УРСС

ВСЯ ВЫСШАЯ МАТЕМАТИКА

М.Л.Краснов
А.И.Киселев
Г.И.Макаренко
Е.В.Шикин
В.И.Заляпин

6

ВСЯ ВЫСШАЯ МАТЕМАТИКА

М.Л.Краснов
А.И.Киселев
Г.И.Макаренко
Е.В.Шикин
В.И.Заляпин

6

Рекомендовано
Министерством образования
Российской Федерации
в качестве учебника для студентов
высших технических учебных заведений

УРСС

Москва • 2003

ББК 22.1я73

**Красиов Михаил Леонтьевич, Киселев Александр Иванович,
Макаренко Григорий Иванович, Шикин Евгений Викторович, Заляпин Владимир Ильич**

Вся высшая математика: Учебник. Т. 6. — М.: Едиториал УРСС, 2003. — 256 с.

ISBN 5–354–00386–5

Предлагаемый учебник впервые вышел в свет в виде двухтомника сначала на английском и испанском языках в 1990 году, а затем на французском. Он пользуется большим спросом за рубежом.

В 1999 году книга стала лауреатом конкурса по созданию новых учебников Министерства образования России.

Этот учебник адресован студентам высших учебных заведений (в первую очередь будущим инженерам и экономистам) и охватывает практически все разделы математики, но при этом представляет собой не набор разрозненных глав, а единое целое.

Шестой том включает в себя материал по вариационному исчислению, линейному программированию, вычислительной математике и теории сплайнов.

Издательство «Едиториал УРСС». 117312, г. Москва, пр-т 60-летия Октября, 9.
Лицензия ИД №05175 от 25.06.2001 г. Подписано к печати 25.04.2003 г.
Формат 70×100/16. Тираж 5000 экз. Печ. л. 16. Заказ №64.

Отпечатано в ООО «Арт-диал». 129110, г. Москва, ул. Б. Переяславская, 46.

ISBN 5–354–00270–2 (Полное произведение)
ISBN 5–354–00386–5 (Том 6)

Издательство УРСС
научная и учебная литература

Тел./факс: 7(095)135–44–23
Тел./факс: 7(095)135–42–46
E-mail: URSS@URSS.ru
Каталог изданий в *Internet:* http://URSS.ru

ОТ АВТОРОВ

Этот том отличается от всех предыдущих тем, что только один из его разделов — «Вариационное исчисление» — включает наборы задач. Все остальные разделы свободны от каких бы то ни было упражнений. Это объясняется тем, что разделы, отведенные под численные методы, линейное программирование и сплайны, представляют собой необходимое теоретическое и алгоритмическое преддверие вычислительного практикума, который естественно опирается на использование компьютеров, в том числе и персональных. Отбор заданий для такого практикума, описание соответствующего программного обеспечения, анализ полученных результатов и наиболее часто встречающихся ошибок — задача, бесспорно важная, но заметно выходящая за рамки данного издания как по объему, так и по специфике. Предлагать же задачи с громоздкими вычислениями для счета «на руках» — вещь идеологически неправильная, особенно при наличии значительного числа всевозможных программных средств (пакетов).

ВАРИАЦИОННОЕ ИСЧИСЛЕНИЕ. НЕОБХОДИМЫЕ УСЛОВИЯ

Задача поиска экстремумов функции одной или нескольких переменных

$$f(x) = f(x_1, x_2, \ldots, x_n) \longrightarrow \text{extr}$$

является одной из важнейших. Она заключается в отыскании оптимальных значений аргумента $x_{\min}$ и/или $x_{\max}$ и соответствующих им значений функции

$$f_{\min} = f(x_{\min}), \quad f_{\max} = f(x_{\max}) \quad \text{таких, что} \quad f_{\min} \leqslant f(x), \quad f_{\max} \geqslant f(x)$$

для всех допустимых (выделенных естественными требованиями, предъявляемыми к области определения функции, или особо оговоренными условиями) значений аргумента x. Если множество ограничений D достаточно «массивно», то говорят о задаче поиска *абсолютного* или *глобального экстремума* функции $f(x)$ на множестве D. Если же множество D включает в себя только «ближайшие» к $x_{\min}$ или $x_{\max}$ точки $\left(D = \{x: \|x - x_{\max}\| < \delta\}\right.$ или $\left.D = \{x: \|x - x_{\min}\| < \delta\}\right)$, то говорят о задаче поиска *локального экстремума*.

Исследованию задач, подобных по постановке вышеизложенным, с тем отличием, что объектом минимизации являются не функции, а более общие отображения — *функционалы*, — посвящено вариационное исчисление. При этом естественными аргументами оптимизации являются аргументы функционалов — функции. Так, например, если ставится задача поиска *наискорейшего пути* из одного города в другой, то *минимизируемым* функционалом будет время, аргументом минимизации — путь, а ограничения выделяют доступные путешественнику пути, ведущие из одного города в другой. Другим примером может служить задача об облете самолетом области *наибольшей площади*. Здесь *максимизируемый* функционал — площадь облетаемой области, аргумент максимизации — траектория полета, ограничения — например, запас горючего.

Интересно отметить, что несмотря на современный математический аппарат, использующий последние достижения функционального анализа, топологии, теории обыкновенных дифференциальных уравнений и уравнений в частных производных и изощренную технику, основные идеи и принципы решения новых, неклассических экстремальных задач поразительно похожи на идеи и принципы решения задач классических, известных со времен Ферма, Эйлера и Лагранжа. В частности, основополагающий *принцип множителей Лагранжа*, позволяющий сводить исследование экстремальной задачи с ограничениями к исследованию задачи на экстремум без ограничений, оказывается эффективным в самых разнообразных ситуациях.

Глава XLIX

ЭКСТРЕМУМЫ ФУНКЦИОНАЛОВ

Мы начнем изложение с рассмотрения вопросов, касающихся описания основных объектов, с которыми нам придется иметь дело в дальнейшем, и их свойств. Этими объектами являются *функциональные пространства и функционалы*, определенные на элементах этих пространств.

§ 1. Некоторые сведения и понятия из функционального анализа

1.1. Функциональные пространства

Пусть $x(t)$, $y(t)$, $z(t)$, ..., $u(t)$, ... — функции, заданные на отрезке $[a, b]$ и обладающие некоторыми оговоренными заранее свойствами — непрерывностью, дифференцируемостью и т. д. Их сумма, разность и произведение на число определены на этом же отрезке и обладают теми же свойствами. В этом случае будем говорить, что определено *функциональное пространство*, элементами (точками, векторами) которого являются функции. Если дополнительно для любой пары $x(t)$, $y(t)$ функций определено число $\|x - y\|$, называемое *расстоянием* и обладающее свойствами обычного расстояния, т. е.

1. $\|x - 0\| = \|x\| \geqslant 0$, $\|x\| = 0 \iff x(t) \equiv 0$,
2. $\|ax\| = |a| \cdot \|x\|$,
3. $\|x + y\| \leqslant \|x\| + \|y\|$,

то функциональное пространство называется *нормированным*.

Наличие в функциональном пространстве расстояния позволяет обычным образом определить *окрестность* элементов этого пространства — ε-окрестностью элемента x_0 является множество элементов x этого пространства, таких, что

$$\|x - x_0\| < \varepsilon.$$

Как и в обычном анализе, можно определить сходимость последовательности элементов, т. е. ввести понятие предела

$$x = \lim_{n \to \infty} x_n \iff \forall \varepsilon > 0 \quad \exists n_0 \colon \forall n > n_0 \quad \|x - x_n\| < \varepsilon.$$

Заметим, что на одном и том же множестве функций расстояние можно задавать по-разному. При этом будут получаться различные *нормированные* пространства. Близость элементов в них будет пониматься по-разному и, соответственно, по-разному будет выглядеть предел последовательности элементов.

Важными являются следующие два примера.

Пример 1. Рассмотрим на отрезке $[a, b]$ совокупность непрерывно дифференцируемых функций. Определим расстояние между функциями $x(t)$ и $y(t)$ соотношением

$$\|x - y\| = \max_{t\in[a,b]} |x(t) - y(t)|. \tag{1}$$

◀ Легко проверить, что свойства 1–3 для таким образом определенного расстояния выполняются. Близкими в смысле расстояния (1) будут функции, модуль разности которых невелик. К примеру, функции $x(t) = 1 + \varepsilon \sin Mt$ и $y(t) = 1$ близки в указанном смысле при достаточно малом ε и любом M, так как

$$\|x(t) - y(t)\| = \|\varepsilon \sin Mt\| = \max_{t\in[a,b]} |\varepsilon \sin Mt| \leqslant \varepsilon. \ ▶$$

Пример 2. Определим на этом же множестве (непрерывно дифференцируемых функций) другое расстояние

$$\|x - y\| = \max_{t\in[a,b]} \left\{|x(t) - y(t)|, |x'(t) - y'(t)|\right\}. \tag{2}$$

◀ Как и выше, легко проверить, что требования 1–3 выполнены. Близкими в смысле расстояния (2) будут функции, которые близки на отрезке $[a, b]$ *вместе со своими производными*. Функции, рассмотренные выше, при фиксированном значении ε могут быть сделаны сколь угодно далекими за счет выбора M, так как

$$\|x(t) - y(t)\| = \|\varepsilon \sin Mt\| = \max_{t\in[a,b]} \left\{|\varepsilon \sin Mt|, |M\varepsilon \cos Mt|\right\} = M\varepsilon$$

при $M \geqslant 1$. ▶

Заметим, что расстояние (2) «сильнее» расстояния (1) — из близости в смысле расстояния (2) следует близость в смысле расстояния (1). Обратное, как свидетельствует рассмотренный пример, неверно.

Выбор способа вычисления расстояния на том или ином множестве функций, как правило, определяется содержанием задачи, приводящим к рассмотрению именно этого класса функций.

Множество непрерывных функций с расстоянием (1) обычно обозначается $C_{[a,b]}$. Множество гладких (непрерывно дифференцируемых) функций с расстоянием (2) обозначается $C^1_{[a,b]}$. Символами $KC_{[a,b]}$ и $KC^1_{[a,b]}$ соответственно будем обозначать пространства *кусочно-непрерывных* и *кусочно-гладких* функций. В первом из них расстояние будем находить из соотношения (1), во втором — из (2).

1.2. Функционалы

Пусть $\mathbf{B}$ — некоторое функциональное пространство с введенным на множестве его элементов расстоянием[1)] и пусть задано правило f, в соответствии с которым каждому элементу некоторого подмножества $\mathbf{D} \subseteq \mathbf{B}$ поставлено в соответствие число,

$$f\colon \mathbf{D} \longrightarrow \mathbf{R}.$$

Этот факт можно записывать обычным образом $y = f(x)$ и говорить, что задан *функционал*, область определения которого есть множество $\mathbf{D} \subseteq \mathbf{B}$.

Данное определение является очевидным обобщением понятия числовой функции и так же, как для обычных числовых функций, для функционалов вводятся понятия непрерывности, дифференцируемости и т. п.

Так, например, функционал $y = f(x)$ называется *непрерывным* в $x_0 \in \mathbf{D}$, если из $x_0 = \lim\limits_{n\to\infty} x_n$ следует, что $f(x_0) = \lim\limits_{n\to\infty} f(x_n)$.

[1)] Ниже под $\mathbf{B}$ можно понимать одно из пространств, описанных в п. 1.1.

Функционал $f(x)$ называется *линейным*, если он определен на всем $\mathbf{B}$, непрерывен в каждой точке и обладает свойством *аддитивности*, т. е. для любых x и y выполняется равенство $f(x+y) = f(x) + f(y)$. В функциональном анализе устанавливается, что линейный функционал обладает свойствами

— *однородности*: $f(ax) = af(x)\ \forall\ a \in \mathbf{R},\ x \in \mathbf{B}$, и

— *ограниченности*: $\exists\ C$ — постоянная, такая, что $|f(x)| \leqslant C \cdot \|x\|\ \forall\ x \in \mathbf{B}$.

Свойства ограниченности и непрерывности линейного функционала взаимозаменяемы — из одного из них следует другое.

Функционал в $C^1_{[0,1]}$, задаваемый соотношением

$$f(x) = x\left(\frac{1}{2}\right) + x'\left(\frac{1}{2}\right),$$

линеен.

Вот еще несколько примеров линейных функционалов:

— в пространстве $C_{[0,1]}$

$$f(x) = \int_0^1 x(t) \sin \pi t\ dt,$$

$$f(x) = \int_0^1 x(t)t^2\ dt,$$

$$f(x) = x(0) + x(1);$$

— в пространстве $C^1_{[0,1]}$

$$f(x) = \int_0^1 x(t)g(t) + x'(t)h(t)\ dt, \quad \text{здесь} \quad g(t),\ h(t) \quad \text{непрерывны},$$

$$f(x)c = \alpha x(0) + \beta x(1), \quad f(x) = \alpha x'\left(\frac{1}{2}\right).$$

Функционал в $C_{[0,1]}$, задаваемый равенством $f(x) = x'\left(\frac{1}{2}\right)$, линейным не является, несмотря на то, что обладает свойствами аддитивности и однородности.

В дальнейшем нам в пространствах $C^1_{[a,b]}$, $KC^1_{[a,b]}$ придется иметь дело с функционалами, заданными соотношениями

$$f(x) = \int_a^b L(t, x(t), x'(t))\ dt \text{ — } \textit{классический интегральный функционал},$$

$$f(x) = l(x(a), x(b)) \text{ — } \textit{классический терминальный функционал},$$

$$f(x) = \int_a^b L(t, x(t), x'(t))\ dt + l(x(a), x(b)) \text{ — } \textit{функционал Больца},$$

где $L(u, v, w)$, $l(\tau, s)$ — заданные функции.

1.3. Экстремумы функционалов

Пусть $\mathbf{B}$ — функциональное пространство, $f(x)$ — функционал, определенный на некотором множестве $\mathbf{D} \subseteq \mathbf{B}$. Будем говорить, что элемент $x_{\min} \in \mathbf{D}$ доставляет функционалу $f(x)$ наименьшее на множестве $\mathbf{D}$ значение (*минимум*), если

$$\forall\, x \in \mathbf{D} \quad f(x_{\min}) \leqslant f(x). \tag{3}$$

Записывать этот факт принято так

$$x_{\min} = \operatorname{argmin} f(x)\big|_{x \in \mathbf{D}}.$$

Если в формуле (3) поставить строгий знак неравенства, то можно говорить о *строгом* минимуме. Аналогично определяются элемент $x_{\max}$, доставляющий функционалу *максимум*, и *строгий* максимум функционала. Обозначение

$$x_{\max} = \operatorname{argmax} f(x)\big|_{x \in \mathbf{D}}.$$

Максимум и *минимум* функционала на $\mathbf{D}$ объединяют общим термином *экстремумы* функционала, точки называют точками экстремума, значения функционала — экстремальными значениями. Так определенные экстремумы называют еще *абсолютными* или *глобальными* экстремумами в $\mathbf{D}$. Если множество $\mathbf{D}$ представляет собой окрестность элемента $x_{\min}$, т. е. множество элементов $\mathbf{B}$, близких в смысле определенного в $\mathbf{B}$ расстояния к элементу $x_{\min}$, то говорят, что этот элемент доставляет функционалу *локальный* или *относительный* минимум. В этом случае пишут

$$x_{\min} = \operatorname{arglocmin} f(x) \iff f(x_{\min}) \leqslant f(x) \quad \forall\, x \in \mathbf{D}\colon \|x - x_{\min}\| < \delta.$$

Локальный максимум определяют аналогично.

В случае, когда элементы пространства $\mathbf{B}$ непрерывно дифференцируемы, различают *сильный* и *слабый* локальные экстремумы. О сильном экстремуме говорят тогда, когда близость функций из $\mathbf{B}$ описывается расстоянием, определенным соотношением (1), о слабом — в случае использования понятия близости, основанного на соотношении (2).

Ясно, что глобальный в $\mathbf{D}$ экстремум является в то же время и локально сильным, и локально слабым экстремумом. Столь же очевидно, что всякий сильный экстремум одновременно является и слабым. Обратные утверждения неверны. В частности, слабый экстремум не обязан быть сильным (пример, иллюстрирующий эту возможность, приведен во введении к гл. LIV).

Поэтому условия, *из которых следует существование* «старшего» экстремума (достаточные условия экстремума), обеспечивают существование и «младших». В то время как условия, *следующие из существования* «младших» экстремумов (необходимые условия экстремума), необходимы и для существования старших.

§ 2. Необходимые условия экстремума

Исследование экстремальных задач для функционалов мы начнем с установления *необходимых условий слабого локального экстремума.*

Пусть функционал $f(x)$ определен в слабой окрестности элемента $x_0 \in C^1_{[a,b]}$ и пусть x_0 — элемент слабого экстремума (для определенности — минимума) этого функционала. Пусть $h \in C^1_{[a,b]}$ и величина (норма) $\|h\|$ мала. Тогда приращение

$\Delta_h f(x)$ функционала $f(x)$ в x_0, соответствующее приращению h, будет неотрицательным для любых h

$$f(x_0+h)-f(x_0)\geqslant 0 \iff \Delta_h f(x)=f(x_0+h)-f(x_0)\geqslant 0. \qquad (1)$$

Соотношение (1) и есть искомое необходимое условие. Отметим, что в случае задачи на максимум знаки неравенства $\geqslant$ в (1) нужно заменить на $\leqslant$.

2.1. Вариации функционалов

Выделим те функционалы, для которых приращение допускает специальное представление.

Назовем функционал *дифференцируемым* (по Фреше) в x_0, если его приращение представимо в виде

$$\Delta_h f(x)=\delta_h f(x_0)+r(x_0, h), \qquad (2)$$

где $\delta_h f(x_0)$ — линейный по h функционал, а функционал-остаток $r(x_0, h)$ является величиной бесконечно малой в сравнении с главной (линейной относительно приращения аргумента h) частью приращения функционала

$$\lim_{\|h\|\to 0}\frac{|r(x_0, h)|}{\|h\|}=0.$$

Таким образом, дифференцируемые функционалы — это функционалы, мало отличающиеся от линейных в окрестности точки x_0.

Главная (линейная относительно приращения аргумента h) часть приращения функционала $\delta_h f(x_0)$ называется *вариацией по Фреше* функционала $f(x)$ в точке x_0.

Пусть h — некоторое приращение аргумента, а λ — число. *Функцией Лагранжа* назовем функцию $\Lambda(\lambda)$, задаваемую при фиксированных h и x_0 равенством

$$\Lambda(\lambda)=f(x_0+\lambda h).$$

Если для любого приращения h существует производная функции Лагранжа в точке $\lambda=0$

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

то она называется *лагранжевой вариацией* функционала $f(x)$ в точке x_0.

Заметим, что из существования *вариации по Фреше* следует, что существует *лагранжева вариация* и они тождественны.

◀ Заменяя в числителе дроби под знаком предела

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

приращение функционала его выражением из (2) и учитывая, что $\delta_{\lambda h} f(x_0)=\lambda\delta_h f(x_0)$ в силу линейности по h вариации Фреше, получаем

$$\Lambda'(0)=\delta_h f(x_0)+\lim_{\lambda\to 0}\frac{r(\lambda h, x_0)}{\lambda}=\delta_h f(x_0)+\lim_{\lambda\to 0}\frac{r(\lambda h, x_0)\|h\|}{\|\lambda h\|}=\delta_h f(x_0). \blacktriangleright$$

Обратное уже неверно — существование лагранжевой вариации не обеспечивает существование вариации по Фреше.

Пример 1. Действительно, рассмотрим функционал $f(x)$, заданный соотношением

$$f(x) = \sqrt[3]{x^3(a) + x^3(b)}.$$

◀ В точке $x(t) \equiv 0$ имеем

$$\frac{d}{d\lambda}\sqrt[3]{(\lambda h(a))^3 + (\lambda h(b))^3}\Big|_{\lambda=0} = \sqrt[3]{h^3(a) + h^3(b)}.$$

Полученное выражение не является линейным относительно допустимых приращений $h(t)$, поэтому вариации по Фреше этого функционала в рассматриваемой точке не существует. ▶

Рассмотрим еще несколько примеров.

Пример 2. Пусть $l(s, \tau)$ — непрерывно дифференцируемая функция двух переменных, $x(t) \in C^1_{[a,b]}$. Рассмотрим терминальный функционал а $C^1_{[a,b]}$, задаваемый соотношением

$$f(x) = l(x(a), x(b)).$$

◀ Для лагранжевой вариации получаем

$$\Lambda'(0) = \frac{d}{d\lambda} l(x(a) + \lambda h(a), x(b) + \lambda h(b))\Big|_{\lambda=0} = \frac{\partial l}{\partial x(a)} h(a) + \frac{\partial l}{\partial x(b)} h(b).$$

Для нахождения вариации по Фреше следует рассмотреть приращение функционала и выделить его линейную по h часть

$$\Delta_h l = l(x(a) + h(a), x(b) + h(b)) - l(x(a), x(b)).$$

В силу непрерывной дифференцируемости $l(s, \tau)$ ее приращение можно заменить дифференциалом

$$\Delta_h l = \frac{\partial l}{\partial x(a)} h(a) + \frac{\partial l}{\partial x(b)} h(b) + r(x, h).$$

Легко устанавливается, что остаточный член r мал в сравнении с линейными по h слагаемыми, откуда и следует дифференцируемость по Фреше. Как и было замечено выше, обе вариации совпадают. ▶

Пример 3. Пусть $L(u, v, w)$ — непрерывно дифференцируемая функция трех переменных, $x(t) \in C^1_{[a,b]}$. Рассмотрим интегральный функционал в $C^1_{[a,b]}$, задаваемый соотношением

$$f(x) = \int_a^b L(t, x(t), x'(t))\, dt. \tag{3}$$

Покажем, что он дифференцируем по Фреше и найдем его вариацию.

◀ Имеем

$$f(x + h) - f(x) = \int_a^b \left[L(t, x(t) + h(t), x'(t) + h'(t)) - L(t, x(t), x'(t))\right] dt.$$

Заменяя под знаком интеграла приращение функции $L(u, v, w)$ дифференциалом, приходим к соотношению

$$\Delta_h f(x) = \int_a^b \left[\frac{\partial L}{\partial x} h(t) + \frac{\partial L}{\partial x'} h'(t)\right] dt + \int_a^b R\, dt,$$

линейная часть которого и дает выражение для вариации интегрального функционала (3)

$$\delta_h f(x) = \int_a^b \left[\frac{\partial L}{\partial x} h(t) + \frac{\partial L}{\partial x'} h'(t)\right] dt.$$

Лагранжева вариация будет такой же в силу установленной дифференцируемости рассматриваемого функционала. ▶

Пример 4. Пусть функции $L(u, v, w)$, $l(s, \tau)$ непрерывно дифференцируемы, $x(t) \in C^1_{[a,b]}$. Рассмотрим функционал Больца в $C^1_{[a,b]}$, задаваемый соотношением

$$f(x) = \int_a^b L(t, x(t), x'(t))\, dt + l(x(a), x(b)). \tag{4}$$

◀ Выкладки и рассуждения, подобные приведенным выше, позволяют заключить, что функционал Больца в оговоренных выше условиях дифференцируем и его вариация дается выражением

$$\delta_h f(x) = \int_a^b \left[\frac{\partial L}{\partial x}h(t) + \frac{\partial L}{\partial x'}h'(t)\right] dt + \frac{\partial l}{\partial x(a)}h(a) + \frac{\partial l}{\partial x(b)}h(b). \blacktriangleright$$

2.2. Теорема Ферма

Теорема Ферма (необходимое условие локального экстремума). *Пусть функционал $f(x)$ удовлетворяет следующим условиям:*

— *определен в точке $x_0 \in \mathbf{D}$ и некоторой ее окрестности,*

— *дифференцируем в точке x_0,*

— *достигает в точке x_0 экстремального значения.*

Тогда вариация функционала в этой точке тождественно равна нулю для любого приращения h.

◀ Пусть для определенности x_0 — точка минимума функционала $f(x)$. Тогда точка $\lambda = 0$ будет точкой минимума для функции Лагранжа $\Lambda(\lambda)$. Необходимым условием экстремума функции Лагранжа является обращение в точке экстремума ее производной в нуль

$$\Lambda'(0) = \delta_h f(x_0) = 0. \blacktriangleright$$

Полученное необходимое условие экстремума носит абстрактный характер и по форме не отличается от аналогичных условий для функций одного или нескольких переменных. Наша ближайшая задача состоит в конкретизации этого (и полученных в следующем пункте) условий для интегральных функционалов.

2.3. Старшие вариации и условия старших порядков

Фактически теорема Ферма следует из возможности представить приращение функционала в окрестности возможной точки экстремума в виде

$$\Delta_h f(x) = f(x+h) - f(x) = \delta_h f(x) + r,$$

где первое слагаемое $\delta_h f(x)$ — главное в том смысле, что именно оно определяет знак приращения. Мы установили, что в точке экстремума это слагаемое обращается в нуль, и для получения дополнительной информации хотелось бы уметь уточнять, что же представляет собой остаток — функционал $r(x, h)$. В классическом анализе подобное уточнение осуществляется за счет введения старших дифференциалов. Мы поступим аналогичным образом, вводя аналог старших дифференциалов — старшие вариации.

Функция $f(x, y)$ двух переменных (x, y), являющихся элементами функционального пространства $\mathbf{B}$, называется *билинейным функционалом*, если выполнены следующие условия:

1) при каждом фиксированном значении переменной y функционал $f(x,y) = f_y(x)$ — линейный,

2) при каждом фиксированном значении переменной x функционал $f(x,y) = f_x(y)$ — линейный,

3) функционал $f(x, y)$ непрерывен по совокупности переменных (x, y).

Последнее свойство равносильно следующему

$$\exists\, C > 0 \quad \forall\, x, y \quad |f(x, y)| \leqslant C\|x\| \cdot \|y\|.$$

Билинейный функционал $f(x, y)$ называется *симметричным*, если для любых значений x и y выполняется равенство $f(x, y) = f(y, x)$.

Функционал $q(x)$ называется *квадратичным*, если найдется билинейный функционал $f(x, y)$ такой, что $q(x) = f(x, x)$. Различные билинейные функционалы могут порождать один и тот же квадратичный.

Отметим два важных свойства квадратичного функционала, следующих из его определения:

$$q(\lambda x) = \lambda^2 q(x),$$

$$\exists\, C > 0 \quad |q(x)| \leqslant C\|x\|^2.$$

Примерами квадратичных функционалов могут служить следующие.

1. Квадратичная форма в R^n

$$x = (x_1, x_2, \ldots, x_n), \quad q(x) = \sum_{i=1,j=1}^{n} a_{ij} x_i x_j.$$

2. Квадратичный терминальный функционал в $C_{[a,b]}$

$$x = x(t), \quad q(x) = \alpha x^2(a) + \beta x(a)x(b) + \gamma x^2(b).$$

3. Квадратичный терминальный функционал в $C^1_{[a,b]}$

$$x = x(t), \quad q(x) = \alpha_{11}x^2(a) + \alpha_{12}x(a)x'(a) + \alpha_{22}x'^2(a) + \beta_{11}x^2(b) + \beta_{12}x(b)x'(b) + \\ + \beta_{22}x'^2(b) + \gamma_{11}x(a)x(b) + \gamma_{12}x(a)x'(b) + \gamma_{21}x'(a)x(b) + \gamma_{22}x'(a)x'(b).$$

4. Квадратичный интегральный функционал в $C^1_{[a,b]}$

$$x = x(t), \quad q(x) = \int_a^b \left[P(t)x'^2(t) + 2R(t)x'(t)x(t) + Q(t)x^2(t)\right] dt.$$

Второй вариацией функционала $f(x)$ назовем вариацию функционала $\delta_h f(x)$ (первой вариации функционала $f(x)$), рассматриваемого как функция переменной x,

$$\delta^2_{hh} f(x) = \delta_h(\delta_h f(x)).$$

Вторая вариация является квадратичным относительно h функционалом и представляет собой аналог второго дифференциала. Как и в обычном анализе, ее использование позволяет уточнить представление приращения функционала в точке. Заметим, что существование второй вариации влечет за собой существование второй производной функции Лагранжа $\Lambda(\lambda)$ в нуле (второй лагранжевой вариации) и их тождественность. Поэтому для функции Лагранжа в точке x, где у нее существуют первая и вторая вариации, имеет место двучленная формула Тейлора

$$\Delta_{\lambda h} f(x) = f(x + \lambda h) - f(x) = \delta_h f(x)\lambda + \delta^2_{hh} f(x)\frac{\lambda^2}{2} + r_1(x, h).$$

Полагая $\lambda = 1$, приходим к равенству

$$f(x+h) - f(x) = \delta_h f(x) + \delta_{hh}^2 f(x)\frac{1}{2} + r_1(x, h),$$

где остаток $r_1(x, h)$ мал в сравнении с $\|h\|^2$, если только $\|h\| \to 0$. Искомое разложение получено.

Теорема (необходимое условие экстремума второго порядка). *Пусть функционал $f(x)$ дважды дифференцируем и достигает экстремума в точке x_0. Тогда в этой точке вторая вариация должна быть знакопостоянной — неотрицательной, если в точке x_0 достигается минимум, и неположительной, если максимум.*

◀ В точке экстремума (для определенности, минимума) должно иметь место неравенство $f(x+h) - f(x) \geqslant 0$ для всех достаточно малых h, и одновременно должна обращаться в нуль первая вариация. Поэтому (в силу малости остатка $r_1(x, h)$) знак разности $f(x+h) - f(x)$ определяет знак второй вариации. ▶

Это же разложение служит источником достаточных условий экстремума функционала — ясно, что *знакопостоянство* второй вариации определяет знакопостоянство приращения функционала для достаточно малых (настолько малых, что остаток $r_1(x, h)$ не влияет на знак приращения) $\|h\|$. Эти соображения позволяют сформулировать следующую теорему.

Теорема (достаточное условие экстремума). *Пусть функционал $f(x)$ дважды дифференцируем в точке x_0 и пусть в этой точке выполнены условия:*

— *первая вариация обращается в нуль для всех достаточно малых приращений h,*

— *вторая вариация знакопостоянна для всех достаточно малых приращений h.*

Тогда, если вторая вариация положительна, то в точке x_0 функционал $f(x)$ достигает локального минимума, если отрицательна — локального максимума.

Если вторая вариация в точке x_0 для различных приращений может принимать значения разных знаков, то в этой точке экстремума нет.

Если вторая вариация в точке x_0 для некоторых ненулевых приращений может принимать нулевое значение — требуется дополнительное исследование. Здесь возможно как наличие экстремума, так и его отсутствие.

Везде в дальнейшем мы ограничиваемся изучением необходимых условий экстремума для классических интегральных и интегро-терминальных функционалов.

Упражнения

1. Найдите норму[2)] элемента $x(t)$ в пространствах $C_{[a,b]}$ и $C^1_{[a,b]}$ соответственно:

а) $x(t) = \dfrac{\sin(n^2 t)}{n}$, $\quad n = 1, 2, 10, 100, 1000, \quad t \in [0, \pi]$;

б) $x(t) = \dfrac{\sin(nt)}{n^2}$, $\quad n = 1, 2, 10, 100, 1000, \quad t \in [0, \pi]$;

[2)] Т. е. расстояние до элемента $\Theta(t) \equiv 0$.

в) $x(t) = \dfrac{\cos(nt)}{n^2+1}$, $n = 1, 2, 10, 100, 1000$, $t \in [0, 2\pi]$;

г) $x(t) = \sin\dfrac{t}{n}$, $n = 1, 2, 10, 100, 1000$, $t \in [0, 1]$.

2. Найдите расстояние между элементами $x(t)$ и $y(t)$ в пространствах $C_{[a,b]}$ и $C^1_{[a,b]}$:

а) $x(t) = \ln t$, $y(t) = t$, $t \in [e^{-1}, e]$;

б) $x(t) = \sin 2t$, $y(t) = \sin t$, $t \in \left[0, \dfrac{\pi}{2}\right]$.

3. Найдите значения функционалов $f(x)$ в указанных точках $x(t)$:

а) $f(x) = x(0)$, $x(t) \in C_{[-1,1]}$, $x(t) = t^2$, $x(t) = 3t^3 + 2t^2 - t + 5$;

б) $f(x) = x(0) + x'(0) - x(0)\cdot x'(0)$, $x(t) \in C^1_{[-1,1]}$, $x(t) = \sin \pi t$, $x(t) = t^2$;

в) $f(x) = \int\limits_0^1 x(t)\,dt$, $x(t) \in C_{[0,1]}$, $x(t) = t^2$, $x(t) = \sin \pi t$;

г) $f(x) = \int\limits_0^1 (x^2(t) + x'^2(t))\,dt$, $x(t) \in C^1_{[0,1]}$, $x(t) = t^2$, $x(t) = \sin \pi t$.

4. Покажите, что функционал $l(x) = x(t_0)$, $x(t) \in C_{[a,b]}$, $a \leqslant t_0 \leqslant b$, линеен.

5. Докажите утверждение: если функционал $l(x)$ линеен и для любого элемента $x(t)$ выполняется

$$\lim_{\|x\|\to 0} \frac{l(x)}{\|x\|} = 0,$$

то функционал $l(x)$ тождественно равен нулю.

6. Исследуйте непрерывность следующих функционалов:

а) $f(x) = x(0)$, $x(t) \in C_{[-1,1]}$;

б) $f(x) = |x(0)|$, $x(t) \in C_{[-1,1]}$;

в) $f(x) = x(0) + x'(0) - x(0)\cdot x'(0)$, $x(t) \in C^1_{[-1,1]}$;

г) $f(x) = \int\limits_0^1 x(t)\,dt$, $x(t) \in C_{[0,1]}$;

д) $f(x) = \int\limits_0^1 [x^2(t) + x'^2(t)]\,dt$, $x(t) \in C_{[0,1]}$;

е) $f(x) = \int\limits_0^1 [x^2(t) + x'^2(t)]\,dt$, $x(t) \in C^1_{[0,1]}$;

ж) $f(x) = \int\limits_0^1 |x'(t)|\,dt$, $x(t) \in C^1_{[0,1]}$.

7. Найдите приращение функционала $f(x)$ в точке $x_0(t)$, отвечающее приращению аргумента $\Delta x = h(t)$:

а) $f(x) = x(0)$, $x(t) \in C_{[-1,1]}$, $x_0(t) = t^2$, $h(t) = 0{,}1 \sin\dfrac{5\pi t}{2}$;

б) $f(x) = x^2(0), \quad x(t) \in C_{[-1,1]}, \quad x_0(t) = t, \quad h(t) = 0{,}1 \sin \dfrac{5\pi t}{2}$;

в) $f(x) = \int\limits_0^1 x(t)\, dt, \quad x(t) \in C_{[0,1]}, \quad x_0(t) = t^2, \quad h(t) = 0{,}1 \sin \dfrac{5\pi t}{2}$;

г) $f(x) = \int\limits_0^1 [x^2(t) + x'^2(t)]\, dt, \quad x(t) \in C_{[0,1]}, \quad x_0(t) = t, \quad h(t) = 0{,}1 \sin \dfrac{5\pi t}{2}$.

8. Проверьте дифференцируемость следующих функционалов:

а) $f(x) = x(0), \quad x(t) \in C_{[-1,1]}$;

б) $f(x) = x(0), \quad x(t) \in C^1_{[-1,1]}$;

в) $f(x) = x(0) \cdot x'(0), \quad x(t) \in C^1_{[-1,1]}$;

г) $f(x) = x^2(0), \quad x(t) \in C_{[-1,1]}$;

д) $f(x) = \int\limits_0^1 x(t)\, dt, \quad x(t) \in C_{[0,1]}$;

е) $f(x) = \int\limits_0^1 |x(t)|\, dt, \quad x(t) \in C^1_{[0,1]}$.

9. Докажите, что линейный функционал дифференцируем.

10. Докажите, что функционал $f(x) = |x(0)|$, $x(t) \in C_{[0,1]}$ недифференцируем.

11. Известно, что функционал $f(x)$ — дифференцируем. Будет ли дифференцируем функционал $f^2(x)$?

12. Известно, что функция $\varphi(u)$ дифференцируема как функция переменной u. Будет ли дифференцируем функционал $f(x) = \varphi(x(0))$ в пространстве $C_{[-1,1]}$? А в $C^1_{[-1,1]}$?

13. Известно, что функция двух переменных $\varphi(u, v)$ дифференцируема. Будет ли дифференцируем функционал $f(x) = \varphi(x(-1), x(1))$ в пространстве $C_{[-1,1]}$?

14. Известно, что функция двух переменных $\varphi(u, v)$ дважды непрерывно дифференцируема. Будет ли дифференцируем функционал

$$f(x) = \int\limits_0^1 \varphi(t, x(t))\, dt, \quad x(t) \in C_{[0,1]}?$$

15. Докажите, что линейный функционал не имеет экстремумов, если только он не тождественный ноль.

16. Исследуйте необходимые условия экстремума для функционалов упражнений 11–14.

17. Докажите, что функция $x(t) \equiv 0$ является точкой минимума функционала

$$f(x) = \int\limits_0^1 (x^2 + t^2)\, dt, \quad x(t) \in C_{[0,1]}.$$

18. Исследуйте необходимые условия экстремума для функционала

$$f(x) = \int\limits_0^1 x^2(t - x)\, dt, \quad x(t) \in C_{[0,1]}.$$

Докажите, что экстремаль $x(t) \equiv 0$ не является точкой минимума этого функционала.

Ответы

1. a) $\frac{1}{n}, n$; b) $\frac{1}{n^2}, \frac{1}{n}$; c) $\frac{1}{n^2+1}, \frac{n}{n^2+1}$; d) $\sin\frac{1}{n}, \frac{1}{n^2}$. **2.** a) $e-1, e-1$; b) $1, 3$. **3.** a) $0, \quad 5$; b) $\pi, \quad 0$; c) $\frac{1}{3}, \frac{2}{\pi}$; d) $\sin\frac{23}{15}, \frac{\pi^2+1}{2}$. **6.** (a)–(d) и (f) непрерывны, (e) и (g) — нет. **7.** a) 0; b) 0; c) $\frac{0{,}04}{\pi}$; d) $\sin\frac{\pi^2}{32}+\frac{4}{125\pi^2}+\frac{41}{200} \approx 0{,}51667$. **8.** (a)–(e) — дифференцируемы, (f) — нет. **11.** Да. **12.** Да. **13.** Да. **14.** Да.

16. 11. $f(x_0)=0 \cup \delta_h f(x_0)=0$; 12. $\varphi'(x(0))=0$; 13. $\frac{\partial\varphi}{\partial u}(x(-1), x(1))=0 \cap \frac{\partial\varphi}{\partial v}(x(-1), x(1))=0$; 14. $\frac{\partial\varphi}{\partial x}(t, x)=0$.

17. Очевидно следует из задачи 16.14 и того обстоятельства, что

$$\int_0^1 (x^2+t^2)\,dt \geqslant \int_0^1 t^2\,dt = \frac{1}{3} \quad \forall x(t).$$

18. Заметим, что функция $x_0(t) \equiv 0$ является экстремалью рассматриваемого функционала, при этом $f(x_0)=0$. Однако, например,

$$f(\sqrt{t}) = -\frac{1}{15} < f(x_0)$$

и, следовательно, $x_0(t) \equiv 0$ не доставляет функционалу глобального минимума. Не является она и точкой локального минимума, так как для достаточно малых значений ε функции $x_\varepsilon(t)$, принимающие значение $1-\varepsilon$ на промежутке $[0, \varepsilon]$ и ноль на оставшейся части промежутка $(\varepsilon, 1]$, близки в $C_{[0,1]}$ к функции $x_0(t) \equiv 0$, в то же время $f(x_\varepsilon(t)) < 0$.

Глава I

ПРОСТЕЙШАЯ ЗАДАЧА КЛАССИЧЕСКОГО ВАРИАЦИОННОГО ИСЧИСЛЕНИЯ

Простейшей задачей классического вариационного исчисления обычно называют задачу отыскания экстремума интегрального функционала

$$f(x) = \int_a^b L(t, x(t), x'(t))\, dt \qquad (*)$$

на множестве дифференцируемых функций, проходящих через заданные точки на плоскости (t, x)

$$x(a) = x_a, \quad x(b) = x_b. \qquad (1)$$

Мы будем различать две задачи для интегрального функционала $(*)$ с граничными условиями (1):

— задачу о *слабом* экстремуме: функции предполагаются непрерывно дифференцируемыми на отрезке $[a, b]$, близость функций понимается как близость в пространстве $C^1_{[a,b]}$,

— задачу о *сильном* экстремуме: функции предполагаются кусочно непрерывно дифференцируемыми на отрезке $[a, b]$, близость функций понимается как близость в пространстве $C[a, b]$.

Граничные условия (1) называются условиями *закрепления концов*, а поставленные задачи носят название задач с *закрепленными концами*.

§ 1. Лемма Лагранжа и уравнение Эйлера

Рассмотрим задачу

$$f(x) = \int_a^b L(t, x(t), x'(t))\, dt \longrightarrow \text{extr}, \quad x(a) = x_a, \quad x(b) = x_b, \quad x(t) \in C^1_{[a,b]}.$$

Пусть $x_0(t)$ — точка экстремума. Слабая окрестность этой точки состоит из гладких функций $x(t)$, мало отличающихся вместе со своими производными на промежутке $[a, b]$ от $x_0(t)$ и ее производной соответственно, и принимающих на концах отрезка те же значения, что и $x_0(t)$. Следовательно, допустимые приращения

$h(t) = x(t) - x_0(t)$ — это гладкие же функции, принимающие на концах отрезка нулевые значения. Необходимое условие экстремума в этой задаче выглядит так: при любом допустимом приращении $h(t)$ функция $x_0(t)$ должна удовлетворять соотношению

$$\delta_h f(x) = \int_a^b \left[\frac{\partial L}{\partial x} h(t) + \frac{\partial L}{\partial x'} h'(t)\right] dt = 0 \quad \forall\, h(a) = h(b) = 0.$$

Проинтегрируем второе слагаемое в выражении, стоящем под знаком интеграла, по частям, предполагая, что подынтегральная функция $L(t, x(t), x'(t))$ дважды дифференцируема. Имеем

$$\int_a^b \frac{\partial L}{\partial x'} h'(t)\, dt = h(t) \frac{\partial L}{\partial x'}\bigg|_a^b - \int_a^b \frac{d}{dt}\frac{\partial L}{\partial x'} h(t)\, dt.$$

Учитывая теперь, что внеинтегральный член обращается в нуль (за счет граничных условий для $h(t)$), получаем соотношение

$$\delta_h f(x) = \int_a^b \left[\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}\right] h(t)\, dt = 0 \quad \forall\, h(a) = h(b) = 0, \tag{1}$$

которому должна удовлетворять функция $x_0(t)$.

Для дальнейших преобразований соотношения (1) нам понадобится утверждение, носящее название *основной леммы вариационного исчисления.*

Лемма Лагранжа. *Пусть функция $h(t)$ непрерывно дифференцируема на промежутке $[a, b]$ и обращается в нуль на концах этого промежутка, а функция $H(t)$ — определена и непрерывна на $[a, b]$. Пусть, кроме того, для любой функции $h(t)$, обладающей указанными свойствами, равен нулю интеграл*

$$\int_a^b H(t)h(t)\, dt = 0.$$

Тогда функция $H(t) \equiv 0 \quad \forall\, t \in [a, b]$.

◀ Допустим, что утверждение леммы неверно и на отрезке $[a, b]$ есть по крайней мере одна точка t_0, в которой функция $H(t)$ не обращается в нуль. Пусть для определенности $H(t_0) > 0$. Тогда в силу своей непрерывности функция $H(t)$ будет положительна во всех точках некоторой δ-окрестности точки t_0. Возьмем функцию $h(t)$, заданную соотношением

$$h(t) = \begin{cases} \sin^2 \dfrac{\pi}{2\delta}(t - t_0 + \delta) & \forall\, |t - t_0| < \delta, \\ 0 & \forall\, |t - t_0| > \delta. \end{cases}$$

Она удовлетворяет требованиям, предъявляемым условием леммы к функциям $h(t)$, и для нее рассматриваемый интеграл должен быть равен нулю. Однако

$$\int_a^b H(t)h(t)\, dt = \int_{t_0-\delta}^{t_0+\delta} H(t)h(t)\, dt > 0$$

в силу положительности функций $H(t)$ и $h(t)$ на промежутке $(t_0-\delta, t_0+\delta)$. Полученное противоречие доказывает лемму. ▶

Применим эту лемму к соотношению (1). В силу сделанных допущений функция

$$H(t) = \frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'}$$

непрерывна и, как это следует из леммы Лагранжа, тождественно равна нулю, $H(t) \equiv 0$. Это означает следующее: *если функция $x_0(t)$ в задаче с закрепленными концами доставляет экстремум интегральному функционалу, то она должна удовлетворять соотношению*

$$\boxed{\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial x'} = 0.} \tag{2}$$

Это обыкновенное дифференциальное уравнение относительно функции $x_0(t)$. Впервые оно было получено Эйлером и носит его имя.

Суммируя вышеизложенное, отметим, что необходимое условие слабого экстремума в задаче с закрепленными концами — это краевая задача для уравнения Эйлера. Функция $x_0(t) = \operatorname{arglocextr} f(x)$ должна быть решением этой краевой задачи.

Уравнение Эйлера получено *в предположении существования решения* рассматриваемой задачи, поэтому если у краевой задачи для уравнения Эйлера решений нет, то нет решений и у экстремальной задачи. В то же время, наличие решений у указанной краевой задачи (у этой задачи могут существовать и другие решения, не являющиеся точками экстремума интегрального функционала) не обеспечивает разрешимости исходной экстремальной задачи, и этот вопрос требует дополнительного исследования.

Решения краевой задачи для уравнения Эйлера принято называть *экстремалями* исходной вариационной задачи.

Замечание. Не следует путать *экстремали*, т. е. только «подозреваемые» на экстремум функции, с *аргументами экстремума*, т. е. с функциями, доставляющими экстремум.

§ 2. Интегрирование уравнения Эйлера

Рассмотрим несколько частных ситуаций, когда уравнение Эйлера может быть разрешено в квадратурах.

1. *Функция $L(t, x(t), x'(t)) = L(t, x(t))$ не зависит от $x'(t)$.*

Уравнение Эйлера в этом случае принимает вид

$$\frac{\partial L}{\partial x} = 0$$

и является не дифференциальным, а обычным конечным уравнением, связывающим значения переменных x и t, т. е. неявной формой задания кривой $x(t)$. Если точки (a, x_a) и (b, x_b) лежат на этой кривой, то она — экстремаль. В противном случае у исходной экстремальной задачи решений нет.

2. *Функция $L(t, x(t), x'(t)) = L(t, x'(t))$ не зависит от $x(t)$.*

В этом случае уравнение Эйлера принимает вид

$$\frac{d}{dt}\frac{\partial L}{\partial x'} = 0$$

и легко интегрируется

$$\frac{\partial L}{\partial x'} = \text{const}. \tag{1}$$

Последнее соотношение носит название *закон сохранения импульса.*

Замечание. Такое название связано со следующей простой механической аналогией: если точка массы m движется в потенциальном поле с потенциалом $U(t, x)$, то ее потенциальная энергия равна $U(t, x)$, а кинетическая — $T = m(x')^2/2$. В этом случае функция Лагранжа задается выражением $L = T - U$, а уравнение движения точки может быть получено из принципа Гамильтона — траектория движения $x(t)$ такова, что доставляет стационарное значение функционалу

$$S = \int_0^\tau L(t, x, x')\, dt,$$

называемому *действием.* Уравнение Эйлера для этого функционала — это просто закон Ньютона: $mx'' = -f$, где сила $f = U_x'$. Производная $\partial L/\partial x'$ при этом равна mx', т.е. импульсу системы.

3. *Функция* $L(t, x(t), x'(t)) = L(x(t), x'(t))$ *не зависит от* t.

Предположим сначала, что экстремаль $x(t)$ монотонна на промежутке $[a, b]$. В этом случае минимизируемый функционал может быть модифицирован следующим образом: принимая в подынтегральном выражении переменную t за функцию — $t = t(x)$ — сделаем замену

$$\int_a^b L(x, x')\, dt\Big|_{t=t(x)} = \int_{x_a}^{x_b} t'(x) L\left(x, \frac{1}{t'(x)}\right) dt.$$

Ясно, что если функция $x(t)$ доставляет экстремум интегральному функционалу, то обратная функция $t = t(x)$ доставляет экстремум функционалу

$$S = \int_{x_a}^{x_b} t'(x) L\left(x, \frac{1}{t'(x)}\right) dt.$$

Этот функционал зависит лишь от $t'(x)$ и x, поэтому уравнение Эйлера может быть записано в виде

$$\frac{d}{dt'}\left[t'(x) L\left(x, \frac{1}{t'(x)}\right)\right] = \text{const},$$

откуда, проделав обратную замену, получим первый интеграл уравнения Эйлера

$$\boxed{x' \frac{\partial L}{\partial x'} - L = \text{const}.} \tag{2}$$

Это важное для приложений соотношение носит название *закон сохранения энергии.*

Замечание. По причинам, аналогичным обсужденным в предыдущем замечании, выражение, стоящее в левой части равенства (2),

$$x' \frac{\partial T}{\partial x'} - T + U = \frac{mx'^2}{2} + U(x),$$

представляет собой полную энергию системы.

Непосредственным дифференцированием легко проверить, что соотношение (2) является первым интегралом уравнения Эйлера и в общей ситуации.

4. $L(t, x(t), x'(t))$ — *функция общего вида.*

Посмотрим, как будет выглядеть *полное* уравнение Эйлера (2) § 1 после разворачивания производной по переменной t. По правилу дифференцирования сложной функции и в предположении существования второй производной $x''(t)$ получим

$$\boxed{L_x - \frac{d}{dt}L_{x'} = L_x - L_{tx'} - x'L_{xx'} - x''L_{x'x'} = 0.} \tag{3}$$

Символом L_{uv} обозначена вторая производная функции L:

$$L_{uv} = \frac{\partial^2 L}{\partial u\, \partial v}.$$

Уравнение (3) — обыкновенное дифференциальное уравнение второго порядка относительно функции $x(t)$, если только производная $L_{x'x'}$ не равна тождественно нулю.

§ 3. Примеры

Пример 1 (задача о кратчайшем расстоянии между точками на плоскости). На плоскости переменных (t, x) среди всех гладких кривых, соединяющих заданные точки $A(a, x_a)$ и $B(b, x_b)$, найти кривую с наименьшей длиной.

◀ Решение этой задачи хорошо известно — кратчайшее расстояние между двумя точками на плоскости реализуется на отрезке прямой, соединяющей эти точки. Нам эта задача интересна с точки зрения иллюстрации изложенных выше методов.

Как известно, длина дуги кривой $x = x(t)$ на промежутке $[a, b]$ дается интегралом

$$f(x) = \int_a^b \sqrt{1 + x'^2(t)}.$$

Подынтегральная функция не зависит от $x(t)$. Из закона сохранения импульса (1) § 2

$$L_{x'} = \frac{x'}{\sqrt{1 + x'^2}} = \text{const},$$

откуда немедленно следует, что $x' = \text{const}$. Единственная экстремаль рассматриваемой задачи — прямая линия, проходящая через точки A и B. ▶

Пример 2 (пример Вейерштрасса). Функционал

$$f(x) = \int_{-1}^{1} t^2 x'^2(t)\, dt, \quad x(-1) = -1, \quad x(1) = 1$$

не имеет экстремалей *на множестве гладких* на промежутке $[-1, 1]$ функций. Убедимся в этом.

◀ Как и выше, имеет место закон сохранения импульса

$$t^2 x'(t) = \text{const},$$

откуда получаем, что экстремали рассматриваемой задачи описываются равенством

$$x(t) = \frac{C}{t} + C_1.$$

Но при $C = 0$ экстремали не удовлетворяют граничным условиям, а при $C \neq 0$ не принадлежат рассматриваемому классу функций. В классе гладких функций у рассматриваемого функционала экстремумов нет. Заметим, что это не исключает наличия решений в других классах функций. ▶

Пример 3. Вариационный принцип Ферма в геометрической оптике утверждает, что свет в неоднородной среде распространяется так, что время прохождения расстояния от точки A до точки B минимально.

Предполагая, что скорость света в воздухе линейно зависит от высоты, найдем траектории распространения света.

◀ Пусть $x(t)$ — траектория распространения света. Тогда время T прохождения светом пути от точки A до точки B дается соотношением

$$T = \int_a^b \frac{\sqrt{1+x'^2}}{kx}\,dt,$$

где k — коэффициент пропорциональности, определяющий скорость распространения света на высоте x: $v_{\text{света}}(x) = kx$. Поскольку подынтегральная функция не зависит от t, можно воспользоваться законом сохранения энергии (2) § 2. После несложных выкладок получим уравнение

$$x\sqrt{1+x'^2} = C,$$

где C — произвольная постоянная. Полагая $x'(t) = \operatorname{ctg}\varphi(t)$, приходим к соотношению

$$x(t) = \frac{C}{\sqrt{1+\operatorname{ctg}^2\varphi(t)}},$$

или

$$x(\varphi) = C\sin\varphi.$$

Дифференцируя последнее по t и учитывая, что $x'(t) = \operatorname{ctg}\varphi(t)$, находим

$$C\cos\varphi\frac{d\varphi}{dt} = \operatorname{ctg}\varphi \iff t'(\varphi) = C\sin\varphi \iff t(\varphi) = -C\cos(\varphi) + C_1.$$

Это соотношение вместе с полученным выше выражением для $x(\varphi)$ задает семейство окружностей, которые и являются искомыми траекториями. Постоянные C, C_1 могут быть определены из условия прохождения траектории через заданные точки. ▶

§ 4. Задача Больца. Условия трансверсальности

Рассмотрим задачу о слабом экстремуме функционала несколько более общего вида, чем классический функционал (*). Пусть теперь $f(x)$ — функционал Больца. Задачу

$$f(x) = \int_a^b L(t, x(t), x'(t))\,dt + l(x(a), x(b)) \longrightarrow \text{extr}, \quad x(t) \in C^1_{[a,b]},$$

назовем *задачей Больца*. Пусть $x_0(t)$ — точка экстремума. Слабая окрестность этой точки состоит из гладких функций $x(t)$, мало отличающихся вместе с производной на промежутке $[a, b]$ от $x_0(t)$ и ее производной соответственно. Допустимые приращения $h(t) = x(t) - x_0(t)$ — гладкие функции, мало отличающиеся от нуля, но в отличие от случая задачи с закрепленными концами, уже, вообще говоря, $h(a) \neq 0$, $h(b) \neq 0$. Необходимое условие экстремума в этой задаче выглядит так: при любом допустимом приращении $h(t)$ функция $x_0(t)$ должна удовлетворять соотношению

$$\delta_h f(x) = \int_a^b \left[\frac{\partial L}{\partial x}h(t) + \frac{\partial L}{\partial x'}h'(t)\right]dt + \frac{\partial l}{\partial x(a)}h(a) + \frac{\partial l}{\partial x(b)}h(b) = 0.$$

Как и в классической задаче, проинтегрируем второе слагаемое в выражении, стоящем под знаком интеграла, по частям, предполагая, что функция $L(t, x(t), x'(t))$ дважды дифференцируема. Имеем

$$\int_a^b \frac{\partial L}{\partial x'}h'(t)\,dt = h(t)\frac{\partial L}{\partial x'}\Big|_a^b - \int_a^b \frac{d}{dt}\frac{\partial L}{\partial x'}h(t)\,dt.$$

Здесь внеинтегральный член в нуль не обращается, и мы получаем соотношение

$$\int_a^b \left[L_x - \frac{d}{dt}L_{x'}\right] h(t)\,dt + h(a)\left[\frac{\partial l}{\partial x(a)} - L_{x'}\big|_{t=a}\right] + h(b)\left[\frac{\partial l}{\partial x(b)} + L_{x'}\big|_{t=b}\right] = 0, \qquad (1)$$

которому должна удовлетворять функция $x_0(t)$ для любых $h(t)$.

В частности, равенство (1) тождественно выполняется для всех $h(t)$, обращающихся в нуль на концах отрезка $[a, b]$. Как было показано выше, отсюда следует, что $x_0(t)$ — экстремаль, т. е. $x_0(t)$ удовлетворяет уравнению Эйлера (2) § 1. При этом формула (1) принимает вид

$$h(a)\left[\frac{\partial l}{\partial x(a)} - L_{x'}\big|_{t=a}\right] + h(b)\left[\frac{\partial l}{\partial x(b)} + L_{x'}\big|_{t=b}\right] = 0.$$

Рассмотрим последнее тождество на всех $h(t)$ таких, что $h(a) \neq 0$, $h(b) = 0$. Тогда

$$h(a)\left[\frac{\partial l}{\partial x(a)} - L_{x'}\big|_{t=a}\right] = 0 \quad \Longrightarrow \quad \frac{\partial l}{\partial x(a)} = L_{x'}\big|_{t=a}.$$

Аналогично, для всех $h(t)$ таких, что $h(a) = 0$, $h(b) \neq 0$, должно выполняться

$$h(b)\left[\frac{\partial l}{\partial x(b)} + L_{x'}\big|_{t=b}\right] = 0 \quad \Longrightarrow \quad \frac{\partial l}{\partial x(b)} = -L_{x'}\big|_{t=b}.$$

Таким образом, *необходимое условие слабого экстремума для функционала Больца имеет тот же характер, что и необходимое условие в классической задаче — функция, доставляющая экстремум функционалу Больца, должна являться решением краевой задачи для уравнения Эйлера* (2) § 1. *Краевые условия задаются равенствами*

$$\boxed{L_{x'}\big|_{t=a} = \frac{\partial l}{\partial x(a)}, \quad L_{x'}\big|_{t=b} = -\frac{\partial l}{\partial x(b)}} \qquad (2)$$

и называются *условиями трансверсальности*. Если терминальная часть функционала Больца отсутствует ($l(s, \tau) \equiv 0$), то условия (2) имеют совсем простой вид:

$$L_{x'}\big|_{t=a} = L_{x'}\big|_{t=b} = 0$$

и во многих прикладных задачах означают просто ортогональность экстремальных траекторий граничным вертикалям $t = a$ и $t = b$[1]. Если на одном из концов отрезка задано условие закрепления, то в качестве краевых условий для уравнения Эйлера выступают условие закрепления на одном конце и условие трансверсальности на другом.

В заключение заметим, что везде выше предполагалась непрерывная дифференцируемость функции $l(s, \tau)$, задающей терминальную часть функционала Больца.

---

[1] Очень часто интегральный функционал, подлежащий исследованию, задается как интеграл по длине дуги, что приводит к экстремальным задачам для функционалов вида

$$\int_a^b U(t, x)\sqrt{1 + x'^2}\,dt.$$

В этом случае

$$L_{x'} = U(t, x)\frac{x'}{\sqrt{1 + x'^2}}$$

и условия (2) эквивалентны тому, что $x'(a) = x'(b) = 0$.

§ 5. Простейшая задача классического вариационного исчисления. Необходимое условие Лежандра

Вернемся опять к рассмотрению простейшей задачи классического вариационного исчисления — задаче об экстремуме интегрального функционала на множестве гладких функций, удовлетворяющих условиям закрепления на концах промежутка. Как и выше, будем предполагать что функция $L(u, v, w)$ дважды непрерывно дифференцируема. Докажем, что при этом предположении интегральный функционал обладает второй вариацией.

◀ Для определения второй вариации рассмотрим приращение первой

$$\Delta_h(\delta_h f(x)) = \int_a^b \left[L_x(t, x+h, x'+h)h + L_{x'}(t, x+h, x'+h)h'\right] dt - \int_a^b \left[L_x(t, x, x')h + L_{x'}(t, x, x')h'\right] dt.$$

Объединяя интегралы и заменяя приращения производных дифференциалами (существование вторых производных у функции $L(u, v, w)$ позволяет это сделать) получим

$$\Delta_h(\delta_h f(x)) = \int_a^b \left[L_{xx}h^2 + 2L_{xx'}hh' + L_{x'x'}h'^2\right] dt + R,$$

где $R = R(x, h)$ — функционал, образовавшийся от интегрирования остаточных членов формулы Тейлора, полученных при замене приращений производных дифференциалами. Можно показать, что $R = o(\|h\|^2)$ при $\|h\| \to 0$. Отсюда заключаем, что у интегрального функционала существует вторая вариация, задаваемая выражением

$$\delta^2_{hh} f(x) = \int_a^b \left[L_{xx}h^2 + 2L_{xx'}hh' + L_{x'x'}h'^2\right] dt. \blacktriangleright$$

Воспользуемся теперь специальной формой второго слагаемого под знаком интеграла и преобразуем подынтегральное выражение.

Интегрируя $2L_{xx'}hh' = L_{xx'}(h^2)'$ по частям и учитывая условие закрепления концов, из которого следует, что допустимые приращения удовлетворяют нулевым граничным условиям, получим

$$\int_a^b L_{xx'}(h^2)'\, dt = L_{xx'}(h^2)\Big|_a^b - \int_a^b \frac{d}{dt}L_{xx'}h^2\, dt = -\int_a^b \frac{d}{dt}L_{xx'}h^2\, dt.$$

Итак, вторая вариация интегрального функционала может быть записана в виде

$$\boxed{\delta^2_{hh} f(x) = \int_a^b \left[\left(L_{xx} - \frac{d}{dt}L_{xx'}\right)h^2 + L_{x'x'}h'^2\right] dt.}$$

Теорема (необходимое условие Лежандра). *Пусть выполнены оговоренные выше условия, обеспечивающие существование второй вариации интегрального функционала, и пусть в точке x_0 функционал достигает своего экстремального значения. Тогда $\forall\, t \in [a, b]$ выполняется неравенство*

$$L_{x'x'}\big(t, x_0(t), x_0'(t)\big) \geqslant 0,$$

если x_0 — точка минимума, и неравенство

$$L_{x'x'}\big(t, x_0(t), x_0'(t)\big) \leqslant 0,$$

если x_0 — точка максимума.

◀ Пусть для определенности x_0 — точка минимума. Как было установлено, вторая вариация в точке минимума должна быть неотрицательна. Допуская, что $L_{x'x'}(t, x_0(t), x_0'(t)) < 0$ в некоторой точке τ отрезка $[a, b]$, заключаем, что в силу непрерывности $L_{x'x'}$ это неравенство остается справедливым и в некоторой окрестности точки τ. Но если мы теперь в качестве $h(t)$ возьмем функцию, равную нулю вне этой окрестности, маленькую по модулю внутри этой окрестности и с большой производной[2], то в выражении для второй вариации основной вклад в интеграл будет вносить слагаемое $L_{x'x'}h'^2$. В силу сделанного допущения об отрицательности $L_{x'x'}$ это приводит к отрицательности второй вариации в точке минимума, чего не может быть. ▶

Упражнения

1. Запишите уравнение Эйлера и найдите экстремали функционалов в $C^1_{[a,b]}$:

а) $f(x) = \int_1^2 (x'^2 - 2tx)\, dt, \quad x(1) = 0, \quad x(2) = -1;$

б) $f(x) = \int_1^3 (3t - x)x\, dt, \quad x(1) = 1, \quad x(3) = \frac{9}{2};$

в) $f(x) = \int_0^{2\pi} (x'^2 - x^2)\, dt, \quad x(0) = 1, \quad x(2\pi) = 1;$

г) $f(x) = \int_1^2 (x'^2 + 2xx' + x^2)\, dt, \quad x(1) = 1, \quad x(2) = 0;$

д) $f(x) = \int_0^1 \sqrt{x(1 + x'^2)}\, dt, \quad x(0) = x(1) = \frac{\sqrt{2}}{2};$

[2] В качестве такой функции можно взять, например, функцию

$$h(t) = \begin{cases} \sqrt{\delta}\, \sin^2 \dfrac{\pi}{2\delta}(t - t_0 + \delta) & \forall\, |t - t_0| < \delta, \\ 0 & \forall\, |t - t_0| \geqslant \delta \end{cases}$$

при достаточно малом δ.

е) $f(x) = \int_0^1 (x'^2 + t)\,dt, \quad x(0) = 1, \quad x(1) = 2;$

ж) $f(x) = \int_0^1 (x'^2 + x^2)\,dt, \quad x(0) = 0, \quad x(1) = 1.$

2. Запишите уравнение Эйлера, условия трансверсальности и найдите экстремали функционалов в $C^1_{[a,b]}$:

а) $f(x) = \int_0^1 (x'^2 + x^2)\,dt;$

б) $f(x) = \int_0^{2\pi} (x'^2 - x^2)\,dt;$

в) $f(x) = \int_0^1 (tx' + x'^2)\,dt;$

г) $f(x) = \int_0^1 x'^2\,dt + x'^2(0) - 2x^2(1);$

д) $f(x) = \int_0^1 (x'^2 + x^2)\,dt - 2\,\mathrm{sh}\,1 \cdot x(1);$

е) $f(x) = \int_0^{\pi} (x'^2 + x^2 - 4x\sin t)\,dt + 2x^2(0) + 2x(\pi) - x^2(\pi);$

ж) $f(x) = \int_0^3 4x'^2x^2\,dt + x^4(0) - 8x(3);$

з) $f(x) = \int_0^1 e^{t+1}(x'^2 + 2x^2)\,dt + 2x(1)\cdot(x(0) + 1);$

и) $f(x) = \int_0^1 e^x x'^2\,dt + 4e^{x(0)} + 32e^{-x(1)};$

к) $f(x) = \int_0^{\pi/2} (x'^2 - x^2)\,dt + x^2(0) - x^2\left(\frac{\pi}{2}\right) + 4x\left(\frac{\pi}{2}\right).$

3. Найдите экстремали функционалов и проверьте выполнение условия Лежандра:

а) $f(x) = \int_0^2 (x'^4 + x'^2)\,dt, \quad x(0) = 1, \quad x(2) = 5;$

б) $f(x) = \int_{-1}^1 (t^2x'^2 + 12x^2)\,dt, \quad x(-1) = -1, \quad x(1) = 1;$

в) $f(x) = \int_0^1 (x'^2 - xx'^3)\,dt, \quad x(0) = 0, \quad x(1) = 0;$

г) $f(x) = \int\limits_0^a x'^3\,dt, \quad x(0) = 0, \quad x(a) = b;$

д) $f(x) = \int\limits_\alpha^\beta \varphi(x)\sqrt{1 + x'^2}\,dt, \quad x(\alpha) = x_\alpha, \quad x(\beta) = x_\beta, \quad \varphi(x) > 0.$

Ответы

1. а) $x'' + t = 0, \quad x(t) = \dfrac{t(1-t^2)}{6}$; б) не существует; $3t - 2x = 0$, $x(t) = 1{,}5t$ — не удовлетворяет граничным условиям; в) $x'' + x = 0, \quad x(t) = \cos t + \alpha \cdot \sin t, \quad \forall\, \alpha \in \mathbb{R}$;
г) $x'' - x = 0, \quad x(t) = \dfrac{\operatorname{sh}(2-t)}{\operatorname{sh} 1}$; д) $1 + x'^2 = C \cdot x$, $x_1(t) = \dfrac{1 + (3+2\sqrt{2})(2x-1)^2}{4(\sqrt{2}+1)}$,
$x_2(t) = \dfrac{1 + (3-2\sqrt{2})(2x-1)^2}{4(\sqrt{2}-1)}$; е) $x'' = 0, \quad x = t + 1$; ж) $x'' - x = 0, \quad x(t) = \dfrac{\operatorname{sh} t}{\operatorname{sh} 1}$.

2. а) $x'' - x = 0, \quad x'(0) = 0, \quad x'(1) = 0, \quad x(t) \equiv 0$; б) $x'' + x = 0, \quad x'(0) = 0,$
$x'(2\pi) = 0, \quad x(t) \equiv 0$; в) $x'' = -0{,}5, \quad x'(0) = 0, \quad x'(1) = -0{,}5, \quad x(t) = -\dfrac{1}{4}t^2 + \alpha\ \forall\, \alpha \in \mathbb{R}$;
г) $x'' = 0, \quad x'(0) = x(0), \quad x'(1) = x(1), \quad x(t) \equiv 0$; д) $x'' - x = 0, \quad x'(0) = 0, \quad x'(1) = \operatorname{sh} 1,$
$x(t) = \operatorname{ch} t$; е) $x'' - x = -2\sin t, \quad x'(0) = 2x(0), \quad x'(\pi) = -1 - x(\pi), \quad x(t) = \dfrac{1}{3} \cdot e^{-t} + \sin t.$
ж) $x''x = -x'^2, \quad 2x'(0)x^2(0) = 4x^3(0), \quad x'(3)x^2(3) = 1, \quad x_1(t) = \sqrt{t+1}, \quad x_2(t) = \sqrt[6]{\dfrac{4}{3}t^2}$;
з) $x'' + x' - 2x = 0, \quad x'(0)e = x(1), \quad x'(1)e^2 = x(0) + 1, \quad x(t) = -\dfrac{e^t}{e^3 + 1}$;
и) $2x'' = -x'^2, \quad x'(0) = 2, \quad x'(1) = 16e^{-x(1)}, \quad x(t) = 2\ln(1+t)$; к) $x'' + x = 0,$
$x'(0) = x(0), \quad x'\left(\dfrac{\pi}{2}\right) = x\left(\dfrac{\pi}{2}\right) - 2, \quad x(t) = \sin t + \cos t.$

3. а) $x(t) = 2t + 1$, условие $L_{x'x'} \geqslant 0$ выполнено; б) $x(t) = t^3$, условие $L_{x'x'} \geqslant 0$ не выполнено;
в) $x(t) \equiv 0$, условие $L_{x'x'} \geqslant 0$ выполнено; г) $x(t) = \dfrac{b}{a}t$, условие $L_{x'x'} \geqslant 0$ выполнено;
д) уравнение Эйлера имеет вид $\varphi = c \cdot \sqrt{1 + x'^2}$, условие $L_{x'x'} = \dfrac{\varphi}{(1 + x'^2)^{3/2}} \geqslant 0$ выполнено.

Глава LI

ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ С ОГРАНИЧЕНИЯМИ. ПРИНЦИП ЛАГРАНЖА

§ 1. Принцип Лагранжа для задач с ограничениями-равенствами

Пусть в функциональном пространстве $\mathbf{B}$ заданы функционалы $f(x), g_1(x), \ldots, g_m(x)$, определенные на одном и том же множестве $\mathbf{D}$. Рассмотрим задачу

$$f(x) \longrightarrow \text{extr}, \quad x \in \mathbf{D}, \tag{1}$$

при дополнительном требовании, что x удовлетворяет еще и условиям $g_i(x) = 0$, $i = 1, 2, \ldots, m$. Ясно, что соображения гл. L при исследовании этой задачи оказываются уже непригодными. Наличие дополнительных ограничений на допустимые значения переменной x приводит к тому, что неравенство $\Delta_h f(x) \geqslant 0$ будет выполняться не для *любых* приращений h, а только для тех, которые не выводят переменную x за пределы множества

$$\mathbf{G} = \{x \in \mathbf{D}:\ g_i(x) = 0,\ i = 1, 2, \ldots, m\}, \tag{2}$$

т. е. для таких h, которые для допустимого x удовлетворяют условиям $g_i(x + h) = 0$, $i = 1, 2, \ldots, m$. Отсюда вытекает, что в точке экстремума и вариация функционала $f(x)$ будет обращаться в нуль не тождественно для *любых* приращений аргумента, а только для *описываемых указанными соотношениями*. Заметим, что если функционалы $g_i(x + h) = 0$, $i = 1, 2, \ldots, m$, дифференцируемы, то указанное ограничение влечет за собой равенство нулю вариаций этих функционалов для *допустимых приращений* h (из $g_i(x + h) = g_i(x) = 0$ заключаем, что $g_i(x + h) - g_i(x) = 0$).

Оказывается, справедливо и обратное утверждение.

Лемма (о допустимых приращениях аргумента в задаче с ограничениями-равенствами). *Если функционал $f(x)$ достигает экстремума в задаче* (1) *с ограничениями* (2) *на элементе x_0, то множество допустимых приращений аргумента в этой точке описывается соотношениями $\delta_h g_i(x_0) = 0$, $i = 1, 2, \ldots, m$.*

Соотношения $\delta_h g_i(x_0) = 0$, $i = 1, 2, \ldots, m$, удовлетворяются тождественно при любом h, если x_0 — стационарная точка функционалов $g_i(x)$. Этот случай мы в дальнейшем из рассмотрения исключим, предполагая, что точка экстремума функционала $f(x)$ для функционалов $g_i(x)$ стационарной не является.

Используя лемму, заключаем, что

если x_0 — точка экстремума в задаче (1)–(2), то в этой точке обращается в нуль вариация функционала $f(x)$ на всех приращениях h таких, что $\delta_h g_i(x_0) = 0$, $i = 1, 2, \ldots, m$.

Для дальнейшего нам понадобится еще один простой факт из функционального анализа.

Теорема (о линейной зависимости линейных функционалов). *Пусть $f(x), l_1(x), l_2(x), \ldots, l_m(x)$ — линейные в $\mathbf{B}$ функционалы, обладающие тем свойством линейной независимости. Пусть для всякого элемента $x \in \mathbf{B}$, на котором все функционалы $l_j(x) = 0$, $j = 1, 2, \ldots, m$, функционал $f(x)$ также обращается в нуль. Тогда существуют постоянные $\lambda_1, \lambda_2, \ldots, \lambda_m$ такие, что функционал $f(x)$ представим в виде линейной комбинации функционалов $l_j(x)$ с коэффициентами λ_j*

$$f(x) \equiv \lambda_1 l_1(x) + \lambda_2 l_2(x) + \ldots + \lambda_m l_m(x).$$

◀ В силу линейной независимости функционалов $l_j(x)$ ни один из них не является тождественно нулевым и, следовательно, существует совокупность из m векторов пространства $\mathbf{B}$, такая, что определитель матрицы $(l_{ij}) = (l_i(x_j))$ не обращается в нуль. При этом для любого элемента $x \in \mathbf{B}$ можно указать постоянные $k_1, k_2, \ldots, k_m$, обладающие тем свойством, что на элементе $x_0 = x + k_1 x_1 + k_2 x_2 + \ldots + k_m x_m$ все функционалы $l_j(x)$, $j = 1, 2, \ldots, m$, обращаются в нуль. Действительно, если это так, то имеют место соотношения

$$\begin{aligned}
l_1(x_0) &= l_1(x) + k_1 l_1(x_1) + k_2 l_1(x_2) + \ldots + k_m l_1(x_m) = 0,\\
l_2(x_0) &= l_2(x) + k_1 l_2(x_1) + k_2 l_2(x_2) + \ldots + k_m l_2(x_m) = 0,\\
l_3(x_0) &= l_3(x) + k_1 l_3(x_1) + k_2 l_3(x_2) + \ldots + k_m l_3(x_m) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
l_m(x_0) &= l_m(x) + k_1 l_m(x_1) + k_2 l_m(x_2) + \ldots + k_m l_m(x_m) = 0,
\end{aligned}$$

представляющие собой систему линейных алгебраических уравнений относительно постоянных k_i. Поскольку определитель этой системы не равен нулю, она однозначно разрешима и в качестве указанных значений k_i могут быть взяты решения этой системы. По правилу Крамера

$$k_i = \frac{\Delta_i}{\Delta},$$

где определители Δ_i и Δ даются выражениями

$$\Delta_i = \begin{vmatrix}
l_{11} & l_{12} & \ldots & l_1(x) & \ldots & l_{1m}\\
l_{21} & l_{22} & \ldots & l_2(x) & \ldots & l_{2m}\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
l_{m1} & l_{m2} & \ldots & l_m(x) & \ldots & l_{mm}
\end{vmatrix}, \quad \Delta = \begin{vmatrix}
l_{11} & l_{12} & \ldots & l_{1i} & \ldots & l_{1m}\\
l_{21} & l_{22} & \ldots & l_{2i} & \ldots & l_{2m}\\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\
l_{m1} & l_{m2} & \ldots & l_{mi} & \ldots & l_{mm}
\end{vmatrix}$$

(здесь $l_{ij} = l_i(x_j)$). Раскладывая определитель Δ_i по столбцу с номером i, получим выражение для k_i

$$k_i = \sum_{j=1}^{m} \frac{\Delta_{ij}}{\Delta} l_j(x) = \sum_{j=1}^{m} k_{ij} l_j(x),$$

где Δ_{ij} — алгебраическое дополнение к элементу l_{ij} определителя Δ_i.

По условию теоремы (все функционалы $l_i(x)$ обращаются в нуль на элементе $x_0 = x + \sum_{j=1}^{m} k_j x_j$) функционал f обращается в нуль на элементе x_0

$$f(x_0) = f(x) + \sum_{i=1}^{m} k_i f(x_i) = f(x) + \sum_{i=1}^{m} \sum_{j=1}^{m} k_{ij} l_j(x) f(x_i) =$$

$$= f(x) + \sum_{j=1}^{m} l_j(x) \left[\sum_{i=1}^{m} k_{ij} f(x_i) \right] = 0.$$

Из последнего соотношения заключаем, что

$$f(x) = -\sum_{j=1}^{m} l_j(x) \left[\sum_{i=1}^{m} k_{ij} f(x_i) \right] = \sum_{j=1}^{m} \lambda_j l_j(x)$$

с постоянными $\lambda_j = -\sum_{i=1}^{m} k_{ij} f(x_i)$. ▶

Определение. *Функционалом Лагранжа, ассоциированным с экстремальной задачей* (1)–(2), *будем называть функционал, задаваемый соотношением*

$$\boxed{F(x, \lambda_1, \lambda_2, \ldots, \lambda_m) = f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x) + \ldots + \lambda_m g_m(x);} \qquad (3)$$

постоянные $\lambda_1, \lambda_2, \ldots, \lambda_m$ называются *множителями Лагранжа*.

Теперь мы можем сформулировать необходимое условие экстремума в задаче (1) с ограничениями (2).

Теорема (необходимое условие экстремума при наличии ограничений-равенств). *Пусть x_0 — точка экстремума в задаче*

$$f(x) \longrightarrow \text{extr}, \quad x \in G = \{x \in D:\ g_i(x) = 0,\ i = 1, 2, \ldots, m\},$$

где функционалы $f(x), g_1(x), \ldots, g_m(x)$ дифференцируемы, функционалы $g_1(x), \ldots, g_m(x)$ линейно независимы и точка x_0 не является стационарной ни для одного из функционалов $g_1(x), \ldots, g_m(x)$. Тогда существуют однозначно определяемые постоянные $\lambda_1, \lambda_2, \ldots, \lambda_m$ такие, что x_0 — стационарная точка функционала Лагранжа (3).

◀ Из леммы вытекает, что если x_0 — точка экстремума в рассматриваемой задаче, то в этой точке для всех приращений h, обращающих в нуль вариации функционалов-ограничений $g_i(x)$, должна обращаться в нуль и вариация функционала $f(x)$. Но все указанные вариации — линейные по h функционалы, поэтому из теоремы о пропорциональности линейных функционалов следует, что вариация $\delta_h f(x_0)$ является линейной комбинацией вариаций $\delta_h g_i(x_0)$ с некоторыми однозначно определяемыми коэффициентами $\mu_1, \mu_2, \ldots, \mu_m$:

$$\delta_h f(x_0) = \mu_1 \delta_h g_1(x_0) + \mu_2 \delta_h g_2(x_0) + \ldots + \mu_m \delta_h g_m(x_0).$$

Учитывая, что вариация обладает свойством аддитивности, перепишем последнее соотношение в виде

$$\delta_h\left[f(x_0) + \lambda_1 g_1(x_0) + \lambda_2 g_2(x_0) + \ldots + \lambda_m g_m(x_0)\right] = 0,$$

где $\lambda_i = -\mu_i$.

Последнее равенство означает, что точка x_0 является стационарной точкой функционала Лагранжа (3) с указанными значениями постоянных. ►

Доказанная теорема замечательным образом сводит исследование подозреваемых на экстремум точек в задаче на *экстремум с ограничениями* для функционала $f(x)$ к исследованию *безусловных* стационарных точек функционала Лагранжа $F(x)$, к которым следует присоединить и стационарные точки функционалов-ограничений.

§ 2. Ограничения-равенства в задаче Больца. Классическая изопериметрическая задача

Применим развитую в § 1 теорию к построению необходимых условий слабого экстремума в задаче Больца с ограничениями-равенствами, задаваемыми интегральными и терминальными функционалами.

Пусть $f(x)$ — функционал Больца, $g_i(x)$, $i = 1, 2, \ldots, m$, — интегральные, а $r_j(x)$, $j = 1, 2, \ldots, k$, — терминальные функционалы, определенные в $C^1_{[a,b]}$ и удовлетворяющие там условиям, обеспечивающим справедливость теоремы о множителях Лагранжа. Функция Лагранжа для задачи

$$f(x) = \int_a^b L(t, x, x')\,dt + r(x(a), x(b)) \longrightarrow \text{extr}, \tag{1}$$

$$g_i(x) = \int_a^b G_i(t, x, x')\,dt = c_i, \quad i = 1, 2, \ldots, m, \tag{2}$$

$$r_j(x) = r_j(x(a), x(b)) = p_j, \quad j = 1, 2, \ldots, k, \tag{3}$$

имеет вид

$$F(x, \lambda_1, \lambda_2, \ldots, \lambda_m, \mu_1, \mu_2, \ldots, \mu_k) = \int_a^b \left[L + \sum_{i=1}^m \lambda_i G_i\right] dt + r + \sum_{j=1}^m \mu_j r_j$$

и является функционалом Больца

$$F(x, \lambda_1, \lambda_2, \ldots, \lambda_m, \mu_1, \mu_2, \ldots, \mu_k) = \int_a^b F_{\text{int}}\,dt + r_{\text{term}},$$

интегральная часть которого определяется функцией

$$F_{\text{int}}(t, x, \lambda_1, \lambda_2, \ldots, \lambda_m) = L(t, x, x') + \sum_{i=1}^m \lambda_i G_i(t, x, x'),$$

а терминальная — функцией

$$r_{\text{term}}(x(a), x(b), \mu_1, \mu_2, \ldots, \mu_k) = r(x(a), x(b)) + \sum_{j=1}^{m} \mu_j r_j(x(a), x(b)).$$

Условия стационарности функционала Больца, полученные выше, представляют собой краевую задачу для уравнения

$$\frac{\partial F_{\text{int}}}{\partial x} - \frac{d}{dt}\frac{\partial F_{\text{int}}}{\partial x'} = 0,$$

дополненную условиями трансверсальности

$$\left.\frac{\partial F_{\text{int}}}{\partial x'}\right|_{t=a} = \frac{\partial r_{\text{term}}}{\partial x(a)}, \qquad \left.\frac{\partial F_{\text{int}}}{\partial x'}\right|_{t=b} = -\frac{\partial r_{\text{term}}}{\partial x(b)}.$$

Эти условия, вместе с исходными ограничениями (2)–(3), в принципе, позволяют найти все фигурирующие в задаче константы и указать подозреваемые на экстремум функции.

Частным случаем рассматриваемой задачи является классическая ***изопериметрическая задача*** об экстремуме интегрального функционала при наличии ограничений-равенств, задаваемых интегральными же функционалами. Одной из наиболее известных ***изопериметрических задач***, давшей название всему классу, является задача Дидоны.

Найти гладкую линию заданной длины, закрепленную в двух точках прямой и ограничивающую вместе с отрезком прямой наибольшую площадь.

◀ Выбирая в качестве фигурирующей в условии прямой на плоскости переменных (t, x) ось абсцисс и располагая точки закрепления концов искомой кривой симметрично относительно начала координат, приходим к следующей формализации задачи Дидоны.

$$\int_{-a}^{a} x(t)\, dt \longrightarrow \max$$

при условии

$$\int_{-a}^{a} \sqrt{1 + x'^2(t)}\, dt = l, \qquad x(-a) = x(a) = 0.$$

В дальнейшем будем считать, что длина искомой кривой удовлетворяет условию $2a < l < \pi a$[1]. Интегральная часть функционала Лагранжа дается соотношением

$$F_{\text{int}} = x(t) + \lambda\sqrt{1 + x'^2(t)}$$

и не зависит от t. Поэтому имеет место закон сохранения энергии

$$x'(F_{\text{int}})'_x - F_{\text{int}} = \frac{\lambda x'^2}{\sqrt{1 + x'^2(t)}} - x - \lambda\sqrt{1 + x'^2(t)} = C.$$

[1] Левое неравенство естественно — невозможно соединить точки, расстояние между которыми равно $2a$, линией, длина которой меньше $2a$. Правое же позволяет избежать некоторых технических сложностей, возникающих из-за того, что линия длины, большей чем πa, может пересекаться вертикалями (прямыми, параллельными оси ординат) более чем в одной точке.

Элементарные выкладки приводят к уравнению первого порядка

$$(C+x)\sqrt{1+x'^2(t)} = -\lambda,$$

интегрируя которое, получаем

$$\frac{dx}{dt} = \frac{\sqrt{\lambda-(C+x)^2}}{C+x} \quad \Longrightarrow$$

$$\Longrightarrow \quad (x+C)^2 + (t+C_1)^2 = \lambda.$$

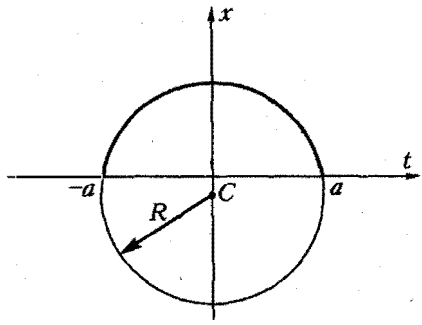


Рис. 1

Граничные условия дают

$$C_1 = 0, \quad C = \pm\sqrt{\lambda - a^2},$$

т. е. экстремали — дуги окружностей, центры которых лежат на оси ординат (рис. 1). Знак ± показывает, что симметричная относительно оси абсцисс дуга доставляет функционалу то же значение. ▶

В заключение отметим, что оставшаяся неопределенной величина множителя Лагранжа может быть легко найдена, если использовать тот факт, что длина искомой дуги известна.

§ 3. Необходимые условия экстремума в задаче со свободно скользящими концами

Рассмотрим интегральный функционал в пространстве гладких на отрезке $[a, b]$ функций и предположим, что один из концов $x(t)$ закреплен, а другой скользит вдоль линии, заданной явным $x = \varphi(t)$ или неявным $\psi(t, x) = 0$ образом (рис. 2).

Поставим задачу поиска экстремума этого функционала с указанным дополнительным ограничением на поведение искомой функции. Заметим, что тогда соответствующий предел интегрирования зависит от положения скользящего конца и подлежит определению вместе с функцией $x(t) \in C^1_{[a,b]}$.

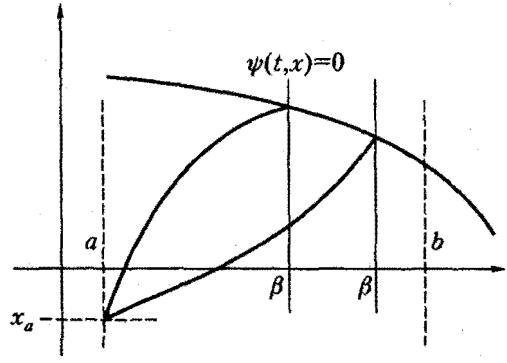


Рис. 2

Пусть, для определенности, подвижным будет правый конец искомой траектории. Применим общие соображения § 2 для получения необходимых условий экстремума, предварительно заметив, что *если $x_0(t)$ — решение задачи, то в точке β_0 — абсциссе правого конца, определяемой решением, производная функции $f(x_0; \beta)$ по параметру β должна обращаться в нуль* — обычное необходимое условие экстремума числовой функции одной переменной.

Итак, исследованию подлежит задача, носящая название *задачи с подвижным правым концом* и состоящая в следующем

$$f(x) = \int_a^\beta L(t, x, x')\,dt \longrightarrow \text{extr}, \quad x(a) = x_a, \quad \psi(\beta, x(\beta)) = 0.$$

Рассматривая краевое условие на правом конце как терминальное ограничение-равенство, построим функционал Лагранжа и запишем условие стационарности этого функционала

$$F_{\text{int}} = L(t, x, x'), \quad r_{\text{term}} = \mu\psi(\beta, x(\beta)).$$

Уравнение Эйлера в этой задаче такое же, как и в задаче с закрепленными концами. Интерес представляют условия трансверсальности на правом конце

$$L_{x'}\big|_{t=\beta} = -\mu\frac{\partial}{\partial x(\beta)}\psi(\beta, x(\beta)).$$

Условие стационарности функционала Лагранжа по переменному правому концу β приводит к равенству

$$\left[L(t, x, x') + \mu\left(\frac{\partial\psi(t, x(t))}{\partial t} + x'(t)\frac{\partial\psi(t, x(\beta))}{\partial x(\beta)}\right)\right]\Bigg|_{t=\beta} = 0,$$

которое вместе с предыдущим позволяет исключить множитель Лагранжа и получить *условие трансверсальности* в задаче с подвижным правым концом

$$\left[L(t, x, x') - x'(t)L_{x'}(t, x, x')\right]\big|_{t=\beta} = \left.\frac{\psi_t}{\psi_x}L_{x'}\right|_{t=\beta}.$$

В частном случае, когда уравнение связи на правом конце задано явно $x(\beta) = \varphi(\beta)$, последнее соотношение принимает вид

$$\left[L(t, x, x') - x'(t)L_{x'}(t, x, x')\right]\big|_{t=\beta} = -\left.\frac{d\varphi(t)}{dt}L_{x'}\right|_{t=\beta}.$$

Как и выше, для функционалов вида

$$f(x) = \int_a^\beta v(t, x)\sqrt{1 + x'^2}\,dt$$

условие трансверсальности означает ортогональность экстремали условию $\psi(t, x) = 0$ (рис. 3).

В качестве примера рассмотрим задачу о поиске кратчайшего расстояния от заданной точки до заданной кривой на плоскости.

◀ Будем предполагать, что линия задана явным уравнением $x = \varphi(t)$. Не ограничивая общности наших рассмотрений, поместим точку в начало координат. Тогда получим следующую вариационную задачу:

найти функцию, доставляющую минимум функционалу

$$\int_0^\beta \sqrt{1+x'^2(t)}\,dt$$

и удовлетворяющую условиям

$$x(0)=0, \quad x(\beta)=\varphi(\beta).$$

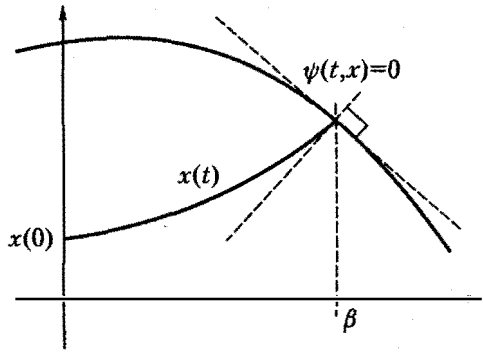


Рис. 3

Уравнение Эйлера в этой задаче

$$L_x-\frac{d}{dt}L_{x'}=0 \implies \frac{x'}{\sqrt{1+x'^2(t)}}=\text{const}$$

приводит к прямолинейным экстремалям $x'(t)=\text{const}=C \implies x(t)=Ct+k$. Граничное условие на левом конце дает $k=0$, а условие трансверсальности превращается в условие

$$\sqrt{1+C^2}-\frac{C}{\sqrt{1+C^2}}C=-\frac{C}{\sqrt{1+C^2}}\varphi'(\beta),$$

из которого получаем

$$C\varphi'(t)\big|_{t=\beta}=-1.$$

Последнее означает, что экстремаль $x(t)=Ct$ должна пересекать линию $x=\varphi(t)$ под прямым углом. ▶

Упражнения

1. Постройте функционал Лагранжа, ассоциированный с рассматриваемой экстремальной задачей, и найдите экстремали в $C^1_{[a,b]}$:

а) $f(x)=\int_0^\pi x'^2\,dt \Rightarrow \min, \quad \int_0^\pi x^2\,dt=1, \quad x(0)=0, \quad x(\pi)=0;$

б) $f(x)=\int_0^1 x'^2\,dt \Rightarrow \min, \quad \int_0^1 x(t)\,dt=3, \quad x(0)=1, \quad x(1)=6;$

в) $f(x)=\int_0^1 (x'^2+t^2)\,dt \Rightarrow \min, \quad \int_0^1 x^2(t)\,dt=2, \quad x(0)=0, \quad x(1)=0;$

г) $f(x)=\int_0^1 x'^2\,dt \Rightarrow \min, \quad \int_0^1 t\cdot x(t)\,dt=0, \quad x(0)=0, \quad x(1)=1;$

д) $f(x)=\int_0^1 x'^2\,dt \Rightarrow \min, \quad \int_0^1 x(t)\,dt=1, \quad \int_0^1 t\cdot x(t)\,dt=1, \quad x(0)=0, \quad x(1)=0.$

2. Найдите экстремали функционалов в следующих задачах:

а) $f(x)=\int_0^\xi x'^2\,dt \Rightarrow \text{extr}, \quad x(0)=0, \quad x(\xi)=-\xi-1;$

б) $f(x)=\int\limits_0^{\xi} x'^2\,dt \Rightarrow \text{extr}, \quad x(0)=0, \quad x(\xi)=\dfrac{2}{1-\xi};$

в) $f(x)=\int\limits_0^{\xi} \sqrt{1+x'^2}\,dt \Rightarrow \text{extr}, \quad x(0)=0, \quad x(\xi)=\dfrac{1}{\xi^2};$

г) $f(x)=\int\limits_0^{\xi} (x'^2+t^2)\,dt \Rightarrow \text{extr}, \quad x(0)=0, \quad x(\xi)=1;$

д) $f(x)=\int\limits_0^{\xi} \dfrac{\sqrt{1+x'^2}}{x}\,dt \Rightarrow \text{extr}, \quad x(0)=1, \quad x(\xi)=\xi-1.$

3. Запишите условие трансверсальности для задачи

$$f(x)=\int\limits_{t_0}^{\xi} \varphi(t,x)\sqrt{1+x'^2}\,dt \Rightarrow \text{extr}, \quad \varphi(t,x)\neq 0, \quad x(t_0)=x_0, \quad x(\xi)=\psi(\xi),$$

и поясните его геометрический смысл.

4. Запишите условие трансверсальности для задачи

$$f(x)=\int\limits_{t_0}^{\xi} \varphi(t,x)e^{\operatorname{arctg} x'}\sqrt{1+x'^2}\,dt \Rightarrow \text{extr}, \quad \varphi(t,x)\neq 0, \quad x(t_0)=x_0, \quad x(\xi)=\psi(\xi),$$

и поясните его геометрический смысл.

5. Найдите расстояние между параболой $x=t^2$ и прямой $t-x=5$.

6. Найдите расстояние от точки $M_0(1,0)$ до эллипса $4t^2+9y^2=36$.

7. Найдите расстояние от прямой $t+x=4$ до окружности $t^2+x^2=1$.

Ответы

1. а) $x_k(t)=\pm\sqrt{\dfrac{2}{\pi}}\sin kt$, $k=1,2,\ldots$. В силу $\int\limits_0^{\pi} x_k'^2\,dt=k^2$, принимаем $k=1$;

б) $x(t)=3t^2+2t+1$; в) $x_m(t)=\pm 2\sin \pi m t$, $m=0,1,\ldots$; г) $x(t)=\dfrac{5}{2}t^3-\dfrac{3}{2}t$;

д) $x(t)=60t^3-96t^2+36t$. **2.** а) $\xi=1$, $x(t)=-2t$; б) $\xi=0{,}5$, $x(t)=\pm 4t$; в) $\xi=2^{1/6}$, $x(t)=\dfrac{t}{\sqrt{2}}$; г) решений нет; д) $\xi=2$, $x(t)=\sqrt{2-(t-1)^2}$. **3.** $x'(\xi)\cdot\psi'(\xi)=-1$ — условие ортогональности экстремали к линии $\psi(t)$. **4.** $\dfrac{\psi'(\xi)-x'(\xi)}{1+\psi'(\xi)\cdot x'(\xi)}=-1$ — экстремаль пересекает линию $\psi(t)$ под углом $\dfrac{\pi}{4}$. **5.** $\int\limits_{\xi}^{\eta}\sqrt{1+x'^2}\,dt$, $x(\xi)=\xi^2$, $x(\eta)=\eta-5$, $\xi=\dfrac{1}{2}$, $\eta=\dfrac{23}{8}$, $x(t)=-t+\dfrac{3}{4}$, $d=\dfrac{19\sqrt{2}}{8}$. **6.** $d=\dfrac{4}{\sqrt{2}}$. **7.** $d=2\sqrt{2}-1$.

Глава LII

ВЕКТОРНЫЕ ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ

Естественным обобщением рассмотренных выше вариационных задач являются задачи для функционалов, зависящих от векторных функций скалярной переменной. К ним приводят, например, задачи, поставленные для кривых в $\mathbb{R}^n$, которые заданы параметрическими уравнениями. Аналоги простейшей задачи классического вариационного исчисления, задачи Больца и задачи с ограничениями-равенствами исследуются буквально так же, как и в одномерном случае — принцип Лагранжа переносится на многомерную ситуацию дословно, вплоть до обозначений.

Новые задачи в многомерном случае возникают за счет расширения типа накладываемых на отыскиваемые кривые ограничений — голономных (конечных) и дифференциальных связей. Замечательным обстоятельством является то, что и эти задачи могут быть исследованы с помощью принципа Лагранжа. В этой главе мы остановимся на *постановке* основной векторной задачи — задачи Лагранжа — и *технике* исследования необходимых условий экстремума для нее.

Начнем рассмотрение указанной задачи с краткого обзора основных результатов для упомянутых выше аналогов классических одномерных вариационных задач.

§ 1. Простейшая векторная задача с закрепленными концами

Пусть $x(t) = (x_i(t))\big|_{i=1}^n$ — вектор-столбец, компоненты которого — непрерывно дифференцируемые на отрезке $[a, b]$ функции. Естественным образом (т. е. покомпонентно) определенные операции сложения и умножения на число превращают совокупность таких функций в линейное функциональное пространство. Если определить теперь расстояние между двумя вектор-функциями $x(t)$ и $y(t)$ как

$$\|x - y\|_C = \max_{1 \leqslant i \leqslant n} \{\max_t |x_i(t) - y_i(t)|\},$$

то мы получим линейное нормированное пространство с *сильно* определенным расстоянием. *Слабое* расстояние задается соотношением

$$\|x - y\|_{C^1} = \max_{1 \leqslant i \leqslant n} \left\{\max \left(\max_t |x_i(t) - y_i(t)|,\ \max_t |x_i'(t) - y_i'(t)|\right)\right\}.$$

Таким образом мы получаем возможность говорить об окрестностях фиксированной функции $x_0(t)$ — сильной и слабой, в зависимости оттого, какое определение близости мы используем.

Пусть далее $x'(t)$ — вектор-столбец производных и

$$L(t, x, x') = L\left(t, x_1, x_2, \ldots, x_n, x_1', x_2', \ldots, x_n'\right),$$

$$l(x(a), x(b)) = l\left(x_1(a), x_2(a), \ldots, x_n(a), x_1(b), x_2(b), \ldots, x_n(b)\right)$$

— скалярные функции $2n + 1$ и $2n$ переменных соответственно, непрерывно дифференцируемые по своим аргументам — $L(t, x, x')$ дважды, а $l(x(a), x(b))$ — единожды. Функционал Больца определим соотношением

$$f(x) = \int_a^b L(t, x, x')\, dt + l(x(a), x(b)),$$

в котором интегральная и терминальная части, как и в одномерном случае, задаются функциями $L(t, x, x')$ и $l(x(a), x(b))$. Задача

$$\boxed{f(x) = \int_a^b L(t, x, x')\, dt \longrightarrow \text{extr}, \quad x(a) = x_a, \quad x(b) = x_b}$$

называется *простейшей векторной экстремальной задачей*. Здесь x_a и x_b — заданные векторы-столбцы из $\mathbb{R}^n$.

Условия, наложенные на интегральный функционал, обеспечивают его дифференцируемость. Выкладки, полностью повторяющие одномерные, дают следующее выражение вариации интегрального функционала

$$\delta_h f(x) = \int_a^b \sum_{i=1}^n \left(L_{x_i} h_i + L_{x_i'} h_i'\right) dt.$$

Учитывая условия закрепления на концах отрезка, проинтегрируем слагаемые вида $L_{x_i'} h_i'$ по частям

$$\int_a^b L_{x_i'} h_i'\, dt = -\int_a^b h_i \frac{d}{dt} L_{x_i'}\, dt$$

и перепишем выражение для вариации в виде

$$\delta_h f(x) = \int_a^b \sum_{i=1}^n \left(L_{x_i} - \frac{d}{dt} L_{x_i'}\right) h_i\, dt. \tag{1}$$

Необходимое условие экстремума в рассматриваемой задаче состоит в тождественном равенстве нулю вариации (1) в точке экстремума для всех допустимых приращений аргумента $h(t) = (h_i(t))$, $i = 1, 2, \ldots, n$, $h_i(a) = h_i(b) = 0$. Возьмем функцию $h(t)$

вида

$$h(t) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ h_i(t) \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

где $h_i(t)$ — произвольная гладкая функция, обращающаяся в нуль на концах отрезка. Для таких $h(t)$ условие равенства нулю вариации (1) влечет за собой обращение в нуль выражения $L_{x_i} - \frac{d}{dt}L_{x_i'}$ в точке экстремума.

Таким образом, мы приходим к заключению, что *если $x_0(t)$ — векторная функция, доставляющая экстремум интегральному функционалу в задаче с закрепленными концами, то $x_0(t)$ должна быть решением следующей краевой задачи*

$$\begin{cases} L_{x_1} - \dfrac{d}{dt}L_{x_1'} = 0, \\ L_{x_2} - \dfrac{d}{dt}L_{x_2'} = 0, \\ \ldots\ldots\ldots\ldots\ldots \\ L_{x_n} - \dfrac{d}{dt}L_{x_n'} = 0, \end{cases}$$

$$x_1(a) = x_{1a}, \quad x_2(a) = x_{2a}, \quad \ldots, \quad x_n(a) = x_{na},$$

$$x_1(b) = x_{1b}, \quad x_2(b) = x_{2b}, \quad \ldots, \quad x_n(b) = x_{nb}.$$

Система полученных обыкновенных дифференциальных уравнений носит название *системы уравнений Эйлера.*

Совершенно аналогично строим вариацию для функционала Больца

$$\delta_h f(x) = \int_a^b \sum_{i=1}^n \left(L_{x_i}h_i + L_{x_i'}h_i'\right) dt + \sum_{i=1}^n \left(l_{x_i(a)}h_i(a) + l_{x_i(b)}h_i(b)\right).$$

Поскольку условия закрепления концов в векторной задаче Больца

$$f(x) = \int_a^b L(t, x, x')\, dt + l(x(a), x(b)) \longrightarrow \text{extr},$$

вообще говоря, отсутствуют, то интегрирование по частям слагаемых, содержащих производные функций $h_i(t)$, приводит к следующему выражению для вариации

$$\delta_h f = \sum_{i=1}^n \left[\int_a^b \left(L_{x_i} - \frac{d}{dt}L_{x_i'}\right)h_i\, dt + \left(l_{x_i(a)} - L_{x_i'}\right)h_i(a) + \left(l_{x_i(b)} + L_{x_i'}\right)h_i(b)\right].$$

Теперь, используя, как и в простейшей векторной задаче, покомпонентно-нулевые векторы $h(t)$ приращений аргумента, приходим к необходимому условию экстрему-

ма: *если $x_0(t)$ — функция, доставляющая экстремум в векторной задаче Больца, то $x_0(t)$ должна удовлетворять системе уравнений Эйлера и условиям трансверсальности, задаваемым соотношениями*

$$\left.\frac{\partial L}{\partial x_i'}\right|_{t=a} = \frac{\partial l}{\partial x_i(a)}, \quad \left.\frac{\partial L}{\partial x_i'}\right|_{t=b} = -\frac{\partial l}{\partial x_i(b)}, \quad i = 1, 2, \ldots, n.$$

Для получения необходимых условий экстремума в векторой задаче с ограничениями-равенствами, задаваемыми интегральными и терминальными функционалами, нужно использовать лемму о пропорциональности линейных функционалов (не зависящую от размерности задачи) и сформулированные выше необходимые условия для задачи Больца. Не вдаваясь в подробности, сформулируем результат:

если $x_0(t)$ — функция, доставляющая экстремум в векторной задаче

$$f(x) = \int_a^b L(t, x, x')\, dt + r(x(a), x(b)) \longrightarrow \text{extr},$$

с ограничениями, задаваемыми равенствами

$$g_k(x) = \int_a^b G_k(t, x, x')\, dt = 0, \quad k = 1, \ldots, m, \quad r_j(x(a), x(b)) = 0, \quad j = 1, \ldots, s,$$

то существуют постоянные λ_k, $k = 1, \ldots, m$, μ_j, $j = 1, \ldots, s$, такие, что $x_0(t)$ является стационарной точкой функционала Лагранжа

$$F(x) = \int_a^b \left[f(x) + \sum_{k=1}^{m} \lambda_k G_k(t, x, x')\right] dt + r(x(a), x(b)) + \sum_{j=1}^{s} \mu_j r_j(x(a), x(b)).$$

Конечно, все оговорки, сделанные в одномерном случае[1)], должны быть сделаны и в векторном случае.

§ 2. Векторная задача с подвижными концами

Задача с подвижными концами в принципе также может быть исследована аналогично одномерному случаю, однако тут ситуация в отношении техники исследования является несколько более сложной. Чтобы не загромождать изложение выкладками, рассмотрим случай $n = 2$, вводя для этого случая специальные обозначения

$$x_1(t) \equiv x(t), \quad x_2(t) \equiv y(t).$$

Пусть $\varphi_k(\tau, u, v)$, $k = 1, 2$ — гладкие функции трех переменных. Рассмотрим задачу об экстремуме интегрального функционала

$$\int_a^\beta L(t, x, y, x', y')\, dt \longrightarrow \text{extr}.$$

[1)] Должны выполняться предположение отом, что $x_0(t)$ не является стационарной точкой функционалов-ограннчений, и предположение о независимости последних.

Будем считать, что левый конец искомой траектории удовлетворяет условиям закрепления $x(a) = x_a$, $y(a) = y_a$, а правый скользит по множеству $\Phi \in \mathbb{R}^3$, задаваемому соотношениями

$$\varphi_1(\beta, x(\beta), y(\beta)) = 0, \quad \varphi_2(\beta, x(\beta), y(\beta)) = 0, \tag{1}$$

либо соотношением

$$\varphi(\beta, x(\beta), y(\beta)) = 0. \tag{2}$$

Замечание. В общем случае количество ограничений может быть любым, однако оно не должно превышать количества компонент отыскиваемой траектории, так как это, как правило, приводит к вырождению рассматриваемой экстремальной задачи — либо оказывается, что условия несовместны, либо они жестко фиксируют положение правого конца — и новой задачи мы не получаем.

В случае ограничений (1) терминальная часть функционала Лагранжа будет иметь вид

$$r_{\text{term}} = \mu_1\varphi_1(\beta, x(\beta), y(\beta)) + \mu_2\varphi_2(\beta, x(\beta), y(\beta)),$$

что приводит к следующим условиям трансверсальности

$$\left.\frac{\partial L}{\partial x'}\right|_{t=\beta} + \mu_1\frac{\partial\varphi_1}{\partial x(\beta)} + \mu_2\frac{\partial\varphi_2}{\partial x(\beta)} = 0, \quad \left.\frac{\partial L}{\partial y'}\right|_{t=\beta} + \mu_1\frac{\partial\varphi_1}{\partial y(\beta)} + \mu_2\frac{\partial\varphi_2}{\partial y(\beta)} = 0.$$

Условие стационарности функционала Лагранжа по скользящему правому концу запишется в виде

$$L|_{t=\beta} + \sum_{k=1}^{2}\mu_k\left(\frac{\partial\varphi_k}{\partial\beta} + \frac{\partial\varphi_k}{\partial x(\beta)}x'(\beta) + \frac{\partial\varphi_k}{\partial y(\beta)}y'(\beta)\right) = 0.$$

Рассмотрим три эти соотношения как однородную систему линейных уравнений относительно переменных $(1, \mu_1, \mu_2)$. Поскольку система имеет ненулевое решение, то ее определитель обращается в нуль

$$\begin{vmatrix} L & \varphi_{1\beta} + \varphi_{1x}x' + \varphi_{1y}y' & \varphi_{2\beta} + \varphi_{2x}x' + \varphi_{2y}y' \\ L_{x'} & \varphi'_{1x} & \varphi_{2x} \\ L_{y'} & \varphi'_{1y} & \varphi_{2y} \end{vmatrix} = 0.$$

Это и есть искомое условие трансверсальности в задаче с подвижным концом. В случае одного ограничения (2) аналогичные рассуждения приводят к двум условиям трансверсальности, которые могут быть записаны в виде

$$\begin{vmatrix} L & \varphi_\beta + \varphi_x x' + \varphi_y y' \\ L_{x'} & \varphi_x \end{vmatrix} = 0, \quad \begin{vmatrix} L & \varphi_\beta + \varphi_x x' + \varphi_y y' \\ L_{y'} & \varphi_y \end{vmatrix} = 0.$$

Для задачи в пространстве n-компонентных вектор-функций наличие k ограничений на положение скользящего конца порождает $n - k + 1$ условий трансверсальности.

В заключение отметим, что если векторная задача получена параметризацией скалярной, то она обладает некоторыми специфическими особенностями — во-первых, интегральная часть зависит только от фазовых переменных $x(t)$, во-вторых, уравнения Эйлера оказываются зависимыми.

◀ Действительно, пусть исходная вариационная задача имеет вид

$$\int_{x_a}^{x_b} L(x, y, y')\,dx \longrightarrow \text{extr}.$$

Полагая $x = x(t)$, $y = y(t)$, $a \leqslant t \leqslant b$, $x(a) = x_a$, $x(b) = x_b$, получаем векторную задачу

$$\int_a^b L\left(x, y, \frac{y'}{x'}\right) x'\, dt = \int_a^b \bar{L}(x, x', y, y')\, dt \longrightarrow \text{extr}.$$

Функция $\bar{L}(x, x', y, y')$ является однородной функцией первого порядка по переменным x', y': $\bar{L}(x, \lambda x', y, \lambda y') = \lambda \bar{L}(x, x', y, y')$. Дифференцируя последнее соотношение по параметру λ в точке $\lambda = 1$, приходим к тождеству Эйлера

$$x'\bar{L}_{x'}(x, x', y, y') + y'\bar{L}_{y'}(x, x', y, y') = \bar{L}(x, x', y, y'),$$

откуда следует, что первые интегралы уравнений Эйлера связаны. Поэтому общее решение системы необходимых условий будет содержать произвольную функцию, которая отвечает различным способам параметризации исходной задачи. ▶

§ 3. Задача Лагранжа: дифференциальные и фазовые ограничения

Рассмотрим векторную экстремальную задачу для интегрального функционала

$$\int_a^b L(t, x, x')\, dt \longrightarrow \text{extr},$$

предполагая, что компоненты искомой траектории $x_j(t)$, $j = 1, 2, 3, \ldots, n$, являются решениями системы обыкновенных дифференциальных уравнений

$$\begin{cases} \varphi_1(t, x, x') = 0, \\ \varphi_2(t, x, x') = 0, \\ \ldots\ldots\ldots\ldots \\ \varphi_k(t, x, x') = 0. \end{cases} \tag{1}$$

Разумно считать, что количество уравнений k не превосходит количества компонент n отыскиваемой траектории: если уравнений больше, то система может оказаться несовместной, а если столько же, то новой задачи не возникает — выражая из уравнений производные $x'(t)$ и подставляя их в исследуемый функционал, приходим к векторной экстремальной задаче, уже рассмотренной в предыдущих параграфах. Поэтому в дальнейшем будем считать $k < n$.

Отметим важный частный случай ограничений (1), когда функции $\varphi_j(t, x, x')$ (все или некоторые) не зависят от значений производных, $\varphi_j(t, x, x') = \psi_j(t, x)$. В этом случае говорят о *фазовых или конечных* ограничениях. В механике такие связи называются *голономными*, в отличие от дифференциальных, которые там называются *неголономными*.

Дифференциальные и фазовые ограничения не являются альтернативными к ранее рассмотренным типам ограничений — дополнительно к связям (1) на компоненты искомого вектора $x(t)$ могут быть наложены ограничения всех других типов. Такая векторная экстремальная задача называется *задачей Лагранжа*. Она содержит в себе

в качестве частных случаев все рассмотренные выше экстремальные задачи — как одномерные, так и векторные.

Именно при исследовании этой задачи Лагранж впервые сформулировал и использовал принцип, называемый сегодня *правилом множителей Лагранжа*, который затем был распространен с вариационных задач на задачи об экстремуме функций нескольких переменных. Сущность этого принципа мы имели возможность проследить на рассмотренных ранее задачах — он позволяет сводить исследование задачи на экстремум с ограничениями к исследованию безусловных стационарных значений специальным образом построенного функционала — функционала Лагранжа.

В общей задаче Лагранжа этот принцип также применим, однако новые ограничения (1) требуют модификации понятия *множитель Лагранжа* и модификации правила построения функционала Лагранжа.

Наводящими соображениями (которые могут быть аккуратно формализованы и превращены в строгое доказательство), приводящими к правилу множителей Лагранжа в общей задаче Лагранжа, могут служить следующие.

Зафиксируем точку t на отрезке $[a, b]$ и рассмотрим ограничение $\varphi(t, x, x') = 0$ в этой точке. Оно представляет собой ограничение-равенство, задаваемое терминальным функционалом, и потому может быть включено в функционал Лагранжа с некоторым множителем ν. Поскольку в каждой точке t отрезка возникает свое ограничение-равенство, то и множитель Лагранжа ν должен зависеть от точки — $\nu = \nu(t)$. Но точек на отрезке очень «много», поэтому сумма терминальных слагаемых, которую мы должны были бы внести в функционал Лагранжа

$$S(\nu, \varphi) = \sum_{a \leqslant t \leqslant b} \nu(t)\varphi(t, x, x'),$$

переходит в интеграл

$$\bar{S}(\nu, \varphi) = \int_a^b \nu(t)\varphi(t, x, x')\, dt,$$

и мы приходим к правилу множителей Лагранжа в задаче с дифференциальными и/или фазовыми ограничениями, которое сформулируем в виде теоремы, предварительно модифицировав определение функционала Лагранжа.

Рассмотрим задачу

$$f(x) = \int_a^b L(t, x, x')\, dt \longrightarrow \text{extr}$$

с ограничениями

$$\begin{aligned} g_i(x) &= \int_a^b G_i(t, x, x')\, dt, \quad i = 1, 2, \ldots, m, \\ \varphi_s(t, x, x') &= 0, \quad s = 1, 2, \ldots, k_1, \\ \psi_l(t, x) &= 0, \quad l = 1, 2, \ldots, k_2, \quad k_1 + k_2 < n. \end{aligned} \tag{2}$$

Заметим, что граничные условия $x(a) = x_a$, $x(b) = x_b$ должны быть согласованы с фазовыми ограничениями

$$\psi_l(t, x_a)\big|_{t=a} = 0, \quad \psi_l(t, x_b)\big|_{t=b} = 0, \quad l = 1, 2, \ldots, k_2.$$

Определение. *Функционалом Лагранжа*, ассоциированным со сформулированной выше экстремальной задачей, назовем функционал $F(x, \lambda, \mu, \nu)$, задаваемый соотношением

$$F(x, \lambda, \mu, \nu) = \int_a^b F_{\text{int}}(t, x, x', \lambda, \mu, \nu)\, dt + r_{\text{term}}(x(a), x(b)),$$

в котором интегральная часть определяется функцией $F_{\text{int}}(t, x, x', \lambda, \mu, \nu)$

$$F_{\text{int}} = L(t, x, x') + \sum_{i=1}^{m} \lambda_i G_i(t, x, x') + \sum_{j=1}^{k_1} \mu_j(t)\varphi_j(t, x, x') + \sum_{s=1}^{k_2} \nu_s(t)\psi_s(t, x),$$

а терминальная строится обычным образом, как было показано выше.

Числа λ_i и функции $\mu_j(t)$, $\nu_s(t)$ *называются множителями Лагранжа* рассматриваемой задачи.

Теорема Эйлера—Лагранжа (правило множителей Лагранжа для экстремальной задачи с дифференциальными и фазовыми ограничениями). *Пусть $x_0(t)$ доставляет экстремум интегральному функционалу в задаче Лагранжа с дифференциальными* (1) *и фазовыми* (2) *ограничениями, где задающие ограничения функции непрерывно дифференцируемы и удовлетворяют условиям, обеспечивающим разрешимость системы неявных уравнений* (1)–(2) *относительно некоторого набора из k компонент вектора $x_0(t)$. Тогда существуют множители Лагранжа λ_i и $\mu_j(t)$, $\nu_s(t)$, такие, что $x_0(t)$ является стационарной точкой для функционала Лагранжа.*

3.1. Пример — задача Чаплыгина

В качестве примера рассмотрим задачу Чаплыгина.

Самолет, двигающийся с постоянной относительно воздуха собственной скоростью $|\mathbf{u}| = u$, должен за заданное время T облететь территорию максимальной площади. Предполагается, что скорость ветра постоянна как по величине, так и по направлению и равна v, $v < u$.

◀ Выберем систему координат (x, y) так, чтобы ось абсцисс была направлена вдоль направления скорости ветра. Пусть положение самолета в момент времени t описывается координатами $(x(t), y(t))$. Будем считать, что самолет вылетает из точки, расположенной на оси ординат, и возвращается в эту же точку: $x(0) = x(T) = 0$, $y(0) = y(T) = y_T$. Площадь, охваченная траекторией самолета, дается выражением

$$S(x, y) = \frac{1}{2} \int_0^T (xy' - x'y)\, dt.$$

Из условия задачи следует, что во всех точках искомой траектории должно иметь место соотношение

$$(x' + v)^2 + y'^2 = u^2.$$

Таким образом, мы приходим к задаче об отыскании максимума функционала

$$S(x, y) = \frac{1}{2}\int_0^T (xy' - x'y)\, dt \longrightarrow \max$$

с ограничениями на искомую траекторию, задаваемыми соотношениями

$$(x' + v)^2 + y'^2 = u^2, \quad x(0) = x(T) = 0, \quad y(0) = y(T) = y_T.$$

Функционал Лагранжа

$$F(x) = \int_0^T \left[\frac{1}{2}(xy' - x'y) + \mu(t)\left((x' + v)^2 + y'^2 - u^2\right)\right] dt$$

порождает следующую систему уравнений Эйлера

$$\frac{1}{2}y' - \frac{d}{dt}\left(-\frac{1}{2}y + 2\mu(t)(x' + v)\right) = 0,$$

$$-\frac{1}{2}x' - \frac{d}{dt}\left(\frac{1}{2}x + 2\mu(t)y'\right) = 0.$$

Интегрируя почленно по t и присоединяя к интегралам уравнений Эйлера дифференциальное ограничение, приходим к системе трех уравнений с тремя неизвестными функциями — $x(t)$, $y(t)$, $\mu(t)$

$$\begin{cases} y - 2\mu(x' + v) = C_1, \\ -x - 2\mu y' = C_2, \\ (x' + v)^2 + y'^2 = u^2. \end{cases}$$

Сделаем замену переменных в уравнениях системы, сдвигая ось абсцисс на величину $-C_2$, а ось ординат на C_1. Эта замена — параллельный перенос осей — приводит к тому, что в новых координатах произвольные постоянные обращаются в нуль. Считая, что мы с самого начала выбрали расположение осей именно так, перепишем систему в виде

$$\begin{cases} y = 2\mu(x' + v), \\ -x = 2\mu y', \\ (x' + v)^2 + y'^2 = u^2. \end{cases}$$

Разделив первое уравнение на второе почленно, исключим из системы функцию $\mu(t)$. В результате придем к соотношению

$$-\frac{y}{x} = \frac{x'}{y'} + \frac{v}{y'},$$

которое влечет за собой равенство

$$x' + v = -\frac{yy'}{x}.$$

Вследствие полученного соотношения последнее уравнение системы примет вид

$$y' = \frac{ux}{\sqrt{x^2+y^2}} \quad \Longrightarrow \quad \frac{v}{y'} = \frac{v}{ux}\sqrt{x^2+y^2}.$$

Учитывая, что $\frac{x'}{y'} = \frac{dx}{dy}$, приходим к обыкновенному дифференциальному уравнению первого порядка, однородному относительно x и y, которое легко интегрируется подстановкой $\frac{x}{y} = z(y)$. Общий интеграл этого уравнения имеет вид

$$vy + u\sqrt{x^2+y^2} = C.$$

Записывая последнее соотношение в полярных координатах и вводя новую произвольную постоянную $\overline{C} = \frac{C}{u}$, получим

$$r = \frac{\overline{C}}{1 + \frac{v}{u}\sin\varphi}.$$

Это уравнение представляет собой полярное уравнение семейства линий второго порядка с эксцентриситетом $\varepsilon = \frac{v}{u}$. По условию скорость ветра меньше скорости самолета, т. е. эксцентриситет меньше единицы — искомой траекторией облета максимальной площади за заданное время является эллипс, большая ось которого перпендикулярна направлению ветра. ▶

3.2. Пример — задача о брахистохроне

Исторически первой вариационной задачей, которая привлекла внимание математиков, была *задача о брахистохроне*, поставленная И. Бернулли.

> Среди всех кривых, соединяющих две данные точки плоскости, найти ту, двигаясь по которой под действием силы тяжести, материальная точка попадет из начальной точки в конечную за кратчайшее время.

Кривая, вдоль которой точка скорее всего скатывается из начальной точки в конечную называется *брахистохроной*.

Мы рассмотрим эту задачу и ее модификацию — задачу о брахистохроне со свободным правым концом — предварительно обсудив возможные их формализации.

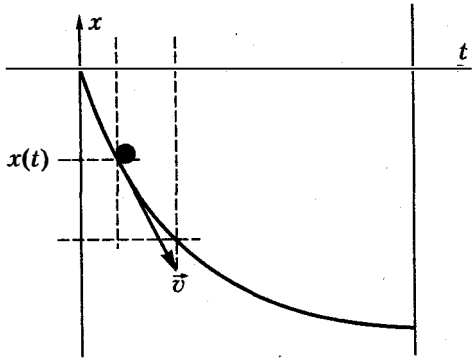


Рис. 1

Первая формализация позволяет поставить задачу о брахистохроне как классическую вариационную задачу для интегрального функционала.

Пусть $x(t)$ — траектория скорейшего скатывания на плоскости (t, x). Предположим, что начальная скорость точки равна нулю и скатывание осуществляется без трения. Тогда ее скорость v на высоте x полностью определяется высотой $x + h$, с которой точка начала свое движение: кинетическая энергии точки должна быть равна изменению потенциальной при спуске с высоты $x + h$ на уровень x

$$\frac{mv^2}{2} = mgh \quad \Longrightarrow \quad v^2 = 2gh,$$

где g — ускорение свободного падения. Поместим начальную точку в начало (рис. 1) координат. Тогда зависимость скорости скатывания от высоты x будет иметь вид

$$v^2 = -2gx \quad \Longrightarrow \quad v = \sqrt{-2gx}.$$

Элемент дуги Δs на промежутке $[t, t + \Delta t]$ точка пройдет за время ΔT, даваемое соотношением

$$\Delta T = \frac{\sqrt{1 + x'^2(t)}}{v(t)}\, dt = \frac{\sqrt{1 + x'^2(t)}}{\sqrt{-2gx}}\, dt.$$

Если конечная точка брахистохроны фиксирована, то мы приходим к задаче

$$T(x) = \int_0^{t_{\text{кон}}} \frac{\sqrt{1 + x'^2(t)}}{\sqrt{-2gx}}\, dt \longrightarrow \min, \quad x(0) = 0, \quad x(t_{\text{кон}}) = x_{\text{кон}},$$

которая является одномерной гладкой классической вариационной задачей с закрепленными концами.

Если же положение правого конца брахистохроны не задано, а фиксирована только вертикаль, которой точка должна достигнуть за кратчайшее время, то мы получаем задачу со свободным правым концом.

Другая, упомянутая выше, формализация выглядит следующим образом.

Пусть на фазовой плоскости (x, y) положение точки в момент времени t задается соотношениями $x = x(t)$, $y = y(t)$. При этом в точке (x, y) скорость v должна удовлетворять соотношению

$$v^2 = x'^2 + y'^2 = -2gy.$$

В начальной точке, которую мы, как и выше, поместим в начало координат, должно выполняться условие $x(0) = y(0) = 0$, в конечной же точке траектории, в зависимости от рассматриваемой модификации задачи, будут иметь место соотношения

$$x(T) = t_{\text{кон}}, \quad y(T) = x_{\text{кон}},$$

если речь идет о задаче с закрепленными концами, и

$$x(T) = t_{\text{кон}},$$

если речь идет о достижения точкой фиксированной вертикали. В обоих случаях T — подлежащее определению *время скорейшего скатывания*, которое и является минимизируемым функционалом

$$T(x, y) = \int_0^T dt = T \longrightarrow \min.$$

Эта формализация привела нас к векторной задаче Лагранжа с дифференциальным ограничением $x'^2 + y'^2 + 2gy = 0$.

Необходимые условия экстремума в этой задаче, в соответствии с теоремой Эйлера—Лагранжа, это условия стационарности функционала Лагранжа

$$F(x, y, \lambda, \mu(t)) = \int_0^T \left[1 + \mu(t)\left(x'^2 + y'^2 + 2gy\right)\right] dt + r(x(T), y(T), \lambda),$$

где терминальная часть имеет вид

$$r(x(T), y(T), \lambda) = \lambda_1(x(T) - t_{\text{кон}}) + \lambda_2(y(T) - x_{\text{кон}})$$

в случае задачи с закрепленными концами и

$$r(x(T), \lambda) = \lambda(x(T) - t_{\text{кон}})$$

для задачи со свободным правым концом. Завершают этот список граничные условия в начальной точке

$$x(0) = y(0) = 0.$$

Для рассматриваемой задачи получаем:

— уравнения Эйлера

$$-\frac{d}{dt}(2x'\mu(t)) = 0, \quad 2\mu(t)g = \frac{d}{dt}(2y'\mu(t)),$$

— условия трансверсальности

$$2\mu(T)x'(T) = -\lambda, \quad 2\mu(T)y'(T) = 0,$$

— условие стационарности по T

$$1 + \mu(T)\left(x'^2(T) + y'^2(T) + 2gy(T)\right) + \lambda x'(T) = 0.$$

Вводя обозначение $\mu(t)y'(t) = z(t)$, запишем получившуюся систему дифференциальных условий в виде

$$\begin{cases} \mu x' = C_1, \\ \mu y' = z, \\ z' = \mu g, \\ x'^2 + y'^2 + 2gy = 0. \end{cases} \tag{3}$$

Выражая производные x' и y' из первых двух уравнений системы и подставляя в последнее, приходим к соотношению

$$C_1^2 + z^2 + 2\mu gy = 0,$$

продифференцировав которое почленно, получим равенство

$$2z'z + 4\mu\mu' gy + 2\mu^2 gy' = 0. \tag{4}$$

Перемножим второе и третье уравнения системы (3)

$$\mu^2 gy' = z'z$$

и подставим в (4)

$$\mu y' + \mu' y = 0 \implies \frac{\mu'}{\mu} = -\frac{y'}{y} \implies \mu(t) = \frac{C_2}{y}.$$

Далее

$$x' = \frac{C_1}{C_2}y = Cy.$$

Последнее уравнение системы (3) можно разрешить относительно функции $y(t)$

$$C^2y^2 + y'^2 + 2gy = 0 \quad \Longrightarrow \quad y' = \sqrt{-2gy - C^2y^2} = \frac{g}{C}\sqrt{1 - \frac{C^2}{g}\left(y + \frac{g}{C^2}\right)^2}.$$

Это уравнение первого порядка с разделяющимися переменными легко интегрируется

$$\arcsin \frac{C^2}{g}\left(y + \frac{g}{C^2}\right) = t + C_3 \quad \Longrightarrow \quad y(t) = \frac{g}{C^2}\sin(t + C_3) - \frac{g}{C^2}.$$

Используя теперь первое уравнение системы (3), получим

$$x(t) = -\frac{g}{C}\cos(t + C_3) - \frac{g}{C}t + C_4.$$

Из граничных условий $x(0) = y(0) = 0$ в начальной точке $t = 0$ заключаем, что $C_3 = \frac{\pi}{2}$, $C_4 = 0$, откуда уравнение искомой кривой имеет вид

$$x(t) = \frac{g}{C}(\sin t - t), \quad y(t) = \frac{g}{C^2}(\cos t - 1). \tag{5}$$

Следовательно, семейство экстремалей в задаче о брахистохроне — это семейство циклоид, проходящих через начало координат (рис. 2).

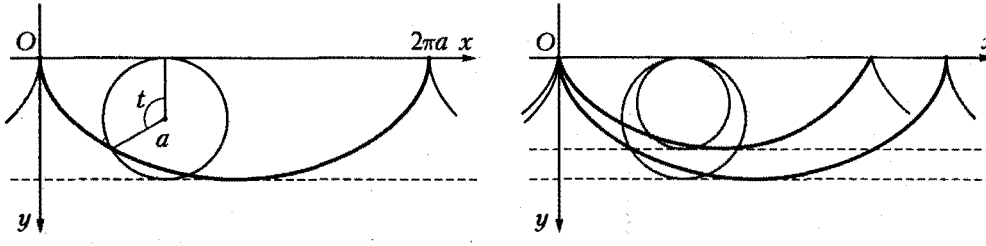


Рис. 2

Произвольная постоянная может быть определена либо из условия прохождения циклоиды через конечную точку, либо из условия трансверсальности на правом конце, которое в данном случае превращается в условие ортогональности брахистохроны вертикали $x = t_{\text{кон}}$.

Упражнения

1. Найдите экстремали функционалов:

а) $f(x, y) = \int\limits_1^2 (x'^2 + y^2 + y'^2)\,dt, \quad x(1) = 1, \quad x(2) = 2, \quad y(1) = 0, \quad y(2) = 1;$

б) $f(x, y) = \int\limits_0^{\pi} (2xy - 2x^2 + x'^2 - y'^2)\,dt, \quad x(0) = 0, \quad x(\pi) = 1, \quad y(0) = 0, \quad y(\pi) = -1;$

в) $f(x, y) = \int\limits_0^{\pi/2} (x'^2 - 2xy + y'^2)\,dt, \quad x(0) = 0, \quad x\left(\frac{\pi}{2}\right) = 1, \quad y(0) = 0, \quad y\left(\frac{\pi}{2}\right) = 1;$

г) $f(x, y) = \int\limits_0^1 (x'^2 + 2x + y'^2)\, dt, \quad x(0) = 1, \quad x(1) = \frac{3}{2}, \quad y(0) = 0, \quad y(1) = 1;$

д) $f(x, y) = \int\limits_0^{\pi/2} (x'y' - xy)\, dt, \quad x(0) = 0, \quad x\left(\frac{\pi}{2}\right) = 1, \quad y(0) = 0, \quad y\left(\frac{\pi}{2}\right) = -1;$

е) $f(x, y) = \int\limits_0^1 (x'y' + xy)\, dt, \quad x(0) = 0, \quad x(1) = e, \quad y(0) = 0, \quad y(1) = \frac{1}{e}.$

2. Установите, что если лагранжиан $L(t, x, x', y, y')$ не зависит от x, то уравнения Эйлера допускают первый интеграл $\dfrac{\partial L}{\partial x'} = C$.

3. Установите, что если лагранжиан $L(t, x, x', y, y')$ не зависит от t, то уравнения Эйлера допускают первый интеграл $L - x' \dfrac{\partial L}{\partial x'} - y' \dfrac{\partial L}{\partial y'} = C$.

4. Запишите и исследуйте необходимое условие экстремума в задаче:

$$f(x, y) = \int\limits_0^{\xi} (x'^2 + y'^2 + 2xy)\, dt \Rightarrow \text{extr}, \quad x(0) = 0, \quad y(0) = 0$$

и точка $M(\xi, x, y)$ перемещается по плоскости $t = \xi$.

5. Найдите экстремали функционала

$$f(x, y) = \int\limits_0^{\xi} \sqrt{1 + x'^2 + y'^2}\, dt,$$

удовлетворяющие условиям

$$x(0) = y(0) = 0, \quad x(\xi) = 2\xi - 3, \quad y(\xi) = \xi + 1.$$

Поясните геометрический смысл рассматриваемой экстремальной задачи.

6. Запишите условия трансверсальности для экстремальной задачи

$$f(x, y) = \int\limits_{t_0}^{\xi} \varphi(t, x, y)\sqrt{1 + x'^2 + y'^2}\, dt \Rightarrow \text{extr}$$

с ограничениями

$$x(t_0) = x_0, \quad y(t_0) = y_0, \quad \psi(\xi, x(\xi), y(\xi)) = 0$$

и поясните геометрический смысл полученных условий.

7. Найдите кратчайшее расстояние от точки $M(1, 1, 1)$ до сферы $t^2 + x^2 + y^2 = 1$.

8. Найдите кратчайшее расстояние от точки $M(0, 0, 3)$ до параболоида $y = t^2 + x^2$.

9. Найдите кратчайшее расстояние между эллипсоидом $\dfrac{t^2}{25} + \dfrac{x^2}{16} + \dfrac{y^2}{9} = 1$ и сферой $t^2 + x^2 + y^2 = 4$.

10. Найдите экстремали функционалов в следующих задачах с фазовыми ограничениями:

а) $f(x, y) = \int_0^1 (x'^2 + y'^2)\, dt \Rightarrow \text{extr},$

$x^2 + y^2 = 1, \quad x(0) = 1, \quad x(1) = 0, \quad y(0) = 0, \quad y(1) = 1;$

б) $f(x, y) = \int_0^1 (x'^2 + y'^2 + t^3)\, dt \Rightarrow \text{extr},$

$x - 2y + 3t = 0, \quad x(0) = 2, \quad x(1) = 1, \quad y(0) = 1, \quad y(1) = 2;$

в) $f(x, y) = \int_0^1 (x'^2 + y'^2 + 1)\, dt \Rightarrow \text{extr},$

$x + y - 2t^2 = 0, \quad x(0) = 0, \quad x(1) = 2, \quad y(0) = 0, \quad y(1) = 0;$

г) $f(x, y) = \int_0^{\pi/2} (x^2 + y^2 - x'^2 - y'^2 + \cos t)\, dt \Rightarrow \text{extr},$

$x - y = 2 \sin t, \quad x(0) = 1, \quad x\left(\frac{\pi}{2}\right) = 1, \quad y(0) = 1, \quad y\left(\frac{\pi}{2}\right) = -1.$

11. Найдите экстремали функционалов в следующих задачах с дифференциальными ограничениями:

а) $f(x, y) = \int_0^1 (x'^2 + 2y'^2 + y^2)\, dt \Rightarrow \text{extr},$

$y' = x, \quad x(0) = -2, \quad x(1) = -\frac{1}{e}, \quad y(0) = 1, \quad y(1) = 0;$

б) $f(x, y) = \int_0^1 (x'^2 + y'^2)\, dt \Rightarrow \text{extr},$

$x' = y, \quad x(0) = 0, \quad x(1) = 1, \quad y(0) = 1, \quad y(1) = 0;$

в) $f(x, y) = \int_0^{\pi} (x'^2 - y'^2)\, dt \Rightarrow \text{extr},$

$x' = y - \cos t, \quad x(0) = 0, \quad x(\pi) = 0, \quad y(0) = 0, \quad y(\pi) = \frac{\pi}{2};$

г) $f(x, y) = \int_0^{\pi/4} \left(\frac{y^2}{2} - 2t^2x^2\right) dt \Rightarrow \text{extr}, \quad x' = 2y - 4tx, \quad x(0) = 0, \quad x\left(\frac{\pi}{4}\right) = 1;$

д) $f(x, y) = \int_0^1 (x'^2 + y'^2)\, dt \Rightarrow \text{extr}, \quad x' = y - x, \quad x(0) = 0, \quad x(1) = 1.$

Ответы

1. а) $\begin{cases} x(t) = t, \\ y(t) = \dfrac{\operatorname{sh}(t-1)}{\operatorname{sh} 1}; \end{cases}$ б) $\begin{cases} x(t) = \alpha \sin t - \dfrac{t}{\pi} \cos t, \\ y(t) = \alpha \sin t + \dfrac{1}{\pi}(2 \sin t - t \cos t); \end{cases}$ в) $\begin{cases} x(t) = \sin t, \\ y(t) = \sin t; \end{cases}$

г) $\begin{cases} x(t) = \sin t, \\ y(t) = -\sin t; \end{cases}$ д) $\begin{cases} x(t) = -t, \\ y(t) = -3t; \end{cases}$ е) $\begin{cases} x(t) = e^t, \\ y(t) = e^{-t}. \end{cases}$

4. Необходимые условия имеют вид:
— уравнения Эйлера $x'' = y$; $y'' = x$,
— условия трансверсальности на правом конце $x'(\xi) = y'(\xi) = 0$,
— условие стационарности на правом конце $x(\xi) \cdot y(\xi) = \mu$,
— граничные условия на левом конце $x(0) = y(0) = 0$.
Отсюда

если $\xi \neq \frac{\pi}{2} + n\pi$, то $\begin{cases} x(t) = 0, \\ y(t) = 0; \end{cases}$ если $\xi = \frac{\pi}{2} + n\pi$, то $\begin{cases} x(t) = C\sin t, \\ y(t) = -C\sin t, \end{cases}$ $C \in \mathbb{R}$.

5. Экстремали — $\begin{cases} x(t) = -t, \\ y(t) = 2t, \end{cases}$ $\xi = 1$, $d = \sqrt{5}$. Рассматриваемая задача — задача о нахождении расстояния от начала координат до прямой, заданной уравнениями $x = 2\xi - 3$, $y = \xi + 1$.

6. Условия трансверсальности записываются в форме

$$\psi_x L - L_{x'}(\psi_t + x'\psi_x + y'\psi_y)\Big|_{t=\xi} = 0, \quad \psi_y L - L_{y'}(\psi_t + x'\psi_x + y'\psi_y)\Big|_{t=\xi} = 0.$$

Учитывая специфический вид лагранжиана, перепишем эти условия в виде

$$\psi_x(1 + x'^2 + y'^2) - x'\psi_t - x'^2\psi_x - x'y'\psi_y = 0, \quad \psi_y(1 + x'^2 + y'^2) - y'\psi_t - x'y'\psi_x - y'^2\psi_y = 0.$$

Отсюда легко получаем

$$\psi_x y' - \psi_y x' = 0 \quad \Rightarrow \quad \frac{y'}{x'} = \frac{\psi_y}{\psi_x}.$$

Далее, например, первое из указанных соотношений дает

$$\psi_x(1+x'^2+y'^2) - x'\psi_t - x'^2\psi_x - x'y'\psi_y + y'^2\psi_y - y'^2\psi_y = \psi_x - x'\psi_t - y'(y'\psi_x - x'\psi_y) = \psi_x - x'\psi_t = 0.$$

Объединяя полученные соотношения, приходим к условию, которому должна удовлетворять экстремаль рассматриваемой задачи в точках поверхности-ограничения:

$$\frac{\psi_x}{x'} = \frac{\psi_y}{y'} = \frac{\psi_t}{1}.$$

Заметим, что вектор с компонентами $\{\psi_t, \psi_x, \psi_y\}$ — это вектор нормали к поверхности, заданной соотношением $\psi(t, x, y) = 0$, а вектор с компонентами $\{1, x', y'\}$ — это касательный вектор к линии-экстремали. Полученное выше соотношение означает, таким образом, коллинеарность указанных векторов — экстремаль пересекает поверхность-ограничение под прямым углом.

7. Экстремали — $t = x = y$ $\xi = \frac{1}{\sqrt{3}}$, $d = \sqrt{3} - 1$. **8.** $d = \sqrt{\frac{11}{2}}$. **9.** $d = 1$.

10. а) $\begin{cases} x(t) = \cos\frac{\pi}{2}t, \\ y(t) = \sin\frac{\pi}{2}t; \end{cases}$ б) $\begin{cases} x(t) = 2 - t, \\ y(t) = 1 + t; \end{cases}$ в) $\begin{cases} x(t) = t^2 + t, \\ y(t) = t^2 - t; \end{cases}$ г) $\begin{cases} x(t) = \cos t + \sin t, \\ y(t) = \cos t - \sin t. \end{cases}$

11. а) $\begin{cases} x(t) = (t-2)e^{-t}, \\ y(t) = (1-t)e^{-t}; \end{cases}$ б) решений нет; в) $\begin{cases} x(t) = \frac{1}{2}t\sin t, \\ y(t) = \frac{1}{2}(\sin t - t\cos t); \end{cases}$

г) $\begin{cases} x(t) = \sin 2t, \\ y(t) = \cos 2t + 2t\sin 2t; \end{cases}$ д) $\begin{cases} x(t) = \dfrac{\operatorname{sh}\sqrt{2}t}{\operatorname{sh}\sqrt{2}}, \\ y(t) = \dfrac{\sqrt{2}\operatorname{ch}\sqrt{2}t + \operatorname{sh}\sqrt{2}t}{\operatorname{sh}\sqrt{2}}. \end{cases}$

Глава LIII

ФУНКЦИОНАЛЫ ОТ ФУНКЦИЙ НЕСКОЛЬКИХ ПЕРЕМЕННЫХ

В этой главе мы коротко рассмотрим, как развитая выше для функций одной переменной теория переносится на экстремальные задачи с функционалами, зависящими от функций нескольких переменных.

§ 1. Обозначения и допущения

Пусть Ω — ограниченная область в $\mathbb{R}^n$ с кусочно-гладкой границей $\partial\Omega$. Координаты точек в $\mathbb{R}^n$ будем обозначать буквой t, понимая под этим символом вектор-столбец с компонентами t_i, $i = 1, 2, \ldots, n$. Пусть $x(t) = x(t_1, t_2, \ldots, t_n)$ — функция n переменных, определенная и непрерывная в области Ω вплоть до границы вместе со своими частными производными[1], которые будем обозначать так

$$\frac{\partial x(t)}{\partial t_i}, \quad x_{t_i};$$

для градиента функции $x(t)$ — вектора-столбца ее частных производных — примем обозначение $\operatorname{grad} x(t)$.

Для n-кратного интеграла от функции n переменных по области Ω будем использовать обозначение

$$\int \ldots \int_{\Omega} \ldots \int \varphi(t_1, \ldots, t_n)\, dt_1 \ldots dt_n = \int_{\Omega} \varphi(t)\, dt.$$

Если функция $\varphi(t)$ рассматривается в точках ω границы $\partial\Omega$ области Ω, то интеграл от этой функции по границе будет обозначаться следующим образом

$$\int_{\partial\Omega} \varphi(t)\, d\omega.$$

Скалярный вариант формулы Гаусса—Остроградского имеет вид

$$\int_{\Omega} \frac{\partial x(t)}{\partial t_i}\, dt = \int_{\partial\Omega} x(t) \cos\theta_i\, d\omega, \tag{1}$$

[1] Под производными на границе будем понимать предел производных изнутри области в соответствующей точке границы

$$\left.\frac{\partial x}{\partial t_i}\right|_{\omega\in\partial\Omega} = \lim_{t\to\omega} \left.\frac{\partial x}{\partial t_i}\right|_{t\in\Omega}.$$

где θ_i — угол, составленный внешней к $\partial\Omega$ нормалью с положительным направлением оси Ot_i.

Для дальнейшего нам понадобится формула интегрирования по частям для кратных интегралов, являющаяся следствием формулы (1).

Лемма (формула интегрирования по частям). *Если функции $u(t)$ и $v(t)$ определены и непрерывны в области Ω вместе со своими производными, то имеет место соотношение*

$$\int\limits_{\Omega} \frac{\partial u(t)}{\partial t_i} v(t)\, dt = -\int\limits_{\Omega} u(t) \frac{\partial v(t)}{\partial t_i}\, dt + \int\limits_{\partial\Omega} u(t)v(t) \cos\theta_i\, dt. \tag{2}$$

◀ Действительно, для любых гладких в области Ω функций $u(t)$ и $v(t)$ имеет место тождество

$$\frac{\partial}{\partial t_i}(u(t)v(t)) = \frac{\partial u(t)}{\partial t_i} v(t) + u(t) \frac{\partial v(t)}{\partial t_i}.$$

Интегрируя его по области Ω, получим интегральное тождество

$$\int\limits_{\Omega} \frac{\partial}{\partial t_i}(u(t)v(t))\, dt = \int\limits_{\Omega} \frac{\partial u(t)}{\partial t_i} v(t)\, dt + \int\limits_{\Omega} u(t) \frac{\partial v(t)}{\partial t_i}\, dt.$$

Используя формулу Гаусса—Остроградского для вычисления левого интеграла, придем к равенству

$$\int\limits_{\Omega} \frac{\partial}{\partial t_i}(u(t)v(t))\, dt = \int\limits_{\partial\Omega} u(t)v(t) \cos\theta_i\, dt,$$

подставляя которое в интегральное тождество, получаем искомую *формулу интегрирования по частям* для кратных интегралов. ▶

Если $\psi(\omega)$ — функция, определенная во всех точках границы и непрерывная там, то будем говорить, что функция $x(t)$, определенная внутри области Ω, *удовлетворяет на границе условию закрепления*, если во всех точках границы имеет место равенство

$$x(t)\big|_{t=\omega} = \psi(\omega). \tag{3}$$

Введем расстояние на множестве гладких в Ω функций соотношениями:

— *сильное расстояние*

$$\|x(t) - y(t)\| = \max_{t\in\Omega} |x(t) - y(t)|,$$

— *слабое расстояние*

$$\|x(t) - y(t)\| = \max_{1\leqslant i\leqslant n} \max_{t\in\Omega} |x_{t_i}(t) - y_{t_i}(t)|,$$

и тем самым определим сильную и слабую окрестности функции $x(t)$, а вместе с ними придадим смысл всем понятиям, связанным с близостью гладких функций, определенных в Ω.

Пусть, наконец, $L(t, x, \operatorname{grad} x)$ — функция $2n+1$ переменной, дважды непрерывно дифференцируемая по совокупности переменных. Под *интегральным* функционалом

на множестве гладких в Ω функций будем понимать функционал, задаваемый соотношением

$$f(x) = \int\limits_{\Omega} L(t, x, \operatorname{grad} x)\, dt = \int\limits_{\Omega} L(t, x, x_{t_1}, x_{t_2}, \ldots, x_{t_n})\, dt.$$

Терминальным назовем функционал

$$r(x) = \int\limits_{\partial\Omega} R(t, x)\, d\omega,$$

где функция $R(t, x)$ предполагается непрерывно дифференцируемой по переменным (x, t).

§ 2. Простейшая задача для функционалов от функций нескольких переменных

Рассмотрим экстремальную задачу

$$f(x) = \int\limits_{\Omega} L(t, x, x_{t_1}, x_{t_2}, \ldots, x_{t_n})\, dt \longrightarrow \operatorname{extr}, \quad x(t)\big|_{t\in\partial\Omega} = \psi(\omega),$$

являющуюся аналогом простейшей задачи с закрепленными концами. Построим вариацию минимизируемого функционала, которая при принятых выше допущениях, очевидно, существует. Выкладки, аналогичные проведенным для одномерного случая, дают

$$\delta_h f(x) = \int\limits_{\Omega} \left[L_x(t, x, \operatorname{grad} x)h + \sum_{i=1}^{n} L_{x_{t_i}} h_{t_i} \right] dt.$$

Преобразуем второе слагаемое, учитывая, что $h(t)\big|_{t\in\partial\Omega} = 0$. Для этого рассмотрим

$$\int\limits_{\Omega} L_{x_{t_i}} h_{t_i}\, dt = \int\limits_{\Omega} L_{x_{t_i}} \frac{\partial h(t)}{\partial t_i}\, dt. \tag{1}$$

Применяя к интегралу (1) формулу интегрирования по частям, получим

$$\int\limits_{\Omega} L_{x_{t_i}} \frac{\partial h(t)}{\partial t_i}\, dt = -\int\limits_{\Omega} h(t) \frac{\partial}{\partial t_i} L_{x_{t_i}}\, dt + \int\limits_{\partial\Omega} h(t) L_{x_{t_i}} \cos\theta_i\, dt.$$

Последнее слагаемое обращается в нуль за счет обращения в нуль функции $h(t)$ на границе области Ω, что дает

$$\int\limits_{\Omega} L_{x_{t_i}} \frac{\partial h(t)}{\partial t_i}\, dt = -\int\limits_{\Omega} h(t) \frac{\partial}{\partial t_i} L_{x_{t_i}}\, dt.$$

Таким образом, для вариации интегрального функционала получаем выражение

$$\delta_h f(x) = \int\limits_{\Omega} \left[L_x(t, x, \operatorname{grad} x) - \sum_{i=1}^{n} \frac{\partial}{\partial t_i} L_{x_{t_i}} \right] h(t)\, dt. \tag{2}$$

Нетрудно доказать, что из равенства нулю вариации (2) для любых гладких $h(t)$, обращающихся в нуль на границе области следует, что *в точке экстремума* $x_0(t)$ *выполняется соотношение*

$$\boxed{L_x(t, x, \operatorname{grad} x) - \sum_{i=1}^{n} \frac{\partial}{\partial t_i} L_{x_{t_i}} = 0.} \tag{3}$$

Уравнение (3) называется *уравнением Эйлера—Остроградского.*

Краевая задача для этого уравнения с граничным условием

$$x(t)\big|_{t \in \partial\Omega} = \psi(\omega)$$

является необходимым условием экстремума в данной задаче.

Рассмотрим пример, иллюстрирующий введенные понятия.

Пример (малые колебания упругой мембраны). Рассмотрим в $\mathbb{R}^3$ с координатами (x, y, z) распределенную по некоторой поверхности систему материальных точек, эволюционирующую с течением времени t по закону $z = z(x, y, t)$. Будем считать, что положение равновесия системы задается равенством $z = z(x, y, t) = 0$, а потенциальная энергия участка $\Delta x \Delta y$ пропорциональна изменению площади поверхности в сравнении с положением равновесия

$$\Delta U(z) = k(x, y)\Delta s = k(x, y)\left(\sqrt{1 + z_x^2 + z_y^2} - 1\right)\Delta x \Delta y.$$

Такая система материальных точек (упругая поверхность) называется в механике *мембраной.*

◀ Пусть форма мембраны в положении равновесия задается областью Ω. Функция $z = z(x, y, t)$ определена в этой области и описывает положение каждой точки (x, y) мембраны в каждый момент времени t. Если поверхностная плотность мембраны задается функцией $\rho(x, y)$, то для потенциальной и кинетической энергий деформированной мембраны получаем

$$U = \int\limits_{\Omega} k(x, y)\left(\sqrt{1 + z_x^2 + z_y^2} - 1\right) dx\, dy, \quad T = \frac{1}{2} \int\limits_{\Omega} \rho(x, y) z_t^2\, dx\, dy.$$

Если дополнительно предположить, что на мембрану действует внешняя сила (возможно также меняющаяся со временем) с плотностью $\varphi(x, y, t)$, то, учитывая, что ее потенциальная энергии дается соотношением

$$U_\varphi = -\int\limits_{\Omega} \varphi(x, y, t) z(x, y, t)\, dx\, dy,$$

получим для потенциальной энергии мембраны следующее выражение

$$U = \int\limits_{\Omega} \left[k(x, y)\left(\sqrt{1 + z_x^2 + z_y^2} - 1\right) - \varphi(x, y, t) z(x, y, t)\right] dx\, dy.$$

Предположим, что мембрана совершает «малые» колебания, так что

$$|\operatorname{grad}_{x,y} z(x, y, t)| \ll 1, \quad |\operatorname{grad}_{x,y} z(x, y, t)|^4 = o(|\operatorname{grad}_{x,y} z(x, y, t)|^2),$$

и для корня в выражении потенциальной энергии получим приближение

$$k(x, y)\left(\sqrt{1 + z_x^2 + z_y^2} - 1\right) = \frac{1}{2} k(x, y)(z_x^2 + z_y^2) + o(|\operatorname{grad}_{x,y} z(x, y, t)|^2).$$

Пренебрегая бесконечно малыми величинами, порядок которых выше второго, получим соотношение для потенциальной энергии «мало деформированной» мембраны

$$U = \int\limits_{\Omega} \left[k(x, y)(z_x^2 + z_y^2) - \varphi(x, y, t) z(x, y, t)\right] dx\, dy.$$

Принцип наименьшего действия (принцип Гамильтона) утверждает, что рассматриваемая система эволюционирует во времени так, что *действие*, т. е. величина S, задаваемая соотношением

$$S(z) = \int\limits_0^t (T - U)\, dt,$$

принимает наименьшее значение. Таким образом задача о малых колебаниях мембраны оказывается задачей об экстремуме функционала $S(z)$, зависящего от функции трех переменных. Подынтегральное выражение в функционале $S(z)$ дается выражением

$$\Delta(T-U)=\frac{1}{2}\left(k(x,y)z_x^2+k(x,y)z_y^2-\varphi(x,y,t)z-\rho(x,y)z_t^2\right).$$

Отсюда уравнение Эйлера—Остроградского имеет вид

$$\varphi(x,y,t)+\frac{\partial}{\partial x}(kz_x)+\frac{\partial}{\partial y}(kz_y)=\rho(x,y)\frac{\partial^2 z}{\partial t^2}.$$

Это и есть искомый закон колебаний мембраны. ▶

§ 3. Условие трансверсальности для функционалов, зависящих от функций нескольких переменных

Рассмотрим теперь аналог задачи Больца

$$f(x)=\int\limits_{\Omega}L(t,x,\operatorname{grad}x)\,dt+\int\limits_{\partial\Omega}R(t,x)\,d\omega\longrightarrow\operatorname{extr}.$$

Из-за отсутствия дополнительных ограничений на поведение искомой функции на границе, вариация для этого функционала будет даваться выражением

$$\delta_h f(x)=\int\limits_{\Omega}\left(L_x-\sum_{i=1}^{n}\frac{\partial}{\partial t_i}L_{x_{t_i}}\right)h(t)\,dt+\int\limits_{\partial\Omega}\left(\sum_{i=1}^{n}L_{x_{t_i}}\cos\theta_i+\frac{\partial R}{\partial x}\right)h(t)\,dt,$$

из которого следует, что наряду с уравнением Эйлера—Остроградского в точке экстремума еще должно выполняться *естественное условие трансверсальности*

$$\boxed{\sum_{i=1}^{n}L_{x_{t_i}}\cos\theta_i\big|_{t\in\partial\Omega}=-\frac{\partial R}{\partial x}.}$$

Упражнения

1. Докажите, что единственной экстремалью в задаче

$$f(x)=\iint\limits_{\mathbb{D}}\sin\left(\frac{\partial x}{\partial\tau}\right)\cdot\exp\left(\frac{\partial x}{\partial\tau}\right)dt\,d\tau\Rightarrow\operatorname{extr},$$

$$\mathbb{D}=\{(t,\tau):\ 0\leqslant t,\tau\leqslant 1\},\quad x(t,0)=0,\quad x(t,1)=1;$$

является функция $x(t,\tau)=\tau$.

2. Запишите уравнение Эйлера—Остроградского для задачи

$$f(x)=\iint\limits_{\mathbb{D}}\left[\left(\frac{\partial x}{\partial t}\right)^2-\left(\frac{\partial x}{\partial\tau}\right)^2\right]dt\,d\tau\Rightarrow\operatorname{extr},\quad x(t,\tau)\big|_{(t,\tau)\in\partial\mathbb{D}}=\varphi(\omega).$$

3. Запишите уравнение Эйлера—Остроградского для задачи

$$f(x)=\iint\limits_{\mathbb{D}}\left[\left(\frac{\partial x}{\partial t}\right)^2+\left(\frac{\partial x}{\partial\tau}\right)^2\right]dt\,d\tau\Rightarrow\operatorname{extr},\quad x(t,\tau)\big|_{(t,\tau)\in\partial\mathbb{D}}=\varphi(\omega).$$

4. Запишите уравнение Эйлера—Остроградского для задачи

$$f(x) = \iint\limits_D \left[\left(\frac{\partial x}{\partial t}\right)^2 + \left(\frac{\partial x}{\partial \tau}\right)^2 - 2x\psi(t, \tau)\right] dt\, d\tau \Rightarrow \text{extr}, \quad x(t, \tau)|_{(t,\tau)\in\partial D} = \varphi(\omega).$$

5. Запишите уравнение Эйлера—Остроградского для задачи

$$f(x) = \iint\limits_D \left[\alpha(t, \tau)\left(\frac{\partial x}{\partial t}\right)^2 + \beta(t, \tau)\left(\frac{\partial x}{\partial \tau}\right)^2 + \gamma(t, \tau)x^2 + 2x\psi(t, \tau)\right] dt\, d\tau \Rightarrow \text{extr},$$

$$x(t, \tau)|_{(t,\tau)\in\partial D} = \varphi(\omega).$$

Ответы

2. $\dfrac{\partial^2 x}{\partial t^2} - \dfrac{\partial^2 x}{\partial \tau^2} = 0$. **3.** $\dfrac{\partial^2 x}{\partial t^2} + \dfrac{\partial^2 x}{\partial \tau^2} = 0$. **4.** $\dfrac{\partial^2 x}{\partial t^2} + \dfrac{\partial^2 x}{\partial \tau^2} = \psi(t, x)$.

5. $\dfrac{\partial}{\partial t}\left(\alpha(t, \tau)\dfrac{\partial x}{\partial t}\right) + \dfrac{\partial}{\partial \tau}\left(\beta(t, \tau)\dfrac{\partial x}{\partial \tau}\right) - \gamma(t, \tau)x = \psi(t, \tau)$.

Глава LIV

НЕОБХОДИМЫЕ УСЛОВИЯ СИЛЬНОГО ЭКСТРЕМУМА

В предыдущих главах достаточно подробно был рассмотрен вопрос о необходимых условиях экстремума для интегральных или интегро-терминальных функционалов, определенных на гладких функциях, при различных дополнительных предположениях о поведении функций на границе области определения. Напомним, что мы различаем *слабый, сильный и глобальный экстремумы*. Как уже отмечалось выше, если мы ограничиваем область определения рассматриваемых функционалов *гладкими* функциями, то наличие в некоторой точке $x_0(t)$ *глобального* экстремума влечет наличие в этой же точке и «младших» экстремумов — *сильного и слабого*. Аналогично, *сильный* экстремум влечет за собой наличие *слабого*. Поэтому *необходимые условия слабого экстремума* на множестве гладких функций являются одновременно *необходимыми условиями сильного и глобального экстремумов* на этом множестве.

Однако поскольку *необходимые* условия *недостаточны* для обеспечения существования экстремума, они существования ни одного из указанных экстремумов нам не гарантируют. Необходимые условия выделяют подозреваемые на экстремум функции, но о каком экстремуме идет речь, не говорят. Если окажется, что необходимое условие *слабого* экстремума выполняется на каких-то функциях, то вполне может оказаться, что среди этих функций есть доставляющие функционалу *слабый* экстремум, и нет — доставляющих сильный.

Пример. Функция $x_0(t) \equiv 0$ доставляет функционалу

$$f(x) = \int_0^\pi x^2(t)\left(1 - (x'(t))^2\right) dt.$$

слабый локальный минимум $f(x_0) = 0$.

◀ В «слабой» δ-окрестности функции $x_0(t) \equiv 0$ находятся функции, принимающие малые значения вместе со своими производными. Значит, можно подобрать δ так, что производные всех функций из δ-окрестности будут принимать значения, меньшие 1. Но тогда, в силу неотрицательности подынтегрального выражения, интеграл будет принимать только неотрицательные значения. В то же время сильный экстремум в этой точке не достигается, так как легко придумать пример функции, маленькой по модулю, но с большой производной, т. е. функции из «сильной» окрестности, для которой значение интеграла будет меньше нуля. Например, в качестве такой функции можно взять

$$x(t) = \frac{\sin 9t}{3}.$$

Тогда получим

$$f(x) = \int_0^\pi \frac{\sin^2 9t}{9}(1 - 9(\cos^2 9t))\, dt = \int_0^\pi \frac{1 - \cos 18t}{18}\, dt - \int_0^\pi \frac{1 - \cos 36t}{8}\, dt = -\frac{5\pi}{72}. \blacktriangleright$$

Поэтому изучение необходимых условий «старших», например *сильного*, экстремумов представляет самостоятельный интерес.

§ 1. Условие Вейерштрасса в простейшей задаче

Рассмотрим задачу

$$\boxed{f(x) = \int_a^b L(t, x, x')\,dt \longrightarrow \operatorname{extr}, \quad x \in C^1_{[a,b]}, \quad x(a) = x_a, \quad x(b) = x_b.}$$

Пусть $x_0(t)$ — гладкая на отрезке $[a, b]$ функция, в которой достигается сильный экстремум (для определенности минимум), т. е. значение интегрального функционала в точке $x_0(t)$ сравнивается со значениями в соседних с $x_0(t)$ в сильном смысле точками $x(t)$

$$\|x_0(t) - x(t)\| = \max_{t \in [a,b]} |x_0(t) - x(t)|,$$

и выполняется неравенство

$$f(x_0) \leqslant f(x) \quad \forall\, x\colon \|x_0(t) - x(t)\| < \delta.$$

Имеет место следующее утверждение.

Теорема (необходимое условие Вейерштрасса сильного экстремума). *Пусть $L(u, v, w)$ дважды непрерывно дифференцируема и $x_0(t)$ — функция из $C^1_{[a,b]}$, доставляющая сильный минимум в простейшей задаче классического вариационного исчисления. Тогда $x_0(t)$ — экстремаль, т. е. $x_0(t)$ является решением краевой задачи для уравнения Эйлера. Кроме того, в каждой точке кривой $x_0(t)$ при любом действительном λ выполняется неравенство*

$$\boxed{E(t, x, x', \lambda) = L(t, x, \lambda) - L(t, x, x') - (\lambda - x')L_{x'}(t, x, x') \geqslant 0.} \qquad (1)$$

◀ Сразу же заметим, что экстремальность точки сильного экстремума следует из сделанных выше замечаний относительно соотношения «старшего» и «младшего» экстремумов.

Дальнейшие рассуждения обычны для вариационного исчисления — следует рассмотреть приращение функционала на достаточно богатом множестве соседних с экстремальной точек и использовать для этих точек условие (для определенности будем вести речь о минимуме) неотрицательности приращения. В качестве такого множества соседних с экстремальной точек мы рассмотрим функции, отличающиеся от экстремальной локально, т. е. в малой окрестности точки t_0 отрезка $[a, b]$, параметризовав эти функции специальным образом.

Пусть $\delta > 0$ — произвольное, сколь угодно малое число, $\lambda \in \mathbb{R}$, t_0 — внутренняя точка отрезка $[a, b]$. Выберем $\delta > 0$ настолько малым, чтобы промежуток $[t_0 - \delta, t_0 + \sqrt{\delta}]$ целиком лежал внутри отрезка $[a, b]$. Рассмотрим семейство функций $\bar{h}_\delta(t)$ [1], задаваемых равенством (рис. 1)

$$\bar{h}_\delta(t) = \begin{cases} \lambda\delta + \lambda(t - t_0), & t \in [t_0 - \delta, t_0], \\ \lambda\delta - \lambda\sqrt{\delta}(t - t_0), & t \in [t_0, t_0 + \sqrt{\delta}]. \end{cases}$$

[1] Функции $\bar{h}_\delta(t)$ и $h_\delta(t)$ называются игольчатыми вариациями Вейерштрасса.

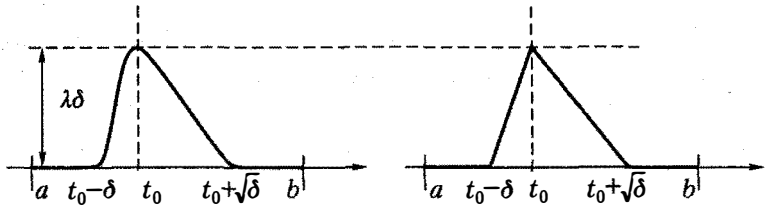


Рис. 1

Все $\bar{h}_\delta(t)$ определены на промежутке $[a, b]$ и обращаются в нуль на концах этого промежутка для любых положительных δ и произвольных λ. Сгладим теперь угловые точки функций $\bar{h}_\delta(t)$ так, чтобы они стали гладкими на $[a, b]$. Функции $h_\delta(t)$ определим как сглаженные $\bar{h}_\delta(t)$.

Рассмотрим приращение интегрального функционала, доставляемое функциями $h_\delta(t)$ в точке сильного экстремума,

$$f(x_0 + h_\delta(t)) - f(x_0) = \int_a^b \left[L(t, x_0 + h, x_0' + h') - L(t, x_0, x_0')\right] dt =$$

$$= \int_{t_0-\delta}^{t_0} \Delta_h L(t, x_0, x_0')\, dt + \int_{t_0}^{t_0+\sqrt{\delta}} \Delta_h L(t, x_0, x_0')\, dt,$$

где под $\Delta_h L(t, x_0, x_0')$ понимается приращение функции $L(t, x_0, x_0')$

$$\Delta_h L(t, x_0, x_0') = L(t, x_0 + h, x_0' + h') - L(t, x_0, x_0').$$

При $\delta \to 0$ первый из интегралов с точностью до бесконечно малых порядка, более высокого чем δ, может быть представлен в виде

$$\int_{t_0-\delta}^{t_0} \Delta_h L(t, x_0, x_0')\, dt = \int_{t_0-\delta}^{t_0} \left[L(t, x_0 + h, x_0' + \lambda) - L(t, x_0, x_0')\right] dt =$$

$$= \delta\left[L(t, x_0, x_0' + \lambda) - L(t, x_0, x_0')\right] + o(\delta).$$

Рассмотрим теперь второй интеграл. В отличие от первого в нем можно заменить приращение подынтегральной функции дифференциалом [2)], интегрируя, как обычно, второе слагаемое по частям и учитывая, что $x_0(t)$ — экстремаль, получим

$$\int_{t_0}^{t_0+\sqrt{\delta}} \Delta_h L(t, x_0, x_0')\, dt = \int_{t_0}^{t_0+\sqrt{\delta}} \left[L_x h + L_{x'} h'\right] dt + o(\delta) = -h(t_0) L_{x'} + o(\delta) = -\lambda\delta L_{x'} + o(\delta).$$

Объединяя результаты, заключаем, что для любого λ и сколь угодно маленького положительного δ приращение функционала $f(x)$ представимо (с точностью до бесконечно малых порядка более высокого, чем δ) в виде

$$f(x + h_\delta) - f(x) = \delta\left[L(t, x_0, x_0' + \lambda) - L(t, x_0, x_0') - \lambda L_{x'}(t, x_0, x_0')\right]\Big|_{t=t_0},$$

откуда, учитывая произвольность точки t_0, получаем утверждение теоремы. ▶

[2)] В первом интеграле приращение третьего аргумента, т. е. λ, вообще говоря, маленьким не является, в то время как во втором приращения всех аргументов маленькие — $h = \lambda\delta$, $h' = \lambda\sqrt{\delta}$ (рис. 2).

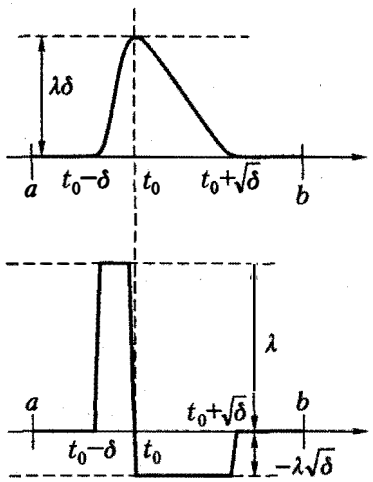


Рис. 2

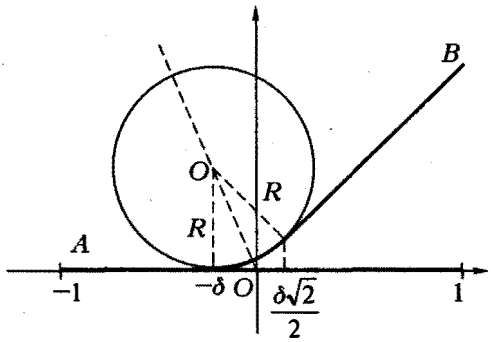


Рис. 3

§ 2. Расширение простейшей задачи. Условия Вейерштрасса—Эрдмана

Уже при изучении экстремальных задач для функций одной и нескольких переменных мы сталкиваемся с ситуацией, когда в исходной постановке задача решения не имеет, но если прибегнуть к расширению множества допустимых аргументов функции, то в этом более широком множестве таковые находятся. Аналогичная ситуация имеет место и в вариационных задачах.

Рассмотрим пример, иллюстрирующий эту ситуацию. Пусть функционал $f(x)$ задан соотношением

$$f(x) = \int_{-1}^{1} x^2(t)\left(1 - (x'(t))^2\right)\, dt.$$

Изучим задачу об экстремуме этого функционала в классе гладких на промежутке $[-1, 1]$ функций, удовлетворяющих граничным условиям $x(-1) = 0$, $x(1) = 1$. Очевидно, что для любой такой кривой $f(x) \geqslant 0$. В то же время легко указать функции, для которых значение рассматриваемого функционала сколь угодно близко к нулю. Действительно, рассмотрим функцию $x_\delta(t)$, задаваемую соотношением

$$x_\delta(t) = \begin{cases} 0 & \forall\, t:\ -1 \leqslant t < -\delta, \\ t & \forall\, t:\ \dfrac{\delta\sqrt{2}}{2} < t \leqslant 1, \\ s_\delta(t) & \forall\, t:\ -\delta \leqslant t \leqslant \dfrac{\delta\sqrt{2}}{2}, \end{cases}$$

где $s_\delta(t)$ — дуга окружности, центр которой лежит на биссектрисе угла AOB и которая касается сторон этого угла (рис. 3). Ее радиус $R = \delta(\sqrt{2} + 1)$. При любом

положительном δ функция $x_\delta(t)$ принадлежит пространству $C^1_{[-1,1]}$ и

$$f(x_\delta) = \int_{-1}^{1} x_\delta^2(t)(1 - (x_\delta'(t))^2)\,dt = \int_{-\delta}^{\delta\sqrt{2}/2} s_\delta^2(t)(1 - (s_\delta'(t))^2)\,dt \leqslant$$

$$\leqslant \int_{-\delta}^{\delta\sqrt{2}/2} s_\delta^2(t)\,dt \leqslant \frac{\delta^3(2+\sqrt{2})}{4} \to 0, \quad \delta \to 0.$$

Однако гладкой (т. е. из $C^1_{[-1,1]}$) функции, на которой функционал принял бы нулевое значение, не существует.

Как мы видим, в назначенном классе функций рассматриваемая экстремальная задача решения не имеет. Естественно посмотреть, не будет ли искомый экстремум достигаться в каком-нибудь другом, более широком классе.

В нашем примере такое расширение класса допустимых функций легко указать — достаточно наряду с гладкими допустить к рассмотрению и кусочно-гладкие функции. Напомним, что под кусочно-гладкой функцией понимается функция непрерывная и непрерывно дифференцируемая на промежутке $[a, b]$ за исключением, быть может, конечного числа точек, в которых она может иметь конечные разрывы производной — изломы с правыми и левыми касательными (рис. 4).

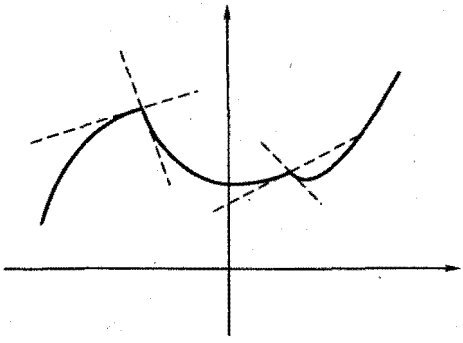


Рис. 4

Интегральный функционал будет определен на этом классе функций и примет нулевое значение в точке (рис. 5)

$$x_0(t) = \begin{cases} 0 & \forall\, t\colon\ -1 \leqslant t < 0, \\ t & \forall\, t\colon\ \ 0 < t \leqslant 1. \end{cases}$$

Ситуация, проиллюстрированная этим примером, типична в том смысле, что если сильный экстремум достигается на классе *кусочно-гладких* функций, то он совпадает с точной нижней (верхней) гранью значений рассматриваемого функционала на множестве *гладких* функций.

Это утверждение составляет содержание следующей теоремы.

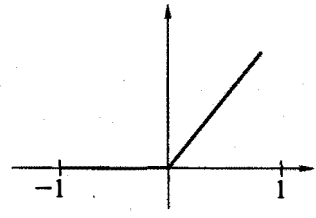


Рис. 5

Теорема (о скруглении углов). *Точные нижние (верхние) грани значений функционала*

$$f(x) = \int_a^b L(t, x(t), x'(t))\,dt$$

на множестве кусочно-гладких функций, удовлетворяющих граничным условиям

$$x(a) = x_a, \quad x(b) = x_b, \tag{1}$$

и на множестве гладких функций, удовлетворяющих тем же граничным условиям, совпадают, если только функция $L(u, v, w)$ непрерывна.

◀ Обозначим через $f_{\inf}(KC^1)$ точную нижнюю грань значений функционала $f(x)$ на множестве кусочно-гладких функций, удовлетворяющих граничным условиям (1); $f_{\inf}(C^1)$ определяется аналогично. В силу того, что всякая гладкая функция одновременно является и кусочно-гладкой, $f_{\inf}(KC_1) \leqslant f_{\inf}(C_1)$. Пусть $f_{\inf}(KC^1) < f_{\inf}(C^1)$. Тогда $\forall\, \varepsilon > 0$ найдется кусочно-гладкая функция $y_\varepsilon(t)$ такая, что $f(y_\varepsilon) < f_{\inf}(C^1) - \varepsilon$. Скругляя возможные изломы, для любого, сколь угодно малого числа δ можно построить гладкую функцию x_δ, отличающуюся от кусочно-гладкой (в смысле сильного расстояния) очень мало и такую, что значение функционала $f(x_\delta)$ будет отличаться от $f(y_\varepsilon)$ не более чем на δ. Но тогда $f(x_\delta) < f_{\inf}(C_1) - (\varepsilon - \delta)$, что невозможно при $|\delta| < \varepsilon$, ибо противоречит определению точной нижней грани. ▶

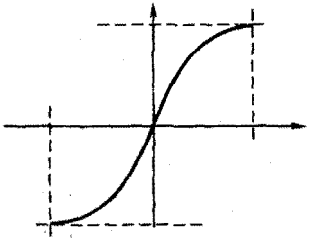


Рис. 6

Однако легко привести пример ситуации, когда расширение области определения функционала с гладких на кусочно-гладкие функции к решению задачи не приводит.

Вспомним пример Вейерштрасса. Рассмотренная там задача не имеет решения на множестве гладких функций, хотя как мы видели, функционал ограничен снизу, и можно показать, что точная нижняя грань его значений равна нулю. Действительно, рассмотрим последовательность функций, задаваемых соотношением

$$x_n(t) = \frac{\operatorname{arctg} nt}{\operatorname{arctg} n}, \quad n = 1, 2, \ldots$$

(рис. 6). Нетрудно убедиться в том, что

$$f(x_n) = \int_{-1}^{1} t^2 (x_n'(t))^2\, dt = \int_{-1}^{1} \frac{t^2 n^2 + 1 - 1}{\operatorname{arctg}^2 n(1 + n^2 t^2)^2}\, dt =$$

$$= \int_{-1}^{1} \frac{dt}{\operatorname{arctg}^2 n(1 + n^2 t^2)} - \int_{-1}^{1} \frac{dt}{\operatorname{arctg}^2 n(1 + n^2 t^2)^2} \leqslant \int_{-1}^{1} \frac{dt}{\operatorname{arctg}^2 n(1 + n^2 t^2)} =$$

$$= \frac{2}{n \operatorname{arctg} n} \to 0, \quad n \to \infty.$$

Однако экстремум[3)] этого функционала не достигается и на кусочно-гладких функциях — здесь подобного расширения области допустимых функций недостаточно.

В случае, когда сильный экстремум достигается на кусочно-гладкой функции, неравенство (1) § 1 остается справедливым во всех точках непрерывности производной. В точках же разрыва (τ_i) должно выполняться дополнительное условие

[3)] В рассматриваемой задаче минимум достигается на *кусочно-непрерывной* функции, принимающей значение $+1$ для положительных и -1 для отрицательных значений аргумента.

(*условие Вейерштрасса—Эрдмана*):

$$\begin{aligned} &L_{x'}(\tau_i, x(\tau_i), x'(\tau_i-)) = L_{x'}(\tau_i, x(\tau_i), x'(\tau_i+)),\\ &L_x(\tau_i, x(\tau_i), x'(\tau_i-)) - x'(\tau_i-)L_{x'}(\tau_i, x(\tau_i), x'(\tau_i-)) =\\ &\quad = L_x(\tau_i, x(\tau_i), x'(\tau_i+)) - x'(\tau_i+)L_{x'}(\tau_i, x(\tau_i), x'(\tau_i+)). \end{aligned}$$

При этом между точками разрыва производной функция $x_0(t)$ удовлетворяет уравнению Эйлера — *кусочно-гладкая функция, доставляющая сильный экстремум, является кусочно-экстремальной*, т. е. склеена из кусков экстремалей.

◀ Пусть для простоты экстремум достигается на функции $x_0(t)$ с одним изломом в точке t_0 (рис. 7). Рассмотрим задачу о сильном экстремуме на промежутке $[a, t_0]$. Если он достигается на гладкой функции $x_1(t) \neq x_0(t)$, $t \in [a, t_0]$, то функция

$$y(t) = \begin{cases} x_1(t), & t \in [a, t_0], \\ x_0(t), & t \in [t_0, b], \end{cases}$$

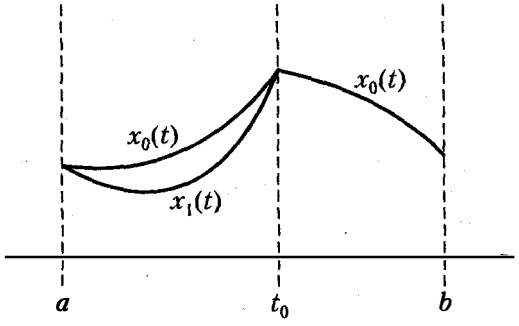


Рис. 7

является кусочно-гладкой и доставляет рассматриваемому функционалу «лучшее» значение, чем $x_0(t)$, что противоречит экстремальности последней. ▶

Кривая, в каждой точке излома которой выполняются условия Вейерштрасса—Эрдмана, а каждый гладкий кусок является экстремалью, называется *ломаной экстремалью*, а задачи с такими экстремалями называются *разрывными*; при этом, конечно, имеется в виду разрыв *производной*, а не функции, на которой достигается экстремум.

Упражнения

1. Найдите ломаные экстремали функционалов либо установите, что таковых нет:

а) $f(x) = \int_0^2 (x'^4 - 6x'^2)\,dt, \quad x(0) = 0, \quad x(2) = 0;$

б) $f(x) = \int_0^2 x'^2(x' - 1)^2\,dt, \quad x(0) = 0, \quad x(2) = 1;$

в) $f(x) = \int_0^4 (x' - 1)^2(x' + 1)^2\,dt, \quad x(0) = 0, \quad x(4) = 2;$

г) $f(x) = \int_0^1 (x'^2 - x^2)\,dt, \quad x(0) = 0, \quad x(1) = 1;$

д) $f(x) = \int_{-1}^{1} (x'^2 + 2tx - x^2)\, dt, \quad x(-1) = -1, \quad x(1) = 0;$

е) $f(x) = \int_{-1}^{1} x^2(1 - x'^2)\, dt, \quad x(-1) = 0, \quad x(1) = 1.$

Ответы

1. а) Точка излома $\tau = 1$, $x_1(t) = \begin{cases} \sqrt{3}t, & t \in [0, 1), \\ -\sqrt{3}(t-2), & t \in [1, 2), \end{cases}$

$x_2(t) = \begin{cases} -\sqrt{3}t, & t \in [0, 1), \\ \sqrt{3}(t-2), & t \in [1, 2); \end{cases}$ б) точка излома $\tau = 1$, $x_1(t) = \begin{cases} 0, & t \in [0, 1), \\ 1-t, & t \in [1, 2), \end{cases}$

$x_2(t) = \begin{cases} t, & t \in [0, 1), \\ 1, & t \in [1, 2]; \end{cases}$ в) 1-й вариант — точка излома $\tau = 1$, $x(t) = \begin{cases} 0, & t \in [0, 1), \\ t-2, & t \in [1, 4]; \end{cases}$

2-й вариант — точка излома $\tau = 3$, $x(t) = \begin{cases} 0, & t \in [0, 3), \\ -t+6, & t \in [3, 4]; \end{cases}$ г) ломаных экстремалей нет; д) ломаных экстремалей нет; е) точка излома $\tau = 0$, $x(t) = \begin{cases} 0, & t \in [-1, 0), \\ t, & t \in [0, 1]. \end{cases}$

ЛИНЕЙНОЕ ПРОГРАММИРОВАНИЕ

Важным частным случаем задач оптимизации являются задачи оптимизации линейных функций при наличии линейных же ограничений — равенств и/или неравенств. Ясно, что с точки зрения классических методов исследования экстремальных задач эта задача интереса не представляет: производные линейной функции нигде в нуль не обращаются, сужение линейной функции на множество, описываемое линейными соотношениями, является линейной функцией.

Эти два обстоятельства позволяют сделать заключение о том, что возможный экстремум может располагаться лишь в так называемых крайних точках множества ограничений. Поэтому поиск экстремума линейной функции при наличии линейных ограничений является по сути своей перебором крайних точек множества ограничений.

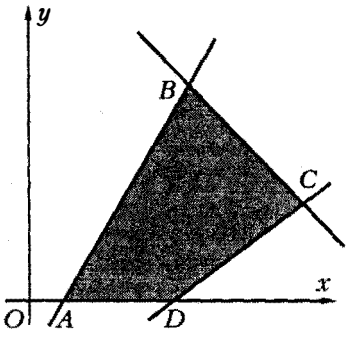


Рис. 1

Пример.

$$f(x, y) = x + y \to \min,$$

$$2x - y \leqslant 1, \quad x - 2y \geqslant 1, \quad 3x + 2y \leqslant 6, \quad y \geqslant 0.$$

◀ Множество ограничений в этой задаче представляет собой четырехугольник ABCD, изображенный на рис. 1. Экстремум заданной функции (в частности минимум) может достигаться лишь в точках A, B, C или D, являющихся вершинами четырехугольника ограничений. Сравнив значение $f(x, y)$ в указанных точках, легко определяем, что $\operatorname{argmin} f = A\left(\frac{1}{2}, 0\right)$. ▶

Рассмотренный пример, конечно, тривиален. Использованная при его исследовании методика полного перебора крайних точек множества ограничений является единственно разумной стратегией решения задачи.

Однако она становится малоэффективной, а в подавляющем большинстве случаев, к которым приводят реальные постановки прикладных задач, просто непригодной, из-за огромного числа подлежащих проверке крайних точек. Возникает потребность в новых методах, позволяющих усовершенствовать полный перебор крайних точек и сделать его эффективным.

Построению подобных методов и посвящена настоящая глава.

Глава LV

ЭЛЕМЕНТЫ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

§ 1. Постановка задачи

Пусть $\mathbf{x} \in \mathbb{R}^n$ и $f(\mathbf{x})$, $f_i(\mathbf{x})$, $i = 1, \ldots, m$, и $g_k(\mathbf{x})$, $k = 1, \ldots, l$, — линейные функции

$$f(\mathbf{x}) = c_1 x_1 + \ldots + c_n x_n = \sum_{i=1}^{n} c_i x_i,$$

$$f_i(\mathbf{x}) = f_{i1} x_1 + \ldots + f_{in} x_n = \sum_{j=1}^{n} f_{ij} x_j, \qquad i = 1, \ldots, m,$$

$$g_k(\mathbf{x}) = g_{k1} x_1 + \ldots + g_{kn} x_n = \sum_{s=1}^{n} g_{ks} x_s, \qquad k = 1, \ldots, l.$$

Общая задача линейного программирования заключается в нахождении экстремума (для определенности — минимума) функции $f(\mathbf{x})$

$$c_1 x_1 + \ldots + c_n x_n \to \min \tag{1}$$

при наличии ограничений-равенств

$$g_k(\mathbf{x}) = a_k, \quad k = 1, \ldots, l, \tag{2}$$

и ограничений-неравенств

$$f_i(\mathbf{x}) \geqslant b_i, \quad i = 1, \ldots, m. \tag{3}$$

В матричных обозначениях задача (1)–(3) может быть записана в следующем виде:

$$\begin{cases} \mathbf{C}^T \mathbf{x} \to \min, \\ \mathbf{F}\mathbf{x} \geqslant \mathbf{b}, \\ \mathbf{G}\mathbf{x} = \mathbf{a}, \end{cases} \tag{4}$$

где

$$\mathbf{F} = (f_{ij})_{i=1,\ldots,m}^{j=1,\ldots,n}, \quad \mathbf{G} = \|g_{ij}\|_{i=1,\ldots,l}^{j=1,\ldots,n}, \quad \mathbf{C} = (c_i)_{i=1}^{n}, \quad \mathbf{a} = (a_k)_{k=1}^{l}, \quad \mathbf{b} = (b_i)_{i=1}^{m}.$$

Для дальнейшего нам будет удобно ввести в рассмотрение так называемую *каноническую задачу* линейного программирования, в которой ограничения (2) и (3) имеют специальный вид.

Каноническая (стандартная) задача линейного программирования состоит в нахождении минимума функции $f(\mathbf{x})$ при наличии ограничений-равенств

$$g_k(\mathbf{x}) = a_k, \quad k = 1, \ldots, l,$$

и ограничений-неравенств, состоящих в требовании неотрицательности переменных x_j. Последнее записывается так

$$\mathbf{x} \geqslant 0. \tag{5}$$

Ограничения (5) являются естественными в большинстве прикладных задач, допускающих постановку в виде задачи линейного программирования, и потому их целесообразно выделить в явном виде из общих ограничений-неравенств (3). Ограничения же (3), не совпадающие с (5), могут быть легко исключены из рассмотрения, как показывают приводимые ниже рассуждения.

Действительно, пусть i-ое из ограничений (3) имеет вид

$$f_{i1}x_1 + f_{i2}x_2 + \ldots + f_{in}x_n \gtrless b_i. \tag{6}$$

Рассмотрим дополнительные переменные $\xi_1, \xi_2, \ldots, \xi_p$ (по числу ограничений-неравенств, не совпадающих с ограничениями вида (5)) такие, что

$$f_{i1}x_1 + f_{i2}x_2 + \ldots + f_{in}x_n + h_i\xi_i = b_i.$$

При этом, если в (6) был знак «$\geqslant$» — переменная ξ_i вводится с коэффициентом $h_i = -1$, в противном случае $h_i = 1$. В обоих случаях переменные ξ_i удовлетворяют условию неотрицательности. Далее, поскольку может оказаться, что не все «старые» переменные (x_i) удовлетворяют условию неотрицательности, рассмотрим еще одну группу дополнительных переменных $\eta_1, \eta_2, \ldots, \eta_q$ (q — число «старых» переменных, не удовлетворяющих условию неотрицательности) таких, что

$$x_i = x_i^+ - \eta_i,$$

где $x_i^+ \geqslant 0$ и $\eta_i \geqslant 0$.

При помощи двух этих простых приемов общая задача линейного программирования может быть всегда сведена к канонической задаче, но уже в пространстве более высокой размерности.

Это делает каноническую задачу стандартной формой постановки общей задачи линейного программирования.

Всюду в дальнейшем мы будем рассматривать задачу линейного программирования в канонической постановке, предполагая дополнительно, что правые части ограничений (2) неотрицательны: $a_j \geqslant 0$. Если это не так, то для того, чтобы добиться выполнения указанного условия, достаточно умножить соответствующее соотношение на -1.

§ 2. Геометрия множества ограничений. Терминология

Рассмотрим в пространстве $\mathbb{R}^n$ множество M, описываемое ограничениями (2), (5) § 1 канонической задачи линейного программирования

$$M = \{\mathbf{x} \in \mathbb{R}^n:\ g_k(\mathbf{x}) = a_k,\ k = 1, 2, \ldots, l,\ x_j \geqslant 0,\ j = 1, 2, \ldots, n\}. \tag{1}$$

Заметим, что это множество выпукло.

◀ Действительно, пусть $\mathbf{x}_1, \mathbf{x}_2 \in M$ и $\lambda+\mu=1,\ \lambda \geqslant 0,\ \mu \geqslant 0$. Тогда для $\mathbf{x}=\lambda\mathbf{x}_1+\mu\mathbf{x}_2$ имеем $\mathbf{x} \geqslant 0$ и

$$\mathrm{G}\mathbf{x}=\lambda\mathrm{G}\mathbf{x}_1+\mu\mathrm{G}\mathbf{x}_2=\lambda\mathbf{a}+\mu\mathbf{a}=\mathbf{a},$$

т. е. $\mathbf{x} \in M$. ▶

Определение. *Крайней точкой (вершиной) множества ограничений* M называется точка $\mathbf{x}^*$, не представимая в виде выпуклой линейной комбинации других точек этого множества.

Другими словами, точка множества M будет крайней, если не существует отрезка, целиком лежащего в M и такого, что $\mathbf{x}^*$ — его внутренняя точка.

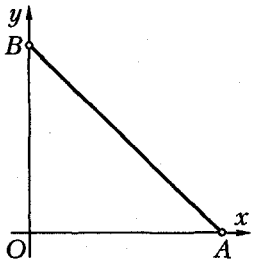


Рис. 1

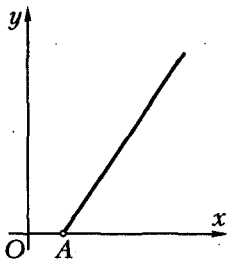


Рис. 2

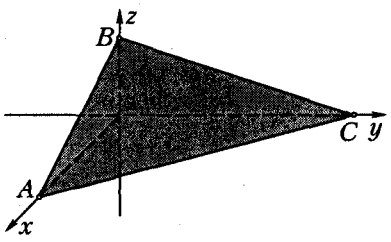


Рис. 3

Примеры. Рассмотрим в пространстве $\mathbb{R}^2$ множество M, задаваемое соотношениями

$$x+y=1, \quad x \geqslant 0, \quad y \geqslant 0$$

(рис. 1). Крайними точками этого множества будут точки A(1, 0) и B(0, 1).

Для множества M в $\mathbb{R}^2$, задаваемого соотношениями

$$x-y=1, \quad x \geqslant 0, \quad y \geqslant 0$$

(рис. 2), крайняя точка — A(1, 0).

Для

$$M \in \mathbb{R}^3, \quad M=\{(x, y, z):\ 2x+y+3z=1,\ x \geqslant 0,\ y \geqslant 0,\ z \geqslant 0\}$$

крайние точки — $A\left(\frac{1}{2}, 0, 0\right)$, $B\left(0, 0, \frac{1}{3}\right)$ и $C(0, 1, 0)$ (рис. 3). ▶

Крайние точки играют главную роль в рассматриваемой оптимизационной задаче. Как мы увидим, именно они представляют интерес в качестве возможных кандидатов на роль точек минимума.

Следующие теоремы дают исчерпывающее описание крайних точек множества ограничений M.

Теорема 1. *Точка* $\mathbf{x}^* \in M$ *является крайней точкой этого множества тогда и только тогда, когда столбцы матрицы* G, *отвечающие ненулевым компонентам* $\mathbf{x}^*$, *линейно независимы.*

◀ **Достаточность.** Пусть $\mathbf{x}^*$ — точка, удовлетворяющая ограничениям (1) и пусть $x_1^*, x_2^*, \ldots, x_k^*$ — ее ненулевые (т. е. положительные!) компоненты (без ограничения общности можно считать, что ненулевыми являются первые k компонент $\mathbf{x}^*$). Запишем ограничения-равенства из (1) в виде

$$\mathrm{G}\mathbf{x}^* = x_1^*\mathbf{g}_1 + x_2^*\mathbf{g}_2 + \ldots + x_k^*\mathbf{g}_k = \mathbf{a}, \tag{2}$$

где столбцы $\mathbf{g}_1, \ldots, \mathbf{g}_k$ матрицы G, отвечающие ненулевым компонентам точки $\mathbf{x}^*$, линейно независимы по условию теоремы. Если точка $\mathbf{x}^*$ не является крайней точкой множества M, то найдутся точки $\mathbf{x}_1, \mathbf{x}_2 \in M$, такие, что

$$\mathbf{x}^* = \frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2.$$

В силу условий $\mathbf{x}_1 \geqslant 0$ и $\mathbf{x}_2 \geqslant 0$ заключаем, что все компоненты этих точек с номерами, большими k, нулевые. Поскольку $\mathbf{x}_1 \in M$ и $\mathbf{x}_2 \in M$, для каждой них имеет место соотношение (2)

$$x_1^1\mathbf{g}_1 + \ldots + x_1^k\mathbf{g}_k = \mathbf{a}, \quad x_2^1\mathbf{g}_1 + \ldots + x_2^k\mathbf{g}_k = \mathbf{a},$$

что в силу линейной независимости векторов $\{\mathbf{g}_j\}_{j=1}^k$, возможно только если

$$x_1^j = x_2^j = x_j^*, \quad j = 1, 2, \ldots, k,$$

т. е. $\mathbf{x}^*$ — крайняя точка.

Необходимость. Пусть $\mathbf{x}^*$ — *крайняя точка множества* ограничений и пусть векторы $\{\mathbf{g}_j\}_{j=1}^k$ из соотношения (2) линейно зависимы. Тогда существует вектор $\boldsymbol{\lambda}$, не все компоненты которого нули, такой, что

$$\mathrm{G}\boldsymbol{\lambda} = \lambda_1\mathbf{g}_1 + \ldots + \lambda_k\mathbf{g}_k = \mathbf{0}. \tag{3}$$

Положим $\alpha > 0$ и, умножая обе части соотношения (3) на α, построим точки $\mathbf{x}_+$ и $\mathbf{x}_-$

$$\mathbf{x}_+ = \mathbf{x}^* + \alpha\boldsymbol{\lambda}, \quad \mathbf{x}_- = \mathbf{x}^* - \alpha\boldsymbol{\lambda}.$$

Отметим, что число α всегда можно выбрать так, чтобы все компоненты точек $\mathbf{x}_+$ и $\mathbf{x}_-$ были неотрицательны. Для этого достаточно взять $\alpha < \min\limits_{(1 \leqslant i \leqslant k)} \frac{x_i^*}{|\lambda_i|}$. Далее

$$\mathrm{G}\mathbf{x}_\pm = \mathrm{G}\mathbf{x}^* \pm \alpha\mathrm{G}\boldsymbol{\lambda} = \mathrm{G}\mathbf{x}^* = \mathbf{a}$$

и, следовательно, $\mathbf{x}_\pm \in M$. *Наконец, легко видеть,* что

$$\mathbf{x}^* = \frac{1}{2}\mathbf{x}_+ + \frac{1}{2}\mathbf{x}_-.$$

Но последнее соотношение противоречит тому, что точка $\mathbf{x}^*$ — крайняя точка M. ▶

Теорема 2. *Если функция $f(\mathbf{x})$ на множестве ограничений M достигает своего наименьшего значения, то она принимает то же значение в одной из крайних точек этого множества.*

◀ Пусть $\mathbf{x}_0 = \operatorname{argmin} f(\mathbf{x})\big|_{\mathbf{x}\in M}$ не является крайней точкой множества ограничений M, и пусть $x_0^1, \ldots, x_0^k$ — ее ненулевые (т. е. положительные) компоненты. Построим точки $\mathbf{x}_+^1$ и $\mathbf{x}_-^1$

$$\mathbf{x}_\pm^1 = \mathbf{x}_0 \pm \alpha\boldsymbol{\lambda},$$

где $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^k$ — вектор коэффициентов из соотношения (3), а $\alpha > 0$ и такое, что точки $\mathbf{x}_\pm$ удовлетворяют ограничениям (1). Из того, что

$$f(\mathbf{x}_\pm^1) = f(\mathbf{x}_0) \pm \alpha f(\boldsymbol{\lambda}) \geqslant f(\mathbf{x}_0),$$

заключаем, что $f(\boldsymbol{\lambda}) = 0$ и, следовательно, значение минимизируемой функции $f(\mathbf{x})$ в построенных точках такое же, как и в точке $\mathbf{x}_0$.

Среди компонент вектора $\boldsymbol{\lambda}$ всегда найдется хотя бы одна положительная. Действительно, $\mathbf{x}_0$ крайней точкой не является, следовательно векторы $\mathbf{g}_i$ из соотношения (3) линейно зависимы и, значит, не все λ_i нули. Если они все отрицательны — умножим (3) на (-1). Возьмем

$$\alpha = \min_{(s:\ \lambda_s>0)} \frac{x_0^s}{\lambda_s}.$$

При этом значении α точка $\mathbf{x}_-^1$ удовлетворяет ограничениям (1) и имеет на одну ненулевую компоненту меньше, чем точка $\mathbf{x}_0$. Если точка $\mathbf{x}_-^1$ оказывается крайней, то доказательство завершено, если же нет, повторим выписанную процедуру, взяв за исходную точку $\mathbf{x}_-^1$, и придем к точке $\mathbf{x}_-^2$, которая обладает всеми свойствами точки $\mathbf{x}_0$, но имеет уже на две ненулевые компоненты меньше. Поскольку ненулевых компонент у точки $\mathbf{x}_0$ конечное число, процесс завершится в крайней точке множества M, где функция $f(\mathbf{x})$ принимает такое же значение, как и в исходной точке минимума $\mathbf{x}_0$. ▶

Заметим, что установив необходимые и достаточные условия того, что точка $\mathbf{x} \in M$ является крайней (теорема 1), мы еще не умеем отвечать на вопрос, а существуют ли вообще крайние точки у множества ограничений M.

Теорема 3. *Если множество ограничений M непусто, то у него существует по крайней мере одна крайняя точка.*

◀ Введем вспомогательные переменные $v_1, v_2, \ldots, v_l$ и рассмотрим пространство $\mathbb{R}^{n+l}$ размерности $n + l$ с переменными

$$\mathbf{w} = \begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ v_1 \\ \vdots \\ v_l \end{pmatrix}.$$

Сформулируем задачу линейного программирования

$$\varphi(\mathbf{w}) = v_1 + v_2 + \ldots + v_l \to \min, \tag{4}$$

$$\begin{cases} g_{11}x_1 + \ldots + g_{1n}x_n + v_1 = a_1, \\ g_{21}x_1 + \ldots + g_{2n}x_n + v_2 = a_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ g_{l1}x_1 + \ldots + g_{ln}x_n + v_l = a_l, \\ x_i \geqslant 0;\ v_j \geqslant 0. \end{cases} \tag{5}$$

Отметим следующие, очевидные для задачи (4)–(5) обстоятельства:

1. *Множество ограничений M_w, задаваемое соотношениями* (5), *не пусто.* Действительно

$$\mathbf{w}_a = \begin{pmatrix} 0 \\ \mathbf{a} \end{pmatrix} \in M_w.$$

2. *Точка $\mathbf{w}_a$ — крайняя точка множества ограничений M_w.* Действительно, если переписать соотношения (5) в форме (3), то мы получим

$$a_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \ldots + a_l \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{a},$$

откуда, в силу линейной независимости единичных векторов

$$\mathbf{e}_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad i = 1, \ldots, n+l,$$

из теоремы 1 следует искомое.

3. *Функция $\varphi(\mathbf{w})$ достигает на множестве M_w своего наименьшего значения и это значение равно нулю.* Действительно, в силу неотрицательности компонент v_j, $\varphi(\mathbf{w}) \geqslant 0$. По условию теоремы множество M непусто, поэтому существует по крайней мере одна точка $\mathbf{x} \in M$ и $\mathbf{w} = \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix} \in M_w$; имеем $\varphi(\mathbf{w}) = 0$.

По теореме 2 то же минимальное значение достигается и в одной из крайних точек множества M_w, например, в точке $\mathbf{w}^* = \begin{pmatrix} \mathbf{x}^* \\ \mathbf{v}^* \end{pmatrix}$. В силу $\mathbf{v}^* \geqslant 0$ и $\varphi(\mathbf{w}^*) = \varphi(\mathbf{v}^*) = 0$ заключаем, что $\mathbf{v}^* = 0$. При этом точка $\mathbf{x}^* \in M$ — крайняя точка множества M. Действительно, если это не так, то найдутся точки $\mathbf{x}_1$, $\mathbf{x}_2 \in M$ такие, что $\mathbf{x}^* = \frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2$. Отсюда немедленно следует, что $\mathbf{w}^* = \frac{1}{2}\mathbf{w}_1 + \frac{1}{2}\mathbf{w}_2$, где $\mathbf{w}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{0} \end{pmatrix}$, а это вступает в противоречие с тем фактом, что $\mathbf{w}^*$ — крайняя точка множества M_w. ▶

Заметим для дальнейшего, что задача (4)–(5) всегда имеет решение, но не обязательно $\varphi_{\min} = 0$. Это обстоятельство может служить индикатором пустоты множества ограничений исходной задачи (2) § 1–(1). Если оказалось, что в задаче (4)–(5) $\varphi_{\min} \neq 0$, то это влечет за собой пустоту множества M, описываемого ограничениями (1).

Подведем некоторые итоги изложенных рассмотрений.

Множество ограничений (1) является выпуклым множеством (может быть пустым, может быть неограниченным).

Если это множество не пусто, то, независимо от того, ограничено оно или нет, оно содержит *не более чем конечное число* крайних точек. В них может достигаться экстремум минимизируемой функции $f(\mathbf{x})$, если он есть. И если пути ответа на вопрос о пустоте множества ограничений M описаны выше, то ответа на вопрос, есть ли решение у рассматриваемой канонической задачи линейного программирования, у нас пока нет. Из общих соображений понятно, что эта задача может и не иметь решения — в случае неограниченности множества M вполне может оказаться, что $f(\mathbf{x})$ неограничена снизу и принимает на M сколь угодно большие по модулю отрицательные значения. Однако если на M функция $f(\mathbf{x})$ ограничена снизу, то экстремум достигается и, как уже было отмечено выше, в одной из крайних точек множества M.

Итак, для успешного решения задачи линейного программирования необходимо уметь

— находить крайние точки множества ограничений M,

— *разумно* сравнивать значения функции $f(\mathbf{x})$ в этих точках (как уже отмечалось выше, крайних точек конечное, но, вполне может быть, очень большое число, и поэтому полный перебор всех крайних точек не является эффективным средством выбора оптимальной (или оптимальных)),

— отвечать на вопрос о существовании решения задачи линейного программирования с тем, чтобы напрасно не тратить усилия на поиски несуществующего экстремума.

Решению этих проблем мы посвятим следующие разделы.

В заключение отметим, что исторически при исследовании и решении задач линейного программирования сложилась своя терминология, связанная с прикладной спецификой рассматриваемых проблем и несколько отличающаяся от принятой в теории экстремальных задач.

Сам термин «программирование» в экстремальных задачах связан с экономической природой задач, решаемых рассматриваемыми методами и является неудачным переводом английского термина «programming» — планирование. Отсюда и другие термины, прижившиеся в русскоязычной литературе по линейному программированию:

- *план* — любая точка множества ограничений M,
- *опорный план* —крайняя точка множества ограничений,
- *оптимальный план* — точка множества M, в которой функция $f(\mathbf{x})$ достигает своего минимума.

§ 3. Симплекс-метод решения задачи линейного программирования

3.1. Процедура перебора крайних точек множества ограничений

Рассмотрения § 2 позволили сделать вывод о конечности числа крайних точек множества M — оно не превышает числа линейно независимых систем среди столбцов

матрицы ограничений-равенств G. Если это число невелико, то не представляет труда перебрать все крайние точки с целью нахождения оптимальной. Однако в реальных прикладных задачах это число становится настолько большим, что полный перебор не под силу даже современным быстродействующим ЭВМ.

Возникает задача построения такого алгоритма перехода от одной крайней точки множества M к другой, который бы перебирал не все вершины, а только некоторые, при этом желательным свойством подобного перебора является свойство релаксации — каждая следующая точка должна быть «не хуже» предыдущей, т. е. после точки $\mathbf{x}_1$ со значением функции $f_1 = f(\mathbf{x}_1)$ алгоритм приводил бы нас в точку $\mathbf{x}_2$, такую, что $f_2 = f(\mathbf{x}_2) \leqslant f(\mathbf{x}_1)$.

В дальнейшем будем предполагать, что $r = l$ (линейно зависимые соотношения исключены из системы ограничений равенств (1) § 2) и $r < n$ (задача минимизации содержательна в том смысле, что множество ограничений M содержит «много» точек).

Крайнюю точку $\mathbf{x}$ множества M будем называть *невырожденной*, если число ее ненулевых компонент равно r.

Без ограничения общности можем считать, что это первые r ее компонент. В соотношении (2) § 2 им отвечает система линейно независимых векторов $\{\mathbf{g}_i\}_{i=1}^r$, образующих базис в $\mathbb{R}^r$. Поскольку $\mathbf{x} \in M$, имеем

$$x_j > 0, \quad j = 1, \ldots, r, \quad x_1\mathbf{g}_1 + \ldots + x_r\mathbf{g}_r = \mathbf{a}.$$

В силу базисности системы $\{\mathbf{g}_i\}_{i=1}^r$ всякий столбец матрицы G может быть представлен линейными комбинациями векторов $\mathbf{g}_i$

$$\forall\, j = r+1, \ldots, n \quad \exists\, \lambda_j = (\lambda_i^j)_{i=1}^r\colon\ \lambda_1^j\mathbf{g}_1 + \ldots + \lambda_r^j\mathbf{g}_r = \mathbf{g}_j.$$

Пусть $\alpha > 0$ — некоторое число. Вычитая из первого из приведенных соотношений второе, умноженное на α, получим

$$(x_1 - \alpha\lambda_1^j)\mathbf{g}_1 + \ldots + (x_r - \alpha\lambda_r^j)\mathbf{g}_r + \alpha\mathbf{g}_j = \mathbf{a}, \quad j = r+1, r+2, \ldots, n. \qquad (1)$$

Лемма. *Для всех $j = r+1, r+2, \ldots, n$, можно указать значение α такое, что точки $\mathbf{y}_j$ с компонентами $y_i^j = x_i - \alpha\lambda_i^j$, $i = 1, \ldots, r$, $y_j^j = \alpha$, $y_k = 0$, $r < k < n$, $k \neq j$, удовлетворяют ограничениям задачи линейного программирования* (1) § 2. *Если при этом хотя бы одно из значений $\lambda_i^j > 0$, то соответствующая точка $\mathbf{y}_j \neq \mathbf{x}$ — крайняя точка множества ограничений M.*

◀ Если все числа $\lambda_i^j \leqslant 0$, то утверждение очевидно, так как при любом $\alpha > 0$ соотношение (1) означает, что $\mathbf{y}_j$ удовлетворяет ограничениям-равенствам, а условие $y_i^j \geqslant 0$ выполняется автоматически. Отметим при этом, что ни одна из точек $\mathbf{y}_j$ не будет крайней точкой множества M, так как векторы $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r, \mathbf{g}_j$ — отвечающие ненулевым компонентам $\mathbf{y}_j$ линейно зависимы.

Если же среди чисел λ_i^j есть по крайней мере одно положительное, то выбрав

$$\alpha_s = \min \frac{x_i}{\lambda_i^j},$$

где минимум берется по всем тем значениям индексов i, для которых $\lambda_i^j > 0$, и достигается при некотором $i = s$, получим s-ю компоненту соответствующего $\mathbf{y}_j$ равной нулю

(может быть и не только s-ю). Система векторов $\mathbf{g}_1, \ldots, \mathbf{g}_{s-1}, \mathbf{g}_{s+1}, \ldots, \mathbf{g}_r, \mathbf{g}_j$ будет линейно независимой, ибо в противном случае мы наряду с разложением вектора $\mathbf{g}_j$ по указанной системе

$$\mathbf{g}_j = \lambda_1^j \mathbf{g}_1 + \ldots + \lambda_r^j \mathbf{g}_r$$

содержащей вектор $\mathbf{g}_s$ с ненулевым (!) коэффициентом λ_s^j, будем иметь разложение $\mathbf{g}_j$, не содержащее $\mathbf{g}_s$, что невозможно в силу единственности разложения по базису. ▶

Отметим, что приведенная выше лемма позволяет при некоторых условиях (наличие хотя бы одного $\lambda_i^s > 0$) перейти от *невырожденной* крайней точки $\mathbf{x}$ множества M к другой (может быть, уже вырожденной!) крайней точке этого множества. Если исходная точка $\mathbf{x}$ вырождена (некоторые из $x_i = 0$, $i = 1, 2, \ldots, r$), то лемма все равно остается справедливой в первой своей части. Однако во второй части уже нельзя утверждать, что $\mathbf{y}_j \neq \mathbf{x}$, и возможно так называемое зацикливание процедуры перебора (способов преодоления этой трудности существует много).

3.2. Пересчет значений минимизируемой функции

Пусть $\mathbf{x}$ — некоторая крайняя точка множества ограничений M и $f(\mathbf{x})$ — значение минимизируемой функции в этой точке. Пусть $\mathbf{y}_j$ — точка, существование которой гарантируется леммой п. 3.1,

$$\mathbf{y} = \begin{pmatrix} x_1 - \alpha\lambda_1^j \\ \vdots \\ x_r - \alpha\lambda_r^j \\ 0 \\ \vdots \\ 0 \\ \alpha \\ 0 \end{pmatrix}.$$

Очевидно, что

$$f(\mathbf{y}_j) = f(\mathbf{x}) - \alpha \cdot f(\lambda_j) + \alpha \cdot c_j = f(\mathbf{x}) - \alpha[f(\lambda_j) - c_j], \tag{2}$$

где $j = r+1, r+2, \ldots, n$.

Соотношение (2) показывает, как изменяется значение минимизируемой функции f при переходе от точки $\mathbf{x}$ к точке $\mathbf{y}_j$.

Сразу отметим, что если найдется хотя бы одно значение j_0, для которого $f(\lambda_{j_0}) - c_{j_0} > 0$, то значение функции f в точке $\mathbf{y}_{j_0}$ «лучше», чем в исходной точке $\mathbf{x}$, так как в силу $\alpha > 0$ $f(\mathbf{y}_{j_0}) < f(\mathbf{x})$. При этом возможны следующие две ситуации (см. лемму):

1. $\mathbf{y}_{j_0}$ — точка множества M, не являющаяся крайней. При этом все $\lambda_i^{j_0} \leqslant 0$ и, следовательно, α могут принимать неограниченно большие значения. Функция f в точках M в этом случае принимает сколь угодно большие по модулю отрицательные значения. *Экстремума в рассматриваемой задаче линейного программирования нет.*
2. $\mathbf{y}_{j_0}$ — крайняя точка множества M. При этом $\alpha > 0$ и $f(\mathbf{y}_{j_0}) < f(\mathbf{x})$, и, следовательно, мы перешли из крайней точки $\mathbf{x}$ в «лучшую» $\mathbf{y}_{j_0}$.

Если же для всех значений j разности $f(\lambda_j) - c_j \leqslant 0$, то имеет место следующая теорема.

Теорема об оптимальной крайней точке. *Если* $\forall\, j = r+1, \ldots, n \;\; f(\lambda_j) - c_j \leqslant 0$, *то* $\mathbf{x} = \operatorname{argmin} f\big|_M$.

◀ Действительно, если $\mathbf{y} \in M$ — произвольная точка множества ограничений задачи линейного программирования, то

$$y_1\mathbf{g}_1 + \ldots + y_n\mathbf{g}_n = \mathbf{a}.$$

Так как $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r$ — линейно независимый набор столбцов матрицы G, отвечающих ненулевым компонентам крайней точки $\mathbf{x}$, то имеют место соотношения

$$\mathbf{g}_j = \lambda_1^j\mathbf{g}_1 + \ldots + \lambda_r^j\mathbf{g}_r, \tag{3}$$

где $j = r+1, r+2, \ldots, n$.

Подставляя их в предыдущее равенство, получим

$$\mathbf{a} = \sum_{i=1}^{r} y_i\mathbf{g}_i + \sum_{j=r+1}^{n} y_j\mathbf{g}_j = \sum_{i=1}^{r} y_i\mathbf{g}_i + \sum_{j=r+1}^{n} y_j \sum_{i=1}^{r} \lambda_i^j\mathbf{g}_j = \sum_{i=1}^{r} \left(y_i + \sum_{j=r+1}^{n} y_j\lambda_i^j\right)\mathbf{g}_j.$$

Сравнивая последнее соотношение с равенством (2) § 2, заключаем, что в силу линейной независимости векторов $\{\mathbf{g}_i\}_{i=1}^r \;\forall\, i = 1, 2, \ldots, r$, имеет место равенство

$$y_i + \sum_{j=r+1}^{n} y_j\lambda_i^j = x_i.$$

Вычислим значение минимизируемой функции $f(\mathbf{x})$ в точке $\mathbf{x}$

$$f(\mathbf{x}) = \sum_{i=1}^{r} c_i x_i = \sum_{i=1}^{r} c_i\left(y_i + \sum_{j=r+1}^{n} y_j\lambda_i^j\right) = \sum_{i=1}^{r} c_i y_i + \sum_{i=r+1}^{n} y_j \sum_{i=1}^{r} c_i\lambda_i^j.$$

Заметим, что $f(\lambda_j) = \sum\limits_{i=1}^{r} c_i\lambda_i^j \leqslant c_j$. Поэтому

$$f(\mathbf{x}) = \sum_{i=1}^{r} c_i y_i + \sum_{j=r+1}^{n} y_j f(\lambda_j) \leqslant \sum_{i=1}^{n} c_i y_i = f(\mathbf{y}).$$

Последнее соотношение завершает доказательство. ▶

3.3. Последовательность вычислений. Симплекс-таблицы

Суммируя изложенное в предыдущих разделах, приходим к следующей последовательности вычислений, приводящей к ответу на вопрос об оптимуме в задаче линейного программирования (1)–(2)–(5) § 1.

1. Фиксируем некоторую точку $\mathbf{x}_1$, являющуюся крайней точкой множества ограничений (2)–(5) § 1.

   Например, можно применить метод введения вспомогательных переменных, описанный в § 2, теорема 3. Одновременно при этом будет установлено, не является ли множество ограничений пустым.

2. В соответствии с соотношением (1) и леммой пункта 3.1 из крайней точки $\mathbf{x}_1$ переходим к другой крайней точке $\mathbf{x}_2$. Если хотя бы одно из чисел $\lambda_i^j > 0$ и $f(\lambda_j) -$

$c_j > 0$, то точка $\mathbf{x}_2$ оказывается «лучше» точки $\mathbf{x}_1$ и принимаем ее за исходную. Если же все $\lambda_i^j \leqslant 0$ и $f(\lambda_j) - c_j > 0$, то оптимума нет.

3. Наконец, если оказывается, что для всех j $f(\lambda_j) - c_j \leqslant 0$, то, в соответствии с теоремой пункта 3.2, точка, в которой это произошло, является искомой точкой минимума. Поскольку крайних точек конечное число, за конечное число шагов мы или найдем оптимум, или установим, что исследуемая задача решений не имеет.

Рассмотренный выше алгоритм легко реализуем, при этом проблемы, связанные с зацикливанием, на практике легко могут быть обойдены.

Сердцевиной алгоритма является процедура пункта **2** перехода от одной крайней точки к другой, на которой мы остановимся подробнее. Если $\mathbf{x}_1$ — исходная крайняя точка и по крайней мере одно из чисел $\lambda_i^j > 0$, то полагая $\alpha = \min(x_i/\lambda_i^j)$, где минимум берется по всем положительным λ_i^j, мы получаем возможность указать другую крайнюю точку $\mathbf{y}$, компоненты которой даются соотношениями

$$\mathbf{y} = (x_1 - \alpha\lambda_1^q, \ldots, x_{p-1} - \alpha\lambda_{p-1}^q, \underset{\substack{\downarrow \\ p\text{-я позиция}}}{0}, x_{p+1} - \alpha\lambda_{p+1}^q, 0, \ldots, 0, \underset{\substack{\downarrow \\ q\text{-я позиция}}}{\alpha}, 0, \ldots, 0)$$

в предположении, что указанный выше минимум достигается при $i = p$ и $j = q$. Базис, соответствующий точке $\mathbf{y}$, получен из базиса, отвечающего точке $\mathbf{x}$, удалением вектора $\mathbf{g}_p$ и включением вместо него вектора $\mathbf{g}_q$. Для продолжения процесса минимизации необходимо найти новые значения коэффициентов $\widehat{\lambda}_i^j$, т. е. разложение столбцов матрицы ограничений по новому базису. Заметим, что для $\mathbf{g}_q$ имеет место представление

$$\mathbf{g}_q = \lambda_1^q\mathbf{g}_1 + \ldots + \lambda_p^q\mathbf{g}_p + \ldots + \lambda_r^q\mathbf{g}_r.$$

Отсюда видно, что

$$\mathbf{g}_p = \frac{\mathbf{g}_q - \lambda_1^q\mathbf{g}_1 - \ldots - \lambda_{p-1}^q\mathbf{g}_{p-1} - \lambda_{p+1}^q\mathbf{g}_{p+1} - \ldots - \lambda_r^q\mathbf{g}_r}{\lambda_p^q}.$$

Заменяя теперь в разложениях (3) вектор $\mathbf{g}_p$ правой частью последнего соотношения, получим

$$\mathbf{g}_{r+i} = \left(\lambda_1^i - \frac{\lambda_p^i}{\lambda_p^q}\lambda_1^q\right)\mathbf{g}_1 + \ldots + \left(\lambda_{p-1}^i - \frac{\lambda_p^i}{\lambda_p^q}\lambda_{p-1}^q\right)\mathbf{g}_{p-1} + \ldots + \frac{\lambda_p^i}{\lambda_p^q}\mathbf{g}_q. \qquad (*)$$

Аналогично пересчитывается значение целевой функции в точке $\mathbf{y}$.

Как правило, изложенная последовательность вычислений организуется как процедура преобразования так называемых симплекс-таблиц, соответствующих рассматриваемой задаче.

Пусть исходная матрица ограничений G такова, что первые ее r столбцов образуют единичную матрицу (если это не так — введем вспомогательные переменные или перенумеруем имеющиеся). В этом случае в качестве исходной крайней точки, с которой начинается процесс минимизации, можно взять точку $\mathbf{x}_1 = \mathbf{a}$, компоненты которой суть правые части ограничений $\mathbf{Gx} = \mathbf{a}$.

Исходная симплекс-таблица строится следующим образом (см. табл. 1): матрица коэффициентов системы ограничений-равенств задачи линейного программирования дополняется двумя строками — верхней и нижней — и левым столбцом, которые содержат соответственно:

Таблица 1

	c_1	c_2	...	c_r	c_{r+1}	...	c_{n-1}	c_n	a
1	1	0	...	0	λ_1^{r+1}	...	λ_1^{n-1}	λ_1^n	a_1
2	0	1	...	0	λ_2^{r+1}	...	λ_2^{n-1}	λ_2^n	a_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	0	0	...	1	λ_r^{r+1}	...	λ_r^{n-1}	λ_r^n	a_r
	0	0	0	0	$z_{r+1} - c_{r+1}$		$z_{n-1} - c_{n-1}$	$z_n - c_n$	$f(\mathbf{a})$

- верхняя строка — строку коэффициентов c_i, $i = 1, 2, \ldots, n$, минимизируемой функции $f(x)$;
- нижняя строка — значения $f(\mathbf{g}_i) - c_i$, где $\mathbf{g}_i$ — i-й столбец системы ограничений равенств;
- левый столбец — номера базисных столбцов матрицы ограничений.

Для сокращения записи в таблице приняты следующие обозначения:

$$f(\mathbf{g}_j) = z_j = \lambda_1^j c_1 + \lambda_2^j c_2 + \ldots + \lambda_r^j c_r,$$
$$f(\mathbf{a}) = a_1 c_1 + a_2 c_2 + \ldots + a_r c_r.$$

Пример 1. Рассмотрим задачу

$$f(\mathbf{x}) = -3x_1 + 2x_2 - 2x_3 + 2x_4 - x_5 \to \min,$$
$$\begin{cases} x_1 + x_2 - x_3 = 1, \\ -x_2 + x_3 + x_4 = 1, \\ x_2 + x_3 + x_5 = 2, \\ x_j \geqslant 0, \quad j = 1, 2, \ldots, 5. \end{cases}$$

◀ Ее симплекс-таблица будет иметь следующий вид (табл. 2).

Таблица 2

	−3	2	−2	2	−1	x_1
1	1	1	−1	0	0	1
4	0	−1	1	1	0	1
5	0	1	1	0	1	2
	0	−8	6	0	0	−3

Отметим, что мы не стали перенумеровывать переменные так, чтобы первые r столбцов образовывали единичную матрицу. Поэтому крайний левый столбец содержит не номера 1, 2, 3, как это было бы, если бы мы произвели перенумерацию, а номера 1, 4 и 5, отвечающие базисным столбцам матрицы системы ограничений G. Нижняя строка таблицы 2 заполняется следующим образом — вектор коэффициентов $\mathbf{c}$ скалярно умножается на вектор, у которого в позициях, отвечающих базисным (в нашем случае — 1, 4, 5) стоят компоненты соответствующего столбца матрицы G, а в остальных позициях — нули, и из результата вычитается c_j. Число −8, например, получается так:

$$-8 = \big[1 \cdot (-3) + 0 \cdot 2 + 0 \cdot (-2) + (-1) \cdot 2 + 1 \cdot (-1)\big] - 2.$$

Аналогично заполнены остальные позиции нижней строки, кроме последней. Число, стоящее в нижнем правом углу таблицы, представляет собой значение минимизируемой функции в исходной крайней точке:

$$f(\mathbf{a}) = 1 \cdot (-3) + 0 \cdot 2 + 0 \cdot (-2) + 1 \cdot 2 + 2 \cdot (-1) = -3.$$

Просмотр нижней строки определяет дальнейшие действия по пересчету таблицы. Если все элементы в этой строке неположительны — точка $\mathbf{x}_1$ является точкой минимума. В противном случае следует отыскать в таблице положительные компоненты, отвечающие положительным значениям $z_j - c_j$. В рассматриваемом примере четвертый столбец ($q = 3$) содержит положительную разность $z_3 - c_3 = 6$. Это означает, что третий столбец матрицы ограничений должен быть включен в число базисных. Столбец, подлежащий исключения из базиса, должен теперь быть определен из условия $x_i/\lambda_i^q \to \min\limits_i$.

В нашем случае имеем

$$\frac{x_4}{\lambda_4^3} = \frac{1}{1} = 1, \quad \frac{x_5}{\lambda_5^3} = \frac{2}{1} = 2,$$

и, следовательно, $p = 4$, т.е. исключению подлежит четвертый столбец. Теперь следует произвести пересчет симплексной таблицы. Это делается (в соответствии с формулами (*)) так: строка симплексной таблицы, отвечающая номеру столбца, который подлежит исключению, заменяется строкой, все элементы которой получены делением на ключевой элемент λ_p^q. В нашем случае $\lambda_4^3 = 1$, т.е. строка, помеченная номером 4 в нашей таблице переходит в новую таблицу, изменив только свой номер. Новый номер этой строки будет равен q, в рассматриваемом случае 3 (табл. 3). Другие строки сохраняют свои номера и преобразуются путем умножения полученной строки на элемент выбранного столбца, отвечающего преобразуемой строке, и вычитания результата из преобразуемой строки.

Таблица 3

	−3	2	−2	2	−1	$\mathbf{x}_2$
1						
3	0	−1	1	1	0	1
5						

Таблица 4

	−3	2	−2	2	−1	$\mathbf{x}_2$
1	1	0	0	1	0	2
3	0	−1	1	1	0	1
5	0	2	0	−1	1	1
	0	−6	0	−4	0	**−9**

В нашем случае первой строке отвечает множитель -1, последней — 1. Преобразование этих строк дает следующий результат

$$n = 1 : (1\ 1\ -1\ 0\ 0\ 1) - (-1)(0\ -1\ 1\ 1\ 0\ 1) = (1\ 0\ 0\ 1\ 0\ 2),$$
$$n = 5 : (0\ 1\ 1\ 0\ 1\ 2) - 1 \cdot (0\ -1\ 1\ 1\ 0\ 1) = (0\ 2\ 0\ -1\ 1\ 1).$$

Новая симплексная таблица — табл. 4.

Последняя строка получена при помощи той же последовательности вычислений, что и при построении исходной симплекс-таблицы. Поскольку все элементы последней строки $z_j - c_j$ удовлетворяют условию неположительности, процедура минимизации закончена. Вектор $\mathbf{x}_2 = \operatorname{argmin} f(\mathbf{x})|_G$ дается соотношением

$$\mathbf{x}_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix};$$

$f(\mathbf{x}_2) = -9$ — наименьшее значение минимизируемой функции. ▶

Рассмотрим еще два примера, иллюстрирующих процедуру преобразования симплекс-таблиц.

Пример 2.

$$f(\mathbf{x}) = -6x_1 - 8x_2 \to \min,$$
$$\begin{cases} 2x_1 + 5x_2 + x_3 = 20, \\ 12x_1 + 6x_2 + x_4 = 72, \\ x_j \geqslant 0, \quad j = 1, \ldots, 4. \end{cases}$$

◀ Исходная симплексная таблица — табл. 5.

Таблица 5

	−6	−8	0	0	x_1
3	2	5	1	0	20
4	12	6	0	1	72
	6	8	0	0	0

Таблица 6

	−6	−8	0	0	x_2
2	$\frac{2}{5}$	1	$\frac{1}{5}$	0	4
4	$\frac{48}{5}$	0	$-\frac{6}{5}$	1	48
	$\frac{14}{5}$	0	$-\frac{8}{5}$	0	−32

Таблица 7

	−6	−8	0	0	x_3
2	0	1	$\frac{1}{4}$	$-\frac{1}{24}$	2
1	1	0	$-\frac{1}{8}$	$\frac{5}{48}$	5
	0	0	$-\frac{5}{4}$	$-\frac{7}{24}$	−46

Элементарный подсчет $\left(\frac{20}{5} = 4 < \frac{20}{2},\ \frac{72}{12} < \frac{72}{6}\right)$ показывает, что включению в базис подлежит столбец 2, а исключению — столбец 3. Преобразованная симплекс-таблица — табл. 6.

Наличие в последней строке положительного элемента свидетельствует о возможности дальнейшего «улучшения» минимизируемой функции. Включению в базис подлежит первый столбец, исключению — четвертый $\left(4 : \frac{2}{5} = 10 > 48 : \frac{48}{5} = 5\right)$. Таблица 7 является итоговой симплекс-таблицей, так как выполнен критерий оптимальности: $z_j - c_j \leqslant 0$.

Оптимальное решение — вектор $x_3 = \begin{pmatrix} 5 \\ 2 \\ 0 \\ 0 \end{pmatrix}$, наименьшее значение — $f(x_3) = -46$. ▶

Пример 3.

$$f(x) = -x_1 - 4x_4 \to \min,$$

$$\begin{cases} -x_1 - 2x_2 + 2x_3 + x_4 + 5x_5 = 13, \\ -2x_1 + 2x_2 + 4x_4 + x_5 = 5, \\ x_1 - x_2 + x_3 - x_4 + 2x_5 = 5, \\ x_j \geqslant 0, \quad j = 1, \ldots, 5. \end{cases}$$

◀ На этом примере мы проиллюстрируем прием, связанный с введением искусственных переменных.

Рассмотрим вспомогательную задачу линейного программирования

$$f(w) = f(x, u) = u_1 + u_2 + u_3 \to \min,$$

Таблица 8

	1	1	1	0	0	0	0	0	w_1
1	1	0	0	−1	−2	2	1	5	13
2	0	1	0	−2	2	0	4	1	5
3	0	0	1	1	−1	1	−1	2	5
	0	0	0	−2	−1	3	4	8	23

$$\begin{cases} u_1 - x_1 - 2x_2 + 2x_3 + x_4 + 5x_5 = 13, \\ u_2 - 2x_1 + 2x_2 + 4x_4 + x_5 = 5, \\ u_3 + x_1 - x_2 + x_3 - x_4 + 2x_5 = 5, \\ x_j \geqslant 0, \quad j = 1, \ldots, 5, \qquad u_i \geqslant 0, \quad i = 1, 2, 3. \end{cases}$$

Для этой задачи исходная симплекс-таблица будет иметь следующий вид (табл. 8). Элементарный подсчет дает следующий результат — исключению подлежит второй (искусственный!) вектор базиса, включению — седьмой (он же четвертый столбец в исходной матрице ограничений).

Преобразованная симплекс-таблица — табл. 9.

Таблица 9

	1	1	1	0	0	0	0	0	w_2
1	1	$-\frac{1}{4}$	0	$-\frac{1}{2}$	$-\frac{5}{2}$	2	0	$\frac{19}{4}$	$\frac{47}{4}$
7	0	$\frac{1}{4}$	0	$-\frac{1}{2}$	$\frac{1}{2}$	0	1	$\frac{1}{4}$	$\frac{5}{4}$
3	0	$\frac{1}{4}$	1	$\frac{1}{2}$	$-\frac{1}{2}$	1	0	$\frac{9}{4}$	$\frac{25}{4}$
	0	−1	0	0	−3	3	0	7	18

Здесь ключевой элемент $\lambda_1^5 = \frac{19}{4}$ — исключаем первый вектор искусственного базиса, включаем — восьмой (он же пятый столбец исходной матрицы).

Очередная симплекс-таблица дана в табл. 10.

Таблица 10

	1	1	1	0	0	0	0	0	w_3
8	$\frac{4}{19}$	$-\frac{1}{19}$	0	$-\frac{2}{19}$	$-\frac{10}{19}$	$\frac{8}{19}$	0	1	$\frac{47}{19}$
7	$-\frac{1}{19}$	$\frac{5}{19}$	0	$-\frac{9}{19}$	$\frac{12}{19}$	$-\frac{2}{19}$	1	0	$\frac{12}{19}$
3	$-\frac{9}{19}$	$\frac{7}{19}$	1	$\frac{14}{19}$	$\frac{13}{19}$	$\frac{1}{19}$	0	0	$\frac{13}{19}$
	$-\frac{28}{19}$	$-\frac{12}{19}$	0	$\frac{14}{19}$	$\frac{13}{19}$	$\frac{1}{19}$	0	0	$\frac{13}{19}$

Ключевой элемент — λ_3^4. На этом этапе из базиса исключается последний остающийся в нем искусственный вектор и включается первый столбец матрицы ограничений. Симплекс-таблица, отвечающая этим преобразованиям, приведена ниже (табл. 11). Следует обратить внимание на то, что первая строка, содержащая коэффициенты минимизируемой функции, на этом этапе изменяется — коэффициенты вспомогательной задачи заменяются коэффициентами исходной.

Таблица 11

	0	0	0	−1	0	0	−4	0	x_1
8 (5)	*	*	*	0	$-\frac{3}{7}$	$\frac{3}{7}$	0	1	$\frac{18}{7}$
7 (4)	*	*	*	0	$\frac{15}{14}$	$-\frac{1}{14}$	1	0	$\frac{15}{14}$
4 (1)	*	*	*	1	$\frac{13}{14}$	$\frac{1}{14}$	0	0	$\frac{13}{14}$
	*	*	*	0	$-\frac{73}{14}$	$\frac{3}{14}$	0	0	$-\frac{73}{14}$

Таблица 12

	...	...	...	−1	0	0	−4	0	x_2
3				0	−1	1	0	$\frac{7}{3}$	6
4				0	1	0	1	$\frac{1}{6}$	$\frac{3}{2}$
1				1	1	0	0	$-\frac{1}{6}$	$\frac{1}{2}$
				0	−5	0	0	$-\frac{3}{6}$	$-\frac{13}{2}$

Весь блок данных, отвечающих искусственным переменным, помечен звездочками, так как в дальнейших вычислениях эти данные не участвуют. В крайнем левом столбце в скобках указана нумерация базисных векторов как столбцов исходной матрицы ограничений. Дальнейшие вычисления дают: ключевой элемент — λ_1^3, и очередную симплекс-таблицу (табл. 12).

На ней, в соответствии с критерием оптимальности (все элементы $z_j - c_j \leqslant 0$) вычисления заканчиваются

$$\operatorname{argmin} f(\mathbf{x})|_G = \mathbf{x}_2, \quad \mathbf{x}_2 = \begin{pmatrix} \frac{1}{2} \\ 0 \\ 6 \\ \frac{3}{2} \\ 0 \end{pmatrix}, \quad f(\mathbf{x}_2) = -\frac{13}{2}. \blacktriangleright$$

ВЫЧИСЛИТЕЛЬНАЯ МАТЕМАТИКА

Развитие вычислительной техники, ее доступность и кажущаяся простота применения резко расширили возможности исследователей в разрешении сложных прикладных задач. Однако эффективное использование современного парка компьютеров невозможно без осознанного владения основами численных методов и идеологией их приложения к решению конкретных научных и инженерно-технических проблем. Вычисления сами по себе не представляют особой ценности без осмысления их результативности и адекватности изучаемому процессу.

Авторы не ставили себе целью сделать читателя этой книги квалифицированным вычислителем — алгоритмистом или программистом — наша цель скромнее: мы хотим показать пользователю, как следует ставить задачу численного анализа, чего следует опасаться при реализации вычислительного процесса и как грамотно интерпретировать полученные результаты.

Начнем изложение с обзора основных понятий, связанных с вычислениями.

Глава LVI

ПОГРЕШНОСТИ ВЫЧИСЛЕНИЙ

§ 1. Погрешности

Пусть a — некоторое число, $\tilde{a}$ — другое число, в некотором смысле близкое к первому и заменяющее его в расчетах (типичные примеры: $a = \pi$, $\tilde{a} = 3{,}14$, или $a = \sqrt{2}$, $\tilde{a} = 1{,}41$ и т. п.).

Абсолютной погрешностью Δa представления числа a числом $\tilde{a}$ назовем модуль разности между точным (a) и приближенным ($\tilde{a}$) значениями

$$\Delta a = |a - \tilde{a}|.$$

Предельной абсолютной погрешностью назовем положительное число Δ_a такое, что

$$\Delta a \leqslant \Delta_a. \tag{1}$$

Из определения видно, что предельная абсолютная погрешность указывает границы отличия a от $\tilde{a}$: точное (но, может быть, неизвестное) значение a заключено в пределах от $\tilde{a} - \Delta_a$ до $\tilde{a} + \Delta_a$. Этот факт обычно коротко фиксируют так

$$a = \tilde{a} \pm \Delta_a.$$

Ясно, что погрешность (абсолютная погрешность, предельная абсолютная погрешность) не очень хороший показатель точности замены числа a его приближением $\tilde{a}$: предельная абсолютная погрешность в 1 км при измерении расстояния от Москвы до Челябинска — это блестящий результат, в то время как та же величина абсолютной предельной погрешности при измерении длины кремлевской стены является уже довольно низким показателем точности измерений.

Введем в рассмотрение понятия *относительной* и *предельной относительной* погрешностей, позволяющие учесть масштаб величины при оценке точности ее замены приближенным значением.

Относительной погрешностью $\delta(a)$ представления числа a числом $\tilde{a}$ назовем отношение

$$\delta(a) = \frac{\Delta a}{|\tilde{a}|}, \quad a \neq 0,$$

Предельной относительной погрешностью назовем положительное число δ_a такое, что

$$\delta(a) = \frac{\Delta a}{|\tilde{a}|} \leqslant \delta_a. \tag{2}$$

Предельная относительная погрешность указывает относительные границы отличия a от $\tilde{a}$: точное (но, может быть, неизвестное) значение a заключено в пределах от $\tilde{a}(1-\delta_a)$ до $\tilde{a}(1+\delta_a)$. Этот факт обычно коротко фиксируют так

$$a = \tilde{a}(1 \pm \delta_a).$$

Определение предельной относительной погрешности позволяет установить связь между предельной абсолютной и предельной относительной погрешностями

$$\Delta_a = |\tilde{a}|\delta_a.$$

Тесно связанным с понятием *погрешности* является понятие *значащей цифры* числа.

Напомним, что всякое действительное число может быть представлено десятичной дробью, конечной или бесконечной. Позиционная форма записи действительного числа предполагает, что значение каждого десятичного знака определяется его местом: если число a записано в виде

$$a = \overline{a_n a_{n-1} \dots a_2 a_1 a_0, a_{-1} a_{-2} \dots a_{-m} \dots}, \tag{3}$$

где a_j, $j = 0, \pm 1, \pm 2, \dots$, — цифры 0, 1, ..., 9, то это означает, что

$$a = 10^n \cdot a_n + \dots + 10^2 \cdot a_2 + 10^1 \cdot a_1 + 10^0 \cdot a_0 + a_{-1} \cdot 10^{-1} + \dots + a_{-m} \cdot 10^{-m} + \dots;$$

при этом предполагается, что $a_n \neq 0$.

Пусть число

$$a = \overline{a_n a_{n-1} \dots a_2 a_1 a_0, a_{-1} a_{-2} \dots a_{-m} \dots}$$

заменено приближением $\tilde{a}$ путем отбрасывания некоторого количества десятичных знаков после запятой:

$$\tilde{a} = \overline{a_n a_{n-1} \dots a_k \dots a_0, a_{-1} a_{-2} \dots a_{-m}}. \tag{4}$$

и/или заменой отброшенных знаков нулями

$$\tilde{a} = \overline{a_n a_{n-1} \dots a_k 0 \dots 0{,}0 \dots 0}.$$

Значащей цифрой числа $\tilde{a}$ (4) будем называть всякую цифру a_j, не равную нулю, и цифру 0, если она является сохраненным десятичным знаком числа a[1].

Когда в десятичной записи числа ненулевым десятичным знакам предшествуют одни нули (например, 0,000354), то они значащими не считаются.

Замечание. Если, к примеру, мы используем для вычислений вместо числа $a = 123{,}45607803$ число $\tilde{a}_1 = 123{,}456078$, то все цифры последнего — значащие, если же в качестве приближения мы возьмем число $\tilde{a}_2 = 123{,}45607800$, то значащими в этом случае будут все цифры, за исключением последнего нуля (предпоследний нуль — значащая цифра!). Заметим, что с рассматриваемой точки зрения числа $\tilde{a}_1$ и $\tilde{a}_2$ — различны, и это странновато, так как они, на первый взгляд, одинаковы. Дело в том, что нули в числе справа могут быть значащими цифрами, а могут просто обозначать *место*, т. е. разряд соответствующего десятичного знака. Особенно неприятна подобная неопределенность при работе с целыми числами — число 123 000 может оказаться как точным (все цифры значащие), так и приближенным — результатом, например, отбрасывания трех последних знаков у числа 123 456.

Чтобы избежать неясности при работе с такими числами, используют так называемую *показательную нормализованную* форму записи чисел, в которой число представляется в виде

$$\boxed{a = \pm M \cdot 10^p,}$$

[1] В частности, всякий нуль, если он заключен между значащими цифрами, является значащей цифрой.

где $M = \overline{0,a_{-1}a_{-2}\ldots a_m\ldots}$ — положительное число из промежутка $(0, 1]$, все цифры которого значащие и $a_{-1} \neq 0$, p — целое число (M называется мантиссой, p — порядком числа a). При такой форме записи числа $\tilde{a}_1$ и $\tilde{a}_2$ легко различимы

$$\tilde{a}_1 = 0{,}123456078 \cdot 10^3, \quad \tilde{a}_2 = 0{,}1234560780 \cdot 10^3,$$

равно как и упомянутые выше целые числа

$$123\,000 = 0{,}123 \cdot 10^6, \quad 123\,000 = 0{,}123000 \cdot 10^6$$

(первое — приближенное, второе — точное).

Пусть a — некоторое число, порядок которого равен $n+1$ (т. е. целая часть a содержит $n+1$ знак), и $\tilde{a}$ — его приближенное значение. Если для некоторого целого положительного r выполняется неравенство

$$\Delta_a = |a - \tilde{a}| \leqslant 10^{n-r},$$

то говорят, что у приближения $\tilde{a}$ r *верных* значащих цифр. (Пример: $a = \pi$, $\tilde{a} = 3{,}14$, у $\tilde{a}$ три верных значащих цифры.)

Отметим, что если цифра приближения является *верной значащей цифрой*, то это не значит, что она обязательно *совпадает* с соответствующим десятичным знаком точного числа. (Хотя в подавляющем большинстве случаев это так.)

Пример. $a = 3{,}000$, $\tilde{a} = 2{,}999$. Здесь $\Delta_a = 0{,}001 = 10^{-3}$, и в соответствии с данным определением у приближения $\tilde{a}$ четыре верных значащих цифры, однако ни одна из них не совпадает с соответствующими десятичными знаками числа a. ▶

Аналогично определяется количество верных значащих цифр приближения *после запятой.*

§ 2. Эволюция погрешностей в процессе вычислений

Важное значение при проведении приближенных вычислений имеет правильное понимание соотношения погрешностей исходных данных и результата. Задача, которую мы будем изучать в этом разделе, можно сформулировать следующим образом:

Что можно сказать о погрешности результата вычислений при известных погрешностях чисел (операндов), используемых в вычислениях?

Общий ответ на этот вопрос может быть получен с помощью следующего утверждения.

Лемма (формула конечных приращений). *Пусть* $y = f(\mathbf{x}) = f(x_1, x_2, \ldots, x_N)$ — *гладкая (непрерывно дифференцируемая) в ограниченной и замкнутой области* $\mathbf{D} \subset \mathbb{R}^N$ *функция* N *переменных. Тогда для произвольных точек* $\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^N$ *таких, что отрезок, их соединяющий, целиком лежит в* $\mathbf{D}$, *имеет место формула конечных приращений Лагранжа:*

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}) = \sum_{i=1}^{N} \frac{\partial f(\xi)}{\partial x_i}(\tilde{x}_i - x_i), \tag{1}$$

где ξ — *некоторая точка в* $\mathbf{D}$, *лежащая на отрезке, соединяющем точки* $\mathbf{x}$ *и* $\tilde{\mathbf{x}}$.

◀ Пусть λ — произвольное действительное число. Рассмотрим функцию переменной λ, которая при фиксированных $\mathbf{x}$ и $\widetilde{\mathbf{x}}$ задается соотношением

$$\varphi(\lambda) = f(\mathbf{x} + \lambda(\widetilde{\mathbf{x}} - \mathbf{x})).$$

Эта функция дифференцируема в любой точке $0 \leqslant \lambda \leqslant 1$, и имеет место формула

$$\varphi'(\lambda) = \lim_{\Delta\lambda \to 0} \frac{\varphi(\lambda + \Delta\lambda) - \varphi(\lambda)}{\Delta\lambda} = \lim_{\Delta\lambda \to 0} \frac{1}{\Delta\lambda} \sum_{i=1}^{N} \frac{\partial f(\mathbf{x} + \lambda(\widetilde{\mathbf{x}} - \mathbf{x}))}{\partial x_i} (\widetilde{x}_i - x_i) = df(\xi),$$

где $\xi = \mathbf{x} + \lambda(\widetilde{\mathbf{x}} - \mathbf{x})$. Применяя к функции $\varphi(\lambda)$ формулу конечных приращений Лагранжа на отрезке $\lambda \in [0, 1]$, получим искомое соотношение (1). ▶

Теперь можно сформулировать некоторые правила, описывающие эволюцию погрешностей в процессе вычислений.

Погрешность алгебраической суммы (выражения вида $\pm a_1 \pm a_2 \pm \ldots \pm a_N$)

Утверждение. *Предельная абсолютная погрешность алгебраической суммы равна сумме предельных абсолютных погрешностей слагаемых.*

◀ Рассмотрим случай двух слагаемых. Пусть функция $f(x_1, x_2) = \pm x_1 \pm x_2$ и мы хотим вычислить значение этой функции при $x_1 = a_1$, $x_2 = a_2$ по приближенным значениям $\widetilde{a}_i$, $i = 1, 2$. Предельные абсолютные погрешности слагаемых предполагаем известными и равными соответственно Δ_{a_i}, $i = 1, 2$. Поскольку $\left|\frac{\partial f}{\partial x_i}\right| = 1$, то равенство (1) дает

$$\boxed{\Delta_f = \left|\frac{\partial f}{\partial x_1}\Delta_{a_1} + \frac{\partial f}{\partial x_2}\Delta_{a_2}\right| \leqslant \Delta_{a_1} + \Delta_{a_2}.} \ ▶$$

Отсюда, между прочим, следует, что точность в определении суммы-разности определяется точностью наименее точного слагаемого. А поскольку предельная абсолютная погрешность регламентирует количество верных значащих цифр числа, то при сложении-вычитании у слагаемых следует оставлять столько значащих цифр (знаков), сколько их у наименее точного слагаемого (практически оставляют на один-два знака больше). Бессмысленно удерживать все верные значащие цифры у более точных слагаемых — точности результата они не повышают.

И еще одно важное замечание, которое полезно иметь в виду. Пусть числа a_1 и a_2 близки, так что их разность мала: $|a_1 - a_2| \ll 1$. Тогда при любой, сколь угодно малой *абсолютной погрешности* слагаемых, *относительная погрешность* результата резко ухудшается в сравнении с относительными погрешностями δ_{a_1} и δ_{a_2}. Соотношение

$$\delta_{a_1 - a_2} = \frac{\Delta_{a_1} + \Delta_{a_2}}{|\widetilde{a}_1 - \widetilde{a}_2|}$$

показывает, что при малых абсолютных и относительных погрешностях операндов $\widetilde{a}_1$ и $\widetilde{a}_2$ относительная погрешность результата может оказаться значительной за счет малого знаменателя.

Пример. $\widetilde{a}_1 = 123{,}456$; $\widetilde{a}_2 = 123{,}455$; $\Delta_{\widetilde{a}_1} = \Delta_{\widetilde{a}_2} = 10^{-4}$. При этом, как легко установить, $\delta_{a_1} = \delta_{a_2} \leqslant 10^{-8}$. Оба операнда имеют по шесть верных значащих цифр. Разность равна 0,001 и ее абсолютная погрешность $2 \cdot 10^{-4} < 10^{-3}$, так что единственная значащая цифра результата оказывается верной. Относительная погрешность результата равна в этом случае $\frac{2 \cdot 10^{-4}}{0{,}001} = 0{,}2$, что в $2 \cdot 10^7$ раз больше относительной погрешности операндов. ▶

Таким образом, при проведении вычислений подобного рода следует иметь у операндов значительно больше верных значащих цифр, чем мы хотим получить в результате. Если такой возможности нет, то процедуру вычислений следует модифицировать с целью исключения вычитания близких чисел.

Погрешность алгебраического произведения (выражения вида $\frac{a_1 a_2 \dots a_N}{b_1 b_2 \dots b_M}$)

Будем в дальнейшем предполагать, что среди алгебраических сомножителей отсутствуют равные нулю.

Утверждение. *Предельная относительная погрешность алгебраического произведения не превосходит суммы предельных относительных погрешностей множителей.*

◀ Рассмотрим функцию

$$f(x_1, x_2) = x_1 \cdot x_2.$$

Вычисляя значение этой функции в точке $x_1 = a_1$, $x_2 = a_2$ по приближенным значениям $\widetilde{a}_i$, $i = 1, 2$, будем считать известными предельные относительные погрешности операндов δ_{a_i}, $i = 1, 2$. Рассмотрим

$$\delta_f = \frac{\Delta_f}{|f(\widetilde{a}_1, \widetilde{a}_2)|}.$$

Для оценки числителя применим к функции $f(\widetilde{a}_1, \widetilde{a}_2)$ формулу Лагранжа (1). Имеем

$$\Delta_f = \left| \frac{\partial f(\xi_1, \xi_2)}{\partial x_1} \Delta_{a_1} + \frac{\partial f(\xi_1, \xi_2)}{\partial x_2} \Delta_{a_2} \right| \leqslant |\xi_2| \Delta_{a_1} + |\xi_1| \Delta_{a_2}. \qquad (2)$$

Здесь (ξ_1, ξ_2) — точка, лежащая на отрезке, соединяющем (a_1, a_2) и $(\widetilde{a}_1, \widetilde{a}_2)$. Поэтому $|\xi_i| \leqslant |\widetilde{a}_i| + \Delta_{a_i}$. Учитывая это обстоятельство, разделим обе части неравенства (2) на $f(\widetilde{a}_1, \widetilde{a}_2) = \widetilde{a}_1 \cdot \widetilde{a}_2$ и получим

$$\delta_f \leqslant \frac{|\xi_2| \Delta_{a_1} + |\xi_1| \Delta_{a_2}}{|\widetilde{a}_1 \cdot \widetilde{a}_2|} \leqslant \delta_{a_1} + \delta_{a_2} + 2\delta_{a_1} \delta_{a_2}.$$

Учитывая, что величина произведения $2\delta_{a_1}\delta_{a_2}$ является малой более высокого порядка, чем каждый из сомножителей

$$\delta_{a_1 a_2} \leqslant \delta_{a_1} + \delta_{a_2},$$

последним слагаемым можно пренебречь.

Аналогичные рассмотрения для функции $f(x_1, x_2) = \frac{x_1}{x_2}$ приводят к соотношению

$$\delta_f \leqslant \delta_{a_1} \frac{|\widetilde{a}_2|}{|\xi_2|} + \delta_{a_2} \frac{|\widetilde{a}_2|^2 |\xi_1|}{|\xi_2|^2 |\widetilde{a}_1|},$$

откуда, пренебрегая отличием отношений $\frac{|\widetilde{a}_i|}{|\xi_i|}$ от единицы, заключаем, что

$$\boxed{\delta_{a_1/a_2} \leqslant \delta_{a_1} + \delta_{a_2}.} \blacktriangleright$$

§ 3. Законы больших чисел и вероятностная оценка суммарной погрешности

Полученные выше правила учета погрешностей операндов мало пригодны для оценки погрешности результата в ситуации, когда количество операций очень велико. В этом случае вспомогательная задача — скрупулезный учет эволюции погрешностей операндов — может превратиться в задачу не менее сложную, чем основная решаемая вычислительная задача. Кроме того, в случае независимости погрешностей операндов, они имеют тенденцию к самопроизвольному компенсированию — реальные суммарные погрешности оказываются значительно ниже полученных выше предельных. Некоторые общие соображения о поведении суммарных погрешностей, учитывающие указанные обстоятельства, могут быть получены, если обратиться к теории вероятностей.

Основным теоретико-вероятностным фактом, который нам понадобятся в этом пункте, является *центральная предельная теорема*.

Центральная предельная теорема. *Пусть $\xi_1, \xi_2, \ldots, \xi_N$ — независимые случайные величины с нулевым средним и среднеквадратичными отклонениями σ_i, удовлетворяющими условию $\sigma_i \leqslant \sigma \ \forall i$. Тогда при $N \to \infty$ имеет место соотношение*

$$P\left(\frac{\sum_{i=1}^{N} \xi_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}} < t\right) \to F(t), \tag{1}$$

где $F(t)$ — стандартная функция Лапласа

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-t^2/2}\, dt.$$

Пусть мы вычисляем значение функции $f(a_1, a_2, \ldots, a_N)$ по приближенным данным $a_i \approx \widetilde{a}_i$, $\Delta(a_i) = \widetilde{a}_i - a_i$ — независимые случайные погрешности соответствующих аргументов, относительно которых будем предполагать, что в среднем они равны нулю (систематическая ошибка отсутствует) и равномерно распределены на промежутке $|\Delta(a_i)| \leqslant \Delta_i \leqslant \Delta$. При этих предположениях условия сформулированной выше теоремы выполняются для последовательности случайных величин $\xi_i = M_i \cdot \Delta(a_i)$, где $M_i = \sup \frac{\partial f}{\partial a_i}$.

Применим для оценки погрешности $\Delta f = f(\widetilde{a}_1, \widetilde{a}_2, \ldots, \widetilde{a}_N) - f(a_1, a_2, \ldots, a_N)$ формулу Лагранжа и получим, что

$$\Delta f = \sum_{i=1}^{N} \frac{\partial f}{\partial a_i}(\widetilde{a}_i - a_i) \leqslant \sum_{i=1}^{N} M_i \Delta(a_i) = \sum_{i=1}^{N} \xi_i.$$

Поскольку $\sigma_i = \frac{|M_i|\Delta_i}{\sqrt{3}} \leqslant \frac{|M_i|\Delta}{\sqrt{3}}$, то из соотношения (1) вытекает, что

$$|\Delta f| = \Delta_f \leqslant t_\alpha \frac{\Delta}{\sqrt{3}} \sqrt{\sum_{i=1}^{N} M_i^2}. \qquad (2)$$

Здесь t_α — решение уравнения

$$2F(t_\alpha) - 1 = \sqrt{\frac{2}{\pi}} \int_0^{t_\alpha} e^{-t^2/2}\, dt = \alpha,$$

отвечающее надежности α.

Таким образом, мы получили следующую оценку погрешности результата:

если предельная абсолютная погрешность операндов не превосходит величины Δ, *то предельная абсолютная погрешность результата в подавляющем большинстве случаев (с вероятностью* α *— т. е. в* $\alpha \cdot 100$ % *случаев) удовлетворяет неравенству* (2).

Таблица наиболее употребительных значений α и соответствующих им значений t_α приведена ниже.

α	0,9	0,95	0,99	0,995	0,9973	0,999
t_α	1,65	1,95	2,59	2,89	3,0	3,3

Для погрешности результата сложения-вычитания N чисел, предельная абсолютная погрешность которых не превышает Δ, рассмотрения предыдущего пункта дают оценку порядка $N \cdot \Delta$, в то время как из (2) следует, что, например, в 99 случаях из 100 эта погрешность не превысит $1{,}5 \cdot \sqrt{N}$.

§ 4. Источники погрешностей

При рассмотрении практически любой прикладной задачи неизбежно огрубление реальной ситуации — математическое описание, как правило, представляет собой идеализированный взгляд на исследуемую проблему. Кроме того, возникающие в процессе вычислений погрешности также искажают результаты. В связи с этим возникает вопрос о соответствии реальной ситуации результатов, полученных при исследовании математической модели. Здесь необходимо учитывать, по крайней мере, три группы причин, могущих вызвать отличие результатов модельного исследования от истинного течения процесса:

— причины, вызванные упрощающими математическую постановку задачи предположениями, в том числе неточность задания начальных данных и параметров задачи (эти причины порождают *погрешность постановки задачи*);

— причины, вызванные необходимостью завершить математическую процедуру, требующую для своего завершения бесконечного числа шагов, после конечного числа шагов (эти причины порождают *погрешность метода*);

— причины, вызванные погрешностями вычислений и округлений (эти причины порождают *вычислительную погрешность*).

По поводу последней хотелось бы отметить, что при использовании вычислительных машин вычислительной погрешности невозможно избежать *в принципе*. Дело в том, что действительные числа, которыми оперирует математика, непредставимы в реальном вычислительном устройстве. Всякая вычислительная машина оперирует с *конечным набором чисел*, представляющим собой машинный аналог множества действительных чисел.

Пренебрегая несущественными с рассматриваемой точки зрения подробностями, всякое машинное число можно представить в виде

$$a = q^p \cdot 0{,}a_1a_2 \ldots a_t, \qquad (1)$$

где q — основание системы счисления, принятой для данной машины, p — порядок числа, $0{,}a_1a_2 \ldots a_t$ — мантисса (число называется *нормализованным*, если $a_1 \neq 0$), t — разрядность машины. (В современных машинах q, как правило, принимается равным 2, 8, 10 или 16, t — от 16 до 128.)

Рассмотрим модельную машину с $q = 10$, $-8 \leqslant p \leqslant 8$, $t = 5$. Как порядок, так и мантисса хранятся в запоминающем устройстве в виде целого числа со знаком. В соответствии с представлением (1), числа вне диапазона от $-10^8 \cdot 0{,}99999$ до $10^8 \cdot 0{,}99999$ непредставимы в нашей машине, как и числа, по модулю меньшие $10^{-8} \cdot 0{,}00001$. Кроме того, множество чисел, которыми оперирует наша машина — дискретно, причем шаг дискретизации неравномерен; он зависит от порядка — для чисел порядка p шаг дискретизации равен $10^{-p} \cdot 0{,}00001$.

В силу указанной выше специфики, машинная арифметика отличается от обычной — не выполняются, вообще говоря, законы ассоциативности сложения и умножения, дистрибутивности умножения относительно сложения, произведение ненулевых чисел может дать нуль и т. п. Уже ввод исходных данных в машину приводит к появлению погрешностей, не говоря о погрешностях округления результатов арифметических операций, связанных с конечной разрядностью машинных чисел. Вычислительная погрешность — непременный атрибут машинных вычислений.

Проиллюстрируем вышеизложенное несколькими простыми примерами.

Пример 1. Найти длину экватора Земли.

◀ Огрубляя реальную ситуацию, будем считать Землю шаром, радиус R которого примем равным 6 371 км. При этом экватор — это дуга большого круга (т. е. круга, лежащий в плоскости, проходящей через центр шара) и ее длина дается соотношением

$$L_{\text{экватора}} = 2\pi R.$$

Точность конечного результата будет определяться следующими обстоятельствами:

— насколько форма Земли отлична от шара и насколько радиус этого шара отличен от 6 371;

— сколько десятичных знаков числа π (а это, как известно, бесконечная десятичная непериодическая дробь) мы возьмем для вычислений;

— как будет организован процесс вычислений.

Первые два обстоятельства порождают *погрешность постановки задачи*, последнее — *вычислительную погрешность*.

В нашей модельной машине исходные данные будут иметь вид

$$2 = 10^1 \cdot 0{,}20000, \quad \pi = 10^1 \cdot 0{,}31416, \quad R = 10^4 \cdot 0{,}6371.$$

В результате умножений получим

$$L_{\text{экватора}} = 2\pi R \approx 40\,030 \text{ м}.$$

Поскольку абсолютная погрешность наихудшего сомножителя (а это приближение к значению R) может быть принята равной 10^1, его относительная погрешность будет не больше чем $0{,}16 \cdot 10^{-2}$, относительная погрешность числа π, связанная с ограниченной разрядностью нашей машины, равна 10^{-5}. Отсюда — абсолютная погрешность результата $\Delta_L = 10^2$ и значащая цифра 3 результата сомнительна. Окончательный результат: при принятых допущениях длина экватора может быть принята равной 40 000 км. ▶

Пример 2. Заменяя километровый участок рельсового пути, рабочие ошиблись и положили взамен изношенного рельс на 1 м длиннее. Насколько отклонится замененный участок от прямолинейного?

◀ **Модель I.** Предполагая, что рельс жестко закреплен на концах километрового участка пути, будем считать что прогиб рельса, вызванный увеличением его длины, может быть описан геометрией сторон равнобедренного треугольника (рис. 1) с основанием 1 км и боковыми сторонами $a = b = \frac{1}{2}1{,}001$ км. При этом максимальное отклонение рельса от прямолинейного участка определяется соотношением

$$h = \sqrt{\left(\frac{1}{2}1{,}001\right)^2 - \left(\frac{1}{2}\right)^2} = \frac{1}{2}10^{-2}\sqrt{20{,}01}.$$

Для вычисления значения радикала воспользуемся соотношением

$$\sqrt{20{,}01} = \sqrt{20{,}25 - 0{,}24} = 4{,}5\sqrt{1 - \frac{4}{75}},$$

применяя для нахождения последнего радикала разложение в ряд Тейлора

$$\sqrt{1 - \frac{4}{75}} = 1 - \frac{1}{2}\cdot\frac{4}{75} + \dots .$$

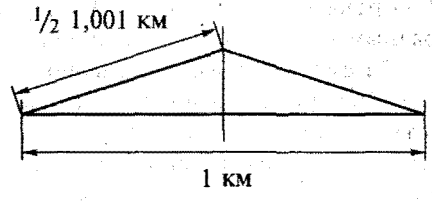


Рис. 1. Модель I прогиба удлиненного рельса

Погрешность метода возникает здесь, когда процесс нахождения суммы ряда мы заменяем процедурой нахождения суммы конечного числа его членов. При проведении вычислений в десятичных дробях образуется вычислительная погрешность из-за замены дроби $\frac{4}{75} = 0{,}05333\dots$ конечным ее отрезком.

В нашей машине эти вычисления имеют вид:

исходные данные

$$\frac{1}{2} = 10^0 \cdot 0{,}50000, \quad 4{,}5 = 10^1 \cdot 0{,}45000, \quad \frac{4}{75} = 10^{-1} \cdot 0{,}53333,$$

удерживая в ряду Тейлора первые два слагаемых, получим

$$h \approx 0{,}0219.$$

Аккуратный анализ погрешностей показывает, что последняя цифра сомнительна, так что следует считать $h \approx 0{,}021 \pm 0{,}01$. Полученный результат любопытен с содержательной точки зрения — если километровый участок удлинить всего на 1 м, то отклонение от прямой составит 21 м. Более точные выкладки здесь бессмысленны из-за значительной погрешности постановки задачи.

В рассматриваемой задаче возможна и другая постановка.

Модель II. Предполагая, что рельс жестко закреплен на концах километрового участка пути, будем считать прогиб рельса, вызванный увеличением его длины, дугой окружности некоторого радиуса R, так что хорда длиной 1 км стягивает дугу длиной 1,001 км. Определению подлежит так называемая *стрелка* дуги x (рис. 2).

Расчетные соотношения для этой модели имеют вид

$$0{,}5 = R\cdot\sin\alpha, \quad R - x = R\cdot\cos\alpha, \quad 2\alpha\cdot R = 1{,}001.$$

Для определения прогиба получаем равенство

$$x = \frac{1}{2}\operatorname{tg}\frac{\alpha}{2},$$

а котором величина α подлежит определению из уравнения

$$\sin\alpha = \frac{\alpha}{1{,}001}.$$

Для решения последнего разложим в ряд Тейлора левую часть

$$\alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} - \ldots = \frac{\alpha}{1{,}001}$$

и, удерживая конечное количество слагаемых в левой части, определим величину α. Например, оставляя слева слагаемые до третьего порядка включительно, придем к уравнению

$$\alpha - \frac{\alpha^3}{3!} = \frac{\alpha}{1{,}001},$$

имеющему очевидный корень $\alpha = 0$ и интересующий нас в рассматриваемой ситуации положительный корень $\alpha = \sqrt{\frac{6}{1001}}$. Заменяя, вообще говоря, бесконечную процедуру извлечения квадратного корня конечным числом операций (например, конечным числом итераций, или, как и выше, суммированием конечного числа членов ряда Тейлора), получим значение $\alpha \approx 0{,}0774$. Подставим его в выражение для стрелки прогиба. Получим $x \approx 19$ м, что несколько точнее результата, полученного на первой модели.

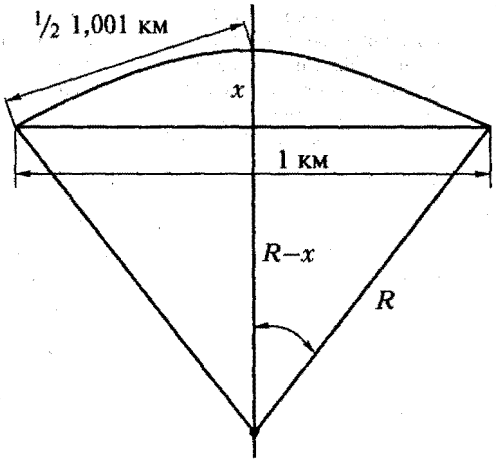


Рис. 2. Модель II прогиба удлиненного рельса

Здесь *погрешность метода* складывается из погрешности в решении уравнения, погрешности извлечения корня и погрешности вычисления значения функции tg. ▶

Завершая обзор возможных источников погрешностей при численном анализе, сделаем два замечания, важных для понимания особенностей вычислительного процесса.

Замечание 1. Весьма значительна роль *способа организации вычислений* в условиях конечной разрядности вычислительного устройства и наличия ошибок округления. Нижеследующий пример показывает вычислительную абсурдность теоретически верных рассуждений.

Пусть в нашей модельной вычислительной машине (характеризуемой, напомним, пятиразрядной сеткой для мантиссы и порядками представляемых чисел в диапазоне от -8 до 8) мы хотим воспользоваться для вычисления значения sin 6,284 всюду сходящимся рядом

$$\sin x = x - \frac{x^3}{3!} + \ldots + (-1)^{2n-1}\frac{x^{2n-1}}{(2n-1)!} + \ldots .$$

Замечая, что рассматриваемый ряд — знакопеременный, ограничимся для проведения вычислений тринадцатью его членами, используя то обстоятельство, что отброшенный остаток не превышает по модулю первого отброшенного члена и, следовательно, не оказывает значительного влияния на сумму

$$r_{14} \leqslant \frac{x^{27}}{27!} \leqslant 3{,}2 \cdot 10^{-7}.$$

Процесс сложения в нашей машине будет иметь вид

$$\sin 6{,}284 \approx 0{,}62840 \cdot 10 - 0{,}41358 \cdot 10^2 + 0{,}81658 \cdot 10^2 - 0{,}76776 \cdot 10^2 + 0{,}42108 \cdot 10^2 - 0{,}15116 \cdot 10^2 + \\ + 0{,}38264 \cdot 10 - 0{,}71952 + 0{,}10446 - 0{,}12060 \cdot 10^{-2} + 0{,}11340 \cdot 10^{-3} - \ldots = 0{,}00069.$$

Полученный таким образом результат не содержит ни одной верной значащей цифры — точное значение равно 0,00081469...! Дело тут в том, что погрешности округления при промежуточных вычислениях значительно превышают сам результат, что, естественно, негативно сказывается на эффективности вычислений. Для вычисления искомого значения в нашей машине избранный алгоритм совершенно непригоден. Этот пример показывает, что *погрешность метода* при неверной организации процесса вычислений может оказаться весьма значительной.

Замечание 2. Не менее важной является проблема *устойчивости* процесса вычислений. Предыдущие рассмотрения дают нам пример теоретически хорошего, но плохо (с вычислительной точки зрения) примененного алгоритма. Это исправимо и за счет некоторых дополнительных ухищрений (двойная

точность, пересчет значений функции в области малых значений аргумента и т. п.) можно добиться высокой эффективности предложенного алгоритма. Однако встречаются задачи, результат в которых не может быть получен с удовлетворительной точностью никаким способом — это задачи, чувствительные к незначительному изменению исходных данных и погрешностям вычислений, так называемые *некорректно поставленные задачи*.

Пусть мы хотим решить систему линейных уравнений

$$\begin{cases} 100{,}001x + 99{,}999y = 1\,100{,}009, \\ 99{,}999x + 100{,}001y = 1\,099{,}991, \end{cases}$$

точное решение которой $x = 10$, $y = 1$. В нашей модельной машине указанная система (за счет округлений, вызванных разрядностью) примет вид

$$\begin{cases} 100x + 99{,}999y = 1\,100, \\ 99{,}999x + 100y = 1\,100, \end{cases}$$

и ее решение, которое легко может быть получено, (например, методом подстановки или каким-либо другим) есть $x = y = 5{,}5$, что ни качественно, ни количественно не похоже на истинное решение исходной системы. Как уже было отмечено выше, причина столь разительного несоответствия — неустойчивость исходной задачи.

Глава LVII

ЛИНЕЙНЫЕ УРАВНЕНИЯ

Большое количество задач численного анализа сводится к решению систем алгебраических уравнений, чаще — линейных, несколько реже — нелинейных. Правильное понимание особенностей этих важных задач, умение понять (а тем более — построить) и адекватно применить алгоритмы эффективного их решения — необходимый элемент образования любого исследователя, использующего методы численного анализа в своей практике. Как уже было отмечено выше, несмотря на кажущуюся простоту и незамысловатость этих задач, уже в простейших ситуациях бездумный счет может обернуться полнейшей бессмыслицей.

§ 1. Линейные уравнения — основные сведения

Пусть $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ — вектора-столбцы с действительными компонентами, $\mathbf{A} = \left(a_i^j\right)_{i=1,\ldots,m}^{j=1,\ldots,n}$ — матрица формата $m \times n$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_1^1 & a_1^2 & \ldots & a_1^n \\ a_2^1 & a_2^2 & \ldots & a_2^n \\ \ldots & \ldots & \ldots & \ldots \\ a_m^1 & a_m^2 & \ldots & a_m^n \end{bmatrix}.$$

Системой m линейных уравнений с n неизвестными называется совокупность уравнений вида

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + \ldots + a_1^n x_n = b_1, \\ a_2^1 x_1 + a_2^2 x_2 + \ldots + a_2^n x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_m^1 x_1 + a_m^2 x_2 + \ldots + a_m^n x_n = b_m. \end{cases} \tag{1}$$

Решением системы (1) называется упорядоченный набор n действительных чисел $\gamma_1, \ldots, \gamma_n$, которые при одновременной подстановке в каждое из уравнений системы вместо неизвестных $x_1, \ldots, x_n$ обращают эти уравнения в тождества.

Система называется *определенной*, если она обладает единственным решением, и *неопределенной*, если решений у нее более одного. В любом из этих случаев система называется *совместной*, в противном случае — *несовместной*.

Система называется *однородной*, если столбец правых частей системы — нулевой. Заметим, что однородная система всегда совместна — в качестве решения можно взять нулевой столбец.

Система (1) может быть записана в виде матричного линейного уравнения

$$\boxed{\mathbf{A} \cdot \mathbf{x} = \mathbf{b}.}$$

Она же может быть представлена линейным векторным уравнением

$$\boxed{x_1\mathbf{a}^1 + x_2\mathbf{a}^2 + \ldots + x_n\mathbf{a}^n = \mathbf{b},}$$

где через $\mathbf{a}^i$, $i = 1, 2, \ldots, n$ обозначены столбцы матрицы A.

Матрица A называется *матрицей коэффициентов системы* или просто *матрицей системы*. Матрица $\overline{\mathrm{A}}$, которая получается из матрицы системы добавлением к ней столбца свободных членов называется *расширенной матрицей системы*: $\overline{\mathrm{A}} = [\mathrm{A}|\mathbf{b}]$.

Две системы называются *эквивалентными*, если множества их решений совпадают.

Система называется *квадратной*, если количество уравнений системы m равно количеству неизвестных n.

Следующие теоремы дают полную информацию о совместности и определенности системы линейных уравнений (1).

Теорема Кронекера—Капелли. *Система m линейных уравнений с n неизвестными совместна тогда и только тогда, когда ранг расширенной матрицы системы совпадает с рангом матрицы коэффициентов. При этом система определена, если этот общий ранг равен количеству неизвестных, и неопределена, если ранг меньше количества неизвестных.*

Теорема Крамера. *Квадратная система однозначно разрешима тогда и только тогда, когда определитель матрицы системы не равен нулю. Решение системы при этом может быть получено по формулам:*

$$x_1 = \frac{\Delta(x_1)}{\Delta}, \quad x_2 = \frac{\Delta(x_2)}{\Delta}, \quad \ldots, \quad x_n = \frac{\Delta(x_n)}{\Delta},$$

которые называются *формулами Крамера*. Здесь через Δ обозначен определитель матрицы системы, а через $\Delta(x_i)$ — определитель матрицы, которая получается из матрицы системы заменой i-го столбца столбцом свободных членов

$$\Delta(x_i) = \begin{vmatrix} a_1^1 & a_1^2 & \ldots & b_1 & \ldots & a_1^n \\ a_2^1 & a_2^2 & \ldots & b_2 & \ldots & a_2^n \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_n^1 & a_n^2 & \ldots & b_n & \ldots & a_n^n \end{vmatrix}.$$

Сформулируем несколько очевидных следствий:

— однородная система имеет ненулевое решение тогда и только тогда, когда ее ранг меньше числа неизвестных,

— квадратная однородная система имеет ненулевое решение тогда и только тогда, когда ее определитель равен нулю,

— если квадратная система совместна для любой правой части, то ее определитель отличен от нуля.

Используемые в настоящее время методы решения линейных систем условно классифицируются следующим образом:

— конечные (точные) методы — методы, позволяющие за *конечное число* шагов или получить *точное* решение рассматриваемой системы, или установить ее неразрешимость,

— бесконечные (итерационные) методы — методы, дающие возможность получить *приближенное* решение за *конечное* число шагов, количество которых, как правило, определяет точность приближения,

— методы стохастического моделирования (Монте-Карло) — методы, позволяющие получить *приближенное* решение с заданной степенью *надежности* за конечное число шагов.

§ 2. Линейные уравнения — метод исключения

Все точные методы решения систем линейных уравнений представляют собой тот или иной вариант *метода последовательного исключения* неизвестных, называющегося еще *методом Гаусса*. Мы рассмотрим здесь основные положения этого метода и некоторые особенности его численной реализации.

Чтобы не загромождать изложение, проиллюстрируем основные этапы метода Гаусса на примере системы трех уравнений с тремя неизвестными

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + a_1^3 x_3 = b_1, \\ a_2^1 x_1 + a_2^2 x_2 + a_2^3 x_3 = b_2, \\ a_3^1 x_1 + a_3^2 x_2 + a_3^3 x_3 = b_3, \end{cases} \tag{1}$$

в которой коэффициент a_1^1 будем считать не равным нулю. Идея предлагаемого метода состоит в следующем — с помощью элементарных преобразований, преобразующих систему (1) в эквивалентную, приведем ее к треугольному виду

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + a_1^3 x_3 = b_1, \\ \widetilde{a}_2^2 x_2 + \widetilde{a}_2^3 x_3 = \widetilde{b}_2, \\ \widetilde{a}_3^3 x_3 = \widetilde{b}_3. \end{cases} \tag{2}$$

Эта часть преобразований носит название *прямого хода* метода исключения. Система (2) теперь может быть легко решена: из последнего уравнения находим значение x_3, подставляем во второе и определяем x_2, и, наконец, подставляя найденные значения x_3 и x_2 в первое уравнение, находим x_1. Эта часть преобразований носит название *обратного хода* метода.

Преобразования прямого хода выглядят следующим образом — поскольку $a_1^1 \neq 0$, то, умножая первое уравнение системы на a_2^1/a_1^1 и a_3^1/a_1^1 соответственно и вычитая их из второго и третьего, приходим к системе

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + a_1^3 x_3 = b_1, \\ a_2^{2,1} x_2 + a_2^{3,1} x_3 = b_2^1, \\ a_3^{2,1} x_2 + a_3^{3,1} x_3 = b_3^1, \end{cases}$$

в которой неизвестная x_1 исключена из всех уравнений системы кроме первого. Теперь, если только $a_2^{2,1} \neq 0$, процедуру можно повторить, исключая переменную x_2

из третьего уравнения, чем и заканчивается прямой ход метода. Если же $a_2^{2,1} = 0$, а $a_3^{2,1} \neq 0$, то поменяв местами второе и третье уравнения системы, исключим из последнего уравнения переменную x_3. Если же $a_2^{2,1} = a_3^{2,1} = 0$, то система будет иметь вид

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + a_1^3 x_3 = b_1, \\ a_2^{3,1} x_3 = b_2^1, \\ a_3^{3,1} x_3 = b_3^1, \end{cases}$$

и дальнейшее исследование очевидно.

Легко видеть, что вышеизложенная процедура переносится на произвольные системы (1) § 1, при этом расчетные соотношения прямого хода, приводящие к исключению переменной с номером $q = 1, 2, \ldots, n-1$ выглядят следующим образом — если переменные с номерами, предшествующими q, уже исключены, то ступенчатая система, эквивалентная исходной, имеет вид

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + a_1^3 x_3 + \ldots + a_1^q x_q + \ldots + a_1^n x_n = b_1, \\ a_2^{2,1} x_2 + \ldots\ldots\ldots + a_2^{q,1} x_q + \ldots + a_2^{n,1} x_n = b_2^1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_q^{q,q-1} x_q + \ldots + a_q^{n,q-1} x_n = b_q^{q-1}, \\ a_{q+1}^{q,q-1} x_q + \ldots + a_{q+1}^{n,q-1} x_n = b_{q+1}^{q-1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_m^{q,q-1} x_1 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + a_m^{n,q-1} x_n = b_m^{q-1}, \end{cases}$$

и исключение переменной с номером q осуществляется по формулам

$$a_i^{j,q} = a_i^{j,q-1} - \frac{a_i^{q,q-1}}{a_q^{q,q-1}} a_q^{j,q-1}, \tag{3}$$

$$b_i^q = b_i^{q-1} - \frac{a_i^{q,q-1}}{a_q^{q,q-1}} b_q^{q-1}, \quad i = q+1, \ldots, m, \quad j = q, q+1, \ldots, n, \tag{4}$$

которые приводят исходную систему (мы рассматриваем здесь случай $m = n$) к следующему, треугольному виду

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 + \ldots + a_1^q x_q + \ldots + a_1^n x_n = b_1, \\ a_2^{2,1} x_2 + \ldots + a_2^{q,1} x_q + \ldots + a_2^{n,1} x_n = b_2^1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_q^{q,q-1} x_q + \ldots + a_q^{n,q-1} x_n = b_q^{q-1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_n^{n,n-1} x_n = b_n^{n-1}. \end{cases} \tag{5}$$

Обратный ход метода реализуется в соответствии с соотношениями

$$x_q = \frac{b_q^{q-1} - \sum\limits_{s=q+1}^{n} a_q^{s,q-1} x_s}{a_q^{q,q-1}}, \tag{6}$$

где $q = n, n-1, \ldots, 1$, очевидно следующими из (5).

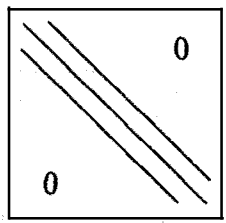
2.1. Трехдиагональные матрицы — метод прогонки

Системы линейных уравнений, возникающие при дискретизации краевых задач для обыкновенных дифференциальных уравнений, при построении интерполяционных сплайнов и т. п., как правило, имеют специальный вид — большое количество коэффициентов матрицы системы оказывается равным нулю. В этом случае использование метода исключения в том виде, как он был описан выше, нецелесообразно — большая часть операций с элементами матрицы окажется бессмысленной переработкой нулей. И поэтому имеет смысл так модифицировать процедуру последовательного исключения, чтобы учесть специфическую структуру матрицы системы и избежать нерезультативного счета.

Мы рассмотрим здесь модификацию метода последовательного исключения, применяемую для решения *трехдиагональных* систем ленточного типа

$$\begin{cases} a_1^1 x_1 + a_1^2 x_2 = b_1, \\ a_2^1 x_1 + a_2^2 x_2 + a_2^3 x_3 = b_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_k^{k-1} x_{k-1} + a_k^k x_k + a_k^{k+1} x_{k+1} = b_k, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{n-1}^{n-2} x_{n-2} + a_{n-1}^{n-1} x_{n-1} + a_{n-1}^n x_n = b_{n-1}, \\ a_n^{n-1} x_{n-1} + a_n^n x_n = b_n, \end{cases} \tag{7}$$

в которых не равны нулю только коэффициенты при переменных, стоящих на главной диагонали, а также над и под ней.



В этом случае исключение неизвестных осуществляется так: из первого уравнения выражаем x_1 через x_2

$$x_1 = \frac{b_1}{a_1^1} - \frac{a_1^2}{a_1^1} x_2 = M^1 + L^1 x_2$$

и подставляем во второе. При этом во втором уравнении останутся только переменные x_2 и x_3. Аналогично предыдущему шагу, выражаем из этого уравнения x_2 через x_3

$$x_2 = \frac{b_2 - a_2^1 M^1}{a_2^2 + a_2^1 L^1} - \frac{a_2^3}{a_2^2 + a_2^1 L^1} x_3 = M^2 + L^2 x_3.$$

Продолжая последовательно эту процедуру, на k-м шаге выражаем переменную x_k через x_{k+1}

$$x_k = \frac{b_k - a_k^{k-1} M^{k-1}}{a_k^k + a_k^{k-1} L^{k-1}} - \frac{a_k^{k+1}}{a_k^k + a_k^{k-1} L^{k-1}} x_{k+1} = M^k + L^k x_{k+1}. \tag{8}$$

Когда мы таким образом доберемся до последнего уравнения системы, то получим серию соотношений вида

$$\begin{cases} x_1 = M^1 + L^1 x_2, \\ x_2 = M^2 + L^2 x_3, \\ \dots\dots\dots\dots\dots\dots \\ x_{n-1} = M^{n-1} + L^{n-1} x_n, \\ x_n = M^n, \end{cases} \tag{9}$$

заканчивающуюся выражением, которое определяет переменную x_n. Если теперь пройти соотношения (9) в обратном порядке, то мы получаем возможность определить последовательно все переменные

$$x_n \Longrightarrow x_{n-1} \Longrightarrow x_{n-2} \Longrightarrow \dots \Longrightarrow x_1.$$

Фактически, для осуществления описанной процедуры, необходимо уметь определять коэффициенты M^k и L^k для всех значений $k = 1, 2, \dots, n$. А это можно быстро и эффективно сделать, используя реккурентные соотношения (8), полученные выше

$$\boxed{M^1 = \frac{b_1}{a_1^1}, \quad L^1 = -\frac{a_1^2}{a_1^1}, \quad M^k = \frac{b_k - a_k^{k-1} M^{k-1}}{a_k^k + a_k^{k-1} L^{k-1}}, \quad L^k = -\frac{a_k^{k+1}}{a_k^k + a_k^{k-1} L^{k-1}}.} \tag{10}$$

Описанная процедура называется *методом прогонки* решения трехдиагональных систем. *Прямой ход* прогонки — расчет прогоночных коэффициентов M^k и L^k в соответствии с формулами (10), *обратный* — нахождение неизвестных x_i в соответствии с расчетными формулами (9).

Утверждение. *Если система* (7) *является системой с доминирующей главной диагональю, т. е.* $|a_k^k| > |a_k^{k-1}| + |a_k^{k+1}|$, *то она однозначно разрешима и знаменатели прогоночных коэффициентов не обращаются в нуль, так что соотношения* (10) *пригодны для проведения расчетов.*

§ 3. Линейные уравнения — итерационные методы

Основная идея итерационных методов решения уравнений и систем уравнений (не обязательно линейных) может быть описана следующим образом: пусть $A(x) = b$ — подлежащее решению уравнение (система уравнений). Переписав его в виде $\Phi(x) = x$, возьмем в качестве начального значения неизвестного $x = x_0$ и построим последовательность $x_1, x_2, \dots, x_N, \dots$ по правилу

$$x_N = \Phi(x_{N-1}). \tag{1}$$

Если у этой последовательности при $N \to \infty$ существует предел, то он будет решением уравнения $\Phi(x) = x$, а следовательно, и исходного уравнения $A(x) = b$.

Сходимость итерационной процедуры (1) может быть обеспечена, например, условием близости двух последовательных итераций — *расстояние между* x_{N-1} *и* x_N *с ростом* N *должно уменьшаться достаточно быстро, например, со скоростью геометрической прогрессии*

$$\|x_N - x_{N-1}\| < C \cdot q^N, \quad 0 < q < 1.$$

◀ В этом случае, как легко проверить, выполняется критерий Коши сходимости последовательности $\{x_N\}_{N=1,2,\ldots}$:

$$\|x_{N+m}-x_N\| = \|x_{N+m}-x_{N+m-1}+x_{N+m-1}-x_{N+m-2}+x_{N+m-2}-\ldots+x_{N+1}-x_N\| \leqslant$$

$$\leqslant \sum_{i=N+1}^{N+m} \|x_i - x_{i-1}\| \leqslant C \sum_{i=N+1}^{N+m} q^i \leqslant C\frac{q^{N+1}}{1-q} \to 0, \quad N \to \infty. \blacktriangleright$$

3.1. Метод простой итерации для линейных систем

Рассмотрим квадратную систему линейных уравнений, записанную в матричной форме

$$\mathbf{A}\cdot\mathbf{x} = \mathbf{b}.$$

Преобразовав ее к виду

$$\mathbf{x} = \Phi\cdot\mathbf{x} + \beta,$$

построим итерационный процесс

$$\mathbf{x}_N = \Phi\cdot\mathbf{x}_{N-1} + \beta, \tag{2}$$

использование которого для решения исходной системы носит название *метода простой итерации* решения системы линейных уравнений.

Заметим, что сходимость построенных по правилу (2) итераций зависит от свойств матрицы Φ и начального приближения $\mathbf{x}_0$.

Напомним определение нормы матрицы, согласованной с нормой вектора.

Нормой матрицы $\Phi = (\varphi_i^j)$, согласованной с нормой вектора $\mathbf{x}$, называется наибольший «коэффициент растяжения» вектора $\mathbf{x}$ под действием определяемого ею преобразования

$$\|\Phi\| = \sup_{\mathbf{x}} \frac{\|\Phi\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Так, например, в $\mathbb{R}^n$ с обычной евклидовой нормой вектора, задаваемой соотношением $\|\mathbf{x}\| = \sqrt{\sum x_i^2}$, норма матрицы определяется равенством

$$\|\Phi\| = \sqrt{\max_{1\leqslant i\leqslant n} \lambda^i_{\Phi^T\cdot\Phi}},$$

где $\lambda^i_{\Phi^T\cdot\Phi}$ — i-е собственное значение матрицы $\Phi^T\cdot\Phi$.

Для дальнейшего нам понадобится еще два выражения для нормы матрицы:

$$\|\Phi\|_C = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |\varphi_i^j|$$

— норма матрицы, согласованная с нормой вектора, задаваемой соотношением

$$\|\mathbf{x}\|_C = \max_{1\leqslant i\leqslant n} |x_i|,$$

и

$$\|\Phi\|_S = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |\varphi_i^j|$$

— норма, согласованная с нормой вектора, задаваемой соотношением

$$\|\mathbf{x}\|_S = \sum_{i=1}^{n} |x_i|.$$

Достаточное условие сходимости предложенного выше итерационного процесса теперь может быть сформулировано в терминах нормы матрицы следующим образом

Теорема. *Система уравнений*

$$\mathbf{x} = \Phi \cdot \mathbf{x} + \beta$$

однозначно разрешима и итерационный процесс, построенный на ее основе сходится к решению, если только $\|\Phi\| < 1$.

◀ Рассмотрим $\mathbf{x}_N = \Phi\mathbf{x}_{N-1} + \beta$, $\mathbf{x}_{N-1} = \Phi\mathbf{x}_{N-2} + \beta$, откуда для разности $\|\mathbf{x}_N - \mathbf{x}_{N-1}\|$ получим

$$\|\mathbf{x}_N - \mathbf{x}_{N-1}\| = \|\Phi\mathbf{x}_{N-1} + \beta - \Phi\mathbf{x}_{N-2} - \beta\| \leqslant \|\Phi\mathbf{x}_{N-1} - \Phi\mathbf{x}_{N-2}\| \leqslant$$

$$\leqslant \|\Phi\| \cdot \|\mathbf{x}_{N-1} - \mathbf{x}_{N-2}\| = q\|\mathbf{x}_{N-1} - \mathbf{x}_{N-2}\|,$$

где $q = \|\Phi\| < 1$. Из неравенства $\|\mathbf{x}_N - \mathbf{x}_{N-1}\| \leqslant q\|\mathbf{x}_{N-1} - \mathbf{x}_{N-2}\|$ индуктивно заключаем, что

$$\|\mathbf{x}_N - \mathbf{x}_{N-1}\| \leqslant q^{N-1}\|\mathbf{x}_1 - \mathbf{x}_0\| = C \cdot q^{N-1}.$$

Отсюда, в силу сделанного выше замечания, вытекает сходимость последовательности итераций $\mathbf{x}_N$

$$\forall\, \mathbf{x}_0\colon\ \exists \lim_{N\to\infty} \mathbf{x}_N = \mathbf{X}.$$

Предельный вектор $\mathbf{X}$ является решением рассматриваемой системы. Покажем, что оно единственно. Допустим противное: существуют два различных решения $\mathbf{X}$ и $\mathbf{Y}$. Тогда должно выполняться равенство

$$\mathbf{Y} - \mathbf{X} = \Phi(\mathbf{Y} - \mathbf{X})$$

и, следовательно,

$$\|\mathbf{Y} - \mathbf{X}\| \leqslant \|\Phi\| \cdot \|\mathbf{Y} - \mathbf{X}\| < \|\mathbf{Y} - \mathbf{X}\|,$$

что невозможно, так как $\mathbf{X} \neq \mathbf{Y}$. ▶

Заметим, что при выполнении условия теоремы итерации сходятся со скоростью геометрической прогрессии *для произвольного* начального приближения. Сформулированное условие достаточно, но не необходимо — можно привести примеры сходящихся итерационных процессов, для которых это условие не выполнено.

Поскольку теорема доказана для абстрактной нормы, то она остается справедливой для каждой из трех норм, о которых шла речь выше.

Последнее замечание позволяет сформулировать простое правило построения сходящегося итерационного процесса решения системы

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}. \tag{3}$$

Теорема (о достаточном условии сходимости специального метода простой итерации). *Пусть система линейных уравнений* (3) *является системой с доминирующей главной диагональю, т. е.*

$$\forall\, i\colon \sum_{j\neq i}^{n} |a_i^j| < |a_i^i|.$$

Тогда специальный итерационный процесс, задаваемый соотношениями

$$\begin{cases} x_1^N = \dfrac{b_1}{a_1^1} - \sum\limits_{j\neq 1} \dfrac{a_1^j}{a_1^1} x_j^{N-1}, \\ x_2^N = \dfrac{b_2}{a_2^2} - \sum\limits_{j\neq 2} \dfrac{a_2^j}{a_2^2} x_j^{N-1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_i^N = \dfrac{b_i}{a_i^i} - \sum\limits_{j\neq i} \dfrac{a_i^j}{a_i^i} x_j^{N-1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^N = \dfrac{b_n}{a_n^n} - \sum\limits_{j\neq n} \dfrac{a_n^j}{a_n^n} x_j^{N-1}, \end{cases}$$

сходится со скоростью геометрической прогрессии для любого начального приближения.

3.2. Метод Зейделя для линейных систем

Один из вариантов метода простой итерации, носящий название метода Зейделя, является достаточно привлекательным за счет такой организации итерационного процесса, при которой i-я компонента очередной итерации вычисляется с использованием предшествующих ей и компонент решения, уже вычисленных на этом шаге.

Точнее, пусть $\mathbf{x}_s$ — s-я итерация решения

$$\mathbf{x}_s = \begin{pmatrix} x_1^s \\ x_2^s \\ \ldots \\ x_n^s \end{pmatrix}.$$

Следующую, $s+1$-ю итерацию решения системы $\mathrm{A}\cdot\mathbf{x} = \mathbf{b}$, будем искать следующим образом

$$\begin{cases} a_1^1 x_1^{s+1} + a_1^2 x_2^s \quad + \ldots + a_1^n x_n^s \quad = b_1, \\ a_2^1 x_1^{s+1} + a_2^2 x_2^{s+1} + \ldots + a_2^n x_n^s \quad = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_n^1 x_1^{s+1} + a_n^2 x_2^{s+1} + \ldots + a_n^n x_n^{s+1} = b_n, \end{cases}$$

Если матрицу A системы представить в виде суммы нижней треугольной A_Δ и верхней треугольной — A^Δ матриц

$$\mathrm{A}_\Delta = \begin{pmatrix} a_1^1 & 0 & 0 & \ldots & 0 \\ a_2^1 & a_2^2 & 0 & \ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_n^1 & a_n^2 & a_n^3 & \ldots & a_n^n \end{pmatrix}, \quad \mathrm{A}^\Delta = \begin{pmatrix} 0 & a_1^2 & a_1^3 & \ldots & a_1^n \\ 0 & 0 & a_2^3 & \ldots & a_2^n \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix},$$

то предлагаемая процедура может быть представлена в виде

$$\mathbf{x}_{s+1} = \Phi\cdot\mathbf{x}_s + \beta,$$

где

$$\Phi = -A_{\Delta}^{-1} \cdot A^{\Delta}, \quad \beta = A_{\Delta}^{-1} \cdot \mathbf{b}.$$

Можно показать, что метод Зейделя сходится, если только матрица A системы удовлетворяет условию *усиленного доминирования главной диагонали*,

$$\sum_{j \neq i} |a_i^j| \leqslant q \cdot |a_i^i|, \quad 0 < q < 1.$$

§ 4. Точность численного решения систем линейных уравнений

Предложенные выше методы принципиально пригодны для решения систем линейных уравнений. Однако при их реализации могут возникнуть сложности, сводящие на нет все усилия по нахождению решения с приемлемой точностью. Эти сложности связаны с организацией вычислительных процедур и эволюцией погрешностей в процессе вычислений, с одной стороны, и со структурой решаемой системы уравнений, с другой. Здесь мы коротко остановимся на поведении погрешностей и реакции системы на малые возмущения правой части.

4.1. Выбор главного элемента

Как уже отмечалось выше, в методе последовательного исключения неизвестных на каждом шаге осуществляется деление на диагональный элемент матрицы. Ясно, что погрешность результата будет тем больше, чем меньше значение этого коэффициента. Если, в частности при осуществлении процедуры (3)–(4) § 2, мы не позаботимся о том, чтобы отношения

$$\frac{a_i^{q,q-1}}{a_q^{q,q-1}}$$

были по возможности малыми (по крайней мере, не превышали по модулю единицы), то нам не избежать катастрофического накопления погрешностей округления и, как следствие, значительной потери точности решения. Но чтобы эти отношения были как можно меньше, необходимо, чтобы знаменатель $a_q^{q,q-1}$ был как можно бо́льшим.

Оказывается, только за счет перестановки уравнений системы и/или перенумерации неизвестных на каждом шаге всегда можно добиться того, чтобы величина коэффициента $a_q^{q,q-1}$ стала максимально возможной и все отношения $\frac{a_i^{q,q-1}}{a_q^{q,q-1}}$ удовлетворяли оговоренному выше условию. Такая модификация метода исключения носит название *метода последовательного исключения неизвестных с выбором главного элемента* и позволяет избежать накопления ошибок округления в машине.

Рассмотрим пример, иллюстрирующий высказанные выше соображения. Пусть решению подлежит система двух уравнений с двумя неизвестными

$$\begin{cases} 1{,}41x_1 + 173{,}2x_2 = 174{,}61, \\ 21\,234x_1 + 1\,541x_2 = 22\,775, \end{cases}$$

точное решение которой, как легко проверить, $x_1 = x_2 = 1$. Как и выше, будем предполагать, что вычисления проводятся в машине, оперирующей с мантиссой в пять десятичных знаков.

Проводя исключение переменной x_1 из второго уравнения, получим

$$\left(\frac{21\,234}{1{,}41}\cdot 1{,}41 - 21\,234\right)x_1 + \left(\frac{21\,234}{1{,}41}\cdot 173{,}2 - 1541\right)x_2 = \left(\frac{21\,234}{1{,}41}\cdot 174{,}61 - 22\,775\right)$$

или, с учетом разрядности нашей модельной машины,

$$1\cdot 10^0 x_1 + 2{,}6069\cdot 10^6 x_2 = 2{,}6068\cdot 10^6.$$

Пренебрегая первым слагаемым, которое должно было обратиться в нуль, но из-за наличия погрешностей округления «уцелело», получим

$$x_2 = 0{,}99996.$$

Тогда из первого уравнения имеем

$$1{,}41x_1 = 174{,}61 - 173{,}19 = 1{,}42 \quad\Longrightarrow\quad x_1 = 1{,}0071.$$

Таким образом, уже в случае всего двух уравнений точность метода исключения оказалась недостаточно высокой, и понятно почему — величина отношения

$$\frac{a_2^1}{a_1^1} = \frac{21\,234}{1{,}41} = 15\,060$$

недопустимо велика.

Посмотрим, что дает процедура выбора главного элемента. Переставив уравнения системы

$$\begin{cases} 21\,234x_1 + 1\,541x_2 = 22\,775, \\ 1{,}41x_1 + 173{,}2x_2 = 174{,}61, \end{cases}$$

и исключая переменную x_1 из второго уравнения, получаем

$$\left(\frac{1{,}41}{21\,234}\cdot 21\,234 - 1{,}41\right)x_1 + \left(\frac{1{,}41}{21\,234}\cdot 1\,541 - 173{,}2\right)x_2 = \left(\frac{1{,}41}{21\,234}\cdot 22\,775 - 174{,}61\right)$$

или, с учетом разрядности нашей модельной машины

$$1{,}7310\cdot 10^2 x_2 = 1{,}7310\cdot 10^2,$$

откуда

$$x_2 = 1 \quad\Longrightarrow\quad x_1 = 1.$$

В этом случае отношение

$$\frac{a_2^1}{a_1^1} = \frac{1{,}41}{21\,234} = 6{,}6403\cdot 10^{-5}$$

мало, и за счет его малости погрешности округления эффективно подавляются, что приводит к повышению точности получаемого решения.

В заключение отметим, что при использовании итерационных методов накопления погрешностей округления, вообще говоря, не происходит — на каждом шаге предыдущее приближение может рассматриваться как начальное для итерационного процесса и, следовательно, точность получаемого решения определяется точностью, с которой проводятся вычисления на этом шаге. Однако в реальных вычислительных процессах итерационные процедуры могут приводить к неожиданным переполнениям порядков чисел за счет большого количества шагов процедуры.

4.2. Возмущения правой части. Обусловленность матрицы

Пусть при решении системы линейных уравнений

$$A \cdot \mathbf{x} = \mathbf{b}$$

правые части по каким-то причинам оказались заданными неточно и вместо исходной мы фактически решаем систему

$$A \cdot \mathbf{x} = \mathbf{b} + \varepsilon.$$

Насколько решение исходной, «точно» заданной системы, может отличаться от решения системы заданной «приближенно»? Умение отвечать на этот вопрос, т. е. умение оценивать влияние точности задания правых частей системы на точность решения, крайне важно. Ведь если малые погрешности в задании правых частей оказывают существенное влияние на на точность решения, то может оказаться, что уже погрешности округления, связанные с представлением коэффициентов системы в вычислительной машине, меняют ее настолько, что используемые для нахождения решения процедуры становятся просто бессмысленными.

Удобной характеристикой этого влияния служит так называемое *число обусловленности* матрицы A системы, оценивающее относительную погрешность решения в сравнении с относительной погрешностью задания правых частей. Пусть

$$\delta_{\mathbf{x}} = \frac{\|\mathbf{x}^* - \mathbf{x}\|}{\|\mathbf{x}^*\|}$$

— относительная погрешность решения (здесь $\mathbf{x}$ — решение точно заданной системы, $\mathbf{x}^*$ — возмущенной) и

$$\delta_{\mathbf{b}} = \frac{\|\varepsilon\|}{\|\mathbf{b} + \varepsilon\|}$$

— относительная погрешность задания правых частей системы.

Числом обусловленности матрицы A *системы называется число* $\nu(A)$, *задаваемое соотношением*

$$\nu(A) = \sup_{\varepsilon} \frac{\delta_{\mathbf{x}}}{\delta_{\mathbf{b}}}. \tag{1}$$

Поскольку

$$\|\mathbf{x}^* - \mathbf{x}\| \leqslant \|A^{-1}\| \cdot \|\varepsilon\|, \quad \|\mathbf{b} + \varepsilon\| = \|A \cdot \mathbf{x}^*\| \leqslant \|A\| \cdot \|\mathbf{x}^*\|,$$

то для числа обусловленности получаем

$$\boxed{\nu(A) = \|A\| \cdot \|A^{-1}\|.}$$

Из определения (1) следует, что *число обусловленности* матрицы A системы можно трактовать как своего рода «коэффициент усиления» матрицы. Он показывает как система реагирует на возмущения — чем больше это число, тем сильнее искажения решения:

$$\boxed{\delta_{\mathbf{x}} = \nu(A) \cdot \delta_{\mathbf{b}}.}$$

Пример, рассмотренный во введении, показывает, что простое применение *принципиально правильных* алгоритмов недостаточно. Предварительно следует убедиться в том, что матрица системы хорошо обусловлена. В противном случае следует предпринять некоторые дополнительные усилия по *регуляризации* рассмотренных выше процедур (с целью придания им смысла).

Глава LVIII

НЕЛИНЕЙНЫЕ УРАВНЕНИЯ И СИСТЕМЫ

Задача, к рассмотрению которой мы приступаем в этой главе, может быть в самом общем виде сформулирована следующим образом: пусть функции $F_1(x), F_2(x), \ldots, F_n(x)$ определены в некоторой области пространства и в ней нужно найти точку $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$, в которой выполняются равенства

$$\begin{cases} F_1(x^0) = F_1(x_1^0, x_2^0, \ldots, x_n^0) = 0, \\ F_2(x^0) = F_2(x_1^0, x_2^0, \ldots, x_n^0) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ F_n(x^0) = F_n(x_1^0, x_2^0, \ldots, x_n^0) = 0. \end{cases}$$

Мы видели выше, что некоторые проблемы, связанные с численной реализацией хорошо известных методов решения систем, возникают уже в случае *линейных* функций F_j. В рассматриваемой ситуации к тому же нет универсальных решающих процедур, что задачу численного анализа не облегчает. Мы рассмотрим здесь основные идеи решения подобных задач и возможности их численной реализации.

§ 1. Нелинейные уравнения. Метод половинного деления

Рассмотрим нелинейное уравнение

$$F(x) = 0. \tag{1}$$

Предполагая, что функция $F(x)$ непрерывна и принимает на концах некоторого промежутка $[a, b]$ значения разных знаков, заключаем, что на этом промежутке у нее имеется корень (может быть и не один).

Пусть $\alpha = \frac{a+b}{2}$ — середина отрезка $[a, b]$. Мы можем уточнить положение искомого корня x^0 следующим образом: если $F(\alpha)F(a) < 0$ — корень лежит в левой половине отрезка, если же $F(\alpha)F(b) < 0$ — в правой. Это рассуждение позволяет нам вдвое сократить длину промежутка, содержащего искомый корень. Применяя многократно изложенное рассуждение, на N-м шаге получим промежуток $[a_N, b_N]$, такой, что $x^0 \in [a_N, b_N]$ и $|b_N - a_N| = \frac{b-a}{2^N}$. Взяв в качестве x^0 любую точку x_N промежутка $[a_N, b_N]$, заметим, что $|x^0 - x_N| \leqslant \frac{b-a}{2^N}$. Это означает, что таким способом искомый корень может быть определен с любой степенью точности.

Первоначальная локализация корня на отрезке $[a, b]$ — это отдельная задача, которую мы здесь рассматривать не будем, однако заметим, что перебором некоторых

значений функции $F(x)$ можно найти точки, в которых она принимает значения противоположных знаков, и тем самым локализовать корень.

Пример. Решить уравнение $F(x) = x^3 + 3x + 1 = 0$.

◀ Заметив, что $F(0) = 1 > 0$, а $F(-1) = -3 < 0$ заключаем, что искомый корень лежит на промежутке $[-1, 0]$. Последовательность вычислений по уточнению значения корня выглядит так

$$
\begin{array}{llll}
\alpha_1 = -0{,}50000, & F(\alpha_1) = -0{,}62500 \cdot 10^0, & \Longrightarrow & x^0 \in [\alpha_1; 0], \\
\alpha_2 = -0{,}25000, & F(\alpha_2) = 0{,}23437 \cdot 10^0, & \Longrightarrow & x^0 \in [-0{,}5; \alpha_2], \\
\alpha_3 = -0{,}37500, & F(\alpha_3) = -0{,}17773 \cdot 10^0, & \Longrightarrow & x^0 \in [\alpha_3; -0{,}25], \\
\alpha_4 = -0{,}31250, & F(\alpha_4) = 0{,}31980 \cdot 10^{-1}, & \Longrightarrow & x^0 \in [-0{,}375; \alpha_4], \\
\alpha_5 = -0{,}34380, & F(\alpha_5) = -0{,}71870 \cdot 10^{-1}, & \Longrightarrow & x^0 \in [\alpha_5; -0{,}3125], \\
\alpha_6 = -0{,}33063, & F(\alpha_6) = -0{,}28033 \cdot 10^{-1}, & \Longrightarrow & x^0 \in [\alpha_6; -0{,}3125], \\
\alpha_7 = -0{,}32157, & F(\alpha_7) = 0{,}20373 \cdot 10^{-2}, & \Longrightarrow & x^0 \in [-0{,}33063; \alpha_7], \\
\alpha_8 = -0{,}32610, & F(\alpha_8) = -0{,}12980 \cdot 10^{-1}, & \Longrightarrow & x^0 \in [\alpha_8; -0{,}32157], \\
\alpha_9 = -0{,}32384, & F(\alpha_9) = -0{,}54819 \cdot 10^{-2}, & \Longrightarrow & x^0 \in [\alpha_9; -0{,}32157], \\
\alpha_{10} = -0{,}32271, & F(\alpha_{10}) = -0{,}17376 \cdot 10^{-2}, & \Longrightarrow & x^0 \in [\alpha_{10}; -0{,}32157], \\
\alpha_{11} = -0{,}32214, & F(\alpha_{11}) = -0{,}15019 \cdot 10^{-3}, & \Longrightarrow & x^0 \in [-0{,}32271; \alpha_{11}],
\end{array}
$$

.....

Если теперь в качестве искомого корня взять середину последнего промежутка, то можно гарантировать, что мы нашли корень $x^0 \approx -0{,}322$ рассматриваемого уравнения с точностью не худшей, чем $\varepsilon_x = \frac{b-a}{2^N} = 10^{-3}$. ▶

Замечание. В рассматриваемом примере он единственен, в силу $F'(x) = 3x^2 + 3 > 0$, что влечет монотонность функции $F(x)$.

Отметим несколько особенностей практической реализации этого простого и надежного алгоритма.

Первая связана с выбором момента окончания вычислений — когда можно считать, что точность найденного приближения уже достаточна? Естественно останавливать процедуру при выполнении условия

$$\frac{b-a}{2^N} \leqslant \varepsilon_x. \tag{2}$$

При этом, правда, может оказаться, что значение функции не будет близко к нулю. Действительно, если градиент функции в окрестности нуля велик, то точность локализации корня, не худшая ε_x, соответствующей малости *невязки* $\Delta F = F\left(\frac{a_N+b_N}{2}\right)$ не гарантирует; последняя может при этом быть сколь угодно большой. Поэтому когда вычислителя интересуют не только значение корня, но и значения функции в окрестности корня, наряду с проверкой условия (2) полезно одновременно проверять и условие малости *невязки*

$$\left|F\left(\frac{a_N + b_N}{2}\right)\right| \leqslant \varepsilon_F.$$

Другая особенность связана с ограниченной точностью машинных вычислений. При малом градиенте функции $F(x)$ в окрестности ее нуля за счет ограниченной точности вычислений возможно «перепутывание знаков» — неверно будет определен знак функции в очередной точке и, тем самым, неверно будет произведена очередная локализация корня. В этой ситуации, если требуемая точность в определении корня еще не достигнута, можно воспользоваться вычислениями с двойной точностью. Но в любом случае катастрофической потери точности не происходит и существующих разрядных сеток реальных вычислительных машин обычно бывает достаточно для удовлетворения требований по достижению точности нахождения корня в подавляющем большинстве практически встречающихся задач.

Наконец, последняя особенность предлагаемой процедуры состоит в том, что она совершенно неприменима для нахождения корня четной кратности — в этом случае функция $F(x)$ не меняет знака в искомой точке, что исключает возможность использования приведенных выше рассуждений.

Отметим в заключение достаточно высокую трудоемкость рассмотренной процедуры — для достижения точности локализации корня, не худшей 10^{-3}, на промежутке длиной 1 нам потребовалось 11 шагов.

§ 2. Нелинейные уравнения. Метод хорд

Рассмотренный выше метод половинного деления может быть усовершенствован, если на каждом шаге в качестве приближения к искомому корню брать не середину очередного промежутка локализации, а точку пересечения хорды, соединяющей концы графика функции $y = F(x)$ на этом промежутке, с осью абсцисс. Другими словами, на каждом шаге следует заменить график функции $F(x)$ хордой, стягивающей его концы (рис. 1), и в качестве приближения к корню взять точку пересечения этой хорды с осью абсцисс.

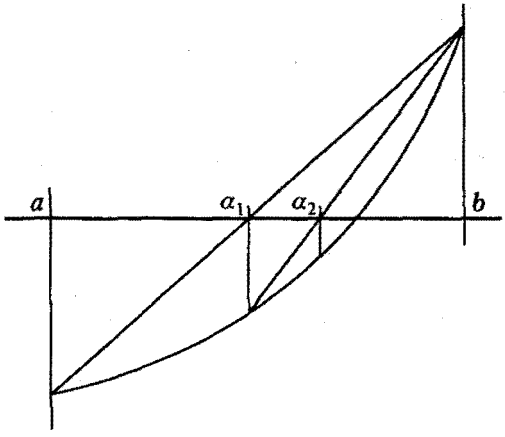


Рис. 1. Метод хорд

Подробнее — пусть в предположениях § 1 $F(a) = f_a$, $F(b) = f_b$. Хорда, стягивающая дугу графика функции $F(x)$ на промежутке $[a, b]$, задается уравнением

$$y - f_a = \frac{f_b - f_a}{b - a}(x - a),$$

и, поскольку $f_a f_b < 0$, точка α_1 пересечения этой хорды с осью абсцисс лежит внутри отрезка $[a, b]$ и находится из условия $y = 0$:

$$-f_a = \frac{f_b - f_a}{b - a}(\alpha_1 - a) \quad \Longrightarrow \quad \alpha_1 = \frac{af_b - bf_a}{f_b - f_a}.$$

Теперь рассматриваем промежуток, один из концов которого α_1, а другой — тот из концов a и b, в котором знак функции противоположен знаку функции $F(x)$ в точке α_1, и в качестве искомого корня вновь берем точку пересечения хорды, стягивающей концы дуги графика функции $F(x)$ на этом промежутке, с осью абсцисс.

Можно надеяться, что повторяя эту процедуру достаточно долго, мы получим значение искомого корня с приемлемой точностью быстрее, чем методом половинного деления.

§ 3. Нелинейные уравнения. Метод касательных (метод Ньютона)

Дальнейшее усовершенствование рассмотренных выше методов решения нелинейного уравнения (1) § 1 может быть получено за счет использования вместо хорд, стяги-

вающих дугу графика функции, касательных к этому графику. Точнее, если x_N — приближение к корню уравнения $F(x) = 0$, полученное на N-м шаге, то в качестве следующего приближения x_{N+1} предлагается взять точку пересечения *касательной* к графику функции $F(x)$ в точке x_N с осью абсцисс (рис. 2):

$$x_{N+1} = x_N - \frac{F(x_N)}{F'(x_N)}. \qquad (1)$$

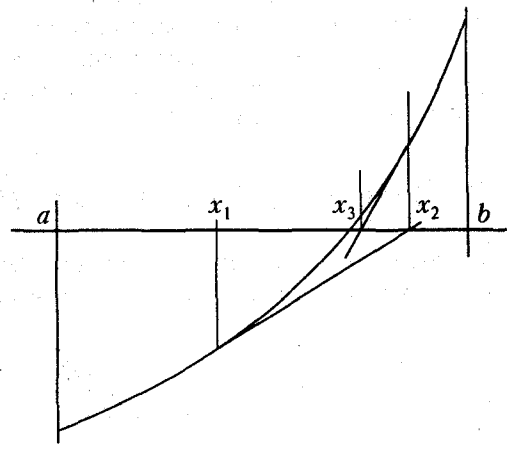


Рис. 2. Метод касательных

Этот метод привлекательнее вышеизложенных, так как он более эффективен и позволяет находить в том числе и корни четной кратности. Правда, в случае плохо локализуемых корней (когда у уравнения имеются близкие несовпадающие корни) в реализации метода могут возникнуть определенные трудности. Однако они могут быть преодолены за счет более тщательного выбора начального приближения.

На практике метод Ньютона широко используется для вычисления радикалов $\sqrt[k]{\alpha}$. Задача нахождения $\sqrt[k]{\alpha}$ эквивалентна задаче решения уравнения

$$F(x) = x^k - \alpha = 0,$$

для которого расчетные соотношения (1) принимают вид

$$x_{N+1} = \frac{k-1}{k} x_N + \frac{\alpha}{k x_N^{k-1}}.$$

Хорошо известный алгоритм извлечения квадратного корня является частным случаем ($k = 2$) этих соотношений.

Метод Ньютона является представителем широкой группы *итерационных* методов решения нелинейных уравнений, речь о которых пойдет ниже.

§ 4. Нелинейные уравнения. Метод простой итерации

Рассмотрим уравнение

$$x = \Phi(x)$$

и, выбрав некоторое начальное значение x_0, построим последовательность точек $\{x_N\}$ так, чтобы

$$x_N = \Phi(x_{N-1}). \qquad (1)$$

Если построенная таким способом последовательность сходится

$$\exists\, x^0\colon\ x^0 = \lim_{N\to\infty} x_N,$$

то ее предел x^0 является искомым корнем исходного уравнения.

Процесс отыскания корня, порождаемый соотношением (1), называется методом *простой итерации*. Важнейший вопрос — о сходимости итерационного процесса (1).

Теорема (достаточное условие сходимости метода простой итерации). *Пусть функция $\Phi(x)$ определена на промежутке $[a, b]$, принимает значения из этого промежутка и непрерывно дифференцируема на нем. Пусть, кроме того, на $[a, b]$ выполняется неравенство*

$$|\Phi'(x)| \leqslant q < 1, \quad q > 0.$$

Тогда из любой начальной точки $x_0 \in [a, b]$ итерационный процесс (1) сходится к точке $x^0 \in [a, b]$, являющейся единственным корнем рассматриваемого уравнения.

◀ Пусть $x_0 \in [a, b]$ — точка, запускающая итерационный процесс (1). Тогда $\forall\, N$ и любого натурального p выполняется соотношение

$$|x_{N+p} - x_N| = |x_{N+p} - x_{N+p-1} + x_{N+p-1} - x_{N+p-2} + \ldots - x_{N+1} + x_{N+1} - x_N| \leqslant$$
$$\leqslant \sum_{i=1}^{p} |x_{N+i} - x_{N+i-1}|.$$

Но

$$|x_{N+i} - x_{N+i-1}| = |\Phi(x_{N+i-1}) - \Phi(x_{N+i-1})|$$

и в силу теоремы Лагранжа о конечных приращениях

$$|\Phi(x_{N+i-1}) - \Phi(x_{N+i-2})| = |\Phi'(\xi)| \cdot |x_{N+i-1} - x_{N+i-2}| \leqslant q|x_{N+i-1} - x_{N+i-2}|.$$

Отсюда легко получаем, что

$$|\Phi(x_{N+i-1}) - \Phi(x_{N+i-1})| \leqslant q^{N+i-1}|x_1 - x_0|.$$

Тем самым,

$$|x_{N+p} - x_N| \leqslant \sum_{i=1}^{p} q^{N+i-1}|x_1 - x_0| = |x_1 - x_0| \sum_{i=1}^{p} q^{N+i-1} = |x_1 - x_0| q^N \frac{1 - q^p}{1 - q}.$$

Учитывая, что $|x_1 - x_0| \leqslant |b - a|$ и $q < 1$, окончательно получаем

$$|x_{N+p} - x_N| < (b - a)\frac{q^N}{1 - q} \to 0 \quad \forall\, p,\ N \to \infty.$$

В силу критерия Коши из последнего соотношения следует сходимость последовательности x_N при произвольном начальном приближении x_0 к решению рассматриваемого уравнения.

Покажем теперь, что это решение единственно и не зависит от выбора начального приближения. Пусть x^0 и y^0 — пределы итерационной последовательности (1), отвечающие начальным приближениям x_0 и y_0 соответственно, $x_0 \neq y_0$. Тогда

$$x^0 = \lim_{N\to\infty} x_N, \quad x_N = \Phi(x_{N-1}), \quad y^0 = \lim_{N\to\infty} y_N, \quad y_N = \Phi(y_{N-1})$$

и, следовательно,

$$|x_N - y_N| = |\Phi(x_{N-1}) - \Phi(y_{N-1})| \leqslant q|x_{N-1} - y_{N-1}| \leqslant \ldots \leqslant q^N|x_0 - y_0|.$$

Отсюда легко усмотреть, что $\lim x_N = \lim y_N$ при $N \to \infty$. ▶

Заметим, что выкладки теоремы позволяют получить оценку точности приближенного значения корня, полученного методом простой итерации, в зависимости от количества итераций. Если x^0 — корень исследуемого уравнения, то

$$x^0 = \Phi(x^0), \quad x_N = \Phi(x_{N-1}) \quad \Longrightarrow \quad |x^0 - x_N| = |\Phi(x^0) - \Phi(x_{N-1})| < q|x^0 - x_{N-1}|.$$

Аналогично

$$|x^0 - x_{s-1}| = |\Phi(x^0) - \Phi(x_{s-2})| < q|x^0 - x_{s-2}| \quad \forall\, s,$$

откуда окончательно заключаем, что

$$|x^0 - x_N| < q^N |x^0 - x_0| < q^N (b - a). \tag{2}$$

Количество итераций, необходимое для достижения заданной точности ε_x, может теперь быть определено из неравенства (2). Если начальное приближение x_0 выбрано удачно, то метод простой итерации позволяет достигать достаточно высокой точности при относительно малом числе итераций.

Для метода Ньютона из доказанной теоремы можно получить простые достаточные условия сходимости. Действительно, в этом случае

$$\Phi(x) = x - \frac{F(x)}{F'(x)},$$

и условие, полученное выше, принимает вид

$$|\Phi'(x)| \leqslant q < 1 \quad \Longrightarrow \quad \left|\frac{F(x)F''(x)}{F'^2(x)}\right| < 1.$$

Поскольку $F(x)$ близко к нулю, для выполнения последнего условия достаточно, чтобы вторая производная функции $F(x)$ была ограничена, а первая была бы не очень маленькой.

Пример. Рассмотрим уравнение $x^3 + 3x + 1 = 0$, которое в § 1 мы решили методом половинного деления, достигнув за 11 шагов точности 10^{-3}. Запишем его в виде (1), пригодном для итераций, положив

$$x_N = -\frac{x_{N-1}^3 + 1}{3}$$

и взяв в качестве начального приближения $x_0 = -0{,}5$. Последовательность итераций будет иметь следующий вид

$$x_0 = -0{,}5 \implies x_1 = -0{,}29167 \implies x_2 = -0{,}32506 \implies$$
$$\implies x_3 = -0{,}32188 \implies x_4 = -0{,}32222 \implies x_5 = -0{,}32218 \implies \ldots .$$

Точность 10^{-3} достигнута уже на 4-й итерации.

Скорость достижения корня регулируется не столько начальным приближением, сколько величиной q, ограничивающей производную. В рассматриваемом примере $|\Phi'(x)| = x^2$, и q — величина порядка 10^{-1}, а это в соответствии с (2) обеспечивает достижение требуемой точности (в данном случае — 10^{-3}) за 3–4 итерации. Положим, например, $x_0 = 0$. При этом

$$x_1 = -0{,}33333 \implies x_2 = -0{,}32099 \implies x_3 = -0{,}32231 \implies$$
$$\implies x_4 = -0{,}32217 \implies x_5 = -0{,}32219 \implies x_6 = -0{,}32218 \implies \ldots .$$

и требуемая точность достигнута уже на третьей итерации. ▶

Если условие $q < 1$ не будет выполнено, то итерационный процесс может как сходиться, так и не сходиться к корню. Условия теоремы носят *достаточный*, а не необходимый характер — можно привести примеры сходящихся итерационных процессов, для которых указанное условие не выполняется. Эти примеры возможны в ситуации, когда в точках x_N модуль производной $|\Phi'(x)|$ оказывается как больше, так и меньше 1. Если же $q > 1$ во всей области, где эволюционирует итерационный процесс, то он, конечно, сходиться не будет.

Продолжение примера. Отметим, что для рассмотренного уравнения мы могли бы построить и другой итерационный процесс, например,

$$x_N = \frac{1}{x_{N-1}^2 + 3},$$

который получается из исходного уравнения с помощью следующих рассуждений

$$x^3+3x+1=0 \implies x^2+3+\frac{1}{x}=0 \implies x=-\frac{1}{x^2+3}.$$

Для него итерационный процесс также довольно быстро сходится к искомому корню

$$x_0=0 \implies x_1=-0{,}33333 \implies x_2=-0{,}32143 \implies$$
$$\implies x_3=-0{,}32224 \implies x_4=-0{,}32218 \implies x_5=-0{,}32219 \implies \ldots . \blacktriangleright$$

§ 5. Системы нелинейных уравнений

В общей ситуации, когда стоит задача решения системы нелинейных уравнений, круг основных идей и методов остается тем же, что и для случая одного уравнения — это итерационные методы и их модификации. Однако собственно вычислительная сторона реализации упомянутых процедур становится существенно сложнее в техническом плане.

Рассмотрим систему

$$\begin{cases} F_1(x)=F_1(x_1, x_2, \ldots, x_n)=0, \\ F_2(x)=F_2(x_1, x_2, \ldots, x_n)=0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ F_n(x)=F_n(x_1, x_2, \ldots, x_n)=0. \end{cases}$$

Переписав ее в эквивалентном виде

$$\begin{cases} x_1=\Phi_1(x_1, x_2, \ldots, x_n), \\ x_2=\Phi_2(x_1, x_2, \ldots, x_n), \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n=\Phi_n(x_1, x_2, \ldots, x_n), \end{cases}$$

построим итерационный процесс

$$\begin{cases} x_1^{N+1}=\Phi_1\left(x_1^N, x_2^N, \ldots, x_n^N\right), \\ x_2^{N+1}=\Phi_2\left(x_1^N, x_2^N, \ldots, x_n^N\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{N+1}=\Phi_n\left(x_1^N, x_2^N, \ldots, x_n^N\right), \end{cases}$$

условия сходимости которого можно получить аналогично тому, как это было сделано в случае одного уравнения.

На основе предложенного итерационного процесса может быть построена его модификация, являющаяся аналогом метода Зейделя для линейных систем:

$$\begin{cases} x_1^{N+1}=\Phi_1\left(x_1^N, x_2^N, \ldots, x_n^N\right), \\ x_2^{N+1}=\Phi_2\left(x_1^{N+1}, x_2^N, \ldots, x_n^N\right), \\ x_3^{N+1}=\Phi_2\left(x_1^{N+1}, x_2^{N+1}, \ldots, x_n^N\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{N+1}=\Phi_n\left(x_1^{N+1}, x_2^{N+1}, \ldots, x_{n-1}^{N+1}, x_n^N\right). \end{cases}$$

Аналогом метода касательных в рассматриваемой ситуации будет метод Ньютона, для описания которого нам удобно использовать следующие векторно-матричные обозначения:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \quad \mathbf{F}(x) = \begin{bmatrix} F_1(x) \\ F_2(x) \\ \dots \\ F_n(x) \end{bmatrix}.$$

Символом $\frac{\partial \mathbf{F}}{\partial x}$ обозначим матрицу Якоби первых производных векторной функции $\mathbf{F}$

$$\frac{\partial \mathbf{F}}{\partial x} = \left(\frac{\partial F_i}{\partial x_j}\right) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \dots & \frac{\partial F_1}{\partial x_n} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \dots & \frac{\partial F_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial F_n}{\partial x_1} & \frac{\partial F_n}{\partial x_2} & \dots & \frac{\partial F_n}{\partial x_n} \end{pmatrix}.$$

В этих обозначениях для уравнения $\mathbf{F}(x) = 0$ итерационный метод Ньютона принимает вид

$$x^{N+1} = x^N - \left[\frac{\partial \mathbf{F}(x^N)}{\partial x}\right]^{-1} \mathbf{F}(x^N),$$

где $\left[\frac{\partial \mathbf{F}}{\partial x}\right]^{-1}$ — матрица, обратная к матрице Якоби.

Можно доказать следующую теорему, обобщающую на систему условия сходимости итераций, полученных по методу Ньютона для одного уравнения

Теорема. *Пусть в некоторой окрестности решения x^0 системы уравнений $\mathbf{F}(x) = 0$ норма матрицы $\left[\frac{\partial \mathbf{F}}{\partial x}\right]^{-1}$ ограничена и ограничены вторые частные производные $\frac{\partial^2 \mathbf{F}}{\partial x_i \partial x_j}$. Тогда итерации, построенные по методу Ньютона, сходятся к решению системы.*

В заключение отметим, что основные вычислительные усилия метода Ньютона чаще всего связаны с пересчетом матрицы, обратной к матрице Якоби — вычисление производных и обращение матрицы Якоби, как правило, оказывается более громоздкой и трудоемкой процедурой, чем вычисление значений функции. На практике часто поступают так — фиксируют достаточное близкое к корню приближение x^* и строят итерации по правилу

$$x^{N+1} = x^N - \left[\frac{\partial \mathbf{F}(x^*)}{\partial x}\right]^{-1} \cdot \mathbf{F}(x^N),$$

тем самым несколько снижая скорость сходимости итераций, но зато значительно уменьшая трудоемкость каждой итерации.

Глава LIX

ВЫЧИСЛЕНИЕ ЗНАЧЕНИЙ ФУНКЦИЙ

При вычислении значений функции $y = f(x)$ с помощью вычислительных устройств часто бывает удобно пользоваться процедурой вычисления значений другой функции $\varphi(x)$, в некотором смысле «похожей» на вычисляемую, но «более удобной с вычислительной точки зрения». В зависимости от того, как формализуются понятия «похожести» и «удобства вычислений», строится конкретная постановка задачи.

Пример 1. Пусть функция $y = f(x)$ задана таблицей своих значений, т. е. известны значения $f(x_i) = f_i$ функции в некоторых точках $x_0 < x_1 < \ldots < x_n$. Задача состоит в вычислении значения функции $f(x)$ в точке $x \neq x_i$. Можно сказать, что в данном случае «удобство» вычислений состоит в *возможности* вычисления значения функции $f(x)$. Пусть функция $\varphi(x)$ такова, что способ вычисления ее значения в точке $x \neq x_i$ известен, и пусть, кроме того, «похожесть» $\varphi(x)$ и $f(x)$ регламентируется требованием совпадения их значений в тех точках, где заданы значения $y = f(x)$: $\varphi(x_i) = f(x_i)$ $\forall\, i = 0, 1, \ldots, n$. Полагая $f(x) = \varphi(x)$, мы получаем возможность вычисления значений искомой функции и в тех точках, где она не задана своими табличными значениями. Если $x_0 < x < x_n$ то говорят о задаче *интерполяции*, в противном случае — о задаче *экстраполяции* (рис. 1).

Ясно, что без дополнительных ограничений на функцию $\varphi(x)$ так сформулированная задача имеет бесконечно много решений — через данные n точек (x_i, f_i) на плоскости можно провести сколько угодно различных кривых. Обычно, с целью сужения множества возможных решений, в рассматриваемой задаче конкретизируют класс интерполирующих (экстраполирующих) функций — ограничиваются, скажем, многочленами или тригонометрическими многочленами заданной степени и т. п. и таким способом делают задачу осмысленной с математической точки зрения. ▶

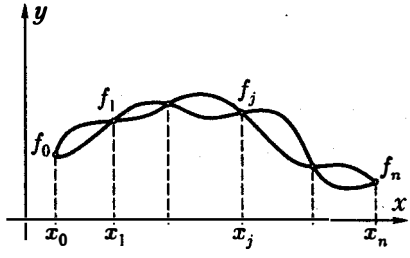


Рис. 1. Интерполяция

Пример 2. Пусть в ситуации, описанной предыдущим примером, значения функции $y = f(x)$ получены с некоторыми погрешностями δ_i, так что табличные значения имеют вид

$$\tilde{f}(x_i) = f(x_i) + \delta_i, \quad i = 0, 1, \ldots, n,$$

где $f(x_i)$ — точные (но неизвестные!) значения функции $f(x)$. Заменяя функцию $f(x)$ функцией $\varphi(x)$, уже бессмысленно требовать совпадения значений этих функций в точках x_i — удовлетворение этого требования приводит к необоснованно сложным выкладкам по построению $\varphi(x)$ и громоздкому выражению для нее, и в то же время не улучшает точности представления функции $f(x)$ функцией $\varphi(x)$ из-за неточного задания табличных значений $\tilde{f}$. В этой ситуации разумнее выбирать функцию $\varphi(x)$ как можно более простой, а близость $f(x)$ и $\varphi(x)$ понимать как близость в смысле некоторого критерия $\Psi(f, \varphi) \to \min$, описывающего суммарное отклонение функции $\varphi(x)$ от $f(x)$ в тех точках, где заданы табличные значения. В качестве критерия $\Psi(f, \varphi)$ близости можно взять, например, один из следующих

$$\Psi_1(f, \varphi) = \sum_{i=0}^{n} |f(x_i) - \varphi(x_i)| \longrightarrow \min, \quad \Psi_2(f, \varphi) = \sum_{i=0}^{n} |f(x_i) - \varphi(x_i)|^2 \longrightarrow \min,$$

$$\Psi_3(f, \varphi) = \max_{0 \leqslant i \leqslant n} |f(x_i) - \varphi(x_i)| \longrightarrow \min.$$

Ясно, что кроме указанных выше, возможны и другие меры близости. ▶

Так сформулированная задача называется *задачей сглаживания*. Отметим, что понятие «простоты» аппроксимирующей функции при постановке конкретной задачи сглаживания нуждается в уточнении.

§ 1. Интерполяция многочленами

Пусть на отрезке $[a, b]$ определена функция $f(x)$, обладающая некоторыми, оговоренными ниже, свойствами гладкости и пусть в точках $a = x_0 < x_1 < \ldots x_n = b$ заданы ее значения $f_i = f(x_i)$, $i = 0, 1, \ldots, n$. Построим многочлен наименьшей возможной степени, принимающий в точках x_i те же значения, что и функция $f(x)$

$$P_m(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m, \quad P_m(x_i) = f(x_i). \qquad (1)$$

Имеет место

Теорема. *Поставленная задача однозначно разрешима для любого набора значений $\{f_i\}_{i=1}^n$ в классе многочленов, степень которых не превышает n.*

◀ Условия (1) приводят к системе линейных уравнений

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n = f_0, \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n = f_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n = f_n \end{cases}$$

относительно коэффициентов $\{a_k\}_{k=0}^n$ искомого многочлена. Ее определитель

$$W = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}$$

— это определитель Вандермонда. Он не обращается в нуль, если только все узлы интерполяции (точки x_i) различны — система совместна и имеет единственное решение.

В то же время решить поставленную задачу многочленом, степень которого ниже n, вообще говоря, нельзя. Если бы это было возможно при любых исходных данных, то система $n + 1$ уравнения с $m + 1$ неизвестным, $n > m$, была бы всегда разрешима. Однако это не так — ранг матрицы системы в силу $x_i \neq x_j$, $i \neq j$, равен $m + 1$, но всегда можно придумать такой исходный набор f_i, что ранг расширенной матрицы будет больше и, следовательно, система окажется несовместной.

Многочлены же, степень которых превышает n, из указанной системы не могут быть определены однозначно. ▶

Многочлен $P_n(x)$, принимающий в заданных $n + 1$ точке те же значения, что и функция $f(x)$, называется ее интерполяционным многочленом Лагранжа. Как показано выше, интерполяционный многочлен Лагранжа существует для произвольных интерполяционных данных и определяется ими единственным образом.

1.1. Каноническое представление интерполяционного многочлена

Пусть $\Lambda_i(x)$ — многочлен степени не выше n, обладающий тем свойством, что

$$\Lambda_i(x_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \tag{2}$$

Как сказано выше, условия (2) определяют функцию $\Lambda_i(x)$ однозначно — это интерполяционный многочлен Лагранжа, принимающий в узле с номером i значение 1, а во всех прочих узлах значение 0 (рис. 2).

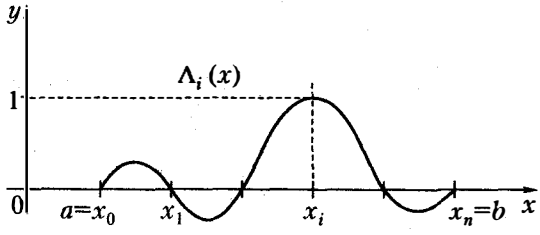


Рис. 2. Базисный интерполяционный многочлен

Функции $\Lambda_i(x)$ образуют *базис Лагранжа, ассоциированный с рассматриваемой интерполяционной задачей.*

Укажем явный вид многочленов $\Lambda_i(x)$. Для этого заметим, что поскольку $\Lambda_i(x)$ обращается в нуль в n узлах x_k, $k \neq i$, и является многочленом степени не выше n, то он представим в виде

$$\Lambda_i(x) = C(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n).$$

Постоянную C определим из условия $\Lambda_i(x_i) = 1$ и окончательно получим

$$\boxed{\Lambda_i(x) = \frac{(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)}.} \tag{3}$$

Теперь заметим, что линейная комбинация многочленов $\Lambda_i(x)$ с произвольными коэффициентами f_i является многочленом, степень которого не превосходит n и в силу условий (2) в точке x_j принимает значение f_j

$$L_n(x_j) = f_0\Lambda_0(x_j) + f_1\Lambda_1(x_j) + \dots + f_n\Lambda_n(x_j) = f_j,$$

откуда, и в силу единственности, функция $L_n(x)$ — интерполяционный многочлен Лагранжа.

Каноническим представлением интерполяционного многочлена Лагранжа $L_n(x)$ назовем следующую форму его записи

$$\boxed{L_n(x) = f_0\Lambda_0(x) + f_1\Lambda_1(x) + \dots + f_n\Lambda_n(x) = \sum_{i=1}^{n} f_i\Lambda_i(x).} \tag{4}$$

Заметим в заключение, что функции $\Lambda_i(x)$ — стандартные многочлены, определяемые только выбором узлов интерполяции и не зависящие от значений f_i интерполируемой функции.

1.2. Точность интерполяции

Интерполяционный многочлен Лагранжа в узлах интерполяции по определению совпадает с интерполируемой функцией. Если интерполируемая функция — многочлен, степень которого не превышает n, то интерполяционный многочлен тождественно совпадает с ней во всех точках отрезка интерполяции.

Если же интерполируемая функция не является многочленом, то при $x \neq x_i$ уже, вообще говоря, $L_n(x) \neq f(x)$ и возникает вопрос о *качестве интерполяции*, т. е. о том, насколько велика может быть разница между значениями интерполяционного многочлена и интерполируемой функции.

Имеет место следующая

Теорема. *Если функция $f(x)$ обладает на промежутке $[a, b]$ производными до $n+1$ порядка включительно и $\left|f^{(n+1)}(x)\right| \leqslant C \ \forall\, x \in [a, b]$, то*

$$|L_n(x) - f(x)| \leqslant \frac{C}{(n+1)!}|\omega_n(x)|, \tag{5}$$

где $\omega_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$.

Оценка (5) позволяет управлять точностью замены функции $f(x)$ ее интерполяционным многочленом за счет выбора узлов интерполяции x_i. Известно, что если в качестве узлов интерполяции взять нули многочленов Чебышева $T_{n+1}(x)$ на промежутке $[a, b]$, то правая часть оценки (5) достигнет своего наименьшего значения

$$\max |\omega_n(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Любое другое распределение узлов интерполяции приводит к худшей оценке точности.

Для практики важен случай равноотстоящих узлов

$$x_0 = a, \quad x_j = x_0 + jh, \quad x_n = b, \quad h = \frac{b-a}{n}, \quad j = 1, 2, \ldots, n.$$

Тогда

$$\omega_n(x) = (x - x_0) \ldots (x - x_0 - nh) = (b-a)^{n+1} t\left(t - \frac{1}{n}\right) \ldots (t-1),$$

где $t \in [0, 1]$, $t = \frac{x - x_0}{nh}$.

Отметим некоторые частные соотношения, связанные с полученной общей оценкой.

1. *Линейная интерполяция с равноотстоящими узлами.*

В этом случае $n = 1$, $|t(t-1)| \leqslant 0{,}25 \ \forall\, t \in [0, 1]$. Интерполяционный многочлен имеет вид

$$L_1(x) = f(a)\frac{b-x}{b-a} + f(b)\frac{x-a}{b-a},$$

а оценка (5) —

$$\boxed{|L_1(x) - f(x)| \leqslant c\frac{(b-a)^2}{8}.}$$

2. *Параболическая интерполяция с равноотстоящими узлами.*

Здесь $n = 2$, $|t(t-0{,}5)(t-1)| \leqslant \frac{\sqrt{3}}{36}$ $\forall\, t \in [0, 1]$.

Интерполяционный многочлен имеет вид

$$L_2(x) = f(a)\Lambda_0(x) + f\left(\frac{a+b}{2}\right)\Lambda_1(x) + f(b)\Lambda_2(x),$$

где

$$\Lambda_0(x) = \frac{2}{(b-a)^2}\left(x - \frac{a+b}{2}\right)(x-b),$$

$$\Lambda_1(x) = -\frac{4}{(b-a)^2}(x-a)(x-b),$$

$$\Lambda_2(x) = \frac{2}{(b-a)^2}(x-a)\left(x - \frac{a+b}{2}\right).$$

Оценка (5):

$$\boxed{|L_2(x) - f(x)| \leqslant c\frac{(b-a)^3}{216}\sqrt{3}.}$$

3. *Кубическая интерполяция с равноотстоящими узлами.*

Здесь $n = 3$, $\left|t\left(t-\frac{1}{3}\right)\left(t-\frac{2}{3}\right)(t-1)\right| \leqslant \frac{1}{81}$ $\forall\, t \in [0, 1]$.

$$L_3(x) = f(a)\Lambda_0(x) + f\left(\frac{2a+b}{3}\right)\Lambda_1(x) + f\left(\frac{a+2b}{3}\right)\Lambda_2(x) + f(b)\Lambda_3(x),$$

где функции $\Lambda_j(x)$ могут быть найдены как и выше.

Оценка (5):

$$\boxed{|L_3(x) - f(x)| \leqslant c\frac{(b-a)^4}{1944}.}$$

§ 2. Интерполяция кусочно-полиномиальными функциями

Многочленная интерполяция, вкратце рассмотренная выше, реализует следующие идеи: *простой функцией*, заменяющей искомую, назначается многочлен, а *критерием близости* является совпадение интерполирующего многочлена и интерполируемой функции в заданных точках области определения. На практике для достижения разумной точности интерполяции приходится прибегать к многочленам довольно высоких степеней, что не всегда удобно. Аппарат кусочно-полиномиальных функций является с этой точки зрения более привлекательным.

Пусть на отрезке $[a, b]$ заданы точки $a = x_0 < x_1 < \ldots < x_n = b$, разбивающие этот отрезок на n частей. Функция $S(x)$, определенная на этом отрезке, называется *сплайном r-го порядка*, если на каждом элементе $\Delta_i = [x_i, x_{i+1}]$ разбиения отрезка $[a, b]$ она является многочленом степени не выше r. В точках $\{x_j\}_{j=1}^{n-1}$ эти многочлены определенным образом «склеиваются», при этом качество склейки регулируется *дефектом* сплайна — мы говорим, что рассматриваемый сплайн имеет дефект q, если в каждой из точек сопряжения у функции $S(x)$ существует не менее $r - q$ непрерывных производных.

2.1. Сплайны первого порядка дефекта 1

Как следует из данного выше определения, сплайн первого порядка дефекта 1 — это кусочно-линейная на промежутке $[a, b]$ функция, с изломами в узлах $\{x_j\}_{j=1}^{n-1}$ (рис. 3).

Поскольку на каждом элементе разбиения $\Delta_i = [x_i, x_{i+1}]$ линейная функция однозначно определяется своими значениями в концах, задача интерполяции функции $f(x)$ по ее значениям в узлах $\{x_j\}_{j=0}^{n}$ кусочно-линейной функцией $S_1^1(x)$ всегда однозначно разрешима.

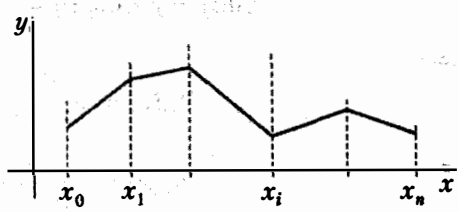


Рис. 3. Сплайн 1-го порядка дефекта 1

Введем в рассмотрение «единичные» интерполяционные функции

$$\Lambda_i(x), \quad i = 0, 1, \ldots, n,$$

определяемые условиями: $\Lambda_i(x)$ определена на $[a, b]$ и линейна на каждом элементе разбиения, в узлах $\{x_j\}_{j=0}^{n}$ функция $\Lambda_i(x)$ удовлетворяет условиям

$$\Lambda_i(x_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Легко видеть, что функции $\Lambda_i(x)$ задаются соотношениями

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

для всех внутренних узлов $i = 1, 2, \ldots, n-1$ и соотношениями

$$\Lambda_0(x) = \begin{cases} \dfrac{x - x_1}{x_0 - x_1}, & x \in [x_0, x_1], \\ 0, & x \notin [x_0, x_1], \end{cases} \qquad \Lambda_n(x) = \begin{cases} \dfrac{x - x_{n-1}}{x_n - x_{n-1}}, & x \in [x_{n-1}, x_n], \\ 0, & x \notin [x_{n-1}, x_n] \end{cases}$$

для крайних узлов. Графики функций $\Lambda_i(x)$ представлены на рис. 4.

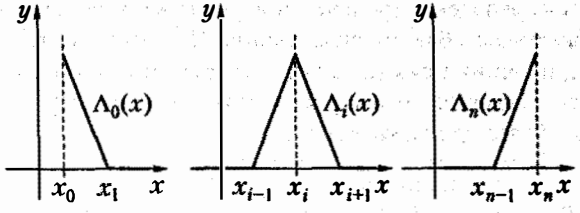


Рис. 4. Функции $\Lambda_i(x)$

В соответствии с принятой нами терминологией они образуют базис Лагранжа, ассоциированный с рассматриваемой интерполяционной задачей. Сплайн $S_1^1(x)$, принимающий в точках $\{x_j\}_{j=0}^{n}$ заданные значения f_j, может быть представлен в виде

$$S_1^1(x) = f_0\Lambda_0(x) + f_1\Lambda_1(x) + \ldots + f_n\Lambda_n(x).$$

Точность интерполяции *гладкой* функции $f(x)$ кусочно-линейной $S_1^1(x)$ не хуже $\varepsilon = M\frac{h^2}{8}$, где $h = \max\limits_i |x_{i+1} - x_i|$, $M = \max\limits_{[a,b]} |f'(x)|$. Этот результат немедленно следует из общей оценки точности интерполяции § 1.

2.2. Сплайны третьего порядка дефекта 2

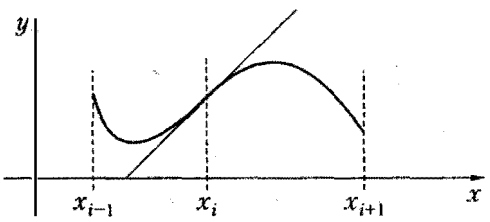


Рис. 5. Сплайн 3-го порядка дефекта 2

Функция $S_3^2(x)$ является сплайном третьего порядка, если она на каждом элементе разбиения совпадает с некоторым многочленом, степень которого не выше третьей. То обстоятельство, что дефект сплайна равен двум, означает, что в точках x_j склейки двух различных многочленов у них есть общая касательная (рис. 5).

Сплайны третьего порядка дефекта 2 являются естественным аппаратом для решения задачи интерполяции функции вместе со значениями ее производной. Действительно, заметим, что выполняется следующее

Утверждение. *Если на промежутке $[x_i, x_{i+1}]$ задана произвольная гладкая функция $f(x)$, то существует единственный многочлен, степень которого не превышает третьей, принимающий на концах отрезка те же значения, что и $f(x)$, и производная которого на концах отрезка принимает те же значения, что и производная $f'(x)$.*

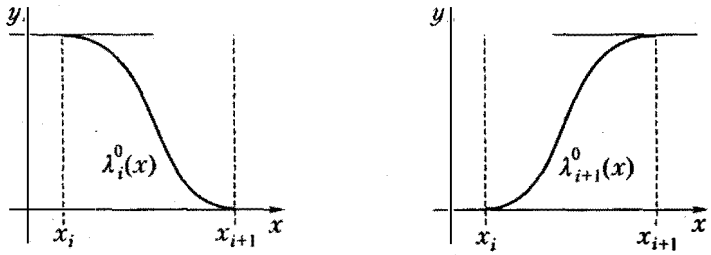


Рис. 6. Функции $\lambda_i^0(x)$ и $\lambda_{i+1}^0(x)$

◀ Для доказательства этого факта заметим, что функция $\lambda_i^0(x)$, принимающая в точке x_i значение 1, в точке x_{i+1} — значение 0, и производные которой в этих точках равны нулю, может быть определена однозначно (рис. 6), если известно, что она является многочленом степени не выше третьей. Действительно, функция

$$\lambda_i^0(x) = -2\left(\frac{x - x_{i+1}}{x_i - x_{i+1}}\right)^3 + 3\left(\frac{x - x_{i+1}}{x_i - x_{i+1}}\right)^2$$

удовлетворяет всем предъявленным выше требованиям. Если предположить, что существует еще одна функция, удовлетворяющая тем же требованиям, то их разность будет многочленом, степени не выше третьей, принимающим на концах отрезка нулевые значения месте со своей производной. Несложные выкладки показывают, что такой многочлен является тождественным нулем. Аналогично устанавливаем, что

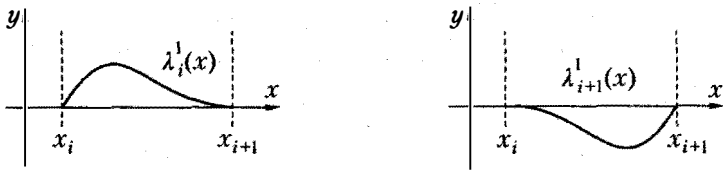


Рис. 7. Функции $\lambda_i^1(x)$ и $\lambda_{i+1}^1(x)$

однозначно определяемая функция

$$\lambda_{i+1}^0(x) = -2\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^3 + 3\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^2$$

принимает в точке x_i значение 0, в точке x_{i+1} — значение 1, а производные ее в этих точках равны нулю. Так же легко можно построить однозначно определенные функции $\lambda_i^1(x)$ и $\lambda_{i+1}^1(x)$ (рис. 7), принимающие на концах рассматриваемого промежутка нулевые значения и производные которых в соответствующих концах равны соответственно 0 и 1.

$$\lambda_i^1(x) = (x_i - x_{i+1})\left[\left(\frac{x - x_{i+1}}{x_i - x_{i+1}}\right)^3 - \left(\frac{x - x_{i+1}}{x_i - x_{i+1}}\right)^2\right],$$

$$\lambda_{i+1}^1(x) = -(x_i - x_{i+1})\left[\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^3 - \left(\frac{x - x_i}{x_{i+1} - x_i}\right)^2\right].$$

Теперь справедливость доказываемого утверждения очевидна — многочлен $P_3(x)$, задаваемый соотношением

$$P_3(x) = f(x_i)\lambda_i^0(x) + f'(x_i)\lambda_i^1(x) + f(x_{i+1})\lambda_{i+1}^0(x) + f'(x_{i+1})\lambda_{i+1}^1(x),$$

решает поставленную задачу. ▶

Пусть функция $f(x)$ определена и непрерывно дифференцируема на отрезке $[a, b]$ и задано разбиение $a = x_0 < x_1 < \ldots < x_n = b$. Рассмотрим задачу *построения на заданном разбиении отрезка $[a, b]$ кусочно-кубической функции, интерполирующей значения функции $f(x)$ и ее производной.*

Приведенные выше соображения показывают, что эта задача всегда однозначно разрешима и ее решение дается сплайном

$$S_3^2(x) = \sum_{i=0}^{n} f(x_i) \cdot \Lambda_i^0(x) + \sum_{i=0}^{n} f'(x_i) \cdot \Lambda_i^1(x), \qquad (1)$$

где функции $\Lambda_i^0(x)$ и $\Lambda_i^1(x)$ — «единичные» сплайны, ассоциированные соответственно с задачами $f(x_j) = 0$, $j \neq i$, $f(x_i) = 1$, $f'(x_j) = 0$, $j = 0, 1, \ldots, n$ и $f(x_j) = 0$, $j = 0, 1, \ldots, n$, $f'(x_i) = 1$, $f'(x_j) = 0$, $j \neq i$. Легко понять (рис. 8 и 9), что они являются склейкой функций $\lambda_i^0(x)$ ($\lambda_i^1(x)$) на промежутках, смежных с i-м узлом разбиения. Особо отметим, что на каждом элементе разбиения сумма (1) содержит не более четырех слагаемых, так как функции $\Lambda_i^0(x)$ и $\Lambda_i^1(x)$ отличны от нуля только на промежутках, смежных с i-ым узлом разбиения.

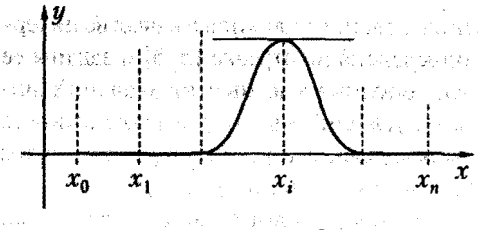


Рис. 8. Функции $\Lambda_i^0(x)$

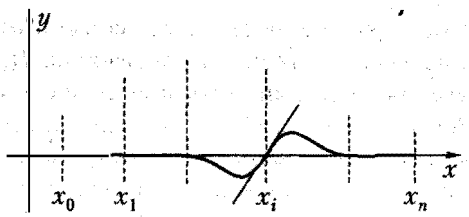


Рис. 9. Функции $\Lambda_i^1(x)$

2.3. Сплайны третьего порядка дефекта 1

Функция $S_3^1(x)$ является сплайном третьего порядка дефекта 1, если на каждом промежутке разбиения она является многочленом степени не выше третьей и в узлах разбиения x_i эти многочлены склеены так, что эта функция обладает непрерывными производными до второго порядка включительно.

Заметим, что для однозначного задания сплайна третьего порядка дефекта 2, необходимо задать его значения, вообще говоря, во всех узлах $\{x_i\}$, $i = 1, 2, \ldots, n$, а также наложить еще два дополнительных условия, которые, как правило, связаны с поведением сплайна в крайних точках отрезка.

◀ Как показывает соотношение (1), кусочно-кубическая на промежутке $[a, b]$ функция может быть однозначно определена набором из $2n + 2$ чисел — $n + 1$ значения сплайна и $n + 1$ значения его первых производных в узлах. Если считать значения сплайна заданными, то нам не будет хватать еще $n + 1$ значения его производных. Условие непрерывности вторых производных в узлах — это $n - 1$ связь, наложенная на $n + 1$ недостающий параметр. Следует добавить еще два условия, что, конечно, не гарантирует разрешимости (и, тем более, однозначной) поставленной задачи, но позволяет надеяться на таковую. ▶

Замечание. Система условий

$$\frac{d^2}{dx^2} S_3^1(x_i^-) = \frac{d^2}{dx^2} S_3^1(x_i^+), \quad i = 1, 2, \ldots, n-1,$$

где через $f(x^-)$ и $f(x^+)$ обозначаются значения функции $f(x)$ в точке x_i слева и справа соответственно, представляет собой систему $n - 1$ линейного уравнения с $n + 1$ неизвестным и добавление еще двух условий уравнивает количество неизвестных и количество уравнений.

Эти дополнительные к значениям сплайна в узлах условия порождаются характером решаемой задачи, имеющейся информацией о поведении сплайна и, как правило, являются условиями одного из следующего типов:

— задаются значения первых производных в концах отрезка

$$\frac{d}{dx} S_3^1(a) = f_0^1, \quad \frac{d}{dx} S_3^1(b) = f_n^1,$$

— задаются значения вторых производных в концах отрезка

$$\frac{d^2}{dx^2} S_3^1(a) = f_0^2, \quad \frac{d^2}{dx^2} S_3^1(b) = f_n^2,$$

— задаются условия равенства первых и вторых производных в концах отрезка

$$\frac{d}{dx} S_3^1(a) = \frac{d}{dx} S_3^1(b), \quad \frac{d^2}{dx^2} S_3^1(a) = \frac{d^2}{dx^2} S_3^1(b).$$

Вышеизложенное делает естественным использование сплайнов в качестве интерполирующих функций. Если функция $f(x)$ определена на отрезке $[a, b]$ и заданы ее значения в узлах интерполяционной сетки, то, дополнив массив этих значений любой парой перечисленных выше дополнительных условий, мы получим возможность построить единственный сплайн третьего порядка дефекта 1, удовлетворяющий этим условиям и принимающий в узлах сетки те же значения, что и интерполируемая функция. В заключение заметим, что влияние, оказываемое дополнительными условиями на сплайн, с удалением от концов промежутка ослабевает.

§ 3. Дробно-рациональная интерполяция

Более гибким инструментом интерполяции, в сравнении с многочленами и сплайнами, являются дробно-рациональные функции. Они оказываются особенно удобными в ситуациях, когда следует адекватно отразить свойства интерполируемой функции, связанные с выпуклостью, резкими скачками на маленьких промежутках или, если интерполируемая функция оказывается неограниченной на конечном промежутке, а также в некоторых других случаях.

Пусть x_i, $i = 1, 2, \ldots, n$, — заданные узлы интерполяции и $f(x_i)$ — соответствующие значения интерполируемой функции. Положим

$$f(x) \approx \frac{\alpha_p x^p + \alpha_{p-1} x^{p-1} + \ldots + \alpha_1 x + \alpha_0}{\beta_q x^q + \beta_{q-1} x^{q-1} + \ldots + \beta_1 x + \beta_0} = R(x),$$

где $p + q = n - 1$, и найдем коэффициенты α_s и β_j из условий $f(x_i) = R(x_i)$, $i = 1, 2, \ldots, n$. Последние приводят к системе n линейных уравнений

$$\begin{cases} f(x_1) \sum\limits_{s=0}^{q} \beta_s x_1^s - \sum\limits_{s=0}^{p} \alpha_s x_1^s = 0, \\ f(x_2) \sum\limits_{s=0}^{q} \beta_s x_2^s - \sum\limits_{s=0}^{p} \alpha_s x_2^s = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ f(x_n) \sum\limits_{s=0}^{q} \beta_s x_n^s - \sum\limits_{s=0}^{p} \alpha_s x_n^s = 0 \end{cases}$$

относительно $(p+1)+(q+1) = p+q+2 = n+1$ неизвестного коэффициента, которая всегда разрешима.

Рассмотрим функцию

$$F(x, y) = y \sum_{s=0}^{q} \beta_s x^s - \sum_{s=0}^{p} \alpha_s x_s.$$

В предположении, что имеет место дробно-рациональная интерполяция, заключаем, что $F(x, y) \equiv 0 \ \forall\, x, y = R(x)$. Поэтому определитель

$$\Delta = \begin{vmatrix} y & xy & x^2y & \dots & x^qy & 1 & x & x^2 & \dots & x^p \\ f(x_1) & x_1f(x_1) & x_1^2f(x_1) & \dots & x_1^qf(x_1) & 1 & x_1 & x_1^2 & \dots & x_1^p \\ f(x_2) & x_2f(x_2) & x_2^2f(x_2) & \dots & x_2^qf(x_2) & 1 & x_2 & x_2^2 & \dots & x_2^p \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ f(x_n) & x_nf(x_n) & x_n^2f(x_n) & \dots & x_n^qf(x_n) & 1 & x_n & x_n^2 & \dots & x_n^p \end{vmatrix} \equiv 0$$

для любых x и y. Его разложение по первой строке дает нам возможность найти коэффициенты многочлена $F(x, y)$ и, тем самым, коэффициенты дробно-рациональной интерполяции.

§ 4. Сглаживание и метод наименьших квадратов

Если значения интерполируемой функции известны неточно

$$\tilde{f}(x_i) = f(x_i) + \delta_i, \quad i = 0, 1, \dots, n,$$

то интерполяция не очень целесообразна и усилия, потраченные на построение интерполяционного многочлена (или сплайна), могут оказаться неадекватными результату. В этой ситуации более разумным представляется подбор некоторой функции $\varphi(x)$, пусть и не проходящей через все точки $(x_i, f(x_i))$, но достаточно близкой к ним и в то же время имеющей более «простую» структуру, чем интерполяционный многочлен высокой степени (удобно представима в аналитической форме, легко вычислима и т. п.).

Высказанные выше соображения могут быть формализованы по-разному. Мы рассмотрим здесь следующую постановку:

пусть $\varphi(x, \alpha_1, \alpha_2, \dots, \alpha_r)$ — известная функция переменной x, определенная на отрезке $[a, b]$ при любых значениях параметров α_j, $j = 1, 2, \dots, r$. Будем говорить, что $\varphi(x, \alpha_1, \alpha_2, \dots, \alpha_r)$ является *МНК-приближением* (т. е. *приближением метода наименьших квадратов*) для $f(x)$, если существует такой набор параметров α_j, $j = 1, 2, \dots, r$, что

$$S(\alpha_1, \alpha_2, \dots, \alpha_r) = \sum_{i=1}^{n} \left(\varphi(x_i, \alpha_1, \alpha_2, \dots, \alpha_r) - \tilde{f}(x_i)\right)^2 \to \min. \qquad (1)$$

Критерий (1) приводит к системе уравнений

$$\left\{ \frac{\partial}{\partial \alpha_i} S(\alpha_1, \alpha_2, \dots, \alpha_r) = 0, \quad i = 1, 2, \dots, r, \right. \qquad (2)$$

для нахождения параметров $\alpha_1, \alpha_2, \dots, \alpha_r$. Если сглаживающая функция зависит от параметров линейно, то система (2) также линейна и может быть легко разрешена.

Заметим, что найденная таким способом функция $\varphi(x, \alpha_1, \alpha_2, \dots, \alpha_r)$ не обязана ни в каком разумном смысле (кроме оговоренного!) походить на истинную функцию $f(x)$. Для тех же исходных данных можно «назначить» другую функцию, по критерию (1) построить систему (2), разрешить ее и получить другую функцию, сглаживающую те же данные. Предлагаемый метод *из заданного класса* функций выбирает наилучшую в смысле критерия (1), но не дает никаких указаний по *выбору* этого класса.

Рассмотрим несколько примеров.

4.1. Линейное сглаживание

Найдем линейную функцию $\varphi(x, \alpha_1, \alpha_2) = \alpha_1 x + \alpha_2$, которая в смысле критерия (1) наилучшим образом сглаживает экспериментальные данные $\tilde{f}(x_i) = f(x_i) + \delta_i$, $i = 0, 1, 2, \ldots, n$. С целью упрощения выкладок будем считать в дальнейшем, что выполнено условие $\sum x_i = 0$. Для этого достаточно ввести новую независимую переменную

$$X = x - \frac{1}{n+1} \sum_0^n x_i.$$

Система (2) запишется в виде

$$\begin{cases} \alpha_1 \sum_{i=0}^n x_i^2 = \sum_{i=0}^n x_i \tilde{f}_i, \\ \alpha_2 = \frac{1}{n+1} \sum_{i=0}^n \tilde{f}_i, \end{cases}$$

откуда

$$\alpha_1 = \frac{\sum x_i \tilde{f}_i}{\sum x_i^2}, \quad \alpha_2 = \frac{1}{n+1} \sum \tilde{f}_i.$$

Попробуем теперь ответить на вопрос о том, насколько хорошо нам удалось решить задачу сглаживания, имея в виду следующее: пусть истинная кривая задается соотношением $y = f(x)$, значения $\tilde{f}(x_i)$, взятые нами для построения сглаживающей функции, не являются точными и не сильно отличаются от истинных

$$\tilde{f}(x_i) = f(x_i) + \delta_i, \quad i = 0, 1, \ldots, n,$$

здесь δ_i — ошибки измерения (рис. 10).

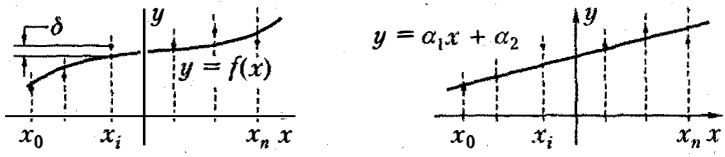


Рис. 10. Экспериментальные точки и сглаживающая прямая

Применив метод наименьших квадратов к значениям $\tilde{f}(x_i)$, мы построим прямую (рис. 10) $y = \alpha_1 x + \alpha_2$, минимизирующую сумму квадратов отклонений от экспериментальных точек. Эта прямая должна представлять функцию $y = f(x)$ в том смысле, что в любой точке отрезка $[a, b]$ неизвестное значение этой функции мы будем заменять на значение найденной линейной (рис. 11). Хотелось бы уметь отвечать на вопрос о том, как велика может быть разница $\Delta(x) = |\alpha_1 x + \alpha_2 - f(x)|$, если известны параметры ошибок измерения δ_i. Исчерпывающий ответ на поставленный вопрос дает анализ вероятностных характеристик ошибок δ_i. Здесь же мы получим

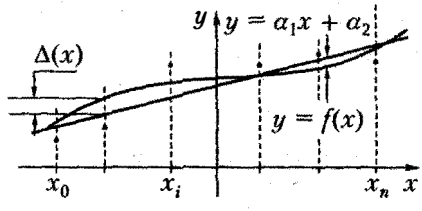


Рис. 11. Погрешность сглаживания

некоторые результаты, носящие качественный характер, предполагая только, что все ошибки измерений равномерно ограничены — $|\delta_i| \leqslant \varepsilon$, $i = 0, \ldots, n$, — и истинная зависимость линейна — $f(x) = A_1 x + A_2$.

При этих допущениях модель измерений имеет вид

$$\tilde{f}(x_i) = A_1 x + A_2 + \delta_i, \quad i = 0, 1, \ldots, n.$$

Из соотношений (2) получаем

$$\alpha_1 = \frac{\sum x_i \tilde{f}_i}{\sum x_i^2} = \frac{\sum x_i (A_1 x + A_2 + \delta_i)}{\sum x_i^2} = A_1 \frac{\sum x_i^2}{\sum x_i^2} + A_2 \frac{\sum x_i}{\sum x_i^2} + R_{\alpha_1},$$

$$\alpha_2 = \frac{1}{n+1} \sum \tilde{f}_i = \frac{1}{n+1} \sum (A_1 x + A_2 + \delta_i) = \frac{A_1}{n+1} \sum x_i + A_2 + R_{\alpha_2}.$$

Учитывая, что $\sum x_i = 0$, перепишем эти соотношения в виде

$$\alpha_1 - A_1 = \frac{\sum x_i \delta_i}{\sum x_i^2}, \quad \alpha_2 - A_2 = \frac{1}{n+1} \sum \delta_i.$$

В силу неравенства Коши—Буняковского

$$\left| \sum x_i \delta_i \right| \leqslant \sqrt{\sum x_i^2} \cdot \sqrt{\sum \delta_i^2},$$

поэтому

$$|\alpha_1 - A_1| \leqslant \frac{\sqrt{\sum \delta_i^2}}{\sqrt{\sum x_i^2}} \leqslant \frac{\varepsilon}{\sqrt{\frac{1}{n+1} \sum x_i^2}}, \quad |\alpha_2 - A_2| \leqslant \varepsilon.$$

Заметим, что величина

$$\sigma_x = \sqrt{\frac{1}{n+1} \sum x_i^2}$$

характеризует диапазон изменения независимой переменной.

Пусть x^* — произвольная точка из отрезка $[a, b]$. В этой точке полученные выше оценки дают

$$\Delta(x^*) = \left| \alpha_1 x^* + \alpha_2 - A_1 x^* - A_2 \right| \leqslant |\alpha_1 - A_1| \cdot |x^*| + |\alpha_2 - A_2| \leqslant \varepsilon \left(\frac{|x^*|}{\sigma_x} + 1 \right).$$

Эта оценка позволяет сделать следующие заключения: сглаживание тем качественнее, чем шире диапазон изменения переменной, при этом точность сглаживания выше в середине промежутка изменения независимой переменной и ухудшается к краям; уменьшение величины ε, т. е. повышение точности исходных данных, увеличивает точность сглаживания (впрочем, это интуитивно ясно и без всяких выкладок). Еще раз отметим, что приведенные выше рассуждения носят качественный характер и по необходимости грубы. Более точное исследование требует более тонких методов.

4.2. Линейное по параметрам сглаживание

Совершенно аналогично вышеизложенному строятся процедуры сглаживания любыми другими, не обязательно линейными по независимой переменной функциями. Важно только, чтобы они были *линейными по параметрам*

$$\varphi(x, \alpha_1, \alpha_2, \ldots, \alpha_r) = \sum_{j=1}^{r} \alpha_j \varphi_j(x).$$

При этом критерий (1) приводит к системе (2) линейных уравнений относительно неизвестных параметров, которая может быть эффективно разрешена.

Следует только иметь в виду вот какое важное обстоятельство — если сглаживающие функции $\varphi_j(x)$ обладают свойством ортогональности на интерполяционной сетке $a = x_0 < x_1 < \ldots < x_n = b$

$$\sum_{s=1}^{r} \varphi_j(x_s) \cdot \varphi_k(x_s) = 0, \quad j \neq k,$$

то система (2) имеет очень простую структуру, а полученные параметры оказываются независимыми. В противном случае параметры зависимы, что может вызвать некоторые затруднения при анализе их точности.

§ 5. Интерполяция функций двух переменных

Пусть D — замкнутое ограниченное множество на плоскости, $P_{i,j} = (x_i, y_j)$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, — точки из D, выбранные некоторым специальным образом, и $\varphi(x, y; \alpha_1, \alpha_2, \ldots, \alpha_n)$ — класс функций, каждая из которых однозначно задается значениями параметров α_j. Рассмотрим задачу нахождения параметров α_j так, чтобы функция $\varphi(x, y; \alpha_1, \alpha_2, \ldots, \alpha_n)$ (и, возможно, ее производные) принимала в узлах $P_{i,j}$ такие же значения, как и заданная на D функция $f(x, y)$.

Ясно, что без дополнительных оговорок относительно соотношения между количеством параметров и количеством интерполяционных ограничений поставленная задача является неопределенной. Кроме того, в рассматриваемом случае важную роль играет не только количество узлов интерполяции, но и их расположение, существенно влияющее на разрешимость поставленной задачи. Полное решение задачи в случае произвольных функций $\varphi(x, y; \alpha_1, \alpha_2, \ldots, \alpha_n)$ достаточно затруднительно.

Мы здесь рассмотрим некоторые частные случаи общей теории. В качестве D будут выступать множества простой геометрической структуры — треугольник, прямоугольник или их объединение (в конечном числе). В качестве интерполирующих функций будем рассматривать многочлены или функции, являющиеся склейкой многочленов. Основным требованием к избранному классу функций является требование *однозначной разрешимости* задачи интерполяции в этом классе функций.

Пусть $f_{i,j}^{p,q}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, $p, q = 0, 1, \ldots$, — интерполяционные данные (это могут быть значения $f_{i,j} = f_{i,j}^{0,0}$ интерполируемой функции или значения $f_{i,j}^{p,q}$, $p \neq 0 \cup q \neq 0$, ее производных).

Основной интерполяционной задачей мы назовем задачу нахождения многочлена $M(x, y)$, который в заданных узлах $P_{i,j}$ принимает заданные значения $f_{i,j}^{0,0}$ и производные которого принимают в узлах интерполяции заданные значения $f_{i,j}^{p,q}$, $p \neq 0 \cup q \neq 0$,

$$M(P_{i,j}) = f_{i,j}^{0,0}, \quad \frac{\partial^{p+q}}{\partial x^p \partial y^q} M(P_{i,j}) = f_{i,j}^{p,q}.$$

Как и в случае одномерной интерполяции, *базисом Лагранжа, ассоциированным с основной интерполяционной задачей*, назовем совокупность многочленов $\Lambda_{i,j}^{p,q}(x, y)$, удовле-

творяющих «единичным» исходным интерполяционным данным

$$\frac{\partial^{s+r}}{\partial x^s \partial y^r}\Lambda_{i,j}^{k,l}(P_{u,v}) = \begin{cases} 1, & i = u \cap j = v \cap s = k \cap r = l, \\ 0, & i \neq u \cup j \neq v \cup s \neq k \cup r \neq l. \end{cases}$$

Легко установить, что если базис Лагранжа существует и единствен, то для произвольных интерполяционных данных поставленная задача однозначно разрешима и искомый интерполяционный многочлен может быть представлен в виде

$$M(x, y) = \sum_{i,j,p,q} f_{i,j}^{p,q}\Lambda_{i,j}^{p,q}.$$

Перейдем к рассмотрению конкретных интерполяционных схем.

5.1. Прямоугольная интерполяция. Четырехузловая схема

Пусть D — прямоугольник на плоскости, стороны которого параллельны осям координат (рис. 12, слева). Выберем в качестве узлов интерполяции вершины прямоугольника $P_{0,0} = (x_0, y_0)$, $P_{0,1} = (x_0, y_1)$, $P_{1,0} = (x_1, y_0)$, $P_{1,1} = (x_1, y_1)$. Пусть $f_{0,0}$, $f_{0,1}$, $f_{1,0}$, $f_{1,1}$ — произвольные числа. Тогда

существует единственный многочлен, линейный по x при каждом y, и линейный по y при каждом x, принимающий в узлах $P_{i,j}$ заданные значения $f_{i,j}$.

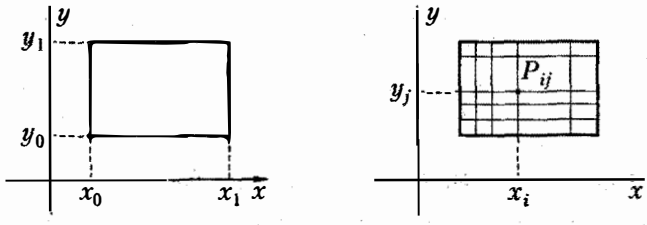


Рис. 12. Узлы прямоугольной интерполяции

◄ Функции $\Lambda_{i,j}(x, y)$ (рис. 13), определенные на прямоугольнике D соотношениями

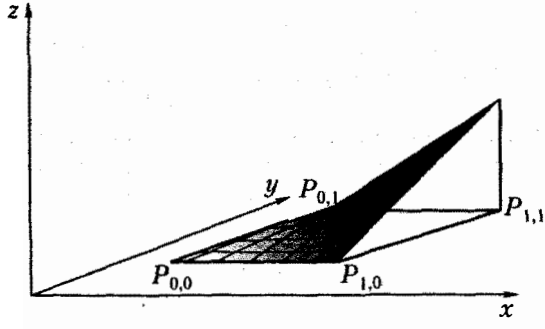


Рис. 13. Базисная функция $\Lambda_{1,1}(x, y)$

$$\Lambda_{i,j}(x, y) = \Lambda_i(x)\Lambda_j(y),$$

где $\Lambda_s(t)$ — базисные функции Лагранжа (являющиеся многочленами первой степени), ассоциированные с двухточечной интерполяционной задачей на промежутках $[x_0, x_1]$ и $[y_0, y_1]$ соответственно, образуют базис Лагранжа, ассоциированный с рассматриваемой интерполяционной задачей. Поэтому многочлен $M_{1,1}(x, y)$, задаваемый соотношением

$$M_{1,1}(x, y) = \sum_{i=0,j=0}^{1} f_{i,j}\Lambda_{i,j}(x, y), \qquad (1)$$

определен однозначно, в узле $P_{i,j}$ принимает значение $f_{i,j}$ и, следовательно, является решением исследуемой задачи. ▶

Пусть теперь в вершинах прямоугольника наряду со значениями функции

$$f_{i,j} = f_{i,j}^{0,0}$$

заданы еще и значения ее частных производных

$$\frac{\partial f}{\partial x}(P_{i,j}) = f_{i,j}^{1,0}, \quad \frac{\partial f}{\partial y}(P_{i,j}) = f_{i,j}^{0,1}.$$

Тогда

существует единственный многочлен, кубический по x при каждом y, и кубический по y при каждом x, который принимает в узлах $P_{i,j}$ заданные значения $f_{i,j}^{0,0}$, а его производные заданные значения $f_{i,j}^{r,s}$.

◀ Достаточно заметить, что функции

$$\Lambda_{i,j}^{r,s}(x, y) = \lambda_i^r(x)\lambda_j^s(y),$$

где $\lambda_i^r(t)$ — многочлены, степени которых не выше третьей и которые на концах отрезка $[t_0, t_1]$ удовлетворяют условиям

$$\lambda_i^0(t_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad \frac{d\lambda_i^0}{dt}(t_j) = 0, \quad \lambda_i^1(t_j) = 0, \quad \frac{d\lambda_i^1}{dt}(t_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

образуют базис Лагранжа, ассоциированный с рассматриваемой интерполяционной задачей. Искомый интерполяционный многочлен дается соотношением

$$M_{3,3}(x, y) = \sum_{i,j,r,s=0}^{1} f_{i,j}^{r,s}\Lambda_{i,j}^{r,s}(x, y). \ ▶$$

5.2. Прямоугольная интерполяция. Многоузловая схема

Пусть D — прямоугольник на плоскости, стороны которого параллельны осям координат,

$$D = \{(x, y):\ a \leqslant x \leqslant b,\ c \leqslant y \leqslant d\}.$$

Разобьем стороны прямоугольника вдоль осей OX и OY на n и m частей соответственно и нанесем на него прямоугольную сетку, отвечающую разбиению (рис. 12, справа). В узлах сетки $P_{i,j} = (x_i, y_j)$, $i = 0, 1, \ldots, n$, $j = 0, 1, 2, \ldots, m$, зададим числа $f_{i,j}$ и будем искать многочлен от двух переменных, который в точках $P_{i,j}$ принимает значения $f_{i,j}$.

Как и в предыдущем пункте, легко убедиться в том, что поставленная интерполяционная задача однозначно разрешима для любых исходных данных $f_{i,j}$ и ее решение дается многочленом $M_{n,m}(x, y)$, степень которого не выше n по x при каждом значении y и не выше m по y при каждом значении x.

◀ Доказательство немедленно следует из того факта, что функции

$$\Lambda_{i,j}(x, y) = \Lambda_i(x)\Lambda_j(y), \quad i = 0, 1, \ldots n,\ j = 0, 1, \ldots, m,$$

где $\Lambda_s(\cdot)$, $s = 0, 1, \ldots, r$ — базис Лагранжа, ассоциированный с соответствующей r-точечной интерполяционной задачей. Поскольку $\Lambda_s(t_k) = \delta_{i,k}$,

$$\Lambda_{i,j}(x_k, y_l) = \begin{cases} 1, & i = k \cap j = l, \\ 0, & i \neq k \cup j \neq l, \end{cases}$$

и, следовательно, последние образуют базис Лагранжа, ассоциированный с рассматриваемой интерполяционной задачей. Поэтому многочлен

$$M_{n,m}(x, y) = \sum_{i,j=0}^{n,m} f_{i,j}\Lambda_{i,j}(x, y)$$

решает поставленную задачу. ▶

В приложениях часто используют рассмотренные интерполяционные многочлены с $n = m = 2$. В этом случае мы располагаем девятью узлами интерполяции, расположенными по периметру прямоугольника, за исключением центрального. Интерполяционный многочлен имеет вид

$$M_{2,2} = a_{0,0} + a_{1,0}x + a_{0,1}y + a_{1,1}xy + a_{2,0}x^2 + a_{0,2}y^2 + a_{2,1}x^2y + a_{1,2}xy^2 + a_{2,2}x^2y^2,$$

и заданные в узлах значения интерполируемой функции позволяют однозначно найти коэффициенты $a_{i,j}$.

В то же время наличие центрального узла не всегда удобно, и возникает потребность в достаточно простом по структуре интерполяционном многочлене, его не использующем. Покажем, как можно модифицировать приведенную выше схему с целью исключения центрального узла.

Потребуем от интерполяционного многочлена, чтобы, оставаясь многочленом степени не выше второй *по каждой переменной*, он был бы многочленом степени не выше третьей *по совокупности переменных*. Такой многочлен не содержит члена четвертого порядка $a_{2,2}x^2y^2$ и может быть записан в виде

$$\widetilde{M}_{2,2} = a_{0,0} + a_{1,0}x + a_{0,1}y + a_{1,1}xy + a_{2,0}x^2 + a_{0,2}y^2 + a_{2,1}x^2y + a_{1,2}xy^2.$$

Докажем, что модифицированная таким образом интерполяционная задача однозначно разрешима.

◀ Пусть $P_{i,j}$, $i, j = 0, 1, 2$, — узлы интерполяции и $\Lambda_{i,j}(x, y)$ — базис Лагранжа, ассоциированный с девятиточечной интерполяционной задачей на прямоугольнике. Пусть далее числа $\alpha_{i,j}$ выбраны так, что функции

$$\widetilde{\Lambda}_{i,j}(x, y) = \Lambda_{i,j}(x, y) - \alpha_{i,j}\Lambda_{1,1}(x, y)$$

являются многочленами степени не выше второй по каждой переменной, и степени не выше третьей по совокупности переменных. Из условия равенства нулю коэффициента при x^2y^2 в разности $\Lambda_{i,j}(x, y) - \alpha_{i,j}\Lambda_{1,1}(x, y)$ числа $\alpha_{i,j}$ могут быть найдены однозначно. Легко видеть, что при этом

$$\widetilde{\Lambda}_{i,j}(P_{k,l}) = \begin{cases} 1, & i = k \cap j = l, \\ 0, & i \neq k \cup j \neq l, \end{cases}$$

во всех узлах $P_{i,j} \neq P_{1,1}$ периметра и функции $\widetilde{\Lambda}_{i,j}(x, y)$ образуют базис Лагранжа, ассоциированный с модифицированной (т.е. восьмиточечной) задачей. Далее убеждаемся, что многочлен

$$\widetilde{M}_{2,2}(x, y) = \sum_{i,j\neq 1} f_{i,j}\widetilde{\Lambda}_{i,j}(x, y)$$

решает поставленную задачу. ▶

5.3. Треугольная интерполяция

В случае прямоугольной интерполяции центральную роль сыграло то обстоятельство, что соответствующие функции Лагранжа оказались просто произведением функций одной переменной, а это очевидное следствие того факта, что прямоугольник — прямое произведение отрезков.

Поскольку треугольник — двумерный аналог отрезка (в том смысле, что отрезок — выпуклая оболочка двух несовпадающих точек, а треугольник — выпуклая оболочка трех точек, не лежащих на одной прямой), то есть смысл попробовать перенести интерполяционную конструкцию Лагранжа на двумерный случай.

Отметим, что многочлен двух переменных, степень которого по совокупности переменных не превосходит n, задается соотношением

$$M_n(x, y) = \sum_{i+j=0}^{n} a_{i,j} x^i y^j$$

и однозначно определяется набором своих коэффициентов $a_{i,j}$, которых всего $\frac{1}{2}(n+1)(n+2)$. Есть основания ожидать, что задавая соответствующее количество начальных интерполяционных данных — значений интерполируемой функции — в некоторых точках треугольника, мы сможем однозначно найти коэффициенты $a_{i,j}$ и, тем самым, решить задачу треугольной многочленной интерполяции. Как будет показано ниже, за счет специального выбора узлов интерполяции эта программа действительно может быть реализована.

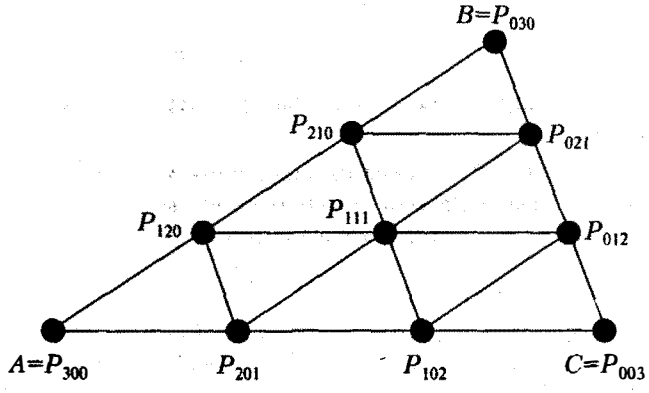


Рис. 14. Узлы треугольной интерполяции

Пусть D — треугольник, вершины A, B, C которого имеют координаты (x_a, y_a), (x_b, y_b), (x_c, y_c) соответственно. Разделив каждую из сторон треугольника на n *одинаковых* частей, проведем через точки деления прямые, параллельные сторонам (рис. 14). Точки пересечения этих прямых друг с другом и со сторонами треугольниками образуют вместе с вершинами регулярную сеть из $\frac{1}{2}(n+1)(n+2)$ узлов. Всякий узел $P_{\alpha\beta\gamma}$ этой сети однозначно представим в виде

$$P_{\alpha\beta\gamma} = \frac{\alpha}{n} A + \frac{\beta}{n} B + \frac{\gamma}{n} C,$$

где (α, β, γ) — целочисленные барицентрические координаты, т. е. такие, что

$$0 \leqslant \alpha \leqslant n, \quad 0 \leqslant \beta \leqslant n, \quad 0 \leqslant \gamma \leqslant n, \quad \alpha + \beta + \gamma = m.$$

При этом координаты узла $P_{\alpha\beta\gamma}$ даются соотношениями

$$x_{\alpha\beta\gamma} = \frac{\alpha}{n}x_a + \frac{\beta}{n}x_b + \frac{\gamma}{n}x_c, \quad y_{\alpha\beta\gamma} = \frac{\alpha}{n}y_a + \frac{\beta}{n}y_b + \frac{\gamma}{n}y_c.$$

Например, вершины треугольника в этой системе записываются в виде

$$A = P_{n00}, \quad B = P_{0n0}, \quad C = P_{00n}.$$

Взяв узлы $P_{\alpha\beta\gamma}$ в качестве узлов интерполяции, докажем,

что существует единственный многочлен, степень которого по совокупности переменных не превышает n и который принимает в этих узлах заданные значения $f_{\alpha\beta\gamma}$.

Доказательство проведем, построив базис Лагранжа, ассоциированный с рассматриваемой интерполяционной задачей.

Лемма (о представлении многочлена, обращающегося в нуль вдоль прямой). *Пусть $M(x, y)$ — многочлен двух переменных, степень которого не превышает n. Предположим, что на прямой $L = \{(x, y)\colon N_1x + N_2y + N_3 = 0\}$ и сам многочлен $M(x, y)$, и первые его r производных по нормали к прямой обращаются в нуль*

$$\left.\frac{d^sM(x, y)}{d\xi^s}\right|_{(x,y)\in L} = 0, \quad s = 1, 2, \ldots, r.$$

Тогда многочлен $M(x, y)$ представим в виде

$$M(x, y) = (N_1x + N_2y + N_3)^{r+1}Q(x, y),$$

где $Q(x, y)$ — многочлен, степень которого не превосходит $n - r + 1$.

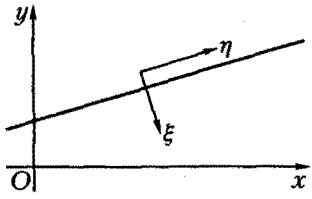


Рис. 15. Координатная система (ξ, η)

◀ Введем на плоскости xOy новую координатную систему, выбрав в качестве одной из осей направление вдоль прямой L, а в качестве второй — направление нормали к L (рис. 15). Новые координаты (ξ, η) связаны со старыми (x, y) соотношениями

$$\xi = N_1x + N_2y + N_3, \quad \eta = -N_2x + N_1y.$$

В новых координатах многочлен $M(x, y)$ превратится в многочлен $T(\xi, \eta)$, степень которого не превышает n и для которого условие леммы примет вид

$$\left.\frac{d^sT(\xi, \eta)}{d\xi^s}\right|_{\xi=0} = 0, \quad s = 1, 2, \ldots, r.$$

Раскладывая $T(\xi, \eta)$ по формуле Тейлора в окрестности точки $\xi = 0$, получаем

$$T(\xi, \eta) = T(0, \eta) + T'(0, \eta)\xi + \frac{T''(0, \eta)}{2!}\xi^2 + \ldots + \frac{T^{(n)}(0, \eta)}{n!}\xi^n.$$

Из условия равенства нулю при $\xi = 0$ многочлена $T(\xi, \eta)$ и первых его r производных вытекает, что

$$T(\xi, \eta) = \xi^{r+1}\left(\frac{T^{(r+1)}(0, \eta)}{(r+1)!} + \ldots + \frac{T^{(n)}(0, \eta)}{2}\xi^{n-r+1}\right).$$

Возвращаясь в последнем соотношении к старым координатам (x, y), получаем утверждение леммы. ▶

Чтобы не загромождать изложение, построение базиса Лагранжа, ассоциированного с рассматриваемой интерполяционной задачей, проведем для случая $n = 3$. Для других значений n выкладки аналогичны.

Пусть $\Lambda_{\alpha\beta\gamma}(x, y)$ — элементупомянутого базиса, отвечающий узлу $P_{\alpha\beta\gamma}$. Функция $\Lambda_{\alpha\beta\gamma}(x, y)$ является многочленом степени не выше третьей по совокупности переменных и в узлах сетки удовлетворяет условиям

$$\Lambda_{\alpha\beta\gamma}(P_{ijk}) = \begin{cases} 1, & \alpha = i \cap \beta = j \cap \gamma = k, \\ 0, & \alpha \neq i \cup \beta \neq j \cup \gamma \neq k. \end{cases}$$

Если стороны треугольника ABC заданы уравнениями

$$AB\colon\ N_1^{AB}x + N_2^{AB}y + N_3^{AB} = 0,$$

$$BC\colon\ N_1^{BC}x + N_2^{BC}y + N_3^{BC} = 0,$$

$$AC\colon\ N_1^{AC}x + N_2^{AC}y + N_3^{AC} = 0,$$

то параллельные им линии описываются уравнениями, отличающимися от приведенных только свободными членами. Например,

$$P_{201}P_{021}\colon\ N_1^{AB}x + N_2^{AB}y + \widetilde{N}_3^{P_{201}P_{021}} = 0,$$

$$P_{102}P_{012}\colon\ N_1^{AB}x + N_2^{AB}y + \widetilde{N}_3^{P_{102}P_{012}} = 0.$$

Функция $\Lambda_{300}(x, y)$ отвечает узлу $P_{300} = A$. Следовательно, она тождественно обращается в нуль вдоль стороны BC, и по лемме заключаем, что

$$\Lambda_{300}(x, y) = \left(N_1^{BC}x + N_2^{BC}y + N_3^{BC}\right)Q_2(x, y),$$

где $Q_2(x, y)$ — некоторый многочлен, степень которого не выше 2. Вдоль стороны $P_{102}P_{120}$ функция $\Lambda_{300}(x, y)$ представима в виде

$$\Lambda_{300}(x, y)\big|_{P_{102}P_{120}} = CQ_2(x, y)\big|_{P_{102}P_{120}}$$

и обращается в нуль в узлах P_{102}, P_{120}, P_{120}. Поэтому

$$Q_2(x, y)\big|_{P_{102}P_{120}} \equiv 0,$$

и по лемме заключаем, что

$$Q_2(x, y) = \left(N_1^{BC}x + N_2^{BC}y + N_3^{P_{102}P_{120}}\right)Q_1(x, y).$$

Рассматривая $\Lambda_{300}(x, y)$ вдоль стороны $P_{201}P_{210}$, аналогично вышеизложенному заключаем, что

$$Q_1(x, y) = \left(N_1^{BC}x + N_2^{BC}y + N_3^{P_{201}P_{210}}\right) \cdot \text{const}.$$

Определяя постоянную из условия $\Lambda_{300}(A) = 1$, окончательно получаем

$$\Lambda_{300}(x, y) = \frac{l^{BC}(x, y)l^{P_{102}P_{120}}(x, y)l^{P_{201}P_{210}}(x, y)}{l^{BC}(x_a, y_a)l^{P_{102}P_{120}}(x_a, y_a)l^{P_{201}P_{210}}(x_a, y_a)},$$

где

$$l^{BC}(x, y) = N_1^{BC}x + N_2^{BC}y + N_3^{BC},$$

$$l^{P_{102}P_{120}}(x, y) = N_1^{BC}x + N_2^{BC}y + N_3^{P_{102}P_{120}},$$

$$l^{P_{201}P_{210}}(x, y) = N_1^{BC}x + N_2^{BC}y + N_3^{P_{201}P_{210}}.$$

Аналогичные рассуждения позволяют построить и остальные функции $\Lambda_{\alpha\beta\gamma}(x, y)$. Так, например, функция $\Lambda_{210}(x, y)$ дается соотношением

$$\Lambda_{210}(x, y) = \frac{l^{BC}(x, y) l^{P_{102}P_{120}}(x, y) l^{AC}(x, y)}{l^{BC}(x_{210}, y_{210}) l^{P_{102}P_{120}}(x_{210}, y_{210}) l^{AC}(x_{210}, y_{210})},$$

а функция $\Lambda_{111}(x, y)$ соотношением

$$\Lambda_{111}(x, y) = \frac{l^{AB}(x, y) l^{BC}(x, y) l^{AC}(x, y)}{l^{AB}(x_{111}, y_{111}) l^{BC}(x_{111}, y_{111}) l^{AC}(x_{111}, y_{111})}.$$

Теперь легко убедиться в том, что для произвольных интерполяционных данных $f_{\alpha\beta\gamma}$, заданных в узлах $P_{\alpha\beta\gamma}$, многочлен

$$M(x, y) = \sum_{\alpha+\beta+\gamma=3} f_{\alpha\beta\gamma} \Lambda_{\alpha\beta\gamma}(x, y)$$

решает поставленную задачу.

5.4. Треугольная интерполяция. Частные случаи

В приложениях редко прибегают к треугольной интерполяции, порядок которой выше трех. Причем для случая $n = 3$ рассмотренная выше интерполяционная схема подвергается модификации с целью исключения внутреннего узла P_{111}. В этом пункте мы построим явные выражения для функций $\Lambda_{\alpha\beta\gamma}(x, y)$ при $n = 1, 2, 3$.

Все построения проведем для стандартного треугольника

$$D_{uv} = \{(u, v):\ u \geqslant 0,\ v \geqslant 0,\ u + v \leqslant 1\},$$

заметив, что если $\widetilde{\Lambda}_{\alpha\beta\gamma}(u, v)$ — базис Лагранжа, ассоциированный с интерполяционной задачей на треугольнике D_{uv}, а $W(x, y)$ — аддитивное и, вообще говоря, неоднородное преобразование, переводящее треугольник D_{xy} в треугольник D_{uv}

$$W(x, y):\ \begin{cases} u(x, y) = u_1 x + u_2 y + u_0, \\ v(x, y) = v_1 x + v_2 y + v_0, \end{cases}$$

то функции

$$\Lambda_{\alpha\beta\gamma}(x, y) = \widetilde{\Lambda}_{\alpha\beta\gamma}(u(x, y), v(x, y))$$

образуют базис Лагранжа, ассоциированный с интерполяционной задачей на треугольнике D_{xy}.

Если треугольник D_{xy} задан своими вершинами $A(x_a, y_a)$, $B(x_b, y_b)$, $C(x_c, y_c)$ и переводится преобразованием $W(x, y)$ с сохранением ориентации в треугольник D_{uv} (рис. 16), то несложные выкладки приводят к соотношениям

$$u(x, y) = \frac{\begin{vmatrix} x - x_a & y - y_a \\ x_b - x_a & y_b - y_a \end{vmatrix}}{2S}, \quad v(x, y) = \frac{\begin{vmatrix} x - x_a & y - y_a \\ x_c - x_a & y_c - y_a \end{vmatrix}}{2S},$$

где S — площадь треугольника D_{xy}, так что

$$2S = \begin{vmatrix} x_b - x_a & y_b - y_a \\ x_c - x_a & y_c - y_a \end{vmatrix}.$$

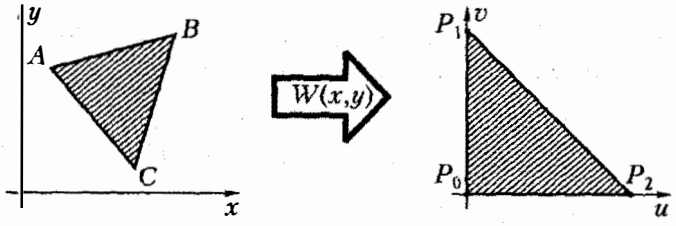


Рис. 16. Отображение D_{xy} в D_{uv}

Если же треугольник D_{xy} задан не координатами вершин, а уравнениями сторон

$$AB:\ N_1^{AB}x + N_2^{AB}y + N_3^{AB} = 0,$$

$$BC:\ N_1^{BC}x + N_2^{BC}y + N_3^{BC} = 0,$$

$$AC:\ N_1^{AC}x + N_2^{AC}y + N_3^{AC} = 0,$$

то функции $u(x, y)$ и $v(x, y)$ представимы в виде

$$u(x, y) = C_1\left(N_1^{AB}x + N_2^{AB}y + N_3^{AB}\right), \quad v(x, y) = C_2\left(N_1^{BC}x + N_2^{BC}y + N_3^{BC}\right),$$

где постоянные C_i легко могут быть найдены из условий $u(x_c, y_c) = 1$ и $v(x_b, y_b) = 1$ соответственно.

Перейдем непосредственно к построению. Для $n = 1$ соображения предыдущего пункта дают

$$\widetilde{\Lambda}^1_{00}(u, v) = 1 - u - v, \quad \widetilde{\Lambda}^1_{010}(u, v) = v, \quad \widetilde{\Lambda}^1_{001}(u, v) = u.$$

Переход на плоскость x●y очевиден.

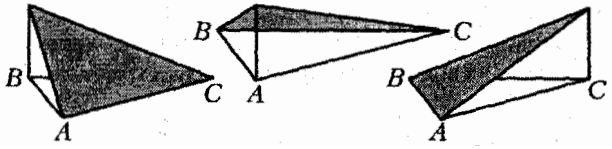


Рис. 17. Базис Лагранжа, ассоциированный с треугольной интерполяцией

Функции Λ^1_{ijk} показаны на рис. 17.

В случае $n = 2$ получим

$$\widetilde{\Lambda}^2_{200}(u, v) = 2(u+v-1)(u+v-0{,}5), \quad \widetilde{\Lambda}^2_{020}(u, v) = 2v(v-0{,}5), \quad \widetilde{\Lambda}^2_{002}(u, v) = 2u(u-0{,}5),$$

$$\widetilde{\Lambda}^2_{110}(u, v) = 4(1-u-v)v, \quad \widetilde{\Lambda}^2_{011}(u, v) = 4uv, \quad \widetilde{\Lambda}^2_{101}(u, v) = 4u(1-u-v).$$

Заметим, что функции $\widetilde{\Lambda}^2_{\alpha\beta\gamma}(u, v)$ легко могут быть выражены через $\widetilde{\Lambda}^1_{\alpha\beta\gamma}(u, v)$. Так, например,

$$\widetilde{\Lambda}^2_{200} = \widetilde{\Lambda}^1_{100}(2\widetilde{\Lambda}^2_{100} + 1), \quad \widetilde{\Lambda}^2_{110} = 4\widetilde{\Lambda}^1_{100}\Lambda^1_{010}.$$

Для случая $n = 3$ (нумерация узлов приведена на рис. 18) все выкладки аналогичны рассмотренным выше и дают, например, следующее

$$\widetilde{\Lambda}^3_{300}(u, v) = 0{,}5\widetilde{\Lambda}^1_{100}(3\widetilde{\Lambda}^1_{100}+1)(3\widetilde{\Lambda}^1_{100}+2), \quad \widetilde{\Lambda}^3_{210}(u, v) = 1{,}5\widetilde{\Lambda}^1_{010}(3\widetilde{\Lambda}^1_{100}+1)(3\widetilde{\Lambda}^1_{100}+2),$$

$$\widetilde{\Lambda}^3_{102}(u, v) = 1{,}5\widetilde{\Lambda}^1_{001}(3\widetilde{\Lambda}^1_{001}-1)3\widetilde{\Lambda}^1_{100}, \quad \widetilde{\Lambda}^3_{111}(u, v) = 9\widetilde{\Lambda}^1_{001}\widetilde{\Lambda}^1_{010}\widetilde{\Lambda}^1_{100}.$$

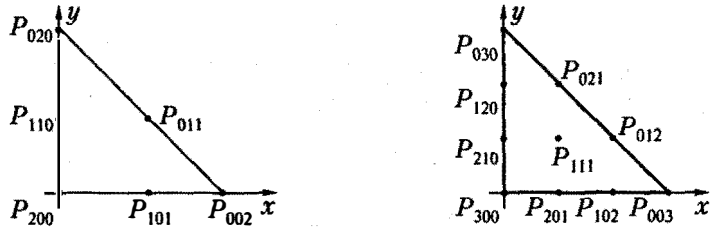


Рис. 18. Узлы интерполяции для $n = 2$ и $n = 3$

5.5. Треугольная интерполяция — исключение среднего узла в десятиузловой схеме

Как и для прямоугольной интерполяции, в некоторых ситуациях полезно иметь треугольную интерполяционную схему, отвечающую $n = 3$, с исключенным внутренним узлом. Пусть $\Lambda_{\alpha\beta\gamma}(x, y)$ — базис Лагранжа, ассоциированный с десятиточечной интерполяционной задачей на треугольнике. Всякий многочлен степени не выше трех, определенный на этом треугольнике, однозначно представим в виде

$$Q(x, y) = \sum_{\alpha+\beta+\gamma=3} Q(P_{\alpha\beta\gamma})\Lambda_{\alpha\beta\gamma}(x, y).$$

Построим базис Лагранжа, ассоциированный с девятиточечной интерполяционной задачей, полученной из десятиточечной исключением внутреннего узла P_{111}.

Если $T_{ijk}(x, y)$, $i + j + k = 3$, $i, j, k \geqslant 0$, $ijk = 0$, — одна из искомых базисных функций, то, поскольку она является многочленом степени не выше трех, имеет место представление

$$T_{ijk}(x, y) = \sum_{\alpha+\beta+\gamma=3} T_{ijk}(P_{\alpha\beta\gamma})\Lambda_{\alpha\beta\gamma}(x, y).$$

А так как функции $T_{ijk}(x, y)$, $i + j + k = 3$, $i, j, k \geqslant 0$, $ijk = 0$, образуют базис Лагранжа, то

$$T_{ijk}(P_{\alpha\beta\gamma}) = \begin{cases} 1, & \alpha = i \cap \beta = j \cap \gamma = k, \\ 0, & \alpha \neq i \cup \beta \neq j \cup \gamma \neq k. \end{cases}$$

Учитывая последнее[1] обстоятельство, перепишем предыдущее соотношение в виде

$$T_{ijk}(x, y) = T_{ijk}(P_{111})\Lambda_{111}(x, y) + \Lambda_{ijk}.$$

Отсюда заключаем, что искомые функции однозначно определяются заданием чисел $t_{ijk} = T_{ijk}(P_{111})$, являющихся их значениями во внутреннем узле. Значение произвольной функции $f(x, y)$, определенной на треугольнике, интерполируется при этом по формуле

$$f(P_{111}) = \sum_{i \cdot j \cdot k = 0} f(P_{ijk})t_{ijk}.$$

В заключение отметим, что некоторые дополнительные соображения, связанные с оптимальностью интерполяции во внутреннем узле, приводят к следующим классам интерполирующих функций — многочлены, степень которых не превышает трех и которые не содержат либо произведения xy, либо произведения x^2y, либо произведения xy^2.

[1] Отметим, что это соотношение не регламентирует значений функций $T_{ijk}(x, y)$ во внутреннем узле P_{111}.

5.6. Заключительные замечания

Рассмотренные выше соображения, связанные с интерполяцией функций двух переменных, допускают обобщение как на области произвольной конфигурации, так и на интерполяционные процедуры, использующие не только значения функции в отмеченных узлах, но и значения некоторых ее производных. Остановимся коротко на идейной стороне этих обобщений.

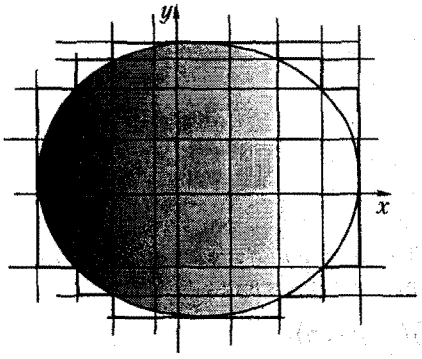


Рис. 19. Замена области D областью $\overline{D_\varepsilon}$

Пусть $D \subset \mathbb{R}^2$ — квадрируемая область на плоскости. Прямыми, параллельными осям координат, разобьем эту область на прямоугольники (рис. 19) и заменим область D на область $\overline{D_\varepsilon}$, представляющую собой приближение к D и являющуюся объединением прямоугольников, со сторонами, параллельными координатным осям. При этом $D \subset \overline{D_\varepsilon}$ и для произвольного положительного ε площади этих областей отличаются не более чем на величину ε.

Пусть далее $f(x, y)$ — функция двух переменных, определенная в D. Проинтерполировав ее на каждом из прямоугольников области $\overline{D_\varepsilon}$ так, как это было описано выше, возьмем в качестве интерполяционной структуры на всей области D *сужение склейки интерполяционных многочленов* в $\overline{D_\varepsilon}$ на D. Если по каким-то соображениям важными представляются значения интерполируемой функции на границе ∂D области D, то разбиение следует производить так, чтобы вершины граничных прямоугольников попали на границу ∂D.

Аналогичным образом триангуляция (рис. 20) области $D \subset \mathbb{R}^2$ позволяет строить интерполяционные структуры на базе треугольников как *склейки интерполяционных многочленов*.

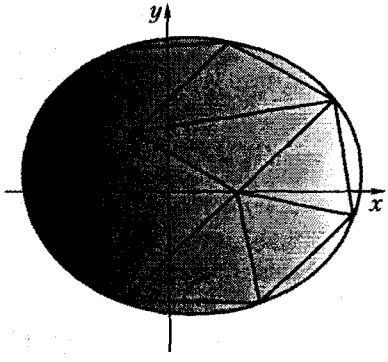


Рис. 20. Одна из возможных триангуляций области D

В обоих случаях узлами интерполяции служат узлы разбиения области D.

Использование значений производных интерполируемой функции наряду со значениями самой функции тоже может быть осуществлено двумя способами — либо прямым использованием одномерных интерполяционных структур Эрмита по каждой из координат (в случае прямоугольников), либо построением двумерного аналога эрмитовой интерполяции (для треугольников) аналогично тому, как это было сделано выше для лагранжевой интерполяции.

Глава LX

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Многие прикладные задачи (механика, физика, химия, радиофизика и т. д.) приводят к необходимости вычисления интеграла

$$I = \int \dots \int_D \dots \int f(x_1, \dots, x_n)\, dx_1 \dots dx_n, \tag{1}$$

где предполагается, что функция $f(x) = f(x_1, \dots, x_n)$ определена в области $D \subset \mathbb{R}^n$ и обладает там достаточно хорошими свойствами, обеспечивающими его существование, может быть, и в несобственном смысле.

В простейших случаях интеграл удается вычислить при помощи элементарных преобразований и использования формулы Ньютона—Лейбница. Однако элементарные первообразные существуют не для любых подынтегральных функций, кроме того, $f(x)$ зачастую задается таблично, алгоритмом или каким-либо другим способом, исключающим использование первообразных. Но даже если элементарная первообразная и существует, вполне может оказаться что ее вычисление требует больших затрат вычислительных ресурсов или является нецелесообразным по каким-либо иным соображениям.

Изучение способов и методов прямого (не использующего аппарат первообразных и формулу Ньютона—Лейбница) вычисления интеграла (1) является достаточно интересной с теоретической и важной с практической точки зрения задачей. Мы начнем ее рассмотрение с исследования процедур вычисления однократных интегралов от функций одной переменной.

§ 1. Квадратурные формулы

Пусть функция $f(x)$ определена и кусочно-непрерывна на конечном отрезке $[a, b]$ действительной прямой. При этих условиях интеграл

$$I = \int_a^b f(x)\, dx \tag{1}$$

существует в собственном смысле.

Пусть на отрезке $[a, b]$ заданы r точек $a \leqslant x_1 < x_2 < \dots < x_r \leqslant b$ и пусть, кроме того, задан упорядоченный набор r чисел $\omega_1, \omega_2, \dots, \omega_r$.

Квадратурной формулой для вычисления интеграла (1) назовем выражение

$$I_r = \sum_{i=1}^{r} \omega_i f(x_i)$$

такое, что

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{r} \omega_i f(x_i). \tag{2}$$

Величина

$$R_r(f) = \int_a^b f(x)\,dx - \sum_{i=1}^{r} \omega_i f(x_i)$$

называется *остаточным членом* квадратурной формулы, так что

$$I = \int_a^b f(x)\,dx = \sum_{i=1}^{r} \omega_i f(x_i) + R_r(f).$$

Точки x_i называются *узлами*, а числа ω_i — *весами* квадратурной формулы.

Как узлы, так и веса квадратурной формулы являются величинами, характерными для данной квадратурной формулы, и не зависят от подынтегральной функции.

Если для некоторого класса функций оказывается, что $R_r(f) \equiv 0$, то квадратурная формула (2) называется *точной на этом классе функций*. Если концы отрезка включены в число узлов квадратурной формулы, то формула называется *замкнутой*, в противном случае — *открытой*.

Квадратурная формула (2) называется *простой*. Пусть отрезок $[a, b]$ разбит на N подотрезков

$$[a, b] = \bigcup_{j=1}^{N} \Delta_j, \quad \Delta_j = [\alpha_j, \alpha_{j+1}], \quad j = 1, \ldots, N-1, \quad \alpha_1 = a, \ \alpha_N = b,$$

и пусть на каждом подотрезке Δ_j для вычисления интеграла (1) используется квадратура (2)

$$I_j = \int_{\alpha_j}^{\alpha_{j+1}} f(x)\,dx \approx \sum_{i=1}^{r} \omega_i^j f(x_i^j),$$

так что

$$I = \int_a^b f(x)\,dx \approx \sum_{j=1}^{N} \sum_{i=1}^{r} \omega_i^j f(x_i^j).$$

В этом случае квадратура носит название *составной*. Конечно, возможно использование на различных подотрезках различных простых квадратур.

Наиболее распространенный способ построения квадратурных формул для интеграла (1) состоит в замене подынтегральной функции $f(x)$ «похожей» на нее функцией $\varphi(x)$, по каким-то соображениям более привлекательной в вычислительном отношении, и такой, что

$$I = \int_a^b f(x)\,dx \approx \int_a^b \varphi(x)\,dx.$$

Так может быть получено подавляющее большинство используемых в настоящее время квадратур.

С учетом этих предварительных замечаний перейдем к построению конкретных формул.

§ 2. Квадратуры Ньютона—Котеса

Широкий класс простых и часто используемых квадратурных формул может быть получен заменой подынтегральной функции ее интерполяционным многочленом.

Зададим на отрезке $[a, b]$ произвольным образом r различных точек $a \leqslant x_1 < x_2 < \ldots < x_r \leqslant b$ и построим интерполяционный многочлен Лагранжа с узлами интерполяции в этих точках

$$L_{r-1}(x) = \sum_{i=1}^{r} f(x_i)\Lambda_i(x)$$

(здесь $\Lambda_i(x)$ — базисные интерполяционные многочлены, задаваемые соотношениями

$$\Lambda_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j},$$

зависящие от выбора узлов интерполяции на отрезке $[a, b]$ и не зависящие от интерполируемой функции).

Подставляя в интеграл вместо функции $f(x)$ интерполяционный многочлен $L_{r-1}(x)$, получим

$$\int_a^b f(x)\,dx \approx \int_a^b L_{r-1}(x)\,dx = \int_a^b \sum_{i=1}^{r} f(x_i)\Lambda_i(x) = \sum_{i=1}^{r} \omega_i f(x_i). \qquad (1)$$

Здесь через ω_i обозначены величины, даваемые соотношениями

$$\omega_i = \int_a^b \Lambda_i(x)\,dx, \quad i = 1, 2, \ldots, r.$$

В силу неравенства (5) из § 1 гл. LIX

$$|L_{r-1}(x) - f(x)| \leqslant \frac{C}{r!}|\omega_{r-1}(x)|, \quad \omega_{r-1}(x) = (x - x_1)(x - x_2)\ldots(x - x_r)$$

заключаем, что для остаточного члена $R_r(f)$ квадратуры (1) имеет место оценка

$$|R_r(f)| \leqslant \frac{C}{r!} \int_a^b |\omega_{r-1}(x)|\, dx. \tag{2}$$

В случае равномерного разбиения отрезка $x_i = a + i\frac{b-a}{r-1}$ формулы (1) называются квадратурными формулами *Ньютона—Котеса.*

Использование указанных квадратур для вычисления интегралов приводит к вычислению значений подынтегральной функции в узлах квадратурной формулы. На первый взгляд может показаться, что чем выше степень многочлена, интерполирующего подынтегральную функцию, тем выше точность вычисления интеграла. Однако это не совсем так. Наличие вычислительных погрешностей и ошибок округления при вычислении значений $f(x_i)$ может привести к значительным ошибкам.

Действительно, пусть предельная абсолютная погрешность

$$\Delta_i = |\tilde{f}(x_i) - f(x_i)| \leqslant \delta.$$

Тогда в силу того, что

$$\sum_{i=1}^{r} \Lambda_i(x) \equiv 1 \quad \forall\, x \in [a, b],$$

заключаем

$$\Delta I = \delta \sum_{i=1}^{r} |\omega_i| = \delta \sum_{i=1}^{r} \int_a^b |\Lambda_i|\, dx \geqslant \delta \int_a^b \left| \sum_{i=1}^{r} \Lambda_i \right| dx = \delta(b-a)$$

и, следовательно, существует неустранимая погрешность вычисления интеграла, пропорциональная длине промежутка интегрирования и погрешности вычисления значений функции.

Кроме того, известно, что с ростом r среди весов квадратур (1) будут встречаться сколь угодно большие по абсолютной величине числа, что приводит к росту вычислительных ошибок и делает использование многоузловых квадратур совершенно неприемлемым в реальных расчетах.

Поэтому на практике используется следующий прием — промежуток интегрирования разбивается на подотрезки

$$[a, b] = \bigcup_{j=1}^{N} \Delta_j, \quad \Delta_j = [\alpha_j, \alpha_{j+1}], \quad j = 1, \ldots, N-1, \quad \alpha_1 = a,\ \alpha_N = b,$$

на каждом из которых применяют квадратуру (1) с относительно малым числом узлов. Такие составные квадратурные формулы оказываются весьма эффективным инструментом вычисления интегралов.

Рассмотрим некоторые частные случаи соотношений (1).

1. *Формула прямоугольников.*

Положим в (1) $r = 1$. Тогда интерполяционный многочлен $L_0(x) \equiv \text{const} = f(x_1)$, и квадратура принимает вид

$$I_1 = (b-a) f(x_1),$$

т. е. площадь криволинейной трапеции (рис. 1) заменяется площадью прямоугольника с основанием $b-a$ и высотой $f(x_1)$.

Выбор узла x_1 может быть осуществлен произвольным образом. Если в качестве узла x_1 берется середина отрезка $[a, b]$, то получается *одноузловая открытая* формула Ньютона—Котеса

$$I_1 = (b-a) f\left(\frac{a+b}{2}\right).$$

Составная формула прямоугольников, как правило, строится на равномерном разбиении отрезка $[a, b]$. Пусть $h = \frac{b-a}{N}$, $\alpha_j = a + jh$, $j = 0, 1, \ldots, N$. Тогда

$$\boxed{I \approx h\left[f\left(\frac{\alpha_0+\alpha_1}{2}\right) + f\left(\frac{\alpha_1+\alpha_2}{2}\right) + \ldots + f\left(\frac{\alpha_{N-1}+\alpha_N}{2}\right)\right].}$$

2. *Формула трапеций.*

При $r = 2$, $x_1 = a$, $x_2 = b$ получаем

$$\omega_1 = \int_a^b \frac{x-b}{a-b}\,dx = \frac{b-a}{2}, \quad \omega_2 = \int_a^b \frac{x-a}{b-a}\,dx = \frac{b-a}{2},$$

и искомая квадратура запишется в виде

$$I_2 = \frac{b-a}{2}[f(a) + f(b)].$$

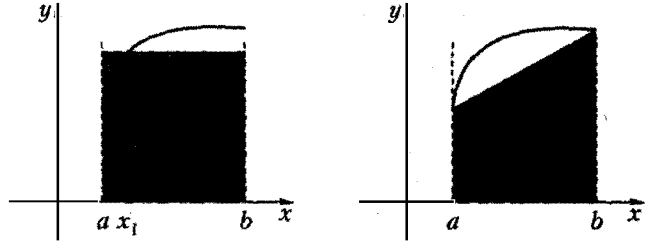


Рис. 1. Формулы прямоугольников и трапеций

Она называется *простой формулой трапеций* и геометрически соответствует (рис. 1) замене площади криволинейной трапеции площадью трапеции, образованной прямыми $x = a$, $x = b$ и хордой, соединяющей концы дуги графика функции $f(x)$ на промежутке $[a, b]$. *Составная* формула трапеций имеет вид

$$\boxed{I \approx \frac{h}{2}[f(\alpha_0) + 2f(\alpha_1) + \ldots + 2f(\alpha_{N-1}) + f(\alpha_N)]}$$

и получается применением простой формулы на каждом из подотрезков $[\alpha_j, \alpha_{j+1}]$.

3. *Формула парабол (формула Симпсона).*

Полагая $r = 3$, $x_1 = a$, $x_2 = \frac{a+b}{2}$, $x_3 = b$, получаем, как и выше, что

$$\omega_1 = \frac{b-a}{6}, \quad \omega_2 = \frac{2}{3}(b-a), \quad \omega_3 = \frac{b-a}{6},$$

и искомая квадратура принимает вид

$$I_3 = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Поскольку $L_2(x)$ — многочлен второй степени, то геометрически содержание квадратуры Симпсона состоит в замене площади криволинейной трапеции площадью фигуры, ограниченной прямыми $x = a$, $x = b$ и параболой, проходящей через концы дуги графика функции $f(x)$ (рис. 2).

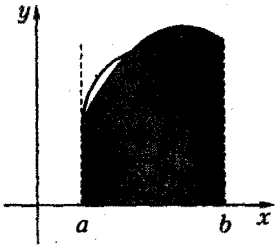


Рис. 2. Формула Симпсона

Для получения составной формулы Симпсона разобьем отрезок $[a, b]$ на $2N$ одинаковых частей узлами α_j, $j = 0, 1, \ldots, 2N$, и применим простую квадратуру парабол на каждом из промежутков $[\alpha_{2j-2}, \alpha_{2j}]$. Обозначая через h длину промежутка разбиения, получим

$$\boxed{I \approx \frac{h}{3}\left[f(\alpha_0) + 4f(\alpha_1) + 2f(\alpha_2) + \ldots + 2f(\alpha_{2N-2}) + 4f(\alpha_{2N-1}) + f(\alpha_N)\right].}$$

§ 3. Точность простейших квадратур Ньютона—Котеса

Оценивание точности полученных выше квадратурных формул на заданных классах подынтегральных функций может быть осуществлено использованием соотношения (2) § 2.

Рассмотрим простую квадратуру (1) § 2, построенную по r равноотстоящим узлам x_i, $i = 1, 2, \ldots, r$. В этом случае $x_1 = a$, $x_i = x_1 + ih$, $x_r = b$, $h = \frac{b-a}{r-1}$ и функция $\omega_{r-1}(x)$ имеет вид

$$\omega_{r-1}(x) = \prod_{i=1}^{r}\left[x - x_1 - (i-1)h\right] = \prod_{i=1}^{r}\left[h(r-1)\right]^r\left[\frac{x-x_1}{h(r-1)} - \frac{i-1}{r-1}\right].$$

Вводя новую нормированную переменную $t = \frac{x-x_1}{h(r-1)}$, $0 \leqslant t \leqslant 1$, и учитывая, что $h(r-1) = b-a$, получаем

$$\omega_{r-1}(x) = (b-a)^r \prod_{i=1}^{r}\left[t - \frac{i-1}{r-1}\right].$$

Теперь оценка (2) § 2 может быть записана в форме

$$|R_r(f)| \leqslant \frac{C(b-a)^r}{r!}\int_0^1 \left|t\left(t - \frac{1}{r-1}\right)\ldots(t-1)\right| dx.$$

Предполагается, что подынтегральная функция r раз непрерывно дифференцируема на промежутке интегрирования и постоянная C такова, что $|f^{(r)}(x)| \leqslant C$ $\forall\, x \in [a, b]$.

Оценки остаточных членов полученных выше простых квадратурных формул примут вид:

— *для формулы прямоугольников на классе гладких (непрерывно дифференцируемых) функций*

$$|R_1(f)| \leqslant \max_{[a,b]} |f'(x)| \frac{(b-a)^2}{4},$$

— *для формулы трапеций на классе дважды непрерывно дифференцируемых функций*

$$|R_2(f)| \leqslant \max_{[a,b]} |f''(x)| \frac{(b-a)^3}{12},$$

— *для формулы парабол на классе трижды непрерывно дифференцируемых функций*

$$|R_3(f)| \leqslant \max_{[a,b]} |f'''(x)| \frac{(b-a)^4}{192}.$$

Заметим, что если класс подынтегральных функций изменить, то изменятся и оценки точности квадратурных формул. Например, формула прямоугольников на классе *дважды непрерывно дифференцируемых функций* допускает следующую оценку точности

$$|R_1(f)| \leqslant \max_{[a,b]} |f''(x)| \frac{(b-a)^3}{24}.$$

◀ В силу наличия второй производной для функции $f(x)$ получаем

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

поэтому

$$\int_a^b f(x)\,dx = f\left(\frac{a+b}{2}\right)(b-a) + R_1(f).$$

Здесь $R_1(f)$ дается выражением

$$R_1(f) = f'\left(\frac{a+b}{2}\right)\int_a^b \left(x - \frac{a+b}{2}\right) dx + \frac{f''(\xi)}{2!}\int_a^b \left(x - \frac{a+b}{2}\right)^2 dx,$$

где первое слагаемое равно нулю, а второе дает указанную оценку. ▶

Аналогичным образом можно установить, что формула Симпсона на классе функций, *обладающих четырьмя непрерывными производными*, также допускает отличную от приведенной выше оценку точности

$$|R_3(f)| \leqslant \max_{[a,b]} |f^{(4)}(x)| \frac{(b-a)^5}{2\,880}.$$

Полученные соотношения показывают, что квадратура тем точнее, чем меньше длина промежутка интегрирования. Поэтому для повышения точности квадратурных формул на практике используются составные квадратуры.

Легко установить, что составная формула прямоугольников на классе гладких функций допускает оценку

$$\left|R_1^{\text{сост}}(f)\right| \leqslant \max_{[a,b]} \left|f'(x)\right| \frac{(b-a)^2}{4N},$$

а на классе дважды непрерывно дифференцируемых функций оценку

$$\left|R_1^{\text{сост}}(f)\right| \leqslant \max_{[a,b]} \left|f''(x)\right| \frac{(b-a)^3}{24N^2}.$$

Аналогичные соображения для составной формулы трапеций на классе гладких функций позволяют записать оценку точности в виде

$$\left|R_2^{\text{сост}}(f)\right| \leqslant \max_{[a,b]} \left|f''(x)\right| \frac{(b-a)^3}{12N^2}.$$

Так же можно получить оценки точности и для других составных квадратур.

Приведенные соотношения формально утверждают, что точность квадратуры может быть неограниченно увеличена за счет увеличения величины N — числа элементов разбиения промежутка интегрирования. Однако в реальном вычислительном процессе повышение точности квадратуры с увеличением числа N подотрезков интегрирования происходит до определенного предела, за которым это увеличение приводит к ухудшению точности используемой составной квадратуры — разбиение промежутка интегрирования на малые подпромежутки позволяет повысить точность квадратурной формулы на каждом подотрезке, но за счет увеличения их количества растет и суммарная вычислительная ошибка.

◀ Проследим качественно это явление на примере составной формулы прямоугольников

$$I \approx I_N = h\left[f(\alpha_1) + f(\alpha_2) + \ldots + f(\alpha_N)\right],$$

где буквой h обозначен шаг разбиения промежутка интегрирования $h = \frac{b-a}{N}$. Предельная абсолютная погрешность величины I_N дается соотношением

$$\Delta_{I_N} = \Delta_h \sum_{i=1}^{N} f(\alpha_i) + h \sum_{i=1}^{N} \Delta_{f(\alpha_i)}.$$

Замечая, что $\sum_{i=1}^{N} f(\alpha_i) = \frac{I_N}{h}$ и обозначая через Δ наибольшую из погрешностей вычисления значений функции $f(x)$ в точках α_i, получаем

$$\boxed{\Delta_{I_N} \sim I_N \frac{\Delta_h}{h} + \Delta(b-a).}$$

Из последнего видно, что уменьшение шага h влечет увеличение погрешности вычисления интеграла. ▶

В заключение отметим, что r-узловая квадратура Ньютона—Котеса является *точной* на классе многочленов, степень которых не превышает $r-1$. Это очевидно следует из того обстоятельства, что всякий многочлен степени $r-1$ и ниже тождествен своему r-узловому интерполяционному многочлену Лагранжа. Легко показать, что это

свойство является определяющим для указанных квадратур — веса соответствующих квадратурных формул могут быть получены методом неопределенных коэффициентов при заданных узлах.

◀ Действительно, пусть узлы x_i, $i = 1, 2, \ldots, r$, искомой квадратуры

$$I_r(f) = \sum_{i=1}^{r} \omega_i f(x_i)$$

заданы[1]. Тогда из условия точности квадратуры на классе многочленов, степень которых не превосходит $r - 1$, для весов получаем систему линейных уравнений

$$\begin{cases} \int_a^b 1\, dx = \sum_{i=1}^{r} \omega_i \cdot 1, \\ \int_a^b x\, dx = \sum_{i=1}^{r} \omega_i \cdot x_i, \\ \int_a^b x^2\, dx = \sum_{i=1}^{r} \omega_i \cdot x_i^2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \int_a^b x^{r-1}\, dx = \sum_{i=1}^{r} \omega_i \cdot x_i^{r-1}, \end{cases}$$

которая, таким образом, всегда однозначно разрешима, так как ее определитель — определитель Вандермонда — не равен нулю при различных узлах x_i. ▶

§ 4. Квадратуры Гаусса

Рассмотрения § 3 естественно приводят к мысли попытаться улучшить качество получаемых квадратур за счет специального выбора узлов при том же их количестве.

Возможны разные понимания того, что такое «более качественная квадратура», мы же в основу последующих рассуждений положим следующее — ***квадратура с фиксированным числом узлов тем качественнее, чем выше степень многочленов, для которых она точна.*** Такое понимание качества квадратуры связано с тем, что многочлен более высокой степени, вообще говоря, лучше приближает функцию, а поэтому можно ожидать, что квадратуры, точные на классе многочленов более высокой степени, будут иметь меньшую погрешность

Элементарный анализ показывает, что построить r-узловую квадратуру, точную на многочленах степени выше или равной $2r$, невозможно.

---

[1] Для квадратур Ньютона—Котеса

$$x_1 = a, \quad x_i = x_1 + (i-1)h, \quad h = \frac{b-a}{r-1}.$$

◀ Действительно, если такая квадратура существует, то для функции

$$P_{2r}(x) = (x - x_1)^2 (x - x_2)^2 \dots (x - x_r)^2,$$

где x_i — узлы квадратуры, должно выполняться равенство

$$\int_a^b P_{2r}(x)\, dx = \sum_{i=1}^{r} \omega_i P_{2r}(x_i),$$

что невозможно, так как интеграл слева положителен в силу положительности подынтегральной функции, а сумма справа тождественно равна нулю при любом выборе узлов x_i и весов ω_i. ▶

Поскольку r-узловая квадратура определяется $2r$ параметрами — узлами и весами, — то разумно искать r-узловую квадратуру, точную на многочленах степени не выше $(2r - 1)$-й. Оказывается, такие квадратуры существуют. Они называются *квадратурами Гаусса* и могут быть построены для всех r.

Для упрощения выкладок рассмотрим процедуру построения упомянутых квадратур для интеграла

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

переход к которому от интеграла

$$\int_a^b f(x)\, dx$$

осуществляется посредством замены переменной интегрирования

$$x = \frac{b-a}{2} t + \frac{b+a}{2}.$$

При этом

$$\int_a^b f(x)\, dx = \frac{b-a}{2} \int_{-1}^{1} f\left(\frac{b-a}{2} t + \frac{b+a}{2}\right) dt.$$

Пусть искомая r-узловая квадратура имеет вид

$$G_r(f) = \sum_{i=1}^{r} \omega_i f(x_i).$$

Поскольку мы хотим, чтобы она была точной на классе многочленов степени не выше $2r - 1$, то для каждого из значений $s = 0, 1, \dots, 2r - 1$, должно иметь место соотношение

$$G_r(x^s) = \omega_1 \cdot x_1^s + \omega_2 \cdot x_2^s + \dots + \omega_r \cdot x_r^s =$$

$$= \int_{-1}^{1} x^s\, dx = \begin{cases} 0, & \text{если } s \text{ — нечетное,} \\ \dfrac{2}{s+1}, & \text{в противном случае.} \end{cases} \qquad (1)$$

Совокупность соотношений (1) образует систему $2r$ уравнений относительно неизвестных весов ω_i, $i = 1, 2, \ldots, r$, и узлов x_i, $i = 1, 2, \ldots, r$.

Система (1) может быть эффективно решена способом *скользящего суммирования*: пусть

$$W_r(x) = (x - x_1)(x - x_2) \ldots (x - x_r) = x^r + p_{r-1}x^{r-1} + \ldots + p_1 x + p_0$$

— многочлен степени r, корни которого совпадают с узлами искомой квадратуры. Умножим первое уравнение системы (1) на p_0, второе — на p_1, третье — на p_2 и так вплоть до r-го уравнения системы, которое умножим на 1. Складывая получившиеся соотношения почленно, получим

$$\omega_1 W_r(x_1) + \omega_2 W_r(x_2) + \ldots + \omega_r W_r(x_r) = 2p_0 + \frac{2}{3}p_2 + \ldots + \frac{2}{s+1}p_s + \ldots .$$

Поскольку по определению $W_r(x_i) = 0 \quad \forall\, i = 1, 2, \ldots, r$, последнее соотношение приводит к равенству, связывающему коэффициенты многочлена $W_r(x)$

$$2p_0 + \frac{2}{3}p_2 + \ldots + \frac{2}{s+1}p_s + \ldots = 0.$$

Продолжая процесс, умножим теперь второе уравнение системы (1) на p_0, третье — на p_1, четвертое — на p_2 и так вплоть до $(r+1)$-го уравнения системы, которое умножим на 1. Сложив эти соотношения почленно, получим второе уравнение для коэффициентов p_s

$$\omega_1 x_1 W_r(x_1) + \omega_2 x_2 W_r(x_2) + \ldots + \omega_r x_r W_r(x_r) = \frac{2}{3}p_1 + \ldots + \frac{2}{s+1}p_{s-1} + \ldots = 0.$$

Последовательно сдвигая суммирование на один шаг вперед, мы таким образом получим систему из r линейных уравнений относительно коэффициентов p_s характеристического многочлена $W_r(x)$.

Можно показать, что *для любого r эта система однозначно разрешима, причем все корни $W_r(x)$ различны и лежат на промежутке* $[-1, 1]$.

Взяв в качестве узлов искомой квадратуры корни многочлена $W_r(x)$, соответствующие им веса ω_i легко можно найти из системы (1). Они оказываются неотрицательными и обладающими симметрией относительно среднего узла. Последнее означает, что

$$\omega_i = \omega_{r-i+1} \quad \forall\, i.$$

Проиллюстрируем изложенное на простых примерах.

Одноузловая квадратура Гаусса

Одноузловая квадратура Гаусса (точная в классе линейных функций) определяется узлом $x_1 = 0$ и весом $\omega_1 = 2$

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

и совпадает с формулой прямоугольников из семейства квадратур Ньютона—Котеса.

Двухузловая квадратура Гаусса

При $r = 2$ получаем $W_2(x) = (x - x_1)(x - x_2) = x^2 + p_1 x + p_0$. Для этого случая система уравнений (1) запишется в виде

$$\begin{cases} \omega_1 + \omega_2 = 2, \\ \omega_1 x_1 + \omega_2 x_2 = 0, \\ \omega_1 x_1^2 + \omega_2 x_2^2 = \dfrac{2}{3}, \\ \omega_1 x_1^3 + \omega_2 x_2^3 = 0. \end{cases}$$

Скользящее суммирование приводит к системе для коэффициентов p_0 и p_1

$$\begin{cases} 2 \cdot p_0 + 0 \cdot p_1 + \dfrac{2}{3} \cdot 1 = 0, \\ 0 \cdot p_0 + \dfrac{2}{3} \cdot p_1 + 0 \cdot 1 = 0, \end{cases}$$

откуда получаем $p_0 = -\frac{1}{3}$, $p_1 = 0$ и

$$W_2(x) = x^2 - \frac{1}{3} = \left(x + \frac{1}{\sqrt{3}}\right)\left(x - \frac{1}{\sqrt{3}}\right).$$

Теперь легко определяем веса $\omega_1 = \omega_2 = 1$ и искомую двухузловую квадратуру Гаусса

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

Трехузловая квадратура Гаусса

Полагая $r = 3$ и $W_3(x) = (x - x_1)(x - x_2)(x - x_3) = x^3 + p_2 x^2 + p_1 x + p_0$, получаем

$$\left\{\begin{array}{l|lll} \omega_1 + \omega_2 + \omega_3 = 2, & p_0 & & \\ \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0, & p_1 & p_0 & \\ \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2 = \dfrac{2}{3}, & p_2 & p_1 & p_0 \\ \omega_1 x_1^3 + \omega_2 x_2^3 + \omega_3 x_3^3 = 0, & 1 & p_2 & p_1 \\ \omega_1 x_1^4 + \omega_2 x_2^4 + \omega_3 x_3^4 = \dfrac{2}{5}, & & 1 & p_2 \\ \omega_1 x_1^5 + \omega_2 x_2^5 + \omega_3 x_3^5 = 0. & & & 1 \end{array}\right.$$

Скользящее суммирование приводит к системе для коэффициентов p_0, p_1 и p_2

$$\begin{cases} 2 \cdot p_0 + 0 \cdot p_1 + \dfrac{2}{3} \cdot p_2 + 0 \cdot 1 = 0, \\ 0 \cdot p_0 + \dfrac{2}{3} \cdot p_1 + 0 \cdot p_2 + \dfrac{2}{5} \cdot 1 = 0, \\ \dfrac{2}{3} \cdot p_0 + 0 \cdot p_1 + \dfrac{2}{5} \cdot p_2 + 0 \cdot 1 = 0, \end{cases}$$

откуда получаем $p_0 = p_2 = 0$, $p_1 = -\frac{3}{5}$ и

$$W_3(x) = x^3 - \frac{3}{5}x = x\left(x + \sqrt{\frac{3}{5}}\right)\left(x - \sqrt{\frac{3}{5}}\right).$$

Теперь легко определяем веса $\omega_1 = \omega_3 = \frac{5}{9}$, $\omega_2 = \frac{8}{9}$ и искомую квадратуру Гаусса

$$\int_{-1}^{1} f(x)\,dx \approx \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right).$$

Точность квадратур Гаусса на классе $2r$ непрерывно дифференцируемых на промежутке $[a, b]$ функций может быть оценена соотношением

$$|R_r(f)| \leqslant \frac{2^{2r+1}(r!)^4}{(2r!)^3(2r+1)} \max_{[a,b]} |f^{(2r)}(x)|, \tag{2}$$

откуда для полученных выше квадратур получаем

$$|R_1(f)| \leqslant \frac{1}{3} \max_{[a,b]} |f''(x)|,$$

$$|R_2(f)| \leqslant \frac{1}{135} \max_{[a,b]} |f^{(4)}(x)|,$$

$$|R_3(f)| \leqslant \frac{1}{15\,750} \max_{[a,b]} |f^{(6)}(x)|.$$

§ 5. Квадратуры специального назначения

Рассмотренные выше квадратурные формулы дают неплохие результаты в случае «хороших» подынтегральных функций, т. е. таких, которые хорошо аппроксимируются многочленами. Если же функция плохо приближается многочленами или такова, что интеграл в собственном смысле не существует, то использование квадратур Ньютона—Котеса или Гаусса неэффективно или просто невозможно.

В этой ситуации обычно поступают следующим образом: для построения квадратуры для вычисления интеграла

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

подынтегральную функцию представляют в виде произведения

$$f(x) = q(x)\varphi(x)$$

где функция $\varphi(x)$ «хорошая» — аппроксимируется многочленами с достаточной степенью точности, а функция $q(x)$ — «весовая», в ней собраны все особенности подынтегральной функции $f(x)$ и она обладает тем свойством, что ее моменты

$$q_s = \int_a^b x^s q(x)\,dx, \quad s = 0, 1, 2, \ldots,$$

легко вычисляются аналитически. Если теперь положить

$$\varphi(x) \approx \sum_{i=1}^{r} \varphi_i x^i,$$

то для рассматриваемого интеграла при этом получаем

$$\int_a^b f(x)\,dx \approx \int_a^b q(x) \sum_{i=1}^{r} \varphi_i x^i\,dx = \sum_{i=1}^{r} \varphi_i \int_a^b q(x) x^i\,dx = \sum_{i=1}^{r} \varphi_i q_i.$$

Эти соотношения являются источником построения конкретных квадратур. При этом возможно построение как интерполяционных квадратур с фиксированными узлами, так и аналогов квадратур Гаусса со свободными узлами.

Рассмотрим несколько примеров.

Интегрирование быстроосциллирующих функций

Довольно часто приходится сталкиваться с проблемой вычисления интегралов вида

$$S(M) = \int_{-1}^{1} \varphi(x) \sin(Mx)\,dx, \quad C(M) = \int_{-1}^{1} \varphi(x) \cos(Mx)\,dx,$$

где M — большое число, так что подынтегральная функция на промежутке интегрирования является сильно колеблющейся. Для удовлетворительного интегрирования функций такого вида с помощью квадратур Ньютона—Котеса или квадратур Гаусса придется использовать квадратуры с большим числом узлов.

В то же время, заменив функцию $\varphi(x)$ ее интерполяционным многочленом Лагранжа

$$\varphi(x) \approx \sum_{i=1}^{r} \varphi(x_i)\Lambda_i(x),$$

и вычислив моменты тригонометрических функций

$$q_i^{\sin} = \int_{-1}^{1} \Lambda_i(x) \sin(Mx)\,dx, \quad q_i^{\cos} = \int_{-1}^{1} \Lambda_i(x) \cos(Mx)\,dx,$$

получим r-узловые квадратуры

$$S(M) \approx \sum_{i=1}^{r} q_i^{\sin} \varphi(x_i), \quad C(M) \approx \sum_{i=1}^{r} q_i^{\cos} \varphi(x_i),$$

позволяющие вычислять указанные интегралы с относительно высокой степенью точности уже при достаточно малых значениях r.

Например, для интегралов $C(M)$ при $r = 2$, положив $x_1 = -1$, $x_2 = 1$, получаем

$$\Lambda_1(x) = \frac{1-x}{2}, \quad \Lambda_2(x) = \frac{x+1}{2},$$

$$q_i^{\cos} = \int_{-1}^{1} \Lambda_i(x) \cos(Mx)\,dx = \int_{-1}^{1} \frac{1 \pm x}{2} \cos(Mx)\,dx = \frac{\sin M}{M},$$

и искомая двухузловая квадратура имеет вид

$$\int_{-1}^{1} \varphi(x) \cos(Mx)\, dx \approx \frac{\sin M}{M}(\varphi(-1) + \varphi(1)).$$

Она будет точной на классе линейных функций, а в классе дважды непрерывно дифференцируемых допускает следующую оценку погрешности:

$$|R_1(\varphi)| \leqslant \max_{[-1,1]} |f''(x)| \frac{C}{M}$$

(эта оценка груба, но при $M \to \infty$ верна по порядку).

Интегрирование неограниченных функций. Квадратурные формулы Гаусса—Эрмита (Гаусса—Чебышева)

Пусть подынтегральная функция неограничена на промежутке интегрирования, который без ограничения общности будем считать отрезком $[0, 1]$, и имеет в точке $x = 0$ интегрируемую особенность типа $f(x) \sim C/x^{\alpha}$, $0 < \alpha < 1$, так что интеграл

$$I = \int_0^1 f(x)\, dx$$

сходится. Положим

$$f(x) = \varphi(x)\frac{1}{x^{\alpha}},$$

где $\varphi(x)$ — хорошо аппроксимируемая многочленами на промежутке $[0, 1]$ функция. Тогда, положив

$$\varphi(x) \approx \sum_{i=1}^{r} \varphi(x_i)\Lambda_i(x),$$

получим r-узловую квадратуру для вычисления рассматриваемого несобственного интеграла

$$\int_0^1 f(x)\, dx = \int_0^1 \varphi(x)\frac{dx}{x^{\alpha}} \approx \sum_{i=1}^{r} \varphi(x_i) \int_0^1 \Lambda_i(x)\frac{dx}{x^{\alpha}} = \sum_{i=1}^{r} \varphi(x_i) q_i^{(\alpha)},$$

где $q_i^{(\alpha)}$ — веса квадратуры, даваемые соотношениями

$$q_i^{(\alpha)} = \int_0^1 \Lambda_i(x)\frac{dx}{x^{\alpha}}.$$

Рассмотрим в качестве примера интегралы вида

$$\int_{-1}^{1} \varphi(x)\frac{dx}{\sqrt{1-x^2}}, \tag{1}$$

имеющие в концах промежутка интегрируемые особенности типа $\frac{C}{(1\pm x)^{1/2}}$. Предполагается, что функция $\varphi(x)$ — «хорошая» на $[-1, 1]$. Полагая, как и выше,

$$\varphi(x) \approx \varphi(-1)\frac{1-x}{2} + \varphi(1)\frac{1+x}{2},$$

получаем $(i = 1, 2)$

$$q_i = \int_{-1}^{1} \Lambda_i(x)\frac{dx}{\sqrt{1-x^2}} = \int_{-1}^{1} \frac{1 \pm x}{2}\frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2},$$

и искомая квадратура имеет вид

$$\boxed{\int_{-1}^{1} \varphi(x)\frac{dx}{\sqrt{1-x^2}} \approx \frac{\pi}{2}(\varphi(-1) + \varphi(1)).} \tag{2}$$

Если отказаться от фиксации узлов квадратуры, то их можно найти, потребовав, чтобы отыскиваемая квадратура была точной на классе многочленов максимально возможной в рассматриваемой ситуации степени. Соответствующие квадратуры для интегралов (1) называются *квадратурными формулами Гаусса—Эрмита.*

Можно показать, что узлы r-узловой квадратуры Гаусса—Эрмита являются корнями многочлена Чебышева[2] $T_r(x) = \cos(r \arccos x)$ и задаются соотношениями

$$i = 1, 2, \ldots, r, \quad x_i = \cos\frac{2i-1}{2r}\pi,$$

а веса равны $q_1 = q_2 = \ldots = q_r = \pi/r$, так что r-узловая квадратура Гаусса—Чебышева имеет вид

$$\int_{-1}^{1} \varphi(x)\frac{dx}{\sqrt{1-x^2}} \approx \frac{\pi}{r}\sum_{i=1}^{r} \varphi(x_i).$$

Для остаточного члена в этом случае получаем

$$|R_r(\varphi)| \leqslant \max_{[-1,1]} |\varphi^{(2r)}(x)|\frac{\pi}{(2r)!\,2^{r-1}}.$$

Эти квадратуры точны на классе многочленов, степень которых не превышает $2r - 1$ и могут быть построены для любого r.

В частности, при $r = 2$ приходим к двухузловой квадратуре Гаусса—Чебышева

$$\boxed{\int_{-1}^{1} \varphi(x)\frac{dx}{\sqrt{1-x^2}} \approx \frac{\pi}{2}\left[\varphi\left(-\frac{\sqrt{2}}{2}\right) + \varphi\left(\frac{\sqrt{2}}{2}\right)\right],}$$

отличной от полученной выше двухузловой квадратуры (2) с фиксированными узлами.

[2] В связи с этим обстоятельством разумно указанные квадратуры именовать квадратурами *Гаусса—Чебышева.*

Интегрирование на неограниченном промежутке

Если промежуток интегрирования неограничен, а подынтегральная функция такова, что интеграл сходится, то последний можно вычислить, ограничив промежуток интегрирования и применив на нем рассмотренные выше квадратуры:

$$\int_a^\infty f(x)\,dx = \int_a^b f(x)\,dx + R_{\text{int}} = \sum_{i=1}^{r} \omega_i f(x_i) + R_r(f) + R_{\text{int}}.$$

В силу сходимости интеграла, остаточный член R_{int} при $b \to \infty$ стремится к нулю. Выбирая достаточно большое значение b, можно добиться малого отличия интеграла по промежутку $[a, b]$ от интеграла по неограниченному промежутку.

Другой способ построения квадратур в этой ситуации предполагает, что подынтегральная функция асимптотически ведет себя как

$$f(x) \sim q(x)\varphi(x), \quad x \to \infty,$$

где функция $\varphi(x)$ хорошо аппроксимируется многочленами, а функция $q(x)$ быстро стремится к нулю при $x \to \infty$. Заменив $\varphi(x)$ многочленом и подсчитав моменты q_i, получим квадратуру

$$\int_a^\infty f(x)\,dx = \int_a^\infty q(x)\varphi(x)\,dx \approx \sum_{i=1}^{r} q_i \varphi(x_i).$$

Этот способ оказывается особенно эффективным для интегралов вида

$$\int_0^\infty e^{-x}\varphi(x)\,dx \quad \text{и} \quad \int_{-\infty}^\infty e^{-x^2}\varphi(x)\,dx,$$

часто встречающихся в приложениях.

Если, как и выше, узлы квадратур не фиксировать, а искать, требуя, чтобы отыскиваемая квадратура была точной на классе многочленов максимально возможной степени, то для указанных интегралов мы получим квадратурные формулы, точные для многочленов степени не превышающей $2r - 1$. Узлы этих квадратур оказываются корнями многочленов Лагерра и многочленов Эрмита соответственно.

Приведем конкретные квадратуры Гаусса—Лагерра (для $r = 2$)

$$\boxed{\int_0^\infty e^{-x}\varphi(x)\,dx \approx \frac{2+\sqrt{2}}{4}\varphi(2-\sqrt{2}) + \frac{2-\sqrt{2}}{4}\varphi(2+\sqrt{2})}$$

и Гаусса—Эрмита ($r = 3$)

$$\boxed{\int_{-\infty}^\infty e^{-x^2}\varphi(x)\,dx \approx \frac{\sqrt{\pi}}{6}\left[\varphi\left(-\sqrt{\frac{3}{2}}\right) + 4\varphi(0) + \varphi\left(\sqrt{\frac{3}{2}}\right)\right].}$$

§ 6. Кубатурные формулы для кратных интегралов

В случае кратных интегралов основные идеи построения формул для их приближенного вычисления, которые называются *кубатурными формулами*, остаются теми же, что и выше. Однако трудоемкость описанных процедур быстро растет с ростом размерности задачи, т. е. кратности интеграла, подлежащего вычислению.

Чтобы не загромождать изложение выкладками, везде ниже мы будем рассматривать процедуры построения кубатурных формул для двойных интегралов. Перенос рассмотрений и выводов на случай более высоких размерностей принципиальных затруднений не вызывает.

Интерполяционные формулы

Рассмотрим интеграл

$$I = \iint\limits_D f(x, y)\, dx\, dy \tag{1}$$

от непрерывной ограниченной функции двух переменных $f(x, y)$, определенной в ограниченной области $D \subset \mathbb{R}^2$ с кусочно гладкой границей ∂D.

Пусть подынтегральная функция $f(x, y)$ заменена в области D интерполяционным соотношением

$$f(x, y) \approx \sum_{i=1}^{r} f(P_i)\Lambda_i(x, y)$$

так, что $P_i \in D$, $i = 1, 2, \ldots, r$, — узлы интерполяции, а функции $\Lambda_i(x, y)$ образуют базис Лагранжа для избранной интерполяционной системы в области D:

$$\Lambda_i(P_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Тогда для интеграла (1) получим

$$\iint\limits_D f(x, y)\, dx\, dy \approx \iint\limits_D \sum_{i=1}^{r} f(P_i)\Lambda_i(x, y)\, dx\, dy = \sum_{i=1}^{r} f(P_i) \iint\limits_D \Lambda_i(x, y)\, dx\, dy.$$

Последнее соотношение служит источником различных кубатурных формул для вычисления интегралов (1).

Так, например, если воспользоваться трехузловой треугольной интерполяционной схемой для функции $f(x, y)$ в области D, приходим к кубатурной формуле

$$\iint\limits_D f(x, y)\, dx\, dy \approx \sum_{i=1}^{r} f(P_i)\omega_i,$$

где узлы интерполяции P_i — вершины триангуляционной схемы для области интегрирования, а ω_i — сумма площадей треугольников с общей вершиной P_i.

Повторное интегрирование

Если предположить, что область D — «правильная» в одном из координатных направлений (всякая прямая, параллельная одному из координатных направлений, пересекает границу области не более чем в двух точках или по целому отрезку), то можно

прибегнуть к расстановке пределов интегрирования по области D

$$I = \iint\limits_D f(x, y)\,dx\,dy = \int\limits_a^b I(x)\,dx = \int\limits_a^b dx \int\limits_{c(x)}^{d(x)} f(x, y)\,dy$$

и последующему использованию одномерных квадратур для получающегося повторного интеграла

$$I(x) = \int\limits_{c(x)}^{d(x)} f(x, y)\,dy \approx \sum_{j=1}^{n(x)} f(x, y_j)\omega_j(x),$$

$$I = \int\limits_a^b I(x)\,dx \approx \sum_{i=1}^{N} I(x_i)\nu_i,$$

и окончательно

$$I = \iint\limits_D f(x, y)\,dx\,dy \approx \sum_{i=1}^{N} \sum_{j=1}^{n(x_i)} f(x_i, y_j)\nu_i\omega_j(x_i).$$

Используя различные квадратуры для однократных интегралов, мы можем таким способом построить для двойного интеграла различные кубатурные формулы.

Глава LXI

ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

§ 1. Постановка задачи

Пусть функция $f(x)$ определена, непрерывна и непрерывно дифференцируема на промежутке $[a, b]$ и точка x лежит на этом промежутке, в частности, может совпадать с одним из его концов. Будем предполагать, что нам известны точно вычисленные значения функции $f(x)$ в точках x_i, $i = 0, 1, \ldots, n$, $x_0 = a$, $x_n = b$. Требуется найти значение производной $f'(x)$ в точке x.

Соотношение

$$f'(x) = \sum_{i=s}^{r} c_i f(x_i) + R(f), \quad 0 \leqslant s < r \leqslant n, \tag{1}$$

будем называть *формулой численного дифференцирования*. Точность формулы численного дифференцирования описывается величиной $R(f)$, которая называется *остаточным членом* формулы.

Формулу численного дифференцирования будем называть *точной на заданном классе функций*, если для любой функции f этого класса $R(f) \equiv 0$

$$f'(x) = \sum_{i=s}^{r} c_i f(x_i).$$

Обозначим через h *диаметр разбиения*, $h = \max\limits_j |x_{j+1} - x_j|$. Говорят, что формула численного дифференцирования имеет *порядок аппроксимации* m, если

$$\left| f'(x) - \sum_{i=s}^{r} c_i f(x_i) \right| = |R(f)| \leqslant \text{const} \cdot h^m.$$

Например, простейшей формулой численного дифференцирования является формула

$$f'(x) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} + R,$$

дающая значение производной в точке $x_i \leqslant x \leqslant x_{i+1}$. Легко видеть, что она является точной на классе линейных функций. Действительно, если $f(x) = \alpha x + \beta$, то $f'(x) \equiv \alpha$. С другой стороны, $f(x_{i+1}) - f(x_i) = \alpha(x_{i+1} - x_i)$ и $R(\alpha x + \beta) = 0$.

Далее, раскладывая числитель по формуле Тейлора в точке x, получаем

$$f(x_{i+1}) = f(x) + f'(x)(x_{i+1} - x) + \delta_{i+1}(h), \quad f(x_i) = f(x) + f'(x)(x_i - x) + \delta_i(h),$$

и, следовательно,

$$R = f'(x) - \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = O(h),$$

т. е. эта формула имеет первый порядок точности.

§ 2. Метод неопределенных коэффициентов. Первая производная

Пусть фиксированы узлы $x_s, x_{s+1}, \ldots, x_r$, по значениям функции $f_j = f(x_j)$ в которых построена формула (1) § 1 численного дифференцирования. Эта формула полностью определяется своими коэффициентами c_j, которые можно найти, потребовав, чтобы искомая формула была *точной на классе многочленов, степень которых не превосходит* $N = r - s$ (количество неизвестных коэффициентов будет равно количеству условий, и есть основания надеяться, что указанные условия позволят однозначно определить коэффициенты).

Полагая последовательно $f(x) = x^k$, $k = 0, 1, 2, \ldots, r - s$, приходим к системе уравнений относительно неизвестных коэффициентов c_j

$$\begin{cases} c_s 1 + c_{s+1} 1 + \ldots + c_r 1 = 0, \\ c_s x_s + c_{s+1} x_{s+1} + \ldots + c_r x_r = 1, \\ c_s x_s^2 + c_{s+1} x_{s+1}^2 + \ldots + c_r x_r^2 = 2x, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ c_s x_s^{r-s} + c_{s+1} x_{s+1}^{r-s} + \ldots + c_r x_r^{r-s} = (r-s)x^{r-s-1}. \end{cases}$$

Определитель системы — определитель Вандермонда — не обращается в нуль, если узлы x_j попарно различны. Следовательно, при произвольном выборе попарно различных точек x_j и точки x эта система однозначно разрешима.

Проиллюстрируем изложенные выше соображения. Пусть точка x расположена между узлами с номерами i и $i+1$. Различных формул для нахождения производной функции f в точке x, точных на заданном классе многочленов, можно построить много, выбирая различные $r - s + 1$ узлов из имеющихся в нашем распоряжении $n+1$. Так, формулы численного дифференцирования, точные на классе линейных функций, могут быть получены $C_{n+1}^2 = \frac{(n+1)n}{2}$ способами.

Если в качестве узловых точек взять любые две соседние точки, получим систему

$$\begin{cases} c_j + c_{j+1} = 0, \\ c_j x_j + c_{j+1} x_{j+1} = 1, \end{cases}$$

из которой следует

$$c_{j+1} = -c_j = \frac{1}{x_{j+1} - x_j},$$

и искомая формула, уже упоминавшаяся выше простейшая формула численного дифференцирования

$$f'(x) = \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} + R_1.$$

Если же в качестве узловых взять две произвольные точки $x_s < x_r$, то приходим к формуле

$$f'(x) = \frac{f(x_r) - f(x_s)}{x_r - x_s} + R_2.$$

Заметим, что приведенные формулы пригодны для *любой* точки x отрезка $[a, b]$.

Исследование остаточного члена R полученных формул численного дифференцирования показывает, что лучшей будет формула, использующая два соседних узла, между которыми заключена точка x.

◀ Действительно, если дополнительно предположить, что функция $f(x)$ дважды непрерывно дифференцируема на промежутке $[a, b]$, то, заменяя значения $f(x_r)$ и $f(x_s)$ по формуле Тейлора, для остаточного члена получим

$$R = f'(x) - \frac{f(x_r) - f(x_s)}{x_r - x_s} = \frac{1}{2}\left[f''(\xi_r)(x - x_r)^2 - f''(\xi_s)(x - x_s)^2\right].$$

Отсюда

$$|R| \leqslant M\left[(x - x_r)^2 + (x - x_s)^2\right].$$

Правая часть достигает наименьшего значения при $x_s = x_i < x_r = x_{i+1}$. ▶

Если точка x, в которой ищется производная, совпадает с одним из узлов x_i, то мы получаем три различные формулы численного дифференцирования:

— *правая* формула

$$\boxed{f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} + R_{\text{пр}},}$$

— *левая* формула

$$\boxed{f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} + R_{\text{лев}},}$$

— *центральная* формула

$$\boxed{f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{x_{i+1} - x_{i-1}} + R_{\text{центр}}.}$$

Все эти формулы могут быть с равным успехом использованы для нахождения производных, однако предпочтение следует отдать все же *центральной*. Дело в том, что для часто встречающегося в приложениях случая *равномерного* разбиения промежутка $[a, b]$, эта формула в сравнении с прочими обладает дополнительными привлекательными свойствами.

Во-первых, *более высокой степенью аппроксимации*.

◀ Если $x_{i+1} - x_i = h \ \forall\, i$, то центральная формула принимает вид

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + R_{\text{центр}}.$$

Разложение функции $f(x)$ в точках $x_{i+1} = x_i + h$, $x_{i-1} = x_i - h$ по формуле Тейлора в этом случае может быть представлено в следующей форме

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

Для остаточного члена формулы численного дифференцирования (в предположении существования второй производной) получаем $R_{\text{центр}} = O(h^2)$, в то время как нецентральные формулы при тех же предположениях имеют порядок аппроксимации $O(h)$. ▶

Во-вторых, является *точной на классе многочленов, степень которых не превышает двух.*

◀ Если $f(x) = \alpha x^2 + \beta x + \gamma$, то, с одной стороны,

$$f'(x_i) = 2\alpha x_i + \beta = 2\alpha(x_0 + ih) + \beta,$$

а с другой

$$\frac{f(x_{i+1}) - f(x_{i-1})}{2h} = \alpha(x_{i+1} + x_{i-1}) + \beta =$$
$$= \alpha(x_0 + ih + h + x_0 + ih - h) + \beta = 2\alpha(x_0 + ih) + \beta. \blacktriangleright$$

Аналогичным образом могут быть получены формулы численного дифференцирования более высоких порядков аппроксимации, точные на классах многочленов более высоких степеней.

§ 3. Метод неопределенных коэффициентов. Старшие производные

Та же техника может быть использована и для построения формул численного дифференцирования порядков выше первого. В качестве иллюстрации получим формулы численного нахождения вторых производных.

Двучленных формул (т. е. формул вида $f''(x) = c_1 f(x_s) + c_2 f(x_r)$) для второй производной не существует, так как такая формула может быть получена из условия ее точности для линейных функций, что приводит к очевидному и абсолютно бесполезному соотношению

$$f''(x) = 0 \cdot f(x_s) + 0 \cdot f(x_r) \equiv 0.$$

Пусть искомая формула использует значения функции в точках $x_s < x_p < x_r$ и, следовательно, имеет вид

$$f''(x) = c_s f(x_s) + c_p f(x_p) + c_r f(x_r) + R.$$

Условие точности этой формулы на классе многочленов степени не выше второй приводит к системе

$$\begin{cases} c_s + c_p + c_r = 0, \\ c_s x_s + c_p x_p + c_r x_r = 0, \\ c_s x_s^2 + c_p x_p^2 + c_r x_r^2 = 2, \end{cases}$$

решение которой

$$c_s = \frac{2}{(x_s - x_p)(x_s - x_r)}, \quad c_p = \frac{2}{(x_p - x_s)(x_p - x_r)}, \quad c_r = \frac{2}{(x_r - x_s)(x_r - x_p)},$$

так что искомая формула дается соотношением

$$f''(x) = \frac{2f(x_s)}{(x_s - x_p)(x_s - x_r)} + \frac{2f(x_p)}{(x_p - x_s)(x_p - x_r)} + \frac{2f(x_r)}{(x_r - x_s)(x_r - x_p)} + R.$$

Для дальнейшего анализа полученного соотношения с целью упрощения выкладок ограничимся случаем равномерного разбиения промежутка $[a, b]$. В этом случае $x_i = x_0 + ih$ для любого номера i, и предыдущая формула запишется в виде

$$f''(x) = \frac{2}{h^2}\left[\frac{f(x_s)}{(s-p)(s-r)} + \frac{f(x_p)}{(p-s)(p-r)} + \frac{f(x_r)}{(r-s)(r-p)}\right] + R.$$

Полагая

$$f(x_i) = f(x) + f'(x)(x_i - x) + \frac{f''(x)}{2}(x_i - x)^2 + \alpha_i,$$

где $\alpha_i = o(h^2)$, после несложных выкладок получим

$$|R| \leqslant \frac{\alpha_s}{(s-p)(s-r)} + \frac{\alpha_p}{(p-s)(p-r)} + \frac{\alpha_r}{(r-s)(r-p)}.$$

Можно показать, что если точка x, в которой ищется вторая производная, расположена между узлами с номерами i и $i+1$, то наименьший по величине остаточный член отвечает расчетным узлам $x_s = x_{i-1}$, $x_p = x_i$, $x_r = x_{i+1}$. Аналогичное утверждение справедливо и для точек x, являющихся узлами сетки. Таким образом, наилучшая расчетная формула численного нахождения второй производной для равномерного разбиения дается соотношением

$$\boxed{f''(x) = \frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2} + R,}$$

где $x_i \leqslant x < x_{i+1}$.

§ 4. Интерполяционные формулы численного дифференцирования

Другим распространенным способом построения формул численного дифференцирования является *интерполяционный* способ. Функция $f(x)$ на промежутке $[a, b]$ заменяется каким-нибудь интерполяционным соотношением

$$f(x) = \sum_{i=1}^{k} f(x_i)\Lambda_i(x) + R_k$$

и в качестве производной q-го порядка в точке $x \in [a, b]$ принимается значение соответствующей производной интерполяционного выражения в этой точке

$$f^{(q)}(x) = \sum_{i=1}^{k} f(x_i)\Lambda_i^{(q)}(x) + R_k^{(q)}.$$

Если функция интерполируется алгебраическим многочленом, то получающиеся формулы будут точными на классе многочленов, степень которых не превышает степени интерполяционного многочлена. Как легко проверить, они совпадают с формулами, получающимися методом неопределенных коэффициентов. Для других интерполяционных соотношений получаются, вообще говоря, другие формулы.

В вычислительной практике хорошо зарекомендовали себя формулы численного дифференцирования, получающиеся дифференцированием интерполяционных сплайнов.

Чтобы проиллюстрировать возможности метода, получим, например, формулы численного дифференцирования в некоторой точке x промежутка $[a, b]$ по трем значениям функции $f(x)$ в соседних точках x_{i-1}, x_i, x_{i+1} на равномерной сетке $x_0 = a$, $x_i = x_0 + ih$. Полагая

$$f(x) = L_2(x) + R_2(f) = f(x_{i-1})\Lambda_{i-1}(x) + f(x_i)\Lambda_i(x) + f(x_{i+1})\Lambda_{i+1}(x) + R_2(f),$$

где

$$\Lambda_{i-1}(x) = \frac{(x - x_i)(x - x_{i+1})}{2h^2},$$

$$\Lambda_i(x) = -\frac{(x - x_{i-1})(x - x_{i+1})}{h^2},$$

$$\Lambda_{i+1}(x) = \frac{(x - x_{i-1})(x - x_i)}{2h^2},$$

получим расчетную формулу

$$f'(x) \approx \frac{f(x_{i-1})}{2h^2}[(x - x_i) + (x - x_{i+1})] - \frac{f(x_i)}{h^2}[(x - x_{i-1}) + (x - x_{i+1})] + \\ + \frac{f(x_{i+1})}{2h^2}[(x - x_{i-1}) + (x - x_i)].$$

В частности, для левого, центрального и правого узлов отсюда получаем следующие соотношения

$$\boxed{\begin{aligned} f'(x_{i-1}) &= \frac{4f(x_i) - 3f(x_{i-1}) - f(x_{i+1})}{2h}, \\ f'(x_i) &= \frac{f(x_{i+1}) - f(x_{i-1})}{2h}, \\ f'(x_{i+1}) &= \frac{f(x_{i-1}) - 4f(x_i) + 3f(x_{i+1})}{2h}. \end{aligned}}$$

Приведенные формулы точны на классе многочленов, степень которых не превышает двух.

§ 5. Неустойчивость процедур численного дифференцирования

Полученные выше формулы численного дифференцирования теоретически дают достаточно хорошие результаты. Однако при машинной реализации нас, как обычно, поджидают некоторые осложнения. Если значения функции $f(x)$ на промежутке $[a, b]$ получены с погрешностями[1] $\Delta f(x)$

$$\bar{f}(x) = f(x) + \Delta f(x),$$

[1] Наличие погрешностей, как уже отмечалось, может быть связано, во-первых, с ошибками в определении значений функции, а во-вторых, с округлением и представлением значений функции в машине.

то для производной $f'(x)$ получаем

$$\frac{d}{dx}\bar{f}(x) = \frac{d}{dx}f(x) + \frac{d}{dx}\Delta f(x).$$

Но даже если ошибка $|\Delta f(x)|$ очень мала, может оказаться что ее производная достаточно велика[2], что сводит на нет все наши усилия по нахождению производной $f'(x)$.

Если подобная *высокочастотная составляющая* погрешности отсутствует, если функция ведет себя «достаточно хорошо»[3] и вычислительные погрешности не очень велики, то использование полученных в предыдущих параграфах формул численного дифференцирования дает практически приемлемые результаты.

В противном случае требуются специальные вычислительные процедуры по нейтрализации влияния погрешностей на результат — *регуляризующие алгоритмы*. Основная идея последних состоит в *фильтрации* высокочастотной составляющей погрешности $\Delta f(x)$, т. е. замене неточно измеренной функции $\bar{f}(x)$ некоторой другой функцией $g(x)$, которая, с одной стороны, принимает значения, близкие к значениям функции $\bar{f}(x)$, а с другой стороны, меняется достаточно медленно. Если такая функция построена, то можно положить

$$f'(x) \approx g'(x),$$

без опасений используя для вычисления $g'(x)$ полученные выше формулы численного дифференцирования.

Одним из широко распространенных методов регуляризации является следующий: функцию $g(x)$ отыскивают из условия минимума функционала

$$G(\lambda, g) = \int_a^b \left[\bar{f}(x) - g(x)\right]^2 dx + \lambda \int_a^b \left[g'(x)\right]^2 dx$$

при некотором значении *параметра регуляризации* $\lambda \geq 0$. При этом малость первого слагаемого обеспечивает близость значений функций $\bar{f}(x)$ и $g(x)$, а малость второго — гладкость функции $g(x)$.

---

[2] Например, положим $f(x) = x$, $\Delta f(x) = \varepsilon \sin Mx$. Тогда $\bar{f}'(x) = f'(x) + M\varepsilon \sin Mx$ и ясно, что для сколь угодно малого ε можно взять настолько большое значение M, что истинное значение производной будет отличаться от найденного очень сильно.

[3] Например, хорошо приближается многочленом невысокой степени.

Глава LXII

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ. ЗАДАЧА КОШИ

Пусть функция $y(x)$ удовлетворяет на промежутке $[a, b]$ обыкновенному дифференциальному уравнению первого порядка

$$\frac{dy}{dx} = f(x, y) \tag{1}$$

и в начальной точке $x = a$ принимает заданное значение y_a

$$y(a) = y_a. \tag{2}$$

Задача отыскания решения уравнения (1), удовлетворяющего условию (2), носит название *задачи Коши*. В исключительных случаях решение этой задачи может быть получено *в квадратурах*, т. е. может быть выражено через элементарные функции и/или интегралы от элементарных функций. Как правило, это классические модельные ситуации, хорошо изученные в теоретическом плане.

В рядовой же прикладной ситуации единственным средством решения указанной задачи является численный анализ.

В любой процедуре численного решения можно выделить следующие основные этапы:

— *дискретизация* задачи — замена всех фигурирующих в задаче функций дискретным набором чисел и перевод дифференциального уравнения, подлежащего решению, в некоторые соотношения, связывающие эти числа,

— разработка *метода решения* полученного дискретного аналога и реализация его в вычислительном устройстве,

— установление *соответствия найденного решения дискретной задачи искомому точному*.

Прежде чем переходить к описанию этих этапов, напомним основные факты из теории обыкновенных дифференциальных уравнений.

§ 1. Свойства решений задачи Коши

1. *Существование и единственность*

Известно, что если правая часть уравнения

$$\frac{dy}{dx} = f(x, y) \tag{1}$$

удовлетворяет некоторым условиям регулярности в окрестности точки $x = a$, то можно гарантировать существование и единственность его решения в некоторой подокрестности этой точки. Сформулируем простейшее легко проверяемое условие подобного рода.

Теорема. *Если правая часть уравнения* (1) *непрерывна по совокупности переменных, дифференцируема по переменной y и производная $\frac{\partial f}{\partial y}$ ограничена, то задача Коши*

$$\frac{dy}{dx} = f(x, y), \quad y(a) = y_a. \tag{2}$$

однозначно разрешима в некоторой окрестности начальной точки.

Заметим, что *непрерывность* правой части обеспечивает только *разрешимость* задачи (2), но *единственности* не гарантирует. Для единственности нужно чтобы выполнялось условие ограниченности упомянутой выше производной или какое-нибудь другое условие подобного рода, например, условие Липшица:

существует постоянная C такая, что $\forall\, x \in (a, a + \delta)$ функция $f(x, y)$ удовлетворяет неравенству

$$|f(x, y_1) - f(x, y_2)| \leqslant C|y_1 - y_2|.$$

2. *Продолжаемость решений*

Теорема, сформулированная выше, носит *локальный характер* — существование и единственность решения гарантируются только в окрестности начальной точки. Может оказаться, что его невозможно продолжить на весь отрезок $[a, b]$, даже если правая часть уравнения (1) удовлетворяет условиям теоремы существования и единственности во всех точках этого отрезка. Условия продолжаемости решения описываются следующей альтернативой.

Теорема. *Если правая часть уравнения* (1) *непрерывна по совокупности переменных, дифференцируема по переменной y и производная $\frac{\partial f}{\partial y}$ ограничена в любой ограниченной области, то*

либо решение задачи Коши продолжается неограниченно вправо от начальной точки до ∞,

либо при некотором конечном значении $\bar{x}$ решение задачи Коши имеет вертикальную асимптоту.

Последняя возможность иллюстрируется следующим примером

$$\frac{dy}{dx} = y^2 + 1, \quad y(0) = 0.$$

Решение этой задачи $y = \operatorname{tg} x$ правее точки $\bar{x} = \pi/2$, в которой у функции $\operatorname{tg} x$ вертикальная асимптота, не продолжается.

3. *Зависимость от параметров*

Важным для дальнейшего является вопрос о зависимости решения рассматриваемой задачи от различных параметров, в том числе от начальных условий, правой части и т. п.

Теорема. *Если правая часть уравнения*

$$\frac{dy}{dx} = f(x, y, \mu)$$

непрерывна по совокупности переменных x, y, μ, дифференцируема по переменной y и производная $\frac{\partial f}{\partial y}$ равномерно ограничена относительно параметра μ, то решение этого уравнения $y = y(x, \mu)$, удовлетворяющее начальному условию $y(a) = y_a$, непрерывно зависит от параметра μ.

Заметим, что наличие непрерывной зависимости решения от параметра (в том числе от начальных условий) гарантирует нам, что малые ошибки в задании параметра (начальных условий) вызывают малое изменение решения на всем *конечном* промежутке $[a, b]$, на котором мы решаем задачу. Однако, если решение продолжается на всю числовую прямую, то может оказаться, что малые изменения в задании параметра (начальных условий), не вызывающие больших изменений решения на конечном промежутке, вызывают значительные изменения при $x \to \infty$.

Решение, которое мало меняется при достаточно малом изменении *начальных условий* для сколь угодно больших значений аргумента, называется *устойчивым*.

Для дальнейшего важно, что устойчивость решения задачи Коши — внутреннее свойство задачи.

4. *Дифференцируемость решения*

Гладкость правой части рассматриваемого уравнения определяет гладкость решения. Сформулируем более строго.

Теорема. *Если в окрестности начальной точки (a, y_a) правая часть уравнения (1) n раз непрерывно дифференцируема, то решение задачи Коши (2) обладает в некоторой окрестности начальной точки $(n + 1)$-й непрерывной производной.*

---

5. *Интегральное уравнение*

В заключение этого краткого обзора отметим еще, что задача Коши (2) эквивалентна задаче решения интегрального уравнения

$$y(x) = y(a) + \int_a^x f(x, y)\, dx. \tag{3}$$

Теорема. *Если выполнены условия теоремы существования и единственности, решение задачи (2) является решением интегрального уравнения (3), и наоборот, решение интегрального уравнения (3) является решением задачи Коши (2). Обе задачи при этом однозначно разрешимы.*

---

§ 2. Дискретизация задачи Коши

2.1. Конечно-разностные схемы

Разобьем промежуток $[a, b]$ точками $a = x_0 < x_1 < \ldots < x_i < \ldots < x_N = b$ на N частей.

Полученное разбиение назовем *сеткой* на промежутке $[a, b]$, точки x_i — *узлами* сетки.

Сетку будем называть *равномерной*, если $\forall\, i$ выполняется равенство $x_{i+1} - x_i = h$.

Величину $h = \frac{b-a}{N}$ назовем *шагом* сетки. Если сетка не является равномерной, то под шагом сетки будем понимать величину $h = \max |x_{i+1} - x_i|$.

Вектор U^N с компонентами u_i, $i = 0, 1, \ldots, N$, назовем *сеточной функцией*.

Если некоторая функция $\varphi(x)$ определена на указанном промежутке, то ее *сеточным аналогом* назовем упорядоченную совокупность значений этой функции в узлах сетки — вектор Φ^N, задаваемый соотношением

$$\Phi^N = \begin{bmatrix} \varphi(x_0) \\ \varphi(x_1) \\ \ldots\ldots\ldots \\ \varphi(x_N) \end{bmatrix}.$$

Как уже отмечалось, процедура дискретизации исходной задачи Коши состоит в замене всех фигурирующих в задаче функций некоторыми сеточными функциями и в замене дифференциального уравнения расчетными соотношениями для их компонент.

Например, если заменить производную правой формулой численного дифференцирования

$$y'(x_s) \approx \frac{y(x_{s+1}) - y(x_s)}{h},$$

а правую часть $f(x, y)$ значением $f(x_s, y(x_s))$, то исходная задача нахождения функции $y(x)$ будет сведена к задаче отыскания сеточной функции[1] U^N, компоненты которой удовлетворяют расчетным соотношениям

$$\frac{u_{s+1} - u_s}{h} = f(x_s, u_s), \quad s = 0, 1, 2, \ldots, N-1,$$

и начальному условию

$$u_0 = y_a.$$

Для той же задачи можно получить другой дискретный аналог, заменив производную, другой, например, центральной формулой численного дифференцирования

$$y'(x_s) \approx \frac{y(x_{s+1}) - y(x_{s-1})}{2h}$$

и/или взяв в качестве правой части полусумму значений $f(x_{s+1}, y_{s+1})$ и $f(x_{s-1}, y_{s-1})$.

---

[1] Здесь важно понимать, что сеточная функция U^N, являющаяся решением приведенной ниже системы уравнений *не является, вообще говоря, сеточным аналогом решения задачи Коши*, т. е.

$$y(x_s) \neq u_s.$$

Подобные процедуры дискретизации исходной задачи Коши могут быть представлены следующим образом:

Значение производной в точке x_s заменяется процедурой численного дифференцирования (см. гл. LXI)

$$y'(x_s) = \sum_{i=0}^{N} \alpha_i^s y(x_i),$$

а значения правой части $f(x, y)$ в этой же точке каким-нибудь интерполяционным соотношением вида

$$f(x_s, y_s) = \sum_{i=0}^{N} \beta_i^s f(x_i, y_i).$$

При этом получается задача:

найти сеточную функцию U^N, определяемую условиями

$$\sum_{i=0}^{N} \alpha_i^s u_i = \sum_{i=0}^{N} \beta_i^s f(x_i, u_i), \quad s = 0, 1, 2, \ldots, N,$$

$$u_0 = y_a,$$

которую и принимают в качестве дискретного аналога исходной.

Полученный таким образом дискретный аналог задачи Коши называется ее *конечно-разностным* аналогом, или *конечно-разностной схемой*.

Однако это не единственно возможный путь дискретизации задачи Коши. Многие из широко использующихся в приложениях дискретизаций могут быть получены, например, следующим образом.

На сетке $a = x_0 < x_1 < \ldots < x_N = b$ (неважно — равномерной или нет), используя интегральный вариант формулировки задачи Коши, запишем очевидное тождество

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y(x))\, dx.$$

Если $y(x)$ — точное решение, то правая часть уравнения $f(x, y(x)) = F(x)$ является функцией переменной x. Заменяя $F(x)$ интерполяционным многочленом Лагранжа, построенным по некоторой наперед заданной системе узлов, и интегрируя его на промежутке $[x_i, x_{i+1}]$, получим равенство, которое и служит источником процедур дискретизации, вообще говоря, отличающихся от приведенных выше.

2.2. Формулы Адамса

Группа методов, известных под названием *методов Адамса*, получается, когда используется интерполяция правой части $f(x, y(x))$ по $q+1$ предшествующей узлу x_i точке — в качестве узлов интерполяции берутся узлы сетки $x_i, x_{i-1}, \ldots, x_{i-q}$.

При $q = 0$ интерполяционный многочлен Лагранжа задается равенством

$$L_0(x) = F(x_i) \equiv \text{const}$$

и мы получаем расчетную схему *метода Адамса нулевого порядка*, широко известную еще под названием *метода Эйлера*,

$$\frac{u_{i+1} - u_i}{h} = f(x_i, u_i).$$

Известное начальное условие $u_0 = y_a$ позволяет рекуррентно получить все значения сеточной функции, являющейся решением рассматриваемой дискретной задачи.

При $q = 1$ получаем

$$L_1(x) = F(x_i)\frac{x - x_{i-1}}{h} - F(x_{i-1})\frac{x - x_i}{h},$$

откуда следуют *расчетные формулы Адамса первого порядка*

$$\frac{u_{i+1} - u_i}{h} = \frac{3}{2}f(x_i, u_i) - \frac{1}{2}f(x_{i-1}, u_{i-1}).$$

Здесь уже для запуска вычислительного процесса недостаточно начального условия — требуется еще знание значения u_1, которое может быть задано, например, с использованием схемы Эйлера $u_1 = u_0 + hf(x_0, u_0)$, или из каких-нибудь других соображений.

Формулы Адамса более высоких порядков для запуска вычислительной процедуры требуют задания дополнительно ровно q значений $u_1, u_2, \ldots, u_q$. Их можно задавать, используя, например, схемы Адамса низших порядков.

2.3. Формулы Рунге—Кутта

Другой путь использования интегрального соотношения для построения дискретных аналогов задачи Коши состоит в следующем — на промежутке $[x_i, x_{i+1}]$ по некоторому наперед заданному набору узлов $x_i = x_i^0 < x_i^1 < x_i^2 < \ldots < x_i^q = x_{i+1}$ построим интерполяционный многочлен Лагранжа для правой части уравнения

$$F(x) = f(x, y(x)) = \sum_{s=0}^{q} f\left(x_i^s, y(x_i^s)\right)\Lambda_s(x)$$

и проинтегрируем его почленно. Это даст нам расчетные соотношения

$$\frac{u_{i+1} - u_i}{h} = \sum_{s=0}^{q} \omega_s^i \cdot f\left(x_i^s, y(x_i^s)\right),$$

в которых

$$\omega_s^i = \int_{x_i}^{x_{i+1}} \Lambda_s(x).$$

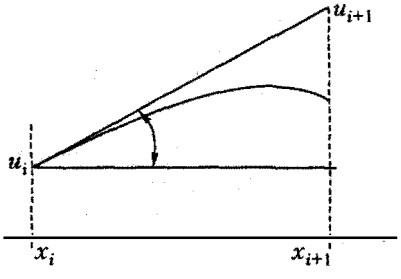


Рис. 1. Метод Эйлера — $u_{i+1} = u_i + hf(x_i, u_i)$

Простейшей формулой Рунге—Кутта является уже встречавшаяся выше формула Эйлера (она же формула Адамса нулевого порядка). Она соответствует значению $q = 0$. Отметим прозрачный геометрический смысл этой формулы (рис. 1): если значение искомой функции известно в некотором узле x_i, то в качестве значения в следующем узле сетки x_{i+1} берется то, которое получается линеаризацией функции на промежутке $[x_i, x_{i+1}]$ — на указанном промежутке график искомой функции заменяется касательной в точке x_i.

При $q = 1$ правая часть $f(x, y(x))$ интерполируется линейной функцией, так что

$$\int_{x_i}^{x_{i+1}} f(x, y(x))\,dx \approx \frac{h}{2}\big(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\big)$$

и соответствующие расчетные соотношения записываются в виде

$$u_{i+1} = u_i + \frac{h}{2}\big(f(x_i, u_i) + f(x_{i+1}, u_{i+1})\big).$$

Однако они неудобны для расчетов, так как при известном значении u_i тяжело поддаются решению в явном виде относительно u_{i+1}. Поэтому обычно поступают следующим образом: сначала используют формулу Эйлера для получения приближения к значению функции в точке x_{i+1}, а затем производят уточнение по приведеиной выше формуле, так что вся процедура строится на основе двух расчетных формул

$$u^1_{i+1} = u_i + hf(x_i, u_i), \quad u_{i+1} = u_i + \frac{h}{2}\big(f(x_i, u_i) + f(x_{i+1}, u^1_{i+1})\big),$$

которые и составляют содержание метода Рунге—Кутта первого порядка.

Геометрический смысл этого метода также достаточно прозрачен (рис. 2). Искомая функции линеаризуется на промежутке $[x_i, x_{i+1}]$ — ее график заменяется на указанном промежутке прямой линией, тангенс угла наклона которой принимается равным среднему арифметическому тангенсов углов наклона касательных в точка (x_i, u_i) и (x_{i+1}, u^1_{i+1}) и в качестве значения u_{i+1} берется значение этой линейной функции в точке x_{i+1}.

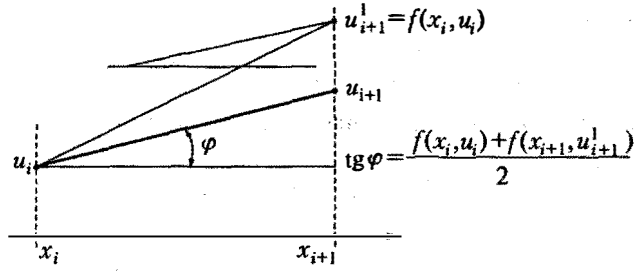


Рис. 2. Метод Рунге—Кутта

Общая структура алгоритмов Рунге—Кутта такова — на основе q-членного интерполяционного соотношения на сетке $x_i = x^0_i < x^1_i < x^2_i < \ldots < x^q_i = x_{i+1}$ строится расчетная формула

$$u_{i+1} = u_i + \sum_{s=1}^{q} \omega_s f^s,$$

для которой значения f^s последовательно уточняются с помощью расчетных формул низших порядков:

$$f^1 = hf(x_i, u_i),$$
$$f^2 = hf\big(x^2_i, u_i + \alpha_{21} f^1\big),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$f^q = hf\big(x_{i+1}, u_i + \alpha_{q1} f^1 + \alpha_{q2} f^2 + \ldots + \alpha_{qq-1} f^{q-1}\big).$$

§ 3. Сходимость

Осуществленное в § 2 сведение исходной дифференциальной задачи к дискретной относительно сеточной функции U^N имеет смысл только при условии, что решение дискретной задачи мало отличается от точного решения дифференциальной.

Если $y(x)$ — точное решение исходной задачи, Y^N — его сеточный аналог и U^N — решение дискретного аналога исходной задачи, то требуют, чтобы Y^N и U^N были близки, причем с измельчением разбиения (т. е. при $N \to \infty$ или, что то же, при $h \to 0$) эта близость становилась бы пренебрежимо малой.

Формализация приведенных соображений требует, чтобы на множестве сеточных функций была определена *близость* двух сеточных функций. Это можно сделать различными способами. Мы изберем следующий — будем оценивать близость двух сеточных функций U^N и V^N величиной $d(U^N, V^N)$, задаваемой соотношением

$$d(U^N, V^N) = \max_{0 \leqslant i \leqslant N} |u_i - v_i|.$$

Ясно, что чем меньше величина $d(U^N, V^N)$, тем ближе значения сравниваемых сеточных функций во всех узлах сетки одновременно.

Пусть исходная дифференциальная задача однозначно разрешима на промежутке[2] $[a, b]$ и для каждого значения $N = 1, 2, \ldots$ соответствующая дискретная задача также однозначно разрешима.

Будем говорить, что решение U^N дискретной задачи *сходится* к точному решению исходной задачи, если

$$\lim_{N \to \infty} d(U^N, Y^N) = 0.$$

Рассмотрим примеры, иллюстрирующие введенные понятия.

Пример 1. Для задачи Коши

$$y' = y, \quad y(0) = 1$$

на промежутке $[0, X]$ построим дискретный аналог, используя схему Эйлера

$$\frac{u_{s+1} - u_s}{h} = u_s, \quad s = 0, 1, \ldots, N, \quad u_0 = 1.$$

Точное решение исходной задачи Коши дается функцией $y(x) = e^x$. Дискретный аналог рассматриваемой задачи описывается расчетными соотношениями

$$u_0 = 1, \quad u_{s+1} = u_s(1 + h), \quad s = 0, 1, \ldots, N,$$

из которых легко получаем, что для любого значения N решение U^N дискретной задачи существует и дается сеточной функцией

$$U^N = \begin{bmatrix} 1 \\ 1 + h \\ (1 + h)^2 \\ \cdots\cdots\cdots \\ (1 + h)^N \end{bmatrix} = \begin{bmatrix} 1 \\ 1 + \dfrac{X}{N} \\ \left(1 + \dfrac{X}{N}\right)^2 \\ \cdots\cdots\cdots\cdots \\ \left(1 + \dfrac{X}{N}\right)^N \end{bmatrix}.$$

---

[2] Т. е. локально однозначно разрешима и решение продолжается вплоть до точки b.

Дискретный аналог точного решения дается сеточной функцией Y^N

$$Y^N = \begin{bmatrix} 1 \\ e^h \\ e^{2h} \\ \cdots\cdots \\ e^{Nh} \end{bmatrix} = \begin{bmatrix} 1 \\ e^{\frac{1}{N}X} \\ e^{\frac{2}{N}X} \\ \cdots\cdots \\ e^X \end{bmatrix}.$$

Отсюда

$$\max_s |u_s - y(x_s)| = \max_s \left|(1+h)^s - e^{sh}\right| = \left|(1+h)^N - e^X\right| \to 0, \quad N \to \infty,$$

и, следовательно, для рассмотренной дискретизации имеет место *сходимость* решения дискретного аналога к точному решению. ▶

Пример 2. Для той же задачи на промежутке $[0, X]$ построим дискретный аналог, заменив производную левым трехточечным разностным отношением (см. § 4 гл. LXI):

$$\frac{4u_s - 3u_{s-1} - u_{s+1}}{2h} = u_{s-1}, \quad s = 1, \ldots, N-1, \quad u_0 = 1. \tag{1}$$

Заметим, что в этом случае дискретный аналог имеет бесконечное множество решений — мы имеем N неизвестных величин u_s и только $N-1$ уравнение для их определения. Как правило, в подобной ситуации дополнительно к соотношениям (1) задают значение u_1, замыкая тем самым систему (1) и обеспечивая ее однозначную разрешимость[3].

Заметим, что в силу линейности и однородности рассматриваемой системы, вместе с любыми двумя решениями u_s^1 и u_s^2, ее решением будет и их линейная комбинация с произвольными коэффициентами C_1 и C_2

$$u_s(C_1, C_2) = C_1 u_s^1 + C_2 u_s^2.$$

Будем искать решение системы (1) в виде геометрической прогрессии $u_s = \lambda^s$. Подставляя u_s в (1), для определения λ получаем уравнение

$$\lambda^2 - 4\lambda + (3 + 2h) = 0,$$

корни которого

$$\lambda_1 = 2 + \sqrt{1 - 2h}, \quad \lambda_2 = 2 - \sqrt{1 - 2h}.$$

Решение исследуемой системы уравнений получаем в виде

$$u_s = C_1 \lambda_1^s + C_2 \lambda_2^s.$$

Постоянные C_1 и C_2 определим из начальных условий

$$\begin{cases} u_0 = C_1 + C_2, \\ u_1 = C_1 \lambda_1 + C_2 \lambda_2. \end{cases}$$

Из этих соотношений следует

$$C_1 = \frac{u_1 - \lambda_2}{\lambda_1 - \lambda_2}, \quad C_2 = \frac{\lambda_1 - u_1}{\lambda_1 - \lambda_2},$$

откуда решение рассматриваемого дискретного аналога дифференциальной задачи запишется в виде

$$u_s = \frac{u_1 - \lambda_2}{\lambda_1 - \lambda_2} \lambda_1^s + \frac{\lambda_1 - u_1}{\lambda_1 - \lambda_2} \lambda_2^s.$$

Для исследования сходимости полученного решения к точному решению исходной задачи, нам понадобится следующее соотношение[4]

$$\sqrt{1 - 2h} = 1 - h + o(h^2), \quad h \to 0.$$

---

[3] Из каких соображений следует задавать u_1? Если никакой дополнительной информации о решении у нас нет, то вполне приемлемым способом задания u_1 является, например, способ, использующий для определения u_1 разностную схему, рассмотренную выше

$$u_1 = u_0 + h \cdot f(x_0, u_0) = 1 + h \cdot 1,$$

хотя возможны и другие подходы.

[4] Имеющее место в силу формулы Тейлора.

Используя последнее, заключаем, что при $h \to 0$

$$\lambda_1 = 3 - h + o(h^2), \quad \lambda_2 = 1 + h + o(h^2), \quad \lambda_1 - \lambda_2 = 2 - 2h + o(h).$$

Рассмотрим, как ведет себя решение разностной задачи при $h \to 0$, $s \to \infty$. Для λ_1^s имеем

$$\lambda_1^s \approx (3-h)^s = 3^s \left(1 - \frac{h}{3}\right)^s \to 3^s e^{-x_s/3}.$$

Точно так же для λ_2^s получаем

$$\lambda_2^s \approx (1+h)^s \to e^{x_s}.$$

Для дальнейшего анализа нам нужно задать недостающее значение u_1. Положим, например, $u_1 = 1+h$, что хорошо согласуется со свойствами точного решения исходной задачи. Асимптотические соображения, аналогичные вышеизложенным, показывают, что коэффициенты C_1 и C_2 в этом случае асимптотически ведут себя как $\text{const} \cdot h^2 + o(h^2)$ и $1 + o(h)$ соответственно.

Суммируя вышеизложенное, заключаем, что при $h \to 0$, $s \to \infty$ решение разностного аналога не только не сходится к точному решению исходной задачи, но и, так как первое слагаемое неограниченно растет с ростом s

$$C_1 \cdot \lambda_1^s \approx \text{const} \cdot \frac{1}{s^2} \cdot 3^s e^{-x_s/3} \to \infty,$$

вообще не имеет предела.

Более тонкий анализ показывает, что это обстоятельство никак не связано с выбором значения u_1, как могло бы показаться на первый взгляд, а является *внутренним свойством* избранного метода дискретизации. Ясно, что для решения рассматриваемой задачи Коши подобные схемы непригодны и важно уметь определять, имеет или нет место сходимость решения дискретного аналога к точному решению исследуемой задачи Коши. ▶

§ 4. Аппроксимация. Устойчивость

В примере, рассмотренном в § 3, нам удалось обнаружить отсутствие сходимости благодаря тому, что мы построили и проанализировали решение дискретного аналога в явном виде. На практике это, как правило, невозможно, и хотелось бы иметь другие способы анализа наличия или отсутствия сходимости, не использующие формул для решения.

Ясно, что заменяя дифференциальную задачу дискретной разностной, нужно постараться сделать ее в известном смысле как можно более похожей на исходную. Тогда есть основания надеяться, что и решение дискретной задачи тоже будет мало отличаться от решения исходной дифференциальной. Посмотрим с этой точки зрения на разобранные в § 3 примеры.

Придадим, в первую очередь, точный смысл понятию *близости дискретной и дифференциальной задач*.

Пусть $y(x)$ — точное решение задачи Коши

$$y'(x) = f(x, y), \quad y(a) = y_a,$$

для численного анализа которой мы хотим использовать некоторую разностную схему. Подставим сеточный аналог точного решения $Y^N = (y(x_i))$ в разностную схему. Поскольку решение сеточной задачи, вообще говоря, не тождественно решению исходной, то в результате такой подстановки образуется *невязка* Δ^N — разница между результатом подстановки компонент $Y^N = (y(x_i))$ в левую и правую части сеточных уравнений. Чем меньше невязка, тем лучше разностная схема *аппроксимирует* исходную задачу Коши. Поскольку невязка — сеточная функция, то ее малость эквивалентна малости величины

$$d(\Delta^N, 0) = \max_s |\delta_s|,$$

которой мы и будем оценивать качество *аппроксимации* задачи Коши разностной схемой.

Будем говорить, что разностная схема *аппроксимирует* задачу Коши, если

$$d(\Delta^N, 0) = \max_s |\delta_s| \to 0, \quad h \to 0.$$

Пусть, например, мы решили использовать для решения задачи Коши разностную схему

$$\frac{u_{s+1} - u_s}{h} = f(x_s, u_s), \quad u_0 = y_a.$$

Подставляя в каждое из уравнений компоненты $y(x_i)$ сеточного аналога Y^N точного решения и используя тот факт, что

$$y_{i+1} - y_i = y'(x_i)h + O(h^2), \quad i = 1, 2, \ldots, n-1, \quad \text{и} \quad y'(x_i) = f(x_i, y_i),$$

получим следующую систему соотношений

$$\begin{aligned}
&y_0 - y_a = 0,\\
&\frac{y_1 - y_0}{h} - f(x_0, y_0) = O(h),\\
&\frac{y_2 - y_1}{h} - f(x_s, y_s) = O(h),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\frac{y_N - y_{N-1}}{h} - f(x_{N-1}, y_{N-1}) = O(h).
\end{aligned}$$

В этом случае невязка имеет вид

$$\Delta^N = \begin{bmatrix} 0 \\ O(h) \\ \ldots\ldots \\ O(h) \end{bmatrix}.$$

Первое уравнение (начальное условие исходной задачи Коши) на решении удовлетворяется точно, остальные — с точностью до $O(h)$, т. е. разница между левыми и правыми частями разностных уравнений стремится к нулю при $h \to 0$. Следовательно, рассмотренная разностная схема аппроксимирует исходную задачу Коши.

Если выполняется условие

$$d(\Delta^N, 0) = \max_s |\delta_s| = O(h^r),$$

то говорят, что *имеет место аппроксимация r-го порядка.*

В рассмотренном примере порядок аппроксимации равен 1.

Легко убедиться в том, что и для разностной схемы (1) § 3 аппроксимация имеет место.

Напомним, что для этой задачи $y_0 = 1$ и мы положили $y_1 = y_0 + hf(x_0, y_0) = 1 + h$. Полагая

$$y(x_i) = y(x_{i-1}) + y'(x_{i-1})h + \frac{1}{2}y''(x_{i-1})h^2 + O(h^3),$$

$$y(x_{i+1}) = y(x_{i-1}) + y'(x_{i-1})2h + \frac{1}{2}y''(x_{i-1})(2h^2) + O(h^3)$$

и учитывая, что $y'(x_{i-1}) = y_{i-1}$, получим

$$y_0 - y_a = 0, \quad y_1 - 1 - h = O(h^2)$$

и для всех $s = 1, 2, \ldots, N-1$

$$\frac{4y_i - 3y_{i-1} - y_{i+1}}{2h} - y_{i-1} = O(h^2).$$

Таким образом, эта разностная схема аппроксимирует задачу Коши со вторым порядком.

Тем не менее, как мы видели выше, решение построенного дискретного варианта исходной задачи *к точному не сходится.*

Причина этого неприятного обстоятельства связана с тем, что наличие *аппроксимации* не исключает, пусть и малых, но отличий используемой разностной схемы от разностной схемы, которой удовлетворяет точное решение исходной дифференциальной задачи Коши. При этом в одних случаях наличие подобных малых отличий вызывает малые же отличия соответствующих решений, а в других — значительные. Ясно, что наличие значительных ошибок в решении разностной задачи при малых ее возмущениях, делает схему непригодной для вычислений. Подчеркнем, что указанное обстоятельство связано только со свойствами взятой для решения разностной схемы.

Будем говорить, что разностная схема *устойчива*, если малые ее возмущения вызывают малые изменения решения.

Таким образом, пригодными для реализации вычислительного процесса могут считаться только устойчивые схемы.

Если надлежащим образом формализовать описанные выше понятия, то может быть доказана следующая фундаментальная теорема.

Теорема (аппроксимация + устойчивость = сходимость). *Если разностная схема аппроксимирует исходную задачу Коши и обладает устойчивостью, то имеет место сходимость решения разностной схемы к точному решению задачи Коши.*

---

§ 5. Системы обыкновенных дифференциальных уравнений

Пусть теперь $y(x)$, $f(x, y) = f(x; y_1, y_2, \ldots, y_n)$ — векторные функции, задаваемые соотношениями

$$y(x) = \begin{bmatrix} y_1(x) \\ y_2(x) \\ \cdots\cdots \\ y_n(x) \end{bmatrix}, \quad f(x, y) = \begin{bmatrix} f_1(x, y) \\ f_2(x, y) \\ \cdots\cdots\cdots \\ f_n(x, y) \end{bmatrix}.$$

Для задачи Коши

$$y'(x) = f(x, y), \quad y(a) = y_a$$

все рассмотрения предыдущих разделов сохраняют силу, если скалярные функции и их сеточные аналоги всюду понимать как векторные.

Например, расчетные соотношения метода Эйлера превращаются в систему рекуррентных соотношений

$$\begin{cases} u_{1i+1} = u_{1i} + hf_1(x_i, u_{1i}, u_{2i}, \ldots, u_{ni}), \\ u_{2i+1} = u_{2i} + hf_2(x_i, u_{1i}, u_{2i}, \ldots, u_{ni}), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ u_{ni+1} = u_{ni} + hf_n(x_i, u_{1i}, u_{2i}, \ldots, u_{ni}), \end{cases}$$

дополненную начальными условиями

$$u_{10} = y_{1a}, \quad u_{20} = y_{2a}, \quad \ldots, \quad u_{n0} = y_{na}.$$

Метод Рунге—Кутта первого порядка принимает форму

$$\begin{cases} u_{1i+1} = u_{1i} + \dfrac{h}{2} \cdot \left(f_1(x_i, u_{1i}, u_{2i}, \ldots, u_{ni}) + f_1\left(x_{i+1}, u^1_{1i+1}, u^1_{2i+1}, \ldots, u^1_{ni+1}\right)\right), \\ u_{2i+1} = u_{2i} + \dfrac{h}{2} \cdot \left(f_2(x_i, u_{1i}, u_{2i}, \ldots, u_{ni}) + f_2\left(x_{i+1}, u^1_{1i+1}, u^1_{2i+1}, \ldots, u^1_{ni+1}\right)\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ u_{ni+1} = u_{ni} + \dfrac{h}{2} \cdot \left(f_n(x_i, u_{1i}, u_{2i}, \ldots, u_{ni}) + f_n\left(x_{i+1}, u^1_{1i+1}, u^1_{2i+1}, \ldots, u^1_{ni+1}\right)\right), \end{cases}$$

где значения u^1_{ji+1}, $j = 1, 2, \ldots, n$, подлежат определению из системы уравнений метода Эйлера, приведенной выше.

Понятия, связанные с близостью сеточных вектор-функций, сходимостью, аппроксимацией и устойчивостью, переформулируются на случай систем очевидным образом. Все основные рассуждения и результаты, будучи переформулированы надлежащим образом, остаются справедливыми.

§ 6. Задача Коши для уравнений второго порядка

Дифференциальные уравнения, порядок которых выше первого, введением вспомогательных переменных легко могут быть сведены к системе уравнений первого порядка и тем самым, по крайней мере формально, проблема численного решения таких уравнений сводится к использованию методов и процедур, разработанных для систем.

Общая структура такого подхода следующая. Полагая $v_1(x) = y$, $v_2(x) = y'$, трансформируем задачу Коши для уравнения второго порядка

$$y'' = f(x, y, y'), \quad y(a) = y_a, \quad y'(a) = y^1_a$$

в задачу Коши для системы двух уравнений

$$\begin{cases} v_1' = v_2, \\ v_2' = f(x, v_1, v_2), \\ v_1(a) = y_a, \\ v_2(a) = y^1_a, \end{cases}$$

численное решение которой может быть получено описанными выше способами.

Однако уравнения второго порядка занимают особое место. Они широко распространены в приложениях и часто обладают специфической структурой, учет которой при разработке методов численного решения позволяет получать алгоритмы более эффективные, чем общеупотребительные универсальные.

Рассмотрим задачу

$$y'' = f(x, y, y'), \quad y(a) = y_a, \quad y'(a) = y_a^1.$$

Ее дискретный аналог может быть получен, например, прямой заменой производных разностными отношениями. Полагая

$$y'' = \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2}, \quad y' = \frac{y(x_{i+1}) - y(x_i)}{h},$$

приходим к расчетным соотношениям

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - f\left(x_i, u_i, \frac{u_{i+1} - u_i}{h}\right) = 0, \quad i = 1, 2, \ldots, N-1,$$

дополняя которые начальными условиями

$$u_0 = y_a, \quad \frac{u_1 - u_0}{h} = y_a^1,$$

получим неявную систему уравнений относительно компоненты u_{i+1} при известных компонентах u_{i-1}, u_i.

В общей ситуации этот подход не лучше сведения задачи Коши для уравнения к задаче Коши для системы. Однако если уравнение обладает специальными свойствами, то их учет может дать некоторые преимущества. Так, например, для часто встречающегося класса уравнений второго порядка вида

$$y'' = f(x, y), \tag{1}$$

правая часть которых не зависит явно от первой производной, разработаны специальные методы, являющиеся более эффективными в реализации, чем упомянутые выше.

Наиболее распространенными среди них являются методы, называемые *методами Штермера*, которые могут быть получены из следующих соображений. Рассмотрим решение $y(x)$ задачи Коши на промежутке $[x_i, x_{i+1}]$. Дважды интегрируя тождество (1), получим

$$y(x) - y(x_i) - y'(x_i)(x - x_i) = \int_{x_i}^{x} dx \int_{x_i}^{x} f(t, y(t))\, dt. \tag{2}$$

Зафиксировав теперь $q+1$ узел $x_i, x_{i-1}, x_{i-2}, \ldots, x_{i-q}$, заменим функцию $f(x, y(x))$ ее интерполяционным многочленом Лагранжа

$$f(x, y(x)) = \sum_{s=0}^{q} f(x_{i-s}, y(x_{i-s}))\Lambda_s(x).$$

Полагая в (2) $x = x_{i+1}$, заменим $y'(x_i)$ левым разностным отношением и проинтегрируем интерполяционный многочлен. Полученные расчетные соотношения называются *явными формулами Штермера* q-го порядка,

$$u_{i+1} - 2u_i + u_{i-1} = \sum_{s=0}^{q} f(x_{i-s}, y(x_{i-s}))\omega_s.$$

Здесь коэффициенты ω_s даются формулами

$$\omega_s = \int_{x_i}^{x_{i+1}} dx \int_{x_i}^{x} \Lambda_s(t)\, dt = \int_{x_i}^{x_{i+1}} (x_{i+1} - t)\Lambda_s(t)\, dt.$$

Неявные формулы Штермера q-го порядка получаются из аналогичных соображений, за исключением того, что интерполяция правой части осуществляется по системе узлов, включающих узел x_{i+1},

$$u_{i+1} - 2u_i + u_{i-1} = \sum_{s=0}^{q} f(x_{i+1-s}, y(x_{i+1-s}))\omega_s.$$

При $q = 0$, например, получаются следующие явное

$$u_{i+1} - 2u_i + u_{i-1} = h^2 f(x_i, u_i), \quad i = 1, 2, \ldots, N - 1,$$

и неявное

$$u_{i+1} - 2u_i + u_{i-1} = h^2 f(x_{i+1}, u_{i+1}), \quad i = 1, 2, \ldots, N - 1,$$

расчетные соотношения.

Глава LXIII

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ. КРАЕВЫЕ ЗАДАЧИ

Наряду с рассмотренной в гл. LXII задачей Коши часто приходится решать и другие задачи, связанные с обыкновенными дифференциальными уравнениями. Среди них одно из центральных мест занимают *краевые*, или *граничные* задачи, в которых из бесконечного множества решений обыкновенного дифференциального уравнения требуется выделить решения, удовлетворяющие определенным условиям на концах промежутка, на котором рассматривается уравнение.

Мы ограничимся рассмотрением простейшей краевой задачи для уравнения второго порядка, разрешенного относительного старшей производной.

§ 1. Краевая задача для уравнения второго порядка

Пусть $y(x)$ — некоторая дважды непрерывно дифференцируемая на промежутке $[a, b]$ функция и $f(x, v, w)$ — заданная функция трех переменных.

Краевой задачей для уравнения

$$y'' = f(x, y, y') \tag{1}$$

назовем задачу нахождения решения этого уравнения, принимающего на концах промежутка $[a, b]$ заданные значения

$$y(a) = y_a, \quad y(b) = y_b. \tag{2}$$

Краевая задача называется *линейной*, если $f(x, y, y') = p(x)y' + q(x)y + r(x)$. Линейная краевая задача называется *однородной*, если $r(x) \equiv y_a = y_b = 0$, и *полуоднородной*, если $y_a = y_b = 0$.

Не всякая краевая задача разрешима, а если и разрешима — то не всегда однозначно. Например, задача

$$y'' = -y, \quad y(0) = 0, \quad y(\pi) = \alpha,$$

рассматриваемая на промежутке $[0, \pi]$, не имеет решений при $\alpha \neq 0$ и имеет бесконечное множество решений, даваемых соотношением $y(x) = C \sin x$, при $\alpha = 0$.

Дополнительные условия, наложенные на правую часть уравнения, позволяют обеспечить однозначную разрешимость рассматриваемой задачи. Мы ограничимся здесь формулировкой трех теорем, носящих достаточно общий характер.

Теорема 1 (существование и единственность решения краевой задачи). *Пусть функция $f(x, v, w)$ непрерывна по совокупности переменных и обладает ограниченными частными производными на промежутке $[a, b]$*

$$\left|\frac{\partial f}{\partial v}\right| \leqslant \nu_v, \quad \left|\frac{\partial f}{\partial w}\right| \leqslant \nu_w,$$

причем постоянные ν_v и ν_w удовлетворяют условию

$$\nu_v \frac{(b-a)^2}{8} + \nu_w \frac{(b-a)}{2} < 1.$$

Тогда краевая задача (1)–(2) на $[a, b]$ однозначно разрешима.

Теорема 2 (существование и единственность решения линейной краевой задачи). *Пусть коэффициенты линейной краевой задачи $q(x)$ и $r(x)$ непрерывны на промежутке $[a, b]$, а коэффициент $p(x)$ непрерывно дифференцируем. Если выполнено условие*[1)]

$$q(x) - \frac{1}{2(b-a)} \frac{dp(x)}{dx} \geqslant c > -\frac{8}{\pi(b-a)^2} \quad \forall\, x \in [a, b],$$

то линейная краевая задача однозначно разрешима при любой $r(x)$.

В приложениях иногда оказывается полезной следующая *теорема единственности.*

Теорема 3 (единственность решения линейной краевой задачи). *Пусть коэффициенты линейной краевой задачи $p(x)$, $q(x)$ и $r(x)$ непрерывны на промежутке $[a, b]$ и удовлетворяют там условию*

$$q(x) \geqslant \frac{p(x)^2}{4}.$$

Тогда линейная краевая задача имеет не более одного решения.

Везде ниже, при обсуждении подходов к численному анализу краевых задач, мы будем предполагать, что исследуемая краевая задача однозначно разрешима.

§ 2. Метод стрельбы

При решении краевой задачи (1)–(2) § 1 естественно использовать один из изложенных в гл. LXII методов дискретизации дифференциального уравнения. Для компонент u_i сеточной функции U^N, определенной на сетке $a = x_0 < x_1 < \ldots < x_N = b$, мы получим при этом систему уравнений вида

$$u_{i+1} = \Phi(u_{i-s}, u_{i-s+1}, \ldots, u_i),$$

[1)] Это условие может быть заменено другим, близким к нему

$$q(x) - \frac{1}{2(b-a)} \frac{dp(x)}{dx} > -\frac{\pi}{(b-a)^2} \quad \forall\, x \in [a, b].$$

дополненную двумя граничными условиями $u_0 = y_a$ и $u_N = y_b$. В отличие от аналогичных систем для задачи Коши, которые можно было рекуррентно разрешить, последовательно определяя значения u_{i+1} по уже найденным $u_{i-s}, u_{i-s+1}, \ldots, u_i$ в соответствии с приведенным расчетным соотношением, в рассматриваемой ситуации этого добиться не удается из-за наличия условия на правом конце отрезка $[a, b]$.

Один из приемов, позволяющий получить решение исследуемой задачи, состоит в следующем: наряду с краевой задачей (1)–(2) § 1 рассматривается задача Коши

$$y'' = f(x, y, y'), \quad y(a) = y_a, \quad y'(a) = \xi,$$

где значение параметра ξ задано пока произвольно. Решая эту задачу, сравнивают значение, которое принимает найденное решение $y(x; \xi)$ на правом конце отрезка с заданным на этом конце граничным условием. Если окажется, что $y(b; \xi) = y_b$ — то найденное решение является одновременно и решением краевой задачи.

Такое совпадение, как правило, крайне маловероятно. Меняя значение параметра ξ, будем добиваться того, чтобы разница между $y(b; \xi)$ и y_b уменьшалась. Отсюда и название метода — *метод стрельбы*. Оно вызвано тем, что изменяя значения параметра ξ, мы меняем угловой коэффициент решения уравнения $y'' = f(x, y, y')$, выходящего из точки (a, y_a) (рис. 1), тем самым изменяя «точку попадания снаряда» — ординату правого конца траектории. Сравнение значения полученного решения с заданным граничным в правом конце отрезка является основанием для «корректировки» стрельбы, в зависимости от того, что мы наблюдаем — «недолет» ($y(b; \xi) < y_b$) или «перелет» ($y(b; \xi) > y_b$).

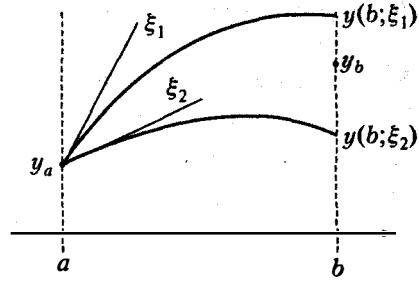


Рис. 1. Стрельба: $\xi = \xi_1$ — перелет, $\xi = \xi_2$ — недолет

Описанный выше направленный перебор значений параметра ξ может быть реализован в эквивалентной (и в некоторых ситуациях более эффективной) форме решения уравнения

$$y(b; \xi) = y_b.$$

Используя любой метод решения нелинейных уравнений (см. гл. LVIII), получим значение ξ_0, решающее поставленную задачу.

Применение метода бисекции, например, приводит к следующей последовательности вычислений:

1) если ξ_1 и ξ_2 — значения параметра, отвечающие недолету ($y(b; \xi) - y_b < 0$) и перелету ($y(b; \xi) - y_b > 0$) соответственно, то искомое значение (в силу предполагаемого наличия непрерывной зависимости решения от начальных условий) лежит на промежутке $[\xi_1, \xi_2]$,

2) полагаем

$$\xi_3 = \frac{\xi_1 + \xi_2}{2},$$

решаем задачу Кошу с начальными условиями

$$y(a) = y_a, \quad y'(a) = \xi_3,$$

3) сравнивая $y(b; \xi_3)$ и y_b обычным образом определяем следующее приближение ξ_4,

4) продолжаем процедуру до достижения требуемой точности.

Для линейных краевых задач

$$y'' = p(x)y' + q(x)y + r(x), \quad y(a) = y_a, \quad y(b) = y_b,$$

число решаемых вспомогательных задач Коши в сравнении с общим случаем может быть резко сокращено.

Действительно, пусть $\varphi(x)$ — решение задачи Коши для однородного уравнения

$$y'' = p(x)y' + q(x)y, \quad y(a) = 0, \quad y'(a) = 1.$$

Отметим, что в силу предполагаемой однозначной разрешимости краевой задачи, $\varphi(b) \neq 0$, так как в противном случае у однородной краевой задачи с нулевыми граничными условиями существует нетривиальное решение (с $\varphi'(a) \neq 0$). Пусть, далее, $\psi(x)$ — решение задачи Коши для неоднородного уравнения

$$y'' = p(x)y' + q(x)y + r(x), \quad y(a) = y_a, \quad y'(a) = 0.$$

Рассмотрим функцию $y_c(x) = C\varphi(x) + \psi(x)$ с некоторой, пока неопределенной постоянной C. При любых значениях этой постоянной функция $y_c(x)$ является решением исходного неоднородного уравнения, удовлетворяет на левом конце промежутка левому граничному условию исследуемой краевой задачи

$$y_c(a) = C\varphi(a) + \psi(a) = 0 + y_a = y_a,$$

а на правом конце отрезка принимает значение

$$y_c(b) = C\varphi(b) + \psi(b).$$

Если выбрать теперь значение постоянной C так, чтобы для функции $y_c(x)$ выполнялось правое граничное условие[2)]

$$y_c(b) = y_b \quad \Longrightarrow \quad C = \frac{y_b - \psi(b)}{\varphi(b)},$$

то полученная функция будет решением рассматриваемой краевой задачи.

Приведенные рассуждения показывают, что в случае линейной краевой задачи придется решить всего две вспомогательные задачи Коши.

В заключение отметим, что метод стрельбы хорошо ведет себя в реализации, если промежуток $[a, b]$ не слишком велик и искомое решение не сильно осциллирует на нем. В противном случае он становится неустойчивым и дает неприемлемые результаты[3)]. Качественно это явление можно проследить на примере следующей линейной краевой задачи[4)]

$$y'' = \omega^2 y, \quad y(0) = y_0, \quad y(l) = y_l.$$

[2)] В силу $\varphi(b) \neq 0$ это всегда можно сделать.

[3)] Даже если исходная краевая задача нечувствительна к вариации граничных условий.

[4)] Точное решение которой дается формулой

$$y(x) = y_0 \frac{\operatorname{sh} \omega(l - x)}{\operatorname{sh} \omega l} + y_l \frac{\operatorname{sh} \omega x}{\operatorname{sh} \omega l}.$$

Как легко можно проверить, малые варнации граничных условий вызывают малые изменения решения.

Решение вспомогательной задачи Коши

$$y'' = \omega^2 y, \quad y(0) = y_a, \quad y'(0) = \xi$$

может быть записано в виде

$$y_\xi(x) = y_0 \operatorname{ch} \omega x + \frac{\xi}{\omega} \operatorname{sh} \omega x,$$

что на правом конце промежутка дает

$$y_\xi(b) = y(b, \xi) = y_0 \operatorname{ch} \omega l + \frac{\xi}{\omega} \operatorname{sh} \omega l.$$

Если значение ξ_0, при котором $y(l, \xi) = y_l$, определено неточно, скажем с погрешностью $\delta\xi$, так что $\bar{\xi}_0 = \xi_0 + \delta\xi$, то эта погрешность вызовет в точке $x = l$ возмущение Δy_l, даваемое соотношением

$$\Delta y_l = y(b, \bar{\xi}) - y(b, \xi) = \frac{\operatorname{sh} \omega l}{\omega} \delta\xi,$$

которое показывает, что с ростом l и/или ω возмущение решения на правом конце промежутка неограниченно растет, что, конечно, делает использование метода стрельбы бессмысленным.

§ 3. Линейные краевые задачи. Прогонка

В случае линейной краевой задачи

$$y'' = p(x)y' + q(x)y + r(x), \quad y(a) = y_a, \quad y(b) = y_b,$$

возможен и прямой путь решения, не использующий вспомогательных задач Коши. Дело в том, что в рассматриваемой ситуации дискретизация задачи на сетке $a = x_0 < x_1 < \ldots < x_N = b$ приводит к *системе линейных уравнений* относительно компонент сеточной функции U^N, которая может быть эффективно разрешена существующими методами (см. гл. LVII). Часто матрицы получающихся систем имеют специфическую (в частности, ленточную) структуру, что позволяет для их решения использовать специально разработанные процедуры.

Если, например, заменить производные стандартными разностными формулами

$$y'' = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \quad y' = \frac{y_{i+1} - y_{i-1}}{2h},$$

а коэффициенты $p(x)$, $q(x)$ и $r(x)$ их сеточными аналогами

$$P^N = (p(x_i)), \quad Q^N = (q(x_i)), \quad R^N = (r(x_i)),$$

то мы получим систему линейных уравнений с трехдиагональной матрицей

$$\begin{cases} u_0 = y_a, \\ \left(1 - p_i \dfrac{h}{2}\right) u_{i+1} - (2 + q_i h^2) u_i + \left(1 + p_i \dfrac{h}{2}\right) u_{i-1} = r_i h^2, \quad i = 1, 2, \ldots, N-1, \\ u_N = y_b, \end{cases}$$

которую можно экономно и быстро решить методом прогонки.

§ 4. Вариационные методы решения краевых задач

Многие задачи естествознания допускают описание исследуемых процессов и явлений с точки зрения вариационных принципов, каждый из которых утверждает, что из допустимых течений процесса фактически реализуются лишь те, которые доставляют экстремальное[5] значение некоторому функционалу, имеющему обычно определенный физический смысл. В качестве примеров отметим принцип Ферма в геометрической оптике, принцип наименьшего действия в динамике консервативных систем, принцип виртуальных перемещений в неконсервативной динамике и другие. При этом траектория исследуемого процесса для описывающего процесс функционала удовлетворяет необходимому условию экстремума. Поскольку прикладные вариационные принципы описываются, как правило, интегральными функционалами

$$J(y) = \int_a^b L(x, y, y')\,dx,$$

необходимое условие экстремума приводит к краевой задаче

$$\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} = 0, \quad y(a) = y_a, \quad y(b) = y_b, \tag{1}$$

для уравнения, которое называется *уравнением Эйлера*.

Таким образом задачу минимизации функционала можно свести к исследованию краевой задачи для дифференциального уравнения. А можно попробовать поступить наоборот — *заменить краевую задачу задачей минимизации некоторого функционала и использовать для решения последней хорошо разработанный аппарат численной минимизации функционалов*. Это и есть основная идея вариационных методов решения краевых задач.

Рассмотрим основные этапы ее реализации.

4.1. Сведение краевой задачи к вариационной

В первую очередь нужно уметь указывать функционал, для которого исследуемая краевая задача является необходимым условием экстремума. Ниже мы ограничимся рассмотрением линейных краевых задач

$$y'' = p(x)y' + q(x)y + r(x), \quad y(a) = y_a, \quad y(b) = y_b. \tag{2}$$

Заметим, что любая такая задача может быть представлена в стандартной форме

$$\boxed{-\frac{d}{dx}\left(P(x)\frac{dy}{dx}\right) = Q(x)y + R(x), \quad y(a) = y_a, \quad y(b) = y_b.}$$

◀ Положим

$$P(x) = \exp\left\{-\int_a^x p(x)\,dx\right\}.$$

Замечая, что при этом

$$\frac{dP}{dx} = -p(x)\cdot P(x),$$

[5] Более глубокий анализ показывает, что стационарное.

умножим обе части уравнения (1) на $P(x)$ и свернем производную произведения

$$-P(x)y'' - p(x)P(x)y' = -\frac{d}{dx}\left(P(x)\frac{dy}{dx}\right).$$

Полагая

$$Q(x) = P(x)q(x), \quad R(x) = P(x)r(x),$$

получим искомое. ▶

Рассмотрим теперь функционал, задаваемый соотношением

$$\boxed{J(y) = \frac{1}{2}\int_a^b \left[-P(x)y'^2 + Q(x)y^2 + 2R(x)y\right]dx.} \tag{3}$$

Легко проверить, что он именно тот, который мы ищем. Действительно,

$$\frac{\partial L}{\partial y} = \frac{1}{2}\frac{\partial}{\partial y}\left[-P(x)y'^2 + Q(x)y^2 + 2R(x)y\right] = Q(x)y + R(x),$$

$$\frac{\partial L}{\partial y'} = \frac{1}{2}\frac{\partial}{\partial y'}\left[-P(x)y'^2 + Q(x)y^2 + 2R(x)y\right] = -P(x)\frac{dy}{dx},$$

$$\frac{d}{dx}\frac{\partial L}{\partial y'} = \frac{d}{dx}\left(-P(x)\frac{dy}{dx}\right).$$

Необходимое условие экстремума для этого функционала дается задачей

$$Q(x)y + R(x) - \frac{d}{dx}\left(-P(x)\frac{dy}{dx}\right) = 0, \quad y(a) = y_a, \quad y(b) = y_b,$$

что совпадает с исследуемой краевой задачей.

4.2. Метод Ритца

Сведя краевую задачу к вариационной, будем решать последнюю. Эффективным прямым методом поиска экстремума функционалов является *метод Ритца*. Суть его в следующем: пусть y^* — функция, доставляющая экстремум исследуемому функционалу $J(y)$, а система функций $\varphi_1(x), \varphi_2(x), \ldots, \varphi_N(x), \ldots$ такова, что функция y^* представима в виде

$$y^* = y_1\varphi_1(x) + y_2\varphi_2(x) + \ldots + y_N\varphi_N(x) + \ldots.$$

Рассмотрим конечный отрезок этого ряда[6]

$$y_N^* = y_1\varphi_1(x) + y_2\varphi_2(x) + \ldots + y_N\varphi_N(x)$$

и вспомогательную задачу поиска экстремума функции N переменных

$$\Psi(y_1, \ldots, y_N) \to \text{extr},$$

задаваемой соотношением

$$\Psi(y_1, \ldots, y_N) = J(y_N^*) = J\left(y_1\varphi_1(x) + y_2\varphi_2(x) + \ldots + y_N\varphi_N(x)\right).$$

Последняя задача может быть решена стандартными методами — численными или аналитическими. Определив значения коэффициентов y_i, за решение исследуемой вариационной задачи примем функцию[7] y_N^*.

[6] При больших значениях N мало отличающийся от y^*.

[7] При некоторых дополнительных предположениях относительно коэффициентов уравнения можно доказать, что так полученное решение сходится к точному при $N \to \infty$. При конечных значениях N y_N^*

4.3. Реализация метода Ритца для линейных краевых задач

Как мы убедились выше, в случае линейных краевых задач функционал (3) имеет специфическую структуру — он является квадратичным. В этой ситуации эффективным способом решения конечномерной экстремальной задачи является исследование системы необходимых условий экстремума

$$\frac{d\Psi(y_1, \ldots, y_N)}{dy_j} = 0, \quad j = 1, 2, \ldots, N,$$

которая оказывается системой линейных уравнений относительно неизвестных y_j.

При практической реализации метода важное значение имеет выбор функций $\varphi_j(x)$. Во-первых, функции $\varphi_j(x)$ должны быть подобраны так, чтобы их линейная комбинация удовлетворяла граничным условиям при любых N и y_j, во-вторых они должны «правильно» отражать характер поведения искомого решения.

Первое требование обычно удовлетворяется включением в систему функций $\varphi_j(x)$ функции $\varphi_0(x)$, удовлетворяющей заданным граничным условиям

$$\varphi_0(a) = y_a, \quad \varphi_0(b) = y_b,$$

так, чтобы все прочие удовлетворяли нулевым граничным условиям[8]. Что касается второго требования, то оно носит несколько мистический характер. Имеется в виду, что функции $\varphi_j(x)$ должны подбираться с использованием априорной информации о решении и так, чтобы приемлемая точность приближения искомого решения достигалась при относительно малом числе слагаемых. Удачный выбор этих функций позволяет получать относительно точное решение с малыми вычислительными затратами. При неудачном выборе функций $\varphi_j(x)$ приходится брать большие значения N, что приводит к необходимости решения систем высоких порядков, а это обстоятельство, в свою очередь, повышает трудоемкость процедуры решения.

4.4. Система уравнений метода Ритца

Пусть каким-то образом выбраны функция $\varphi_0(x)$, описывающая граничные условия, и функции $\varphi_j(x)$, обращающиеся в нуль на концах промежутка. Полагая $y_0 = 1$, запишем искомое приближение в виде

$$y_N^* = \sum_{i=0}^{N} y_i\varphi_i(x) = \varphi_0(x) + y_1\varphi_1(x) + y_2\varphi_2(x) + \ldots + y_N\varphi_N(x).$$

Подставляя в функционал (3), получим

$$J(y_N^*) = \frac{1}{2}\int_a^b \left[-P(x)\left(\sum_{i=0}^{N} y_i\varphi_i'(x)\right)^2 + Q(x)\left(\sum_{i=0}^{N} y_i\varphi_i(x)\right)^2 + \right.$$

$$\left. + 2R(x)\left(\sum_{i=0}^{N} y_i\varphi_i(x)\right)\right] dx,$$

и y^* будут, конечно, отличаться, но хочется надеяться, что не очень сильно. Во всяком случае, за счет увеличения N сколь угодно высокой точности всегда можно добиться.

[8] Часто в приложениях в качестве $\varphi_0(x)$ берется линейная функция

$$\varphi_0(x) = y_a\frac{x-b}{a-b} + y_b\frac{x-a}{b-a},$$

а в качестве $\varphi_j(x)$ функции $(x-a)^s(x-b)^l$, $s+l=j$.

откуда искомая система уравнений

$$\frac{\partial}{\partial y_j} J(y_N^*) = 0, \quad j = 1, 2, \ldots, N,$$

запишется в виде

$$\begin{cases} y_1 A_{11} + y_2 A_{12} + \ldots + y_N A_{1N} = B_1, \\ y_1 A_{21} + y_2 A_{22} + \ldots + y_N A_{2N} = B_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_1 A_{N1} + y_2 A_{N2} + \ldots + y_N A_{NN} = B_N. \end{cases} \tag{4}$$

Коэффициенты A_{ij} и правые части B_j вычисляются по формулам

$$A_{ij} = \int_a^b \left[-P(x)\varphi_i'(x)\varphi_j'(x) + Q(x)\varphi_i(x)\varphi_j(x) \right] dx, \tag{5}$$

$$B_j = \int_a^b \left[P(x)\varphi_0'(x)\varphi_j'(x) - Q(x)\varphi_0(x)\varphi_j(x) - R(x)\varphi_j(x) \right] dx. \tag{6}$$

Пример 1 (одночленная аппроксимация). Рассмотрим задачу

$$y'' = y + \sin x, \quad y(0) = 0, \quad y(\pi) = 0,$$

точное решение которой дается функцией $y = -\frac{1}{2}\sin x$. Эта задача описывает необходимое условие минимума функционала

$$J(y) = \frac{1}{2}\int_0^\pi (y'^2 + y^2 + 2y\sin x)\, dx, \quad y(0) = 0, \quad y(\pi) = 0.$$

◀ Полагая $\varphi_j(x) = x^j(\pi - x)$, $j = 1, 2, \ldots$, будем искать решение задачи минимизации в виде

$$y_N^*(x) = y_1 x(\pi - x) + y_2 x^2(\pi - x) + \ldots + y_N x^N(\pi - x).$$

Заметим, что при всех значениях N и любых y_j функция $y_N^*(x)$ удовлетворяет граничным условиям. Ограничиваясь первым слагаемым, получаем $y_1^*(x) = y_1 x(\pi - x)$ и

$$\Psi(y_1) = \frac{1}{2}\int_0^\pi \left[(y_1(\pi - 2x))^2 + (y_1 x(\pi - x))^2 + 2y_1 x(\pi - x)\sin x \right] dx.$$

Коэффициент y_1 найдем из условия $\frac{d\Psi}{dy_1} = 0$, которое приводит к уравнению

$$y_1 \int_0^\pi \left[(\pi - 2x)^2 + x^2(\pi - x)^2 \right] dx = -\int_0^\pi x(\pi - x)\sin x\, dx.$$

Из него получаем

$$\frac{\pi^3(\pi^2 + 10)}{30} y_1 = -4,$$

что дает значение $y_1 = -0{,}194779$ и выражение для искомого приближения $y_1^*(x)$

$$y_1^*(x) = -0{,}194779x(\pi - x).$$

На рис. 2 приведены эскизы точного и приближенного решений. Максимальная разница между найденным приближением и точным решением равна 0,019, что составляет около 4 %. Таким образом, вполне допустимая для большинства практических целей точность достигнута уже при $N = 1$. ▶

Пример 2 (несколько слагаемых). В задаче

$$y'' = y + 34\sin 4x, \quad y(0) = 0, \quad y\left(\frac{\pi}{2}\right) = 0,$$

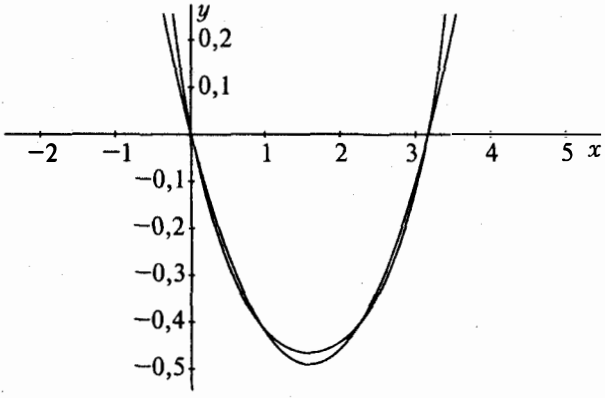


Рис. 2. Точность метода Ритца для задачи $y'' = y + \sin x$, $y(0) = y(\pi) = 0$

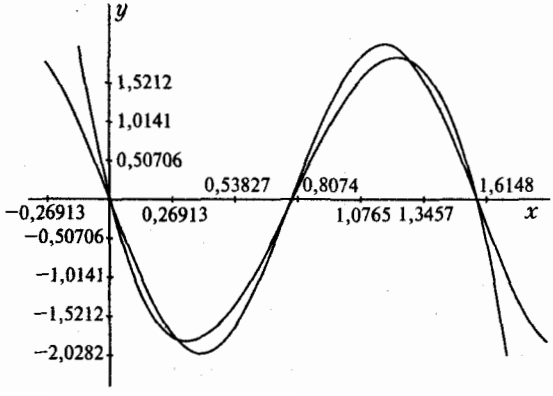


Рис. 3. Точность метода Ритца для задачи $y'' = y + \sin 4x$, $y(0) = y(\pi/2) = 0$

очень похожей на предыдущую, одним слагаемым уже нельзя достичь хорошей точности аппроксимации решения.

◀ Положим $N = 3$, $\varphi_i(x) = x^i\left(\frac{\pi}{2} - x\right)$,

$$y_3^* = y_1 x\left(\frac{\pi}{2} - x\right) + y_2 x^2\left(\frac{\pi}{2} - x\right) + y_1 x^3\left(\frac{\pi}{2} - x\right).$$

Система уравнений метода Ритца примет вид

$$\begin{cases} 1{,}611y_1 + 1{,}265y_2 + 1{,}181y_3 = \ \ 0{,}000, \\ 1{,}265y_1 + \ \ 1{,}5y_2 + 1{,}723y_3 = -5{,}000, \\ 1{,}181y_1 + 1{,}723y_2 + 2{,}254y_3 = -7{,}865, \end{cases}$$

откуда для коэффициентов y_i получим

$$y_1 \approx -7{,}768, \quad y_2 \approx 9{,}890, \quad y_3 \approx 0,$$

и искомое решение запишется в виде

$$y_3^* \approx -7{,}768x\left(\frac{\pi}{2} - x\right) + 9{,}890x^2\left(\frac{\pi}{2} - x\right),$$

в то время как точное решение дается формулой

$$y(x) = -2\sin 4x.$$

Графики точного решения и решения, полученного по методу Ритца, приведены на рис. 3.

Поясним, в чем тут дело. Точное решение плохо аппроксимируется многочленами низкой степени — для достижения приемлемой точности требуется многочлен, степень которого не ниже третьей, и чтобы в рассматриваемой ситуации добиться более высокой степени точности, нужно взять еще по крайней мере пару слагаемых. ▶

4.5. Кусочно-линейные аппроксимации

Удобным аппаратом построения конечномерных приближений в рассматриваемой ситуации оказываются кусочно-полиномиальные функции. Универсальность аппарата кусочно-полиномиальных аппроксимаций и специфическая структура получающихся при их применении систем уравнений обусловили широкое распространение именно такого способа выбора базисных функций[9].

Рассмотрим линейную краевую задачу

$$-\frac{d}{dx}\left(P(x)\frac{dy}{dx}\right) = Q(x)y + R(x), \quad y(a) = y_a, \quad y(b) = y_b.$$

Зададим на промежутке $[a, b]$ сетку $a = x_0 < x_1 < \ldots < x_N = b$ и на этой сетке определим кусочно-линейный базис Лагранжа (см. гл. LIX) соотношениями

$$\Lambda_0 = \begin{cases} \dfrac{x - x_1}{x_0 - x_1}, & x \in [x_0, x_1], \\ 0, & \text{иначе}, \end{cases} \qquad \Lambda_N = \begin{cases} \dfrac{x - x_{N-1}}{x_N - x_{N-1}}, & x \in [x_{N-1}, x_N], \\ 0, & \text{иначе}, \end{cases}$$

$$\Lambda_j = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x \in [x_{j-1}, x_j], \\ \dfrac{x - x_{j+1}}{x_j - x_{j+1}}, & x \in [x_j, x_{j+1}], \\ 0, & \text{иначе}. \end{cases}$$

Теперь для реализации метода Ритца положим

$$\varphi_j(x) = \Lambda_j(x), \quad j = 1, 2, \ldots, N-1,$$

и

$$\varphi_0(x) = y_a\Lambda_0(x) + y_b\Lambda_N(x).$$

Решение задачи ищем в виде

$$y_N^*(x) = \sum_{i=0}^{N} y_i\Lambda_i(x), \quad y_0 = y_a, \quad y_N = y_b.$$

Заметим, что всякая функция описанного вида является кусочно-линейной функцией с изломами в точках (x_i, y_i), так что искомые коэффициенты в данном случае имеют прозрачный смысл — это просто значения решения в узлах сетки.

Для построения системы (4) рассмотрим выражения (5), дающие ее коэффициенты. Поскольку функции Λ_i локализованы на промежутке $[x_{i-1}, x_{i+1}]$ (рис. 4), произведения $\Lambda_i'\Lambda_j'$ и $\Lambda_i\Lambda_j$ отличны от нуля только при условии $|i - j| \leqslant 1$. Поэтому в j-м уравнении системы не обращаются в нуль только коэффициенты A_{j-1j}, A_{jj} и A_{jj+1}.

[9] Прямые методы, использующие кусочно-полиномиальные базисы, получили название методов *конечных элементов*.

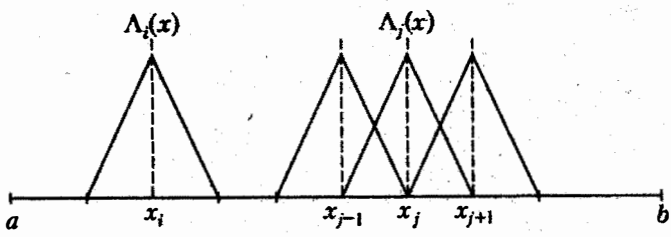


Рис. 4. Графики функций Λ_i, Λ_{j-1}, Λ_j и Λ_{j+1} при $|i-j|>1$

Система метода Ритца оказывается в этом случае трехдиагональной системой

$$\begin{cases} y_0 = y_a, \\ A_{j-1j}y_{j-1j} + A_{jj}y_{jj} + A_{jj+1}y_{jj+1} = B_j, \quad j = 1, 2, \ldots, N-1, \\ y_N = y_b, \end{cases}$$

коэффициенты которой могут быть найдены из соотношений

$$A_{j-1j} = \int_{x_{j-1}}^{x_j} \left[-P\Lambda'_{j-1}\Lambda'_j + Q\Lambda_{j-1}\Lambda_j \right] dx,$$

$$A_{jj+1} = \int_{x_j}^{x_{j+1}} \left[-P\Lambda'_{j+1}\Lambda'_j + Q\Lambda_{j+1}\Lambda_j \right] dx,$$

$$A_{jj} = \int_{x_{j-1}}^{x_{j+1}} \left[-P(\Lambda'_j)^2 + Q(\Lambda_j)^2 \right] dx.$$

В простейших случаях эти коэффициенты вычисляются по указанным формулам непосредственным интегрированием. Однако обычно для их вычисления строят квадратуры, исходя из кусочно-линейной аппроксимации коэффициентов P, Q и R уравнения.

Существующие методы решения трехдиагональных систем (уже упоминавшаяся выше прогонка и другие, аналогичные) делают этот вариант выбора базисных функций весьма привлекательным и эффективным в реализации.

Глава LXIV

УРАВНЕНИЯ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Методы решения задач математической физики, оставаясь принципиально такими же, как для обыкновенных дифференциальных уравнений, претерпевают тем не менее в сравнении с рассмотренными в предыдущих главах значительные идейные и технические изменения.

В теоретическом плане это связано, во-первых, с резко возрастающим разнообразием постановок задач в случае нескольких независимых переменных, а во-вторых, со сложностью исследования однозначной разрешимости большинства важных прикладных задач.

На практике указанные обстоятельства приводят к значительно более сложным в смысле обоснования и существенно более громоздким в плане технической реализации процедурам дискретизации. Немалые сложности представляет и разработка процедур решения полученных дискретных аналогов — системы уравнений, как правило, имеют высокий порядок и при уменьшении параметра дискретизации (например, шага сетки) число обусловленности (см. гл. LVII) их растет, что тоже исследования не облегчает. А отсюда и сложности изучения сходимости дискретных аналогов решений к точным.

Тем не менее, «скелет» численных методов остается прежним — дискретизация задачи, получение решения дискретного аналога и исследование того, насколько это решение похоже на искомое точное.

Ниже будут рассмотрены основные этапы построения и реализации вычислительных процедур исследования простейших классических задач математической физики в случае двух независимых переменных — пространственных для уравнений стационарных (эллиптического типа) и временно-пространственных для уравнений эволюционных (гиперболического и параболического типов).

§ 1. Основные уравнения

1.1. Классификация

Многие задачи математической физики описываются квазилинейными уравнениями в частных производных второго порядка

$$A_{11}\frac{\partial^2 z(x, y)}{\partial x^2} + 2A_{12}\frac{\partial^2 z(x, y)}{\partial x \partial y} + A_{22}\frac{\partial^2 z(x, y)}{\partial y^2} = B,$$

коэффициенты которых, вообще говоря, зависят от x, y, z, $\frac{\partial z}{\partial x}$, $\frac{\partial z}{\partial y}$.

Уравнение[1] называется *гиперболическим*, если

$$A_{11}A_{22} - A_{12}^2 < 0,$$

параболическим, если

$$A_{11}A_{22} - A_{12}^2 = 0,$$

и *эллиптическим*, если

$$A_{11}A_{22} - A_{12}^2 > 0.$$

Параболические и гиперболические уравнения описывают, как правило, нестационарные, т. е. эволюционирующие во времени процессы — переходные процессы, процессы распространения возмущений и т. п.

Для эллиптических уравнений типично описание стационарных процессов — установившиеся температурные и электростатические поля, упругие деформации и т. д.

Для выделения из совокупности всех решений уравнения какого-то определенного должна быть сформулирована *задача*. Другими словами, должны быть заданы некоторые дополнительные условия, выделяющие именно то решение, которое интересует исследователя. Для эволюционных уравнений это, как правило, *начальные* и *граничные* условия, либо некоторая их комбинация, для стационарных — только *граничные*.

1.2. Начально-граничная задача для волнового уравнения

В качестве представителя гиперболических уравнений ограничимся рассмотрением одномерного *волнового уравнения*[2]

$$\frac{\partial^2 z(x,t)}{\partial t^2} = c^2 \frac{\partial^2 z(x,t)}{\partial x^2} + f(x,t), \tag{1}$$

и следующей задачи для него:

Найти функцию $z(x,t)$, являющуюся решением уравнения (1) *и удовлетворяющую начальным*

$$z(x,0) = \varphi(x), \quad \frac{\partial z(x,0)}{\partial t} = \psi(x), \quad 0 \leqslant x \leqslant l,$$

и граничным

$$z(0,t) = \mu(t), \quad z(l,t) = \nu(t), \quad t \geqslant 0,$$

условиям.

Известно, что сформулированная задача однозначно разрешима, если начальные и граничные условия естественным образом согласованы. При этом решение $z(x,t)$ непрерывно зависит от начальных и граничных условий.

---

[1] Заметим, что в случае уравнения линейного, его тип зависит только от точки области, в случае квазилинейного — еще и от рассматриваемого решения.

[2] Чтобы подчеркнуть эволюционный характер этого и параболического уравнений, одну из независимых переменных обозначаем буквой t.

1.3. Начально-граничная задача для уравнения теплопроводности

Типичным представителем параболических уравнений является *уравнение теплопроводности*

$$\frac{\partial z(x, t)}{\partial t} = c^2 \frac{\partial^2 z(x, t)}{\partial x^2} + f(x, t). \tag{2}$$

Для него рассмотрим задачу:

Найти функцию $z(x, t)$, являющуюся решением уравнения (2) *и удовлетворяющую начальному*

$$z(x, 0) = \varphi(x), \quad 0 \leqslant x \leqslant l,$$

и граничным

$$z(0, t) = \mu(t), \quad z(l, t) = \nu(t), \quad 0 \leqslant t \leqslant T,$$

условиям.

Так поставленная задача однозначно разрешима при наличии естественной согласованности начального и граничных условий.

1.4. Задача Дирихле для уравнения Пуассона

Для уравнения эллиптического типа

$$\frac{\partial}{\partial x}\left(p^2(x, y)\frac{\partial z(x, y)}{\partial x}\right) + \frac{\partial}{\partial y}\left(q^2(x, y)\frac{\partial z(x, t)}{\partial y}\right) = -f\ (x, y) \tag{3}$$

типичной является следующая *внутренняя краевая задача*:

В ограниченной области Ω с кусочно-гладкой границей $\partial\Omega$ найти функцию $z(x, y)$, удовлетворяющую уравнению (3) *везде внутри Ω, непрерывную вплоть до границы $\partial\Omega$ и принимающую на границе заданные значения*

$$z(x, y)\big|_{(x,y)\in\partial\Omega} = \varphi(\omega)\big|_{\omega\in\partial\Omega}.$$

Решение этой задачи существует, единственно и непрерывно зависит от граничных условий, которые предполагаются непрерывными.

Мы будем обсуждать свойства задач эллиптического типа на примере уравнения Пуассона

$$\frac{\partial^2 z(x, y)}{\partial x^2} + \frac{\partial^2 z(x, t)}{\partial y^2} = -f(x, y).$$

§ 2. Двумерные сетки и сеточные функции

Пусть на плоскости переменных (x, y) задана ограниченная область Ω с границей $\partial\Omega$. *Сеткой* на Ω или дискретным аналогом Ω^{NM} области назовем совокупность $N \times M$ точек $P_{11}, P_{12}, \ldots, P_{NM}$, принадлежащих области Ω и/или ее границе $\partial\Omega$. *Сеточной функцией* на Ω^{NM} назовем упорядоченный набор $N \times M$ чисел, $U^{NM} = (u_{ij})$, трактуемых в дальнейшем как значения в точках P_{ij} некоторой функции $u(x, y)$. Если функция двух переменных $f(x, y)$ определена в области Ω, то ее сеточным аналогом назовем сеточную функцию $f_{ij} = f(P_{ij})$.

Сетка на Ω может быть определена различными способами, соответственно и для функции $f(x, y)$, определенной в этой области, по-разному можно строить сеточные аналоги[3].

Мы рассмотрим здесь наиболее употребительные способы дискретизации двумерной области.

2.1. Прямоугольные сетки

Пусть проекции области Ω на координатные оси Ox и Oy даются промежутками $[a, b]$ и $[c, d]$ соответственно. Зададим на них одномерные сетки $a = x_0 < x_1 < \ldots < x_Q = b$ и $c = y_0 < y_1 < \ldots < y_R = d$ и положим $P_{ij} = (x_i, y_j)$. Сетка Ω^{NM} (рис. 1) состоит из всех тех узлов P_{ij}, которые лежат внутри области Ω или на ее границе $\partial\Omega$. Узел P_{ij} называется внутренним, если смежные с ним узлы $P_{i\pm1\,j}$ и $P_{ij\pm1}$ лежат внутри Ω или на ее границе $\partial\Omega$. Все прочие узлы называются граничными[4].

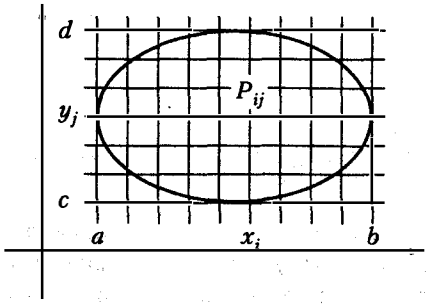


Рис. 1. Прямоугольная сетка

В приложениях распространены координатно-равномерные сетки,

$$x_{i+1} - x_i = \frac{b-a}{N} = h, \quad y_{j+1} - y_j = \frac{d-c}{M} = \tau.$$

2.2. Треугольные сетки

Пусть область Ω триангулирована (рис. 2), т. е. разбита некоторым образом на треугольники так, что их вершины лежат внутри области и/или на границе. В качестве узлов сетки Ω^N берутся узлы триангуляции. Сеточные функции и сеточные аналоги функций, заданных на Ω, определяются обычным образом.

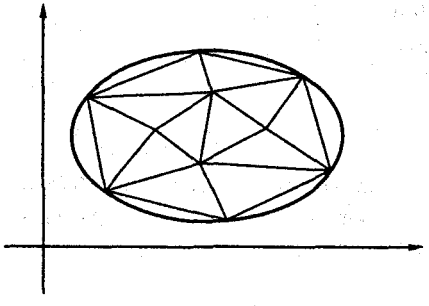


Рис. 2. Нерегулярная триангуляция

Распространенным вариантом триангуляции является такой — на Ω строится прямоугольная сетка с последующим делением каждой прямоугольной ячейки на две треугольные. На границе с некоторым шагом ω задаются точки, соединяя которые с граничными узлами прямоугольной сетки, триангуляцию области завершают (рис. 3).

Для областей, граница которых состоит из дуг окружностей, бывает удобно использовать дискретизацию области и ее границы прямоугольной сеткой в полярных координатах, что в декартовых индуцирует радиально-полярные сетки. Пример такой сетки приведен на рис. 4.

---

[3] Восстановление функции $f(x, y)$ по ее сеточному аналогу — задача теории интерполяции, см. гл. LX.

[4] При таком способе дискретизации области возникает, как правило, проблема аппроксимации граничных условий, заданных на $\partial\Omega$, условиями в граничных узлах Ω^{NM}.

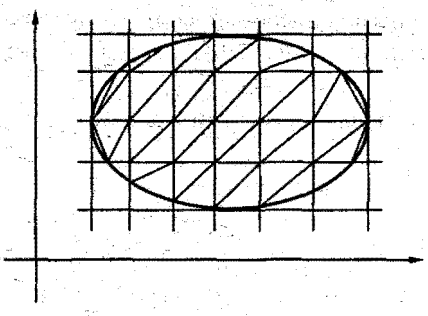


Рис. 3. Триангуляция на прямоугольной сетке

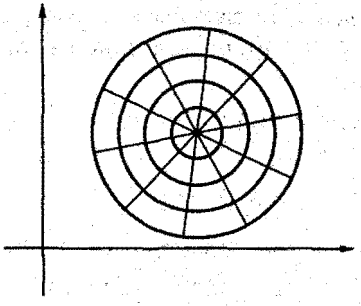


Рис. 4. Радиально-полярная сетка

§ 3. Дискретизация задачи

Для получения дискретного аналога рассматриваемых задач нужно заменить дифференциальные уравнения, начальные и граничные условия их разностными аналогами. Это можно сделать, как уже отмечалось выше, многими способами.

Для сокращения записи будем пользоваться в дальнейшем следующими обозначениями

$$z(x_i, y_j) = z_{ij}, \quad \frac{\partial z(x_i, y_j)}{\partial x} = z_{x,ij}, \quad \frac{\partial z(x_i, y_j)}{\partial y} = z_{y,ij}.$$

Аналогичные обозначения будем использовать и для старших производных $z_{xx,ij}$, $z_{yy,ij}$ и $z_{xy,ij} = z_{yx,ij}$. В этих обозначениях простейшие уравнения математической физики примут следующий вид

$$\begin{aligned} z_{tt} &= c^2 z_{xx} + f \quad \text{— волновое,} \\ z_t &= c^2 z_{xx} + f \quad \text{— теплопроводности,} \\ z_{xx} + z_{yy} &= -f \quad \text{— Пуассона.} \end{aligned}$$

3.1. Дискретизация уравнений. Шаблоны и расчетные соотношения

Начнем с дискретизации уравнений. Существует несколько различных путей. Можно заменить производные, фигурирующие в уравнениях, формулами численного дифференцирования, а функции — их сеточными аналогами, положив, например,

$$z_{x,ij} = \frac{z_{ij} - z_{i-1,j}}{h}, \quad z_{xx,ij} = \frac{z_{i+1,j} - 2z_{ij} + z_{i-1,j}}{h^2};$$

аналогично и для производных по другим переменным. Здесь мы использовали левую формулу численного дифференцирования для замены первых производных, однако возможно использование и других соотношений, например, центрального

$$z_{t,ij} = \frac{z_{i,j+1} - z_{i,j-1}}{2\tau}$$

или правого

$$z_{t,ij} = \frac{z_{i,j+1} - z_{ij}}{\tau}.$$

Аналогом исследуемого дифференциального уравнения при этом будем считать систему расчетных соотношений, получающихся из исходных дифференциальных уравнений заменой фигурирующих в них производных и функций их сеточными аналогами.

Например, при таком подходе волновое уравнение заменится расчетными соотношениями[5]

$$\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\tau^2} = c^2 \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} + f_{ij},$$

а для уравнения теплопроводности получим

$$\frac{u_{ij} - u_{i,j-1}}{\tau} = c^2 \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} + f_{ij}.$$

Если для замены первой производной по времени воспользоваться центральным разностным отношением, то уравнение теплопроводности заменится уже другими расчетными соотношениями

$$\frac{u_{i,j+1} - u_{i,j-1}}{2\tau} = c^2 \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} + f_{ij}.$$

После того как дискретизация уравнения тем или иным образом осуществлена, следует убедиться в том, что полученные расчетные соотношения *аппроксимируют* исходное уравнение. Для этого нужно изучить поведение невязки, возникающей при подстановке точного решения уравнения в разностный аналог, при измельчении сетки — $\tau \to 0$, $h \to 0$ (см. гл. LXIII). Предполагая, что $z(x, t)$ достаточно гладкая, получаем

$$z_{i+1,j} = z_{ij} + z_{x,ij}h + \frac{1}{2}z_{xx,ij}h^2 + O(h^3),$$

$$z_{i-1,j} = z_{ij} - z_{x,ij}h + \frac{1}{2}z_{xx,ij}h^2 + O(h^3),$$

$$z_{i,j+1} = z_{ij} + z_{t,ij}\tau + \frac{1}{2}z_{tt,ij}\tau^2 + O(\tau^3),$$

$$z_{i,j-1} = z_{ij} - z_{t,ij}\tau + \frac{1}{2}z_{tt,ij}\tau^2 + O(\tau^3).$$

Подставляя эти выражения, например, в полученный выше дискретный аналог волнового уравнения, для невязки получим

$$\frac{z_{i,j+1} - 2z_{ij} + z_{i,j-1}}{\tau^2} - \frac{z_{i+1,j} - 2z_{ij} + z_{i-1,j}}{h^2} - f_{ij} = O(h) + O(\tau),$$

т. е. исследуемая разностная схема аппроксимирует волновое уравнение и порядок аппроксимации равен 1 по каждой из переменных (h, τ).

Аналогичные выкладки для первой дискретизации уравнения теплопроводности дают

$$\frac{z_{ij} - z_{i,j-1}}{\tau} - c^2 \frac{z_{i+1,j} - 2z_{ij} + z_{i-1,j}}{h^2} - f_{ij} = O(\tau) + O(h),$$

т. е. опять имеет место аппроксимация исходного уравнения разностной схемой.

---

[5] В расчетных соотношениях фигурирует некоторая сеточная функция u_{ij}, вообще говоря, не тождественная z_{ij}, а только, может быть, похожая на нее (а может быть и нет — если, например, нет аппроксимации или сходимости).

Однако не следует думать, что наличие такой аппроксимации есть автоматическое следствие наличия аппроксимации производных, фигурирующих в уравнении. Как показывает пример Дюфора—Франкела, это, вообще говоря, не так.

Пример. Рассмотрим дискретизацию уравнения теплопроводности, задаваемую расчетными соотношениями

$$\frac{u_{i,j+1}-u_{i,j-1}}{2\tau}=c^2\frac{u_{i+1,j}-u_{i,j+1}-u_{i,j-1}+u_{i-1,j}}{h^2}-f_{ij}.$$

◀ Для невязки получаем

$$\Delta_{\tau,h}(z)=c^2\left[\frac{\tau^2}{2h^2}z_{tt}+O\left(\frac{\tau^4}{h^2}\right)+O(h)\right].$$

Отсюда видно, что эта дискретизация аппроксимирует исходное уравнение только, если $\tau\to 0$ быстрее, чем $h\to 0$,

$$\lim_{\tau,h\to 0}\frac{\tau}{h}=0.$$

В противном случае либо аппроксимации нет вообще (при $\frac{h}{\tau}\to 0$), либо эта расчетная схема аппроксимирует гиперболическое уравнение (при $\frac{\tau}{h}\to\alpha=\text{const}$)

$$u_t+c^2\alpha^2u_{tt}=c^2u_{xx}+f.\ \blacktriangleright$$

Назовем *шаблоном* расчетных соотношений совокупность узлов сетки, которые фигурируют в расчетных соотношениях. Так, например, для дискретного аналога волнового уравнения, полученного выше, шаблоном служат пять узлов с номерами $(i-1,j)$, (i,j), $(i,j+1)$, $(i,j-1)$, $(i+1,j)$ — это так называемый шаблон-крест (рис. 5 а). Для приведенных дискретизаций уравнения теплопроводности шаблонами расчетных соотношений служат наборы $(i-1,j)$, (i,j), $(i,j-1)$, $(i+1,j)$ — Т-образный шаблон (рис. 5 б) и $(i,j+1)$, (i,j), $(i,j-1)$, $(i+1,j)$, $(i-1,j)$ — крест (рис. 5 а) соответственно.

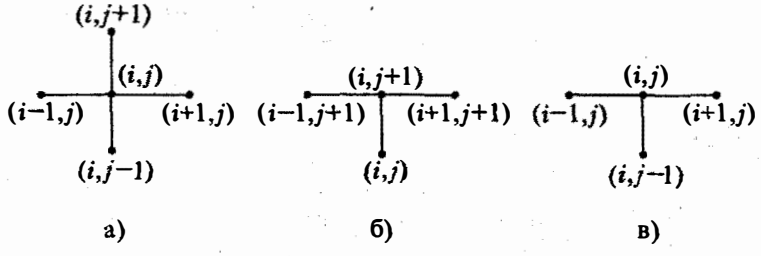


Рис. 5. Шаблоны расчетных соотношений

Конечно, возможно использование аналогов, связанных и с иными шаблонами. Так, однопараметрическое семейство схем, задаваемых соотношением

$$\frac{u_{i,j+1}-u_{ij}}{\tau}=c^2\frac{\lambda[u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}]+(1-\lambda)[u_{i+1,j}-2u_{ij}+u_{i-1,j}]}{h^2}+f_{ij},$$

регулируется параметром λ и при каждом его значении дает некоторые конкретные расчетные соотношения для уравнения теплопроводности. Шаблон этих схем (при $\lambda\neq 0,\neq 1$) — шеститочечная совокупность узлов $(i-1,j)$, (i,j), $(i+1,j)$, $(i-1,j+1)$, $(i,j+1)$, $(i+1,j+1)$ (рис. 6).

Одним из распространенных методов построения расчетных соотношений является метод неопределенных коэффициентов — сеточный аналог уравнения отыскивают, требуя, чтобы он на заданном шаблоне аппроксимировал исходное.

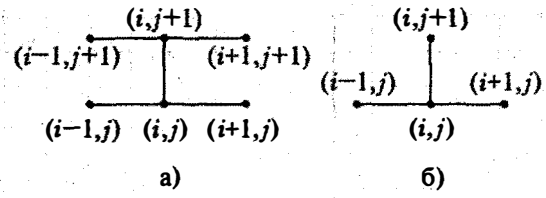


Рис. 6. Шаблоны расчетных соотношений

Проиллюстрируем этот метод на примере уравнения Пуассона. Выберем некоторый шаблон, например, крест $(i, j+1)$, (i, j), $(i, j-1)$, $(i+1, j)$, $(i-1, j)$, и положим

$$z_{xx,ij} + z_{yy,ij} \approx \alpha_{-10} z_{i-1,j} + \alpha_{00} z_{ij} + \alpha_{10} z_{i+1,j} + \alpha_{0-1} z_{i,j-1} + \alpha_{01} z_{i,j+1}.$$

Предполагая теперь, что функция $z(x, y)$ — достаточно гладкая, для невязки получим

$$\Delta_{\tau,h}(z) = z_{ij}\Big(\sum \alpha_{sl}\Big) + h z_x(\alpha_{10} - \alpha_{-10}) + \frac{h^2}{2} z_{xx}(\alpha_{10} + \alpha_{-10}) - z_{xx} +$$

$$+ \tau z_\tau(\alpha_{01} - \alpha_{0-1}) + \frac{\tau^2}{2} z_{\tau\tau}(\alpha_{01} + \alpha_{0-1}) - z_{\tau\tau} + O(h^2).$$

Чтобы имела место аппроксимация, для коэффициентов α_{sl} должно выполняться условие

$$\begin{cases} \sum \alpha_{sl} = 0, \\ \alpha_{10} - \alpha_{-10} = 0, \\ \alpha_{10} + \alpha_{-10} = \dfrac{2}{h^2}, \\ \alpha_{01} - \alpha_{0-1} = 0, \\ \alpha_{01} + \alpha_{0-1} = \dfrac{2}{\tau^2}, \end{cases}$$

откуда возникает искомая схема

$$\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\tau^2} + \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} = -f_{ij},$$

аппроксимирующая уравнение Пуассона со вторым порядком.

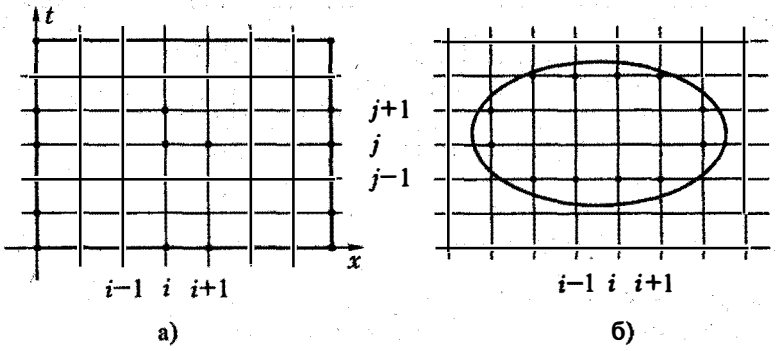
3.2. Дискретизация граничных условий

Как уже было отмечено выше, построение дискретного аналога дифференциальной *задачи для уравнения в частных производных* наряду с дискретизацией собственно уравнения предусматривает дискретизацию начальных и граничных условий.

Рассмотрим для примера начально-граничную задачу для уравнения теплопроводности

$$\begin{cases} z_t = c^2 z_{xx} + f, \\ z(x, 0) = \varphi(x), \quad 0 \leqslant x \leqslant l, \\ z(0, t) = \mu(t), \quad z(l, t) = \nu(t), \quad 0 \leqslant t \leqslant T. \end{cases}$$

Прямоугольная (h, τ)-сетка задает естественную дискретизацию границы (рис. 7 а), на которой начальную $\varphi(x)$ и граничные $\mu(t)$ и $\nu(t)$ функции заменяют некоторыми



а) б)

Рис. 7. Дискретизация граничных условий

сеточными функциями, а когда сетка естественно укладывается на границу области — просто их сеточными аналогами. При этом в качестве дискретного аналога рассматриваемой задачи получаем, например[6], задачу:

найти сеточную функцию $U^{NM} = (u_{ij})$, *удовлетворяющую уравнениям*

$$\frac{u_{i,j+1} - u_{ij}}{\tau} = c^2 \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} + f_{ij}$$

во всех внутренних узлах сетки $i = 1, 2, \ldots, N-1$, $j = 1, 2, \ldots, M$, *и принимающую в граничных узлах значения*

$$u_{0j} = \mu_j, \quad u_{Nj} = \nu_j, \quad j = 0, 1, \ldots, M,$$

$$u_{i,0} = \varphi_i, \quad i = 0, 1, \ldots, N.$$

В случае начально-граничной задачи для волнового уравнения дискретизация первого начального ($z(x, 0) = \varphi(x)$) и граничных условий и присоединение их к расчетным соотношениям, полученным дискретизацией уравнения, осуществляются аналогично. Второе граничное условие ($z_t(x, 0) = \psi(x)$) может быть дискретизировано, например, так: принимая во внимание, что с точностью до бесконечно малых более высокого порядка

$$z_{i1} \approx z_{i0} + z_{t,i0}\tau,$$

положим для искомой сеточной функции

$$u_{i1} = \varphi_i + \tau\psi_i, \quad \varphi_i = \varphi(x_i), \quad \psi_i = \psi(x_i).$$

Присоединяя эти соотношения к расчетным, получим дискретизацию задачи.

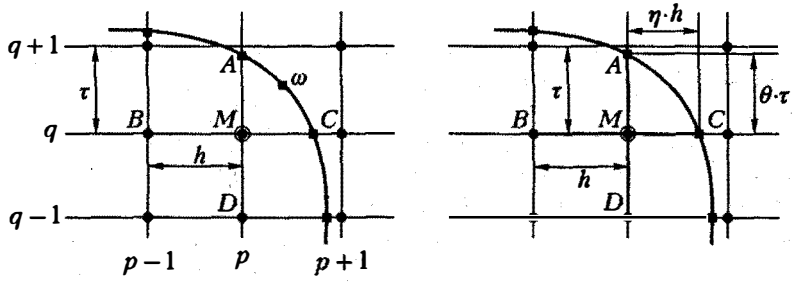
Несколько более сложная ситуация складывается с дискретизацией граничных условий, когда сетка не ложится на границу области (рис. 7 б). Так будет, например, в случае задачи Дирихле для уравнения Пуассона в области Ω

$$\begin{cases} z_{xx} + z_{yy} = -f(x, y), \\ z(x, y)\big|_{\partial\Omega} = \varphi(\omega). \end{cases}$$

[6] При другом выборе дискретизирующих уравнение соотношений получится другая дискретизация задачи.

В каждом внутреннем узле сетки мы можем записать расчетные соотношения, полученные дискретизацией уравнения. К ним следует присоединить еще соотношения, связывающие значения искомой функции в граничных узлах со значениями во внутренних. Это можно сделать несколькими различными способами.

Во-первых, можно просто перенести значения функции с границы $\partial\Omega$ в ближайшие граничные узлы, полагая, например, в граничном узле $u(M) = \varphi(\omega)$, где в качестве ω может быть взята любая точка на дуге границы $\partial\Omega$ между точками A и C (рис. 8 а). В этом случае система расчетных соотношений пополняется граничными условиями $u(M) = u_{pq} = \varphi_{pq}$, которые ее и замыкают, делая количество неизвестных равным количеству уравнений.



а) б)

Рис. 8. Дискретизация граничных условий

Во-вторых, можно значение искомой функции в граничной точке M проинтерполировать. Линейная интерполяция по горизонтальным узлам (B — внутреннему и C — лежащему на границе) дает в точке M

$$u(M) = \frac{\eta}{1+\eta}u(B) + \frac{1}{1+\eta}u(C).$$

Здесь $0 < \eta \leqslant 1$ — коэффициент, характеризующий расстояние от границы до точки M: $|C - M| = \eta h$. Аналогично, интерполяция по вертикальным узлам (D — внутреннему и A — лежащему на границе) дает в точке M

$$u(M) = \frac{\theta}{1+\theta}u(A) + \frac{1}{1+\theta}u(D), \quad |A - M| = \theta\tau.$$

Возможны и другие интерполяционные соотношения. При таком подходе система расчетных соотношений пополняется равенствами типа

$$u_{pq} = \lambda_A\varphi(A) + \lambda_B u_{p-1,q} + \lambda_C\varphi(C) + \lambda_D u_{p,q-1}$$

с некоторыми коэффициентами λ, которые ее и замыкают.

В-третьих, можно распространить расчетные соотношения, полученные дискретизацией уравнения, на граничные узлы. Правда, их приходится несколько модифицировать, учитывая что расчетный шаблон, (в рассматриваемой ситуации пятиточечный «крест»), искажен границей (рис. 8 б). Для этого достаточно воспользоваться формулами численного дифференцирования на неравномерной сетке. Полагая, например, (см. гл. LXII, § 3) в граничных узлах

$$z_{xx,pq} = \frac{2}{h^2}\left[\frac{z_{p-1,q}}{1+\eta} - \frac{z_{pq}}{\eta} + \frac{\varphi(C)}{\eta(1+\eta)}\right], \quad z_{yy,pq} = \frac{2}{\tau^2}\left[\frac{z_{p,q-1}}{1+\theta} - \frac{z_{pq}}{\theta} + \frac{\varphi(A)}{\theta(1+\theta)}\right],$$

дополним расчетные соотношения

$$\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\tau^2} + \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} = -f_{ij},$$

полученные дискретизацией уравнения Пуассона и справедливые во всех внутренних узлах сетки, указанными выше, справедливыми в граничных узлах. Тем самым, система расчетных соотношений замкнется.

§ 4. Устойчивость. Сходимость. Решение сеточных задач

Следующий этап — исследование полученных дискретных аналогов рассматриваемых дифференциальных задач и изучение степени близости решений этих задач к точным.

Как мы уже отмечали в случае обыкновенных дифференциальных уравнений (гл. LXII, § 3–4), здесь важно, чтобы построенная дискретная задача была устойчивой относительно возмущений правых частей и начальных условий. Дело в том, что решение сеточной задачи $U = u_{ij}$ и точное решение дифференциальной задачи $Z = z_{ij}$ являются решениями похожих дискретных уравнений — при наличии *аппроксимации* правые части этих уравнений отличаются тем меньше, чем мельче сетка. Следовательно, чтобы решение сеточной задачи мало отличалось от точного, нужно, чтобы невязка не вызывала больших возмущений решения сеточного аналога.

Подробнее, пусть сеточный аналог исследуемой задачи записан в виде

$$LU = F, \tag{1}$$

где $U = u_{ij}$ — искомая сеточная функция, L — сеточный аналог дифференциального оператора, включающий в себя начальные и граничные условия, и F — правая часть сеточного аналога. Если $Z = z_{ij}$ — сеточный аналог точного решения, то

$$LZ = F + \delta_{z,h\tau}.$$

Наличие аппроксимации означает стремление невязки к нулю ($\delta_{z,h\tau} \to 0$), а сходимость — стремление к нулю разности ΔZ ($\Delta Z = Z - U \to 0$) при $h, \tau \to 0$. Для рассматриваемых задач математической физики оператор L линеен и потому

$$L(Z - U) = LZ - LU = \delta_{z,h\tau} \quad \Longrightarrow \quad L(\Delta Z) = \delta_{z,h\tau}. \tag{2}$$

Отсюда ясно, что для сходимости решения дискретного аналога изучаемой задачи при измельчении сетки ($h, \tau \to 0$) к точному нужно, чтобы, во-первых, была малой невязка $\delta_{z,h\tau}$, т. е. имела место уже упоминавшаяся аппроксимация, а во-вторых, чтобы при малой невязке (являющейся правой частью рассматриваемого (1)–(2) сеточного уравнения) решение этого уравнения тоже было малым. Последнее — внутреннее свойство сеточного аналога — и есть требование *устойчивости*.

Замечание. В случае обыкновенных дифференциальных уравнений мы видели, что одной аппроксимации для сходимости недостаточно. Отметим, что одной устойчивости тоже недостаточно. Рассмотренная выше схема Дюфора—Фраикела аппроксимирует уравнение теплопроводности не при любых соотношениях между шагами сетки (h, τ), однако можно показать, что устойчивостью она обладает всегда.

Отсутствие устойчивости делает расчетные соотношения непригодными для вычислений, и нужно еще на этапе построения дискретных аналогов быть уверенным в их

эффективности. В вычислительной практике разработаны и широко используются приемы установления наличия или отсутствия устойчивости разностных схем, важнейшими из которых являются так называемые *спектральные признаки устойчивости*. Аккуратная их формулировка требует формализации использованных выше понятий «малости» сеточных функций и их «близости».

Остановимся на процедуре решения сеточных задач. В случае начально-граничной задачи для уравнения теплопроводности выше был получен следующий дискретный аналог

$$\frac{u_{i,j+1}-u_{ij}}{\tau}=c^2\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2}+f_{ij},$$

$$u_{0j}=\mu_j,\quad u_{Nj}=\nu_j,\quad j=0,1,\ldots,M,$$

$$u_{i,0}=\varphi_i,\quad i=0,1,\ldots,N.$$

Можно показать, что эта схема устойчива при условии

$$c^2\frac{\tau}{h^2}\leqslant\frac{1}{2}$$

и, следовательно, может быть использована для получения решения рассматриваемой задачи. Взятые расчетные соотношения используют четырехточечный $\perp$-образный шаблон, так что три значения функции U^{NM} на временно́м слое $t=t_j$ позволяют легко найти ее значение в центральном узле на следующем слое $t=t_j+1$ (рис. 6 б). Поскольку значения на нижнем слое и на боковых границах заданы, то, продвигаясь от слоя к слою, мы последовательно найдем все значения искомой сеточной функции.

Такие схемы, допускающие рекуррентное нахождение значений U^{NM}, называются *явными*.

Если же расчетные соотношения не дают возможности рекуррентно переходить от слоя к слою, то они называются *неявными*. Такая ситуация возникает, например, при использовании для решения этой же задачи расчетных соотношений

$$\frac{u_{i,j+1}-u_{ij}}{\tau}=c^2\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{h^2}+f_{ij}$$

на четырехточечном T-образном шаблоне. Эта схема безусловно (т. е. при любом соотношении между параметрами (h,τ)-сетки) устойчива. Полученная задача представляет собой систему линейных уравнений, которую можно решить, к примеру, методом прогонки.

Для этой же задачи дискретизация, порождаемая заменой временно́й производной центральным разностным отношением

$$\frac{u_{i,j+1}-u_{i,j-1}}{2\tau}=c^2\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2}+f_{ij},$$

оказывается безусловно неустойчивой и, следовательно, для получения решения непригодной.

В случае начально-краевой задачи для волнового уравнения расчетные соотношения

$$\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{\tau^2}=c^2\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2}+f_{ij},$$

дополненные начальными и граничными условиями, порождают схему, устойчивую лишь при выполнении условия

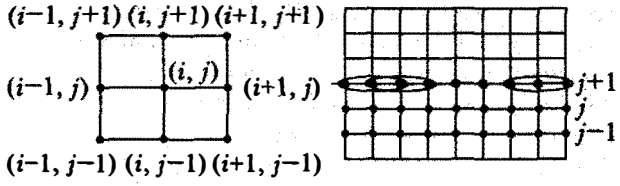
$$c\frac{\tau}{h}\leqslant\tau<1.$$

Эта схема использует шаблон-крест и является трехслойной. Схема — явная. Начальные и граничные условия позволяют реккурентно определить значения искомой сеточной функции на слое с номером $j+1$, используя значения на двух нижних слоях j и $j-1$.

Для этой же задачи может быть предложена безусловно устойчивая расчетная схема, получающаяся аппроксимацией второй пространственной производной линейной комбинацией вторых разностных отношений на слоях с номерами $j-1$, j и $j+1$:

$$u_{xx,ij} \approx \frac{1}{4}\left[\left(\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2}\right)+2\left(\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2}\right)+\right.$$
$$\left.+\left(\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{h^2}\right)\right].$$

Эти расчетные соотношения используют девятиточечный шаблон-ящик (рис. 9 а). Схема неявная, трехслойная. Эффективным методом решения полученной системы уравнений является метод прогонки — при известных на двух нижних слоях значениях искомой функции, расчетные соотношения образуют на $(j+1)$-м слое трехдиагональную систему линейных уравнений (рис. 9 б).



а) б)

Рис. 9. Шаблон-ящик для волнового уравнения

В случае задачи Дирихле для уравнения Пуассона мы имеем в каждом внутреннем узле расчетные соотношения

$$\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2}+\frac{u_{i,j-1}-2u_{ij}+u_{i,j+1}}{\tau^2}=-f_{ij},$$

к которым добавляются соотношения в граничных узлах, полученные одним из указанных выше способов дискретизации граничных условий. Количество полученных линейных уравнений совпадает с количеством неизвестных компонент отыскиваемой сеточной функции. Однако эта неявная схема для непрямоугольных областей Ω приводит к системе линейных уравнений, разрешить которую оказываются довольно трудно, особенно в случае малых h и τ. Здесь с успехом могут быть использованы итерационные методы решения систем. Однако более эффективным оказывается подход, использующий идею *установления*, когда решение задачи Дирихле рассматривается как установившееся решение двумерного уравнения теплопроводности. Вместо задачи Дирихле решают краевую задачу для уравнения теплопроводности до того момента T, когда его эволюция во времени становится пренебрежимо малой, и принимают полученное решение за решение исследуемой задачи.

В заключение отметим, что в случае краевой задачи для уравнений эллиптического типа, как и в случае краевых задач для обыкновенных дифференциальных уравнений, возможно использование вариационных соображений, основанных на том факте, что

условие стационарности интегрального функционала описывается краевой задачей для уравнения Эйлера—Остроградского.

Для рассмотренной выше задачи таким функционалом является функционал, задаваемый соотношением

$$F(z) = \frac{1}{2} \iint\limits_{\Omega} L(x, y, z, z_x, z_y)\, dx\, dy = \frac{1}{2} \iint\limits_{\Omega} \left(z_x^2 + z_y^2 - 2zf(x, y)\right) dx\, dy.$$

Уравнение Эйлера—Остроградского для него имеет вид

$$L_z - (L_{z_x})_x - (L_{z_y})_y = 0 \quad \Longrightarrow \quad z_{xx} + z_{yy} = -f,$$

и задача поиска стационарных точек указанного функционала при условии

$$z(x, y)\big|_{\partial\Omega} = \varphi(\omega)$$

совпадает с задачей Дирихле для уравнения Пуассона.

ТЕОРИЯ СПЛАЙНОВ

Кривые и поверхности, встречающиеся в практических задачах, часто имеют довольно сложную форму, не допускающую универсального аналитического задания в целом при помощи элементарных функций. Поэтому их собирают из сравнительно простых гладких фрагментов — отрезков (кривых) или вырезков (поверхностей), каждый из которых может быть вполне удовлетворительно описан при помощи элементарных функций одной или двух переменных. При этом вполне естественно потребовать, чтобы гладкие функции, которые используются для построения частичных кривых или поверхностей, имели схожую природу, например, были бы многочленами одинаковой степени. А чтобы получающаяся в результате кривая или поверхность оказалась достаточно гладкой, необходимо быть особенно внимательным в местах стыковки соответствующих фрагментов. Степень многочленов выбирается из простых геометрических соображений и, как правило, невелика. Для гладкого изменения касательной вдоль всей составной кривой достаточно описывать стыкуемые кривые при помощи многочленов третьей степени, кубических многочленов. Коэффициенты таких многочленов всегда можно подобрать так, чтобы кривизна соответствующей составной кривой была непрерывной.

Кубические сплайны, возникающие при решении одномерных задач, можно приспособить к построению фрагментов составных поверхностей. И здесь вполне естественно появляются бикубические сплайны, описываемые при помощи многочленов третьей степени по каждой из двух переменных. Работа с такими сплайнами требует уже значительно большего объема вычислений. Но правильно организованный процесс позволит учесть непрерывно нарастающие возможности вычислительной техники в максимальной степени.

Глава LXV

СПЛАЙНЫ

§ 1. Сплайн-функции

Рис. 1

Пусть на отрезке $[a, b]$ задана *сетка* ω

$$a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b. \tag{1}$$

Точки x_0 и x_m называются *граничными узлами* сетки ω, а точки $x_1, \ldots, x_{m-1}$ — ее *внутренними узлами* (рис. 1). Сетка называется *равномерной*, если расстояния между любыми двумя соседними узлами одинаковы.

Функция $S(x)$, заданная на отрезке $[a, b]$, называется *сплайном порядка* $p+1$ (*степени* p), если эта функция

1) на каждом из отрезков

$$\Delta_i = [x_i, x_{i+1}], \quad i = 0, 1, \ldots, m-1,$$

является многочленом заданной степени $p \geqslant 2$, то есть может быть записана в виде

$$S(x) = S_i(x) = \sum_{k=0}^{p} a_k^{(i)} (x - x_i)^k, \quad i = 0, 1, \ldots, m-1,$$

и

2) $p-1$ раз непрерывно дифференцируема на отрезке $[a, b]$, то есть

$$S(x) \in C^{p-1}[a, b].$$

Замечание. Индекс (i) у чисел $a_k^{(i)}$ указывает на то, что набор коэффициентов, которым определяется функция $S(x)$, на каждом частичном отрезке Δ_i свой.

На каждом из отрезков Δ_i, $i = 0, 1, \ldots, m-1$, сплайн $S(x)$ является многочленом степени p и определяется на этом отрезке $p+1$ коэффициентом. Всего частичных отрезков — m. Значит, для того, чтобы полностью определить сплайн, необходимо найти $(p+1)m$ чисел

$$a_k^{(i)}, \quad k = 0, 1, \ldots, p, \quad i = 0, 1, \ldots, m-1.$$

Условие $S(x) \in C^{p-1}[a, b]$ означает непрерывность функции $S(x)$ и ее производных $S'(x), S''(x), \ldots, S^{p-1}(x)$ во всех внутренних узлах сетки ω. Число таких узлов $m - 1$. Тем самым, для отыскания коэффициентов всех многочленов получается $p(m - 1)$ условий (уравнений).

Для полного определения сплайна недостает $(p + 1)m - p(m - 1) = m + p$ условий (уравнений). Выбор дополнительных условий определяется характером рассматриваемой задачи, а иногда и просто — желанием пользователя.

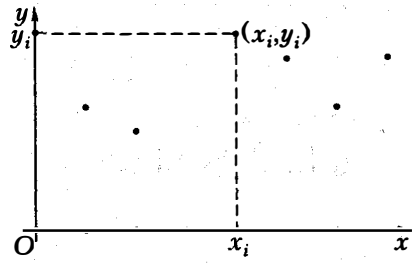


Рис. 2

Наиболее часто рассматриваются задачи интерполяции и сглаживания, когда требуется построить тот или иной сплайн по заданному массиву точек на плоскости (x_i, y_i), $i = 0, 1, \ldots, m$ (рис. 2).

В задачах *интерполяции* требуется, чтобы график сплайна проходил через точки (x_i, y_i), $i = 0, 1, \ldots, m$, что накладывает на его коэффициенты $m + 1$ дополнительных условий (уравнений). Остальные $p - 1$ условий (уравнений) для однозначного построения сплайна чаще всего задают в виде значений младших производных сплайна на концах рассматриваемого отрезка $[a, b]$ — *граничных (краевых) условий*. Возможность выбора различных граничных условий позволяет строить сплайны, обладающие самыми разными свойствами.

В задачах *сглаживания* сплайн строят так, чтобы его график проходил вблизи точек (x_i, y_i), $i = 0, 1, \ldots, m$, а не через них. Меру этой близости можно определять по-разному, что приводит к значительному разнообразию сглаживающих сплайнов.

Описанные возможности выбора при построении сплайн-функций далеко не исчерпывают всего их многообразия. И если первоначально рассматривались только кусочно полиномиальные сплайн-функции, то по мере расширения сферы их приложений стали возникать сплайны, «склеенные» и из других элементарных функций.

1.1. Интерполяционные кубические сплайны

Постановка задачи интерполяции

Пусть на отрезке $[a, b]$ задана сетка ω

$$a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b.$$

Рассмотрим набор чисел

$$y_0, y_1, \cdots, y_{m-1}, y_m.$$

Задача. Построить гладкую на отрезке $[a, b]$ функцию $\sigma(x)$, которая принимает в узлах сетки ω заданные значения, то есть

$$\sigma(x_i) = y_i, \quad i = 0, 1, \ldots, m - 1, m.$$

Замечание. Сформулированная задача интерполяции состоит в восстановлении гладкой функции, заданной таблично (рис. 2). Ясно, что такая задача имеет множество различных решений. Накладывая на конструируемую функцию дополнительные условия, можно добиться необходимой однозначности.

В приложениях часто возникает необходимость приблизить функцию, заданную аналитически,

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

при помощи функции с предписанными достаточно хорошими свойствами. Например, в тех случаях, когда вычисление значений заданной функции $f(x)$ в точках отрезка $[a, b]$ связано со значительными трудностями и/или заданная функция $f(x)$ не обладает требуемой гладкостью, удобно воспользоваться другой функцией, которая достаточно хорошо приближала бы заданную функцию и была лишена отмеченных ее недостатков.

Задача интерполяции функции. Построить на отрезке $[a, b]$ гладкую функцию $\sigma(x)$, совпадающую в узлах сетки ω с заданной функцией $f(x)$.

Определение интерполяционного кубического сплайна

Интерполяционным кубическим сплайном $S(x)$ на сетке ω называется функция, которая

1) на каждом из отрезков

$$[x_i, x_{i+1}], \quad i = 0, 1, \ldots, m-1,$$

представляет собой многочлен третьей степени,

$$S(x) = S_i(x) = a_0^{(i)} + a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3,$$

2) дважды непрерывно дифференцируема на отрезке $[a, b]$, то есть принадлежит классу $C^2[a, b]$, и

3) удовлетворяет условиям

$$S(x_i) = y_i, \quad i = 0, 1, \ldots, m. \tag{2}$$

На каждом из отрезков $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, m-1$, сплайн $S(x)$ является многочленом третьей степени и определяется на этом отрезке четырьмя коэффициентами. Всего отрезков — m. Значит, для того, чтобы полностью определить сплайн, необходимо найти $4m$ чисел

$$a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, \quad i = 0, 1, \ldots, m-1.$$

Условие

$$S(x) \in C^2[a, b]$$

означает непрерывность функции $S(x)$ и ее производных $S'(x)$ и $S''(x)$ во всех внутренних узлах сетки ω. Число таких узлов — $m - 1$. Тем самым, для отыскания коэффициентов всех многочленов получается еще $3(m-1)$ условий (уравнений). Вместе с условиями (2) получается

$$3(m-1) + (m+1) = 4m - 2$$

условия (уравнения).

Граничные (краевые) условия

Два недостающих условия задаются в виде ограничений на значения сплайна и/или его производных на концах промежутка $[a, b]$. При построении интерполяционного кубического сплайна наиболее часто используются краевые условия следующих четырех типов.

А. Краевые условия 1-го типа.

$$S'(a) = f'(a), \quad S'(b) = f'(b)$$

— на концах промежутка $[a, b]$ задаются значения первой производной искомой функции.

Б. Краевые условия 2-го типа.

$$S''(a) = f''(a), \quad S''(b) = f''(b)$$

— на концах промежутка $[a, b]$ задаются значения второй производной искомой функции.

В. Краевые условия 3-го типа.

$$S'(a) = S'(b), \quad S''(a) = S''(b)$$

называются *периодическими*. Выполнения этих условий естественно требовать в тех случаях, когда интерполируемая функция является периодической с периодом $T = b - a$.

Г. Краевые условия 4-го типа.

$$S'''(y, x_1 - 0) = S'''(y, x_1 + 0), \quad S'''(y, x_{m-1} - 0) = S'''(y, x_{m-1} + 0)$$

требуют особого комментария.

Комментарий. Во внутренних узлах сетки третья производная функции $S(x)$, вообще говоря, разрывна. Однако число разрывов третьей производной можно уменьшить при помощи условий 4-го типа. В этом случае построенный сплайн будет трижды непрерывно дифференцируем на промежутках $[x_0, x_2]$ и $[x_{m-1}, x_m]$.

Построение интерполяционного кубического сплайна

Опишем способ вычисления коэффициентов кубического сплайна, при котором число величин, подлежащих определению, равно $m + 1$, а не $4m$.

На каждом из промежутков

$$[x_i, x_{i+1}], \quad i = 0, 1, \ldots, m - 1,$$

интерполяционная сплайн-функция ищется в следующем виде

$$\boxed{S(x) = S_i(x) = y_i(1-t)^2(1+2t) + y_{i+1}t^2(3-2t) + n_i h_i t(1-t)^2 - n_i t^2(1-t).} \quad (3)$$

Здесь

$$h_i = x_{i+1} - x_i, \quad t = \frac{x - x_i}{h_i},$$

а числа n_i, $i = 0, 1, \ldots, m$, являются решением системы линейных алгебраических уравнений, вид которой зависит от типа краевых условий.

Для **краевых условий 1-го и 2-го типов** эта система имеет следующий вид

$$\begin{cases} 2n_0 + \mu_0^* n_1 = c_0^*, \\ \lambda_i n_{i-1} + 2n_i + \mu_i n_{i+1} = c_i, \quad i = 1, 2, \ldots, m-1, \\ \lambda_m^* n_{m-1} + 2n_m = c_m^*, \end{cases}$$

где

$$c_i = 3\left(\mu_i \frac{y_{i+1} - y_i}{h_i} + \lambda_i \frac{y_i - y_{i-1}}{h_{i-1}}\right).$$

Коэффициенты μ_0^*, c_0^*, λ_m^*, c_m^* зависят от выбора краевых условий.

Краевые условия 1-го типа:

$$\mu_0^* = 0, \quad c_0^* = 2y_0',$$
$$\lambda_m^* = 0, \quad c_m^* = 2y_m'.$$

Краевые условия 2-го типа:

$$\mu_0^* = 1, \quad c_0^* = 3\frac{y_1 - y_0}{h_0} - \frac{h_0}{2}y_0'',$$

$$\lambda_m^* = 1, \quad c_m^* = 3\frac{y_m - y_{m-1}}{h_{m-1}} + \frac{h_{m-1}}{2}y_m''.$$

В случае **краевых условий 3-го типа** система для определения чисел n_i, $i = 1, 2, \ldots, m$, записывается так

$$\begin{cases} 2n_1 + \mu_1 n_2 + \lambda_1 n_m = c_1, \\ \lambda_i n_{i-1} + 2n_i + \mu_i n_{i+1} = c_i, \quad i = 2, \ldots, m-1, \\ \mu_m n_1 + \lambda_m n_{m-1} + 2n_m = c_m. \end{cases}$$

Число неизвестных в последней системе равно m, так как из условия периодичности вытекает, что $n_0 = n_m$.

Для **краевых условий 4-го типа** система для определения чисел n_i, $i = 1, 2, \ldots, m-1$, имеет вид

$$\begin{cases} (1+\gamma_0)n_1 + \gamma_0 n_2 = c_1^*, \\ \lambda_i n_{i-1} + 2n_i + \mu_i n_{i+1} = c_i, \quad i = 2, \ldots, m-2, \\ \gamma_m n_{m-2} + (1+\gamma_m)n_{m-1} = c_{m-1}^*, \end{cases}$$

где

$$\gamma_0 = \frac{h_0}{h_1}, \qquad c_1^* = \frac{1}{3}c_1 + 2\gamma_0\frac{y_2 - y_1}{h_1},$$

$$\gamma_m = \frac{h_{m-1}}{h_m}, \quad c_{m-1}^* = \frac{1}{3}c_{m-1} + 2\gamma_m\frac{y_{m-1} - y_{m-2}}{h_{m-2}}.$$

По найденному решению системы числа n_0 и n_m можно определить при помощи формул

$$n_0 = 2\left(\frac{y_1 - y_0}{h_0} - \gamma_0^2\frac{y_2 - y_1}{h_1}\right) - (1-\gamma_0^2)n_1 + \gamma_0^2 n_2,$$

$$n_m = 2\left(\frac{y_m - y_{m-1}}{h_{m-1}} - \gamma_m^2\frac{y_{m-1} - y_{m-2}}{h_{m-2}}\right) - (1-\gamma_m^2)n_{m-1} + \gamma_m^2 n_{m-2}.$$

Важное замечание. Матрицы всех трех линейных алгебраических систем являются матрицами с диагональным преобладанием. Такие матрицы невырождены, и потому каждая из этих систем имеет единственное решение.

Теорема. *Интерполяционный кубический сплайн, удовлетворяющий условиям* (2) *и краевому условию одного из перечисленных четырех типов, существует и единствен.*

Таким образом, построить интерполяционный кубический сплайн — это значит найти его коэффициенты $n_0, n_1, \ldots, n_m$.

Когда коэффициенты сплайна найдены, значение сплайна $S(x)$ в произвольной точке отрезка $[a, b]$ можно найти по формуле (3). Однако для практических вычислений больше подходит следующий алгоритм нахождения величины $S(x)$.

Пусть $x \in [x_i, x_{i+1}]$. Сначала вычисляются величины A и B по формулам

$$\boxed{\begin{aligned} A &= -2\frac{y_{i+1} - y_i}{x_{i+1} - x_i} + n_i + n_{i+1}, \\ B &= -A + \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - n_i, \end{aligned}}$$

а затем находится величина $S(x)$:

$$\boxed{S(x) = y_i + (x - x_i)[n_i + t(B + tA)],}$$

где, как обычно,

$$t = \frac{x - x_i}{x_{i+1} - x_i}.$$

Применение этого алгоритма существенно сокращает вычислительные затраты на определение величины $S(x)$.

Советы пользователю

Выбор граничных (краевых) условий и узлов интерполяции позволяет в известной степени управлять свойствами интерполяционных сплайнов.

А. Выбор граничных (краевых) условий.

Выбор граничных условий является одной из центральных проблем при интерполяции функций. Он приобретает особую важность в том случае, когда необходимо обеспечить высокую точность аппроксимации функции $f(x)$ сплайном $S(x)$ вблизи концов отрезка $[a, b]$. Граничные значения оказывают заметное влияние на поведение сплайна $S(x)$ вблизи точек a и b, и это влияние по мере удаления от них быстро ослабевает.

Выбор граничных условий часто определяется наличием дополнительных сведений о поведении аппроксимируемой функции $f(x)$.

Если на концах отрезка $[a, b]$ известны значения первой производной $f'(x)$, то естественно воспользоваться краевыми условиями 1-го типа.

Если на концах отрезка $[a, b]$ известны значения второй производной $f''(x)$, то естественно воспользоваться краевыми условиями 2-го типа.

Если есть возможность выбора между краевыми условиями 1-го и 2-го типов, то предпочтение следует отдать условиям 1-го типа.

Если $f(x)$ — периодическая функция, то следует остановиться на краевых условиях 3-го типа.

В случае, если никакой дополнительной информации о поведении аппроксимируемой функции нет, часто используют так называемые естественные граничные условия

$$S''(a) = 0, \quad S''(b) = 0.$$

Однако следует иметь ввиду, что при таком выборе граничных условий точность аппроксимации функции $f(x)$ сплайном $S(x)$ вблизи концов отрезка $[a, b]$ резко снижается. Иногда используются краевые условия 1-го или 2-го типа, но не с точными

значениями соответствующих производных, а с их разностными аппроксимациями. Точность такого подхода невысока.

Практический опыт расчетов показывает, что в рассматриваемой ситуации наиболее целесообразным является выбор граничных условий 4-го типа.

Б. Выбор узлов интерполяции.

Если третья производная $f'''(x)$ функции терпит разрыв в некоторых точках отрезка $[a, b]$, то для улучшения качества аппроксимации эти точки следует включить в число узлов интерполяции.

Если разрывна вторая производная $f''(x)$, то для того, чтобы избежать осцилляции сплайна вблизи точек разрыва, необходимо принять специальные меры. Обычно узлы интерполяции выбирают так, чтобы точки разрыва второй производной попадали внутрь промежутка $[x_i, x_{i+1}]$, такого, что

$$h_i = \alpha \min\{h_{i-1}, h_{i+1}\},$$

где $\alpha \ll 1$. Величину α можно выбрать путем численного эксперимента (часто достаточно положить $\alpha = 0{,}01$).

Существует набор рецептов по преодолению трудностей, возникающих при разрывной первой производной $f'(x)$. В качестве одного из самых простых можно предложить такой: разбить отрезок аппроксимации на промежутки, где производная непрерывна, и на каждом из этих промежутков построить сплайн.

Выбор интерполяционной функции (плюсы и минусы)

Подход 1-й. Интерполяционный многочлен Лагранжа

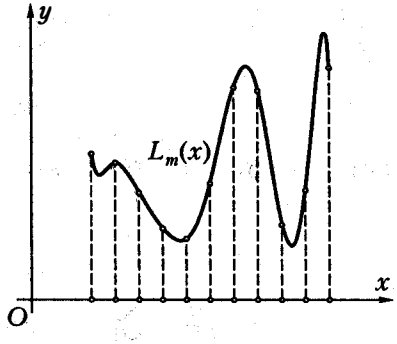


Рис. 3

По заданному массиву

$$(x_i, y_i), \quad i = 0, 1, \ldots, m-1, m$$

(рис. 3) *интерполяционный многочлен Лагранжа* определяется формулой

$$L_m(x) = \sum_{i=0}^{m} y_i \frac{\varphi_m(x)}{(x - x_i)\varphi'_m(x_i)},$$

где

$$\varphi_m(x) = \prod_{i=0}^{m} (x - x_i).$$

Свойства интерполяционного многочлена Лагранжа целесообразно рассматривать с двух противоположных позиций, обсуждая основные достоинства отдельно от недостатков.

Основные достоинства 1-го подхода:

1) график интерполяционного многочлена Лагранжа проходит через каждую точку массива,
2) конструируемая функция легко описывается (число подлежащих определению коэффициентов интерполяционного многочлена Лагранжа на сетке ω равно $m + 1$),
3) построенная функция имеет непрерывные производные любого порядка,
4) заданным массивом интерполяционный многочлен определен однозначно.

Основные недостатки 1-го подхода:

1) степень интерполяционного многочлена Лагранжа зависит от числа узлов сетки, и чем больше это число, тем выше степень интерполяционного многочлена и, значит, тем больше требуется вычислений,

2) изменение хотя бы одной точки в массиве требует полного пересчета коэффициентов интерполяционного многочлена Лагранжа,

3) добавление новой точки в массив увеличивает степень интерполяционного многочлена Лагранжа на единицу и также приводит к полному пересчету его коэффициентов,

4) при неограниченном измельчении сетки степень интерполяционного многочлена Лагранжа неограниченно возрастает.

Поведение интерполяционного многочлена Лагранжа при неограниченном измельчении сетки вообще требует особого внимания.

Комментарии

А. О приближении непрерывной функции многочленом.

Известно (Вейерштрасс, 1885 год), что всякая непрерывная (а тем более гладкая) на отрезке функция может быть как угодно хорошо приближена на этом отрезке многочленом.

Опишем этот факт на языке формул. Пусть $f(x)$ — функция, непрерывная на отрезке $[a, b]$. Тогда для любого $\varepsilon > 0$ найдется такой многочлен $P_n(x)$, что для любого x из промежутка $[a, b]$ будет выполняться неравенство (рис. 4)

$$|P_n(x) - f(x)| < \varepsilon.$$

Отметим, что многочленов даже одной степени, приближающих функцию $f(x)$ с указанной точностью, существует бесконечно много.

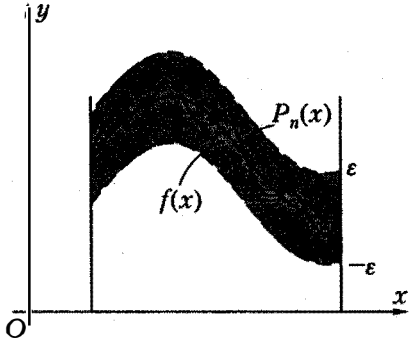


Рис. 4

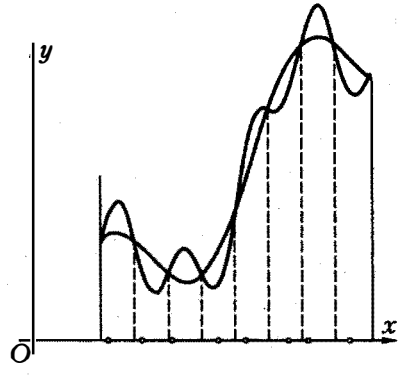


Рис. 5

Построим на отрезке $[a, b]$ сетку ω. Ясно, что ее узлы, вообще говоря, не совпадают с точками пересечения графиков многочлена $P_n(x)$ и функции $f(x)$ (рис. 5). Поэтому для взятой сетки многочлен $P_n(x)$ не является интерполяционным.

Б. Об интерполировании непрерывной функции.

При аппроксимации непрерывной функции интерполяционным многочленом Лагранжа его график не только не обязан быть близким графику функции $f(x)$ в каждой точке отрезка $[a, b]$, но может уклоняться от этой функции как угодно сильно. Приведем два примера.

Пример 1 (Рунге, 1901 год). При неограниченном увеличении числа узлов для функции

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

на отрезке $[-1, 1]$ выполняется предельное равенство (рис. 6)

$$\lim_{m \to \infty} \max_{-1 \leqslant x \leqslant 1} |f(x) - L_m(x)| = \infty. \blacktriangleright$$

Пример 2 (Бернштейн, 1912 год). Последовательность интерполяционных многочленов Лагранжа

$$\{L_m(x)\},$$

построенных на равномерных сетках ω_m для непрерывной функции $f(x) = |x|$ на отрезке $[-1, 1]$, с возрастанием числа узлов m *не стремится* к функции $f(x)$ (рис. 7). $\blacktriangleright$

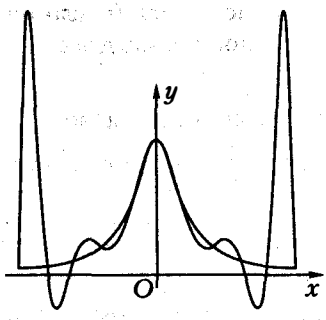


Рис. 6

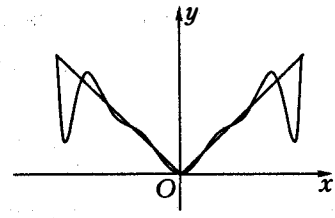


Рис. 7

Подход 2-й. Кусочно-линейная интерполяция

При отказе от гладкости интерполируемой функции соотношение между числом достоинств и числом недостатков можно заметно изменить в сторону первых.

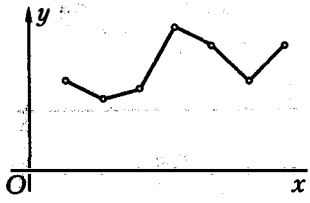


Рис. 8

Построим кусочно-линейную функцию путем последовательного соединения точек (x_i, y_i) прямолинейными отрезками (рис. 8).

Основные достоинства 2-го подхода:

1) график кусочно-линейной функции проходит через каждую точку массива,
2) конструируемая функция легко описывается (число подлежащих определению коэффициентов соответствующих линейных функций для сетки (1) равно $2m$),
3) заданным массивом построенная функция определена однозначно,
4) степень многочленов, используемых для описания интерполяционной функции, не зависит от числа узлов сетки (равна 1),
5) изменение одной точки в массиве требует вычисления четырех чисел (коэффициентов двух прямолинейных звеньев, исходящих из новой точки),
6) добавление дополнительной точки в массив требует вычисления четырех коэффициентов.

Кусочно-линейная функция достаточно хорошо ведет себя и при измельчении сетки.

Основной недостаток 2-го подхода:

аппроксимирующая кусочно-линейная функция не является гладкой: первые производные терпят разрыв в узлах сетки (*узлах интерполяции*).

Подход 3-й. Сплайн-интерполяция

Предложенные подходы можно объединить так, чтобы число перечисленных достоинств обоих подходов сохранилось при одновременном уменьшении числа недостатков. Это можно сделать путем построения гладкой интерполяционной сплайн-функции степени p.

Основные достоинства 3-го подхода:

1) график построенной функции проходит через каждую точку массива,
2) конструируемая функция сравнительно легко описывается (число подлежащих определению коэффициентов соответствующих многочленов для сетки (1) равно $m(p+1)$),
3) заданным массивом построенная функция определена однозначно,
4) степень многочленов не зависит от числа узлов сетки и, следовательно, не изменяется при его увеличении,
5) построенная функция имеет непрерывные производные до порядка $p-1$ включительно,
6) построенная функция обладает хорошими аппроксимационными свойствами.

Краткая справка. Предложенное название — *сплайн* — не является случайным — введенные нами гладкие кусочно-полиномиальные функции и чертежные сплайны тесно связаны.

Рассмотрим гибкую идеально тонкую линейку, проходящую через расположенные на плоскости (x, y) опорные точки массива (x_i, y_i), $i = 0, 1, \ldots, m-1, m$ (рис. 9). Согласно закону Бернулли—Эйлера линеаризованное уравнение изогнутой линейки имеет вид

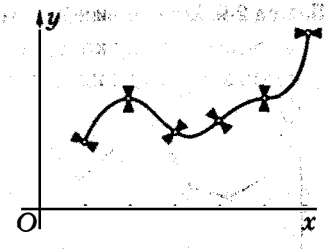


Рис. 9

$$EIS''(x) = -M(x),$$

где $S(x)$ — изгиб, $M(x)$ — *изменяющийся линейно* от опоры к опоре изгибающий момент, EI — жесткость линейки.

Функция $S(x)$, описывающая форму линейки, является многочленом третьей степени между каждыми двумя соседними точками массива (опорами) и дважды непрерывно дифференцируема на всем промежутке $[a, b]$.

Комментарий. Об интерполировании непрерывной функции

В отличие от интерполяционных многочленов Лагранжа, последовательность интерполяционных кубических сплайнов на равномерной сетке всегда сходится к интерполируемой непрерывной функции, причем с улучшением дифференциальных свойств этой функции скорость сходимости повышается.

Пример. Для функции

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

кубический сплайн на сетке с числом узлов $m = 6$ дает *погрешность аппроксимации* того же порядка, что и интерполяционный многочлен $L_5(x)$, а на сетке с числом узлов $m = 21$ эта погрешность настолько мала, что в масштабе обычного книжного рисунка просто не может быть показана (рис. 10) (интерполяционный многочлен $L_{20}(x)$ дает в этом случае погрешность около 10 000 %). ▶

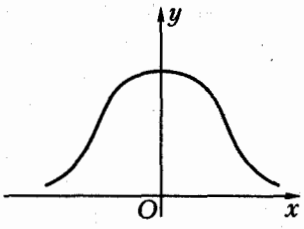


Рис. 10

Свойства интерполяционного кубического сплайна

А. Аппроксимационные свойства кубического сплайна.

Аппроксимационные свойства интерполяционного сплайна зависят от гладкости функции $f(x)$ — чем выше гладкость интерполируемой функции, тем выше порядок аппроксимации и при измельчении сетки тем выше скорость сходимости.

Если интерполируемая функция $f(x)$ непрерывна на отрезке $[a, b]$, то есть

$$f(x) \in C^0[a, b],$$

то

$$\boxed{\|f(x) - S(x)\|_C = \max_{x\in[a,b]} |f(x) - S(x)| \to 0}$$

при

$$h = \max_{0\leqslant i\leqslant N-1} h_i \to 0.$$

Если интерполируемая функция $f(x)$ имеет на отрезке $[a, b]$ непрерывную первую производную, то есть

$$f(x) \in C^1[a, b],$$

а $S(x)$ — интерполяционный сплайн, удовлетворяющий граничным условиям 1-го или 3-го типа, то при $h \to 0$ имеем

$$\boxed{\|f(x) - S(x)\|_C = o(h), \quad \|f'(x) - S'(x)\|_C = o(1).}$$

В этом случае не только сплайн сходится к интерполируемой функции, но и производная сплайна сходится к производной этой функции.

В случае, если

$$f(x) \in C^4[a, b],$$

сплайн $S(x)$ аппроксимирует на отрезке $[a, b]$ функцию $f(x)$, а его первая и вторая производные аппроксимируют соответственно функции $f'(x)$ и $f''(x)$,

$$\boxed{\begin{aligned} |f(x) - S(x)|_C &= o(h^4), \\ |f'(x) - S'(x)|_C &= o(h^3), \\ |f''(x) - S''(x)|_C &= o(h^2). \end{aligned}}$$

Б. Экстремальное свойство кубического сплайна.

Интерполяционный кубический сплайн обладает еще одним полезным свойством. Рассмотрим следующий пример.

Пример. Построить функцию $f(x)$, минимизирующую функционал

$$J(f) = \int_a^b (f''(x))^2\, dx$$

на классе функций из пространства $C^2[a, b]$, графики которых проходят через точки массива (x_i, y_i), $i = 0, 1, \ldots, m$.

◀ Среди всех функций, проходящих через опорные точки $(x_i, f(x_i))$ и принадлежащих указанному пространству, именно кубический сплайн $S(x)$, удовлетворяющий краевым условиям

$$S''(a) = S''(b),$$

доставляет экстремум (минимум) функционалу $J(f)$. ▶

Замечание 1. Часто именно это экстремальное свойство берут в качестве определения интерполяционного кубического сплайна.

Замечание 2. Интересно отметить, что интерполяционный кубический сплайн обладает описанным выше экстремальным свойством на очень широком классе функций, а именно, на классе $\overline{W}_2^2[a, b]$.

1.2. Сглаживающие кубические сплайны

О постановке задачи сглаживания

Пусть заданы сетка

$$a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b$$

и набор чисел

$$y_0, y_1, \ldots, y_{m-1}, y_m.$$

Комментарий к исходным данным

На практике часто приходится иметь дело со случаем, когда значения y_i в массиве

$$(x_i, y_i), \quad i = 0, 1, \ldots, m,$$

заданы с некоторой погрешностью. Фактически это означает, что для каждого $i = 0, 1, \ldots, m$, указан интервал

$$(c_i, d_i) \quad \text{или} \quad (y_i - \delta_i, y_i + \delta_i)$$

и любое число из этого интервала может быть взято в качестве значения y_i. Величины y_i удобно интерпретировать, например, как результаты измерений некоторой функции $y(x)$ при заданных значениях переменной x, содержащие случайную погрешность. При решении задачи восстановления функции по таким ее «экспериментальным» значениям вряд ли целесообразно использовать интерполяцию, поскольку интерполяционная функция будет послушно воспроизводить причудливые осцилляции, обусловленные случайной компонентой в массиве $\{y_i\}$. Более естественным является подход, основанный на процедуре сглаживания, призванной как-то уменьшить элемент случайности в результате измерений. Обычно в таких задачах требуется найти функцию, значения которой при $x = x_i$, $i = 0, 1, \ldots, m$, попадали бы в соответствующие интервалы и которая обладала бы, кроме того, достаточно хорошими свойствами. Например, имела бы непрерывные первые и вторые производные, или же ее график был бы не слишком сильно искривлен, то есть не имел бы сильных осцилляций.

Задача подобного рода возникает и тогда, когда по заданному (точно) массиву

$$(x_i, y_i), \quad i = 0, 1, \ldots, m,$$

требуется построить функцию, которая проходила бы не через заданные точки, а вблизи них и к тому же изменялась достаточно плавно. Другими словами, искомая функция как бы сглаживала заданный массив, а не интерполировала его.

Пусть заданы сетка ω

$$a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b$$

и два набора чисел

$$y_0, y_1, \cdots, y_{m-1}, y_m$$

и

$$\Delta_0 > 0, \quad \Delta_1 > 0, \quad \ldots, \quad \Delta_{m-1} > 0, \quad \Delta_m > 0.$$

Задача. Построить гладкую на отрезке $[a, b]$ функцию $\sigma(x)$, значения которой в узлах сетки ω отличались от чисел y_i на заданные величины $\Delta_i > 0$

$$|\sigma(x_i) - y_i| < \Delta_i, \quad i = 0, 1, \ldots, m-1, m.$$

Замечание. Сформулированная задача сглаживания состоит в восстановлении гладкой функции, заданной таблично. Ясно, что такая задача имеет множество различных решений. Накладывая на конструируемую функцию дополнительные условия, можно добиться необходимой однозначности.

Определение сглаживающего кубического сплайна

Сглаживающим кубическим сплайном $S(x)$ на сетке ω называется функция, которая

1) на каждом из отрезков

$$[x_i, x_{i+1}], \quad i = 0, 1, \ldots, m-1,$$

представляет собой многочлен третьей степени,

$$\boxed{S(x) = S_i(x) = a_0^{(i)} + a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3,}$$

2) дважды непрерывно дифференцируема на отрезке $[a, b]$, то есть принадлежит классу $C^2[a, b]$,

3) доставляет минимум функционалу

$$\boxed{J(f) = \int_a^b \left(f''(x)\right)^2 \, dx + \sum_{i=0}^{m} \frac{1}{\rho_i}(f(x_i) - y_i)^2,} \tag{4}$$

где y_i и $\rho_i > 0$ — заданные числа,

4) удовлетворяет граничным условиям одного из трех указанных ниже типов.

Граничные (краевые) условия

Граничные условия задаются в виде ограничений на значения сплайна и его производных в граничных узлах сетки ω.

А. Граничные условия 1-го типа.

$$S'(a) = y_0', \quad S'(b) = y_m'$$

— на концах промежутка $[a, b]$ задаются значения первой производной искомой функции.

Б. Граничные условия 2-го типа.

$$S''(a) = 0, \quad S''(b) = 0$$

— вторые производные искомой функции на концах промежутка $[a, b]$ равны нулю.

В. Граничные условия 3-го типа.

$$S(a) = S(b), \quad S'(a) = S'(b), \quad S''(a) = S''(b)$$

называются *периодическими.*

Теорема. *Кубический сплайн $S(x)$, минимизирующий функционал (4) и удовлетворяющий краевым условиям одного из указанных трех типов, определен однозначно.*

Определение. *Кубический сплайн, минимизирующий функционал $J(f)$ и удовлетворяющий граничным условиям i-го типа, называется сглаживающим сплайном i-го типа.*

Замечание. На каждом из отрезков $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, m-1$, сплайн $S(x)$ является многочленом третьей степени и определяется на этом отрезке четырьмя коэффициентами. Всего отрезков — m. Значит, для того, чтобы полностью определить сплайн, необходимо найти $4m$ чисел

$$a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, \quad i = 0, 1, \ldots, m-1.$$

Условие

$$S(x) \in C^2[a, b]$$

означает непрерывность функции $S(x)$ и ее производных $S'(x)$ и $S''(x)$ во всех внутренних узлах сетки ω. Число таких узлов — $m-1$. Тем самым, для отыскания коэффициентов всех многочленов получается $3(m-1)$ условий (уравнений).

Построение сглаживающего кубического сплайна

Опишем способ вычисления коэффициентов кубического сплайна, при котором число величин, подлежащих определению, равно $2m+2$.

На каждом из промежутков

$$[x_i, x_{i+1}], \quad i = 0, 1, \ldots, m-1,$$

сглаживающая сплайн-функция ищется в следующем виде

$$\boxed{S(x) = S_i(x) = z_i(1-t) + z_{i+1}t - \frac{h_i^2}{6}t(1-t)[(2-t)n_i + (1+t)n_{i+1}].}$$

Здесь

$$h_i = x_{i+1} - x_i, \quad t = \frac{x - x_i}{h_i},$$

а числа z_i и n_i, $i = 0, 1, \ldots, m$, являются решением системы линейных алгебраических уравнений, вид которой зависит от типа краевых условий.

Опишем сначала, как находятся величины n_i.

Для **краевых условий 1-го и 2-го типов** система линейных уравнений для определения величин μ_i записывается в следующем виде

$$\begin{aligned}
a_0 n_0 + b_0 n_1 + c_0 n_2 &= g_0, \\
b_0 n_0 + a_1 n_1 + b_1 n_2 + c_1 n_3 &= g_1, \\
c_{i-2} n_{i-2} + b_{i-1} n_{i-1} + a_i n_i + b_i n_{i+1} + c_i n_{i+2} &= g_i, \qquad i = 2, 3, \ldots, m-2, \\
c_{m-3} n_{m-3} + b_{m-2} n_{m-2} + a_{m-1} n_{m-1} + b_{m-1} n_m &= g_{m-1}, \\
c_{m-2} n_{m-2} + b_{m-1} n_{m-1} + a_m n_m &= g_m,
\end{aligned}$$

где

$$a_i = \frac{1}{3}(h_{i-1}+h_i)+\frac{1}{h_{i-1}^2}\rho_{i-1}+\left(\frac{1}{h_{i-1}}+\frac{1}{h_i}\right)^2\rho_i+\frac{1}{h_i^2}\rho_{i+1}, \quad i=1,2,\ldots,m-1,$$

$$b_i = \frac{1}{6}h_i-\frac{1}{h_i}\left[\left(\frac{1}{h_{i-1}}+\frac{1}{h_i}\right)\rho_i+\left(\frac{1}{h_i}+\frac{1}{h_{i+1}}\right)\rho_{i+1}\right], \quad i=1,2,\ldots,m-2, \tag{5}$$

$$c_i = \frac{1}{h_i h_{i+1}}\rho_{i+1}, \quad i=1,2,\ldots,m-3,$$

$$g_i = \frac{y_{i+1}-y_i}{h_i}-\frac{y_i-y_{i-1}}{h_{i-1}}, \quad i=1,2,\ldots,m-1$$

($\rho_i \geqslant 0$ — известные числа).

Коэффициенты

$$a_0,\ b_0,\ c_0,\ g_0,\ c_{m-2},\ b_{m-1},\ a_m,\ g_m$$

зависят от выбора граничных условий.

Граничные условия 1-го типа:

$$a_0 = \frac{h_0}{3}+\frac{1}{h_0^2}(\rho_0+\rho_1),$$

$$b_0 = \frac{h_0}{6}-\frac{1}{h_0}\left(\frac{1}{h_0}+\frac{1}{h_1}\right)\rho_1-\frac{1}{h_0^2}\rho_0,$$

$$c_0 = \frac{1}{h_0 h_1}\rho_1,$$

$$g_0 = \frac{y_1-y_0}{h_0}-y_0',$$

$$a_m = \frac{h_{m-1}}{3}+\frac{1}{h_{m-1}^2}(\rho_{m-1}+\rho_m),$$

$$b_{m-1} = \frac{h_{m-1}}{6}-\frac{1}{h_{m-1}}\left(\frac{1}{h_{m-1}}+\frac{1}{h_{m-2}}\right)\rho_{m-2}-\frac{1}{h_{m-1}^2}\rho_m,$$

$$c_{m-2} = \frac{1}{h_{m-2}h_{m-1}}\rho_{m-1},$$

$$g_m = y_m'-\frac{y_m-y_{m-1}}{h_{m-1}}.$$

Граничные условия 2-го типа:

$$a_0=1, \quad b_0=1, \quad c_0=0, \quad g_0=0,$$
$$a_m=1, \quad b_{m-1}=1, \quad c_{m-2}=0, \quad g_m=0.$$

В случае граничных условий 3-го типа система для определения чисел n_i, $i = 1, 2, \ldots, m$, записывается так:

$$c_{i-2}n_{i-2}+b_{i-1}n_{i-1}+a_i n_i+b_i n_{i+1}+c_i n_{i+2}=g_i, \quad i=1,2,\ldots,m,$$

причем все коэффициенты вычисляются по формулам (5) (величины с индексами k и $m+k$ считаются равными: $n_0 = n_m$, $h_0 = h_m$, $a_0 = a_m$ и т. д.).

Важное замечание. Матрицы систем невырождены и потому каждая из этих систем имеет единственное решение.

Если числа n_i найдены, то величины z_i легко определяются по формулам

$$z_i = y_i - \rho_i D_i, \quad i = 1, 2, \ldots, m,$$

где

$$D_0 = \frac{1}{h_0}(n_1 - n_0),$$

$$D_i = \frac{1}{h_i}(n_{i+1} - n_i) - \frac{1}{h_{i-1}}(n_i - n_{i-1}), \quad i = 1, 2, \ldots, m-1,$$

$$D_m = -\frac{1}{h_{m-1}}(n_m - n_{m-1}).$$

В случае периодических граничных условий

$$h_m = h_0, \quad n_m = n_0, \quad n_1 = n_{m+1}$$

и

$$D_0 = D_m = \frac{1}{h_m}(n_1 - n_m) - \frac{1}{h_{m-1}}(n_m - n_{m-1}).$$

Выбор весовых коэффициентов

Выбор весовых коэффициентов ρ_i, входящих в функционал (4), позволяет в известной степени управлять свойствами сглаживающих сплайнов.

Если все $\rho_i = 0$, то $z_i = y_i$ и сглаживающий сплайн оказывается интерполяционным. Это, в частности, означает, что чем точнее заданы величины y_i, тем меньше должны быть соответствующие весовые коэффициенты. Если же необходимо, чтобы сплайн прошел через точку (x_k, y_k), то отвечающий ему весовой множитель ρ_k следует положить равным нулю.

В практический вычислениях наиболее важным является выбор величин ρ_i.

Пусть Δ_i — погрешность измерения величины y_i. Тогда естественно потребовать, чтобы сглаживающий сплайн $\sigma(x)$ удовлетворял условию

$$|\sigma(x_i) - y_i| \leqslant \Delta_i \tag{6}$$

или, что то же,

$$\rho_i |D_i| \leqslant \Delta_i.$$

В простейшем случае весовые коэффициенты ρ_i можно задать, например, формулой

$$\rho_i = c\Delta_i,$$

где c — некоторая достаточно малая постоянная. Однако такой выбор весов ρ_i не позволяет использовать «коридор», обусловленный погрешностями величин y_i. Более рациональный, но и более трудоемкий алгоритм определения величин ρ_i может выглядеть следующим образом.

Если на k-й итерации величины $D_i^{(k)}$ найдены, то полагают

$$\rho_i^{(k+1)} = \begin{cases} \dfrac{\Delta_i}{|D_i^{(k)}|} & \text{при} \quad |D_i^{(k)}| > \varepsilon, \\ 0 & \text{при} \quad |D_i^{(k)}| \leqslant \varepsilon, \end{cases}$$

где ε — малое число, которое выбирается экспериментально с учетом разрядной сетки компьютера, значений Δ_i и точности решения системы линейных алгебраических уравнений.

Если на k-й итерации в точке x_i нарушилось условие (6), то последняя формула обеспечит уменьшение соответствующего весового коэффициента ρ_i. Если же

$$|\sigma^{(k)}(x_i) - y_i| < \Delta_i,$$

то на следующей итерации $\rho_i^{(k+1)} > \rho_i^{(k)}$. Увеличение ρ_i приводит к более полному использованию «коридора» (6) и, в конечном счете, более плавно изменяющемуся сплайну.

Немного теории

А. Обоснование формул для вычисления коэффициентов интерполяционного кубического сплайна. Введем обозначения

$$S'(x_i) = m_i, \quad i = 0, 1, \ldots, m,$$

где m_i — неизвестные пока величины. Их число равно $m + 1$.

Сплайн, записанный в форме

$$\boxed{S(x) = S_i(x) = y_i(1-t)^2(1+2t) + y_{i+1}t^2(32-2t) + m_i h_i t(1-t)^2 - m_i t^2(1-t),} \quad (7)$$

где

$$h_i = x_{i+1} - x_i, \quad t = \frac{x - x_i}{h_i},$$

удовлетворяет условиям интерполяции

$$S(x_i) = y_i, \quad i = 0, 1, \ldots, m,$$

и непрерывен на всем промежутке $[a, b]$: положив в формуле (7) $t = 0$ и $t = 1$, получим соответственно

$$S(x_i) = y_i, \quad S(x_{i+1}) = y_{i+1}.$$

Кроме того, он имеет на промежутке $[a, b]$ непрерывную первую производную: продифференцировав соотношение (7) и положив $t = 0$ и $t = 1$, получим соответственно

$$S'(x_i) = m_i, \quad S'(x_{i+1}) = m_{i+1}.$$

Покажем, что числа m_i можно выбрать так, чтобы сплайн-функция (7) имела на отрезке $[a, b]$ непрерывную вторую производную.

Вычислим на промежутке $[x_{i-1}, x_i]$ вторую производную сплайна (7):

$$S''(x) = S''_{i-1}(x) = (y_i - y_{i-1})\frac{6 - 12t}{h_{i-1}^2} + m_{i-1}\frac{-4 + 6t}{h_{i-1}} + m_i\frac{-2 + 6t}{h_{i-1}}.$$

В точке $x_i - 0$ (при $t = 1$) имеем

$$S''(x_i - 0) = S''_{i-1}(x_i) = -6\frac{y_i - y_{i-1}}{h_{i-1}^2} - 4\frac{m_{i-1}}{h_{i-1}} - 2\frac{m_i}{h_{i-1}}.$$

Вычислим на промежутке $[x_i, x_{i+1}]$ вторую производную сплайна (7):

$$S''(x) = S''_i(x) = (y_{i+1} - y_i)\frac{6 - 12t}{h_i^2} + m_i\frac{-4 + 6t}{h_i} + m_{i+1}\frac{-2 + 6t}{h_i}.$$

В точке $x_i + 0$ (при $t = 0$) имеем

$$S''(x_i + 0) = S''_i(x_i) = -6\frac{y_{i+1} - y_i}{h_i^2} - 4\frac{m_i}{h_i} - 2\frac{m_{i+1}}{h_i}.$$

Из условия непрерывности второй производной во внутренних узлах сетки ω

$$S''(x_i - 0) = S''(x_i + 0), \quad i = 1, 2, \ldots, m - 1,$$

получаем $m - 1$ соотношение

$$\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = 3\left(\mu_i\frac{y_{i+1} - y_i}{h_i} + \lambda_i\frac{y_i - y_{i-1}}{h_{i-1}}\right), \quad i = 1, 2, \ldots, m - 1,$$

где

$$\mu_i = \frac{h_{i-1}}{h_{i+1} + h_i}, \quad \lambda_i = 1 - \mu_i = \frac{h_i}{h_{i+1} + h_i}.$$

Добавляя к этим $m - 1$ уравнениям еще два, вытекающих из краевых условий, получаем систему из $m + 1$ линейного алгебраического уравнения с $m + 1$ неизвестной m_i, $i = 0, 1, \ldots, m$.

Система уравнений для вычисления величин m_i в случае краевых условий 1-го и 2-го типов имеет вид

$$\begin{cases} 2m_0 + \mu_0^* m_1 = c_0^*, \\ \lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = c_i, \quad i = 2, \ldots, m - 1, \\ \lambda_m^* m_{m-1} + 2m_m = c_m^*, \end{cases}$$

где

$$c_i = 3\left(\mu_i\frac{y_{i+1} - y_i}{h_i} + \lambda_i\frac{y_i - y_{i-1}}{h_{i-1}}\right)$$

и

$$\mu_0^* = 0, \quad c_0^* = 2y'_0, \quad \lambda_m^* = 0, \quad c_m^* = 2y'_m$$

(*краевые условия 1-го типа*),

$$\mu_0^* = 1, \quad c_0^* = 3\frac{y_1 - y_0}{h_0} - \frac{h_0}{2}y''_0,$$

$$\lambda_m^* = 1, \quad c_m^* = 3\frac{y_m - y_{m-1}}{h_{m-1}} 2y''_m$$

(*краевые условия 2-го типа*).

Для периодических краевых условий (*краевые условия 3-го типа*) сетку ω удлиняют еще на один узел и полагают

$$y_m = y_0, \quad y_{m+1} = y_1,$$
$$m_m = m_0, \quad m_{m+1} = m_1, \quad h_m = h_0.$$

Тогда система для определения величин m_i будет иметь вид

$$\begin{cases} 2m_1 + \mu_1 m_2 + \lambda_1 m_m = c_1, \\ \lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = c_i, \quad i = 2, \ldots, m-1, \\ \mu_0 m_1 + \lambda_m m_{m-1} + 2m_m = c_m. \end{cases}$$

Для того чтобы получить систему уравнений для определения чисел m_i в случае краевых условий 4-го типа, найдем на отрезке $[x_i, x_{i+1}]$ третью производную сплайна (7)

$$S'''(x) = S_i'''(x) = \frac{6}{h_i^2}\left(m_{i+1} + m_i - 2\frac{y_{i+1} - y_i}{h_i}\right)$$

и потребуем ее непрерывности во втором и $(m-1)$-м узлах сетки.

Имеем

$$\frac{1}{h_0^2}\left(m_1 + m_0 - 2\frac{y_1 - y_0}{h_0}\right) = \frac{1}{h_1^2}\left(m_2 + m_1 - 2\frac{y_2 - y_1}{h_1}\right),$$

$$\frac{1}{h_{m-1}^2}\left(m_m + m_{m-1} - 2\frac{y_m - y_{m-1}}{h_{m-1}}\right) = \frac{1}{h_{m-2}^2}\left(m_{m-1} + m_{m-2} - 2\frac{y_{m-1} - y_{m-2}}{h_{m-2}}\right).$$

Из последних двух соотношений получаем недостающие два уравнения, отвечающие *краевым условиям 4-го типа*:

$$m_0 + (1 - \gamma_0^2)m_1 - \gamma_0^2 m_2 = 2\left(\frac{y_1 - y_0}{h_0} - \gamma_0^2\frac{y_2 - y_1}{h_1}\right),$$

$$-\gamma_m^2 m_{m-2} + (1 - \gamma_m^2)m_{m-1} + m_m = 2\left(\frac{y_m - y_{m-1}}{h_{m-1}} - \gamma_m^2\frac{y_{m-1} - y_{m-2}}{h_{m-2}}\right).$$

Исключая из уравнений

$$m_0 + (1 - \gamma_0^2)m_1 - \gamma_0^2 m_2 = 2\left(\frac{y_1 - y_0}{h_0} - \gamma_0^2\frac{y_2 - y_1}{h_1}\right)$$

и

$$\lambda_1 m_0 + 2m_1 + \mu_1 m_2 = 3\left(\mu_1\frac{y_2 - y_1}{h_1} + \lambda_1\frac{y_1 - y_0}{h_0}\right)$$

неизвестное m_0, а из уравнений

$$-\gamma_m^2 m_{m-2} + (1 - \gamma_m^2)m_{m-1} + m_m = 2\left(\frac{y_m - y_{m-1}}{h_{m-1}} - \gamma_m^2\frac{y_{m-1} - y_{m-2}}{h_{m-2}}\right)$$

и

$$\lambda_{m-1} m_{m-2} + 2m_{m-1} + \mu_{m-1} m_m = 3\left(\mu_{m-1}\frac{y_m - y_{m-1}}{h_{m-1}} + \lambda_{m-1}\frac{y_{m-1} - y_{m-2}}{h_{m-2}}\right)$$

неизвестное m_m, в результате получим систему уравнений

$$\begin{cases} (1 + \gamma_0)m_1 + \gamma_0 m_2 = c_1^*, \\ \lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = c_i, \quad i = 2, \ldots, m-2, \\ \gamma_m m_{m-2} + (1 + \gamma_m)m_{m-1} = c_{m-1}^*. \end{cases}$$

Отметим, что число неизвестных в этой системе равно $m-1$.

Б. Обоснование формул для вычисления коэффициентов сглаживающего кубического сплайна. Введем обозначения

$$S(x_i) = z_i, \quad S''(x_i) = n_i, \quad i = 0, 1, \ldots, m,$$

где z_i и n_i — неизвестные пока величины. Их число равно $2m + 2$.

Сплайн-функция, записанная в форме

$$\boxed{S(x) = S_i(x) = z_i(1-t) + z_{i+1}t - \frac{h_i^2}{6}t(1-t)[(2-t)n_i + (1+t)n_{i+1}],} \tag{8}$$

где

$$h_i = x_{i+1} - x_i, \quad t = \frac{x - x_i}{h_i},$$

непрерывна на всем промежутке $[a, b]$: положив в этой формуле $t = 0$ и $t = 1$, получим соответственно

$$S(x_i) = z_i, \quad S(x_{i+1}) = z_{i+1}.$$

Покажем, что числа z_i и n_i можно выбрать так, чтобы сплайн, записанный в форме (8), имел на промежутке $[a, b]$ непрерывную первую производную.

Вычислим первую производную сплайна $S(x)$ на промежутке $[x_{i-1}, x_i]$:

$$S'(x) = S'_{i-1}(x) = \frac{-z_{i-1} + z_i}{h_{i-1}} - \frac{h_{i-1}}{6}\left((2 - 6t + 3t^2)n_{i-1} + (1 - 3t^2)n_i\right).$$

В точке $x_i - 0$ (при $t = 1$) имеем

$$S'(x_i - 0) = S'_{i-1}(x_i) = \frac{-z_{i-1} + z_i}{h_{i-1}} + \frac{h_{i-1}}{6}(n_{i-1} + 2n_i).$$

Вычислим первую производную сплайна $S(x)$ на промежутке $[x_i, x_{i+1}]$:

$$S'(x) = S'_i(x) = \frac{-z_i + z_{i+1}}{h_i} - \frac{h_i}{6}\left((2 - 6t + t^2)n_i + (1 - 3t^2)n_{i+1}\right).$$

В точке $x_i + 0$ (при $t = 0$) имеем

$$S'(x_i + 0) = S'_i(x_i) = \frac{-z_i + z_{i+1}}{h_i} - \frac{h_i^2}{6}(2n_i + n_{i+1}).$$

Из условия непрерывности первой производной сплайна во внутренних узлах сетки ω

$$S'(x_i - 0) = S'(x_i + 0), \quad i = 1, 2, \ldots, m - 1,$$

получаем $m - 1$ соотношение

$$-z_{i-1} - 2z_i + z_{i+1} = \frac{h_{i-1}^2}{6}(n_{i-1} + 2n_i) + \frac{h_i^2}{6}(2n_i + n_{i+1}), \quad i = 1, 2, \ldots, m - 1.$$

Эту связь удобно записать в матричной форме

$$\mathbf{AM} = 6\mathbf{Hz}.$$

Здесь использованы следующие обозначения

$$\mathbf{z} = \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_m \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} n_0 \\ n_1 \\ \vdots \\ n_m \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & & & & & & \\ 0 & u_1 & h_1 & & & & & \\ & h_1 & u_2 & h_2 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & h_{m-3} & u_{m-2} & h_{m-2} & & \\ & & & & h_{m-2} & u_{m-1} & 0 & \\ & & & & & 0 & 1 & \end{pmatrix},$$

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 & & & & & \\ v_0 & -w_0 & v_1 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & v_i & -w_i & v_{i+1} & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & v_{m-2} & -w_{m-2} & v_{m-1} & \\ & & & & & 0 & 0 & \end{pmatrix},$$

где

$$u_i = 2(h_{i-1} + h_i), \quad v_i = \frac{1}{h_i}, \quad w_i = \frac{1}{h_i} + \frac{1}{h_{i+1}}.$$

Кроме того, сплайн $S(x)$ на промежутке $[a, b]$ имеет непрерывную вторую производную: продифференцировав соотношение (8) и положив $t = 0$ и $t = 1$, получим соответственно

$$S''(x_i) = n_i, \quad S''(x_{i+1}) = n_{i+1}.$$

Еще одно матричное соотношение получается из условия минимума функционала (4). Имеем

$$\mathbf{z} + \mathbf{R}\mathbf{H}^T\mathbf{M} = \mathbf{y}, \tag{9}$$

где

$$\mathbf{R} = \begin{pmatrix} \rho_0 \\ \rho_1 \\ \vdots \\ \rho_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

Два последних матричных равенства можно рассматривать как линейную систему $2m+2$ линейных алгебраических уравнений относительно $2m+2$ неизвестных z_i и n_i, $i = 0, 1, \ldots, m$. Заменяя в первом равенстве столбец $\mathbf{z}$ его выражением, полученным из соотношения (9), приходим к матричному уравнению

$$(\mathbf{A} + 6\mathbf{H}\mathbf{R}\mathbf{H}^T)\mathbf{M} = 6\mathbf{H}\mathbf{y}$$

для определения столбца M. Это уравнение имеет единственное решение вследствие того, что матрица $\mathbf{A} + 6\mathbf{H}\mathbf{R}\mathbf{H}^T$ всегда невырождена. Найдя его, мы легко определяем $\mathbf{z}$.

Замечание. Элементы трехдиагональных матриц A и H определяются только параметрами сетки ω (ее шагами h_i) и не зависят от величин y_i.

Линейное пространство кубических сплайн-функций

Множество кубических сплайнов, построенных на отрезке $[a, b]$ по сетке ω с $m+1$ узлом, является линейным пространством размерности $m+3$:

1) сумма двух кубических сплайнов, построенных по сетке ω, и произведение кубического сплайна, построенного по сетке ω, на произвольное число также являются кубическими сплайнами, построенными по этой сетке,

2) любой кубический сплайн, построенный по сетке ω из $m+1$ узла, полностью определяется $m+1$ значением величин y_i в этих узлах и двумя граничными условиями — всего $m+3$ параметрами.

Выбрав в этом пространстве базис, состоящий из $m+3$ линейно независимых сплайнов $\sigma_i(x)$, $i = 1, \ldots, m+3$, мы можем записать произвольный кубический сплайн $\sigma(x)$ в виде их линейной комбинации

$$\sigma(x) = \sum_{i=1}^{m+3} \alpha_i \sigma_i(x),$$

причем единственным образом.

Замечание. Подобное задание сплайна широко распространено в вычислительной практике. Особенно удобным является базис, состоящий из так называемых кубических B-сплайнов (базовых, или фундаментальных, сплайнов). Применение B-сплайнов позволяет существенно снизить требования к объему памяти компьютера.

B-сплайны.

B-сплайном нулевой степени, построенным на числовой прямой по сетке ω, называется функция вида

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

B-сплайн степени $k \geqslant 1$, построенный на числовой прямой по сетке ω, определяется посредством рекуррентной формулы

$$B_i^{(k)}(x) = \frac{x - x_i}{x_{i+1} - x_i} B_i^{(k-1)}(x) + \frac{x_{i+k+1} - x}{x_{i+k+1} - x_{i+1}} B_{i+1}^{(k-1)}(x).$$

Графики B-сплайнов первой $B_i^{(1)}(x)$ и второй $B_i^{(2)}(x)$ степеней представлены на рис. 11 и 12 соответственно.

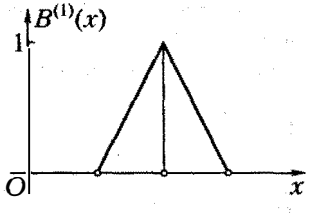


Рис. 11

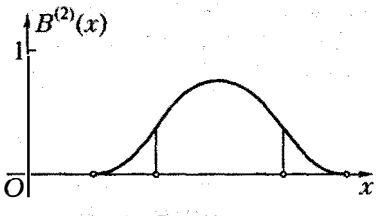


Рис. 12

B-сплайн произвольной степени k может быть отличен от нуля только на некотором отрезке (определяемом $k+2$ узлами).

Кубические B-сплайны удобнее нумеровать так, чтобы сплайн $B_i^{(3)}(x)$ был отличен от нуля на отрезке $[x_{i-2}, x_{i+2}]$.

Приведем формулу для кубического сплайна третьей степени для случая равномерной сетки (с шагом h). Имеем

$$B_i^{(3)}(x) = \begin{cases} \frac{1}{6}\left(\frac{x-x_i}{h} - 2\right)^3, & x \in [x_i, x_{i+1}], \\ \frac{2}{3} - \frac{1}{2}\left[\left(\frac{x-x_i}{h}\right)^3 - \left(\frac{x-x_i}{h}\right)^2\right], & x \in [x_{i+1}, x_{i+2}], \\ \frac{2}{3} + \frac{1}{2}\left[\left(\frac{x-x_i}{h}\right)^3 - \left(\frac{x-x_i}{h}\right)^2\right], & x \in [x_{i+2}, x_{i+3}], \\ \frac{1}{6}\left(2 - \frac{x-x_i}{h}\right)^3, & x \in [x_{i+3}, x_{i+4}], \\ 0 & \text{в остальных случаях.} \end{cases}$$

Типичный график кубического B-сплайна представлен на рис. 13.

Замечание. Функция $B_i^{(3)}(x)$:

а) дважды непрерывно дифференцируема на отрезке $[a, b]$, то есть принадлежит классу $C^2[a, b]$, и

б) отлична от нуля только на четырех последовательных отрезках $[x_{i-2}, x_{i-1}]$, $[x_{i-1}, x_i]$, $[x_i, x_{i+1}]$, $[x_{i+1}, x_{i+2}]$.

Отрезок $[x_{i-2}, x_{i+2}]$ называется *носителем* функции $B_i^{(3)}(x)$.

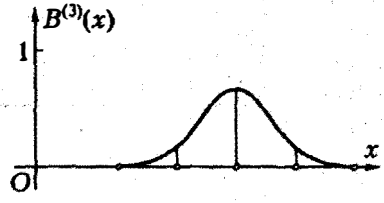


Рис. 13

Дополним сетку ω вспомогательными узлами

$$x_{-3} < x_{-2} < x_{-1} < a, \quad b < x_{m+1} < x_{m+2} < x_{m+3},$$

взятыми совершенно произвольно.

По расширенной сетке ω^*

$$x_{-3} < x_{-2} < x_{-1} < a < x_1 < \ldots$$
$$\ldots < x_m < x_{m+1} < x_{m+2} < x_{m+3}$$

можно построить семейство из $m+3$ кубических B-сплайнов:

$$B_i^{(3)}(x), \quad i = -1, 0, \ldots, m, m+1.$$

Это семейство образует базис в пространстве кубических сплайнов на отрезке $[a, b]$. Тем самым, произвольный кубический сплайн $S(x)$, построенный на отрезке $[a, b]$ по сетке ω из $m+1$ узла, может быть представлен на этом отрезке в виде линейной комбинации

$$S(x) = \sum_{i=-1}^{m+1} b_i B_i^{(3)}(x).$$

Условиями задачи коэффициенты b_i этого разложения определяются однозначно.

В случае, когда заданы значения y_i функции в узлах сетки и значения y'_0 и y'_m первой производной функции на концах сетки (задача интерполяции с граничными

условиями первого рода), эти коэффициенты вычисляются из системы следующего вида

$$\begin{cases} b_{-1}B'_{-1}(x_0) + b_0B'_0(x_0) + b_{-1}B'_1(x_0) = y'_0, \\ b_{i-1}B'_{i-1}(x_i) + b_iB'_i(x_i) + b_{i+1}B'_{i+1}(x_i) = y_i, \quad i = 0, 1, \ldots, m, \\ b_{m-1}B'_{m-1}(x_m) + b_mB'_m(x_m) + b_{m+1}B'_{m+1}(x_m) = y'_m. \end{cases}$$

После исключения величин b_{-1} и b_{m+1} получается линейная система с неизвестными $b_0, \ldots, b_m$ и трехдиагональной матрицей. Условие

$$\rho = \max_{|i-j|=1} \frac{h_i}{h_j} < \frac{1+\sqrt{3}}{2} \approx 1{,}366$$

обеспечивает диагональное преобладание и, значит, возможность применения метода прогонки для ее разрешения.

Замечание 1. *Линейные системы аналогичного вида возникают при рассмотрении и других задач интерполяции.*

Замечание 2. *В сравнении с алгоритмами, описанными в разделе 1.1, применение B-сплайнов в задачах интерполяции позволяет уменьшить объем хранимой информации, то есть существенно снизить требования к объему памяти компьютера, хотя и приводит к увеличению числа операций.*

Построение сплайновых кривых при помощи сплайн-функций

Выше рассматривались массивы, точки которых были занумерованы так, что их абсциссы образовывали строго возрастающую последовательность. Например, случай, изображенный на рис. 14, когда у разных точек массива одинаковые абсциссы, не допускался. Это обстоятельство определяло и выбор класса аппроксимирующих кривых (графики функций), и способ их построения.

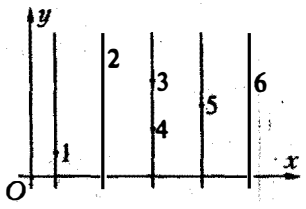


Рис. 14

Однако предложенный выше метод позволяет достаточно успешно строить интерполяционную кривую и в более общем случае, когда нумерация точек массива

$$\mathbf{P} = \{\mathbf{P}_i(x_i, y_i),\ i = 0, 1, \ldots, m\}$$

и их расположение на плоскости, как правило, не связаны (рис. 15). Более того, ставя задачу построения интерполяционной кривой, можно считать заданный массив неплоским, то есть

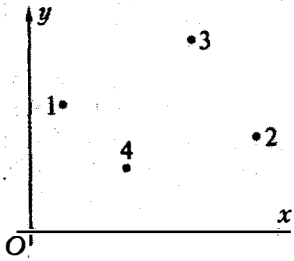


Рис. 15

$$\mathbf{P} = \{\mathbf{P}_i(x_i, y_i, z_i),\ i = 0, 1, \ldots, m\}.$$

Ясно, что для решения этой общей задачи необходимо существенно расширить класс допустимых кривых, включив в него и замкнутые кривые, и кривые, имеющие точки самопересечения, и пространственные кривые. Такие кривые удобно описывать при помощи параметрических уравнений

$$x = x(t), \quad y = y(t), \quad z = z(t), \quad \alpha \leqslant t \leqslant \beta.$$

Потребуем дополнительно, чтобы функции $x(t)$, $y(t)$ и $z(t)$ обладали достаточной гладкостью, например, принадлежали классу $C^1[\alpha, \beta]$ или классу $C^2[\alpha, \beta]$.

Для отыскания параметрических уравнений кривой, последовательно проходящей через все точки массива, поступают следующим образом.

1-*й шаг.* На произвольно взятом отрезке $[\alpha, \beta]$ изменения параметра t вводится вспомогательная сетка

$$\alpha = t_0 < t_1 < \ldots < t_{m-1} < t_m = \beta,$$

число узлов которой совпадает с числом точек в массиве **P**.

2-*й шаг.* По заданному массиву **P** строятся три новых вспомогательных массива (в плоском случае два)

$$X = \{(t_i, x_i),\ i = 0, 1, \ldots, m\},$$
$$Y = \{(t_i, y_i),\ i = 0, 1, \ldots, m\},$$
$$Z = \{(t_i, z_i),\ i = 0, 1, \ldots, m\}$$

(рис. 16 (плоский случай)).

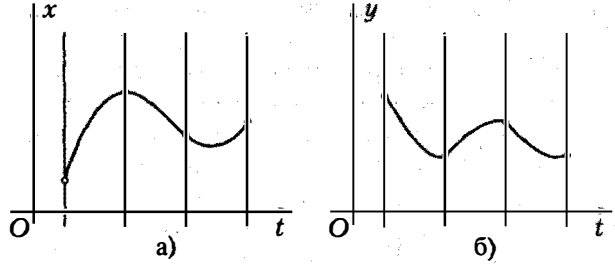


Рис. 16

3-*й шаг.* Для каждого из массивов X, Y и Z находятся соответствующие интерполяционные сплайн-функции $x(t)$, $y(t)$ и $z(t)$.

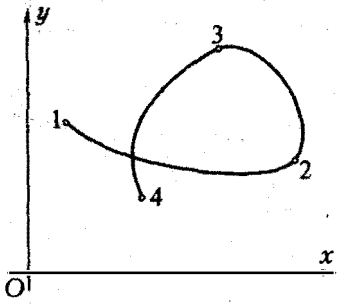


Рис. 17

В результате мы получаем параметрические уравнения

$$x = x(t), \quad y = y(t), \quad z = z(t), \quad \alpha \leqslant t \leqslant \beta,$$

кривой, проходящей через точки $\mathbf{P}_i$, $i = 0, 1, \ldots, m$ (см. рис. 17 для плоского случая).

Замечание 1. Предложенный подход позволяет строить замкнутые интерполяционные кривые (при $\mathbf{P}_0 = \mathbf{P}_m$); для этого при построении координатных функций $x(t)$, $y(t)$ и $z(t)$ нужно использовать граничные условия 3-го типа.

Замечание 2. Полученная кривая будет гладкой, но не обязательно регулярной, так как возможность одновременного обращения в нуль производных

$$x'(t^*) = 0, \quad y'(t^*) = 0, \quad z'(t^*) = 0$$

для некоторого $t^* \in (\alpha, \beta)$ исключать нельзя. Кроме того, эта кривая может иметь точки самопересечения.

Замечание 3. Построение сглаживающих сплайновых кривых проводится практически так же, как и построение интерполяционных сплайновых кривых.

§ 2. Геометрические сплайны

Во многих задачах требование того, чтобы конструируемая кривая (или поверхность) однозначно проектировалась на прямую (или плоскость), является слишком жестким. Расширяя допустимые классы кривых и поверхностей, естественно обратиться к более общему способу описания их частичных фрагментов. В качестве нового, более общего способа задания кривых и поверхностей удобно взять параметрический способ.

Параметрическое задание кривой или поверхности имеет известные преимущества перед другими методами, в частности, потому, что он не накладывает практически никаких ограничений на множество вершин в опорном массиве.

Вместе с тем, этот метод требует и большей осторожности: для того, чтобы составная кривая (или поверхность) была достаточно регулярной, необходимо быть очень внимательным, особенно в местах стыковки.

Выбор многочленов третьей степени для описания координатных функций проектируемых кривых геометрически вполне обоснован: координатные функции должны быть сравнительно простыми и, одновременно, обеспечивающими разумную гладкость.

При построении составных поверхностей полезно использовать результаты решения соответствующих одномерных задач, в частности, ограничиться бикубическим описанием координатных функций. Это заметно упростит рассмотрение задачи создания поверхности сложной формы.

Общую задачу можно сформулировать так: по заданному множеству вершин

$$\mathbf{P} = \{\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_{m-1}, \mathbf{P}_m\}$$

с учетом их нумерации построить гладкую кривую, которая, плавно изменяясь, последовательно проходила бы вблизи этих вершин и удовлетворяла некоторым дополнительным условиям. Эти условия могут иметь различный характер. Например, можно потребовать, чтобы искомая кривая проходила через все заданные вершины или, проходя через заданные вершины, касалась заданных направлений, являлась замкнутой или имела заданную регулярность и т. п.

При отыскании подходящего решения задачи приближения важную роль играет ломаная, звенья которой соединяют соседние вершины заданного набора. Эту ломаную называют *контрольной*, *опорной* или *управляющей*, а ее вершины — *контрольными*, *опорными* или *управляющими* (рис. 18). Во многих случаях она довольно точно показывает, как будет проходить искомая кривая, что особенно полезно при решении задачи сглаживания. Каждая вершина заданного массива является либо *внутренней*, либо *граничной* (*концевой*). В массиве $\mathbf{P}$ вершины $\mathbf{P}_i$, $i = 1, \ldots, m-1$, — внутренние, а вершины $\mathbf{P}_0$ и $\mathbf{P}_m$ — граничные (концевые).

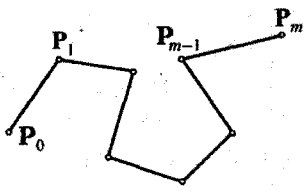


Рис. 18

В отличие от ситуации, рассматриваемой в разделе «Сплайн-функции», здесь на множество вершин не накладывается никаких ограничений — они могут быть заданы как на плоскости, так и в пространстве, их взаимное расположение может быть совершенно произвольным, некоторые из вершин могут совпадать и т. д. Поэтому описание нужной кривой следует искать в более общей, параметрической форме,

например, в следующем виде

$$\boxed{\mathbf{R}(t) = \sum a_i(t)\mathbf{P}_i,} \qquad (1)$$

где $a_i(t)$ — некоторые функциональные коэффициенты, подлежащие определению.

Если количество вершин в заданном множестве P достаточно велико, то найти универсальные функциональные коэффициенты $a_i(t)$, как правило, довольно затруднительно. Если же универсальные коэффициенты $a_i(t)$ все же найдены, то часто оказывается, что они, наряду с нужными свойствами, обладают и такими, которые не всегда удовлетворительно согласуются с ожидаемым поведением соответствующей кривой (например, кривая, описываемая уравнением (1) с этими коэффициентами, может осциллировать или отклоняться от заданного множества, местами очень заметно).

Для успешного решения поставленной задачи приближения, весьма удобно привлечь кривые, составленные из элементарных фрагментов. В случае, если эти элементарные фрагменты строятся по единой, сравнительно простой схеме, такие составные кривые принято называть *сплайновыми кривыми*.

Параметрические уравнения каждого элементарного фрагмента ищутся в виде (1) с той лишь разницей, что всякий раз привлекается только часть заданных вершин множества P, а соответствующие коэффициенты имеют одинаковую природу: часто используются многочлены одинаковой степени, рациональные дроби, экспоненты и др.

Для описания элементарных кривых и вычисления их геометрических характеристик (информация о которых необходима при состыковке) в качестве функциональных коэффициентов обычно используются многочлены невысоких степеней, второй или третьей, в первую очередь потому, что они сравнительно просто вычисляются. Конечно, привлекая многочлены больших степеней, можно описывать весьма сложные кривые. Однако у таких многочленов много коэффициентов, физический и геометрический смысл которых трудно понять. Кроме того, использование многочленов высокой степени может вызвать нежелательные колебания результирующей кривой.

Наибольшее распространение получили методы конструирования составных кривых, в которых используются кубические многочлены. Выбор кубических многочленов при построении фрагментов в качестве функциональных коэффициентов позволяет учесть и дифференциальные, и геометрические требования, накладываемые на искомую кривую.

В этом разделе мы расскажем о кривых Безье и B-сплайновых кривых. А также остановимся, совсем коротко, на описании сплайновых поверхностей.

2.1. Кривые Безье

Параметрические уравнения кривой Безье

По заданному массиву вершин

$$\mathbf{P} = \{\mathbf{P}_i(x_i, y_i, z_i),\ i = 0, 1, \ldots, m\}$$

(*элементарная*) *кривая Безье степени* m определяется при помощи векторного уравнения, имеющего следующий вид

$$\boxed{\mathbf{R}(t) = \sum_{i=0}^{m} B_i^m(t)\mathbf{P}_i, \quad 0 \leqslant t \leqslant 1,}$$

где

$$\boxed{B_i^m(t) = \binom{m}{i} t^i (1-t)^{m-i} = \frac{m!}{i!(m-i)!} t^i (1-t)^{m-i}}$$

— *многочлены Бернштейна.*

Матричная запись параметрических уравнений, описывающих кривую Безье,

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} x_0 & \dots & x_m \\ y_0 & \dots & y_m \\ z_0 & \dots & z_m \end{pmatrix} \begin{pmatrix} \mu_{00} & \dots & \mu_{0m} \\ \dots & \dots & \dots \\ \mu_{m0} & \dots & \mu_{mm} \end{pmatrix} \begin{pmatrix} t^0 \\ \vdots \\ t^m \end{pmatrix},$$

$$0 \leqslant t \leqslant 1,$$

где

$$\mu_{ij} = (-1)^{j-i} \binom{m}{j} \binom{j}{i}.$$

В случае, если промежуток изменения параметра произволен $a \leqslant t \leqslant b$, уравнение кривой Безье имеет следующий вид

$$\mathbf{R}(t) = \sum_{i=0}^{m} B_i^m \left(\frac{t-a}{b-a} \right) \mathbf{P}_i.$$

Важный частный случай. При $m = 3$ имеем элементарную кубическую кривую Безье, определяемую четырьмя вершинами и описываемую уравнением

$$\boxed{\mathbf{R}(t) = \big(((1-t)\mathbf{P}_0 + 3t\mathbf{P}_1)(1-t) + 3t^2\mathbf{P}_2\big)(1-t) + t^3\mathbf{P}_3, \quad 0 \leqslant t \leqslant 1,}$$

или, в матричной форме

$$\boxed{\mathbf{R}(t) = \mathbf{PMT}, \quad 0 \leqslant t \leqslant 1,}$$

где

$$\mathbf{R}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} \mathbf{P}_0 & \mathbf{P}_1 & \mathbf{P}_2 & \mathbf{P}_3 \end{pmatrix} = \begin{pmatrix} x_0 & x_1 & x_2 & x_3 \\ y_0 & y_1 & y_2 & y_3 \\ z_0 & z_1 & z_2 & z_3 \end{pmatrix},$$

$$\mathbf{M} = \begin{pmatrix} 1 & -3 & 3 & 1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{T} = \begin{pmatrix} t^0 \\ t^1 \\ t^2 \\ t^3 \end{pmatrix}.$$

Матрица **M** называется *базисной матрицей кубической кривой Безье.*

Свойства кривых Безье

При описании элементарных кривых Безье в качестве функциональных весовых множителей берутся многочлены Бернштейна

$$B_i^m(t) = \binom{m}{i} t^i (1-t)^{m-i} = \frac{m!}{i!(m-i)!} t^i (1-t)^{m-i}.$$

Свойства, которыми они обладают, оказывают существенное влияние на поведение элементарных кривых Безье. Укажем некоторые из них.

Многочлены Бернштейна:

1^+) неотрицательны,

2^+) в сумме составляют 1

$$\sum_{i=0}^{m} B_i^m(t) = 1,$$

3^+) не зависят от вершин массива $\mathbf{P}$ (универсальны).

Основные свойства кривых Безье

Элементарная кривая Безье, порожденная массивом $\mathbf{P}$:

1^+) является гладкой кривой, в частности, первую и вторую производные радиус-вектора $\mathbf{R}(t)$ можно записать так

$$\dot{\mathbf{R}}(t) = \frac{d}{dt}\mathbf{R}(t) = m\sum_{i=0}^{m-1}(\mathbf{P}_{i+1} - \mathbf{P}_i)B_i^{m-1}(t),$$

$$\ddot{\mathbf{R}}(t) = \frac{d^2}{dt^2}\mathbf{R}(t) = m(m-1)\sum_{i=0}^{m-2}(\mathbf{P}_{i+2} - 2\mathbf{P}_{i+1} + \mathbf{P}_i)B_i^{m-2}(t),$$

2^+) начинается в первой вершине $\mathbf{P}_0$ массива $\mathbf{P}$, $\mathbf{R}(0) = \mathbf{P}_0$, касаясь отрезка $\mathbf{P}_0\mathbf{P}_1$ опорной ломаной,

$$\dot{\mathbf{R}}(0) = m(\mathbf{P}_1 - \mathbf{P}_0),$$

и заканчивается в последней его вершине $\mathbf{P}_m$, $\mathbf{R}(1) = \mathbf{P}_m$, касаясь отрезка $\mathbf{P}_{m-1}\mathbf{P}_m$ опорной ломаной,

$$\dot{\mathbf{R}}(1) = m(\mathbf{P}_m - \mathbf{P}_{m-1})$$

(рис. 19),

3^+) лежит в выпуклой оболочке, порожденной массивом $\mathbf{P}$ (рис. 20),

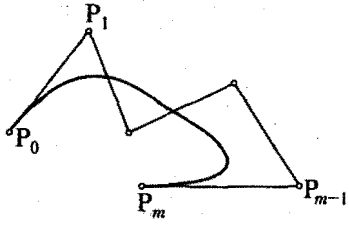


Рис. 19

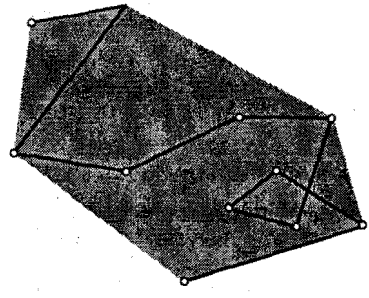
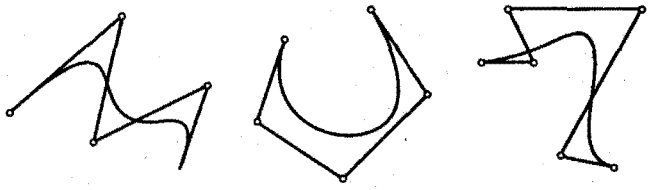


Рис. 20

4^+) симметрична — при перемене порядка вершин массива на противоположный,

$$\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_{m-1}, \mathbf{P}_m \to \mathbf{P}_m, \mathbf{P}_{m-1}, \ldots, \mathbf{P}_1, \mathbf{P}_0,$$

не изменяет своей формы,



а) б) в)

Рис. 21

5^+) «повторяет» опорную ломаную (рис. 21) (в частности, число точек пересечения кривой Безье с произвольной прямой не больше числа точек пересечения с этой прямой опорной ломаной).

6^+) в случае, если опорные вершины $\mathbf{P}_0, \ldots, \mathbf{P}_m$ лежат на одной прямой, кривая Безье совпадает с отрезком $\mathbf{P}_0\mathbf{P}_m$,

7^+) в случае, если опорные вершины $\mathbf{P}_0, \ldots, \mathbf{P}_m$ лежат в одной плоскости, кривая Безье лежит в этой же плоскости,

8^-) степень функциональных коэффициентов напрямую связана с количеством вершин в массиве (на единицу больше) и растет при его увеличении,

9^-) при добавлении в массив хотя бы одной вершины возникает необходимость полного пересчета параметрических уравнений элементарной кривой Безье,

10^-) изменение хотя бы одной вершины в массиве приводит к заметному изменению всей кривой Безье,

11^-) априорные сведения о расположении кривой Безье (принадлежность выпуклой оболочке заданного массива вершин) являются достаточно грубыми (на рис. 22 показан вид кубических кривых Безье для массива из четырех вершин на плоскости при разном порядке их нумерации; нетрудно видеть, что, находясь в одном и том же выпуклом четырехугольнике и пытаясь повторить ход соответствующих опорных ломаных, эти кривые сильно разнятся).

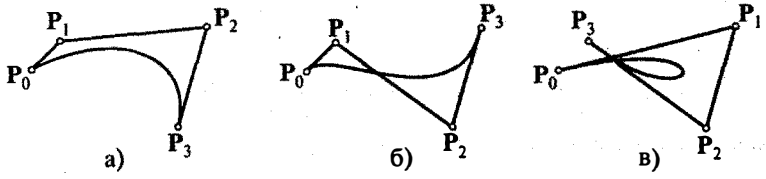


Рис. 22

2.2. B-сплайновые кривые

Параметрические уравнения элементарной кубической B-сплайновой кривой

По заданному массиву

$$\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$$

(*элементарная*) *кубическая B-сплайновая кривая* определяется при помощи векторного уравнения, имеющего следующий вид

$$\mathbf{R}(t) = \frac{(1-t)^3}{6}\mathbf{P}_0 + \frac{3t^3 - 6t^2 + 4}{6}\mathbf{P}_1 + \frac{-3t^3 + 3t^2 + 3t + 1}{6}\mathbf{P}_2 + \frac{t^3}{6}\mathbf{P}_3, \quad 0 \leqslant t \leqslant 1.$$

Матричная запись параметрических уравнений, описывающих элементарную кубическую B-сплайновую кривую

$$\boxed{\mathbf{R}(t) = \mathbf{PMT}, \quad 0 \leqslant t \leqslant 1,}$$

где

$$\mathbf{R}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}, \quad \mathbf{P} = (\ \mathbf{P}_0 \quad \mathbf{P}_1 \quad \mathbf{P}_2 \quad \mathbf{P}_3\ ) = \begin{pmatrix} x_0 & x_1 & x_2 & x_3 \\ y_0 & y_1 & y_2 & y_3 \\ z_0 & z_1 & z_2 & z_3 \end{pmatrix},$$

$$\mathbf{M} = \frac{1}{6}\begin{pmatrix} 1 & -3 & 3 & -1 \\ 4 & 0 & -6 & 3 \\ 1 & 3 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} t^0 \\ t^1 \\ t^2 \\ t^3 \end{pmatrix}.$$

Матрица $\mathbf{M}$ называется *базисной матрицей B-сплайновой кривой.*

Свойства элементарных кубических B-сплайновых кривых

Свойства функциональных весовых множителей

$$\boxed{\begin{aligned} n_0(t) &= \frac{(1-t)^3}{6}, & n_1(t) &= \frac{3t^3 - 6t^2 + 4}{6}, \\ n_2(t) &= \frac{-3t^3 + 3t^2 + 3t + 1}{6}, & n_3(t) &= \frac{t^3}{6} \end{aligned}}$$

оказывают существенное влияние на поведение элементарной кубической B-сплайновой кривой. Укажем некоторые из них.

Функциональные коэффициенты $n_i(t)$:

1^+) неотрицательны,

2^+) в сумме составляют 1,

3^+) не зависят от точек массива $\mathbf{P}_0$, $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ (универсальны).

Элементарная кубическая B-сплайновая кривая:

1^+) лежит в выпуклой оболочке, порожденной вершинами $\mathbf{P}_0$, $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ опорной ломаной, и, как правило, не проходит ни через одну из них,

2^+) касательная в начальной точке

$$\frac{1}{6}(\mathbf{P}_0 + 4\mathbf{P}_1 + \mathbf{P}_2)$$

параллельна отрезку $\mathbf{P}_0\mathbf{P}_2$, а в концевой точке

$$\frac{1}{6}(\mathbf{P}_1 + 4\mathbf{P}_2 + \mathbf{P}_3)$$

— отрезку $\mathbf{P}_1\mathbf{P}_3$ (рис. 23).

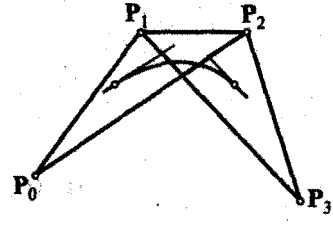


Рис. 23

Составные кубические B-сплайновые кривые

(*Составной*) *кубической B-сплайновой кривой*, определяемой массивом

$$\mathbf{P}_0, \ldots, \mathbf{P}_m, \quad m \geqslant 3,$$

называется кривая γ, которую можно представить в виде объединения элементарных кубических B-сплайновых кривых $\gamma^{(1)}, \ldots, \gamma^{(m-2)}$,

$$\gamma = \gamma^{(1)} \cup \ldots \cup \gamma^{(m-2)};$$

(i)-я кривая $\gamma^{(i)}$ описывается параметрическим уравнением следующего вида

$$\boxed{\begin{gathered}\mathbf{R}^{(i)}(t) = (\ \mathbf{P}_{i-1} \quad \mathbf{P}_i \quad \mathbf{P}_{i+1} \quad \mathbf{P}_{i+2}\ )\,\mathbf{M}\mathbf{T}, \\ 0 \leqslant t \leqslant 1, \quad i = 1, \ldots, m-2,\end{gathered}}$$

где $\mathbf{M}$ — базисная матрица кубической B-сплайновой кривой.

Единая параметризация. Рассматривая составную кривую γ как целое, более естественно пользоваться единой параметризацией.

Наиболее простой является параметризация с равноотстоящими целочисленными узлами. Для массива из $m+1$ опорных вершин составная B-сплайновая кривая строится из $m-2$ элементарных фрагментов. Если каждый из них определен на единичном отрезке, то длина общего промежутка изменения параметра должна быть равной $m-2$. Взяв 0 за начальную точку, получаем отрезок $[0, m-2]$. В этом случае узлы параметризации определяются по формуле

$$t_1 = 0, \quad t_{i+1} = t_i + 1, \quad i = 1, \ldots, m-3.$$

Описанный выбор отрезка параметризации позволяет записать уравнение составной кубической B-сплайновой кривой γ следующим образом

$$\mathbf{R} = \mathbf{R}(t), \quad t_1 \leqslant t \leqslant t_{m-2},$$

где

$$\mathbf{R} = \mathbf{R}^{(i)}(t) = (\ \mathbf{P}_{i-1} \quad \mathbf{P}_i \quad \mathbf{P}_{i+1} \quad \mathbf{P}_{i+2}\ )\,\mathbf{M} \begin{pmatrix} (t-t_i)^0 \\ (t-t_i)^1 \\ (t-t_i)^2 \\ (t-t_i)^3 \end{pmatrix},$$

$$t_i \leqslant t \leqslant t_{i+1}, \quad i = 1, \ldots, m-3,$$

— параметрическое векторное уравнение (i)-й элементарной кубической B-сплайновой кривой $\gamma^{(i)}$.

Свойства составной кубической B-сплайновой кривой.

Составная кубическая B-сплайновая кривая, порожденная массивом $\mathbf{P}_0, \ldots, \mathbf{P}_m$, $m \geqslant 3$:

1^+) является C^2-гладкой кривой в промежутке $[0, t_{m-2}]$, в точке стыка элементарных кривых $\gamma^{(i)}$ и $\gamma^{(i+1)}$ выполняются равенства (рис. 24)

$$\mathbf{R}^{(i)}(t_{i+1}) = \mathbf{R}^{(i+1)}(t_{i+1}) = \frac{1}{6}(\mathbf{P}_i + \mathbf{P}_{i+1} + \mathbf{P}_{i+2}),$$

$$\dot{\mathbf{R}}^{(i)}(t_{i+1}) = \dot{\mathbf{R}}^{(i+1)}(t_{i+1}) = \frac{1}{2}(-\mathbf{P}_i + \mathbf{P}_{i+2}),$$

$$\ddot{\mathbf{R}}^{(i)}(t_{i+1}) = \ddot{\mathbf{R}}^{(i+1)}(t_{i+1}) = \mathbf{P}_i - 2\mathbf{P}_{i+1} + \mathbf{P}_{i+2},$$

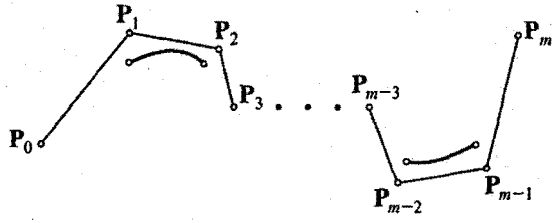


Рис. 24

2^+) как правило, не проходит ни через одну из точек заданного массива,

3^+) лежит в объединении $m-2$ выпуклых оболочек, порожденных четверками вершин (рис. 25)

$$\mathbf{P}_{i-1}, \mathbf{P}_i, \mathbf{P}_{i+1}, \mathbf{P}_{i+2}, \quad i = 1, \ldots, m-2,$$

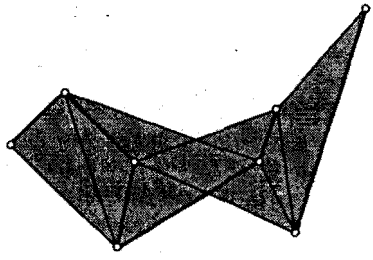


Рис. 25

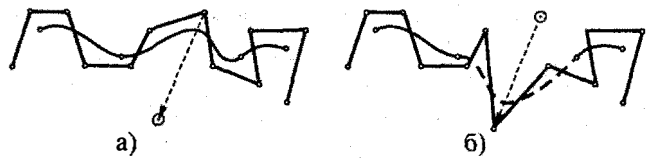
Рис. 26

4^+) «повторяет» контрольную ломаную (рис. 26) (в частности, число точек пересечения составной кубической B-сплайновой кривой с произвольной прямой не больше числа точек пересечения с этой прямой контрольной ломаной),

5^+) если опорные вершины $\mathbf{P}_0, \ldots, \mathbf{P}_m$ массива лежат на одной прямой, то составная кубическая B-сплайновая кривая также лежит на этой прямой (между вершинами $\mathbf{P}_0$ и $\mathbf{P}_m$),

6^+) если опорные вершины $\mathbf{P}_0, \ldots, \mathbf{P}_m$ массива лежат в одной плоскости, то составная кубическая B-сплайновая кривая также лежит в этой плоскости,

7^+) изменение одной вершины в массиве приводит к изменению только части кривой: при изменении вершины $\mathbf{P}_i$ нужно пересчитать параметрические уравнения только четырех элементарных кривых $\gamma^{(i-2)}$, $\gamma^{(i-1)}$, $\gamma^{(i)}$, $\gamma^{(i+1)}$ (рис. 27),



а) б)

Рис. 27

8^+) при добавлении в массив одной вершины возникает необходимость пересчета параметрических уравнений только четырех элементарных кривых, в описании которых эта вершина участвует.

Замечание. На взаимное расположение вершин в массиве не накладывается никаких ограничений: они могут и совпадать. Однако следует иметь в виду, что в подобных случаях кривая может потерять свою регулярность. Впрочем, если номера совпадающих вершин сильно разнятся, то никакой потери регулярности не происходит.

Случай, когда совпадают две или три первых (последних) вершины, рассматривается ниже.

Кратные и воображаемые вершины.

Составная кубическая B-сплайновая кривая, как правило, не проходит ни через одну вершину определяющего ее массива. Однако расположение ее начальной и конечной точек всегда известно: начальная точка $\mathbf{R}^{(1)}(0)$ составной кривой γ лежит в треугольнике $\mathbf{P}_0\mathbf{P}_1\mathbf{P}_2$, а конечная $\mathbf{R}^{(m)}(1)$ — в треугольнике $\mathbf{P}_{m-2}\mathbf{P}_{m-1}\mathbf{P}_m$. Подбором вспомогательных вершин и построением дополнительных элементарных кривых можно добиться того, чтобы начальная точка новой составной кривой выходила на отрезок $\mathbf{P}_0\mathbf{P}_1$, располагалась ближе к вершине $\mathbf{P}_0$ и даже совпадала с ней. Аналогичных результатов можно добиться и для конечной точки. Обычно это проводится путем использования *кратных* или *воображаемых* вершин.

А. Двойные вершины. Положим $\mathbf{P}_{-1} = \mathbf{P}_0$ и $\mathbf{P}_{m+1} = \mathbf{P}_m$ и построим две новые элементарные кривые $\gamma^{(0)}$ и $\gamma^{(m-1)}$, задав их параметрическими уравнениями следующего вида

$$\mathbf{R}^{(0)}(t) = (n_0(t) + n_1(t))\mathbf{P}_0 + n_2(t)\mathbf{P}_1 + n_3(t)\mathbf{P}_2,$$
$$\mathbf{R}^{(m-1)}(t) = n_0(t)\mathbf{P}_{m-2} + n_1(t)\mathbf{P}_{m-1} + (n_2(t) + n_3(t))\mathbf{P}_m,$$

где $0 \leqslant t \leqslant 1$.

С учетом кривых $\gamma^{(0)}$ и $\gamma^{(m-1)}$ новая составная кубическая B-сплайновая кривая

$$\gamma^* = \gamma^{(0)} \cup \gamma^{(1)} \cup \ldots \cup \gamma^{(m-2)} \cup \gamma^{(m-1)}$$

будет начинаться в точке

$$\mathbf{R}^{(0)}(0) = \frac{5}{6}\,\mathbf{P}_0 + \frac{1}{6}\,\mathbf{P}_1,$$

касаясь отрезка $\mathbf{P}_0\mathbf{P}_1$,

$$\dot{\mathbf{R}}_2(0) = \frac{1}{2}\,(\mathbf{P}_1 - \mathbf{P}_0),$$

и заканчиваться в точке

$$\mathbf{R}^{(m-1)}(1) = \frac{1}{6}\,\mathbf{P}_{m-1} + \frac{5}{6}\,\mathbf{P}_m,$$

касаясь отрезка $\mathbf{P}_{m-1}\mathbf{P}_m$,

$$\dot{\mathbf{R}}^{(m-1)}(1) = \frac{1}{2}\,(\mathbf{P}_m - \mathbf{P}_{m-1})$$

(рис. 28). Кроме того, кривая γ^* будет иметь в двух этих точках нулевую кривизну.

Б. Тройные вершины. Положим $\mathbf{P}_{-2} = \mathbf{P}_{-1} = \mathbf{P}_0$ и $\mathbf{P}_{m+2} = \mathbf{P}_{m+1} = \mathbf{P}_m$ и возьмем в качестве двух новых элементарных кривых $\gamma^{(-1)}$ и $\gamma^{(m)}$ прямолинейные отрезки

$$\mathbf{R}^{(-1)}(t) = \left(1 - \frac{t^3}{6}\right)\mathbf{P}_0 + \frac{t^3}{6}\,\mathbf{P}_1,$$
$$\mathbf{R}^{(m)}(t) = \frac{(1-t)^3}{6}\,\mathbf{P}_{m-1} + \left(1 - \frac{(1-t)^3}{6}\right)\mathbf{P}_m,$$

где $0 \leqslant t \leqslant 1$ (рис. 29).

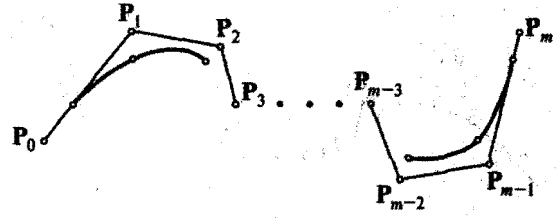


Рис. 28

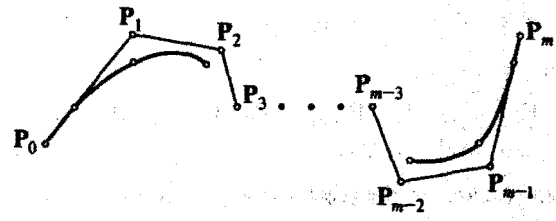


Рис. 29

С учетом кривых $\gamma^{(0)}$ и $\gamma^{(m-1)}$ новая составная B-сплайновая кривая

$$\gamma^{**} = \gamma^{(-1)} \cup \gamma^{(0)} \cup \gamma^{(1)} \cup \ldots \cup \gamma^{(m-2)} \cup \gamma^{(m-1)} \cup \gamma^{(m)}$$

будет начинаться в вершине $\mathbf{P}_0$ и заканчиваться в вершине $\mathbf{P}_m$.

в. Воображаемые вершины. Подбором дополнительных вершин $\mathbf{P}_{-1}$ и $\mathbf{P}_{m+1}$ к массиву

$$\mathbf{P}_0, \ldots, \mathbf{P}_m, \quad m \geqslant 3,$$

можно добиться выполнения различных условий на концах составной кривой.

Например, составная B-сплайновая кривая, построенная по новому массиву

$$\mathbf{P}_{-1}, \mathbf{P}_0, \ldots, \mathbf{P}_m, \mathbf{P}_{m+1},$$

где

$$\mathbf{P}_{-1} = (\mathbf{P}_0 - \mathbf{P}_1) + \mathbf{P}_0, \quad \mathbf{P}_{m+1} = (\mathbf{P}_m - \mathbf{P}_{m-1}) + \mathbf{P}_m,$$

будет начинаться в вершине $\mathbf{P}_0$, касаясь отрезка $\mathbf{P}_0\mathbf{P}_1$,

$$\dot{\mathbf{R}}^{(0)}(0) = \mathbf{P}_1 - \mathbf{P}_0,$$

и заканчиваться в вершине $\mathbf{P}_m$, касаясь отрезка $\mathbf{P}_{m-1}\mathbf{P}_m$,

$$\dot{\mathbf{R}}^{(m-1)}(1) = \mathbf{P}_m - \mathbf{P}_{m-1}$$

(рис. 30). Кривизны новой кривой в точках $\mathbf{P}_0$ и $\mathbf{P}_m$, вообще говоря, отличны от нуля.

Замечание. Дополнительные вершины $\mathbf{P}_{-1}$ и $\mathbf{P}_{m+1}$ можно выбрать так, чтобы в концах новой составной кривой первые или вторые производные радиус-векторов $\mathbf{R}^{(0)}(t)$ и $\mathbf{R}^{(m-1)}(t)$ кривых $\gamma^{(0)}$ и $\gamma^{(m-1)}$ совпадали с заданными значениями (при $t = 0$ и $t = 1$ соответственно).

Построение замкнутой кривой.

Чтобы по заданному массиву

$$\mathbf{P}_0, \ldots, \mathbf{P}_m, \quad m \geqslant 3,$$

построить C^2-гладкую замкнутую кривую, достаточно выбрать дополнительные вершины $\mathbf{P}_{m+1}$, $\mathbf{P}_{m+2}$, $\mathbf{P}_{m+3}$ по правилу

$$\mathbf{P}_{m+1} = \mathbf{P}_0, \quad \mathbf{P}_{m+2} = \mathbf{P}_1, \quad \mathbf{P}_{m+3} = \mathbf{P}_2$$

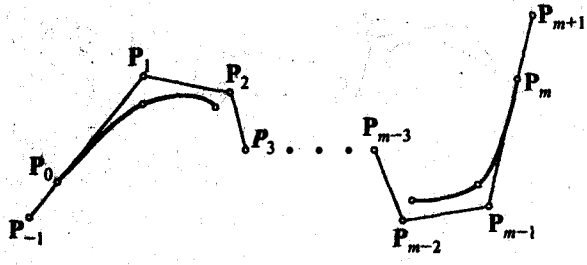


Рис. 30

и рассмотреть массив

$$\mathbf{P}_0,\ \mathbf{P}_1,\ \mathbf{P}_2,\ \ldots\ ,\ \mathbf{P}_m,\ \mathbf{P}_{m+1}=\mathbf{P}_0,\ \mathbf{P}_{m+2}=\mathbf{P}_1,\ \mathbf{P}_{m+3}=\mathbf{P}_2$$

(при условии, что $t_{i+1}=t_i+1$, $t_1=0$).

2.3. Параметрические уравнения бикубической поверхности Безье

Элементарная *бикубическая* поверхность Безье определяется шестнадцатью вершинами

$$\begin{array}{cccc}\mathbf{P}_{00}, & \mathbf{P}_{01}, & \mathbf{P}_{02}, & \mathbf{P}_{03},\\ \mathbf{P}_{10}, & \mathbf{P}_{11}, & \mathbf{P}_{12}, & \mathbf{P}_{13},\\ \mathbf{P}_{20}, & \mathbf{P}_{21}, & \mathbf{P}_{22}, & \mathbf{P}_{23},\\ \mathbf{P}_{30}, & \mathbf{P}_{31}, & \mathbf{P}_{32}, & \mathbf{P}_{33}.\end{array}$$

Используя многочлены Бернштейна, эту поверхность можно задать так

$$\boxed{\mathbf{R}(u,v)=\sum_{i=0}^{3}B_i^3(u)\left(\sum_{j=0}^{3}B_j^3(v)\mathbf{P}_{ij}\right),\quad 0\leqslant u,\quad v\leqslant 1,}$$

или, в матричной форме,

$$\mathbf{R}(u,v)=\begin{pmatrix}B_0^3(u) & B_1^3(u) & B_2^3(u) & B_3^3(u)\end{pmatrix}\begin{pmatrix}\mathbf{P}_{00} & \mathbf{P}_{01} & \mathbf{P}_{02} & \mathbf{P}_{03}\\ \mathbf{P}_{10} & \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13}\\ \mathbf{P}_{20} & \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23}\\ \mathbf{P}_{30} & \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33}\end{pmatrix}\begin{pmatrix}B_0^3(v)\\ B_1^3(v)\\ B_2^3(v)\\ B_3^3(v)\end{pmatrix}.$$

Последнюю формулу часто записывают и так

$$\boxed{\mathbf{R}(t)=\mathbf{U}^T\mathbf{M}^T\mathbf{P}\mathbf{M}\mathbf{V},\quad 0\leqslant u,\quad v\leqslant 1,}$$

где

$$\mathbf{R}(u,v)=\begin{pmatrix}x(u,v)\\ y(u,v)\\ z(u,v)\end{pmatrix},\quad \mathbf{U}=\begin{pmatrix}u^0\\ u^1\\ u^2\\ u^3\end{pmatrix},\quad \mathbf{V}=\begin{pmatrix}v^0\\ v^1\\ v^2\\ v^3\end{pmatrix},$$

$$\mathbf{P}=\begin{pmatrix}\mathbf{P}_{00} & \mathbf{P}_{01} & \mathbf{P}_{02} & \mathbf{P}_{03}\\ \mathbf{P}_{10} & \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13}\\ \mathbf{P}_{20} & \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23}\\ \mathbf{P}_{30} & \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33}\end{pmatrix},\quad \mathbf{M}=\begin{pmatrix}1 & -3 & 3 & 1\\ 0 & 3 & -6 & 3\\ 0 & 0 & 3 & -3\\ 0 & 0 & 0 & 1\end{pmatrix}.$$

Матрица M называется *базисной матрицей бикубической поверхности Безье.*

Свойства элементарных поверхностей Безье

Свойства элементарных поверхностей Безье являются прямыми следствиями свойств элементарных кривых Безье.

Элементарная поверхность Безье, порожденная массивом $\mathbf{P}$:

1^+) является гладкой поверхностью, в частности, первые производные радиус-вектора $\mathbf{R}(u, v)$ можно записать так

$$\mathbf{R}_u(u, v) = \frac{\partial}{\partial u}\mathbf{R}(u, v) = m\sum_{i=0}^{m-1}\sum_{j=0}^{n}(\mathbf{P}_{i+1,j} - \mathbf{P}_{i,j})B_i^{m-1}(u)B_j^n(v),$$

$$\mathbf{R}_v(u, v) = \frac{\partial}{\partial v}\mathbf{R}(u, v) = n\sum_{i=0}^{m}\sum_{j=0}^{n-1}(\mathbf{P}_{i,j+1} - \mathbf{P}_{i,j})B_i^m(u)B_j^{n-1}(v);$$

если векторы $\mathbf{R}_u(u, v)$ и $\mathbf{R}_v(u, v)$ неколлинеарны, то определен единичный вектор нормали поверхности Безье $\mathbf{N}(u, v)$,

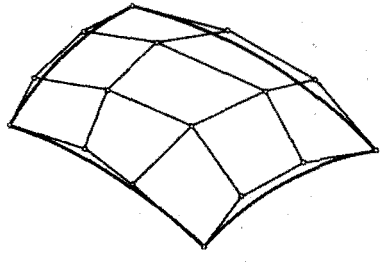


Рис. 31

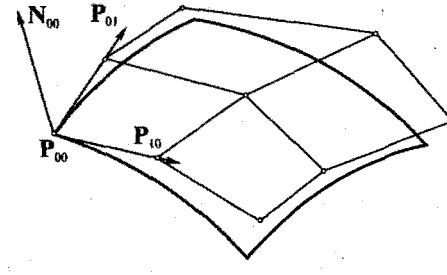


Рис. 32

2^+) граничные кривые элементарной поверхности Безье суть элементарные кривые Безье соответствующих степеней, их опорные ломаные — границы опорной многогранной поверхности (опорного графа), в частности, граничная кривая, описываемая радиус-вектором $\mathbf{R}(0, v)$, является элементарной кривой Безье n-й степени с опорным массивом вершин $\mathbf{P}_{00}, \ldots, \mathbf{P}_{0n}$ (рис. 31); все четыре угловые вершины опорного многогранника $\mathbf{P}_{00}$, $\mathbf{P}_{0n}$, $\mathbf{P}_{m0}$ и $\mathbf{P}_{mn}$ лежат на поверхности Безье,

$$\mathbf{R}(0, 0) = \mathbf{P}_{00}, \quad \mathbf{R}(1, 0) = \mathbf{P}_{m0}, \quad \mathbf{R}(0, 1) = \mathbf{P}_{0n}, \quad \mathbf{R}(1, 1) = \mathbf{P}_{mn},$$

и поверхность касается угловых граней опорного многогранника, например, для угловой вершины $\mathbf{P}_{00}$ ($u = v = 0$), $\mathbf{R}(0, 0) = \mathbf{P}_{00}$, выполняются равенства

$$\mathbf{R}_u(0, 0) = m(\mathbf{P}_{10} - \mathbf{P}_{00}), \quad \mathbf{R}_v(0, 0) = n(\mathbf{P}_{01} - \mathbf{P}_{00})$$

(касательные векторы элементарной поверхности Безье в угловой вершине коллинеарны звеньям соответствующих граничных опорных ломаных, исходящих из этой вершины (рис. 32)),

$$\mathbf{R}_{uv}(0, 0) = mn\big((\mathbf{P}_{11} - \mathbf{P}_{10}) - (\mathbf{P}_{01} - \mathbf{P}_{00})\big)$$

(вектор скручивания $\mathbf{R}_{uv}(0, 0)$ в угловой вершине $\mathbf{P}_{00}$ только множителем отличается от вектора скручивания билинейной поверхности, порожденной четверкой вершин

$$\mathbf{P}_{01}, \quad \mathbf{P}_{01}, \quad \mathbf{P}_{10}, \quad \mathbf{P}_{11},$$

и оценивает степень отклонения опорной вершины $\mathbf{P}_{11}$ от касательной плоскости элементарной поверхности Безье в вершине $\mathbf{P}_{00}$),

$$\mathbf{N}(0, 0) = \frac{(\mathbf{P}_{10} - \mathbf{P}_{00}) \times (\mathbf{P}_{01} - \mathbf{P}_{00})}{|(\mathbf{P}_{10} - \mathbf{P}_{00}) \times (\mathbf{P}_{01} - \mathbf{P}_{00})|}$$

(вектор нормали перпендикулярен плоскости треугольника $\mathbf{P}_{00}\mathbf{P}_{01}\mathbf{P}_{10}$) (для каждой из трех других угловых вершин $\mathbf{P}_{01}$ ($u = 0$, $v = 1$), $\mathbf{P}_{10}$ ($u = 1$, $v = 0$), $\mathbf{P}_{11}$ ($u = v = 1$) выполняются аналогичные соотношения),

3^+) элементарная поверхность Безье лежит в выпуклой оболочке, порожденной массивом $\mathbf{P}$,

4^+) элементарная поверхность Безье «повторяет» опорную многогранную поверхность,

5^+) если все вершины массива $\mathbf{P}$ лежат в одной плоскости, то определяемая этим массивом элементарная поверхность Безье представляет собой плоский криволинейный четырехугольник, лежащий в этой плоскости,

6^-) изменение хотя бы одной вершины в массиве приводит к заметному изменению всей поверхности Безье.

Предметный указатель

К

Л

М

Н

О

П

Р

С

Т

У

Ф

Ц

Ч

Ш

Э

Оглавление