Введение в теорию приближенных вычислений

v. 0.63

Н.Р. Беляев, И.В. Танатаров

20 октября 2011 г.

Оглавление

Введение

Настоящее пособие представляет собой немного расширенный конспект лекций семестрового курса по численному анализу для физиков. В курсе дается введение в математический аппарат численных методов, которые стоят за всевозможными алгоритмами и схемами, повседневно использующимися при численном счете. Излагаемые методы и алгоритмы важны в первую очередь для численного счета на компьютере, без которого в настоящее время может обойтись лишь очень (!) талантливый физик. Их понимание должно помочь эффективно работать с различными пакетами символьной алгебры[1]. Предполагается, что студенты прошли курсы математического анализа и линейной алгебры. Слушателям, имеющим навыки программирования, полезно будет попробовать реализовать некоторые из излагаемых методов самостоятельно.

Курс делится на две основные части, разного стиля изложения. В первую часть входит теория линейной интерполяции, численное дифференцирование, и весьма кратко методы решения дифференциальных уравнений, обыкновенных и в частных производных. Интерполяция важна в первую очередь как основа, на которой строится все дальнейшее изложение. В частности, конечно-разностные методы решения дифференциальных уравнений, которые в основном в этой части рассматриваются, базируются на формулах интерполяции. Вся первая часть использует элементарный математический аппарат, в рамках первого года математического анализа и элементов линейной алгебры.

Во вторую часть выделен материал, который в большей или меньшей степени опирается на понятие рядов Фурье. В третьей главе дается небольшое введение в функциональный анализ, вводится понятие гильбертова простран-

[1] Пакет символьной алгебры (computer algebra system, CAS) это программа, которая манипулирует математическими выражениями в символьном виде. Есть платные приложения, например широко используемые Mathematica, Maple, MatLab; бесплатные Maxima, Sage. Также представляет интерес ресурс wolframalpha, который работает как веб-интерфейс к ядру системы Mathematica, хотя его функциональность этим не ограничивается.

ства; дается краткое введение в общую теорию ортогональных полиномов. На ее основе далее излагается среднеквадратичное приближение и численное интегрирование, включая общую квадратурную формулу Гаусса. Завершает вторую часть глава по интегральным уравнениям и вариационным методам решения задач математической физики.

Настоящий курс был создан *Беляевым Николаем Романовичем*, и "откатывался" на факультете в 2000х годах. Исходный текст основан в основном на его конспекте, набранном неизвестными студентами. Работа по пересчету и исправлению опечаток плавно перетекла в переработку значительной части курса.

Вследствие естественного ограничения на объем охваченного материала, изложение иногда носит более справочный и инженерный характер, чем хотелось бы. Особенно это касается таких обширных областей, как численное решение уравнений математической физики, которому посвящены многие монографии. Подробное оглавление и градация материала по сложности должны помочь самостоятельно изучающим предмет. Так, некоторые дополнительные главы вынесены в приложения. Названия разделов и пунктов, которые немного сложнее основной части, или немного менее необходимы, помечены звездочкой (вся вторая часть – со звездочкой). Небольшие куски текста вспомогательного характера набраны мелким шрифтом, а примечания вынесены в сноски. Начало и конец доказательств обозначаются значками "◂" и "▸" соответственно. Значок "◂ . . . ▸" означает, что доказательство в этом месте опущено. Многие из опущенных выводов слушателям и читателям предлагается провести самостоятельно, что отмечается значком "✯". Список литературы приводится в конце каждой главы и общий в конце пособия.

Игорь Танатаров
08.2011

Часть I

Глава 1

Теория интерполяции

Interpolation theory

В основе любых численных методов лежит вопрос аппроксимации функций. К аппроксимации есть два различных подхода, которые можно назвать интерполяцией и (весьма условно) среднеквадратичным приближением. Каждый имеет свои преимущества и используется в большом количестве конкретных задач, таких как численное интегрирование или решение уравнений математической физики. Первая глава посвящена простейшему из двух подходов – теории интерполяции. В следующей главе мы используем этот аппарат для построения численных методов решения дифференциальных уравнений.

Пусть у нас есть некоторое семейство функций

$$\overline{\Phi}_n = \{\Phi(x; a_0, \ldots, a_n),\ a_i \in \mathbb{R}\}, \tag{1.1}$$

параметризуемое $n+1$ параметрами $\{a_i\}_{i=0}^n$. *Задача интерполяции (interpolation problem)* для $\overline{\Phi}_n$ состоит в том, чтобы найти такие значения параметров a_i (и, таким образом, уникальную функцию семейства Φ), чтобы для заданной $n+1$ пары чисел $\{(x_i, f_i)\}_{i=0}^n$, с попарно различными x_i, выполнялось

$$\Phi(x_i; a_0, \ldots, a_n) = f_i, \quad i = 0, \ldots, n. \tag{1.2}$$

Точки x_i будем называть *узлами* интерполяции, их совокупность *сеткой интерполяции*, а множество пар (x_i, f_i) – *сеткой данных*[1].

Таким образом, интерполяция – это, грубо говоря, такое приближение функции, значения которого в заданном наборе точек в точности совпадают со значениями в них исходной функции.

[1] Англоязычная терминология: the pairs (x_i, f_i) are *support points*, locations x_i are *support abscissas*, f_i *support ordinates*, the set $\{x_i\}$ is *the support*.

Если Φ зависит от параметров a_i линейно

$$\Phi(x; a_0, \ldots, a_n) = a_n\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x), \tag{1.3}$$

то семейство $\overline{\Phi}_n$ представляет собой соответствующее линейное пространство функций, и мы получаем задачу *линейной интерполяции*.

Этот класс задач включает в себя *полиномиальную интерполяцию*

$$\varphi_k(x) = x^k,$$

а также *тригонометрическую интерполяцию*

$$\varphi_k(x) = e^{ikx}.$$

В прошлом полиномиальная интерполяция часто использовалась для интерполяции значений функций, заданных таблично. В связи с развитием вычислительной техники, такая необходимость сейчас практически отпала. До сих пор, однако полиномиальная интерполяция важна как основа ряда широко применяемых методов численного интегрирования. В последнее время полиномиальная и рациональная интерполяция (см. ниже) также используются при построении "экстраполяционных" методов численного интегрирования, решения дифференциальных уравнений и родственных им задач.

Тригонометрическая интерполяция широко используется в численном Фурье-анализе. В этом контексте особенно важным является так называемое "быстрое преобразование Фурье".

В класс задач линейной интерполяции также входит *сплайн-интерполяция*. В частном случае *кубических сплайнов* в качестве функций φ_i берутся дважды непрерывно дифференцируемые на (x_0, x_n) функции, на каждом из промежутков (x_i, x_{i+1}) совпадающие с некоторым полиномом третьей степени.

Из нелинейных задач интерполяции следует отметить две: *рациональную интерполяцию*

$$\Phi(x; a_0, \ldots, a_n, b_0, \ldots, b_m) = \frac{a_0 + a_1x + \ldots + a_nx^n}{b_0 + b_1x + \ldots + b_mx^m},$$

и *экспоненциальную интерполяцию*

$$\Phi(x; a_0, \ldots, a_n, \lambda_0, \ldots, \lambda_n) = a_0e^{\lambda_0 x} + \ldots + a_ne^{\lambda_n x}.$$

1.1 Линейная интерполяция, системы Чебышёва

Пусть мы имеем дело с линейной интерполяцией (1.3)

$$\Phi(x; a_0, \ldots, a_n) = a_n\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x),$$

и ограничимся, для определенности, вещественными функциями.

Пусть всякая конечная совокупность $\varphi_i(x)$ линейно-независима[2]. Тогда семейство функций $\overline{\Phi}_n$ (1.1) представляет собой линейное пространство размерности $n+1$. Его элементы $\Phi(x; a_0, \ldots, a_n) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$ называют *обобщенными многочленами* по системе функций φ_i. В случае, если $\varphi_i = x^i$, обобщенные многочлены превращаются в обычные.

Интерполяционный многочлен $\varphi(x)$, который интерполирует функцию $f(x)$ в узлах $\{x_i\}_{i=0}^n$, по определению удовлетворяет условию

$$\varphi(x_j) \equiv \sum_{i=0}^{n} a_i\varphi_i(x_j) = f(x_j) \quad \text{для } j = 0, 1, \ldots, n, \tag{1.4}$$

Эта неоднородная система уравнений для коэффициентов a_i имеет решение тогда и только тогда, когда определитель системы, $\|\varphi_i(x_j)\|_{i,j=0}^n$, отличен от нуля.

Говорят, что линейная задача интерполяции поставлена корректно на некотором промежутке $[a, b]$, если обобщенный интерполяционный многочлен $\Phi(x) \in \overline{\Phi}_n$, удовлетворяющей условиям (1.4), существует и единственен для любых $\{x_i \in [a, b]\}_{i=0}^n$. Таким образом, она корректна тогда и только тогда, когда неоднородная система уравнений для a_i (1.4) разрешима при любых $\{x_i\}_{i=0}^n$. Это так, если соответствующая *однородная* система *не имеет* нетривиальных решений:

$$\forall\ x_0, x_1, \ldots, x_n \in [a, b] \quad \Delta = \begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} \neq 0. \tag{1.5}$$

Переформулируем: на $[a, b]$ нет таких $(n+1)$ точек, для которых бы сумма $\sum\limits_{i=0}^{n} a_i\varphi_i(x)$ обращалась в ноль при $\sum a_i^2 \neq 0$.

Переформулируем еще раз: всякая функция $\Phi(x) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$, где $\sum a_i^2 \neq 0$, имеет не более чем n нулей на $[a, b]$.

[2] В этом случае говорят, что $\{\varphi_i\}_{i=0}^{\infty}$ является бесконечной линейно-независимой системой.

Def.: Если всякая функция вида $\Phi(x) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$, где $\sum a_i^2 \neq 0$, имеет на $[a,b]$ не более чем n нулей, то система $(n+1)$ функции $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ называется *системой Чебышёва*[3] *(Chebyshev system)* на $[a,b]$.

Таким образом, мы пришли к теореме:

T^0: Чтобы для любых[4] $f(x) \in R_{[a,b]}$ и $x_0, x_1, x_2, \ldots, x_n \in [a,b]$ существовал обобщенный интерполяционный многочлен $\Phi(x) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$, необходимо и достаточно, чтобы система $\{\varphi_i(x)\}_{i=0}^{n}$ была системой Чебышёва. При этом интерполяционный многочлен единственен.

Коэффициенты a_i дает формула Крамера:

$$a_k = \frac{\Delta_k}{\Delta}, \quad \Rightarrow \quad \Phi(x) = \frac{1}{\Delta}\sum_k \Delta_k\varphi_k(x).$$

Раскладывая Δ_k по k-му столбцу, получим $\Delta_k = \sum_i f(x_i)\Delta_{ik}$, так что

$$\Phi(x) = \frac{1}{\Delta}\sum_{i,k} f(x_i)\Delta_{ik}\varphi_k(x) = \sum_i f(x_i)\Phi_i(x), \quad \text{где} \quad \Phi_i(x) = \sum_k \frac{\Delta_{ik}}{\Delta}\varphi_k(x) \tag{1.6}$$

– функции, не зависящие от f.

По построению $\Phi(x_j) = f(x_j)$, следовательно[5]

$$\Phi_i(x_j) = \delta_{ij}, \quad i,j = 0,1,\ldots,n. \tag{1.7}$$

Следует обратить внимание, что, хотя система Чебышёва на заданном промежутке всегда линейно-независима, обратное совсем не обязательно верно. Так, в соответствии с определением, пара функций $\{x, x^2\}$ образует систему Чебышёва на $[1,2]$, но не образует систему Чебышёва на $[-1,1]$. Возникает вопрос: каковы *достаточные* условия, чтобы система $\{\varphi_i(x)\}_0^n$ была системой Чебышёва на $[a,b]$? Ответ на него дает теорема:

T^0.: Если функции $\varphi_i(x)$ $(n+1)$ раз дифференцируемы на $[a,b]$ и для любого $k \leq n$ определитель Вронского[6] системы $\{\varphi_i\}_{i=0}^{k}$ не обращается в ноль на $[a,b]$, то система $\{\varphi_i(x)\}_{i=0}^{n}$ — система Чебышёва. ◂ . . . ▸ ☆

---

[3] Чебышёв Пафнутий Львович, 1821-1894. Считается одним из основоположников теории приближения функций, работал и преподавал в Москве и С.-Петербурге. Член Петербургской, Берлинской, Болонской, Лондонской академий наук. Встречаются также варианты транслитерации Chebychev, Chebyshov, Tchebycheff, Tschebyscheff. Обратная транслитерация и замена “ё” на “е” в русскоязычной литературе привела к потере буквы ё, так что фамилия часто произносится неправильно.

[4] $R_{[a,b]}$ – линейное пространство веществennозначных функций, определенных на $[a,b]$.

[5] Равенство также очевидно если посмотреть на формулу для $\Phi_i(x_j)$ и вспомнить что такое алгебраическое дополнение. Этим условием можно также воспользоваться, чтобы сразу строить интерполяционный многочлен по некоторой конкретной системе функций, вместо того, чтобы последовательно решать систему (1.4).

[6] Определитель Вронского $W[\varphi_1, \ldots, \varphi_k]$ системы функций $\{\varphi_i\}_{i=1}^{k}$определяется как определитель матрицы $(k+1)\times(k+1)$, элементы которой равны $W_{ij} = d^i\varphi_j/dx^i$, $i,j = 0,\ldots,k$.

1.2 Полиномиальная интерполяция

Исторически важным толчком к исследованию вопроса о приближении функций стала *аппроксимационная теорема Вейерштрасса*, сформулированная им в 1885:

$\mathbf{T}^0$: Пусть $f(x)$ – непрерывная функция на $[a,b]$, где $-\infty < a < b < \infty$. Тогда для всякого $\varepsilon > 0$ существует алгебраический многочлен $p(x)$, такой что

$$|f(x) - p(x)| < \varepsilon \quad \forall x \in [a,b].$$

Иначе говоря, *всякую непрерывную функцию на* $[a,b]$ *можно сколь угодно точно равномерно аппроксимировать многочленом*[7]. Поэтому логична постановка задачи полиномиальной интерполяции, в которой в качестве функций φ_i берутся степенные функции $\varphi_i = x^i$.

Будем обозначать через Π_n множество степенных функций степени не выше n, а через Π множество всех степенных функций конечной степени:

$$\Pi_n = \left\{1, x, x^2, \ldots, x^n\right\}, \quad \Pi \equiv \Pi_\infty = \left\{1, x, x^2, \ldots, x^n, \ldots\right\}. \tag{1.8}$$

Тогда множество $\overline{\Pi}_n$ всех многочленов степени не выше n есть линейная оболочка Π_n:

$$\overline{\Pi}_n = \operatorname{span}\{\Pi_n\}, \quad \overline{\Pi} \equiv \overline{\Pi}_\infty = \operatorname{span}\{\Pi\}. \tag{1.9}$$

Полагая $\varphi_i(x) = x^i$, $i = 0, \ldots, n$, получаем задачу полиномиальной интерполяции. Понятно, что в этом случае $\{\varphi_i(x)\}_0^n = \Pi_n$ – система Чебышёва, и поэтому задача является корректной. Ее явная формулировка:

Табличным способом задана функция $y(x)$*:*

$$x_i \in [a,b]; \quad y(x_i) = y_i, \quad \textit{для} \quad i = 0, \ldots, n;$$

Требуется найти $P_n(x) \in \overline{\Pi}_n$*, такой чтобы*

$$P_n(x_i) = y_i, \quad \textit{для} \quad i = 0, 1, \ldots, n.$$

Методы построения такого многочлена делятся на три основные группы:

1. Формула Лагранжа;
2. Итерационные методы;
3. Разностные методы.

[7] Подробнее см. п.3.1.9.

1.2.1 Интерполяционная формула Лагранжа

Построим интерполяционный многочлен функции f по системе $\{x^i\}_{i=0}^n$ в виде (1.6):

$$P_n(x) = \sum_{j=0}^{n} f(x_j) L_j(x),$$

где $L_j(x) \in \overline{\Pi}_n$ – многочлен степени не выше n, однозначно определяемый $(n+1)$ условиями (1.7). Многочлены $L_j(x)$ называют вспомогательными или *базисными многочленами Лагранжа*. Легко убедиться, что их можно представить в виде

$$L_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} = \frac{(x-x_0)(x-x_1)\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_n)}{(x_j-x_0)(x_j-x_1)\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_n)}. \quad (1.10)$$

Видно, что условие (1.7) выполнено, а значит единственный $L_j(x)$ – найден.

Формула

$$P_n(x) = \sum_{j} f(x_j) \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} \quad (1.11)$$

называется *интерполяционной формулой Лагранжа*[8] (interpolation polynomial in the Lagrange form), а P_n в этой форме – *интерполяционным многочленом Лагранжа*.

Можно еще ввести

$$\omega_n(x) \equiv \omega_n(x; x_0, \dots, x_n) = \prod_{i=0}^{n} (x - x_i). \quad (1.12)$$

Тогда P_n можно представить в виде

$$P_n(x) = \sum_{j} f(x_j) \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)}. \quad (1.13)$$

Пример построения интерполяционного многочлена

Контрольная функция $y = x^3$.

а) Пусть функция задана таблицей

i	0	1	2
x_i	1	2	3
y_y	1	8	27.

[8] Жозеф Луи Лагранж, Joseph-Louis Lagrange, 1736-1813. Родился в Турине, в Италии, работал в Берлине,потом в Париже. Один из основателей, среди прочего, вариационного исчисления и аналитической механики (см. уравнения Эйлера-Лагранжа).

Построим интерполяционный многочлен второй степени $P_2(x)$:

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{1}{2}(x^2-5x+6); \\ L_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = -(x^2-4x+3); \\ L_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{1}{2}(x^2-3x+2); \end{aligned}$$

тогда $P_2(x) = 1 \cdot \frac{1}{2}(x^2-5x+6) - 8(x^2-4x+3) + 27 \cdot \frac{1}{2}(x^2-3x+2) = 6x^2-11x+6$.

Посчитаем значения P_2 в нескольких контрольных точках:

x	1.9	2.3	2.8
$P_2(x)$	6.76	12.44	22.24
x^3	6.859	12.167	21.952

б) Добавим точку $(0,0)$ к исходной таблице, чтобы построить P_3:

i	0	1	2	3
x_i	0	1	2	3
y_i	0	1	8	27

Тогда

$$L_0(x) = -\frac{1}{6}(x-1)(x-2)(x-3); \qquad L_1(x) = \frac{1}{2}x(x-2)(x-3);$$
$$L_2(x) = -\frac{1}{2}x(x-1)(x-3); \qquad L_3(x) = \frac{1}{6}x(x-1)(x-2).$$

и $P_3(x) = \frac{1}{2}x(x-2)(x-3) - 4x(x-1)(x-3) + \frac{9}{2}x(x-1)(x-2) = x^3$.

То, что интерполяционный многочлен просто совпадает с контрольной функцией, неудивительно, т.к. контрольная функция есть многочлен 3й степени. Кстати, можно заметить, что L_0 можно было не вычислять.

Как можно было убедиться, недостаток такой схемы заключается в том, что при переходе к построению интерполяционного многочлена более высокой степени практически всю работу приходится проделывать с начала. Это имеет значение в том случае, если исходный массив данных очень велик, и для интерполяции не обязательно использовать все точки. Тогда используются итерационные методы, которые подбирают нужное число узлов и степень многочлена исходя из требования точности интерполяции.

Формула Лагранжа полезна в ситуации, когда, например, необходимо решить много задач интерполяции на одной и той же фиксированной системе узлов x_i, но для разных f_i.

1.2.2 Итерационные методы, алгоритм Невилла

Вместо того, чтобы решать задачу интерполяции всю сразу, можно сначала рассмотреть задачу для некоторых поднаборов узлов интерполяции. Потом из решений задач для разных поднаборов можно составить решение исходной задачи.

Пусть наша полная сетка данных $\{x_i, f_i\}_{i=0}^n$. Обозначим через

$$P_{i_0 i_1 \dots i_{k-1}}(x) \in \overline{\Pi}_{k-1}$$

многочлен, являющийся решением задачи интерполяции

$$P_{i_0 i_1 \dots i_{k-1}}(x_{i_j}) = f_{i_j} \quad \text{для} \quad j = 0, 1, \dots, k-1 < n,$$

которая образуют некоторый поднабор исходной сетки интерполяции Будем говорить для краткости, что он "интерполирует функцию f в узлах $\{x_{i_0}, \dots, x_{i_{k-1}}\}$", или через узлы $\{x_{i_0}, \dots, x_{i_{k-1}}\}$, или на сетке $\{x_{i_0}, \dots, x_{i_{k-1}}\}$.

При таком обозначении индексы соответствуют узлам, в которых $P_{\dots}$ интерполирует f; например P_{1673} по определению есть многочлен степени 3, который интерполирует f на подсетке $\{x_1, x_3, x_6, x_7\}$; порядок индексов, очевидно, значения не имеет.

Несложно увидеть, что из двух таких многочленов, каждый из которых интерполирует сетку в k узлах, из которых $k-1$ узел совпадают, можно сконструировать многочлен, который интерполирует сетку в $k+1$ узле:

$$P_{i_0 i_1 \dots i_k}(x) = \frac{(x - x_{i_0}) P_{i_1 \dots i_k}(x) - (x - x_{i_k}) P_{i_0 i_1 \dots i_{k-1}}(x)}{x_{i_k} - x_{i_0}}. \tag{1.14}$$

◂ Обозначим правую часть (1.14) через $R(x)$. Это многочлен степени k. Прямой подстановкой $x = x_{i_0}, \dots, x_{i_k}$ в (1.14), убеждаемся в том, что

$$\begin{aligned} R(x_{i_0}) &= P_{i_0 \dots i_{k-1}}(x_{i_0}) = f_{i_0}; \\ R(x_{i_k}) &= P_{i_1 \dots i_k}(x_{i_k}) = f_{i_k}; \\ R(x_{i_j}) &= \frac{(x_{i_j} - x_{i_0}) f_{i_j} - (x_{i_j} - x_{i_k}) f_{i_j}}{x_{i_k} - x_{i_0}} = f_{i_j} \text{ для } j = 1, \dots, k-1. \end{aligned} \tag{1.15}$$

Так как интерполяционный многочлен для узлов $x_{i_0}, \dots, x_k$ определяется однозначно, то он совпадает с $R(x)$, ч.и т.д.▸

Метод Невилла (Neville's algorithm) предназначен[9] для определения значения интерполяционного многочлена в заданной точке x. Строится он на основе рекурсии (1.14), в которой в качестве поднаборов узлов берутся наборы

[9] E. H. Neville, 1889-1961, английский математик

последовательных узлов: вначале строятся интерполяционные многочлены через последовательные пары узлов, из них – через последовательные тройки, потом через четверки, и так далее. Если пронумеровать все узлы в порядке возрастания, то схему можно нарисовать в таком виде

	$k=0$	1	2	3	
x_0	$P_{00}=f_0$				
		P_{01}			
x_1	$P_{11}=f_1$		P_{012}		
		P_{12}		P_{0123}	
x_2	$P_{22}=f_2$		P_{123}		$\ddots$
		P_{23}		$\ddots$	
x_3	$P_{33}=f_3$		$\ddots$		
$\vdots$	$\vdots$	$\ddots$			

Здесь в k-ом столбце получаются значения многочленов степени k, которые интерполируют сетку в последовательных $k+1$ узлах, из предыдущего столбца по рекурсии (1.14), которая в данном случае выглядит как

$$P_{m\ldots n}(x)=\frac{1}{x_n-x_m}\begin{vmatrix} x-x_m & P_{m\ldots n-1}(x) \\ x-x_n & P_{m+1\ldots n}(x) \end{vmatrix}. \tag{1.16}$$

Здесь подразумевается, что $m<n$; если $n-m=1$, то P_{jj} в правой части следует понимать как многочлен степени 0, интерполирующий сетку в узле j, то есть число $P_{jj}=f_j$.

Таким образом, строится последовательность значений в заданной точке x многочленов, которые интерполируют f на все возрастающей сетке узлов. Последнее число – значение интерполяционного многочлена на полной сетке. На практике, однако, вычисление может быть остановлено на любом промежуточном этапе, по достижении требуемой точности. Это главное преимущество в использовании итерационных методов перед непосредственным счетом формулы Лагранжа в случае большого числа узлов.

Пример

Посчитаем значение $y(x)$ в точке $x=2.3$ для контрольной функции $y=x^3$, заданной таблично в точках $x=1,2,3,4,5$. Табличку можно нарисовать следующим

образом

i	x_i	$x - x_i$	$k=0$	1	2	3	4
1	1	1.3	1				
				10.1			
2	$\boxed{2}$	**0.3**	8		12.44		
				13.7		12.167	
3	3	−0.7	27		**11.81**		12.167
				1.1		$\underline{12.167}$	
4	4	−1.7	64		**15.38**		
				39.7			
5	$\boxed{5}$	**−2.7**	125				

Очередной шаг вычислений (числа в большой табличке выделены так же) выглядит как[10]

$$\frac{1}{\boxed{5}-\boxed{2}}\begin{vmatrix} \mathbf{0.3} & \mathbf{11.81} \\ \mathbf{-2.7} & \mathbf{15.38} \end{vmatrix} = \underline{12.167}.$$

Получившееся число 12.167 есть значение интерполяционного полинома через узлы x_2, x_3, x_4, x_5 в точке $x=2.3$. Таким образом, наглядно видно как сходится, если сходится, ряд значений в x при увеличении количества точек. На третьем шаге получаем значение 12.167 интерполяционного полинома через четыре точки, т.е. полинома третьей степени, который совпадает с исходной функцией. Поэтому это значение совпадает для разных наборов точек и совпадает с точным значением 2.3^3.

1.2.2.1 Метод Эйткена*

Метод Эйткена[11] *(Aitken's algorithm)* отличается от метода Невилла только тем, в каком порядке выбираются подпоследовательности узлов. На первом шаге строится интерполяция через пары узлов $01, 02, 03, \ldots, 0n$, на втором через тройки $012, 013, \ldots, 01n$, и так далее. Рекуррентная формула (1.16) принимает вид

$$P_{0\ldots n}(x) = \frac{1}{x_{n-1} - x_n}\begin{vmatrix} x - x_{n-1} & P_{0\ldots n-2,n-1}(x) \\ x - x_n & P_{0\ldots n-2,n}(x) \end{vmatrix}.$$

[10] В знаменателе стоит разность $(x_i - x_j)$, а не их индексов, хотя в данном примере они и совпадают!

[11] Alexander Craig Aitken, 1895–1967, новозеландский математик, писатель, музыкант и счетчик.

Пример: посчитаем значение $y(x)$ в точке $x = 2.3$ для той же контрольной функции $y = x^3$, заданной таблично

i	x_i	$x - x_i$	y_i	P_{0i}	P_{01i}	P_{012i}
0	1	1.3	1			
1	*2*	**0.3**	8	**10.1**		
2	3	−0.7	27	17.9	12.44	
3	4	−1.7	64	28.3	12.83	12.167
4	*5*	**−2.7**	125	**41.3**	13.22	12.167

Очередной шаг вычислений выглядит так:

$$\frac{1}{5-2}\left|\begin{array}{cc} \mathbf{0.3} & \mathbf{10.1} \\ \mathbf{-2.7} & \mathbf{41.3} \end{array}\right| = 13.22$$

Метод сейчас практически не используется, в отличие от метода Невилла и его модификаций, и важен разве что с исторической точки зрения, как классический пример итерационных методов интерполяции.

1.2.3 Разделенные разности, формула Ньютона

Если нам нужны значения интерполяционного многочлена не в одной точке, а в нескольких, то варианты алгоритма Невилла становятся неэффективны, т.к. для каждой точки надо проводить вычисление заново. В этом случае удобнее использовать интерполяционную формулу Ньютона, которая сразу дает явное выражение для самого́ интерполяционного многочлена.

Заметим, что два многочлена $P_{i_0 i_1 \ldots i_k}(x)$ и $P_{i_0 i_1 \ldots i_{k-1}}(x)$ (в развернутых обозначениях предыдущего параграфа) отличаются на многочлен степени k, который обращается в ноль в узлах $x_{i_0}, x_{i_1}, \ldots, x_{i_{k-1}}$, так как оба они по построению интерполируют соответствующие точки. Значит есть такое число $f_{i_0 i_1 \ldots i_k}$, что

$$P_{i_0 i_1 \ldots i_k}(x) = P_{i_0 i_1 \ldots i_{k-1}}(x) + f_{i_0 i_1 \ldots i_k}(x - x_{i_0})(x - x_{i_1}) \ldots (x - x_{i_{k-1}}).$$

Видно, что $f_{i_0 i_1 \ldots i_k}$ есть старший коэффициент многочлена $P_{i_0 i_1 \ldots i_k}(x)$, потому логично и обозначить эту величину теми же индексами.

Переписывая так же $P_{i_0 i_1 \ldots i_{k-1}}$ и далее по рекурсии, и учитывая что $P_{i_0} \equiv f_{i_0}$, получаем ньютоновское представление интерполяционного многочлена $P_{i_0 i_1 \ldots i_k}(x)$,

или *интерполяционную формулу Ньютона*[12] (или Ньютона-Грегори)

$$P_{i_0 i_1 \dots i_k}(x) = \underbrace{f_{i_0} + f_{i_0 i_1}(x - x_{i_0}) + \dots}_{P_{i_0 i_1 \dots i_{k-1}}(x)} + f_{i_0 i_1 \dots i_k}(x - x_{i_0})(x - x_{i_1}) \dots (x - x_{i_{k-1}}). \quad (1.17)$$

Коэффициенты $f_{i_0 i_1 \dots i_k}$ называются *разделенными разностями (divided differences)*.

Почему, видно вот откуда. Как было отмечено, разделенная разность $f_{i_0 i_1 \dots i_k}$ является старшим коэффициентом многочлена $P_{i_0 i_1 \dots i_k}(x)$. Поэтому, приравнивая в итерационной формуле (1.14) коэффициенты при старших степенях x, получим рекурсию

$$f_{i_0 i_1 \dots i_k} = \frac{f_{i_1 \dots i_k} - f_{i_0 \dots i_{k-1}}}{x_{i_k} - x_{i_0}}, \quad (1.18)$$

из вида которой ясно, откуда взялось название[13].

Пронумеруем узлы так, чтобы $i_j = j$ для $j = 0, \dots, n$. Тогда, используя соотношение (1.18), можно считать разделенные разности алгоритмом типа Невилла, по схеме

	$k=0$	1	2	…
x_0	$\mathbf{f_0}$			
		$\mathbf{f_{01}}$		
x_1	f_1		$\mathbf{f_{012}}$	
		f_{12}	⋮	⋱
x_2	f_2	⋮		
⋮	⋮			

(1.19)

Коэффициенты, дающие решение задачи интерполяции, оказываются на верхней диагонали (они выделены полужирным шрифтом). При этом имеет смысл строить табличку не слева направо, а по диагоналям $f_1 - f_{01}$, $f_2 - f_{12} - f_{012}$, и так далее. Таким образом, при достройке k-той по счету диагонали мы будем получать многочлен, интерполирующий первые k точек сетки $\{x_i, f_i\}_{i=0}^n$:

$$P_{01\dots k}(x) = f_0 + f_{01}(x - x_0) + \dots + f_{01\dots k}(x - x_0)(x - x_1) \dots (x - x_{k-1}). \quad (1.20)$$

Если при этом узлы $x_0, \dots, x_n$ пронумерованы в порядке возрастания, то получаем формулу Ньютона для *интерполяции вперед* – каждая итерация добавляет к набору узлов, в которых интерполяционный многочлен интерполирует

[12] По имени Исаака Ньютона, Isaac Newton, 1643-1727, который развивал и численные методы, среди прочего, для решения разных прикладных задач.

[13] Так как интерполяционный многочлен однозначно определяется сеткой данных, вне зависимости от порядка точек, то разделенные разности инвариантны относительно перестановки индексов.

исходную функцию, еще один узел справа. Аналогично, если узлы пронумеровать в порядке убывания, то получим формулу Ньютона для *интерполяции назад*. Основываясь на общем рекуррентном соотношении (1.18), можно вывести и другие ее вариации, придуманные Гауссом, Стирлингом и Бесселем. В частности, полезной может оказаться формула Ньютона для случая, когда узлы добавляются поочередно с правого и левого концов промежутка.

1.2.3.1 Эквидистантные узлы

Счет немного упрощается в задаче интерполяции на эквидистанстных (равноотстоящих) узлах. Пусть шаг сетки h: $x_i = x_0 + ih$. Пронумеровав узлы в порядке возрастания, получим $x_n - x_m = h(n-m)$. Тогда при вычислении k-той колонки (1.19), в соответствии с рекуррентной формулой (1.18), будем иметь

$$f_{m\dots m+k} = \frac{f_{m+1\dots m+k} - f_{m\dots m+k-1}}{x_{m+k} - x_m} = \frac{f_{m+1\dots m+k} - f_{m\dots m+k-1}}{kh}.$$

Значит элементы k-той колонки можно представить в виде

$$f_{m\dots m+k} = \frac{\Delta_{m\dots m+k}}{k!h^k},$$

где Δ – новые разности, которые вычисляются по той же схеме (1.19), но без деления на $(x_n - x_m)$:

$$\Delta_{m\dots m+k} = \Delta_{m+1\dots m+k} - \Delta_{m\dots m+k-1}, \quad \Delta_j = f_j.$$

Они являются функциями только f_i и считаются по аналогичной схеме, которую здесь повернем против часовой стрелки на $\pi/4$:

i	$k=0$	$k=1$	$k=2$	$k=3$	$\dots$
0	$\mathbf{f_0}$	$\mathbf{\Delta_{01}}$	$\mathbf{\Delta_{012}}$	$\mathbf{\Delta_{0123}}$	
1	f_1	Δ_{12}	Δ_{123}		$\ddots$
2	f_2	Δ_{23}		$\ddots$	
3	f_3		$\ddots$		
$\vdots$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	
	f_i	$\hat{\Delta} f_i$	$\hat{\Delta}^2 f_i$	$\hat{\Delta}^3 f_i$	$\dots$

(1.21)

Как видно из процесса построения, разности $\Delta_{m\dots m+k}$ можно представить как k-кратное действие на f_m *разностного оператора* (difference operator) $\hat{\Delta}$. Этот оператор, действуя на элемент произвольной последовательности $\{A_i\}_{i=0}^N$, дает приращение на следующем шаге:

$$\hat{\Delta} A_i = A_{i+1} - A_i. \tag{1.22}$$

Интерполяционная формула Ньютона через разделенные разности тогда принимает вид

$$P_{01\ldots k}(x) = f_0 + \frac{\hat{\Delta} f_0}{h}(x - x_0) + \frac{\hat{\Delta}^2 f_0}{2h^2}(x - x_0)(x - x_1) + \\ + \ldots + \frac{\hat{\Delta}^k f_0}{k!\, h^k}(x - x_0)(x - x_1)\ldots(x - x_{k-1}). \tag{1.23}$$

Несложно показать, что в пределе $h \to 0$, когда все узлы совпадают с x_0, коэффициенты $\hat{\Delta}^k f_0 / h^k$ переходят в производные

$$\frac{\hat{\Delta}^k f_0}{h^k} \to f^{(k)}(x_0)$$

и интерполяционная формула Ньютона (1.20), (1.23) переходит в ряд Тейлора☆. Таким образом, она является естественным эквивалентом ряда Тейлора в *конечных разностях*[14]. В дальнейшем шляпку над Δ мы будем опускать.

Пример

Контрольная функция (опять) $y = x^3$, на эквидистантных узлах $1, 2, 3, 4$. Сначала считаем разности Δ по принятой схеме (один из шагов показан, $\underline{19} = \mathbf{27} - \mathbf{8}$)

i	x_i	$\Delta^0 y_i = y_i$	$\Delta^1 y_i$	$\Delta^2 y_i$	$\Delta^3 y_i$
1	2	**8**	$\underline{19}$	18	6
2	3	**27**	37	24	
3	4	64	61		
4	5	125			

(1.24)

То, что $\Delta^3 y_i$ – одинаковые константы, – не случайность. Т.к. контрольная функция $y = x^3$, то интерполяционный многочлен 3й степени с ней совпадает.

С помощью построенных разностей вычислим интерполяционный многочлен и 2.3^3 в последовательных приближениях.

$P_1 = \Delta^0 y_1 = y_1 = 8$;

$P_{12} = 8 + \frac{\Delta^1 y_1}{h}(x - x_1) = 8 + 19(x - 2) = 19x - 30$;

$P_{123} = 19x - 30 + \frac{\Delta^2 y_1}{2h^2}(x - x_1)(x - x_2) = 19x - 30 + \frac{18}{2}(x - 2)(x - 3) = 9x^2 - 26x + 24$;

$P_{1234} = 9x^2 - 26x + 24 + \frac{6}{6}(x - 2)(x - 3)(x - 4) = x^3$.

Подставляя $x = 2.3$, получаем последовательные приближения:

$P_1(2.3) = 8$; $P_{12}(2.3) = 13.7$, $P_{123}(2.3) = 11.81$, $P_{1234}(2.3) = 12.167$.

[14] Также отсюда получаем один из способов численного счета производной произвольного порядка.

1.2.4 Погрешность полиномиальной интерполяции

1.2.4.1 Погрешность интерполяционной формулы Лагранжа

Пусть функция $f(x)$ — $(n+1)$ раз непрерывно дифференцируема.

Рассмотрим

$$\varphi(x) = f(x) - P_n(x) - K \cdot \omega_n(x), \quad \text{где} \quad \omega_n(x) = (x - x_0) \dots (x - x_n).$$

Тогда $\varphi(x_0) = \varphi(x_1) = \dots = \varphi(x_n) = 0$.

Зафиксируем постоянную K из требования чтобы $\varphi(x)$ обращалась в ноль в еще одной точке $\tilde{x} \neq x_0, \dots, x_n$:

$$K = \frac{f(\tilde{x}) - P_n(\tilde{x})}{(\tilde{x} - x_0) \dots (\tilde{x} - x_n)}.$$

Тогда:

$\varphi(x)$ имеет $(n+2)$ корня на $[a, b]$;

$\varphi'(x)$ имеет $(n+1)$ корня на $[a, b]$;

$\dots \quad \dots \quad \dots$

$\varphi^{(n+1)}(x)$ имеет один корень на $[a, b]$: $\exists \xi \in [a, b] \,|\, \varphi^{(n+1)}(\xi) = 0$.

Но $\varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K(n+1)!$, значит $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$ и

$$f(\tilde{x}) - P_n(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (\tilde{x} - x_0) \dots (\tilde{x} - x_n).$$

Тогда, т.к. $\tilde{x}$ можно выбрать произвольно на $[a, b]$, опускаем тильду и для $\forall x \in [a, b]$ получаем

$$\begin{aligned} |f(x) - P_n(x)| \quad &\leq \quad \frac{\sup\limits_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right|}{(n+1)!} |(x - x_0) \dots (x - x_n)| = \\ &= \frac{\sup\limits_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right|}{(n+1)!} |\omega_n(x)| . \end{aligned} \tag{1.25}$$

Это и есть оценка погрешности формулы Лагранжа. Выбором $x_0, \dots, x_n$ можно попытаться уменьшить $|\omega_n(x)|$ и тем самым повысить точность интерполяции.

1.2.4.2 Распределение погрешности по отрезку

Вопрос: Если узлы интерполяции зафиксированы, то для каких промежутков изменения аргумента остаточный член будет меньше?

Ограничимся случаем равноотстоящих узлов:

$$x_{k+1} - x_k = h \quad \text{для} \quad k = 0, \ldots, n-1. \tag{1.26}$$

Тогда $x_k = x_0 + kh$ и
$\omega_n(x) = (x-x_0)(x-x_0-h)\ldots(x-x_0-nh) = h^{n+1}t(t-1)\ldots(t-n)$, где $t = (x-x_0)/h$.

Введя

$$\psi_n(t) = t(t-1)\ldots(t-n),$$

получим $\omega_n(x) = h^{n+1}\psi_n(t)$. Когда $x \in [x_0, x_n]$, $t \in [0, n]$. Свойства $\psi_n(t)$:

1. при $n = 2k$ функция $\psi_n(t)$ нечетна относительно $t = n/2$;
   при $n = 2k+1$ функция $\psi_n(t)$ четна относительно $t = n/2$;

2. $\psi_n(t+1) = \frac{t+1}{t-n}\psi_n(t)$. Значит, если разбить отрезок $[0, n]$ на участки $[0,1], [1,2], \ldots, [n-1, n]$, то значение функции на каждом отрезке будет получаться из значения на предыдущем отрезке умножением на $\frac{t+1}{t-n}$.

   Так как при $t \in [0, \frac{n-1}{2}]$ верно $|\frac{t+1}{t-n}| < 1$, то экстремальные значения к середине отрезка будут убывать.

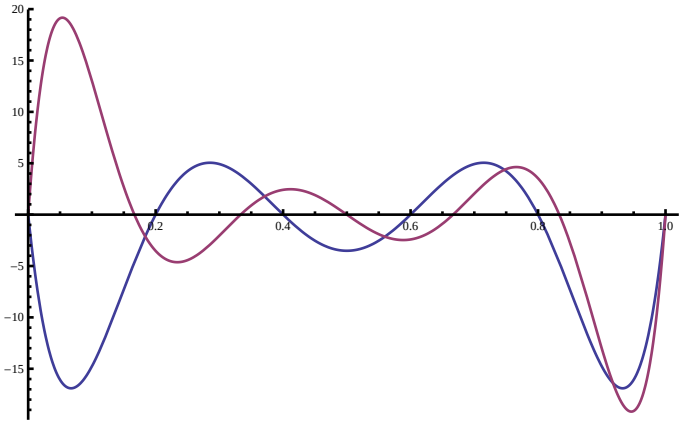


Рис. 1.1: $\psi_6(t)$ (синяя, симметричная) и $\psi_7(t)/5$ (красная, асимметричная), масштабированные на промежуток $[0, 1]$.

Вывод: Оценка остаточного члена формулы Лагранжа будет особенно велика при $x \notin [x_0, x_n]$, т.е. при экстраполировании погрешности будут велики. При интерполировании точность выше к середине промежутка, см. рис. 1.1.

1.2.4.3 Феномен Рунге

Мы видели (1.25), что погрешность приближения многочленами порядка n пропорциональна $(n+1)$-ой производной интерполируемой функции на интервале, причем при равноотстоящих (эквидистантных) узлах ошибка растет на концах интервала, что выражается в характерных осцилляциях. При этом погрешность вне промежутка быстро растет, так что для экстраполяции полученные формулы на эквидистантных узлах вообще мало пригодны. Можно однако подумать, что внутри интервала, при достаточно больших n, "для большинства разумных функций" ошибки становятся маленькими. Покажем, что это на самом деле не всегда так.

Для примера, рассмотрим бесконечно гладкую функцию

$$f(x) = \frac{1}{1+25x^2}.$$

Ее производные в $x=1$ легко считаются[15]

n	1	2	3	4	5
$\lvert f^{(n)}(1)\rvert$	0.07	0.21	0.79	3.6	19.8

Как видно, значения производных быстро растут с n. Поэтому при повышении степени интерполяционного многочлена на $[-1,1]$ (и, соответственно, числа точек сетки данных) его погрешность тоже растет, и можно показать, что для равноотстоящих узлов

$$\lim_{n\to\infty}\left(\max_{x\in[-1,1]}|f(x)-P_n(x)|\right)=\infty.$$

Этот эффект называют феноменом Рунге[16] (Runge's phenomenon) интерполяции многочленами высокой степени (см. рис. 1.2).

Еще об ошибке при приближении многочленами*

Обсудим эффект более подробно. Рассмотрим аналитическую функцию, т.е. функцию, ряд Тейлора для которой

$$y(x)=\sum_{k=0}^{n}\frac{(x-x_0)^k}{k!}y^{(k)}(x_0)+\frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi),\quad \text{где}\quad \xi\in[x_0,x],$$

[15] Например, для счета простых выражений, которые можно записать в строчку в синтаксисе программы "Математика" (хелп тут) можно воспользоваться бесплатным ресурсом wolframalpha.com. Так, пятую производную дает строка "N[D[(1+25 x^2)^(-1),{x,5}]/.x->1]"

[16] Карл Рунге, Carl David Tolmé Runge, 1856-1927. Немецкий математик, физик и спектрограф. Один из создателей метода Рунге-Кутта численного решения дифференциальных уравнений.

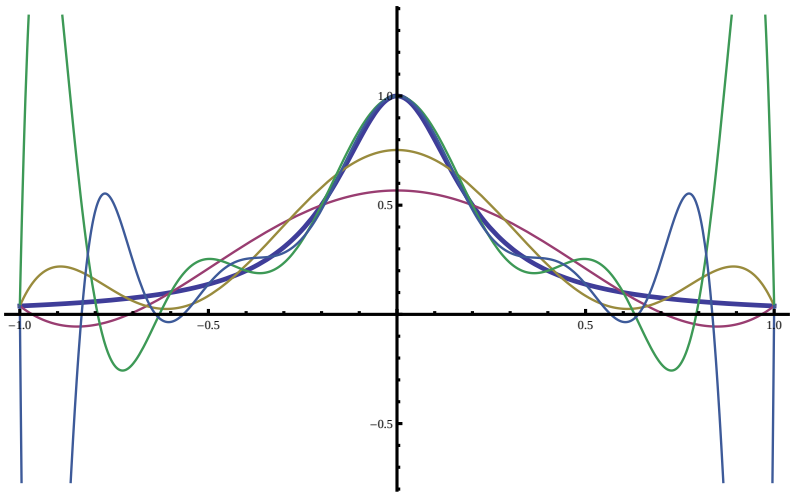


Рис. 1.2: $f(x)$ и многочлены степеней $n = 5, 7, 10, 12$, которые интерполируют ее на промежутке $[-1, 1]$, на равноотстоящих узлах. Видно, что при повышении порядка появляются характерные резкие осцилляции вблизи краев промежутка.

сходится во всех точках интересующей нас области. Если она еще и целая, т.е. ряд Тейлора сходится всюду в конечной части комплексной плоскости, как $\sin x$, e^x, $P_n(x)$, то действительно, все производные достаточно высокого порядка малы. Но если функция имеет особенность в конечной части комплексной плоскости, как почти всякая случайно взятая функция, то ряд Тейлора имеет конечный радиус сходимости R. А это означает, что верхняя грань n-ой производной в этой области растет как $n!$.

В качестве примера можно рассмотреть функцию $y = \ln x$. Для нее

$$y' = 1/x, \ \ldots, \ y^{(n)} = (-1)^{n-1}(n-1)!x^{-n}.$$

Таким образом, даже если кривая $y = \ln x$ выглядит гладкой вблизи некоторых значений x, тем не менее, когда n становится достаточно большим, производные в этой точке делаются очень большими по величине и ведут себя как $n!$.

Это общий случай: для "большинства функций" некоторые из производных высокого порядка имеют тенденцию расти как $n!$. Это относится и к многочленам, производные которых растут до n-й производной как $a_0 n!$, после чего только обращаются в ноль. Ограниченными производными обладают лишь некоторые целые функции (хотя не все!). Но если функция является целой, то можно ожидать, что задачу можно решить аналитически, и необходимость в интерполяции сомнительна.

Таким образом, интерполяция Лагранжа на равноотстоящих узлах не дает, в общем случае, последовательности полиномов Вейерштрасса, которая бы равномерно сходилась к f.

Погрешности (осцилляции но концах промежутка) могут быть уменьшены подбором узлов интерполяции, не эквидистантных, а сгущающихся к концу промежутка. Классическим примером являются нули полиномов Чебышёва, которые асимптотически сгущаются к концу интервала $[-1,1]$ как $(1-x^2)^{-1/2}$. О них пойдет речь в следующих двух параграфах.

1.2.4.4 Многочлены Чебышёва

Раскрывая скобки в формуле (Эйлера-)Муавра[17] (De Moivre's formula)

$$\cos n\varphi + i\sin n\varphi = e^{in\varphi} = (\cos\varphi + n\sin\varphi)^n,$$

видим, что $\cos n\varphi$ и $\sin n\varphi$ – вещественные многочлены с целыми коэффициентами от $\cos\varphi$ и $\sin\varphi$. При этом в вещественную часть входят только четные степени $\sin\varphi$, поэтому $\cos n\varphi$ является многочленом одного лишь аргумента – $\cos\varphi$ (так как $\sin^2\varphi = 1-\cos^2\theta$), и степени не выше n. Эти многочлены называются *многочленами Чебышёва*[18] *(Chebyshev polynomials)*:

$$T_n(x) = \cos(n\arccos x) \quad x\in[-1,1]. \tag{1.27}$$

Первые два полинома очевидны

$$T_0 = 1, \quad T_1 = x;$$

остальные удобно вычислять с помощью рекуррентного соотношения:

$$\begin{aligned}&\cos(n+1)\theta + \cos(n-1)\theta = \Re(e^{i(n+1)\theta} + e^{i(n-1)\theta}) = \Re(e^{in\theta}\cdot 2\cos\theta) = 2\cos\theta\cos n\theta,\\ &\theta = \arccos x \quad \Rightarrow \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).\end{aligned} \tag{1.28}$$

Из него видно, что $T_n(x)$ это многочлен x степени n, со старшим коэффициентом 2^{n-1}. Таким образом, система $\{T_i(x)\}_0^n$ является системой Чебышёва.

[17] Абрахам де Муавр, Abraham de Moivre, 1667-1754. Французский математик, занимался аналитической геометрией и теорией вероятностей. Будучи гугенотом, после эдикта Фонтенбло 1685 года эмигрировал в Лондон. Дружил с Исааком Ньютоном.

[18] Или многочленами Чебышёва первого рода; есть еще второго рода, см. п.3.2.4.4.

Несколько первых полиномов для справки:

$$\begin{aligned}
T_0(x) &= 1; & T_5(x) &= 16x^5 - 20x^3 + 5x;\\
T_1(x) &= x; & T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1;\\
T_2(x) &= 2x^2 - 1; & T_7(x) &= 64x^7 - 112x^5 + 56x^3 - 7x;\\
T_3(x) &= 4x^3 - 3x; & T_8(x) &= 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1\\
T_4(x) &= 8x^4 - 8x^2 + 1; & T_9(x) &= 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x.
\end{aligned}$$

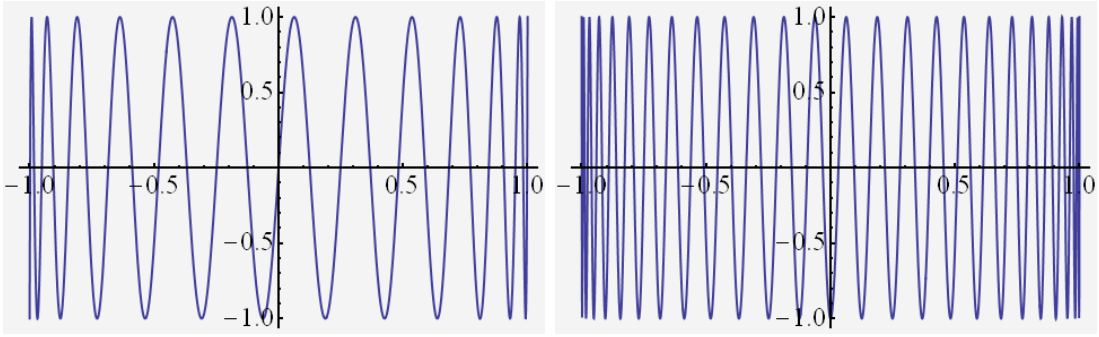


Рис. 1.3: Полиномы Чебышёва $T_{25}(x)$ и $T_{50}(x)$

Нули $T_n(x)$ находим из $\cos(n \arccos x) = 0 \;\Rightarrow\; n \arccos x = \frac{(2k-1)\pi}{2} \;\Rightarrow$

$$x_k = \cos \frac{(2k-1)\pi}{2n},$$

и значит на промежутке $[-1, 1]$ многочлен $T_n(x)$ имеет ровно n корней (т.е. иначе говоря, все его корни лежат на $[-1, 1]$): меняя k от n до 1, получаем все эти n нулей на $[-1, 1]$ в порядке возрастания.

Плотность нулей растет к концам промежутка, и легко видеть, что асимптотически, при $n \to \infty$, она ведет себя как $(1 - x^2)^{-1/2}$:

$$\rho = \frac{\Delta k}{\Delta x} \sim \left(\frac{dx}{dk}\right)^{-1} = \left(\frac{\pi}{n} \cdot \sin \frac{(2k+1)\pi}{2n}\right)^{-1} = \frac{n/\pi}{\sqrt{1 - x^2}}.$$

При этом $\max |T_n(x)| = 1$ и этот максимум достигается в $n + 1$ точке на промежутке $[-1, 1]$, включая концы промежутка:

$$x_{extr}^{(n,k)} = \cos \frac{\pi k}{n}; \quad \text{для} \quad k = 0, \ldots, n.$$

1.2.4.5 Выбор узлов интерполяции

Если в качестве отрезка интерполяции взять $[-1,1]$, а узлы интерполяции $\{x_i\}_0^n$ взять в нулях многочлена Чебышёва степени $(n+1)$, то

$$\omega_n(x) = \prod_{i=0}^{n}(x - x_i) = \frac{1}{2^n}T_{n+1}(x) \;\Rightarrow\; \sup|\omega_n(x)| = \frac{1}{2^n}.$$

Тогда распределение погрешности по отрезку будет в точности соответствовать $T_{n+1}(x)$ (см. рис. 1.3), а для оценки модуля погрешности (1.25) получаем

$$|f(x) - P_n(x)| \le \frac{\sup|f^{(n+1)}(x)|}{(n+1)!\,2^n}.$$

При этом оказывается, что выбором $x_0, x_1, \ldots, x_n$ величину $\sup|\omega_n(x)|$ больше уже нельзя уменьшить.

◂ От противного. Предположим, что $\Omega_n(x)$ – приведенный многочлен степени $n+1$, все нули которого лежат на $[-1,1]$, и максимальное абсолютное значение которого на этом интервале меньше чем 2^{-n}. Тогда

$$f_n = \omega_n(x) - \Omega_n(x) \in \overline{\Pi}_n,$$

т.к. старшие слагаемые ω_n и Ω_n сокращаются. Из предположения следует, что в точках экстремума ω_n (а это экстремумы T_{n+1}) верно

$$\left|\Omega_n(x_{extr}^{(n+1,k)})\right| < \left|\omega_n(x_{extr}^{(n+1,k)})\right| = 2^{-n},$$

следовательно $f_n > 0$ в точках максимума ω_n, и $f_n < 0$ в точках минимума ω_n. Всего точек экстремума T_{n+1} на $[-1,1]$ – $n+2$, включая концы промежутка. Тогда f_n имеет $n+1$ перемен знака, и так как это многочлен степени n, он обращается тождественно в ноль. ▸

Поэтому многочлены Чебышева называют *многочленами, наименее уклоняющимися от нуля*, и интерполяция с узлами в нулях T_{n+1} имеет наименьшую возможную ошибку на промежутке.

Для произвольной непрерывной, и даже бесконечно гладкой на промежутке, функции f, однако, это вовсе не означает, что при $n \to \infty$ максимальная ошибка стремится к нулю: даже при оптимальном выборе узлов равномерной сходимости к f нет.

Для отрезка $[a,b]$ замена $z=(2x-a-b)/(b-a)$ переводит его в отрезок $[-1,1]$ и тогда имеем наилучшую оценку в целом по отрезку $[a,b]$ в виде:

$$|f(x) - P_n(x)| \le \frac{\sup|f^{(n+1)}(x)|}{(n+1)!}\,\frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{1.29}$$

1.2.5 Численное дифференцирование

Пусть функция $y(x)$ задана таблицей своих значений в точках $x_0, x_1, x_2, \ldots, x_n$:

$$y(x_i) = y_i \quad \text{для} \quad i = 0, 1, \ldots, n.$$

Требуется: найти производную $y'(x)$ в точках $x_0, x_1, x_2, \ldots, x_n$.

Пример вывода формул

Покажем на простом примере, как можно получить формулы численного дифференцирования исходя из формул интерполяции.

Пусть интерполяционный многочлен для $y(x)$ построен, и пусть это полином второй степени, интерполирующий функцию в точках x_0, x_1, x_2. Возьмем его выражение в том виде, который получается в методе Ньютона при интерполяции вперед и равноотстоящих узлах:

$$P(x) = P_2(x) = C_0 + C_1(x - x_0) + C_2(x - x_0)(x - x_1), \text{ где}$$
$$C_0 = y_0; \quad C_1 = \tfrac{1}{h}\Delta_{01} = \tfrac{1}{h}(y_1 - y_0); \quad C_2 = \tfrac{1}{2h^2}\Delta_{012} = \tfrac{1}{2h^2}(y_2 - 2y_1 + y_0).$$

Тогда $P'(x) = C_1 + C_2(2x - x_0 - x_1)$ и

$$y_0' \equiv P'(x_0) = C_1 + C_2(x_0 - x_1),$$
$$y_0'' \equiv P''(x_0) = C_2.$$

Подставляя $C_{1,2}$ и учитывая что узлы равноотстоящие, получаем

$$y_0' = \frac{1}{h}(y_1 - y_0) + \frac{1}{2h^2}(y_2 - 2y_1 + y_0)(-h) = \frac{1}{2h}(2y_1 - 2y_0 - y_2 + 2y_1 - y_0),$$

и окончательно

$$y_0' = \frac{-3y_0 + 4y_1 - y_2}{2h}, \qquad y_0'' = \frac{y_2 - 2y_1 + y_0}{2h^2}. \tag{1.30}$$

Получены формулы для первой и второй производных по трем точкам с использованием правых разностей (разностей вперед).

Пример оценки точности

Для оценки точности (1.30) рассмотрим разложение функции $y = f(x)$ в ряд Тейлора.

$$f(x + \varepsilon) = f(x) + \varepsilon f'(x) + \frac{\varepsilon^2}{2!}f''(x) + \frac{\varepsilon^3}{3!}f'''(x) + \ldots \tag{1.31}$$

Положим $x = x_0$, $\varepsilon = h$, тогда $f(x_0 + h) = f(x_1) = y_1$:

$$y_1 = y_0 + hf'(x_0) + \frac{h^2}{2}f''(x_0) + \frac{h^3}{6}f'''(x_0) + \dots$$

Положим $x = x_0$, $\varepsilon = 2h$, тогда $f(x_0 + 2h) = f(x_2) = y_2$:

$$y_2 = y_0 + 2hf'(x_0) + 2h^2f''(x_0) + \frac{4h^3}{3}f'''(x_0) + \dots$$

Из этих двух уравнений исключаем $f''(x_0)$:

$$4y_1 - y_2 = 3y_0 + 2hf'(x_0) - \frac{2}{3}h^3f'''(x_0) + \dots,$$

откуда получаем выражение для $y'_0 \equiv f'(x_0)$:

$$y'_0 = \frac{-3y_0 + 4y_1 - y_2}{2h} + \frac{h^2}{3}f'''(x_0) + \dots. \tag{1.32}$$

Последнее слагаемое $\frac{h^2}{3}y'''_0$ отличает это выражение от ранее полученного (1.30) и определяет величину ошибки.

1.2.5.1 Различные формулы

Все эти формулы можно вывести,☆ используя соответствующие формулы интерполяции, или, вместе с погрешностью, через разложение в ряд Тейлора.

1. y'_0 по трем точкам

$$\begin{aligned} y'_0 &= \frac{-y_2 + 4y_1 - 3y_0}{2h} + \frac{h^2}{3}y'''_0 \quad \text{– правые разности;} \\ y'_0 &= \frac{y_1 - y_{-1}}{2h} - \frac{h^2}{6}y'''_0 \quad \text{– центральные разности;} \\ y'_0 &= \frac{3y_0 - 4y_{-1} + y_{-2}}{2h} + \frac{h^2}{3}y'''_0 \quad \text{– левые разности.} \end{aligned}$$

2. y'_0 по четырем точкам

$$\begin{aligned} y'_0 &= \frac{2y_3 - 9y_2 + 18y_1 - 11y_0}{6h} - \frac{h^3}{4}y_0^{(4)} \quad \text{– правые разности;} \\ y'_0 &= \frac{11y_0 - 18y_{-1} + 9y_{-2} - 2y_{-3}}{6h} + \frac{h^3}{4}y_0^{(4)} \quad \text{– левые разности.} \end{aligned}$$

3. y'_0 по пяти точкам

$$y'_0 = \frac{-y_2 + 8y_1 - 8y_{-1} + y_{-2}}{12h} + \frac{h^4}{30}y_0^{(5)} \quad \text{– центральные разности.}$$

4. y_0'' по трем точкам

$$\begin{aligned} y_0'' &= \frac{y_2 - 2y_1 + y_0}{h^2} - hy_0''' \quad \text{– правые разности;} \\ y_0'' &= \frac{y_1 - 2y_0 + y_{-1}}{h^2} - \frac{h^2}{12}y_0^{(4)} \quad \text{– центральные разности;} \\ y_0'' &= \frac{y_0 - 2y_{-1} + y_{-2}}{h^2} + hy_0''' \quad \text{– левые разности.} \end{aligned}$$

1.2.5.2 О точности счета производных

В пространстве непрерывно дифференцируемых функций расстояние между функциями определим как

$$\rho[x(t), y(t)] = \max_{[a,b]} |x(t) - y(t)|.$$

Рассмотрим две функции: $x(t)$ и $\tilde{x}_n(t) = x(t) + \frac{1}{n}\sin\left[n^2(t-a)\right]$. Тогда на достаточно большом промежутке

$$\rho[x, \tilde{x}_n] = \max_{[a,b]} \left|\frac{1}{n}\sin\left[n^2(t-a)\right]\right| = \frac{1}{n}.$$

Но $\tilde{x}_n'(t) = x'(t) + n\cos\left[n^2(t-a)\right]$ и поэтому

$$\rho[x', \tilde{x}_n'] = \max_{[a,b]} \left|n\cos\left[n^2(t-a)\right]\right| = n.$$

Как видно, для сколь угодно близких функций расстояние между их производными может быть сколь угодно велико.

Из иллюстраций для распределения погрешности интерполяции по отрезку (рис.1.1 для равноотстоящих узлов и рис.1.3 для нулей полиномов Чебышёва) видно, что ошибка интерполяции Лагранжа качественно ведет себя похоже на $\sin nx$, так что приведенное тривиальное рассуждение вполне иллюстрирует общую проблему.

При увеличении n, даже если узлы не равноотстоящие, так что абсолютная величина ошибки минимальна, "частота" осцилляций растет, и с каждым дифференцированием ошибка приближения увеличивается в $\sim n$ раз.

Вывод: *Если есть возможность не применять формулы численного дифференцирования, то надо воспользоваться этой возможностью.*

С другой стороны, то же рассуждение приводит к тому соображению, что при интегрировании знакопеременная ошибка "сглаживается" и величина ошибки уменьшается. Поэтому приближенные формулы численного интегрирования, построенные на основе формул интерполяции, работают довольно хорошо.

Также для численного решения дифференциальных уравнений в ряде случаев уравнения вначале сводятся к интегральным, а затем, при помощи формул интерполяции, строятся их приближенные решения. Более подробно этот вопрос будет рассмотрен в соответствующих разделах.

1.3 Тригонометрическая интерполяция*

Во второй части работы 1885 года Вейерштрасс доказал свою вторую аппроксимационную теорему, которая отличается от первой лишь тем, что речь в ней идет не об алгебраических, а о тригонометрических полиномах:
T^0: *Пусть* $f(x)$ *– непрерывная функция на* $[0, 2\pi]$. *Тогда для всякого* $\varepsilon > 0$ *существует тригонометрический многочлен*[19]

$$\Psi(x) = \frac{A_0}{2} + \sum_{n=1}^{M} (A_n \cos nx + B_n \sin nx), \tag{1.33}$$

такой что

$$|f(x) - \Psi(x)| < \varepsilon \quad \forall x \in [0, 2\pi]. \tag{1.34}$$

Рассмотрим теперь задачу линейной интерполяции (1.3) сетки данных

$$\{x_k, f_k\}_{k=0}^{N-1}, \quad \text{где} \quad x_k \in [0, 2\pi],$$

тригонометрическим многочленом. То есть, ищем коэффициенты A_i и B_i, такие чтобы

$$\Psi(x_k) = f_k, \quad k = 0, 1, \ldots, N-1.$$

Здесь и далее мы будем полагать, что число узлов нечетно[20]

$$N = 2M + 1.$$

Также ограничимся наиболее важным случаем равноотстоящих узлов

$$x_k = \frac{2\pi k}{N}, \quad k = 0, 1, \ldots, N-1. \tag{1.35}$$

[19] Фактически он представляет собой частичную сумму некоторого ряда Фурье (см. пп. 1.3.4 и 3.1.10)

[20] В случае $N = 2M$ нужно убрать из суммы слагаемое $B_M \sin Mx$ – так останется столько же коэффициентов, сколько узлов. Вычисления будут немного другие, но но существу ничего не изменится.

1.3.1 Фазовые многочлены

Задачу интерполяции (1.33,1.34) на сетке (1.35), перейдя к комплексной записи, можно свести к задаче о нахождении *фазового многочлена* степени $(N-1)$

$$p(x) = \beta_0 + \beta_1 e^{ix} + \ldots + \beta_{N-1} e^{i(N-1)x}, \tag{1.36}$$

с N комплексными коэффициентами β_i, такими что

$$p(x_k) = f_k, \quad k = 0, 1, \ldots, N-1. \tag{1.37}$$

◂ В самом деле, на нашей сетке

$$e^{-inx_k} = e^{-2\pi i nk/N} = e^{-2\pi i (N-n)k/N} = e^{i(N-n)x_k}. \tag{1.38}$$

Поэтому, переписывая экспоненты в $p(x_k)$ через формулу Муавра, получим (помним, что $N = 2M+1$),

$$\begin{aligned} f_k = p(x_k) &= \beta_0 + \sum_{n=1}^{N-1} \beta_n e^{inx_k} = \beta_0 + \sum_{n=1}^{M} \beta_n e^{inx_k} + \sum_{n=M+1}^{N-1} \beta_n e^{inx_k} = \\ &= \left\{ \begin{array}{c} \text{во второй сумме} \\ \text{делаем замену } l = N-n \\ \text{и учитываем (1.38)} \end{array} \right\} = \beta_0 + \sum_{n=1}^{M} \left[\beta_n e^{inx_k} + \beta_{N-n} e^{-inx_k} \right] = \\ &= \underbrace{\beta_0}_{A_0/2} + \sum_{n=1}^{M} \left[\underbrace{(\beta_n + \beta_{N-n})}_{A_n} \cos nx_k + \underbrace{i(\beta_n - \beta_{N-n})}_{B_n} \sin nx_k \right]. \end{aligned}$$

Получили условие задачи тригонометрической интерполяции, и заодно выразили A_k и B_k через β_i. ▸

Обратные соотношения:

$$\beta_0 = \frac{A_0}{2}, \quad \beta_n = \frac{1}{2}(A_n - iB_n), \quad \beta_{N-n} = \frac{1}{2}(A_n + iB_n), \quad n = 1, \ldots, M.$$

Заметим, что хотя $p(x_k) = \Psi(x_k) = f_k$ для всех узлов, это вовсе не означает, что $p(x) = \Psi(x)$. На самом деле это две разные функции, а соответствующие интерполяционные задачи эквивалентны лишь в том смысле, что по решению одной из них можно построить решение второй.

1.3.2 Дискретное преобразование Фурье (DFT)

Перейдя к новой переменной

$$\omega = e^{ix}, \quad \omega_k = e^{ix_k} = e^{2\pi i k/N},$$
$$P(\omega) = \beta_0 + \beta_1 \omega + \ldots + \beta_{N-1} \omega^{N-1},$$

видим, что задача свелась к интерполяции многочленами[21], откуда немедленно следует существование и единственность решения. В этом случае, правда, многочлены комплексные, а узлы расположены в комплексной плоскости, на единичной окружности. Но так как мы говорим о задаче *линейной* интерполяции, все сказанное выше о многочлене Лагранжа остается верным и в комплексном случае.

Теперь можно было бы вывести коэффициенты β_i из интерполяционного многочлена Лагранжа, но мы пойдем другим путем. Во-первых, обратим внимание на то, что[22]

$$\omega_n^j = \omega_j^n, \qquad \omega_n^{-j} = \overline{\omega_n^j}, \quad \text{для} \quad 0 \le j, n \le N-1. \tag{1.39}$$

Во-вторых, верно такое равенство

$$\sum_{k=0}^{N-1} \omega_k^j \overline{\omega_k^n} = N \cdot \delta_{jn} \tag{1.40}$$

◂ Величина ω_{j-n}, будучи одним из корней степени N из единицы, является корнем многочлена

$$\omega^N - 1 = (\omega - 1)\left(\omega^{N-1} + \omega^{N-2} + \ldots + 1\right).$$

Но приравнивая его нулю, получаем или $\omega_{j-n} = 1$, т.е. $j = n$ (и тогда сумма в (1.40) равна N), или

$$0 = \sum_{k=0}^{N-1} \omega_{j-n}^k = \sum_{k=0}^{N-1} \omega_k^{j-n} = \sum_{k=0}^{N-1} \omega_k^j \overline{\omega_k^n}. \quad \blacktriangleright$$

Равенство (1.40) можно интерпретировать следующим образом. Определим N-мерные комплексные вектора

$$w^{(n)} = (1, \omega_1^n, \ldots, \omega_{N-1}^n), \quad n = 0, 1, \ldots, N-1,$$

и введем для них обычное скалярное произведение

$$(u, v) = \sum_{j=0}^{N-1} u_j \overline{v_j}.$$

Тогда (1.40) означает, что вектора $\{w^{(n)}\}_{n=0}^{N-1}$ образуют ортогональный (хотя не нормированный) базис

$$\left(w^{(j)}, w^{(n)}\right) = N \cdot \delta_{jn}, \tag{1.41}$$

[21] Так как для всех $0 \le j, k \le N-1$, $j \ne k$ верно $\omega_j \ne \omega_k$.

[22] Чертой сверху мы будем здесь и далее обозначать комплексное сопряжение.

по которому можно разложить $N-$вектор $f = (f_0, f_1, \ldots, f_{N-1})$:

$$f_k = \beta_0 + \beta_1\omega_k^1 + \ldots + \beta_{N-1}\omega_k^{N-1} = \sum_{j=0}^{N-1}\beta_j\omega_k^j = \sum_{j=0}^{N-1}\beta_j\left(w^{(j)}\right)_k,$$
$$\Rightarrow \quad f = \sum_{j=0}^{N-1}\beta_j w^{(j)},$$

а коэффициенты разложения β_j находятся тривиально, так как базис ортогональный (1.41):

$$\beta_j = \frac{(f, w^{(j)})}{(w^{(j)}, w^{(j)})} = \frac{1}{N}\sum_{k=0}^{N-1} f_k\overline{\omega_k^j} = \frac{1}{N}\sum_{k=0}^{N-1} f_k e^{-2\pi i\, jk/N}.$$

Таким образом, решение задачи тригонометрической интерполяции дается коэффициентами

$$\beta_j = \frac{1}{N}\sum_{k=0}^{N-1} f_k e^{-2\pi i\, jk/N}, \quad j = 0, 1, \ldots, N-1. \tag{1.42}$$

Такое преобразование дискретной комплексной функции f называется *дискретным преобразованием Фурье*[23] *(discrete Fourier transform, DFT)*.

1.3.3 Быстрое преобразование Фурье (FFT)

Итак, задача интерполяции N равноотстоящих узлов на $[0, 2\pi]$ тригонометрическим многочленом приводит к необходимости вычислять дискретные преобразования Фурье (DFT) от f (1.42).

DFT и его разновидности[24] применяются повсеместно для обработки сигналов, сжатия (в lossy форматы) изображений (jpeg), аудио (mp3), видео (theora). Такое их широкое применение обусловлено, в первую очередь, наличием эффективного алгоритма счета коэффициентов, который называется *быстрым преобразованием Фурье (fast Fourier transform, FFT)*. Если прямое вычисление сумм вида (1.42) требует $O(N^2)$ операций, то FFT дает возможность считать (1.42), используя лишь $O(N \ln N)$ операций.

Рассмотрим один из подходов (*the Cooley-Tukey method*, 1965) в своей простейшей постановке, когда

$$N = N_p \equiv 2^p, \quad p \in \mathbb{N},$$

[23] Жан Батист Жозеф Фурье, Jean Baptiste Joseph Fourier; 1768-1830. Французский математик и физик; участвовал в Египетском походе Наполеона (1798–1801) в составе Легиона культуры.

[24] Наиболее широко распространено, пожалуй, дискретное косинусное преобразование (DCT). В нем четная вещественная дискретная функция раскладывается по косинусам.

и вернемся для этого к задаче интерполяции фазовыми многочленами на сетке равноотстоящих узлов предыдущего пункта.

$$p(x) = \sum_{k=0}^{N-1} \beta_k e^{ikx} \ : \quad p(x_j) = f_j, \quad x_j = \frac{2\pi j}{N}, \quad j = 0, 1, \ldots, N-1.$$

Пусть у нас уже есть два интерполяционных фазовых многочлена $q(x)$ и $r(x)$, степени $(N_{p-1}-1)$ каждый, такие что $q(x)$ интерполирует все точки сетки с четным индексом, а $r(x-2\pi/N)$ интерполирует все точки с нечетным индексом:

$$q(x_{2n}) = f_{2n}, \quad r(x_{2n}) = r(x_{2n-1} - 2\pi/N) = f_{2n+1}, \quad n = 0, 1, \ldots, N_1 - 1.$$

Такое определение $r(x)$ приводит к тому, что задачи интерполяции для нахождения $q(x)$ и $r(x)$ поставлены на одной и той же сетке, а отличаются только значениями функции в узлах.

Используя, что

$$e^{iNx_k/2} = e^{iN_{p-1}x_k} = e^{2\pi i\, N_{p-1}k/N_p} = e^{i\pi k} = (-1)^k,$$

мы можем представить полный интерполяционный многочлен как линейную комбинацию $q(x)$ и $r(x)$:

$$p(x) = q(x) \cdot \left(\frac{1 + e^{iNx/2}}{2}\right) + r(x - 2\pi/N) \cdot \left(\frac{1 - e^{iNx/2}}{2}\right). \tag{1.43}$$

Так как $e^{iNx/2}$ это (фазовый) многочлен степени $N/2$, то $p(x)$ – степени $N-1$, как и должно быть.

Для построения q и r применимо то же рассуждение, с той разницей, что число узлов интерполяционной задачи для них в два раза меньше. Рассуждая так же и далее, получаем *p-шаговую рекуррентную схему для интерполяции $N = 2^p$ узлов.*

Проиллюстрируем процесс последовательного построения интерполяционных многочленов для случая $p = 3$. На каждом этапе мы комбинируем a многочленов степени $(b-1)$, каждый из которых интерполирует f в b узлах, в $a/2$ многочленов степени $(2b-1)$, каждый из которых интерполирует f в $2b$ узлах:

Н.у.	Шаг 1	Шаг 2	Шаг 3
0(0) 1(1)	$0,4 \to 0\ (0,4)$	$0,2 \to 0\ (0,2,4,6)$	$0,1 \to 0\ (0,\ldots,7).$
2(2) 3(3)	$1,5 \to 1\ (1,5)$	$1,3 \to 1\ (1,3,5,7)$	
4(4) 5(5)	$2,6 \to 2\ (2,6)$		
6(6) 7(7)	$3,7 \to 3\ (3,7)$		

Здесь каждый многочлен на каждом этапе обозначен своим номером r (начиная с нуля), а узлы, в которых он интерполирует f, пишем в скобках. Начинаем процедуру с восьми многочленов нулевой степени $1, 2, \ldots, 8$, равных просто числам f_k.

В общем случае рекуррентная схема выглядит следующим образом: На шаге m (m=$1, \ldots, p$) строим

$$R = 2^{p-m} = \frac{N}{2^m} = \frac{N}{M}$$

фазовых многочленов степени $(M-1)$, каждый из которых интерполирует $M = 2^m$ точек сетки:

$$p_r^{(m)} = \beta_{r,0}^{(m)} + \beta_{r,1}^{(m)} e^{ix} + \ldots + \beta_{r,2-1}^{(m)} e^{(M-1)ix}, \qquad r = 0, 1, \ldots, R-1.$$

Они комбинируются из $2R$ многочленов степени $(M/2-1)$, интерполирующих каждый свой набор из $M/2$ точек сетки, по рекурсии (1.43):

$$2p_r^{(m)}(x) = p_r^{(m-1)}(x) \cdot \left(1 + e^{iMx/2}\right) + p_{R+r}^{(m-1)}(x - 2\pi/M) \cdot \left(1 - e^{iMx/2}\right).$$

Здесь нижний индекс r, которым нумеруются многочлены, соответствует первой (по порядку) из точек сетки, которую он интерполирует. В таких обозначениях $p_r^{(m)}$ оказываются так упорядоченными, что пара, интерполирующая равноудаленные перемежающиеся узлы отличается значением нижнего индекса на R. Поэтому во втором слагаемом и стоит p_{R+r}.

Для коэффициентов β тогда несложно вывести☆рекурсию в форме

$$\begin{array}{ll} 2\beta_{r,j}^{(m)} = & \beta_{r,j}^{(m-1)} + \beta_{R+r,j}^{(m-1)} \cdot e^{-2\pi i \cdot j/2^m}, \\ 2\beta_{r,M+j}^{(m)} = & \beta_{r,j}^{(m-1)} - \beta_{R+r,j}^{(m-1)} \cdot e^{-2\pi i \cdot j/2^m}, \end{array} \qquad \left\{ \begin{array}{l} r = 0, 1, \ldots, R-1, \\ j = 0, 1, \ldots, M/2-1. \end{array} \right.$$

Процедура начинается с ввода “многочленов” нулевой степени

$$\beta_{k,0}^{(0)} = f_k, \qquad k = 0, \ldots, N-1$$

и заканчивается коэффициентами

$$\beta_{0,j}^{(p)} = \beta_j, \qquad j = 0, \ldots, N-1.$$

1.3.4 Ряды и интегралы Фурье. Связь с DFT*

Напомним здесь формальные определения рядов и интегралов Фурье, с тем чтобы показать, как они связаны с дискретным преобразованием Фурье (DFT).

Ряд Фурье

Всякая функция $f(x)$, непрерывная на $[-\pi, \pi]$, и такая что $f(-\pi) = f(\pi)$, может быть на этом промежутке представлена в виде своего *ряда Фурье*[25]

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)\,. \tag{1.44}$$

Коэффициенты a_k и b_k легко найти, домножая ряд на $\sin nx$ или $\cos nx$ и интегрируя почленно от $-\pi$ до π:

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} d\xi \cos k\xi\ f(\xi); \quad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} d\xi \sin k\xi\ f(\xi).$$

Переписав в (1.44) синусы и косинусы по формуле Эйлера-Муавра, получим *ряд Фурье в комплексной форме*,

$$f(x) = \sum_{k=-\infty}^{+\infty} c_k e^{ikx}, \quad \text{где} \quad c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\xi\, e^{-ik\xi} f(\xi). \tag{1.45}$$

Понятно, что если ряд Фурье сходится к функции $f(x)$ на $[-\pi, \pi]$, то снаружи этого промежутка он будет сходится к периодически повторенному, до бесконечности, тому же куску $f(x)|_{x\in[-\pi,\pi]}$. То есть ряд Фурье непрерывной функции f сходится к f на всей вещественной оси тогда и только тогда, когда f периодична, с периодом 2π.

Что же если мы хотим рассматривать функции, периодичные с другим периодом? Очевидно, задача сводится к предыдущей путем замены переменных. Пусть $f(x) = f(x + 2l)$. Тогда введя $t = x \cdot \pi/l$, получим, что $f(t \cdot l/\pi)$, как функция t, периодична с периодом 2π и представима в виде ряда Фурье (1.45). Переписывая его в терминах x, получаем

$$f(x) = \sum_{n=-\infty}^{+\infty} c_n e^{i\pi n\cdot x/l}, \quad \text{где} \quad c_n = \frac{1}{2l}\int_{-l}^{l} d\xi\, e^{-i\pi n\cdot \xi l} f(\xi). \tag{1.46}$$

Интегральное преобразование Фурье

Тогда мы можем устремить l к бесконечности, и получить "ряд" (в кавычках потому что на самом деле получается интеграл, см. ниже) для вовсе не периодической функции, определенной на $(-\infty, \infty)$. Для этого удобно переписать

[25] Здесь не будем приводить никаких пояснений и доказательств. В следующем разделе этот вопрос весьма подробно обсуждается (см. п. 3.1.10).

(1.46) в виде

$$f(x) = \sum_{n=-\infty}^{+\infty} \frac{1}{2l} \int_{-l}^{l} d\xi\, f(\xi) e^{-i\lambda_n (x-\xi)}, \quad \text{где} \quad \lambda_n = n\frac{\pi}{l}.$$

Здесь сумма представляет собой интегральную сумму Римана, и в пределе

$$l \to \infty: \qquad 1 \equiv \Delta n = \Delta\lambda_n \cdot \frac{l}{\pi} \to d\lambda \cdot \frac{l}{\pi} \quad \Rightarrow \quad \frac{1}{2l} \to \frac{d\lambda}{2\pi},$$

так что получаем *повторный интеграл Фурье*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\lambda \int_{-\infty}^{+\infty} d\xi\, e^{i\lambda(x-\xi)} f(\xi). \tag{1.47}$$

То же самое в пошаговой записи дает нам *интегральное преобразование Фурье* – обратное и прямое (в том порядке что они записаны):

$$f(x) = \int_{-\infty}^{+\infty} d\lambda\, e^{i\lambda x} F(\lambda), \quad \text{где} \quad F(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\xi\, e^{-i\lambda\xi} f(\xi). \tag{1.48}$$

DFT как дискретное приближение ряда и интеграла Фурье

Вернемся к функциям, периодичным с периодом $2l$ и рядам Фурье для них (1.46). Интеграл для коэффициентов перепишем вначале, с помощью замены переменных $\xi = \eta - l$, как

$$c_n = \frac{1}{2l} \int_{-l}^{l} d\xi\, f(\xi) e^{-i\pi n\xi/l} = \left\{ e^{i\pi n} = (-1)^n \right\} = \frac{(-1)^n}{2l} \int_{0}^{2l} d\eta\, \tilde{f}(\eta) e^{-i\pi n\eta/l},$$

где $\tilde{f}(\eta) = f(\eta - l)$. Его, очевидно, можно вычислить приближенно, заменяя соответствующей интегральной суммой. В простейшем случае, для равномерного разбиения промежутка на N участков, имеем

$$(-1)^n c_n \approx \frac{1}{2l} \sum_{j=0}^{N-1} \frac{2l}{N} \tilde{f}(x_j) e^{-i\pi n x_j/l} = \left\{ x_j = 2l \cdot \frac{j}{N} \right\} = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{f}(x_j) e^{-2\pi i \cdot kn/N}.$$

Пришли в точности к дискретному преобразованию Фурье (1.42).

Что касается интегрального преобразования, то отличие для него заключается только в том, что оно содержит еще один предельный переход, в котором пределы интегрирования $\int_{-l}^{l} \ldots$ стремятся к бесконечности. При этом DFT дает его приближение на дискретном наборе точек.

Таким образом, DFT выступают как дискретные аппроксимации коэффициентов рядов Фурье и для преобразования Фурье.

1.4 Интерполяция Эрмита*

Пусть у нас есть набор не совпадающих узлов $\{x_i\}_{i=0}^m$, упорядоченных по возрастанию

$$x_0 < x_1 < \ldots < x_m,$$

и таблица чисел

$$f_i^{(k)} \quad \text{для} \quad i = 0, \ldots, m, \quad k = 0, 1, \ldots, n_i - 1.$$

Интерполяционная задача Эрмита (Hermite interpolation problem) для этих данных состоит в нахождении многочлена

$$P \in \overline{\Pi}_n, \quad n = \sum_{i=0}^{m} n_i - 1,$$

который удовлетворяет следующим условиям интерполяции:

$$P^{(k)}(x_i) = f^{(k)}(x_i), \quad i = 0, 1, \ldots, m; \quad k = 0, 1, \ldots, n_i - 1. \tag{1.49}$$

Эта задача отличается от обычной задачи интерполяции тем, что в i-том узле задается не только значение самого полинома, но и значения его первых $n_i - 1$ производных – так что всего на узел приходится n_i условий. Рассмотренная ранее полиномиальная интерполяция есть частный случай с $n_i = 1$.

Всего, таким образом, задается $\sum_{i=0}^m n_i = n + 1$ условий на $n + 1$ коэффициент искомого многочлена, так что можно ожидать, что поставленная задача имеет и единственное решение.

◂ Докажем единственность. Пусть $P_{1,2}(x)$ это два многочлена степени не выше n, удовлетворяющих (1.49). Тогда их разность $Q = P_1 - P_2$ также есть многочлен степени не выше n. При этом

$$Q^{(k)}(x_i) = 0, \quad k = 0, 1, \ldots, n_i - 1, \quad i = 0, 1, \ldots, m,$$

то есть x_i есть по крайней мере n_i-кратный корень Q, а значит всего у него $\sum n_i = n + 1$ корней, с учетом кратности. Следовательно, Q есть тождественный ноль, и единственность доказана.▸

Существование докажем в следующем пункте конструктивно, построив решение.

1.4.1 Обобщенные многочлены Лагранжа

Решение этой системы линейных уравнений, очевидно, линейно по неоднородности f, и следовательно записывается в виде

$$P(x) = \sum_{i=0}^{m} \sum_{k=0}^{n_i-1} f_i^{(k)} L_{ik}(x). \tag{1.50}$$

Многочлены $L_{ik}(x)$ называют *обобщенными многочленами Лагранжа* (см. (1.10)). Они должны удовлетворять соотношениям

$$L_{ik}^{(\sigma)}(x_j) = \delta_{ij}\delta_{\sigma k}, \quad \text{для} \quad k, \sigma = 0, 1, \ldots, n_i - 1, \tag{1.51}$$

для того чтобы построенный многочлен (1.50) в самом деле являлся решением задачи (1.49).

Обобщенные многочлены Лагранжа определяются следующим образом. Вначале введем вспомогательные многочлены

$$l_{ik} = \frac{(x - x_i)^k}{k!} \prod_{\substack{j=0 \\ j \neq i}}^{m} \left(\frac{x - x_j}{x_i - x_j} \right)^{n_j}, \quad 0 \leq i \leq m,\ 0 \leq k \leq n_i.$$

Определим через них

$$L_{i,n_i-1}(x) := l_{i,n_i-1}(x), \quad i = 0, 1, \ldots, m,$$

а остальные L_{ik}, с меньшими k, по рекурсии

$$L_{ik}(x) := l_{ik}(x) - \sum_{\nu=k+1}^{n_i-1} l_{ik}^{(\nu)}(x_i) L_{i\nu}(x). \tag{1.52}$$

Докажем по индукции, что для таким образом определенных L_{ik} выполняются условия (1.51).

◂ Выпишем все случаи, когда производные степени $\sigma < n_j$ от l_{ik} в узлах равны нулю или единице:

$$l_{ik}^{(\sigma)}(x_i) = \delta_{k\sigma} \quad \text{для} \quad \sigma \leq k, \tag{1.53}$$

$$l_{ik}^{(\sigma)}(x_j) = 0 \quad \text{для} \quad i \neq j. \tag{1.54}$$

База. Пусть $k = n_i - 1$. Тогда $L_{ik} = l_{i,n_i-1}$, и

$$L_{i,n_i-1}^{(\sigma)}(x_i) = l_{i,n_i-1}^{(\sigma)}(x_i) = \delta_{n_i-1,\sigma}; \qquad L_{i,n_i-1}^{(\sigma)}(x_j) = 0 \quad j \neq i.$$

Таким образом, база

$$L_{i,n_i-1}^{(\sigma)}(x_j) = \delta_{ij}\delta_{n_i-1,\sigma}$$

доказана.

Предположим теперь, что (1.51) выполняется для всех $k = q, \dots, n_i - 1$. Покажем, что в таком случае то же будет верно и для $k = q - 1$. В самом деле, тогда согласно (1.52)

$$L^{(\sigma)}_{i,q-1}(x_j) = l^{(\sigma)}_{i,q-1}(x_j) - \sum_{\nu=q}^{n_i-1} l^{(\nu)}_{i,q-1}(x_i) L^{(\sigma)}_{i\nu}(x_j) = l^{(\sigma)}_{i,q-1}(x_j) - \delta_{ij} \sum_{k=q}^{n_i-1} l^{(\nu)}_{i,q-1}(x_i)\delta_{\sigma k}.$$

Если $\sigma < q < n_i$, то все слагаемые в сумме равны нулю, и

$$L^{(\sigma)}_{i,q-1}(x_j) = l^{(\sigma)}_{i,q-1}(x_j) = \left\{ \begin{array}{c} \text{из} \\ (1.53,1.54) \end{array} \right\} = \delta_{ij}\delta_{q-1,\sigma}.$$

Если же $q \le \sigma < n_i$, то член суммы, отличный от нуля (у которого $\nu = \sigma$), сокращается с первым слагаемым, и получаем тот же результат. Таким образом, для всех возможных i, j, σ доказываемое условие выполняется и для $k = q-1$. Поэтому оно верно для всех $k = 0, \dots, n_i - 1$, и формула (1.51) доказана. ▸

1.4.2 Погрешность

Пусть задана $(n+1)$ раз дифференцируемая на $[a, b]$ функция $f(x)$, и известны ее значения и первые несколько производных в наборе узлов $\{x_i \in [a, b]\}_{i=0}^{m}$, так что можно поставить задачу интерполяции Эрмита с $f_i^{(k)} = f^{(k)}(x_i)$.

Тогда ошибка интерполяции дается формулой

$$\forall\ x \in [a, b] \quad f(x) - P(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\hat{\omega}(x), \quad \text{где} \quad \xi \in [a, b], \tag{1.55}$$

$$\hat{\omega}(x) = (x - x_0)^{n_0}(x - x_1)^{n_1} \dots (x - x_m)^{n_m}. \tag{1.56}$$

Доказательство по существу повторяет проведенное выше для интерполяционного полинома Лагранжа (1.25):

◂ Рассмотрим

$$\varphi(x) = f(x) - P(x) - K \cdot \hat{\omega}(x).$$

Каждый узел x_j является по построению корнем $\varphi(x)$ кратности n_j; всего корней φ, с учетом кратности, $\sum n_j = n + 1$.

Зафиксируем постоянную K из требования чтобы $\varphi(x)$ обращалась в ноль в еще одной точке $\tilde{x} \ne x_0, \dots, x_n$:

$$K = \frac{f(\tilde{x}) - P(\tilde{x})}{\hat{\omega}(\tilde{x})}.$$

Тогда $\varphi(x)$ имеет $(n + 2)$ корня на $[a, b]$; $\varphi'(x)$ имеет $(n + 1)$ корня; и так далее; $\varphi^{(n+1)}(x)$ имеет один корень на $[a, b]$:

$$\exists\ \xi \in [a, b] \mid \quad \varphi^{(n+1)}(\xi) = 0.$$

Но $P(x) \in \overline{\Pi}_n$, поэтому

$$0 = \varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K(n+1)!.$$

Выражая отсюда K, получаем

$$f(\tilde{x}) - P(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\hat{\omega}(\tilde{x}).$$

Т.к. $\tilde{x}$ можно выбрать произвольно на $[a,b]$, опускаем тильду и для $\forall x \in [a,b]$ получаем приведенное выше выражение для ошибки. ▸

Если положить $n_i = N$, и выбрать узлы в нулях многочлена Чебышева, то получим

$$\sup \hat{\omega} = \sup(\omega_n)^N = \frac{1}{2^{Nn}},$$

так что для оценки ошибки сверху

$$|f(x) - P_n(x)| \le \frac{\sup |f^{(n+1)}(\xi)|}{(n+1)! \cdot 2^{nN}}.$$

При достаточно большом N (а на практике уже часто и для $N = 2$) знаменатель подавит рост с n производных в числителе, если только они конечны, так что при $n \to \infty$ получим равномерную сходимость $P_n \to f$.

1.5 Сплайн-интерполяция*

1.5.1 Разные сплайны. Постановка задачи.

Под сплайном обычно понимается определенная в некоторой области D кусочно-полиномиальная функция, т.е. функция, для которой существует разбиение D на подобласти, такое, что внутри каждого элемента разбиения функция представляет собой многочлен некоторой степени n. Кроме того, эта функция непрерывна в области D вместе с производными до k-го порядка включительно $(k \le n)$.

Мы рассмотрим один простой частный случай задачи на сплайн-интерполяцию – интерполяцию кубическими многочленами в классе C^2. Это наиболее широко используемый на практике вариант.

Постановка задачи кусочно-кубической интерполяции:
На отрезке $[a,b]$ вещественной оси задана сетка $a = x_0 < x_1 < \ldots < x_n = b$, в узлах которой заданы значения $\{f_k\}_{k=0}^n$ функции $f(x)$, определенной на $[a,b]$. На отрезке $[a,b]$ необходимо найти функцию $g(x)$, удовлетворяющую требованиям:

1. $g(x)$ непрерывна вместе со своими производными до второго порядка включительно
$$g(x) \in C^2[a,b]; \tag{1.57}$$

2. На каждом из отрезков $[x_{k-1}, x_k]$ функция $g(x)$ является кубическим многочленом:
$$g(x)\Big|_{[x_{k-1},x_k]} \equiv g_k(x) = \sum_{i=0}^{3} a_{ik}(x - x_k)^i; \tag{1.58}$$

3. В узлах сетки $g(x)$ интерполирует $f(x)$:
$$g(x_k) = f_k \quad \text{для } k = 0, 1, \ldots, n; \tag{1.59}$$

4. На концах промежутка выполняются граничные условия "свободного провисания"
$$g''(a) = g''(b) = 0. \tag{1.60}$$

1.5.2 Кубические сплайны

Для нахождения $g(x)$ воспользуемся условиями (1–4).

Непрерывность второй производной: так как вторая производная функции $g(x)$ линейна на каждом отрезке сетки $[x_{i-1}, x_i]$, для $i = 1, \ldots, n$, то на каждом из этих промежутков мы можем записать:
$$g''(x) \equiv g_i''(x) = m_{i-1}\frac{x_i - x}{h_i} + m_i\frac{x - x_{i-1}}{h_i} \quad \text{для} \quad x \in [x_{i-1}, x_i],\ i = 1, \ldots, n, \tag{1.61}$$
где $h_i = x_i - x_{i-1}$, $m_i = g''(x_i)$.

Проинтегрируем равенство (1.61) дважды и, перегруппировав постоянные интегрирования, запишем
$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + A_i\frac{x_i - x}{h_i} + B_i\frac{x - x_{i-1}}{h_i}. \tag{1.62}$$

Константы A_i, B_i получим из условий (3): $g(x_{i-1}) = f_{i-1}$ и $g(x_i) = f_i$. Подставляя $x = x_{i-1}, x_i$ в (1.62), получим
$$f_{i-1} = m_{i-1}\frac{h_i^2}{6} + A_i; \quad f_i = m_i\frac{h_i^2}{6} + B_i. \tag{1.63}$$

Тогда из (1.62) получаем выражения для g_i и ее производной g_i'в виде

$$\begin{aligned} g_i(x) &= m_{i-1}\frac{(x_i-x)^3}{6h_i} + m_i\frac{(x-x_{i-1})^3}{6h_i} + \\ &+\left(f_{i-1} - m_{i-1}\frac{h_i^2}{6}\right)\frac{x_i-x}{h_i} + \left(f_i - m_i\frac{h_i^2}{6}\right)\frac{x-x_{i-1}}{h_i}; \end{aligned} \tag{1.64}$$

$$\begin{aligned} g_i'(x) &= -m_{i-1}\frac{(x_i-x)^2}{2h_i} + m_i\frac{(x-x_{i-1})^2}{2h_i} + \\ &+\frac{f_i-f_{i-1}}{h_i} - \frac{m_i-m_{i-1}}{6}h_i. \end{aligned} \tag{1.65}$$

Осталось использовать непрерывность в узлах первой производной g. Из (1.65) получаем односторонние производные в точках $x_1, \ldots, x_n$:

$$\begin{aligned} g'(x_i-0) \equiv g_i'(x_i) &= \frac{h_i}{6}m_{i-1} + \frac{h_i}{3}m_i + \frac{f_i-f_{i-1}}{h_i}; \\ g'(x_i+0) \equiv g_{i+1}'(x_i) &= -\frac{h_{i+1}}{3}m_i - \frac{h_{i+1}}{6}m_{i+1} + \frac{f_{i+1}-f_i}{h_{i+1}}. \end{aligned}$$

Приравнивая, получаем систему $(n-1)$ уравнений для m_i, $i = 0, \ldots, n$:

$$\frac{h_i}{6}m_{i-1} + \frac{h_i+h_{i+1}}{3}m_i + \frac{h_{i+1}}{6}m_{i+1} = \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i}. \tag{1.66}$$

Дополняя ее условиями $m_0 = m_n = 0$, которые следуют из граничного условия (1.60), получаем замкнутую систему для $\{m_i\}_{i=1}^{n-1}$. В матричном виде она записывается как

$$Am = Hf. \tag{1.67}$$

Здесь A – квадратная матрица $(n-1)\times(n-1)$

$$A = \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \ldots & 0 & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \ldots & 0 & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{h_{n-2}+h_{n-1}}{3} & \frac{h_{n-1}}{6} \\ 0 & 0 & 0 & \ldots & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} \end{pmatrix}, \tag{1.68}$$

m и f вектора

$$m = \begin{pmatrix} m_1 \\ \vdots \\ m_{n-1} \end{pmatrix}, \; f = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix},$$

а H – матрица $(n+1)\times(n-1)$

$$H=\begin{pmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1}+\frac{1}{h_2}\right) & \frac{1}{h_2} & \dots & 0 & 0 \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_2}+\frac{1}{h_3}\right) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\left(\frac{1}{h_{n-1}}+\frac{1}{h_n}\right) & \frac{1}{h_n} \end{pmatrix}.$$

Матрица A положительно определена и неособенная.

◂ Пусть есть произвольная ненулевая матрица $B=(b_{ij})$ и λ – некоторое ее собственное значение, соответствующее собственному вектору $x=(x_1,\dots,x_n)$. Перенумеруем так оси координат, чтобы координаты x_i были в порядке убывания абсолютных значений:

$$|x_1|\ge|x_2|\ge\dots\ge|x_1|.$$

Тогда $x_1\neq 0$, и из первой строки $Ax=\lambda x$ имеем

$$b_{11}x_1+\dots+b_{1n}x_n=\lambda x_1.$$

Деля на x_1 и учитывая последовательность неравенств для x_j, получим

$$|\lambda-b_{11}|\le\sum_{i>1}|b_{i1}|, \tag{1.69}$$

то есть собственное значение лежит на комплексной плоскости в круге с центром в b_{11} и радиуса

$$R=\sum_{i>1}|b_{i1}|.$$

Это утверждение называют (первой) теоремой Гершгорина о локализации собственных значений, а круг – кругом Гершгорина. Понятно, что теорема Гершгорина применима для каждого собственного значения, так что все собственные значения матрицы лежат каждое в своем круге Гершгорина (которые, конечно, могут пересекаться).

В нашем случае центры кругов Гершгорина для матрицы A лежат на положительной части вещественной оси $a_{ii}>0$, и кроме того выполняется условие строгого диагонального преобладания

$$a_{ii}>\sum_{j\neq i}|a_{ij}|.$$

Поэтому радиус каждого ее круга Гершгорина больше чем абсцисса его центра, и этот круг не включает в себя начало координат. Учитывая, что все собственные значения симметричной матрицы вещественны, видим, что все они и положительны, и не обращаются в нуль. ▸

Поэтому коэффициенты $m_1,m_2,\dots,m_{n-1}$ определяются из системы (1.67) однозначно. Следовательно, сплайн-функция $g(x)$ также однозначно восстанавливается по формулам (1.64) и задача о нахождении кусочно-кубической функции $g(x)$ имеет единственное решение. Так как матрица трехдиагональна, то решение быстро вычисляется методом прогонки (см. п.2.4.3).

В явном виде g_i получаем, подставив решение системы (1.67) m_i в (1.64).

1.5.3 Сплайн как решение вариационной задачи

Кубические сплайн-функции обладают очень важным свойством, которое обусловливает высокую эффективность сплайн-интерполяции. А именно, рассмотрим класс функций $W_2^2[a,b]$, имеющих интегрируемые на $[a,b]$ с квадратом вторые производные:

$$\int_a^b dx(u''(x))^2 < \infty.$$

Поставим задачу отыскания функции $u(x) \in W_2^2[a,b]$, которая интерполирует в узлах $\{x_i\}_0^n$ нашу функцию $f(x)$ и минимизирует функционал

$$\Phi[u] = \int_a^b dx\,[u''(x)]^2 : \tag{1.70}$$

$$I = \inf\left\{\Phi[u]\,\middle|\, u(x_k) = f_k \quad \text{для } k = 0,\ldots,n; \quad u(x) \in W_2^2[a,b]\right\} \tag{1.71}$$

Утверждается, что минимум такого функционала достигается на кусочно-кубической сплайн-функции $g(x)$, которую мы только что построили.

◂ Рассмотрим величину

$$\Phi[u-g] = \int_a^b dx\,[u''(x) - g''(x)]^2.$$

Интегрируя по частям и используя что $g''(a) = g''(b) = 0$, получим

$$\begin{aligned}
\Phi[u-g] &= \int_a^b dx\Big([u'']^2 - [g'']^2 + 2g''(g''-u'')\Big) = \Phi[u] - \Phi[g] + 2\int_a^b g''d(g'-u') = \\
&= \Phi[u] - \Phi[g] + 2\,[g''(g'-u')]_a^b - 2\int_a^b (g'-u')g'''dx = \\
&= \Phi[u] - \Phi[g] - 2\int_a^b (g'-u')g'''dx
\end{aligned}$$

Но на каждом из промежутков (x_{i-1}, x_i), $i = 1,\ldots,n$ функция g''' есть константа $g''' = c_i$ и тогда

$$\Phi[u-g] = \Phi[u] - \Phi[g] - 2\sum_{i=0}^{n} c_i\,[g(x) - u(x)]\Big|_{x_{i-1}}^{x_i} = \Phi[u] - \Phi[g].$$

Значит

$$\Phi[g] = \Phi[u] - \Phi[u-g] \le \Phi[u]$$

и кубический сплайн g минимизирует заданный функционал, ч. и. т. д.

Единственность доказывается от противного: пусть

$$\Phi[g] = \Phi[u] \quad \Rightarrow \quad \Phi[g-u] = 0 \quad \Rightarrow \quad u'' = g'',$$

тогда $u(x)$ так же как и g является сплайном, причем с теми же коэффициентами m_i во всех узлах, и следовательно тождественно совпадает с g. ▸

Основываясь на (1.71), можно дать другое, *эквивалентное определение кусочно-кубической сплайн-функции*: это такая функция из класса $W_2^2[a,b]$, которая принимает в узлах сетки заданные значения и минимизирует функционал (1.70).

Такое свойство сплайн-функций интересно тем, что функционал $\Phi[u]$ можно интерпретировать как аналог потенциальной энергии упругого стержня, закрепленного в точках плоскости (x_k, f_k), и на кубических сплайнах реализуется минимум этой энергии.

1.5.4 Вариации граничных условий*

До сих пор мы ограничились рассмотрением кубических сплайнов, удовлетворяющих граничным условиям (1.60), которые представляют собой условия "свободного провисания" интерполяционной кривой в точках a и b. Однако на практике часто бывают известны наклоны интерполяционной кривой в граничных точках. Тогда становится естественным применение условий

$$g'(a) = f'_0; \quad g'(b) = f'_n. \tag{1.72}$$

Если мы знаем кривизну кривой в точках a и b, то естественны условия

$$g''(a) = f''_0; \quad g''(b) = f''_n. \tag{1.73}$$

Если же об интерполируемой функции известно априори, что она периодична с периодом $b-a$, то $f_0 = f_n$ и следует применить граничные условия

$$g'(a) = g'(b); \quad g''(a) = g''(b). \tag{1.74}$$

Каким же образом изменится система линейных алгебраических уравнений (1.67) при наличии такого ряда граничных условий? В простейшем случае (1.60) мы пополняли систему (1.66) равенствами $m_0 = m_n = 0$.

Условия (1.72) приведут нас, с учетом (1.65), к равенствам

$$\frac{2}{3}m_0 + \frac{1}{3}m_1 = \frac{2}{h_1}\left(\frac{f_1 - f_0}{h_1} - f'_0\right), \quad \frac{1}{3}m_{n-1} + \frac{2}{3}m_n = \frac{2}{h_n}\left(f'_n - \frac{f_n - f_{n-1}}{h_n}\right);$$

условия (1.73) к равенствам

$$m_0 = f_0'', \quad m_n = f_n'';$$

и наконец, условие периодичности сплайна (1.74) к

$$m_0 = m_n, \quad \frac{h_1 + h_n}{3} m_n + \frac{h_n}{6} m_{n-1} + \frac{h_1}{g} m_1 = \frac{f_1 - f_0}{h_1} - \frac{f_n - f_{n-1}}{h_n}.$$

Этими равенствами и следует пополнить систему (1.67). Конечно, возможно комбинировать условия различных типов в точках a и b.

Следует отметить, что кубические сплайны с различными типами краевых условий все равно обеспечивают минимум функционала (1.70), только уже не на всем классе функций $W_2^2[a,b]$, а на подмножестве этого класса, состоящем из функций, удовлетворяющих данному краевому условию.

1.5.5 Ошибка кубической сплайн-интерполяции**

Проиллюстрируем на простейшем примере технику получения оценок сходимости кубических сплайн-функций и их производных. Ограничимся для простоты граничными условиями (1.72) и, чтобы не усложнять формул, только равномерными сетками $h_i = h$. Эти ограничения по ходу доказательств не играют существенной роли. Также предположим, что функция f дважды дифференцируема.

Рассмотрим уравнение $Am = Hf$ (1.67) для вторых производных сплайна в узловых точках m_i. В случае равномерной сетки матрицы A и H сильно упрощаются:

$$A = hB; \qquad B = \begin{pmatrix} \frac{2}{3} & \frac{1}{6} & 0 & \dots & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \dots & 0 & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & 0 & \dots & \frac{1}{6} & \frac{2}{3} \end{pmatrix}; \qquad H = \frac{1}{h} \begin{pmatrix} 1 & -2 & 1 & \dots & 0 & 0 \\ 0 & 1 & -2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -2 & 1 \end{pmatrix}.$$

Напомним следующее понятие. *Модулем непрерывности* $\omega(h, \varphi)$ для непрерывной функции $\varphi(x)$, заданной на отрезке $[a,b]$, называется величина

$$\omega(h, \varphi) = \sup_{\substack{x', x'' \in [a,b] \\ |x' - x''| \le h}} \left| \varphi(x') - \varphi(x'') \right|.$$

Она представляет собой максимальное изменение (колебание) функции $\varphi(x)$ на отрезке длины h в рамках $[a,b]$.

Обозначим вектор $\frac{1}{h}Hf$ через d. Тогда разделив уравнение $Am = Hf$ на h, получим $Bm = d$. Вычтя справа и слева Bd, запишем

$$m - d = B^{-1}(I - B)\,d. \tag{1.75}$$

Из явного вида H следует, что

$$d_j = \frac{f_{j+1} - 2f_j + f_{j-1}}{h^2} = \frac{1}{h}\left(\frac{f_{j+1} - f_j}{h} - \frac{f_j - f_{j-1}}{h}\right).$$

Это вторая центральная разность функции f. Так как первые разности, стоящие в скобках, равны[26] значениям f' на промежутках x_{j-1}, x_j и x_j, x_{j+1} соответственно, то аналогично

$$\exists\, \xi_j \in (x_{j-1}, x_{j+1}) \quad | \quad d_j = f''(\xi_j).$$

Тогда, т.к.

$$|d_{j-1} - d_j| = |f''(\xi_{j-1}) - f''(\xi_j)|, \quad \text{где} \quad \xi_j \in (x_j, x_{j+2}), \quad \Rightarrow \quad \xi_j - \xi_{j-1} \in (h, 3h),$$

то

$$\left|\Big((I-B)d\Big)_j\right| = \left|\frac{d_{j-1} + d_{j+1}}{6} - \frac{d_j}{3}\right| = \left|\frac{d_{j-1} - d_j}{6} + \frac{d_{j+1} - d_j}{6}\right| \le$$
$$\le \frac{1}{6}\Big(|d_{j-1} - d_j| + |d_j - d_{j+1}|\Big) \le \frac{1}{3}\omega(3h, f'') \le \omega(h, f''). \tag{1.76}$$

В последнем неравенстве мы использовали выпуклость $\omega(h)$, которая достаточно очевидна[27].

Норму матрицы[28] B^{-1} оценим из теоремы Гершгорина о локализации собственных значений для B. В соответствии с ней, все собственные значения B лежат на комплексной плоскости в круге с центром в $2/3$ и радиуса $1/3$, а значит вещественны (вследствие симметрии B) и больше $1/3$. Поэтому все собственные значения B^{-1} не больше 3 и $\|B^{-1}\| \le 3$. Тогда из (1.75) и (1.76)[29]

$$|m_j - d_j| \le \|B^{-1}\| \cdot \max\left|\Big((I-B)d\Big)_j\right| \le 3\omega(h, f'').$$

[26] По теореме Лагранжа, из французских теорем анализа.

[27] Если колебание функции f на промежутке длины h не превышает δ, то колебание f на промежутке длины $3h$ не превышает 3δ.

[28] Норма матрицы A как оператора определяется как $\|A\| = \sup\limits_x \frac{\|Ax\|}{\|x\|}$, так что $\forall x\ \|Ax\| \le \|A\| \cdot \|x\|$.

[29] Здесь фактически записано неравенство $\|m - d\| \le \|B\|^{-1} \cdot \|(I-B)d\|$, где в качество нормы для векторов используется максимум-норма $\|a\|_{max} = \max\limits_j |a_j|$

Также, прямо из определения ω,

$$|f''(x_j) - d_j| \le \omega(h, f''),$$

так что

$$|f''(x_j) - m_j| \le |f''(x_j) - d_j| + |d_j - m_j| \le 4\omega(h, f'').$$

Тогда для всякого $x \in (x_{j-1}, x_j)$ будет верно:

$$|f''(x) - m_j| \le |f''(x) - f''(x_j)| + |f''(x_j) - m_j| \le \omega(h, f'') + 4\omega(h, f'') = 5\omega(h, f''),$$

и используя (1.61), получаем

$$|f''(x) - g''(x)| \le \frac{x_j - x}{h} |f''(x) - m_{j-1}| + \frac{x - x_{j-1}}{h} |f''(x) - m_j| \le 5\omega(h, f'').$$

Ввиду равенства $f(x_j) = g(x_j)$, на каждом интервале (x_{j-1}, x_j) найдется точка η_j, для которой $f'(\eta_j) = g'(\eta_j)$. Следовательно

$$|f'(x) - g'(x)| = \left| \int_{\eta_j}^{x} dx \big[f''(x) - g''(x) \big] \right| \le 5h\omega(h, f'').$$

Повторяя интегрирование, находим

$$|f(x) - g(x)| \le \frac{5}{2} h^2 \omega(h, f'').$$

Таким образом, мы получили оценки сходимости самого сплайна, а также его первой и второй производной. Неравенства такого вида, выражающие величину наилучшего приближения некоторой функции f заданным классом функций через ее модуль непрерывности, называют в теории приближений неравенствами типа Джексона-Стечкина. Отметим, что выведенные оценки сохраняются для граничных условий (1.73) и (1.74).

С возрастанием гладкости функции $f(x)$ оценки ходимости улучшаются. Однако при помощи кубических сплайнов нельзя добиться сходимости выше чем $O(h^4)$, если, конечно, функция $f(x)$ не является кубическим многочленом.

1.5.6 Кусочно-кубическая интерполяция со сглаживанием*

Рассмотрим *задачу о гладком восполнении* функции, которая определена на сетке: $a = x_0 < x_1 < \ldots < x_n = b$. Однако теперь значения функции $\tilde{f}_i$ в узлах сетки возмущены некоторой погрешностью. В этом случае не имеет смысла

строить интерполяционную функцию, которая в узлах в точности совпадает с заданными значениями. Более того, следует построить функцию, которая проходила бы вблизи заданных значений более "плавно" чем интерполяционная. Такие функции называют уже не интерполяционными, а сглаживающими. Потребуем, чтобы искомая сглаживающая функция $g(x)$ минимизировала на классе $W_2^2[a,b]$ функционал

$$\tilde{\Phi}[u] = \int_a^b dx\,[u''(x)]^2 + \sum_{k=0}^{n} p_k \left[u(x_k) - \tilde{f}_k\right]^2, \tag{1.77}$$

где $p_k > 0$ – некий набор констант. В функционале $\tilde{\Phi}[u]$ скомбинированы интерполяционные условия прохождения кривой вблизи заданных значений и условие минимальности "изгибания" функции. Чем больше весовые коэффициенты p_k, тем больший вклад в функционал вносят интерполяционные условия, тем ближе к заданным значениям проходит сглаживающая функция. В пределе $p_k \to \infty$ получаем задачу на среднеквадратичное приближение (п.3.3.1), в противоположном предельном случае $p_k \to 0$ получаем задачу на сплайн-интерполяцию.

Решением вариационной задачи (1.77) является кубический сплайн, т. е. функция, удовлетворяющая требованиям (1.57), (1.58) и (1.60) (но без требования интерполяции (1.59)).

◂ Пусть $u_0 \in W_2^2[a,b]$ – решение задачи. Построим сплайн $g(x)$ такой, что

$$g(x_k) = u_0(x_k) \quad \text{для} \quad k = 0, 1, \ldots, n.$$

Тогда второе слагаемое в (1.77) одинаково для функций $g(x)$ и $u_0(x)$, и при этом $\tilde{\Phi}[u] \le \tilde{\Phi}[g]$, так что

$$\int_a^b dx\,[u_0''(x)]^2 \le \int_a^b dx\,[g''(x)]^2. \tag{1.78}$$

Но, как было показано в п.1.5.3, $g(x)$ – единственная функция, дающая при интерполяции $u_0(x)$ минимум выражения $\int_a^b dx\,[u'']^2$. Поэтому $u_0 \equiv g$. ▸

Итак, минимум функционала $\tilde{\Phi}[u]$ достаточно искать в классе кубических сплайнов. Так как кубический сплайн однозначно определяется множеством его значений $\{\mu_k\}_0^n$, принимаемых в узлах $\{x_k\}_0^n$, то минимизация $\tilde{\Phi}[u]$ сводится к нахождению минимума функции от переменных $\mu_0, \mu_1, \ldots, \mu_n$.

Мы уже знаем, что $g''(x)$ – кусочно-линейная функция (1.61). Подставляя

это в (1.77), получаем

$$\int\limits_{x_{k-1}}^{x_k} dx\big[u''(x)\big]^2 = \int\limits_{x_{k-1}}^{x_k} dx\left[m_{k-1}\frac{x_k-x}{h_k} + m_k\frac{x-x_{k-1}}{h_k}\right]^2 = \frac{h_k}{3}\{m_{k-1}^2 + m_{k-1}m_k + m_k^2\}.$$

Тогда весь интеграл представляется в виде

$$\int\limits_a^b dx\big[u''(x)\big]^2 = \sum_{k=1}^n \frac{h_k}{3}\{m_{k-1}^2 + m_{k-1}m_k + m_k^2\} =$$
$$= \ldots = \sum_{k=1}^n m_k\left(\frac{h_k}{6}m_{k-1} + \frac{h_k+h_{k+1}}{3}m_k + \frac{h_{k+1}}{6}m_{k+1}\right) = \sum(Am)_k m_k \equiv (Am, m),$$

где под m мы понимаем n-мерный вектор, составленный из коэффициентов m_i, и для функционала получаем

$$\tilde{\Phi}[g] = (Am, m) + \sum_{k=0}^n p_k\left(\mu_k - \tilde{f}_k\right)^2.$$

Из (1.67) $Am = H\mu$, так что m линейно выражается через вектор μ, и поэтому $\tilde{\Phi}[g]$ есть положительно определенная квадратичная форма от μ. Ее экстремумом может быть только минимум, необходимым условием которого является

$$\frac{\partial\tilde{\Phi}}{\partial\mu_s} \equiv \frac{\partial}{\partial\mu_s}(Am, m) + 2p_s\left(\mu_s - \tilde{f}_s\right) = 0 \quad \text{для } s = 0, 1, \ldots, n. \tag{1.79}$$

Но матрицы A и H не зависят от μ, и $A = A^+$, поэтому в силу (1.67) $Am = H\mu$ и

$$\frac{\partial(Am, m)}{\partial\mu_s} = 2\left(\frac{\partial Am}{\partial\mu_s}, m\right) = 2\left(\frac{\partial H\mu}{\partial\mu_s}, m\right) = 2\left(\frac{\partial\mu}{\partial\mu_s}, H^+m\right) = 2\left(H^+m\right)_s.$$

Тогда в векторной форме условие минимума (1.79) принимает вид

$$H^+m + P\mu = P\tilde{f}, \tag{1.80}$$

где

$$\tilde{f} = \begin{pmatrix} \tilde{f}_0 \\ \vdots \\ \tilde{f}_n \end{pmatrix}; \quad P = \begin{pmatrix} p_0 & 0 & \ldots & 0 \\ 0 & p_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_n \end{pmatrix}. \tag{1.81}$$

Умножая (1.80) слева на HP^{-1} и опять учитывая что $H\mu = Am$, получим

$$\left(A + HP^{-1}H^+\right)m = H\tilde{f}. \tag{1.82}$$

Можно показать, что матрица системы (1.82) пятидиагональна, симметрична и со строгим диагональным преобладанием ◂…▸. Тогда, так же как и в случае

кубических сплайнов, из теоремы Гершгорина о локализации собственных значений следует, что она положительно определена и неособенна. Система (1.82) решается также методом прогонки, модифицированным для пятидиагональных матриц. После того, как вектор m определен, необходимо найти вектор сеточных значений сглаживающего сплайна по формуле, которая следует из (1.80): $\mu = \tilde{f} - P^{-1}H^{+}m$. Затем по формуле (1.64) восстановить сплайн $g(x)$.

1.5.7 Гладкие восполнения*

Рассмотрим другую идею построения гладкого восполнения сеточных функций, которая несколько отличается от методов интерполяции и теории сплайн-функций, однако тоже весьма эффективна.

Опишем очень кратко процесс построения интерполяционной функции произвольного класса гладкости C^p. Пусть $x_1 < x_2 < \ldots < x_n$ – некоторая фиксированная сетка, в узлах которой известны значения функции $f_1, f_2, \ldots, f_n$, и пусть n достаточно велико. Фиксируем $p \ll n$.

Вначале построим функцию класса гладкости C^p на отрезке $[x_1, x_2]$ следующим образом. Обозначим через $P_0(x) \in \overline{\Pi}_p$ интерполяционный многочлен Лагранжа, интерполирующий функцию в узлах $x_1, x_2, \ldots, x_{p+1}$, а через $P_1(x) \in \overline{\Pi}_p$ – многочлен Лагранжа, интерполирующий f в узлах $x_2, x_3, \ldots, x_{p+2}$. Построим многочлен $Q_1(x) \in \overline{\Pi}_{2p+1}$, который удовлетворяет $2p+2$ условиям

$$\left.\frac{d^k Q_1}{dx^k}\right|_{x=x_1} = \left.\frac{d^k P_0}{dx^k}\right|_{x=x_1}, \quad \left.\frac{d^k Q_1}{dx^k}\right|_{x=x_2} = \left.\frac{d^k P_1}{dx^k}\right|_{x=x_2}, \quad k = 0, 1, \ldots, p.$$

Понятно, что такой многочлен существует, и $2p+2$ записанными условиями определен однозначно.

К примеру, в случае $p = 1$ это будет многочлен третьей степени, который можно записать в виде

$$Q_1(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)^2 + a_3(x - x_1)^3,$$

где коэффициенты выражаются через разделенные разности (1.18) ◂…✩▸

$$a_0 = f_1, \quad a_1 = \frac{f_2 - f_1}{x_2 - x_1} \equiv f_{21}, \quad a_2 = -\frac{x_3 - x_1}{x_2 - x_1} f_{321}, \quad a_3 = \frac{x_3 - x_1}{(x_2 - x_1)^2} f_{321}.$$

Положим теперь что интерполяционная функция $g(x)$ на отрезке (x_1, x_2) равна $Q_1(x)$. На отрезке (x_2, x_3) она строится совершенно аналогично, если за начало отсчета принять точку x_2. Такой процесс построения $g(x)$ можно

продолжить до интервала (x_{n-p-1}, x_{n-p}). Построенный интерполянт $g(x)$ есть кусочно-полиномиальная функция класса гладкости C^p; степень полиномов на интервалах сетки равна $2p+1$.

Если положить $p = 2$, что соответствует гладкости кубических сплайнов, то мы будем иметь дело с многочленами пятой степени. Степень многочленов выше на два, но размерность системы уравнений, которую нужно решать для построения каждого кусочка $g(x)$, всего лишь 6, в отличие от задачи сплайн-интерполяции, где число уравнений пропорционально полному числу узлов сетки.

Построение интерполяционных функций многих переменных мало чем отличается от одномерного случая, к которому все и сводится.

Гладкие восполнения обладают хорошими аппроксимационными свойствами. Именно, при интерполировании функции многих переменных $f \in C^q$ восполнением $g(x)$ класса C^p, где $p \geq q$, для самой функции и ее производных справедливы равенства ◂…▸

$$\left|D^k(f-g)\right| \leq C(p)h^{q-|k|}\sup_x \max_{|a|=q}\left|D^a f(x)\right|,$$

где мультииндекс $k = (k_1, k_2, \ldots, k_r)$ таков, что $C(p)$ зависит только от p и не зависит от шага сетки h и функции f.

О задачах аппроксимации

Проблема интерполяции величин, заданных на дискретном множестве точек, на всю область определения функции непрерывного аргумента тесно связана с построением вариационно-разностных схем и непрерывного представления решений разностных задач. В самом деле, при решении различных дифференциальных или интегральных уравнений, как правило, осуществляется процесс дискретизации оператора и решения задачи с помощью подходящих методов проектирования. При этом решение разностной задачи обычно представляет собой приближенное решение исходной задачи на дискретном множестве точек.

Предположим, что разностная задача решена, и мы располагаем информацией о приближенном решении этой задачи. Дальнейшее связано с интерполяцией полученных данных на всю область определения решения исходной задачи. Естественно, что при такой интерполяции должны быть соблюдены

некоторые условия, а именно: если решение разностного уравнения получено с определенной степенью точности, то порядок интерполяции данных должен согласовываться с порядком аппроксимации разностного уравнения и быть не ниже последнего. Если мы располагаем дополнительной информацией о погрешностях приближенного решения, то интерполяцию приближенного решения можно осуществить не по точным данным, а с учетом возможных погрешностей в узлах. Тогда априорная информация о гладкости решения в некоторых случаях позволит даже уточнить приближенное решение задачи, полученное с помощью тех или иных разностных методов. Конечно, проблема интерполяции данных имеет и самостоятельное значение.

Алгоритмы интерполяции функций по точным данным, определенным на дискретном множестве точек, как правило, основаны на использовании интерполяционных многочленов Лагранжа. При этом относительно интерполируемой функции $f(x)$ вводится априорное предположение о том, что она обладает производными до некоторого порядка.

Другая, близкая к проблеме интерполяции, задача возникает в том случае, когда значения заданной функции $f(x)$ известны в узловых точках x_k не точно, а с некоторой погрешностью, максимальная величина которой для каждой точки задается в качестве априорной информации. В этом случае задача состоит в построении такой кривой, которая бы в известном смысле наилучшим образом аппроксимировала функцию, заданную со случайными погрешностями в узловых точках. Такая задача обычно решается на основе метода наименьших квадратов (см. п.3.3.1).

Задача сплайн-интерполяции, как мы видели, относится (очевидно) к классу интерполяционных задач, однако ее естественное обобщение со сглаживанием больше похоже на наименьшие квадраты, и занимает промежуточное положение. Эти задачи также переформулируются на вариационном языке, и потому естественно комбинируются с соответствующими методами решения дифференциальных, интегральных и интегро-дифференциальных уравнений.

Сплайны, в частности так называемые сплайны Безье[30] (Bézier splines) широко используются во всевозможных приложениях, связанных с векторной графикой (Adobe Illustrator/Flash/Photoshop, PostScript, шрифты, игры etc.; дизайн самолетов, машин). Безье сплайн – кривая, получаемая при интерполяции полиномами в форме Безье (Bézier curves). Полиномами Безье называются по-

[30] О применении сплайнов Безье см. например Bezier and B-spline Technology.

линомы Бернштейна (Bernstein polynomials)

$$b_{\nu,n}(x) = C_n^\nu x^\nu (1-x)^{n-\nu}, \quad \nu = 1, \dots, n, \quad n \in \mathbb{N}, \tag{1.83}$$

ограниченные на промежуток $[0, 1]$.

Литература к теории интерполяции

- Stoer J., Bulirsch R. *Introduction to numerical analysis*, 2002, [1]. Весьма основательная книжка, написана на хорошем уровне, но вполне еще для физиков. Все доказывается, хотя не очень подробно. Структура изложения глав об аппроксимации здесь в значительной мере соответствует подходу в этой книжке. Нет раздела по уравнениям мат физики. Перевода на русский, похоже, пока нет.

Дополнительно:

1. Н.С. Бахвалов, Н.П. Жидков, Г.М, Кобельков, *Численные методы* [2]. Классический учебник по численным методам. Довольно подробный, много слов. Весьма ценный общими рассуждениями и конкретными рекомендациями авторов – ветеранов численного счета.

2. Н.Н. Калиткин, *Численные методы*, [3]. Еще один классический учебник. Содержит материал от интерполяции до интегральных уравнений.

3. Р.В. Хемминг, *Численные методы* [4].

4. Б.П. Демидович, И.А. Марон, Э.З. Шувалова, *Численные методы анализа*, [5].

Глава 2

Дифференциальные уравнения

Numerical Differential Equations

В этой главе мы рассмотрим различные методы решения дифференциальных уравнений, обыкновенных и в частных производных. Способы решения обыкновенных дифференциальных уравнений, которые излагаются здесь, получаются дискретизацией исходной задачи и выводятся из формул интерполяции или разложением в ряд Тейлора. Из множества различных методов решения уравнений в частных производных мы кратко рассмотрим конечно-разностные методы, которые получаются в результате естественного обобщения на многомерный случай аналогичных методов решения одномерных краевых задач.

2.1 Обыкновенные дифференциальные уравнения. Введение

Обыкновенные дифференциальные уравнения, (ordinary differential equations, ODEs) – уравнения для функции одной переменной, в которое входят ее производные.

Численные методы предназначены для нахождения *частных* решений дифференциальных уравнений, а не общих. Поэтому для корректной постановки задачи уравнение необходимо дополнить начальными или граничными условиями.

Задача Коши (Cauchy problem, initial value problem) – дифференциальное уравнение с дополнительными условиями на решение, заданными в одной точке. Если интерпретировать независимую переменную как время в динамиче-

ской задаче, то это будут начальные условия. Для уравнения первого порядка это

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0. \tag{2.1}$$

Если дополнительное условие задается в нескольких точках, то это *краевая задача (boundary value problem)*. Например, для уравнения второго порядка

$$y'' = f(x, y, y'), \quad y(a) = 0, \quad y'(b) = \alpha y(b).$$

При численном решении задач Коши и краевых задач используются существенно различные методы.

При решении задачи Коши применяются:

- **Одношаговые методы**, в которых для нахождения следующей точки на кривой $y = y(x)$ требуется информация лишь об одном предыдущем шаге. К таким методам относятся метод Эйлера и методы Рунге-Кутта.
- **Многошаговые методы**, в которых для отыскания следующей точки кривой $y = y(x)$ требуется информация более чем об одной из предыдущих точек. Их усовершенствованной разновидностью являются методы прогноза и коррекции, к числу которых относятся методы Милна, Адамса-Башфорта, и пр.

При решении краевых задач используются **методы стрельбы**, в которых задачу пытаются свести к задаче Коши, и **конечно-разностные методы**.

Погрешности численного счета

Всякий метод численного решения дифференциальных уравнений дает некоторый компромисс между точностью решения и устойчивостью.

Потери точности определяется следующими источниками погрешностей, связанных с численной аппроксимацией:

1. Ошибка входных данных. Так как от метода она не зависит, мы ее рассматривать не будем.
2. Погрешность округления, вследствие ограниченной разрядности представления чисел.
3. Погрешность усечения, возникающая из-за того, что для аппроксимации функций используются не бесконечные ряды, а их конечные частные суммы.

Погрешностью распространения называют погрешность, являющуюся результатом накопления всех погрешностей, появившихся на предыдущих этапах вычислений.

Если метод обеспечивает малость конечной ошибки распространения при условии малости ошибки во входных данных, то говорят, что он устойчив (numerically stable). Мы далее не будем обсуждать устойчивость, а здесь заметим лишь, что ни один алгоритм не может быть пригоден на все случаи жизни, а при численном решении дифференциальных уравнений всегда следует обращать внимание на оценку ошибки на разных этапах и проверку решения.

2.2 Задача Коши. Одношаговые методы

Мы здесь рассмотрим для простоты уравнения первого порядка. Уравнения более высокого порядка можно представить как уравнения первого порядка для векторной функции, и соответствующие обобщения довольно прямолинейны.

2.2.1 Методы Эйлера и Адамса

Решаем задачу Коши (2.1)

$$y'(x) = f(x, y), \quad y(a) = y_a, \quad x \in [a, b].$$

Разложим $y(x)$ вблизи некоторой точки $x = x_i \in [a, b]$ в ряд Тейлора:

$$y(x_i + h) = y(x_i) + hy'(x_i) + \frac{h^2}{2!}y''(x_i) + \dots \tag{2.2}$$

I. Если ограничиться двумя слагаемыми, то получим

$$y(x_i + h) = y(x_i) + hy'(x_i) = y(x_i) + hf(x_i, y(x_i)).$$

Тогда, разбив весь интервал изменений аргумента на n промежутков с шагом h набором узлов $\{x_i\}_1^{n-1}$, получим соотношение для нахождения $y(x)$ в каждом последующем узле

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = 1, 2, \dots, n-1, \tag{2.3}$$

которое определяет *метод Эйлера (Euler method)* (его также называют методом ломаных). Мы, таким образом, заменили дифференциальное уравнение разностным. Ошибка на шаге – порядка отброшенных членов в ряде Тейлора h^2y''. Недостаток заключается в очевидном накоплении систематической погрешности (см. рисунок).

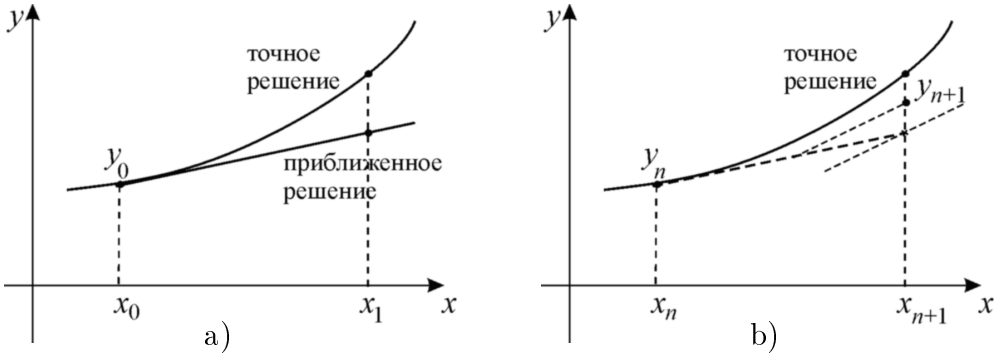


Рис. 2.1: Ошибка в методе Эйлера (a) растет с каждым шагом. В модифицированном методе (b) главная часть ее устраняется благодаря тому, что берется не значение производной в исходной точке, а полусумма значений на концах шага.

II. Главный член отброшенного ряда есть $h^2y''(x_i)/2$, и в первом приближении из $y_i' = \frac{y_{i+1}-y_i}{h}$ имеем $y_i'' = \frac{y_{i+1}'-y_i'}{h}$. Подставляя производные из уравнения, получаем

$$y_{i+1} = y_i + hf(x_i, y_i) + \frac{h^2}{2}\frac{1}{h}\big(f(x_i, y_i) - f(x_{i+1}, y_{i+1})\big).$$

Заменяя теперь y_{i+1} в аргументе функции f на $y_i + hf(x_i, y_i)$, мы пренебрежем слагаемыми того же порядка малости по h, что уже отброшенные, и получим формулу *модифицированного метода Эйлера*

$$y_{i+1} = y_i + \frac{h}{2}\Big[f(x_i, y_i) + f\big(x_i + h, y_i + hf(x_i, y_i)\big)\Big], \quad i = 1, 2, \ldots, n-1. \tag{2.4}$$

Погрешность метода порядка h^3y'''.

III. Другое разбиение промежутка $[x_i, x_{i+1}]$ дает метод *Адамса*:

$$y_{i+1} = y_i + hf\Big(x_i + \frac{h}{2}, y_i + \frac{h}{2}f(x_i, y_i)\Big), \quad i = 1, 2, \ldots, n-1. \tag{2.5}$$

Погрешность – порядка h^3y'''. Эта формула проверяется, так же как и предыдущая, непосредственно – разложением правой части в ряд Тейлора по обоим переменным вблизи (x_i, y_i) до слагаемых порядка h^2:

$$y_{i+1} \approx y_i + h\big[f + \tfrac{h}{2}f_x + \tfrac{h}{2}f_y y_x\big]_{x_i,y_i} = y_i + hy' + \frac{h^2}{2}\big[f_x + f_y y'\big]_{x_i,y_i} = y_i + hy_i' + \frac{h^2}{2}y_i''$$

Повышение точности достигается сохранением большего числа слагаемых в ряде Тейлора.

2.2.2 Методы Рунге-Кутта (Runge–Kutta)

Чтобы удержать в ряде Тейлора n-ю производную, ее надо как-то вычислить. Например, чтобы вычислить третью производную, необходимо иметь значения второй производной, по меньшей мере, в двух точках, т.е. надо знать наклон кривой и в какой-то промежуточной точке интервала (x_i, x_{i+1}). Чем выше порядок сохраняемой производной, тем в большем количестве точек на промежутке (x_i, x_{i+1}) нужно вычислять правую часть уравнения, f. *Порядком метода* называется порядок наивысшей из оставляемых производных в ряде Тейлора.

Так как все высшие производные так или иначе будут линейно выражаться через значения f в разных точках, то их учет сводится к некоторому сложному усреднению наклона $y(x)$, посчитанного в этих точках:

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i. \tag{2.6}$$

Семейство методов Рунге-Кутта[1] дает набор формул для расчета координат внутренних точек и выбора относительных весов в этих точках в рекуррентном виде: для s шагов

$$\begin{aligned}
k_1 &= f(x_n, y_n);\\
k_2 &= f(x_n + c_2 h, y_n + a_{21} h k_1);\\
k_3 &= f(x_n + c_3 h, y_n + a_{31} h k_1 + a_{32} h k_2);\\
&\vdots\\
k_s &= f(x_n + c_s h, y_n + a_{s1} h k_1 + a_{s2} h k_2 + \cdots + a_{s,s-1} h k_{s-1}),\\
&\text{где} \quad \sum_{j=1}^{i-1} a_{ij} = c_i.
\end{aligned} \tag{2.7}$$

Конкретный метод семейства характеризуется числом этапов (шагов, stages[2]) $s \in \mathbb{N}$, и коэффициентами $\{a_{ij}\}_{1 \le j < i \le s}$ и $\{b_i\}_{i=1}^{s}$. Они обычно вписываются в мне-

[1] По именам немецких математиков Carl Runge (1856—1927) и Martin Wilhelm Kutta (1867–1944).

[2] Это число шагов, или этапов для нахождения y_{n+1} при известном y_n. Для того, чтобы не возникало путаницы с шагами h разбиения отрезка $[a, b]$, а также с многошаговыми методами, о которых речь будет позже, а шаги имеют другой смысл, мы будет использовать здесь слово “этапы”, а методы называть, к примеру, четырехэтапным – хоть это и звучит не очень красиво.

моническую табличку (таблица Бутчера, Butcher tableau)

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & & \ddots & & \\ c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} & \\ \hline & b_1 & b_2 & \dots & b_{s-1} & b_s \end{array}$$

Рассмотренным выше методам Эйлера, модифицированному методу Эйлера и Адамса соответствуют таблички

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

Наборы коэффициентов получаются следующим образом. Исходя из дифференциального уравнения, мы можем разложить $y(x)$ в ряд Тейлора по h вблизи $x = x_n$:

$$y_{n+1} = y_n + hf_n + \frac{h^2}{2}\Big(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\Big)f + \frac{h^3}{3!}\Big(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\Big)^2 f + \dots.$$

Так же мы можем разложить y_{n+1} из (2.6,2.7). В общем случае формулы получаются весьма громоздкие, из-за того, что k_i с ростом i зависят от h все более сложным образом. Приравнивая эти два разложения тождественно до членов некоторого порядка p включительно, получим равенства для всех коэффициентов при выражениях вида $h^\alpha f^\beta f_{xxy} f_{yy}$, и в итоге – некоторую нелинейную систему уравнений на коэффициенты a_{ij}, b_k.

Заметим, что число неизвестных коэффициентов определяется числом этапов метода: для $s = 2$ получим три свободных коэффициента (a_{21}, b_1 и b_2), для $s = 3$ их 6, для четырехэтапного метода 10, и так далее. Между тем число уравнений (и их алгебраическая сложность) резко растет с повышением p. Для методов третьего порядка получается 4 уравнения, и потому можно построить трехэтапные методы третьего порядка, и два параметра остаются свободными; а для методов 4 порядка имеем 8 уравнений – можно построить четырехэтапный метод с двумя свободными параметрами. При повышении порядка число уравнений растет быстрее, и метод пятого порядка должен иметь по крайней мере 6 этапов, шестого порядка – 7, дальше еще больше.

Фиксируя удобным образом имеющиеся свободные параметры (например полагая их равными нулю), получим один из возможных наборов a_{ij}, b_l, c_k.

Необходимо отметить, что точность метода порядка s выше, чем точность метода более низкого порядка, только если решение *имеет* производную порядка s. В противном случае повышение порядка метода повышает громоздкость и сложность, но отнюдь не повышает точность решения.

2.2.3 RK4

В классическом методе 4го порядка, который вследствие распространенности называют просто методом Рунге-Кутта (RK4 or "*the* Runge–Kutta method"), удерживаются члены ряда Тейлора, включая $\sim h^4$. Его коэффициенты образуют табличку

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

Ошибка на шаге имеет порядок h^5, и значит полная ошибка $\sim h^4$. При понижении порядка ухудшается точность, а при повышении ухудшается точность подсчета высших производных, которые используются (см. численное дифференцирование).

Рекуррентные формулы в явном виде☆

$$y_{i+1} = y_i + h\frac{k_1 + 2k_2 + 2k_3 + k_4}{6}; \qquad \begin{aligned} k_1 &= f(x_i, y_i)\,; \\ k_2 &= f\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_1}{2}\Big); \\ k_3 &= f\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_2}{2}\Big); \\ k_4 &= f(x_i + h, y_i + hk_3)\,. \end{aligned} \tag{2.8}$$

А это еще одна разновидность метода Рунге-Кутта 4 порядка:

$$y_{i+1} = y_i + h\frac{k_1 + 4k_3 + k_4}{6}; \qquad \begin{aligned} k_1 &= f(x_i, y_i)\,; \\ k_2 &= f\Big(x_i + \tfrac{h}{4}, y_i + h\tfrac{k_1}{4}\Big); \\ k_3 &= f\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_2}{2}\Big); \\ k_4 &= f(x_i + h, y_i + h(k_1 - 2k_2 + 2k_3))\,; \end{aligned}$$

0				
1/4	1/4			
1/2	0	1/2		
1	1	−2	2	
	1/6	0	2/3	1/6

Погрешности обоих формул $\Delta y_{i+1} = \frac{h^5}{120} y^{(5)}(\xi)$.

Пример.

Рассмотрим задачу Коши $y' = 2x^2 + 2y$, $y(0) = 1$, $x \in [0, 1]$.
Точное ее решение $y = \frac{3}{2}e^{2x} - x^2 - x - \frac{1}{2}$.
Численное решение проведено различными методами с шагом 0.1:

x_i	Эйлер	Мод. Эйлер	Рунге-Кутта	Точное
0	1.0000	1.0000	1.0000	1.0000
0.1	1.2000	1.2210	1.2221	1.2221
0.5	2.5569	2.7680	2.8274	2.8274
1	7.0472	8.0032	8.5834	8.5836

2.2.4 Решение систем ОДУ*

Пусть есть уравнение $y'' = g(x, y, y')$. Обозначив $y' = z$, получим, добавив начальные условия, задачу Коши

$$\begin{cases} z' = g(x, y, z) & y(x_0) = y_0 \\ y' = f(x, y, z); & z(x_0) = z_0. \end{cases} \tag{2.9}$$

где $f(x, y, z) = z$. Понятно что таким образом любые уравнения, более высокого порядка чем первый, сводятся к системам уравнений первого порядка. Последние тогда решаются обычными методами, модифицированными для переменных-векторов, в нашем случае $\tilde{y}(x) = (y(x), z(x))$.

Расчетные формулы метода Рунге-Кутта RK4 для решения (2.9) получаются весьма прямолинейным обобщением:

$$y_{i+1} = y_i + h\frac{k_1 + 2k_2 + 2k_3 + k_4}{6}; \qquad z_{i+1} = z_i + h\frac{l_1 + 2l_2 + 2l_3 + l_4}{6}; \tag{2.10}$$

$$\begin{aligned} k_1 &= f(x_i, y_i, z_i)\,; & l_1 &= g(x_i, y_i, z_i)\,; \\ k_2 &= f\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_1}{2}, z_i + h\tfrac{l_1}{2}\Big)\,; & l_2 &= g\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_1}{2}, z_i + h\tfrac{l_1}{2}\Big)\,; \\ k_3 &= f\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_2}{2}, z_i + h\tfrac{l_2}{2}\Big)\,; & l_3 &= g\Big(x_i + \tfrac{h}{2}, y_i + h\tfrac{k_2}{2}, z_i + h\tfrac{l_2}{2}\Big)\,; \\ k_4 &= f(x_i + h, y_i + hk_3, z_i + hl_3)\,; & l_4 &= g(x_i + h, y_i + hk_3, z_i + hl_3)\,. \end{aligned}$$

По сути, то же самое, что было для уравнения первого порядка.

2.2.5 Формула Рунге для локальной погрешности

Если порядок метода p, то при вычислении с шагом h ошибка на шаге будет порядка h^{p+1}:

$$\Delta_h y = y_{true} - y_h = Ch^{p+1}, \tag{2.11}$$

где постоянная C зависит от метода и уравнения. Тогда для шага kh, где k может быть произвольным положительным множителем, получаем ошибку $Ch^{p+1} \cdot k^{p+1}$, и $|y_h - y_{kh}| = Ch^{p+1}|1 - k^{p+1}|$. Таким образом, сравнивая результаты вычислений с разным шагом, мы получаем оценку для локальной погрешности. При уменьшении шага вдвое (при $k=1/2$) получаем

$$\Delta_h y = \frac{|y_h - y_{h/2}|}{2^{p+1} - 1} 2^{p+1} = \frac{|y_h - y_{h/2}|}{1 - 2^{-(p+1)}} \tag{2.12}$$

Это (первая) формула Рунге, для апостериорной[3] оценки погрешности на шаге, или локальной погрешности.

При большом шаге h велика локальная погрешность на шаге, а если шаг мал, то увеличивается время счета и возрастает ошибка накопления. Поэтому к выбору шага следует относиться осторожно.

Если для оценки погрешности пользоваться формулой Рунге, то дважды вычисляется y_j и объем вычислений возрастает втрое (так как на каждый шаг h добавляется два шага по $h/2$). Зато таким образом можно оценивать ошибку на каждом шаге, и если она слишком велика, уменьшать шаг и добиваться нужной точности. Можно проверку делать не на каждом шаге, а время от времени. Другой способ оценки ошибки на шаге, использующийся в разновидностях методов Рунге-Кутта – совмещать в вычислении на шаге методы порядка p и $p-1$.

Однако оценка и регулировка ошибки на шаге в целом более естественно производится в методах прогноза и коррекции.

2.2.6 Общая характеристика одношаговых методов

1. В основе методов лежит разложение функций в ряд Тейлора.

2. Не требуется вычисление производных через конечные разности – вычисляется лишь функция, стоящая в правой части уравнения. Для $RK4$, четвертого порядка, она вычисляется 4 раза за шаг.

3. Для получения информации в новой точке, надо иметь данные лишь в одной предыдущей точке. Это свойство принято называть свойством "самостартования". Следствие – относительная простота программной реализации.

4. Свойство "самостартования" позволяет гибко управлять шагом.

[3] Апостериорная: ошибка оценивается *после* счета y_h и $y_{h/2}$.

2.3 Задача Коши. Методы прогноза и коррекции

2.3.1 Многошаговые методы

Напомним, что мы рассматриваем решение уравнения вида

$$y' = f(x, y) \quad \text{с начальным условием} \quad y(x_0) = y_0.$$

Для того, чтобы повысить точность и вычислить дальнейшие производные искомой функции, необходимо знать ее значения в большем количестве точек. В одношаговых методах для вычисления $y_{i+1}^{(n)}$ использовалось $(n-1)$ дополнительных точек на интервале (x_i, x_{i+1}). В *многошаговых методах* для этого используются результаты счета на предыдущих шагах – в точках $x_i, x_{i-1}, x_{i-2}, \ldots$ Поэтому для старта метода предварительно необходимо вычислить значения y в достаточном количестве первых точек (с помощью одношаговых методов). Если для счета y'_{n+1} используется значения f в s точках, то будем иметь s-шаговый метод.

Для вывода формул используются различные приближенные решения интегрального уравнения

$$y_{i+k} = y_{i-p} + \int_{x_{i-p}}^{x_{i+k}} f(x, y)dx. \tag{2.13}$$

Пример: первая формула Милна

Рассмотрим уравнение (2.13) с $k = 1$, $p = 3$.

$$y_{i+1} = y_{i-3} + \int_{x_{i-3}}^{x_{i+1}} f(x, y)dx. \tag{2.14}$$

Считаем, что значения y, а значит и $f(x, y)$, на предыдущих шагах $i, i-1, \ldots$ известны, и нам нужно получить приближенное значение y_{i+1}. Для этого заменим $f(x, y)$ под интегралом на интерполяционный многочлен, который интерполирует f в четырех точках $x_{i-3}, x_{i-2}, x_{i-1}, x_i$ (таким образом, имеем четырехшаговый метод). Полагаем, что шаг фиксирован $x_i = x_0 + ih$.

Используем интерполяционную формулу Ньютона (1.23) для интерполяции назад. Она получается из формулы для интерполяции вперед заменой $h \to (-h)$ и правых разностей (разностей вперед) левыми (разностями назад), в которых

индексы $i+k$ заменены на $i-k$:

$$P_+^{(3)} = f_i + \frac{\Delta f_0}{h}(x-x_i) + \frac{\Delta^2 f_0}{2h^2}(x-x_i)(x-x_{i+1}) + \frac{\Delta^3 f_0}{3!\,h^3}(x-x_i)(x-x_{i+1})(x-x_{i+2});$$
$$P_-^{(3)} = f_0 - \frac{\Delta_- f_0}{h}(x-x_i) + \frac{\Delta_-^2 f_0}{2h^2}(x-x_i)(x-x_{i-1}) - \frac{\Delta_-^3 f_0}{3!\,h^3}(x-x_i)(x-x_{i-1})(x-x_{i-2}). \tag{2.15}$$

Здесь нижний индекс минус у Δ_- обозначает, что это левые разности

$$\Delta_- f_0 = f_{i-1} - f_i; \quad \Delta_-^2 f_0 = \Delta_-(f_{i-1} - f_i) = f_{i-2} - 2f_{i-1} + f_i; \quad \dots$$

Подставляя (2.15) в интеграл (2.14), видим, что первые два слагаемых дают

$$\int_{x_i-3h}^{x_i+h} dx\,(x-x_i) = \int_{-3h}^{h} d\xi\,\xi = -4h^2;$$
$$\int_{x_i-3h}^{x_i+h} dx\,(x-x_i)(x-x_{i-1}) = \int_{-3h}^{h} d\xi\,\xi(\xi+h) = \frac{16}{3}h^3,$$

а третье слагаемое при интегрировании дает ноль, так как при равноотстоящих узлах оно представляет собой нечетную функцию относительно середины промежутка $[x_i - 3h, x_i + h]$.

Собирая все вместе и собирая подобные, получаем

$$y_{i+1} - y_{i-3} \approx \int_{x_i-3h}^{x_i+h} dx\,P_-^{(3)} = f_i \cdot 4h - \frac{f_{i-1}-f_i}{h}\cdot(-4h^2) + \frac{f_{i-2}-2f_{i-1}+f_i}{2h^2}\cdot\frac{16}{3}h^3;$$
$$\Rightarrow \qquad y_{i+1} \approx y_{i-3} + \frac{4h}{3}\left(2f_i - f_{i-1} + 2f_{i-2}\right) \tag{2.16}$$

Это *явная (explicit)* формула, т.к. y_{i+1} явно выражается через y_i, y_{i-1}, y_{i-2}. *Неявную (implicit)* формулу мы получим, если проинтегрируем какой-либо интерполяционный полином, в число узлов интерполяции которого входит и x_{i+1}. Например, так называемую вторую формулу Милна получим аналогично первой, интегрированием от x_{i-1} до x_{i+1} по формуле Симпсона[☆](или, что то же самое, интегрированием в этих пределах интерполяционного многочлена через три точки x_{i-1}, x_i, x_{i+1}):

$$y_{i+1} = y_{i-1} + \frac{h}{3}\left(f_{i+1} + 4f_i + f_{i-1}\right). \tag{2.17}$$

Тогда y_{i+1} получим, решая уравнение.

Разные явные и неявные формулы можно вывести, например, выбирая разные k и p в (2.13) и разные наборы узлов интерполяции. Так, метод Адамса-Башфорта (Adams-Bashforth) порядка s получим, полагая $p = 0$, $k = 1$, и заменяя f интерполяционным многочленом через точки $x_i, \dots, x_{i-s+1}$ – это явные

методы. Метод Адамса-Мултона (Adams-Moulton) порядка s получим, полагая $p = 0$, $k = 1$, и заменяя f интерполяционным многочленом через точки $x_{i+1}, \ldots, x_{i-s+2}$ – это неявные методы.

Общий и очевидный недостаток неявных методов в том, что для нахождения y_{i+1} на каждом шаге необходимо решать уравнение относительно этой величины, так как в правой части (см. (2.16)) стоит $f_{i+1} = f(x_{i+1}, y_{i+1})$. Однако неявные формулы оказываются более устойчивыми, а потому более универсальными.

Большинство практически используемых многошаговых методов принадлежат семейству линейных методов (linear multistep methods), которые описываются формулой

$$\alpha_0 y_i + \alpha_1 y_{i+1} + \ldots + \alpha_k y_{i+k} = h\left[\beta_0 f(x_i, y_i) + \beta_1 f(x_{i+1}, y_{i+1}) + \ldots + \beta_k f(x_{i+k}, y_{i+k})\right],$$

причем как правило ограничиваются $k \leq 3$.

2.3.2 Методы прогноза и коррекции

В *методах прогноза и коррекции (predictor-corrector methods)* оказывается возможным совместить простоту и скорость явных многошаговых методов с устойчивостью неявных. Идея заключается в том, что выбираются *два* разных многошаговых метода – один явный (формула прогноза) а второй неявный (формула коррекции).

На каждом шаге $y_i \to y_{i+1}$ вычисления производятся в следующем порядке:

1. По явной формуле прогноза вычисляем y_{i+1} в нулевом приближении – $y_{i+1}^{(0)}$;
2. Подставляя в $f(x, y)$, получаем правую часть уравнения в нулевом приближении: $f_{i+1}^{(0)} = f(x_{i+1}, y_{i+1}^{(0)})$;
3. Приближенное значение $f_{i+1}^{(0)}$ подставляем в правую часть формулы коррекции и получаем следующее приближение $y_{i+1}^{(1)}$;
4. Разница $\left|y_{i+1}^{(1)} - y_{i+1}^{(0)}\right|$ дает оценку ошибки на шаге.

Можно превратить эту схему в итерационную: если ошибка на шаге не достаточно мала, то уточненное после коррекции значение $f_{i+1}^{(1)}$ можем опять вставить в правую часть формулы коррекции, и повторять цикл (2-4), пока ошибка не станет достаточно маленькой. Однако следует отметить, что такая схема малоэффективна: сколько бы итераций не было сделано, решение будет сходится к

точному решению дискретизированной задачи, задаваемой формулой коррекции – а не к точному решению исходного уравнения. Поэтому в оптимальном режиме работы метода прогноза и коррекции каждый шаг заканчивается после первой коррекции.

Проиллюстрируем методы прогноза и коррекции на паре примеров.

2.3.2.1 Метод Милна (Milne's method).

Для прогноза используем явную формулу, выведенная ранее (2.16) интегрированием интерполяционного многочлена через четыре точки $x_i, \ldots, x_{i-3}$ от x_{i-3} до x_{i+1} (первая формула Милна):

$$y_{i+1} = y_{i-3} + \frac{4h}{3}\left(2f_i - f_{i-1} + 2f_{i-2}\right). \tag{2.18}$$

Ошибка на шаге[4] $\frac{28}{90}h^5 y^{(5)}(\xi)$, где $\xi \in (x_i, x_{i+1})$.

Подставляя в эту формулу f_i, f_{i-1}, f_{i-2}, вычисленные на предыдущих шагах, получим y_{i+1} в нулевом приближении, и соответственно $f_{i+1} \equiv f(x_{i+1}, y_{i+1})$ в нулевом приближении.

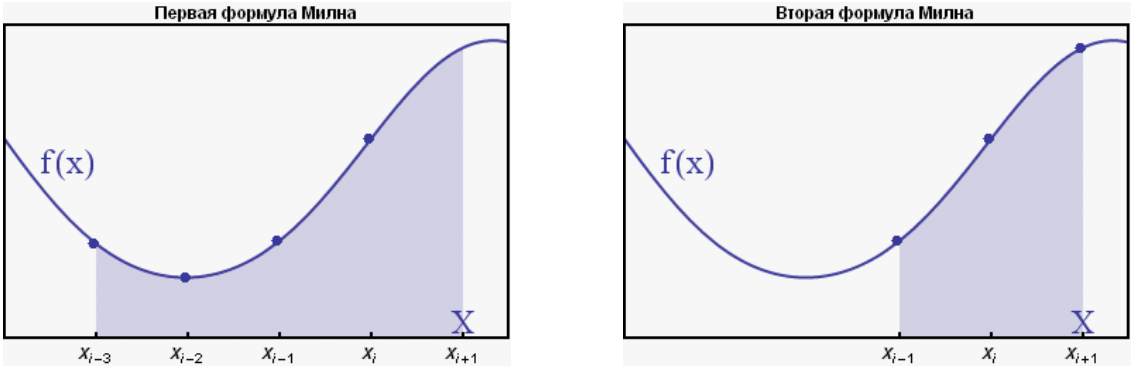


Рис. 2.2: Схема работы формул прогноза (слева) и коррекции (справа) метода Милна. Точками обозначены узлы интерполяции, а площадь затемненной области дает приближенное значение интеграла в (2.13).

В качестве формулы коррекции возьмем вторую формулу Милна (2.17), которая получается интегрированием от x_{i-1} до x_{i+1} по формуле Симпсона (то есть интегрированием в этих пределах интерполяционного многочлена через

[4]Здесь и далее остаточные члены приводятся лишь для справки. Выводятся они интегрированием ошибки интерполяции Ларганжа. На практике для оценки точности используется разница между y_{i+1} при последовательных итерациях.

три точки x_{i-1}, x_i, x_{i+1}):

$$y_{i+1} = y_{i-1} + \frac{h}{3}\left(f_{i+1} + 4f_i + f_{i-1}\right). \tag{2.19}$$

Ошибка на шаге равна $-\frac{1}{90}h^5 y^{(5)}$.

Подставляя в правую часть только что вычисленное значение f_{i+1} в нулевом приближении, получим откорректированное значение y_{i+1}. Дальше можно продолжить итерации, используя лишь формулу коррекции.

Недостаток метода Милна – неустойчивость. Ошибка имеет тенденцию к экспоненциальному росту, что свойственно всем формулам, построенным на основе формулы Симпсона.

2.3.2.2 Метод Адамса-Башфорта (Adams-Bashforth m.)

Так называют метод прогноза и коррекции, построенный на прогнозе по формуле Адамса-Башфорта 4 порядка и коррекции по формуле Адамса-Мултона 4 порядка (здесь присутствует небольшая путаница в терминологии).

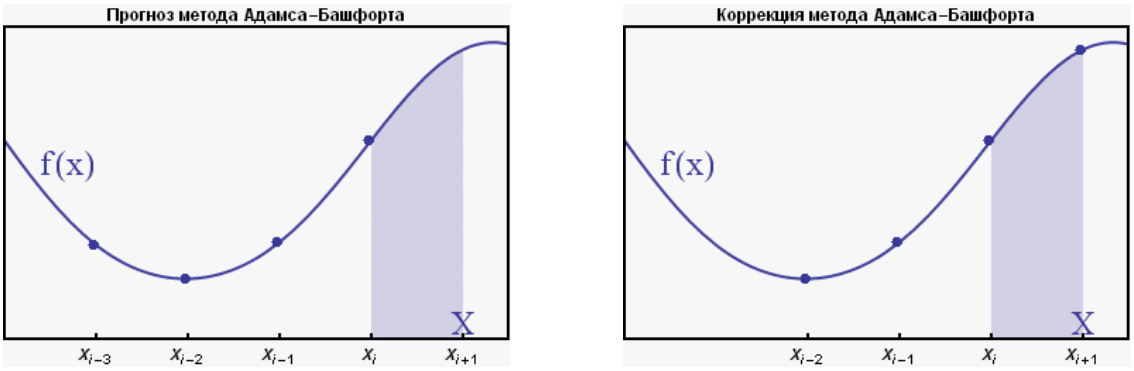


Рис. 2.3: Схема работы формул прогноза (слева) и коррекции (справа) метода Адамса-Башфорта.

Прогноз получаем интегрированием обратной интерполяционной формулы Ньютона* через точки $x_{i-3}, \ldots, x_i$ от x_i до x_{i+1}:

$$y_{i+1} = y_i + \frac{h}{24}\left(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}\right) + \frac{251}{720}h^5 y^{(5)}; \tag{2.20}$$

Коррекция (строится так же, но теперь интерполяция через $x_{i-2}, \ldots, x_{i+1}$)*

$$y_{i+1} = y_i + \frac{h}{24}\left(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}\right) - \frac{19}{720}h^5 y^{(5)}. \tag{2.21}$$

Метод устойчив – ошибка не имеет тенденции к экспоненциальному росту, – и широко используется.

Метод Хэмминга (Хемминга, Hamming's m.)* Прогноз по Милну:

$$y_{i+1}^{(0)} = y_{n-3} + \frac{4h}{3}\left(2f_i - f_{i-1} + 2f_{i-2}\right) + \frac{28}{90}h^5 y^{(5)}.$$

Уточнение прогноза:

$$\bar{y}_{i+1}^{(0)} = y_{i+1}^{(0)} + \frac{112}{121}(y_i - y_i^{(0)}); \quad \left(\bar{y}_{i+1}^{(0)}\right)' = f\left(x_{i+1}, \bar{y}_{i+1}^{(0)}\right).$$

Коррекция☆

$$y_{i+1} = \frac{1}{8}\left(9y_i - y_{i-2} + 3h\left(f_{i+1} + 2f_i - f_{i-1}\right)\right) - \frac{1}{40}h^5 y^{(5)}.$$

Метод устойчив, он позволяет оценивать погрешности на стадиях прогноза и коррекции и устранять их.

2.3.3 Общая характеристика и выбор алгоритма

1. Методы не относятся к числу "самостартующих" , так как для осуществления очередного шага необходимо иметь информацию о нескольких предыдущих шагах.

2. Одношаговые методы и методы прогноза и коррекции обеспечивают примерно одинаковую точность. Однако методы прогноза и коррекции позволяют оценить погрешность на каждом шаге и, в силу этого, эффективный шаг здесь может быть больше.

3. При оптимально выбранном шаге, который для методов RK и для методов прогноза и коррекции одного порядка по порядку одинаков, в первом случае функция f вычисляется s раз за шаг, а во втором – 2 раза за шаг (прогноз и коррекция). Поэтому, например, для $s = 4$ методы прогноза и коррекции требуют в среднем примерно вдвое меньше машинного времени, чем методы Рунге-Кутта сравнимой точности.

Также следует заметить, что многошаговые формулы допускают достаточно прямолинейное обобщение на случай неравноотстоящих узлов, и соответствующие методы прогноза и коррекции поэтому могут использоваться с переменным шагом.

Выбор алгоритма решения задачи Коши. Современные реализации как методов Рунге-Кутта, так и многошаговых, обходят недостатки простейших алгоритмов и в среднем оказываются примерно одинаково эффективны. Критерий выбора того или иного метода для своего круга задач в конечном счете один – практика.

Можно однако попробовать сформулировать некоторые общие замечания в том случае, если делается собственная реализация.

- **A.** Чем выше порядок точности метода, тем более точным должен быть полученный результат. Однако конечно-разностные аналоги производных по мере повышения порядка ведут себя все хуже и хуже. Из-за этого погрешности методов при переходе от четвертого-пятого порядков точности к более высоким порядкам практически не убывают, при том что громоздкость формул повышается существенно. Поэтому если решение предполагается достаточно гладким, часто ограничиваются методами четвертого порядка.

- **B.** Метод Рунге-Кутта следует выбирать, если вычисление $f(x, y)$ не вызывает трудностей, время счета несущественно, и основные ожидаемые затраты времени – на подготовку задачи к счету. Иначе говоря, если есть надежда решить уравнение не слишком задумываясь, потому что методы Рунге-Кутта довольно просто реализуются.

- **C.** Методы прогноза и коррекции следует выбирать, если функция $f(x, y)$ сложна, а время счета и эффективность являются существенными факторами, или если метод Рунге-Кутта оказывается неустойчивым. Тогда реализация одного из методов прогноза и коррекции будет оправдана.

2.4 Краевые задачи

Численные методы решения краевых задач можно разбить на две группы:

- **A.** Методы, основанные на замене решения краевой задачи решением нескольких задач Коши.

- **B.** Методы, в которых используются конечно-разностные формы дифференциального уравнения.

Изложение (для простоты) будем вести на примере задачи

$$y'' = f(x, y'); \qquad x \in (a, b); \qquad y(a) = A, \quad y(b) = B. \tag{2.22}$$

2.4.1 Методы стрельбы (shooting m.)

A. Если дифференциальное уравнение второго порядка является линейным, т.е. имеет вид

$$y'' = f_1(x)y' + f_2(x)y + f_3(x) \quad \text{и} \quad y(a) = A,\ y(b) = B,$$

то краевую задачу можно свести к задаче Коши.

Пусть $y_1(x)$ – решение задачи Коши с начальными условиями $y(a) = A,\ y'(a) = \alpha_1$, а $y_2(x)$ – решение задачи Коши с начальными условиями $y(a) = A,\ y'(a) = \alpha_2$. Тогда если $y_1(b) = \beta_1$, а $y_2(b) = \beta_2$, то решение

$$y(x) = \frac{1}{\beta_1 - \beta_2}\left[(B - \beta_2)y_1(x) - (B - \beta_1)y_2(x)\right]$$

удовлетворяет и исходному уравнению и исходным граничным условиям.

B. Если решается нелинейное дифференциальное уравнение, то решение можно свести к последовательному решению ряда задач Коши, последовательно вводя в граничные условия различные значения α, такие что $y(a) = A,\ y'(a) = \alpha$, и стремясь найти решения, удовлетворяющие условию $y(b) = B$.

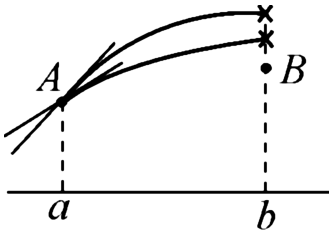


Рис. 2.4: Shooting...

Методы стрельбы: для решения нелинейного уравнения, последовательно вводя в граничные условия задачи Коши с $y(a) = A$ различные значения $\alpha = y'(a)$, ищем решения, удовлетворяющие условию $y(b) = B$.

2.4.2 Конечно-разностные методы (finite differences m.)

Эти методы основаны на сведении краевой задачи к системе алгебраических уравнений путем замены производных конечными разностями.

Так, для двухточечной задачи

$$y'' = f(x, y, y'); \quad y(a) = A, \quad y(b) = B \quad \text{на} \quad [a, b]$$

можно промежуток $[a,b]$ разбить на n равных частей: $x_i = x_0 + ih$, где $i = 1, 2, \ldots, n$, $h = (b-a)/n$; $x_0 = a$, $x_n = b$. В узлах x_i надо получить значения решения y_i.

Пользуясь конечно-разностными выражениями для производных, можно представить дифференциальное уравнение в виде разностного уравнения. Например, используя центральные разности через три точки

$$y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h}, \quad y''(x_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}. \tag{2.23}$$

Если записать эти разностные уравнения для каждого узла $i = 1, \ldots, n-1$ при двух граничных условиях, то получим систему $n-1$ уравнений с $n-1$ неизвестными.

Если дифференциальное уравнение линейное, то и полученная система линейна и решаема.

Если дифференциальное уравнение не линейно, то и полученная система не линейна и все зависит от конкретной задачи.

Способ построения разностных схем покажем на примере задачи:

$$y'' - p(x)y = f(x); \qquad y(a) = A, \quad y(b) = B. \tag{2.24}$$

1) Заменяя y'' в каждом узле центральными разностями через три точки (2.23), получим

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - p_i y_i = f_i, \quad i = 1, 2, \ldots, n-1; \quad y_0 = A,\ y_n = B. \tag{2.25}$$

Это *трехдиагональная система* $n-1$ уравнений с $n-1$ неизвестными.

2) Ошибка на шаге:

$$r_i \equiv \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - y_i'' = \frac{1}{h^2}\left\{ y_i + hy' + \frac{h^2}{2}y'' + \frac{h^3}{3!}y''' + \frac{h^4}{4!}y^{(4)} + \ldots + \right.$$
$$\left. + y_i - hy' + \frac{h^2}{2}y'' - \frac{h^3}{3!}y''' + \frac{h^4}{4!}y^{(4)} + \ldots - 2y_i \right\} - y_i'' = \frac{h^2}{12}y_i^{(4)} + O(h^4 y^{(6)}).$$

Тогда

$$y_i'' = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{h^2}{12}y_i^{(4)} + O(h^4 y^{(6)}). \tag{2.26}$$

Подставляя четвертую производную через центральные разности, приходим к

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{y_{i+2} - 4y_{i+1} + 6y_i - 4y_{i-1} + y_{i-2}}{12h^2} - p_i y_i = f_i, \quad i = 1, \ldots, n-1. \tag{2.27}$$

Это *пятидиагональная система* $n-1$ уравнений с $n-1$ неизвестными. Здесь значения y_{-1} и y_{n+1} надо вычислять экстраполяцией соответствующей точности или заменить в крайних точках разности на левые/правые.

3) С другой стороны, из уравнения имеем $y'' = f + py \Rightarrow y^{(4)} = (f+py)''$ и тогда

$$y_i^{(4)} = \frac{(f_{i+1} + p_{i+1}y_{i+1}) - 2(f_i + p_i y_i) + (f_{i-1} + p_{i-1}y_{i-1})}{h^2}. \tag{2.28}$$

Таким образом, опять получаем *трехдиагональную систему*

$$\begin{aligned} &\tfrac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}) - \\ &-\tfrac{1}{12}\left[(f_{i+1}+p_{i+1}y_{i+1}) - 2(f_i + p_i y_i) + (f_{i-1}+p_{i-1}y_{i-1})\right] - p_i y_i = f_i, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, 2, \ldots, n-1, \end{aligned} \tag{2.29}$$

но более высокой степени точности.

Аналогично оценивая погрешность, получим $r_i^{(1)} = -\frac{h^4 y_i^{(6)}}{240} + O(h^6 y_i^{(8)})$ и можно построить новую разностную схему.

2.4.3 Метод прогонки решения трехдиагональной системы линейных уравнений

Для решения трехдиагональной системы уравнений вида (2.25)

$$\begin{cases} a_{10}y_0 + a_{11}y_1 + a_{12}y_2 = b_1; & \\ a_{11}y_1 + a_{22}y_2 + a_{23}y_3 = b_2; & y_0 = A; \\ \quad\vdots & y_n = B. \\ a_{n-1,n-2}y_{n-2} + a_{n-1,n-1}y_{n-1} + a_{n-1,n}y_n = b_{n-1}; & \end{cases}$$

используется метод прогонки, который по сути представляет собой метод исключения Гаусса, который максимально просто работает для трехдиагональных систем. Метод состоит из двух частей.

1. Прямая прогонка.

Полагаем $y_0 = c_0 y_1 + \varphi_0$, где $c_0 = 0$ и $\varphi_0 = A$, так что $y_0 = A$.

Подставляем y_0 в первое уравнение системы, и выражаем y_1 через y_2, находя тем самым коэффициенты c_1 и φ_1:

$$y_1 = c_1 y_2 + \varphi_1.$$

Подставляем y_1 во второе уравнение $a_{21}y_1 + a_{22}y_2 + a_{23}y_3 = b_2$, и выражаем y_2 через y_3, находя тем самым коэффициенты c_2 и φ_2:

$$y_2 = c_2 y_3 + \varphi_2.$$

Продолжая в том же духе, на последнем шаге получим c_{n-1} и φ_{n-1} в

$$y_{n-1} = c_{n-1} y_n + \varphi_{n-1}.$$

Так вычислили наборы чисел $\{c_i\}_1^{n-1}$ и $\{\varphi_i\}_1^{n-1}$.

2. Обратная прогонка. Из второго граничного условия имеем $y_n = B$. Подставляя в последнее из полученных соотношений $y_{n-1} = c_{n-1}y_n + \varphi_{n-1}$, получаем y_{n-1}, это подставляем в предыдущее, и так далее. В итоге получаем решение $y_{n-1}, y_{n-2}, \ldots, y_1$.

Всю схему, в применении к (2.24), можно записать так:

$$\begin{array}{ccccccccccc} & & f_1,\, p_1 & & f_2,\, p_2 & & \ldots & & f_{n-1},\, p_{n-1} & & \\ & & \downarrow & & \downarrow & & & & \downarrow & & \\ \begin{array}{c} c_0 = 0 \\ \varphi_0 = A \end{array} & \to & c_1,\, \varphi_1 & \to & c_2,\, \varphi_2 & \to & \ldots & \to & c_{n-1},\, \varphi_{n-1} & & \\ & & \downarrow & & \downarrow & & & & \downarrow & & \\ & & y_1 & \leftarrow & y_2 & \leftarrow & \ldots & \leftarrow & y_{n-1} & \leftarrow y_n = B & \end{array}$$

2.5 Дифференциальные уравнения в частных производных (Numerical PDEs)

Дифференциальные уравнения в частных производных (partial differential equations, PDEs) – то же что и *уравнения мат. физики*: уравнения для функции или функций нескольких переменных, содержащие их частные производные по этим переменным.

Мы ограничимся методами решения линейных уравнений второго порядка с двумя переменными. Увеличение количества переменных или повышение порядка, по существу, ничего не меняют.

Приближенные методы решения уравнений мат. физики делятся грубо на две большие группы:

- **A.** Методы, в которых, приближенное решение получается аналитически в виде частичной суммы некоторого бесконечного ряда. К таким относится метод разделения переменных, который рассматривается в курсе методов математической физики. Сюда же относятся вариационные методы решения краевых задач (метод Ритца) и близкий к ним метод Галёркина[5].

- **B.** Численные методы. Среди них наиболее применяемыми являются *разностные методы (finite difference methods)*, благодаря универсальности и наличию хорошо разработанной теории. Их мы очень кратко и рассмотрим в этой главе.

[5]Неполный список всевозможных методов можно посмотреть на вики, например здесь или тут.

2.5.1 Разностные методы

Для применения разностного метода в области изменения переменных G вводят некоторую сетку (соответственно метод еще иногда называют методом сеток). Для каждого узла сетки записывается дифференциальное уравнение, в котором все производные заменяются разностями (или другими алгебраическими комбинациями) значений функции в соседних узлах сетки. В узлах на границе области ∂G уравнения получаются аналогичным образом из начальных или граничных условий. Получающуюся при этом систему алгебраических уравнений называют *разностной схемой*. Решая ее, находим приближенное решение в узлах сетки. Тогда решение во всей области G можно получить интерполяцией (часто используют сплайны).

В каждом конкретном случае возникают вопросы:

1. Разрешима ли разностная схема? Если дифференциальное уравнение линейное, то и разностная схема будет линейной. В общем случае это не так, и схема может быть неразрешимой, и следовательно непригодной.

2. Если схема разрешима, то какова погрешность метода, т.е. насколько ее решение близко к точному? Погрешность возникает в связи с дискретизацией задачи, и источниками ее являются, очевидно,

   (a) замена дифференциального уравнения разностной схемой;

   (b) снос граничных условий с границы ∂G на граничные узлы сетки.

   Ее можно получить, рассматривая распространение этих ошибок по разностной схеме.

3. Отсюда следует и вопрос о сходимости метода, то есть можно ли, сгущая сетку, получить решение сколь угодно близкое к точному.

Мы рассмотрим только корректно поставленные задачи, когда для определенного класса начальных и граничных условий решение (в данном классе функций) существует, единственно и непрерывно зависит от этих данных[6].

[6] Существуют физически интересные задачи, являющиеся некорректно поставленными: обратные задачи теплопроводности, задачи на развитие турбулентности, и другие.

2.5.2 Разностная аппроксимация

2.5.2.1 Сетка

Рассмотрим уравнение для функции двух переменных $u(x,y)$ вида

$$Lu(x,y) = f(x,y), \tag{2.30}$$

где f – заданная функция, а L – дифференциальный оператор. Для линейного уравнения второго порядка он имеет вид:

$$L = a\partial_x^2 + b\partial_y^2 + c\partial_x\partial_y + d\partial_x + e\partial_y + g, \tag{2.31}$$

где a, b, c, d, e, g – функции x и y, и использована сокращенная запись

$$\frac{\partial}{\partial x} \equiv \partial_x, \quad \frac{\partial^n}{\partial x^n} \equiv \partial_x^n.$$

Пусть это уравнение решается в области G, и граничные условия на границе области $\partial G \equiv \Gamma$ имеют вид $u|_\Gamma = \varphi$.

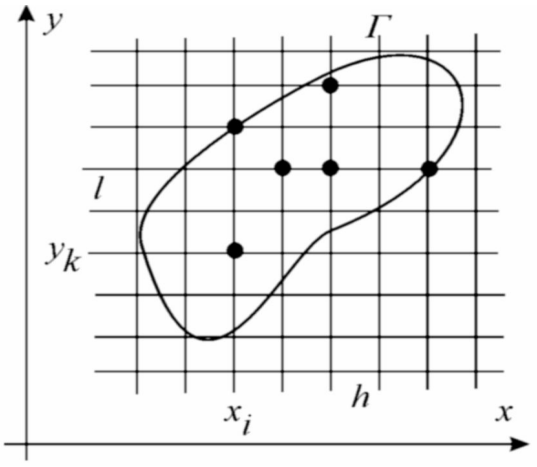


Рис. 2.5: Область G с границей Γ, сеткой и узлами.

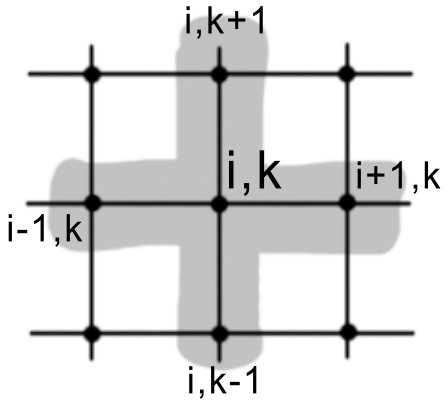


Рис. 2.6: Нумерация узлов, и шаблон при замене производных центральными разностями по трем точкам.

Введем прямоугольную *сетку* узлов, с шагом h по оси x и l по оси y, и в каждом узле (i,k) заменим производные центральными разностями через три точки:

$$\partial_x u|_{i,k} \approx \Delta_x u_{i,k} = \frac{u_{i+1,k} - u_{i-1,k}}{2h}; \qquad \partial_y u|_{i,k} \approx \Delta_y u_{i,k} = \frac{u_{i,k+1} - u_{i,k-1}}{2l};$$
$$\partial_x^2 u|_{i,k} \approx \frac{u_{i+1,k} - 2u_{i,k} + u_{i-1,k}}{h^2}; \qquad \partial_y^2 u|_{i,k} \approx \frac{u_{i,k+1} - 2u_{i,k} + u_{i,k-1}}{l^2};$$
$$\partial_x\partial_y u|_{i,k} \approx \Delta_y \frac{u_{i,k+1} - u_{i,k-1}}{2l} = \frac{u_{i+1,k+1} - u_{i-1,k+1} - u_{i+1,k-1} + u_{i-1,k-1}}{4hl}.$$

Подставляя вместо производных разности в уравнение (2.30) для каждого узла, получим вместо дифференциального уравнения систему конечно-разностных уравнений. Если $c = 0$, то в нем не будет слагаемых, пропорциональных $u_{i\pm1,k\pm1}$, и сгруппировав остальные, получим

$$A_{ik}u_{i,k+1} + B_{ik}u_{i,k-1} + C_{ik}u_{i+1,k} + D_{ik}u_{i-1,k} + E_{ik}u_{ik} = f_{ik}, \quad \text{для} \quad (i,k) \in G. \quad (2.32)$$

Набор узлов, значения u в которых используются для записи одного уравнения схемы, называют *шаблоном* (см. рис. 2.6). Если $c \neq 0$, то в шаблон будут входить и диагональные узлы. Если заменять производными более точными разностями – больше чем через три точки, – то шаблон будет содержать еще большее число узлов.

Будем для простоты считать, что граничные узлы сетки в точности попадают на границу области ∂G. Тогда уравнения в граничных узлах получаем, полагая

$$u_{ik} = \varphi_{ik} \quad \text{для} \quad (i,k) \in \partial G.$$

2.5.2.2 Задача Коши для уравнений гиперболического типа

Постановка задачи:

$$\begin{gathered} a\partial_x^2 u - b\partial_y^2 u + d\partial_x u + e\partial_y u = f, \\ u|_{y=0} = \varphi(x),\ \partial_y u|_{y=0} = \psi(x); \quad x \in \mathbb{R}. \end{gathered} \quad (2.33)$$

Заменяя производные конечными разностями, получаем в каждом узле уравнение (2.32). Зная значения функции u в двух последовательных горизонтальных рядах, из этой системы можно найти ее значения в следующем ряду.

При этом первые два ряда рассчитываются так (см. рис. 2.7):

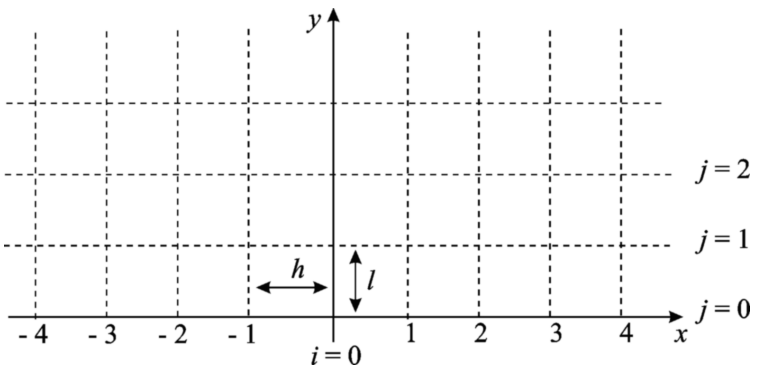


Рис. 2.7: Задача Коши. Начальные условия заданы на $y = 0$

- Значения в нулевом ряду известны и заданы граничным условием $u_{i0} = \varphi_i$;
- Можно найти значения в ряду $j=1$ двумя способами:
  - a) $\psi_i = \partial_y u\big|_{y=0} = \frac{u_{i1}-u_{i0}}{l} + O(l) \Rightarrow u_{i1} = u_{i0} + \psi_i l + O(l^2)$;
  - b) $\psi_i = \partial_y u\big|_{y=0} = \frac{u_{i,1}-u_{i,-1}}{2l} + O(l^2) \Rightarrow u_{i,-1} := u_{i,1} - 2l\psi_i + O(l^3)$;
    подставляем в уравнение (2.31) для $j \equiv k = 0$:

$$A_{i0}u_{i,1} + B_{i0}(u_{i,1} - 2l\psi_i) + C_{i0}u_{i+1,0} + D_{i0}u_{i-1,0} + E_{i0}u_{i,0} = f_{i0} \quad \Rightarrow$$
$$u_{i,1} = \frac{1}{A_{i0}+B_{i0}}\left[f_{i0} + 2lB_{i0}\psi_i - C_{i0}u_{i+1,0} - D_{i0}u_{i-1,0} - E_{i0}u_{i,0}\right];$$

    при этом погрешность порядка l^3.

Замечание: Величина h/l, вообще говоря, не может быть произвольной:

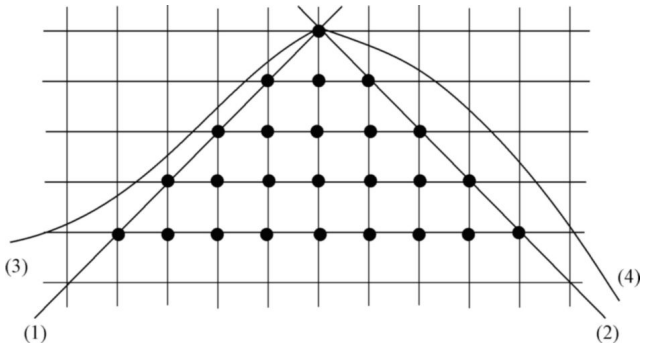


Рис. 2.8: Лучами (1) и (2) определяется так называемый треугольник определённости разностной схемы. Характеристики данного уравнения (3) и (4) определяют треугольник определенности дифференциального уравнения.

T^0: Для сходимости метода сеток треугольник определенности дифференциального уравнения должен находится внутри треугольника определённости разностной схемы ◂…▸.

При достаточно гладких коэффициентах эти условия является также и достаточным.

2.5.2.3 Краевые задачи для уравнений гиперболического типа.

Рассмотрим **первую краевую задачу**[7] – то же уравнение (2.33) с граничными условиями

$$u\big|_{y=0} = \varphi(x), \quad \partial_y u\big|_{y=0} = \psi(x);$$
$$u\big|_{x=0} = \Phi_1(y), \quad u\big|_{x=1} = \Phi_2(y); \quad x \in [0,1].$$

[7] Англоязычная классификация краевых условий (boundary conditions, b.c.): I рода – Dirichlet (or first type) b.c.; II – Neuman (or second type) b.c; III – Robin (or third type) b.c.; смешанные – mixed type b.c.

Решаем ее так:

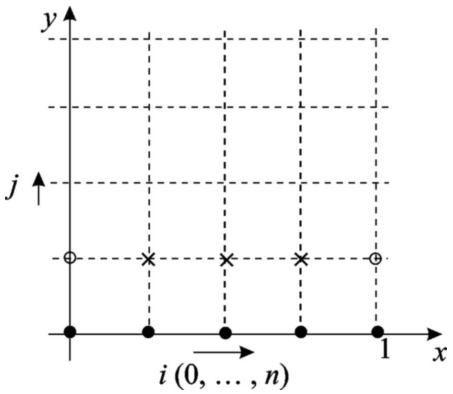


Рис. 2.9: Сетка для краевой задачи.

1. В нулевом ряду $j = 0$ из гран. условия имеем $u_{i,0} = \varphi_i$;
2. Во внутренних узлах первого ряда $j = 1$ находим значения $u_{i,1}$ так же как в предыдущем случае.
3. В граничных узлах первого ряда из граничных условий имеем
   $u_{01} = \Phi_1(y_1), \quad u_{n,1} = \Phi_2(y_1)$.
4. Значения $u_{i,j+1}$ находим, повторяя пункты 2 и 3.

Третья краевая задача: то же уравнение (2.33) с гран. условиями

$$\begin{aligned} u|_{y=0} &= \varphi(x); & (\partial_x u + \beta_1 u)|_{x=0} &= F_1(y); \\ \partial_y u|_{y=0} &= \psi(x); & (\partial_x u + \beta_2 u)|_{x=1} &= F_2(y). \end{aligned}$$

Граничные условия на $x = 0, 1$ можно переписать, заменив производные разностями:

$$\begin{cases} \frac{u_{1,k} - u_{0,k}}{h} + \beta_1 u_{0,k} = F_1(y_k); \\ \frac{u_{n,k} - u_{n-1,k}}{h} + \beta_2 u_{n,k} = F_2(y_k). \end{cases} \quad \Rightarrow \quad \begin{aligned} u_{0,k} &= \frac{u_{1,k} - hF_1(y_k)}{1 - h\beta_1}; \\ u_{n,k} &= \frac{u_{n-1,k} + hF_2(y_k)}{1 + h\beta_2}. \end{aligned}$$

Таким образом, по сравнению с первой краевой задачей отличие схемы решения только в пункте 3.

Для **второй краевой задачи** применима схема решения третьей краевой задачи, но с $\beta_1 = \beta_2 = 0$.

2.5.3 Метод неопределенных коэффициентов

Рассмотрим другой способ замены дифференциального уравнения (2.30)

$$Lu = fu$$

системой конечно-разностных уравнений.

Введем опять сетку узлов на плоскости (x, y), на этот раз не обязательно квадратную. Для каждого узла (i, k) записываем уравнение следующим образом. Перенумеруем узлы сетки: узел (i, k) назовем нулевым, некоторые N

узлов вокруг него пронумеруем от 1 до N (см. к примеру рис. 2.10), и составим линейную комбинацию с неопределенными коэффициентами $\sum_{j=0}^{N} c_j u_j$. Значения коэффициентов определим из условия, чтобы эта линейная комбинация наилучшим образом приближала Lu. Таким образом получим разностное уравнение

$$\sum_{j=0}^{N} c_j u_j = f_0. \tag{2.34}$$

Этот способ несколько отличается от метода замены производных их конечно-разностными аналогами. Преимущество его состоит в том, что он применим не только для прямоугольных сеток, но и для сеток весьма сложной конфигурации.

Мы рассмотрим его на нескольких примерах аппроксимации двумерного уравнения Пуассона, с $L = \Delta \equiv \partial_x^2 + \partial_y^2$, для различных сеток:

$$(\partial_x^2 + \partial_y^2)u = f. \tag{2.35}$$

Таким образом, мы будем строить дискретные аналоги оператора Лапласа (discrete Laplace operators) на разных решетках.

2.5.3.1 Пример. Квадратная решетка, 5 узлов

Выберем систему узлов как показано на рисунке 2.10, и будем искать аппроксимацию L в виде разностного оператора $\tilde{L}$:

$$\tilde{L}u\big|_0 = c_0 u_0 + c_1 u_1 + c_2 u_2 + c_3 u_3 + c_4 u_4.$$

С учетом симметрии L и системы узлов, сразу полагаем $c_1 = c_2 = c_3 = c_4$, так что

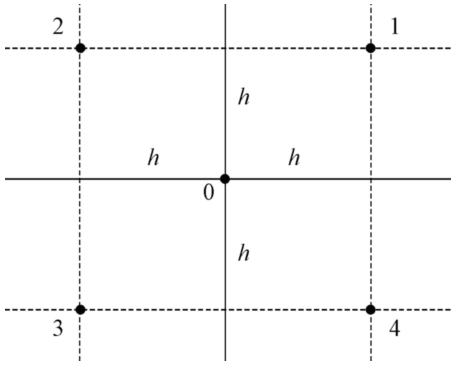


Рис. 2.10: Квадратная решетка, 5 узлов.

(здесь и далее подразумевается, что и дифференциальный L и разностный $\tilde{L}$

операторы действуют на u в точке 0, и этот индекс опускается)

$$\tilde{L}u = c_0 u_0 + c_1(u_1+u_2+u_3+u_4).$$

Значения $u_{1,2,3,4}$ получим из ряда Тейлора[8]:

$$\begin{aligned}
u_1 =& u(x_0+h, y_0+h) = \\
&= u_0 + h(\partial_x+\partial_y)u + \frac{h^2}{2}(\partial_x+\partial_y)^2 u + \frac{h^3}{3!}(\partial_x+\partial_y)^3 u + O(h^4); \\
u_2 =& u(x_0-h, y_0+h) = \\
&= u_0 + h(-\partial_x+\partial_y)u + \frac{h^2}{2}(-\partial_x+\partial_y)^2 u + \frac{h^3}{3!}(-\partial_x+\partial_y)^3 u + O(h^4); \\
u_3 =& u(x_0-h, y_0-h) = \\
&= u_0 - h(\partial_x+\partial_y)u + \frac{h^2}{2}(\partial_x+\partial_y)^2 u - \frac{h^3}{3!}(\partial_x+\partial_y)^3 u + O(h^4); \\
u_4 =& u(x_0+h, y_0-h) = \\
&u_0 + h(\partial_x-\partial_y)u + \frac{h^2}{2}(\partial_x-\partial_y)^2 u + \frac{h^3}{3!}(\partial_x-\partial_y)^3 u + O(h^4).
\end{aligned}$$

При сложении все слагаемые, нечетные по h, сократятся. Слагаемые $(\partial_x \pm \partial_y)^n$ с чередующимися знаками тоже уходят. Таким образом, получаем

$$\tilde{L}u\big|_0 = (c_0+4c_1)u_0 + 4c_1\left\{\frac{h^2}{2}(\partial_x^2+\partial_y^2)u + O(h^4)\right\}.$$

Требуем чтобы первые два слагаемых давали оператор Лапласа:

$$\begin{cases} c_0 + 4c_1 = 0 \\ 4c_1\frac{h^2}{2} = 1 \end{cases} \Rightarrow c_1 = \frac{1}{2h^2},\ c_0 = -\frac{2}{h^2}.$$

Таким образом получили дискретное приближение оператора Лапласа

$$\begin{aligned}
&\tilde{L}u\big|_0 = \tfrac{1}{2h^2}[u_1 + u_2 + u_3 + u_4 - 4u_0]; \\
&\tilde{L}u\big|_0 = \Delta u\big|_0 + O(h^2)
\end{aligned}$$

Подставляя его в уравнение Пуассона (2.35), получаем уравнение в узле

$$u_1 + u_2 + u_3 + u_4 - 4u_0 = 2h^2 f_0 + O(h^4).$$

Как видно, точность в узле $\sim h^4$.

2.5.3.2 Повышение точности. Квадратная решетка, 9 узлов

Теперь рассмотрим аппроксимацию с другой системой узлов, из 9 точек (см. рис. 2.11). Приближение Lu ищем в виде

$$\tilde{L}u = c_0 u_0 + c_1(u_1+u_2+u_3+u_4) + c_2(u_5+u_6+u_7+u_8). \tag{2.36}$$

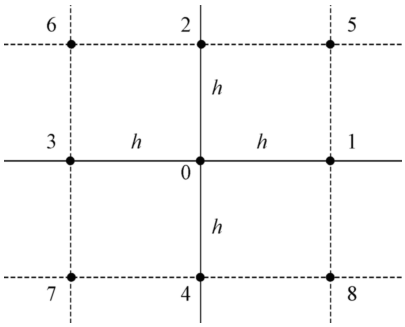


Рис. 2.11: Квадратная решетка, 9 узлов.

Поступая так же, как и в предыдущем случае, раскладываем $u_{1,\dots,8}$ в ряд Тейлора, но на этот раз до слагаемых $\sim h^6$ включительно. Вначале посчитаем $u_{5,6,7,8}$ (это узлы на диагоналях, те же что считались в предыдущем примере). Понятно, что вследствие симметрии набора узлов слагаемые, нечетные по h сократятся, поэтому их выписывать нет смысла.

$$
\begin{aligned}
u_5 =& u(x_0+h, y_0+h) = u_0 + h(\dots) + \frac{h^2}{2}(\partial_x+\partial_y)^2 u + h^3(\dots) + \\
&+ \frac{h^4}{4!}(\partial_x+\partial_y)^4 u + h^5(\dots) + \frac{h^6}{6!}(\partial_x+\partial_y)^6 u + O(h^7) \\
u_6 =& u(x_0-h, y_0+h) = u_0 + h(\dots)u + \frac{h^2}{2}(-\partial_x+\partial_y)^2 u + h^3(\dots) + \\
&+ \frac{h^4}{4!}(-\partial_x+\partial_y)^4 u + h^5(\dots) + \frac{h^6}{6!}(-\partial_x+\partial_y)^6 u + O(h^7) \\
&\dots
\end{aligned}
$$

Слагаемые $(\partial_x \pm \partial_y)^{2k}$ с чередующимися знаками тоже уходят. Поэтому, раскрывая скобки $(\partial_x \pm \partial_y)^{2k}$, оставляем лишь слагаемые, четные по производным $\partial_{x,y}$. Складывая $u_{5,6,7,8}$, в итоге получаем

$$
\begin{aligned}
u_5+u_6+u_7+u_8 = 4\Big\{ & u_0 + \frac{h^2}{2}(\partial_x^2+\partial_y^2)u + \frac{h^4}{4!}(\partial_x^4+6\partial_x^2\partial_y^2+\partial_y^4)u + \\
& + \frac{h^6}{6!}(\partial_x^6+15\partial_x^4\partial_y^2+15\partial_x^2\partial_y^4+\partial_y^6)u + O(h^8)\Big\}.
\end{aligned}
$$

Точно так же считаем $u_{1,2,3,4}$, только еще проще. Для $u_{1,2}$ имеем

$$
\begin{aligned}
u_1 = u(x_0+h, y_0) =& u_0 + h(\dots) + \frac{h^2}{2!}\partial_x^2 u + h^3(\dots) + \frac{h^4}{4!}\partial_x^4 u + h^5(\dots) + \frac{h^6}{6!}\partial_x^6 u + O(h^7); \\
u_2 = u(x_0, y_0+h) =& u_0 + h(\dots) + \frac{h^2}{2!}\partial_y^2 u + h^3(\dots) + \frac{h^4}{4!}\partial_y^4 u + h^5(\dots) + \frac{h^6}{6!}\partial_y^6 u + O(h^7);
\end{aligned}
$$

[8] Подразумевается, что значения производных также берутся в узле 0.

а $u_{3,4}$ получаются из них заменой h на $(-h)$. Складывая, получаем

$$u_1+u_2+u_3+u_4 = 4u_0+2\left\{\frac{h^2}{2!}(\partial_x^2+\partial_y^2)u+\frac{h^4}{4!}(\partial_x^4+\partial_y^4)u+\frac{h^6}{6!}(\partial_x^6+\partial_y^6)u+O(h^8)\right\}.$$

Таким образом,

$$\begin{aligned}\tilde{L}u = (c_0+4c_1+4c_2)u_0 + c_2\left\{h^4\partial_x^2\partial_y^2 u+\frac{h^6}{12}(\partial_x^4\partial_y^2+\partial_x^2\partial_y^4)u\right\}+\\ +(c_1+2c_2)\left\{h^2(\partial_x^2+\partial_y^2)u+\frac{h^4}{12}(\partial_x^4+\partial_y^4)u+\frac{2h^6}{6!}(\partial_x^6+\partial_y^6)u\right\}+O(c_ih^8).\end{aligned}$$

Приравнивая часть с младшими степенями производных оператору Лапласа, получаем систему

$$\begin{cases} c_0+4c_1+c_2 = 0,\\ h^2(c_1+2c_2) = 1.\end{cases} \tag{2.37}$$

Как видно, система неполна, и например c_2 может быть свободной переменной. Выберем ее так, чтобы производные 4й степени образовывали квадрат оператора $\Delta=\partial_x^2+\partial_y^2$. Это будет при $c_2 = 1/(6h^2)$ – тогда

$$c_2\partial_x^2\partial_y^2+\frac{1}{12}(c_1+2c_2)(\partial_x^4+\partial_y^4) = \frac{1}{12h^2}(\partial_x^2+\partial_y^2)^2,$$

а из производных 6й степени также можно выделить Δ как множитель:

$$\begin{aligned}\tilde{L}u+O(h^6) &= (\partial_x^2+\partial_y^2)u+\frac{h^2}{12}(\partial_x^2+\partial_y^2)^2u+\frac{2h^4}{6!}\left\{\partial_x^6+\partial_y^6+5(\partial_x^4\partial_y^2+\partial_x^2\partial_y^4)\right\}u =\\ &= (\partial_x^2+\partial_y^2)u+\frac{h^2}{12}(\partial_x^2+\partial_y^2)^2u+\frac{2h^4}{6!}\left\{(\partial_x^2+\partial_y^2)^3+2\partial_x^2\partial_y^2(\partial_x^2+\partial_y^2)\right\}u =\\ &= \left\{1+\frac{h^2}{12}\Delta+\frac{2h^4}{6!}\left(\Delta^2+2\partial_x^2\partial_y^2\right)\right\}\Delta u =\\ &= \left\{1+\frac{h^2}{12}\Delta+\frac{2h^4}{6!}\left(\Delta^2+2\partial_x^2\partial_y^2\right)\right\}f.\end{aligned} \tag{2.38}$$

Как видно, нам удалось вынести оператор Δ за скобки не только для слагаемых четвертого порядка $\sim\partial^4$, но и шестого порядка. Это обусловлено высокой симметрией и оператора Лапласа и выбранной решетки. Благодаря этому, заменяя $\Delta u=f$ из уравнения, получаем разностную систему точности $\sim h^6$. При $c_2 = \frac{1}{6h^2}$ из (2.37) имеем $c_1 = \frac{4}{6h^2}$ и $c_0 = -\frac{20}{6h^2}$. Подставляя эти коэффициенты в (2.36), получаем

$$\tilde{L}u\big|_0 = \frac{1}{6h^2}\left\{-20u_0+4(u_1+u_2+u_3+u_4)+(u_5+u_6+u_7+u_8)\right\}.$$

Приравнивая к (2.38), окончательно имеем уравнение

$$\begin{aligned}4(u_1+u_2+u_3+u_4)+(u_5+u_6+u_7+u_8)-20u_0 =\\ = 6h^2f_0+\frac{h^4}{2}(\Delta f)_0+\frac{h^6}{60}\left(\partial_x^4f+4\partial_x^2\partial_y^2f+\partial_y^4f\right)_0+O(h^8).\end{aligned}$$

Если функция $f(x,y)$ задана не аналитически, а только в узлах сетки, то заменив $(\Delta f)_0 = \frac{1}{2h^2}(f_5 + f_6 + f_7 + f_8 - 4f_0)$ (результат предыдущего пункта, с точностью $\sim h^4$), получим

$$4(u_1 + u_2 + u_3 + u_4) + (u_5 + u_6 + u_7 + u_8) - 20u_0 = \\ = \frac{h^2}{4}(f_5 + f_6 + f_7 + f_8 + 20f_0) + O(h^6).$$

2.5.3.3 Еще два примера*

Треугольная решетка. Возьмем теперь треугольную решетку (см. рис. 2.12). Пронумеровав от одного до шести всех ближайших соседей узла 0, которые находятся в вершинах правильного шестиугольника, по рассмотренной схеме получим☆

$$u_1 + u_2 + u_3 + u_4 + u_5 + u_6 - 6u_0 = \frac{3h^2}{2}f_0 + \frac{3h^4}{32}(\Delta f)_0 + O(h^6).$$

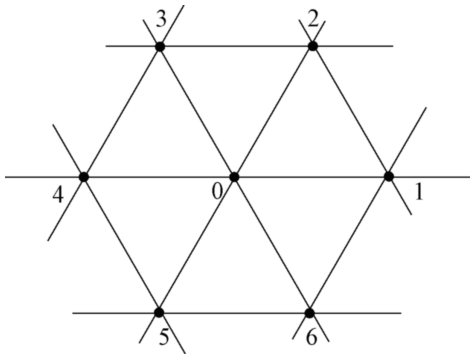


Рис. 2.12: Треугольная решетка, 7 узлов.

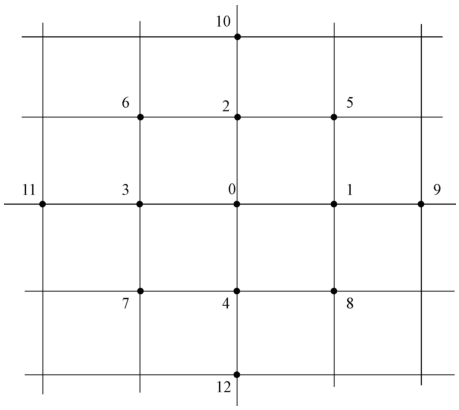


Рис. 2.13: Квадратная решетка, 13 узлов.

Квадратная решетка, 13 узлов. Рассмотрим уравнение

$$\partial_x^4 u + 2\partial_x^2\partial_y^2 u + \partial_y^4 u = f(x,y).$$

Дополнив систему девяти узлов пункта 2.3.2 еще четырьмя точками с координатами $(0, \pm 2)$ и $(\pm 2, 0)$ относительно узла 0 (см. рис.2.13), получим☆ разностную схему

$$20u_0 - 8(u_1 + u_2 + u_3 + u_4) + 2(u_5 + u_6 + u_7 + u_8) + \\ +(u_9 + u_{10} + u_{11} + u_{12}) = h^4 f_0 + O(h^6).$$

Замечание: Важно помнить, что при построении разностной схемы, аппроксимирующей дифференциальное уравнение с той или иной точностью, надо предполагать, что решение имеет необходимое количество производных, а это накладывает условия на коэффициенты уравнения, на область, и на функции, входящие в краевые условия. В противном случае повышение точности аппроксимации приводит к усложнению работы, но отнюдь не улучшает результат.

2.5.4 Об аппроксимации граничных условий

Решая методом сеток краевую задачу для дифференциального уравнения в частных производных, мы заменяем область G с границей Γ сеточной областью G^* с границей Γ^*:

$$\{G, \Gamma\} \to \{G^*, \Gamma^*\}$$

С целью уменьшения погрешностей, возникающей за счет этой замены, желательно:

- стараться осуществить выбор сетки так, чтобы ее узлы были как можно более близки к Γ;
- для нахождения значений в узлах, близких к границе, следует пользоваться интерполяционными или экстраполяционными формулами;
- в приграничных узлах использовать другие конечно-разностные уравнения, чем во внутренних узлах.

Литература

Обыкновенные дифференциальные уравнения

- Т. Шуп, *Решение инженерных задач на ЭВМ* [6];
- Wikipedia: Numerical ordinary differential equations, MathWorld: ODEs;
- J. Butcher, *Numerical methods for ordinary differential equations* [7], Вторая глава, “Numerical differential equations methods”, служит очень хорошим и основательным введением в методы решения ОДУ.

Дополнительно:
Б.П. Демидович, И.А. Марон, Э.З. Шувалова *Численные методы анализа* [5];
Р.В. Хемминг *Численные методы* [4];
Н.Н. Калиткин *Численные методы* [3], и другие.

Уравнения мат. физики

- И.С. Березин, Н.П. Жидков, *Методы вычислений* [9]. Том 2, глава 9 – ОДУ, глава 10 – метод сеток.
- Научно-образовательный сайт EqWorld.

Дополнительно:
Н.Н. Калиткин, *Численные методы* [3], глава 9;
А.Н. Тихонов, А.А Самарский, *Уравнения математической физики* [10], дополнение I.

Часть II

Глава 3

Ортогональные полиномы и численное интегрирование

3.1 Функциональные пространства
Functional Spaces

Вспомним базовые сведения из линейной алгебры, и немного продвинемся дальше, с тем чтобы иметь представление об основных структурах в бесконечномерных пространствах. В заключение мы придем к понятию гильбертова пространства, центрального для дальнейших приложений и для квантовой механики.

3.1.1 Группа и поле

Def.: *Группа (Group).* Множество G является группой по отношению ко внутренней бинарной[1] операции "$\cdot$" , если[2]

- $\forall a,b \in G \quad a \cdot b \in G$ (замкнутость, closure);
- $\forall a,b,c \in G \quad (a \cdot b) \cdot c = a \cdot (b \cdot c)$ (ассоциативность, associativity);
- $\exists e \in G \mid \forall a \in G \; a \cdot e = a$ ($\exists$ (правый) нейтральный элемент, identity element);

[1] Бинарная операция – операция, которая принимает два аргумента, и возвращает один результат.

[2] Используем мультипликативную запись, в которой групповая операция называется (абстрактным) умножением и обозначается значком "$\cdot$"; если обозначать ее значком “+”, получим аддитивную запись.

- $\forall a \in G\ \exists b \in G \mid a \cdot b = e$ ($\exists$ (правый) обратный элемент, inverse element)[3]
- Если также выполняется коммутативность (commutativity) $\forall a, b \in G \quad a \cdot b = b \cdot a$, то группа – абелева (Abelian group).

Примеры групп:

1. целые числа $\mathbb{Z}$ относительно сложения (с нейтральным элементом 0);
2. рациональные $\mathbb{Q}$, вещественные $\mathbb{R}$ и комплексные $\mathbb{C}$ числа относительно сложения (с нейтральным элементом 0) и умножения (без нуля, с нейтральным элементом 1);
3. группы преобразований, если в качестве групповой операции принять суперпозицию: трансляций (абелева), поворотов в $\mathbb{R}^n$ (не коммутативная), Лоренца в $\mathbb{R}^4_{1,3}$ (также некоммутативная)...

Def.: *Поле (Field).* Множество F с двумя бинарными операциями "+" ("сложение") и "×" ("умножение"), называется полем, если оно

- образует коммутативную группу по сложению (нейтральный элемент назовем нулем);
- все его ненулевые элементы образуют коммутативную группу по умножению;
- и выполняется свойство дистрибутивности (distributivity)

$$a \times (b + c) = a \times b + a \times c.$$

Примеры полей: $\mathbb{Q}, \mathbb{R}, \mathbb{C}$; $\mathbb{Z}_p$ – множество вычетов по модулю простого числа p.

Множество матриц $n \times n$ не является полем из-за некоммутативности умножения. Такая, более общая, структура называется *кольцом*[4].

[3] Из свойств 3 и 4 следует, что левые единица и обратный элемент существуют и совпадают с правыми:

$$\begin{cases} a \cdot e = a \\ a \cdot a^{-1} = e \end{cases} \Rightarrow \begin{cases} e \cdot a = a \cdot e = a \\ a^{-1} \cdot a = e \Leftrightarrow (a^{-1})^{-1} = a. \end{cases}$$

$$\blacktriangleleft \begin{cases} (a^{-1})^{-1} \equiv x \Leftrightarrow a^{-1} \cdot x = e \Rightarrow a = a \cdot e = a \cdot (a^{-1} \cdot x) = (a \cdot a^{-1}) \cdot x = e \cdot x = x \Leftrightarrow (a^{-1})^{-1} = a \\ e \cdot a = (a \cdot a^{-1}) a = a \cdot (a^{-1} a) = a \cdot e = a. \end{cases} \blacktriangleright$$

Единственность единицы и обратного элемента тоже легко доказываются.

[4] См. подробнее в приложении B.1.

3.1.2 Линейные пространства

Def.: *Линейное (векторное) пространство V над полем M (Linear space / vector space V over field M)* – непустое множество, в котором введены две операции, внутренняя (векторное сложение, "+") и внешняя (умножение на скаляр из поля M), такие что

- V образует абелеву группу по отношению к сложению;
- Эта группа образует "модуль над кольцом[5]" M, то есть выполняются
  1. $\forall a \in V,\ \forall m, n \in M\ (mn)a = m(na)$ (ассоциативность);
  2. $\forall a, b \in V,\ \forall m \in M\ m(a+b) = ma + mb$ (дистрибутивность);
  3. $\forall a \in V,\ \forall m, n \in R\ (m+n)a = ma + na$ (дистрибутивность);
  4. $\forall a \in V\ 1a = a$ (умножение на единицу).

Нас интересуют в первую очередь пространства над полями $\mathbb{R}$ и $\mathbb{C}$.

Пусть $\{a_i\}_{i=1}^n$ – набор (система) векторов в линейном пространстве V. Линейной комбинацией этой системы векторов называется сумма вида $\sum\limits_{i=1}^{n} \alpha_i a_i$, где $\alpha_i \in M$.

Система $\{a_i\}_{i=1}^n$ является линейно-независимой (linear-independent), если

$$\text{из} \quad \sum_{i=1}^{n} \alpha_i a_i = 0 \quad \text{следует, что} \quad \alpha_i = 0\ \forall i$$

Конечная система $\{a_i\}_{i=1}^n$ называется полной (complete) в V, если

$$\forall x \in V\ \exists \{\alpha_i\}_{i=1}^n \mid \quad x = \sum \alpha_i a_i.$$

Базис (basis) V – полная линейно-независимая система векторов из V. Число векторов в базисе есть размерность этого пространства[6].

Если для любого $n \in \mathbb{N}$ существует набор линейно-независимых векторов $\{a_i\}_{i=1}^n$, то пр.-во называется *бесконечномерным*. Иначе говоря, бесконечномерное пр.-во – такое пространство, в котором не существует конечного базиса.

[5] *Модулем над кольцом R, или R-модулем (module over ring R, R-module structure)*, называется абелева группа A с умножением на элементы кольца R с единицей, таким что выполняются приведенные 4 аксиомы.

[6] Несложно показать, что такое определение корректно и размерность не зависит от выбора базиса.

Классические примеры бесконечномерных пространств – это *функциональные пространства* (пространства функций). Если конечные последовательности $\{y_i\}_1^n$ можно трактовать, как функции, заданные на дискретном наборе точек $\{x_i\}_1^n$, и они образуют конечномерные линейные пространства, то следует ожидать, что функции, определенные (например) на непрерывных множествах, т.е. на бесконечном числе точек, образуют бесконечномерные линейные пространства.

Примеры линейных функциональных пространств

1. $C^0_{[a,b]}$ – множество непрерывных (continuous) функций, заданных на $[a,b]$;
2. $C^k_{[a,b]} \equiv C^k[a,b]$ – множество функций класса гладкости[7] k на $[a,b]$;
3. $R_{[a,b]}$ – множество функций, интегрируемых (integrable) на $[a,b]$ по Риману
$$\int_a^b dx\, f(x) < \infty;$$
4. $R^2_{[a,b]}$ – множество вещественных функций, квадратично интегрируемых на $[a,b]$ по Риману:
$$\int_a^b dx\, f^2(x) < \infty;$$
5. Множество комплекснозначных функций, с интегрируемым на $[a,b]$ квадратом модуля;
6. Множество $\overline{\Pi}_n$ многочленов степени не выше n;
7. *Векторы состояния*, которые задают состояния квантовомеханической системы, также образуют бесконечномерное пространство. В координатном представлении такой вектор задается *волновой функций* $\psi(x)$.

Структуры в линейных пространствах

В линейном пространстве можно ввести следующие конструкции

Скалярное произведение (Inner product): $\forall a,b \in V\ \exists\,(a,b) \in M$

[7] Говорят.ю что функциия f – класса гладкости k ($k \in \mathbb{N}$), $f \in C^k$, если ее k-тая производная существует и непрерывна. C^0 это непрерывные функции; функции класса C^∞ называют бесконечно-гладкими, или просто гладкими (smooth). В дальнейшем мы будем использовать обозначение $C^k_{[a,b]}$ также в ином смысле.

- $(a,b)=(b,a)$ если $M=\mathbb{R}$, или $(a,b)=\overline{(b,a)}$ если $M=\mathbb{C}$
- $(\alpha_1 a_1+\alpha_2 a_2, b)=\alpha_1(a_1,b)+\alpha_2(a_2,b)$
- $(a,a)\geq 0 \quad$ и $\quad\{(a,a)=0 \Leftrightarrow a=e\}$ ($e\equiv 0$ – нейтральный элемент V)

Обозначается ab, $a\cdot b$, (a,b), $(a;b)$, $\langle a,b\rangle$, $\langle a|b\rangle$.

Норма (Norm): $\forall a\in V\ \exists\, \|a\|\in\mathbb{R}\ |$

- $\|a\|\geq 0 \quad$ и $\quad\{\|a\|=0 \Leftrightarrow a=e\}$
- $\forall\alpha\in M \quad \|\alpha\cdot a\|=|\alpha|\cdot\|a\|$
- $\forall a,b\in V \quad \|a+b\|\leq\|a\|+\|b\|$ (неравенство треугольника)

Линейное пространство с введенной в нем нормой называют *нормированным*.

Метрика (Metric): $\forall a,b\in V\ \exists\,\rho(a,b)\in\mathbb{R}\ |$

- $\rho(a,b)\geq 0 \quad$ и $\quad\{\rho(a,b)=0 \Leftrightarrow a=b\}$
- $\rho(a,b)=\rho(b,a)$
- $\forall c\in V \quad \rho(a,b)\leq\rho(a,c)+\rho(b,c)$ (неравенство треугольника)

Линейное пространство с введенной в нем метрикой называется, очевидно, *метрическим*.

3.1.3 Евклидово пространство

В линейном пространстве со скалярным произведением всегда можно определить (индуцировать/породить скалярным произведением) норму как

$$\|a\| := \sqrt{(a,a)}. \tag{3.1}$$

Тогда все аксиомы нормы будут выполняться. Доказательство первых двух очевидно, покажем для третьей (для простоты рассматриваем пространство над $\mathbb{R}$):
◂$\forall a\ (a,a)\geq 0 \Rightarrow \forall a,b\in V,\ \lambda\in\mathbb{R}\ \ (a+\lambda b, a+\lambda b)=\lambda^2(b,b)+2\lambda(a,b)+(b,b)\geq 0$. Для положительной определенности относительно λ (это ведь квадратное уравнение) имеем $(a,a)(b,b)-(a,b)^2\geq 0$ и значит

$$|(a,b)|\leq\|a\|\cdot\|b\| \quad \text{– } \textit{неравенство Коши-Буняковского(-Шварца).} \tag{3.2}$$

Тогда $\|a+b\|^2 = (a+b, a+b) = (a,a)+(b,b)+2(a,b) \le \|a\|^2+\|b\|^2+2\|a\|\,\|b\| = (\|a\|+\|b\|)^2$ и имеем неравенство треугольника $\|a+b\| \le \|a\| + \|b\|$, ч.и т.д. ▸

Def.: Векторное пространство E со скалярным произведением, порождающим норму, называется *евклидовым пространством (Inner product space)*[8].

Система векторов $\{a_i \in E\}_{i=1}^n$ называется:
ортогональной (orthogonal), если $\forall\, i \neq j \;\; (a_i, a_j) = 0$, и
ортонормированной (orthonormal), если $\forall\, i,j \;\; (a_i, a_j) = \delta_{ij}$.

Матрица Грама[9] G системы векторов $\{a_i\}_{i=1}^n$ евклидового пространства составляется из попарных скалярных произведений:

$$G_{ij} = (a_i, a_j). \tag{3.3}$$

Ее определитель $G[a_1, \ldots, a_n]$ – определитель Грама системы $\{a_i\}_{i=1}^n$.

T^0: Система $\{a_i\}_{i=1}^n$ линейно-независима $\Leftrightarrow$ $G[a_1, \ldots, a_n] \neq 0$.

◂ Пусть $detG = 0$. Тогда строки G линейно-зависимы, т.е.
$\exists\, \{\alpha_i\}_{i=1}^n$, отличные от нуля, такие что $\sum\limits_i \alpha_i G_{ij} = 0 \quad \forall\, j \;\Leftrightarrow\; \sum\limits_i \alpha_i (a_i, a_j) = 0 \quad \forall\, j.$
Домножим обе части на α_j и просуммируем по j: получим $\left(\sum \alpha_i a_i\right)^2 = 0$
и значит $\sum \alpha_i a_i = 0$, т.е. $\{a_i\}_{i=1}^n$ линейно зависима.

Обратно: пусть система $\{a_i\}_{i=1}^n$ линейно зависима.
Значит $\exists\, \{\beta_i\}_{i=1}^n$, отличные от нуля, такие что $\sum\limits_i \beta_i a_i = 0$.
Домножим на a_j: $\sum\limits_i \beta_i G_{ij} = 0 \quad \forall\, j$. Значит строки G линейно зависимы и $detG = 0$.

Доказано, что $detG = 0 \;\Leftrightarrow$ система $\{a_i\}_{i=1}^n$ линейно зависима, и значит теорема верна. ▸

Примеры евклидовых и нормированных пространств

1. n-мерное координатное пространство R^n, элементами которого служат наборы действительных чисел $x = (x_1, x_2, \ldots, x_n)$, со скалярным произведением

$$(a,b) \equiv a \cdot b = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n. \tag{3.4}$$

является евклидовым[10];

[8] Также его называют предгильбертовым; вообще же следует иметь в виду, что термин “Е. пр-во” может иметь разные значения, и употребляется в разных (хотя и близких) смыслах.

[9] Jørgen Pedersen Gram (1850 – 1916), датский математик. Обратите внимание, что с одним "м"!; матрица Грама / Gram matrix / Gramian названа его именем.

[10] Немного подробнее об англоязычной терминологии. В ней термины scalar/dot product и Euclidean space обычно резервируют именно для этого пространства, со скалярным произведением (3.4), которое называется scalar/dot product. Inner product является обобщением такого скалярного произведения на абстрактные, возможно бесконечномерные, векторные пространства.

2. Пространство $C^2_{[a,b]}$ вещественных непрерывных функций на $[a,b]$, со скалярным произведением

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

также евклидово (проверьте* что аксиомы "$\cdot$" выполняются)

3. Пространство непрерывных функций на $[a,b]$ с максимум-нормой

$$\|f\| = \max_{x\in[a,b]} |f(x)| \tag{3.5}$$

является *нормированным* (проверьте,* что аксиомы нормы выполняются), но не евклидовым, т.к. в нем нельзя ввести соответствующее скалярное произведение – нарушается линейность.

Норма и полунорма

Рассмотрим пространство $R^2_{[a,b]}$ вещественных квадратично интегрируемых функций на $[a,b]$. В нем можем ввести скалярное произведение как

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

Оно индуцирует конечную норму

$$\|f\|^2 = \int_a^b dx\, f^2(x) < \infty \tag{3.6}$$

откуда вследствие неравенства Коши-Буняковского (3.2) следует, что введенное скалярное произведение в самом деле существует для всех $f,g \in R^2_{[a,b]}$. Линейность и коммутативность (f,g) очевидна, однако для выполнения условия положительной определенности должно выполняться

$$\|f\|^2 = 0 \quad \Rightarrow \quad f = 0.$$

Между тем, $\int dx f^2 = 0$ тогда и только тогда, когда функция f равна нулю *почти всюду*. Почти всюду – всюду на рассматриваемом множестве, за исключением множества лебеговой меры ноль. Множество точек на $\mathbb{R}$ имеет *лебегову меру ноль (is a Lebesgue null set)*, если для $\forall\,\varepsilon > 0$ существует счетная система покрытий этого множества промежутками с суммарной мерой $\mu < \varepsilon$ (мера промежутка $[a,b]$ и интервала (a,b) на $\mathbb{R}$ есть их длина $(b-a)$). Точка имеет

лебегову меру ноль. Множество, состоящее из счетного числа точек, имеет лебегову меру ноль. Канторово множество – пример множества несчетного числа точек с лебеговой мерой, равной нулю.

Таким образом, существует целый класс функций, не равных тождественно нулю, для которых $\|f\|^2 = 0$. Такая операция, удовлетворяющая всем аксиомам нормы, за исключением положительной определенности в части

$$\|f\| = 0 \;\Rightarrow\; f = 0,$$

называется не нормой, а *полунормой*. Если мы хотим рассматривать евклидово пространство (с положительно определенной нормой), нужно отождествить все функции, отличные лишь на множествах лебеговой меры ноль.

Это и будет подразумеваться везде ниже, когда будет идти речь о нормах вида (3.6) в пространствах не непрерывных функций.

3.1.4 Метрика. Метрическое пространство

Метрику (т.е. расстояние между элементами) можно ввести не только в линейном пространстве, но и на произвольном множестве. Так, положив для элементов произвольного множества

$$\rho(x,y) = \begin{cases} 0 & \text{если } x = y, \\ 1 & \text{если } x \neq y, \end{cases}$$

получим, очевидно, метрическое пространство.

Множество действительных чисел с расстоянием

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

образует одномерное метрическое пространство R^1, которое представляет собой предмет рассмотрения в математическом анализе.

В метрическом пространстве можно определить *предел последовательности*:

$$\{f_n\}_{n=1}^{\infty} \to f_0 \;\Leftrightarrow\; \rho(f_n, f_0) \underset{n\to\infty}{\longrightarrow} 0. \tag{3.7}$$

Единственность предела следует из неравенства треугольника.

Имея предел, можно ввести и понятие непрерывности функции, переводящей метрическое пространство в метрическое: $f(x)$ непрерывна в точке x, если для всякой последовательности $\{x_n\}$, такой что $x_n \to x$, верно $f(x_n) \to f(x)$.

Последовательность $\{f_n\}_{n=1}^{\infty}$ называют *фундаментальной (или последовательностью, сходящейся в себе, или последовательностью Коши / Cauchy sequence)*, если

$$\forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad | \quad \forall m, n > N \quad \rho(f_m, f_n) < \varepsilon. \tag{3.8}$$

Из неравенства треугольника следует, что всякая сходящаяся последовательность фундаментальна:

◂ Последовательность $\{f_n\}$ сходится к f_0:

$$\rho(f_n, f_0) \to 0 \Leftrightarrow \forall\, \varepsilon > 0\ \exists\, N \in \mathbb{N} \,|\, \forall n > N \qquad \rho(f_n, f_0) < \varepsilon/2$$
$$\Rightarrow \forall\, \varepsilon > 0\ \exists\, N \in \mathbb{N} \,|\, \forall m, n > N \qquad \rho(f_n, f_m) \le \rho(f_n, f_0) + \rho(f_m, f_0) < \varepsilon \quad \blacktriangleright$$

Обратное, вообще говоря, не обязательно верно.

Def.: *Полное пространство (complete space)* – метрическое пространство, в котором всякая фундаментальная последовательность имеет предел (т.е. сходится к элементу этого же пространства)[11].

Как известно из анализа, множество рациональных чисел $\mathbb{Q}$ неполно в метрике $\rho(q_1, q_2) = |q_1 - q_2|$, но добавлением к нему предельных точек всех фундаментальных последовательностей его можно *дополнить*[12] до множества вещественных чисел $\mathbb{R}$. $\mathbb{R}$ уже является полным, и кроме того, $\mathbb{Q}$ в нем всюду плотно (об этом подробнее см. ниже по тексту), т.е.

$$\forall \varepsilon > 0,\ x \in \mathbb{R} \quad \exists q \in \mathbb{Q} \,| \quad |x - q| < \varepsilon.$$

Точно так же, для любого метрического пространства верна теорема

T^0: Всякое метрическое пространство R имеет *пополнение* R^*, т.е. полное метрическое пространство, такое что R является его подпространством и всюду плотно в R^*. При этом оно единственно с точностью до изоморфизма ◂ . . . ▸.

Простейший способ построить пополнение следующий:

Назовем две фундаментальные последовательности $\{x_n\}$ и $\{y_n\}$ в R эквивалентными[13], если $\rho(x_n, y_n) \to 0$. И пусть тогда элементами пространства R^*

[11] Надо обратить внимание, что термин “полнота” также используется в математике в очень разных смыслах, и в каждом конкретном случае нужно понимать что именно имеется в виду. Так, например, мы уже вводили понятие полноты для конечных систем векторов линейного пространства.

[12] См. построение вещественных чисел по Кантору

[13] Отношение эквивалентности на произвольном множестве есть бинарное отношение, такое что 1) $a \sim a$; 2) $a \sim b \ \Rightarrow b \sim a$; 3) $a \sim b$, $b \sim c \Rightarrow a \sim c$. Класс эквивалентности – множество элементов, которые эквивалентны друг другу.

будут классы эквивалентности всех фундаментальных последовательностей из R. Выберем из каждого класса по представителю, скажем $\{x_n\}$ из класса x^* и $\{y_n\}$ из класса y^*, и определим метрику в R^* как $\rho(x^*, y^*) = \lim\limits_{n\to\infty}\rho(x_n, y_n)$. При этом элементам x пространства R соответствуют классы эквивалентности последовательностей, эквивалентных последовательности констант $x_n = x$.

3.1.5 Банахово пространство

В нормированном пространстве метрика может быть индуцирована нормой

$$\rho(a, b) := \|a - b\| \tag{3.9}$$

и тогда сходимость будет определяться как сходимость по норме.

При этом метрика приобретает дополнительные свойства, а именно удовлетворяет тождеству параллелограмма[14]

$$\rho(a, b) = \rho(a + c, b + c) \quad \text{и} \quad \rho(\alpha a, \alpha b) = |\alpha| \cdot \rho(a, b),$$

и обратно: если метрика удовлетворяет тождеству параллелограмма, то она индуцирует норму и скалярное произведение ◂ . . . ▸☆

Def.: *Банахово пространство (Banach space)* – нормированное линейное векторное пространство, полное по метрике, порождённой нормой.

Функции сложения, умножения на число и скалярного произведения (в евклидовом пространстве) *непрерывны*, т.е.

$$\begin{cases} x_n \to x \\ y_n \to y \end{cases} \quad \Rightarrow \quad \begin{cases} \alpha x_n + \beta y_n \to \alpha x + \beta y \quad \forall \alpha, \beta \in M \\ (x_n, y_n) \to (x, y). \end{cases} \tag{3.10}$$

Докажем для скалярного произведения (докажите остальное☆).

◂ Положим $x_n = x + u_n$, $y_n = y + v_n$. Из сходимости имеем $\|u_n\| \to 0$, $\|v_n\| \to 0$. Тогда

$$(x, y) - (x_n, y_n) = -(x, v_n) - (u_n, y) + (u_n, v_n)$$

и из неравенства Коши-Буняковского (3.2) и неравенства треугольника

$$|(x, y) - (x_n, y_n)| \le \|x\|\,\|v_n\| + \|u_n\|\,\|y\| + \|u_n\|\,\|v_n\| \to 0,$$

т.е. $(x_n, y_n) \to (x, y)$ ▸

[14] Это одна из возможных эквивалентных формулировок.

Примеры метрических и банаховых пространств

Приведем несколько примеров пространств с разными метриками. Конечномерные метрические пространства всегда полны[15], а полноту или неполноту бесконечномерных пока доказывать не будем.

1. n-мерное пространство наборов n действительных чисел $x = (x_1, x_2, \ldots, x_n)$, с евклидовыми нормой и метрикой

$$\|x\|^2 = \sum_{k=1}^{n} x_k^2 \quad \Leftrightarrow \quad \rho(x,y) = \left(\sum_{k=1}^{n} (y_k - x_k)^2 \right)^{1/2}. \tag{3.11}$$

   Эта норма индуцирует скалярное произведение $(x,y) = \sum x_k y_k$, в результате чего получаем евклидово пространство R^n.

2. То же множество с *L^p-нормой*, которая индуцирует *p-метрику*:

$$\|x\| = \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} \Leftrightarrow \rho(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^p \right)^{1/p}, \quad \text{где} \quad p \geq 1, \tag{3.12}$$

   обозначается $R_p^{(n)}$. Предыдущий вариант есть его частный случай с $p = 2$. Другой частный случай $p = 1$ дает расстояние "taxicab geometry"[16]:

$$\rho(x,y) = \sum_{k=1}^{n} |x_k - y_k|. \tag{3.13}$$

   Предельный случай $p \to \infty$ дает равномерную (uniform) норму (ее также называют максимум-нормой, по очевидной причине) и соответствующую метрику

$$\rho(x,y) = \max_{1 \leq k \leqslant n} |x_k - y_k|. \tag{3.14}$$

   Неравенство треугольника в этом случае сводится к неравенству Минковского

$$\left(\sum_{k=1}^{n} |a_k + b_k|^p \right)^{1/p} \leq \left(\sum_{k=1}^{n} |a_k|^p \right)^{1/p} + \left(\sum_{k=1}^{n} |b_k|^p \right)^{1/p}. \tag{3.15}$$

3. *Пространство квадратично суммируемых комплекснозначных последовательностей l^2*

$$x = (x_1, x_2, \ldots, x_n, \ldots) \mid \quad \sum_{k=1}^{\infty} |x_k|^2 < \infty, \tag{3.16}$$

[15] Для евклидовых пространств полнота следует из эквивалентности сходимости по норме и покоординатной (см. конспект по высшей алгебре), и полноты $\mathbb{R}$.

[16] Это расстояние, которое проезжает такси от одного перекрестка до другого в городе, где дороги образованы прямоугольной сеткой улиц.

с расстоянием

$$\rho(x,y) = \left(\sum_{k=1}^{\infty} |x_k - y_k|^2 \right)^{1/2}. \tag{3.17}$$

Такая метрика[17], так же как и первый приведенный пример, порождает норму $\|x\|^2 = \sum |x_k|^2$ и скалярное произведение $(x,y) = \sum_{k=1}^{\infty} x_k \overline{y_k}$. Неравенство треугольника следует из неравенства Коши-Буняковского (3.2). Полноту мы докажем позже (п.3.1.8).

4. Множество $C_{[a,b]}$ всех непрерывных действительных функций, определенных на $[a,b]$, с *равномерной метрикой*

$$\rho(f,g) = \max_{t\in[a,b]} |g(t) - f(t)| \tag{3.18}$$

тоже образует метрическое пространство, которое обозначим так же как и исходное множество функций – $C_{[a,b]}$. Этой метрике соответствует равномерная норма (3.5). Аксиомы метрики и нормы проверяются непосредственно. Это бесконечномерный вариант равномерной метрики, рассмотренной в примере 2.

$C_{[a,b]}$ полно ◂…▸☆и следовательно образует банахово пространство.

5. Пространство $C^2_{[a,b]}$ непрерывных действительных функций на $[a,b]$, с *квадратичной метрикой*, соответствующей норме (3.6)

$$\rho(f,g) = \left(\int_a^b dt \left[g(t) - f(t) \right]^2 \right)^{1/2}. \tag{3.19}$$

$C^2_{[a,b]}$ не полно, а значит может быть дополнено до полного (см. п.3.1.10)

Равномерная (3.18) и квадратичная (3.19) метрика, вместе с соответствующими нормами (3.5) и (3.6), нас далее будут интересовать в первую очередь.

3.1.6 Ряды Фурье в бесконечномерных пространствах

Fourier series in infinite-dimensional spaces

Рассмотрим полное линейное пространство B с метрикой и нормой, индуцированными скалярным произведением. Оно, очевидно, и евклидово и банахово.

[17] По аналогии с конечномерными $R_p^{(n)}$, можно определить целое семейство пространств p-суммируемых последовательностей l^p, с p-метрикой:

$$\sum_{k=1}^{\infty} |x_k|^p < \infty, \quad p \geq 1; \quad \longrightarrow \quad \rho(x,y) := \left(\sum_{k=1}^{\infty} |x_k - y_k|^p \right)^{1/p}.$$

Пусть $\{a_i\}_1^\infty$ – ортогональная система, т.е. $(a_i, a_j) = 0$ при $i \neq j$. В метрическом пространстве B можно определить бесконечный ряд

$$\sum_{i=1}^{\infty} a_i \tag{3.20}$$

Как обычно, мы говорим что ряд (3.20) сходится, если сходится последовательность частичных сумм ряда $\sum_1^n a_i$ при $n \to \infty$.

Ряд $\sum^\infty a_i$ сходится $\Leftrightarrow$ последовательность $\sum^n a_i$ фундаментальна $\Leftrightarrow$

$\Leftrightarrow \ \forall \varepsilon > 0 \ \exists N \in \mathbb{N} \,|\, \forall\, n \geq N,\, p \geq 1 \ \|a_{n+1} + \ldots + a_{n+p}\| < \varepsilon \ \Leftrightarrow$

$\Leftrightarrow \ \|a_{n+1}\|^2 + \ldots + \|a_{n+p}\|^2 < \varepsilon^2$, так как для конечной суммы $\|a_1 + \ldots + a_n\|^2 = \|a_1\|^2 + \ldots + \|a_n\|^2$.

$\Leftrightarrow$ числовая последовательность $\sum^n \|a_i\|^2$ фундаментальна $\Leftrightarrow$

$\Leftrightarrow$ числовой ряд $\sum^\infty \|a_i\|^2$ сходится.

Таким образом, если $\{f_i\}_1^\infty$ – ортонормированная система, т.е. $(f_i, f_j) = \delta_{ij} \ \forall\, i, j$, то сходимость ряда по этой ортонормированной системе

$$\sum_{k=1}^{\infty} \xi_k f_k \tag{3.21}$$

эквивалентна сходимости числового ряда $\sum \xi_k^2$ (это для пространства над $\mathbb{R}$, над $\mathbb{C}$ будет очевидно $\sum |\xi_k|^2$).

- Пусть ряд (3.21) сходится, и его сумма есть x. Тогда рассмотрим скалярное произведение $(\sum_1^n \xi_k f_k, f_p)$. С одной стороны, при достаточно большом n только одно слагаемое суммы отлично от нуля, и $(\sum_1^n \xi_k f_k, f_p) = \xi_p \underset{n\to\infty}{\longrightarrow} \xi_p$. С другой стороны, из непрерывности скалярного произведения имеем $(\sum_1^n \xi_k f_k, f_p) \underset{n\to\infty}{\longrightarrow} (x, f_p)$.Тогда вследствие единственности предела $\xi_k = (x, f_k)$.

  Числа ξ_k называются *коэффициентами Фурье* x по ортонормированной системе $\{f_k\}$, а ряд (3.21) с такими коэффициентами

  $$x = \sum_{k=1}^{\infty} (x, f_k)\, f_k \tag{3.22}$$

  называется ***рядом Фурье*** для x по ортонормированной системе $\{f_k\}$.

  Таким образом, если ряд (3.21) сходится, то он есть ряд Фурье для своей суммы x. При этом[18]

  $$\|x - \sum^n \xi_k f_k\|^2 = (x - \sum \xi_k f_k)^2 = (x, x) - 2\sum \xi_k (x, f_k) + \sum \xi_k^2 = \|x\|^2 - \sum \xi_k^2,$$

[18] Здесь и далее для скалярного квадрата *вектора (функции)* f используется сокращенное обозначение $(f, f) \equiv f^2$. Следует его отличать от квадрата *числа* x, для которого используется то же обозначение.

и т.к. левая часть стремится к нулю при $n \to \infty$, то выполняется

$$\textit{равенство Парсеваля:} \qquad \sum_{k=1}^{\infty} \xi_k^2 = \|x\|^2. \tag{3.23}$$

- Пусть теперь есть $x \in B$.
  Построим набор коэффициентов $\xi_k = (x, f_k)$ и ряд Фурье (3.22) для x по системе f_k. Т.к. $\|x - \sum^n \xi_k f_k\|^2 = \|x\|^2 - \sum^n \xi_k^2$, то

  $$\sum_{k=1}^{n} \xi_k^2 \le \|x\|^2$$

  – числовая последовательность частичных сумм $\sum^n \xi_k^2$ неубывающая и ограничена сверху, а потому сходится. Поэтому ряд Фурье для любого вектора $x \in B$ сходится всегда.

  При этом он сходится к самому элементу x, если в последнем неравенстве при $n \to \infty$ достигается равенство, так что выполняется равенство Парсеваля (3.23).

  Если для ортонормированной системы $\{f_k\}$ и $\forall\, x \in B$ верно (3.23), то система $\{f_k\}$ называется *замкнутой*. Тогда каждый элемент пространства представим в виде своего ряда Фурье по $\{f_k\}$.

Остается открытым вопрос о *существовании* в B замкнутой системы функций и о том, является ли заданная система замкнутой. Это не следует из аксиом (предположений об изучаемом пространстве), которые были до сих пор введены. Оказывается, что существование замкнутой системы связано со свойством *сепарабельности* пространства B.

3.1.7 Сепарабельность и замкнутые системы

Сепарабельность

Def.: Множество A *всюду плотно (dense)* в метрическом пространстве $M \supset A$, если

$$\forall m \in M\ \forall \varepsilon > 0 \quad \exists a \in A \quad | \quad \rho(m - a) < \varepsilon.$$

Так, множество рациональных чисел $\mathbb{Q}$ всюду плотно в $\mathbb{R}$ (в метрике $\rho(x, y) = |x - y|$). При этом оно *счетно (countable)*[19].

[19] Множество счетно, если оно равномощно подмножеству $\mathbb{N}$. Один из способов пересчитать все рациональные числа показан на картинке.

Пространство M называется *сепарабальным (separable)*, если в нем существует счетное всюду плотное множество. Так, $\mathbb{R}$ сепарабельно.

Свойство сепарабельности, по существу, означает, что пространство в определенном смысле "не слишком большое". Банахово пространство всех ограниченных последовательностей с супремум-нормой $\|a\| = \sup\limits_i |a_i|$– один из примеров несепарабельных пространств.

$\mathbf{T}^0$: Если пространство B сепарабельно, то всякая ортонормированная система содержит конечное или счетное число векторов.

◂ Пусть w и v – два вектора из множества попарно ортонормированных векторов.

Тогда $\|w - v\| = \sqrt{(w-v, w-v)} = \sqrt{2}$. В силу сепарабельности $\forall\, w\, \exists\, u_k \mid \|u_k - w\| \le \varepsilon$. Зафиксируем $\varepsilon < 1/\sqrt{2}$. Предположим, что заданному k соответствует два разных вектора из ортонормированной системы – w и v. Тогда $\|w - v\| \le \|w - u_k\| + \|v - u_k\| < \sqrt{2}$, чего быть не может. Таким образом, каждому значению индекса k, который нумерует счетную систему $\{u_k\}$, соответствует не более одного вектора из ортонормированной системы. Значит последняя счетна (или, как частный случай, конечна), ч.и т.д. ▸

Замкнутые и полные системы векторов

Ортонормированная система $\{f_k\}_1^\infty$ называется *полной*, если

$$\text{из} \quad (x, f_k) = 0\ \forall\, k \quad \text{следует, что} \quad x = 0.$$

Если $\{f_k\}_1^\infty$ полна, то разные вектора имеют разные ряды Фурье.

◂ Пусть коэффициенты Фурье x есть ξ_k, а y – η_k. Если $\xi_k = \eta_k\ \forall k$, то $(x-y, f_k) = 0\ \forall k$ и из полноты $x - y = 0 \quad \Rightarrow \quad x = y$ ▸

$\mathbf{T}^0$: В сепарабельном пространстве замкнутая система полна, а полная замкнута.

◂ Пусть $\{f_k\}$ замкнута. Тогда из $\xi_k \equiv (x, f_k) = 0\ \forall\, k$ вследствие замкнутости имеем $\|x\|^2 = \sum \xi_k^2 = 0$ и значит $x = 0$. Таким образом, $\{f_k\}$ полна.

Пусть $\{f_k\}$ полна. Тогда рассмотрим последовательность $x - \sum^n \xi_k f_k$: $(\sum^n \xi_k f_k, f_p)$ при достаточно больших n равно ξ_p и значит в пределе $n \to \infty$ имеем $(x - \sum^n \xi_k f_k, f_p) = 0\ \forall p$. Тогда из полноты $x = \sum^\infty \xi_k f_k$ и $\sum^\infty \xi_k^2 = \|x\|^2$, т.е. система замкнута. ▸

Сепарабельность и замкнутые системы

T⁰: Если пространство B сепарабельно, то, ортонормировав плотное в нем счетное множество $\{u_i\}_{i=1}^{\infty}$, мы получим замкнутую счетную систему $\{w_i\}_{i=1}^{\infty}$.

◂ Вначале из последовательности $\{u_i\}$ извлечем линейно-независимую подпоследовательность $\{\tilde{u}_i\}$. Для этого пройдем по $\{u_i\}$ по порядку, отбрасывая те элементы, что являются линейной комбинацией предыдущих оставленных, и ноль. После этого ортогонализуем последовательность $\{\tilde{u}_i\}$ стандартной процедурой Грама-Шмидта:

$$w_1 = \frac{\tilde{u}_1}{\|u_1\|};\ w_2 = \frac{\tilde{u}_2 - (\tilde{u}_2, w_1)w_1}{\|\tilde{u}_2 - (\tilde{u}_2, w_1)w_1\|},\ \ldots,\ w_n = \frac{\tilde{u}_n - \sum_{i=1}^{n-1}(\tilde{u}_n, w_i)w_i}{\left\|\tilde{u}_n - \sum_{i=1}^{n-1}(\tilde{u}_n, w_i)w_i\right\|},\ \ldots \tag{3.24}$$

По построению всякий вектор w_i есть конечная линейная комбинация векторов u_j и наоборот.

Покажем, что $\{w_i\}_{i=1}^{\infty}$ полна. Предположим, что для некоторого $f \in B$ верно $(f, w_i) = 0\ \forall i$. Но тогда верно и $(f, u_i) = 0\ \forall i$. Система $\{u_i\}$ плотна в B, так что для всякого $\varepsilon > 0$ существует u_i, такое что $\|f - u_i\| < \varepsilon$. Тогда

$$\|f\|^2 = (f, f) = (f, f) - (u_i, f) = (f - u_i, f) \le \|f - u_i\| \cdot \|f\| < \varepsilon\|f\|,$$

и значит $\|f\| = 0$, то есть $f = 0$, и система $\{w_i\}$ полна, а следовательно и замкнута. ▸

Верно и обратное утверждение: если в пространстве B существует замкнутая система $\{w_i\}$, то конечные суммы вида $\sum c_i w_i$ с рациональными коэффициентами образуют счетное[20] всюду плотное в B множество. Таким образом, верна теорема:

T⁰: Сепарабельность B эквивалентна существованию в нем замкнутой системы векторов.

3.1.8 Гильбертово пространство (Hilbert space)

Посмотрим теперь, что нужно для того, что в бесконечномерном линейном пространстве, как и в привычном нам конечномерном, всякий вектор можно было “разложить по базису”. Для этого нужно чтобы в нем существовала замкнутая ортонормированная система векторов, и чтобы всякий ряд Фурье каждого

[20] Счетность его доказывается с использованием идеи схемы, использующейся для доказательства счетности $\mathbb{Q}$.

элемента пространства по ней сходился к этому элементу. Собирая свойства пространства, который для этого необходимы, и которые обсуждались в настоящем разделе, получим определение гильбертова пространства.

Гильбертово пространство (Hilbert space) – **"п.л.е.б.с.": полное линейное евклидово бесконечномерное сепарабельное пространство**[21]. Обозначается оно H.

Очевидно, гильбертово пространство является одновременно евклидовым, метрическим, нормированным и банаховым, причем метрика и норма индуцированы в нем скалярным произведением. В указанном смысле оно является самым "хорошим" из бесконечномерных. Когда мы говорим о разложении по базису в гильбертовом пространстве, мы подразумеваем разложение в (возможно бесконечный) ряд Фурье по соответствующей замкнутой системе, которая представляет собой *базис Шаудера (Schauder basis)*[22].

Между каждым элементом H и его рядом Фурье по замкнутой системе, коэффициенты которого образуют квадратично суммируемую последовательность, есть взаимно однозначное соответствие. Т.е. есть взаимно-однозначное соответствие между элементами H и пространства квадратично суммируемых последовательность l^2 (3.16). Как несложно увидеть, это взаимно-однозначное соответствие сохраняет линейные операции, определенные в этих пространствах, и скалярное произведение[23]. Значит оно осуществляет изоморфим евклидовых (бесконечномерных в данном случае) пространств.

Значит *H изоморфно l^2 и уникально с точностью до изоморфизма*[24]. Отсюда немедленно следует полнота l^2.

Таким образом, l^2 можно рассматривать как "координатную реализацию" H, так же как n-мерное координатное пространство со скалярным произведением $\sum^n x_i y_i$ представляет собой координатную реализацию евклидова n-мерного пространства, заданного аксиоматически.

В квантовой механике необходимо, чтобы вектор состояния системы можно было разложить по некоторому базису. Поэтому полагается, что вектора состояния образуют гильбертово пространство.

[21] Иногда в определение не включают требования бесконечномерности и сепарабельности. Конечномерные гильбертовы пространства в таком смысле совпадают с евклидовыми.

[22] Можно также определить *базис Гамеля (Hamel basis)*, как такую систему элементов пространства V, что каждый вектор из V может быть представлен в виде *конечной* суммы по этой системе. В бесконечномерных пространствах это совершенно разные базисы.

[23] В l^2 это скалярное призведение, индуцирующее норму и метрику (3.17).

[24] Несепарабельные гильбертовы пространства не уникальны.

3.1.9 Теоремы Вейерштрасса

Какие функциональные пространства являются гильбертовыми? Для того, чтобы ответить на этот вопрос, рассмотрим вначале множество $C_{[a,b]}$ непрерывных функций на $[a,b]$.

Первая аппроксимационная теорема Вейерштрасса[25] (Weierstrass' fundamental theorem of approximation theory):

T⁰: Пусть $f(x)$ – непрерывная функция на $[a,b]$, где $-\infty < a < b < \infty$. Тогда для всякого $\varepsilon > 0$ существует алгебраический многочлен $p(x)$, такой что

$$|f(x) - p(x)| < \varepsilon \quad \forall x \in [a,b]. \tag{3.25}$$

Другими словами, всякую непрерывную функцию на $[a,b]$ можно сколь угодно точно *равномерно* аппроксимировать многочленом.

Последовательность $f_n(x)$ сходится (поточечно) к $f(x)$ при $n \to \infty$ если

$$\forall x \in [a,b],\ \varepsilon > 0 \quad \exists N \mid \forall n > N \quad |f(x) - f_n(x)| < \varepsilon.$$

Последовательность $f_n(x)$ сходится равномерно к $f(x)$ при $n \to \infty$ если

$$\forall \varepsilon > 0 \quad \exists N \mid \forall n > N \quad |f(x) - f_n(x)| < \varepsilon \quad \forall x \in [a,b].$$

Из равномерной последовательности следует поточечная, но не наоборот.

◀ 1. Вначале заметим, что всякую непрерывную функцию f можно сколь угодно точно приблизить ломаной, которая в точках $x_0 < x_1 < \ldots < x_m$ совпадает с f (при достаточном числе узлов). Соответствующую кусочно-линейную (полигональную) функцию можно записать как

$$g^{(m)}(x) = g_1(x) + \sum_{i=1}^{m-1} [g_{i+1}(x) - g_i(x)]\,\Theta(x - x_i),$$

где $g_i(x)$ – прямая[26], совпадающая с $f(x)$ в двух соседних узлах, x_i и x_{i-1}, а $\Theta(x)$ есть функция единичной ступеньки

$$\Theta(x) = \begin{cases} 0 & \text{если} \quad x < 0, \\ 1 & \text{если} \quad x \geq 0. \end{cases}$$

[25] Карл Вейерштрасс, Karl Theodor Wilhelm Weierstrass (1815 – 1897). Статья с этим результатом была опубликована им в 1885, в возрасте 70 (!) лет.

[26] Можно выписать и в явном виде

$$g_i(x) = f(x_i) + \frac{x - x_i}{x_i - x_{i-1}} [f(x_i) - f(x_{i-1})],$$

В самом деле, тогда при $x \in (x_{j-1}, x_j)$ функция $\Theta(x - x_i)$ будет отлична от нуля если $x_i < x$, то есть при $i = 1, 2, \ldots, j-1$. Подставляя, получаем

$$x \in (x_{j-1}, x_j): \quad g^{(m)}(x) = g_1 + (g_2 - g_2) + \ldots + (g_j - g_{j-1}) = g_j(x).$$

Тогда задача сводится к задаче о приближении многочленами единичной ступеньки $\Theta(x)$ – потому что тогда и $g^{(m)}(x)$ можно будет сколь угодно точно приблизить многочленами

2. Для дальнейшего нам понадобится неравенство Бернулли

$$(1+x)^n \geq 1 + nx, \quad n = 0, 1, \ldots; \quad x > -1. \tag{3.26}$$

◂ Доказательство по индукции. При $n = 0$ из (3.26) имеем $1 \geq 1$, что верно. Пусть (3.26) верно для $n = k$. Тогда

$$(1+x)^{k+1} \geq (1+x)(1+kx) = 1 + (k+1)x + kx^2 \geq 1 + (k+1)x,$$

то есть (3.26) верно и для $n = k+1$. Неравенство доказано. ▸

3. Пусть

$$q_n(x) = (1 - x^n)^{2^n}.$$

Покажем, что

$$\lim_{n\to\infty} q_n(x) = \Theta\Big(\frac{1}{2} - x\Big) \equiv \begin{cases} 1 & \text{если} \quad 0 \leq x < 1/2; \\ 0 & \text{если} \quad 1/2 < x \leq 1. \end{cases}$$

◂ При $x \in [0, 1/2)$ имеем из неравенства Бернулли (3.26)

$$1 \geq q_n = (1 - x^n)^{2^n} \geq 1 - x^n 2^n = 1 - (2x)^n,$$

значит из теоремы о двух милиционерах (sandwich theorem)

$$x \in [0, 1/2): \quad \lim_{n\to\infty} q_n(x) = 1.$$

При $x \in (1/2, 1)$ из того же неравенства

$$\frac{1}{q_n(x)} = \frac{1}{(1 - x^n)^{2^n}} = \left(1 + \frac{x^n}{1 - x^n}\right)^{2^n} \geq 1 + \frac{(2x)^n}{1 - x^n} > (2x)^n,$$

откуда

$$0 < q_n(x) < 1/(2x)^n$$

и так как $2x > 1$, то опять по теореме о двух милиционерах

$$x \in (1/2, 1): \quad \lim_{n\to\infty} q_n(x) = 0. \qquad \blacktriangleright$$

Для доказательства поточечной аппроксимации f многочленами этого достаточно. Доказательство равномерной сходимости требует еще некоторых рассуждений.

Пусть, для определенности, $\tilde{x} \in (0, 1/2)$. Тогда если для этого $\tilde{x}$ и заданного ε существует N, такое что для всякого $n > N$ верно $|q_n(\tilde{x}) - 1| < \varepsilon$, то и для всех $x < \tilde{x}$ это также верно, и получаем равномерную сходимость на $[0, 1/2 - \delta] \bigcup [1/2 + \delta, 1]$ для $\forall \delta > 0$.

4. Кусочно-линейная функция $g^{(m)}(x)$ записывается через единичные ступеньки в каждом узле x_i, каждая из которых аппроксимируется равномерно многочленами на $[a, b] \backslash O_\delta(x_i)$ для $\forall \delta > 0$.

Для достаточно малых δ (меньше расстояния между соседними узлами) функция $g^{(m)}(x)$ в δ-окрестности точки x_i, $O_\delta(x_i)$, имеет вид

$$g^{(m)}(x) = f(x_i) + \big[\alpha + \beta\Theta(x - x_i)\big](x - x_i).$$

Если $Q_{i,n}(x)$ – последовательность многочленов, которая приближает $\Theta(x - x_i)$, то последовательность многочленов $P_n^{(m)}$, которая приближает $g^{(m)}$, можно представить в виде

$$P_n^{(m)}(x) = f(x_i) + \big[\alpha + \beta Q_{i,n}(x)\big](x - x_i) + R_n^{(m)}(x),$$

где $R_n^{(m)}$ – последовательность многочленов, равномерно сходящаяся к нулю в этой окрестности. Тогда

$$|g^{(m)} - P_n^{(m)}| \le \Big|\beta\big[\Theta(x - x_i) - Q_{i,n}\big](x - x_i)\Big| + |R_n| = O(1) \cdot |x - x_i| + |R_n|$$

и первое слагаемое может быть сделано сколь угодно малым в $O_\delta(x_i)$ выбором достаточно малого δ. А при $n \to \infty$ второе слагаемое стремится к нулю равномерно на $[x_{i-1} + \delta_1, x_{i+1} - \delta_1]$ для $\forall \delta_1 > 0$. Таким образом, $g^{(m)}$ равномерно аппроксимируется многочленами и на всякой достаточно малой окрестности каждого узла x_i, а значит – и на всем промежутке $[a, b]$.

5. Итак, всякая непрерывная функция f равномерно аппроксимируется на промежутке полигональной функцией $g^{(m)}(x)$, а $g^{(m)}$ в свою очередь равномерно аппроксимируется многочленами $P_n^{(m)}$. Так как

$$|f - P_n^{(n)}| \le |f - g_n| + |g_n - P_n^{(n)}|,$$

то теорема Вейерштрасса доказана. ▶

Теорему можно переформулировать так:

T⁰: Множество многочленов плотно в пространстве $C_{[a,b]}$ непрерывных функций на $[a,b]$ с равномерной метрикой

$$\rho(f,g) = \max_{x\in[a,b]} |f(x) - g(x)|.$$

Но из равномерной сходимости следует и сходимость по всякой норме L^p с $p > 1$:

$$|f_n(x) - p(x)| < \varepsilon \ \forall x \in [a,b], \quad \Rightarrow \quad \left(\int_a^b dx\, |f_n - p|^\alpha \right)^{1/\alpha} < \varepsilon \cdot |b-a|^{1/\alpha}.$$

Поэтому множество многочленов плотно в $C_{[a,b]}$ по норме L^p с любым $p > 1$, и в частности по квадратичной норме (3.6).

При этом в множестве многочленов плотно множество многочленов с рациональными коэффициентами $\Pi_{\mathbb{Q}}$ (также в равномерной метрике и следовательно в любой L^p), а последнее счетно, что доказывается по обычной схеме. Значит $C^p_{[a,b]}$ сепарабельно.

Из всех этих метрических пространств нас больше всего интересует $C^2_{[a,b]}$, так как в нем можно определить скалярное произведение.

Вторая аппроксимационная теорема Вейерштрасса формулируется практически так же как и первая:

T⁰: Множество тригонометрических многочленов (1.33) плотно в пространстве $C_{[0,2\pi]}$ непрерывных функций на $[0,2\pi]$ с равномерной метрикой.

Тогда, естественно, множество тригонометрических многочленов плотно и в пространстве непрерывных функций с квадратичной метрикой $C^2_{[0,2\pi]}$.

Первая и вторая теоремы Вейерштрасса эквивалентны. Проще всего показать, что первая следует из второй. Синус и косинус – целые функции, ряд Тейлора для которых сходится на всей комплексной плоскости. Поэтому всякий тригонометрический многочлен всегда можно приблизить с любой наперед заданной точностью алгебраическим многочленом – частичной суммой его ряда Тейлора. Доказательство обратного не столь тривиально, но несколько более слабое утверждение мы докажем, рассматривая полиномы Чебышёва в п.3.2.4.4. Вторая теорема может быть доказана и независимо, аналогично первой.

3.1.10 Пространство L^2

В метрическом пространстве $C^2_{[a,b]}$ вещественных непрерывных функций, определенных на $[a,b]$, с метрикой и нормой

$$\|x\| = \left(\int_a^b dt\, x^2(t) \right)^{1/2}, \qquad \rho(x,y) = \|x-y\|, \tag{3.27}$$

ведем скалярное произведение

$$(x,y) = \int_a^b dt\, x(t)y(t), \tag{3.28}$$

и таким образом получим бесконечномерное евклидово пространство.

Как мы видели, из аппроксимационной теоремы Вейерштрасса следует, что оно сепарабельно. Оказывается однако, что $C^2_{[a,b]}$ не полно. Это несложно показать, рассматривая последовательность функций

$$\varphi_n(t) = \begin{cases} -1 & \text{при} \quad t < -1/n; \\ nt & \text{при} \quad |t| < 1/n; \\ +1 & \text{при} \quad t > 1/n, \end{cases}$$

которая фундаментальна в квадратичной метрике $\blacktriangleleft \ldots \blacktriangleright^{\star}$, но, очевидно, сходится при $n \to \infty$ к разрывной ступеньке.

Как мы знаем (п.3.1.4), неполное пространство всегда можно дополнить до полного, добавив к нему элементы, к которым сходятся его фундаментальные последовательности. Обозначим пополнение C^2 через C^{2^*}. Так как счетное множество многочленов с рациональными коэффициентами $\Pi_{\mathbb{Q}}$ плотно в C^2, а C^2 плотно в C^{2^*}, то $\Pi_{\mathbb{Q}}$ плотно в C^{2^*}, и последнее сепарабельно*. Таким образом, C^{2^*} является, по построению, полным, бесконечномерным и сепарабельным.

Правда, так как C^{2^*} шире чем C^2, оно содержит разрывные функции, и значит положительная определенность нормы во второй части может быть нарушена (п.3.1.3). Для ее восстановления необходимо отождествить все функции, равные поточечно почти всюду – расстояние между которыми равно нулю. То, что получается в результате, называется пространством L^2. Это уже гильбертово пространство.

Базисы L^2

В соответствии с первой теоремой Вейерштрасса, в C^2, а значит и в L^2, плотно множество алгебраических многочленов. Поэтому множество

$$\Pi = \{1, x, x^2, \ldots, x^n, \ldots\},$$

образующее базис многочленов, является и базисом L^2, так что всякая функция $f(x) \in L^2$ может быть представлена в виде ряда по многочленам, который сходится к $f(x)$ в норме L^2:

$$f(x) = \sum_{n=0}^{\infty} f_n x^n, \qquad \int_a^b dx \left(f(x) - \sum_{n=0}^{N} f_n x^n \right)^2 \underset{N\to\infty}{\longrightarrow} 0$$

Правда, элементы П не образуют ортогональную систему, поэтому раскладывать по ним неудобно.

Точно так же, по второй теореме Вейерштрасса, в $L^2[0, 2\pi]$ плотно множество тригонометрических многочленов (1.33):

$$T = \{1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots, \sin nx, \cos nx, \ldots\}$$

Поэтому синусы и косинусы являются базисом L^2, причем ортогональным (несложно проверить) в смысле (3.28). Ряд Фурье (в рассмотренном общем смысле) функции $f(x) \in L^2[0, 2\pi]$ это и есть обычный *тригонометрический ряд Фурье* (1.44):

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) .$$

Коэффициенты a_k и b_k находятся домножением ряда на $\sin nx$ или $\cos nx$ и интегрированием почленно от 0 до 2π.

Интегралы Римана и Лебега

Норма, метрика и скалярное произведение в L^2 индуцированы метрикой C^2: если $\{x_n\}$ – фундаментальная последовательность в C^2, сходящаяся к $x \in L^2$, то

$$\|x\|_{L^2} = \lim_{n\to\infty} \|x_n\|_{C^2} = \lim_{n\to\infty} \left(\int_a^b dt\, x_n^2(t) \right)^{1/2} .$$

Этот предел удобно принять за новое определение интеграла,

$$\int_a^b dt\, x(t) := \lim_{n\to\infty} \int_a^b dt\, x_n^2(t),$$

который, в отличие от интеграла Римана, пригоден для всех функций из C^{2*}, например для функции Дирихле

$$D(x) \equiv I_{\mathbb{Q}} = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \in \mathbb{R}\backslash\mathbb{Q} \end{cases} ,$$

которая разрывная во всех точках, но принадлежит C^{2*}. В последнем можно убедиться, построив фундаментальную последовательность функций C^2, которая к ней сходится[27].

Можно показать[28], что такое обобщение интеграла в данном случае эквивалентно более общему *интегралу Лебега*. Формальное определение последнего требует довольно абстрактного введения в теорию меры, и потому вынесено в приложение B. Однако геометрическую идею его построения можно легко понять. Идея построения интеграла Лебега состоит в том, что разбивается на много кусочков не промежуток интегрирования, как в случае интеграла Римана, а интервал значений функции (см. картинку[29]).

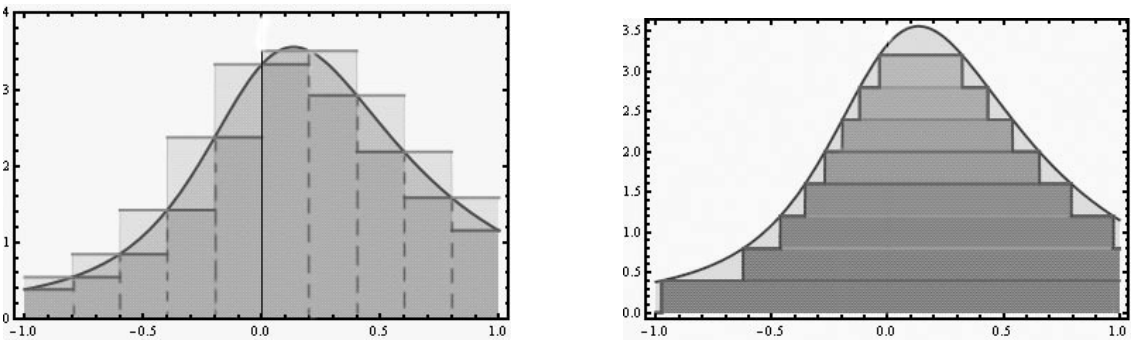


Рис. 3.1: Интегральные суммы для интеграла Римана (слева) и интеграла Лебега (справа).

Такой подход имеет ряд преимуществ, и в частности, позволяет распространить понятие интеграла на более широкий класс функций, в том числе на функции, определенные на произвольных множествах – лишь бы в них можно было ввести *меру*. Так, функция Дирихле $D(x)$ не интегрируема по Риману. Однако она интегрируема по Лебегу и в этом смысле $\int_0^1 D(x) = 1$, потому что $D(x) = 0$ лишь на счетном множестве $\mathbb{Q}$, а единице на гораздо "большем" множестве $\mathbb{R}\backslash\mathbb{Q}$. Как видно, интеграл Лебега в некотором смысле дает более

[27] Это делается следующим образом. Вначале перенумеруем все рациональные числа на $[0,1]$ – $\{q_i\}_{i=1}^{\infty}$. В качестве первого элемента последовательности возьмем полигональный зубчик, равный единице в q_1 и с шириной основания $1/2$. В качестве второго элемента – два таких зубчика вокруг q_1 и q_2, с ширинами основания $1/2^2$. Продолжая в том же духе, получим фундаментальную (это можно проверить) последовательность непрерывных функций, которые представляют собой набор все увеличивающегося числа все более острых зубчиков вокруг рациональных чисел, и которая сходится к $D(x)$.

[28] Это далеко выходит за рамки нашего курса, и интересующихся мы отсылаем к литературе [11], [12].

[29] Также на demonstrations.wolfram.com можно посмотреть демонстрацию.

естественную конструкцию чем интеграл Римана.

Приведем еще пару утверждений об отношении интегралов Лебега и Римана, ограничившись случаем интегралов на вещественной прямой.

T⁰: Функция $f(x)$, определенная на $[a,b]$, $a<b$, интегрируема по Риману т.и т.т., когда она ограничена и непрерывна почти всюду на $[a,b]$.

T⁰: Если функция $f(x)$ интегрируема на $[a,b]$ по Риману, то она интегрируема и по Лебегу, и соответствующие интегралы равны. Таким образом, интеграл Лебега является обобщением интеграла Римана.

Таким образом, $L^2[a,b]$ совпадает с пространством функций, квадратично интегрируемых на $[a,b]$ по Лебегу – отсюда и обозначение.

Тогда пространство[30] $R^2[a,b]$ функций, интегрируемых на $[a,b]$ по Риману с метрикой (3.27), является подпространством L^2, а с другой стороны, содержит C^2:

$$C^2[a,b] \subset R^2[a,b] \subset L^2[a,b].$$

Метрика в L^2 задается так же (3.9), но интеграл надо, очевидно, понимать как интеграл Лебега.

Если мы будем понимать его как обычно, т.е. как интеграл Римана, это означает, что мы будем работать в пространстве R^2. Неприятности, которые нам из-за этого могут встретиться, ограничиваются тем, что, вследствие неполноты R^2, некоторые фундаментальные последовательности функций из R^2 будут сходиться к элементу L^2, не принадлежащему R^2. Но если мы будем работать с достаточно "хорошими" функциями, то можно надеяться, что этого не произойдет. Интегрирование же по Лебегу, как правило, можно произвести, переопределив функцию на множестве лебеговой меры ноль, так чтобы она стала интегрируемой по Риману.

3.2 Ортогональные полиномы
Orthogonal Polynomials

Рассмотрим пространство V вещестсвеннозначных функций $f(x)$, определенных на промежутке $C = [a,b]$, и квадратично интегрируемых (по Лебегу) на нем с *весом* $p(x)$, т.е.

$$\int\limits_C d\mu\, f^2(x) < \infty \quad \forall f(x) \in V, \quad \text{где} \quad d\mu = p(x)dx > 0 \tag{3.29}$$

[30] Оно является *метрическим* пространством, так же как и L^2, только после отождествления функций, расстояние между которыми равно нулю.

будем понимать просто как сокращенное обозначение; и потребуем чтобы все *моменты*

$$\mu \equiv \mu_0 = \int\limits_C d\mu; \qquad \mu_k \equiv \int\limits_C d\mu\, x^k, \quad k = 0, 1, \ldots$$

были конечны, с тем, чтобы V включало в себя все многочлены.

Это пространство строится аналогично L^2: начинаем с пространства непрерывных функций на $[a, b]$, со скалярным произведением

$$(f, g) = \int\limits_C d\mu\, f(x) g(x), \tag{3.30}$$

и индуцированными нормой и метрикой; дополняем его до полного и отождествляем функции, отличные на множестве лебеговой меры ноль. Получаем гильбертово пространство, которое обозначается[31] $L^2_{[a,b]}(p)$.

Как мы знаем, по теореме Вейерштрасса, в $C_{[a,b]}$ есть всюду плотное, в равномерной норме, множество многочленов $\overline{\Pi}$. Тогда, так же как в случае с $L^2(1)$, оно плотно и в $L^2(p)$, т.к.

$$|f - p_n| < \varepsilon \quad \Rightarrow \quad \int d\mu\, |f - p_n| < \varepsilon \cdot \mu_0,$$

и значит система $\{x^n\}_{n=0}^{\infty}$ образует базис в L^2. Ортогонализуя ее, получим систему полиномов, *ортогональных с весом* $p(x)$ *на промежутке* $[a, b]$ (3.30), которая замкнута (и полна) в $L^2(p)$.

Для ортогонализации можно использовать процесс Грама-Шмидта, а можно выразить ортогональные полиномы через моменты μ_k следующим образом. Пусть наша система ортонормирована

$$(p_n, p_m) = \delta_{nm}, \qquad n, m = 0, 1, 2, \ldots \tag{3.31}$$

и обозначим коэффициенты как

$$p_n(x) = k_n^{(n)} x^n + k_{n-1}^{(n)} x^{n-1} + \ldots + k_0^{(n)}, \quad k_n \equiv k_n^{(n)} > 0; \tag{3.32}$$

$$\tilde{p}_n(x) \equiv \frac{p_n}{k_n} = x^n + \kappa_1^{(n)} x^{n-1} + \ldots + \kappa_0^{(n)}, \quad \text{где} \quad \kappa_i^{(n)} = \frac{k_i^{(n)}}{k_n}. \tag{3.33}$$

Здесь k_n – старшие коэффициенты (leading coefficients) p_n, а $\tilde{p}_n$ – *приведенные* (monic) ортогональные полиномы.

Для ортогональности достаточно обеспечить, чтобы для всех n выполнялось

$$(p_n, x^k) = 0, \quad k = 0, 1, \ldots, n-1. \tag{3.34}$$

[31] Для меньшей громоздкости, когда будет понятно какой имеется в виду промежуток $[a, b]$ и вес p, их будем опускать.

Из (3.34) получаем систему уравнений для $\kappa_i^{(n)}$:

$$\sum_{i=0}^{n-1} \sigma_{ik}\kappa_i^{(n)} + \sigma_{nk} = 0 \quad k = 1, \ldots, n-1, \tag{3.35}$$

где коэффициенты

$$\sigma_{ik} = (x^i, x^k) = \int_C d\mu\, x^{i+k} = \mu_{i+k} \tag{3.36}$$

образуют симметричную матрицу $(n \times n)$ – $(\sigma_{ij})_{i,j=0}^{n-1}$. Она представляет собой матрицу Грама для системы функций $\{x^i\}_{i=0}^{n-1}$, которая очевидно линейно-независима, а значит матрица невырождена и ее определитель отличен от нуля. Значит система (3.35) разрешима, из нее можно найти все $\kappa_i^{(n)}$, а старшие коэффициенты тогда определятся из условия нормировки.

Решение можно представить в явном виде через детерминант:

$$p_n(x) = \frac{1}{\sqrt{D_n D_{n-1}}} \begin{vmatrix} \sigma_{00} & \sigma_{01} & \ldots & \sigma_{0,n-1} & 1 \\ \sigma_{10} & \sigma_{11} & \ldots & \sigma_{1,n-1} & x \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{n-1,0} & \sigma_{n-1,1} & \ldots & \sigma_{n-1,n-1} & x^{n-1} \\ \sigma_{n0} & \sigma_{n1} & \ldots & \sigma_{n,n-1} & x^n \end{vmatrix}, \tag{3.37}$$

где $D_n = det[(\sigma_{ij})_{i,j=0}^{n}]$ – определитель Грама системы $\{x^k\}_{k=0}^{n}$. Понятно, что $p_n(x)$ вещественны.

◂ Обозначим определитель в правой части (3.37) через Q_n. Это многочлен степени n. Домножим последний столбец определителя на x^k $(k < n)$ и проинтегрируем с весом p, то получим этот же определитель, в котором последний столбец совпадает с k-тым, то есть

$$\int d\mu Q_n(x) x^k = 0, \quad \text{для} \quad k = 0, 1, \ldots, n-1,$$

так что Q_n ортогонален всем многочленам степени ниже n.

Квадрат нормы Q_n тогда получим, раскрывая один из сомножителей по последнему столбцу:

$$\int d\mu Q_n^2 = \int d\mu Q_n (x^n D_{n-1} + \ldots) = D_{n-1} \int d\mu Q_n x^n = D_{n-1} D_n.$$

Вследствие ортогональности, все слагаемые степени ниже n в первом интеграле дают ноль, а оставшийся интеграл считаем так же, домножая и интегрируя последний столбец Q_n. Таким образом, получили (3.37). ▸

Заметим, что зная базис $\{p_i(x)\}$, можно построить $L^2(p)$ конструктивно, и определить его просто как пространство всех бесконечных сумм $\sum \xi_i p_i(x)$ с квадратично суммируемыми коэффициентами $\sum \xi_i^2 \|p_n\|^2 < \infty$ (множитель $\|p_n\|^2$ возникает, если система не нормирована).

Ортогональные полиномы в комплексной плоскости*

Точно так же можно рассмотреть ортогональные полиномы на произвольной кривой C (не обязательно даже связной) в комплексной плоскости. При этом условие $d\mu > 0$ остается (но теперь "вес" $p(z)$ не обязательно вещественный), скалярное произведение определяется с комплексным сопряжением

$$(f, g) = \int\limits_C d\mu\, f(x)\overline{g(x)},$$

а все рассуждение по существу не меняется. Только моменты μ_k и коэффициенты σ_{ik} в этом случае также становятся комплексные

$$\sigma_{ik} = (z^i, z^k) = \int\limits_C d\mu\, z^i\, \overline{z^k}$$

и образуют соответственно эрмитову матрицу.

Конечно, ортогональные полиномы на произвольной кривой в $\mathbb{C}$ сложно к чему-либо приспособить. Хорошо развита теория ортогональных полиномов на единичном круге и на вещественной оси. Первой мы касаться не будем, и в дальнейшем сосредоточимся на второй.

3.2.1 Экстремальная задача в L^2

T^0: Приведенные ортогональные полиномы являются решениями задачи на минимизацию нормы L^2 среди приведенных полиномов заданной степени.

◂ Действительно, поставим задачу найти приведенный полином степени n, $P_n = x^n + \ldots$, который минимизирует норму L^2, т.е.

$$\int d\mu\, P_n^2 = \min. \tag{3.38}$$

Разложим P_n по ортонормированным полиномам p_n: $P_n = \sum\limits_{i=0}^{n} c_i p_n$. Вследствие ортонормированности

$$\int d\mu\, P_n^2 = \sum_{i=0}^{n} c_i^2,$$

причем $c_n = 1/k_n$, т.к. старший коэффициент P_n должен быть равен 1. Тогда интеграл достигает минимума, когда все остальные коэффициенты c_k равны нулю, т.е. когда $P_n = \tilde{p}_n$ ▸.

3.2.2 Рекуррентные соотношения

T^0: Для ортогональных полиномов на вещественной прямой выполняются *трехчленные рекуррентные соотношения*:

$$xp_n(x) = a_n p_{n+1}(x) + b_n p_n(x) + \tilde{a}_n p_{n-1}(x), \tag{3.39}$$

где

$$a_n = \frac{(xp_{n+1}, p_n)}{(p_{n+1}, p_{n+1})} = \frac{k_n}{k_{n+1}} > 0, \quad b_n = \frac{(xp_n, p_n)}{(p_n, p_n)}, \quad \tilde{a}_n = a_{n-1}\frac{\|p_{n+1}\|^2}{\|p_n\|^2}. \tag{3.40}$$

◂ Разложим $xp_n(x)$ по системе $\{p_i(x)\}_{i=0}^{n+1}$:

$$xp_n(x) = \sum c_k p_k, \quad \text{где} \tag{3.41}$$

$$c_k \equiv \frac{(xp_n, p_k)}{(p_k, p_k)} = \frac{\int d\mu\, xp_n(x)\, p_k(x)}{(p_k, p_k)} = \frac{(xp_k, p_n)}{(p_k, p_k)}. \tag{3.42}$$

Вследствие ортогональности, c_k обращаются в ноль для всех $k < n-1$, т.к в этом случае $xp_k(x)$ есть полином степени меньше n.

Сравнивая коэффициенты при старших степенях в (3.41), видим, что $c_{n+1} = k_n/k_{n+1}$. С другой стороны, из (3.42)

$$c_{n-1} = \frac{(xp_{n-1}, p_n)}{(p_n, p_n)} = \frac{\|p_{n+1}\|^2}{\|p_n\|^2}\frac{(xp_{n-1}, p_n)}{(p_{n+1}, p_{n+1})} = \frac{\|p_{n+1}\|^2}{\|p_n\|^2}c_{n+1} = \frac{\|p_{n+1}\|^2}{\|p_n\|^2}\cdot\frac{k_{n-1}}{k_n}.$$

Оставшийся коэффициент c_n определяется из (3.42) с $k = n$.▸.

Эти соотношения очевидным образом упрощаются в случае если система полиномов ортонормирована $\|p_i\|^2 = 1$.

Стартовые значения рекуррентного соотношения следует выбирать так: из условия нормировки получаем $p_0 = 1/\sqrt{\mu}$, а $p_{-1} = 0$ (проверяется непосредственно из соотношения (3.39) подстановкой $n = 0$, где p_1 получаем из (3.37)).

Теорема Фаварда*: если система полиномов удовлетворяет соотношениям (3.39) с вещественными $a_n > 0$ и b_n, то она образует ортогональную систему по отношению к некоторому весу на вещественной оси ◂...▸.

Если начинать рекуррентную цепь со значений $q_{-1} = -1$, $q_0 = 0$ и использовать те же самые рекуррентные соотношения (3.39) с $a_{-1} = 1$, то получим другую систему полиномов q_n степени $n-1$, которая по теореме Фаварда также является ортогональной с каким-то весом. Эти полиномы называют *ортогональными полиномами второго рода* (p_n тогда – первого рода), и их можно записать в виде ◂...▸

$$q_n(z) = p_0 \int_a^b d\mu\, \frac{p_n(z) - p_n(x)}{z - x}.$$

3.2.3 Свойства нулей

$\mathbf{T}^0$: Все n нулей p_n простые, вещественные и лежат на промежутке (a,b), а нули p_n и p_{n+1} перемежаются.

То есть каждый нуль p_n находится между двумя нулями p_{n+1} (см. рис.3.2).

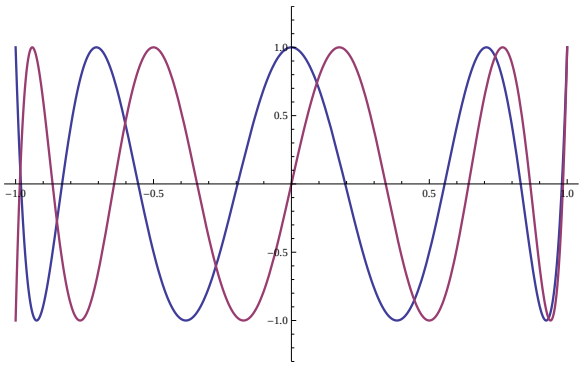


Рис. 3.2: Перемежаемость нулей на примере $T_8(x)$ и $T_9(x)$ (см. п.3.2.4.4)

◂ 1) В самом деле, полином p_n должен иметь n перемен знака на (a,b). Если бы он имел только $m < n$ перемен знака, скажем в точках $y_1, \dots, y_m$, то он не мог бы быть ортогонален полиному $q(x) = \prod_{j=1}^{m}(x - y_j)$ степени $m < n$, т.к. в этом случае $q(x)p_n(x)$ была бы знакопостоянной функцией.

2) Из рекуррентного соотношения (3.39), полагая $x = x_k$, получим

$$0 = a_n p_{n+1}(x_k) + a_{n-1} p_{n-1}(x_k), \tag{3.43}$$

т.е. $p_{n+1}(x_k)$ и $p_{n-1}(x_k)$ имеют противоположные знаки.

3) Теперь докажем перемежаемость нулей по индукции.

База. Из предыдущего понятно, что p_0 – положительная константа, p_1 - линейная функция с нулем на промежутке $[a,b]$, который обозначим как x_0. p_2 – квадратный трехчлен с положительным старшим коэффициентом и обоими нулями на $[a,b]$, так что $p_2(a) > 0$, $p_2(b) > 0$. Из (3.43) следует, что $p_2(x_0) < 0$. Следовательно, два нуля p_2 находятся на (a,x_0) и (x_0,b) соответственно.

Теперь предположим, что мы уже знаем, что нули p_n и p_{n-1} перемежаются, и пусть $x_n < x_{n-1} < \ldots < x_1$ это нули p_n. Значит, каждый из $(n-1)$ нулей p_{n-1} лежит между двумя нулями p_n, и между каждой парой точек x_k, x_{k+1} меняет знак: $\operatorname{sign} p_{n-1}(x_k) = (-1)^{k+1}$ (x_1 лежит справа от всех нулей p_{n-1}, а старший коэффициент p_{n-1} больше нуля, поэтому $p_{n-1}(x_1) > 0$). Тогда из (3.43) сразу следует, что в каждом из $(n-1)$ интервалов (x_k, x_{k+1}) лежит по нулю p_{n+1}, что и означает перемежаемость.

У p_{n+1} есть еще два нуля. То, что они принадлежат промежуткам (a,x_n) и (x_1,b) соответственно, доказывается так же, как доказывалась база. ▸

3.2.4 Классические ортогональные полиномы

Так называемые *"классические"* ортогональные полиномы возникают в математической физике как решения дифференциальных уравнений второго порядка определенного вида и образуют самый простой класс специальных функций. К ним относятся полиномы Эрмита (задача о гармоническом осцилляторе в квантовой механике $E_n = \hbar\omega(n + 1/2)$); Лагерра (радиальная зависимость волновой функции электрона в атоме водорода); Лежандра (решение уравнения Лапласа в сферических координатах) и пр.

Рассмотрим уравнение для $y(z)$ вида

$$\sigma(z)y'' + \tau(z)y' + \lambda y = 0, \tag{3.44}$$

где $\sigma(z)$ и $\tau(z)$ – полиномы не выше второй и первой степени соответственно. Такое уравнение будет называть *уравнением гипергеометрического типа*[32], а его решения – функциями гипергеометрического типа. Слово "гипергеометрический" будем здесь сокращать как "г.".

Покажем, что производные от y удовлетворяют уравнению того же типа. В самом деле, дифференцируя (3.44), получим что для $v_1(z) \equiv y'(z)$ выполняется

$$\begin{aligned} &\sigma(z)v_1'' + \tau_1(z)v_1' + \mu_1 v_1 = 0, \\ &\text{где} \quad \tau_1(z) = \tau(z) + \sigma'(z), \quad \mu_1 = \lambda + \tau'(z). \end{aligned}$$

Так как $\tau(z)$ и $\sigma'(z)$ – полиномы степени не выше первой, то это уравнение г. типа (3.44).

Продолжая дифференцировать по z, получим, что $v_n(z) \equiv y^{(n)}(z)$ удовлетворяет уравнению

$$\begin{aligned} &\sigma(z)v_n'' + \tau_n(z)v_n' + \mu_n v_n = 0, \\ &\text{где} \quad \tau_n = \tau + n\sigma', \quad \mu_n = \lambda + n\tau' + \frac{1}{2}n(n-1)\sigma''. \end{aligned} \tag{3.45}$$

Если при этом $\mu_k \neq 0$ для $k = 0, 1, \ldots, n-1$, то всякое решение (3.45) можно представить в виде $v_n = y^{(n)}$, где $y(z)$ – некоторое решение уравнения (3.44).

Это свойство позволяет построить семейство частных решений (3.44), соответствующих определенным значениям λ. Действительно, уравнение (3.45) при

[32] Потому что это несколько более общая форма так называемого гипергеометрического уравнения.

$\mu_n = 0$ имеет очевидное частное решение $v_n(z) = \text{const}$. Так как $v_n(z) = y^{(n)}(z)$, то это означает, что при

$$\lambda = \lambda_n \equiv -n\tau' - \frac{1}{2}n(n-1)\sigma'' \tag{3.46}$$

существует частное решение уравнения г. типа $y = y_n(z)$, являющееся полиномом степени n. Будет называть такие решения *полиномами г. типа*. Они являются в известном смысле простейшими решениями уравнения (3.44).

3.2.4.1 Формула Родрига

Исходное уравнение (3.44) часто удобно записывать в "самосопряженном" виде

$$\frac{d}{dz}\Big[\sigma(z)\rho(z)\frac{dy}{dz}\Big] + \lambda\rho(z)y = 0, \tag{3.47}$$

где $\rho(z)$ удовлетворяет уравнению

$$[\sigma\rho(z)]' = \tau\rho(z). \tag{3.48}$$

В явном виде из него получаем

$$\rho = \frac{1}{\sigma}\exp\int\frac{\tau dz}{\sigma}. \tag{3.49}$$

Запишем, аналогично (3.47) и (3.48), в самосопряженном виде уравнения для $v_n \equiv y^{(n)}$:

$$\Big[\sigma\rho_n v_n'\Big]' + \mu_n\rho_n v_n = 0, \quad \text{где} \quad (\sigma\rho_n)' = \tau_n\rho_n. \tag{3.50}$$

Используя связь $\tau_n = \tau + n\sigma'$ (3.45), выразим ρ_n через $\rho \equiv \rho_0$:

$$\frac{\rho_n'}{\rho_n} = \frac{\rho'}{\rho} + n\frac{\sigma'}{\sigma},$$

откуда получаем

$$\rho_n = \rho\sigma^n.$$

Тогда $\rho_{n+1} = \sigma\rho_n$, так что уравнение (3.50), с учетом $v_{n+1} = v_n'$, можно переписать как рекуррентное соотношение

$$\rho_n v_n = -\frac{1}{\mu_n}\left(\rho_{n+1}v_{n+1}\right)'.$$

Если решение $y = y_n(z)$ есть полином степени n, то $v_n = \text{const}$, и

$$\rho y \equiv \rho_0 v_0 \sim (\rho_1 v_1)' \sim (\rho_2 v_2)'' \sim \ldots \sim (\rho_n v_n)^{(n)} \sim \rho_n^{(n)} \sim (\rho\sigma^n)^{(n)},$$

так что верна *формула Родрига*

$$y_n(z) = \frac{1}{\xi_n\rho(x)}\frac{d}{dz^n}\Big[\rho(z)\sigma^n(z)\Big], \tag{3.51}$$

где ξ_n есть некоторый постоянный множитель, определяющий нормировку.

3.2.4.2 Ортогональность. Задача Штурма-Лиувилля.

Пусть y_n и y_m – два решения уравнения (3.47), соответствующие $\lambda = \lambda_n$ и $\lambda = \lambda_m$:

$$[\sigma\rho y_n']' = -\lambda_n \rho y_n;$$
$$[\sigma\rho y_m']' = -\lambda_m \rho y_m.$$

Домножая первое уравнение на y_m, второе на y_n, и вычитая одно из другого, в левой части получим

$$y_m [\sigma\rho y_n']' - y_n [\sigma\rho y_m']' = \frac{d}{dz}\Big[\sigma\rho\, W[y_m, y_n]\Big],$$

где

$$W[u, v] = \begin{vmatrix} u & v \\ u' & v' \end{vmatrix}$$

– определитель Вронского.

Тогда интегрируя правую и левую части по z от a до b, получим

$$(\lambda_m - \lambda_n) \cdot (y_n, y_m) = \Big[\sigma\rho\, W[y_n, y_m]\Big]_a^b, \tag{3.52}$$

где скалярное произведение между функциями определено с весом $\rho(z)$ на $[a, b]$:

$$(f, g) = \int_a^b dz\, \rho(z)\, f(z) g(z). \tag{3.53}$$

Это соотношение приводит к двум следствиям.

Пусть мы решаем *задачу Штурм-Лиувилля (Sturm-Liouville problem) для уравнения (3.47)*: найти λ, при которых существуют нетривиальные решения однородной краевой задачи для уравнения с заданными граничными условиями на промежутке $[a, b]$

$$\alpha_1 y'(a) + \beta_1 y(a) = 0; \quad \alpha_2 y'(b) + \beta_2 y(b) = 0 \tag{3.54}$$

и $y_{n,m}$ – два решения (не обязательно полиномиальных), соответствующие $\lambda_{n,m}$.

Вследствие условий (3.54), которые фиксируют линейную связь между y и y' на концах промежутка интегрирования, W при $x = a, b$ обращается в ноль, и правая часть (3.52) равна нулю. Поэтому *решения с* $\lambda_n \neq \lambda_m$, что эквивалентно $n \neq m$, *взаимно ортогональны в смысле (3.53)*:

$$(y_n, y_m) \sim \delta_{nm}.$$

С другой стороны, тот же результат для произвольных *полиномиальных* решений (3.44) получим, если σ и ρ удовлетворяют условию

$$\sigma(z)\rho(z)z^k\Big|_{z=a,b} = 0, \quad \text{для} \quad k = 0, 1, \ldots \tag{3.55}$$

Таким образом, если σ и ρ удовлетворяют условию (3.55), то полиномиальные решения (3.44) образуют на $[a, b]$ ортогональную систему полиномов с весом $\rho(z)$ и следовательно ортогональный базис соответствующего пространства $L^2_{[a,b]}(\rho)$. Их можно записать в виде (3.37), для них выполняются рекуррентные соотношения (3.39) и свойства нулей (п.3.2.3). Эти же решения являются собственными для соответствующей задачи Штурма-Лиувилля.

3.2.4.3 Полиномы Эрмита, Лагерра, Якоби

До настоящего момента мы не фиксировали σ и τ – это лишь должны были быть многочлены не выше второй и первой степени соответственно. Рассмотрим разные σ и τ, и каким системам ортогональных полиномов они соответствуют.

Вначале отметим, что мы всегда можем упростить уравнение (3.44), а) домножив его на какое-то число, фиксируя, таким образом, например, старший коэффициент у σ и б) сделав линейную замену переменной, зафиксировав каких-либо еще два параметра.

Полиномы Эрмита (Hermite). Пусть τ – линейная функция, а σ – константа:

$$\sigma = \sigma_0; \qquad \tau = \tau_0(x - x_0).$$

Интегрируя (3.49) с такими τ и σ, получаем весовую функцию в явном виде:

$$\rho \sim \exp\left(\frac{\tau_0}{\sigma_0}\left(\frac{x^2}{2} - xx_0\right)\right).$$

Очевидно, $\rho\sigma$ обращается в ноль на $\pm\infty$, и только если $\tau_0/\sigma_0 < 0$. Используя свободу в выборе коэффициентов, положим $\sigma_0 = 1$, $\tau_0 = -2$, $x_0 = 0$, так чтобы вес имел максимально простой вид[33]:

$$\rho = e^{-x^2}; \qquad a = -\infty,\ b = +\infty.$$

[33]Весовая функция вообще определяется с точностью до числового множителя, так что он фиксируется из соображений удобства

Полиномы, ортогональные с таким скалярным произведением, называются полиномами Эрмита, и обозначаются $H_n(x)$.

Уравнение для них получаем подстановкой $\sigma = 1$ и $\tau = -2x$ в уравнение общего вида (3.44); параметр λ находим из (3.46), так что уравнение для полиномов Эрмита имеет вид

$$H_n'' - 2xH_n' + 2nH_n = 0.$$

Оно возникает в квантово-механической задаче для одномерного гармонического осциллятора, так что ψ-функция оказывается равна

$$\psi_\omega(\xi) = \text{const} \cdot e^{-\xi^2/2} H_n(\xi), \quad \text{где} \quad \xi = \sqrt{\frac{m\omega}{\hbar}}x; \quad E_n = \hbar\omega(n + 1/2).$$

Интересно также отметить, что функции ψ_ω являются собственными векторами непрерывного преобразования Фурье.

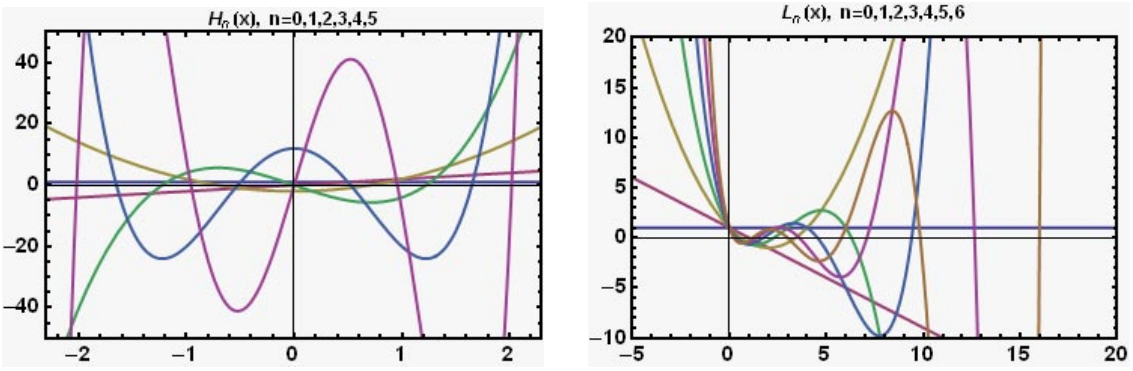


Рис. 3.3: Первые шесть полиномов Эрмита H_n (слева) и семь полиномов Лагерра $L_n^{(0)}$ (справа).

Полиномы Лагерра (Laguerre). Пусть теперь σ – линейная функция:

$$\sigma = \sigma_0(x - x_1); \qquad \tau = \tau_0(x - x_0).$$

Интегрируя (3.49) с такими τ и σ, получаем:

$$\rho\sigma \sim |x - x_1|^{\frac{\tau_0}{\sigma_0}(x_0 - x_1)} e^{\frac{\tau_0}{\sigma_0}x}.$$

Понятно, что $\rho\sigma$ обращается в ноль в x_1 (если степень положительна) и на $\pm\infty$ (в зависимости от знака τ_0/σ_0) . Используя свободу в выборе коэффициентов, фиксируем $\sigma_0 = 1$, $\tau_0 = -1$ и $x_1 = 0$; тогда вес принимает вид:

$$\rho \sim x^\alpha e^{-x}; \qquad a = 0,\ b = +\infty.$$

Полиномы, ортогональные с таким скалярным произведением, называются полиномами Лагерра, и обозначаются $L_n^{(\alpha)}(x)$. Коэффициент $\alpha = x_0 - 1$ остался свободным – он параметризует получившееся семейство полиномов.

Условие $\alpha > -1$, с одной стороны, обеспечивает обращение в ноль $\rho\sigma$ в $x = 0$, а с другой стороны, возникает как требование интегрируемости $L_n^{(\alpha)}$ в нуле – для ортогональности системы полиномов, очевидно, необходимо, чтобы соответствующие попарные скалярные произведения существовали.

Уравнение для полиномов Лагерра получаем так же, подстановкой $\sigma = x$ и $\tau = -(x - \alpha - 1)$ в уравнение общего вида (3.44), а λ находим из (3.46):

$$xL_n^{(\alpha)\prime\prime} + (\alpha + 1 - x)L_n^{(\alpha)\prime} + nL_n^{(\alpha)} = 0; \qquad \alpha > -1.$$

$L_n^{(\alpha)}(x)$ также называют обобщенными или присоединенными (associated) полиномами Лагерра, тогда просто полиномами Лагерра называют $L_n^{(\alpha)}$ с $\alpha = 0$. Они возникают в квантовомеханической задаче о движении в центрально-симметричном кулоновском поле и описывают радиальную зависимость волновой функции электрона в кулоновском поле ядра:

$$\psi_{nlm} \sim Y_l^m(\Theta, \varphi) \cdot e^{-\rho/2}\rho^l L_{n-l-1}^{(2l+1)}(\rho), \quad \text{где} \quad \rho = \frac{2r}{na_0}.$$

Полиномы Якоби (Jacobi). Наконец, пусть σ – квадратный трехчлен:

$$\sigma = \sigma_0(x - x_1)(x - x_2); \qquad \tau = \tau_0(x - x_0).$$

Интегрируя (3.49) с такими τ и σ, получаем:

$$\rho\sigma \sim |x - x_1|^{\beta+1}|x - x_2|^{\alpha+1},$$

где α и β – некоторые числовые коэффициенты, выражающиеся линейно через $x_{0,1,2}$ и τ_0/σ_0.

Функция $\rho\sigma$ обращается в ноль в $x_{1,2}$ если $\alpha, \beta > -1$. Используя свободу в выборе коэффициентов, фиксируем $\sigma_0 = 1$ и $x_{1,2} = \pm 1$; тогда вес принимает вид:

$$\rho = (1 - x)^\alpha(1 + x)^\beta; \qquad a = -1,\ b = +1.$$

Полиномы, ортогональные с таким скалярным произведением, называются полиномами Якоби, и обозначаются $P_n^{(\alpha,\beta)}(x)$ – они характеризуются двумя параметрами.

Выразив τ через α и β и подставив σ и τ в уравнение общего вида, получим уравнение для полиномов Якоби:

$$(1 - x^2)y'' + \big[\beta - \alpha - x(\alpha + \beta + 2)\big]y' + n(n + \alpha + \beta + 1)y = 0.$$

Частным случаем их при $\alpha=\beta$ являются ультрасферические (ultraspherical) полиномы, которые также называют полиномами *Гегенбауэра (Gegenbauer)* C_n^α.

Полиномы Лежандра (Legendre) P_n (или L_n) это простейший частный случай полиномов Якоби (а также ультрасферических п.), при $\alpha=\beta=0$, т.е. они ортогональны с весом

$$\rho(x)=1 \quad \text{на} \quad [-1,1]; \quad (1-x^2)P_n'' - 2xP_n' + n(n+1)P_n = 0.$$

Это уравнение возникает при решении уравнения Лапласа в сферических координатах – через них записывается угловая часть.

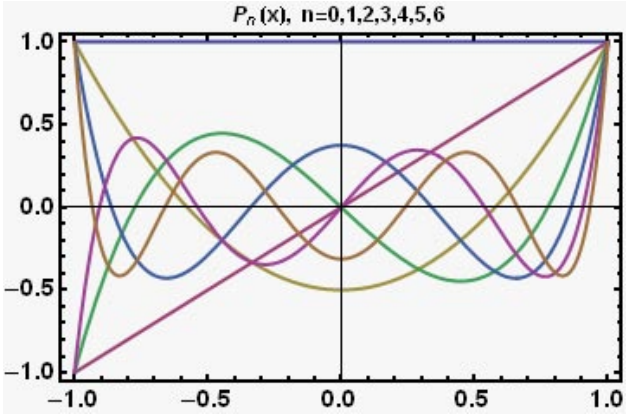
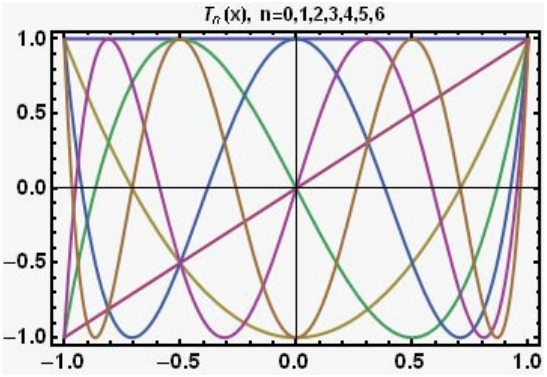


Рис. 3.4: Первые семь полиномов Лежандра на $[-1,1]$

3.2.4.4 Полиномы Чебышёва I и II рода



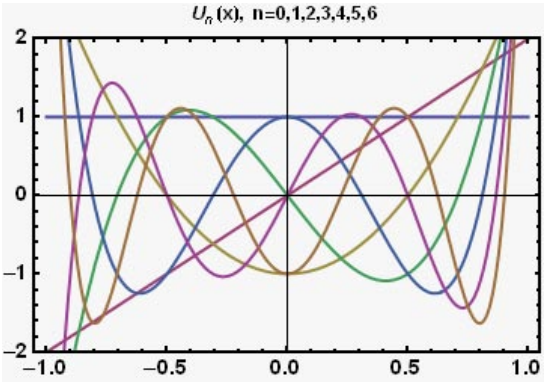


Рис. 3.5: Полиномы Чебышёва T_n и U_n

Полиномы Чебышёва I и II рода также являются частными случаями полиномов Якоби, и получаются при $\alpha=\beta=-1/2$ и при $\alpha=\beta=+1/2$ соответственно.

Полиномы Чебышёва I рода T_n ортогональны с весом

$$\rho(x)=\frac{1}{\sqrt{1-x^2}} \quad \text{на} \quad [-1,1]; \quad (1-x^2)T_n'' - xT_n' + n^2T_n = 0.$$

Полиномы Чебышёва II рода U_n ортогональны с весом

$$\rho(x) = \sqrt{1-x^2} \quad \text{на} \quad [-1,1]; \quad (1-x^2)U_n'' - 3xU_n' + n(n+2)U_n = 0.$$

T_n это это те же самые полиномы, которые мы ввели как $T_n^{tr} = \cos(n \arccos x)$ когда рассматривали интерполяцию Лагранжа в п. 1.2.4.4:

◂ Чтобы это доказать, достаточно убедиться в том, что T_n^{tr} ортогональны в соответствующем скалярном произведении:

$$\begin{aligned}
\int_{-1}^{1} \frac{dx\, T_n^{tr}(x) T_m^{tr}(x)}{\sqrt{1-x^2}} &= \int \frac{dx \cos(n \arccos x) \cos(m \arccos x)}{\sqrt{1-x^2}} = \\
&= \int \frac{d\cos\Theta \, \cos n\Theta \cos m\Theta}{\sin\Theta} = \\
&= \int_0^{\pi} dx \cos\, nx \cos\, mx = \delta_{mn} \begin{cases} \pi & \text{если } n=0; \\ \pi/2 & \text{если } n \geq 1. \end{cases}
\end{aligned}$$

Тогда требуя $T_n \equiv T_n^{tr}$, мы фиксируем нормировку T_n. Видно, что в принятой нормировке $T_n(1) = T_n^{tr}(1) = \cos 0 = 1$. Заодно вывели квадрат нормы. ▸

Как упоминалось ранее (п.1.2.4.5), полиномы Чебышёва I рода также называют полиномами, наименее уклоняющимися от нуля. Это означает, что T_n минимизируют на $[-1,1]$ норму L^∞, т.е. $\max\limits_{[-1,1]} |P_n(x)|$ среди приведенных полиномов степени n.

Полиномы Чебышёва T_n и тригонометрический ряд Фурье* Из первой теоремы Вейерштрасса следует, что множество полиномов всюду плотно в пространстве $C_{[-1,1]}$ непрерывных функций на $[-1,1]$ в метрике $L^2\left[(1-x^2)^{-1/2}\right]$. Поэтому, как мы видели, всякую функцию $f \in C_{[-1,1]}$ можно представить в виде ее ряда Фурье по полиномам Чебышёва I рода T_n:

$$f(x) = \sum_{k=0}^{\infty} f_k T_k(x), \quad T_k(x) = \cos(k \arccos x), \quad f_k = \frac{1}{\|T_k\|^2} \int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} f(x) T_k(x),$$

где $\|T_0\|^2 = \pi$, а для $k > 1$: $\|T_k\|^2 = \pi/2$.

Сделаем замену переменных $x = \cos\Theta$, $\Theta \in [0,\pi]$. Тогда $T_k = \cos(k\Theta)$, и получаем для $F(\Theta) \equiv f(\cos\Theta)$ ряд

$$F(\Theta) = \sum_{k=0}^{\infty} F_k \cos k\Theta,$$

где

$$F_k = f_k = \frac{1}{\|T_k\|^2} \int_0^{\pi} d\Theta F(\Theta) \cos k\Theta.$$

Учитывая, что по построению F четна $F(\Theta) = F(-\Theta)$, видим, что

$$F_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\Theta F(\Theta); \quad F_k = \frac{1}{\pi}\int_{-\pi}^{\pi} d\Theta F(\Theta)\cos k\Theta \quad \text{для} \quad k > 1.$$

Это в точности косинус-преобразование Фурье – тригонометрический ряд Фурье (1.44) для четной функции, для которой все коэффициенты при синусах равны нулю.

Таким образом, ряд по полиномам Чебышёва с точностью до замены переменной совпадает с тригонометрическим рядом Фурье. Из его сходимости следует, что тригонометрические полиномы всюду плотны на множестве $C[0, 2\pi]$ в метрике $L^2(1)$. Вторая теорема Вейерштрасса (см. п.3.1.9) делает еще более сильное утверждение о плотности в равномерной метрике.

И обратно – все хорошие свойства тригонометрических рядов Фурье, таким образом, прямо переносятся на ряды по полиномам Чебышёва.

3.3 Среднеквадратичное приближение
Least squares approximation

3.3.1 Приближение функций, заданных таблично

3.3.1.1 Постановка задачи

Рассмотрим вначале дискретный вариант. Постановка задачи следующая. Пусть у нас есть набор точек $\{x_k\}_1^m$, функция, определенная на этих точках, то есть числа $\{f_k = f(x_k)\}_1^m$, и набор положительных чисел $\{p_k > 0\}_1^m$, называемых весами, или весовыми коэффициентами.

В m-мерном пространстве V_m функций, заданных на сетке $\{x_k\}_{k=1}^m$, введем скалярное произведение и соответствующую норму

$$(f, g) = \sum_{k=1}^{m} p_k f(x_k) g(x_k), \quad \|f\|^2 = (f, f). \tag{3.56}$$

Норма индуцирует метрику

$$\rho^2(f, g) = \|f - g\|^2 = \sum_{k=1}^{m} p_k \left[f(x_k) - g(x_k)\right]^2.$$

Пусть у нас также задана система функций $\{\varphi_i(x)\}_1^n$, сужение которой на $\{x_i\}$ линейно-независимо[34]. Линейная оболочка $\{\varphi_i\}_1^n$ $(n < m)$ образует n-мерное

[34] Простейший пример $\{1, x, \ldots, x^{n-1}\}$, для $n < m$. Если же $n > m$, то такая система линейно-зависима на сетке $\{x_i\}_1^m$, просто потому, что число функций, n, больше размерности пространства m.

подпространство V_m, которое мы обозначим как S. Будем искать элемент этого подпространства, то есть линейную комбинацию $\varphi(x) = \sum\limits_{i=1}^{n} c_i\varphi_i$, для которой отклонение от f в введенной метрике минимально:

$$(f-\varphi, f-\varphi) \equiv \sum_{k=1}^{m} p_k(f_k - \varphi(x_k))^2 = \min. \tag{3.57}$$

Решение задачи на минимизацию (3.57) называется *элементом наилучшего приближения* в S для f.

Такая постановка задачи естественна, к примеру, если у нас есть функция f, заданная дискретно, например результаты каких-то измерений, для которой нужно подобрать аналитическую зависимость. В связи с тем, что у измерения всегда есть какая-то ошибка, делать интерполяцию, и требовать чтобы искомая функция в точках x_i принимала в точности значения f_i, часто оказывается весьма бессмысленно. Гораздо более привлекательным представляется сделать приближение "наименьшими квадратами". При этом, положим, мы больше уверены в точности измерений в одних точках, и меньше в других.

Веса $\{p_k\}$ как раз и определяют важность попадания приближения φ в окрестность каждой точки $\{x_i, f_i\}$: если для некоторого узла i вес p_i больше чем для других, то при равных абсолютных отклонениях φ от f в каждом узле $|f_i - \varphi(x_i)|$ доля (вес) соответствующего слагаемого $p_i(f_i - \varphi(x_i))^2$ в отклонении велика, – так что отклонение минимизируется, когда $|f_i - \varphi(x_i)|$ меньше, чем абсолютные значения отклонения в других узлах.

Задача "на наименьшие квадраты" корректна при $n < m$; при $n = m$ набор $\{\varphi_i\}$ образует базис V_m, по которому можно разложить $f = \sum f_i\varphi_i$, и при $c_i = f_i$ отклонение будет в точности равно нулю, так что задача вырождается в задачу интерполяции; аппроксимация с $n > m$ бессмысленна, т.к. интерполяционных решений и то бесконечно много.

3.3.1.2 Элемент наилучшего приближения

Подпространством конечномерного векторного пространства V называется любое его подмножество L, такое что если $x, y \in L$, то и $\alpha x + \beta y \in L$ для любых чисел α, β.

Ортогональным дополнением к L в V называется множество

$$L_\perp = \{x \in V \,|\forall a \in L \; x \perp a\}.$$

Несложно показать, что оно также является подпространством, и что $\left(L_\perp\right)_\perp = L$.

Т^0: Если L – подпространство V, то $\forall\, x \in V$ может быть представлен в виде

$$x = x_{\|} + x_{\perp}, \quad \text{где} \quad x_{\|} \in L, \quad x_{\perp} \in L_{\perp}$$

и указанное представление единственно. Элемент $x_{\|}$ называется *ортогональной проекцией* x на подпространство L, а $x_{\perp}$ тогда есть ортогональная проекция x на $L_{\perp}$.

◂ Пусть $\{\varphi_i\}_{i=1}^n$ – базис L. Предположим, что x можно представить в виде

$$x = \sum_{i=1}^{n} \xi_i \varphi_i + x_{\perp}, \text{где} \quad x_{\perp} \in L_{\perp}.$$

Домножая правую и левую часть на φ_j, $j = 1, \ldots, n$, получим систему

$$(x, \varphi_j) = \sum_{i=1}^{n} \xi_i (\varphi_i, \varphi_j).$$

Эта система относительно ξ_i разрешима и ее решение единственно, так как ее определитель есть определитель Грама для $\{\varphi_i\}$, которые линейно-независимы. Таким образом, $x_{\|} = \sum \xi_i \varphi_i$, и $x_{\perp} = x - x_{\|}$. ▸

Т^0: *Элемент наилучшего приближения в S для f есть ортогональная проекция f на S.*

◂ Представим f в виде $f = f_{\|} + f_{\perp}$, где $f_{\|} \in S$, $f_{\perp} \in S_{\perp}$. Элемент наилучшего приближения φ и $f_{\|}$ разложим по базису S:

$$\varphi = \sum c_i \varphi_i, \quad f_{\|} = \sum f_i \varphi_i.$$

Тогда

$$\|f - \varphi\|^2 = \left(f_{\perp} + \sum (f_i - c_i)\varphi_i\right)^2 = f_{\perp}^2 + \left(\sum (f_i - c_i)\varphi_i\right)^2,$$

где подразумевается $f^2 \equiv (f, f)$. Первое слагаемое не зависит от коэффициентов c_i, а второе достигает минимума, нуля, т.и т.т., когда $c_i = f_i$. Тогда $\varphi = f_{\|}$, ч.и т.д.

При этом

$$\varphi(x) = \sum_{i=1}^{n} f_i \varphi_i(x), \tag{3.58}$$

а коэффициенты f_i находим из системы

$$(f, \varphi_j) = \sum_{i=1}^{n} f_i (\varphi_i, \varphi_j), \qquad j = 1, \ldots, n. \qquad \blacktriangleright \tag{3.59}$$

Процедура сильно упрощается, если система $\{\varphi_i\}_1^n$ ортогональна в смысле (3.56). Например, если взять $\{x^i\}$ и ортогонализовать в смысле (3.56) с некоторым весом p_k, получим так называемые *ортогональные полиномы дискретной*

переменной[35]. Тогда из (3.58) и (3.59) получим

$$\varphi(x) = \sum_{i=1}^{n} \frac{(f, \varphi_i)}{(\varphi_i, \varphi_i)} \varphi_i(x). \tag{3.60}$$

Пример

Нужно найти наилучшее приближении для $\sin \pi x$ на $x \in [-1, 1]$ среди многочленов степени не выше 3, используя значения функции в точках $-1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ (с весом 1).

Ищем полином $P_3(x) = C_0 + C_1 x + C_2 x^2 + C_3 x^3$. Т.к. наша система функций $\{\varphi_i = x^i\}_{i=0}^3$ не ортогональна, ищем коэффициенты из системы общего вида (3.59):

$$(y, \varphi_i) = \sum_{i=0}^{3} C_i (\varphi_i, \varphi_j) \quad \text{для} \quad j = 0, 1, 2, 3.$$

Значения функций в указанных точках и попарные скалярные произведения:

x	-1	-1/2	0	1/2	1
y	0	-1	0	1	0
φ_0	1	1	1	1	1
φ_1	-1	-1/2	0	1/2	1
φ_2	1	1/4	0	1/4	1
φ_3	-1	-1/8	0	1/8	1

	φ_0	φ_1	φ_2	φ_3
φ_0	5	0	5/2	0
φ_1		5/2	0	17/8
φ_2			17/8	0
φ_3				65/32
y	0	1	0	1/4

Подставляя все это в систему, получим

$$\begin{cases} 0 = 5\,C_0 + 5/2\,C_2 \\ 1 = 5/2\,C_1 + 17/8\,C_3 \\ 0 = 5/2\,C_0 + 17/8\,C_2 \\ 1/4 = 17/8\,C_1 + 65/32\,C_3 \end{cases} \Rightarrow \ldots \Rightarrow \begin{array}{l} C_2 = 0 \\ C_0 = 0 \\ C_1 = 8/3 \\ C_3 = -8/3 \end{array} \Rightarrow P_3(x) = \frac{8}{3}(x - x^3).$$

3.3.2 Приближение в L^2

При большом числе узлов m становится логичнее переформулировать задачу среднеквадратичного приближения в терминах непрерывных функций, заданных на некотором промежутке $[a, b]$:

[35] См. брошюру Никифорова [16] и (3.82).

Пусть на отрезке $[a,b]$ задана некоторая вещественнозначная функция $f(x)$, вес (весовая функция) $p(x) > 0$ и набор линейно-независимых функций $\{\varphi_i(x)\}_1^n$. В пространстве функций определим скалярное произведение как

$$(f,g) = \int_a^b dx\, p(x)\, f(x) g(x).$$

Ищем линейную комбинацию

$$\varphi(x) = \sum_{i=1}^{n} c_i \varphi_i,$$

которая минимизирует отклонение от f в соответствующей такому скалярному произведению метрике:

$$\rho^2(f,\varphi) \equiv \int_a^b dx\, p(x) \left[f(x) - \varphi(x)\right]^2 = \min. \tag{3.61}$$

Задача корректна, если $f, \{\varphi_i\} \in L^2_{[a,b]}(p)$; для простоты можно считать, что все эти функции непрерывны.

Можно поставить задачу и так, что система функций $\{\varphi_i\}$ изначально не задана. Тогда ее еще необходимо подобрать удобным образом.

3.3.2.1 Приближение в гильбертовом пространстве

Понятно, что рассмотренная выше дискретная задача переходит в непрерывную в пределе $m \to \infty$. Приведенные для дискретной задачи рассуждения непосредственно обобщаются на бесконечномерный случай благодаря тому уникальному свойству гильбертова пространства, что в нем, так же как в конечномерных пространствах, всякий вектор можно разложить по базису. Лишь немного меняются определения и усложняются доказательства.

Элементом наилучшего приближения для $f \in H$ в подпространстве[36] $S = \mathrm{span}\{\varphi_1, \ldots, \varphi_n\}$ называется функция $\varphi \in S$, такая что

$$\|f - \varphi\| = \inf_{s \in S} \|f - s\|.$$

Линейным многообразием[37] гильбертова пространства H назовем любое его подмножество L, такое что если $x, y \in L$, то и $\alpha x + \beta y \in L$ для любых чисел α, β.

[36] $\mathrm{span}\{\varphi_1, \ldots, \varphi_n\}$ – линейная оболочка $\{\varphi_i\}_{i=1}^n$.

[37] Бывает и другое определение линейного многообразия.

Подпространство – замнутое (т.е. содержащее все свои предельные точки) линейное многообразие[38].

Всякое конечномерное линейное многообразие замкнуто и следовательно образует подпространство. Замкнув всякое незамкнутое линейное многообразие, получим подпространство. Всякое бесконечномерное подпространство H можно рассматривать как самостоятельное гильбертово пространство.

Ортогональным дополнением к подпространству L в H называется множество

$$L_\perp = \{x \in H \,|\forall a \in L \;\; x \perp a\}.$$

Оно также является подпространством, а ортогональное дополнение к ортогональному дополнению к L совпадает с L ◂…▸.

T^0: Если L – подпространство H, то $\forall\, x \in H$ может быть представлен в виде

$$x = x_{\|} + x_\perp, \quad \text{где} \quad x_{\|} \in L, \quad x_\perp \in L_\perp$$

и указанное представление единственно ◂…▸.

Здесь $x_{\|}$ есть ортогональная проекция x на L, а $x_\perp$ – на $L_\perp$.

T^0: Элемент наилучшего приближения φ для f в S – ортогональная проекция f на S. Он существует и единственен.

◂ Доказательство дословно повторяет приведенное для дискретного случая, и (см. (3.58), (3.59))

$$\varphi(x) = \sum_{i=1}^{n} f_i \varphi_i(x),$$

где f_i есть решение системы линейных уравнений

$$(f, \varphi_j) = \sum_{i=1}^{n} f_i (\varphi_i, \varphi_j), \quad j = 1, \ldots, n. \quad \blacktriangleright \tag{3.62}$$

Если исходная система $\{\varphi_i\}_{i=1}^n$ ортогональна, то система уравнений (3.62) тривиальна $(f, \varphi_j) = \|\varphi_j\|^2 f_j$, и элемент наилучшего приближения (3.58) определяется суммой

$$\varphi = \sum_{i=1}^{n} f_i \varphi_i, \quad \text{где } f_i = \frac{(f, \varphi_i)}{\|\varphi_i\|^2} \equiv \frac{1}{\|\varphi_i\|^2} \int d\xi\, p(\xi) f(\xi) \varphi_i(\xi). \tag{3.63}$$

Если при этом $\{\varphi_i\}_1^n$ является частью бесконечной системы $\{\varphi_i\}_1^\infty$, которая *замкнута*, то этот ряд представляет собой частичную сумму ряда Фурье для f по $\{\varphi_i\}_1^\infty$.

[38] Пример незамкнутого многообразия – $C_{[a,b]}$, так как оно неполно в квадратичной метрике, а значит как подмножество L^2 – незамкнуто.

3.3.2.2 Оценка точности

Квадрат отклонения приближения φ от точной функции f:

$$\delta_n^2 \equiv \|f_\perp\|^2 = (f - \sum_i f_i\varphi_i)^2 = f^2 - 2\sum_i f_i(f,\varphi_i) + \Big(\sum_i f_i\varphi_i\Big)^2 =$$
$$= f^2 - 2\sum_i f_i \sum_j f_j(\varphi_j,\varphi_i) + \Big(\sum_i f_i\varphi_i, \sum_j f_j\varphi_j\Big) = f^2 - \sum_{i,j} f_i f_j(\varphi_i,\varphi_j).$$

В случае ортогональной системы $(\varphi_i,\varphi_j) = \|\varphi_i\|^2\delta_{ij}$ и $f_i = (f,\varphi_i)/\|\varphi_i\|^2$. Тогда выражение сильно упрощается

$$\delta_n^2 = f^2 - \sum_{i=1}^n \frac{f_i^2}{\|\varphi_i\|^2} = (f,f) - \sum_{i=1}^n \frac{(f,\varphi_i)^2}{(\varphi_i,\varphi_i)}. \tag{3.64}$$

В развернутом виде, через интегралы, оно имеет вид

$$\delta_n^2 = \int_a^b dx\, p(x)f^2(x) - \sum_{i=1}^n \frac{\Big(\int_a^b dx\, p(x)f(x)\varphi_i(x)\Big)^2}{\int_a^b dx\, p(x)\varphi_i^2(x)}. \tag{3.65}$$

Для ортонормированной системы $\|\varphi_i\| = 1$ и

$$\delta_n^2 = f^2 - \sum_{i=1}^n f_i^2 = \int_a^b dx\, p(x)f^2(x) - \sum_{i=1}^n \left(\int_a^b dx\, p(x)f(x)\varphi_i(x)\right)^2. \tag{3.66}$$

3.3.2.3 Различные варианты постановки задачи

До сих пор шла речь о задаче, в которой набор функций $\{\varphi_i\}_1^n$ задан, и число их n фиксировано. Если $\varphi_i = x^{i-1}$, то такая задача переформулируется как "построить средне-квадратичное приближение функции f с весом $p(x)$ многочленами степени не выше n". Можно ее решать по общей схеме, а можно сначала ортогонализовать полиномы, и потом раскладывать f по получившейся ортогональной системе полиномов. Эти задачи эквивалентны по сложности, но если вес один из стандартных, то ортогональная система известна заранее (полиномы Якоби, Эрмита и пр. – все сведения о них, что могут понадобиться, есть в справочниках) и задача существенно упрощается. Если же вес нестандартен, но необходимо делать многократные приближения для разных функций, то также выгоднее один раз ортогонализовать $\{\varphi_i\}$, и потом для каждой функции использовать упрощенную формулу (3.63).

Часто возникает задача другого рода: построить приближение "средними квадратами" заданной функции f с весом $p(x)$ функциями из бесконечного

набора $\{\varphi_i\}_1^\infty$ с заданной точностью. Такая постановка задачи имеет смысл, если набор $\{\varphi_i\}_1^\infty$ образует замкнутую систему (возможно, не ортогональную) в соответствующем гильбертовом пространстве. В таком случае решать каждый раз с увеличением n заново систему уравнений с матрицей $n \times n$ очень неэффективно.

Естественно искать последовательные приближения, ортогонализуя систему $\{\varphi_i\}$ и параллельно строя разложение f в ряд Фурье по получившейся ортогонализованной системе. На каждом шаге $n \to n+1$ процедурой Грама-Шмидта строится один вектор $\tilde{\varphi}_n$ ортогональной системы, считается соответствующий член ряда Фурье по ней для f (3.63) и новое слагаемое суммы (3.63) для отклонения. Процесс продолжается до тех пор, пока отклонение не станет меньше заданного значения.

Если речь идет о системе полиномов, и вес $p(x)$ один из стандартных, то следует сразу взять готовую систему классических ортогональных полиномов. Если вес нестандартен, то систему придется строить самому.

Может вполне оказаться такая ситуация, что вес не задан по условию, а мы сами задаем его из каких-то физических соображений. Если речь идет о конечном промежутке, то линейной заменой переменных он сводится к $[-1, 1]$. *В качестве весовой функции нужно выбирать такую $p(x)$, которая больше на участке, на котором для нас более важна точность.* Тогда абсолютное отклонение φ от f на этом участке даст больший вклад в интеграл (3.61), и таким образом, в расстояние между ними по норме $L^2(p)$, так же как и в дискретном случае.

Так, если все участки на отрезке для нас одинаково важны, то можно брать $p = 1$ и пользоваться многочленами Лежандра. Если для нас важнее чтобы функция хорошо аппроксимировалась на концах промежутка, нам больше подойдет $p = (1 - x^2)^{-1/2}$ и многочлены Чебышёва I рода; если больше важна аппроксимация посередине промежутка, то $p = (1-x^2)^{+1/2}$ и многочлены Чебышёва II рода. Если один конец гораздо важнее чем второй – берем полиномы Якоби с разными α и β.

На полубесконечном и бесконечном интервале годятся полиномы Эрмита и Лагерра. Присоединенные полиномы Лагерра $L_n^{(\alpha)}$ позволяют придавать больший вес значениям x вблизи $x = \alpha$, где весовая функция $x^\alpha e^{-x}$ достигает максимума.

3.3.2.4 Примеры

Пример 1. Построить среднеквадратичное приближение функции $f(x) = |x|$ на вещественной оси $(-\infty, \infty)$ многочленами степени не выше 5, с весом $p = e^{-x^2}$.

Воспользуемся полиномами Эрмита, которые ортогональны в соответствующем скалярном произведении

$$\int_{-\infty}^{\infty} dx\, e^{-x^2} H_n(x) H_m(x) = \delta_{nm} \|H_n\|^2.$$

Квадрат нормы и первые шесть полиномов посмотрим[39] в приложении C:

$$\begin{array}{lll} H_0 = 1 & H_3 = 8x^3 - 12x & \\ H_1 = 2x & H_4 = 16x^4 - 48x^2 + 12 & \|H_n\|^2 = \sqrt{\pi}\, 2^n n! \\ H_2 = 4x^2 - 2 & H_5 = 32x^5 - 160x^3 + 120x & \end{array}$$

Наша функция f четная, а в H_n чередуются четные и нечетные полиномы. Поэтому $(f, H_1) = (f, H_3) = (f, H_5) = 0$ и

$$f(x) \approx H_0(x) \frac{(f, H_0)}{\|H_0\|^2} + H_2(x) \frac{(f, H_2)}{\|H_2\|^2} + H_4(x) \frac{(f, H_4)}{\|H_4\|^2}.$$

Считаем сначала

$$(f, x^{2n}) = \int_{-\infty}^{\infty} dx e^{-x^2} |x| x^{2n} = 2 \int_0^{\infty} dx\, x^{2n+1} e^{-x^2} = \int_0^{\infty} d\xi\, \xi^n e^{-\xi} = n!,$$

так что скалярные произведения $(f, 1) = 0! = 1$, $(f, x^2) = 1! = 1$, $(f, x^4) = 2! = 2$. Тогда

$$\begin{aligned} &(f, H_0) = 1; \\ &(f, H_2) = (f, 4x^2 - 2) = 4 - 2 = 2; \\ &(f, H_4) = (f, 16x^4 - 48x^2 + 12) = 16 \cdot 2 - 48 + 12 = -4 \end{aligned}$$

и для среднеквадратичного приближения f, с учетом нормы $\|H_n\|$, получаем

$$\begin{aligned} f \approx \frac{H_0}{\sqrt{\pi}} + \frac{2H_2}{8\sqrt{\pi}} - \frac{4H_4}{128 \cdot 3\sqrt{\pi}} = \frac{3 \cdot 32 + 3 \cdot 8(4x^2 - 2) - (16x^4 - 48x^2 + 12)}{3 \cdot 32\sqrt{\pi}} = \\ = \frac{1}{3 \cdot 32\sqrt{\pi}}(-16x^4 + 16 \cdot 9x^2 + 36) = \frac{1}{24\sqrt{\pi}}(-4x^4 + 36x^2 + 9). \end{aligned}$$

Построив график $f(x)$ и найденного приближения, можно видеть, что аппроксимация правильная.

[39] Или можно взять из справочника, или на wiki, или на MathWorld, или на wolfram|alpha, или по строчке "HermiteH[5,x]" в Mathematica, или по строчке "hermite(5,x)" в Maxima...

Пример 2*. Построить среднеквадратичное приближение функции $f(x) = \frac{e^{x^2}}{1+25x^2}$ на промежутке $[-1,1]$ многочленами степени не выше 8, 12 и 16, с весом $p = 1$.

Для аппроксимации используем полиномы Лежандра, ортогональные на $[-1,1]$ с весом 1. Очевидно, что для решения этой задачи нужно посчитать с большой точностью много нехороших интегралов. Поэтому мы, дабы не перетрудиться, посчитаем все на компьютере.

Во-первых, введем в mathematica функцию f и определим частичную сумму ряда Фурье для нее, с m-того по n-тое слагаемое:

```
In[1]:= f[x_] = E^x^2 / (1 + 25 * x^2);
        LeastSquaresLegendre[x_, n_, m_] :=
          Sum[LegendreP[i, x] * NIntegrate[f[z] * LegendreP[i, z], {z, -1, 1}] * (2 i + 1)/2, {i, n, m}]
```

Во-вторых, считаем частичные суммы ряда Фурье до $n = 8, 12, 16$

```
In[3]:= LSALeg8[x_] = Expand[LeastSquaresLegendre[x, 0, 8]] // Quiet
        LSALeg12[x_] = Expand[LSALeg8[x] + LeastSquaresLegendre[x, 9, 12]] // Quiet
        LSALeg16[x_] = Expand[LSALeg12[x] + LeastSquaresLegendre[x, 13, 16]] // Quiet
```

В-третьих и последних, рисуем графики:

```
In[6]:= Plot[{f[x], LSALeg8[x], LSALeg12[x], LSALeg16[x]},
          {x, -1, 1}, PlotRange → {0, 1}, PlotStyle → Thickness[.004]]
```

Получаем картинку 3.6.

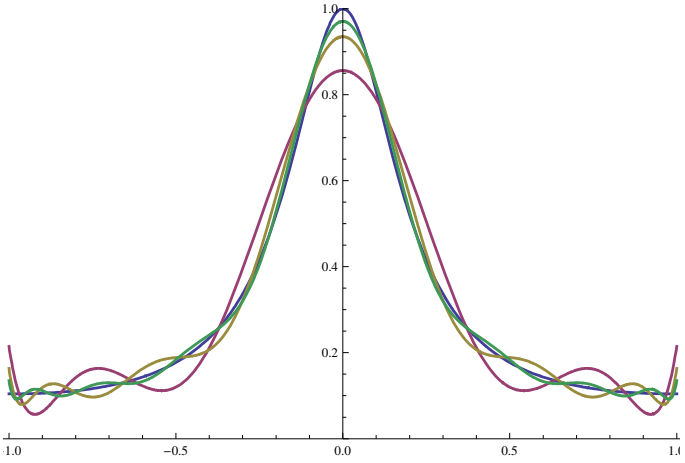


Рис. 3.6: Приближение $f(x)$ полиномами Лежандра степени не выше $8, 12, 16$ в метрике $L^2_{[-1,1]}(1)$.

3.3.3 Еще одна постановка задачи. Пример*

Встречается еще такой вариант постановки задачи о приближении наименьшими квадратами. Обсудим его на примере.

Пусть необходимо построить многочлен, который *равномерно* приближает функцию $f(x) = \ln(17/16 - x/2)$ на отрезке $[-1, 1]$ с точностью $\varepsilon < 0,0005$.

Для приближения выберем полиномы Чебышёва I рода. И вот почему. Можно заметить, что при $|a|<1$

$$\begin{aligned}\ln(1-2a\cos\Theta + a^2) &= \ln[(a-e^{i\Theta})(a-e^{-i\Theta})] = \ln(1-ae^{i\Theta}) + \ln(1-ae^{-i\Theta}) = \\ &= \left\{-ae^{i\Theta} - \frac{a^2}{2}e^{2i\Theta} - \frac{a^3}{3}e^{3i\Theta} + \dots\right\} + \left\{-ae^{-i\Theta} - \frac{a^2}{2}e^{-2i\Theta} - \frac{a^3}{3}e^{-3i\Theta} + \dots\right\} = \\ &= -2\left(a\cos\Theta + \frac{a^2}{2}\cos 2\Theta + \frac{a^3}{2}\cos 3\Theta + \dots\right) = -2\sum_{k=1}^{\infty}\frac{a^k}{k}\cos k\Theta.\end{aligned}$$

Положив здесь $\cos\Theta = x$ и $a = 1/4$, получим наш пример:

$$\ln\left(\frac{17}{16} - \frac{x}{2}\right) = \ln\left(1 - 2x\frac{1}{4} + \frac{1}{4^2}\right) = -2\sum_{n=1}^{\infty}\frac{1}{4^n n}\cos(n\arccos x) = -2\sum_{n=1}^{\infty}\frac{T_n(x)}{4^n n}.$$

Погрешность: теперь не норма L_2 считается мерой отклонения, а просто максимальная разность $|f-\varphi|$ на указанном промежутке $[-1, 1]$.

$$|\delta_n| = \left|2\sum_{k=n+1}^{\infty}\frac{T_k(x)}{4^k k}\right| \le 2\sum_{k=n+1}^{\infty}\frac{1}{4^k k}|T_k(x)| \le \frac{2}{n+1}\sum_{k=n+1}^{\infty}\frac{1}{4^k k} = \frac{2}{3(n+1)4^n}.$$

Если немного повозиться, то можно увидеть, что $|\delta_n| < 0,0005$ при $n \ge 5$.

Тогда подставляя в ряд явные выражения для полиномов Чебышёва I рода, получим

$$\begin{aligned}P_5(x) &= -\frac{1}{2}T_1(x) - \frac{1}{16}T_2(x) - \frac{1}{96}T_3(x) - \frac{1}{512}T_4(x) - \frac{1}{2560}T_5(x) = \\ &= \dots = \frac{31}{512} - \frac{241}{512}x - \frac{7}{64}x^2 - \frac{13}{384}x^3 - \frac{x^4}{64} - \frac{x^5}{160}.\end{aligned}$$

3.4 Численное интегрирование

Numerical integration

3.4.1 Равноотстоящие узлы

3.4.1.1 Формулы Ньютона-Котса

Задача 1. Получить формулы приближенного вычисления $\int_{-\alpha}^{\alpha} f(x)dx$

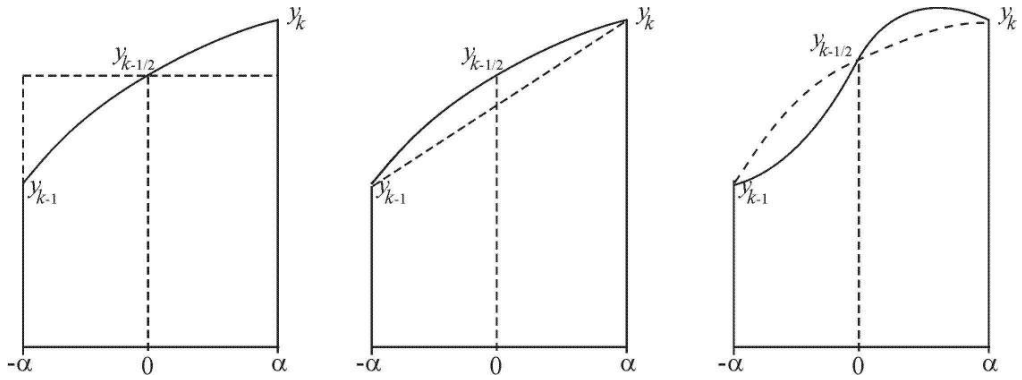


Рис. 3.7: Схема получения формул прямоугольников и пр.

A. Заменяем $f(x)$ константой, т.е. площадь под кривой – прямоугольником высоты $y_{k-1/2}$:

$$S_1 = 2\alpha \cdot y_{k-1/2}. \tag{3.67}$$

B. Заменяем $f(x)$ линейной функцией, т.е. площадь под кривой – трапецией с основаниями y_{k-1} и y_k:

$$S_2 = 2\alpha \frac{y_{k-1} + y_k}{2}. \tag{3.68}$$

C. Заменяем $f(x)$ параболой, которая совпадает с интегрируемой функцией на концах промежутка и в середине.

Ищем $y = ax^2 + bx + c \quad | \qquad y(-\alpha) = y_{k-1}, \quad y(0) = y_{k-1/2}, \quad y(\alpha) = y_k$:

$$\left\{ \begin{array}{rcl} a\alpha^2 - b\alpha + c &=& y_{k-1} \\ c &=& y_{k-1/2} \\ a\alpha^2 + b\alpha + c &=& y_k \end{array} \right. \Rightarrow \left\{ \begin{array}{l} a = \frac{1}{2\alpha^2}(y_{k-1} - 2y_{k-1/2} + y_k) \\ b = \frac{1}{2\alpha}(y_k - y_{k-1}) \\ c = y_{k-1/2}. \end{array} \right.$$

Тогда площадь $S_3 = \int\limits_{-\alpha}^{\alpha} dx(ax^2 + bx + c)$ равна

$$S_3 = \left(\frac{ax^3}{3} + \frac{bx^2}{2} + cx \right) \Big|_{-\alpha}^{\alpha} = a\frac{2\alpha^3}{3} + 2c\alpha = \frac{\alpha}{3}\left(y_{k-1} + 4y_{k-1/2} + y_k \right). \tag{3.69}$$

D. Можно продолжать в том же духе и заменить $f(x)$ на промежутке интерполяционным многочленом степени n, который интерполирует $f(x)$ в $n+1$ узле на промежутке (1.11):

$$f(x_j) = P_n(x_j), \quad -\alpha \le x_0 < x_1 < \ldots < x_n \le \alpha \quad \Rightarrow$$
$$P_n(x) = \sum_{j=1}^{n} f(x_j) L_j(x), \quad \text{где} \quad L_j(x) = \prod_{\substack{i \ne j \\ i=0}}^{n} \frac{x - x_i}{x_j - x_i}$$

это базисные многочлены Лагранжа, которые зависят только от выбора узлов интерполяции $\{x_j\}$, но не от $f(x)$.

Рассмотрим интерполяцию на *равноотстоящих узлах*, на произвольном промежутке интегрирования $[a,b]$

$$x_j = a + jh, \quad h = \frac{b-a}{n}, \quad j = 0,1,\ldots,n.$$

Введем новую переменную t, такую что $x = a+ht$: $L_j(x) = L_j(a+ht) \equiv \lambda_j(t)$. Интегрируя $P_n(x)$ от a до b, получим

$$S_n \equiv \int_a^b dx\, P(x) = \sum_{j=0}^{n} f_j \int_a^b dx\, L_j(x) = h \sum_{j=0}^{n} f_j \int_0^n dt\, \lambda_j(t) = h \sum_{j=0}^{n} f_j \alpha_j,$$

где α_j есть *весовые коэффициенты*, равные

$$\alpha_j = \int_0^n dt\, \lambda_j(t).$$

Их несложно посчитать*.Так как формула должна быть строго верна для $f(x) \equiv 1$, они удовлетворяют условию

$$\sum \alpha_j = n.$$

По построению это рациональные числа. Обозначим через s их общий знаменатель, и введем $\sigma_j = s\alpha_j$. Тогда формула приближенного интегрирования для равноотстоящих узлов, которая называется *квадратурной формулой Ньютона-Котса*[40] принимает вид

$$\int_a^b dx\, f(x) \approx \frac{b-a}{ns} \sum_{j=0}^{n} \sigma_j f_j. \tag{3.70}$$

Для $n = 1,2,\ldots,6$ получаются формулы Ньютона-Котса, приведенные в таб-

[40]По именам Исаака Ньютона и Роджера Котса. Роджер Котс, Roger Cotes (1682-1716) – английский математик и философ, работал с Ньютоном и корректировал второе издание его “Principia”. Фамилия читается так же как у Шерлока Холмса, поэтому часто встречающаяся транслитерация “Котес” неадекватна, и тем более не стоит ставить ударение на второй слог. “Квадратурная” от слова квадратура – интеграл.

личке

n			σ_j					ns	Ошибка	Название
1	1	1						2	$h^3\frac{1}{12}f^{(2)}(\xi)$	Трапеций
2	1	4	1					6	$h^5\frac{1}{90}f^{(4)}(\xi)$	Симпсона / парабол
3	1	3	3	1				8	$h^5\frac{3}{80}f^{(4)}(\xi)$	3/8
4	7	32	12	32	7			90	$h^7\frac{1}{90}f^{(6)}(\xi)$	Милна / Буля
5	19	75	50	50	75	19		288	$h^7\frac{275}{12096}f^{(6)}(\xi)$	–
6	41	216	27	272	27	216	41	840	$h^9\frac{9}{1400}f^{(8)}(\xi)$	Weddle

При бóльших n некоторые из σ_j становятся отрицательными, соответствующие формулы численно неустойчивы, и ими пользоваться не следует. Ошибки приведены для $b-a=1$ и без доказательства, вывод для $n=0,1$ см. в следующем пункте.

Задача 2. Формулы для вычисления $\int\limits_a^b f(x)dx$.

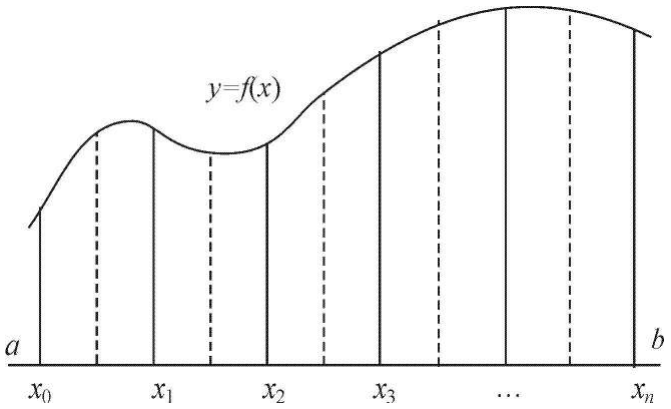


Рис. 3.8: Разбиваем $[a,b]$ на много одинаковых кусочков...

Формулы Ньютона-Котса не работают при большом числе узлов n. Поэтому весь промежуток интегрирования $[a,b]$ разбивается на много кусочков, на каждом их которых применяется формула с небольшим n (обычно 2–4).

Пусть

$$x_j = a + jh, \quad h = \frac{b-a}{m}, \quad j = 0,1,\ldots,m,$$

где m это число подпромежутков, на которых используем формулу интегрирования с n узлами; h – ширина каждого подпромежутка, тогда полуширина $\alpha = (b-a)/2m$.

Суммируя интегралы S_1, S_2, S_3 по всем промужеткам, получим формулы приближенного интегрирования (названия относятся также и к (3.67-3.69)):

A. Прямоугольников (rectangle rule):

$$\int_a^b f(x)dx \approx h\left(y_{1/2} + y_{3/2} + \ldots + y_{m-1/2}\right); \tag{3.71}$$

B. Трапеций (trapezoidal rule):

$$\int_a^b f(x)dx \approx h\left(\frac{y_0}{2} + y_1 + y_2 + \ldots + y_{n-1} + \frac{y_n}{2}\right); \tag{3.72}$$

C. Парабол (Симпсона, Simpson's rule):

$$\int_a^b f(x)dx \approx \frac{h}{6}\Big(y_0 + y_m + 2(y_1 + y_2 + \ldots + y_{m-1}) + 4\left(y_{1/2} + y_{3/2} + \ldots + y_{m-1/2}\right)\Big); \tag{3.73}$$

и так далее.

Число разбиений промежутка m нужно выбирать достаточно большое, чтобы обеспечить требуемую точность. Аналогично можно записать формулы с бóльшими n.

3.4.1.2 Остаточные члены на примере $n = 0, 1$

Формула прямоугольников, $n = 0$. Вернемся к задаче 1 и рассмотрим функцию $\Phi(\alpha) = \int_{-\alpha}^{\alpha} f(x)dx$.

- Она нечетная: $\Phi(\alpha) = -\Phi(-\alpha)$; $\Rightarrow \quad \Phi(0) = \Phi''(0) = \ldots = 0$;
- $\Phi'(\alpha) = f(-\alpha) + f(\alpha)$

Ошибка, допущенная при вычислении интеграла Φ по формуле прямоугольников – это разность $|\Phi(\alpha) - S_1|$, где $S_1 = 2\alpha f(0) = \alpha\Phi'(0)$.

$\Phi(\alpha)$ разложим в ряд Тейлора с остаточным членом в интегральной форме и учтем что она нечетна:

$$\Phi(\alpha) = 0 + \alpha\Phi'(0) + 0 + \frac{1}{2}\int_0^\alpha \Phi'''(t)(\alpha - t)^2 dt = \alpha\Phi'(0) + \frac{1}{2}\int_0^\alpha \Phi'''(t)(\alpha - t)^2 dt.$$

Тогда $|S_1 - \Phi(\alpha)| = \frac{1}{2}\left|\int_0^\alpha \Phi'''(t)(\alpha - t)^2 dt\right|$ и используя теорему о среднем, получаем

$$|S_1 - \Phi(\alpha)| = \frac{1}{2}\left|\Phi'''(\xi)\int_0^\alpha (\alpha - t)^2 dt\right| = \frac{\alpha^3}{6}\left|f''(-\xi) + f''(\xi)\right| = \frac{\alpha^3}{3}\left|f''(\bar{\xi})\right|,$$

где $\overline{\xi} \in (-\alpha, \alpha)$, и мы предположили что f'' – непрерывная функция, хотя то же можно было получить и при более скромных предположениях. Также при последнем переходе использовали теорему о взвешенных средних: $\sum n_i b_i / \sum n_i$ находится между наибольшим и наименьшим b_i.

Получена оценка ошибки на одном интервале (x_k, x_{k+1}). Суммируя по всем интервалам и опять используя теорему о взвешенных средних, получим для ошибки на всем интервале (a, b)

$$R_n^{rect} \le \frac{\alpha^3}{3} \left| f''(\overline{\xi}_1) + \ldots + f''(\overline{\xi}_n) \right| = \frac{\alpha^3}{3} n \left| f''(\hat{\xi}) \right| = \frac{(b-a)^3}{24 n^2} \left| f''(\hat{\xi}) \right|.$$

В итоге получили оценку сверху для остаточного члена формулы прямоугольников

$$R_n^{rect} \le \frac{(b-a)^3}{24\, n^2} \sup_{\xi \in [a,b]} |f''(\xi)|. \tag{3.74}$$

Формула трапеций, $n = 1$. Остаточные члены для нечетных n можно получить интегрированием ошибки интерполяционного полинома Лагранжа. Проиллюстрируем этот способ на простейшем примере $n = 1$ (формула трапеций, два узла интерполяции). Верхнюю оценку для ошибки интерполяционной формулы Лагранжа (1.25) запишем без модуля ω, учитывая ее знакопеременность:

$$R(x) = \frac{\sup |f^{(n+1)}(\xi)|}{(n+1)!} \cdot \omega_n(x) \quad \Rightarrow \quad R|_{n=1} = \frac{\sup |f''(\xi)|}{2} (x - x_0)(x - x_1).$$

Полагая $x_0 = -\alpha$, $x_1 = \alpha$, и интегрируя в этих пределах, получаем ошибку на интервале $[-\alpha, \alpha]$:

$$R_1^{trap} \le \frac{|\sup f''(\xi)|}{2} \int_{-\alpha}^{\alpha} dx (x^2 - \alpha^2) = \frac{|\sup f''(\xi)|}{2} \cdot \frac{4}{3} \alpha^3 = \frac{2\alpha^3}{3} |\sup f''(\xi)|.$$

Полагая теперь $\alpha = (b-a)/2n$ и суммируя по всем промежуткам, получим

$$R_n^{trap} \le n \frac{2(b-a)^3}{3(2n)^3} |\sup f''(\xi)| = \frac{(b-a)^3}{12 n^2} |\sup f''(\xi)|.$$

Заметим, что для четных n такой способ не годится, т.к. ω_n в этом случае функция нечетная, и интеграл от нее обращается в ноль. Это означает, что ошибка квадратурной формулы определяется слагаемыми следующего порядка малости, что и видно из таблички для формул Ньютона-Котса при разных n – погрешности квадратурных формул с $n = 2m$ и $n = 2m + 1$ одного порядка по h.

3.4.2 Квадратурная формула Гаусса

Gaussian quadrature

3.4.2.1 Постановка задачи

Полученные выше формулы численного интегрирования на равноотстоящих узлах – прямоугольников, парабол, и общая формула Ньютона-Котса (3.70) – представляют собой частные случаи квадратурных формул вида $\int dx\, f(x) = \sum \mathrm{w}_i f(x_i) + R$, где суммирование производится по какому-то набору узлов $\{x_i\}$, а весовые коэффициенты w_i не зависят от подынтегральной функции. Формулы прямоугольников и трапеций дают точные выражения (т.е. остаточный член R обращается в ноль) только когда f линейна по x. Формула парабол точна когда f есть многочлен степени не выше двух. Поставим следующую задачу.

Требуется построить формулы, имеющие, при заданном числе узлов n, наивысшую алгебраическую точность, т.е. узлы и коэффициенты квадратурных формул ищутся из условия, чтобы остаточный член формулы обращался в ноль на множестве всех многочленов максимально высокой степени. При этом мы будем рассматривать сразу более общую задачу –интегралы с некоторой *весовой функцией* $p(x) > 0$.

Таким образом, мы ищем наборы *узлов* $\{x_i\}_1^n$ и *весов* $\{\mathrm{w}_i\}_1^n$, такие, чтобы формула

$$\boxed{\int_a^b dx\, p(x) f(x) = \sum_{k=1}^{n} \mathrm{w}_k f(x_k)} \tag{3.75}$$

была строго верна для всех f полиномиального вида максимально возможной степени m

$$f(x) = a_0 + a_1 x + \ldots + a_m x^m = \sum_{i=0}^{m} a_i x^i \in \overline{\Pi}_m. \tag{3.76}$$

Ниже мы увидим, что эта задача имеет и единственное решение для $m = 2n-1$. Соответствующие узлы и веса при подстановке в (3.75) дают *квадратурную формулу Гаусса*.

Покажем для начала, что для $m = 2n$ задача неразрешима. Подставляя в (3.75) $f(x) = x^i$, $i = 0, \ldots, m$ получим систему (здесь $d\mu = p(x)dx$) для узлов x_k и весов w_k

$$\sum_{k=1}^{n} \mathrm{w}_k x_k^i = \int_a^b d\mu x^i \equiv \mu_i, \qquad i = 0, 1, \ldots, m. \tag{3.77}$$

Числа $\mu_i = \int d\mu\, x^i$ – моменты распределения μ, а $\mu \equiv \mu_0 = \int d\mu$ есть мера промежутка интегрирования. В этой нелинейной системе $m+1$ уравнение и $2n$

неизвестных, поэтому она может быть разрешима при $m \le 2n-1$, ч.и т.д.

Для $m = n-1$ квадратурную формулу можно построить для произвольной системы (простых) узлов $\{x_i\}_{i=1}^n$. Покажем это. Построим для $f(x)$ интерполяционный многочлен Лагранжа (1.11) $\tilde{f}$ через точки $\{x_i\}_{i=1}^n$ (так же как и в случае равноотстоящих узлов):

$$\tilde{f}(x) = \sum_{k=1}^{n} f(x_k)\Phi_k(x), \quad \text{где} \quad \Phi_k(x) = \prod_{\substack{i\neq j\\ i=1}}^{n} \frac{x-x_i}{x_j-x_i} \equiv \frac{\psi_{nk}(x)}{\psi_{nk}(x_k)}; \tag{3.78}$$

$$\psi_{nk}(x) = \frac{\omega_n(x)}{x-x_k}; \qquad \omega_n(x) = (x-x_1)(x-x_2)\dots(x-x_n). \tag{3.79}$$

Здесь удобнее и принято использовать обозначения, немного отличные от тех что были раньше: мы нумеровали узлы начиная с нуля, так что n было равно степени полинома, а теперь нумеруем их от единицы, так что $L_k \equiv \Phi_{k+1}$ и k – число узлов интерполяции. Формулы (3.78-3.79) фиксируют обозначения раздела по квадратурной формуле Гаусса.

Легко видеть, что $\Phi_k(x_i) = \delta_{ik}$; по построению $\Phi_k(x), \tilde{f}(x) \in \overline{\Pi}_{n-1}$. Если при этом $f \in \overline{\Pi}_{n-1}$, то значит интерполяционный многочлен совпадает с самой функцией $\tilde{f} \equiv f$ и

$$\int d\mu\, f = \int d\mu\, \tilde{f} = \int d\mu \sum_{k=1}^{n} f(x_k)\Phi_k = \sum_{k=1}^{n} f(x_k) \int d\mu \Phi_k(x).$$

Получили

$$\int d\mu\, f(x) = \sum_{k=1}^{n} f(x_k)\mathrm{w}_k, \quad \text{где} \qquad \mathrm{w}_k = \int d\mu\, \Phi_k(x). \tag{3.80}$$

Таким образом, квадратурная формула для $m = n-1$ построена. Осталось доказать, что можно так выбрать систему узлов, чтобы она была верна для всех многочленов степени $m = (2n-1)$ (как было указано выше, для степени $2n$ задача неразрешима), и выяснить как их выбирать.

3.4.2.2 Узлы и веса

Узлы. Положим, что $f \in \overline{\Pi}_{2n-1}$. Как известно, всякий многочлен $f \in \overline{\Pi}_{2n-1}$ можно поделить в столбик на многочлен меньшей степени $\omega_n \in \overline{\Pi}_n$ (3.79) и представить в виде

$$f(x) = \omega_n(x)q(x) + r(x),$$

где $q(x) \in \overline{\Pi}_{n-1}$ – частное от деления f на ω_n, а $r(x) \in \overline{\Pi}_{n-1}$ – остаток от деления. Но раз $r(x)$ – многочлен степени не выше $n-1$, то при любом наборе узлов для

него верна квадратурная формула (3.80), и учитывая что $\omega_n(x_k)$=0, получаем

$$\int d\mu\, r(x) = \sum r(x_k)\mathrm{w}_k = \sum f(x_k)\mathrm{w}_k.$$

Следовательно

$$\int d\mu f = \int d\mu\big(\omega_n(x)q(x) + r(x)\big) = \sum f(x_k)\mathrm{w}_k + \int d\mu\, \omega_n(x)q(x).$$

Таким образом, квадратурная формула для f будет верна тогда и только тогда, когда функция $\omega_n(x)$ будет ортогональна с весом $p(x)$ всякому полиному $q(x)$ степени $n-1$. Учитывая также, что старший коэффициент ω_n из (3.79) равен единице, получаем, что:
Квадратурная формула (3.75) верна для $\forall f(x) \in \overline{\Pi}_{2n-1}$ *т.и т.т., когда* $\omega_n(x)$ *есть приведенный полином степени* n *из системы полиномов, ортогональных в скалярном произведении* $(f,g) = \int d\mu fg$*:*

$$(f,g) = \int d\mu f(x)g(x); \quad \omega_n(x) = x^n + \kappa_{n-1}^{(n)}x^{n-1} + \ldots + \kappa_0^{(n)}; \quad (\omega_n, \omega_m) \sim \delta_{nm}.$$

Будем обозначать, как обычно, не приведенные ортогональные полиномы из соответствующей системы (3.32) как $p_n = k_n^{(n)}x^n + \ldots k_0^{(n)}$. Тогда $\{x_i\}$ – нули p_n.

Конкретные формулы (3.75) с весами и пределами интегрирования, которые соответствуют классическим ортогональным полиномам, называются теми же именами: формула Гаусса-Лежандра для $p=1$ и Гаусса-Чебышёва для $p = (1-x^2)^{-1/2}$ на $[-1,1]$, Гаусса-Эрмита на $(-\infty,\infty)$ с весом e^{-x^2} и так далее.

Веса. Весовые коэффициенты в общем случае даются формулой (3.80), где в качестве узлов Φ_k следует брать нули ортогональных полиномов $\omega_n \sim p_n$. Для стандартных $p(x)$ такие интегралы считаются аналитически. Однако задачу можно свести и к более простой.

Подставим в квадратурную формулу (3.75) для n узлов в качестве $f(x)$ ортогональные полиномы p_j, с $j = 0,\ldots,n-1$, из ортогональной системы с весом $p(x)$. Тогда

$$\sum_{k=1}^{n} \mathrm{w}_k p_j(x_k) = \int_a^b d\mu p_j(x) = \frac{1}{p_0}\int_a^b d\mu p_j(x)p_0(x) = \frac{\|p_0\|^2}{p_0}\cdot\delta_{0j}.$$

Так как

$$\|p_0\|^2 = p_0^2\int_a^b d\mu = p_0^2\,\mu,$$

то в итоге мы получили систему уравнений для w_k:

$$\sum_{k=1}^{n} w_k p_j(x_k) = \delta_{0j} \cdot \mu p_0, \quad j = 0, 1, \ldots, n-1. \tag{3.81}$$

Эта система разрешима, так как ее матрица A невырождена.

◂ От противного. Если матрица $A = \left[p_j(x_k)\right]$ вырождена, то существует ненулевой вектор $\tilde{w} = (\tilde{w}_0, \ldots, \tilde{w}_{n-1})$, такой что $A\tilde{w} = 0$. Согласно определению A (3.81), это означает, что есть многочлен $\sum \tilde{w}_i p_i(x) \in \overline{\Pi}_{n-1}$, который обращается в ноль в n узлах $x_1, \ldots, x_n$, а значит тождественно равен нулю. Но это противоречит линейной независимости $\{p_i\}$, так что A должна быть невырожденной, ч.и т.д. ▸

3.4.2.3 Веса и дискретная ортогональность

Решение системы (3.81) получим в другом подходе.

Дискретная ортогональность. Пусть $\{p_n\}$ – система ортогональных полиномов на $[a, b]$ с весом $p(x)$. Тогда $p_i p_j$ для $i, j \leq (n-1)$ есть многочлен степени ниже $2n$ и для него верна общая квадратурная формула Гаусса

$$(p_i, p_j) \equiv \int d\mu p_i p_j = \sum_{k=1}^{n} w_k p_i(x_k) p_j(x_k).$$

С другой стороны, вследствие ортогональности, $(p_i, p_j) = \delta_{ij}\|p_i\|^2$, $i, j = 0, 1, \ldots$, поэтому получаем запись *дискретной ортогональности* полиномов p_i:

$$\sum_{k=1}^{n} w_k p_i(x_k) p_j(x_k) = \delta_{ij}\|p_i\|^2, \quad i, j = 0, \ldots, n-1. \tag{3.82}$$

Немного отвлекаясь, заметим, что дискретная ортогональность может быть полезна при решении задачи на среднеквадратичное приближение функции, заданной таблично (3.57). Тогда приближение можно получать сразу в виде суммы по базисным полиномам дискретной переменной, без решения системы уравнений. Неудобство в том, что функция должна быть задана своими значениями в нулях $p_n(x)$, которые для классических ортогональных полиномов всегда не равноотстоящие. Для полиномов Чебышёва T_n формула все равно полезна, потому что их нули просто считаются. Можно и дальше развить эту идею и построить ортогональные полиномы дискретной переменной для заданной системы узлов, в том числе равноотстоящих (см. [16]).

Формула для весов. Равенство (3.82) можно переписать как

$$\sum_{k=1}^{n}\left(\sqrt{\mathrm{w}_k}\frac{p_i(x_k)}{\|p_i\|}\right)\left(\sqrt{\mathrm{w}_k}\frac{p_j(x_k)}{\|p_j\|}\right)=\delta_{ij},\quad i,j=0,\ldots,n-1.$$

Если определить матрицу Q $n\times n$ с элементами

$$Q_{ij}=\sqrt{\mathrm{w}_i}\,\frac{p_j(x_i)}{\|p_j\|},$$

то последнее равенство сводится к $\sum_k Q_{ki}Q_{kj}=\delta_{ij}$, что представляет собой запись матричного уравнения $Q^TQ=I$ (I – единичная матрица). Тогда верно и $QQ^T=I$. Расписав обратно через сумму это равенство, получим $\sum_k Q_{ik}Q_{jk}=\delta_{ij}$, и в явном виде

$$\sum_{k=0}^{n-1}\left(\sqrt{\mathrm{w}_i}\frac{p_k(x_i)}{\|p_k\|}\right)\left(\sqrt{\mathrm{w}_j}\frac{p_k(x_j)}{\|p_k\|}\right)=\delta_{ij},\quad i,j=0,\ldots,n-1.$$

Тогда при $i=j$ получаем

$$\mathrm{w}_i\sum_{k=0}^{n-1}\frac{p_k^2(x_i)}{\|p_k\|^2}=1,$$

и окончательно

$$\mathrm{w}_i=\left\{\sum_{k=0}^{n-1}\frac{p_k^2(x_i)}{\|p_k\|^2}\right\}^{-1}>0. \tag{3.83}$$

Матрица Якоби*. Приведем без вывода*еще один способ получения узлов и весов (Golub-Welsch algorithm).

Для приведенных ортогональных полиномов $\tilde{p}_n(x)$, очевидно, выполняются рекуррентные соотношения вида (3.39). Такие же рассуждения приводят к

$$\tilde{p}_{n+1}=(x-\beta_{n+1})\tilde{p}_n-\gamma_{n+1}^2\tilde{p}_{n-1},\quad \text{где}$$
$$\beta_{n+1}=\frac{(x\tilde{p}_n,\tilde{p}_n)}{(\tilde{p}_n,\tilde{p}_n)}=\frac{(xp_n,p_n)}{(p_n,p_n)};\quad \gamma_{n+1}^2=\frac{(\tilde{p}_n,\tilde{p}_n)}{(\tilde{p}_{n-1},\tilde{p}_{n-1})}.$$

Коэффициенты β и γ считаются аналитически для каждой конкретной системы ортогональных полиномов.

Построим трехдиагональную матрицу (Якоби) $n\times n$

$$J_n=\begin{pmatrix}\beta_1 & \gamma_2 & & & \\ \gamma_2 & \beta_2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \gamma_n \\ & & & \gamma_n & \beta_n\end{pmatrix} \tag{3.84}$$

T^0: Корни x_i, $i=1,\ldots,n$ полинома p_n являются собственными значениями матрицы J_n.

T^0: Если $v_{(i)} = (v_1^{(i)}, \dots, v_n^{(i)})$ – собственный вектор J_n, соответствующий собственному значению x_i, то весовые коэффициенты квадратурной формулы Гаусса равны

$$\mathrm{w}_i = \mu \frac{\left(v_1^{(i)}\right)^2}{\left(v^{(i)}, v^{(i)}\right)} = \frac{\mu \left(v_1^{(i)}\right)^2}{\sum\limits_{k=1}^{n} \left(v_k^{(i)}\right)^2}.$$

3.4.2.4 Остаточный член

Если f – $2n$ раз непрерывно дифференцируемая функция, то верхняя оценка для остаточного члена формулы Гаусса есть

$$R \equiv \int d\mu f(x) - \sum_{k=1}^{n} \mathrm{w}_k f(x_k) = \frac{f^{(2n)}(\xi)}{(2n)!} \|\omega_n\|^2, \qquad \xi \in [a,b]. \tag{3.85}$$

◂Пусть $h \in \overline{\Pi}_{2n-1}$ – решение интерполяционной задачи Эрмита (1.49) для f:

$$h(x_i) = f(x_i), \quad h'(x_i) = f'(x_i), \quad i = 1, 2, \dots, n.$$

Для h строго верна формула Гаусса

$$\int_a^b d\mu h(x) = \sum_{i=1}^{n} \mathrm{w}_i h(x_i) = \sum_{i=1}^{n} \mathrm{w}_i f(x_i),$$

так что ошибку можно переписать в виде интеграла

$$R = \int_a^b d\mu \big(f(x) - h(x)\big).$$

Но погрешность интерполяции Эрмита, в соответствии с (1.55), равна[41]

$$f(x) - h(x) = \frac{f^{(2n)}(\xi)}{(2n)!}(x-x_1)^2(x-x_2)^2 \dots (x-x_n)^2 = \frac{f^{(2n)}(\xi)}{(2n)!}\omega^2(x), \quad \xi \in [a,b].$$

Тогда, интегрируя и используя теорему о среднем, получим (3.85). ▸

Недостаток этой оценки точности заключается в том, что эта оценка сверху как правило является сильно завышенной. Кроме того, она вообще применима только для функций, у которых есть соответствующая производная, в то время как *формула Гаусса сходится при $n \to \infty$ для любой непрерывной на отрезке функции f*. Кроме того, если даже производная и существует, то численно или аналитически считать производную функции порядка $2n$ – дело неблагодарное.

[41] В формуле (1.55) $n+1$ это число условий в задаче интерполяции. В нашем случае значит вместо $n+1$ надо ставить $2n$.

3.4.2.5 Формула Гаусса-Лежандра

Самый простой частный случай $p(x) = 1$, очевидно, представляет также и наибольшую ценность для практического применения. Разберем его подробно.

В этом случае нам нужны *полиномы Лежандра*, которые ортогональны на промежутке $[-1, 1]$ с весом $p{=}1$. Табулированные полиномы Лежандра нормированы так, чтобы $P_n(1) = 1$. Тогда формула Родрига (3.51) для них дает[42]

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right] \tag{3.86}$$

Тогда для приведенных полиномов получаем[43]

$$\tilde{P}_n(x) = \frac{n!}{(2n)!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right] = \frac{2^n(n!)^2}{(2n)!}P_n \tag{3.87}$$

Ортогональные полиномы на $[a, b]$ с весом 1 получаем из P_n и $\tilde{P}_n$ линейной заменой $x = \frac{b+a}{2} + \frac{b-a}{2}t$, которая переводит (-1) в a, а $(+1)$ в b. Таким образом, под производной стоит $(x-a)^n(x-b)^n$, а коэффициент для ω_n точно так же фиксируется нормировкой:

$$\omega_n(x) = \frac{n!}{(2n)!}\frac{d^n}{dx^n}\left[(x-a)^n(x-b)^n\right]. \tag{3.88}$$

Явный вид первых нескольких P_n несложно получить из (3.86) (см. также приложение C):

$$\begin{aligned}
P_0 &= 1 & P_4 &= \tfrac{1}{8}\left[35x^4 - 30x^2 + 3\right]\\
P_1 &= x & P_5 &= \tfrac{1}{8}\left[63x^5 - 70x^3 + 15x\right]\\
P_2 &= \tfrac{1}{2}\left[3x^2 - 1\right] & P_6 &= \tfrac{1}{16}\left[231x^6 - 315x^4 + 105x^2 - 5\right]\\
P_3 &= \tfrac{1}{2}\left[5x^3 - 3x\right] & P_7 &= \tfrac{1}{16}\left[429x^7 - 693x^5 + 315x^3 - 35x\right]
\end{aligned}$$

Узлы. Как видно, P_n это четные и нечетные многочлены, и их нули вплоть до $n{=}5$ получаются как корни (би)квадратных уравнений, и до $n{=}7$ как корни (би)кубических. Начиная с $n = 10$ нули уже находятся только численно, но процесс облегчен тем, что они хорошо локализованы. Вследствие четности-нечетности P_n нули распределены симметрично относительно нуля, и тогда из общей формулы (3.83) для весовых коэффициентов w_k видно, что они также распределены симметрично.

[42] Можно расписать $(x^2-1) = (x+1)(x-1)$ и расписать n-тую производную по формуле Лейбница. Тогда при $x = 1$ все слагаемые в сумме будут равны нулю, кроме тех, в которых n раз дифференцируется $(x-1)^n$. Тогда $\left.[x^2-1]^{(n)}\right|_{x=1} = (1+1)^2 n!$, так что $P_n(1) = 1$.

[43] Так как $(x^{2n})^{(n)} = \frac{(2n)!}{n!}x^n$.

Узлы полиномов на $[a,b]$ получаются масштабированием табулированных нулей полиномов Лежандра P_n на $[-1,1]$ тем же линейным преобразованием, и следовательно они распределены симметрично относительно середины промежутка $[a,b]$.

Весовые коэффициенты*. Для весовых коэффициентов можно использовать формулу (3.80) или (3.83), но в этом частном случае можно получить и еще более короткое выражение.

Из определения $\psi_{nk}(x)$ (3.79) видно, что

$$\psi_{nk}(x_i) = \delta_{ik}\psi_{nk}(x_k); \quad \psi_{nk}(x_k) = \omega_n'(x_k).$$

Так как $\psi_{nk} \in \overline{\Pi}_{n-1}$, то для $\psi_{nk}^2 \in \overline{\Pi}_{2n-2}$ верна квадратурная формула Гаусса (3.75):

$$\int d\mu\, \psi_{nk}^2 = \sum_{i=1}^{n} \mathrm{w}_i \psi_{nk}^2(x_i) = \mathrm{w}_k \psi_{nk}^2(x_k),$$

и следовательно

$$\mathrm{w}_k = \frac{\int d\mu\, \psi_{nk}^2(x)}{\psi_{nk}^2(x_k)} \equiv \frac{\|\psi_{nk}\|^2}{\psi_{nk}^2(x_k)}. \tag{3.89}$$

Это общая формула, но именно при $p=1$ квадрат нормы ψ_{nk} в (3.89) несложно вычислить явно:

$$\int d\mu\, \psi_{nk}^2 = \int \frac{dx\,\omega_n^2}{(x-x_k)^2} = -\int \omega_n^2 d\frac{1}{x-x_k} = -\left.\frac{\omega_n^2}{x-x_k}\right|_a^b + 2\int \frac{dx\,\omega_n\omega_n'}{x-x_k}. \tag{3.90}$$

Последний интеграл можно переписать $\int \frac{dx\,\omega_n\omega_n'}{x-x_k} = \int dx\,\psi_{nk}\omega_n'$. Здесь подынтегральное выражение есть многочлен степени $2n-2$, и поэтому для него верна квадратурная формула (3.80):

$$\int dx\,\psi_{nk}\omega_n' = \sum_i \mathrm{w}_i\omega_n'(x_i)\psi_{nk}(x_i) = \mathrm{w}_k\omega_n'(x_k)\psi_{nk}(x_k) = \mathrm{w}_k\psi_{nk}^2(x_k).$$

Тогда из (3.89) и (3.90) получаем

$$\mathrm{w}_k\psi_{nk}^2(x_k) = \left.\frac{\omega_n^2}{x-x_k}\right|_a^b.$$

Значения $\omega_n(x)|_{x=a,b}$ находим из (3.88). Если в (3.88) расписать производную произведения по формуле Лейбница, то при $x=a$ все слагаемые обратятся в ноль, кроме того, в котором n раз дифференцируется множитель $(x-a)^n$. Поэтому

$$\omega_n(a) = \frac{(n!)^2}{(2n)!}(a-b)^n.$$

Точно так же получаем $\omega_n(b)$, и видим что

$$\omega_n^2(a) = \omega_n^2(b) = \frac{(n!)^4}{[(2n)!]^2}(b-a)^{2n}. \tag{3.91}$$

Подставляя (3.91) во внеинтегральный член, получаем

$$\left.\frac{\omega_n^2}{x-x_k}\right|_a^b = \omega_n^2(b)\left(\frac{1}{b-x_k} - \frac{1}{a-x_k}\right) = \frac{(n!)^4}{[(2n)!]^2}\frac{(b-a)^{2n+1}}{(x_k-a)(b-x_k)}.$$

Тогда для весовых коэффициентов w_k в случае $p(x)=1$ окончательно имеем:

$$\mathrm{w}_k^{Leg}[a,b] = \frac{(n!)^4}{[(2n)!]^2}\frac{(b-a)^{2n+1}}{(x_k-a)(b-x_k)} \cdot \frac{1}{\psi_{nk}^2(x_k)}. \tag{3.92}$$

Последний множитель также можно переписать как $[\omega_n'(x_k)]^{-2}$.

При $a=-1$, $b=1$ из (3.87) получаем $\omega_n \equiv \tilde{P}_n = \frac{2^n(n!)^2}{(2n)!}P_n$, и подставляя в (3.92), имеем

$$\mathrm{w}_k^{Leg}[-1,1] = \frac{2}{1-x_k^2}\left[P_n'(x_k)\right]^{-2}. \tag{3.93}$$

Остаточный член. Подставляем $\omega_n = \tilde{P}_n$ в общую формулу (3.85) для остаточного члена. Учитывая что $\|P_n\|^2 = 2/(2n+1)$, из (3.87) получаем

$$\|\omega_n\|^2 = 2^{2n}\frac{(n!)^4}{[(2n)!]^2}\|P_n\|^2 = \frac{2^{2n+1}}{2n+1}\frac{(n!)^4}{[(2n)!]^2}.$$

Тогда (3.85) сводится к

$$R^{Leg}_{[-1,1]}(f) = \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} f^{(2n)}(\xi), \qquad \xi \in [-1,1]. \tag{3.94}$$

На промежутке $[a,b]$ тогда получаем ◂…▸

$$R(f) = \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)\,[(2n)!]^3} \cdot f^{(2n)}(\xi), \qquad \xi \in [a,b]. \tag{3.95}$$

Приведем выражения для остаточных членов в случае $b-a = 1$ для нескольких конкретных n, и для сравнения выпишем справа остаточные члены формулы Симпсона с тем же n:

n	G.-Legendre	Simpson
2	$\frac{1}{2592} f^{(4)}(\xi)$	$\frac{1}{4\cdot 10^4} f^{(4)}(\xi)$
4	$\frac{1}{2\cdot 10^9} f^{(8)}(\xi)$	$\frac{1}{7\cdot 10^5} f^{(4)}(\xi)$
6	$\frac{1}{5\cdot 10^{15}} f^{(12)}(\xi)$	$\frac{1}{3\cdot 10^7} f^{(4)}(\xi)$.

Как мы уже знаем, высокая степень производной f в столбце Гаусса не есть столь большой повод для радости, как можно было бы думать. Однако скорость уменьшения величины коэффициента при нем действительно говорит об эффективности квадратурной формулы начиная с $n \geq 3$.

Числа. Итак, в случае $p = 1$ квадратурная формула Гаусса на промежутке $[-1,1]$ превращается в формулу Гаусса-Лежандра

$$\int_{-1}^{1} dx\, f(x) = \sum_{k=1}^{n} \mathrm{w}_k f(x_k) + R(f),$$

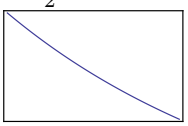
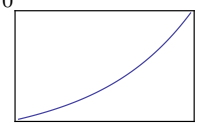
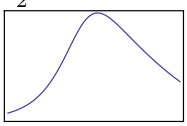
где узлы являются нулями полиномов Лежандра P_n (3.86), веса даются выражением (3.83) или (3.93), а остаточный член формулой (3.94).

Приведем конкретные численные значения для некоторых n:

n	$P_n(x)$	x_i	w_i	R
1	x	0	2	$\frac{1}{3}f^{(2)}$
2	$\frac{1}{2}(3x^2-1)$	$\pm 1/\sqrt{3}$	1	$\frac{1}{135}f^{(4)}$
3	$\frac{1}{2}(5x^3-3x)$	0 $\pm\sqrt{3/5}$	8/9 5/9	$\frac{f^{(6)}}{15\,750}$
4	$\frac{1}{8}(35x^4-30x^2+3)$	$\pm\sqrt{\frac{3-2\sqrt{6/5}}{7}}$ $\pm\sqrt{\frac{3+\sqrt{6/5}}{7}}$	$\frac{18+\sqrt{30}}{36}$ $\frac{18-\sqrt{30}}{36}$	$\frac{f^{(8)}}{3.5\cdot 10^6}$
6	$\frac{231x^6-315x^4+105x^2-5}{16}$	$\pm 0,932\,469\,514\,203$ $\pm 0,661\,209\,386\,466$ $\pm 0,238\,619\,186\,083$	$0,171\,324\,492\,379\,170$ $0,360\,761\,573\,048\,138$ $0,467\,913\,934\,572\,691$	$\frac{f^{(12)}}{6.4\cdot 10^{11}}$

Как было указано выше для общего случая, оценка для остаточного члена не особо удобна для практического применения. Поэтому для оценки точности часто применяется двойной пересчет – вычисление интеграла, скажем, при n=3 и при n=6. Разница дает оценку точности.

Примеры счета нескольких интегралов по формуле Гаусса-Лежандра с $n = 3, 6, 12, 24$ и по формуле Симпсона с $m = 1, 2, 4, 8$ (т.е с примерно тем же – чуть большим – числом точек на всем промежутке, в которых вычисляется подынтегральное выражение) иллюстрируют сравнительную точность счета:

$\int_a^b dx\, f(x)$ $f(x)$	$\int_2^{3.2} \frac{dx\, x}{\sqrt{x^2+1}}$ 	$\int_0^2 \frac{dx\, e^x \arctan(x+1)}{\sqrt{x+2}}$ 	$\int_2^{3.2} \frac{dx\, e^{x^2-x}}{\ln(4x^2+x+2)}$ 
Simpson m = 1	1.98901988112	4.04007836678	2.12292234843
Simpson m = 3	1.98899926573	4.02411689900	1.67097644818
Simpson m = 6	1.98899916887	4.02390926239	1.66074632172
Simpson m = 12	1.98899916408	4.02389612394	1.66073662614
Gauss-L n = 3	1.98899978616	4.02380410716	1.79395322693
Gauss-L n = 6	1.98899916378	4.02389524582	1.65561069898
Gauss-L n = 12	1.98899916378	4.02389524524	1.66075432476
Gauss-L n = 24	1.98899916378	4.02389524524	1.66073754983
все знаки верные	1.98899916378	4.02389524524	1.66073754990

Здесь подчеркнуты неверные разряды. Как видно, численная сложность интегралов идет по нарастающей.

3.4.2.6 Формула Гаусса-Чебышёва

Формула Гаусса-Лежандра хороша для вычисления интегралов от функций без особенностей и на конечном отрезке. Если же у f есть (даже интегрируемые) особенности, то приведенная верхняя оценка для остаточного члена становится бесконечной, а сходимость и точность значительно ухудшаются. В этом случае может иметь смысл включить особенность в вес $p(x)$ и воспользоваться разновидностью квадратурной формулы с этим весом. С другой стороны, если промежуток интегрирования (полу-)бесконечный, то тоже придется пользоваться общей формулой.

Стандартными и используемыми являются, кроме уже рассмотренной формулы Гаусса-Лежандра, формулы Г.-Эрмита (на бесконечном промежутке, с весом e^{-x^2}), Г.-Лагерра (на полубесконечном, с весом e^{-x}) и Г.-Чебышёва – на конечном промежутке $[a,b]$, с весами $\big((b-x)(x-a)\big)^{\pm 1/2}$. Формула Г.-Ч. с весом $1/\sqrt{1-x^2}$ (дальше будем для простоты говорить об интервале $[-1,1]$) хороша, во-первых, тем, что с помощью нее можно эффективно интегрировать функции с корневыми особенностями, а во-вторых, узлы и весовые коэффициенты в ней имеют особенно простой вид.

Рассмотрим теперь формулу Г.-Чебышёва на $[-1,1]$:

$$\int\limits_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx = \sum_{i=1}^{n} \mathrm{w}_i f(x_i) + R(f).$$

В этом случае узлы есть нули полиномов Чебышёва I рода $T_n(x)$, которые с указанным весом ортогональны на $[-1,1]$. Т.к. $T_n = \cos(n \arccos x)$, то нули находим из $n \arccos x = (k+1/2)\pi$:

$$x_k = \cos\frac{(2k+1)\pi}{2n}, \quad k = -1, \dots, -n,$$

или в развернутой записи

$$\{x_i\}_1^n = \Big\{ \cos\frac{\pi}{2n}, \cos\frac{3\pi}{2n}, \cos\frac{5\pi}{2n}, \dots, \cos\frac{(2n-1)\pi}{2n} \Big\}.$$

Веса w_i вычислим по формуле (3.83).

Из (3.2.4.4)

$$\|T_0\|^2 = \pi, \quad \text{а для} \quad n \geq 1 \quad \|T_n\|^2 = \pi/2.$$

Поэтому

$$w_i^{-1} = \frac{\cos^2(0)}{\pi} + \frac{2}{\pi}\sum_{k=1}^{n-1}\cos^2(k\arccos x_i) = \frac{1}{\pi} + \frac{1}{\pi}\sum_{k=1}^{n-1}\left[1+\cos(2k\arccos x_i)\right] =$$
$$= \frac{n}{\pi} + \frac{1}{\pi}\sum_{k=1}^{n-1}\cos(2k\Theta_i), \quad \text{где} \quad \Theta_i = \arccos x_i.$$

Заметим, что, так как x_i есть нули T_n, то $\cos(n\Theta_i)=0$.

Возьмем первый и последний члены из суммы. Сумма косинусов это

$$\cos\alpha + \cos\beta = 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2},$$

так что

$$\cos 2\Theta + \cos[2(n-1)\Theta_i] \sim \cos n\Theta_i = 0.$$

Так же комбинируя попарно оставшиеся члены, получим что суммы пропорциональны полусумме аргументов, то есть $\cos n\Theta_i$ и обращаются в ноль. Поэтому и вся сумма равна нулю и остается просто

$$w_i = \frac{\pi}{n}.$$

Остаточный член получаем из (3.85), учитывая что (для $n>0$) $\|T_n\|^2 = \pi/2$, а старшие коэффициенты равны 2^{n-1} (например из рекуррентных соотношений). Собирая вместе все результаты, получаем формулу Гаусса-Чебышёва в явном виде:

$$\int_{-1}^{1}\frac{f(x)dx}{\sqrt{1-x^2}} = \frac{\pi}{n}\sum_{k=1}^{n} f(x_k) + \frac{2\pi}{(2n)!\,2^{2n}}f^{(2n)}(\xi), \quad x_k = \cos\frac{(2k-1)\pi}{2n}. \tag{3.96}$$

В частности, при $n=5$ остаточный член $R_5(f) = \frac{\pi}{1{,}8\cdot 10^9}f^{(10)}(\xi)$.

3.4.2.7 Другие квадратурные формулы*

Формула Гаусса-Кронрода. Как было указано выше, для проверки точности вычисления интеграла по формуле Гаусса относительно эффективным способом является двойной пересчет. Он хорошо работает, если идет речь скажем о формулах Гаусса-Лежандра с небольшими n, для которых узлы и веса давно посчитаны, и нужно только подставить нужные значения, а также значения функции в этих узлах вычисляются быстро. Но может быть и так, что функция под интегралом сложная, и вычисление ее значений в узлах занимает

большую часть времени расчета. Тогда вычислять интеграл два раза, с разными n, становится не очень эффективно, так как узлы для разных n никогда не совпадают.

Оказывается, что формулу Гаусса в связи с этим можно модифицировать следующим образом. Пусть у нас есть формула для n узлов точности $2n-1$. Если к узлам добавить еще специальным образом выбранные $n+1$ узел, получим формулу точности $3n+1$. Такое обобщение формулы Гаусса называется формулой *Гаусса-Кронрода (Gauss-Kronrod)*. Таким образом, при вычислении с повышением точности, вычисленные ранее значения функции в узлах используются повторно. Разницу двух результатов можно использовать как оценку точности. Приведем узлы и веса для формулы "G7K15" с 15 узлами

Узлы Гаусса		Веса w_i
±0.94910 79123 42759	*	0.12948 49661 68870
±0.74153 11855 99394	*	0.27970 53914 89277
±0.40584 51513 77397	*	0.38183 00505 05119
0.00000 00000 00000	*	0.41795 91836 73469
Узлы Кронрода		
±0.99145 53711 20813		0.02293 53220 10529
±0.94910 79123 42759	*	0.06309 20926 29979
±0.86486 44233 59769		0.10479 00103 22250
±0.74153 11855 99394	*	0.14065 32597 15525
±0.58608 72354 67691		0.16900 47266 39267
±0.40584 51513 77397	*	0.19035 05780 64785
±0.20778 49550 07898		0.20443 29400 75298
0.00000 00000 00000	*	0.20948 21410 84728

Метод Кленшоу-Куртиса. Есть и другие алгоритмы численного интегрирования, не основанные на формуле Гаусса. Среди них можно упомянуть метод *Кленшоу-Куртиса (Clenshaw–Curtis quadrature)*. Идея заключается в следующем. Путь надо вычислить интеграл $\int_{-1}^{1} dx\, f(x)$. Перейдя к новой переменной, $x = \cos\Theta$, его можно переписать как

$$\int_{-1}^{1} dx\, f(x) = \int_{0}^{\pi} d\Theta \sin\Theta f(\cos\Theta).$$

$f(\cos\Theta)$ – четная функция Θ, которую можно разложить в косинус-ряд Фурье[44]

$$f(\cos\Theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos k\Theta,$$

а этот ряд можно проинтегрировать почленно и получить ◂…▸

$$\int_{0}^{\pi} d\Theta \sin\Theta f(\cos\Theta) = a_0 + \sum_{k=1}^{\infty} \frac{2a_{2k}}{1-(2k)^2}.$$

[44] В тригонометрическом ряде Фурье для четной f все коэффициенты при синусах равны нулю

Таким образом, задача свелась к нахождению коэффициентов ряда Фурье

$$a_k = \frac{2}{\pi}\int_0^\pi d\Theta f(\cos\theta)\cos(k\theta).$$

На первый взгляд, это не является упрощением. Однако на самом деле такие интегралы считаются очень быстро благодаря FFT (см. п.1.3.3), а сам ряд быстро сходится. Считая численно интеграл для коэффициентов методом трапеций, мы фактически заменяем его дискретным косинус-преобразованием Фурье (DCT)

$$a_k \approx \frac{2}{N}\left[\frac{f(1)}{2} + \frac{f(-1)}{2}(-1)^k + \sum_{n=1}^{N-1} f\Big(\cos\frac{n\pi}{N}\Big)\cos\frac{nk\pi}{N}\right].$$

Для нужных нам четных коэффициентов имеем DCT порядка $N/2$

$$a_{2k} \approx \frac{2}{N}\left[\frac{f(1)+f(-1)}{2} + f(0)(-1)^k + \sum_{n=1}^{N/2-1}\Big\{f\Big(\cos\frac{n\pi}{N}\Big) + f\Big(-\cos\frac{n\pi}{N}\Big)\Big\}\cos\frac{nk\pi}{N/2}\right],$$

который быстро считается методом FFT (например для $N = 2^p$).

Заметим, что для счета интеграла в итоге используются значения функции в максимумах полинома Чебышёва T_N: $x_k = \pm\cos\frac{n\pi}{N}$. Это неудивительно, если вспомнить, что тригонометрический ряд Фурье по $\Theta = \arccos x$ это то же самое что ряд Фурье по ортогональным полиномам $T_n(x)$.

Алгоритм быстро сходится, имеет точность, сравнимую с методом Гаусса, и последовательные приближения $N \to 2N$ используют перекрывающиеся наборы узлов.

Литература

Функциональный анализ

- А.Н. Колмогоров, С.В. Фомин, *Элементы теории функций и функционального анализа*, [11]. Классический университетский учебник по функциональному анализу.
- В.М. Кадец. *Курс функционального анализа*, [12] Хороший современный учебник.
- В.И. Смирнов, *Курс высшей математики, том 5*, [13]. Пятая часть известного фундаментального курса высшей математики.
- Allan Pincus, *Weierstrass and Approximation Theory*, *J. Approx. Theory*, **107** (2000), 1-66, [14]. Симпатичная статья о теоремах Вейерштрасса.

Ортогональные полиномы

- V. Totik, Orthogonal polynomials, *Surveys in Approximation Theory* (2005), **1**, 70-125, [15]. Обзорная статья по общей теории ортогональных полиномов на вещественной оси. Много всего интересного плюс исчерпывающий список литературы.
- А.Ф. Никифоров, С.К. Суслов, *Классические ортогональные полиномы*, [16]. Интересная брошюра по классическим полиномам, также в ней излагается теория ортогональных полиномов дискретной переменной, цилиндрические функции и другие обобщения.
- Дополнительные справочные сведения и литература см. в приложении C.

Среднеквадратичное приближение

- Березин И.С, Жидков Н.П, *Методы вычислений*, [9]. Последовательное изложение среднеквадратичного приближения и ортогональных полиномов в близком подходе.

Численное интегрирование

- Stoer J. and Bulirsch R. *Introduction to Numerical Analysis*, [1].
- John A. Gubner, Gaussian quadrature and the eigenvalue problem, [17], (см. gubner.ece.wisc.edu);

Сетевые ресурсы: Справочные сведения по ортогональным полиномам – wiki: Orthogonal polynomials, также много на MathWorld и wolfram|alpha; на alglib.sources.ru есть разные статьи, библиотеки и коды; на wiki: gaussian quadrature есть, среди прочего, ссылки на библиотеки и собрания алгоритмов по численному интегрированию.

Глава 4

Интегральные уравнения

Integral Equations

4.1 Постановка задачи и терминология

Рассмотрим линейные интегральные уравнения на вещественной оси: *Фредгольма (Fredholm) I и II рода*

$$\text{I} \quad : \qquad \lambda \int_a^b ds\, K(x,s) y(s) = f(x) \tag{4.1}$$

$$\text{II} \quad : \quad y(x) - \lambda \int_a^b ds\, K(x,s) y(s) = f(x) \tag{4.2}$$

и *Вольтéрра (Volterra) I и II рода*

$$\text{I} \quad : \qquad \lambda \int_a^x ds\, K(x,s) y(s) = f(x) \tag{4.3}$$

$$\text{II} \quad : \quad y(x) - \lambda \int_a^x ds\, K(x,s) y(s) = f(x). \tag{4.4}$$

Функция $K(x,s)$ называется *ядром (kernel)* интегрального уравнения.

Формально уравнение Вольтерра можно рассматривать как частный случай уравнения Фредгольма с ядром, для которого

$$K(x,s) = 0 \quad \text{при } s > x,$$

но вследствие особенностей решения его иногда целесообразно выделять в отдельных класс.

Если правая часть уравнения, $f(x)$, равна нулю, то уравнения называются однородными, иначе – неоднородными.

Ядро называется *эрмитовым (Hermite)*, если $K(x,s)=\overline{K(s,x)}$ (черта означает комплексное сопряжение). В вещественных задачах ядро называется *симметричным (или симметрическим, или symmetric)* если $K(x,s)=K(s,x)$. Мы здесь в основном ограничимся вещественным случаем.

Если $K(x,s)=K(x-s)$, то K – *разностное ядро (difference kernel).*

Ядро называется *полярным (polar)*, если оно представимо в виде

$$K(x,s) = \frac{\Phi(x,s)}{|x-s|^\alpha}, \quad 0<\alpha<1,$$

где $\Phi(x,s)$ непрерывна.

Если ядро представимо в виде конечной суммы

$$K(x,s) = \sum_{i=1}^{n} \psi_i(x)\varphi_i(s), \tag{4.5}$$

то оно называется *вырожденным (degenerate).*

4.2 Метод квадратур. Замена интеграла конечной суммой

Будем предполагать, что $K(x,s)$ и $f(x)$ непрерывны и, более того, имеют непрерывные производные до некоторого порядка. Тогда логично предположить, что и решения уравнения будут иметь производные того же порядка.

Для решения уравнения интеграл, входящий в него, заменяется конечной суммой, при помощи какой-либо квадратурной формулы, т.е.

$$\int_a^b dx\, F(x) = \sum_{j=1}^{n} \mathrm{w}_j F(x_j) + R(F), \tag{4.6}$$

где $x_1, x_2, \ldots, x_n \in [a,b]$ это набор узлов, w_i – весовые коэффициенты, не зависящие от x и F, а $R(F)$ – остаточный член.

Запишем интегральное уравнение, к примеру, неоднородное уравнение Фредгольма II рода (4.2), в каждом узле

$$y(x_i) - \lambda \int_a^b ds\, K(x_i,s)y(s) = f(x_i), \quad i = 1,\ldots,n,$$

и заменим в каждом уравнении интеграл конечной суммой (4.6). Получим систему n алгебраических уравнений с n неизвестными

$$y_i - \lambda \sum_{j=1}^{n} \mathrm{w}_j K_{ij} y_j = f_i, \quad j = 1, \ldots, n, \tag{4.7}$$

где введены очевидные обозначения $y_i = y(x_i)$, $f_i = f(x_i)$ и

$$K(x_i, x_j) = K_{ij}.$$

Решая ее, получим $\{y_i\}_1^n$ – значения решения в узлах $\{x_i\}_1^n$. В случае уравнения Вольтерра матрица алгебраической системы имеет треугольный вид и потому особенно быстро решается (см. п. 4.7).

Функция

$$Y(x) = f(x) + \lambda \sum_{j=1}^{n} \mathrm{w}_j K(x, x_j) y_j$$

в узлах $\{x_i\}_1^n$ принимает значения $\{y_i\}_1^n$, и следовательно, может считаться аналитическим выражением для приближенного решения.

- Если в качестве квадратурной формулы брать формулу прямоугольников, то

$$x_1 = a + \tfrac{h}{2},\ x_2 = a + \tfrac{3h}{2},\ \ldots,\ x_n = a + \tfrac{(2n-1)h}{2};$$
$$\mathrm{w}_1 = \mathrm{w}_2 = \ldots = \mathrm{w}_n = \tfrac{b-a}{n} = h.$$

- Если в качестве квадратурной формулы брать формулу Симпсона, то $n = 2m+1$, $h = \frac{b-a}{2m}$ и

$$x_1 = a,\ x_2 = a + h,\ \ldots,\ x_{2m+1} = a + 2mh;$$
$$\mathrm{w}_1 = \mathrm{w}_{2m+1} = \tfrac{h}{3};\quad \mathrm{w}_2 = \mathrm{w}_4 = \ldots = \mathrm{w}_{2m} = \tfrac{4h}{3};\quad \mathrm{w}_3 = \mathrm{w}_5 = \ldots = \mathrm{w}_{2m-1} = \tfrac{2h}{3}.$$

Следует иметь в виду, что чем более точная квадратурная формула применяется, тем выше требования, предъявляемые к ядру K и правой части f. Попытка применения более точных квадратурных формул для получения более точного приближения при несоблюдении этих требований может привести к обратному эффекту.

Однородное уравнение решается по такой же схеме и приводится к однородной системе линейных уравнений (4.7), с $f = 0$. Нетривиальные решения ее, как известно, существуют, если определитель системы равен нулю. Тогда из этого условия получаем для λ характеристическое уравнение n-й степени. Его корни $\lambda_1, \ldots, \lambda_n$ – собственные значения. Соответствующие им решения алгебраической системы (4.7) с $f = 0$ дают приближенные выражения для (некоторых из) собственных функций ядра $K(x, s)$.

Как известно, структура решений неоднородной системы зависит от того, является ли λ собственным значением однородной системы. Если нет, то решение единственно, иначе решения или не существует, или оно не единственно.

Увеличивая число узлов n, мы можем, в принципе, получить сколь угодно большое число собственных значений и собственных решений системы (4.7). Это наводит на мысль, что интегральные уравнения, грубо говоря, представляют собой в некотором роде бесконечную систему линейных уравнений, со свойствами, похожими на свойства конечных систем. Ниже мы увидим, что действительно, интегральные уравнения определенного класса представимы как операторные в гильбертовом пространстве функций L^2.

Пример

Решается интегральное уравнение

$$y(x) + x \int_0^1 ds\,(e^{xs} - 1)y(s) = e^x - x \tag{4.8}$$

Применим формулу Симпсона сразу для всего промежутка интегрирования: $n{=}3$ и $\int_0^1 f(s)ds \approx \frac{1}{6}\left(f(0) + 4f(1/2) + f(1)\right)$. Тогда имеем

$$y(x) + \frac{x}{6}\left[(e^{x\cdot 0} - 1)y_0 + 4(e^{x/2} - 1)y_{1/2} + (e^x - 1)y_1\right] = e^x - x.$$

Подставляя $x = 0, \frac{1}{2}, 1$, получаем систему

$$\begin{aligned}
x = 0: &\quad y_0 = 1,\\
x = \tfrac{1}{2}: &\quad \tfrac{1}{3}(e^{1/4} + 2)y_{1/2} + \tfrac{1}{12}(e^{1/2} - 1)y_1 = e^{1/2} - \tfrac{1}{2},\\
x = 1: &\quad \tfrac{2}{3}(e^{1/2} - 1)y_{1/2} + \tfrac{1}{6}(e + 5)y_1 = e - 1.
\end{aligned}$$

Решая ее,получим $y_0 = 1,\ y_{1/2} = 0.999,\ y_1 = 0.9996$.

Учитывая, что точное решение исходного уравнения есть $y \equiv 1$, результат довольно хороший.

4.2.1 Некоторые приемы борьбы с особенностями

Рассмотрим их опять на примере неоднородного уравнения Фредгольма II рода (4.2).

1. Пусть $K(x,s)$ – гладкая, а $f(x)$ имеет особенности. Сделаем замену $y(x) = f(x) + z(x)$, и подставляя в уравнение, получим

$$z(x) + f(x) - \lambda \int_a^b ds\,K(x,s)z(s) - \lambda \int_a^b ds\,K(x,s)f(s) = f(x)$$

$$\Rightarrow z(x) - \lambda \int_a^b ds\,K(x,s)z(s) = \lambda \int_a^b ds\,K(x,s)f(s)$$

и правая часть нового уравнения более гладкая, чем была, благодаря "сглаживанию" интегрированием.

2. Пусть $K(x,s)$ и ее производные по s имеют разрывы при $x=s$ (это часто встречающийся случай, например в уравнениях Вольтерра или когда ядро есть функция Грина). Тогда перепишем уравнение так:

$$y(x)\left[1-\lambda\int_a^b ds\,K(x,s)\right]-\lambda\int_a^b ds\,K(x,s)\big(y(s)-y(x)\big)=f(x).$$

В квадратных скобках нет искомой функции и интеграл может быть вычислен явно.

Функция $K(x,s)\big(y(s)-y(x)\big)$, которая теперь стоит под интегралом, имеет класс гладкости на единицу выше, чем исходное ядро $K(x,s)$ при $x=s$.

4.3 Интегральные уравнения как операторные в L^2

4.3.1 Интегральный оператор Фредгольма

Мы видели, что от особенностей неоднородности f и ядра K можно эффективно избавляться переходом к новой искомой функции или новому ядру. Рассмотрим уравнения с достаточно "хорошим" ядром K.

Пусть $K(x,s)$ интегрируемо с квадратом, т.е.

$$M\equiv\int_a^b\int_a^b dx\,ds\,|K(x,s)|^2<\infty, \tag{4.9}$$

и будем искать $y(x)$ в классе функций $L^2[a,b]$.

Сопоставим уравнению (4.1) $\lambda\int ds\,K(x,s)y(s)=f(x)$ линейный оператор A, определяемый так, что

$$\psi=A\,y \quad\Leftrightarrow\quad \psi(x)=\int_a^b ds\,K(x,s)y(s). \tag{4.10}$$

Этот оператор, действующий в бесконечномерном пространстве $L^2[a,b]$, называется *оператором Фредгольма* с ядром K. Уравнения (4.1) и (4.2) с ядром, удовлетворяющим условию (4.9), и с $f(x),y(x)\in L^2[a,b]$ называются уравнениями Фредгольма (в строгом смысле).

Исследование уравнения (4.1) сводится к изучению свойств линейного оператора A (4.10) и *абстрактного уравнения Фредгольма I рода*

$$Ay = \Lambda f, \quad \text{где } \Lambda = 1/\lambda.$$

Таким образом, *однородные* уравнения (4.1-4.4) сводятся к задаче на собственные вектора и собственные значения в пространстве L^2: нужно найти значения λ (собственные значения), при которых уравнение имеет нетривиальное решение (собственный вектор). Собственные значения оператора A называют также собственными значениями ядра K, а его собственные вектора – собственными функциями ядра.

В *неоднородных уравнениях* (4.1-4.4) параметр λ обычно полагается заданным, и задача, грубо говоря, представляет собой бесконечную неоднородную систему линейных уравнений. Существование решений зависит от того, является λ собственным значением оператора A или нет.

В бесконечномерном пространстве для задачи на собственные вектора и значения становятся существенными вопросы сходимости, а следовательно и ограниченности, непрерывности, компактности и пр.

4.3.2 Линейные операторы в L^2

Рассмотрим гильбертово пространство $L^2[a,b]$ со скалярным произведением с весом $p(x)=1$:

$$(f,g) = \int_a^b dx\, f(x)\overline{g(x)}. \tag{4.11}$$

Норма и метрика, как обычно, индуцированы скалярным произведением. В скалярном произведении (4.11), как мы знаем, можно в качестве ортогональной системы в L^2 взять, например, полиномы Лежандра или последовательность тригонометрических полиномов (см. п.3.1.10) $\{\sin kx, \cos kx\}$.

Ограниченные операторы. *Нормой линейного оператора*, действующего в гильбертовом пространстве H, назовем число

$$\|A\| = \sup_{\|y\|=1} \|Ay\|. \tag{4.12}$$

Везде ниже речь будет идти также только о *линейных* операторах.

Имея норму, можно теперь в *пространстве операторов* в H индуцировать метрику как $\rho(A,B) = \|A-B\|$ и определить предел последовательности, бесконечные ряды, сходимость и непрерывность.

Множество B в нормированном пространстве называется ограниченным, если

$$\exists \alpha > 0 \mid \quad \forall\, b \in B \quad \|b\| \le \alpha.$$

Если у оператора A есть конечная норма $\|A\|$, то оператор называется *ограниченным (bounded)*. Таким образом, ограниченный оператор переводит всякое ограниченное множество в ограниченное. Пример неограниченных операторов – дифференциальные операторы[1].

Ограниченность оператора Фредгольма. В силу неравенства Коши-Буняковского, для почти всех x имеем

$$|\psi(x)|^2 = \left| \int_a^b ds\, K(x,s) y(s) \right|^2 \le \int_a^b ds |K(x,s)|^2 \cdot \int_a^b ds |y(s)|^2 = \int_a^b ds |K(x,s)|^2 \cdot \|y\|^2.$$

Интегрируя по x, получаем неравенство

$$\|Ay\|^2 = \int_a^b ds\, |\psi^2(s)| \le \|y\|^2 \cdot \int_a^b \int_a^b dx\, ds |K(x,s)|^2 = \|y\| \cdot M,$$

откуда видим, что $\psi \in L^2$, и получаем теорему:

$\mathbf{T}^0$: Оператор Фредгольма A ограничен, с нормой

$$\|A\| \le \sqrt{M}. \tag{4.13}$$

Непрерывные операторы. Оператор A *непрерывен в точке* y, если

$$\forall \varepsilon > 0\ \exists\, \delta(\varepsilon) \mid \big\{ \|y - z\| < \delta \Rightarrow \|Ay - Az\| < \varepsilon \big\};$$

оператор *непрерывен (continuous)*, если он непрерывен $\forall y \in H$.

$\mathbf{T}^0$: Оператор непрерывен т.и т.т., когда он ограничен $\blacktriangleleft \ldots \blacktriangleright$.

4.3.3 Вырожденное ядро и спектр вырожденного оператора

Пусть ядро является вырожденным, то есть имеет вид (4.5)

$$K(x,s) = \sum_{i=1}^{n} \psi_i(x) \varphi_i(s),$$

где n – конечное число. Не ограничивая общности, можем считать, что функции $\psi_i(x)$ – линейно-независимые[2].

[1] Например, $y = x^{-1/4} \in L^2_{[0,1]}$ (т.е. $\exists \int_0^1 dx\, y^2$), но $y' \sim x^{-5/4}$, и интеграл $\int dx (y')^2$ в нуле расходится. Таким образом, оператор d/dx не имеет конечной нормы L^2.

[2] В противном случае мы всегда можем переписать сумму через линейно-независимую подсистему.

Такому ядру соответствует оператор A, который переводит всё $L^2[a,b]$ в конечномерное подпространство, порожденное векторами $\psi_1, \ldots, \psi_n$:

$$Ay = \int_a^b ds\, K(x,s)y(s) = \sum_{i=1}^n \psi_i(x) \int_a^b ds\, \varphi_i(s)y(s)$$

Будем называть такой оператор *вырожденным*.

Подставляя (4.5) в уравнение (4.2), получим

$$y(x) = f(x) + \lambda \sum_{i=1}^n \psi_i(x) \int_a^b ds\, \varphi_i(s)y(s). \tag{4.14}$$

Перепишем интегралы в терминах скалярного произведения (4.11) и введем обозначения для коэффициентов:

$$y(x) = f(x) + \sum_i C_i\psi_i(x) \quad \text{где} \quad C_i = \lambda \int ds\, \varphi_i(s)y(s) \equiv \lambda(\varphi_i, y). \tag{4.15}$$

Подставляя y в таком виде в скалярное произведение, получим

$$\sum_i C_i\psi_i(x) = \lambda \sum_i \psi_i(x)\Big(\varphi_i, f(x) + \sum_j C_j\psi_j(x)\Big),$$

и в итоге

$$\sum_{i=1}^n \psi_i(x) \left\{ C_i - \lambda \sum_j \alpha_{ij}C_j - \lambda f_i \right\} = 0,$$

$$\text{где} \begin{cases} f_i = (\varphi_i, f) = \int ds\, \varphi_i(s)f(s), \\ \alpha_{ij} = (\varphi_i, \psi_j) = \int ds\, \varphi_i(s)\psi_j(s). \end{cases}$$

Воспользовавшись линейной независимостью ψ_i, получаем *точную конечную* линейную систему уравнений для коэффициентов C_i:

$$C_i - \lambda \sum_{j=1}^n \alpha_{ij}C_j = \lambda f_i, \quad i = 1, 2, \ldots, n. \tag{4.16}$$

Равенство нулю определителя системы дает уравнение для λ – это уравнение для собственных значений оператора Фредгольма, который в рассматриваемом случае вырожден. Соответствующие собственные решения есть решения однородного уравнения. Как видно, их конечное число n.

Таким образом, мы видим, что *вырожденный* оператор имеет лишь *конечное* число собственных значений и векторов, т.е. его спектр состоит из конечного числа точек.

4.3.3.1 Замена ядра на вырожденное

Пусть ψ_n – замкнутая (и полная) ортогональная система в $L^2[a,b]$. Тогда всевозможные попарные произведения $\psi_m(x)\psi_n(s)$ образуют замкнутую систему в пространстве $L^2([a,b]\times[a,b])$, и следовательно всякое квадратично интегрируемое ядро K можно представить в виде ряда Фурье по $\psi_i(x)\psi_j(s)$:

$$K(x,s) = \sum_m \sum_n a_{mn}\psi_m(x)\psi_n(s).$$

Но всякая частичная сумма

$$K_N(x,s) = \sum_{m,n=1}^{N} a_{mn}\psi_m(x)\psi_n(s) \tag{4.17}$$

является *вырожденным* ядром K_N, и соответствует некоторому вырожденному оператору A_N.

С другой стороны, ряд Фурье K_N сходится к K по норме L^2, то есть

$$\int_a^b \int_a^b dx\,ds\,[K(x,s)-K_N(x,s)]^2 \underset{N\to\infty}{\longrightarrow} 0,$$

что в терминах нормы оператора означает

$$\|A - A_N\| \to 0. \tag{4.18}$$

Поэтому точное решение интегрального уравнения можно получить как предел решений соответствующих уравнений с вырожденными ядрами[3].

4.3.4 Компактность и компактные операторы*

Класс операторов, по своим свойствам близких к операторам, действующим в конечномерных пространствах, образуют так называемые *вполне непрерывные*, или *компактные (compact)* операторы[4]. Для их определения напомним понятия компактности последовательностей и множеств в метрических пространствах[5].

Здесь мы ограничимся минимальным набором определений и теорем, а интересующихся доказательствами отошлем к приложению D и литературе.

[3] Подробнее см. ниже, в п.4.3.7.

[4] Примерно в том же смысле что гильбертово пространство, среди бесконечномерных пространств, близко по свойствам к конечномерным пр.-вам тем, что в нем всякий вектор можно разложить по базису.

[5] Компактность наиболее общим образом вводится в топологических пространствах. Метрические линейные пространства, однако, всегда отделимы (хаусдорфовы), и в них определения можно дать проще.

Последовательность y_n называется компактной, если из нее можно выделить сходящуюся подпоследовательность.

Множество K называется компактным (или компактом), если всякая последовательность его элементов компактна в K, т.е. если из нее можно выделить подпоследовательность, сходящуюся к элементу K.

Множество называется *предкомпактным* (или относительно компактным, или предкомпактом), если его замыкание компактно. Т.е. мн-во предкомпактно, если любая последовательность его элементов компактна (не обязательно в K).
$\mathbf{T}^0$: В конечномерном евклидовом пространстве E_n множество компактно тогда и только тогда, когда оно ограничено и замкнуто[6] ◂…▸.

В бесконечномерном пространстве E_∞ всякое компактное множество так же ограничено и замкнуто. Однако для доказательства достаточности существенным является условие конечномерности, и в E_∞ оно не выполняется. Поэтому свойство компактности оказывается более сильным, чем просто требования ограниченности и замкнутости.

Так, возьмем в пространстве l^2 последовательность векторов $\{x_i\}$ такую: $(1,0,0,\ldots)$, $(0,1,0,\ldots)$ и так далее, так что i-й вектор имеет на i-й позиции единицу, а на остальных нули. Тогда $\forall\, i\ \|x_i\| = 1$, так что последовательность ограничена. Однако для любой пары $\|x_i - x_j\| = \sqrt{2}$, так что любая подпоследовательность из бесконечного числа элементов не сходится. Таким образом, все точки $\{x_i\}$ изолированы, и эта последовательность ограничена и замкнута, но не компактна. Эта последовательность лежит на единичной сфере в l^2. Поэтому и единичная сфера, которая очевидно ограничена и замкнута, – не компактна (и не предкомпактна).

Примеры компактов в банаховых пространствах:

- Компактным является любое ограниченное и замкнутое подмножество конечномерного подпространства E_∞.
- *Фундаментальный параллелепипед**("гильбертов кирпич") пространства l^2 – множество точек $(x_1, x_2, \ldots)$ с координатами $|x_i| \le 1/2^{i-1}$, – компактен.

Оператор A называется компактным, если для всякой ограниченной последовательности y_n последовательность Ay_n компактна. Таким образом, ком-

[6] См. теорему Больцано-Вейерштрасса в мат. анализе, в которой это доказывается для числовой прямой.

пактный оператор переводит всякое ограниченное множество в предкомпактное.

Так как компактное и предкомпактное множество ограничено, то компактный оператор непрерывен и ограничен. Однако обратное неверно. Так, единичный оператор переводит единичную сферу, которая ограничена но не компактна, в себя, а потому ограничен но не компактен!

T^0: Линейная комбинация и произведение компактных операторов компактны ◂…▸*

T^0: Если последовательность компактных операторов в банаховом пространстве сходится, то ее предел есть также компактный оператор ◂…▸.

Компактность оператора Фредгольма. Теорема:

T^0. *Оператор Фредгольма* (4.9, 4.10) *компактен.*

◂ Ограниченность уже доказана.

Как мы видели, разложив ядро в ряд Фурье (4.17), оператор Фредгольма A можно представить как предел последовательности вырожденных операторов (4.18) A_N. Но вырожденный оператор A_N, по определению, переводит все пространство $L^2[a,b]$ в конечномерное подпространство, порожденное векторами $\psi_1,\dots,\psi_N$. Поэтому он переводит любую ограниченную последовательность из L^2 в ограниченную последовательность в конечномерном подпространстве L^2, а из последней всегда можно выделить сходящуюся подпоследовательность. Поэтому A_N компактны. Следовательно, по теореме о сходящейся последовательности компактных операторов, A компактен. ▸

4.3.5 Компактные эрмитовы операторы*

Оператор A, действующий в гильбертовом пространстве H, называется *самосопряженным (self-adjoint)*, если $\forall\, x,y\in H\ (Ax,y)=(x,Ay)$.

Обычно в пространствах над $\mathbb{C}$ эти операторы еще называют эрмитовыми (в руссокоязычной и физической литературе). В такой терминологии

эрмитово ядро K соответствует **эрмитовому оператору** A, а

симметричное ядро K – **самосопряженному оператору** A.

Однородное уравнение. Теорема Гильберта-Шмидта распространяет на компактные эрмитовы операторы в гильбертовом пространстве известный факт о приведении матрицы самосопряженного оператора в конечномерном евкли-

довом пространстве к диагональной форме в некотором ортонормированном базисе:

T^0 (Гильберта-Шмидта):

◂…▸* *Для всякого эрмитового компактного оператор A в гильбертовом пространстве H существует (конечная или бесконечная) ортонормированная система собственных векторов $\{\varphi_i\}$, отвечающих вещественным собственным значениям $\{\Lambda_i\}$, такая что всякий вектор $\xi \in H$ представляется единственным образом в виде*

$$\xi = \sum c_k \varphi_k + \xi', \tag{4.19}$$
$$\text{где} \quad A\xi' = 0, \quad \text{так что} \quad A\xi = \sum \Lambda_k c_k \varphi_k,$$
$$\text{и} \quad \lim_{n\to\infty} \Lambda_n \to 0.$$

При этом каждому ненулевому собственному значению соответствует конечное число векторов, а максимальное из собственных значений равно норме оператора $\Lambda_1 = \|A\|$.

Теорема означает, что для всякого эрмитового компактного оператора A существует ортогональный базис из его собственных векторов. Если множество $\{\varphi_i\}$ бесконечно, то оно и является базисом, а если нет, то нужно его дополнить произвольным базисом подпространства, которое A переводит в ноль (т.е. подпространства векторов, соответствующих Λ=0).

Число собственных функций конечно $\Leftrightarrow$ ядро вырождено.

Процедура построения набора такая же как в конечномерном случае (см. теорему о собственном базисе эрмитового оператора). Наибольшее из собственных значений $\Lambda_1 = \|A\|$ и соответствующий ему вектор y_1 есть решения вариационной задачи $\Lambda_1 = \max\{\|Ay\| : y \in H, \|y\|=1\}$.

Следующее по абсолютной величине собственное значение – решение той же вариационной задачи в ортогональном дополнении H_1 к линейной оболочке y_1. Можно показать, что ядро

$$K^{(2)}(x,s) = K(x,s) - \Lambda_1 y_1(x) y_1(s)$$

имеет те же собственные значения и функции, что и $K(x,s)$, за исключением $\{y_1, \Lambda_1\}$. Таким образом можно построить последовательность собственных значений и функций $K(x,s)$. Если на каком-то шаге получаем $K^{(m)} \equiv 0$, то последовательность обрывается и значит ядро вырождено.

В квантовой механике утверждается, что вектор состояния физической системы всегда можно разложить по собственным состояниям операторов наблюдаемых величин. Это означает, что наблюдаемые в квантовой механике представлены эрмитовыми операторами, действующими в гильбертовом пространстве состояний квантовомеханической системы.

Неоднородное уравнение решается разложением по собственным функциям ядра. Пусть набор ортонормированных собственных функций и значений однородного уравнения $\lambda Ay = y$ есть $\{y_i, \lambda_i\}_{i=1}^n$ (конечное n на самом деле не представляет особого интереса, так как соответствует вырожденному ядру и такая задача сводится к конечной системе линейных уравнений). Тогда рассматриваем неоднородное уравнение $y - \lambda Ay = f$ и ищем решения в виде $y = f + \sum c_i y_i$, где суммирование производится по всему набору собственных функций y_i, конечному или нет.

Подставляя в уравнение, получаем

$$f + \sum c_i y_i - \lambda A(f + \sum c_i y_i) = f, \quad \Rightarrow \quad \sum c_i y_i (1 - \lambda/\lambda_i) = \lambda A f.$$

Раскладываем f по системе y_i: $f = \sum f_i y_i + f_\perp$, где $f_i = (f, y_i)$, а $f_\perp$ – проекция f на ортогональное дополнение $H_\perp$ к линейной оболочке $\{y_i\}$, которое непустое если n конечно. При этом $Af_\perp = 0$, иначе норма $\|A\|$ в подпространстве $H_\perp$ была бы отлична от нуля и значит существовал бы еще один собственный вектор. Тогда $\lambda Af = \lambda \sum \lambda_i^{-1} f_i y_i$ и собирая слагаемые, получаем $\sum y_i \lambda_i^{-1} \{c_i(\lambda_i - \lambda) - \lambda f_i\} = 0$. Таким образом, получили систему уравнений (возможно бесконечную)

$$c_i(\lambda_i - \lambda) = \lambda f_i, \quad i = 1, \dots, n.$$

- Если λ не совпадает ни с одним из собственных значений λ_i, то получаем $c_i = \frac{\lambda f_i}{\lambda - \lambda_i}$ и решение уравнения имеет вид

$$y(x) = f(x) + \lambda \sum_{i=1}^{n} \frac{f_i y_i(x)}{\lambda - \lambda_i}.$$

- Если λ совпадает с одним из значений $\lambda = \lambda_p$, то обозначим все собственные вектора, соответствующие λ_p через $y_{p1}, \dots, y_{pm}$. Тогда для c_i с $i \neq p1, \dots, pm$ имеем те же выражения что и раньше, а для $i = p1, \dots, pm$ имеем

$$c_i \cdot 0 = \lambda f_i, \quad i = p1, \dots, pm.$$

  – Тогда если $f_i = 0$ для $i = p1, \dots, pm$, т.е. если неоднородность f ортогональна всем собственным функциям A, соответствующим собственному значению $\lambda = \lambda_p$, то соответствующие коэффициенты c_i *произвольны*, и решение имеет вид

$$y(x) = f(x) + \lambda {\sum_{i=1}^{n}}' \frac{f_i y_i(x)}{\lambda - \lambda_i} + \sum_{i=p1}^{pm} c_i y_i(x).$$

Штрих над суммой означает, что в нее не входят слагаемые с $i = p1, \ldots, pm$.

– Иначе – решений нет.

4.3.6 Теоремы Фредгольма

Для произвольных компактных операторов, вообще говоря, не выполняется теорема Гильберта-Шмидта. Однако, воспользовавшись тем, что компактный оператор всегда можно представить как предел последовательности вырожденных, можно сформулировать некоторые общие свойства. Так, несложно убедиться в том, что если λ есть собственное значение однородного уравнения Фредгольма II рода (4.2)

$$y(x) - \lambda \int_a^b ds\, K(x,s) y(s) = f(x), \tag{4.20}$$

то $\overline{\lambda}$ есть собственное значение сопряженного к нему (или союзного) уравнения

$$z(x) - \overline{\lambda} \int_a^b ds\, \overline{K(s,x)} z(s) = g(x). \tag{4.21}$$

Для уравнений с произвольным ядром, удовлетворяющим условию (4.9), верны три **теоремы Фредгольма** ◂…▸☆:

1. Если λ – собственное значение $K(x,s)$, то для существования решений неоднородного уравнения (4.20) необходимо и достаточно, чтобы неоднородность f была ортогональна всем собственным решениям сопряженного к нему уравнения (4.21) (в скалярном произведении (4.11)).

2. Альтернатива Фредгольма: Если λ не совпадает ни с одним из собственных значений ядра $K(x,s)$, то решения уравнений (4.20,4.21) существуют и единственны при любых непрерывных f, g.

3. Если λ – собственное значение $K(x,s)$, то однородные уравнения (4.20,4.21) имеют одинаковое число линейно-независимых решений.

Решение неоднородного уравнения, если оно существует, можно получить в виде ряда по λ методом последовательных приближений (см. ниже).

Уравнение Фредгольма I рода оказывается неустойчивым по отношению к малым изменениям f и K. Аналитическое решение привлекает особый аппарат

так называемой регуляризации, который мы затрагивать не будем. При численном решении неустойчивость может также возникать из-за погрешностей округления и пр., что надо иметь в виду.

4.3.7 Метод замены ядра на вырожденное
Degenerate kernel method

Благодаря тому, что всякий компактный оператор может быть представлен как предел (по норме) последовательности операторов, соответствующих вырожденным ядрам, точное решение уравнения можно получить как предел решений соответствующих уравнений с вырожденными ядрами.

Есть следующая теорема ◂см. [9], том2, стр 599▸:

$\mathbf{T}^0$: Пусть есть два интегральных уравнения

$$y(x) - \lambda \int_a^b ds\, K(x,s)y(s) = f(x) \quad \text{и} \quad z(x) - \lambda \int_a^b ds\, H(x,s)z(s) = g(x),$$

и существуют константы δ и ε такие, что

$$\forall x \in [a,b] \qquad \int_a^b ds|K(x,s)-H(x,s)|<\delta, \quad |f(x)-g(x)|<\varepsilon,$$

а также выполняются равенства

$$\int_a^b ds|R(x,s,\lambda)| < N, \quad \text{и} \quad \delta|\lambda|(1+|\lambda|N)<1,$$

где R – резольвента (см ниже):

$$y(x) = f(x) + \lambda \int_a^b dsR(x,s,\lambda)f(s).$$

Тогда

$$|y(x) - z(x)| < \frac{\delta|\lambda|(1+|\lambda|N)^2}{1-\delta|\lambda|(1+|\lambda|N)} \cdot \max_{[a,b]}|f(x)| + \varepsilon(1+|\lambda|N).$$

Таким образом, если построить последовательность ядер $H_n(x,s)$, *равномерно* сходящуюся к $K(x,s)$, то последовательность решений $z_n(x)$ уравнений с ядрами $H_n(x,s)$ будет равномерно сходиться к решению $y(x)$ уравнения с $K(x,s)$.

Способы построения вырожденных ядер могут быть самыми разнообразными: ядро $K(x,s)$ можно приближать частичными суммами степенных или

двойных тригонометрических рядов, частичными суммами двойных рядов Фурье по некоторым ортогональным системам функций, можно приближать алгебраическими или тригонометрическими интерполяционными многочленами, рациональными функциями ...

Пример

Решаем то же самое уравнение (4.8):

$$y(x) + x\int\limits_0^1 ds\,(e^{xs}-1)y(s) = e^x - x.$$

Ядро $K(x,s) = x(e^{xs}-1)$ аппроксимируем суммой первых трех членов его разложения в ряд Тейлора

$$K(x,s) \approx H(x,s) = x^2 s + \frac{x^3 s^2}{2} + \frac{x^4 s^3}{6}$$

и рассмотрим интегральное уравнение $z(x) + \int\limits_0^1 ds\, H(x,s)z(s) = e^x - x$. Решение его ищем в виде неоднородность плюс ряд:

$$z(x) = e^x - x + C_1 x^2 + C_2 x^3 + C_3 x^4.$$

Подставляя в уравнение, имеем

$$C_1x^2 + C_2x^3 + C_3x^4 + \int\limits_0^1 ds\left(x^2 s + \frac{x^3 s^2}{2} + \frac{x^4 s^3}{6}\right)(e^s - s + C_1 s^2 + C_2 s^3 + C_3 s^4) = 0.$$

Приравнивая коэффициенты при x^2, x^3, x^4 нулю, получаем систему уравнений

$$\begin{gathered}
C_1 + \int\limits_0^1 ds\, s\,(e^s - s + C_1 s^2 + C_2 s^3 + C_3 s^4) = 0,\\
C_2 + \tfrac{1}{2}\int\limits_0^1 ds\, s^2\,(e^s - s + C_1 s^2 + C_2 s^3 + C_3 s^4) = 0,\\
C_3 + \tfrac{1}{6}\int\limits_0^1 ds\, s^3\,(e^s - s + C_1 s^2 + C_2 s^3 + C_3 s^4) = 0,
\end{gathered}$$

и вычисляя интегралы,

$$\left\{\begin{array}{l}
\frac{5}{4}C_1 + \frac{1}{5}C_2 + \frac{1}{6}C_3 = -\frac{2}{3}\\
\frac{1}{5}C_1 + \frac{13}{6}C_2 + \frac{1}{7}C_3 = \frac{9}{4} - e\\
\frac{1}{6}C_1 + \frac{1}{7}C_2 + \frac{49}{8}C_3 = 2e - \frac{29}{5}
\end{array}\right. \Rightarrow \ldots \Rightarrow \left\{\begin{array}{l}
C_1 = -0.5010\\
C_2 = -0.1671\\
C_3 = -0.0422.
\end{array}\right.$$

Таким образом, приближенное решение имеет вид

$$z(x) = e^x - x - 0.5010\,x^2 - 0.1671\,x^3 - 0.0422\,x^4.$$

Если вычислить значения $z(x)$ в нескольких точках, получим $z(0) = 1.0000$, $z(1/2) = 1.0000$, $z(1) = 1.0080$. Вспоминая, что точное решение есть $y(x) \equiv 1$, видим, что совпадение весьма приличное.

4.4 Метод последовательных приближений

Решение уравнения Фредгольма II рода

$$(I - \lambda A)y = f, \tag{4.22}$$

где I – единичный оператор, можно искать в виде степенного ряда по λ

$$y_\lambda = \varphi_0 + \lambda\varphi_1 + \lambda^2\varphi_2 + \ldots + \lambda^n\varphi_n + \ldots$$

Подставив этот ряд вместо y в (4.22) и приравняв коэффициенты при равных степенях λ слева и справа, получим

$$\varphi_0 = f, \quad \varphi_1 = Af, \quad \ldots, \quad \varphi_n = A\varphi_{n-1} = A^n f, \ \ldots \tag{4.23}$$

В развернутой, интегральной, записи то же самое:

$$\begin{aligned}
\varphi_0(x) &= f(x),\\
\varphi_1(x) &= \int_a^b ds\, K(x,s)\varphi_0(s),\\
\varphi_2(x) &= \int_a^b ds\, K(x,s)\varphi_1(s),\\
&\ldots\\
\varphi_n &= \int_a^b ds\, K(x,s)\varphi_{n-1}(s),\\
&\ldots
\end{aligned}$$

Можно выразить $y_\lambda(x)$ через неоднородность f в виде

$$y_\lambda(x) = f(x) + \lambda \int_a^b ds\, R(x,s,\lambda)f(s),$$

где функция $R(x,s,\lambda)$ называется *резольвентой* ядра K или уравнения.

С другой стороны, этот же результат можно получить, если формально записать решение уравнения (4.22) как

$$y = (I - \lambda A)^{-1} f.$$

Если $\|\lambda A\| < 1$, т.е. $|\lambda| < \frac{1}{\|A\|}$, эта формула действительно определяет решение, поскольку в этом случае оператор $(I - \lambda A)^{-1}$ существует, определен на всем H и ограничен ◂…▸[7]. При этом его можно представить в виде степенного ряда

$$(I - \lambda A)^{-1} = I + \lambda A + \lambda^2 A^2 + \ldots + \lambda^n A^n + \ldots,$$

[7] Это доказывается просто если A эрмитов и имеет собственный базис ортонормированных векторов.

сходимость которого (по норме) обеспечивается условием $|\lambda| < \frac{1}{\|A\|}$. Значит, решение нашего уравнения (4.22) можно записать в виде

$$y = f + \lambda A f + \lambda^2 A^2 f + \ldots + \lambda^n A^n f + \ldots,$$

что совпадает с (4.23).

На самом деле, однако, оказывается, что этот ряд *всегда* сходится к решению уравнения, если только это решение существует. То есть он сходится для всех λ, не являющимися собственными значениями ядра, даже если $|\lambda| > \|A\|^{-1}$.

Если ядро ограничено но модулю $|K|<k$, то оно и интегрируемо

$$M \equiv \int ds\, dx |K|^2 \le (b-a)^2 k^2 \equiv M_m, \quad \Rightarrow \quad \|A\|^2 \le M \le M_m.$$

Тогда ряд сходится (по норме) при $|\lambda| < M_m^{-1/2}$. Можно также показать, что в этом случае он сходится равномерно ◂[9]▸:

$$|y(x) - \varphi_n(x)| \le \frac{\left(|\lambda|\sqrt{M_m}\right)^{n+1}}{1 - |\lambda|\sqrt{M_m}} \max_{[a,b]} |f(x)|.$$

Пример: уравнение Вольтерра II рода

Решаем уравнение $y(x) - \int\limits_0^x ds\, e^{-x-s} y(s) = \frac{1}{2}(e^{-x} + e^{-3x})$.

Ищем решение в виде ряда по λ (в этом примере λ=1), методом последовательных приближений: $y = \varphi_0 + \varphi_1 + \varphi_2 + \ldots$

$$\varphi_0(x) = f(x) = \frac{1}{2}(e^{-x} + e^{-3x});$$

$$\varphi_1(x) = \int\limits_0^x ds\, e^{-x-s}\varphi_0(s) = \frac{1}{2}\int\limits_0^x ds\, e^{-x-s}(e^{-x}+e^{-3x}) = \frac{3e^{-x}-2e^{-3x}-e^{-5x}}{8};$$

$$\varphi_2(x) = \int\limits_0^x ds\, e^{-x-s}\varphi_1(s) = \frac{5e^{-x}-9e^{-3x}+3e^{-5x}+e^{-7x}}{48};$$

$$\varphi_3(x) = \int\limits_0^x ds\, e^{-x-s}\varphi_2(s) = \frac{7e^{-x}-20e^{-3x}+18e^{-5x}-4e^{-7x}-e^{-9x}}{384};$$

$$\varphi_4(x) = \int\limits_0^x ds\, e^{-x-s}\varphi_3(s) = \frac{9e^{-x}-35e^{-3x}+50e^{-5x}-30e^{-7x}+5e^{-9x}+e^{-11x}}{3840}.$$

Складывая все, получаем

$$y_4 = \frac{1}{3840}\left(3839e^{-x} + 5e^{-3x} - 10e^{-5x} + 10e^{-7x} - 5e^{-9x} + e^{-11x}\right).$$

Точное решение уравнения есть $y_\infty(x) = e^{-x}$.

Для сравнения: $y_\infty(0) = 1.00000, \quad y_4(0) = 1.00000,$
$y_\infty(1) = 0.36788, \quad y_4(1) = 0.36783.$

4.5 Метод моментов

Ищем решение уравнения Фредгольма II рода в виде суммы $f(x)$ и линейной комбинации (конечного числа) заранее выбранных линейно-независимых функций $\{\varphi_i(x)\}_{i=1}^n$.

$$y(x) = f(x) + \sum_{i=1}^{n} C_i \varphi_i(x). \tag{4.24}$$

Очень удобно, если система ортогональна, хотя это и не обязательно.

Подставляя (4.24) в уравнение $y - \lambda A y = f$, получаем

$$f + \sum C_i \varphi_i - \lambda A f - \lambda \sum C_i A \varphi_i = f \quad \Rightarrow$$
$$\Phi \equiv \sum C_i \left(\varphi_i - \lambda A \varphi_i\right) - \lambda A f = 0.$$

Домножая скалярно на φ_j, $j = 1, \ldots, n$, получаем систему уравнений для коэффициентов C_i:

$$\sum_{i=1}^{n} C_i \big[(\varphi_i, \varphi_j) - \lambda (A\varphi_i, \varphi_j)\big] - \lambda (Af, \varphi_j) = 0, \quad j = 1, \ldots, n. \tag{4.25}$$

При этом мы заменили векторное уравнение в бесконечномерном пространстве $\Phi = 0$ требованием, чтобы лишь *проекция* Φ на конечномерное подпространство, порожденное функциями $\{\varphi_i\}$, была равна нулю.

Решение системы дает коэффициенты C_i и искомое приближение для $y(x)$. Если система ортогональна, то $(\varphi_i, \varphi_j) = \|\varphi_i\|^2 \delta_{ij}$ и

$$\|\varphi_j\|^2 C_j - \lambda \sum_{i=1}^{n} C_i (A\varphi_i, \varphi_j) = \lambda (Af, \varphi_j).$$

Так как по определению $A\varphi(x) = \int ds\, K(x,s)\varphi(s)$, то скалярные произведения здесь в развернутом виде есть

$$(Af, \varphi_i) = \int_a^b \int_a^b dx\, ds\, K(x,s) f(s) \varphi_i(x).$$

На геометрическом языке, метод заключается в том, что мы выбираем некоторое конечномерное подпространство L_n, смещенное от начала координат на f, и ищем в нем решение конечномерного уравнения, являющегося проекцией исходного (бесконечномерного) уравнения на L_n. При стремлении его размерности n к бесконечности, система $\{\varphi_i\}$ образует базис гильбертова пространства H, а система алгебраических уравнений (4.25) становится эквивалентной исходному векторному.

Метод пригоден для решения как неоднородных, так и однородных уравнений.

Пример

Рассмотрим такую задачу[☆]. Необходимо найти два первых собственных значения и соответствующие им собственные функции однородного интегрального уравнения

$$u(x) - \lambda \int_0^1 ds\, K(x,s)u(s) = 0 \quad \text{с ядром } K(x,s) = \begin{cases} x(1-s) & 0 \le x \le s \le 1, \\ s(1-x) & 0 \le s \le x \le 1. \end{cases}$$

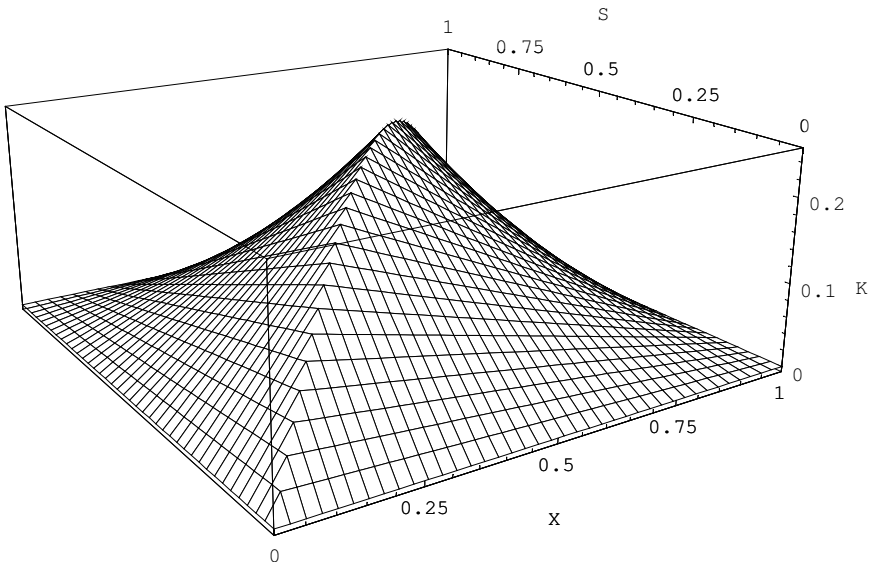


Рис. 4.1: Ядро $K(x,s)$

Решая методом моментов, ищем $u(x)$ в виде $u = A\varphi_1 + B\varphi_2 + C\varphi_3$, и выбераем линейно-независимые функции как

$$\varphi_1(x) = 1, \quad \varphi_2(x) = x(1-x), \quad \varphi_3(x) = x(1-x)(1-2x).$$

Система уравнений для коэффициентов (4.25), после вычисления интегралов, принимает вид

$$\begin{cases} A\left(1 - \frac{\lambda}{12}\right) + \frac{B}{6}\left(1 - \frac{\lambda}{10}\right) &= 0, \\ \frac{A}{6}\left(1 - \frac{\lambda}{10}\right) + \frac{B}{30}\left(1 - \frac{17\lambda}{168}\right) &= 0, \\ \frac{C}{210}\left(1 - \frac{\lambda}{40}\right) &= 0. \end{cases}$$

Характеристическое уравнение получаем, приравняв определитель системы нулю:

$$\left(\lambda^2 - 180\lambda + 1680\right)(\lambda - 40) = 0.$$

Таким образом,

$$\tilde{\lambda}_1 = 9.8751, \quad \tilde{\lambda}_2 = 40, \quad \tilde{\lambda}_3 = 170.1249.$$

Для $\tilde{\lambda}_1$: $A = -0.01176B$, $C = 0$. Число B находим из условия нормировки $\int\limits_0^1 dx\, y^2(x) = 1$, и получаем $\tilde{y}_1(x) = -0.0684 + 5.817\, x(1-x)$.

Для $\tilde{\lambda}_2$: $A = B = 0$, а C произвольное, находим его также из условия нормировки и получаем $\tilde{y}_2(x) = 14.49\, x(1-x)(1-2x)$

Точное решение:

$$\begin{aligned} &\lambda_1 = \pi^2 \approx 9.8696; \qquad y_1 = \sqrt{2}\sin \pi x, \\ &\lambda_2 = 4\pi^2 \approx 39.4784; \quad y_2 = \sqrt{2}\sin 2\pi x. \end{aligned}$$

Сравнивая значения приближенных решений $\tilde{y}_{1,2}$ и точных $y_{1,2}$ в разных точках промежутка $[0,1]$, можно видеть, что приближение первой собственной функции близко к точному решению, в то время как приближение второй значительно хуже.

4.6 Метод наименьших квадратов

Ищем решение уравнения Фредгольма в виде $y(x) = \sum C_i \varphi_i(x)$, где $\{\varphi_i\}$ – заранее выбранная система линейно-независимых функций. Подставляя в уравнение (II рода) $y - \lambda A y = f$, получаем

$$\Phi(C_1, \ldots, C_n; x) \equiv \sum_{i=1}^{n} C_i(\varphi_i - \lambda A \varphi_i) - f = 0.$$

В методе моментов мы аналогичное уравнение $\Phi = 0$ заменяли на приближенное условие, что проекция Φ на выбранное подпространство равно нулю. Теперь же мы будем искать $\{C_i\}$ из условия, чтобы величина Φ была *наименьшей по норме*, т.е. чтобы C_i минимизировали квадратичную форму $\|\Phi\|^2$.

Запишем это условие в виде

$$\begin{aligned} &\tfrac{\partial}{\partial C_i} \sum_{j=1}^{n} (C_j\psi_j - f, C_j\psi_j - f) = 0, \quad i = 1, \ldots, n; \\ &\quad \text{где } \psi_j = \varphi_j - \lambda A \varphi_j. \end{aligned} \tag{4.26}$$

Тогда в вещественной записи получаем алгебраическую систему

$$\sum_{j=1}^{n} C_j(\psi_i, \psi_j) = (f, \psi_i). \tag{4.27}$$

В развернутом виде скалярное произведение под суммой есть

$$\begin{aligned} &(\psi_i, \psi_j) = \int dx\, \psi_i(x)\psi_j(x) = \int dx \big(\varphi_i(x) - \lambda A\varphi_i(x)\big)\big(\varphi_j(x) - \lambda A\varphi_j(x)\big) = \\ &= \int\limits_a^b dx \Big(\varphi_i(x) - \lambda \int\limits_a^b ds\, K(x,s)\varphi_i(s)\Big)\Big(\varphi_j(x) - \lambda \int\limits_a^b ds\, K(x,s)\varphi_j(s)\Big). \end{aligned}$$

Этот метод пригоден и для нахождения собственных функций и собственных значений ядра. Для этого просто полагаем $f = 0$, приравниваем определитель системы нулю, и так далее, как обычно.

Пример

Решаем неоднородное уравнение Фредгольма I рода с таким же ядром как в прошлом примере☆:

$$\int_0^1 ds\, K(x,s)u(s) = x^4 - 2x^3 + x, \quad \text{где } K(x,s) = \begin{cases} x(1-s) & 0 \le x \le s \le 1, \\ s(1-x) & 0 \le s \le x \le 1. \end{cases}$$

К такому уравнению приводит решение задачи об отыскании статической нагрузки, под действием которой струна длины 1, закрепленная на концах $x = 0$ и $x = 1$, примет форму, описываемую неоднородностью $f = x^4 - 2x^3 + x$.

1. Ищем первое приближение в виде $u_1(x) = C_1 + C_2 x$, то есть берем $\varphi_1 = 1$, $\varphi_2 = x$. Используя метод наименьших квадратов, мы минимизируем квадрат нормы величины

$$\Phi_1 = f - Au_1 = f - \int ds\, K(x,s)u_1(x).$$

Считая интегралы

$$-\psi_1 \equiv A\varphi_1 = \int_0^1 ds\, K(x,s)\cdot 1 = \int_0^x ds\, x(1-s) + \int_x^1 ds\, s(1-x) = \ldots = \frac{1}{2}(x - x^2),$$

$$-\psi_2 \equiv A\varphi_2 = \int_0^1 ds\, K(x,s)\cdot s = \int_0^x ds\, xs(1-s) + \int_x^1 ds\, s^2(1-x) = \ldots = \frac{1}{6}(x - x^3),$$

получаем $\Phi_1 = f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x)$.

Систему уравнений (4.26) можно записать как $(\Phi_1, \frac{\partial}{\partial C_i}\Phi_1) = 0$ для $i = 1, 2$, что в явном виде в нашем случае дает

$$\begin{cases} \int\limits_0^1 dx \left[f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x) \right](x^2 - x) = 0, \\ \int\limits_0^1 dx \left[f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x) \right](x^3 - x) = 0. \end{cases}$$

Подставляя $f = x^4 - 2x^3 + x$ и считая интегралы от степеней, получим

$$\begin{cases} \frac{C_1}{3\cdot 4\cdot 5} + \frac{C_2}{4\cdot 5\cdot 6} - \frac{17}{3\cdot 4\cdot 5\cdot 7} = 0 \\ \frac{C_1}{5\cdot 8} + \frac{4C_2}{5\cdot 7\cdot 9} + \frac{17}{5\cdot 7\cdot 8} = 0 \end{cases} \Rightarrow \begin{cases} 14C_1 + 7C_2 = 34 \\ 63C_1 + 32C_2 = 153 \end{cases} \Rightarrow \begin{cases} C_1 = 17/7 \\ C_2 = 0. \end{cases}$$

Таким образом, в первом приближении $u_1(x) = 17/7 \approx 2.4286$

2 Ищем второе приближение в виде $u_2(x) = C_1 + C_2 x + C_3 x^2$, то есть $\varphi_1 = 1$, $\varphi_2 = x$, $\varphi_3 = x^2$. Так же минимизируем квадрат нормы величины

$$\Phi_2 = f - Au_2 = f - \int ds\, K(x,s) u_2(x).$$

Так как

$$-\psi_3 \equiv A\varphi_3 = \int\limits_0^1 ds\, K(x,s)\cdot s^2 = \ldots = \frac{1}{12}(x - x^4),$$

то $\Phi_2 = f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x) + \frac{C_3}{12}(x^4 - x)$ и система уравнений для C_i:

$$\begin{cases} \int\limits_0^1 dx \left[f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x) + \frac{C_3}{12}(x^4 - x) \right](x^2 - x) = 0, \\ \int\limits_0^1 dx \left[f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x) + \frac{C_3}{12}(x^4 - x) \right](x^3 - x) = 0, \\ \int\limits_0^1 dx \left[f + \frac{C_1}{2}(x^2 - x) + \frac{C_2}{6}(x^3 - x) + \frac{C_3}{12}(x^4 - x) \right](x^4 - x) = 0. \end{cases}$$

Считая интегралы, получаем линейную систему

$$\begin{cases} \frac{C_1}{3\cdot4\cdot5} + \frac{C_2}{4\cdot5\cdot6} + \frac{5C_3}{4\cdot6\cdot6\cdot7} - \frac{17}{3\cdot4\cdot5\cdot7} = 0 \\ \frac{C_1}{5\cdot8} + \frac{4C_2}{5\cdot7\cdot9} + \frac{11C_3}{3\cdot5\cdot8\cdot12} - \frac{17}{5\cdot7\cdot8} = 0 \\ \frac{5C_1}{4\cdot6\cdot7} + \frac{11C_2}{3\cdot5\cdot6\cdot8} + \frac{9C_3}{9\cdot12} - \frac{13}{4\cdot5\cdot9} = 0 \end{cases} \Rightarrow \begin{cases} 84C_1 + 42C_2 + 25C_3 = 204 \\ 252C_1 + 128C_2 + 77C_3 = 612 \\ 450C_1 + 231C_2 + 140C_3 = 1092 \end{cases} \Rightarrow \begin{cases} C_1 = 0 \\ C_2 = 12 \\ C_3 = -12. \end{cases}$$

Таким образом, решая ее, получили решение во втором приближении $u_2(x) = 12x(1-x)$. Подстановкой несложно проверить, что $u_2(x)$ является точным решением исходного уравнения.

4.7 Уравнения Вольтерра

Если f и K дифференцируемы и $K(x,x) \neq 0$, то уравнение Вольтерра I рода сводится к уравнению II рода: дифференцируя (4.3) по x и деля уравнение на $K(x,x)$, получаем

$$y(x) + \int\limits_a^x ds \frac{K'_x(x,s)}{K(x,x)} y(s) = \frac{f'(x)}{K(x,x)}.$$

Если $K(x,x) \equiv 0$, то продолжаем дифференцировать исходное уравнение по x, пока m-я производная $K^{(m)}_{x^m}(x,t)\Big|_{x=t}$ не станет отличной от нуля. Тогда уравнение так же сводится к уравнению II рода.

Однородное уравнение Вольтерра II рода (4.4) с непрерывным или полярным ядром $K(x,s)$ *не имеет нетривиальных решений.*

Неоднородное уравнение Вольтерра II рода с непрерывным или полярным ядром $K(x,s)$ и непрерывной $f(x)$ *имеет единственное непрерывное решение* для любого λ.

Это решения можно найти например методом последовательных приближений, причем ряд по λ *всегда сходится* равномерно.

Так, если $k = \max\limits_{R}|K(x,s)|$, где $R = \{a \le s \le x \le b\}$, то норма оператора A ограничена величиной $\sqrt{M} = k(b-a)$ и ◂…▸

$$|y(x) - y_n(x)| \le \sum_{k=n+1}^{\infty} \frac{|\lambda|^k}{M^{k/2}k!} \cdot \max_{[a,b]}|f(x)|.$$

Также можно решать уравнение Вольтерра заменой интеграла конечной суммой. Как было указано выше, получающаяся алгебраическая система имеет треугольный вид и потому относительно быстро решается.

Пример

Один пример решения уравнения Вольтерра был дан в иллюстрации метода последовательных приближений. Решим здесь то же самое уравнение

$$y(x) - \int_0^x ds\, e^{-x-s} y(s) = \frac{1}{2}(e^{-x} + e^{-3x})$$

методом квадратур, используя формулу трапеций с шагом $h = 0.2$, т.е. берем $x_i = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$.

Вычислительная табличка:

x_i	K_{0i}	K_{1i}	K_{2i}	K_{3i}	K_{4i}	K_{5i}	f_i
0	1.00000	0.81873	0.67032	0.54881	0.44993	0.36788	1.00000
1	0.81873	0.67032	0.54881	0.44993	0.36788	0.30119	0.68377
2	0.67032	0.54881	0.44993	0.36788	0.30119	0.24660	0.48576
3	0.54881	0.44993	0.36788	0.30119	0.24660	0.20190	0.35706
4	0.44993	0.36788	0.30119	0.24660	0.20190	0.16530	0.27002
5	0.36788	0.30119	0.24660	0.20190	0.16530	0.13534	0.20883

Применим формулу трапеций, но не в лоб. Вместо того, чтобы подставлять табличку в общую формулу (4.7), при счете интеграла для $x = x_i$ верхний предел интегрирования заменяем на x_i, после чего используем готовое разбиение на i кусочков для формулы трапеций. Так, для $i = 2$, $x_i = 0.4$ получаем

$$f_2 = y_2 + h\Big(\frac{1}{2}K_{00}y_0 + K_{11}y_1 + \frac{1}{2}K_{12}y_2\Big),$$

Таким образом, в системе

$$y_i(1 - A_{ii}K_{ii}) = f_i + \sum_{j=1}^{i-1} A_{ij}K_{ij}y_j$$

коэффициенты A, в отличие от (4.7), имеют два индекса: $A_{0j} = A_{jj} = h/2$, а остальные равны h. Вычисления дают:☆

$$\begin{aligned}
y_0 &= f_0 = 1.00000,\\
y_1 &= \frac{1}{1-\frac{h}{2}K_{11}}\left(f_1 + \frac{h}{2}K_{10}y_0\right) = 0.8206,\\
y_2 &= \frac{1}{1-\frac{h}{2}K_{22}}\left(f_2 + \frac{h}{2}K_{20}y_0 + hK_{21}y_1\right) = 0.6731,\\
y_3 &= \frac{1}{1-\frac{h}{2}K_{33}}\left(f_3 + \frac{h}{2}K_{30}y_0 + h(K_{31}y_1 + K_{32}y_2)\right) = 0.5518,\\
y_4 &= \frac{1}{1-\frac{h}{2}K_{44}}\left(f_4 + \frac{h}{2}K_{40}y_0 + h(K_{41}y_1 + K_{42}y_2 + K_{43}y_3)\right) = 0.4522,\\
y_5 &= \frac{1}{1-\frac{h}{2}K_{55}}\left(f_5 + \frac{h}{2}K_{50}y_0 + h(K_{51}y_1 + K_{52}y_2 + K_{53}y_3 + K_{54}y_4)\right) = 0.3705.
\end{aligned}$$

Для сравнения точного решения $y_\infty = e^{-x}$ с полученным приближенным приведем табличку

x_k	0	0.2	0.4	0.6	0.8	1
y_∞	1.0000	0.8127	0.6703	0.5488	0.4493	0.3679
y_5	1.0000	0.8287	0.6731	0.5418	0.4522	0.3705
δ	0.0000	0.0019	0.0028	0.0030	0.0029	0.0026

4.8 Вариационные методы

Эти методы работают хорошо как в одномерном, так и в трехмерном случаях без существенных модификаций. Поэтому, подразумевая что они годятся для решений уравнений мат. физики, будем для простоты их излагать в основном на примере обыкновенных дифференциальных уравнений.

4.8.1 Вариационные задачи

Исторически так сложилось, что вариационные задачи, возникающие в физике естественным образом из фундаментальных принципов – Ферма, наименьшего действия, и как-либо еще, – прямо решать не умели. Универсальным способом

стало сведение вариационной задачи к соответствующему дифференциальному уравнению Эйлера-Лагранжа, которое уже исследовалось и решалось.

Со временем, однако, стали развиваться как аналитические, так и численные *прямые* методы решения вариационных задач. Поэтому иногда выгодно дифференциальные уравнения переформулировать как вариационные задачи – что в некоторых случаях является просто возвратом к исходной формулировке. Ниже мы рассмотрим несколько примеров прямых численных методов, но вначале вспомним принцип наименьшего действия.

Принцип наименьшего действия. Как мы знаем, истинное движение механической системы, которая описывается обобщенными координатами q_α, реализует экстремум функционала действия

$$S \equiv \int dt\, L\big(q_\alpha(t), \dot{q}_\alpha(t)\big) = \min,$$

где $L(q, \dot{q})$ – функция Лагранжа механической системы.

При описании *поля* $\varphi_q(\mathbf{r})$ (природа индекса $q = 1, \ldots, n$ не имеет значения) функция Лагранжа становится *функционалом* от функции поля

$$L[\varphi] = \int d^3r\, \mathcal{L}(\varphi_q, \dot{\varphi}_q),$$

и принцип наименьшего действия формулируется практически так же:

$$S \equiv \int dt\, L[\varphi] = \int d\Omega\, \mathcal{L}(\varphi_q, \partial_i\varphi)_q = \min,$$

где $d\Omega = dt\, dx\, dy\, dz$ – элемент четырехмерного объема. Обозначим для краткости $(t, x, y, z) = (x_0, x_1, x_2, x_3)$, будем подразумевать суммирование по повторяющимся индексам[8] $q = 1, \ldots, n$ и $i = 0, 1, 2, 3$; также обозначим $\partial_i \equiv \partial/\partial x_i$. Варьирование приводит к уравнениям Эйлера-Лагранжа:

$$\begin{aligned}
0 = \delta S &= \int d\Omega\, \delta\mathcal{L} = \int d\Omega \left\{ \frac{\partial\mathcal{L}}{\partial\varphi_q}\delta\varphi_q + \frac{\partial\mathcal{L}}{\partial\,\partial_i\varphi_q}\delta(\partial_i\varphi_q) \right\} = \\
&= \left\{ \delta(\partial_i\varphi_q) = \partial_i\left(\delta\varphi_q\right) \;\Rightarrow\; \text{по частям} \quad \int d\Omega A\, \delta(\partial_i\varphi_q) = -\int d\Omega\, \delta\varphi_q\, \partial_i A \right\} = \\
&= \int d\Omega\, \delta\varphi_q \left\{ \frac{\partial\mathcal{L}}{\partial\varphi_q} - \partial_i \frac{\partial\mathcal{L}}{\partial\,\partial_i\varphi_q} \right\} \quad \Rightarrow \qquad \frac{\partial\mathcal{L}}{\partial\varphi_q} - \partial_i \frac{\partial\mathcal{L}}{\partial\,\partial_i\varphi_q} = 0 \quad \forall q.
\end{aligned}$$

[8] Формально это соглашение Эйнштейна специальной теории относительности, но в нашем случае оно не имеет никакого отношения к преобразованию Лоренца, и лишь служит сокращению обозначений.

Внеинтегральные члены при интегрировании по частям здесь обращаются в ноль, потому что для $i = 0$ на границах временного промежутка $t_{1,2}$ поле не варьируется, а для $i = \alpha = 1,2,3$ на бесконечности или на пространственных границах оно обращается в ноль или иным образом фиксировано:

$$\int d\Omega A\delta(\partial_0\varphi_q) = \int d^3r\left\{A\delta\varphi_q\big|_{t_1}^{t_2} - \int dt\,\delta\varphi_q\partial_0 A\right\} = -\int d\Omega\delta\varphi_q\partial_0 A;$$
$$\int d\Omega A\delta(\partial_\alpha\varphi_q) = \int dt\,\{\oint dS_\alpha A\delta\varphi_q - \int d^3r\,\delta\varphi_q\partial_\alpha A\} = -\int d\Omega\delta\varphi_q\partial_\alpha A.$$

Задача Штурма-Лиувилля. Решение одномерной задачи Штурм-Лиувилля с граничными условиями I рода

$$-\frac{d}{dx}\Big[k(x)\frac{dy(x)}{dx}\Big] + p(x)y(x) = f(x); \qquad \begin{cases} y(a) = A, \\ y(b) = B. \end{cases}$$

является экстремумом функционала

$$I[y] = \int\limits_a^b dx\left\{k(x)y'^2(x) + p(x)y^2(x) - 2f(x)y(x)\right\} \tag{4.28}$$

на классе функций, удовлетворяющих тем же граничным условиям и квадратично интегрируемых с производной[9]

$$y(a) = A,\ y(b) = B, \quad y(x) \in W_2^1[a,b]:\ \|y\|^2_{W_2^1[a,b]} \equiv \int\limits_a^b dx\,\left[y^2 + (y')^2\right] < \infty.$$

◂ Покажем, варьируя I:

$$\delta I = \int dx\Big\{2ky'\delta y' + 2py\delta y - 2f\delta y\Big\} = \left\{\begin{array}{l} \int dxY\delta y' = \int Y\,d\delta y = \\ = Y\delta y\big|_a^b - \int dY\,\delta y = -\int dx\,Y'\delta y \end{array}\right\} =$$
$$= 2\int dx\Big\{\big[k(x)y'(x)\big]' + p(x)y(x) - f(x)\Big\}\delta y. \qquad \blacktriangleright$$

Обобщение на трехмерный случай прямолинейно:

$$\begin{cases} -\nabla\big[k(\mathbf{r})\nabla u\big] + p(\mathbf{r})u = f, \\ \mathbf{r} \in G; \quad u\big|_{\partial G} = u_0. \end{cases} \Leftrightarrow \begin{cases} I[y] \equiv \int\limits_G d^3r\Big\{k\,(\nabla u)^2 + pu^2 - 2fu\Big\} = \min, \\ u\big|_{\partial G} = u_0. \end{cases} \tag{4.29}$$

[9] Очевидно, это требование нужно, чтобы функционал существовал. То, что решение вариационной задачи оказывается еще и дважды дифференцируемым, при достаточно "хороших" k и p, – утверждение нетривиальное, которое мы не доказываем. Ограничения на p и k мы здесь обсуждать не будем, для простоты можно их считать достаточно гладкими. Заметим, что пространство Соболева W_2^1 является гильбертовым.

В двумерной задаче частный случай $k = 1$, $p = 0$, $f = 0$ дает уравнение Лапласа $\Delta u = 0$, которое в описывает изгиб гибкой мембраны. Соответствующий функционал

$$I[u(x,y)] = \int dx\,dy\,\left[(\partial_x u)^2 + (\partial_y u)^2\right]$$

дает потенциальную энергию мембраны.

4.8.1.1 Метод Галеркина

Допустим, исходная задача для уравнения $Ly = f$ не является вариационной. Это может быть произвольное дифференциальное, интегральное, интегродифференциальное уравнение. В этом случае можно поступить следующим образом.

Во-первых, представим искомую функцию в виде $y = \tilde{y} + y_0$, где y_0 удовлетворяет граничным условиям. Тогда $\tilde{y}$ удовлетворяет соответствующему уравнению

$$\tilde{L}\tilde{y} = \tilde{f}$$

и однородным граничным условиям.

Возьмем теперь систему функций $\{\varphi_i\}$, полную (замкнутую) в том классе функций, в котором мы ищем $\tilde{y}$, и представим $\tilde{y}$ в виде ряда по этой системе, ограничившись n членами:

$$\tilde{y} = \sum_{i=1}^{n} C_i \varphi_i$$

Вместо равенства $\tilde{L}\tilde{y} - \tilde{f} = 0$ (это векторное равенство в бесконечномерном пространстве) потребуем, чтобы была равна нулю проекция вектора $(\tilde{L}\tilde{y} - \tilde{f})$ на линейную оболочку $\{\varphi_1, \ldots, \varphi_n\}$. Вводя удобным образом скалярное произведение (f, g), получаем систему уравнений

$$(\tilde{L}\tilde{y} - \tilde{f}, \varphi_i) = 0, \quad i = 1, 2, \ldots, n. \tag{4.30}$$

Если дифференциальный оператор L линеен, то подставляя разложение $\tilde{y}$ по $\{\varphi_i\}$, получаем систему линейных уравнений для коэффициентов C_i:

$$\sum_{j=1}^{n} C_j (\tilde{L}\varphi_j, \varphi_i) = (\tilde{f}, \varphi_i).$$

Вид скалярного произведения имеет смысл выбирать так, чтобы элементы матрицы системы легко считались численно, а еще лучше аналитически. Если L нелинеен, то и система для C_j будет нелинейной.

В применении к линейным интегральным уравнениям идея метода Галеркина приводит к рассмотренному выше методу моментов.

4.8.2 Метод Ритца на примере одномерной задачи Штурма-Лиувилля

Рассмотрим так называемый метод Ритца решения вариационной задачи на примере одномерной задачи Штурма-Лиувилля (4.28)

$$I[y] = \int_a^b dx \left\{ k(x) y'^2(x) + p(x) y^2(x) - 2 f(x) y \right\} = \min .$$

Граничные условия сразу будем считать однородными, потому что исходную задачу можно свести к однородной, переходя к новой искомой функции.

Пусть $\{\varphi_i\}_1^\infty$ – линейно-независимая система функций, удовлетворяющая граничным условиям. Например, на промежутке $[0,1]$ с однородными граничными условиями I рода такими могут быть системы

$$\begin{aligned} \varphi_k^{(1)} &= \sin\left(k\pi x\right); \\ \varphi_k^{(2)} &= x^k(1-x); \\ \varphi_k^{(3)} &= x(1-x)T_i(2x-1), \ \ldots \end{aligned}$$

Опять будем искать $y(x)$ в виде конечной суммы

$$y \approx y_n = \sum_{k=1}^{n} C_k \varphi_k .$$

Подставляя в функционал I, получим

$$I = \int dx \left\{ k \sum_{i,j=1}^{n} C_i C_j \varphi_i' \varphi_j' + p \sum_{i,j=1}^{n} C_i C_j \varphi_i \varphi_j - 2 f \sum_{i=1}^{n} C_i \varphi_i \right\}.$$

Ищем минимум по отношению к коэффициентам C_k:

$$\frac{\partial I}{\partial C_k} = 0 \quad \Rightarrow \int dx \left\{ k \sum_{i=1}^{n} C_i \varphi_i' \varphi_k' + p \sum_{i=1}^{n} C_i \varphi_i \varphi_k - f \varphi_k \right\} = 0, \quad k = 1, \ldots, n.$$

В итоге получили систему линейных уравнений для C_k

$$\sum C_i \alpha_{ik} = \beta_k, \text{ где } \quad \alpha_{ik} = \int dx \left\{ k \varphi_i' \varphi_k' + p \varphi_i \varphi_k \right\}; \quad \beta_k = \int dx \, f \varphi_k .$$

Решая ее, получаем C_i и y_n. Для сходимости $\{y_n\}$ к точному решению в норме W_2^1 достаточно полноты системы $\{\varphi_i\}$ в этом пространстве ◂…▸

$$\forall y \in W_2^1, \ \varepsilon > 0 \quad \exists \, n, \{C_i\}_1^n \quad | \quad \left\| y - \sum C_i \varphi_i \right\|_{W_2^1} < \varepsilon .$$

Систему желательно выбирать такую, чтобы элементы матрицы α_{ij} считались аналитически или достаточно быстро численно[10].

[10] Есть и следующий момент. Система $\{\varphi_k^{(2)}\}$, например, плоха тем, что у соответствующей матрицы α_{ij} большой разброс собственных значений – $\lambda_{max}^{(n)} / \lambda_{min}^{(n)}$ растет с n быстрее всякой степени, – что плохо сказывается на численном решении системы. В этом плане лучше оказывается $\{\varphi_k^{(3)}\}$, для которой $\lambda_{max}^{(n)} / \lambda_{min}^{(n)} = O(n^\alpha)$.

4.8.3 Вариационно-разностный вариант метода Ритца

Основной недостаток приведенного метода Ритца заключается в том, что даже если система $\{\varphi_i\}$ является ортогональной в каком-либо простом скалярном произведении вида $(f,g) = \int d\mu fg$, то матрица α_{ij} все равно оказывается полностью заполненной, и при большом n соответствующая система решается за время $\sim n^2$. С этим можно справиться следующим образом.

Во-первых, заметим, что при увеличении числа линейно-независимых функций n вовсе не обязательно оставлять первые n функций теми же самыми

$$\{\varphi_1, \varphi_2 \ldots, \varphi_n\} \longrightarrow \{\varphi_1^{(n)}, \varphi_2^{(n)}, \ldots, \varphi_n^{(n)}\}.$$

Тогда вместо условия полноты имеем

$$\forall y \in W_2^1,\ \varepsilon > 0 \quad \exists\, n, C_i \quad | \quad \Big\| y - \sum^{n} C_i \varphi_i^{(n)} \Big\|_{W_2^1} < \varepsilon.$$

Рассмотрим задачу на $[0,1]$. Зададим на этом промежутке точки $0 = x_0 < x_1 < \ldots < x_n = 1$ и будем искать решение в виде функции, линейной на каждом из отрезков $[x_i, x_{i+1}]$, и принимающей на концах $[0,1]$ заданные значения[11]. Это эквивалентно тому, что в качестве линейно-независимых функций $\varphi_i^{(n)}$ на $[0,1]$ мы берем кусочно-линейные "зубцы"

$$\varphi_0^{(n)} = \begin{cases} \frac{x_1 - x}{x_1} & \text{при } x \in [0, x_1], \\ 0 & \text{при } x \geq x_1; \end{cases}$$

$$\varphi_i^{(n)} = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{при } x \in [x_{i-1}, x_i], \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & \text{при } x \in [x_i, x_{i+1}], \\ 0 & \text{в остальных точках}; \end{cases} \qquad \text{для} \quad i = 1, \ldots, n-1 \tag{4.31}$$

$$\varphi_n^{(n)} = \begin{cases} 0 & \text{при } x < x_{n-1}, \\ \frac{x - x_{n-1}}{x_n - x_{n-1}} & \text{при } x \in [x_{n-1}, x_n]. \end{cases}$$

Так как каждый зубец перекрывается только с двумя соседними, то матрица α_{ij} оказывается трехдиагональной, и система быстро решается методом прогонки[12]

[11] Такая полигональная функция использовалась как промежуточный этап для построения аппроксимационного многочлена в доказательстве теоремы Вейерштрасса.

[12] Понятно, что система таких зубцов при $n = 1, 2, \ldots$ образует множество, плотное в множестве непрерывных функций на $[0,1]$ в равномерной метрике, а следовательно и в квадратичной. Плотность в смысле метрики W_2^1 еще нужно доказать.

Вариационно-разностная модификация метода Галеркина

Аналогичным образом можно видоизменить метод Галеркина. Домножим исходное уравнение $Ly - f = 0$ на произвольную функцию ψ и проинтегрируем по промежутку, на котором задана y:

$$\int dx\big[Ly - f\big]\psi = 0.$$

Пусть $Ly = [ky']' + py$ (одномерная задача Штурма-Лиувилля). Проинтегрировав по частям первое слагаемое, получим

$$\Lambda(y,\psi) \equiv \int dx\Big[ky'\psi' + py\psi - f\psi\Big] = 0.$$

Будем искать решение в виде

$$y_n = \sum^{n} C_i\varphi_i^{(n)},$$

где $\varphi_i^{(n)}$ – "зубчатые" функции (4.31), – и потребуем, чтобы интеграл занулялся для всех ψ того же вида. Так как $\Lambda(y,\psi)$ линейно по ψ, то получим линейную систему уравнений для коэффициентов C_i

$$\Lambda(y_n,\varphi_i^{(n)}) = 0, \quad i = 1,\dots,n,$$

матрица которой также является трехдиагональной.

Литература по интегральным уравнениям

- Березин, Жидков, *Методы вычислений*, том 2, [9].
- Колмогоров А.Н. и Фомин С.В *Элементы теории функций и функционального анализа*, [11].
- Смирнов В.И. *Курс высшей математики*, тома IV и V, [13].
- Васильева А.Б. и др. *Дифференциальные и интегральные уравнения, вариационное исчисление в примерах и задачах*, [18]. Очень кратко даны понятия и теоремы, примеры решения задач.
- Васильева А.Б., Тихонов Н.А. *Интегральные уравнения*, [19]. Учебник, изложение подробнее и с доказательствами.
- Polyanin A.D., Manzhirov A. *Handbook of integral equations*, [20]. Огромный справочник по аналитическим и численным методам решений линейных и нелинейных интегральных уравнений, содержит более 2500 интегральных уравнений с решениями.

- Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. *Численные методы*, [2]; главы 9, 10 – вариационные методы.

Приложение А

Задача собственных значений

Задача о собственных векторах и собственных значениях формулируется следующим образом. У нас есть некоторая квадратная матрица (оператор) A, и необходимо найти все числа λ_i и вектора x_i, для которых верно

$$Ax_i = \lambda_i x_i. \tag{A.1}$$

Числа λ_i называются собственными значениями матрицы; вектора, которые им соответствуют – собственными векторами матрицы.

Решение задачи можно разделить на следующие этапы:

1. Нахождение характеристического уравнения;
2. Решение характеристического уравнения;
3. Нахождение собственных векторов по известным собственным значениям.

А.1 Метод Данилевского раскрытия характеристического уравнения

Задачу (A.1) можно переписать как однородное уравнение для вектора x

$$(A - \lambda I)x = 0,$$

которое имеет нетривиальные решения если

$$det(A - \lambda I) = 0.$$

Это характеристическое уравнение – полиномиальное уравнение степени n относительно λ, которое имеет n корней. Вычисление определителя в лоб по определению очень неэффективно и требует $O(n!)$ арифметических операций. Здесь

мы рассмотрим более эффективный метод Данилевского получения характеристического уравнения.

Будем говорить, что матрицы A и A' – B-подобны, если $A' = B^{-1}AB$, где B некоторая невырожденная матрица. Тогда если $Ax = \lambda x$, то

$$A'(B^{-1}x) = B^{-1}ABB^{-1}x = B^{-1}Ax = \lambda(B^{-1}x).$$

Таким образом, собственные значения B-подобных векторов одинаковы, а собственные вектора связаны соотношением $x = Bx'$.

Метод Данилевского состоит в приведении матрицы A к подобной ей матрице F, которая имеет нормальную форму Фробениуса:

$$F = \begin{pmatrix} p_1 & p_2 & p_3 & \dots & p_{n-1} & p_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \tag{A.2}$$

Такая форма хороша тем, что характеристическое уравнение для F, $det(F - \lambda I) = 0$, раскрывая определитель по верхней строке, несложно привести к виду

$$(-1)^n\left[\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \dots - p_{n-1}\lambda - p_n\right] = 0,$$

так что характеристическое уравнение оказывается

$$\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \dots - p_{n-1}\lambda - p_n = 0. \tag{A.3}$$

Способ приведения произвольной матрицы к нормальной форме Фробениуса с помощью преобразования подобия покажем на примере матрицы четвертого порядка.

Пусть есть матрица

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

Возьмем матрицу B_1 с обратной, равные

$$B_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Rightarrow \quad B_1 = -\frac{1}{a_{43}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a_{41} & a_{42} & 1 & a_{44} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(то что это обратные матрицы можно проверить перемножая). Тогда действуя на A преобразованием подобия с матрицей B, получим матрицу с нижней строкой в нужном виде:

$$A_1 = B_1^{-1}AB_1 = \ldots = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Продолжая процесс с

$$B_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ b_{31} & b_{32} & b_{33} & b_{34} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Rightarrow \quad B_2 = -\frac{1}{b_{32}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ b_{31} & 1 & b_{33} & b_{34} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

получим матрицу $A_2 = B_2^{-1}A_1B_2 = \left(c_{ij}\right)$ с двумя нижними строками в нужном виде. Делая последний очевидный шаг, получим матрицу

$$A_3 = B_3^{-1}B_2^{-1}B_1^{-1}AB_1B_2B_3 = B^{-1}AB, \quad \text{где} \quad B = B_1B_2B_3$$

в нормальной форме Фробениуса (A.2). Собственные значения тогда находятся из уравнения (A.3), а собственные вектора из системы $A_3x = \lambda x$:

$$x' = (\lambda^{n-1}, \lambda^{n-2}, \ldots, \lambda^2, \lambda, 1).$$

Тогда Bx' будут соответствующими собственными векторами A.

A.2 Границы собственных значений

Решать характеристическое уравнение проще, если есть какое-то представление о расположении его корней. Кроме того, часто возникает необходимость знать границы собственных значений и не требуется знать сам характеристический многочлен.

Пусть матрица A симметрична, так что существует n действительных собственных значений и ортонормированный базис собственных векторов (в комплексном случае это эрмитова матрица).

Пусть собственные значения упорядочены в порядке убывания

$$\lambda_n \le \lambda_{n-1} \le \ldots \le \lambda_1,$$

а e_i – соответствующие собственные вектора. Тогда для любого $x \in \mathbb{R}^n$

$$x = \sum x_i e_i, \quad \Rightarrow \quad (Ax, x) = \sum \lambda_i x_i^2,$$

так что

$$\lambda_n \sum x_i^2 = \lambda_n (x, x) \le (Ax, x) \le \lambda_1 (x, x) = \lambda_1 \sum x_i^2,$$

и

$$\lambda_n \le \frac{(Ax, x)}{(x, x)} \le \lambda_1.$$

Это неравенство называется неравенством Рэлея. Из него следует, что

$$\lambda_{min} = \inf \frac{(Ax, x)}{(x, x)}, \quad \lambda_{max} = \sup \frac{(Ax, x)}{(x, x)}.$$

A.3 Наибольшее собственное значение

Упорядочим теперь собственные числа в порядке убывания *модулей*

$$|\lambda_1| \ge \lambda_2 \ge \ldots \ge |\lambda_n|.$$

Тогда если подействовать на произвольный вектор $x = \sum x_i e_i$ степенью n матрицы A, то в пределе $n \to \infty$ получим

$$A^n x = \sum x_i \lambda_i^n e_i = \lambda_1^n \sum \left(\frac{\lambda_i}{\lambda_1} \right)^n x_i e_i = \lambda_1^n x_1 e_1 + O\Big((\lambda_i / \lambda_1)^n \Big).$$

Таким образом, при последовательном действии одной и той же матрицей на произвольный вектор x, результат приближается (по направлению) к ее собственному вектору, соответствующему максимальному по модулю собственному значению. Постоянство отношения

$$\lambda_1 = \frac{(A^{n+1} x)_i}{(A^n x)_i}$$

по i является критерием сходимости.

Приложение B

Мера и интеграл Лебега

Интеграл Лебега (Lebesgue integral) – это обобщение интеграла Римана на более широкий класс функций. Все функции, определённые на конечном отрезке числовой прямой и интегрируемые по Риману, являются также интегрируемыми по Лебегу, причём в этом случае оба интеграла равны. Однако, существует большой класс функций, определённых на отрезке и интегрируемых по Лебегу, но неинтегрируемых по Риману. Также интеграл Лебега может иметь смысл для функций, заданных на произвольных множествах. Но для его строгого определения необходимо небольшое введение в теорию меры.

B.1 Мера

Мера на числовой прямой

Мера (measure) интервала $I = (a, b)$, где $a \leq b$, на числовой прямой определяется как $m(I) = (b - a)$ (и равна нулю если интервал пустой $a = b$). Определение не изменится, если интервал заменить отрезком $[a, b]$.

Эта мера удовлетворяет трем свойствам:

1. $m(\varnothing) = 0$;
2. $m(I) \geq 0$;
3. Аддитивность: если $\{I_k\}_{k=1}^{\infty}$ – счетная система попарно непересекающихся интервалов $I_i \cap I_j = 0 \ \forall i, j$, то

$$m\left(\cup_{i=1}^{\infty} I_k\right) = \sum_{i=1}^{\infty} m(I_k). \tag{B.1}$$

Верхняя мера (внешняя мера, outer measure) множества A на числовой прямой – точная нижняя грань сумм длин интервалов, составляющих покрытие множества A (она существует, т.к. сумма длин ограничена снизу):

$$\mu^*(A) = \inf_{A \subset \cup I_k} \sum m(I_k). \tag{B.2}$$

Нижняя мера (внутренняя мера, inner measure) – точная верхняя грань верхней меры дополнения к A до $[a,b]$ по всем отрезкам $[a,b]$:

$$\mu_*(A) = \sup_{[a,b]} \{(b-a) - \mu^*([a,b] \setminus A)\}. \tag{B.3}$$

Множество A называется *измеримым* в смысле Лебега, если $\mu^*(A) = \mu_*(A)$, и общее значение $\mu(A) = \lambda(A) = |A|$ верхней и нижней мер называется *лебеговой мерой (Lebesgue measure)* множества A.

Пусть $F(x)$ – неубывающая, непрерывная слева функция на прямой. Тогда можно определить меру так:

$$\begin{aligned} m(a,b) &= F(b) - F(a+0); \\ m[a,b] &= F(b+0) - F(a); \\ m[a,b) &= F(b+0) - F(a+0); \\ m(a,b] &= F(b) - F(a); \end{aligned} \tag{B.4}$$

Таким образом перечисленные три свойства меры тоже выполняются, и так же как и только что, но основываясь на мере m, можно ввести меру Лебега. Меры, получаемые с помощью разных функций F, называются *мерами Лебега-Стилтьеса (Lebesgue-Stieltjes measures)*

Меры на произвольных множествах

Кольцо и полукольцо в алгебре

Для определения меры на числовой прямой, необходимо было то свойство промежутков и интервалов, что они составляют *полукольцо* относительно операций пересечения и объединения.

Кольцо (Ring): множество R, на котором заданы две бинарные операции "+" (аддитивная операция или сложение) и "×" (мультипликативная операция или умножение), со следующими свойствами:

1. R есть абелева группа относительно "+";

2. Ассоциативность "×": $\forall a,b,c \in R \quad a \times (b \times c) = (a \times b) \times c$;

3. Операция "+" дистрибутивна (distributive) относительно "×":
   $\forall a,b,c \in R \quad a \times (b+c) = a \times b + a \times c,\ (b+c) \times a = b \times a + c \times a.$

Если добавить свойство коммутативности относительно "×" , получим *коммутативное кольцо.*

Добавив еще существование единицы $1 \neq e$ $(\exists 1 \in R | \forall a \in R\ 1 \times a = a)$ и обратимость всех ненулевых элементов $(\forall a \in R,\ a \neq e,\ \exists b |\ a \times b = 1,\ b \equiv a^{-1})$, получим что все ненулевые элементы образуют коммутативную (абелеву) группу по умножению. Такое кольцо называется полем.

Полукольцо, с другой стороны, есть более общая структура. Это кольцо без требования существования обратного элемента относительно "сложения". Оно не является группой ни по одной из бинарных операций. Таким образом, алгебраически полукольцо R можно определить аксиомами

1. Относительно "+" (пересечение множеств; $0 = R$):

   (a) $(a+b)+c = a+(b+c)$

   (b) $0+a = a+0 = a$

   (c) $a+b = b+a$

2. Относительно "×" (объединение множеств; $1 = \varnothing$):

   (a) $(a \times b) \times c = a \times (b \times c)$

   (b) $1 \times a = a \times 1 = a$

3. "Умножение" дистрибутивно относительно "сложения":

   (a) $a \times (b+c) = (a \times b) + (a \times c)$

   (b) $(a+b) \times c = (a \times c) + (b \times c)$

4. $0 \times a = a \times 0 = 0$

Полукольцо в теории множеств

В теории множеств принято такое определение:

Система множеств $\mathcal{A}$ называется *полукольцом (semiring)*, если она

1. содержит пустое множество $\varnothing$: $\mathcal{A} \ni \varnothing$;

2. замкнута по отношению к образованию пересечений: $A, B \subset \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$;

3. $A, A_1 \subset \mathcal{A}$ и $A_1 \subset A \;\Rightarrow$ можно представить A в виде $A = \cup_{i=1}^{n} A_i$, где A_i – попарно непересекающиеся множества на $\mathcal{A}$, первое из которых есть заданное множество A_1.

Лебеговская мера на полукольце

Функция множества $\mu(A)$ на системе множеств $\mathcal{A} \supset A$ называется (конечно-аддитивной) мерой, если

1. ее область определения $\mathcal{A}$ есть полукольцо множеств

2. $\mu(A) \geq 0 \; \forall A$

3. $\mu(A)$ аддитивна, т.е. если $A_i \cap A_j = 0 \; \forall i \neq j$; $i, j = 1, \ldots, n$, то

$$m\left(\cup_1^n A_k\right) = \sum_{i=1}^{n} m(A_k). \tag{B.5}$$

Если расширить третий пункт до счетной аддитивности, т.е. до требования чтобы для любой *счетной* системы попарно непересекающихся множеств $\{A_k\}_{k=1}^{\infty}$ выполнялось $m\left(\cup_1^{\infty} A_k\right) = \sum\limits_{i=1}^{\infty} m(A_k)$, то получим определение *счетно-аддитивной (или σ-аддитивной) меры.*

Она удовлетворяет свойствам меры, перечисленным для длин отрезков в начале предыдущего параграфа.

Дальше можно определить лебеговское продолжение меры, определенной на полукольце, в таком же духе как это делалось для числовой прямой.

B.2 Измеримые функции

Множество Бореля (Borel set) в топологическом пространстве – это любое множество, которое можно получить счетным числом операций объединения и пересечения открытых и закрытых множеств. На числовой прямой это отрезки и интервалы. Всякая мера, определенная на множествах Бореля, называется борелевой мерой.

Пусть X и Y – два произвольных множества, в которых выделены две системы подмножеств $\mathcal{X}$ и $\mathcal{Y}$ соответственно. Функция $y = f(x) : X \mapsto Y$ называется $(\mathcal{X}, \mathcal{Y})$-измеримой, если $\{A \in \mathcal{Y} \;\Rightarrow\; f^{-1}(A) \in \mathcal{X}\}$.

Если в качестве X и Y взять числовые прямые, т.е. рассматривать действительные функции действительного аргумента, то это определение сведется к непрерывности.

Действительная функция $y = f(x) : X \mapsto \mathbb{R}$ *измерима*, если

$\forall c \in \mathbb{R} \quad \{x \in X \mid f(x) < c\}$ есть борелевское множество.

B.3 Интеграл Лебега

Идея построения интеграла Лебега состоит в том, что вместо разбиения области определения подынтегральной функции на части и составления потом интегральной суммы из значений функции на этих частях, на интервалы разбивают её область значений, а затем суммируют с соответствующими весами меры прообразов этих интервалов.

Интеграл Лебега определяется индуктивно, переходя от более простых функций к сложным. Пусть у нас есть пространство X с мерой μ, и на нем определена μ-измеримая функция $f(x)$.

1. Пусть $f(x) = 1_A(x)$ – индикатор некоторого измеримого множества A, т.е. $f(x) = 1$ если $x \in A$ и нулю если $x \notin A$. Тогда по определению
$$\int\limits_X f(x)\mu(dx) \equiv \int\limits_X f(x)d\mu = \mu(A).$$

2. Пусть $f(x)$ – простая функция, т.е. принимающая конечное число значений на X. Она тогда представима в качестве конечной линейной комбинации индикаторов $f(x) = \sum\limits_{i=1}^{n} f_i 1_{F_i}(x)$, где $f_i \in \mathbb{R}$, а $\{F_i\}_1^n$ – конечное разбиение X на измеримые множества. Тогда
$$\int\limits_X f(x)d\mu = \sum_{i=1}^{n} f_i\mu(F_i).$$

3. Пусть теперь $f(x) \geq 0 \ \forall x \in X$ – неотрицательная функция. Рассмотрим все простые функции $\{f_s(x)\}$, такие что $f_s(x) < f(s) \ \forall x \in X$. Тогда
$$\int\limits_X f(x)d\mu = \sup_{f_s} \int\limits_X f_s(x)d\mu.$$

4. Пусть $f(x)$ – функция произвольного знака. Тогда ее можно представить в виде разности неотрицательных функций
$$f(x) = f^+(x) - f^-(x), \quad \text{где}$$
$$f^+(x) = \max\{f(x), 0\},$$
$$f^-(x) = -\min\{0, f(x)\}.$$

и интеграл Лебега определяется как

$$\int\limits_X f(x)d\mu = \int\limits_X f^+(x)d\mu - \int\limits_X f^-(x)d\mu.$$

5. Наконец, если A – произвольное измеримое множество из X, то

$$\int\limits_A f(x)d\mu = \int\limits_X f(x)1_A(x)d\mu.$$

Приложение C

Справочные сведения по классическим ортогональным полиномам

C.1 Полиномы Эрмита H_n

ортогональны на $(-\infty, \infty)$ с весом[1] e^{-x^2}.

Принята нормировка $k_n = 2^n$, так что

$$\|H_n\|^2 = \sqrt{\pi} 2^n n!$$

Первые несколько H_n:

$$\begin{array}{ll} H_0 = 1 & H_4 = 16x^4 - 48x^2 + 12 \\ H_1 = 2x & H_5 = 32x^5 - 160x^3 + 120x \\ H_2 = 4x^2 - 2 & H_6 = 64x^6 - 480x^4 + 720x^2 - 120 \\ H_3 = 8x^3 - 12x & H_7 = 128x^7 - 1344x^5 + 3360x^3 - 1680x \end{array}$$

T^0: Если $f(x)$ кусочно-гладкая функция на $(-\infty, \infty)$ и существует

$$\int_{-\infty}^{\infty} dx|x|e^{-x^2} f^2(x),$$

то ряд Фурье по многочленам Эрмита сходится к $f(x)$ в точках непрерывности и к $\frac{1}{2}[f(x+0) + f(x-0)]$ в точках разрыва.

Коэффициенты ряда $f(x)$ по H_n и отклонение:

$$f_k = \frac{(-1)^k}{2^k k! \sqrt{\pi}} \int_{-\infty}^{\infty} dx e^{-x^2} f(x) H_k(x);$$
$$\delta_n^2 = \int_{-\infty}^{\infty} dx e^{-x^2} f^2(x) - \sum_{k=0}^{n} 2^k k! \sqrt{\pi} f_k^2.$$

[1] Это "физическое" определение, в теории вероятностей используется вес $e^{-x^2/2}$

C.2 Полиномы Лагерра $L_n^{(\alpha)}$

ортогональны на $[0,\infty)$ с весом $x^\alpha e^{-x}$.

Принята нормировка $k_n = (-1)^n/n!$, так что

$$\|L_n^{(\alpha)}\|^2 = \Gamma(\alpha+n+1)/n!$$

Первые полиномы

$$\begin{array}{ll} L_0^{(\alpha)} = 1 & L_0^{(0)} = 1 \\ L_1^{(\alpha)} = -x+\alpha+1 & L_1^{(0)} = -x+1 \\ L_2^{(\alpha)} = \frac{x^2}{2} - (\alpha+2)x + \frac{(\alpha+2)(\alpha+1)}{2} & L_2^{(0)} = \frac{1}{2}[x^2-4x+2] \\ L_3^{(\alpha)} = -\frac{x^3}{6} + \frac{(\alpha+3)x^2}{2} - \frac{(\alpha+3)(\alpha+2)x}{2} + \frac{(\alpha+3)(\alpha+2)(\alpha+1)}{6} & L_3^{(0)} = \frac{1}{6}[-x^3+9x^2-18x+6] \end{array}$$

T^0: Если $f(x)$ кусочно-гладкая функция на $(0,\infty)$ и существует

$$\int_0^\infty dx\, x^{\frac{\alpha}{2}-\frac{1}{4}} e^{-x/2} |f(x)|,$$

то ряд Фурье по многочленам Лагерра сходится к $f(x)$ в точках непрерывности и к $\frac{1}{2}[f(x+0)+f(x-0)]$ в точках разрыва.

Коэффициенты ряда $f(x)$ по $L_n^{(\alpha)}$ и отклонение:

$$f_k = \frac{1}{k!\Gamma(\alpha+k+1)} \int_0^\infty dx\, x^\alpha e^{-x} f(x) L_k^{(\alpha)}(x);$$
$$\delta_n^2 = \int_0^\infty dx\, x^\alpha e^{-x} f^2(x) - \sum_{k=0}^n k!\Gamma(\alpha+k+1) f_k^2.$$

C.3 Полиномы Лежандра P_n

ортогональны на $[-1,1]$ с весом 1.

Принятая нормировка $P_n(1) = 1$;

$$\|P_n\|^2 = \frac{2}{2n+1}.$$

Первые несколько полиномов

$$\begin{array}{ll} P_0 = 1 & P_4 = \frac{1}{8}[35x^4 - 30x^2 + 3] \\ P_1 = x & P_5 = \frac{1}{8}[63x^5 - 70x^3 + 15x] \\ P_2 = \frac{1}{2}[3x^2-1] & P_6 = \frac{1}{16}[231x^6 - 315x^4 + 105x^2 - 5] \\ P_3 = \frac{1}{2}[5x^3-3x] & P_7 = \frac{1}{16}[429x^7 - 693x^5 + 315x^3 - 35x] \end{array}$$

T^0: Если $f(x)$ непрерывна и имеет ограниченную производную на $[-1,1]$, то ряд Фурье по многочленам Лежандра сходится равномерно на $[-1,1]$.

Коэффициенты ряда $f(x)$ по P_n и отклонение:

$$f_k = \frac{2k+1}{2}\int_{-1}^{1} dx\, f(x)P_k(x); \quad \delta_n^2 = \int_{-1}^{1} dx\, f^2(x) - \sum_{k=0}^{n} \frac{2}{2k+1} f_k^2.$$

C.4 Полиномы Чебышева I рода T_n

ортогональны на $[-1,1]$ с весом $(1-x^2)^{-1/2}$.

Принятая нормировка $T_n(1)=1$;

$$\|T_0\|^2 = \pi, \quad \text{а для} \quad n>1 \quad \|T_n\|^2 = \pi/2.$$

Первые несколько T_n:

$$\begin{aligned} T_0 &= 1 & T_4 &= 8x^4 - 8x^2 + 1 \\ T_1 &= x & T_5 &= 16x^5 - 20x^3 + 5x \\ T_2 &= 2x^2 - 1 & T_6 &= 32x^6 - 48x^4 + 18x^2 - 1 \\ T_3 &= 4x^3 - 3x & T_7 &= 64x^7 - 112x^5 + 56x^3 - 7x. \end{aligned}$$

$\mathbf{T}^0$: Если $f(x)$ непрерывно дифференцируема на $[-1,1]$, то ряд Фурье по многочленам Чебышева I рода сходится равномерно на $[-1,1]$.

Коэффициенты ряда $f(x)$ по T_n и отклонение:

$$f_0 = \tfrac{1}{\pi}\int_0^{\pi} d\theta f(\cos\theta); \; f_k = \tfrac{2}{\pi}\int_0^{\pi} d\theta \cos(k\theta) f(\cos\theta); \; k \geq 1$$
$$\delta_n^2 = \int_{-1}^{1} \tfrac{dx\, f^2(x)}{\sqrt{1-x^2}} - \tfrac{\pi}{2}\left[2C_0 + \sum_{k=1}^{n} f_k^2\right].$$

C.5 Полиномы Чебышева II рода U_n

ортогональны на $[-1,1]$ с весом $(1-x^2)^{-1/2}$;

Можно определить как

$$U_n(x) = \frac{T'_{n+1}(x)}{n+1} = \frac{\sin[(n+1)\arccos x]}{\sqrt{1-x^2}}.$$

Тогда $U_n(1)=n+1$ и

$$\|U_n\|^2 = \pi/2.$$

Первые несколько U_n:

$$\begin{aligned} U_0 &= 1 & U_4 &= 16x^4 - 12x^2 + 1 \\ U_1 &= 2x & U_5 &= 32x^5 - 32x^3 + 6x \\ U_2 &= 4x^2 - 1 & U_6 &= 64x^6 - 80x^4 + 24x^2 - 1 \\ U_3 &= 8x^3 - 4x & U_7 &= 128x^7 - 192x^5 + 80x^3 - 8x \end{aligned}$$

$\mathbf{T}^0$: Если $f(x)$ имеет непрерывные производные до третьего порядка включительно на $[-1,1]$, то ряд Фурье по многочленам Чебышева II рода сходится равномерно на $[-1,1]$.

Коэффициенты ряда $f(x)$ по U_n и отклонение:

$$f_k = \frac{2}{\pi}\int\limits_0^\pi d\theta \sin\theta \cos((k+1)\theta) f(\cos\theta);$$
$$\delta_n^2 = \int\limits_{-1}^{1} dx\sqrt{1-x^2} f^2(x) - \frac{\pi}{2}\sum_{k=1}^{n} f_k^2.$$

Полиномы Чебышева II рода минимизируют на $[-1,1]$ норму L^1, т.е. $\int_{-1}^{+1} dx\,|P_n(x)|$, среди приведенных полиномов заданной степени n.

Основные свойства и соотношения для классических ортогональных полиномов

Полиномы	Интервал	Вес	Стандартизация	Уравнение	Ф-ла Родрига	Рекуррентное соотношение
Все	$(a,b)\subset\mathbb{R}$	$p(x)>0$		$\sigma(x)y''(x)+g(x)y'(x)=\lambda\, y(x)$	$\frac{1}{\xi_n p(x)}\frac{d^n}{dx^n}\{p(x)\sigma^n(x)\}$	$xp_n=a_np_{n+1}+b_np_n+a_{n-1}p_{n-1}$ $a_n=\frac{k_n}{k_{n+1}},\ b_n=\|p_n\|^{-2}\int dx\, p(x)\, xp_n^2$
Эрмита Hermite H_n	$(-\infty,\infty)$	e^{-x^2}	$k_n=2^n$ $\|H_n\|^2=2^n n!\sqrt{\pi}$	$\sigma=1,\ g=-2x,\ \lambda_n=-2n:$ $H_n''(x)-2xH_n'(x)+2nH_n(x)=0$	$\xi_n=(-1)^n,$ $\sigma(x)=1$	$H_{n+1}(x)-2xH_n(x)+2nH_{n-1}(x)=0$
Лагерра Laguerre $L_n^{(\alpha)}$	$[0,\infty)$	$x^\alpha e^{-x}$ $\alpha>-1$	$k_n=(-1)^n/n!$ $\|L_n\|^2=\Gamma(\alpha+n+1)/n!$	$\sigma=x,\ g=(\alpha+1-x),\ \lambda_n=-n:$ $xL_n''+(\alpha+1-n)L_n'+nL_n=0$	$\xi_n=(-1)^n$ $\sigma(x)=x$	$L_{n+1}-(x-\alpha-2n-1)L_n+$ $+n(\alpha+n)L_{n-1}=0$
Якоби Jacobi $P_n^{(\alpha,\beta)}$	$[-1,1]$	$(1-x)^\alpha(1+x)^\beta$ $\alpha,\beta>-1$	$P_n^{(\alpha,\beta)}(1)=\binom{n+\alpha}{n}$	$\sigma=1-x^2,\ g=\beta-\alpha-x(\alpha+\beta+2),$ $\lambda_n=-n(n+\alpha+\beta+1)$ $\dots$	$\xi_n=\frac{(-1)^n}{2^n n!},$ $\sigma(x)=1-x^2$	$k_n=2^{-n}\binom{2n+\alpha+\beta}{n}$ $b_n=\frac{(2n+\alpha+\beta+1)(\beta^2-\alpha^2)}{(2n+\alpha+\beta)_3}$
Лежандра Legendre P_n, L_n	– " – " –	$\alpha=\beta=0:$ $p=1$	$P_n(1)=1$ $\|P_n\|^2=\frac{2}{2n+1}$	$\sigma=1-x^2,\ g=-2x,\ \lambda_n=-n(n+1):$ $\frac{d}{dx}\left[(1-x^2)\frac{dP_n}{dx}\right]+n(n+1)P_n=0$	$\xi_n=(-2)^n n!$ $\sigma(x)=1-x^2$	$(n+1)P_{n+1}-(2n+1)xP_n+$ $+nP_{n-1}=0$
Чебышёва I Chebyshev I T_n	– " – " –	$\alpha=\beta=-1/2:$ $p=(1-x^2)^{-1/2}$	$T_n(1)=1$ $\|T_n\|^2=\begin{cases}\pi, & n=0\\ \pi/2, & n\neq 0\end{cases}$	$\sigma=1-x^2,\ g=-x,\ \lambda_n=-n^2:$ $(1-x^2)T_n''-xT'+n^2T_n=0$	$\xi_n=(-2)^n\frac{\Gamma(n+1/2)}{\sqrt{\pi}}$ $\sigma(x)=1-x^2$	$T_{n+1}-2xT_n+T_{n-1}=0$
Чебышёва II Chebyshev II U_n	– " – " –	$\alpha=\beta=+1/2:$ $p=(1-x^2)^{1/2}$	$U_n(1)=n+1$ $\|U_n\|^2=\pi/2$	$\sigma=1-x^2,\ g=-3x,\ \lambda_n=-n(n+2):$ $(1-x^2)U_n''-3xT'+(n^2+2n)U_n=0$	$\xi_n=2(-2)^n\frac{\Gamma(n+3/2)}{(n+1)\sqrt{\pi}}$ $\sigma(x)=1-x^2$	$U_{n+1}-2xU_n+U_{n-1}=0$

Следует иметь в виду, что способы стандартизации и нормировки полиномов могут быть разные, что соответственно меняет некоторые из соотношений, как например коэффициент ξ_n в формуле Родрига и рекуррентные соотношения. Уравнения очевидно не меняются.

Irene Anne Stegun

Более подробные табличные сведения об ортогональных полиномах и еще много всего можно найти в бумажных справочниках, например

- *Abramowitz and Stegun*: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, edited by Milton Abramowitz and Irene A. Stegun, New York: Dover, ISBN 0-486-61272-4.
  Online тут http://www.math.sfu.ca/ cbm/aands/ (html+jpeg).
  Перевод на русский: Абрамовиц М., Стиган И. Справочник по специальным функциям [Наука, 1979].
- Градштейн И.С., Рыжик И.М. "Таблицы интегралов, сумм, рядов и произведений"
- Г. Бейтмен и Д. Эрдейи "Высшие трансцендентные функции, том 2" (теория и формулы);
- Gabor Szegö, "Orthogonal polynomials" (1939), на русском Г. Сегё "Ортогональные многочлены" – классическая книга по ортогональным полиномам.

* Обозначения:
$\binom{a}{b}=\frac{\Gamma(b+1)}{\Gamma(a+1)\Gamma(b-a+1)}$ – аналитическое продолжение биномиальных коэффициентов C_b^a на вещественные или комплексные значения аргументов;
$(x)_n=x(x-1)(x-2)\dots(x-n+1)$ (Pochhammer symbol for the falling factorial).

Приложение D

Компактные операторы в L^2

D.1 Компактность

Класс операторов, по своим свойствам близких к операторам, действующим в конечномерных пространствах, образуют так называемые *вполне непрерывные* или *компактные (compact)* операторы (примерно в том же смысле что гильбертово пространство, среди бесконечномерных пространств, близко по свойствам к конечномерным пр.-вам – в нем всякий вектор можно разложить по базису). Для их определения напомним понятия компактности последовательностей и множеств в метрических пространствах[1].

D.1.1 Компактные множества в конечномерных и бесконечномерных пространствах

Последовательность y_n называется компактной, если из нее можно выделить сходящуюся подпоследовательность.

Множество K называется компактным (или компактом), если всякая последовательность его элементов компактна в K, т.е. если из нее можно выделить подпоследовательность, сходящуюся к элементу K.

Множество называется *предкомпактным* (или относительно компактным, или предкомпактом), если его замыкание компактно. Т.е. мн-во предкомпактно, если любая последовательность его элементов компактна (не обязательно в K).

[1] Компактность наиболее общим образом вводится в топологических пространствах. Метрические линейные пространства, однако, всегда отделимы (хаусдорфовы), и в них определения можно дать проще.

$\mathbf{T}^0$: В конечномерном евклидовом пространстве E_n множество компактно тогда и только тогда, когда оно ограничено и замкнуто[2].

◂ Если K компактно, то оно ограничено: доказывается от противного. Пусть K неограничено. Тогда можно построить последовательность его элементов x_k, таких что $\|x_k\| \to \infty$. Но такая последовательность не сходится, и из нее нельзя выделить сходящуюся подпоследовательность. Значит K не компактно. Противоречие. Доказано.

Если K компактно, то оно замкнуто: также от противного. Пусть K незамкнуто. Тогда существует последовательность элементов K, которая сходится к $x \notin K$. Вследствие единственности предела, тогда K не компактно. Противоречие. Доказано.

Если K ограничено и замкнуто, то оно компактно: опять так же. Пусть K не компактно. Тогда в нем существует бесконечная последовательность x_k, не сходящаяся в K. Так как K ограничено, то и последовательность ограничена. Тогда в конечномерном пространстве можно найти куб размера L, в котором она содержится. Разделим L пополам, тогда куб поделится на 2^n частей (n – размерность пр.-ва). По крайней мере одна из частей содержит бесконечно большое число элементов x_k. Делим ее так же, и продолжаем до бесконечности. В каждом очередном кубе выбирая по точке из x_k, получаем подпоследовательность, которая сходится. Так как K замкнуто, то она сходится в K. Противоречие. Доказано. ▸

В бесконечномерном пространстве E_∞ доказательство необходимости остается в силе – всякое компактное множество ограничено и замкнуто. Однако для доказательства достаточности было существенно условие конечномерности, и в E_∞ она не выполняется. Поэтому свойство компактности оказывается более сильным, чем просто требования ограниченности и замкнутости.

Так, возьмем в пространстве l^2 последовательность векторов $\{x_i\}$ такую: $(1, 0, 0, \ldots)$, $(0, 1, 0, \ldots)$ и так далее, так что i-й вектор имеет на i-й позиции единицу, а на остальных нули. Тогда $\forall\, i\ \|x_i\| = 1$, так что последовательность ограничена. Однако для любой пары $\|x_i - x_j\| = \sqrt{2}$, так что любая подпоследовательность из бесконечного числа элементов не сходится. Таким образом, все точки $\{x_i\}$ изолированы, и эта последовательность ограничена и замкнута, но не компактна. Эта последовательность лежит на единичной сфере в l^2. Поэтому и единичная сфера, которая очевидно ограничена и замкнута, – не

[2] См. теорему Больцано-Вейерштрасса в мат. анализе, в которой это доказывается для числовой прямой.

компактна.

Примеры компактов в банаховых пространствах:

- Компактным является любое ограниченное и замкнутое подмножество конечномерного подпространства E_∞.
- *Фундаментальный параллелепипед* ("гильбертов кирпич") пространства l^2 – множество точек $(x_1, x_2, \ldots)$ с координатами $|x_i| \leq 1/2^{i-1}$.

Полная ограниченность. Пусть M – множество метрического пространства R, а $\varepsilon > 0$. Множество A из R называется ε-сетью для M, если

$$\forall\, x \in M\, \exists\, a \in A \,|\, \rho(x,a) < \varepsilon.$$

Множество называется вполне ограниченным, если для него при любом ε существует конечная ε-сеть.

Множество M, лежащее в в полном метрическом пространств, компактно т. и т.т., когда оно вполне ограничено.

D.1.2 Компактные операторы

Оператор A называется компактным, если для всякой ограниченной последовательности y_n последовательность Ay_n компактна. Таким образом, компактный оператор переводит всякое ограниченное множество в предкомпактное.

Так как компактное множество ограничено, то компактный оператор непрерывен и ограничен. Однако обратное неверно. Так, единичный оператор переводит единичную сферу, которая ограничена но не компактна, в себя, а потому ограничен но не компактен!

$\mathbf{T}^0$: Линейная комбинация и произведение компактных операторов компактны ◂.▸.

$\mathbf{T}^0$: Если последовательность компактных операторов в банаховом пространстве сходится, то ее предел есть также компактный оператор.

◂ Пусть A_N есть последовательность компактных операторов, которая сходится к A, и пусть $\{x_i\}$ – произвольная ограниченная последовательность $\|x_i\| < C$.

Так как A_1 компактен, то из $\{A_1 x_i\}$ можно выделить сходящуюся подпоследовательность. Тогда пусть $\{x_i^{(1)}\}$ это такая подпоследовательность $\{x_i\}$, что $\{A_1 x_i^{(1)}\}$ сходится.

Рассмотрим теперь посл-ть $\{A_2 x_i^{(1)}\}$. Из нее тоже можно выбрать сходящуюся подпосл-ть, вследствие компактности A_2. Пусть $\{x_i^{(2)}\}$ это такая подпосл-ть $\{x_i^{(1)}\}$, что $\{A_2 x_i^{(2)}\}$ сходится. Понятно, что и $\{A_1 x_i^{(2)}\}$ сходится.

Повторяя такую выборку, получим на шаге m последовательность $\{x_i^{(m)}\}$, такую что $\{A_n x_i^{(m)}\}$ сходится для $n=1,2,\ldots,m$.

Возьмем теперь диагональную последовательность $x_1^{(1)}, x_2^{(2)}, \ldots, x_n^{(n)}, \ldots$. Каждый из операторов $A_1, A_2, \ldots$ переводит ее в сходящуюся. Тогда из неравенства треугольника

$$\|Ax_n^{(n)} - Ax_m^{(m)}\| \le \|Ax_n^{(n)} - A_k x_n^{(n)}\| + \|A_k x_n^{(n)} - A_k x_m^{(m)}\| + \|A_k x_m^{(m)} - Ax_m^{(m)}\|.$$

Так как ряд $\{A_N\}$ сходится по норме к A, то при достаточно большом k можно сделать $\|A - A_k\| < \varepsilon C$ для любого ε, так что первое и третье слагаемое меньше ε. Второе слагаемое стремится к нулю, так как $A_k x_n^{(n)}$ сходится, а значит фундаментальна. Поэтому $\{Ax_n^{(n)}\}$ фундаментальна, а значит сходится (банахово пространство полно). Таким образом, A компактен, ч. и т.д. ▸

D.2 Оператор Фредгольма

T^0: *Оператор Фредгольма* (4.9, 4.10)

$$A\,y = \psi \quad \Leftrightarrow \quad \int_a^b ds\, K(x,s) y(s) = \psi(x), \quad | \quad \int_a^b \int_a^b dx\, ds\, |K(x,s)|^2 \equiv M < \infty,$$

компактен.

◂ Во-первых, заметим, что ограниченность A уже доказана (4.13).

Вследствие условия (4.9) и теоремы Фубини, интеграл $\int_a^b ds |K(x,s)|^2$ существует для почти всех x. Иначе говоря, $K(x,s)$ как функция s при почти всех x принадлежит $L^2[a,b]$. Так как произведение интегрируемых с квадратом функций интегрируемо, то интеграл в (4.10) существует для почти всех x, т.е. функция ψ определена почти всюду.

Осталось показать, что оператор A компактен. Как мы видели, разложив ядро в ряд Фурье (4.17), оператор Фредгольма A можно представить как предел последовательности вырожденных операторов (4.18) A_N.

Но вырожденный оператор A_N, по определению, переводит все пространство $L^2[a,b]$ в конечномерное подпространство, порожденное векторами $\psi_1, \ldots, \psi_N$. Поэтому он переводит любую ограниченную последовательность из L^2 в ограниченную последовательность в конечномерном подпространстве L^2, а из по-

следней всегда можно выделить сходящуюся подпоследовательность. Поэтому A_N компактны.

Следовательно, по теореме о сходящейся последовательности компактных операторов, A компактен. ▸

Свойство компактности A является основополагающим для доказательства утверждений о существовании и структуре его собственных значений и векторов.

Литература

[1] Stoer J., Bulirsch R. *Introduction to numerical analysis*, 3ed., Springer, 2002, ISBN 038795452X, 755p.

[2] Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. *Численные методы*, Москва, ФМЛ, 2001, 630 стр.

[3] Калиткин Н.Н. *Численные методы*, Москва, Наука, 1978, 583 стр.

[4] Хемминг Р.В. *Численные методы*, Москва, Наука, 1972. [R.W. Hamming *Numerical methods for scientists and engineers*, Mc Graw-Hill, 1962].

[5] Демидович Б.П., Марон И.А., Шувалова Э.З. *Численные методы анализа*, 3е изд, Москва, Наука, 1967, 368стр.

[6] Т. Шуп. *Решение инженерных задач на ЭВМ*. Москва, Мир, 1982, 238 стр. [Terry E. Shoup. *A practical guide to computer methods for engineers*, Prentice-Hall Inc, Englewood Cliffs, N.J., 1979.].

[7] Butcher J. *Numerical methods for ordinary differential equations*, 2ed., Wiley, 2008, ISBN 0470723351, 482p.

[8] Acton F.S. *Numerical methods that usually work*; rev.ed., MAA, 1990, ISBN 0883854503.

[9] Березин И.С., Жидков Н.П. *Методы вычислений*, в двух томах, 1962, 464+620 стр.

[10] Тихонов А.Н, Самарский А.А. *Уравнения математической физики*, дополнение I.

[11] Колмогоров А.Н, Фомин С.В. *Элементы теории функций и функционального анализа*, 7 изданий 1968–2007 гг.

[12] Кадец В.М. *Курс функционалъного анализа*, Харьков, 2006.

[13] Смирнов В.И. *Курс высшей математики, том 5*, БХВ-Петербург, 24-е изд. 2008 г.

[14] Allan Pincus, Weierstrass and Approximation Theory, *J. Approx. Theory*, **107** (2000), 1-66.

[15] V. Totik, Orthogonal polynomials, *Surveys in Approximation Theory* (2005), **1**, 70-125.

[16] Никифоров А.Ф., Суслов С.К., *Классические ортогоналъные полиномы*, Москва, Знание, 1985 (Новое в жизни, науке, технике; сер. "Математика, кибернетика", №12).

[17] John A. Gubner, Gaussian Quadrature and the Eigenvalue Problem.

[18] Васильева А.Б. и др. *Дифференциалъные и интегралъные уравнения, вариационное исчисление в примерах и задачах.*

[19] Васильева А.Б., Тихонов Н.А. *Интегралъные уравнения.*

[20] Polyanin A.D., Manzhirov A. *Handbook of integral equations*, 2008, ISBN 1584885076, 1143p.