Методы оптимизации

2.МЕТОДЫ ОДНОМЕРНОЙ МИНИМИЗАЦИИ

2.1. ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

В некоторых случаях *ограничения* в *задаче оптимизации* позволяют через один из *параметров оптимизации* выразить остальные и исключить их из *целевой функции*. В результате задача будет сведена к поиску наибольшего или наименьшего (в зависимости от цели оптимизации) значения скалярной действительной функции $f(x)$, $x \in D(f) \subset \mathrm{R}$, выражающей *критерий оптимальности*. Выбирая тот или иной знак перед этой функцией, всегда можно ограничиться лишь поиском ее наименьшего значения в области определения $D(f)$, заданной с учетом ограничений на параметр оптимизации x. Поэтому далее в этой главе будем рассматривать задачу

$$f(x) \to min,\ x \in D(f) \subset \mathrm{R}, \quad (2.1)$$

поиска наименьшего значения $f = f(x_*)$ функции $f(x)$ и точки $x_* \in D(f)$, в которой $f(x)$ принимает это значение. Для краткости будем говорить об ***одномерной минимизации***, имея в виду нахождение наименьшего значения функции $f(x)$ на множестве $D(f)$ и точек, в которых это значение достигается.

Изучение методов одномерной минимизации важно не только для решения задачи (2.1), имеющей самостоятельное значение. Эти методы являются также существенной составной частью методов многомерной минимизации, при помощи которых находят наименьшее значение действительных функций многих переменных.

Пусть область определения $D(f)$ функции $f(x)$ есть промежуток числовой прямой. Напомним, что если $D(f)$ — отрезок и $f(x)$ непрерывна на нем, то она имеет на этом отрезке наименьшее значение. Но при наличии на отрезке точек разрыва функции она может не иметь на нем наименьшего значения. Оно может не существовать и в том случае, когда $D(f)$ является интервалом или полуинтервалом.

Если функция $f(x)$ не имеет на множестве $D(f)$ наименьшего значения, то (2.1) следует заменить формулировкой задачи в виде

$$f(x) \to inf,\ x \in D(f) \subset \mathbb{R} \quad (2.2)$$

Тогда под решением задачи минимизации такой функции на $D(f)$ следует понимать построение последовательности $\{x_n\}$ точек из $D(f)$, для которой существует предел

$$\lim_{n\to\infty} f(x_n) = \inf_{x_n \in D(f)} f(x) = \tilde{f}_*, \quad (2.3)$$

и нахождение этого предела. Если функция $f(x)$ достигает на множестве $D(f)$ своего наименьшего значения f_*, то $\tilde{f}_* = f_*$.

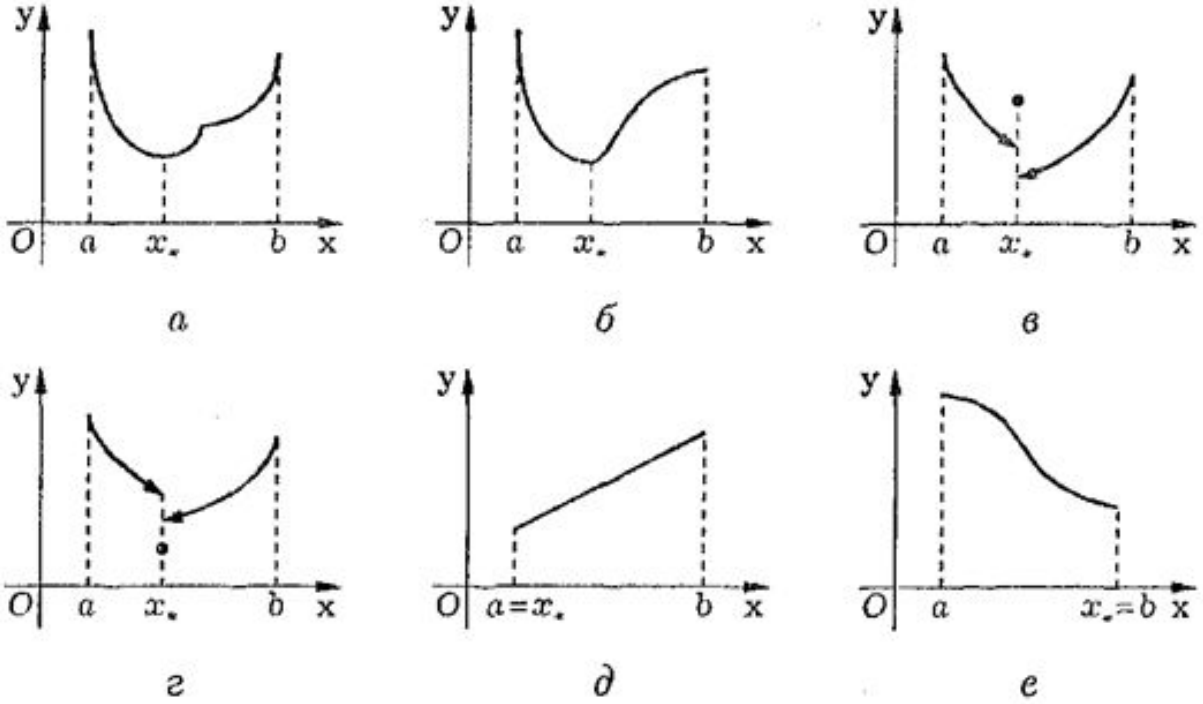
Например, функция $f(x) = 1/x$ на множестве $D(f) = [1, 2)$ не достигает наименьшего значения, хотя и ограничена снизу. Точная нижняя грань $\tilde{f}_*$ функции в данном случае равна $1/2$. В качестве последовательности $\{x_n\}$ точек в полуинтервале [1. 2), для которой справедливо (2.3), можно выбрать $\{2 - 1/n\}$. Тогда

$$f(x_n) = \frac{1}{x_n} = \frac{1}{2 - 1/n} = \frac{n}{2n - 1},$$

и последовательность $\{f(x_n)\}$ сходится к числу $1/2 = \tilde{f}_*$.

Функция может достигать наименьшего значения как в единственной точке, так и на некотором множестве точек, конечном, счетном или несчетном. Например, функция $f(x) = x^4$ определена на всей числовой прямой и достигает своего наименьшего значения $f_* = 0$ в единственной точке $x_* = 0$, которая является ее точкой минимума. Функция $f(x) = x^4 - 2x^2 + 2$ также определена на всей числовой прямой и достигает наименьшего значения $f_* = 1$ в точках $x_* = \pm 1$. Функция $f(x)=cosx$ достигает наименьшего значения на счетном множестве $D_* = \{x \in \mathbb{R}:\ x = \pi + 2\pi k,\ k \in \mathbb{Z}\}$, а функция $f(x) = |x + 1| + |x - 1|$ — на несчетном множестве $D_* = [-1,1]$.

Функцию $f(x)$ называют ***унимодальной функцией*** на отрезке $[a, b]$, если существует такая точка $x_* \in [a, b]$, что функция $f(x)$ в полуинтервале $[a, x_*)$ убывает, а в полуинтервале $(x_*, b]$ возрастает. Примеры графиков унимодальных функций приведены на рис. 2.1.



Точка x_* может быть внутренней точкой отрезка $[a, b]$ (т.е. $a < x_* < b$, см. рис. *а-г*) или совпадать с одним из его концов ($x_* = a$ или $x_* = b$, см. рис. *д,е*). Унимодальная функция не обязательно непрерывна на отрезке $[a, b]$ (см. рис. *в, г*).

Функцию $f(x)$, достигающую на отрезке $[a, b]$ наименьшего значения в единственной точке $x_* \in [a, b]$, убывающую при $x \in [a, x_*]$ и возрастающую при $x \in [x_*, b]$, будем называть ***строго унимодальной*** на отрезке $[a, b]$ (на рисунке строго унимодальными являются все функции, кроме функции на рис. *г*).

Область определения $D(f)$ минимизируемой функции $f(x)$ может состоять из нескольких промежутков, не имеющих граничных точек. В этом случае, чтобы найти значение функции на множестве $D(f)$, достаточно определить наименьшее значение функции в каждом из промежутков, составляющих $D(f)$, а затем, сравнивая, выбрать среди этих значений минимальное.

Если функция дифференцируема в промежутке, то возможно использование необходимого и достаточного условий локального минимума [II]. Однако в прикладных задачах нередки ситуации, когда трудно вычислить производные функции (например, если функция не задана в аналитическом виде). Более того, не исключено, что значения функции известны или могут быть вычислены только в отдельных точках. В таких ситуациях использование необходимого и достаточного условий локального минимума невозможно и следует применять другие методы решения задачи оптимизации. Методы минимизации функции одного переменного, в которых используют значения функции в точках рассматриваемого промежутка и не используют значения ее производных, называют ***методами прямого поиска.***

2.2. ПАССИВНЫЙ И ПОСЛЕДОВАТЕЛЬНЫЙ ПОИСК

Пусть требуется найти наименьшее значение или точную нижнюю грань f_* скалярной действительной функции $f(x)$ одного переменного на отрезке $[a, b]$. Предположим, что задан алгоритм вычисления значения функции для любой точки $x \in [a, b]$. Можно выделить две группы *методов прямого поиска*, соответствующие двум принципиально различным ситуациям:

1) все N точек x_k, $k = 1, 2, \ldots, N$, в которых будут вычислены значения функции, выбирают заранее (до вычисления функции в этих точках);

2) точки x_k выбирают последовательно (для выбора последующей точки используют значения функции, вычисленные в предыдущих точках).

В первом случае поиск значения f_* называют ***пассивным,*** а во втором — ***последовательным.*** Естественно ожидать, что последовательный поиск лучше пассивного. В этом можно убедиться, вспомнив детскую игру, в которой надо найти спрятанную вещь, задавая вопросы и получая на них ответы „да" или „нет". Задавая вопросы последовательно с учетом предыдущих ответов, можно найти спрятанную вещь за меньшее число вопросов (итераций), чем, задав определенное количество заранее подготовленных вопросов сразу.

Так как в прикладных задачах вычисление каждого значения функции может быть достаточно трудоемким, то целесообразно выбрать такую стратегию поиска, чтобы значение f_* с заданной точностью было найдено наиболее экономным путем. Будем считать, что стратегия поиска определена, если:

- определен алгоритм выбора точек x_k, $k = 1, 2, \ldots, N$;
- определено условие прекращения поиска, т.е. условие, при выполнении которого значение f_* считают найденным с заданной точностью.

Для методов пассивного поиска алгоритм выбора точек x_k, $k = 1, 2, \ldots, N$, — это правило, по которому заранее определяют все N точек x_k, $k = 1, 2, \ldots, N$, в которых затем будут вычислены значения функции $f(x)$. Для методов последовательного поиска алгоритм выбора точек x_k — это правило, по которому последовательно определяют каждую следующую точку x_k по информации о расположении точек x_i, $i = 1, 2, \ldots, k$-1, и о вычисленных значениях $f(x_i)$ функции $f(x)$ в этих точках. Выбор очередной точки x_k и вычисление значения $f(x_k)$ называют ***шагом*** последовательного *поиска*.

В методах последовательного поиска количество точек x_k обычно не задают заранее. Однако объективное сравнение различных методов прямого поиска нужно проводить при одинаковом количестве n вычисленных значений функции $f(x)$. После n вычислений обычно указывают интервал (или отрезок) длины l_n, называемый ***интервалом неопределенности***, в котором гарантированно находится точка x_*, соответствующая значению f_*. Условие прекращения вычислений в случае пассивного или последовательного поиска примем одинаковым — выполнение неравенства $l_n \le \varepsilon_*$, где ε_* — заданная наибольшая допустимая длина интервала неопределенности.

Длина l_n зависит как от самого метода прямого поиска P, так и от минимизируемой функции $f(x)$, т.е. $l_n = l_n(P,f)$. Зависимость l_n от n дает оценку скорости сходимости конкретного метода прямого поиска P к искомому значению f_* заданной функции $f(x)$. Различные методы из некоторого множества $\mathcal{P}$ методов прямого поиска сравнивают обычно при выбранном фиксированном значении $n=N$ на некотором достаточно широком классе функций. В качестве такого класса можно выбрать множество *унимодальных функций*, определенных на фиксированном отрезке $X \subset \mathbb{R}$. Для метода прямого поиска $P \in \mathcal{P}$ наихудшую оценку

$$l_N(P) = \max_{f \in \mathcal{F}} l_N(P,f).$$

Если „наихудшей” унимодальной функции не найдется, то оценку принимаем в виде

$$l_N(P) = \sup_{f \in \mathcal{F}} l_N(P,f).$$

Значение $l_N(P)$ представляет собой оценку сверху погрешности вычисления точки $x_* \in X$, соответствующей искомому значению f_* произвольной функции $f \in \mathcal{F}$, которая получена методом прямого поиска $P \in \mathcal{P}$ по N вычисленным значениям этой функции. Метод прямого поиска P^* считаем наилучшим, если $l_N(P^*) = \min_{P \in \mathcal{P}} \max_{f \in \mathcal{F}} l_N(P,f)$, или $l_N(P^*) = \min_{P \in \mathcal{P}} \sup_{f \in \mathcal{F}} l_N(P,f)$.

Этот критерий сравнения методов поиска определяет ***минимаксный метод поиска***. Такой метод является наилучшим для всего множества $\mathcal{F}$ унимодальных функций на отрезке $f \in \mathcal{F}$, том смысле, что он дает наименьшую погрешность вычисления точки x_*, соответствующей значению f_* любой из рассматриваемых функций $X \subset \mathbb{R}_B$. Хотя вполне возможно, что существует некоторый конкретный метод, который для определенной специально подобранной унимодальной функции из множества $\mathcal{F}$ обеспечит еще меньшую погрешность.

Все методы прямого поиска можно строить и сравнивать между собой на отрезке $X = [0, 1]$. Полученные результаты при необходимости нетрудно перенести на случай произвольного отрезка $[a, b]$, так как любую точку отрезка [0,1] можно перевести в соответствующую ей точку отрезка $[a, b]$ растяжением в $b - a$ раз и сдвигом на a.

Если минимизируемая функция $f(x)$ не является унимодальной на отрезке $[a,b]$ (такую функцию называют ***мультимодальной функцией*** на этом отрезке), то, даже если она непрерывна на $[a, b]$, при поиске наименьшего значения f_* функции на отрезке может возникнуть ошибка: будет найдена точка локального минимума, в которой значение функции не f_*, а другое, большее. Чтобы избежать такой ошибки, в процесс минимизации включают предварительный этап, на котором отрезок минимизации разделяют на несколько отрезков, на каждом из которых минимизируемая функция унимодальна. Сравнительный анализ наименьших значений функции на этих отрезках позволяет найти искомое наименьшее значение f_* на всем отрезке минимизации.

Пример 2.1. Рассмотрим один из возможных подходов к выделению из промежутка X в области определения $D(f)$ минимизируемой функции $f(x)$ отрезка $[a, b]$, на котором эта функция является унимодальной. Пусть известна такая точка $x_0 \in X$, что при $x \geq x_0$ функция $f(x)$ сначала убывает, а затем, начиная с пока неизвестного значения $x = x_* \in X$ возрастает, хотя далее в промежутке X могут быть расположены и другие участки немонотонного поведения этой функции. Выберем начальное значение $h > 0$ приращения аргумента x функции $f(x)$, в несколько раз меньшее предполагаемого расстояния между точками x_0 и x_*, и вычислим значения $f(x)$ и $f(x_1)$, где $x_1 = x_0 + h$.

Может оказаться, что $f(x_0) \leq f(x_1)$. Тогда за искомый отрезок $[a, b]$ можно сразу принять отрезок $[x_0, x_1]$. Но можно продолжить вычисления и, используя последовательный поиск, определять значения $f(x_k)$ в точках $x_k = x_0 + h/2^{k-1},\ k = 2,3 \ldots,$ до тех пор, пока не будет выполнено неравенство $f(x_k) < f(x_0)$. Тогда следует принять $[a, b] = [x_0, x_{k-1}]$ (на рис. 2.2 $[a, b] = [x_0, x_2]$, поскольку точка x_* должна быть либо на отрезке $[x_0, x_3]$, либо на отрезке $[x_3, x_2]$). Надо сказать, что при этом можно „не заметить" по крайней мере, еще один отрезок, на котором функция унимодальна (штриховая линия на рис. 2.2).

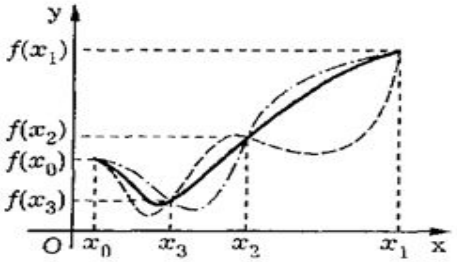


Рис. 2.2

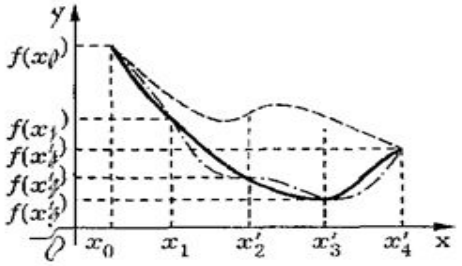


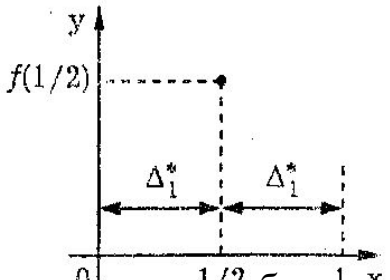
Рис. 2.3

Если $f(x_0) > f(x_1)$, то используя последовательный поиск, вычисляем значения $f(x_k')$ где $x_k' = x_0 + (k-1)h,\ k = 2,3 \ldots,$ пока не будет выполнено неравенство $f(x_{k-1}') \leq f(x_k')$, что позволяет принять $[a, b] = [x_{k-2}', x_k']$ (на рис. 2.3 $[a, b] = [x_2', x_4']$, так как точка x_* должна быть либо на отрезке $[x_2', x_3']$, либо на отрезке $[x_3', x_4']$)

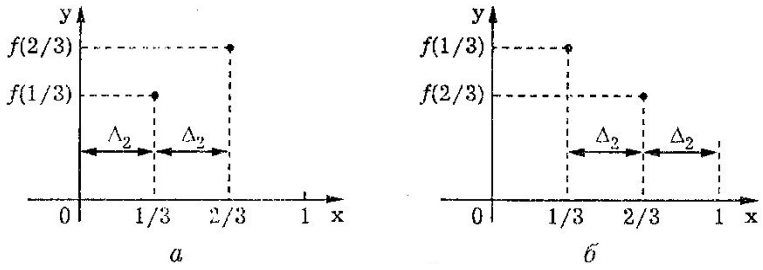
Отметим, что описанный подход не гарантирует нахождения отрезка унимодальности функции. Например, на рис. 2.3 штриховой линией показан график функции, для которой этот подход не позволяет обнаружить искомый отрезок.

2.3. ОПТИМАЛЬНЫЙ И ПАССИВНЫЙ ПОИСК

Пусть требуется путем *пассивного поиска* найти точку $x_* \in [0, 1]$, в которой *унимодальная* на отрезке [0, 1] *функция* $f(x)$ достигает наименьшего значения $f_* = f(x_*)$. *Минимаксный метод поиска*, в котором информация о значениях функции, вычисленных в предшествующих точках, не может быть использована, называют ***оптимальным пассивным поиском***. Рассмотрим алгоритм такого поиска при различном числе N точек, выбираемых на отрезке [0, 1].



Если $N = 1$, то единственную точку целесообразно выбрать в середине отрезка, т.е. принять $x_1 = 1/2$ (рис. 2.4). В этом случае вследствие унимодальности функции $f(x)$ имеем $f_* \le f(1/2)$. Поэтому наименьшая возможная длина *интервала неопределенности* равна $l_1^* = 1$ и можно гарантировать, что выбор в качестве точки $x_* \in [0, 1]$ точки $x_1 = 1/2$ приведет к погрешности не более $\Delta_1^* = l_1^*/2 = 1/2$. При любом ином положении точки x_1 погрешность при выборе $x_* = x_1$ будет $\Delta_1 \ge \Delta_1^*$, так как в действительности точка x_* может лежать на большей части отрезка [0, 1].



Если при $N = 2$ (рис. 2.5) две точки расположить на отрезке [0,1] так, чтобы они делили его на равные части, т.е. выбрать $x_1 = 1/3$ и $x_2 = 2/3$, то точка $x_* \in [0,1]$ будет найдена с точностью $\Delta_2^* = 1/3$, а наименьшая длина интервала неопределенности составит $l_2^* = 2\Delta_2^* = 2/3$. В самом деле, если $f(1/3) < f(2/3)$ (рис. 2.5, *а*), то в силу унимодальности функции $f(x)$ отрезок [2/3,1] можно исключить и считать, что $x_* \in [0,2/3]$ Тогда при выборе $x_* = 1/3$ наибольшая погрешность равна $\Delta_2 = 1/3$ и $f_* \approx f(1/3)$ Если же окажется, что $f(1/3) > f(2/3)$ (рис. 2.5, *б*), то можно исключить отрезок [0,1/3] и считать, что $x_* \in [1/3,1]$. И в этом случае выбор $x_* = 2/3$ приведет к погрешности не более $\Delta_2 = 1/3$, а $f_* \approx f(2/3)$ Заметим, что при $f(1/3) = f(2/3)$ можно исключить любой из указанных отрезков, гарантируя ту же точность нахождения точки $x_* \in [0,1]$. При ином делении отрезка [0, 1] на части двумя точками длина какой-то из его частей будет больше 1/3 и в действительности точка может принадлежать именно этой части, так что получим погрешность $\Delta_2 > \Delta_2^* = 1/3$.

Рассуждая аналогично, можно заключить, что при $N = 3$ нужно также выбирать точки равномерно на отрезке [0,1]: $x_1 = 1/4$, $x_2 = 2/4$, $x_3 = 3/4$, обеспечив точность $\Delta_3^* = 1/4$ нахождения точки $x_* \in [0,1]$ и наименьшую длину $l_3^* = 1/2$ интервала неопределенности. В случае произвольного $N \in \mathbb{N}$ по тем же соображениям надо выбирать точки

$$x_k = \frac{k}{N+1} \in [0,1], \qquad k = \overline{1,N}, \qquad (2.4)$$

обеспечивая точность $\Delta_N^* = 1/(N+1)$ нахождения точки и наименьшую возможную длину

$$\Delta_N^* = \frac{2}{N+1} \qquad (2.5)$$

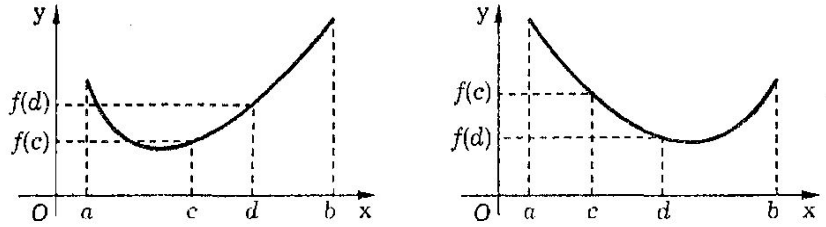
интервала неопределенности. Таким образом, оптимальный пассивный поиск состоит в выборе точек, равномерно расположенных на отрезке. При этом (2.5) дает оценку скорости сходимости пассивного поиска с ростом числа N точек, так как скорость сходимости любого метода прямого поиска можно характеризовать скоростью уменьшения интервала неопределенности с возрастанием N.

Пример 2.2. При заданной наибольшей допустимой длине интервала неопределенности, используя оптимальный пассивный поиск, найдем точку , в которой унимодальная на отрезке [0,1] функция $f(x) = x^3 - x + e^{-x}$ достигает наименьшего на этом отрезке значения*. Из (2.5) следует, что для этого необходимо принять $N = 9$ и в соответствии с (2.4) вычислить значения функции $f(x)$ в точках $x_k = k/10, \quad k = \overline{1,9}$:

x	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$f(x)$	0,81	0,63	0,47	0,33	0,23	0,17	0,14	0,16	0,18

Из представленных результатов вычислений можно сделать вывод, что интервалом неопределенности является интервал (0,6, 0,8), а $x_* = 0{,}7 \pm 0{,}1$. Отметим, что при $\varepsilon_* = 0{,}01$ потребуется принять $N = 199$.

Рассуждения, проведенные выше, попутно обосновывают ***процедуру исключения отрезка***, которую используют во всех методах прямого поиска точки минимума унимодальной функции одного переменного. Эта процедура состоит в следующем. Пусть на отрезке $[a, b]$ числовой прямой расположены две точки c и d, $a < c < d < b$, и известны (или вычислены) значения $f(c)$ и $f(d)$ унимодальной на $[a, b]$ функции $f(x)$. Если $f(c) < f(d)$ (рис. 2.6, *а*), то в силу унимодальности функции $f(x)$ имеем $x_* \in [a, d]$, а отрезок $[d, b]$ можно исключить из дальнейшего рассмотрения. Наоборот, если $f(c) \geq f(d)$ (рис. 2.6, *б*), то $x_* \in [c, b]$, а отрезок $[a, c]$ далее можно не рассматривать.



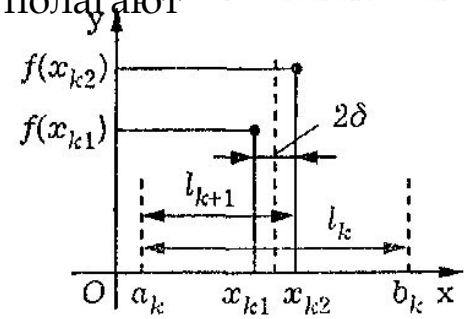
а *б*

Таким образом, в результате применения процедуры исключения отрезка получаем новый отрезок, вложенный в рассматриваемый и заведомо содержащий искомую точку . В методах пассивного поиска применение этой процедуры позволяет оценить наибольшую возможную погрешность нахождения точки . Все рассмотренные далее методы *последовательного поиска* используют процедуру исключения отрезка для выбора нового отрезка на каждом очередном *шаге* такого поиска.

2.4. МЕТОДЫ ПОСЛЕДОВАТЕЛЬНОГО ПОИСКА

Метод дихотомии. Рассмотрим *последовательный поиск* точки $x_* \in [0, 1]$, в которой *унимодальная* на отрезке [0, 1] *функция* $f(x)$ достигает наименьшего значения $f_* = f(x_*)$. Метод прямого поиска, основанный на делении пополам отрезка, на котором находится точка x_*, называют ***методом дихотомии***. Опишем алгоритм этого метода.

Пусть известно, что на k-м *шаге последовательного поиска* $x_* \in [a_k, b_k] \subset [0,1]$ (на первом шаге при $k = 1$ имеем $a_1 = 0$ и $b_1 = 1$). На отрезке $[a_k, b_k]$ длиной l_k выберем две точки $x_{k1} = (a_k + b_k)/2 - \delta$ и $x_{k2} = (a_k + b_k)/2 + \delta$ (рис. 2.7), где $\delta > 0$ — некоторое достаточно малое число. Вычислим значения $f(x_{k1})$ и $f(x_{k2})$ функции $f(x)$ в этих точках и выполним *процедуру исключения отрезка*. В результате получим новый отрезок $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$. Если длина l_k+1 нового отрезка больше заданной наибольшей допустимой длины ε_* *интервала неопределенности*, то алгоритм метода дихотомии переходит к $(k + 1)$-му шагу, повторяя все описанные для k-го шага. Если $l_{k+1} \le \varepsilon_*$, то вычисления прекращают и полагают $x_* = (a_{k+1} + b_{k+1})/2$.



Так как $l_{k+1} = l_k/2 + \delta$, или $l_{k+1} - 2\delta = (l_k - 2\delta)/2$, то

$$l_k - 2\delta = \frac{l_1 - 2\delta}{2^{k-1}}$$

Из этого равенства выводим следующую формулу длины l_k отрезка $[a_к, b_к]$, получаемого на k-м шаге метода дихотомии:

$$l_k = \frac{l_1 - 2\delta}{2^{k-1}} + 2\delta. \qquad (2.6)$$

Из (2.6) следует, что $l_k \to 2\delta$ при $k \to \infty$ но при этом Поэтому выполнение неравенства $l_{k+1} < \varepsilon_*$, означающее достижение заданной точности нахождения точки возможно лишь при условии выбора $2\delta < \varepsilon_*$ Кроме того, нужно учитывать неизбежную погрешность, возникающую при вычислении приближенных значений $\tilde{f}(x)$ функции $f(x)$. Это приводит к дополнительной погрешности Δ_* при нахождении точки x_* (см. 2.7). Поэтому выбор значения δ ограничен и снизу, т.е.

$$\Delta_* < 2\delta < \varepsilon_*. \qquad (2.7)$$

Если эти неравенства нарушаются, то знак разности $\tilde{f}(x_{k1}) - \tilde{f}(x_{k2})$ может не совпадать со знаком разности $(x_{k1}) - f(x_{k2})$, что приводит к ошибочному выполнению процедуры исключения отрезка.

Итак, метод дихотомии — это последовательное построение на каждом $к$-м шаге поиска точек $x_{k1} = (a_k + b_k)/2 - \delta$ и $x_{k1} = (a_k + b_k)/2 + \delta$ симметричных относительно середины отрезка $[a_к, b_к]$ длины $l_к$. После выполнения k-го шага будет выделен отрезок $[a_{к+1}, b_{k+1}]$ длины $l_{к+1}$ и вычислено $N = 2k$ значений функции. Используя формулу (2.6) для длины отрезка *(интервала неопределенности)* и полагая $l_1 = 1$, получаем

$$l_N^d = l_{k+1} = \frac{1 - 2\delta}{2^{k+1}} + 2\delta = \frac{1 - 2\delta}{2^{N/2}} + 2\delta. \qquad (2.8)$$

Сравнивая (2.8) с (2.5), видим, что скорость сходимости метода дихотомии значительно выше скорости сходимости *оптимального пассивного поиска*.

Отметим, что после исключения отрезка на k-м шаге описанного алгоритма точки $x_{к1}$ и $x_{к2}$ принадлежат новому отрезку $[a_{к+1}, b_{k+1}]$, причем одна из них является внутренней для этого отрезка. Но вычисленное в этой точке значение функции $f(x)$ в методе дихотомии не используют для исключения отрезка на следующем шаге, а проводят вычисления в двух новых точках. Существуют методы последовательного поиска, в которых на каждом k-м шаге начиная с $к = 2$ вычисляют лишь одно новое значение функции в точке, принадлежащей отрезку $[a_{к+1}, b_{k+1}]$. Это значение вместе с уже вычисленным на предыдущем шаге значением функции во внутренней точке отрезка $[a_к, b_к]$ используют при выполнении процедуры исключения отрезка на следующем шаге последовательного поиска.

Метод золотого сечения. Как известно, ***золотым сечением отрезка*** называют такое его деление на две неравные части, при котором отношение длины всего отрезка к длине его большей части равно равно отношению длины большей части к длине меньшей.

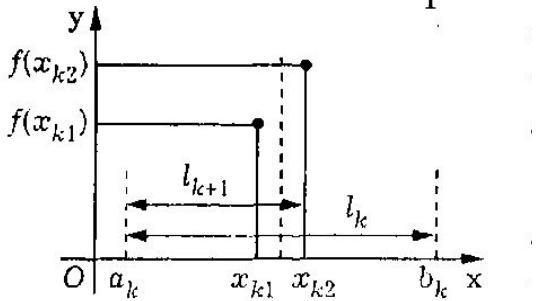
Термин „золотое сечение" ввел Леонардо да Винчи. Золотое сечение широко применяли при композиционном построении многих произведении мирового искусства, в том числе в античной архитектуре и в эпоху Возрождения.

Рассмотрим $к$-й шаг последовательного поиска. Чтобы выполнить процедуру исключения отрезка на этом шаге, отрезок $[a_к, b_k]$ необходимо двумя внутренними точками $x_{к1}$, x_{k2}, $x_{к1} < x_{k2}$, разделить на три части. Эти точки выберем симметрично относительно середины отрезка $[a_к, b_k]$ (рис. 2,8) и так, чтобы каждая из них производила золотое сечение отрезка $[a_к, b_k]$. В этом случае отрезок $[a_{к+1}, b_{k+1}]$ внутри будет содержать одну из точек $x_{к1}$, x_{k2} (другая будет одним из концов отрезка), причем эта точка будет производить золотое сечение отрезка $[a_{к+1}, b_{k+1}]$. Это вытекает из равенства длин отрезков $[a_к, x_{k1}]$ и $[x_{к2}, b_k]$. Таким образом, на $(к + 1)$-м шаге в одной из точек $x_{к+1,1}$ и $x_{к+1,2}$ значение функции вычислять не нужно. При этом отношение l_k/l_{k+1} длин отрезков сохраняется от шага к шагу, т.е.

$$\frac{l_k}{l_{k+1}} = \frac{l_{k+1}}{l_{k+2}} = r = \text{const.} \qquad (2.9)$$

Число r называют ***отношением золотого сечения.***

Последовательный поиск, в котором на k-м шаге каждая из симметрично выбранных на отрезке $[a_k, b_k]$ точек $x_{к1}$, x_{k2} осуществляет золотое сечение этого отрезка, называют ***методом золотого сечения.*** В этом методе каждое исключение отрезка уменьшает оставшийся отрезок в r раз.



Выясним, чему равно отношение золотого сечения. Так как точки $x_{к1}$ и $x_{к2}$, $x_{к1} < x_{к2}$, выбраны симметрично относительно середины отрезка $[a_к, b_к]$, то

$$b_k - x_{k2} = x_{k1} - a_k = l_k - l_{k-1}.$$

(см. рис. 2.8). Для определенности будем считать, что на k-м шаге выбран отрезок $[a_к, x_{к2}]$. Тогда на $(k + 1)$-м шаге одной из точек деления (а именно правой) будет точка $x_{к1}$. Значит, длина l_{k+2} отрезка, выбираемого на $(k + 1)$-м шаге, совпадает с длиной отрезка $[a_к, x_{к1}]$ и верно равенство $l_{k+2} = l_k - l_{k-1}$. Подставляя найденное выражение для l_{k+2} в уравнение (2.9), получаем

$$\frac{l_k}{l_{k+1}} = \frac{l_{k+1}}{l_k - l_{k+1}}$$

или $r = 1/(r - 1)$. Преобразуя это соотношение, приходим к квадратному уравнению $r^2 - r - 1 = 0$, имеющему единственное положительное решение

$$r = \frac{1 + \sqrt{5}}{2} \approx 1{,}618034.$$

Предположим, что отрезком минимизации унимодальной функции $f(x)$ является [0,1], т.е. $a_1 = 0$, $b_1 = 1$ и $l_1 = 1$. На первом шаге последовательного поиска ($k = 1$) на отрезке [0, 1] выбираем две точки $x_{11} = a_1 + (1 - 1/r)b_1 = 1 - 1/r$ и $x_{12} = a_1 + b_1/r = 1/r$, осуществляющие золотое сечение отрезка [0,1]. Вычисляем значения минимизируемой функции в этих точках и выполняем процедуру исключения отрезка. Если, $f(x_{11}) < f(x_{12})$, то выбираем отрезок $[a_1, x_{12}]$, т.е. полагаем $a_2 = a_1 = 0$, $b_2 = x_{12}$; в противном случае выбираем отрезок $[x_{11}, b_1]$, т.е. полагаем $a_2 = x_{11}$, $b_2 = b_1 = 1$. Кроме того, в первом случае принимаем $\tilde{x}_2 = x_{11}$, а во втором случае $\tilde{x}_2 = x_{12}$. Точка $\tilde{x}_2$ — одна из точек, осуществляющих золотое сечение отрезка $[a_2, b_2]$, меньшая в первом случае и большая во втором. Если длина вновь полученного отрезка больше заданной допустимой длины *интервала неопределенности*, то следует перейти ко второму шагу алгоритма, на котором одна из точек x_{21}, x_{22} есть точка $\tilde{x}_2$, а вторую можно найти, например, по формуле $a_2 + b_2 - \tilde{x}_2$. На втором шаге алгоритма вычисляем лишь одно значение функции в точке, симметричной относительно середины отрезка $[a_2, b_2]$. Если же длина l_2 отрезка $[a_2, b_2]$, полученного после первого шага алгоритма, оказалась меньше, то поиск прекращают и полагают $x_* \approx (a_2 + b_2)/2$.

Пусть на κ-м шаге, $\kappa \geq 2$, последовательного поиска по методу золотого сечения выбран отрезок $[a_k, b_k]$ и в нем точка $\tilde{x}_k$, осуществляющая золотое сечение этого отрезка. $f(\tilde{x}_k)$ Значение функции в этой точке уже вычислено на предыдущем шаге. Находим вторую точку $\hat{x}_k$ золотого сечения по формуле $\hat{x}_k = a_k + b_k - \tilde{x}_k$ и вычисляем в ней значение функции. Если $\hat{x}_k < \tilde{x}_k$, то $x_{k1} = \hat{x}_k$ и $x_{k1} = \hat{x}_k$, иначе $x_{k1} = \tilde{x}_k$ и $x_{k2} = \hat{x}_k$.

Пусть для определенности $\hat{x}_k < \tilde{x}_k$ (см. рис. 2.8) и $x_{k1} = \hat{x}_k$, $x_{k2} = \hat{x}_k$. Если $f(x_{k1}) < f(x_{k2})$, то выбираем отрезок $[a_k, x_{k2}]$, т.е. полагаем $a_{k+1}=a_k$, $b_{k+1}=x_{k2}$, $\tilde{x}_{k+1} = x_{k1}$, иначе выбираем отрезок $[x_{k1}, b_k]$, т.е. полагаем $a_{k+1}=x_{k1}$, $b_{k+1}=b_k$, $\tilde{x}_{k+1} = x_{k2}$. Длину l_{k+1} нового отрезка $[a_{k+1}, b_{k+1}]$ сравниваем с ε_* и принимаем решение, продолжать поиск (при $l_{k+1} \geq \varepsilon_*$) или нет (при $l_{k+1} < \varepsilon_*$). В случае прекращения поиска полагаем $x_* \approx a_k + b_k/2$.

Согласно описанию алгоритма, на первом шаге значение функции вычисляют в двух точках, а на каждом из последующих шагов вычисляют лишь одно значение функции. Поэтому после κ шагов алгоритма значение функции будет вычислено в $N = \kappa + 1$ точках. Поскольку после каждого шага интервал неопределенности уменьшается в r раз, то для длины l_{k+1} отрезка $[a_{k+1}, b_{k+1}]$ получаем $l_{k+1} = l_1/r^k = 1/r^k$, а зависимость l_N^z длины интервала неопределенности от количества N вычисленных значений функции выражается формулой

$$l_N^z = l_{k+1} = \frac{1}{r^k} = \frac{1}{r^{N-1}}. \qquad (2.10)$$

Алгоритмы методов золотого сечения и дихотомии аналогичны. Различие состоит лишь в том, что в методе дихотомии расстояние 2δ между внутренними точками x_{k1}, x_{k2} отрезка $[a_k, b_k]$ на каждом κ-м шаге остается неизменным, а в методе золотого сечения оно зависит от номера шага поиска и уменьшается с уменьшением длины $l\kappa$ отрезка по мере возрастания номера шага. Действительно, в методе золотого сечения на κ-м шаге поиска внутренними точками отрезка $[a_k, b_k]$ будут $x_{k1} = a_k + (1 - 1/r)l_k$ и $x_{k2} = a_k + l_k/r$, а расстояние между ними равно

$$x_{k2} - x_{k1} = (2/r - 1)l_k = (\sqrt{5} - 2)l_k \approx 0{,}236068\, l_k.$$

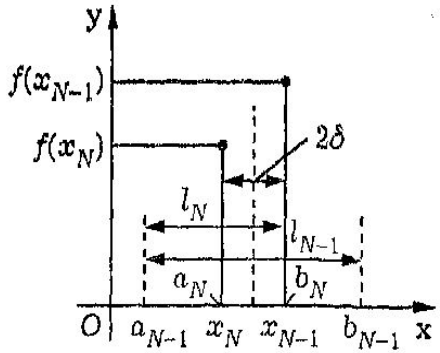
Метод Фибоначчи. Пусть при поиске точки $x_* \in [0, 1]$, в которой унимодальная на отрезке [0, 1] функция $f(x)$ принимает наименьшее на этом отрезке значение, можно вычислить ее значения только в двух точках. Тогда предпочтение следует отдать методу дихотомии при $\delta << 1$, так как он позволит уменьшить интервал неопределенности почти вдвое, а метод золотого сечения — лишь в $r \approx 1{,}618$ раз. Сравнение (2.8) и (2.10) показывает, что при количестве вычисляемых значений функции $N \geq 4$ эффективность метода золотого сечения становится выше, чем метода дихотомии.

Однако при любом заданном общем числе $N > 2$ вычисляемых значений функции можно построить еще более эффективный метод, состоящий из $N - 1$ шагов. Он сочетает преимущество симметричного расположения внутренних точек x_{k1}, x_{k2} на отрезке $[a_k, b_k]$ относительно его середины, реализованное в методах дихотомии и золотого сечения, с возможностью на каждом шаге изменять отношение l_k/l_{k+1} длин сокращаемого и нового отрезков. Как показано при обсуждении метода золотого сечения, в случае выбора внутренних точек симметрично относительно середины отрезка для трех последовательных шагов этого метода выполняется соотношение

$$l_{k-1} = l_k + l_{k+1}, \; \kappa = 2,3,... \qquad (2.11)$$

Построение алгоритма такого метода удобнее начать с последнего шага, но предварительно уточним задачу. Располагая возможностью вычислить в N точках $x_k \in [0,1], \quad k = \overline{1,N}$, значения унимодальной на отрезке функции $f(x)$, необходимо как можно точнее, т.е. с наименее возможной длиной интервала неопределенности, отыскать точку наименьшего значения этой функции на отрезке [0, 1].

При выполнении процедуры исключения отрезка на последнем, $(N - 1)$-м шаге имеем отрезок $[a_{N-1}, b_{N-1}]$ длины l_{N-1} с двумя внутренними точками x_{N-1} и x_N, симметрично расположенными относительно середины отрезка на достаточно малом расстоянии 2δ друг от друга (рис. 2.9). В этих точках вычислены значения $f(x_{N-1})$ и $f(x_N)$ функции $f(x)$. Пусть для определенности $f(x_N) < f(x_{N-1})$, тогда для нового отрезка $[a_N, b_N]$ длины $l_N = l_{N-1}/2 + \delta$ внутренней будет точка x_N, а точка x_{N-1} совпадет с одним из его концов.



В такой ситуации при выборе $= x_N$ длина интервала неопределенности равна пока неизвестной длине l_N отрезка $[a_N, b_N]$. Через l_N можно выразить длину $l_{N-1} = 2l_N —$ 2δ отрезка $[a_{N-1}, b_{N-1}]$. Далее в соответствии с (2.11) получаем

$l_{N-2} = l_{N-1} + l_N = 3l_N - 2\delta$, $l_{N-3} = l_{N-2} + l_{N-1} = 5l_N - 4\delta$

$l_{N-4} = l_{N-3} + l_{N-2} = 8l_N - 6\delta$, $l_{N-5} = l_{N-4} + l_{N-3} = 13l_N - 10\delta$

и в общем виде

$$l_{N-K} = F_{K+2} l_N - 2F_K \delta, \qquad K = \overline{0, N-1}, \tag{2.12}$$

где коэффициенты F_m определены рекуррентным соотношением

$$F_m = F_{m-1} + F_{m-2}, \quad m = \overline{3, N-1}, \quad F_1 = F_2 = 1. \tag{2.13}$$

Так как при $K = N — 1$ длина $l_{N-K} = l_1 = 1$ отрезка $[0,1]$ известна, то из (2.12) можно найти длину интервала неопределенности.

$$l_N^f = \frac{l_1}{F_{N+1}} + 2\delta \frac{F_{N-1}}{F_{N+1}}. \tag{2.14}$$

Суще
(2.14).

Все коэффициенты F_m принадлежат множеству натуральных чисел, и их называют ***числами Фибоначчи****. В табл. 2.1 представлены эти числа до номера $m = 25$.

Таблица 2.1

m	F_m	m	F_m	m	F_m	m	F_m	m	F_m
1	1	6	8	11	89	16	987	21	10946
2	1	7	13	12	144	17	1597	22	17711
3	2	8	21	13	233	18	2584	23	28657
4	3	9	34	14	377	19	4181	24	46368
5	5	10	55	15	610	20	6765	25	75025

Метод, использующий числа Фибоначчи для выбора длин отрезков l_k, а значит, и точек, в которых вычисляют значения минимизируемой функции, называют ***методом Фибоначчи*** (иногда — оптимальным последовательным поиском). Если на первом шаге поиска ($к = 1$, $K = N$ — 1) интервал неопределенности имеет длину l_1, то в соответствии с (2.12) и (2.14) длина l_2 нового отрезка $[a_1, b_2]$ равна

$$l_2 = F_N l_N - 2\delta F_{N-2} = \frac{F_N}{F_{N+1}} l_1 + 2\delta \frac{F_N F_{N-1} - F_N F_{N-2}}{F_{N+1}} = \frac{F_N}{F_{N+1}} l_1 + (-1)^{N+1} \frac{2\delta}{F_{N+1}}.$$

Опишем алгоритм метода, пренебрегая малой величиной δ, т.е. принимая

$$\frac{l_1}{l_2} = \frac{F_{N+1}}{F_N}. \qquad (2.15)$$

Несложно проверить, что в этом случае выполнение процедуры исключения отрезка на последнем, (N — 1)-м шаге поиска приводит к совпадению внутренних точек x_{N-1} и x_N (см. рис. 2.9).

Отметим, что уже при $N = 11$ имеем $F_{12}/F_{11} = 144/89 \approx 1{,}617978$, а при $N = 21$ получаем $F_{22}/F_{21} = 17711/10946 \approx 1{,}618034$, что совпадает с отношением золотого сечения с точностью до 10^{-6}. Таким образом, на первом шаге длина исходного отрезка уменьшается практически так же, как и в методе золотого сечения.

При $l_1 = 1$ из (2.15) находим $l_2 = F_N / F_{N+1}$. Таким образом, учитывая (2.13), заключаем, что на первом шаге выбор точек, симметричных относительно середины отрезка [0,1], можно определить по формулам

$$x_1 = l_2 = \frac{F_N}{F_{N+1}}, \qquad x_2 = 1 - l_2 = 1 - \frac{F_N}{F_{N+1}} = \frac{F_{N-1}}{F_{N+1}}, \quad x_2 < x_1,$$

причем расстояние между ними будет равно

$$d_1 = x_1 - x_2 = \frac{F_N}{F_{N+1}} - \frac{F_{N-1}}{F_{N+1}} = \frac{F_{N-2}}{F_{N+1}}$$

После выполнения на этом шаге процедуры исключения отрезка одна из точек x_1, x_2 будет граничной точкой нового отрезка $[a_2, a_1]$, а другая — его внутренней точкой, которую обозначим . Вторая внутренняя точка на этом отрезке должна быть выбрана симметрично точке относительно ее середины. Аналогично происходит выбор второй внутренней точки нового отрезка на всех последующих шагах поиска.

На $к$-м шаге в соответствии с равенством (2.12), в котором следует положить $K = N$ — $к$, и равенством (2.14) длина отрезка $[a_k, b_k]$ равна $l_k = F_{N+2-k} / F_{N+1}$ и происходит ее уменьшение в $l_k / l_{k+1} = F_{N+2-k} / F_{N+1-k}$ раз. Если внутренние точки на этом отрезке обозначить , то проведенные рассуждения позволяют написать

$$\alpha_k = \alpha_k + \frac{F_{N-k}}{F_{N+1}}, \quad \beta_k = \alpha_k + \frac{F_{N+1-k}}{F_{N+1}}, \quad \alpha_k < \beta_k, \qquad k = \overline{1, N-1}.$$

Подчеркне требуемого количества N вычисляемых значений функции (или количества шагов поиска). Этот параметр необходим для реализации первого шага алгоритма при выборе точек x_{11} и x_{12} деления отрезка $[a_1, b_1]$. Если параметр N по каким-либо причинам не может быть задан заранее, следует использовать другие методы, например дихотомии или золотого сечения.

2.5. СРАВНЕНИЕ МЕТОДОВ ПОСЛЕДОВАТЕЛЬНОГО ПОИСКА

В качестве оценки скорости сходимости методов прямого поиска можно использовать скорость убывания длины *интервала неопределенности* в зависимости от числа n вычисленных значений минимизируемой функции в различных методах.

Из полученных результатов видно, что скорость сходимости метода Фибоначчи при больших значениях n немного выше, чем скорость сходимости метода золотого сечения. Метод золотого сечения качественно “лучше” метода дихотомии. Но скорость сходимости метода дихотомии при больших значениях n выше, чем скорость сходимости метода *оптимального пассивного поиска.*

Итак, метод золотого сечения уступает по скорости сходимости лучшему методу — методу Фибоначчи — примерно в 1,17 раза, но является более гибким, поскольку не требует выбора заранее определенного числа точек, в которых предстоит вычислить значения минимизируемой функции.