кристофеля точна, поэтому

$$\int\limits_a^b \left[\frac{\mathcal{N}_L(x)}{x-x_k}\right]^2\rho(x)dx = \lambda_k\left[\frac{\mathcal{N}_L(x)}{x-x_k}\Big|_{x=x_k}\right]^2 = \lambda_k\left[\mathcal{N}'_L(x_k)\right]^2 > 0 \ ,$$

следовательно $\lambda_k > 0$.

2) связь весов λ_k с моментами $c_l = \int\limits_a^b x^l\rho(x)dx$:

$$\sum_{k=1}^{L} x_k^l\lambda_k = c_l \ , \ l = 0\,,\,1\,,\,\ldots\,,\,2L-1 \ .$$

Свойство становится очевидным, если сосчитать интеграл с весом от степени x^l по формуле Гаусса-Кристофеля.

3) $\sum\limits_{k=1}^{L}\lambda_k = \int\limits_a^b \rho(x)dx$.

Это частный случай свойства 2) при $l = 0$.

4.3.4 Примеры ортогональных полиномов

1) Полиномы Лежандра $P_n(x)$ являются ортогональными на промежутке (-1,1) с весом $\rho(x) = 1$. С точностью до нормировки для них справедливо выражение

$$P_n(x) = \frac{(-1)^n}{2^n n!}\frac{d^n}{dx^n}(1-x^2)^n \ .$$

В частности $P_0 = 1$, $P_1 = x$, $P_2 = \frac{1}{2}(3x^2-1)$.

2) Полиномы Чебышева первого рода

$$T_n = \frac{n}{2}\sum_{m=0}^{[n/2]}\frac{(-1)^m(n-m-1)!}{m!(n-2m)!}(2x)^{n-2m}$$

ортогональны на том же промежутке $[-1,1]$, с весом $\rho = \frac{1}{\sqrt{1-x^2}}$.

3) Полиномы Эрмита H_n ортогональны на промежутке $(-\infty,\infty)$, с весом $\rho(x) = e^{-x^2}$. С точностью до нормировки они имеют вид

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

4) Полиномы Лагерра L_n ортогональны на промежутке $[0,\infty)$, с весом $\rho(x) = e^{-x}$. Их можно представить в виде

$$L_n(x) = \frac{1}{n!}e^x\frac{d^n}{dx^n}(x^n e^{-x}) \ .$$

4.3.5 Погрешность квадратурных формул

Пусть функция $f(x)$ проинтерполирована по её значениям $f(x_i)$ в L точках x_i , $i = 1,2,\ldots,L$, полиномом g_{L-1} :

$$f(x) = g_{L-1}(x) + r(x) \ , \quad g_{L-1}(x) = \sum_{i=1}^{L} f(x_j)\prod_{j\neq i}^{L}\frac{(x-x_j)}{(x_i-x_j)} \ .$$

Погрешность интегрирования R при замене $f(x)$ интерполяционным полиномом g_{L-1} (она же — погрешность соответствующей квадратурной формулы) имеет вид

$$R = \int_a^b f\rho dx - \int_a^b g_{L-1}\rho dx = \int_a^b r(x)\rho dx = \int_a^b \frac{f^{(L)}(\xi(x))}{L!}\mathcal{N}(x)\rho(x)dx \ ,$$

$$\mathcal{N}(x) = \prod_{i=1}^{L}(x-x_i)$$

и если f — полином степени не выше $L-1$, то $f^{(L)} \equiv 0$ и, следовательно, квадратурная формула точна. Для случая равноотстоящих узлов $x_i - x_{i-1} = h$ имеем:

$$\frac{|\mathcal{N}_L(x)|}{L!} \leq h^L \max_k\left[\frac{1}{C_L^k}\right] \leq h^L$$

и, значит

$$|R| \leq h^L ||f^{(L)}||_C \int_a^b \rho(x)dx \ ,$$

и при $\rho = 1$

$$|R| \leq h^L ||f^{(L)}||(b-a) \ .$$

Это довольно грубая оценка, однако она показывает порядок точности по h .

В случае, если узлы не произвольные, а корни ортогонального полинома P_L , то квадратурная формула точна для полиномов степени не превосходящей $2L-1$, хотя полученная оценка этого и не "чувствует". Чтобы улучшить оценку в этом случае поступим следующим образом. Пусть $f \in C^{2L}$. Разложим ее в ряд Тейлора в окрестности некоторой точки x_*:

$$f(x) = \underbrace{\sum_{k=0}^{2L-1} \frac{f^{(k)}(x_*)(x-x_*)^k}{k!}}_{f_1(x)} + \underbrace{\frac{f^{(2L)}(x_*)(x-x_*)^{2L}}{(2L)!} + q(x)}_{f_2(x)} ,$$

тогда

$$R = \int_a^b [f - g_{L-1}(x)]\rho(x)dx = \underbrace{\int_a^b [f_1 - g_{L-1}(x)]\rho(x)dx}_{=0} + \int_a^b f_2(x)\rho(x)dx .$$

Пусть вес $\rho = 1$, оценим последний интеграл отбросив от функции $f_2(x)$ остаток $q(x)$ и выбрав в качестве точки разложения x_* точку $\frac{a+b}{2}$, тогда

$$\int_a^b \frac{f^{(2L)}(x_*)(x-x_*)^{2L}}{(2L)!}dx \le \frac{||f^{(2L)}||_C}{2^{2L}(2L+1)!}(b-a)^{2L+1} ,$$

то есть погрешность R ведет себя как

$$R \sim \frac{||f^{(2L)}||_C}{2^{2L}(2L+1)!}(b-a)^{2L+1} .$$

4.4 Примеры квадратурных формул

В этом пункте мы будем считать, что вес $\rho = 1$, и, что L — число узлов на $[a,b]$.

4.4.1 Число узлов $L = 1$

а) Формула левых прямоугольников: $x_1 = a$, $\int_a^b f(x)dx \approx (b-a)f(a)$.

б) Формула правых прямоугольников: $x_1 = b$, $\int_a^b f(x)dx \approx (b-a)f(b)$.

в) Формула средних (прямоугольников)

— формула наивысшей алгебраической степени точности (она должна быть точной для полиномов не превосходящих степени $2L-1=1$). Построим ее в соответствии с изложенными выше соображениями для формул Гаусса-Кристофеля. Для этого сначала с помощью масштабного преобразования и сдвига переведем отрезок $[a,b]$ в отрезок $[-1,1]$, на котором ортогональными являются полиномы Лежандра P_k . Тогда

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^1 \underbrace{f(\frac{b+a}{2} + \frac{b-a}{2}y)}_{q(y)}dy , \quad x = \frac{b+a}{2} + \frac{b-a}{2}y .$$

Поскольку $P_1(y) = y$, то единственный корень этого полинома точка $y = 0$. Вес λ (по свойству весов $\sum \lambda_i = \int_a^b \rho(x)dx$) равен $\lambda = \int_{-1}^1 dx = 2$, таким образом

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^1 q(y)dy \approx \frac{b-a}{2}2q(0) = (b-a)f(\frac{b+a}{2}) .$$

4.4.2 Число узлов $L = 2$

а) Формула трапеций.

Здесь узлами являются точки $x_1 = a\ ,\ x_2 = b.$ $f(x)$ заменяется интерполяционным полиномом первой степени $p_1(x)$, построенным по этим узлам:

$$f(x)\ \rightarrow\ g_1(x) = \frac{x-b}{a-b}f(a) + \frac{x-a}{b-a}f(b)\ ,$$

$$\int_a^b f(x)dx \approx f(a)\int_a^b \frac{x-b}{a-b}dx + f(b)\int_a^b \frac{x-a}{b-a}dx =$$

$$= \frac{f(a)}{a-b}\int_a^b y\,dy + \frac{f(b)}{b-a}\int_a^b y\,dy = -\frac{f(a)}{(a-b)}\frac{(a-b)^2}{2} + \frac{f(b)}{(b-a)}\frac{(b-a)^2}{2} =$$

$$= (b-a)\frac{f(a)+f(b)}{2}\ .$$

Эта формула разумеется точна для полиномов степени не превосходящей $L-1=1$ (и не больше).

б) Формула Гаусса-Кристофеля для $L = 2$

Для ее получения поступим так же как в случае формулы средних:

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} q(y)dy\ \ ,\quad q(y) = f(\frac{b+a}{2} + \frac{b-a}{2}y)\ .$$

Полином P_2 имеет вид: $P_2 = \frac{1}{2}(3y^2-1)$. Его корни $y_1 = -\frac{1}{\sqrt{3}}\ ,\ y_2 = \frac{1}{\sqrt{3}}$. Веса из симметричности должны быть одинаковы: $\lambda_1 = \lambda_2\ ,\ \lambda_1+\lambda_2 = 2 \Rightarrow \lambda_i = 1$, следовательно, $\int_{-1}^{1} g(y)dy \approx q(-\frac{1}{\sqrt{3}}) + q(\frac{1}{\sqrt{3}})$. Таким образом искомая квадратурная формула имеет вид

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\left[f(\frac{b+a}{2} - \frac{b-a}{2}\frac{1}{\sqrt{3}}) + f(\frac{b+a}{2} + \frac{b-a}{2}\frac{1}{\sqrt{3}})\right]\ .$$

Алгебраическая степень точности M равна $2L-1 = 3$.

4.4.3 Число узлов $L = 3$

Формула Симпсона

Здесь узлами являются точки $x_1 = a\ ,\ x_2 = \frac{a+b}{2}\ ,\ x_3 = b.$ Для удобства вычислений перейдем к отрезку $[-1,1]$ масштабным преобразованием $q(y) = f(\frac{b+a}{2} + \frac{b-a}{2}y)$:

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} q(y)dy\ \ ,\ \ y_1 = -1\ ,\ y_2 = 0\ ,\ y_3 = 1\ .$$

Заменим $q(y)$ интерполяционным полиномом $p_2(y)$:

$$q(y)\ \rightarrow\ p_2(y) = q(-1)\mathcal{L}_1(y) + q(0)\mathcal{L}_2(y) + q(1)\mathcal{L}_3(y)\ ,$$

где

$$\mathcal{L}_1(y) = \frac{(y-0)(y-1)}{(-1-0)(-1-1)} = \frac{y(y-1)}{2}\ \ ,\ \ \mathcal{L}_2(y) = \frac{(y-(-1))(y-1)}{(0-(-1))(0-1)} = \frac{(y+1)(y-1)}{-1}\ ,$$

$$\mathcal{L}_3(y) = \frac{(y-(-1))(y-0)}{(1-(-1))(1-0)} = \frac{(y+1)y}{2} \ .$$

Тогда интеграл $\int\limits_{-1}^{1} p_2(y)dy$ равен

$$q(-1)\int\limits_{-1}^{1}\frac{y(y-1)}{2}dy + q(0)\int\limits_{-1}^{1}\frac{(y+1)(y-1)}{-1}dy + q(1)\int\limits_{-1}^{1}\frac{y(y+1)}{2}dy \ .$$

Сосчитаем веса

$$\lambda_3 = \lambda_1 = \int\limits_{-1}^{1}(\frac{y^2}{2} - \frac{y}{2})dy = \frac{1}{3} \ , \quad \lambda_2 = \int\limits_{-1}^{1}(1-y^2)dy = \frac{4}{3} \ ,$$

таким образом $\int\limits_{-1}^{1} q(y)dy \approx \frac{q(-1)}{3} + \frac{4}{3}q(0) + \frac{q(1)}{3}$. Возвращаясь к исходной функции f, получаем *формулу Симпсона*

$$\int\limits_{a}^{b} f(x)dx \approx \frac{b-a}{6}[f(a) + 4f(\frac{a+b}{2}) + f(b)] \ .$$

Заметим, что эта формула точна и для полиномов третьей степени, хотя построение гарантировало точность лишь до значения $L-1=2$.

Для более точного вычисления интегралов можно строить интерполяционные полиномы все более высокой степени, однако более разумным подходом является разбиение промежутка интегрирования на части и применение на них какого либо из изложенных выше простых способов интегрирования.

4.5 Составные квадратурные формулы

Разобьем промежуток интегрирования $[a,b]$ на N частей $x_0 = a \, , x_1 \, , \ldots \, , x_N = b$ и на каждом промежутке $\Delta_i = [x_i, x_{i-1}]$ применим ту или иную квадратурную формулу и просуммируем по всем промежуткам. Пусть $h_i = x_i - x_{i-1}$. Получаем следующие составные квадратурные формулы

$M = \sum\limits_{i=1}^{N} h_i f(\frac{x_i + x_{i-1}}{2})$;

$T = \sum\limits_{i=1}^{N} h_i \frac{f(x_i)+f(x_{i-1})}{2}$;

$S = \sum\limits_{i=1}^{N} \frac{h_i}{6}[f_{i-1} + 4f(\frac{x_{i-1}+x_i}{2}) + f(x_i)]$.

Любопытно отметить, что $S = \frac{2}{3}M + \frac{1}{3}T$.

Удобно составную формулу Симпсона представлять в виде (при четном числе промежутков)

$$\bar{S}(f) = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + \ \ldots \ + 4f_{N-1} + f_N) \ .$$

Такая запись называется *обобщенной формулой Симпсона.*

4.5.1 Сходимость квадратурных формул

Устремим в составных квадратурных формулах ранг дробления $h = \max h_i$ к нулю. Естественным образом возникают вопросы

1) Стремится ли сумма к интегралу?

2) Если "да", то с какой скоростью?

Ответ на первый вопрос положителен. Поскольку и формула средних M и формула трапеций T — суть интегральные суммы, а для интегрируемой функции интеграл по определению есть предел интегральных сумм. Поскольку формула Симпсона S является линейной комбинацией (с суммой коэффициентов равной 1) формул средних и трапеций, то при ранге дробления стремящимся к нулю, она также стремится к интегралу. Нетрудно доказать сходимость и других квадратурных формул.

Теперь обратимся к вопросу о скорости сходимости. Поскольку формулы трапеций T и средних M точны для полиномов степени не превосходящей 1 , то естественно ожидать, что их погрешность есть $O(h^2)$, а для формулы Симпсона, имеющей алгебраическую степень точности равную трем, погрешность — $O(h^4)$.

Рассмотрим ситуацию детально. Пусть $\bar{x}_i = \frac{x_i + x_{i-1}}{2}$. Разложим $f(x)$ в ряд Тейлора в окрестности точки $\bar{x}$.

$$f(x) = f(\bar{x}_i) + (x - \bar{x}_i)f'(\bar{x}_i) + \frac{1}{2}(x - \bar{x}_i)^2 f''(\bar{x}_i) +$$
$$+ \frac{(x - \bar{x}_i)^3}{3!} f'''(\bar{x}_i) + \frac{(x - \bar{x}_i)^4}{24} f^{(4)}(\bar{x}_i) + \frac{(x - \bar{x}_i)^5}{120} f^{(5)}(\bar{x}_i) + O(h_i^6) \ .$$

Проинтегрируем это разложение по промежутку $[x_{i-1}, x_i]$. Заметим, что при этом все члены Тейлоровского разложения с нечетными степенями $(x - \bar{x}_i)$ пропадут из-за симметрии расположения точки $\bar{x}_i$. Таким образом

$$\int_{x_{i-1}}^{x_i} f(x)dx = h_i f(\bar{x}_i) + \frac{h_i^3}{3!2^2} f''(\bar{x}_i) + \frac{h_i^5}{5!2^4} f^{(4)}(\bar{x}_i) + \frac{h_i^7}{7!2^6} f^{(6)}(\bar{x}_i) \ + \dots \ . \tag{8}$$

Из тейлоровского разложения также нетрудно видеть, что

$$\frac{f(x_i) + f(x_{i-1})}{2} = f(\bar{x}_i) + \frac{h_i^2}{2!2^2} f''(\bar{x}_i) + \frac{h_i^4}{4!2^4} f^{(4)}(\bar{x}_i) + O(h_i^6) \ ,$$

откуда

$$f(\bar{x}_i) = \frac{f(x_i) + f(x_{i-1})}{2} - \frac{h_i^2}{2!2^2} f''(\bar{x}_i) - \frac{h_i^4}{4!2^4} f^{(4)}(\bar{x}_i) - O(h_i^6) \ .$$

Подставляя это выражение в (8), получаем

$$\int_{x_{i-1}}^{x_i} f(x)dx = h_i \frac{f(x_i) + f(x_{i-1})}{2} - \frac{h_i^3}{12} f''(\bar{x}_i) + \frac{h_i^5}{4!2^4} f^{(4)}(\bar{x}_i)[\frac{1}{5} - 1] + O(h_i^6) \ . \tag{9}$$

Далее, поскольку $S = \frac{2}{3}M + \frac{1}{3}T$, то

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{h_i}{3} \left[2f(\bar{x}_i) + \frac{(f(x_i) + f(x_{i-1}))}{2} \right] +$$
$$+ \frac{f^{(4)}(\bar{x}_i) h_i^5}{4!2^4} \left[\frac{2}{3} \cdot \frac{1}{5} - \frac{1}{3} \cdot \frac{4}{5} \right] + O(h_i^6) = S + \frac{f^{(4)}(\bar{x}_i) h_i^5}{4!2^4 3 \cdot 5}[-2] + O(h_i^6) \ . \tag{10}$$

Итого, для равноотстоящих узлов из (8) погрешность составной формулы средних δ_M равна

$$\delta_M = \int_a^b f(x)dx - M = \int_a^b f(x)fdx - \sum_{i=1}^N h_i f(\bar{x}_i) = -\sum_{i=1}^N \frac{h_i^3}{24} f''(\bar{x}_i) + O(h^5) \ ,$$

то есть

$$|\delta_M| \le \frac{1}{24} \sum_{i=1}^N h_i^3 ||f''||_C = \frac{h^2 ||f''||_C}{24} \sum_{i=1}^N h_i = \frac{(b-a)}{24} ||f''||_C h^2 \ .$$

Из (9), аналогично

$$|\delta_T| \le \frac{(b-a)}{12}||f''||_C h^2 \ .$$

Из (10)

$$|\delta_S| \sim \frac{||f^{(4)}||_C h^4}{6!4}(b-a) \ .$$

Здесь имеется в виду составная формула Симпсона S . Для обобщенной формулы Симпсона надо h заменить на $2h$:

$$|\delta_{\bar{S}}| \sim \frac{||f^{(4)}||_C}{6!}\frac{2^4 h^4}{2^2}(b-a) = \frac{M_4}{180}h^4(b-a) \ .$$

4.6 Другие формулы

4.6.1 Сплайн-квадратура

Для приближенного интегрирования можно также использовать сплайны. Именно, интегрируемая функция заменяется сплайном, который и интегрируется.

Пусть $x \in \Delta_i = [x_{i-1}, x_i]$, $h_i = x_i - x_{i-1}$, $\omega = 1 - \bar{\omega} = \frac{x - x_{i-1}}{h_i}$. Применим сплайн S_3^1 для приближенного интегрирования. Заметим, что на промежутке Δ_i его можно представить в виде:

$$S_3^1(x) = \omega f_i + \bar{\omega} f_{i-1} + \frac{h_i^2}{6}[(\omega^3 - \omega)M_i + (\bar{\omega}^3 - \bar{\omega})M_{i-1}] \ .$$

Здесь $M_i = S''(x_i)$. Пусть $S(\omega) = S_3^1(x)$, тогда

$$\int\limits_{x_{i-1}}^{x_i} S_3^1(x)dx = h_i \int\limits_0^1 S(\omega)d\omega \ , \ (dx = h_i d\omega) \ .$$

При этом $\int\limits_0^1 \omega d\omega = \frac{1}{2}$, $\int\limits_0^1 (\omega^3 - \omega)d\omega = -\frac{1}{4}$. Таким образом

$$\int\limits_{x_{i-1}}^{x_i} S_3^1(x)dx = h_i \frac{f_i + f_{i-1}}{2} - h_i^3 \frac{M_i + M_{i-1}}{24}.$$

Последний член в этой формуле "имитирует"поправку к формуле трапеций. Действительно, вторая производная от сплайна, аппроксимирует вторую производную от функции и

$$\frac{h_i^3(M_i + M_{i-1})}{24} \approx \frac{h_i^3}{12} f''(\bar{x}_i) \ ,$$

что представляет собой поправочный член формулы трапеций (см. формулу (9)). Таким образом происходит компенсация ошибки формулы трапеций. Окончательно

$$\int\limits_a^b f(x)dx \approx \sum_{i=1}^{N} \left(h_i \frac{f_i + f_{i-1}}{2} - h_i^3 \frac{M_i + M_{i-1}}{24}\right) \ .$$

Замечание. Сплайн-квадратура не есть квадратурная формула в чистом виде, поскольку она использует не только значения функции, но и вторые производные от сплайна.

4.6.2 Формулы Филона

Пусть $I = \int\limits_a^b f(x)e^{i\omega x}dx$, $|\omega| >> 1/|b-a|$, а $f(x)$ медленно меняющаяся относительно периода $T = 2\pi/\omega$ колебаний, функция. В этом случае подынтегральная функция $f(x)e^{i\omega x}$ имеет много осцилляций на промежутке (a, b) и использование обычных квадратурных формул весьма затруднено, поскольку приходится делить промежуток интегрирования на большое количество частей. Однако нет необходимости заменять всю подынтегральную функцию интерполяционным полиномом. Достаточно эту процедуру проделать лишь с функцией $f(x)$. Итак, заменим f интерполяционным полиномом p. Тогда

$$f(x) \approx p(x) = \sum_{j=0}^{N} \mathcal{L}_j(x) f(x_j) \quad , \quad \mathcal{L}_j(x) = \prod_{k\neq j, k=0}^{N} \frac{(x-x_k)}{(x_j-x_k)} \ ,$$

$$J = \int\limits_a^b p(x)e^{i\omega x}dx = \sum_{j=0}^{N} f(x_j) \int\limits_a^b e^{i\omega x}\mathcal{L}_j(x)dx = \sum_{j=0}^{N} A_j(\omega) f(x_j) \ .$$

Интегралы $A_j(\omega) = \int\limits_a^b e^{i\omega x}\mathcal{L}_j(x)dx$ берутся в элементарных функциях. Получаемые при этом формулы приближенного интегрирования называются *формулами Филона*:

$$I = \int\limits_a^b f(x)e^{i\omega x}dx \approx \sum_{j=0}^{N} A_j(\omega) f(x_j) \ .$$

Задача: Для интегралов $\int\limits_{-1}^{1} \sin\omega x f(x)dx$, $\int\limits_{-1}^{1} \cos\omega x f(x)dx$ получить формулу Филона с тремя узлами: $x_0 = 1$, $x_1 = 0$, $x_2 = 1$.

4.6.3 Составные формулы Филона

Разобьем промежуток $[a, b]$ на N частей $a = x_0 < x_1 < \ldots < x_N = b$ и на каждом промежутке $[x_{k-1}, x_k]$ заменим $f(x)$ интерполяционным полиномом p^k некоторой степени, тогда

$$I = \int\limits_a^b f(x)e^{i\omega x}dx = \sum_{k=1}^{N} \int\limits_{x_{k-1}}^{x_k} f(x)e^{i\omega x}dx \sim J = \sum_{k=1}^{N} \int\limits_{x_{k-1}}^{x_k} p^k(x)e^{i\omega x}dx \ .$$

Пример. Аналог формулы средних.

$$\int\limits_{x_{k-1}}^{x_k} f(x)e^{i\omega x}dx \sim \int\limits_{x_{k-1}}^{x_k} f(\bar{x})e^{i\omega x}dx =$$

$$= f(\bar{x}_k)\frac{e^{i\omega x_k} - e^{i\omega x_{k-1}}}{i\omega} = \frac{2}{\omega} f(\bar{x}_k)e^{i\omega\bar{x}_k} \sin\frac{\omega h_k}{2} \ .$$

Оценим погрешность этой формулы. Представим $f(x)$ приближенно: $f(x) \approx f(\bar{x}_k) + f'(\bar{x}_k)(x - \bar{x}_k)$. Тогда погрешность R приближенно описывается выражением

$$R = \int\limits_{x_0}^{x_N} r(x)e^{i\omega x}dx \approx \sum_{k=1}^{N} f'(\bar{x}_k) \int\limits_{x_{k-1}}^{x_k} (x - \bar{x}_k)e^{i\omega x}dx =$$

$$= \frac{2i}{\omega^2} \sum_{k=1}^{N} f'(\bar{x}_k) \left(\sin\frac{\omega h_k}{2} - \frac{\omega h_k}{2}\cos\frac{\omega h_k}{2} \right) e^{i\omega\bar{x}_k} \ ,$$

т.е. если произведение ωh_k порядка 1, то формула Филона имеет небольшую погрешность, в противном случае погрешность того же порядка, что и значение интеграла.

Глава 5

Поиск минимума

5.1 Случай одной переменной

5.1.1 Метод золотого сечения

Пусть $\Phi(x) : [a, b] \to \mathbf{R}$ и известно, что на промежутке $[a, b]$ функция Φ имеет хотя бы один локальный минимум. Для применения излагаемого ниже метода золотого сечения, от функции $\Phi(x)$ не требуется даже непрерывность, достаточно лишь кусочной непрерывности. Будем пока считать, что Φ имеет на промежутке лишь один локальный минимум (он же и глобальный).

Метод основан на сравнении значений функции в различных точках, с последующим отбрасыванием промежутков, на которых минимум уж точно не может находиться. Ясно, что чтобы осуществлять подобную процедуру, необходимо знать значения функции, вообще говоря, в 4-х точках. Действительно, пусть $a = x_0 < x_1 < x_2 < x_3 = b$, и пусть, скажем, в точке x_2 значение функции наименьшее из этих четырех величин. Тогда минимум Φ заведомо не может находиться на промежутке $[x_0, x_1]$ и поэтому этот промежуток можно отбросить. Теперь на оставшемся промежутке $[x_1, x_3]$ нам известны крайние значения функции и значение в одной внутренней точке. Добавляя новую точку x_4 мы можем повторить сравнение значений Φ и вновь сузить допустимый промежуток. Как наиболее разумно размещать добавляемые точки? Представляется естественным, чтобы деление отрезков происходило подобно предыдущему делению.

x_0_____x_1_____x_2___x_4_____x_3

Это означает, в частности, что внутренние точки должны располагаться симметрично, то есть $|x_1 - x_0| = |x_3 - x_2| = h$. Если длина исходного промежутка равна l , то должно выполняться соотношение

$$\xi = \frac{h}{l} = \frac{x_1 - x_0}{x_3 - x_0} = \frac{x_2 - x_1}{x_3 - x_1} = \frac{l - 2h}{l - h} \ ,$$

откуда

$$\xi = \frac{h}{l} = \frac{1 - 2\frac{h}{l}}{1 - \frac{h}{l}} = \frac{1 - 2\xi}{1 - \xi} \ .$$

Разрешая квадратное уравнение относительно ξ, получаем $\xi = \frac{3-\sqrt{5}}{2} \approx 0,38$, то есть на каждом шаге (за исключением вычисления стартовых внутренних точек x_1 и x_2) отрезок сокращается в $1/(1-\xi) \approx 1,61$ раза и сходимость метода линейная.

Таким образом, для того, чтобы начать процесс золотого сечения, к граничным точкам $x_0 = a$ и $x_3 = b$ добавляются две точки $x_1 = x_0 + \xi(x_3 - x_0)$ $x_2 = x_3 - \xi(x_3 - x_0)$. Затем после отбрасывания точек и добавления новых, на последующих шагах номера точек перемешаны беспорядочно. Дадим им номера j,k,l,m, и пусть $\Phi(x_j) < \Phi(x_{k,l,m})$. При делении по золотому сечению отбрасывается отрезок одним, из концов которого является точка наиболее удаленная от x_j. Пусть этой точкой является x_k (очевидно, что это одна из крайних точек). Затем надо добавить новую точку x_n. Пусть для определенности $x_k < x_j < x_m$. Тогда в силу симметрии расположения внутренних точек она определяется соотношением $x_n = x_k + x_m - x_i$ (т.е. сумма крайних точек минус внутренняя).

Если функция Φ имеет на исходном промежутке несколько локальных минимумов, то метод золотого сечения всё равно сойдется к одному из них, не обязательно к глобальному.

5.1.2 Метод парабол

Если функция обладает достаточной гладкостью (имеет вторую производную) то естественно использовать это обстоятельство при поиске минимума. В этой точке $\Phi'(x) = 0$, и можно искать нуль первой производной, скажем методом Ньютона

$$x_{n+1} = x_n - \frac{\Phi'(x_n)}{\Phi''(x_n)} .$$

Эту формулу легко получить и непосредственно разложив $\Phi(x)$ в ряд Тейлора в точке x_n и ограничившись тремя членами, т.е. аппроксимируя кривую параболой

$$\Phi(x) \approx \Phi(x_n) + (x - x_n)\Phi'(x_n) + \frac{(x - x_n)^2}{2}\Phi''(x_n) .$$

Минимум этой параболы находится как раз в точке x_{n+1}. В связи с этим метод и называется *методом парабол*.

Вычислять и первую и вторую производную на каждом шаге довольно накладно. Поэтому их приближенно заменяют разностными производными, вычисленными с помощью вспомогательного шага h:

$$\Phi'(x_n) \to \frac{\Phi(x_n + h) - \Phi(x_n - h)}{2h} ,$$

$$\Phi''(x_n) \to \frac{\Phi(x_n + h) - 2\Phi(x_n) + \Phi(x_n - h)}{h^2} ,$$

и метод принимает вид

$$x_{n+1} = x_n - \frac{h}{2}\frac{\Phi(x_n + h) - \Phi(x_n - h)}{\Phi(x_n + h) - 2\Phi(x_n) + \Phi(x_n - h)} .$$

Кстати, этот подход эквивалентен замене кривой на интерполяционную параболу, построенную по трем точкам $x_n - h, x_n, x_n + h$ с последующим нахождением минимума этой параболы (точки x_{n+1}).

Замечание. Уместно сравнить методы поиска минимума и методы поиска корня уравнения. Так метод золотого сечения подобен дихотомии. И в том и другом на функцию накладываются минимальные ограничения. Они чрезвычайно просты и надежны, порядок сходимости в обоих методах линейный. Метод парабол в этом смысле подобен методу Ньютона. От функции требуется больше, сходимость быстрее. Поиск минимума по методу парабол соответствует поиску корня по методу секущих.

5.2 Функции многих переменных

Пусть $\Phi : M \to \mathbf{R}$; $M \subset \mathbf{R}^N$ и пусть $\Phi \in C^2_M$. В точках минимума $\frac{\partial \Phi}{\partial x_i} = 0\,, i = 1\,, \ldots\,, N$, как впрочем и в точках максимума и в седловых точках. Но разложение в ряд Тейлора в окрестности невырожденной точки минимума $\mathbf{x}^*$

$$\Phi(\mathbf{x}) = \Phi(\mathbf{x}^*) + \frac{1}{2}\sum_{i,j=1}^{N} \frac{\partial^2 \Phi}{\partial x_i^2}\Delta x_i \Delta x_j + \ldots$$

выделено тем, что квадратичная форма $\sum\limits_{i,j=1}^{N} \frac{\partial^2 \Phi}{\partial x_i^2}\Delta x_i \Delta x_j$ положительно определена (напомним, что квадратичная форма называется положительно определенной, если $\sum a_{ij} z_i z_j \geq \gamma ||\mathbf{z}||^2$, $\gamma > 0$, где $\mathbf{z} = (z_1, z_2 \ldots, z_N)^T$).

5.2.1 Координатный спуск

Процедуру координатного спуска рассмотрим на примере функции двух переменных $\Phi(x.y)$. Пусть (x^0, y^0) — некоторая точка. Зафиксируем переменную y и найдем минимум функции $\Phi(x, y^0)$ каким либо из уже известных способов поиска минимума функции одного переменного (спуск по первой координате). Пусть этот минимум достигается в точке x^1. Зафиксировав это значение найдем минимум функции $\Phi(x^1, y)$ (спуск по второй координате). Пусть он находится в точке y^1. Теперь найдем минимум функции $\Phi(x, y^1)$ (следующий спуск по первой координате) и т.д. Такой метод поиска минимума называется *координатным спуском*. В зависимости от свойств функции и положения начальной точки, процесс может сойтись к экстремальной точке или нет. Отметим достаточный признак сходимости координатного спуска. Если Φ дважды непрерывно дифференцируема в области M, содержащей точку минимума x^* и квадратичная форма $\sum\limits_{i,j=1}^{N} \frac{\partial^2 \Phi}{\partial x_i^2}\Delta x_i \Delta x_j$ положительно определена, то в некоторой окрестности x^* метод координатного спуска сходится к указанному минимуму. Доказательство проведем для случая двух переменных. Пусть $\Phi_{xx} \geq a$, $\Phi_{yy} \geq b$, $|\Phi_{xy}| \leq c$, где $a, b, c > 0$ и $ab > c^2$ в области M (это означает в частности, что матрица вторых производных положительно определена). Будем считать, что точка $A = (x^0, y^0)$ получена в результате спуска по переменной y , т.е. $\Phi_y(A) = 0$. Пусть $|\Phi_x(A)| = \zeta_1$. В точке $B = (x^1, y^0)$ обращается в нуль Φ_x , а модуль Φ_y равен некоторому числу η. Таким образом

$$\zeta_1 = |\Phi_x(A) - \Phi_x(B)| = |\Phi_{xx}(\xi)|\rho(A, B) \geq a\rho(A, B)\ ,$$

$$\eta = |\Phi_y(A) - \Phi_y(B)| = |\Phi_{xy}(\xi')|\rho(A, B) \leq c\rho(A, B)\ ,$$

откуда

$$c\zeta_1 \geq a\eta \ . \tag{1}$$

В точке $C = (x^1, y^1)$: $\Phi_y = 0$, и $\Phi_x = \zeta_2$, при этом

$$\zeta_2 = |\Phi_x(C) - \Phi_x(B)| = |\Phi_{xy}(\tau)|\rho(C,B) \leq c\rho(C,B) \ ,$$

$$\eta = |\Phi_y(C) - \Phi_y(B)| = |\Phi_{yy}(\tau')|\rho(C,B) \geq b\rho(C,B) \ ,$$

и, следовательно,

$$c\eta \geq b\zeta_2 \ . \tag{2}$$

Из (1) и (2) заключаем, что $\zeta_2 \leq q\zeta_1$, где $0 \leq q \leq \frac{c^2}{ab} < 1$. Таким образом с каждым циклом Φ_x уменьшается как минимум в q раз. Аналогично убывает и частная производная по переменной y. Таким образом координатный спуск действительно сходится к точке минимума.

5.2.2 Наискорейший спуск

Спуск можно осуществлять не только вдоль координатных осей, а вообще вдоль любого направления. Пусть $\mathbf{a}$ произвольный единичный вектор, задающий направление. Функция $\varphi_0(t) = \Phi(\mathbf{r}^0 + \mathbf{a}t)$ есть функция одной переменной и ее минимум является минимумом функции $\Phi(\mathbf{r})$ на прямой $\mathbf{r}^0 + \mathbf{a}t$. Если выбрать $\mathbf{a} = \mathbf{a}^0 = -grad\Phi|_{\mathbf{r}^0}$, то $\mathbf{a}$ будет являться направлением наибольшего убывания функции Φ в точке $\mathbf{r}^0$. Осуществим спуск вдоль этого направления (то есть найдем минимум функции $\varphi(t)$). Пусть он находится в точке $\mathbf{r}^1$. Теперь в этой точке выберем новое направление $\mathbf{a}^1 = -grad\Phi|_{\mathbf{r}^1}$ и осуществим спуск вдоль него и т.д. Заметим, кстати, что векторы $\mathbf{a}^0$ и $\mathbf{a}^1$ ортогональны. Описанный метод спуска называется наискорейшим.

Хотя при наискорейшем спуске движение происходит вдоль направления наибольшего убывания функции в текущей точке $\mathbf{r}^k$, однако порядок сходимости остается таким же, как и при покоординатном спуске (при этом приходится в каждой точке $\mathbf{r}^n$ заново считать градиент). Дело здесь в том, что при сдвиге от точки $\mathbf{r}^n$ направление наибыстрейшего убывания функции Φ изменяется. Обычно спуск производят не точно до минимума, а несколько меньше. То есть, если $\varphi_n(t) = \Phi(\mathbf{r}^n + \mathbf{a}^n t)$, и минимум функции $\varphi_n(t)$ достигается в точке t_n , то спуск осуществляется до точки αt_n , где $\alpha < 1$. В "идеале"можно спускаться на бесконечно малую величину и заново корректировать направление. При этом кривая спуска $\mathbf{r}(t)$ будет удовлетворять уравнению

$$\frac{d\mathbf{r}(t)}{dt} = -grad\Phi(\mathbf{r}(t)) \ .$$

Однако интегрирование такой системы уравнений в частных производных представляет собой отдельную непростую задачу.

5.2.3 Метод сопряженных направлений

В окрестности точки минимума функция $\Phi(\mathbf{r})$ ведет себя обычно как квадратичная функция. Будем считать для начала, что она является в точности квадратичной, т.е.

$$\Phi(\mathbf{r}) = \langle \mathbf{r}, A\mathbf{r} \rangle + \langle \mathbf{r}, \mathbf{b} \rangle + c \ .$$

Здесь $\mathbf{b} \in \mathbf{R}^N$ $(\mathbf{C}^n)$, $c \in \mathbf{R}$, A – положительно определенная матрица. Поскольку A положительно определена, то квадратичная форма

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \mathbf{x}, A\mathbf{y} \rangle$$

удовлетворяет всем свойствам скалярного произведения. Введем ортонормированный базис $\{\mathbf{e}^i\}_{i=1}^N$ в линейном пространстве с нормой $\langle \cdot, \cdot \rangle_A$. Будем называть направления, задаваемые им, сопряженными. Если $\mathbf{r}^0$ некоторая точка, то произвольная точка $\mathbf{r}$ представляется в виде $\mathbf{r} = \mathbf{r}^0 + \sum\limits_1^N \alpha_i \mathbf{e}^i$. Тогда

$$\Phi(\mathbf{r}) = \left\langle \mathbf{r}^0 + \sum \alpha_i \mathbf{e}^i, A\mathbf{r}^0 + \sum \alpha_i A\mathbf{e}^i \right\rangle + \left\langle \mathbf{r}^0 + \sum \alpha_i \mathbf{e}^i, \mathbf{b} \right\rangle + \mathbf{c} =$$

$$= \Phi(\mathbf{r}^0) + \sum_{i=1}^N \left[\alpha_i^2 + 2\alpha_i \langle \mathbf{e}^i, \mathbf{r}^0 \rangle + \alpha_i \langle \mathbf{e}^i, \mathbf{b} \rangle \right] \ .$$

В этой сумме отсутствуют перекрестные члены, таким образом спуск вдоль любого направления $\mathbf{e}^i$ минимизирует лишь свой член суммы. Это означает, что осуществив спуск по каждому из сопряженных направлений лишь один раз, мы в точности достигаем минимума. В самой же точке минимума

$$\frac{\partial \Phi}{\partial \alpha_i} = 2\alpha_i + 2\langle \mathbf{e}^i, A\mathbf{r}^0 + \mathbf{b}/2 \rangle = 0 \ , \quad i = 1, 2, \ldots N \ ,$$

откуда $\alpha_i = -\langle \mathbf{e}^i, A\mathbf{r}^0 + \frac{\mathbf{b}}{2} \rangle$.

Построение базиса сопряженных направлений и спуск по ним

Пусть $\mathbf{r}^1$ произвольная точка и $\mathbf{e}^1$ произвольный единичный вектор. Рассмотрим прямую $\mathbf{r} = \mathbf{r}^1 + \alpha \mathbf{e}^1$ и найдем вдоль этой прямой минимум функции $\varphi(\alpha) = \Phi(\mathbf{r}^1 + \alpha \mathbf{e}^1)$. Будем считать, что указанный минимум достигается как раз в точке $\mathbf{r}^1$, то есть $0 = \alpha_1 = -\langle \mathbf{e}^1, A\mathbf{r}^1 + \mathbf{b}/2 \rangle$. Аналогично, пусть на прямой $\mathbf{r}^2 + \alpha \mathbf{e}^1$ минимум достигается в точке $\mathbf{r}^2$ $(0 = \alpha_2 = -\langle \mathbf{e}^1, A\mathbf{r}^2 + \mathbf{b}/2 \rangle)$. Таким образом

$$\alpha_2 - \alpha_1 = \langle \mathbf{e}^1, A(\mathbf{r}^2 - \mathbf{r}^1) \rangle = \langle \mathbf{e}^1, (\mathbf{r}^2 - \mathbf{r}^1) \rangle_A = 0 \ ,$$

и единичный вектор $\mathbf{e}^2 = \frac{\mathbf{r}^2 - \mathbf{r}^1}{\|\mathbf{r}^2 - \mathbf{r}^1\|_A}$ оказывается сопряженным к $\mathbf{e}^1$. Аналогично, пусть имеется m сопряженных векторов $\mathbf{e}^1, \mathbf{e}^2, \ldots, \mathbf{e}^m$ и в двух параллельных m-мерных сопряженных плоскостях $\mathbf{r}^1 + \sum \alpha_i \mathbf{e}^i$ и $\mathbf{r}^2 + \sum \alpha_i \mathbf{e}^i$ минимум достигается в точках $\mathbf{r}^1$ и $\mathbf{r}^2$ соответственно, тогда вектор $\mathbf{r}^2 - \mathbf{r}^1$ сопряжен всем векторам $\{\mathbf{e}^i\}_{i=1}^m$.

Опишем теперь последовательность действий по построению базиса сопряженных направлений. Пусть $\{\mathbf{d}^i\}_{i=1}^N$ стандартный базис и мы построили m сопряженных векторов, при этом выбросив из рассмотрения m векторов стандартного базиса, скажем, с номерами $N-m+1, N-m+2, \ldots, N$. Пусть $\mathbf{r}^0$ произвольная точка, произведем из нее спуск по сопряженным векторам $\{\mathbf{e}^i\}_{i=1}^m$. Попадем при этом в некоторую точку $\mathbf{r}^1$. Из нее произведем спуск по оставшимся векторам $\{\mathbf{d}\}_{i=1}^{N-m}$ стандартного базиса и попадем в точку $\mathbf{r}^2$. Теперь из $\mathbf{r}^2$ спустимся по сопряженным векторам $\{\mathbf{e}^i\}_{i=1}^m$. Попадем при этом в некоторую точку $\mathbf{r}^3$. Точки $\mathbf{r}^1$ и $\mathbf{r}^3$ являются минимумами функции $\Phi(\mathbf{r})$, в двух параллельных гиперплоскостях, задаваемых

сопряженными направлениями. Таким образом $\mathbf{r}^3 - \mathbf{r}^1$ — новое сопряженное направление. Добавим его к уже построенным и выбросим один из векторов $\{\mathbf{d}\}_{i=1}^{N-m}$. Формально все равно какой из них выбрасывать, однако для повышения точности желательно выбрасывать тот, при спуске вдоль которого функция $\Phi(\mathbf{r})$ изменилась меньше всего, даже если случайно он оказался одним из сопряженных. Дело здесь в том, что в этом случае мы не теряем точность в процессе ортогонализации. Заметим, что из точки $\mathbf{r}^3$ необходимо спуститься лишь вдоль нового направления $\mathbf{r}^3 - \mathbf{r}^1$, поскольку по другим сопряженным направлениям спуск уже произведен. Из полученной при этом точки $\mathbf{r}^4$ производится спуск по оставшимся $N - m - 1$ векторам стандартного базиса и т.д. Таким образом если бы не ошибки округления, то для квадратичной функции, произведя $N - 1$ циклов мы бы в точности попали в минимум. Однако именно из-за ошибок округления этого не произойдет и процедуру необходимо повторить некоторое количество раз.

Замечание. Хотя понятие сопряженных направлений было введено только для квадратичной функции, сам описанный процесс можно применять к произвольной функции $\Phi(\mathbf{r})$, поскольку сама процедура основана лишь на поиске минимума вдоль того или иного направления.

Литература

[1] *Н.Н. Калиткин* // Численные методы // Москва, Наука, 1978.

[2] *Н.С.Бахвалов, Н.П.Жидков, Г.М.Кобельков* // Численные методы // Москва — Санкт-Петербург, Лаборатория базовых знаний, 2000.

[3] *Д.Каханер, К.Моулер, С.Неш* // Численные методы и программное обеспечение // Москва, Мир, 1998.

[4] *Дж. Форсайт, М.Малькольм, К.Моулер* // Машинные методы математических вычислений // Москва, Мир, 1980.

[5] *С.Б. Стечкин, Ю.Н. Субботин* // Сплайны в вычислительной математике // Москва, Наука, 1976.

[6] *Дж.Бейкер, П.Грейвс-Моррис* // Аппроксимации Паде // Москва, Мир, 1986.

[7] *Д.Мак-Кракен, У.Дорн* // Численные методы и программирование на ФОРТРАНе // М., Мир, 1977.

[8] *В.В.Вершинин, Ю.С.Завьялов, Н.Н.Павлов* // Экстремальные свойства сплайнов и задача сглаживания // Новосибирск, Наука, 1988.

[9] *А.И.Гребенников* // Метод сплайнов и решение некорректных задач теории приближений // Издательство МГУ, 1983.

[10] *Э.Дулан, Дж.Миллер, У.Шилдерс* // Равномерные численные методы решения задач с пограничным слоем // М., Мир, 1983.

[11] *В.В.Воеводин, Ю.А.Кузнецов* // Матрицы и вычисления // М., Наука, 1984.

[12] *С.Писсанецки* // Технология разреженных матриц // М., Мир, 1988.

[13] *И.С.Березин, Н.П.Жидков* // Методы вычислений. Т.1. // М., Наука, 1966.

[14] *И.С.Березин, Н.П.Жидков* // Методы вычислений. Т.2. // М., Физматгиз, 1962.

[15] *А.Н.Колмогоров, С.И.Фомин* // Элементы теории функций и функционального анализа // М., Наука, 1972.

[16] *Д.К.Фаддеев*// Лекции по алгебре// М., Наука, 1984.

[17] *Г.Е.Шилов* // Математический анализ (функции одного переменного. Часть 3) // М., Наука, 1970.

[18] *Л.Д.Ландау, Е.М.Лифшиц* // Квантовая механика (нерелятивистская теория) // М., Наука, 1989.

[19] *А.Н.Тихонов, А.А.Самарский* // Уравнения математической физики // М., Наука, 1972.

[20] *Г.Корн, Т.Корн* // Справочник по математике // М., Наука, 1984.

Оглавление