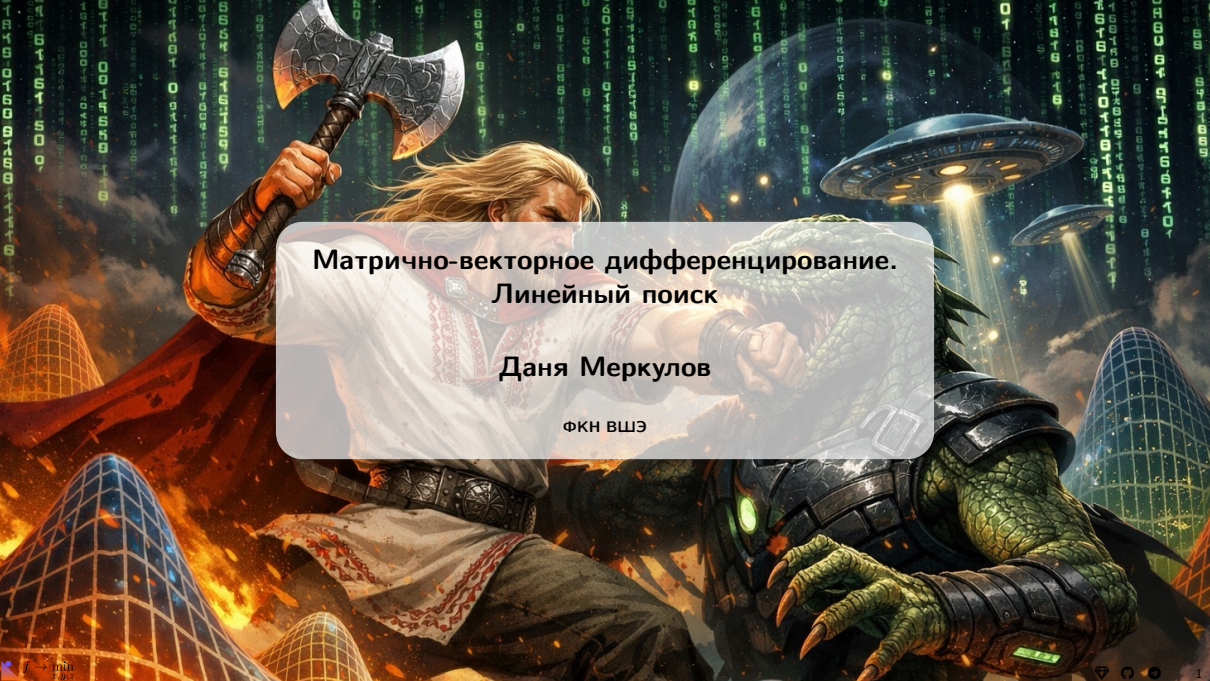


Матрично-векторное дифференцирование. Линейный поиск

Даня Меркулов

ФКН ВШЭ

Матрично-векторное дифференцирование

Градиент

Пусть $f(x) : \mathbb{R}^n \to \mathbb{R}$, тогда вектор, содержащий все частные производные первого порядка:

$$\nabla f(x) = \frac{df}{dx} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Градиент

Пусть $f(x) : \mathbb{R}^n \to \mathbb{R}$, тогда вектор, содержащий все частные производные первого порядка:

$$\nabla f(x) = \frac{df}{dx} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

называется градиентом функции $f(x)$. Этот вектор указывает направление наискорейшего возрастания. Таким образом, вектор $-\nabla f(x)$ указывает направление наискорейшего убывания функции в точке. Кроме того, вектор градиента всегда ортогонален линии уровня в точке.

Example

Для функции $f(x, y) = x^2 + y^2$ градиент равен:

$$\nabla f(x, y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

Он указывает направление наискорейшего возрастания функции.

Question

Как связана норма градиента с крутизной функции?

Гессиан

Пусть $f(x) : \mathbb{R}^n \to \mathbb{R}$, тогда матрица, содержащая все частные производные второго порядка:

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j=1}^{n} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

Гессиан

Пусть $f(x) : \mathbb{R}^n \to \mathbb{R}$, тогда матрица, содержащая все частные производные второго порядка:

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j=1}^{n} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

Если $f : \mathbb{R}^n \to \mathbb{R}^m$, то вторые производные образуют тензор третьего порядка. Его (k)-й срез - гессиан скалярной функции $f_k : \nabla^2 f_k(x)$.

Example

Для функции $f(x, y) = x^2 + y^2$ гессиан равен:

$$H_f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Эта матрица содержит информацию о кривизне функции в разных направлениях.

Question

Как можно использовать гессиан для определения выпуклости или вогнутости функции?

Теорема Шварца

Пусть есть функция $f : \mathbb{R}^n \to \mathbb{R}$. Если смешанные частные производные $\frac{\partial^2 f}{\partial x_i \partial x_j}$ и $\frac{\partial^2 f}{\partial x_j \partial x_i}$ непрерывны на открытом множестве, содержащем точку a, то они равны в ней. То есть,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

Теорема Шварца

Пусть есть функция $f : \mathbb{R}^n \to \mathbb{R}$. Если смешанные частные производные $\frac{\partial^2 f}{\partial x_i \partial x_j}$ и $\frac{\partial^2 f}{\partial x_j \partial x_i}$ непрерывны на открытом множестве, содержащем точку a, то они равны в ней. То есть,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

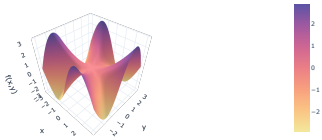
То есть, гессиан симметричен:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \quad \nabla^2 f(x) = (\nabla^2 f(x))^T$$

Эта симметричность упрощает вычисления и анализ, связанные с гессианом в различных приложениях, особенно в оптимизации.

Контрпример Шварца

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} & \text{для } (x,\, y) \neq (0,\, 0), \\ 0 & \text{для } (x,y) = (0,0). \end{cases}$$



Можно проверить, что $\frac{\partial^2 f}{\partial x \partial y}(0,0) \neq \frac{\partial^2 f}{\partial y \partial x}(0,0)$, хотя смешанные частные производные существуют, и в во всех остальных точках симметричность выполняется.

Якобиан

Обобщением понятия градиента на случай векторнозначной функции $f(x) : \mathbb{R}^n \to \mathbb{R}^m$ является следующая матрица:

$$J_f = f'(x) = \frac{df}{dx^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Она содержит информацию о скорости изменения функции по входным переменным.

Question

Можно ли связать эти три определения выше (градиент, якобиан и гессиан) с помощью одного утверждения?

Example

Для функции

$$f(x, y) = \begin{bmatrix} x + y \\ x - y \end{bmatrix},$$

Якобиан равен:

$$J_f(x, y) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Question

Как матрица Якоби связана с градиентом для скалярных функций?

Итог

$$f(x) : X \to Y; \qquad \frac{\partial f(x)}{\partial x} \in G$$

X	Y	G	Название
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$f'(x)$ (производная)
$\mathbb{R}^n$	$\mathbb{R}$	$\mathbb{R}^n$	$\left(\frac{\partial f}{\partial x_i}\right)_{i=1}^{n}$ (градиент)
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{n\times m}$	$\frac{\partial f_i}{\partial x_j}$ (якобиан)
$\mathbb{R}^{m\times n}$	$\mathbb{R}$	$\mathbb{R}^{m\times n}$	$\frac{\partial f}{\partial x_{ij}}$

Аппроксимация Тейлора первого порядка

Аппроксимация Тейлора первого порядка, также известная как линейное приближение, строится вблизи некоторой точки x_0. Если $f : \mathbb{R}^n \to \mathbb{R}$ - дифференцируемая функция, то ее аппроксимация первого порядка задается следующим образом:

$$f^I_{x_0}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0)$$

где:

- $f(x_0)$ - значение функции в точке x_0.

Аппроксимация Тейлора первого порядка

Аппроксимация Тейлора первого порядка, также известная как линейное приближение, строится вблизи некоторой точки x_0. Если $f : \mathbb{R}^n \to \mathbb{R}$ - дифференцируемая функция, то ее аппроксимация первого порядка задается следующим образом:

$$f^I_{x_0}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0)$$

где:

- $f(x_0)$ - значение функции в точке x_0.
- $\nabla f(x_0)$ - градиент функции в точке x_0.

Аппроксимация Тейлора первого порядка

Аппроксимация Тейлора первого порядка, также известная как линейное приближение, строится вблизи некоторой точки x_0. Если $f : \mathbb{R}^n \to \mathbb{R}$ - дифференцируемая функция, то ее аппроксимация первого порядка задается следующим образом:

$$f^I_{x_0}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0)$$

где:

- $f(x_0)$ - значение функции в точке x_0.
- $\nabla f(x_0)$ - градиент функции в точке x_0.

Аппроксимация Тейлора первого порядка

Аппроксимация Тейлора первого порядка, также известная как линейное приближение, строится вблизи некоторой точки x_0. Если $f : \mathbb{R}^n \to \mathbb{R}$ - дифференцируемая функция, то ее аппроксимация первого порядка задается следующим образом:

$$f^I_{x_0}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0)$$

где:

- $f(x_0)$ - значение функции в точке x_0.
- $\nabla f(x_0)$ - градиент функции в точке x_0.

Часто, чтобы упростить теоретический анализ, функцию в окрестности точки заменяют её линейной аппроксимацией.

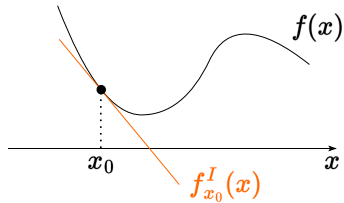


Figure 1: Аппроксимация Тейлора первого порядка в окрестности точки x_0

Аппроксимация Тейлора второго порядка

Аппроксимация Тейлора второго порядка, также известная как квадратичное приближение, использует информацию о кривизне функции. Для дважды дифференцируемой функции $f : \mathbb{R}^n \to \mathbb{R}$ квадратичная аппроксимация в окрестности x_0 имеет вид:

$$f_{x_0}^{II}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

Где $\nabla^2 f(x_0)$ - гессиан функции f в точке x_0.

Аппроксимация Тейлора второго порядка

Аппроксимация Тейлора второго порядка, также известная как квадратичное приближение, использует информацию о кривизне функции. Для дважды дифференцируемой функции $f : \mathbb{R}^n \to \mathbb{R}$ квадратичная аппроксимация в окрестности x_0 имеет вид:

$$f_{x_0}^{II}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

Где $\nabla^2 f(x_0)$ - гессиан функции f в точке x_0.
Когда линейного приближения функции недостаточно, можно рассмотреть замену $f(x)$ на $f_{x_0}^{II}(x)$ в окрестности точки x_0.

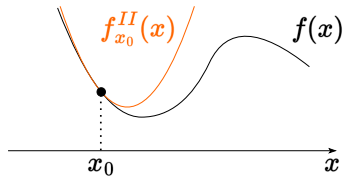


Figure 2: Аппроксимация Тейлора второго порядка в окрестности точки x_0

Дифференциалы

Theorem

Пусть $f : U \to V$, $x \in U$ - внутренняя точка. Пусть $D : U \to V$ - линейный оператор. Мы говорим, что функция f дифференцируема в точке x с производной D, если для всех достаточно малых $h, x + h \in U$ выполняется следующее разложение:

$$f(x + h) = f(x) + D[h] + o(\|h\|)$$

Если не существует линейного оператора D, удовлетворяющего этому разложению, то f не дифференцируема в точке x.

Дифференциалы

После получения дифференциальной записи df мы можем получить градиент, используя следующую формулу:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Дифференциалы

После получения дифференциальной записи df мы можем получить градиент, используя следующую формулу:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Далее, если у нас есть дифференциал в такой форме и мы хотим вычислить вторую производную матричной/векторной функции, мы фиксируем первый дифференциал $dx := dx_1$ (т.е. в вычислениях считаем его константой) и берём дифференциал ещё раз $d(df) = d^2 f(x)$

$$d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx \rangle = \langle H_f(x) dx_1, dx \rangle$$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-T}, dX\rangle$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-T}, dX\rangle$
- $d\left(\mathsf{tr}\ X\right) = \langle I, dX\rangle$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-T}, dX\rangle$
- $d\left(\mathsf{tr}\ X\right) = \langle I, dX\rangle$
- $df(g(x)) = \dfrac{df}{dg} \cdot dg(x)$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-T}, dX\rangle$
- $d\left(\mathsf{tr}\ X\right) = \langle I, dX\rangle$
- $df(g(x)) = \dfrac{df}{dg} \cdot dg(x)$
- $H = (J(\nabla f))^T$

Свойства дифференциалов

Пусть A и B - постоянные матрицы, а X и Y - переменные (или матричные функции).

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-T}, dX\rangle$
- $d\left(\mathsf{tr}\ X\right) = \langle I, dX\rangle$
- $df(g(x)) = \dfrac{df}{dg} \cdot dg(x)$
- $H = (J(\nabla f))^T$
- $d(X^{-1}) = -X^{-1}(dX)X^{-1}$

Матричное дифференцирование. Пример 1

Example

Найти $df, \nabla f(x)$, если $f(x) = \langle x, Ax \rangle - b^T x + c$.

Матричное дифференцирование. Пример 2

Example

Найти $df, \nabla f(x)$, если $f(x) = \ln\langle x, Ax\rangle$.

Матричное дифференцирование. Пример 2

Example

Найти $df, \nabla f(x)$, если $f(x) = \ln\langle x, Ax\rangle$.

1. Заметим, что A должна быть положительно определенной, потому что $\langle x, Ax\rangle$ аргумент логарифма и для любого x аргумент логарифма должен быть положительным. Таким образом, $A \in \mathbb{S}^n_{++}$. Сначала найдем дифференциал:

$$df = d\left(\ln\langle x, Ax\rangle\right) = \frac{d\left(\langle x, Ax\rangle\right)}{\langle x, Ax\rangle} = \frac{\langle dx, Ax\rangle + \langle x, d(Ax)\rangle}{\langle x, Ax\rangle} =$$
$$= \frac{\langle Ax, dx\rangle + \langle x, Adx\rangle}{\langle x, Ax\rangle} = \frac{\langle Ax, dx\rangle + \langle A^T x, dx\rangle}{\langle x, Ax\rangle} = \frac{\langle (A + A^T)x, dx\rangle}{\langle x, Ax\rangle}$$

Матричное дифференцирование. Пример 2

Example

Найти $df, \nabla f(x)$, если $f(x) = \ln\langle x, Ax\rangle$.

1. Заметим, что A должна быть положительно определенной, потому что $\langle x, Ax\rangle$ аргумент логарифма и для любого x аргумент логарифма должен быть положительным. Таким образом, $A \in \mathbb{S}^n_{++}$. Сначала найдем дифференциал:

$$df = d\left(\ln\langle x, Ax\rangle\right) = \frac{d\left(\langle x, Ax\rangle\right)}{\langle x, Ax\rangle} = \frac{\langle dx, Ax\rangle + \langle x, d(Ax)\rangle}{\langle x, Ax\rangle} =$$
$$= \frac{\langle Ax, dx\rangle + \langle x, Adx\rangle}{\langle x, Ax\rangle} = \frac{\langle Ax, dx\rangle + \langle A^T x, dx\rangle}{\langle x, Ax\rangle} = \frac{\langle (A + A^T)x, dx\rangle}{\langle x, Ax\rangle}$$

2. Наша основная цель - получить форму $df = \langle \cdot, dx\rangle$. Имея ввиду $A + A^T = 2A$, получаем:

$$df = \left\langle \frac{2Ax}{\langle x, Ax\rangle}, dx \right\rangle$$

Таким образом, градиент равен $\nabla f(x) = \dfrac{2Ax}{\langle x, Ax\rangle}$

Матричное дифференцирование. Пример 3

Example

Найти $df, \nabla f(X)$, если $f(X) = \langle S, X \rangle - \log\det X$.

Линейный поиск

Задача

Предположим, у нас есть задача минимизации функции $f(x) : \mathbb{R} \to \mathbb{R}$ одной переменной:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Задача

Предположим, у нас есть задача минимизации функции $f(x) : \mathbb{R} \to \mathbb{R}$ одной переменной:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Иногда мы рассматриваем похожую задачу поиска минимума функции на отрезке $[a, b]$:

$$f(x) \to \min_{x \in [a,b]}$$

Задача

Предположим, у нас есть задача минимизации функции $f(x) : \mathbb{R} \to \mathbb{R}$ одной переменной:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Иногда мы рассматриваем похожую задачу поиска минимума функции на отрезке $[a, b]$:

$$f(x) \to \min_{x \in [a,b]}$$

Example

Типичным примером задачи линейного поиска является выбор подходящего шага для алгоритма градиентного спуска:

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$
$$\alpha = \text{argmin } f(x_{k+1})$$

Задача

Предположим, у нас есть задача минимизации функции $f(x) : \mathbb{R} \to \mathbb{R}$ одной переменной:

$$f(x) \to \min_{x \in \mathbb{R}}$$

Иногда мы рассматриваем похожую задачу поиска минимума функции на отрезке $[a, b]$:

$$f(x) \to \min_{x \in [a,b]}$$

Example

Типичным примером задачи линейного поиска является выбор подходящего шага для алгоритма градиентного спуска:

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$
$$\alpha = \text{argmin } f(x_{k+1})$$

Линейный поиск - фундаментальный инструмент оптимизации, который используют для решения других задач оптимизации. Для упрощения предположим, что $f(x)$ *унимодальна*, то есть, неформально говоря, имеет единственную впадину.

Унимодальная функция

Definition

Функция $f(x)$ называется **унимодальной** на отрезке $[a, b]$, если существует такой $x^* \in [a, b]$, что $f(x_1) > f(x_2) \quad \forall a \leq x_1 < x_2 < x^*$ и $f(x_1) < f(x_2) \quad \forall x^* < x_1 < x_2 \leq b$

Унимодальная функция

Definition

Функция $f(x)$ называется **унимодальной** на отрезке $[a, b]$, если существует такой $x^* \in [a, b]$, что $f(x_1) > f(x_2) \quad \forall a \le x_1 < x_2 < x^*$ и $f(x_1) < f(x_2) \quad \forall x^* < x_1 < x_2 \le b$

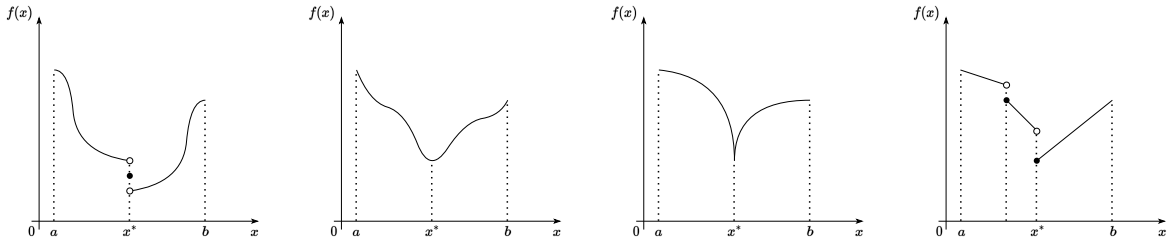


Figure 3: Примеры унимодальных функций

Ключевое свойство унимодальных функций

Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$

Ключевое свойство унимодальных функций

Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$
- Если $f(x_1) \geq f(x_2)$, то $x^* \in [x_1, b]$

Ключевое свойство унимодальных функций

Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$
- Если $f(x_1) \geq f(x_2)$, то $x^* \in [x_1, b]$

Ключевое свойство унимодальных функций

Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$
- Если $f(x_1) \geq f(x_2)$, то $x^* \in [x_1, b]$

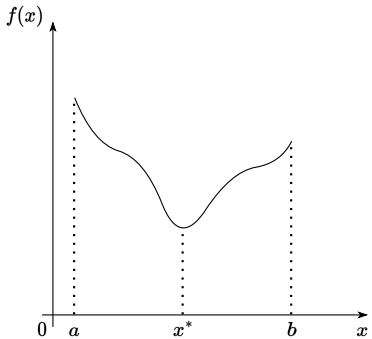
Доказательство Докажем первое утверждение. Предположим, что $f(x_1) \leq f(x_2)$, но $x^* > x_2$. Тогда, поскольку $x_1 < x_2 < x^*$, из определения унимодальности функции $f(x)$ следует, что должно выполняться неравенство $f(x_1) > f(x_2)$. Мы получили противоречие.

Ключевое свойство унимодальных функций

Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$
- Если $f(x_1) \geq f(x_2)$, то $x^* \in [x_1, b]$

Доказательство Докажем первое утверждение. Предположим, что $f(x_1) \leq f(x_2)$, но $x^* > x_2$. Тогда, поскольку $x_1 < x_2 < x^*$, из определения унимодальности функции $f(x)$ следует, что должно выполняться неравенство $f(x_1) > f(x_2)$. Мы получили противоречие.

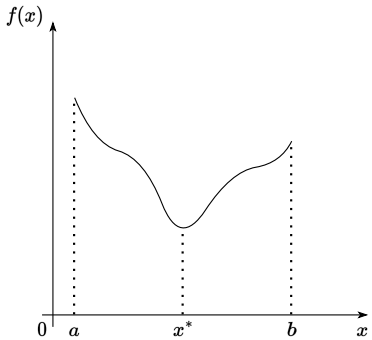


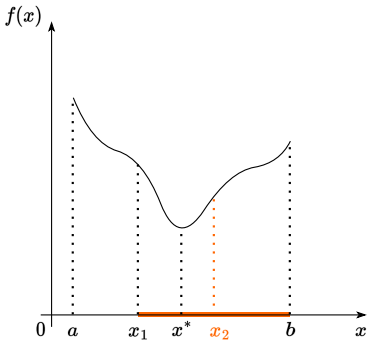
Ключевое свойство унимодальных функций

Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$
- Если $f(x_1) \geq f(x_2)$, то $x^* \in [x_1, b]$

Доказательство Докажем первое утверждение. Предположим, что $f(x_1) \leq f(x_2)$, но $x^* > x_2$. Тогда, поскольку $x_1 < x_2 < x^*$, из определения унимодальности функции $f(x)$ следует, что должно выполняться неравенство $f(x_1) > f(x_2)$. Мы получили противоречие.



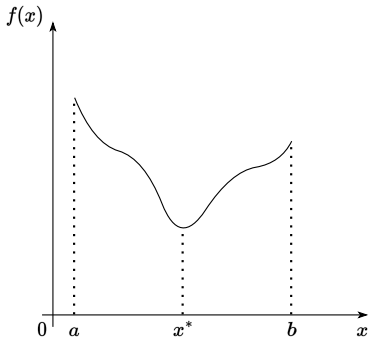


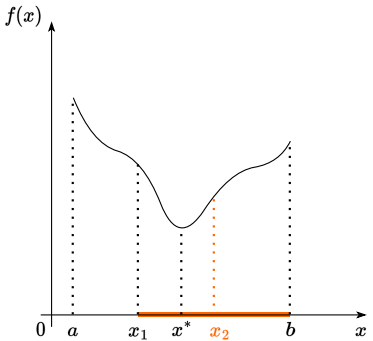
Ключевое свойство унимодальных функций

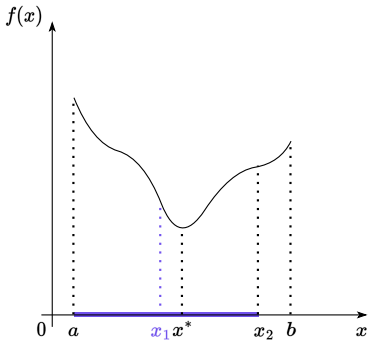
Пусть $f(x)$ является унимодальной функцией на отрезке $[a, b]$. Тогда если $x_1 < x_2 \in [a, b]$, то:

- Если $f(x_1) \leq f(x_2)$, то $x^* \in [a, x_2]$
- Если $f(x_1) \geq f(x_2)$, то $x^* \in [x_1, b]$

Доказательство Докажем первое утверждение. Предположим, что $f(x_1) \leq f(x_2)$, но $x^* > x_2$. Тогда, поскольку $x_1 < x_2 < x^*$, из определения унимодальности функции $f(x)$ следует, что должно выполняться неравенство $f(x_1) > f(x_2)$. Мы получили противоречие.







Метод дихотомии

Мы хотим решить следующую задачу:

$$f(x) \to \min_{x \in [a,b]}$$

Делим отрезок на две равные части и выбираем ту, которая содержит решение задачи, основываясь на ключевом свойстве, описанном выше.
Наша цель после одной итерации метода - локализовать решение в отрезке в два раза меньшей длины.

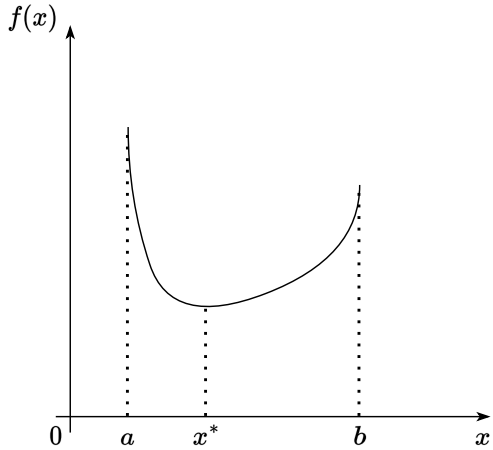


Figure 4: Метод дихотомии для унимодальной функции

Метод дихотомии

Вычисляем значение функции в середине отрезка

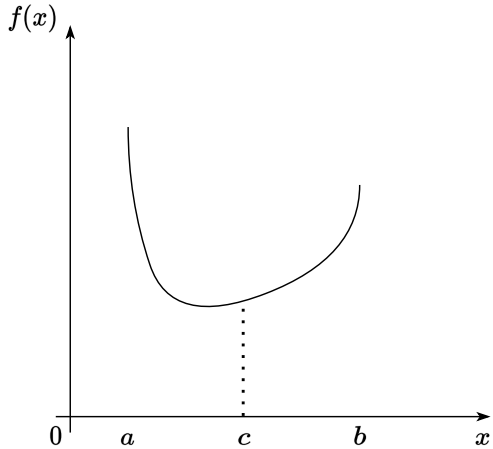


Figure 5: Метод дихотомии для унимодальной функции

Метод дихотомии

Чтобы применить ключевое свойство, мы выполняем еще одно вычисление значения функции.

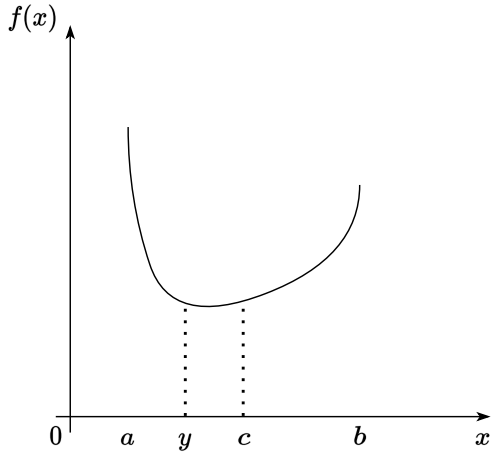


Figure 6: Метод дихотомии для унимодальной функции

Метод дихотомии

Выбираем целевой отрезок. В случае на изображении нас все устраивает, потому что новый отрезок локализации решения является половиной исходного. Так происходит не всегда.

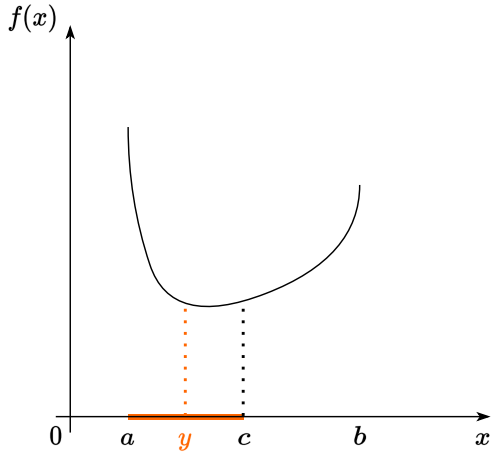


Figure 7: Метод дихотомии для унимодальной функции

Метод дихотомии

Рассмотрим другую унимодальную функцию.

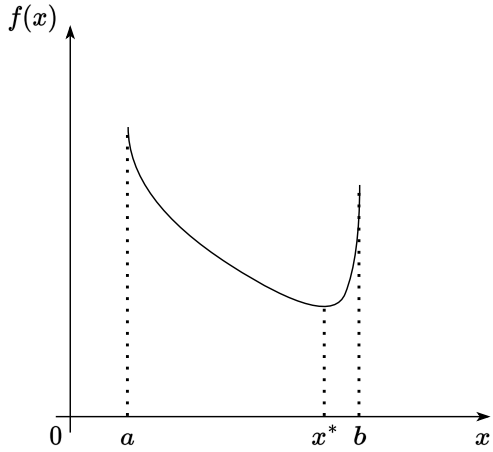


Figure 8: Метод дихотомии для унимодальной функции

Метод дихотомии

Вычисляем значение функции в середине отрезка.

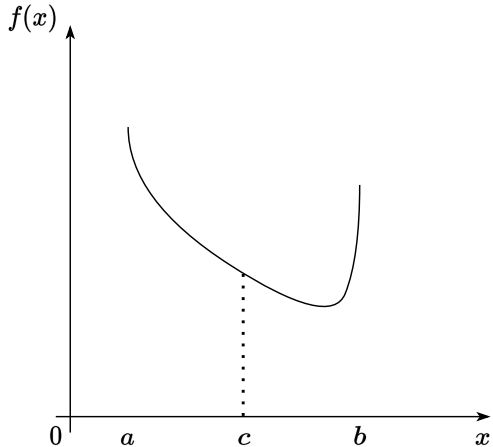


Figure 9: Метод дихотомии для унимодальной функции

Метод дихотомии

Делаем еще одно вычисление значения функции.

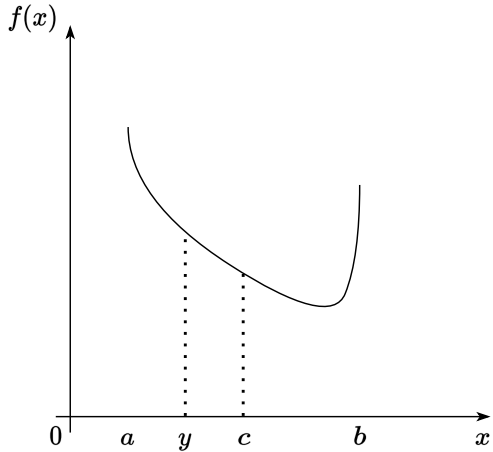


Figure 10: Метод дихотомии для унимодальной функции

Метод дихотомии

Выбираем целевой отрезок. Легко видеть, что полученный отрезок не является половиной исходного. Его длина равна $\frac{3}{4}(b-a)$. Чтобы исправить это, нам нужен еще один шаг алгоритма.

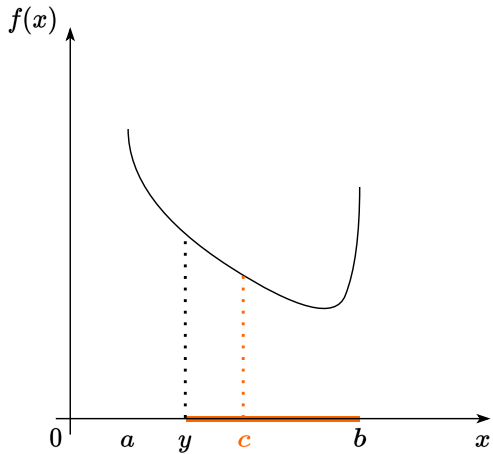


Figure 11: Метод дихотомии для унимодальной функции

Метод дихотомии

После дополнительного вычисления значения функции мы точно получим $\frac{2}{3}\frac{3}{4}(b-a) = \frac{1}{2}(b-a)$

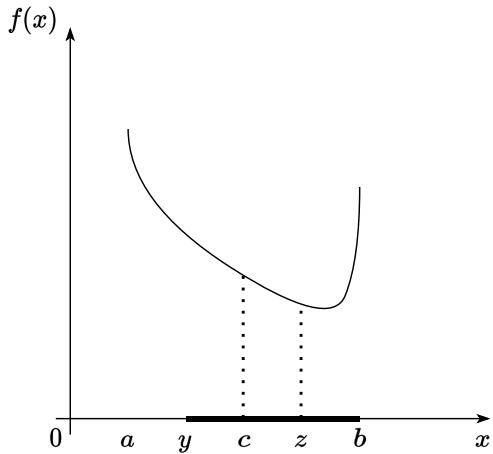


Figure 12: Метод дихотомии для унимодальной функции

Метод дихотомии

В итоге, каждая последующая итерация требует не более двух вычислений значения функции.

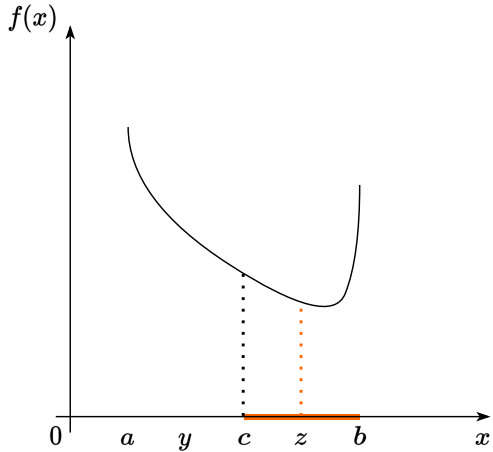


Figure 13: Метод дихотомии для унимодальной функции

Метод дихотомии. Алгоритм

```
def bisection_search(f, a, b, epsilon):
    c = (a + b) / 2.0
    fc = f(c)
    while (b - a) > epsilon:
        y = (a + c) / 2.0
        fy = f(y)
        if fy <= fc:        # минимум точно в [a, c]
            b = c
            c = y
            fc = fy
            continue
        z = (b + c) / 2.0
        fz = f(z)
        if fc <= fz:        # минимум в [y, z]
            a = y
            b = z
        else:               # минимум в [c, b]
            a = c
            c = z
            fc = fz
    return c
```

Метод дихотомии. Оценка

Длина отрезка на k-й итерации:

$$\Delta_k = b_k - a_k = \frac{1}{2^k}(b - a)$$

Метод дихотомии. Оценка

Длина отрезка на k-й итерации:

$$\Delta_k = b_k - a_k = \frac{1}{2^k}(b - a)$$

Для унимодальных функций это верно, если мы выбираем середину отрезка в качестве выхода итерации x_k:

$$|x_k - x_*| \leq \frac{\Delta_k}{2} \leq \frac{1}{2^{k+1}}(b - a) \leq (0.5)^k \cdot \frac{b - a}{2}$$

Метод дихотомии. Оценка

Длина отрезка на k-й итерации:

$$\Delta_k = b_k - a_k = \frac{1}{2^k}(b - a)$$

Для унимодальных функций это верно, если мы выбираем середину отрезка в качестве выхода итерации x_k:

$$|x_k - x_*| \leq \frac{\Delta_k}{2} \leq \frac{1}{2^{k+1}}(b - a) \leq (0.5)^k \cdot \frac{b - a}{2}$$

Заметим, что на каждой итерации мы обращаемся к оракулу (вычисляем значение функции) не более двух раз, поэтому количество вызовов функции равно $N = 2 \cdot k$, что означает:

$$|x_k - x_*| \leq (0.5)^{\frac{N}{2}} \cdot \frac{b - a}{2} \leq (0.707)^N \frac{b - a}{2}$$

Метод дихотомии. Оценка

Длина отрезка на k-й итерации:

$$\Delta_k = b_k - a_k = \frac{1}{2^k}(b-a)$$

Для унимодальных функций это верно, если мы выбираем середину отрезка в качестве выхода итерации x_k:

$$|x_k - x_*| \leq \frac{\Delta_k}{2} \leq \frac{1}{2^{k+1}}(b-a) \leq (0.5)^k \cdot \frac{b-a}{2}$$

Заметим, что на каждой итерации мы обращаемся к оракулу (вычисляем значение функции) не более двух раз, поэтому количество вызовов функции равно $N = 2 \cdot k$, что означает:

$$|x_k - x_*| \leq (0.5)^{\frac{N}{2}} \cdot \frac{b-a}{2} \leq (0.707)^N \frac{b-a}{2}$$

Обозначив правую часть последнего неравенства за ε, мы получаем количество итераций метода, необходимое для достижения точности ε:

$$K = \left\lceil \log_2 \frac{b-a}{\varepsilon} - 1 \right\rceil$$

Метод золотого сечения

Идея очень похожа на метод дихотомии. На отрезке выбираются две точки - левая и правая точки золотого сечения. Ключевая идея метода заключается в том, что на следующей итерации одна из точек останется точкой золотого сечения.

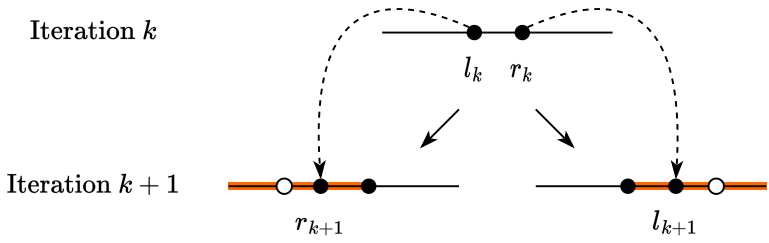


Figure 14: Идея, позволяющая уменьшить количество вызовов функции

Метод золотого сечения. Алгоритм

```
def golden_search(f, a, b, epsilon):
   tau = (sqrt(5) + 1) / 2
   y = a + (b - a) / tau**2
   z = a + (b - a) / tau
   fy = f(y)
   fz = f(z)
   while b - a > epsilon:
      if fy <= fz:
         b = z
         z, fz = y, fy
         y = a + (b - a) / (tau * tau)
         fy = f(y)          # 1 новый вызов
      else:
         a = y
         y, fy = z, fz
         z = a + (b - a) / tau
         fz = f(z)          # 1 новый вызов
   return (a + b) / 2.0
```

Метод золотого сечения. Оценка

$$|x_k - x_*| \leq \frac{b_k - a_k}{2} = \left(\frac{1}{\tau}\right)^N \frac{b-a}{2} \approx 0.618^k \frac{b-a}{2}$$

где $\tau = \frac{\sqrt{5}+1}{2}$.

- Знаменатель геометрической прогрессии для метода золотого сечения **больше**, чем для метода дихотомии: $0.618 > 0.5$.

Метод золотого сечения. Оценка

$$|x_k - x_*| \leq \frac{b_k - a_k}{2} = \left(\frac{1}{\tau}\right)^N \frac{b-a}{2} \approx 0.618^k \frac{b-a}{2}$$

где $\tau = \frac{\sqrt{5}+1}{2}$.

- Знаменатель геометрической прогрессии для метода золотого сечения **больше**, чем для метода дихотомии: $0.618 > 0.5$.
- Количество вызовов функции **меньше** для метода золотого сечения, чем для метода дихотомии: 0.707 больше (значит медленнее), чем 0.618. Для каждой итерации метода дихотомии (кроме первой), функция вызывается не более двух раз, в то время как для метода золотого сечения, она вызывается не более одного раза за итерацию.

Метод параболической интерполяции

Три точки, не лежащие на одной прямой, однозначно определяют параболу, проходящую через них. Идея метода — аппроксимировать функцию такой параболой и в качестве следующего приближения взять точку её минимума. Предположим, у нас есть три точки $x_1 < x_2 < x_3$, такие что отрезок $[x_1, x_3]$ содержит минимум функции $f(x)$. Тогда мы должны решить следующую систему уравнений:

Метод параболической интерполяции

Три точки, не лежащие на одной прямой, однозначно определяют параболу, проходящую через них. Идея метода — аппроксимировать функцию такой параболой и в качестве следующего приближения взять точку её минимума. Предположим, у нас есть три точки $x_1 < x_2 < x_3$, такие что отрезок $[x_1, x_3]$ содержит минимум функции $f(x)$. Тогда мы должны решить следующую систему уравнений:

$$ax_i^2 + bx_i + c = f_i = f(x_i), i = 1, 2, 3$$

Заметим, что эта система линейна, мы должны решить ее относительно a, b, c. Минимум этой параболы вычисляется по формуле:

Метод параболической интерполяции

Три точки, не лежащие на одной прямой, однозначно определяют параболу, проходящую через них. Идея метода — аппроксимировать функцию такой параболой и в качестве следующего приближения взять точку её минимума. Предположим, у нас есть три точки $x_1 < x_2 < x_3$, такие что отрезок $[x_1, x_3]$ содержит минимум функции $f(x)$. Тогда мы должны решить следующую систему уравнений:

$$ax_i^2 + bx_i + c = f_i = f(x_i), i = 1, 2, 3$$

Заметим, что эта система линейна, мы должны решить ее относительно a, b, c. Минимум этой параболы вычисляется по формуле:

$$u = -\frac{b}{2a} = x_2 - \frac{(x_2 - x_1)^2(f_2 - f_3) - (x_2 - x_3)^2(f_2 - f_1)}{2\left[(x_2 - x_1)(f_2 - f_3) - (x_2 - x_3)(f_2 - f_1)\right]}$$

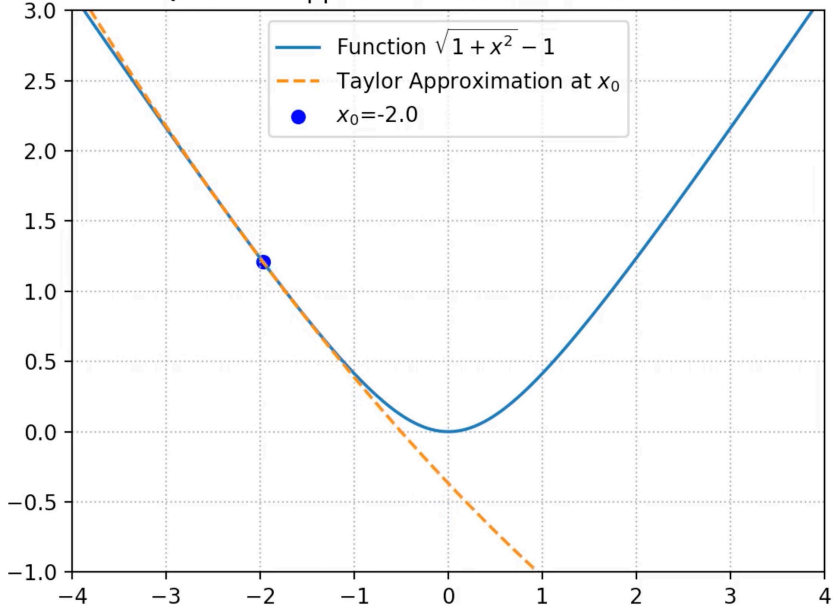
Заметим, что если $f_2 < f_1, f_2 < f_3$, то u будет лежать в $[x_1, x_3]$

Метод параболической интерполяции. Алгоритм [1]

```
def parabola_search(f, x1, x2, x3, epsilon):
   f1, f2, f3 = f(x1), f(x2), f(x3)
   while x3 - x1 > epsilon:
      u = x2 - ((x2 - x1)**2*(f2 - f3) - (x2 - x3)**2*(f2 - f1))/(2*((x2 - x1)*(f2 - f3) - (x2 - x3)*(f2 - f1)))
      fu = f(u)

      if x2 <= u:
         if f2 <= fu:
            x1, x2, x3 = x1, x2, u
            f1, f2, f3 = f1, f2, fu
         else:
            x1, x2, x3 = x2, u, x3
            f1, f2, f3 = f2, fu, f3
      else:
         if fu <= f2:
            x1, x2, x3 = x1, u, x2
            f1, f2, f3 = f1, fu, f2
         else:
            x1, x2, x3 = u, x2, x3
            f1, f2, f3 = fu, f2, f3
   return (x1 + x3)/2
```

[1] Сходимость метода локально сверхлинейная, что означает, что мы можем получить выгоду от использования этого метода только в некоторой окрестности оптимума. *Здесь* доказательство сверхлинейной сходимости порядка 1.32.


Quadratic approximation becomes inaccurate
Function $\sqrt{1+x^2}-1$
Taylor Approximation at x_0
x_0=-2.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−4
−3
−2
−1
0
1
2
3
4

Неточный линейный поиск

Нам не всегда нужно точно решать задачу минимизации. Иногда достаточно найти приближенное решение. Это часто встречается при выборе шага в методах оптимизации.

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$
$$\alpha = \arg\min_{\alpha \geq 0} f(x_{k+1})$$

Неточный линейный поиск

Нам не всегда нужно точно решать задачу минимизации. Иногда достаточно найти приближенное решение. Это часто встречается при выборе шага в методах оптимизации.

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$
$$\alpha = \arg\min_{\alpha \geq 0} f(x_{k+1})$$

Рассмотрим скалярную функцию $\phi(\alpha)$ в точке x_k:

$$\phi(\alpha) = f(x_k - \alpha \nabla f(x_k)), \alpha \geq 0$$

Неточный линейный поиск

Нам не всегда нужно точно решать задачу минимизации. Иногда достаточно найти приближенное решение. Это часто встречается при выборе шага в методах оптимизации.

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$
$$\alpha = \arg\min_{\alpha \geq 0} f(x_{k+1})$$

Рассмотрим скалярную функцию $\phi(\alpha)$ в точке x_k:

$$\phi(\alpha) = f(x_k - \alpha \nabla f(x_k)), \alpha \geq 0$$

Первое приближение $\phi(\alpha)$ в окрестности $\alpha = 0$ равно:

$$\phi(\alpha) \approx \phi_0^I(\alpha) = f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k)$$

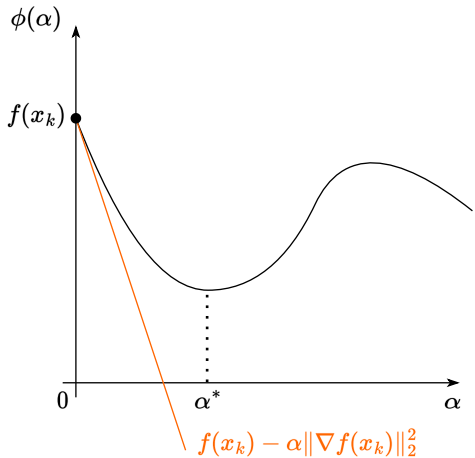


Figure 15: Иллюстрация аппроксимации Тейлора $\phi_0^I(\alpha)$

Неточный линейный поиск. Условие достаточного убывания

Условие неточного линейного поиска, известное как условие Армихо, требует, чтобы α обеспечивало достаточное убывание функции f:

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - c_1 \cdot \alpha \nabla f(x_k)^T \nabla f(x_k)$$

Неточный линейный поиск. Условие достаточного убывания

Условие неточного линейного поиска, известное как условие Армихо, требует, чтобы α обеспечивало достаточное убывание функции f:

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - c_1 \cdot \alpha \nabla f(x_k)^T \nabla f(x_k)$$

для некоторой постоянной $c_1 \in (0, 1)$. Заметим, что установка $c_1 = 1$ соответствует первому приближению Тейлора $\phi(\alpha)$.

Однако этому условию могут соответствовать очень малые значения α, потенциально замедляющие процесс решения. Обычно на практике используется $c_1 \approx 10^{-4}$.

Неточный линейный поиск. Условие достаточного убывания

Условие неточного линейного поиска, известное как условие Армихо, требует, чтобы α обеспечивало достаточное убывание функции f:

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - c_1 \cdot \alpha \nabla f(x_k)^T \nabla f(x_k)$$

для некоторой постоянной $c_1 \in (0, 1)$. Заметим, что установка $c_1 = 1$ соответствует первому приближению Тейлора $\phi(\alpha)$.

Однако этому условию могут соответствовать очень малые значения α, потенциально замедляющие процесс решения. Обычно на практике используется $c_1 \approx 10^{-4}$.

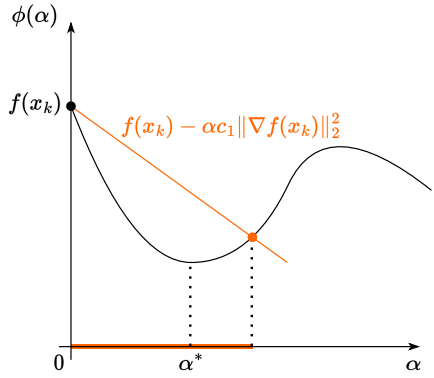


Figure 16: Иллюстрация условия достаточного убывания с коэффициентом c_1

Example

На практике выбор подходящего значения c_1 может быть очень важным. Например, в задачах машинного обучения неправильное значение c_1 может привести к очень медленной сходимости или пропуску минимума.

Неточный линейный поиск. Условия Гольдштейна

Рассмотрим две линейные скалярные функции $\phi_1(\alpha)$ и $\phi_2(\alpha)$:

$$\phi_1(\alpha) = f(x_k) - c_1\alpha\|\nabla f(x_k)\|^2$$

$$\phi_2(\alpha) = f(x_k) - c_2\alpha\|\nabla f(x_k)\|^2$$

Неточный линейный поиск. Условия Гольдштейна

Рассмотрим две линейные скалярные функции $\phi_1(\alpha)$ и $\phi_2(\alpha)$:

$$\phi_1(\alpha) = f(x_k) - c_1\alpha\|\nabla f(x_k)\|^2$$

$$\phi_2(\alpha) = f(x_k) - c_2\alpha\|\nabla f(x_k)\|^2$$

Условия Гольдштейна-Армихо требуют, чтобы функция $\phi(\alpha)$ лежала между $\phi_1(\alpha)$ и $\phi_2(\alpha)$. Обычно $c_1 = \rho$ и $c_2 = 1 - \rho$, с $\rho \in (0, 0.5)$.

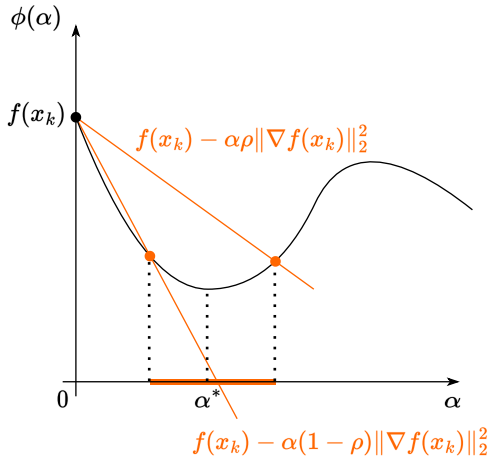


Figure 17: Иллюстрация условий Гольдштейна

Неточный линейный поиск. Условие ограничения на кривизну

Чтобы избежать слишком коротких шагов, вводится дополнительное ограничение:

$$-\nabla f(x_k - \alpha \nabla f(x_k))^T \nabla f(x_k) \geq c_2 \nabla f(x_k)^T (-\nabla f(x_k))$$

Неточный линейный поиск. Условие ограничения на кривизну

Чтобы избежать слишком коротких шагов, вводится дополнительное ограничение:

$$-\nabla f(x_k - \alpha \nabla f(x_k))^T \nabla f(x_k) \geq c_2 \nabla f(x_k)^T (-\nabla f(x_k))$$

для некоторого $c_2 \in (c_1, 1)$. Здесь c_1 из условия Армихо.

Левая часть является производной $\nabla_\alpha \phi(\alpha)$, гарантирующей, что наклон $\phi(\alpha)$ в целевой точке не менее чем в c_2 раз больше начального наклона $\nabla_\alpha \phi(\alpha)(0)$.

Обычно для методов Ньютона и квазиньютоновских методов используется $c_2 \approx 0.9$. Вместе условие достаточного убывания и ограничение на кривизну образуют условия Вульфа.

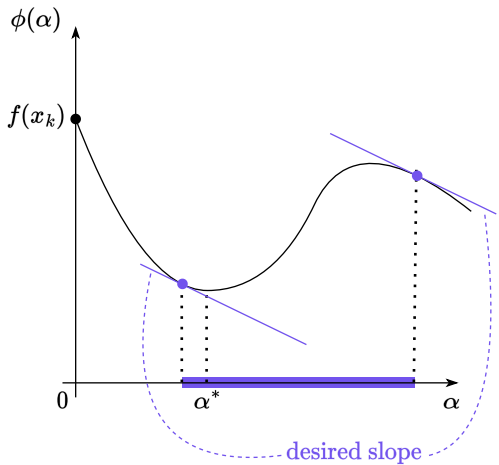


Figure 18: Иллюстрация условия ограничения на кривизну

Неточный линейный поиск. Условия Вульфа

$$-\nabla f(x_k - \alpha \nabla f(x_k))^T \nabla f(x_k) \geq c_2 \nabla f(x_k)^T (-\nabla f(x_k))$$

Вместе, условие достаточного убывания и ограничение на кривизну образуют условия Вульфа.

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ непрерывно дифференцируема, и пусть $\phi(\alpha) = f(x_k - \alpha \nabla f(x_k))$. Предположим, что $\nabla f(x_k)^T p_k < 0$, где $p_k = -\nabla f(x_k)$ - направление спуска. Также предположим, что f ограничена снизу вдоль луча $\{x_k + \alpha p_k \mid \alpha > 0\}$. Мы хотим показать, что для $0 < c_1 < c_2 < 1$, существуют интервалы шагов, удовлетворяющие условиям Вульфа.

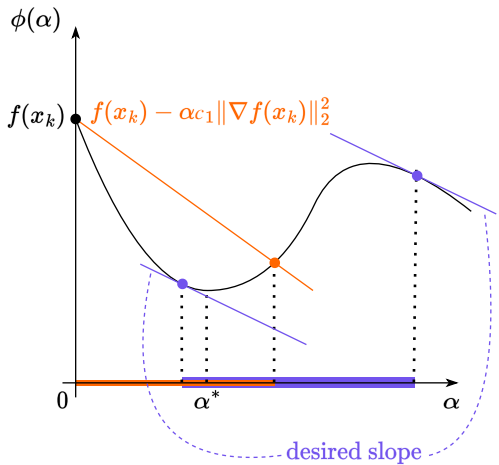


Figure 19: Иллюстрация условий Вульфа

Неточный линейный поиск. Условия Вульфа. Доказательство

1. Поскольку $\phi(\alpha) = f(x_k + \alpha p_k)$ ограничена снизу и $l(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^T p_k$ неограничена снизу (как $\nabla f(x_k)^T p_k < 0$), график $l(\alpha)$ должен пересекать график $\phi(\alpha)$ по крайней мере один раз. Пусть $\alpha' > 0$ будет наименьшим таким значением, удовлетворяющим:

$$f(x_k + \alpha' p_k) \leq f(x_k) + \alpha' c_1 \nabla f(x_k)^T p_k. \qquad (1)$$

Это гарантирует выполнение **условия достаточного убывания**.

Неточный линейный поиск. Условия Вульфа. Доказательство

1. Поскольку $\phi(\alpha) = f(x_k + \alpha p_k)$ ограничена снизу и $l(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^T p_k$ неограничена снизу (как $\nabla f(x_k)^T p_k < 0$), график $l(\alpha)$ должен пересекать график $\phi(\alpha)$ по крайней мере один раз. Пусть $\alpha' > 0$ будет наименьшим таким значением, удовлетворяющим:

$$f(x_k + \alpha' p_k) \leq f(x_k) + \alpha' c_1 \nabla f(x_k)^T p_k. \quad (1)$$

   Это гарантирует выполнение **условия достаточного убывания**.
2. По теореме о среднем значении, существует $\alpha'' \in (0, \alpha')$ такое, что:

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k. \quad (2)$$

   Подставляя $f(x_k + \alpha' p_k)$ из (1) в (2), мы получаем:

$$\alpha' \nabla f(x_k + \alpha'' p_k)^T p_k \leq \alpha' c_1 \nabla f(x_k)^T p_k.$$

Неточный линейный поиск. Условия Вульфа. Доказательство

1. Поскольку $\phi(\alpha) = f(x_k + \alpha p_k)$ ограничена снизу и $l(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^T p_k$ неограничена снизу (как $\nabla f(x_k)^T p_k < 0$), график $l(\alpha)$ должен пересекать график $\phi(\alpha)$ по крайней мере один раз. Пусть $\alpha' > 0$ будет наименьшим таким значением, удовлетворяющим:

$$f(x_k + \alpha' p_k) \leq f(x_k) + \alpha' c_1 \nabla f(x_k)^T p_k. \quad (1)$$

   Это гарантирует выполнение **условия достаточного убывания**.
2. По теореме о среднем значении, существует $\alpha'' \in (0, \alpha')$ такое, что:

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k. \quad (2)$$

   Подставляя $f(x_k + \alpha' p_k)$ из (1) в (2), мы получаем:

$$\alpha' \nabla f(x_k + \alpha'' p_k)^T p_k \leq \alpha' c_1 \nabla f(x_k)^T p_k.$$

Неточный линейный поиск. Условия Вульфа. Доказательство

1. Поскольку $\phi(\alpha) = f(x_k + \alpha p_k)$ ограничена снизу и $l(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^T p_k$ неограничена снизу (как $\nabla f(x_k)^T p_k < 0$), график $l(\alpha)$ должен пересекать график $\phi(\alpha)$ по крайней мере один раз. Пусть $\alpha' > 0$ будет наименьшим таким значением, удовлетворяющим:

$$f(x_k + \alpha' p_k) \leq f(x_k) + \alpha' c_1 \nabla f(x_k)^T p_k. \quad (1)$$

Это гарантирует выполнение **условия достаточного убывания**.

2. По теореме о среднем значении, существует $\alpha'' \in (0, \alpha')$ такое, что:

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k. \quad (2)$$

Подставляя $f(x_k + \alpha' p_k)$ из (1) в (2), мы получаем:

$$\alpha' \nabla f(x_k + \alpha'' p_k)^T p_k \leq \alpha' c_1 \nabla f(x_k)^T p_k.$$

Делим на $\alpha' > 0$, получаем:

$$\nabla f(x_k + \alpha'' p_k)^T p_k \leq c_1 \nabla f(x_k)^T p_k. \quad (3)$$

3. Поскольку $c_1 < c_2$ и $\nabla f(x_k)^T p_k < 0$, неравенство $c_1 \nabla f(x_k)^T p_k < c_2 \nabla f(x_k)^T p_k$ выполняется. Это означает, что существует α'' такое, что:

$$\nabla f(x_k + \alpha'' p_k)^T p_k \leq c_2 \nabla f(x_k)^T p_k. \quad (4)$$

Неравенства (3) и (4) вместе гарантируют выполнение условий Вульфа.

Неточный линейный поиск. Условия Вульфа. Доказательство

1. Поскольку $\phi(\alpha) = f(x_k + \alpha p_k)$ ограничена снизу и $l(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^T p_k$ неограничена снизу (как $\nabla f(x_k)^T p_k < 0$), график $l(\alpha)$ должен пересекать график $\phi(\alpha)$ по крайней мере один раз. Пусть $\alpha' > 0$ будет наименьшим таким значением, удовлетворяющим:

$$f(x_k + \alpha' p_k) \leq f(x_k) + \alpha' c_1 \nabla f(x_k)^T p_k. \quad (1)$$

Это гарантирует выполнение **условия достаточного убывания**.

2. По теореме о среднем значении, существует $\alpha'' \in (0, \alpha')$ такое, что:

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k. \quad (2)$$

Подставляя $f(x_k + \alpha' p_k)$ из (1) в (2), мы получаем:

$$\alpha' \nabla f(x_k + \alpha'' p_k)^T p_k \leq \alpha' c_1 \nabla f(x_k)^T p_k.$$

Делим на $\alpha' > 0$, получаем:

$$\nabla f(x_k + \alpha'' p_k)^T p_k \leq c_1 \nabla f(x_k)^T p_k. \quad (3)$$

3. Поскольку $c_1 < c_2$ и $\nabla f(x_k)^T p_k < 0$, неравенство $c_1 \nabla f(x_k)^T p_k < c_2 \nabla f(x_k)^T p_k$ выполняется. Это означает, что существует α'' такое, что:

$$\nabla f(x_k + \alpha'' p_k)^T p_k \leq c_2 \nabla f(x_k)^T p_k. \quad (4)$$

Неравенства (3) и (4) вместе гарантируют выполнение условий Вульфа.

4. Для сильных условий Вульфа, условие ограничения на кривизну:

$$\left|\nabla f(x_k + \alpha p_k)^T p_k\right| \leq c_2 \left|\nabla f(x_k)^T p_k\right| \quad (5)$$

выполняется, потому что $\nabla f(x_k + \alpha p_k)^T p_k$ отрицательно и ограничено снизу $c_2 \nabla f(x_k)^T p_k$.

Неточный линейный поиск. Условия Вульфа. Доказательство

1. Поскольку $\phi(\alpha) = f(x_k + \alpha p_k)$ ограничена снизу и $l(\alpha) = f(x_k) + \alpha c_1 \nabla f(x_k)^T p_k$ неограничена снизу (как $\nabla f(x_k)^T p_k < 0$), график $l(\alpha)$ должен пересекать график $\phi(\alpha)$ по крайней мере один раз. Пусть $\alpha' > 0$ будет наименьшим таким значением, удовлетворяющим:

$$f(x_k + \alpha' p_k) \leq f(x_k) + \alpha' c_1 \nabla f(x_k)^T p_k. \quad (1)$$

Это гарантирует выполнение **условия достаточного убывания**.

2. По теореме о среднем значении, существует $\alpha'' \in (0, \alpha')$ такое, что:

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k. \quad (2)$$

Подставляя $f(x_k + \alpha' p_k)$ из (1) в (2), мы получаем:

$$\alpha' \nabla f(x_k + \alpha'' p_k)^T p_k \leq \alpha' c_1 \nabla f(x_k)^T p_k.$$

Делим на $\alpha' > 0$, получаем:

$$\nabla f(x_k + \alpha'' p_k)^T p_k \leq c_1 \nabla f(x_k)^T p_k. \quad (3)$$

3. Поскольку $c_1 < c_2$ и $\nabla f(x_k)^T p_k < 0$, неравенство $c_1 \nabla f(x_k)^T p_k < c_2 \nabla f(x_k)^T p_k$ выполняется. Это означает, что существует α'' такое, что:

$$\nabla f(x_k + \alpha'' p_k)^T p_k \leq c_2 \nabla f(x_k)^T p_k. \quad (4)$$

Неравенства (3) и (4) вместе гарантируют выполнение условий Вульфа.

4. Для сильных условий Вульфа, условие ограничения на кривизну:

$$\left|\nabla f(x_k + \alpha p_k)^T p_k\right| \leq c_2 \left|\nabla f(x_k)^T p_k\right| \quad (5)$$

выполняется, потому что $\nabla f(x_k + \alpha p_k)^T p_k$ отрицательно и ограничено снизу $c_2 \nabla f(x_k)^T p_k$.

5. Из-за гладкости f, существует интервал вокруг α'', где выполняются условия Вульфа (и, следовательно, сильные условия Вульфа). Таким образом, доказательство завершено.

Бэктрекинг

Бэктрекинг - это техника для нахождения шага, удовлетворяющего условию Армихо, условиям Гольдштейна или другим критериям неточного линейного поиска. Она начинает с относительно большого шага и итеративно уменьшает его до тех пор, пока не будет выполнено условие.

Бэктрекинг

Бэктрекинг - это техника для нахождения шага, удовлетворяющего условию Армихо, условиям Гольдштейна или другим критериям неточного линейного поиска. Она начинает с относительно большого шага и итеративно уменьшает его до тех пор, пока не будет выполнено условие.

Алгоритм:

1. Выберите начальный шаг, α_0, и параметры $\beta \in (0, 1)$ и $c_1 \in (0, 1)$.

Бэктрекинг

Бэктрекинг - это техника для нахождения шага, удовлетворяющего условию Армихо, условиям Гольдштейна или другим критериям неточного линейного поиска. Она начинает с относительно большого шага и итеративно уменьшает его до тех пор, пока не будет выполнено условие.

Алгоритм:

1. Выберите начальный шаг, α_0, и параметры $\beta \in (0, 1)$ и $c_1 \in (0, 1)$.
2. Проверьте, удовлетворяет ли выбранный шаг выбранному условию (например, условию Армихо).

Бэктрекинг

Бэктрекинг - это техника для нахождения шага, удовлетворяющего условию Армихо, условиям Гольдштейна или другим критериям неточного линейного поиска. Она начинает с относительно большого шага и итеративно уменьшает его до тех пор, пока не будет выполнено условие.

Алгоритм:

1. Выберите начальный шаг, α_0, и параметры $\beta \in (0, 1)$ и $c_1 \in (0, 1)$.
2. Проверьте, удовлетворяет ли выбранный шаг выбранному условию (например, условию Армихо).
3. Если условие выполнено, остановитесь; в противном случае, установите $\alpha := \beta\alpha$ и повторите шаг 2.

Бэктрекинг

Бэктрекинг - это техника для нахождения шага, удовлетворяющего условию Армихо, условиям Гольдштейна или другим критериям неточного линейного поиска. Она начинает с относительно большого шага и итеративно уменьшает его до тех пор, пока не будет выполнено условие.

Алгоритм:

1. Выберите начальный шаг, α_0, и параметры $\beta \in (0, 1)$ и $c_1 \in (0, 1)$.
2. Проверьте, удовлетворяет ли выбранный шаг выбранному условию (например, условию Армихо).
3. Если условие выполнено, остановитесь; в противном случае, установите $\alpha := \beta\alpha$ и повторите шаг 2.

Бэктрекинг

Бэктрекинг - это техника для нахождения шага, удовлетворяющего условию Армихо, условиям Гольдштейна или другим критериям неточного линейного поиска. Она начинает с относительно большого шага и итеративно уменьшает его до тех пор, пока не будет выполнено условие.

Алгоритм:

1. Выберите начальный шаг, α_0, и параметры $\beta \in (0, 1)$ и $c_1 \in (0, 1)$.
2. Проверьте, удовлетворяет ли выбранный шаг выбранному условию (например, условию Армихо).
3. Если условие выполнено, остановитесь; в противном случае, установите $\alpha := \beta\alpha$ и повторите шаг 2.

Шаг α обновляется как

$$\alpha_{k+1} := \beta\alpha_k$$

в каждой итерации до тех пор, пока выбранное условие не будет выполнено.

Example

В задачах машинного обучения линейный поиск с бэктрекингом может использоваться для регулировки скорости обучения. Если функция потерь не уменьшается достаточно, скорость обучения уменьшается мультипликативно до тех пор, пока не будет выполнено, например, условие Армихо.

Численная иллюстрация

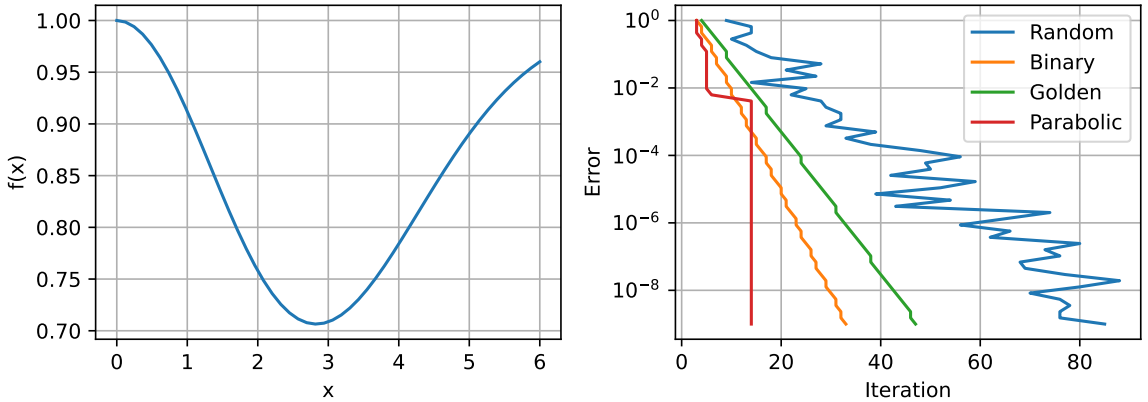
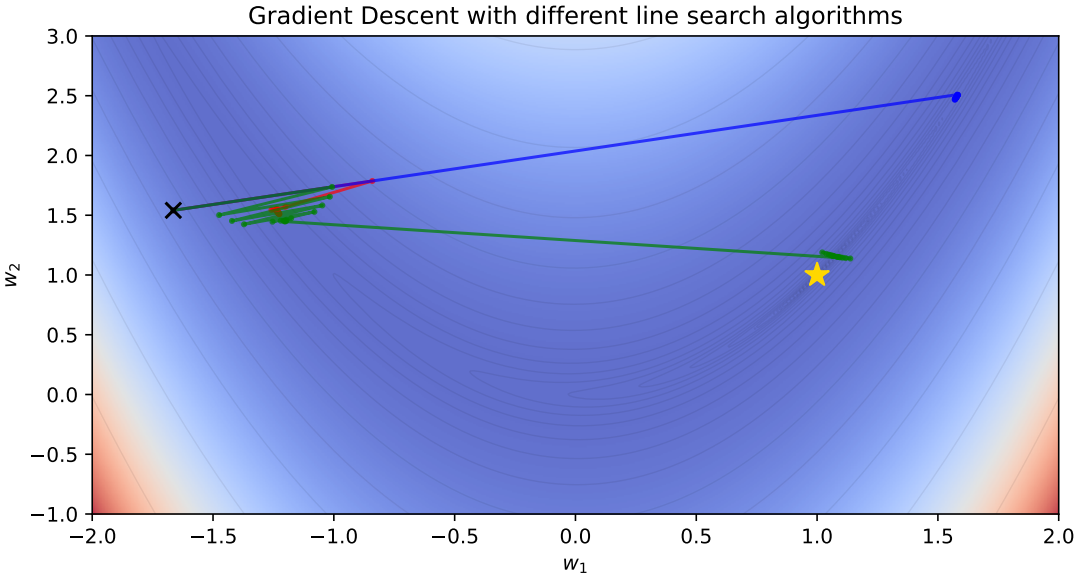


Figure 20: Сравнение различных алгоритмов линейного поиска

Открыть в Colab ♣

Градиентный спуск с линейным поиском



Итоги

Итоги

Определения

1. Унимодальная функция.
2. Метод дихотомии.
3. Метод золотого сечения.
4. Метод параболической интерполяции.
5. Условие достаточного убывания для неточного линейного поиска.
6. Условия Гольдштейна для неточного линейного поиска.
7. Условие ограничения на кривизну для неточного линейного поиска.
8. Градиент функции $f(x) : \mathbb{R}^n \to \mathbb{R}$.
9. Гессиан функции $f(x) : \mathbb{R}^n \to \mathbb{R}$.
10. Якобиан функции $f(x) : \mathbb{R}^n \to \mathbb{R}^m$.
11. Формула для аппроксимации Тейлора первого порядка $f^I_{x_0}(x)$ функции $f(x) : \mathbb{R}^n \to \mathbb{R}$ в точке x_0.
12. Формула для аппроксимации Тейлора второго порядка $f^{II}_{x_0}(x)$ функции $f(x) : \mathbb{R}^n \to \mathbb{R}$ в точке x_0.
13. Связь дифференциала функции df и градиента ∇f для функции $f(x) : \mathbb{R}^n \to \mathbb{R}$.
14. Связь второго дифференциала функции $d^2 f$ и гессиана $\nabla^2 f$ для функции $f(x) : \mathbb{R}^n \to \mathbb{R}$.

Теоремы

1. Метод дихотомии и золотого сечения для унимодальных функций. Скорость сходимости.