Методы оптимизации, ФКН ВШЭ, зима 2017

Семинар 2: вычисление производных и условия оптимальности.

17 января 2017 г.

1 Производная функции

Начнём с напоминания понятия производной.

Для функции одной переменной $f : \mathbb{R} \to \mathbb{R}$ её производная в точке x обозначается $f'(x)$ и определяется из равенства:

$$f(x+h) = f(x) + f'(x)h + o(h) \quad \text{для всех достаточно малых } h.$$

Другими словами, зафиксировав некоторую точку x, мы хотим приблизить изменение функции $f(x+h)-f(x)$ в окрестности этой точки с помощью линейной функции по h, и $f'(x)h$ — наилучший способ это сделать.

Рассмотрим теперь более общую ситуацию.

Пусть U и V суть конечномерные линейные пространства с нормами. Основными примерами таких пространств для нас будут служить числа: $\mathbb{R}$, векторы: $\mathbb{R}^n$ и матрицы: $\mathbb{R}^{n\times m}$, а также их комбинации (декартовы произведения).

Рассмотрим функцию $f : X \to V$, где $X \subseteq U$.

Определение 1 (Дифференцируемость)**.** Пусть $x \in X$ — внутренняя точка множества X, и пусть $L : U \to V$ — линейный оператор. Будем говорить, что функция f *дифференцируема в точке x с производной* L, если для всех достаточно малых $h \in U$ справедливо следующее разложение:

$$f(x+h) = f(x) + L[h] + o(\|h\|). \tag{1}$$

Если для любого линейного оператора $L : U \to V$ функция f не является дифференцируемой в точке x с производной L, то будем говорить, что f *не является дифференцируемой в точке* x. Если точка x не является внутренней точкой множества X, то оставим понятие дифференцируемости функции f в точке x неопределенным.

Замечание 1. Выражение $o(\|h\|)$ имеет стандартное значение:

$$f(x+h) - f(x) - L[h] = o(\|h\|) \quad \overset{\text{def}}{\iff} \quad \lim_{h\to 0} \frac{\|f(x+h) - f(x) - L[h]\|}{\|h\|} = 0.$$

Замечание 2. Поскольку рассматриваемые пространства U и V являются конечномерными (а в конечномерном пространстве все нормы топологически эквивалентны), то не имеет значения, какие конкретно нормы используются в данном выше определении: если функция f является дифференцируемой в точке x с производной L для одного выбора норм, то f также будет дифференцируемой в точке x с производной L для любого другого выбора норм.

Утверждение 1. *Предположим, что функция f дифференцируема в точке x с производной L_1 и также дифференцируема в точке x с производной L_2. Тогда $L_1 = L_2$.*

Таким образом, если функция f является дифференцируемой в точке x, то ее производная L определяется единственным образом. Будем обозначать ее символом $df(x)$.

Замечание 3. Объект df зависит от двух параметров: точка $x \in X$, в которой мы аппроксимируем функцию, и приращение $h \in U$, которое откладывается от зафиксированной точки:

$$df : X \times U \to V, \quad \text{линейный по второму аргументу — по «}h\text{».}$$

Замечание 4. Встречаются разные обозначения производной фунции f в точке x:

$$df(x)[h] \;\equiv\; Df(x)[h] \;\equiv\; Df(x)[\Delta x].$$

Все они обозначают одно и то же. При работе с определением производной, удобно явным образом указывать приращение (h или Δx) в квадратных скобках. При вычислении производных на практике, пользуясь уже известными посчитанными производными и свойствами пересчёта, приращение в квадратных скобках обычно не пишут: $df(x)$ или даже просто df, когда понятно, о чём идёт речь.

Итак, производная функции в точке x — это линейный оператор $df(x)$ который лучше всего аппроксимирует приращение функции:

$$f(x+h) - f(x) \;\approx\; df(x)[h].$$

Ещё одним известным и важным понятием является производная функции по направлению. Оказывается, зная производную функции f мы можем легко посчитать её производную вдоль любого направления h.

Утверждение 2. *Пусть f дифференцируема в точке x. Выберем произвольное направление h. Тогда:*

$$df(x)[h] \;=\; \frac{\partial f(x)}{\partial h} \;\overset{def}{=}\; \lim_{t\to+0} \frac{f(x+th) - f(x)}{t}$$

То есть, чтобы посчитать $\frac{\partial f(x)}{\partial h}$ — производную функции f вдоль направления h, достаточно применить $df(x)[\cdot]$ к этому направлению.

Набор векторов

$$e_i = (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0) \in \mathbb{R}^n, \quad i = 1, \ldots, n$$

называется *стандартным базисом* в $\mathbb{R}^n$.

Если для некоторого i у функции существует (двусторонняя) производная вдоль направления e_i, то её называют *частной производной по i-ой координате*:

$$\frac{\partial f(x)}{\partial x_i} \;\overset{\text{def}}{=}\; \lim_{t\to 0} \frac{f(x+te_i) - f(x)}{t} \;=\; df(x)[e_i].$$

Обратите внимание, что функция может быть недифференцируемой, даже если у неё существуют производные по всем направлениям.

Пример 1. Рассмотрим функцию $f(x) = \|x\|_2 = \sqrt{x^T x}$. Найдём её производную по направлению h в точке $x = 0$:

$$\frac{\partial \|x\|_2}{\partial h}\Big|_{x=0} \;=\; \lim_{t\to+0} \frac{\|0+th\|_2 - \|0\|_2}{t} \;=\; \lim_{t\to+0} \frac{t\|h\|_2}{t} = \|h\|_2.$$

Если бы функция $f(x) = \|x\|_2$ была дифференцируема в нуле, то по утверждению 2:

$$df(0)[h] \;=\; \frac{\partial f(0)}{\partial h} = \|h\|_2,$$

но функция $\|h\|_2$ не является линейной, что противоречит тому, что производная — линейный оператор. Значит $\|x\|_2$ не дифференцируема в нуле, хотя и имеет производные вдоль всех направлений.

Градиент функции, матрица Якоби.

- В случае $U = \mathbb{R}^n,\ V = \mathbb{R}$ линейную функцию $df(x)[h]$ всегда можно представить с помощью скалярного произведения с некоторым вектором:

  $$df(x)[h] = \langle a_x, h\rangle = a_x^T h \quad \text{ где } a_x \in \mathbb{R}^n \text{ — разный для каждого } x.$$

  Вектор a_x называется *градиентом* функции f в точке x и обозначается $\nabla f(x)$.

  В стандартном базисе градиент функции представляется в виде вектора из частных производных:

  $$\nabla f(x) = \left(\frac{\partial f}{\partial x_i}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right) \in \mathbb{R}^n.$$

  Как и все векторы у нас, этот вектор — вектор-столбец.

- В случае $U = \mathbb{R}^{n\times m},\ V = \mathbb{R}$ линейную функцию $df(x)[H]$ всегда можно представить с помощью скалярного произведения с некоторой матрицей:

$$df(x)[H] = \langle A_x, H\rangle = \operatorname{tr}(A_x^T H), \quad A_x, H \in \mathbb{R}^{n\times m}.$$

  Эта матрица также называется *градиентом функции* в точке x: $\nabla f(x) = A_x$ и в стандартном базисе (из матриц, у которых все нули, кроме одной единички) записывается в виде матрицы частных производных:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_{ij}}(x)\right)_{i=1,j=1}^{n,m}$$

- В случае $U = \mathbb{R}^m,\ V = \mathbb{R}^n$, линейный оператор $df(x)[\cdot]$, зафиксировав базисы, всегда можно представить матрицей:

$$df(x)[h] = J_x h, \quad J_x \in \mathbb{R}^{n\times m}.$$

  Матрица J_x называется *матрицей Якоби* функции f. В стандартном базисе она состоит из частных производных:

$$J_x = \left(\frac{\partial f_i}{\partial x_j}(x)\right)_{i=1,j=1}^{n,m}.$$

Утверждение 3 (Дифференциальное исчисление)**.** *Пусть U и V — векторные пространства, X — подмножество U, $x \in X$ — внутренняя точка X. Справедливы следующие свойства:*

1. *(Производная константы) Пусть $f : X \to V$ — постоянная функция, т. е. найдется $v \in V$, что $f(x') = v$ для всех $x' \in X$. Тогда f дифференцируема в точке x, и $df(x) = 0$.*

2. *(Производная тождественной функции) Пусть $f : X \to V$ — тождественная функция, т. е. $f(x') = x'$ для всех $x' \in X$. Тогда f дифференцируема в точке x, и ее производная — также тождественная функция: $df(x)[h] = h$ для всех $h \in U$.*

3. *(Линейность) Пусть $f : X \to V$ и $g : X \to V$ — функции. Если f и g дифференцируемы в точке x, а $c_1, c_2 \in \mathbb{R}$ — числа, то функция $(c_1 f + c_2 g)$ также дифференцируема в точке x, и*

$$d(c_1 f + c_2 g)(x) = c_1 df(x) + c_2 dg(x).$$

4. *(Правило произведения) Пусть $\alpha : X \to \mathbb{R}$ и $f : X \to V$ — функции. Если α и f дифференцируемы в точке x, то функция αf также дифференцируема в точке x, и*

$$d(\alpha f)(x)[h] = (d\alpha(x)[h])f(x) + \alpha(x)(df(x)[h])$$

   для всех $h \in U$.

5. *(Правило композиции) Пусть Y — подмножество V, $f : X \to Y$ — функция. Также пусть W — векторное пространство, $g : Y \to W$ — функция. Если f дифференцируема в точке x, и g дифференцируема в точке $f(x)$, то их композиция $(g \circ f) : X \to W$ (определенная как $(g \circ f)(x) = g(f(x))$) также будет дифференцируема в точке x, и*

$$d(g\circ f)(x) = dg(f(x))\Big[\, df(x) \Big] \qquad \text{или, более подробно,} \qquad d(g\circ f)(x)[h] = dg(f(x))\Big[\, df(x)[h] \Big].$$

6. *(Правило частного) Пусть $\alpha : X \to \mathbb{R}$ и $f : X \to V$ — функции. Если α и f дифференцируемы в точке x, и если α не обращается в ноль на X, то функция $(1/\alpha)f$ также дифференцируема в точке x, и*

$$d\left(\frac{1}{\alpha}f\right)(x)[h] = \frac{\alpha(x)(df(x)[h]) - (d\alpha(x)[h])f(x)}{\alpha(x)^2}.$$

   для всех $h \in U$.

Доказательство. Первые четыре свойства доказываются по определению, а последнее выводится из правил произведения и композиции. □

Заметим, что правило произведения в утверждении 3 установлено только в случае, когда одна из функций является скалярной. Это понятно, поскольку в векторных пространствах определено лишь умножение на скаляр, а не на произвольный элемент векторного пространства. Тем не менее, в некоторых частных случаях правило произведения остается верным даже если обе функции являются не скалярными. Например, справедливо следующее утверждение.

Утверждение 4. *Пусть U — векторное пространство, X — подмножество U, $x \in X$ — внутренняя точка X. Пусть $f : X \to \mathbb{R}^{m\times n}$ и $g : X \to \mathbb{R}^{n\times k}$ — матрично-значные функции. Предположим, что f и g дифференцируемы в точке x. Тогда функция fg также дифференцируема в точке x, и*

$$d(fg)(x)[h] = (df(x)[h])g(x) + f(x)(dg(x)[h]).$$

для всех $h \in U$. (Здесь под операцией умножения подразумевается матричное умножение, поэтому порядок множителей имеет значение.)

Вторая производная, гессиан функции.

Пусть функция $f : X \to V$ дифференцируема в каждой точке $x \in X \subseteq U$.

Рассмотрим производную функции f при фиксированном приращении $h_1 \in U$ как функцию от x:

$$g(x) = df(x)[h_1].$$

Определение 2. Если для функции g в некоторой точке x существует производная, то она называется *второй производной* функции f в точке x:

$$d^2f(x)[h_1, h_2] \stackrel{\text{def}}{=} dg(x)[h_2].$$

. Можно показать, что $d^2f(x)[h1, h2]$ является билинейной функцией по h_1 и h_2.

По аналогии определяются третья: $d^3f(x)[h_1, h_2, h_3]$, четвёртая и производные более высоких порядков.

Если производная $df(x)$ является непрерывной функцией по x, то говорят, что f — *непрерывно дифференцируема*. Если вторая производная $d^2f(x)$ непрерывна по x, то тогда f — *дважды непрерывно дифференцируема.*

Для функций $f : \mathbb{R}^n \to \mathbb{R}$ вторую производную, как и любую билинейную форму, можно представить с помощью матрицы:

$$d^2f(x)[h_1, h_2] = \langle H_x h_1, h_2\rangle = h_2^T H_x h_1, \quad H_x \in \mathbb{R}^{n\times n}.$$

Матрица H_x называется *гессианом* функции f в точке x и обозначается обычно $\nabla^2 f(x)$. В стандартном базисе эта матрица состоит из вторых частных производных:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{i=1,j=1}^{n,n}$$

Для дважды непрерывно дифференцируемой функции её гессиан — симметричная матрица:

$$\nabla^2 f(x) \in \mathbb{S}^n.$$

Формула Тейлора.

Для дважды непрерывно-дифференцируемой функции справедлива формула Тейлора:

$$f(x + h) = f(x) + df(x)[h] + \frac{1}{2}d^2f(x)[h, h] + o(\|h\|^2).$$

Для функции $f : \mathbb{R}^n \to \mathbb{R}$ её можно записать, используя градиент и гессиан:

$$f(x + h) = f(x) + \nabla f(x)^T h + \frac{1}{2}h^T\nabla^2 f(x)h + o(\|h\|^2).$$

Если функция имеет непрерывные производные до порядка k включительно, то формулу Тейлора можно записать до k-ой производной:

$$f(x + h) = f(x) + df(x)[h] + \frac{1}{2!}d^2f(x)[h, h] + \frac{1}{3!}d^3f(x)[h, h, h] + \ldots + \frac{1}{k!}d^kf(x)[h, \ldots, h] + o(\|h\|^k).$$

2 Подсчет табличных производных

1. $d(A)[h] = 0$ (A — фиксированная матрица)
2. $d(c^T x)[h] = c^T h$
3. $d(x^T Ax)[h] = x^T(A + A^T)h$
4. $d(x^T Ax)[h] = 2x^T Ah$ (если $A = A^T$)
5. $d(X^T)[H] = H^T$
6. $d\operatorname{Tr}(X)[H] = \operatorname{Tr}(H)$
7. $d(AXB)[H] = AHB$
8. $d(X^{-1})[H] = -X^{-1}HX^{-1}$
9. $d\operatorname{Det}(X)[H] = \operatorname{Det}(X)\operatorname{Tr}(X^{-1}H)$

Замечание: всюду в дальнейшем $\|\cdot\|$ обозначает евклидову норму для векторов и спектральную (операторную) норму для матриц.

Пример 2 (Линейная функция)**.** Пусть $c \in \mathbb{R}^n$, и пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция $f(x) := c^T x$. Покажем, что f является дифференцируемой в произвольной точке $x \in \mathbb{R}^n$ и найдем ее производную $df(x) : \mathbb{R}^n \to \mathbb{R}$. Для этого зафиксируем произвольное приращение аргумента $h \in \mathbb{R}^n$ и вычислим соответствующее приращение функции:

$$f(x + h) - f(x) = c^T(x + h) - c^T x = c^T h.$$

Заметим, что отображение $h \mapsto c^T h$ является линейным. Значит, для функции f справедливо разложение (1) с $df(x)[h] := c^T h$. Таким образом, функция f является дифференцируемой в произвольной точке $x \in \mathbb{R}^n$ с производной $df(x)[h] = c^T h$.

Пример 3 (Квадратичная форма)**.** Пусть $A \in \mathbb{R}^{n\times n}$, и пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция $f(x) := x^T Ax$. Зафиксируем произвольную точку $x \in \mathbb{R}^n$ и произвольное приращение аргумента $h \in \mathbb{R}^n$ и вычислим соответствующее приращение функции:

$$f(x + h) - f(x) = (x + h)^T A(x + h) - x^T Ax = x^T(A + A^T)h + h^T Ah.$$

Заметим, что отображение $h \mapsto x^T(A+A^T)h$ является линейным, а $h^T Ah = o(\|h\|)$, поскольку для всех $h \in \mathbb{R}^n$ справедлива следующая цепочка неравентв:

$$|h^T Ah| \leq \|h\|\|Ah\| \leq \|A\|\|h\|^2.$$

Здесь первое неравенство следует из неравенства Коши-Буняковского; второе неравенство следует из согласованности матричной и векторной норм. Таким образом, функция f дифференцируема в произвольной точке $x \in \mathbb{R}^n$ с производной $df(x)[h] = x^T(A + A^T)h$.

Пример 4 (Обратная матрица)**.** Пусть $S := \{X \in \mathbb{R}^{n\times n} : \operatorname{Det}(X) \neq 0\}$ — множество всех квадратных невырожденных матриц размера n. Рассмотрим функцию $f : S \to S$, которая для каждой матрицы $X \in S$ возвращает ее обратную: $f(X) := X^{-1}$. Покажем, что f является дифференцируемой в любой точке $X \in S$. Для этого зафиксируем произвольное достаточно малое приращение аргумента $H \in \mathbb{R}^{n\times n}$ (удовлетворяющее $X + H \in S$ и $\|H\| < 1/\|X^{-1}\|$) и рассмотрим соответствующее приращение функции:

$$f(X + H) - f(X) = (X + H)^{-1} - X^{-1} = (X(I_n + X^{-1}H))^{-1} - X^{-1} = ((I_n + X^{-1}H)^{-1} - I_n)X^{-1}.$$

Оценим отдельно $(I_n + X^{-1}H)^{-1}$. Для этого разложим эту матрицу в ряд Неймана:[1]

$$(I_n + X^{-1}H)^{-1} = I_n - X^{-1}H + \sum_{k=2}^{\infty}(-X^{-1}H)^k.$$

[1]Имеется в виду разложение $(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k$, справедливое для любой матрицы $A \in \mathbb{R}^{n\times n}$, такой, что $\|A\| < 1$. Эта формула является обобщением известной формулы для суммы геометрического ряда: $(1 - q)^{-1} = \sum_{k=0}^{\infty} q^k$ для любого $|q| < 1$.

Заметим, что ряд, стоящий в правой части последнего равенства, является абсолютно сходящимся, поскольку $\|X^{-1}H\| < 1$ в силу достаточной малости H. Покажем, что сумма этого ряда есть $o(\|H\|)$:

$$\left\|\sum_{k=2}^{\infty}(-X^{-1}H)^k\right\| \leq \sum_{k=2}^{\infty}\|(-X^{-1}H)^k\| \leq \sum_{k=2}^{\infty}\|X^{-1}\|^k\|H\|^k = \frac{\|X^{-1}\|^2\|H\|^2}{1-\|X^{-1}\|\cdot\|H\|}.$$

Здесь первое неравенство следует из неравенства треугольника для нормы; второе неравенство следует из субмультипликативности нормы; далее вычисляется сумма геометрического ряда. Таким образом,

$$(I_n + X^{-1}H)^{-1} = I_n - X^{-1}H + o(\|H\|).$$

Подставляя это выражение в полученную выше формулу для приращения функции, получаем

$$f(X+H) - f(X) = -X^{-1}HX^{-1} + o(\|H\|).$$

Таким образом, функция f дифференцируема в произвольной точке $X \in S$ с производной $df(X)[H] = -X^{-1}HX^{-1}$.

Замечание 5. Выведенную формулу для производной функции X^{-1} можно очень просто получить с помощью следующего трюка. Рассмотрим дифференциал единичной матрицы $d(I_n)$. С одной стороны, поскольку матрица постоянная, $d(I_n) = 0$. С другой стороны, по правилу произведения, $dI_n = d\left(XX^{-1}\right) = (dX)X^{-1} + Xd(X^{-1})$. Приравняв выражения, получим $d(X^{-1}) = -X^{-1}(dX)X^{-1}$, или, в другой форме, $d(X^{-1})[H] = -X^{-1}HX^{-1}$. Заметим, однако, что приведённое рассуждение не является полным доказательством тождества, так как предполагает, но не доказывает существование дифференциала $d(X^{-1})$.

Пример 5 (Определитель матрицы)**.** Пусть $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ — функция $f(X) := \mathrm{Det}(X)$. Рассмотрим произвольную точку $X \in \mathbb{R}^{n\times n}$ и произвольное приращение аргумента $H \in \mathbb{R}^{n\times n}$. Будем предполагать, что матрица X обратима. Выпишем соответветствующее приращение функции:

$$f(X+H) - f(X) = \mathrm{Det}(X+H) - \mathrm{Det}(X) = \mathrm{Det}(X(I_n + X^{-1}H)) - \mathrm{Det}(X) = \mathrm{Det}(X)(\mathrm{Det}(I_n + X^{-1}H) - 1).$$

Оценим отдельно $\mathrm{Det}(I_n + X^{-1}H)$. Для этого воспользуемся тем, что определитель матрицы равен произведению ее собственных значений. Пусть $\lambda_1(X^{-1}H), \ldots, \lambda_n(X^{-1}H)$ — собственные значения матрицы $X^{-1}H$ (пронумерованные в произвольном порядке и, возможно, комплексные). Заметим, что собственными значениями матрицы $I_n + X^{-1}H$ будут $1 + \lambda_1(X^{-1}H), \ldots, 1 + \lambda_n(X^{-1}H)$. Поэтому

$$\mathrm{Det}(I_n + X^{-1}H) = \prod_{i=1}^{n}[1 + \lambda_i(X^{-1}H)] = 1 + \sum_{i=1}^{n}\lambda_i(X^{-1}H) + \left(\sum_{1\leq i\leq j\leq n}\lambda_i(X^{-1}H)\lambda_j(X^{-1}H) + \ldots\right),$$

где многоточие скрывает сумму всевозможных троек $\lambda_i(X^{-1}H)\lambda_j(X^{-1}H)\lambda_k(X^{-1}H)$, всевозможных четверок и т. д. Заметим, что выражение, стоящее скобках, представляет из себя величину $o(\|H\|)$. Это следует из неравенства треугольника и того факта, что для произвольной матрицы $A \in \mathbb{R}^{n\times n}$ все ее собственные значения не превосходят по модулю ее нормы $\|A\|$. (Действительно, пусть $\lambda \in \mathbb{C}$ — собственное значение матрицы A, и пусть $x \in \mathbb{C}^n \setminus \{0\}$ — соответствующий собственный вектор: $Ax = \lambda x$. Тогда $|\lambda|\|x\| = \|Ax\| \leq \|A\|\|x\|$.) Таким образом,

$$\mathrm{Det}(I_n - X^{-1}H) = 1 + \sum_{i=1}^{n}\lambda_i(X^{-1}H) + o(\|H\|) = 1 + \mathrm{Tr}(X^{-1}H) + o(\|H\|).$$

Подставляя полученное выражение в полученную выше формулу для приращения функции, получаем

$$f(X+H) - f(X) = \mathrm{Det}(X)\,\mathrm{Tr}(X^{-1}H) + o(\|H\|).$$

Таким образом, для любой обратимой матрицы $X \in \mathbb{R}^{n\times n}$ функция f дифференцируема в точке X с производной $df(x)[H] = \mathrm{Det}(X)\,\mathrm{Tr}(X^{-1}H)$.

Замечание 6. Можно показать, что рассматриваемая функция $f(X) = \mathrm{Det}(X)$ будет дифференцируемой всюду на $\mathbb{R}^{n\times n}$, а не только на подмножестве обратимых матриц. Общая формула для производной в этом случае называется *формулой Якоби* и выглядит следующим образом: $df(X)[H] = \mathrm{Tr}(\mathrm{Adj}(X)H)$, где $\mathrm{Adj}(X)$ — присоединенная матрица к X. Заметим, что если X — невырожденная матрица, тогда $\mathrm{Adj}(X) = \mathrm{Det}(X)X^{-1}$ и формула Якоби переходит в доказанную формулу $df(X)[H] = \mathrm{Det}(X)\,\mathrm{Tr}(X^{-1}H)$.

3 Условия оптимальности

Рассмотрим функцию $f : X \to \mathbb{R}$, где $X \subseteq U$ — подмножество нормированного линейного пространства, например, $U = \mathbb{R}^n$ — векторы или $U = \mathbb{R}^{n \times m}$ — матрицы.

Определение 3. Точка x называется *точкой локального минимума* функции f, если существует шар некоторого радиуса $r > 0$, с центром в этой точке: $W = \{z \in U \mid \|z - x\| < r\}$, и выполнено:

$$f(x) \leqslant f(z) \quad \text{для любого } z \in W \cap X.$$

Если для всех $z \in W$ кроме $z = x$ в определении выполняется строгое неравенство, то локальный минимум называется *строгим локальным минимумом.*

Если неравенство выполнено в другую сторону, то точка x называется *локальным максимумом.* Локальные минимумы и максимумы вместе называются *локальными экстремумами.*

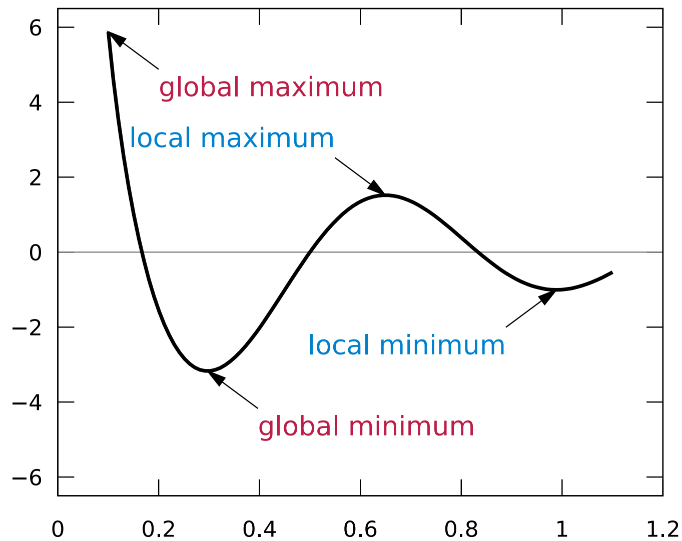

Рис. 1: Пример экстремумов из Википедии.

Если функция является дифференцируемой, то отыскать её локальные экстремумы иногда удаётся с помощью следующих утверждений:

Утверждение 5 (условие оптимальности первого порядка). *Пусть для функции $f : X \to \mathbb{R}$ точка x является точкой локального экстремума.*

Тогда если функция непрерывно-дифференцируема в окрестности этой точки, то её производная в этой точке равна нулю:

$$df(x) = 0.$$

Замечание 7. Равенство $df(x) = 0$ означает, что $df(x)[h] = 0$ для любого $h \in U$.

Замечание 8. Напомним, что для функции векторного аргумента $f : \mathbb{R}^n \to \mathbb{R}$ ее производную всегда можно представить с помощью градиента: $df(x)[h] = \nabla f(x)^T h$. В этом случае равенство производной нулю эквивалентно равенству нулю градиента:

$$df(x) = 0 \qquad \Leftrightarrow \qquad \nabla f(x)^T h = 0 \;\text{ для любого } h \in \mathbb{R}^n \qquad \Leftrightarrow \qquad \nabla f(x) = 0.$$

Замечание 9. Важно понимать, что утверждение 5 является лишь необходимым условием локального минимума. Например, для функции $f(x) = x^2$ производная равна нулю в единственной точке $x = 0$, и эта точка действительно является локальным минимумом функции (в данном случае и глобальным). Но для $f(x) = x^3$, в той же точке $x = 0$ — производная равна нулю, но сама точка не является ни локальным минимумом, ни локальным максимумом.

Определение 4. Точка x называется *стационарной точкой*, если производная в ней обращается в ноль: $df(x) = 0$. Стационарная точка, которая не является ни локальным минимумом, ни локальным максимумом, называется *седловой точкой.*

В случае функции двух переменных «седловая точка» действительно напоминает седло. Но мы используем это понятие для произвольных функций, в значении, указанном выше.

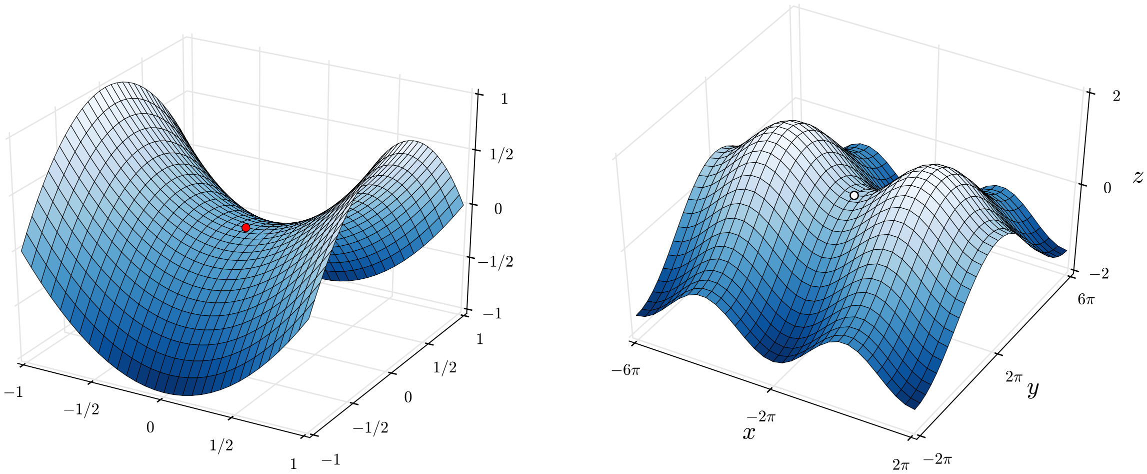

Рис. 2: Примеры седловых точек (слева — красная точка, справа — белая) из Википедии.

Для классификации стационарных точек удобным оказывается следующее:

Утверждение 6 (условия оптимальности второго порядка). *Пусть функции $f : X \to \mathbb{R}$ дважды непрерывно дифференцируема в окрестности стационарной точки $x \in X$. Тогда:*

- *если $d^2 f(x)[h, h] > 0$ для всех $h \in U$, $h \neq 0$, то x — строгий локальный минимум;*
- *если $d^2 f(x)[h, h] < 0$ для всех $h \in U$, $h \neq 0$, то x — строгий локальный максимум;*
- *если существует $h_1, h_2 \in U$ такие, что: $d^2 f(x)[h_1, h_1] > 0$ и $d^2 f(x)[h_2, h_2] < 0$, то x — седловая точка.*

Также, справедливы и необходимые условия оптимальности второго порядка:

- *если x — локальный минимум, то $d^2 f(x)[h, h] \geqslant 0$ для всех $h \in U$;*
- *если x — локальный максимум, то $d^2 f(x)[h, h] \leqslant 0$ для всех $h \in U$.*

Замечание 10. Напомним, что для функции векторного аргумета $f : \mathbb{R}^n \to \mathbb{R}$ её вторую производную всегда можно представить с помощью матрицы — гессиана:

$$d^2 f(x)[h, h] \;=\; h^T \nabla^2 f(x) h.$$

В этом случае утверждение 6 можно переписать в терминах знакоопределённости гессиана:

- если $\nabla^2 f(x) \succ 0$, то x — строгий локальный минимум;
- если $\nabla^2 f(x) \prec 0$, то x — строгий локальный максимум;
- если $\nabla^2 f(x)$ — *неопределённая* матрица (т.е. не выполнено ни $\nabla^2 f(x) \succeq 0$ ни $\nabla^2 f(x) \preceq 0$), то x — седловая точка.
- если x — локальный минимум, то $\nabla^2 f(x) \succeq 0$;
- если x — локальный максимум, то $\nabla^2 f(x) \preceq 0$.

Пример 6. Рассмотрим фунцию $f(x_1, x_2) = x_1^2 - x_2^2 + 2x_1$. Её градиент и гессиан:

$$\nabla f(x_1, x_2) = \begin{bmatrix} 2x_1 + 2 \\ -2x_2 \end{bmatrix}$$

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \quad — \quad \text{всюду постоянная матрица.}$$

Градиент обращается в ноль в единственной точке: $(-1, 0)$.
Гессиан — неопределённая матрица, значит стационарная точка $(-1, 0)$ — седловая.
Локальных экстремумов нет, функция неограниченная:

$$\inf_{x \in \mathbb{R}^2} f(x) = -\infty, \quad \sup_{x \in \mathbb{R}^2} f(x) = +\infty.$$