Министерство науки и образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ

ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

А.С. Курапцев, А. Н. Литвинов,

К.А. Баранцев, Н. В. Ларионов,

Решение дифференциальных уравнений в задачах теоретической физики с использованием методов математического моделирования и оптимизации

Учебное пособие

Санкт-Петербург

2019

Аннотация

Пособие соответствует рабочим программам дисциплин «Математическое моделирование» и «Методы моделирования и оптимизации», относящимся к базовому модулю направления подготовки магистров по направлениям 11.04.02 «Инфокоммуникационные технологии и системы связи» и 11.04.04 «Электроника и наноэлектроника».

Учебное пособие содержит изложение теоретических и методических основ математического моделирования. Приведена общая методология математического моделирования. Представлены основные методы численного моделирования, такие как конечно-разностный метод, метод конечных элементов, методы Монте-Карло. Приведены базовые принципы математической оптимизации.

Пособие предназначено для студентов высших учебных заведений, обучающихся по направлениям подготовки 11.04.02 «Инфокоммуникационные технологии и системы связи» и 11.04.04 «Электроника и наноэлектроника» по дисциплинам «Математическое моделирование» и «Методы моделирования и оптимизации», а также для студентов, обучающихся по специальности 16.04.01 «Техническая физика».

Содержание

Введение

В основе математического моделирования лежитмысленная замена реального физического объекта его образом, который отражает наиболее важные свойства изучаемого объекта.Этот образ изучают при помощи аналитического аппарата и ЭВМ. Математическое моделирование сочетает в себе как преимущества теории, так и эксперимента. Моделирование объекта безопаснее, дешевле и быстрее чем работа непосредственно с самим физическим объектом. Эти преимущества характерны для теоретических исследований. В то же время, современные вычислительные мощности позволяют строить достаточно подробные математические модели, что позволяет отразить изучаемый объект в его полноте. Это преимущество присуще эксперименту.

В настоящее время математическое моделирование активно используется не только в физике и технике, но также и в экологии, экономике, в социально-политических проектах. При построении математической модели предпочтителен путь «от задачи к методу».

Процесс моделирования можно условно разбить на три этапа: модель, алгоритм и программа.

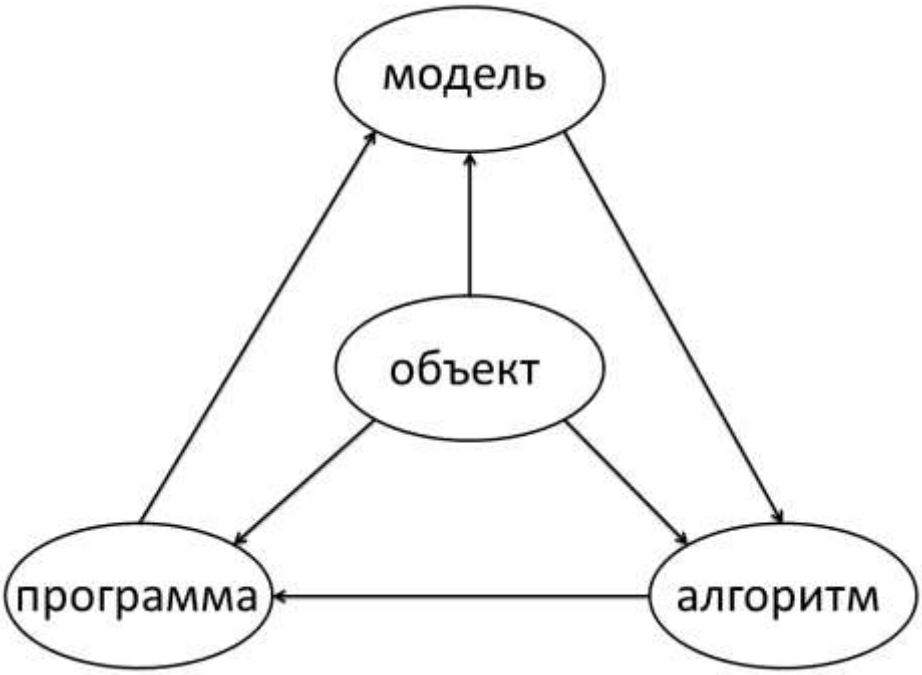

Рис. 1. Триада «модель – алгоритм – программа»

На каждом шаге построении триады, схематично изображенной на рис. 1, возникают погрешности. Перечислим их виды:

1. Погрешность модели. Она обусловлена тем фактом, что ни одна математическая модель не может учесть все физические факторы, оказывающие влияние на реальный объект. По отношению к алгоритму эта погрешность неустранима.
2. Погрешность дискретизации. Эта погрешность обусловлена тем, что при численных расчетах используется не исходная модель, а его аналог, приспособленный для вычислений на ЭВМ. Погрешность дискретизации – это разность решений исходной модели и её численного аналога.
3. Погрешность округления (или вычислительная погрешность). Это отличие точного решения аналога математической модели, приспособленного для ЭВМ, от реального численного решения.

Математическая модель должна удовлетворять следующим требованиям:

1. Полнота. Это требование означает учёт всех факторов, существенно влияющих на объект и определяющих его поведение.
2. Адекватность. Под этим понимают правильное качественное и достаточно точное количественное описание объекта.
3. Точность.
4. Экономичность.
5. Робастность (или устойчивость).
6. Продуктивность (т.е. реальность исходных данных).
7. Наглядность.

К вычислительному алгоритму предъявляются следующие требования:

1. Вычислительный алгоритм должен обеспечивать решение задачи с любой наперед заданной точностью за конечное число операций.
2. Устойчивость. Под устойчивостью алгоритма понимают то, что вычислительная погрешность в процессе работы алгоритма возрастает незначительно. Т.е. вычислительная погрешность должна быть много меньше вычисляемой величины, если последняя отлична от нуля.
3. Алгоритм должен обеспечивать приемлемые значения величин в процессе вычисления (в частности, не меньше машинного нуля и не больше машинной бесконечности).
4. Приемлемое время вычислений и объем требуемой оперативной памяти.

Математические модели могут строиться различными способами:

- Из фундаментальных законов природы (например, из закона сохранения энергии, закона сохранения импульса).
- Из вариационных принципов.
- Использованием аналогий с ранее изученными явлениями, в том числе из других областей науки. Этот способ обусловлен свойством универсальности математических моделей: одна и та же модель может описывать принципиально разные объекты и явления.
- По иерархическому принципу. Т.е. обобщением уже готовых моделей путём учета новых факторов.

Глава 1.Методология математического моделирования

§1.1. Математические модели из фундаментальных законов природы

Способ построения математических моделей из фундаментальных законов природы достаточно распространен. В этом параграфе мы рассмотрим простейшиепримеры моделей, полученные данным способом.

Задача №1.1. Небольшое количество радиоактивного вещества в области 1 окружено толстым слоем свинца (область 2), Рис. 2. В начальный момент времени массы радиоактивного вещества и свинцового экрана равны $M_1(0)$ и $M_2(0)$ соответственно. Длина свободного пробега частиц в области 1 много больше характерных линейных размеров области,$\lambda_1 \gg L_1$, в области 2 верно обратное соотношение,$\lambda_2 \ll L_2$. Определить зависимость масс вещества и свинцового экрана от времени $M_1(t)$, $M_2(t)$.

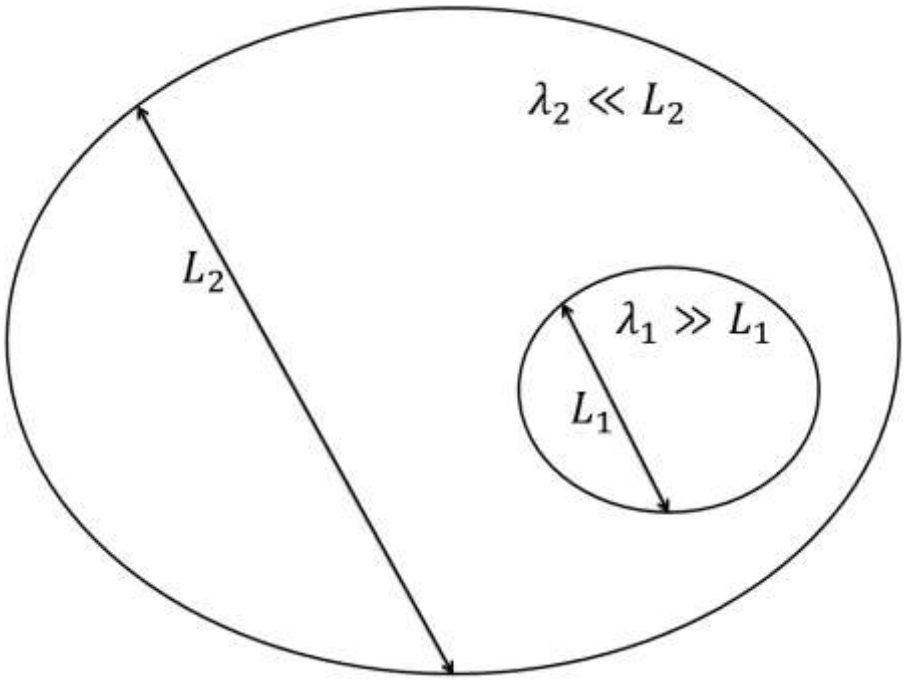

Рис. 2. Иллюстрация к задаче №1.1

Сделаем допущение, что все испущенные частицы радиоактивного вещества свободно проходят область 1 и все поглощаются в области 2. Воспользуемся законом сохранения массы.

$$M_1(0) + M_2(0) = M_1(t) + M_2(t) \qquad (1.1.1)$$

Также используем закон радиоактивного распада.

$$\frac{dN_1(t)}{dt} = -\alpha N_1(t)$$

В этой формуле $N_1(t)$ – количество частиц радиоактивного вещества, оставшихся в области 1 к моменту времени t, α – константа радиоактивного распада. Обозначив массу одной частицы μ, имеем

$$M_1(t) = \mu N_1(t)$$

Следовательно, уравнение для функции $M_1(t)$ выглядит следующим образом

$$\frac{dM_1(t)}{dt} = -\alpha M_1(t) \qquad (1.1.2)$$

В этой задаче математическую модель составляют уравнения (1.1.1) и (1.1.2), условия $\lambda_1 \gg L_1$ и $\lambda_2 \ll L_2$, а также параметры $M_1(0)$, $M_2(0)$ иα.

Уравнение (1.1.2) элементарно решается разделением переменных.

$$M_1(t) = M_1(0)exp(-\alpha t) \qquad (1.1.3)$$

Из закона сохранения массы (1.1.1) находим зависимость $M_2(t)$.

$$M_2(t) = M_2(0) + M_1(0)\big(1 - exp(-\alpha t)\big) \qquad (1.1.4)$$

Нетрудно видеть, что на больших временах $t \gg \alpha^{-1}$асимптотика зависимостей (1.1.3) и (1.1.4) следующая $M_1(t) \to 0$, $M_2(t) \to M_2(0) + M_1(0)$. Это означает, что по истечении достаточно большого промежутка времени (по сравнению со временем распада) все частицы из области 1 перейдут в область 2.

Задача №1.2. Определить закон движения шарика, присоединенного к пружинке (рис. 3). Длина пружинки в свободном состоянии l_0, диаметр витков пружинки D, количество витков N, диаметр проволоки, из которой сделана пружинка,d, масса шарика m. Трением о поверхность и сопротивлением воздуха пренебречь.

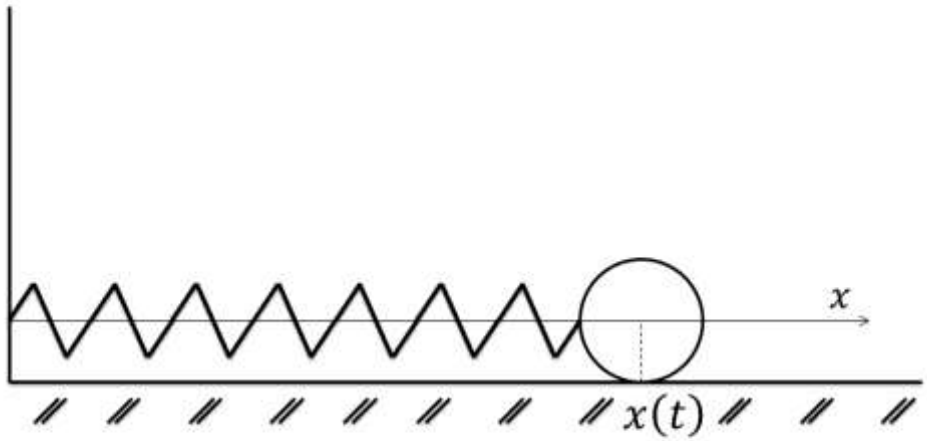

Рис. 3. Иллюстрация к задаче 1.2

В приближении малых отклонений шарика от положения равновесия ($\Delta x = |x - l_0| \ll l_0$) предполагаем линейную зависимость возвращающей силы от смещения $F_x = -k(x - l_0)$. Используем второй закон Ньютона.

$$m\frac{d^2(x-l_0)}{dt^2} = -k(x-l_0) \quad (1.2.1)$$

Решение уравнения (1.2.1) в общем виде даёт

$$x - l_0 = A\sin\omega t + B\cos\omega t \quad (1.2.2)$$

В этой формуле $\omega = \sqrt{k/m}$. Дифференцируя выражение (1.2.2) по времени, найдём проекцию скорости на ось x,$v(t)$.

$$v(t) = A\omega\cos\omega t - B\omega\sin\omega t \quad (1.2.3)$$

Обозначим начальную координату и проекцию начальной скорости на осьx, x_0 и v_0 соответственно. Тогда, подставляя в выражения (1.2.2) и (1.2.3) значение $t = 0$, имеем

$$\begin{cases} x_0 - l_0 = B \\ v_0 = A\omega \end{cases} \quad (1.2.4)$$

Отсюда получаем значения констант A и B. Подставим эти значения в (1.2.2).

$$x(t) = \frac{v_0}{\omega}\sin\omega t + (x_0 - l_0)\cos\omega t + l_0 \quad (1.2.5)$$

Отметим, что в решение (1.2.5) не входит целый ряд параметров объекта: диаметр витков пружинки, их количество, а также диаметр проволоки, из которой сделана пружинка. Кроме того, мы не учитывали трение о поверхность и сопротивление воздуха. Это говорит о том, что построенная математическая модель очень упрощенная и не учитывает многих свойств реального физического объекта. В дальнейшем мы вернемся к рассмотрению этой физической системы при обсуждении иерархического принципа построения моделей.

§1.2. Математические модели из вариационных принципов

Способ построения математических моделей из вариационных принципов также является весьма распространенным. Суть этого способа состоит в том, что из множества вариантов поведения объекта выбираются те, которые удовлетворяют определенному условию. Рассмотрим простой пример.

Задача №2.1. Мотоциклисту требуется доехать из точки А до точки В за минимальное время, но обязательно побывав в любой точке прямой С (рис. 4). Определить траекторию, по которой ему следует двигаться.

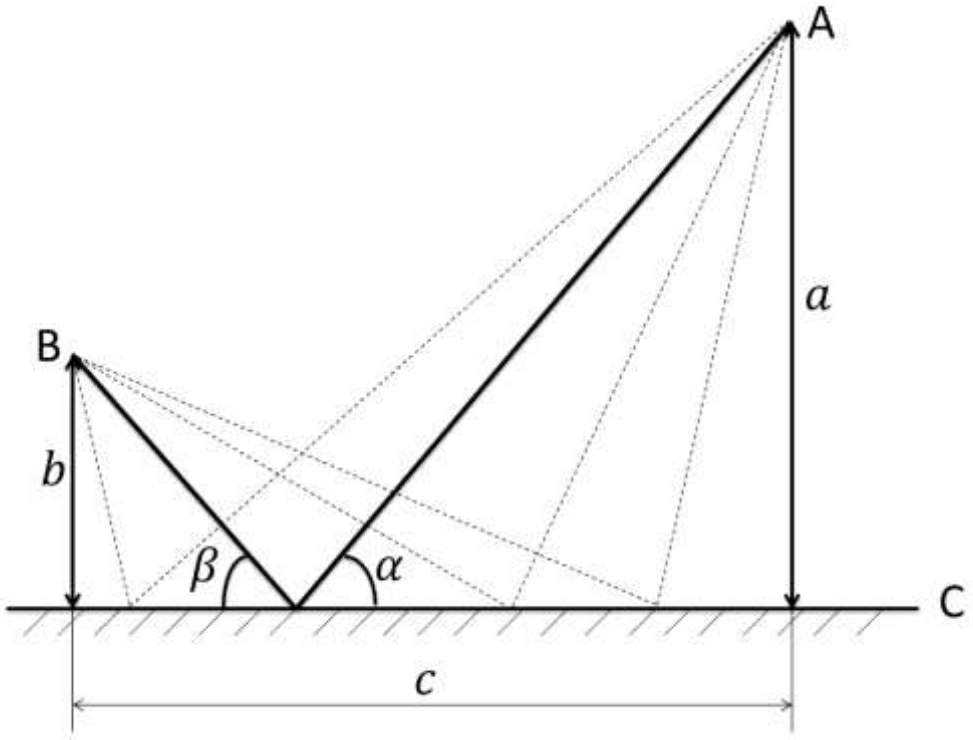

Рис. 4. Иллюстрация к задаче №2.1

На рис. 4 показаны возможные траектории движения мотоциклиста. Очевидно, что углы α и β связаны между собой, потому что если зафиксировать один из углов, это однозначно задаст точку на прямой C, а по ней можно вычислить второй угол. Таким образом, $\beta = \beta(\alpha)$. Следовательно, время пути является функцией только одного угла α.

$$t(\alpha) = \frac{a}{v \sin \alpha} + \frac{b}{v \sin \beta(\alpha)} \tag{1.2.1}$$

В этой формуле v – скорость мотоциклиста. Для нахождения минимума функции $t(\alpha)$ необходимо приравнять нулю её первую производную. Сократив скорость, получаем следующее уравнение

$$\frac{a \cos \alpha}{\sin^2 \alpha} + \frac{b \cos \beta(\alpha)}{\sin^2 \beta(\alpha)} \frac{d\beta}{d\alpha} = 0 \tag{1.2.2}$$

Помимо уравнения (1.2.2) нам требуется уравнение, связывающее углы β и α.

$$c = \frac{a}{\tan \alpha} + \frac{b}{\tan \beta(\alpha)} \tag{1.2.3}$$

Продифференцируем уравнение (1.2.3) по углу α.

$$\frac{a}{\sin^2\alpha}+\frac{b}{\sin^2\beta(\alpha)}\frac{d\beta}{d\alpha}=0 \qquad (1.2.4)$$

Выразив из (1.2.4) производную $d\beta/d\alpha$,и подставив её значение в (1.2.2), нетрудно найти, что $\cos\alpha=\cos\beta(\alpha)$.Так как α и β – острые углы, делаем вывод о том, что$\alpha=\beta$.

Заметим, что по такому же закону движется луч света при отражении от плоской границы раздела двух сред с различными показателями преломления. В оптике принцип наименьшего времени называется принципом Ферма.

Нетрудно видеть, что в рассмотренной задаче зависимость $\beta(\alpha)$ могла быть непосредственно выражена из условия (1.2.3) и затем подставлена в выражение (1.2.1). Однако такой путь привёл бы к более громоздким математическим выкладкам.

Вариационные принципы находят наибольшее применение в тех случаях, когда минимизируемая величина является функционалом, записанным в виде интеграла от некоторой функции.Рассмотрим такую постановку задачи. Пусть $y(x)$ – функция, заданная на отрезке $a<x<b$, и известны её значения на концах этого отрезка $y(a)=A, y(b)=B$. Требуется среди всего множества таких функций определить ту, которая обеспечивает экстремум следующего функционала

$$J[y]=\int_a^b F(x,y,y')dx$$

Здесь $F(x,y,y')$ - произвольная дифференцируемая функция, зависящая от аргумента x, функции y и её производной y'. Обозначим $h(x)$ приращение к функции $y(x)$. Так как значения $y(x)$ на краях отрезка зафиксированы, приращение $h(x)$ должно удовлетворять однородным граничным условиям $h(a)=h(b)=0$. Запишем соответствующее приращение функционала$\Delta J=J[y+h]-J[y]$.

$$\Delta J=\int_a^b F(x,y+h,y'+h')\,dx-\int_a^b F(x,y,y')dx \qquad (1.2.6)$$

Разложив подынтегральную функцию в выражении (1.2.6) в ряд Тейлора, имеем

$$\Delta J=\int_a^b\left[\frac{\partial F(x,y,y\prime)}{\partial y}h+\frac{\partial F(x,y,y\prime)}{\partial y\prime}h'\right]dx+\cdots \qquad (1.2.7)$$

Многоточием в формуле (1.2.7) обозначены члены второго и более высоких порядков поh и h'.Дифференциалом (или вариацией)

функционала $J[y]$ является главная часть приращения, записанного в выражении (1.2.7). Обозначим вариацию δJ.

$$\delta J = \int_a^b \left[\frac{\partial F(x,y,y\prime)}{\partial y} h + \frac{\partial F(x,y,y\prime)}{\partial y\prime} h'\right] dx \quad (1.2.8)$$

Необходимым условием экстремума функционала является равенство нулю его вариации $\delta J = 0$. Для решения этого уравнения проинтегрируем по частям второе слагаемое в выражении (1.2.8).

$$\int_a^b \frac{\partial F(x,y,y')}{\partial y'} h'\, dx = \frac{\partial F(x,y,y')}{\partial y'} h\Big|_a^b - \int_a^b h \frac{d}{dx}\left(\frac{\partial F(x,y,y')}{\partial y'}\right)$$

Внеинтегральное слагаемое в этом выражении обращается в нуль в силу однородных граничных условий для функции $h(x)$. Таким образом, приравнивание к нулю вариации приводит к следующему уравнению

$$\int_a^b h(x) \left[\frac{\partial F(x,y,y\prime)}{\partial y} - \frac{d}{dx}\left(\frac{\partial F(x,y,y\prime)}{\partial y\prime}\right)\right] dx = 0 \quad (1.2.9)$$

Равенство (1.2.9) должно выполняться для любой функции $h(x)$, следовательно, выражение в квадратных скобках должно тождественно равняться нулю. Записав это в более компактном виде, получаем окончательное необходимое условие экстремума функционала $J[y]$.

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0 \quad (1.2.10)$$

Формула (1.2.10) называется уравнением Эйлера-Лагранжа, это основное уравнение вариационного исчисления.

В задачах вариационного исчисления часто требуется найти экстремум функционала не на всём множестве функций $y(x)$, удовлетворяющих определенным граничным условиям, а на множестве таких функций, которые обеспечивают постоянство другого функционала. Такие задачи называются изопериметрическими. Рассмотрим пример изопериметрической задачи.

Задача №2.2. (задача Дидоны). Это исторически первая изопериметрическая задача, она связана с древней легендой об основании города Карфагена на южном побережье Средиземного Моря. При помощи веревки из шкуры быка, концы которой должны были находиться на побережье, требовалось очертить такую кривую, которая охватывала бы максимальную площадь под строительство города.

Будем считать береговую линию прямой. Расположим оси декартовой системы координат как показано на рис. 5.

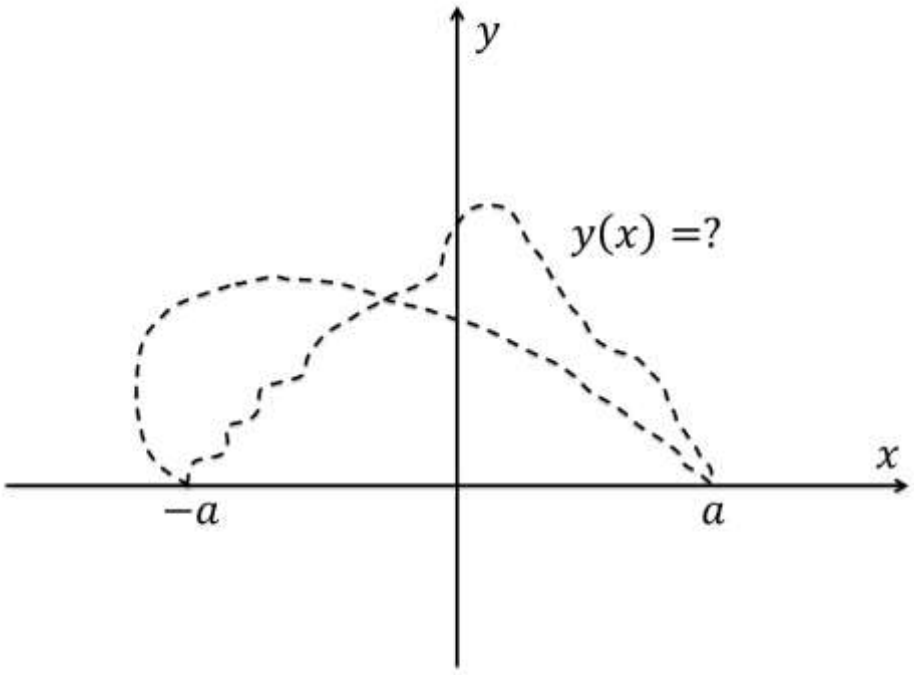

Рис. 5. Иллюстрация к задаче №2.2

Обозначим длину веревки $2l$ ($l > a$). Запишем длину кривой.

$$L[y] = \int_{-a}^{a} \sqrt{1 + y'^2}\, dx = 2l \qquad (1.2.11)$$

$L[y]$ – это функционал, зависящий от функции $y(x)$. Граничные условия для$y(x)$, в соответствии с рис. 5, следующие: $y(-a) = y(a) = 0$. Площадь под графиком функции $y(x)$ равна

$$S[y] = \int_{-a}^{a} y\, dx \to max$$

$S[y]$ также является функционалом. Задача состоит в отыскании такой функции $y(x)$, которая обеспечивает максимизацию функционала площади $S[y]$ при условии постоянства функционала длины $L[y]$. Такая задача решается методом множителей Лагранжа. Составим функционал

$$J[y] = S[y] + \lambda L[y]$$

Здесь λ обозначен множитель Лагранжа. Метод множителей Лагранжа состоит в том, что мы вначале решаем задачу о нахождении экстремума объединенного функционала $J[y]$ на всём множестве функций

$y(x)$, удовлетворяющих заданным граничным условиям, (это делается путём решения соответствующего уравнения Эйлера-Лагранжа), а затем находим параметр λ из условия постоянства функционала (1.2.11).

$$J[y] = \int_{-a}^{a}\left(y + \lambda\sqrt{1 + y'^2}\right) dx \qquad (1.2.12)$$

Подставляя подынтегрального функцию функционала (1.2.12) в уравнение Эйлера-Лагранжа (1.2.10), имеем

$$\frac{\partial}{\partial y}\left(y + \lambda\sqrt{1 + y'^2}\right) - \frac{d}{dx}\left[\frac{\partial}{\partial y'}\left(y + \lambda\sqrt{1 + y'^2}\right)\right] = 0 \qquad (1.2.13)$$

Вычисление частных производных в (1.2.13) даёт выражение

$$1 - \lambda\frac{d}{dx}\left(\frac{y'}{\sqrt{1 + y'^2}}\right) = 0$$

Проинтегрировав последнее выражение, получим

$$x - \lambda\frac{y'}{\sqrt{1+y'^2}} = C_1 \qquad (1.2.14)$$

В формуле (1.2.14) C_1 обозначена константа интегрирования. Выразив из (1.2.14) y', приходим к выражению

$$\frac{dy}{dx} = \pm\frac{x - C_1}{\sqrt{\lambda^2 - (x - C_1)^2}}$$

Интегрируем полученное выражение

$$y = C_2 \pm \int \frac{x-C_1}{\sqrt{\lambda^2-(x-C_1)^2}}\, dx \qquad (1.2.15)$$

Здесь C_2 – константа интегрирования. Остаётся только вычислить интеграл в формуле (1.2.15).

$$y = C_2 \pm \frac{1}{2}\int \frac{d((x - C_1)^2)}{\sqrt{\lambda^2 - (x - C_1)^2}}$$
$$= C_2 \mp \frac{1}{2}\int \frac{d(\lambda^2 - (x - C_1)^2)}{\sqrt{\lambda^2 - (x - C_1)^2}} = C_2 \mp \frac{1}{2} * 2\sqrt{\lambda^2 - (x - C_1)^2}$$

Таким образом, приходим к выражению

$$(x - C_1)^2 + (y - C_2)^2 = \lambda^2 \qquad (1.2.16)$$

Уравнение (1.2.16) описывает окружность с центом(C_1, C_2) и радиусом $|\lambda|$.Без ограничения общности можно полагать $\lambda > 0$. Константы C_1, C_2 и λ находятся из граничных условий и условия постоянства функционала длины (1.2.11). Вначале воспользуемся граничными условиями

$$\begin{cases} (-a - C_1)^2 + C_2^2 = \lambda^2 \\ (a - C_1)^2 + C_2^2 = \lambda^2 \end{cases}$$

Откуда следует $C_1 = 0$ и $C_2 = \pm\sqrt{\lambda^2 - a^2}$. Теперь воспользуемся условием (1.2.11).

$$\int_{-a}^{a} \sqrt{1 + \frac{(x-C_1)^2}{\lambda^2-(x-C_1)^2}}\, dx = 2l \quad (1.2.17)$$

Для того чтобы выразить λ из (1.2.17) необходимо вычислить интеграл, входящий в это выражение.

$$\int_{-a}^{a} \sqrt{1 + \frac{(x - C_1)^2}{\lambda^2 - (x - C_1)^2}}\, dx$$

$$= \int_{-a}^{a} \sqrt{\frac{\lambda^2}{\lambda^2 - (x - C_1)^2}}\, dx = \pm\lambda \int_{-a}^{a} \frac{d(x - C_1)}{\sqrt{\lambda^2 - (x - C_1)^2}}$$

$$= \pm\lambda \left(arcsin \frac{x - C_1}{\lambda} \right)\Big|_{-a}^{a} = \pm 2\lambda\, arcsin \frac{a}{\lambda}$$

Таким образом, получаем трансцендентное уравнение относительно λ.

$$\frac{a}{\lambda} = \sin \frac{l}{\lambda} \qquad (1.2.18)$$

Уравнение (1.2.18) решается численно для конкретных параметров задачи a и l, после чего по найденному значению λ вычисляется константа C_2.

Рассмотрим ещё одну изопериметрическую задачу.

Задача 2.3. Определить форму свободно провисающей нерастяжимой нити, концы которой закреплены в точках a и $-a$ (рис. 6). Линейная плотность нити ρ постоянна по всей длине. Длина нити $L > 2a$.

Эта модель описывает не только форму нити, но и форму свободно провисающих проводов, висящих мостов с шарнирным закреплением на краях и т.д.

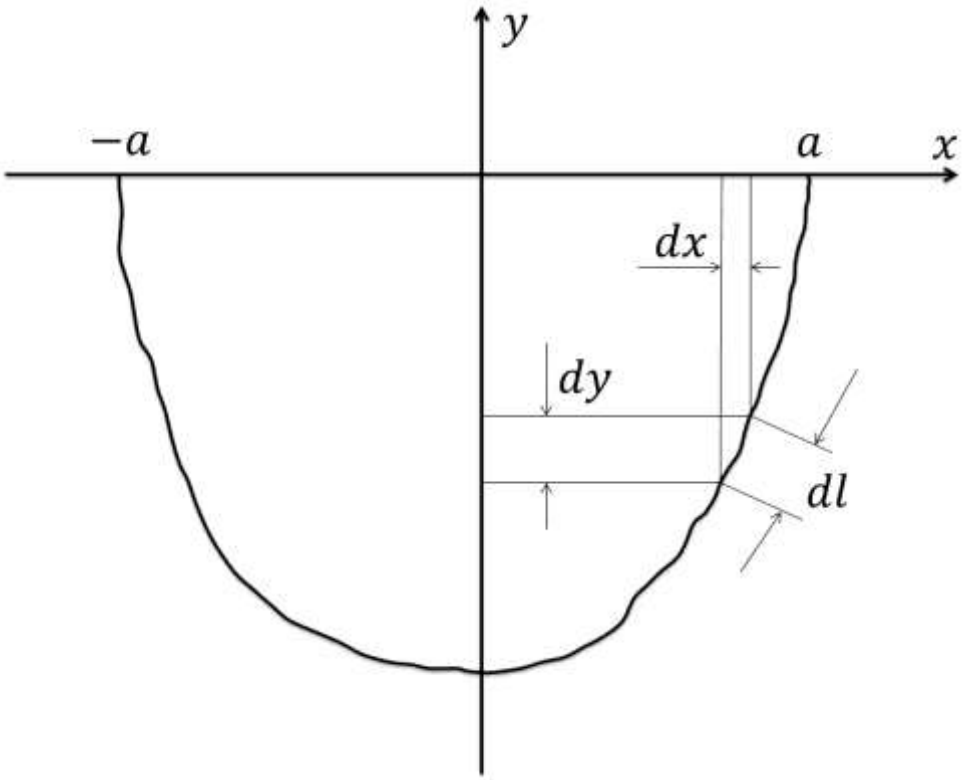

Рис. 6. Иллюстрация к задаче 2.3

В этой задаче необходимо минимизировать функционал потенциальной энергии при условии постоянства длины. Функционал длины записывается так же как и предыдущей задаче

$$L[y] = \int_{-a}^{a} \sqrt{1 + y'^2}\, dx = L \quad (1.2.19)$$

Дифференциальный элемент потенциальной энергии равен $dU = gy * dm = \rho g y * dl = \rho g y \sqrt{1 + y'^2} dx$, где$dm$ – элемент массы, g – ускорение свободного падения. Величины ρ и g являются постоянными множителями, поэтому они не оказывают какого-либо влияния на экстремум функционала. Итак, будем минимизировать функционал потенциальной энергии, делённый наρg.

$$U[y] = \int_{-a}^{a} y\sqrt{1 + y'^2}\, dx \quad (1.2.20)$$

Для решения задачи воспользуемся методом множителей Лагранжа, который мы уже использовали в предыдущей задаче. Составим

функционал $J[y] = U[y] + \lambda L[y]$ и будем решать для него уравнение Эйлера-Лагранжа, которое в этой задаче имеет следующий вид

$$\sqrt{1+y'^2} - \frac{d}{dx}\left\{(y+\lambda)\frac{y'}{\sqrt{1+y'^2}}\right\} = 0 \qquad (1.2.21)$$

Раскроем полную производную в выражении (1.2.21), выразив её через частные производныестандартным способом

$$\frac{d}{dx} = \frac{\partial}{\partial x} + y'\frac{\partial}{\partial y} + y''\frac{\partial}{\partial y'}$$

Тогда уравнение (1.2.21) приводит к следующему

$$\sqrt{1+y'^2} - y'\frac{y'}{\sqrt{1+y'^2}} - y''(y+\lambda)\frac{\sqrt{1+y'^2} - y'\frac{y'}{\sqrt{1+y'^2}}}{1+y'^2} = 0$$

Несмотря на кажущуюся громоздкость этого выражения, оно легко упрощается.

$$\frac{1}{\sqrt{1+y'^2}} - \frac{y''(y+\lambda)}{(1+y'^2)\sqrt{1+y'^2}} = 0$$

Дальнейшее упрощение приводит к достаточно компактному дифференциальному уравнению

$$1 + y'^2 - y''(y+\lambda) = 0 \qquad (1.2.22)$$

Отметим, что дифференциальное уравнение (1.2.22) является нелинейным. Порядок уравнения может быть понижен благодаря тому, что оно не содержит аргумента x. Сделаем замену функции и аргумента. В качестве нового аргумента примем y, а в качестве новой функции $p = y'$. Тогда $y'' = dp/dx = (dp/dy)(dy/dx) = p'p$. Подставляя новые аргумент и функцию в (1.2.22), получим уравнение первого порядка

$$1 + p^2 - pp'(y+\lambda) = 0 \qquad (1.2.23)$$

Уравнение (1.2.23) представляет собой уравнение с разделяющимися переменными. Разделив переменные, приходим к уравнению

$$\frac{pdp}{1+p^2} = \frac{dy}{y+\lambda} \qquad (1.2.24)$$

Далее проинтегрируем (1.2.24), что позволит выразить функцию $p(y)$.

$$\frac{1}{2}\ln(1+p^2) = \ln|y+\lambda| + C'$$

Отсюда нетрудно выразить p.

$$p = \pm\sqrt{C_1'(y+\lambda)^2 - 1} \qquad (1.2.25)$$

В этой формуле $C_1' > 0$ – константа интегрирования. Теперь сделаем обратную замену $p = dy/dx$, и получаем уравнение с разделяющимися переменными для функции $y(x)$.

$$\frac{dy}{dx} = \pm\sqrt{C_1'(y+\lambda)^2 - 1} \qquad (1.2.26)$$

В уравнении (1.2.26) необходимо разделить переменные, а затем проинтегрировать.

$$\pm\frac{dy}{\sqrt{C_1'(y+\lambda)^2 - 1}} = dx$$

$$\pm\frac{1}{\sqrt{C_1'}}\int\frac{d(y+\lambda)}{\sqrt{(y+\lambda)^2 - \frac{1}{C_1'}}} = \int dx$$

Интеграл в левой части этого выражения является табличным.

$$\pm\frac{1}{\sqrt{C_1'}}Arch\left(\sqrt{C_1'}|y+\lambda|\right) = x + C_2'$$

$$|y+\lambda| = \frac{1}{\sqrt{C_1'}}ch\left(\sqrt{C_1'}(x+C_2')\right) \qquad (1.2.27)$$

Далее нужно рассмотреть два случая – когда выражение под модулем в (1.2.27) положительно, и когда оно отрицательно.Первый случай соответствует минимуму потенциальной энергии (гиперболический косинус с ветвями вверх), а второй – максимуму (ветви вниз).

Введём новые константы $C = 1/\sqrt{C_1'} > 0$, $C_1 = -C_2'$.

Тогда, окончательно, получаем решение, соответствующее минимуму потенциальной энергии:

$$y = C\,ch\frac{x-C_1}{C} - \lambda. \qquad (1.2.28)$$

Теперь надо определить константы C, C_1 и λ из граничных условий, а также из условия постоянства длины нити.

Так как функция чётная, нетрудно понять, что $C_1 = 0$ вследствии симметрии задачи. Далее подставим найденный вид функции (1.2.28) в условие постоянства функционала длины (1.2.19). Получим трансцендентное уравнение для определения C.

$$2C\ sh\frac{a}{C} = L \qquad (1.2.29)$$

Воспользовавшись граничным условием $y(a) = 0$, найдём λ.

$$\lambda = \mathrm{C}\ ch\frac{a}{C} \qquad (1.2.30)$$

§1.3. Применение аналогий при построении математических моделей

В начале предыдущего параграфа мы обращали внимание, что решение задачи №2.1 совпадает с законом отражения света из геометрической оптики. В задаче 2.1 мы исходили из того, что время пути должно быть минимальным. В оптике используется аналогичный принцип Ферма. Если бы мы рассмотрели в задаче 2.1 точки A и B по разные стороны отпрямойC, то получили бы траекторию, которую даёт закон Снеллиуса для преломления света. Это является примером того, что одна и та же математическая модель может описывать совершенно разные объектыи явления. Данное свойство математических моделей получило название принцип универсальности. Этот принцип может быть использован при построении математических моделей в тех областях, в которых точные законы природы либо не изучены, либо их не существует вовсе, однако возможно выдвинуть некоторые упрощающие предположения. Это могут быть биологические, экономические, социально-политические модели. Такие модели чаще всего строятся путем применения аналогий с уже изученными объектами, в том числе из других областей. Рассмотрим пример одной из таких моделей.

Задача 3.1. Определить динамику численности биологической популяции (т.е. зависимость количества особей от времени) $N(t)$. В начальный момент времени численность равна $N(0)$, коэффициент рождаемости $\alpha \geq 0$, коэффициент смертности $\beta \geq 0$.

Сделаем предположение о том, что скорость изменения численности популяции пропорциональна самой численности. Тем самым используем аналогию с задачей №1.1 о радиоактивном распаде.

$$\frac{dN(t)}{dt} = \big(\alpha(t) - \beta(t)\big)N(t) \qquad (1.3.1)$$

Разделив переменные в уравнении (1.3.1) и проинтегрировав, получим

$$N(t) = N(0)exp\left(\int_0^t \big(\alpha(t) - \beta(t)\big)dt\right) \qquad (1.3.2)$$

Теперь сделаем упрощающее предположение о том, что коэффициенты рождаемости и смертности не зависят от времени: $\alpha(t) = \alpha_0$, $\beta(t) = \beta_0$. Тогда интеграл в показателе экспоненты в выражении (1.3.2) даёт $(\alpha_0 - \beta_0)t$. Графики зависимости $N(t)$ показаны на рис. 7.

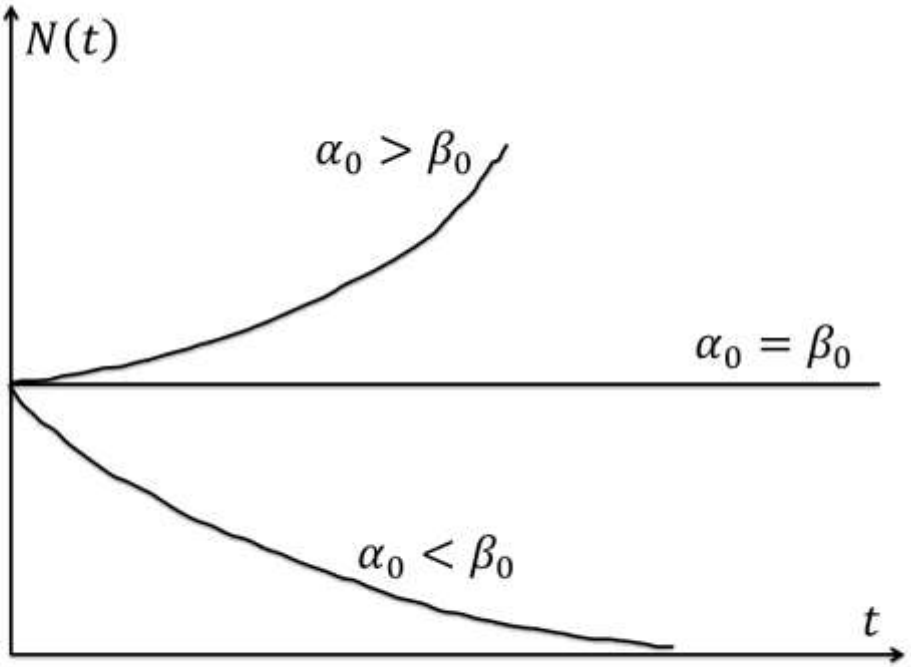

Рис. 7. Иллюстрация к задаче №3.1,график зависимости численности популяции от времени

На рис. 7 видно, что процесс неустойчивый: любое, даже самое малое, отличие коэффициентов рождаемости и смертности приводит либо к вымиранию популяции, либо к неограниченному росту её численности. Рассмотренная модель называется моделью Мальтуса.

Очевидно, что рассмотренная модель является очень сильно упрощенной, и она не учитывает многих факторов, реально влияющих на динамику биологической популяции. К таким факторам относятся ограниченность ресурсов (которая приводит к эффекту насыщения), конкуренция с другими видами, а также саморегуляция коэффициентов рождаемости и смертности (например, если мы интересуемся динамикой численности людей, а не животных). В дальнейшем мы вернемся к рассмотрению этой задачи при обсуждении иерархических моделей, построенных на базе модели Мальтуса.

§1.4. Иерархия математических моделей

Основой иерархического принципа построения математических моделей является дополнение уже готовой модели новыми факторами, оказывающими влияние на исследуемый объект. В этом подходе важно избежать методологической ошибки, когда математическую модель дополняют несущественными факторами, при этом пренебрегая более значимыми.Здесь необходимо руководствоваться физическими соображениями или, если речь идёт о трудноформализуемых или неформализуемых объектах, соображениями логики и здравого смысла.

Вернемся к задаче №1.2, идополним модель некоторыми новыми факторами. Очевидно, что чем больше факторов учесть, тем более громоздкой будет модель. Поэтому для простоты мы рассмотрим добавление к исходной модели различных факторов по отдельности.

1) Пусть на шарик действует внешняя сила $F(x,t)$. Тогда уравнение движения примет следующий вид
$$m\ddot{x} = -kx + F(x,t) \qquad (1.4.1)$$
Рассмотрим, для примера, гармоническую внешнюю силу $F(x,t) = F_0 \sin \omega_1 t$. В этом случае частное решение уравнения (1.4.1), которое описывает установившиеся вынужденные колебания, примет вид $x_1(t) = C \sin \omega_1 t$. Подставляя это частное решение в уравнение (1.4.1), находим $C = F_0/\left(m(\omega^2 - \omega_1^2)\right)$, где $\omega = \sqrt{k/m}$. Видно, что при приближении частоты внешней силы к собственной частоте осциллятора, амплитуда колебаний неограниченно возрастает (наблюдается резонанс). В этой модели мы не учитывали диссипацию механической энергии (трение и сопротивление воздуха). Если бы эти факторы были учтены, то амплитуда колебаний при резонансе была бы конечной.

2) Учтём силу трения $F_{\text{тр}} = -\mu mg sign\dot{x}$. В этом случае уравнение движения принимает вид

$$m\ddot{x} = -kx = -\mu mg sign\dot{x} \tag{1.4.2}$$

Вследствие знакопеременности, уравнение (1.4.2) не сводится к стандартному уравнению колебаний. Проанализируем, однако, динамику полной механической энергии при таких колебаниях. Для того умножим обе части равенства (1.4.2) на$\dot{x}$. Тогда уравнение преобразуется к виду

$$\frac{d}{dt}\left(\frac{m\dot{x}^2}{2} + \frac{kx^2}{2}\right) = -\mu mg\dot{x} sign\dot{x} = -\mu mg|\dot{x}| \tag{1.4.3}$$

Слагаемое в круглых скобках в левой части выражения (1.4.3) представляет собой полную механическую энергию. Видно, что полная механическая энергия убывает с течением временем.

3) Учтём силу сопротивления воздуха $F_{\text{с}} = -k_1\dot{x}$ (закон Стокса), $k_1 > 0$. Уравнение движения при этом имеет следующий вид

$$m\ddot{x} = -kx - k_1\dot{x} \tag{1.4.4}$$

Сделаем замену функции

$$x(t) = \bar{x}(t) exp\left(-\frac{k_1}{2m}t\right) \tag{1.4.5}$$

Вычислив $\dot{x}$ и $\ddot{x}$, приходим к уравнению

$$m\ddot{\bar{x}} = -k_2\bar{x},\ k_2 = k - \frac{k_1^2}{4m} \tag{1.4.6}$$

Характер решения существенно зависит от знака коэффициента k_2. Если $k_2 > 0$, то решение $x(t)$ имеет вид затухающих колебаний

$$x(t)|_{k_2>0} = exp\left(-\frac{k_1}{2m}t\right)(A\sin\omega t + B\cos\omega t),\ \omega = \sqrt{\frac{k_2}{m}} \tag{1.4.7}$$

Константы A и Bнетрудно выразить через начальную координату x_0 и начальную скорость v_0. Для этого необходимо вначале найти скорость, продифференцировав выражение (1.4.7).

$$v(t)|_{k_2>0} = exp\left(-\frac{k_1}{2m}t\right)\left\{\left(-\frac{k_1}{2m}A - B\omega\right)\sin\omega t + \left(-\frac{k_1}{2m}B + A\omega\right)\cos\omega t\right\}$$

A и Bнаходятся из решения следующей системы уравнений

$$\begin{cases} x_0 = B \\ v_0 = -\dfrac{k_1}{2m}B + A\omega \end{cases}$$

Если $k_2 = 0$, колебаний не будет. В этом случае система может перейти через положение равновесия не более чем один раз.

Решение для этого случая выглядит следующим образом

$$x(t)|_{k_2=0} = exp\left(-\frac{k_1}{2m}t\right)\{ct + c_1\}$$

Константы интегрирования выражаются через начальную координату и начальную скорость.

$$c_1 = x_0$$

$$c = v_0 + \frac{k_1 x_0}{2m}$$

Для того чтобы система один раз перешла точку равновесия необходимо выполнение одного из следующих условий

$$\left[\begin{cases} x_0 > 0 \\ v_0 < -\frac{k_1 x_0}{2m} \end{cases}\right.$$

$$\left.\begin{cases} x_0 < 0 \\ v_0 > -\frac{k_1 x_0}{2m} \end{cases}\right.$$

Квадратной скобкой здесь обозначена совокупность, фигурной скобкой – система.

Если $k_2 < 0$, решение описывается линейной комбинацией двух убывающих экспонент

$$x(t)|_{k_2<0} = exp\left(-\frac{k_1}{2m}t\right)\left(A\ exp(-\omega t) + B\ exp(\omega t)\right) \quad (1.4.8)$$

В этой формуле $\omega = \sqrt{-k_2/m}$.

Дифференцируя выражение (1.4.8) получим зависимость скорости от времени

$$v(t)|_{k_2<0} = exp\left(-\frac{k_1}{2m}t\right)\left\{A\left(-\frac{k_1}{2m} - \omega\right)exp(-\omega t) + B\left(-\frac{k_1}{2m} + \omega\right)exp(\omega t)\right\}$$

Константы интегрирования A и B выражаются через начальную координату и скорость следующим образом

$$\begin{cases} A = \frac{x_0}{2} - \frac{v_0}{2\omega} - \frac{k_1 x_0}{4m\omega} \\ B = \frac{x_0}{2} + \frac{v_0}{2\omega} + \frac{k_1 x_0}{4m\omega} \end{cases}$$

Можно показать, что такое решение, независимо от x_0 и v_0, знакопостоянно.

4) Учтем нелинейность пружины, т.е. отклонение зависимости возвращающей силы от смещения от закона Гука. Обозначим положения равновесия l_0. Координата x ограничена снизу произведением толщины проволоки на количество витков dN, что

соответствует полностью сжатой пружинке. Сверху x ограничена величиной πDN (полностью растянутая пружинка). Обозначим силу, необходимую для полного сжатия пружинки $F_{\text{кр1}}$, а для полного растяжения - $F_{\text{кр2}}$. При больших $|x - l_0|$ зависимость силы упругости от координаты $F_{\text{упр}}(x)$ нелинейная. Схематичный график этой зависимости изображен на рис. 8.

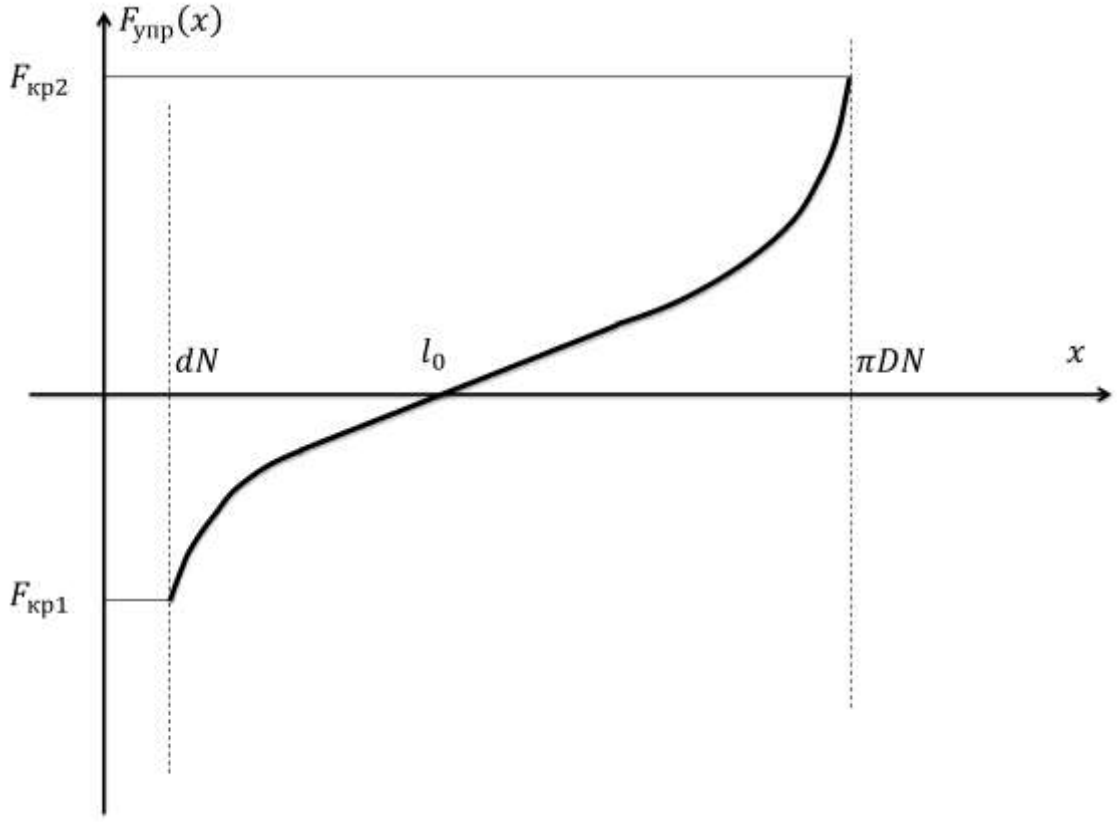

Рис. 8. Схематичная зависимость возвращающей силы пружинки от координаты шарика (см. задачу 1.2)

Введем эффективный коэффициент жёсткости $k(x) = F_{\text{упр}}(x)/(x - l_0)$. $k(x) > 0$. Уравнение движения шарика записывается в виде

$$m\ddot{x} = -k(x)(x - l_0) \qquad (1.4.9)$$

Точное решение полученного уравнения, разумеется, зависит от конкретного вида функции $k(x)$. Получим энергетическое соотношение. Для этого умножим обе части уравнения (1.4.9) на $\dot{x}$. Преобразуем левую часть:

$$m\dot{x}\ddot{x} = \frac{m}{2}\frac{d}{dt}(\dot{x}^2)$$

Преобразуем правую часть:

$$-\dot{x}k(x)(x-l_0) = -\frac{dx}{dt}\frac{d}{dx}\int_{l_0}^{x} k(x')(x'-l_0)\,dx'$$
$$= -\frac{d}{dt}\left(\int_{l_0}^{x} k(x')(x'-l_0)\,dx'\right)$$

Интегрируя полученное уравнение по времени, получаем

$$\frac{m}{2}\dot{x}^2 + \int_{l_0}^{x} k(x')(x'-l_0)\,dx' = const \quad (1.4.10)$$

Уравнение (1.4.10) представляет собой закон сохранения полной механической энергии: первое слагаемое – кинетическая энергия, второе слагаемое – потенциальная. Сохранение полной механической энергии отражает свойство консервативности системы.

Можно и далее продолжать усложнение математической модели путем учёта новых факторов и путём одновременного учёта нескольких факторов из числа рассмотренных.

В качестве второго примера в настоящем параграфе рассмотрим иерархию модели Мальтуса (задача 3.1). Учтём ограниченность ресурсов. Пусть $N_{\text{р}}$ – равновесная численность популяции. Добавим в уравнение для численности популяции множитель $1 - N/N_{\text{р}}$, описывающий насыщение. Тогда уравнение для численности популяции примет вид

$$\frac{dN}{dt} = \alpha\left(1 - \frac{N}{N_{\text{р}}}\right)N \quad (1.4.11)$$

В этом уравнении α – разность между коэффициентами рождаемости и смертности. Примем в дальнейшем $\alpha > 0$.Уравнение (1.4.11) решается методом разделения переменных.

$$\frac{dN}{N\left(1 - \frac{N}{N_{\text{р}}}\right)} = \left(\frac{1}{N_{\text{р}} - N} + \frac{1}{N}\right)dN = \alpha dt$$

Проинтегрировав, получим

$$-\ln\left|N - N_{\text{р}}\right| + \ln N = \alpha t + C$$

Здесь C – константа интегрирования. Из начального условия получаем

$$C = \ln\frac{N(0)}{\left|N(0) - N_{\text{р}}\right|}$$

Окончательно, динамика численности популяции имеет вид

$$N(t) = \frac{N_{\text{р}} N(0) exp(\alpha t)}{N_{\text{р}} - N(0)\left(1 - exp(\alpha t)\right)} \qquad (1.4.12)$$

Для наглядности построим график зависимости $N(t)$при различных $N(0)$.

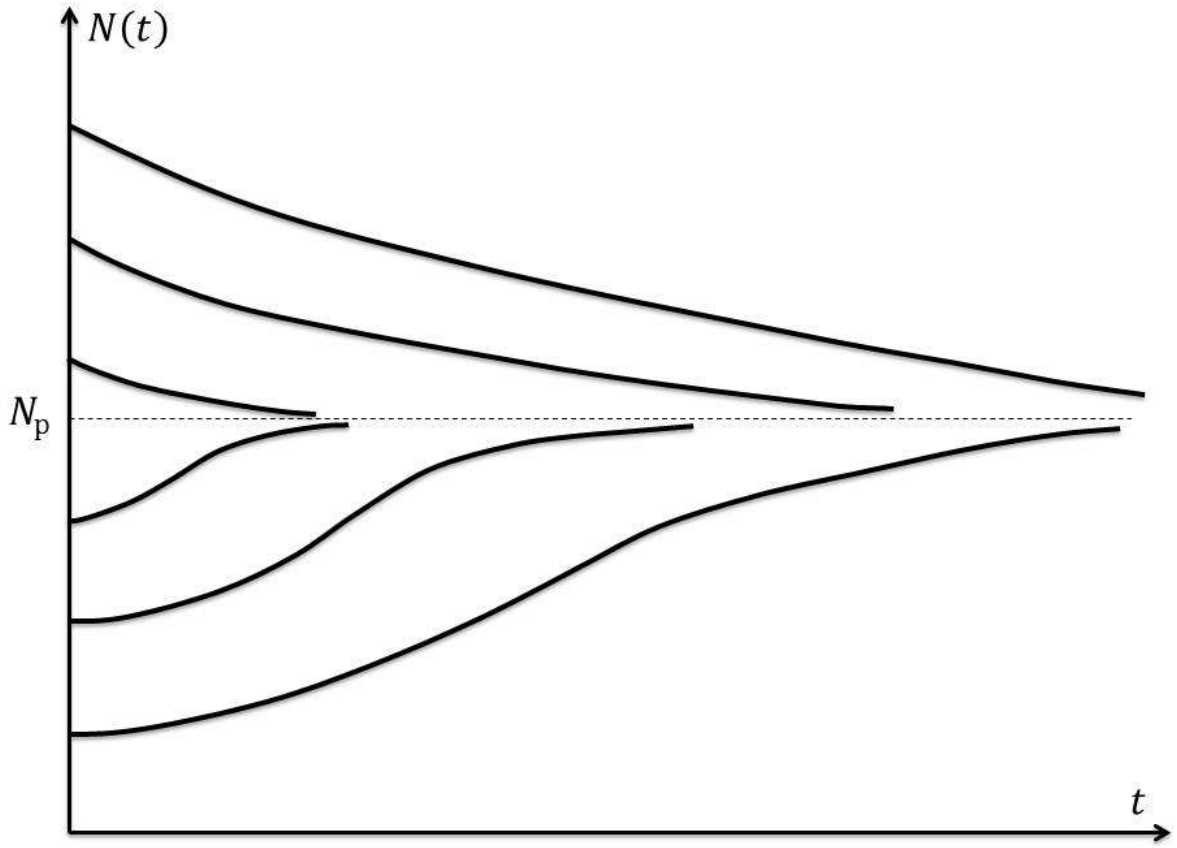

Рис. 9. Динамика численности популяции в модели с насыщением

Изображенные на рис. 9 зависимости называются логистическими кривыми. Они более правдоподобно описывают динамику численности популяции, чем результат модели Мальтуса, однако модель с насыщением (как и любая другая) имеет свои ограничения.

§1.5. Исследование математических моделей

После построения математической модели целесообразно убедиться в том, что она не противоречит хорошо известным законам природы или

здравому смыслу. Например, если модель была построена на основе второго закона Ньютона, можно проверить выполнение закона сохранения энергии. Часто оказывается полезным проанализировать предельные случаи. Правильно построенная модель должна быть устойчива по начальным данным.

В качестве примерарассмотрим уравнение нелинейной теплопроводности.

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k(T)\frac{\partial T}{\partial x}\right) \qquad (1.5.1)$$

В этом уравнении $T > 0$ – температура, $k(T) > 0$ – коэффициент теплопроводности. Требование устойчивости для уравнений в частных производных параболического типа воплощается в принципе максимума и теоремах сравнения.

Рассмотрим задачу Коши для уравнения (1.5.1), т.е. считаем заданным начальное распределение температуры $T(x,0) = T_0(x) \geq 0$. Принцип максимума состоит в том, что максимум решения уравнения (1.5.1) не может превышать максимума начального распределения температуры $T_0(x)$.

$$\max_{t>0,-\infty<x<\infty} T(x,t) \leq \max_{-\infty<x<\infty} T_0(x) \qquad (1.5.2)$$

Физический смысл принципа максимума весьма прост: поток тепла переносит энергию от более горячих участков к более холодным, поэтому максимум начального распределения температуры не может увеличиться со временем.

Для первой краевой задачи в полупространстве $x \geq 0$, принцип максимума записывается в следующем виде

$$\max_{t>0,0<x<\infty} T(x,t) \leq \max\{\max_{0\leq x<\infty} T_0(x), \max_{t\geq 0} T(0,t)\} \qquad (1.5.3)$$

Следствием из принципа максимума являются теоремы сравнения. Их физический смысл заключается в том, что большее тепловое воздействие на один и тот же объект приводит к формированию в нём большего поля температуры.

Сформулируем теоремы сравнения для задачи Коши.Пусть$T^{(1)}(x,t), T(x,t), T^{(2)}(x,t)$ – решения задачи Коши, соответствующие начальным данным $T_0^{(1)}(x), T_0(x), T_0^{(2)}(x)$ соответственно. Тогда если $T_0^{(1)}(x) \leq T_0(x) \leq T_0^{(2)}(x)$ при любом x, то из этого следует $T^{(1)}(x,t) \leq T(x,t) \leq T^{(2)}(x,t)$ для любого x.

Теоремы сравнения для первой краевой задачи в полупространстве выглядят немного иначе. Если для любого $x > 0$ и для любого $t > 0$ выполнены условия

$$\begin{cases} T_0^{(1)}(x) \le T_0(x) \le T_0^{(2)}(x) \\ T^{(1)}(0,t) \le T(0,t) \le T^{(2)}(0,t) \end{cases}$$

то из этого следует $T^{(1)}(x,t) \le T(x,t) \le T^{(2)}(x,t)$ для любого $x > 0$ и для любого $t > 0$.

В литературе можно найти похожие теоремы для уравнений в частных производных эллиптического типа, а также для широкого класса уравнений гиперболического типа.

Глава 2. Конечно-разностные методы решения дифференциальных уравнений

§2.1. Разностные аппроксимации

Рассмотрим достаточно гладкую функциюодного вещественного аргумента$f(x)$, т.е. имеющую достаточное для проведения расчетов количество производных. Из курса математического анализа известно, что производная функции определяется следующим образом

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \qquad (2.1.1)$$

В этой формуле шаг h - величина, устремляемая к нулю, поэтому использование точной формулы (2.1.1) при численном моделировании невозможно. При расчетах на ЭВМ пользуются конечно-разностными аппроксимациями (или для краткости, просто разностными аппроксимациями), выбирая какое-либо достаточно малое, но конечное h. Простейшим примером разностной аппроксимации является правая разностная производная, которая по определению равна

$$f_x = \frac{f(x+h)-f(x)}{h} \qquad (2.1.2)$$

Очевидно, что при любом конечном h значение правой разностной производной отличается от$f'(x)$.

$$f_x = f'(x) + \delta \qquad (2.1.3)$$

Оценим отличие δ правой разностной производной от точного значения производной. Для этого требуется разложить $f(x+h)$ в формуле (2.1.2) в ряд Тейлора.

$$f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(x+\theta h), 0 < \theta < 1 \quad (2.1.4)$$

Тогда воспользовавшись формулами (2.1.2) и (2.1.3), получаем

$$\delta = \frac{h}{2} f''(x+\theta h) = O(h) \quad (2.1.5)$$

Введём определение порядка аппроксимации. Величина p называется порядком разностной аппроксимации, если разность между аппроксимацией и аппроксимируемой величиной есть $O(h^p)$. В соответствии с этим определением, правая разностная производная имеет первый порядок аппроксимации.

Другим важным понятием является шаблон аппроксимации. По определению, шаблоном аппроксимации называется множество точек, в которых необходимо вычислить значение функции для того, чтобы определить значение аппроксимации в одной точке x. Например, шаблон аппроксимации первой разностной производной содержит две точки: $\{x, x+h\}$.

Левая разностная производная определяется следующим образом

$$f_{\bar{x}} = \frac{f(x)-f(x-h)}{h} \quad (2.1.6)$$

Нетрудно показать, что левая разностная производная также имеет первый порядок аппроксимации.

$$\delta_1 = f_{\bar{x}} - f'(x) = O(h) \quad (2.1.7)$$

Шаблон левой разностной производной $\{x-h, x\}$.Определим центральную разностную производную

$$f_{\tilde{x}} = \frac{f_x + f_{\bar{x}}}{2} = \frac{f(x+h)-f(x-h)}{2h} \quad (2.1.8)$$

Определим порядок аппроксимации для центральной разностной производной. Разложим $f_{\tilde{x}}$ в окрестности точки x.

$$f_{\tilde{x}} = \frac{1}{2h}\left\{f(x) + hf'(x) + \frac{h^2}{2}f''(x) - f(x) + hf'(x) - \frac{h^2}{2}f''(x) + O(h^3)\right\}$$

Учитывая свойство $O(h^3)/h = O(h^2)$, получаем

$$f_{\tilde{x}} = f'(x) + O(h^2) \quad (2.1.9)$$

Это означает, что центральная разностная производная имеет второй порядок аппроксимации.

Таким образом, правая и левая разностные производные аппроксимируют $f'(x)$ с первым порядком по h, а центральная разностная производная – со вторым порядком по h. Однако, шаблон аппроксимации для центральной разностной производной $\{x-h, x+h\}$ шире, чем для правой или левой.

Существуют и другие аппроксимации первой производной с более высокими порядками аппроксимации, но чем выше порядок аппроксимации, тем шире её шаблон.

Введём в рассмотрение аппроксимацию второй производной.

$$f_{\bar{x}x} = (f_{\bar{x}})_x = \frac{f_{\bar{x}}(x+h) - f_{\bar{x}}(x)}{h} = \frac{1}{h}\left[\frac{f(x+h)-f(x)}{h} - \frac{f(x)-f(x-h)}{h}\right] = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \quad (2.1.10)$$

Такая аппроксимация имеет второй порядок по h. Докажем это, разложив $f_{\bar{x}x}$ в ряд Тейлора в окрестности точки x.

$$f_{\bar{x}x} = \frac{1}{h^2}\left\{f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) - 2f(x) + f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + O(h^4)\right\}$$

Используя свойство $O(h^4)/h^2 = O(h^2)$, получаем

$$f_{\bar{x}x} = f''(x) + O(h^2) \quad (2.1.11)$$

Нетрудно также доказать, что $f_{x\bar{x}} = f_{\bar{x}x}$, а аппроксимации f_{xx} и $f_{\bar{x}\bar{x}}$ выводят за пределы шаблона $\{x-h, x, x+h\}$. Проанализируем для примера шаблон f_{xx}.

$$f_{xx} = \frac{f_x(x+h) - f_x(x)}{h} = \frac{1}{h}\left(\frac{f(x+2h) - f(x+h)}{h} - \frac{f(x+h) - f(x)}{h}\right)$$

$$f_{xx} = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2}$$

Таким образом, шаблон f_{xx} составляют точки $\{x, x+h, x+2h\}$.

§2.2. Конечно-разностный метод для обыкновенных дифференциальных уравнений. Метод прогонки.

Будем рассматривать обыкновенное дифференциальное уравнение второго порядка следующего вида

$$-\frac{d^2u}{dx^2} + q(x)u = f(x) \quad (2.2.1)$$

Переменную x полагаем в интервале $0 \le x \le 1$. Сформулируем граничные условия, соответствующие первой краевой задаче $u(0) = u(1) = 0$.

Введём по переменной x равномерную сетку: $x_i = ih; i = 0,1,2,\dots,M \equiv \overline{0,M}; h = 1/M$. Значения искомой функции в узлах сетки обозначим $u_i \equiv u(x_i)$. Воспользовавшись рассмотренной выше аппроксимацией второй разностной производной (2.1.10), получаем разностную схему, аппроксимирующую уравнение (2.2.1).

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + q_i u_i = f_i + \delta_i \quad (2.2.2)$$

Погрешность этой разностной схемы $\delta_i = O(h^2)$. Пренебрегая δ_i в уравнении (2.2.2), получаем приближённое равенство

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + q_i u_i \approx f_i \quad (2.2.3)$$

Для того чтобы в дальнейшем работать с точным уравнением, введём сеточную функцию $v_i, i = \overline{0,M}$, удовлетворяющую уравнению

$$-\frac{v_{i+1}-2v_i+v_{i-1}}{h^2} + q_i v_i = f_i; i = \overline{1,M-1} \quad (2.2.4)$$

Уравнение (2.2.4) отличается от приближённого равенства (2.2.3) тем, что вместо приближённого равенства используется точное, но для новой сеточной функции. Граничные условия следующие: $v_0 = v_M = 0$. Таким образом, (2.2.4) совместно с граничными условиями образует систему $M+1$ уравнений. Выпишем из системы (2.2.4) отдельно первое и последнее уравнения (соответствующие $i = 1, M-1$).

$$-\frac{v_2-2v_1}{h^2} + q_1 v_1 = f_1 \quad (2.2.5)$$

$$-\frac{2v_{M-1}+v_{M-2}}{h^2} + q_{M-1} v_{M-1} = f_{M-1} \quad (2.2.6)$$

В уравнения (2.25) и (2.2.6) входят по два значения сеточной функции, в то время как во все остальные уравнения системы (2.2.4) – по три значения. Таким образом, система (2.2.4) имеет следующий вид

$$\begin{cases} b_1 v_1 + c_1 v_2 = f_1 \\ a_i v_{i-1} + b_i v_i + c_i v_{i+1} = f_i; i = \overline{2,M-2} \\ a_{M-1} v_{M-2} + b_{M-1} v_{M-1} = f_{M-1} \end{cases} \quad (2.2.7)$$

Это система линейных алгебраических уравнений с трёхдиагональной матрицей.

$$\begin{pmatrix} b_1 & c_1 & 0 & & 0 & 0 & 0 \\ a_2 & b_2 & c_2 & & 0 & 0 & 0 \\ 0 & a_3 & b_3 & & c_3 & 0 & 0 \\ 0 & 0 & * & * & * & & 0 \\ 0 & 0 & 0 & * & * & & c_{M-2} \\ 0 & 0 & 0 & 0 & a_{M-1} & & b_{M-1} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ * \\ * \\ * \\ v_{M-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ * \\ * \\ * \\ f_{M-1} \end{pmatrix}$$

Самый экономичный с вычислительной точки зрения способ решения систем с трёхдиагональной матрицей – это метод прогонки. Количество элементарных численных операций в этом методе пропорционально первой степени размерности матрицы.

Нетрудно видеть, что первое уравнение системы связывает v_1 и v_2. Это позволяет формально выразить v_1 через v_2. Второе уравнение связывает v_1, v_2 и v_3. Подставив сюда выражение v_1 через v_2, полученное из первого уравнения, получим однозначную связь между v_2 и v_3. Из этого выразим v_2 через v_3. Дальнейшие действия аналогичны: из третьего уравнения выражаем v_3 через v_4 и т.д. Становится понятным, что на каждом шаге i-й элемент столбца неизвестных линейно связан с последующим элементом. Таким образом, в рамках метода прогонки вводятся прогоночные коэффициенты α_i, β_i следующим образом

$$v_i = \alpha_i v_{i+1} + \beta_i \tag{2.2.8}$$

Используя (2.2.8), получаем из второго уравнения системы (2.2.7) (точнее системы из $M-3$ уравнений) следующее выражение

$$a_i(\alpha_{i-1}v_i + \beta_{i-1}) + b_i v_i + c_i v_{i+1} = f_i \tag{2.2.9}$$

Выразим из уравнения (2.2.9) v_i через v_{i+1}.

$$v_i = -\frac{c_i}{a_i\alpha_{i-1}+b_i} v_{i+1} + \frac{f_i - \alpha_i\beta_{i-1}}{a_i\alpha_{i-1}+b_i} \tag{2.2.10}$$

Сравнив полученное выражение с (2.2.8), получаем рекуррентные соотношения для прогоночных коэффициентов.

$$\alpha_i = -\frac{c_i}{a_i\alpha_{i-1}+b_i} \tag{2.2.11}$$

$$\beta_i = \frac{f_i - \alpha_i\beta_{i-1}}{a_i\alpha_{i-1}+b_i} \tag{2.2.12}$$

α_1 и β_1 можно найти из первого уравнения системы (2.2.7), выразив v_1 через v_2. $\alpha_1 = -c_1/b_1$, $\beta_1 = f_1/b_1$, все остальные прогоночные коэффициенты находятся по рекуррентным соотношениям.

После нахождения прогоночных коэффициентов начинается обратный ход метода прогонки. Воспользуемся выражением (2.2.8), подставив в него $i = M-1$. Получим $v_{M-1} = \alpha_{M-1}v_M + \beta_{M-1}$. Воспользовавшись граничным условием$v_M = 0$, получаем, $v_{M-1} = \beta_{M-1}$. По найденномуv_{M-1} вычисляем последовательно v_{M-2}, v_{M-3},…,v_1, используя (2.2.8) как рекуррентное соотношение.

§2.3. Устойчивость разностных схем для обыкновенных дифференциальных уравнений. Жёсткие системы дифференциальных уравнений.

Как упоминалось выше, по мере выполнения алгоритма накапливается вычислительная погрешность. Одним из условий корректности вычислительного алгоритма является его устойчивость. Это означает, что вычислительная погрешность должна возрастать

незначительно, т.е. погрешность должна быть много меньше вычисляемой величины (если последняя отлична от нуля).

Исследование устойчивости разностных схем в общем виде представляет собой сложную задачу. В этом разделе мы рассмотрим вопрос устойчивости на примере задачи Коши для обыкновенных дифференциальных уравнений первого порядка.

$$\frac{du(t)}{dt} = f(t,u); t > 0; u(0) = u_0 \quad (2.3.1)$$

Функцию $f(t,u)$ полагаем непрерывной в области $D: |t| \leq a, |u - u_0| \leq b$. Из этого следует ограниченность функции $f(t,u)$ в области $D|f(t,u)| \leq M$. Также считаем, что функция $f(t,u)$ удовлетворяет условию Липшица по аргументу u, т.е. $f(t,u_1) - f(t,u_2) \leq L|u_1 - u_2|$ где $u_1, u_2 \in D$. Сформулированные условия обеспечивают существование и единственность решения задачи Коши при $t < t_0 = min(a, b/M)$.

Введём равномерную сетку по аргументу t: $\omega_\tau = \{t_n = n\tau,\ n = 0,1,2,\dots\}$. $y_n \equiv y(t_n)$ – сеточная функция, аппроксимирующая точное решение $u(t)$.

Рассмотрим одну из простейших разностных схем, аппроксимирующих уравнение (2.3.1)

$$\frac{y_{n+1} - y_n}{\tau} - f(t_n, y_n) = 0; n = 0,1,2,\dots;\ y_0 = u_0 \quad (2.3.2)$$

Решение задачи Коши по такой разностной схеме получило название метода Эйлера.

Говорят, что метод (или разностная схема) сходится, если выполняется условие $y_n \to u(t_n)$ при $\tau \to 0$ в каждой точке рассматриваемого диапазона. Введём также определение порядка точности метода: разностный метод имеет порядок точности p, если $|y_n - u(t_n)| = O(\tau^p)$ при $\tau \to 0$.

Погрешность метода по определению равна $z_n = y_n - u(t_n)$. Далее будем обозначать для краткости $u_n = u(t_n)$. Таким образом, $y_n = z_n + u_n$. Подставив это выражение в (2.3.2), получим уравнение для погрешности.

$$\frac{z_{n+1} + u_{n+1} - z_n - u_n}{\tau} - f(t_n, u_n + z_n) = 0 \quad (2.3.3)$$

$$\frac{z_{n+1} - z_n}{\tau} = f(t_n, u_n + z_n) - \frac{u_{n+1} - u_n}{\tau} = \psi_n^{(1)} + \psi_n^{(2)} \quad (2.3.4)$$

Здесь введены новые обозначения

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + f(t_n, u_n) \quad (2.3.5)$$

$$\psi_n^{(2)} = f(t_n, u_n + z_n) - f(t_n, u_n) \quad (2.3.6)$$

Величина $\psi_n^{(1)}$ получила название невязки. Можно убедиться, что невязка – это результат подстановки точного решения в разностное уравнение.

Введём понятие порядка аппроксимации метода: разностный метод имеет порядок аппроксимацииp, если $\psi_n^{(1)} = O(\tau^p)$. Доказано, что порядок аппроксимации совпадает с порядком точности разностного метода.

Вычислим порядок аппроксимации метода Эйлера. Для этого необходимо разложить в формуле (2.3.5) величину $u_{n+1} - u_n$ в ряд Тейлора.

$$\psi_n^{(1)} = -\frac{u_n + \tau u_n' + O(\tau^2) - u_n}{\tau} + u_n' = O(\tau)$$

Таким образом, мы показали, что метод Эйлера имеет первый порядок аппроксимации.

Разумеется, существуют разностные схемы, обеспечивающие более высокий порядок аппроксимации. Одним из таких примеров является симметричная схема

$$\frac{y_{n+1}-y_n}{\tau} - \frac{1}{2}\big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\big) = 0; n = 0,1,2, \dots;\ y_0 = u_0 \quad (2.3.7)$$

У симметричной схемы второй порядок аппроксимации.Для того чтобы это показать нужно подставить в уравнение (2.3.7) точное решение дифференциального уравнения (2.3.1).

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + \frac{1}{2}\big(f(t_n, u_n) + f(t_{n+1}, u_{n+1})\big)$$

$$\psi_n^{(1)} = -u_n' - \frac{\tau}{2}u_n'' + O(\tau^2) + \frac{1}{2}(u_n' + u_{n+1}')$$

$$\psi_n^{(1)} = -u_n' - \frac{\tau}{2}u_n'' + \frac{1}{2}\big(u_n' + u_n' + \tau u_n'' + O(\tau^2)\big) = O(\tau^2)$$

Если требуется более высокий порядок точности, то используют многошаговые методы. По определению, линейный m-шаговый разностный метод – это метод, использующий разностную схему вида

$$\frac{a_0 y_n + a_1 y_{n-1} + \dots + a_m y_{n-m}}{\tau} = b_0 f_n + b_1 f_{n-1} + \dots + b_m f_{n-m} \quad (2.3.8)$$

Здесь $n = m, m+1, m+2, \dots;\ a_0 \neq 0$. Таким образом, для использования m-шагового метода необходимо предварительно определить m начальных значений $y_0, y_1, y_2, \dots, y_{m-1}$. Но так как исходное дифференциальное уравнение имеет первый порядок, в задаче задано только одно начальное значение $y_0 = u_0$. Это означает, что $y_1, y_2, \dots, y_{m-1}$ необходимо найти одношаговыми методами.

Если $b_0 = 0$, то такой метод называют явным, а если $b_0 \neq 0$ – неявным. Смысл этих определений состоит в том, что при $b_0 = 0$ значение сеточной функции на каждом шаге может быть явно выражено через

значения сеточной функции и функции f на предыдущих шагах. При $b_0 \neq 0$ такое явное выражение невозможно и необходимо решать систему уравнений. Метод Эйлера, рассмотренный выше, является явным методом, а симметричная схема – неявным.

Нетрудно видеть, что коэффициенты уравнения (2.3.8) определены с точностью до произвольного множителя. Для определённости, примем нормировку $\sum_{k=0}^{m} b_k = 1$. В этой нормировке правая часть (2.3.8) аппроксимирует $f(t,u)$. Частным случаем многошаговых методов являются методы Адамса, в которых $a_0 = -a_1 = 1;\ a_k|_{k=2,3,\ldots,m} = 0$.

Доказано, что линейный m-шаговый метод имеет порядок аппроксимации p, если коэффициенты разностной схемы (2.3.8) удовлетворяют следующим условиям

$$\begin{cases} \sum_{k=1}^{m} k a_k = -1 \\ \sum_{k=1}^{m} k^{l-1}(k a_k + l b_k) = 0, l = 2,3,\ldots,p \\ a_0 = \sum_{k=1}^{m} a_k \\ b_0 = 1 - \sum_{k=1}^{m} b_k \end{cases} \qquad (2.3.9)$$

Первые два соотношения в (2.3.9) образуют систему из p уравнений. Количество неизвестных равно$2m$. Следовательно, максимальный порядок аппроксимации m-шагового метода составляет $2m$.

Устойчивость и сходимость линейного многошагового метода определяется расположением корней характеристического уравнения

$$\sum_{k=0}^{m} a_k q^{m-k} = 0 \qquad (2.3.10)$$

Метод устойчив и сходится, когда $|q| \leq 1$ для всех корней, причём те корни, для которых $|q| = 1$, не должны быть кратными (без доказательства). Выполнение этого условия гарантирует ограниченность решений однородного разностного уравнения, однако в нём никак не задействованы коэффициенты правой частиb_k.

На практике необходимо руководствоваться априорными знаниями о поведении решения. Поясним это на примере задачи Коши

$$\frac{du}{dt} = \lambda u, t > 0, u(0) = u_0, \lambda < 0 \qquad (2.3.11)$$

Не решая само уравнение нетрудно понять, что решение должно либо монотонно убывать при $u_0 > 0$, либо монотонно возрастать при $u_0 < 0$, при этом выходя на горизонтальную асимптоту $u = 0$при$t \to \infty$. Действительно, предположим, например, что $u_0 > 0$. Тогда $du/dt|_{t=0} = 0$, следовательно, первоначально функция должна убывать. Она убывает при всех положительных значениях u. При этом функция u не может стать отрицательной, потому что если функция пересечет ось x, она обязана

будет возрастать. Т.е. не может быть никакого участка, где функция отрицательна и при этом убывает. Это означает, что она всегда положительна и монотонно убывает.

Аналогичными рассуждениями можно доказать, что при $u_0 < 0$ функция монотонно возрастает и выходит на горизонтальную асимптоту $u = 0 \text{при} t \to \infty$.

Применим метод Эйлера.

$$\frac{y_{n+1}-y_n}{\tau} = \lambda y_n; n = 0,1,2, \ldots \quad (2.3.12)$$

Рекуррентное соотношение для этой разностной схемы выглядит следующим образом

$$y_{n+1} = (1 + \tau\lambda)y_n \quad (2.3.13)$$

Из априорного соображения $|y_{n+1}| \leq |y_n|$ получаем условие устойчивости

$$|1 + \tau\lambda| \leq 1 \quad (2.3.14)$$

Раскрыв модуль в (2.3.14), условие устойчивости можно записать в виде

$$0 < \tau \leq 2/|\lambda| \quad (2.3.15)$$

Таким образом, явный метод Эйлера относится к классу условно устойчивых методов, т.е. тех методов, которые являются устойчивыми только при выполнении определенного ограничения на шаг разностной схемы. Те методы, которые являются устойчивыми при любом значении шага, называют абсолютно устойчивыми.

Применим к задаче (2.3.11) неявный метод Эйлера.

$$\frac{y_{n+1}-y_n}{\tau} = \lambda y_{n+1} \quad (2.3.16)$$

Несмотря на то, что этот метод формально относится к классу неявных, из формулы (2.3.16) может быть получено рекуррентное соотношение

$$y_{n+1} = (1 - \tau\lambda)^{-1} y_n \quad (2.3.17)$$

Несложно показать, что условие устойчивости такой разностной схемы $|(1 - \tau\lambda)^{-1}| < 1$ выполняется при любом значении шага.

Для более сложных разностных схем **явные методы являются условно устойчивыми, а среди неявных существуют абсолютно устойчивые**.

При численном решении систем дифференциальных уравнений важным вопросом является жёсткость системы. Это свойство характеризует чувствительность решения к погрешности входных данных. Рассмотрим пример.

$$\begin{cases} \frac{du_1}{dt} + a_1 u_1 = 0,\ a_1 > 0, t > 0 \\ \frac{du_2}{dt} + a_2 u_2 = 0,\ a_2 \gg a_1 \end{cases} \tag{2.3.18}$$

Приведенный пример, конечно, является иллюстративным, потому что очевидно, что два уравнения системы (2.3.18) могут быть решены по отдельности. Применяя метод Эйлера, условие устойчивости получается следующим

$$\begin{cases} \tau a_1 \leq 2 \\ \tau a_2 \leq 2 \end{cases} \tag{2.3.19}$$

Учитывая, что $a_2 \gg a_1$, окончательное условие устойчивости $\tau \leq 2/a_2$. Начиная с определенного значения t, решение системы определяется a_1 (более медленно спадающая экспонента), в то время как устойчивость определяется a_2.

В общем случае системы линейных дифференциальных уравнений $du/dt = Au$, где у матрицыA большой разброс собственных чисел, возникают аналогичные трудности.Проиллюстрируем это на примере системы дифференциальных уравнений

$$\begin{cases} \frac{du_1}{dt} = a_{11} u_1 + a_{12} u_2\ , t > 0 \\ \frac{du_2}{dt} = a_{21} u_1 + a_{22} u_2 \end{cases}$$

Выразим из первого уравнения системы функцию u_2.

$$u_2 = \frac{1}{a_{12}} \frac{du_1}{dt} - \frac{a_{11}}{a_{12}} u_1$$

Теперь подставим это выражение во второе уравнение системы.

$$\frac{1}{a_{12}} \frac{d^2 u_1}{dt^2} - \frac{a_{11}}{a_{12}} \frac{du_1}{dt} = a_{21} u_1 + \frac{a_{22}}{a_{12}} \frac{du_1}{dt} - \frac{a_{11} a_{22}}{a_{12}} u_1$$

Упростив выражение, получим

$$\frac{d^2u_1}{dt^2} - (a_{11} + a_{22})\frac{du_1}{dt} + (a_{11}a_{22} - a_{12}a_{21}) = 0$$

Рассмотрим для примера значения коэффициентов $a_{11} = a_{12} = a_{21} = -1$, $a_{22} = -100$. В этом случае характеристическое уравнение имеет вид

$$\lambda^2 + 101\lambda + 99 = 0$$

Корни характеристического уравнения $\lambda_1 \approx -0.99$, $\lambda_1 \approx -100.01$. Соответственно, общее решение дифференциального уравнения следующее

$$u_1(t) = Ae^{-0.99t} + Be^{-100.01t}$$

Константы A и B определяются из начальных условий. На временах $t \gg 0.01$ поведение функции $u_1(t)$ определяется первым слагаемым, которое содержит медленно убывающую экспоненту. Тем не менее, устойчивость разностного метода будет определяться всеми коэффициентами.

Если $A \neq A(t)$, то система называется жёсткой при выполнении следующих условий:

1) $Re(\lambda_k) < 0, k = 1,2, \dots, m$. λ_k – собственные числа матрицы A.
2) Число жёсткости $s = \max_{1\le k\le m}|Re(\lambda_k)| \,/\, \min_{1\le k\le m}|Re(\lambda_k)|$ достаточно велико.

В случае, когда $A = A(t)$, под числом жёсткости понимают величину $\sup_{t\in(0;T)} s(t)$, где $s(t) = \max_{1\le k\le m}\left|Re\left(\lambda_k(t)\right)\right| \,/\, \min_{1\le k\le m}\left|Re\left(\lambda_k(t)\right)\right|$.

Решение жёсткой системы дифференциальных уравнений состоит из быстро убывающих и медленно убывающих компонент. Начиная с некоторого значения t, решение практически полностью определяется медленно убывающими компонентами, однако быстро убывающие накладывают ограничения на шаг (условия устойчивости). Выходом из такой ситуации является использование неявных методов.

§2.4. Конечно-разностный метод для уравнения теплопроводности

В этом разделе мы рассмотрим применение конечно-разностного метода к уравнениям в частных производных на примере уравнения теплопроводности.

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x,t);\ \ 0 \le x \le 1;\ \ 0 \le t \le T \qquad (2.4.1)$$

Зададим одно начальное и два граничных условия.

$$\begin{cases} u(0,t) = \alpha(t) \\ u(1,t) = \beta(t) \\ u(x,0) = \varphi(x) \end{cases} \qquad (2.4.2)$$

Введём равномерную сетку по обоим аргументам: $x_i = ih; i = 0,1,\dots,M; h = 1/M;\ \ t_n = n\tau; n = 0,1,\dots,N;\ \tau = T/N$. Сеточная функция $u_i^n = u(x_i,t_n)$. Узлы сетки, соответствующие $x_i = 0, x_i = 1$ и $t_n = 0$ называются граничными узлами, а все остальные узлы – внутренними. Множество точек с одинаковым значением n называется слоем.

Воспользуемся явной разностной схемой.

$$\frac{u_i^{n+1} - u_i^n}{\tau} - \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} = f_i^n + \delta_i^n \qquad (2.4.3)$$

Здесь $\delta_i^n = O(\tau)$, $\delta_i^n = O(h^2)$.

Шаблон этой схемы изображён на рис. 10.

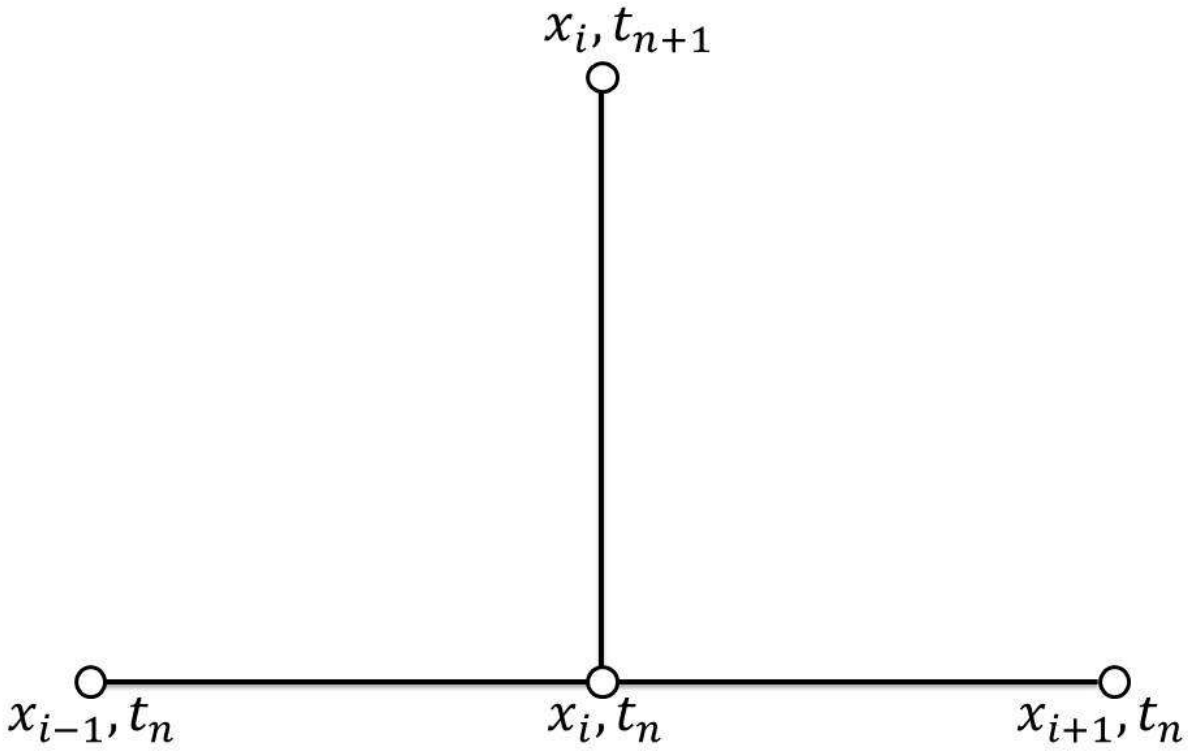

Рис. 10. Шаблон явной аппроксимации для решения уравнения теплопроводности

Заменим сеточною функцию u_i^n на новую сеточную функцию v_i^n, которая удовлетворяет уравнению (2.4.3) безδ_i^n в правой части. Тогда для внутренних узлов сетки $i = 1,2, \dots, M-1; n = 0,1, \dots, N-1$ имеем систему разностных уравнений

$$\frac{v_i^{n+1} - v_i^n}{\tau} - \frac{v_{i+1}^n - 2v_i^n + v_{i-1}^n}{h^2} = f_i^n \qquad (2.4.4)$$

Эту систему необходимо дополнить граничными условиями $v_0^{n+1} = \alpha(t_{n+1})$; $v_M^{n+1} = \beta(t_{n+1})$ для узлов $n = 0,1, \dots, N-1$ и $v_i^0 = \varphi(x_i)$ для узлов $i = 1,2, \dots, M-1$. Нетрудно убедиться в том, что количество неизвестных и уравнений совпадает и равно $(M-1)N$ по количеству внутренних узлов сетки.

Так как схема явная, решение системы удобнее всего находить по слоям.

$$v_i^{n+1} = v_i^n + \frac{\tau}{h^2}(v_{i+1}^n - 2v_i^n + v_{i-1}^n) + \tau f_i^n \qquad (2.4.5)$$

В левую часть выражения (2.4.5) входят переменные $n+1$ –ого слоя, а в правую – n –ого слоя.

Для исследования устойчивости и сходимости разностных схем для уравнений в частных производных используется метод гармоник. В этом методе рассматривается однородное уравнение

$$\frac{u_i^{n+1}-u_i^n}{\tau}=\frac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{h^2} \tag{2.4.6}$$

решение которого ищется в виде гармоники$u_i^n(\varphi)=q^n exp\big(j(ih\varphi)\big)$. Здесь j – мнимая единица, φ и q – вещественные числа. Тогда получаем

$$\frac{q-1}{\tau}=\frac{e^{jh\varphi}-2+e^{-jh\varphi}}{h^2} \tag{2.4.7}$$

Выразим из полученного выражения q.

$$q=1-4\frac{\tau}{h^2}Sin^2\left(\frac{h\varphi}{2}\right) \tag{2.4.8}$$

Условием устойчивости разностной схемы является $|q|\le 1$ для всех φ. В противном случае решение в виде гармоник будет неограниченно возрастать по мере продвижения по слоям.

$$\left|1-4\frac{\tau}{h^2}Sin^2\left(\frac{h\varphi}{2}\right)\right|\le 1\ \forall\varphi\ \leftrightarrow\ \frac{4\tau}{h^2}\le 2 \tag{2.4.9}$$

Окончательно, условие устойчивости принимает следующий вид: $\tau\le 0.5h^2$. Следовательно, явная схема для уравнения теплопроводности является условно устойчивой.

Рассмотрим теперь неявную схему.

$$\frac{v_i^{n+1}-v_i^n}{\tau}-\frac{v_{i+1}^{n+1}-2v_i^{n+1}+v_{i-1}^{n+1}}{h^2}=f_i^{n+1} \tag{2.4.10}$$

Преобразуем выражение (2.4.10), записав в левую часть равенства переменные, относящиеся к $n+1$ слою, а в правую часть – переменные, относящиеся к n слою.

$$-\frac{\tau}{h^2}v_{i+1}^{n+1}+\left(1+2\frac{\tau}{h^2}\right)v_i^{n+1}-\frac{\tau}{h^2}v_{i-1}^{n+1}=$$

$$v_i^n+\tau f_i^{n+1};\ i=1,\dots,M-1 \tag{2.4.11}$$

В этой схеме для нахождения решения на $n+1$ слое по известным значениям переменных на предыдущем слое необходимо решать трёхдиагональную систему линейных алгебраических уравнений (это

наиболее оптимально сделать методом прогонки). Неявная схема обладает первым порядком аппроксимации по τ и вторым поh. Методом гармоник можно доказать, что неявная схема абсолютно устойчива. Для этого нужно подставить решение в виде гармоник $v_i^n(\varphi) = q^n exp\big(j(ih\varphi)\big)$в уравнение

$$\frac{v_i^{n+1} - v_i^n}{\tau} = \frac{v_{i+1}^{n+1} - 2v_i^{n+1} + v_{i-1}^{n+1}}{h^2}$$

При подстановке получим

$$\frac{q^{n+1}exp\big(j(ih\varphi)\big) - q^n exp\big(j(ih\varphi)\big)}{\tau} =$$

$$\frac{q^{n+1}exp\left(j\big((i+1)h\varphi\big)\right) - 2q^{n+1}exp\big(j(ih\varphi)\big) + q^{n+1}exp\left(j\big((i-1)h\varphi\big)\right)}{h^2}$$

Сократив обе части равенства на$q^n exp\big(j(ih\varphi)\big)$, выражаем$q$.

$$q = \left(1 + 4\frac{\tau}{h^2}Sin^2\frac{h\varphi}{2}\right)^{-1}$$

Отсюда видно, что условие устойчивости $|q| \leq 1$выполняется при любых φ, τ и h.

Симметричная схема (Крама – Николсона) определяется следующим образом

$$\frac{v_i^{n+1} - v_i^n}{\tau} - \frac{1}{2}\left[\frac{v_{i+1}^{n+1} - 2v_i^{n+1} + v_{i-1}^{n+1}}{h^2} + \frac{v_{i+1}^n - 2v_i^n + v_{i-1}^n}{h^2}\right] = \frac{f_i^n + f_i^{n+1}}{2} \qquad (2.4.12)$$

Эта схема абсолютно устойчива, при этом обладает вторым порядком аппроксимации как по τ, так и по h. Как и в неявной схеме, в симметричной схеме для нахождения решения на $n+1$ слое по известным значения переменных на предыдущем слое необходимо решать трёхдиагональную СЛАУ.

Существуют другие разностные схемы, среди которых есть и абсолютно неустойчивые. Примером такой схемы является схема «крест» Ричардсона:

$$\frac{v_i^{n+1} - v_i^{n-1}}{2\tau} - \frac{v_{i+1}^n - 2v_i^n + v_{i-1}^n}{h^2} = f_i^n \qquad (2.4.13)$$

Абсолютную неустойчивость этой схемы нетрудно доказать методом гармоник.

§2.5. Метод конечных объёмов для дифференциальных уравнений в частных производных.

Рассмотрим общий вид дифференциального уравнения 2 порядка в частных производных.

$$-\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(b(x,y)\frac{\partial u}{\partial y}\right)+q(x,y)u=f(x,y) \qquad (2.5.1)$$

$a,b>0; q\geq q_m>0; 0\leq x,y\leq 1; u|_\Gamma=\varphi(x,y)$, Γ – граница рассматриваемой области.

Вводим равномерную сетку: $x_i=ih_x; i=0,1,\dots,M;\ h_x=1/M;\ y_j=jh_y; j=0,1,\dots,N;\ h_y=1/N$.

Для упрощения дальнейших преобразований введём функции $w^x=-a\frac{\partial u}{\partial x}$, $w^y=-b\frac{\partial u}{\partial y}$. Тогда уравнение (2.5.1) принимает вид

$$\frac{\partial w^x}{\partial x}+\frac{\partial w^y}{\partial y}+qu=f \qquad (2.5.2)$$

Произведём интегрирование уравнения (2.5.2) по конечному объёму вблизи узла (x_i,y_j): $x_{i-1/2}\leq x\leq x_{i+1/2}$, $y_{j-1/2}\leq y\leq y_{j+1/2}$ (см. рис. 11).

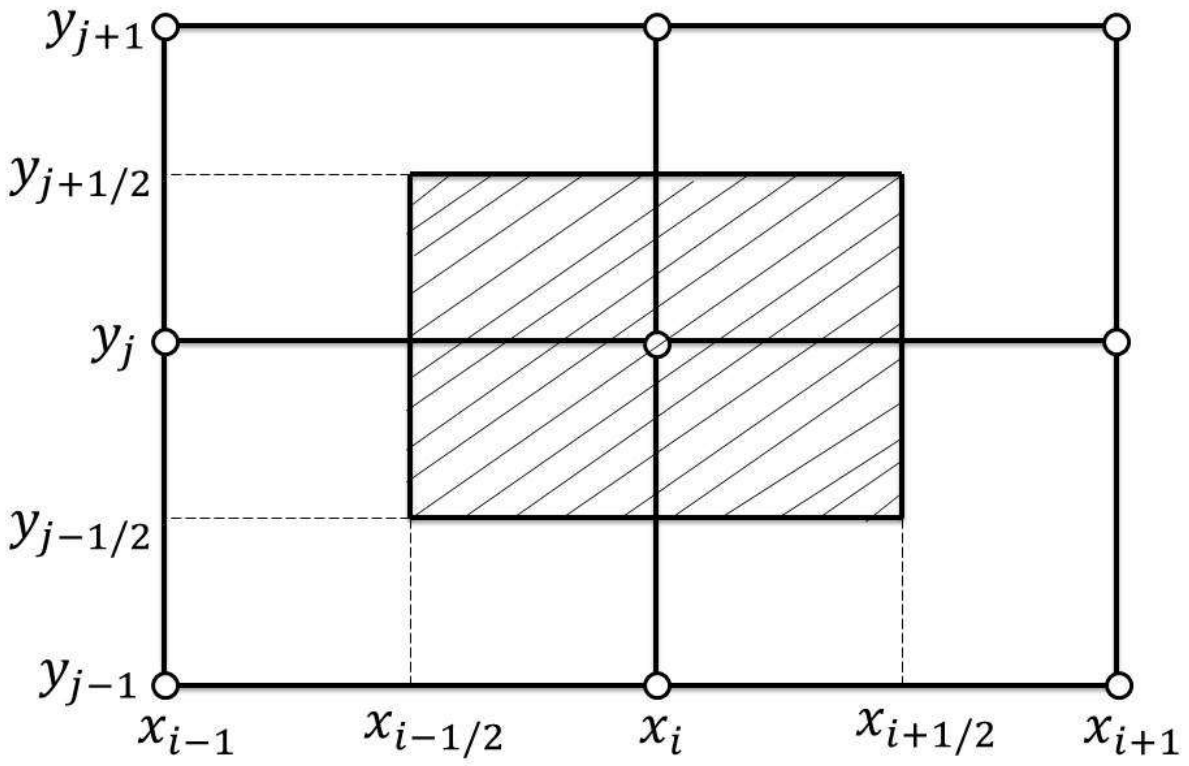

Рис. 11. Конечный объём вблизи узла (x_i, y_j)

$$\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left[w^x_{i+\frac{1}{2}}(y) - w^x_{i-\frac{1}{2}}(y) \right] dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left[w^y_{j+\frac{1}{2}}(x) - w^y_{j-\frac{1}{2}}(x) \right] dx + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q\, u\, dx\, dy = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f\, dx\, dy \qquad (2.5.3)$$

Интегралы, входящие в (2.5.3) вычислим приближенно, используя формулу средних прямоугольников.

$$\left(w^x_{i+\frac{1}{2},j} - w^x_{i-\frac{1}{2},j} \right) h_y + \left(w^y_{i,j+\frac{1}{2}} - w^y_{i,j-\frac{1}{2}} \right) h_x + q_{ij} u_{ij} h_x h_y \approx f_{ij} h_x h_y \quad (2.5.4)$$

В этой формуле принято обозначение $w^x_{i+\frac{1}{2},j} = -\left(a \frac{\partial u}{\partial x} \right)_{i+\frac{1}{2},j}$. Воспользовавшись центральной разностной производной, запишем $w^x_{i+\frac{1}{2},j} \approx -a_{i+\frac{1}{2},j} \frac{u_{i+1,j} - u_{i,j}}{h_x}$. Аналогично для $w^x_{i-\frac{1}{2},j}$, $w^y_{i,j+\frac{1}{2}}$, $w^y_{i,j-\frac{1}{2}}$.

С учётом этих приближений получаем СЛАУ в виде

$$-\frac{1}{h_x}\left(a_{i+\frac{1}{2},j}\frac{v_{i+1,j}-v_{i,j}}{h_x}-a_{i-\frac{1}{2},j}\frac{v_{i,j}-v_{i-1,j}}{h_x}\right)-\frac{1}{h_y}\left(b_{i,j+\frac{1}{2}}\frac{v_{i,j+1}-v_{i,j}}{h_y}-b_{i,j-\frac{1}{2}}\frac{v_{i,j}-v_{i,j-1}}{h_y}\right)+q_{ij}v_{ij}=f_{ij} \quad (2.5.5)$$

В системе (2.5.5) $i=1,2,\dots,M-1; j=1,\dots,N-1$. Граничные условия $v|_\Gamma=\varphi|_\Gamma$. Нетрудно видеть, что разностная схема (2.5.5) имеет пятиточечный шаблон.

Перед тем как перейти к решению системы уравнений необходимо определить порядок нумерации переменных, потому что от этого будет зависеть структура матрицы системы. Для определенности примем схему нумерации как показано на рис. 12.

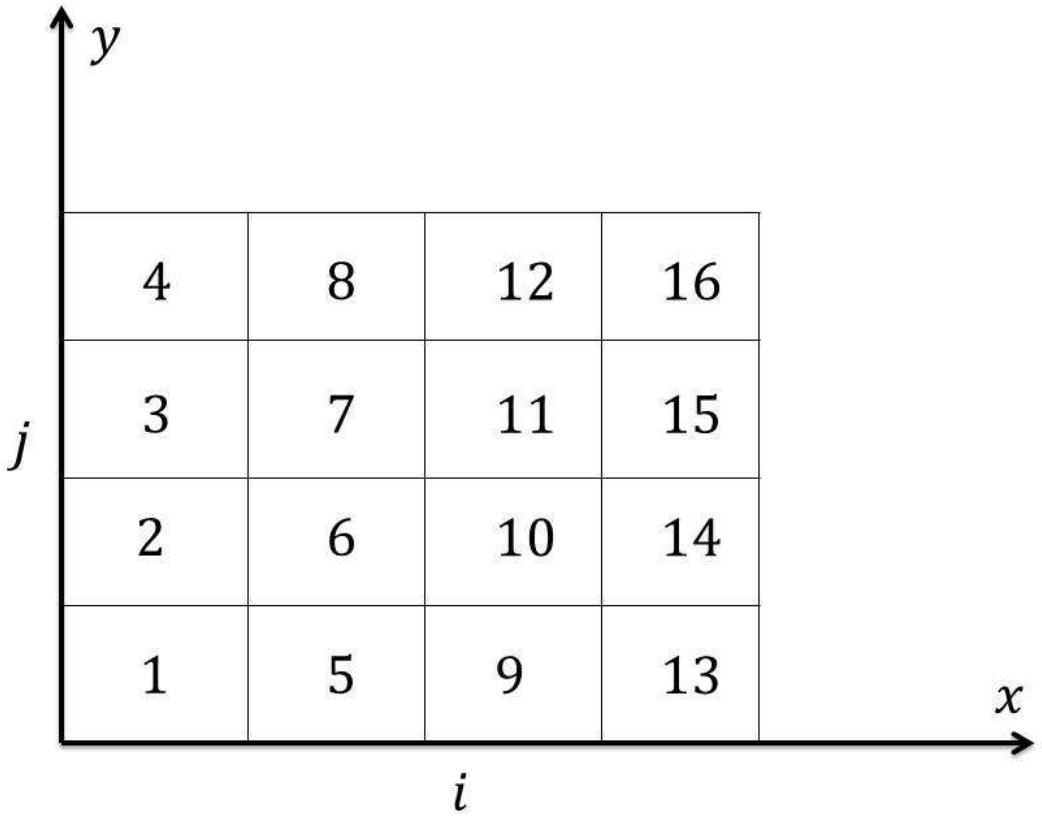

Рис. 12. Схема нумерациионечных объёмов, соответствующих узлам сетки

На рис. 12 изображена схема для $M-1=N-1=4$ (в качестве примера).

Большинство матричных элементов в СЛАУ нулевые. Ненулевые внедиагональные элементы располагаются на тех местах, с которыми граничит та или иная ячейка на рис. 12. Поясним это подробнее.

Рассмотрим первое уравнение СЛАУ. Ячейка с номером 1 граничит с ячейками номер 2 и 5. Следовательно, в первом уравнении отличны от нуля только первый, второй и пятый коэффициенты. Аналогично, во втором уравнении отличны от нуля коэффициенты №1, №2, №3 и №6. На рис. 13 показана структура матрицы СЛАУ, ненулевые элементы отмечены символом ×.

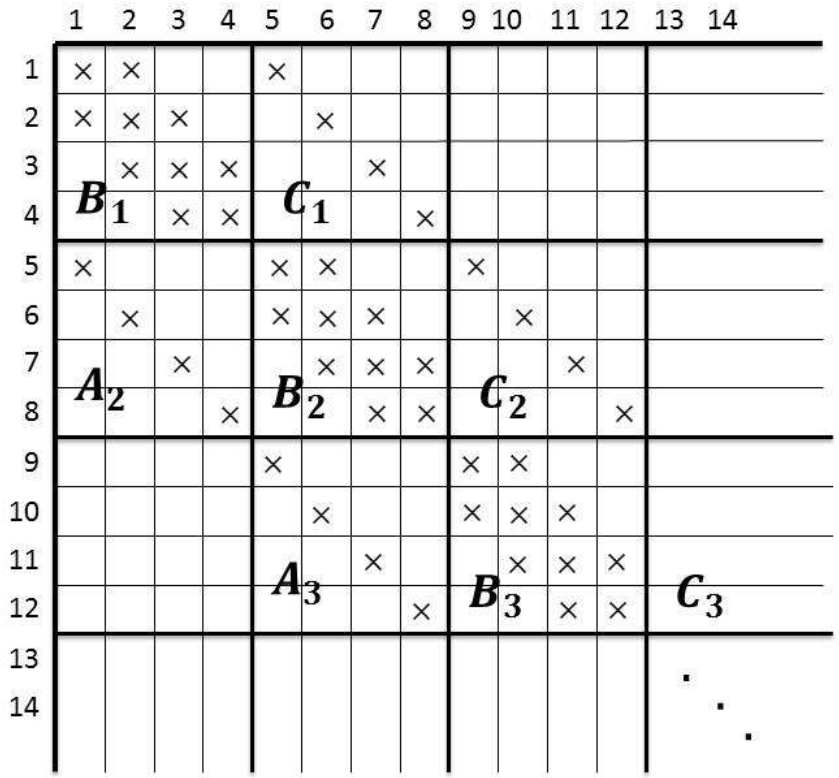

Рис. 13. Структура ненулевых элементов матрицы

На рис. 13 видно, что матрица СЛАУ имеет блочный вид, размер блоков 4×4. Обозначим блоки, стоящие на диагонали общей матрицы $B_1, B_2, \ldots$; блоки, стоящие на нижней квазидиагонали $A_2, A_3, \ldots$; блоки на верхней квазидиагонали $C_1, C_2, \ldots$

СЛАУ такого вида может быть решена с помощью метода матричной прогонки. В рамках этого метода столбцы неизвестных и правой части нужно разбить на блоки (в рассмотренном примере каждый блок состоит из 4 элементов). Обозначим такие блоки $\overline{v_1}$, $\overline{v_2}$,… и $\overline{f_1}$, $\overline{f_2}$,…

Блоки матрицы СЛАУ, показанные на рис. 13, будем обозначать $\overline{\overline{A_i}}$, $\overline{\overline{B_i}}$, $\overline{\overline{C_i}}$. Тогда общая система уравнений может быть представлена в следующем виде

$$\overline{\overline{A_i}}\,\overline{v_{i-1}} + \overline{\overline{B_i}}\,\overline{v_i} + \overline{\overline{C_i}}\,\overline{v_{i+1}} = \overline{f_i},\ i = 1, \dots, M-1 \quad (2.5.6)$$

Здесь $\overline{\overline{A_1}} = \overline{\overline{0}},\ \overline{\overline{C_M}} = \overline{\overline{0}}$. Ход метода матричной прогонки аналогичен ходу обыкновенного метода прогонки за тем исключением, что в матричном варианте вместо прогоночных коэффициентов используются прогоночные матрицы $\overline{\overline{\alpha_i}}$ и столбцы $\overline{\beta_i}$.

$$\overline{v_i} = \overline{\overline{\alpha_i}}\,\overline{v_{i+1}} + \overline{\beta_i} \quad (2.5.7)$$

Рекуррентные соотношения для прогоночных матриц имеют следующий вид

$$\overline{\overline{\alpha_i}} = -\left(\overline{\overline{A_i}}\,\overline{\overline{\alpha_{i-1}}} + \overline{\overline{B_i}}\right)^{-1}\overline{\overline{C_i}} \quad (2.5.8)$$

Количество машинных операций в методе матричной прогонки пропорционально $(N-1)^3(M-1)$.

Глава 3. Метод конечных элементов

§3.1. Кусочно-полиномиальная аппроксимация одномерной функции

Метод конечных элементов является достаточно мощным инструментом численного решения задач математической физики. Основой метода является аппроксимация решения дифференциальной задачи комбинацией функций из некоторого базисного набора.

Вначале рассмотрим аппроксимацию одномерной вещественной функции$f(x)$ кусочно-полиномиальными функциями. Разобьем интервал $x_0 \le x \le x_n$ на n подынтервалов: $[x_i, x_{i+1}](i = 0,1,2, \dots, n-1)$.

Наиболее простой пример кусочно-полиномиальной аппроксимации – это кусочно-линейная аппроксимация.

$$p_1^{(i)}(x) = \alpha_i(x) f_i + \beta_{i+1}(x) f_{i+1}, (x_i \le x \le x_{i+1}) \quad (3.1.1)$$

На отрезке $[x_i, x_{i+1}]$ функция $p_1^{(i)}(x)$ должна удовлетворять граничным условиям$p_1^{(i)}(x_i) = f_i; p_1^{(i)}(x_{i+1}) = f_{i+1}$.Из этих условий можно получить явный вид функций $\alpha_i(x)$ и $\beta_{i+1}(x)$. Подставим граничные условия в выражение (3.1.1):

$$\begin{cases} \alpha_i(x_i)f_i + \beta_{i+1}(x_i)f_{i+1} = f_i \\ \alpha_i(x_{i+1})f_i + \beta_{i+1}(x_{i+1})f_{i+1} = f_{i+1} \end{cases}$$

Очевидно, что для выполнения граничных условий для функции $p_1^{(i)}(x)$ необходимо и достаточно $\alpha_i(x_i) = 1,\ \alpha_i(x_{i+1}) = 0,$ а также $\beta_i(x_i) = 0,\ \beta_i(x_{i+1}) = 1.$ Найдём функцию $\alpha_i(x)$, зная, что она имеет линейный характер $\alpha_i(x) = k_i x + b_i.$ Используем граничные условия для$\alpha_i(x)$.

$$\begin{cases} k_i x_i + b_i = 1 \\ k_i x_{i+1} + b_i = 0 \end{cases}$$

Отсюда $k_i = -1/(x_{i+1} - x_i),\ b_i = x_{i+1}/(x_{i+1} - x_i).$ Таким образом, функция $\alpha_i(x)$ имеет вид

$$\alpha_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} \qquad (3.1.2)$$

Аналогично находится функция $\beta_{i+1}(x)$.

$$\beta_{i+1}(x) = \frac{x - x_i}{x_{i+1} - x_i} \qquad (3.1.3)$$

Здесь $i = 0,1,2, \dots, n-1.$ Функции $p_1^{(i)}(x)$ задаются на небольшом промежутке $x_i \le x \le x_{i+1}$ и называются кусочными базисными функциями. Аппроксимация $p_1(x)$ выражается через полные базисные функции $\varphi_i(x)$, которые определены на всём интервале изменения переменной $x_0 \le x \le x_n$.

$$p_1(x) = \sum_{i=0}^{n} \varphi_i(x) f_i,\ x_0 \le x \le x_n \qquad (3.1.4)$$

$$\varphi_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0}, x_0 \le x \le x_1 \\ 0,\ x_1 \le x \le x_n \end{cases} \qquad (3.1.5)$$

$$\varphi_i(x) = \begin{cases} 0, x_0 \le x \le x_{i-1} \\ \frac{x - x_{i-1}}{x_i - x_{i-1}}, x_{i-1} \le x \le x_i \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, x_i \le x \le x_{i+1} \\ 0, x_{i+1} \le x \le x_n \end{cases} \text{ для } i = 1, \dots, n-1 \qquad (3.1.6)$$

$$\varphi_n(x) = \begin{cases} 0,\ x_0 \le x \le x_{n-1} \\ \frac{x - x_{n-1}}{x_n - x_{n-1}}, x_{n-1} \le x \le x_n \end{cases} \qquad (3.1.7)$$

В рассмотренном примере полные базисные функции имеют пирамидальный вид (см. рис. 14).

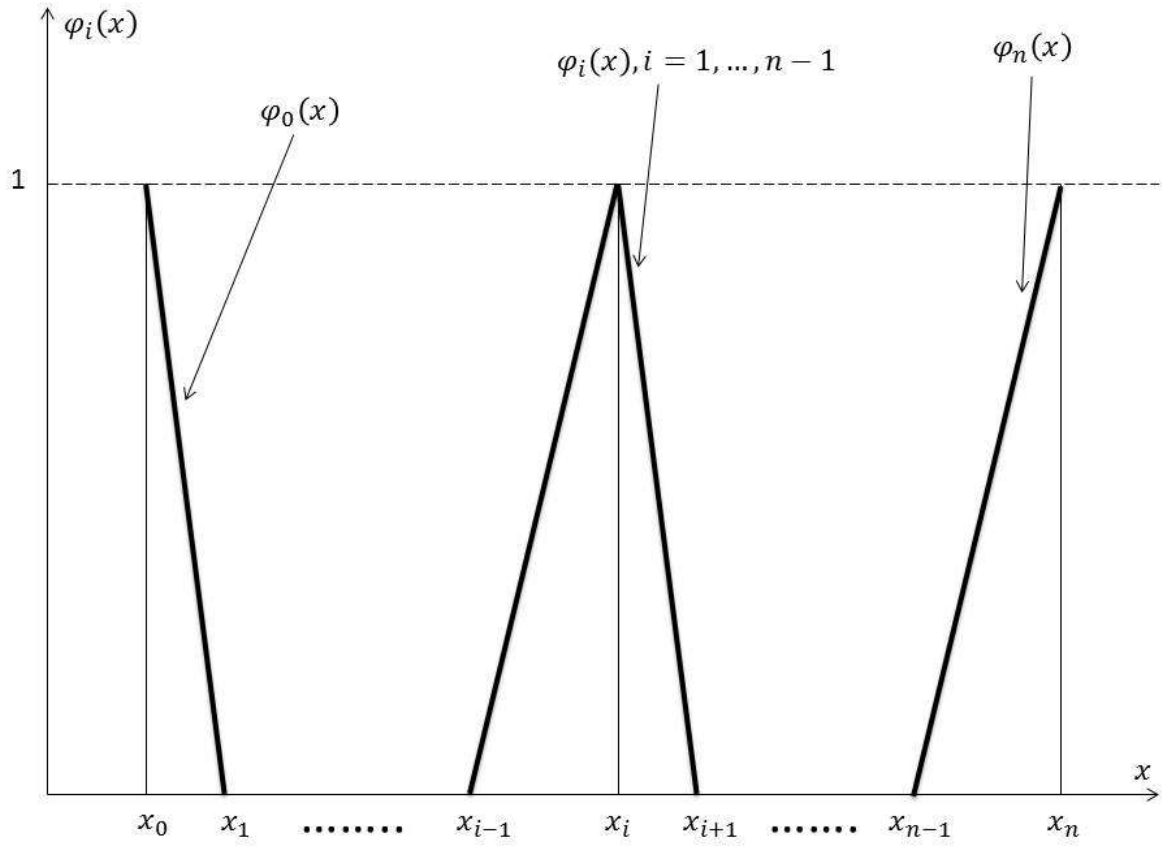

Рис. 14. Полные базисные функции в кусочно-линейной аппроксимации

Нетрудно заметить, что каждая полная базисная функция $\varphi_i(x)$ в рассмотренном примере отлична от нуля только на небольшом интервале изменения x. Этот интервал называют локальным носителем. Для базисных функций с номерами $i = 1,2, \dots, n-1$ локальный носитель $x_{i-1} \le x \le x_{i+1}$, для $i = 0$ - $x_0 \le x \le x_1$, а для $i = n$ - $x_{n-1} \le x \le x_n$.

Кусочно-линейная функция подходит для аппроксимации функции в том случае, когда интерес представляют только значения самой функции. Если требуется соответствие производных, то кусочно-линейная аппроксимация, очевидным образом, не подходит. В этом случае требуется строить более сложную аппроксимацию кубическими полиномами$p_3(x)$.

$$\left.\frac{df}{dx}\right|_{x=x_i} = \left.\frac{dp_3}{dx}\right|_{x=x_i} \&f|_{x=x_i} = p_3|_{x=x_i};\ i = 0,1, \dots, n \qquad (3.1.8)$$

В кусочно-полиномиальной аппроксимации кусочные базисные функции выражаются через значения самой функции $f(x)$ и её производной.

$$p_3^{(i)}(x) = \alpha_i(x)f_i + \beta_{i+1}(x)f_{i+1} + \gamma_i(x)f'_i +$$

$$\delta_{i+1}(x)f'_{i+1};\ x_i \le x \le x_{i+1} \quad (3.1.9)$$

В этой формуле кусочные базисные функции$\alpha_i(x)$, $\beta_{i+1}(x)$, $\gamma_i(x)$ и $\delta_{i+1}(x)$являются кубическими полиномами. На интервале $x \in [x_i, x_{i+1}]$функция $p_3^{(i)}(x)$ должна удовлетворять следующим граничным условиям: $p_3^{(i)}(x_i) = f_i, p_3^{(i)}(x_{i+1}) = f_{i+1}, p_3^{(i)\prime}(x_i) = f'_i, p_3^{(i)\prime}(x_{i+1}) = f'_{i+1}$. Соответственно, коэффициенты кусочных базисных функций определяются из следующих условий

$$\begin{cases} \alpha_i(x_i)=1;\ \alpha_i'(x_i)=0;\ \alpha_i(x_{i+1})=0;\ \alpha_i'(x_{i+1})=0 \\ \beta_{i+1}(x_i)=0;\ \beta_{i+1}'(x_i)=0;\ \beta_{i+1}(x_{i+1})=1;\ \beta_{i+1}'(x_{i+1})=0 \\ \gamma_i(x_i)=0;\ \gamma_i'(x_i)=1;\ \gamma_i(x_{i+1})=0;\ \gamma_i'(x_{i+1})=0 \\ \delta_{i+1}(x_i)=0;\ \delta_{i+1}'(x_i)=0;\ \delta_{i+1}(x_{i+1})=0;\ \delta_{i+1}'(x_{i+1})=1 \end{cases} \quad (3.1.10)$$

Для каждой из четырёх кусочных базисных функций задача нахождения коэффициентов полинома третьей степени сводится к решению системы линейных алгебраических уравнений 4×4. В качестве примера выведем явный вид функции $\alpha_i(x) = a_0 + a_1(x - x_i) + a_2(x - x_i)^2 + a_3(x - x_i)^3$. Используем граничные условия, записанные в первой строке 3.1.10.

$$\begin{cases} a_0 = 1 \\ a_1 = 0 \\ a_0 + a_1(x_{i+1} - x_i) + a_2(x_{i+1} - x_i)^2 + a_3(x_{i+1} - x_i)^3 = 0 \\ a_1 + 2a_2(x_{i+1} - x_i) + 3a_3(x_{i+1} - x_i)^2 = 0 \end{cases}$$

Решив систему, находим$a_2 = -3/(x_{i+1} - x_i)^2$, $a_3 = 2/(x_{i+1} - x_i)^3$.

Таким образом, функция $\alpha_i(x)$ имеет вид

$$\alpha_i(x) = \frac{(x_{i+1}-x)^2[(x_{i+1}-x_i)+2(x-x_i)]}{(x_{i+1}-x_i)^3} \quad (3.1.11)$$

Явный вид остальных кусочных базисных функций находится аналогично.

$$\beta_{i+1}(x) = \frac{(x-x_i)^2[(x_{i+1}-x_i)+2(x_{i+1}-x)]}{(x_{i+1}-x_i)^3} \quad (3.1.12)$$

$$\gamma_i(x) = \frac{(x-x_i)(x_{i+1}-x)^2}{(x_{i+1}-x_i)^2} \quad (3.1.13)$$

$$\delta_{i+1}(x) = \frac{(x-x_i)^2(x-x_{i+1})}{(x_{i+1}-x_i)^2} \quad (3.1.14)$$

Здесь $i = 0,1, \dots, n$. Как и в кусочно-линейной аппроксимации, полные базисные функции $\varphi_i^{(0)}(x)$, $\varphi_i^{(1)}(x)$ выражаются через кусочные базисные функции. Полные базисные функции обладают локальным носителем $x_{i-1} \le x \le x_{i+1}$. Сама аппроксимация записывается в виде

$$p_3(x) = \sum_{i=0}^{n} \left[\varphi_i^{(0)}(x) f_i + \varphi_i^{(1)}(x) f'_i \right] \qquad (3.1.15)$$

Видно, что $p_3^{(i)}$ содержит в качестве степеней свободы, как значения функции f_i, так и её производной f'_i. Такие элементы называются эрмитовыми. Те элементы, которые содержат только значения функции, называются лагранжевыми. Примером лагранжевых элементов являются $p_1^{(i)}(x)$ (кусочно-линейная аппроксимация).

Введение значений производной в качестве дополнительных параметров усложняет задачу, так как увеличивает размерность системы уравнений. В тех задачах, где не требуется определять значения производной, однако требуется плавность аппроксимации, удобно использовать сплайны.

Квадратичный сплайн обеспечивает непрерывность первой производной в каждой внутренней узловой точке $x_i (i = 1, \dots, n-1)$. На подынтервале $x_i \le x \le x_{i+1}$ он имеет следующий вид

$$S_2^{(i)}(x) = f_i + \frac{f_{i+1} - f_i}{x_{i+1} - x_i}(x - x_i) + c_i (x - x_i)(x - x_{i+1}),$$

$$i = 0,1, \dots, n-1 \qquad (3.1.16)$$

Коэффициент c_i, входящий в эту формулу, находится из условия непрерывности первых производных.

$$\frac{d}{dx} S_2^{(i-1)}(x)\Big|_{x=x_i} = \frac{d}{dx} S_2^{(i)}(x)\Big|_{x=x_i} \qquad (3.1.17)$$

В случае равномерного разбиения рассматриваемого диапазона по оси x с шагом h приходим к следующему условия дляc_i

$$c_i + c_{i-1} = \frac{1}{h^2}(f_{i+1} - 2f_i + f_{i-1}), i = 1, \dots, n-1 \qquad (3.1.18)$$

Здесь $n-1$ уравнений и n неизвестных, следовательно, один параметр остаётся свободным. Заметим, что вторая производная квадратичного сплайна $S_2^{(i)\prime\prime}(x) = 2c_i, i = 0,1, \dots, n-1$. Таким образом, задание второй производной в любой точке полностью обеспечивает построение квадратичного сплайна.

У квадратичного сплайна есть недостаток, состоящий в том, что этот сплайн не обладает равномерной сходимостью к аппроксимируемой функции при $h \to 0$. Это означает, что если задать вторую производную, например, на левом конце интервала, то по мере приближения к правому концу квадратичный сплайн начинает осциллировать тем сильнее, чем меньше h.

Указанного недостатка лишён кубический сплайн, обеспечивающий непрерывность первой и второй производных во внутренних узловых точках. На подынтервале$x_i \le x \le x_{i+1}$ вторая производная кубического сплайна имеет следующий вид

$$S_3^{(i)''}(x) = c_i \frac{x_{i+1}-x}{x_{i+1}-x_i} + c_{i+1} \frac{x-x_i}{x_{i+1}-x_i};\ i = 0,1,\dots,n-1 \qquad (3.1.19)$$

В этой формуле коэффициенты c_i и c_{i+1} имеют смысл значений второй производной кубического сплайна в точках x_i и x_{i+1} соответственно. Вид кубического сплайна можно получить, используя следующие условия

$$\left.\begin{matrix} S_3^{(i)}(x_i)=f_i \\ S_3^{(i)}(x_{i+1})=f_{i+1} \end{matrix}\right\}\ i=0,1,\dots,n-1$$
$$S_3^{(i-1)'}(x)=S_3^{(i)'}(x);\ i=1,\dots,n-1 \qquad (3.1.20)$$

Считая разбиение интервала по оси x равномерным с шагом h, получаем

$$S_3^{(i)}(x) = \frac{c_i}{6h}(x_{i+1}-x)^3 + \frac{c_{i+1}}{6h}(x-x_i)^3 + \left(\frac{f_i}{h} - \frac{hc_i}{6}\right)(x_{i+1}-x) + \left(\frac{f_{i+1}}{h} - \frac{hc_{i+1}}{6}\right)(x-x_i);\ \ i = 0,1,\dots,n-1 \qquad (3.1.21)$$

Где

$$c_{i+1} + 4c_i + c_{i-1} = \frac{6}{h^2}(f_{i+1} - 2f_i + f_{i-1}); i = 1,\dots,n-1 \qquad (3.1.22)$$

В выражении (3.1.22) $n-1$ уравнений и $n+1$ неизвестных, т.е. два параметра свободны. На практике часто полагают $c_0 = c_n = 0$.

§3.2. Кусочно-полиномиальная аппроксимация двумерной функции

В этом разделе будем рассматривать функцию двух вещественных переменных $f(x,y)$, заданную на ограниченной области R с границей ∂R. Разобьём область R на конечное число элементов.

Пусть R – прямоугольная область $[x_0, x_m] \times [y_0, y_n]$. Наиболее удобно разбить такую область на прямоугольники $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, $i = 0,1,\dots,m-1; j = 0,1,\dots,n-1$. Для простоты будем считать в дальнейшем разбиение равномерным $x_{i+1} - x_i = h_1;\ y_{j+1} - y_j = h_2$.

Одной из самых простых аппроксимаций двумерной функции на прямоугольном элементе является билинейнаяаппроксимация.

$$p_1^{(i,j)}(x,y) = \alpha_{i,j}(x,y)f_{i,j} + \beta_{i+1,j}(x,y)f_{i+1,j} + \gamma_{i,j+1}(x,y)f_{i,j+1} + \delta_{i+1,j+1}(x,y)f_{i+1,j+1};\ \ x_i \le x \le x_{i+1} \otimes y_j \le y \le y_{j+1} \qquad (3.2.1)$$

Кусочные базисные функции $\alpha_{i,j}(x,y), \beta_{i+1,j}(x,y), \gamma_{i,j+1}(x,y)$ и $\delta_{i+1,j+1}(x,y)$ представляют собой произведение полинома первой степени переменной x на полином первой степени переменной y. Коэффициенты находятся из условий равенства нулю или единице значений кусочных базисных функций в угловых точках элемента (метод аналогичен одномерной аппроксимации). Решив соответствующие системы уравнений для коэффициентов каждой из базисных функций, получим

$$\begin{cases} \alpha_{i,j}(x,y) = \frac{1}{h_1 h_2}(x_{i+1} - x)(y_{j+1} - y) \\ \beta_{i+1,j}(x,y) = \frac{1}{h_1 h_2}(x - x_i)(y_{j+1} - y) \\ \gamma_{i,j+1}(x,y) = \frac{1}{h_1 h_2}(x_{i+1} - x)(y - y_j) \\ \delta_{i+1,j+1}(x,y) = \frac{1}{h_1 h_2}(x - x_i)(y - y_j) \end{cases} \qquad (3.2.2)$$

Здесь $i = 0,1,\dots,m-1; j = 0,1,\dots,n-1$. Явный вид билинейной аппроксимации следующий:

$$p_1(x,y) = \sum_{i=0}^{m}\sum_{j=0}^{n} \varphi_{i,j}(x,y)f_{i,j};\ x_0 \le x \le x_m \otimes y_0 \le y \le y_n. \qquad (3.2.3)$$

В эту формулу входят полные базисные функции $\varphi_{i,j}(x,y)$,которые выражаются через кусочные базисные функции.Как и в одномерном случае, полные базисные функции имеют локальный носитель.

Выпишем полные базисные функции в явном виде для квадратной области размером 1.

$$\varphi_{i,j}(x,y)=\begin{cases}\left[\frac{x}{h}-(i-1)\right]\left[\frac{y}{h}-(j-1)\right]; & i-1\le\frac{x}{h}\le i;\ j-1\le\frac{y}{h}\le j\\ \left[\frac{x}{h}-(i-1)\right]\left[(j+1)-\frac{y}{h}\right]; & i-1\le\frac{x}{h}\le i;\ j\le\frac{y}{h}\le j+1\\ \left[(i+1)-\frac{x}{h}\right]\left[\frac{y}{h}-(j-1)\right]; & i\le\frac{x}{h}\le i+1;\ j-1\le\frac{y}{h}\le j\\ \left[(i+1)-\frac{x}{h}\right]\left[(j+1)-\frac{y}{h}\right]; & i\le\frac{x}{h}\le i+1;\ j\le\frac{y}{h}\le j+1\end{cases} \quad (3.2.4)$$

Здесь $1\le i,j\le m-1; mh=1$. В выражении (3.2.4) выписан вид функции на тех участках, где она отлична от нуля.

В двумерном случае, как и в одномерном, можно строить более сложные аппроксимации, которые обеспечивают непрерывность не только самой функции, но и её производных. В литературе описано большое количество таких аппроксимаций, см., например, [4].

Рассмотрим теперь многоугольную область. Самым распространенным разбиением области на конечные элементы для этого случая является разбиение на треугольные элементы (рис. 15).

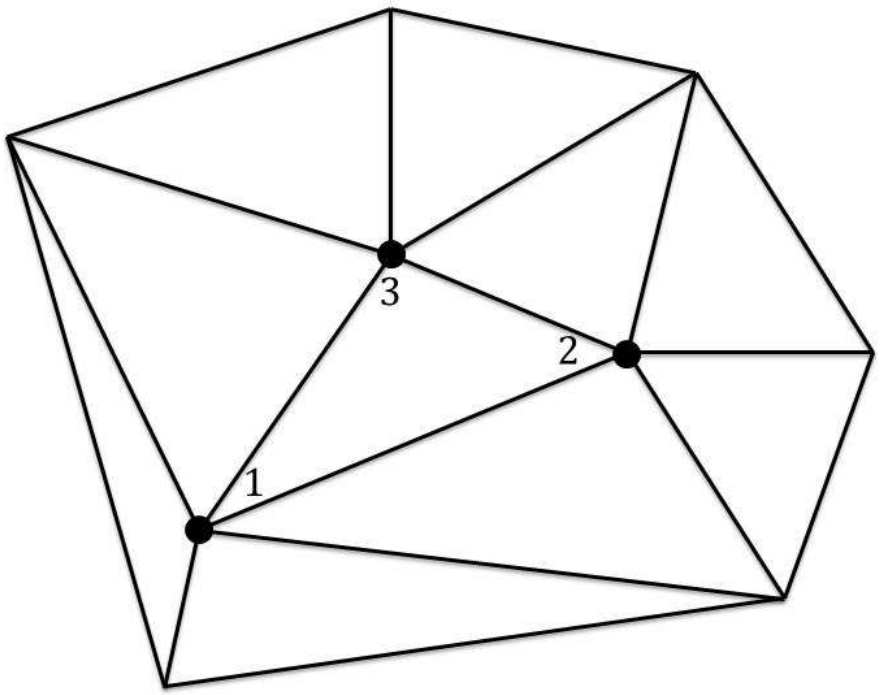

Рис. 15. Разбиение двумерной многоугольной области на треугольные конечные элементы

На треугольнике с вершинами $(x_i, y_i)(i = 1,2,3)$ линейная аппроксимация имеет следующий вид

$$p_1^{(k)}(x, y) = \sum_{i=1}^{3} \alpha_i(x, y) f_i \tag{3.2.5}$$

Кусочные базисные функции, входящие в эту формулу, могут быть несложным образом получены из тех соображений, что каждая из трёх кусочных базисных функций принимает значение 1 в одной из вершин треугольника и 0 в двух других вершинах.

$$\begin{cases} \alpha_1(x, y) = \frac{1}{C_{123}}(x_2 y_3 - x_3 y_2 + (y_2 - y_3)x - (x_2 - x_3)y) \\ \alpha_2(x, y) = \frac{1}{C_{123}}(x_3 y_1 - x_1 y_3 + (y_3 - y_1)x - (x_3 - x_1)y) \\ \alpha_3(x, y) = \frac{1}{C_{123}}(x_1 y_2 - x_2 y_1 + (y_1 - y_2)x - (x_1 - x_2)y) \end{cases} \tag{3.2.6}$$

Здесь $|C_{123}|$ есть удвоенная площадь треугольника. Полная базисная функция относительно какого-либо узла получается путём суммирования частей, связанных с примыкающими к этой вершине треугольниками. Полные базисные функции для этой аппроксимации имеют пирамидальный вид. Например, для узла №1 (см. рис. 15) полная базисная функция состоит из 5 частей, так как к этому узлу примыкают 5 треугольников.

§3.3. Вариационная формулировка дифференциальных уравнений

В главе 1 было выведено уравнение Эйлера, являющееся необходимым условием экстремума функционала вида

$$I(u) = \int_a^b F(x, u, u')dx \tag{3.3.1}$$

Уравнение Эйлера выглядит следующим образом

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) = 0 \tag{3.3.2}$$

Среди решений дифференциального уравнения (3.3.2) необходимо выбрать то решение, которое удовлетворяет заданным граничным условиям задачи.

Таким образом, задача решения уравнения (3.3.2) с заданными граничными условиями равносильна задаче на экстремум функционала (3.3.1) на множестве функций $u(x)$, которые удовлетворяют заданным

граничным условиям. В случае, когда u – функция одного аргумента, (3.3.2) является обыкновенным дифференциальным уравнением. Однако вариационную формулировку имеют не только обыкновенные дифференциальные уравнения, но также и уравнения в частных производных.

Обобщим подход на двумерный случай. Рассмотрим функционал

$$I_1(u) = \iint F\big(x, y, u, u_x, u_y\big)dxdy \qquad (3.3.3)$$

Интегрирование в этой формуле проводится по некоторой ограниченной области R, функция u полагается непрерывной на этой области вместе со всеми производными до второго порядка включительно. Значения функции на границе области $u|_{\partial R}$ считаются заданными.

Для того чтобы вывести необходимое условие экстремума функционала (3.3.3), как и в одномерном варианте, вводим приращение к функции $u{:}h(x, y)$ – произвольная дважды непрерывно дифференцируемая функция, которая обращается в нуль на границе рассматриваемой области $h(x, y)|_{\partial R} = 0$. Запишем приращение функционала

$$\Delta I_1 = I_1(u + h) - I_1(u) = \iint \big[F\big(x, y, u + h, u_x + h_x, u_y + h_y\big) - F\big(x, y, u, u_x, u_y\big)\big]dxdy \qquad (3.3.4)$$

Интегрирование проводится по области R, далее для краткости будем подразумевать это по умолчанию. Разложим приращение в ряд Тейлора.

$$\Delta I_1 = \iint \left[\frac{\partial F(x,y,u,u_x,u_y)}{\partial u} h + \frac{\partial F(x,y,u,u_x,u_y)}{\partial u_x} h_x + \frac{\partial F(x,y,u,u_x,u_y)}{\partial u_y} h_y\right] dxdy + \cdots \qquad (3.3.5)$$

В формуле (3.3.5) многоточием обозначены члены разложения второго и более высоких порядков поh, h_x, h_y.

Вариация функционалаδI_1 является первым членом разложения приращения функционала в ряд Тейлора.

$$\delta I_1 = \iint \left[\frac{\partial F(x,y,u,u_x,u_y)}{\partial u} h + \frac{\partial F(x,y,u,u_x,u_y)}{\partial u_x} h_x + \frac{\partial F(x,y,u,u_x,u_y)}{\partial u_y} h_y\right] dxdy \qquad (3.3.6)$$

Необходимым условием экстремума является равенство нулю вариации функционала. Второе и третье подынтергальные слагаемые в

(3.3.6) для дальнейших преобразований удобнее представить в следующем виде

$$\begin{cases} h_x F_{u_x} = \frac{\partial}{\partial x}\left(hF_{u_x}\right) - h\frac{\partial}{\partial x}F_{u_x} \\ h_y F_{u_y} = \frac{\partial}{\partial y}\left(hF_{u_y}\right) - h\frac{\partial}{\partial y}F_{u_y} \end{cases} \quad (3.3.7)$$

Далее воспользуемся формулой Грина

$$\iint \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy = \oint P dx + Q dy \quad (3.3.8)$$

В формуле (3.3.8) двойной интеграл берётся по области R, а контурный – по её границе ∂R. По формуле (3.3.8) преобразуются первые слагаемые в выражениях системы (3.3.7).

При подстановке выражений (3.3.7) в (3.36) и последующем использовании формулы Грина возникает контурный интеграл по∂Rвида

$$\oint hF_{u_x}dy - hF_{u_y}dx$$

Этот интеграл обращается в нуль, потому что на∂Rфункцияh всюду принимает нулевое значение. В результате получим следующее выражение для вариации функционала

$$\delta I_1 = \iint \left(F_u - \frac{\partial}{\partial x}F_{u_x} - \frac{\partial}{\partial y}F_{u_y}\right) h(x,y)dxdy \quad (3.3.9)$$

Заметим, что необходимым условием экстремума функционала является равенство нулю вариации для любой функции $h(x,y)$ (удовлетворяющей однородным граничным условиям). Это условие выполняется только тогда, когда выражение в круглых скобках в (3.3.9) тождественно обращается в нуль. Итак, уравнение Эйлера-Лагранжа в двумерном случае имеет вид

$$\frac{\partial}{\partial x}F_{u_x} + \frac{\partial}{\partial y}F_{u_y} - F_u = 0 \quad (3.3.10)$$

Очевидно, что различные виды функции $F\left(x, y, u, u_x, u_y\right)$ приводят к различным дифференциальным уравнениям. Например, $F = 0.5\left(u_x^2 + u_y^2\right)$ приводит к уравнению Лапласа $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$.

Можно провести аналогичные выводы необходимого условия экстремума для более сложных функционалов. Рассмотрим некоторые из них (без доказательств).

1. Функционал, зависящий от двух функций одного и того же аргумента

$$I_2(u,v) = \int_{x_0}^{x_1} F\big(x, u(x), v(x), u'(x), v'(x)\big)dx \qquad (3.3.11)$$

Граничные условия $u(x_0), u(x_1), v(x_0), v(x_1)$ заданы. Необходимое условие экстремума для такого функционала имеет вид

$$\begin{cases} \frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) = 0 \\ \frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v'}\right) = 0 \end{cases} \qquad (3.3.12)$$

2. Наличие высших производных (ограничимся рассмотрением второй производной)

$$I_3(u) = \int_{x_0}^{x_1} F\big(x, u(x), u'(x), u''(x)\big)dx \qquad (3.3.13)$$

Значения $u(x_0), u'(x_0), u(x_1), u'(x_1)$ заданы. Необходимое условие экстремума

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial u''}\right) = 0 \qquad (3.3.14)$$

3. Условный экстремум. Ищется экстремум функционала

$$I_4(u) = \int_{x_0}^{x_1} F\big(x, u(x), u'(x)\big)dx \qquad (3.3.15)$$

при условии постоянства другого функционала

$$\int_{x_0}^{x_1} G\big(x, u(x), u'(x)\big)dx = \alpha = const \qquad (3.3.16)$$

В этом случае необходимое условие экстремума имеет вид

$$\frac{\partial(F+\lambda G)}{\partial u} - \frac{d}{dx}\frac{\partial(F+\lambda G)}{\partial u'} = 0 \qquad (3.3.17)$$

Где параметр λ определяется из условия (3.3.16).

Граничные условия в вариационных задачах бывают главные и естественные. Главные граничные условия – это когда функция задана на границе рассматриваемой области и не может там варьироваться. Вариационная задача с главными граничными условиями называется простейшей вариационной задачей. Если в задаче не заданы значения функции на границе области, то такая задача называется вариационной задачей со свободными концами. Граничные условия, которые вытекают из задачи минимизации функционала в задаче со свободными концами,

называются естественными граничными условиями. Продемонстрируем вывод естественных граничных условий для задачи минимизации функционала (3.3.1) на множестве функций $u(x)$, концы которых лежат на вертикалях $x = a$ и $x = b$. Как и ранее, вводим приращение функции $h(x)$, но теперь уже мы не накладываем на $h(x)$ однородные граничные условия (такое условие требовалось в задаче с фиксированными концами для того, чтобы обеспечить выполнение граничных условий). Запишем приращение функционала.

$$I(u+h) - I(u) = \int_a^b [F(x, u+h, u'+h') - F(x, u, u')]dx = \int_a^b [F_u h + F_{u'} h']dx + \cdots \quad (3.3.18)$$

Многоточием здесь обозначены члены разложения второго и более высоких порядков, а выписанное слагаемое в правом равенстве является вариацией функционала. Проинтегрируем по частям одно из слагаемых в вариации.

$$\int_a^b F_{u'} h' dx = h F_{u'}|_a^b - \int_a^b h \frac{d}{dx}(F_{u'})dx \quad (3.3.19)$$

Заметим, что внеинтегральное слагаемое в выражении (3.3.19) не равно нулю как это было в простейшей вариационной задаче. Вариация функционала имеет вид

$$\delta I = \int_a^b \left[F_u - \frac{d}{dx} F_{u'}\right] h(x)dx + F_{u'}|_{x=b} h(b) - F_{u'}|_{x=a} h(a) \quad (3.3.20)$$

Необходимое условие экстремума функционала $\delta I = 0 \; \forall h(x)$. Отсюда имеем следующую систему условий

$$\begin{cases} F_u - \frac{d}{dx} F_{u'} = 0 \\ F_{u'}|_{x=b} = F_{u'}|_{x=a} = 0 \end{cases} \quad (3.3.21)$$

Первое условие в (3.3.21) является хорошо известным уравнением Эйлера-Лагранжа, а второе и третье – это естественные граничные условия. Эти граничные условия получили такое название, потому что они вытекают непосредственно из задачи на экстремум функционала, а не задаются в самой задаче.

Существует большой класс дифференциальных задач, в которых значения функции на границе рассматриваемой области не заданы явно, и при этом граничные условия не совпадают с естественными граничными условиями функционала, который соответствует дифференциальному уравнению. В таких случаях необходимо определённым образом модифицировать функционал так, чтобы для нового функционала

граничные условия задачи были естественными. Нетрудно показать, что необходимым условием экстремума функционала

$$J(u) = \int_{x_0}^{x_1} F(x,u,u')dx + g_1(x,u)|_{x=x_1} - g_0(x,u)|_{x=x_0} \quad (3.3.22)$$

является уравнение Эйлера (3.3.2) совместно со следующими естественными граничными условиями

$$\begin{cases} \left(\frac{\partial F}{\partial u'} + \frac{\partial g_0}{\partial u}\right)\Big|_{x=x_0} = 0 \\ \left(\frac{\partial F}{\partial u'} + \frac{\partial g_1}{\partial u}\right)\Big|_{x=x_1} = 0 \end{cases} \quad (3.2.23)$$

Таким образом, можно, подобрав функции $g_0(x,u)$ и $g_1(x,u)$, свести дифференциальную задачу с произвольными граничными условиями к вариационной задаче со свободными концами для функционала (3.3.22). Например, уравнение $u'' + f(x) = 0$ с граничными условиями $(-u' + \alpha u)|_{x=x_0} = 0;\ (u' + \beta u)|_{x=x_1} = 0$соответствует задаче со свободными концами для функционала

$$J_1(u) = \int_{x_0}^{x_1} \left[\frac{1}{2}u'^2 - f(x)u\right]dx + \left[\frac{1}{2}\beta u^2\right]\Big|_{x=x_1} - \left[\frac{1}{2}\alpha u^2\right]\Big|_{x=x_0} \quad (3.2.24)$$

В заключение данного параграфа рассмотрим обобщение на двумерный случай, т.е. функционал вида (3.3.3). Естественные граничные условия в двумерном случае принимают вид

$$\left[\frac{\partial F}{\partial u_x}\frac{dy}{d\sigma} - \frac{\partial F}{\partial u_y}\frac{dx}{d\sigma}\right]\Big|_{\partial R} = 0 \quad (3.2.25)$$

Здесь $d\sigma$ обозначен элемент длины дуги вдоль границы ∂R.

Если функционал зависит от двух функций, каждая из которых зависит от двух аргументов, то к уравнению Эйлера-Лагранжа для первой функции u добавляется такое же уравнение для второй функции v. При этом к граничному условию (3.2.25) добавляется такое же условие с заменой uнаv.

В тех случаях, когда u и v не заданы на ∂R явно, и граничные условия задачи не совпадают с естественными граничными условиями функционала, необходимо модифицировать функционал (как это было в одномерном случае).

§3.4. Метод Ритца

В данном параграфе будет изложен классический прямой подход к решению дифференциальных задач методом конечных элементов. В рамках этого метода первым шагом является вариационная формулировка дифференциальной задачи.

Требуется решить вариационную задачу $\delta I(v) = 0 (I(v)$ – функционал) на пространстве функций $v(\boldsymbol{X}) \in H$, где H – бесконечномерное функциональное пространство, $\boldsymbol{X} = (x_1, x_2, \dots, x_m)$ – вектор, состоящий из m аргументов.

Базисом пространства H является бесконечный набор функций, поэтому для численного счета необходимо ограничиться рассмотрением некоторого конечномерного подпространства $K_N \subset H$, N – размерность рассматриваемого функционального подпространства. Выберем некоторый базис пространства K_N: $\varphi_i(\boldsymbol{X})(i = 1,2, \dots, N)$. Приближенное решение задачи $U(\boldsymbol{X})$ будем искать в виде разложения по выбранному базису.

$$U(\boldsymbol{X}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{X}) \quad (3.4.1)$$

Коэффициенты разложения α_i вычисляются из тех соображений, чтобы значение функционала $I(U)$ было стационарно относительно этих значений, т.е.

$$\frac{\partial}{\partial \alpha_i} I\left(\sum_{j=1}^{N} \alpha_j \varphi_j\right) = 0, (i = 1,2, \dots, N) \quad (3.4.2)$$

Это общая схема метода Ритца. Рассмотрим теперь пример применения этого метода.

$$\begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - 2 = 0; \ -\frac{\pi}{2} < x < \frac{\pi}{2}; \ -\frac{\pi}{2} < y < \frac{\pi}{2} \\ u\left(x, \pm\frac{\pi}{2}\right) = 0; u\left(\pm\frac{\pi}{2}, y\right) = 0 \end{cases} \quad (3.4.3)$$

Первым шагом является вариационная формулировка задачи. Граничные условия в задаче (3.4.3) заданы в явном виде, поэтому требуется только задать функционал $I(v)$, для которого уравнение Эйлера-Лагранжа совпадает с дифференциальным уравнением в (3.4.3). Нетрудно убедиться, что этот функционал имеет вид

$$I(v) = \iint \left[\frac{1}{2}\left(v_x^2 + v_y^2\right) + 2v\right] dxdy \quad (3.4.4)$$

Интегрирование в выражении (3.4.4) ведется по двумерной области $R = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Разобьем область R на конечные элементы. Так как область квадратная, удобно разбить её на $(T+1)^2$ квадратных элементов путем $2T$ равнорасположенных внутренних линий сетки, параллельных осям. Количество базисных функций совпадает с количеством внутренних точек $N = T^2$.

В тех задачах, где значения искомой функции на границе заданы явно (главные граничные условия), эти граничные условия накладываются на функциональное пространство H. Таким образом, в задаче (3.4.3) нам необходимо рассматривать базисные функции, обращающиеся в нуль на границе $\varphi_{ij}(x,y)\big|_{\partial R} = 0$.

Воспользуемся билинейной аппроксимацией на прямоугольных элементах, которая была изложена в §3.2. Несложно понять, что коэффициенты разложения искомой функции по такому базису совпадают со значениями этой функции во внутренних узлах сетки.

$$U(x,y) = \sum_{i,j=1}^{T} U_{ij}\varphi_{ij}(x,y) \quad (3.4.5)$$

Здесь $U_{ij} \equiv U(x_i, y_j)$. В соответствии с (3.4.2) имеем систему уравнений

$$\frac{\partial}{\partial U_{ij}} I\left(\sum_{k,l=1}^{T} U_{kl}\,\varphi_{kl}(x,y)\right) = 0;\ (i,j = 1,\dots,T) \quad (3.4.6)$$

Используя явный вид функционала (3.4.4), преобразуем полученное выражение.

$$\frac{\partial}{\partial U_{ij}} \iint \left[\frac{1}{2}\left(\sum_{k,l=1}^{T} U_{kl}\,\frac{\partial \varphi_{kl}(x,y)}{\partial x}\right)^2 + \frac{1}{2}\left(\sum_{k,l=1}^{T} U_{kl}\,\frac{\partial \varphi_{kl}(x,y)}{\partial y}\right)^2 + 2\sum_{k,l=1}^{T} U_{kl}\,\varphi_{kl}(x,y)\right] dxdy = 0 \quad (3.4.7)$$

Раскрыв частные производные $\partial/\partial U_{ij}$ в (3.4.7), можно упростить это выражение.

$$\iint \left[\left(\sum_{k,l=1}^{T} U_{kl}\,\frac{\partial \varphi_{kl}}{\partial x}\right)\left(\frac{\partial \varphi_{ij}}{\partial x}\right) + \left(\sum_{k,l=1}^{T} U_{kl}\,\frac{\partial \varphi_{kl}}{\partial y}\right)\left(\frac{\partial \varphi_{ij}}{\partial y}\right) + 2\varphi_{ij}\right] dxdy = 0 \quad (3.4.8)$$

Из формулы (3.4.8) легко получается явный вид системы линейных алгебраических уравнений.

$$\sum_{k,l=1}^{T} U_{kl} \iint \left\{ \left(\frac{\partial \varphi_{kl}}{\partial x}\right)\left(\frac{\partial \varphi_{ij}}{\partial x}\right) + \left(\frac{\partial \varphi_{kl}}{\partial y}\right)\left(\frac{\partial \varphi_{ij}}{\partial y}\right) \right\} dxdy + 2 \iint \varphi_{ij}\, dxdy = 0 \quad (3.4.9)$$

Коэффициенты полученной системы уравнений вычисляются при подстановке в (3.4.9) явного вида базисных функций билинейной аппроксимации и вычисления интегралов. В результате такой подстановки получим

$$3U_{ij} - \frac{1}{3}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1} U_{kl} + 2h^2 = 0 \ \ (i,j = 1, \dots, T) \quad (3.4.10)$$

В соответствии с граничными условиями $U_{kl} = 0$ при $k,l = 0$ или $k,l = T+1$. Решив систему (3.4.10), находим U_{ij}, и, подставляя эти значения в билинейную аппроксимацию, находим $U(x,y)$. Приведем результаты для узловых точек, отмеченных на рис. 16.

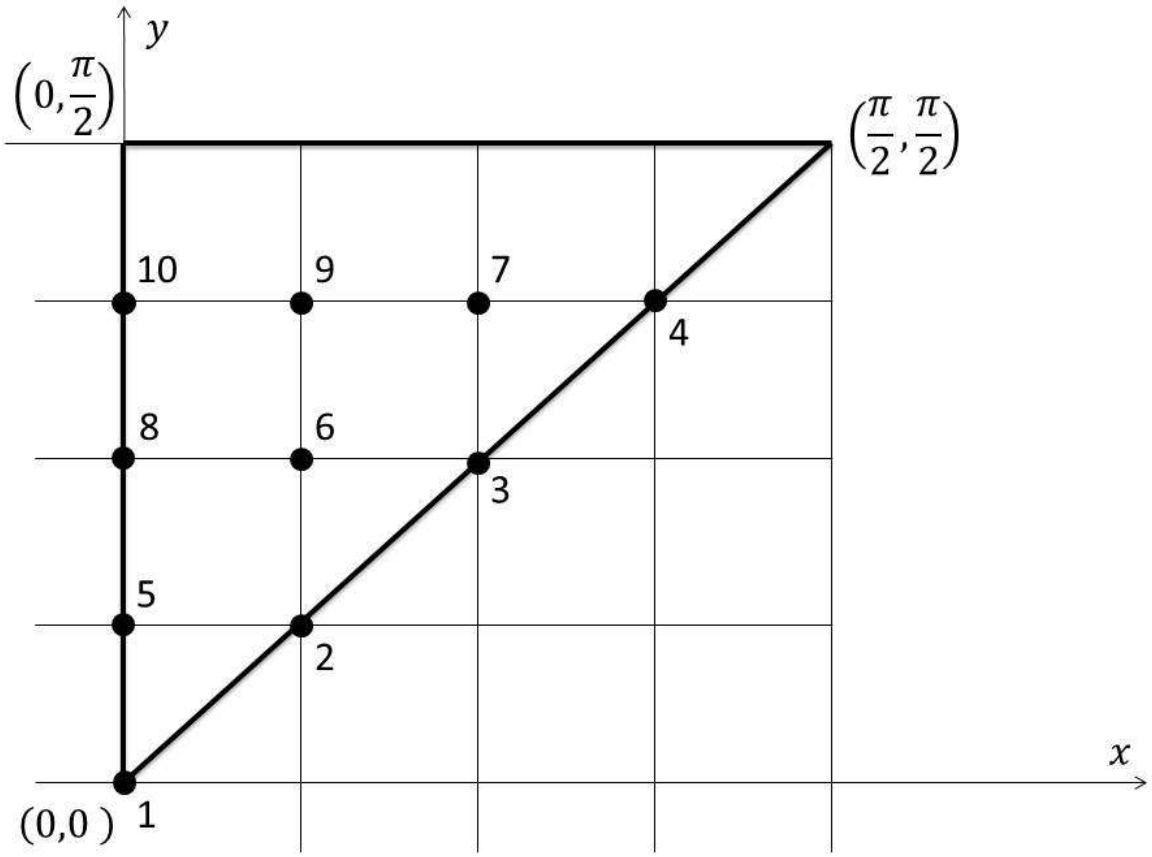

Рис. 16. Внутренние узловые точки

№ точки	МетодРитца(T=15)	Точноерешение
1	-1.459	-1.454
2	-1.308	-1.304

3	-0.897	-0.894
4	-0.362	-0.359
5	-1.381	-1.376
6	-1.078	-1.075
7	-0.559	-0.556
8	-1.135	-1.132
9	-0.660	-0.658
10	-0.692	-0.690

Ввиду симметрии рассматриваемой задачи, значения искомой функции на всей области R могут быть воспроизведены по значениям функции на 1/8 части области.

Для тех задач, где главные граничные условия не заданы, а вместо них заданы естественные граничные условия, необходимо расширить пространство базисных функций. К примеру, если бы в задаче (3.4.3) были заданы граничные условия

$$\begin{cases} \left.\frac{\partial u}{\partial y}\right|_{y=\pm\pi/2} = 0 \quad \left(|x| \le \frac{\pi}{2}\right) \\ \left.\frac{\partial u}{\partial x}\right|_{x=\pm\pi/2} = 0 \quad \left(|y| \le \frac{\pi}{2}\right) \end{cases} \tag{3.4.11}$$

(нетрудно убедиться, что эти граничные условия являются естественными для функционала (3.4.4)), то необходимо рассматривать базисные функции $\varphi_{ij}(x,y)$, которые могут не обращаться в нуль на границе области R. Для билинейной аппроксимации базисные функции могут быть отличны от нуля не более чем на двух приграничных элементах по каждой из координат.

§3.5. Метод Галеркина

Рассмотренный в предыдущем параграфе метод Ритца требует вариационной формулировки дифференциальной задачи. Однако метод конечных элементов может использоваться для решения более широкого класса задач, чем те, которые допускают вариационную формулировку. Основное преимущество метода Галеркина состоит в том, что этот метод не требует вариационной формулировки.

Будем рассматривать дифференциальную задачу с однородными граничными условиями.

$$Au(\boldsymbol{X}) = f, \boldsymbol{X} \in R^m, u|_{\partial R} = 0 \qquad (3.5.1)$$

Здесь A – некоторый дифференциальный оператор. Функция u, которая в точности удовлетворяет дифференциальному уравнению с указанными граничными условиями, называется точным решением.

Скалярно умножим обе части дифференциального уравнения (3.5.1) на некоторую пробную функцию $v \in H$, где H – пространство функций, обращающихся в нуль на границе $v|_{\partial R} = 0$.

$$(Au, v) = (f, v) \; \forall \; v \in H \qquad (3.5.2)$$

Выражение (3.5.2) называется слабой формой соответствующего дифференциального уравнения. Очевидно, что любая функция, удовлетворяющая (3.5.1), удовлетворяет слабой форме. Обратное утверждение, вообще говоря, требует строгого доказательства. Здесь мы не будем его приводить, но поясним на примере: если рассмотреть пространство трехмерных векторов, то из того, что скалярное произведение какого-либо вектора на любой вектор из трехмерного пространства равно нулю, следует, что этот вектор нулевой.

Если дифференциальный оператор имеет порядок $2k$, то наиболее удобная слабая форма (или галеркинская форма) получается в результате k интегрирований по частям скалярного произведения (Au, v).

Поясним это на примере дифференциального оператора

$$A = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} = -\Delta. \qquad (3.5.3)$$

Скалярное произведение (Au, v) преобразуется к виду

$$(Au, v) = \iint \left(-\frac{\partial^2 u}{\partial x^2} - -\frac{\partial^2 u}{\partial y^2} \right) v dx dy =$$

$$\iint \nabla u * \nabla v dx dy - \oint \frac{\partial u}{\partial \boldsymbol{n}} * v d\gamma \qquad (3.5.4)$$

В выражении (3.5.4) применена формула Грина (многомерный аналог интегрирования по частям). Двойные интегралы берутся по области R, а контурный – по её границе ∂R. $\boldsymbol{n}$– единичный вектор нормали к∂R. Так как пробные функции на границе обращаются в нуль, контурный интеграл в (3.5.4) также равен нулю. В оставшемся двойном интеграле раскроем скалярное произведение двух градиентов, используя декартову систему

координат.В результатегалеркинскаяформа оператора A записывается в виде

$$a(u,v) = \iint \left[\left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial v}{\partial x}\right) + \left(\frac{\partial u}{\partial y}\right)\left(\frac{\partial v}{\partial y}\right)\right] dxdy. \quad (3.5.5)$$

Следующим шагом является ограничение бесконечномерного пространства пробных функций H его конечномерным подпространством $K_N \subset H$, N – размерность этого подпространства. Приближенное решениеU (аппроксимация Галеркина) находится из решения следующей системы уравнений:

$$a(U, \varphi_l) = (f, \varphi_l),\ l = 1,2, \dots, N. \quad (3.5.6)$$

В этом выраженииφ_l – базисные функции функционального пространства K_N. В случае однородных граничных условий приближенное решение ищется в виде линейной комбинации базисных функций.

$$U(\boldsymbol{X}) = \sum_{l=1}^{N} \alpha_i \varphi_i(\boldsymbol{X}) \quad (3.5.7)$$

В двумерном случае суммирование по l предполагает двойную сумму по индексам i и j (так как сетка двумерная).Используя билинейные базисные функции, имеем

$$U(x,y) = \sum_{i,j=1}^{T} U_{ij} \varphi_{ij}(x,y) \quad (3.5.8)$$

Коэффициенты разложения искомой функции $U(x,y)$ по билинейным базисным функциям $\varphi_{ij}(x,y)$ совпадают со значениями функции U в узлах сетки – U_{ij}.

Таким образом, для оператора A, заданного формулой (3.5.3), мы приходим к системе уравнений дляU_{ij}, подставив разложение (3.5.8) в слабую форму (3.5.6).

$$\iint \left[\left(\frac{\partial U}{\partial x}\right)\left(\frac{\partial \varphi_{ij}}{\partial x}\right) + \left(\frac{\partial U}{\partial y}\right)\left(\frac{\partial \varphi_{ij}}{\partial y}\right)\right] dxdy = \iint f(x,y)\varphi_{ij}\, dxdy;\ \ i,j = 1, \dots, T \quad (3.5.9)$$

При подстановке следует поменять индексы суммирования в разложении искомой функции U по базисным функциям. Интегралы по пространственным координатам в левой части равенства несложно вычисляются, после чего получаем в явном виде систему уравнений для U_{ij}. Размерность этой системы равна T^2.

Если граничные условия неоднородны, то приближенное решение ищут в виде

$$U(\boldsymbol{X}) = W(\boldsymbol{X}) + \sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{X}) \qquad (3.5.10)$$

Где $W(\boldsymbol{X})$ – функция, удовлетворяющая неоднородным граничным условиям.

В заключение подчеркнем, что в методе Галеркина не требуется вариационной формулировки дифференциальной задачи в отличие от метода Ритца. Это делает метод Галеркина более универсальным.

Глава 4. Методы Монте-Карло

§4.1. Преобразования случайных величин

Методы Монте-Карло – это методы численного решения математических задач при помощи моделирования случайных величин. Основная идея этих методов состоит в сведении задачи к расчету математических ожиданий каких-либо случайных величин.

Таким образом, если требуется вычислить некоторую скалярную величину a, мы строим случайную величину ξ, такую что $M\xi = a$. После этого вычисляем N независимых реализаций случайной величины ($N \gg 1$) $\xi_1, \xi_2, \dots, \xi_N$. Далее используем статистическую оценку математического ожидания

$$a \approx \frac{1}{N}(\xi_1 + \xi_2 + \cdots + \xi_N) \qquad (4.1.1)$$

Очевидно, что существует бесконечно много случайных величин, удовлетворяющих условию $M\xi = a$. На практике необходимо выбрать наиболее удобную из них. Помимо этого, необходим удобный способ вычисления реализаций выбранной случайной величины.

Рассмотрим вначале моделирование дискретных случайных величин. Пусть некоторая дискретная случайная величина может принимать n значений с различными вероятностями.

$$\begin{pmatrix} x_1 x_2 \dots x_n \\ p_1 p_2 \dots p_n \end{pmatrix}; \;\; p_i = P\{\xi = x_i\} \qquad (4.1.2)$$

Разделим интервал числовой оси $0 \le y \le 1$ на n подынтервалов с длинами $\Delta_i = p_i$ $(i = 1,2, \dots, n)$ (см. рис. 17).

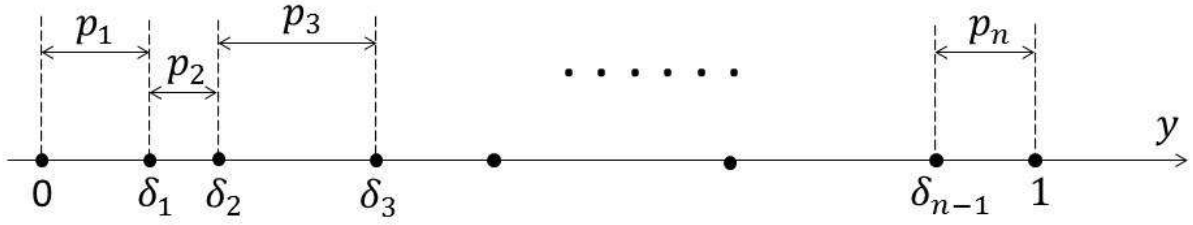

Рис. 17. Разбиение интервала числовой оси на подынтервалы, длины который равны вероятностям, с которыми дискретная случайная величина принимает возможные значения

Случайная величина ξ, определенная формулой (4.1.2) может быть смоделирована на основе стандартной случайной величины γ (непрерывная равномерно распределенная случайная величина на интервале $[0;1]$) следующим образом: $\xi = x_i$ когда $\gamma \in [\delta_{i-1}; \delta_i]$. Доказательство этого утверждения очень простое: $P\{\xi = x_i\} = P\{\gamma \in [\delta_{i-1}; \delta_i]\} = \Delta_i = p_i$.

Таким образом, из стандартной случайной величины можно смоделировать любую дискретную случайную величину, принимающую конечное количество значений.

Изложенный метод можно обобщить и на случай дискретных случайных величин, которые могут принимать бесконечное количество значений.

$$\begin{pmatrix} x_1 x_2 \dots x_n \dots \\ p_1 p_2 \dots p_n \dots \end{pmatrix};\ x_n = x_n(n), p_n = p_n(n) \quad (4.1.3)$$

Вначале упорядочим возможные значения случайной величины по убыванию вероятностей. Будем считать для простоты, что в (4.1.3) такое упорядочение уже проведено. Затем выберем n_0 значений случайной величины, таких, что$p_1 + \dots + p_{n_0}$ близко к 1. При $i \le n_0$ моделирование случайной величины проводится по тому же механизму, что и для дискретной случайной величины, принимающей конечное число значений. При $i > n_0$ значения случайной величины вычисляются непосредственно по заданной формуле.

Перейдем к моделированию непрерывных случайных величин. Пусть теперь ξ – непрерывная случайная величина, принимающая значения в диапазоне $x \in (a; b)$. Плотность вероятности $p(x) > 0$. Тогда функция распределения этой случайной величины

$$F(x) = \int_a^x p(u)du \tag{4.1.4}$$

Случайная величина ξ может быть смоделирована из стандартной случайной величины γ по формуле

$$F(\xi) = \gamma \tag{4.1.5}$$

Доказательство этого также несложное. Функция $F(x)$ возрастает на всем интервале $x \in (a; b)$, $F(a) = 0, F(b) = 1$. Следовательно, уравнение $F(\xi) = \gamma$ имеет один корень при любом γ (см. рис. 18).

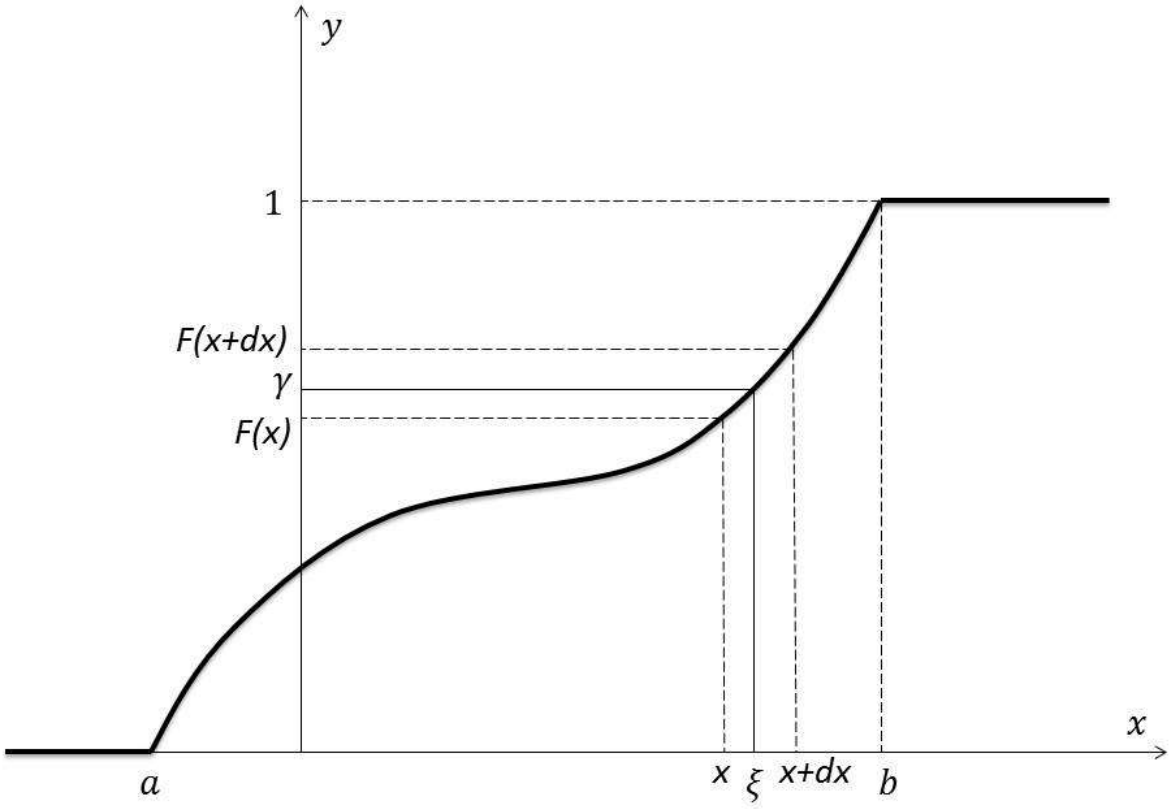

Рис. 18. График функции распределения непрерывной случайной величины

Заметим, что вероятность $P\{x < \xi < x + dx\} = P\{F(x) < \gamma < F(x + dx)\}$. Ввиду того, что стандартная случайная величина γ равномерно распределена на интервале $[0; 1]$, $P\{F(x) < \gamma < F(x + dx)\} = F(x + dx) - F(x) = p(x)dx$. Таким образом, мы доказали, что величина ξ, рассчитанная по формуле (4.1.5) имеет плотность вероятности $p(x)$. Для нахождения ξ в явном виде необходимо аналитически решить уравнение (4.1.5), т.е. найти обратную функцию к $F(\xi)$.

$$\xi = G(\gamma) \tag{4.1.6}$$

Если уравнение (4.1.5) не решается аналитически, то можно найти обратную функцию численно.

В качестве примера можно показать, что случайная величина с плотностью вероятности $p(x) = \alpha e^{-\alpha(x-x_0)}$ выражается через стандартную случайную величину по формуле $\xi = x_0 - (1/\alpha) ln(1-\gamma)$.

Комбинацией двух изложенных методов моделирования дискретных и непрерывных случайных величин можно смоделировать любую случайную величину с произвольной плотностью вероятности.

Далее перейдём к моделированию многомерной случайной величины $Q(\xi_1, \xi_2, \dots, \xi_n)$. Если координаты этой многомерной случайной величины $\xi_1, \xi_2, \dots, \xi_n$ независимы (т.е. функция распределения факторизуется $F_Q(x_1, \dots, x_n) = F_1(x_1) * \dots * F_n(x_n)$), то каждую координату ξ_i можно моделировать по-отдельности.

$$F_i(\xi_i) = \gamma_i; \;\; i = 1,2, \dots, n \tag{4.1.7}$$

Примером многомерной случайной величины с независимыми координатами является случайная точка $Q(x, y, z)$, равномерно распределенная в объёме прямоугольного параллелепипеда $\prod\{0 \le x \le a; \;\; 0 \le y \le b; \; 0 \le z \le c\}$. Плотность распределения многомерной величины $p_Q(x, y, z) = 1/abc$ в объёме параллелепипеда и 0 вне этого объёма. Плотность распределения каждой из координат может быть вычислена с помощью интегрирования по двум другим переменным. $p_Q(x) = \int_0^c dz \int_0^b dy \, p_Q(x, y, z) = 1/a$ при $x \in [0; a]$ и 0 при других x. $p_Q(y)$ и $p_Q(z)$ вычисляются аналогично. Таким образом, можно сделать вывод о независимости координат случайной точки $Q(x, y, z)$.

Если координаты многомерной случайной величины зависимы, ситуация усложняется. В этом случае $p_Q(x_1, \dots, x_n)$ выражается через условные плотности вероятностей.

$$p_Q(x_1, \dots, x_n) = p_1(x_1) p_2(x_2|x_1) p_3(x_3|x_1, x_2) \dots \; p_n(x_n|x_1, \dots, x_{n-1}) \tag{4.1.8}$$

В этой формуле

$$p_1(x_1) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_Q dx_2 \dots dx_n$$

$$p_2(x_2|x_1) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_Q dx_3 \dots dx_n [p_1(x_1)]^{-1}$$

$$p_3(x_3|x_1, x_2) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_Q dx_4 \dots dx_n [p_1(x_1) p_2(x_2|x_1)]^{-1}$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$p_{n-1}(x_{n-1}|x_1, \dots, x_{n-2}) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_Q dx_n [p_1(x_1) \dots p_{n-2}(x_{n-2}|x_1, \dots, x_{n-3})]^{-1}$$

$$p_n(x_n|x_1, \dots, x_{n-1}) = p_Q [p_1(x_1) \dots p_{n-1}(x_{n-1}|x_1, \dots, x_{n-2})]^{-1}$$

Условная функция распределения выражается через условную плотность вероятности.

$$F_i(x_i|x_1, \dots, x_{i-1}) = \int_{-\infty}^{x_i} p_i(x|x_1, \dots, x_{i-1}) dx \qquad (4.1.9)$$

Координаты многомерной случайной величины моделируются посредством решения системы уравнений

$$\begin{cases} F_1(\xi_1) = \gamma_1 \\ F_2(\xi_2|\xi_1) = \gamma_2 \\ \dots\dots\dots\dots \\ F_n(\xi_n|\xi_1, \dots, \xi_{n-1}) = \gamma_n \end{cases} \qquad (4.1.10)$$

Доказательство. Зафиксируем переменные $\xi_1 = x_1, \dots, \xi_{i-1} = x_{i-1}$. Тогда $F_i(\xi_i|x_1, \dots, x_{i-1}) = \gamma_i$. Условная вероятность $P\{x_i < \xi_i < x_i + dx_i|x_1, \dots, x_{i-1}\} = p_i(x_i|x_1, \dots, x_{i-1})dx_i$. С учетом этого$P\{x_1 < \xi_1 < x_1 + dx_1, \dots, x_n < \xi_n < x_n + dx_n\} = P\{x_1 < \xi_1 < x_1 + dx_1\} P\{x_2 < \xi_2 < x_2 + dx_2|\xi_1 = x_1\} \dots P\{x_n < \xi_n < x_n + dx_n|\xi_1 = x_1, \dots, \xi_{n-1} = x_{n-1}\} = p_1(x_1)dx_1 p_2(x_2|x_1)dx_2 \dots p_n(x_n|x_1, \dots, x_{n-1})dx_n = p_Q(x_1, \dots, x_n)dx_1 \dots dx_n$.

Принцип приведенного доказательства похож на доказательство метода моделирования одномерной случайной величины. Однако в данном случае используются условные плотности вероятности и функции распределения.

Во многих практических задачах удается существенно упростить моделирование координат многомерной случайной величины, сделав замену переменных. При этом плотность распределения преобразуется согласно правилу

$$p_{Q'}(y_1, \dots, y_n) = p_Q(x_1, \dots, x_n) \left|\frac{\partial(x_1,\dots,x_n)}{\partial(y_1,\dots,y_n)}\right| \qquad (4.1.11)$$

Где $x_1, \dots, x_n$ – старые координаты, $y_1, \dots, y_n$ – новые координаты.

Рассмотрим пример: случайная точка Q равномерно распределена в шаре $x^2 + y^2 + z^2 < R^2$. $p_Q(x, y, z) = [(4/3)\pi R^3]^{-1}$. Очевидно, что декартовы координаты не являются независимыми.К примеру, если мы зафиксируем x, то область изменения двух других координат будет представлять собой круг, радиус которого зависит от x. Перейдём к сферическим координатам $x = r\sin\theta\cos\varphi\,; y = r\sin\theta\sin\varphi\,; z = r\cos\theta\,; 0 \le r < R; 0 \le \theta < \pi; 0 \le \varphi < 2\pi$. Согласно формуле (4.1.11) $p_Q(r, \theta, \varphi) = [(4/3)\pi R^3]^{-1} r^2 \sin\theta$, откуда видно, что $p_Q(r, \theta, \varphi)$ факторизуется, т.е. $p_Q(r, \theta, \varphi) = p_Q(r)p_Q(\theta)p_Q(\varphi)$, где $p_Q(r) = 3r^2/R^3; p_Q(\theta) = 0.5\sin\theta;$ $p_Q(\varphi) = 1/2\pi$. Таким образом, r, θ, φ моделируются независимо:

$$\begin{cases} \int\limits_0^{r_Q} \dfrac{3r^2}{R^3}\,dr = \gamma_1 \\ \int\limits_0^{\theta_Q} \dfrac{\sin\theta}{2}\,d\theta = \gamma_2 \\ \int\limits_0^{\varphi_Q} \dfrac{d\varphi}{2\pi} = \gamma_3 \end{cases}$$

Проведя интегрирование, выражаем в явном виде сферические координаты случайной точки $r_Q = R\sqrt[3]{\gamma_1};\ \theta_Q = arcCos(1 - 2\gamma_2);\ \varphi_Q = 2\pi\gamma_3$.

§4.2. Простейший метод Монте-Карло

Одним из важнейших применений метода Монте-Карло является вычисление интегралов. Рассмотрим двумерный интеграл по некоторой области G

$$I = \iint f(P)p(P)dP \qquad (4.2.1)$$

В этой формуле $P = (x, y)$ – точка на плоскости, $dP = dxdy$, $p(P)$ – плотность распределения некоторой случайной величины.

Отметим, что любой интеграл $\iint f(P)dP$ можно представить в виде (4.2.1). Действительно, пусть P равномерно распределена по области G, т.е. $p_1(P) = 1/S_G$, где S_G – площадь области G. Введем функцию $f_1(P) = S_G f(P)$. Нетрудно видеть, что $\iint f(P)dP = \iint f_1(P)p_1(P)dP$.

Пусть Q – случайная точка с плотностью распределения $p(P)$. $Z = f(Q)$. Тогда математическое ожидание случайной величины Z равно искомому интегралу (4.2.1) $MZ = I$. Таким образом, для вычисления I требуется вначале разыграть N независимых реализаций случайной точки Q: $Q_1, Q_2, \dots, Q_N$. Затем рассчитываются величины $Z_1 = f(Q_1), Z_2 = f(Q_2), \dots, Z_N = f(Q_N)$. Оценкой интеграла (4.2.1) является среднее арифметическое $Z_i (i = 1,2, \dots, N)$.

$$\Theta_N = \frac{1}{N}\sum_{i=1}^{N} Z_i \qquad (4.2.2)$$

Как линейная комбинация случайных величин Θ сама является случайной величиной. $M\Theta_N = (1/N)\sum_{i=1}^{N} M(Z_i) = (1/N)\sum_{i=1}^{N} I = I$. $D\Theta_N = (1/N^2)\sum_{i=1}^{N} D(Z_i) = (1/N^2)\sum_{i=1}^{N} D(Z) = D(Z)/N$.

Рассмотрим пример:

$$I = \int_0^{+\infty} f(x)e^{-kx}dx\,,\ k > 0 \qquad (4.2.3)$$

Вводя $p(x) = ke^{-kx}$, $f_1(x) = (1/k)f(x)$, интеграл приводится к виду (4.2.1). Обозначим ξ случайную величину с плотностью распределения $p(x)$. $\Theta_N = N^{-1}\sum_{i=1}^{N} f_1(\xi_i) = (kN)^{-1}\sum_{i=1}^{N} f(\xi_i) = (kN)^{-1}\sum_{i=1}^{N} f[-(1/k)ln(1-\gamma_i)]$. Последнее равенство было написано с учётом выражения экспоненциально распределенной случайной величины через стандартную.

Рассмотрим вопрос о статистической погрешности метода. Последовательность одинаково распределенных случайных величин подчиняется центральной предельной теореме.

$$\lim_{N\to\infty} P\left\{x_1 < \frac{1}{\sqrt{ND(Z)}}\sum_{i=1}^{N}(Z_i - I) < x_2\right\} = \frac{1}{\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-t^2/2}dt \qquad (4.2.4)$$

Это означает, что при достаточно большихN величина Θ_N может считаться нормально распределенной с математическим ожиданием I и дисперсией $D\Theta_N = D(Z)/N$.

Как правило, $D(Z)$ не известно заранее. Поэтому используют оценку дисперсии

$$D(Z) = M(Z^2) - \big(M(Z)\big)^2 \approx \frac{1}{N}\sum_{i=1}^{N} Z_i^2 - \left[\frac{1}{N}\sum_{i=1}^{N} Z_i\right]^2 \qquad (4.2.5)$$

Таким образом, для оценки дисперсии необходимо наряду с вычислением суммы случайных реализаций Z_i вычислять также сумму их квадратов.

§4.3. Геометрический метод Монте-Карло

Задача по-прежнему состоит в вычислении интеграла вида (4.2.1).

$f(P)$– функция, заданная на двумерной области G. $0 \le f(P) \le c$. Обозначим $\tilde{G}$ трёхмерную область пространства: $\tilde{G} = G \times [0; c]$. $\tilde{Q}$ – случайная точка в области $\tilde{G}$ с плотностью распределения $\tilde{p}(x, y, z) = (1/c)p(x, y)$. Проекцию $\tilde{Q}$ на плоскость XOY обозначим $Q(\xi, \eta)$ – это случайная точка с плотностью распределения $p(x, y)$, потому что третья координата ζ случайной точки $\tilde{Q}$ распределена равномерно. $p_\zeta(z) = 1/c$.

Разыгрывается набор независимых случайных реализаций случайной точки $\tilde{Q}$:$\tilde{Q}_1, \tilde{Q}_2, \dots, \tilde{Q}_N$. Количество точек, оказавшихся ниже поверхности $z = f(x, y)$ обозначимν ($\nu < N$), см. рис. 19.

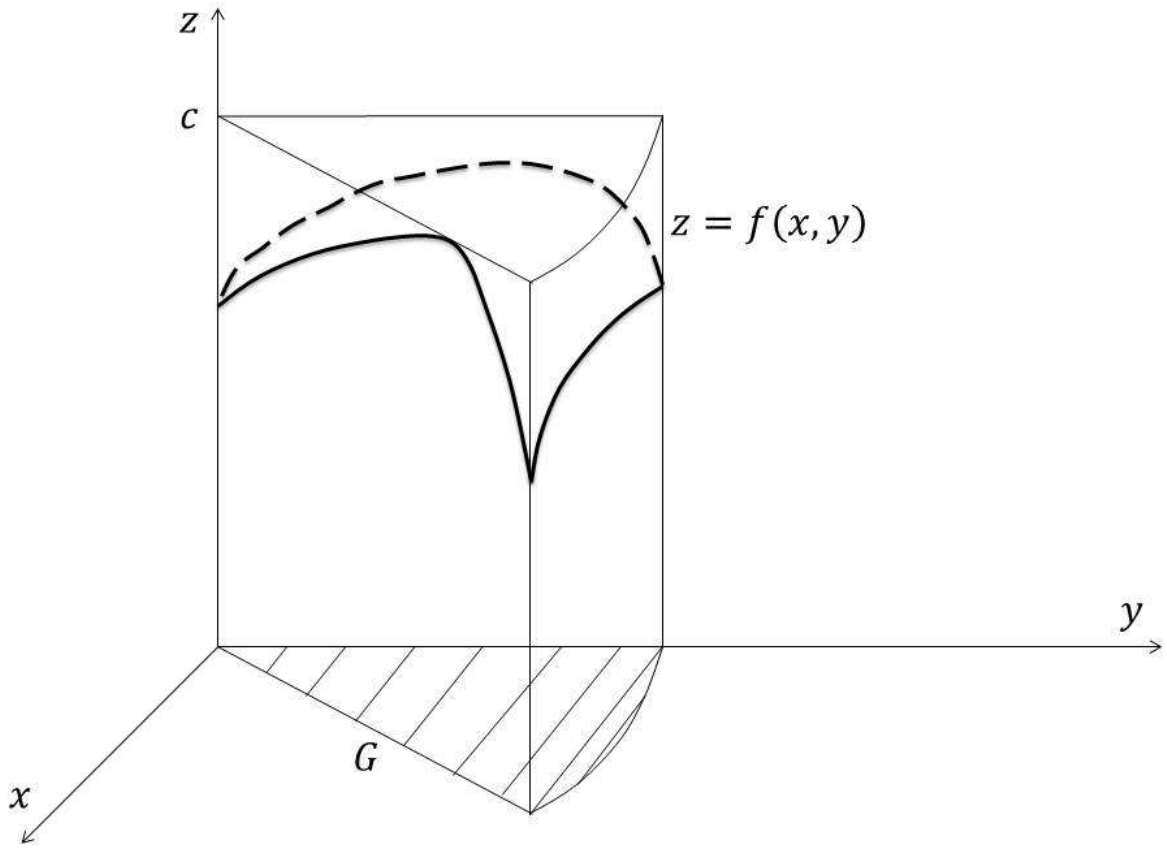

Рис. 19. Иллюстрация к геометрическому методу Монте-Карло

Вероятность того, что точка $\tilde{Q}$ окажется ниже поверхности $f(x,y)$ равна $p = P\{\zeta < f(\xi,\eta)\} = \iint dxdy \int_0^{f(x,y)} \tilde{p}(x,y,z)dz = \iint dxdy \int_0^{f(x,y)} (1/c)p(x,y)dz = (1/c)I$. Двойной интеграл подразумевается по области G. Таким образом, интеграл (4.2.1) $I = cp$. При достаточно большихN вероятность $p \approx \nu/N$. Таким образом, оценкой интеграла I служит величина

$$\widetilde{\Theta}_N = c\frac{\nu}{N} \qquad (4.3.1)$$

Геометрический метод является обобщением метода вычисления объёма, в котором $p(x,y) = const = 1/S_G$ (S_G – площадь области G). В этом случае $\nu/N \approx V/V_{\tilde{G}}$, где V – объём пространства, ограниченный поверхностью $f(x,y)$.

Сравним теперь точность простейшего и геометрического методов Монте-Карло. Дисперсия усредняемой величины в простейшем методе Монте-Карло равна

$$DZ = \iint f^2(P)p(P)dP - I^2 \quad (4.3.2)$$

В геометрическом методе усредняется величина

$$\tilde{Z} = \begin{cases} c, \zeta < f(\xi, \eta) \\ 0, \zeta \geq f(\xi, \eta) \end{cases} \quad (4.3.3)$$

Оценка (4.3.1) связана с $\tilde{Z}$ соотношением $\widetilde{\Theta}_N = (1/N)\sum_{i=1}^{N} \tilde{Z}_i$. Математическое ожидание $M(\tilde{Z}^2) = c^2 P\{\zeta < f(\xi, \eta)\} = c^2(I/c) = cI$. Следовательно, дисперсия $D(\tilde{Z}) = M(\tilde{Z}^2) - I^2 = cI - I^2$.

Проведём оценку дисперсии усредняемой величины в простейшем методе Монте-Карло (4.3.2): так как $0 \leq f(x, y) \leq c$ имеем $DZ = \iint f^2(P)p(P)dP - I^2 \leq c \iint f(P)p(P)dP - I^2 = cI - I^2 = D(\tilde{Z})$. Таким образом, простейший метод Монте-Карло точнее геометрического.

Однако простейший метод Монте-Карло может оказаться более трудоёмким, чем геометрический в смысле затрат машинного времени. Поэтому целесообразно оценить трудоёмкость методов. Из предельной теоремы при $x_2 = -x_1 = x$ следует выражение

$$\lim_{N\to\infty} P\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(\xi_i - a)\right| < x\sqrt{\frac{D\xi}{N}}\right\} = \Phi(x) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt \quad (4.3.4)$$

Здесь ξ – случайная величина с математическим ожиданием a. При достаточно большихN вероятность $P\{|\tilde{\xi}_N - a| < x\sqrt{D\xi/N}\} \approx \Phi(x)$, где $\tilde{\xi}_N = (\xi_1 + \xi_2 + \cdots + \xi_N)/N$. Зададим коэффициент доверия β (его иначе называют доверительной вероятностью). По нему находится доверительный интервалx_β из условия $\Phi(x_\beta) = \beta$. $P\{|\tilde{\xi}_N - a| < x_\beta\sqrt{D\xi/N}\} \approx \beta$. Значению $\beta = 0.5$ соответствует $x_\beta = 0.6745$. Величину $r_N = 0.6745\sqrt{D\xi/N}$ называют вероятностной ошибкой. Т.е. ошибки, большие r_N и меньшие r_N, равновероятны.

Зафиксируем вероятностную ошибку $r_N = \varepsilon$. Тогда количество реализаций случайной величины, которое необходимо вычислить для достижения заданной точности, равно $N' = DZ(0.6745/\varepsilon)^2$ для простейшего метода и $N'' = D(\tilde{Z})(0.6745/\varepsilon)^2$ для геометрического. Обозначим t' и t'' - время, требуемое для расчёта одной реализации случайной величины в простейшем и геометрическом методах соответственно. Тогда полное время равно $t'N' = t'D'(0.6745/\varepsilon)^2$ для простейшего метода и $t''N'' = t''D''(0.6745/\varepsilon)^2$. Видно, что полное время вычислений пропорционально произведению времени расчёта одной

реализации случайной величины на дисперсию. Это произведение называют трудоёмкостью алгоритма.

Глава 5. Методы оптимизации

§5.1. Методы минимизации функции одной переменной

Суть математической оптимизации состоит в нахождении минимального (или максимального) значения некоторой функции $f(x_1, x_2, \dots, x_n)$, заданной на множестве X, причем на аргументы функции могут быть наложены определенные условия. Для определенности в настоящей главе будем понимать под оптимизацией нахождение минимального значения функции, т.к. задача о нахождении максимума сводится к задаче о нахождении минимума элементарной заменой функции $f \to -f$.

Рассмотрим вначале минимизацию функции одной вещественной переменной $f(x)$ на отрезке $a \le x \le b$. Для численного решения такой задачи разработан ряд методов, наиболее распространенные из которыхмы перечислим в данном параграфе.

метод деления отрезка пополам

Данный метод позволяет найти локальный минимум функции $f(x)$, см. рис. 20.

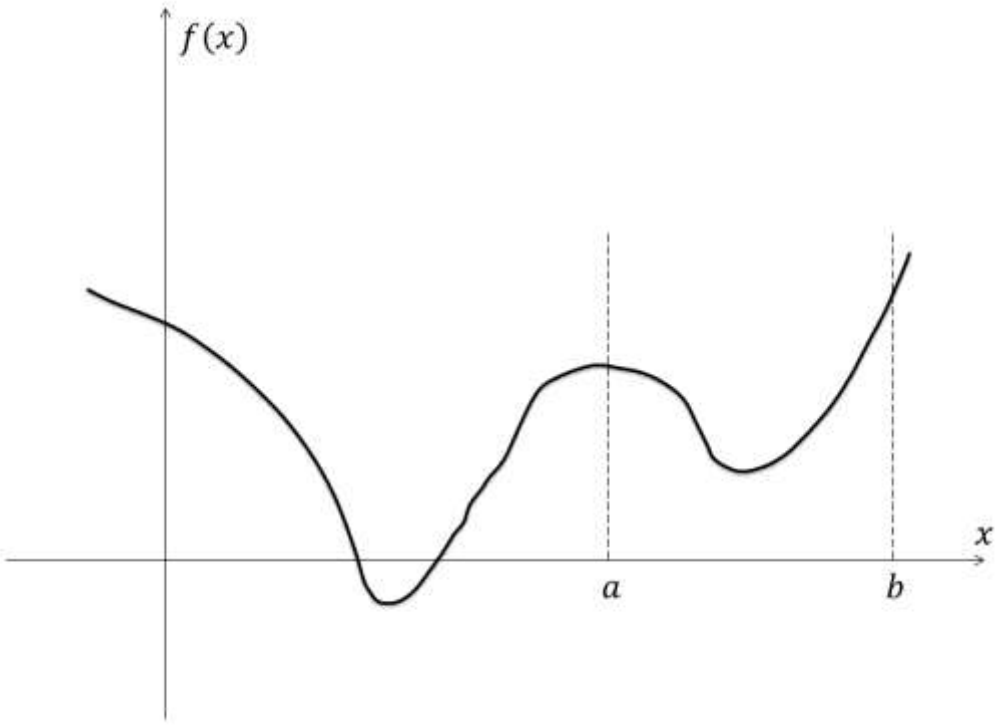

Рис. 20. Иллюстрация к методам нахождения локального минимума функции одной вещественной переменной

Вначале выбираются две точки из отрезка $[a; b] x_1 = (a + b - \delta)/2$, $x_2 = (a + b + \delta)/2$. Здесь δ – параметр метода, $0 < \delta < b - a$. Затем значения функции в этих двух точках $f(x_1)$ и $f(x_2)$ сравниваются друг с другом.

Если $f(x_1) \leq f(x_2)$, то для дальнейшего шага полагаем $a_1 = a;\ b_1 = b$.

Если $f(x_1) > f(x_2)$, то полагаем $a_1 = x_1;\ b_1 = b$.

Так как минимум функции на отрезке $[a; b]$ единственный, отрезок $[a_1; b_1]$ в любом случае его содержит. Длина отрезка $[a_1; b_1]$ равна

$$b_1 - a_1 = \frac{b-a-\delta}{2} + \delta \qquad (5.1.1)$$

Далее с вновь построенным отрезком $[a_1; b_1]$ проделываются те же самые операции, которые были проделаны с отрезком $[a; b]$, в результате чего получаем отрезок $[a_2; b_2]$ и т.д. На k-ом шаге длина отрезка будет равна

$$b_k - a_k = \frac{b-a-\delta}{2^k} + \delta > \delta \qquad (5.1.2)$$

Критерий остановки вычислений: $b_k - a_k < \varepsilon$, где ε – наперед заданная точность. Заметим, что в этом методе обязательно требуется выполнение условия $\varepsilon > \delta$, в противном случае критерий остановки никогда не будет достигнут. Из критерия остановки нетрудно найти номер итерации, когда этот критерий впервые будет достигнут.

$$k_{min} = \left[\left[\log_2\left(\frac{b-a-\delta}{\varepsilon-\delta}\right)\right]\right] + 1 \qquad (5.1.3)$$

В выражении (5.1.3) двойными квадратными скобками обозначена целая часть числа.

Свое название этот метод получил благодаря тому, что при малых δ точки x_1 и x_2 делят отрезок $[a; b]$ примерно пополам, и на каждом последующем шаге аналогично.

метод золотого сечения

Этот метод позволяет решить задачу с требуемой точностью ε, затратив для этого меньшее количество вычислений значения функции, чем предыдущий метод.

Золотым сечением отрезка $[a; b]$ называется разбиение этого отрезка на две части так, чтобы отношение длины всего отрезка к длине большей части было равно отношению длины большей части к меньшей части.

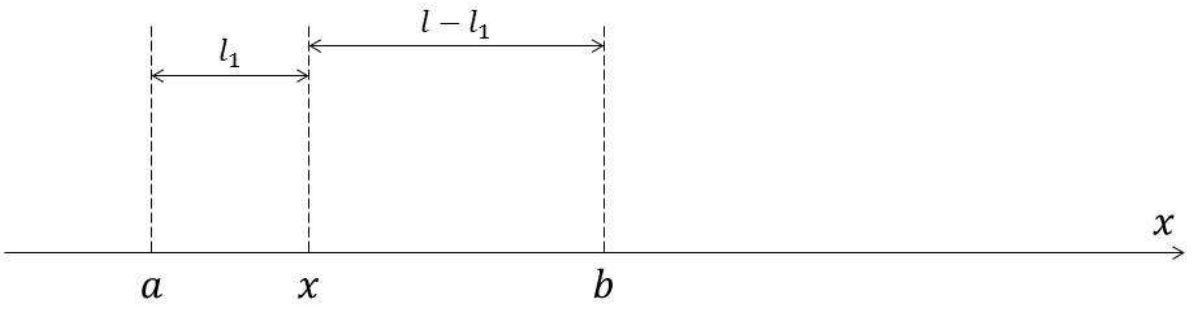

Рис. 21. Иллюстрация к методу золотого сечения

Очевидно, что золотое сечение отрезка можно провести двумя точками, симметричными относительно середины отрезка. Найдем положения этих точек. Пусть точка золотого сечения x лежит левее середины отрезка (см. рис. 21). Составим пропорцию

$$\frac{l}{l-l_1} = \frac{l-l_1}{l_1} \qquad (5.1.4)$$

Выражаем отсюда $l_1 = 0.5\left(3-\sqrt{5}\right)l$. Если предположить точку x правее середины отрезка, то результат был бы $l_2 = 0.5\left(\sqrt{5}-1\right)l$.

Теперь приступим к описанию самого метода. Вначале производится золотое сечение отрезка $[a;b]$ точками x_1 и x_2. $x_1 = a + (b-a) * 0.381966011 \ldots, x_2 = a + (b-a) * 0.618033989 \ldots$

У золотого сечения есть одно важное свойство: точка x_1 является одной из точек золотого сечения для отрезка $[a;x_2]$. В этом нетрудно убедиться, раскрыв пропорцию $(x_2-a)/(x_1-a) = (x_1-a)/(x_2-x_1)$. Аналогично, точка x_2 производит золотое сечение отрезка $[x_1,b]$.

Предполагается, что функция $f(x)$ имеет один минимум на отрезке $[a;b]$. На первом шаге полагаем $a_1 = a, b_1 = b$. Производим золотое сечение отрезка $[a_1;b_1]$ точками x_1 и x_2.

Если $f(x_1) \le f(x_2)$, то полагаем $a_2 = a_1;\ b_2 = x_2;\ \overline{x_2} = x_1$.

Если $f(x_1) > f(x_2)$, то $a_2 = x_1;\ b_2 = b_1;\ \overline{x_2} = x_2$.

Так как на отрезке $[a_1;b_1]$ только один минимум, вновь построенный отрезок $[a_2;b_2]$ в любом случае его содержит. Длина отрезка $[a_2;b_2]$ равна

$$b_2 - a_2 = \frac{\sqrt{5}-1}{2}(b-a) \qquad (5.1.5)$$

Отметим, что отрезок $[a_2;b_2]$ содержит точку $\overline{x_2}$, значение функции в которой было вычислено на предыдущем шаге $f(\overline{x_2}) = min\{f(x_1), f(x_2)\}$. Точка $\overline{x_2}$ производит золотое сечение отрезка $[a_2;b_2]$. Это свойство позволяет не производить золотое сечение на каждом шаге заново, что уменьшает объем вычислений.

Далее с отрезком $[a_2;b_2]$ проделываем те же операции, что и с отрезком $[a_1;b_1]$. В результате получаем отрезок $[a_3;b_3]$ и т.д. На n-ом шаге длина отрезка будет равна

$$b_n - a_n = \left(\frac{\sqrt{5}-1}{2}\right)^{n-1}(b-a) \qquad (5.1.6)$$

Критерием остановки является выполнение неравенства $b_n - a_n < \varepsilon$, где ε – наперед заданная точность. Как было сказано выше, основным преимуществом этого метода по сравнению с методом деления отрезка пополам является то, что он позволяет достичь требуемой точности за

меньшее количество действий. Однако стоит отметить также и недостаток метода золотого сечения – число $\sqrt{5}$ неизбежно будет задано приближенно, а погрешность накапливается с ростом числа итераций. Из этой ситуации можно найти выход, если на каждом k-ом шаге производить золотое сечение отрезка $[a_k; b_k]$, содержащего $\overline{x_k}$ из предыдущего шага, не по формуле $x_{k+1} = a_k + b_k - \overline{x_k}$, а непосредственно.

метод ломаных

Этот метод позволяет находить глобальный минимум функции $f(x)$ на отрезке $[a; b]$.Для использования метода ломаных, функция $f(x)$ должна удовлетворять условию Липшица на отрезке $[a; b]$, т.е. $\exists L > 0$: $|f(x) - f(y)| \leq L|x - y|$ $\forall x, y \in [a; b]$.

В рассматриваемом методе ломаные строятся относительно какой-либо точки функции $f(x)$. Пусть y – некоторая зафиксированная точка на $[a; b]$ (см. рис. 22). Строим функцию $g(x, y) = f(y) - L|x - y|$. Аргументом функции g считается x, а y будем считать параметром. Тогда g представляет собой ломаную, состоящую из двух лучей с началом в точке $\big(y, f(y)\big)$. Для значений $x < y$ угловой коэффициент L, для $x > y$ угловой коэффициент – L.

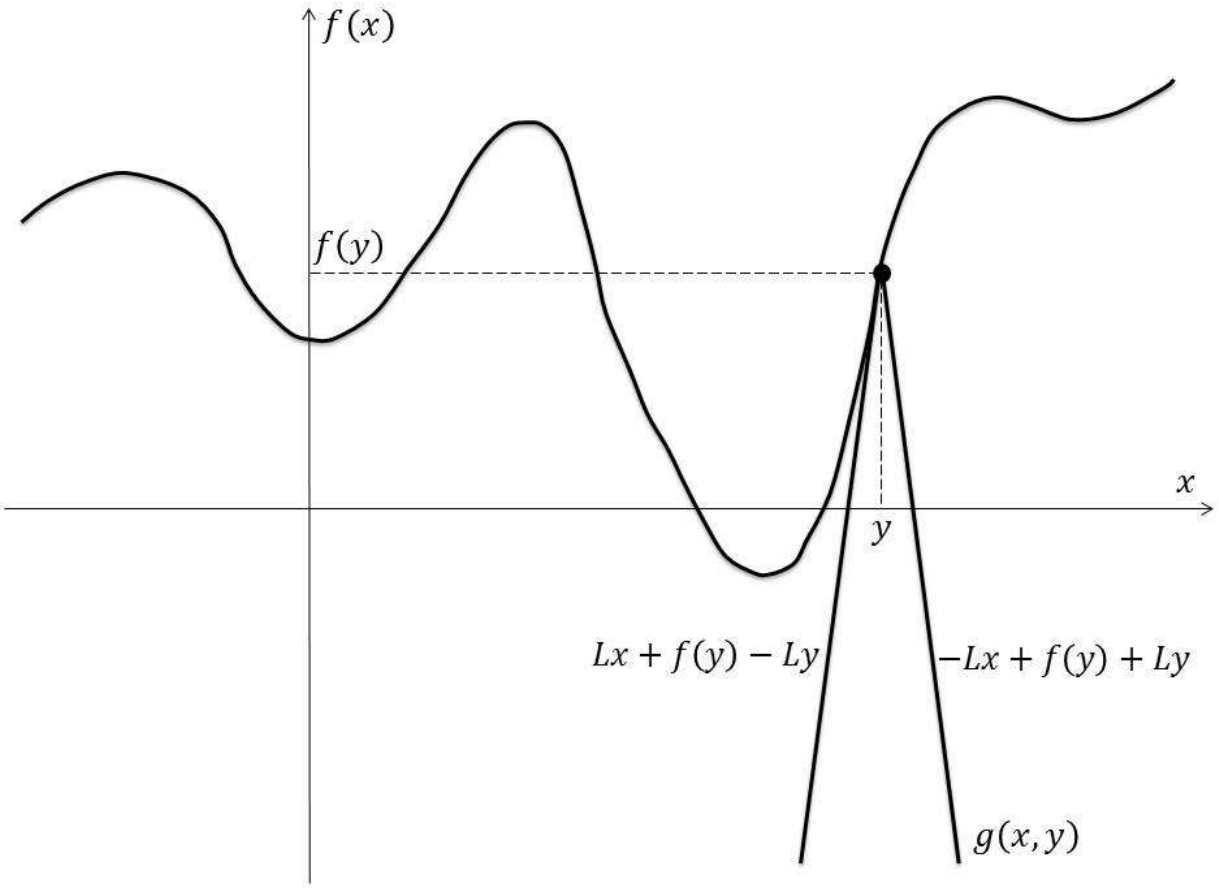

Рис. 22. Иллюстрация к построению ломаных

Приступим теперь к описанию самого метода отыскания глобального минимума (см. рис. 23). Вначале выбирается некоторая начальная точка $x_0 \in [a;b]$. Затем по этой точке строится ломаная $g(x,x_0) = f(x_0) - L|x - x_0| = p_0(x)$. После этого определяется точка x_1 из условия $p_0(x_1) = \min_{x\in[a;b]} p_0(x)$. Очевидно, что при таком построении либо $x_1 = a$ либо $x_1 = b$.

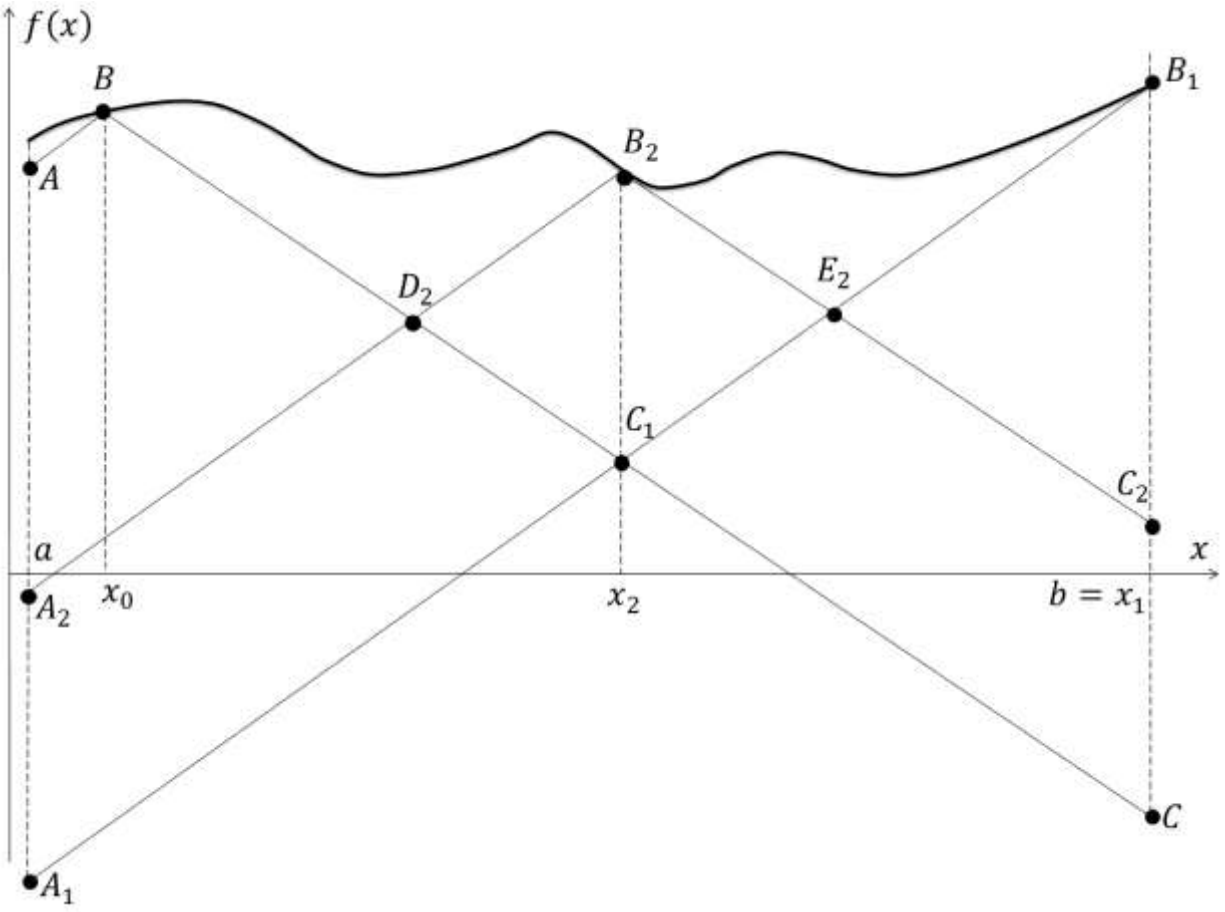

Рис. 23. Иллюстрация к методу ломаных, $p_0(x)$ *– ломаная* ABC, $x_1 = b$, $g(x, x_1)$ *– прямая* A_1B_1, $p_1(x)$ *– ломаная* ABC_1B_1, $g(x, x_2)$ *– ломаная* $A_2B_2C_2$, $p_2(x)$ *– ломаная* $ABD_2B_2E_2B_1$

Составим функцию $p_1(x) = max\{g(x, x_1); p_0(x)\}$. Далее определяется точка x_2 из условия $p_1(x_2) = \min_{x\in[a;b]} p_1(x)$, после чего строится функция $p_2(x) = max\{g(x, x_2); p_1(x)\}$. Процесс продолжается по этой схеме далее аналогично, т.е. находится точка x_3, строится функция $p_3(x)$ и т.д.

Так как на любом шаге $p_n(x) = max\{g(x, x_n); p_{n-1}(x)\}$, нетрудно понять, что $p_{n-1}(x) \le p_n(x)$. Ломаные$p_n(x)$ аппроксимируют функцию $f(x)$ снизу. Таким образом, задача минимизации функции сводится к задаче минимизации ломаной, что осуществляется простым перебором ее вершин.

методы покрытий

Суть методов покрытий состоит в том, что строится система отрезков, покрывающих отрезок $[a; b]$, а затем вычисляются значения функции вподходящим образом выбранных точках этих отрезков.

Простейшим примером является метод равномерного перебора. В рамках этого метода производится равномерное разбиение отрезка $[a; b]$ – $x_1 = a + 0.5h, x_2 = x_1 + h, \dots, x_{i+1} = x_i + h = x_1 + ih, \dots, x_{n-1} = x_1 + (n-2)h, x_n = min\{x_1 + (n-1)h; b\}$.

Функция $f(x)$ должна удовлетворять условию Липшица, а шаг $h = 2\varepsilon/L$, где ε – заданная точность, т.е. $\min_{1 \le i \le n} f(x_i) \le f(x) + \varepsilon \, \forall f(x) \in Q(L)$. Q – пространство рассматриваемых функций, L – константа из условия Липшица.

Более удобной модификацией описанного метода является метод последовательного перебора. Здесь $x_1 = a + 0.5h, x_{i+1} = x_i + h + (f(x_i) - F_i)/L, \quad i = 1,2, \dots, n-2; \; x_n = min\{x_{n-1} + h + L^{-1}(f(x_{n-1}) - F_{n-1}); b\}$. В этих выражениях $h = 2\varepsilon/L$, $F_i = \min_{1 \le j \le i} f(x_j)$, а n определяется условием $x_{n-1} < b - 0.5h \le x_{n-1} + h + (f(x_{n-1}) - F_{n-1})/L$.

В методе последовательного перебора на каждом последующем шаге используются значения функции, вычисленные на предыдущих шагах. Это позволяет достичь требуемой точности за меньшее количество действий по сравнению с методом равномерного перебора.

Существует также большое количество модификаций изложенных здесь методов, однако мы ограничимся рассмотренными методами минимизации функции одной переменной.

§5.2. Условный экстремум многомерных функций. Правило множителей Лагранжа

В этом параграфе мы будем рассматривать функцию n аргументов. Для краткости, совокупность этих аргументов будем обозначать вектором-строкой $x = (x^1, \dots, x^n)$. Причем не все аргументы будут независимые, т.е. $x \in X \ne E^n$. Это означает, что переменные $x^1, \dots, x^n$ удовлетворяют некоторым условиям.

Рассмотрим вначале случай, когда на аргументы функции наложены ограничения типа равенства – $x \in X = \{x \in E^n : g_1(x) = 0, \dots, g_s(x) = 0\}$. Всего ограничений типа равенства s штук. Наиболее прямой способ поиска экстремума в этом случае – это выразить некоторые переменные через остальные (из условий)

$$\begin{gathered} x^1 = \varphi_1(x^{p+1}, \dots, x^n) \\ \dots \\ x^p = \varphi_p(x^{p+1}, \dots, x^n) \end{gathered} \quad (5.2.1)$$

В выражениях (5.2.1) $x^1, \dots, x^p$ – зависимые переменные (их p штук), $x^{p+1}, \dots, x^n$ – независимые переменные (их $n-p$ штук). С учетом (5.2.1) задача на условный экстремум функции $f(x)$ превращается в задачу на безусловный экстремум функции $g(x^{p+1}, \dots, x^n) = f\big(\varphi_1(x^{p+1}, \dots, x^n), \dots, \varphi_p(x^{p+1}, \dots, x^n), x^{p+1}, \dots, x^n\big)$. Главный недостаток такого подхода состоит в том, что не всегда удается в явном виде выразить одни переменные через другие.

Более общим подходом является использование правила множителей Лагранжа. В рамках этого подхода составляется функция Лагранжа

$$L\big(x, \bar{\lambda}\big) = \lambda_0 f(x) + \sum_{j=1}^{s} \lambda_j g_j(x) \quad (5.2.2)$$

Функция Лагранжа зависит от переменных $x_1, \dots, x_n$, а также от множителей Лагранжа $\lambda_0, \lambda_1, \dots, \lambda_s$.

Правило множителей Лагранжа говорит о том, что если x_* - точка локального экстремума функции $f(x)$ на множестве X, то существует такой набор множителей Лагранжа $\lambda_0^*, \dots, \lambda_s^*$, что $\bar{\lambda}^* = (\lambda_0^*, \dots, \lambda_s^*) \neq 0, \lambda_0^* \geq 0$ и

$$\left.\frac{\partial L\big(x, \overline{\lambda^*}\big)}{\partial x^i}\right|_{x=x_*} = \lambda_0^* \frac{\partial f(x_*)}{\partial x^i} + \sum_{j=1}^{s} \lambda_j^* \frac{\partial g_i(x_*)}{\partial x^i} = 0; i = 1, \dots, n \quad (5.2.3)$$

Система уравнений (5.2.3) является необходимым условием экстремума.

Рассмотрим *пример*. Найдем точки экстремума функции $f(x, y) = x$ на множестве $X: g(x, y) = x^3 - y^2 = 0$. Раскроем необходимое условие экстремума для этого случая

$$\begin{cases} \lambda_0 \frac{\partial f(x,y)}{\partial x} + \lambda_1 \frac{\partial g(x,y)}{\partial x} = 0 \\ \lambda_0 \frac{\partial f(x,y)}{\partial y} + \lambda_1 \frac{\partial g(x,y)}{\partial y} = 0 \\ g(x, y) = 0 \\ (\lambda_0, \lambda) \neq 0 \end{cases} \quad (5.2.4)$$

В выражениях (5.2.4) звездочка опущена для краткости. Подстановка в (5.2.4) явного вида функции приводит к системе уравнений

$$\begin{cases} \lambda_0 + \lambda_1 3x^2 = 0 \\ -2\lambda_1 y = 0 \\ x^3 - y^2 = 0 \end{cases} \quad (5.2.5)$$

При этом необходимо учитывать, что $(\lambda_0, \lambda) \neq 0, \lambda_0 \geq 0$. Из второго уравнения системы (5.2.5) следует, что, либо $\lambda_1 = 0$, либо $y = 0$.

Первый вариант $\lambda_1 = 0$ приводит к тому, что из первого уравнения $\lambda_0 = 0$, а это противоречит требованию $(\lambda_0, \lambda) \neq 0$.

Второй вариант $y = 0$ приводит к тому, что из третьего уравнения $x = 0$, затем из первого уравнения $\lambda_0 = 0$. Таким образом, точка $(0; 0)$ является точкой, подозрительной на экстремум, и ей соответствует вектор множителей Лагранжа $\bar{\lambda} = (0, \lambda_1)$.

Теперь перейдем к рассмотрению более общего случая, когда аргументы функции связаны не только ограничениями типа равенства, но также и ограничениями типа неравенства, т.е. $X = \{x \in E^n: g_1(x) \leq 0, \dots, g_m(x) \leq 0, g_{m+1}(x) = 0, \dots, g_s(x) = 0\}$. Количество ограничений-равенств m, ограничений-неравенств $s - m$. Функция Лагранжа вводится так же, как и в предыдущем случае, согласно (5.2.2). Но ограничения на $\bar{\lambda}$ другие: $\lambda_0 \geq 0 \lambda_1 \geq 0, \dots, \lambda_m \geq 0$. Обобщением правила множителей Лагранжа на случай ограничений типа неравенства является теорема Каруша-Джона. Эта теорема говорит о том, что если x_* - точка локального минимума, то существует такой набор множителей Лагранжа $\bar{\lambda}^* = (\lambda_0^*, \dots, \lambda_s^*)$: $\bar{\lambda}^* \neq 0, \lambda_0^* \geq 0, \lambda_1^* \geq 0, \dots, \lambda_m^* \geq 0, L(x_*, \bar{\lambda}^*) = 0, \lambda_i^* g_i(x_*) = 0, i = 1, \dots, m$. Таким образом, точки локального экстремума находятся из следующих условий

$$\begin{cases} \lambda_0 f'(x) + \sum_{j=1}^{s} \lambda_j g'_j(x) = 0, \ i = 1, \dots, m \\ \lambda_i g_i(x) = 0, \ i = 1, \dots, m \\ g_i(x) \leq 0, \ i = 1, \dots, m \\ g_i(x) = 0, \ i = m+1, \dots, s \end{cases} \quad (5.2.6)$$

Вторая строка в (5.2.6) – это так называемые условия дополняющей нежесткости. Отметим, что точки локального максимума определяются так же, лишь за тем исключением, что $\lambda_0 \leq 0$.

Рассмотрим *пример*. Найдем точки локального экстремума функции $f(u) = x$ на множестве $X = \{u = (x, y) \in E^2: g_1(u) = -x \leq 0, g_2(u) = x^2 - y \leq 0, g_3(u) = y - 2x^2 \leq 0\}$. Составляем функцию Лагранжа $L(u, \bar{\lambda}) = \lambda_0 x + \lambda_1(-x) + \lambda_2(x^2 - y) + \lambda_3(y - 2x^2); \ \lambda_1, \lambda_2, \lambda_3 \geq 0, \ \bar{\lambda} \neq 0$. Используем условия (5.2.6), получаем

$$\begin{cases} \lambda_0 - \lambda_1 + 2x(\lambda_2 - 2\lambda_3) = 0 \\ -\lambda_2 + \lambda_3 = 0 \\ \lambda_1(-x) = 0 \\ \lambda_2(x^2 - y) = 0 \\ \lambda_3(y - 2x^2) = 0 \\ -x \leq 0 \\ x^2 - y \leq 0 \\ y - 2x^2 \leq 0 \end{cases} \quad (5.2.7)$$

В системе (5.2.7) первое уравнения – это результат дифференцирования функции Лагранжа по x, второе – поy, с третьего по пятое – это условия дополняющей нежесткости. Последние три неравенства из условия задачи.

Приступим к решению. Из третьего уравнения системы (5.2.7) следует, что, либо $\lambda_1 = 0$, либо $x = 0$.

Рассмотрим вначале вариант, когда $x = 0$. Тогда из седьмого неравенства получается$-y \leq 0$, а из восьмого $y \leq 0$. Таким образом, $y = 0$. Далее из первого равенства следует $\lambda_0 - \lambda_1 = 0$, т.е. $\lambda_1 = \lambda_0$, из второго равенства $\lambda_2 = \lambda_3$. Остается только учесть требования $\lambda_1, \lambda_2, \lambda_3 \geq 0$ и $\bar{\lambda} \neq 0$. Эти требования могут быть выполнены, если $\lambda_0 \geq 0$. Это означает, что точка $(0; 0)$ – это точка экстремума (минимума). С учетом того, что $x \geq 0$ из шестого неравенства системы (5.2.7), точка $(0; 0)$ является также точкой глобального минимума.

Теперь рассмотрим второй вариант, когда $\lambda_1 = 0$, $x \neq 0$. С учетом неравенства $-x \leq 0$ из системы (5.2.7), делаем вывод, что $x > 0$. Из седьмого и восьмого неравенств следует $x^2 \leq y \leq 2x^2$, причем хотя бы одно из неравенств должно быть строгим из-за того что $x > 0$. Здесь тоже нужно рассмотреть оба варианта.

Предположим, что $x^2 < y$. Тогда из четвертого уравнения системы (5.2.7) следует, что $\lambda_2 = 0$. Далее из второго уравнения $\lambda_3 = 0$. Тогда из первого $\lambda_0 = 0$ (помним, что мы рассматриваем случай, когда $\lambda_1 = 0$). В итоге получаем $\bar{\lambda} = 0$, что противоречит требованию теоремы Каруша-Джона $\bar{\lambda} \neq 0$.

Предположим тогда, что $y < 2x^2$. Тогда из пятого уравнения $\lambda_3 = 0$, далее из второго $\lambda_2 = 0$. С учетом того, что $\lambda_1 = 0$, из первого уравнения получим $\lambda_0 = 0$, что также противоречит требованию $\bar{\lambda} \neq 0$.

Таким образом, в этой задаче один минимум в точке $(0; 0)$.

§5.3. Градиентный метод

Рассмотрим задачу минимизации функции нескольких переменных $f(x) \to inf;\quad x \in X \equiv E^n$. Считаем, что функция $f(x)$ непрерывно дифференцируема на E^n. x – это n-мерный вектор. Рассмотрим приращение функции

$$f(x+h) - f(x) = \langle f'(x), h\rangle + o(h) \qquad (5.3.1)$$

В эту формулу входит «о» малое, определяемое в многомерном случае следующим образом

$$\lim_{|h|\to 0} \frac{o(h)}{|h|} = 0 \qquad (5.3.2)$$

Если $f'(x) \neq 0$, то при достаточно малых $|h|$ основная часть приращения функции определяется скалярным произведением $\langle f'(x), h\rangle$. Неравенство Коши-Буняковского говорит о том, что

$$-|f'(x)| * |h| \leq \langle f'(x), h\rangle \leq |f'(x)| * |h| \qquad (5.3.3)$$

Основной интерес для градиентного метода представляет случай, когда одно из неравенств в выражении (5.3.3) становится равенством. При условии $f'(x) \neq 0, \langle f'(x), h\rangle = |f'(x)| * |h| \Leftrightarrow h = \alpha f'(x), \quad \alpha \geq 0. \langle f'(x), h\rangle = -|f'(x)| * |h| \Leftrightarrow h = -\alpha f'(x).$

При $f'(x) \neq 0$ направление наибыстрейшего возрастания функции $f(x)$ совпадает с направлением градиента, т.е. $f'(x)$, а направление наискорейшего убывания – с направлением антиградиента, т.е. $-f'(x)$.

Перейдем к алгоритму градиентного метода. Выбирается точка начального приближения x_0. Затем строим последовательность точек $\{x_k\}$: $x_{k+1} = x_k - \alpha_k f'(x_k), \alpha_k > 0, \quad k = 0,1,2, \ldots \alpha_k$ называют шагом градиентного метода. Запишем приращение функции $f(x_{k+1}) - f(x_k) = f\big(x_k - \alpha_k f'(x_k)\big) - f(x_k) = \langle f'(x_k), -\alpha_k f'(x_k)\rangle + o\big(\alpha_k f'(x_k)\big) = -\alpha_k |f'(x_k)|^2 + o(\alpha_k)$. Предполагается, что α_k достаточно малы и $f'(x_k) \neq 0$.

Если на каком-либо шаге окажется $f'(x_k) = 0$, то процесс прекращается, x_k – стационарная точка.

Существуют различные варианты градиентного метода, отличающиеся друг от друга способом выбора α_k. Рассмотрим два таких варианта.

метод наискорейшего спуска

Введем функцию $g_k(\alpha) = f\big(x_k - \alpha f'(x_k)\big)$, $\alpha > 0$. Луч $x_k - \alpha f'(x_k)$ направлен по антиградиенту функции f в точке x_k. Шаг градиентного метода α_k определяется из условия $g_k(\alpha_k) = inf_{\alpha \geq 0}\, g_k(\alpha)$, $\alpha_k \geq 0$.

вариант с постоянным шагом

Этот вариант часто используется на практике. Берется $\alpha_k = \alpha > 0$, и при этом на каждом шаге проверяют условие $f(x_{k+1}) < f(x_k)$. Если это условие не выполняется, то уменьшают α.

Метод наискорейшего спуска имеет достаточно наглядный геометрический смысл. Точка x_{k+1} лежит на луче $L_k = \{x: x = x_k - \alpha f'(x_k), \alpha \geq 0\}$ в точке его касания поверхности уровня $\Gamma_{k+1} = \{x \in E^n: f(x) = f(x_{k+1})\}$. При этом сам луч L_k перпендикулярен к поверхности уровня $\Gamma_k = \{x \in E^n: f(x) = f(x_k)\}$ (см. рис. 24). Докажем это утверждение.

Пусть $x = x(t),\ a \leq t \leq b$ – параметрическое уравнение кривой, описывающей поверхность уровня Γ_k. Это означает, что $\forall t \in [a; b] f\big(x(t)\big) = f(x_k) = const$. Выберем t_0 так чтобы $x(t_0) = x_k$. Тогда

$$\frac{d}{dt} f\big(x(t)\big) = \langle f'\big(x(t)\big), \dot{x}(t)\rangle = 0\ \forall t \in [a; b] \qquad (5.3.4)$$

В частности, при $t = t_0,\ \langle f'(x_k), \dot{x}(t_0)\rangle = 0$, т.е. градиент (или антиградиент) перпендикулярен касательной поверхности уровня Γ_k в точке x_k. Таким образом, мы доказали, что $L_k \perp \Gamma_k$.

Далее воспользуемся условием $g_k(\alpha_k) = inf_{\alpha \geq 0}\, g_k(\alpha)$. Продифференцируем это выражение поα. Получим $g_k'(\alpha_k) = -\langle f'\big(x_k - \alpha_k f'(x_k)\big), f'(x_k)\rangle = -\langle f'(x_{k+1}), f'(x_k)\rangle = 0$.

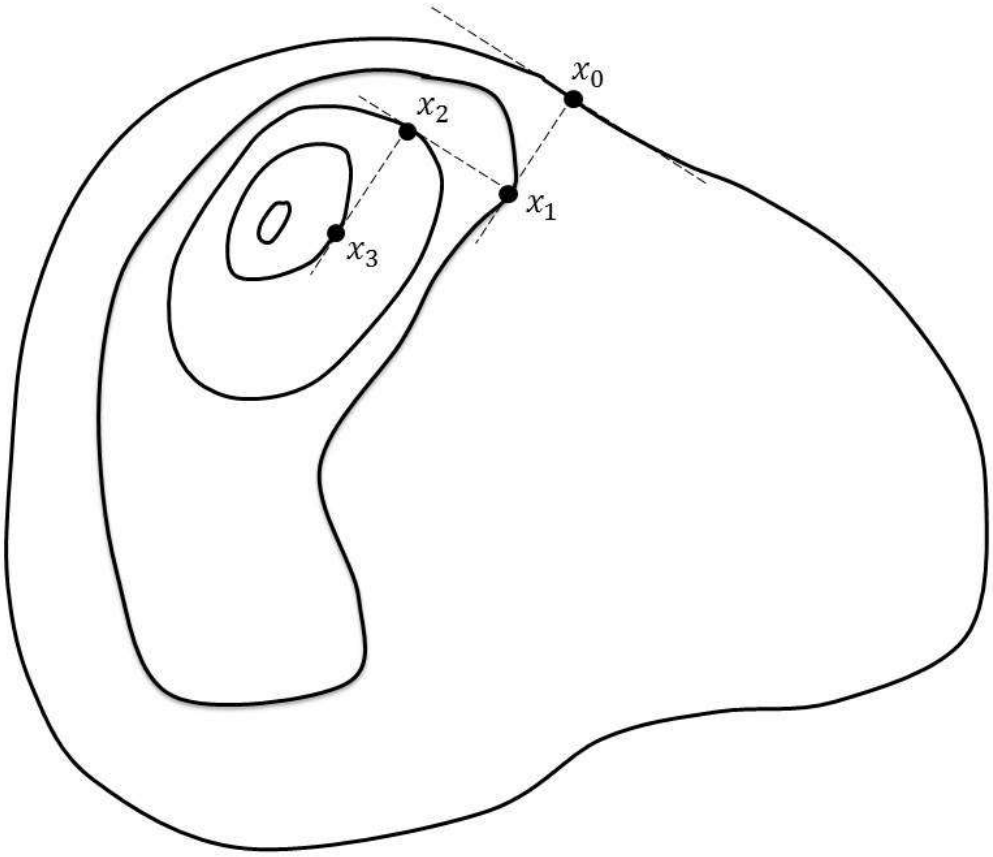

Рис. 24. Геометрический смысл метода наискорейшего спуска, замкнутыми кривыми контурами обозначены поверхности уровня

Так как вектор $f'(x_{k+1})$ перпендикулярен поверхности уровня Γ_{k+1} в точке x_{k+1}, мы можем сделать вывод о том, что направление $f'(x_k)$ и, следовательно, луч L_k – это касательная к Γ_{k+1} в точке x_{k+1}.

Градиентный метод медленно сходится, если поверхности уровня сильно вытянуты, и функция имеет так называемый овражный характер. В этом случае точки $\{x_k\}$ оказываются поочередно, то на одном, то на другом склоне оврага. Для ускорения сходимости градиентного метода в случае овражного минимума применяют специальные методики.

§5.4. Метод проекции градиента

Метод проекции градиента применяют в том случае, когда экстремум функции n ищется не во всем n-мерном пространстве E^n, а на некотором его подпространстве. Т.е. $f(x) \to inf; x \in X \subset E^n$. Функция $f(x)$считается

непрерывно дифференцируемой на множестве X, а само множество X должно быть выпуклым и замкнутым. Применение градиентного метода при такой постановке задачи затруднено тем, что на любом шаге точка x_{k+1} может не принадлежать множеству X.

В методе проекции градиента точку x_{k+1} вычисляют путем проецирования точки $x_k - \alpha_k f'(x_k)$ на множество X. Определение проекции точки на множество выглядит следующим образом: проекцией точки $u \in E^n$ на множество $X \subset E^n$ называется точка $w \in X$: $|u - w| = \inf_{v \in X} |u - v|$. Проекцию точки на множество будем обозначать как $w = P_X(u)$.

Приступим теперь к описанию метода. Вначале выбирается точка начального приближения $x_0 \in X$. Далее строится последовательность точек $\{x_k\}$ по следующему правилу

$$x_{k+1} = P_X\big(x_k - \alpha_k f'(x_k)\big), k = 0,1,2, \dots;\ \alpha_k > 0 \quad (5.4.1)$$

Если на каком-либо шаге окажется, что $x_{k+1} = x_k$, то процесс прекращается. В этом случае x_k – точка, подозрительная на экстремум. Для того чтобы выяснить, является ли эта точка точкой экстремума, необходимо проанализировать поведение функции $f(x)$ в окрестности x_k.

Как и в градиентном методе, в методе проекции градиента можно получать различные модификации метода в зависимости от способа выбора шага α_k. Рассмотрим два таких варианта.

Введем функцию $g_k(\alpha) = f\left(P_X\big(x_k - \alpha f'(x_k)\big)\right),\ \alpha \geq 0$. Шаг α_k выбирается исходя из условия

$$g_k(\alpha_k) = inf_{\alpha \geq 0}\, g_k(\alpha),\ \alpha_k \geq 0 \quad (5.4.2)$$

Этот вариант переходит в метод наискорейшего спуска из градиентного метода при$X = E^n$. Однако на практике не всегда удобно определять шаг α_k из условия (5.4.2). Поэтому более часто используется вариант метода проекции градиента с постоянным шагом.

В варианте с постоянным шагом выбирают $\alpha_k = \alpha > 0$. При этом на каждом шаге проверяют, выполнено ли условие монотонности $f(x_{k+1}) < f(x_k)$. Если это условие нарушается, то уменьшают α.

§5.5. Метод покоординатного спуска

Преимущество метода покоординатного спуска в сравнении с градиентным методом состоит в том, что здесь не требуется вычисление производной минимизируемой функции.

Рассматривается задача $f(x) \to inf; \quad x \in X \equiv E^n$. Обозначим единичный вектор вдоль i-ой координатной оси $e_i = (0, \dots, 0, 1, 0, \dots, 0)$. Единица стоит на i-ом месте. Разумеется, количество таких ортов равно размерности пространства n.

Вначале выбирают точку начального приближения x_0и начальный шаг $\alpha_0 > 0$.Метод покоординатного спуска, как и предыдущие методы, является итерационным. Пусть на некотором k-ом шаге ($k \geq 0$)точка $x_k \in E^n$и шаг $\alpha_k > 0$. Построим векторы $p_k = e_{i_k}$, где$i_k = k - n\left[\left[k/n\right]\right] + 1$, двойными квадратными скобками обозначена целая часть числа. По сути, p_k – это циклический перебор координатных ортов, т.е. $p_0 = e_1$, $p_1 = e_2$,…,$p_{n-1} = e_n$, $p_{n+1} = e_2$,…, $p_{2n-1} = e_n$, $p_{2n} = e_1$,…

Затем проверяют, выполняется ли условие

$$f(x_k + \alpha_k p_k) < f(x_k) \quad (5.5.1)$$

Если выполняется, то полагают $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_{k+1} = \alpha_k$. Если же условие (5.5.1) не выполняется, то проверяют, выполняется ли условие

$$f(x_k - \alpha_k p_k) < f(x_k) \quad (5.5.2)$$

Если условие (5.5.2) выполняется, то полагают $x_{k+1} = x_k - \alpha_k p_k$, $\alpha_{k+1} = \alpha_k$. Итерация называется удачной, если выполнено хотя бы одно из условий (5.5.1) или (5.5.2). Если же не выполняется ни (5.5.1), ни (5.5.2), тогда $(k+1)$-ю итерацию называют неудачной. В этом случае полагают $x_{k+1} = x_k$, при этом шаг дробится.

$$\alpha_{k+1} = \begin{cases} \lambda\alpha_k, \ i_k = n, x_k = x_{k-n+1} \\ \alpha_k, i_k \neq n \ \text{или} \ x_k \neq x_{k-n+1} \ \text{или} \ 0 \leq k \leq n-1 \end{cases} \quad (5.5.3)$$

В этом выражении $\lambda \in (0; 1)$ – параметр метода.Условие (5.5.3) означает, что если среди последних n итераций не оказалось ни одной удачной, то шаг дробится. Это общая схема метода покоординатного спуска для случая когда $X \equiv E^n$.

Если рассматривается задача минимизации функции не наE^n, а на некотором ограниченном многомерном параллелепипеде $X = \{(x^1, \dots, x^n): a_i \leq x^i \leq b_i, i = 1, \dots, n\}$, $a_i \leq b_i$, то наряду с условиями (5.5.1) и (5.5.2) проверяют условия $x_k + \alpha_k p_k \in X$и $x_k - \alpha_k p_k \in$

Xсоответственно. Например, если выполнено (5.5.1) и $x_k + \alpha_k p_k \in X$, то полагают $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_{k+1} = \alpha_k$. В остальном все так же.

§5.6. Метод покрытия в многомерных задачах

Отличительной чертой метода покрытий является то, что он позволяет находить глобальный минимум. При этом вычисление производной не требуется.

Рассмотрим задачу $f(x) \to inf,\ x \in \Pi = \{x = (x^1, \dots, x^n)\colon a_i \le x^i \le b_i, i = 1, \dots, n\}$. Функция $f(x)$ удовлетворяет условию Липшица $|f(x) - f(y)| \le L|x - y|_\infty \forall x, y \in \Pi$. Здесь $L > 0$, а $|x - y|_\infty = \max_{1 \le i \le n} \left|x^i - y^i\right|$. В принципе, можно воспользоваться любой другой нормой $|x - y|_p$, $1 \le p < \infty$. Однако, если функция удовлетворяет условию Липшица с какой-либо нормой, то она будет также удовлетворять условию Липшица со всеми остальными нормами. Это называется условием эквивалентности норм.

Введем на множестве П сетку Π_h, составленную из точек $x_{i_1,\dots,i_n} = \left(x^1_{i_1}, x^2_{i_2}, \dots, x^j_{i_j}, \dots, x^n_{i_n}\right)$. У этих точек j-ая координата $x^j_{i_j}$ образована по правилу $x^j_1 = a_j + 0.5h,\ x^j_2 = x^j_1 + h, \dots,\ x^j_{i+1} = x^j_i + h, \dots,\ x^j_{m_j - 1} = x_1 + \left(m_j - 2\right)h$, $x^j_{m_j} = min\left\{x^j_1 + \left(m_j - 2\right)h; b_j\right\}$. Здесь $h = 2\varepsilon/L$ – шаг метода, ε – наперед заданная точность, а число m_j определяется условием$x^j_{m_j - 1} < b_j - 0.5h \le x^j_1 + (m_j - 1)/h$.

Приближением нижней грани f_*является величина $F_h = \min_\Pi f\left(x_{i_1}, \dots, x_{i_n}\right)$, которая в простейшем варианте метода находится путем простого перебора точек сетки. Нетрудно показать, что $F_h \in [f_*, f_* + \varepsilon]$. Элементарные объемы $\Pi_{i_1,\dots,i_n} = \left\{x \in E^n\colon \left|x - x_{i_1,\dots,i_n}\right|_\infty \le h/2\right\}$покрывают весь объем П. Из условия Липшица следует, что $F_h - f_* \le 0.5Lh = \varepsilon$.

Существует вариант последовательного перебора, который требует меньшего количества вычислений по сравнению с простым перебором. Этот метод не требует вычисления значений функции во всех точках сетки. В этом варианте вначале выбирают произвольную точку v_1и считают значение функции в ней $F_1 = f(v_1)$. Затем выбирают

определенный набор точек$v_2, \dots, v_k$и вычисляют $F_k = \min_{1 \le i \le k} f(v_i)$. При $k \ge 2 F_k = min\{F_{k-1}; f(v_k)\}$. Обозначим v_{j_k}ту из точек набора $v_1, \dots, v_k$, в которой достигается равенство $f(v_{j_k}) = F_k$.

Ввиду того, что в методе последовательного перебора значения функции вычисляются не во всех точках сетки, некоторые точки определенным образом исключаются. Механизм такого исключения будет описан немного ниже.

Берется точка $v_{k+1} \in \Pi_h$, которая не исключалась из рассмотрения и в которой не вычислялось значение функции ранее. Вычисляем $F_{k+1} = min\{F_k; f(v_{k+1})\} = \min_{1 \le i \le k+1} f(v_i)$. Здесь существует два возможных варианта – либо $F_{k+1} = f(v_{k+1}) < F_k$, либо $F_{k+1} = F_k \le f(v_{k+1})$.

В первом случае, когда $F_{k+1} = f(v_{k+1}) < F_k$, полагают $v_{j_{k+1}} = v_{k+1}$, и из дальнейшего рассмотрения исключаем точку v_{j_k}, а также точки $x_{i_1,\dots,i_n}$: $|x_{i_1,\dots,i_n} - v_{j_k}| \le (F_k - F_{k+1})/L$. В этих точках значения функции не могут оказаться меньше, чем F_{k+1}. Докажем это. $f(x_{i_1,\dots,i_n}) - F_{k+1} = f(x_{i_1,\dots,i_n}) - f(v_{j_k}) + F_k - F_{k+1} \ge -L|x_{i_1,\dots,i_n} - v_{j_k}| + F_k - F_{k+1}$. Так как $|x_{i_1,\dots,i_n} - v_{j_k}| \le (F_k - F_{k+1})/L$, $f(x_{i_1,\dots,i_n}) - F_{k+1} \ge 0$.

Во втором случае, когда $F_{k+1} = F_k \le f(v_{j_k})$, полагаем $v_{j_{k+1}} = v_{j_k}$, и из дальнейшего перебора исключаем точку v_{k+1}, а также точки $x_{i_1,\dots,i_n}$: $|x_{i_1,\dots,i_n} - v_{k+1}| \le (f(v_{k+1}) - F_k)/L$.Докажем, что значения функции в исключенных точках не могут быть меньше, чем F_{k+1}. $f(x_{i_1,\dots,i_n}) - F_{k+1} = f(x_{i_1,\dots,i_n}) - F_k = f(x_{i_1,\dots,i_n}) - f(v_{k+1}) + f(v_{k+1}) - F_k \ge -L|x_{i_1,\dots,i_n} - v_{k+1}| + f(v_{k+1}) - F_k \ge 0$.

Процесс останавливается тогда, когда на каком-либо шаге не находится такая точка v_{k+1}, которая не исключалась из рассмотрения на предыдущих шагах, и в которой не вычислялось значение функции.

Заключение

В настоящем пособии рассмотрены основные разделы математического моделирования.

Первая глава настоящего пособия посвящена базовым принципам построения математических моделей.Рассмотрены модели, получаемые на

основе фундаментальных законов природы, вариационного принципа, а также по иерархическому принципу и с помощью аналогий. Данная глава оснащена достаточно большим количеством примеров математических моделей, прежде всего получаемых из вариационного принципа. Эти примеры могут быть полезны для студентов, которые не изучали вариационное исчисление ранее. Рассмотрен также вопрос об исследовании математических моделей на примере принципа максимума и его следствий.

Вторая глава посвящена конечно-разностным методам для численного решения дифференциальных уравнений. В начале этой главы приведены различные разностные аппроксимации для производных, а также разностные схемы для дифференциальных уравнений. Изложено применение конечно-разностных методов, как к обыкновенным дифференциальным уравнениям, так и к уравнениям в частных производных. Рассматриваются вопросы устойчивости разностных схем.

В третьей главе изложен метод конечных элементов. Описаны различные типы базисных функций, прежде всего кусочно-полиномиальные. Дано описание лагранжевых и эрмитовых элементов. Изложен классический метод конечных элементов – метод Ритца. Дано определение слабой формы уравнения, рассмотрен метод Галеркина.

Четвертая глава посвящена статистическим методам Монте-Карло. Рассматриваются методы разыгрывания случайных величин, как дискретных, так и непрерывных. Излагаются два метода Монте-Карло для вычисления интегралов – простейший и геометрический. Проводится сравнительный анализ этих методов, дается понятие трудоемкости алгоритма.

В пятой главе изложены основные методы оптимизации, т.е. методы нахождения минимального (или максимального) значения функции на заданном множестве. Анализируются как функции одной переменной, так и нескольких. Особое внимание уделено минимизации функции на множестве с ограничениями. Описано правило множителей Лагранжа для нахождения условного экстремума многомерных функций. Изложены основные методы численной оптимизации многомерных функций, такие как градиентный метод, метод проекции градиента, метод покоординатного спуска, а также различные формы метода покрытий.

Пособие предназначено для студентов старших курсов, изучающих такие дисциплины как «Математическое моделирование», «Методы моделирования и оптимизации».

Литература

[1] Самарский А.А., Михайлов А.П. «Математическое моделирование. Идеи. Методы. Примеры.»

[2] Гельфанд И.М., Фомин С.В. «Вариационное исчисление»

[3] Самарский А.А.,Гулин А.В. «Численные методы»

[4]Митчел Э., Уэйт Р. «Метод конечных элементов для уравнений с частными производными

[5] Соболь И.М. «Численные методы Монте-Карло»

[6] Васильев Ф.П. «Методы оптимизации»

[7] ТихоновА.Н., Самарский А.А. «Уравнения математической физики».