Семинар 1: Матрично-векторное дифференцирование

1 Матрично-векторное дифференцирование

1.1 Теория

Для вычисления большинства производных, которые возникают на практике, достаточно лишь небольшой таблицы стандартных производных и правил преобразования. Удобнее всего оказывается работать в терминах «дифференциала» — с ним можно не задумываться о промежуточных размерностях, а просто применять стандартные правила.

Замечание: В этом разделе описана сама техника матрично-векторного дифференцирования. Для более подробного описания математической теории, лежащей в основе этой техники, см. раздел A.

Правила преобразования

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

Таблица стандартных производных

$$d\langle A, X\rangle = \langle A, dX\rangle$$
$$d\langle Ax, x\rangle = \langle (A + A^T)x, dx\rangle$$
$$d\langle Ax, x\rangle = 2\langle Ax, dx\rangle \qquad (\textbf{если } A = A^T)$$
$$d(\mathrm{Det}(X)) = \mathrm{Det}(X)\langle X^{-T}, dX\rangle$$
$$d(X^{-1}) = -X^{-1}(dX)X^{-1}$$

Здесь A, B — фиксированные матрицы; α — фиксированный скаляр; X, Y — произвольные дифференцируемые матричные функции (согласованные по размерностям, чтобы все операции имели смысл); ϕ — произвольная дифференцируемая скалярная функция.

Одним из самых важных является правило **производной композиции**. Пусть $g(Y)$ и $f(X)$ — две дифференцирумые функции, и мы знаем выражения для их дифференциалов: $dg(Y)$ и $df(X)$. Чтобы посчитать производную *композиции* $\phi(X) := g(f(X))$, как и в скалярном случае, нужно:

- взять выражение посчитанного дифференциала $dg(Y)$;
- подставить в него вместо Y значение $f(X)$, а вместо dY значение $df(X)$.

Пример

Рассмотрим функцию $\phi(x) := \ln\langle Ax, x\rangle$, где $A \in \mathbb{S}^n_{++}$. В данном случае

$$g(y) := \ln(y), \quad dg(y) = \frac{dy}{y}; \qquad f(x) := \langle Ax, x\rangle, \quad df(x) = 2\langle Ax, dx\rangle.$$

Подставляем формально в $dg(y)$ вместо y выражение для $f(x) = \langle Ax, x\rangle$, а вместо dy выражение для $df(x) = 2\langle Ax, dx\rangle$:

$$d\phi(x) = \frac{2\langle Ax, dx\rangle}{\langle Ax, x\rangle} \qquad (\text{В нотации с «}D\text{»-большим:} \quad D\phi(x)[h] = \frac{2\langle Ax, h\rangle}{\langle Ax, x\rangle}).$$

Обычно, все возникающие на практике матрично-векторные функции составлены с помощью табличных функций и стандартных операций над ними. Благодаря универсальности приведённых правил, дифференцировать сколь угодно сложные функции такого типа становится настолько же просто, как и дифференцировать одномерные функции.

Полученное в конце концов выражение нужно привести к одному из канонических видов:

Вход \ **Выход**	Скаляр	Вектор	Матрица
Скаляр	$df(x) = f'(x)dx$ ($f'(x)$: скаляр; dx: скаляр)	–	–
Вектор	$df(x) = \langle \nabla f(x), dx \rangle$ ($\nabla f(x)$: вектор; dx: вектор)	$df(x) = J_f(x)dx$ ($J_f(x)$: матрица; dx: вектор)	–
Матрица	$df(X) = \langle \nabla f(X), dX \rangle$ ($\nabla f(X)$: матрица; dX: матрица)	–	–

Случаи, отмеченные «–», нас интересовать не будут. Объект $\nabla f(x)$ (вектор для функции векторного аргумента и матрица для функции матричного аргумента) называется **градиентом**. Матрица $J_f(x)$ называется **матрицей Якоби**.

Найти *вторую производную* функции $f(X)$ можно по следующему «алгоритму»:

- посчитать первую производную функции; зафиксировать в выражениии для $df(X)$ приращение dX — обозначить его как dX_1;
- посчитать производную для функции $g(X) = df(X)$, считая dX_1 фиксированным (константа). Новое приращение обозначать dX_2.

Пример

Вернёмся к функции $\phi(x) = \ln\langle Ax, x\rangle$, где $A \in \mathbb{S}^n_{++}$. Мы уже посчитали её первую производную: $d\phi(x) = \frac{2\langle Ax, dx\rangle}{\langle Ax, x\rangle}$. Обозначим dx за dx_1 и рассмотрим новую функцию:

$$g(x) = \frac{2\langle Ax, dx_1\rangle}{\langle Ax, x\rangle}$$

Найдём производную $g(x)$, считая, что dx_1 — константный вектор:

$$\begin{aligned} d^2\phi(x) &= d\left(\frac{2\langle Ax, dx_1\rangle}{\langle Ax, x\rangle}\right) = \frac{d(2\langle Ax, dx_1\rangle)\langle Ax, x\rangle - 2\langle Ax, dx_1\rangle d\langle Ax, x\rangle}{\langle Ax, x\rangle^2} \\ &= \frac{2\langle Adx_1, dx_2\rangle\langle Ax, x\rangle - 2\langle Ax, dx_1\rangle 2\langle Ax, dx_2\rangle}{\langle Ax, x\rangle^2} = \left\langle \left(\frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^T A}{\langle Ax, x\rangle^2}\right) dx_1, dx_2 \right\rangle. \end{aligned}$$

(В нотации с D-большим: $D^2\phi(x)[h_1, h_2] = \left\langle \left(\frac{2A}{\langle Ax,x\rangle} - \frac{4Axx^T A}{\langle Ax,x\rangle^2}\right) h_1, h_2 \right\rangle$.)

Для второй производной каноническая форма для скалярной функции векторного аргумента

$$d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx_2\rangle.$$

Матрица $\nabla^2 f(x)$ называется **гессианом**. Для дважды непрерывно дифференцируемых функций гессиан является симметричной матрицей.

1.2 Задачи

Задача 1 (Квадратичная функция)**.** Найти первую и вторую производные $df(x)$ и $d^2 f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c, \qquad x \in \mathbb{R}^n,$$

где $A \in \mathbb{S}^n$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.

Решение. Найдем первую производную:

$$\boxed{df(x)} = d\left(\frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c\right) = \frac{1}{2}d\langle Ax, x\rangle - d\langle b, x\rangle = \frac{1}{\cancel{2}}\cancel{2}\langle Ax, dx\rangle - \langle b, dx\rangle \boxed{= \langle Ax - b, dx\rangle}.$$

Заметим, что $df(x)$ уже записан в канонической форме $df(x) = \langle \nabla f(x), dx\rangle$, поэтому

$$\boxed{\nabla f(x) = Ax - b}.$$

Теперь найдём вторую производную:

$$\boxed{d^2 f(x)} = d\langle Ax - b, dx_1\rangle = \langle d(Ax - b), dx_1\rangle = \langle d(Ax), dx_1\rangle \boxed{= \langle Adx_2, dx_1\rangle}.$$

Чтобы найти гессиан, приведем $d^2 f(x)$ к канонической форме $d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx_2\rangle$:

$$d^2 f(x) = \langle Adx_1, dx_2\rangle \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = A}.$$

Задача 2 (Куб евклидовой нормы)**.** Найти первую и вторую производные $df(x)$ и $d^2 f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) := \frac{1}{3}\|x\|_2^3, \quad x \in \mathbb{R}^n.$$

Решение. Найдем первую производную:

$$\boxed{df(x)} = d\left(\frac{1}{3}\|x\|_2^3\right) = \frac{1}{3}d\langle x,x\rangle^{3/2} = \frac{1}{\cancel{3}}\frac{\cancel{3}}{2}\langle x,x\rangle^{1/2}d\langle x,x\rangle = \frac{1}{\cancel{2}}\|x\|_2(\cancel{2}\langle x,dx\rangle)\boxed{= \|x\|_2\langle x,dx\rangle}.$$

Чтобы найти градиент, приведем $df(x)$ к канонической форме $df(x) = \langle \nabla f(x), dx\rangle$:

$$df(x) = \langle \|x\|_2 x, dx\rangle \quad \Rightarrow \quad \boxed{\nabla f(x) = \|x\|_2 x}.$$

Теперь найдем вторую производную:

$$\begin{aligned}\boxed{d^2 f(x)} &= d(\|x\|_2\langle x,dx_1\rangle) = \underbrace{d(\|x\|_2)}_{=d(\langle x,x\rangle^{1/2})}\langle x,dx_1\rangle + \|x\|_2 d\langle x,dx_1\rangle \\ &= \left(\frac{1}{\cancel{2}}\langle x,x\rangle^{-1/2}(\cancel{2}\langle x,dx_2\rangle\right)\langle x,dx_1\rangle + \|x\|_2\langle dx_2,dx_1\rangle \\ &\boxed{= \|x\|_2^{-1}\langle x,dx_2\rangle\langle x,dx_1\rangle + \|x\|_2\langle dx_2,dx_1\rangle}.\end{aligned}$$

Чтобы найти гессиан, приведем $d^2f(x)$ к канонической форме $d^2f(x) = \langle \nabla^2 f(x)dx_1, dx_2\rangle$:

$$\begin{aligned}d^2f(x) &= \|x\|_2^{-1}\langle dx_1,x\rangle\langle x,dx_2\rangle + \|x\|_2\langle dx_1,dx_2\rangle \\ &= \langle(\|x\|_2^{-1}xx^T + \|x\|_2 I_n)dx_1, dx_2\rangle \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = \|x\|_2^{-1}xx^T + \|x\|_2 I_n}.\end{aligned}$$

Отметим, что полученная формула для гессиана (и второй производной) верна только при $x \neq 0$, поскольку значение $\|x\|_2^{-1}$ не определено для $x = 0$. Такое ограничение возникло из-за того, что в самом начале мы воспользовались правилом произведения, и у нас возникла производная $d(\|x\|_2)$, которая не существует в точке $x = 0$. Тем не менее, можно показать, что рассматриваемая функция f является всюду дважды непрерывно дифференцируемой, и ее вторая производная в точке $x = 0$ равна нулю. Таким образом, можно сказать, что полученная формула, на самом деле, верна для любых значений x, с оговоркой, что в точке $x = 0$ значение $\|x\|_2^{-1}xx^T$ надо понимать как 0 (предел при $x \to 0$).

Задача 3 (Евклидова норма)**.** Найти первую и вторую производные $df(x)$ и $d^2f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) := \|x\|_2, \quad x \in \mathbb{R}^n \setminus \{0\}.$$

Решение. Найдем первую производную:

$$\boxed{df(x)} = d(\|x\|_2) = d(\langle x,x\rangle^{1/2}) = \frac{1}{2}\langle x,x\rangle^{-1/2}d\langle x,x\rangle = \frac{1}{\cancel{2}}\|x\|_2^{-1}\cancel{2}\langle x,dx\rangle\boxed{= \|x\|_2^{-1}\langle x,dx\rangle}.$$

Чтобы найти градиент, приведем $df(x)$ к канонической форме $df(x) = \langle \nabla f(x), dx\rangle$:

$$df(x) = \langle \|x\|_2^{-1}x, dx\rangle \quad \Rightarrow \quad \boxed{\nabla f(x) = \|x\|_2^{-1}x}.$$

Теперь найдем вторую производную:

$$\begin{aligned}\boxed{d^2f(x)} &= d(\|x\|_2^{-1}\langle x,dx_1\rangle) = d(\|x\|_2^{-1})\langle x,dx_1\rangle + \|x\|_2^{-1}d\langle x,dx_1\rangle \\ &= -\|x\|_2^{-2}d(\|x\|_2)\langle x,dx_1\rangle + \|x\|_2^{-1}\langle dx_2,dx_1\rangle \\ &= -\|x\|_2^{-2}(\|x\|_2^{-1}\langle x,dx_2\rangle)\langle x,dx_1\rangle + \|x\|_2^{-1}\langle dx_2,dx_1\rangle \\ &\boxed{= \|x\|_2^{-1}\langle dx_2,dx_1\rangle - \|x\|_2^{-3}\langle x,dx_2\rangle\langle x,dx_1\rangle}.\end{aligned}$$

Чтобы найти гессиан, приведем $d^2 f(x)$ к канонической форме $d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx_2 \rangle$:

$$
\begin{aligned}
d^2 f(x) &= \|x\|_2^{-1}(\langle dx_1, dx_2\rangle - \|x\|_2^{-2}\langle dx_1, x\rangle\langle x, dx_2\rangle) \\
&= \langle \|x\|_2^{-1}(I_n - \|x\|_2^{-2} x x^T) dx_1, dx_2 \rangle \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = \|x\|_2^{-1}(I_n - \|x\|_2^{-2} x x^T)}.
\end{aligned}
$$

Задача 4 (Логистическая функция)**.** Найти первую и вторую производные $df(x)$ и $d^2 f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) := \ln(1 + \exp(\langle a, x \rangle)), \quad x \in \mathbb{R}^n,$$

где $a \in \mathbb{R}^n$.

Решение. Найдем первую производную:

$$
\begin{aligned}
\boxed{df(x)} &= d(\ln(1 + \exp(\langle a, x\rangle))) = \left\{ d(\ln(x)) = \frac{dx}{x} \right\} = \frac{d(1 + \exp(\langle a, x\rangle))}{1 + \exp(\langle a, x\rangle)} = \frac{d(\exp(\langle a, x\rangle))}{1 + \exp(\langle a, x\rangle)} \\
&= \{ d(\exp(x)) = \exp(x) dx \} = \frac{\exp(\langle a, x\rangle) d\langle a, x\rangle}{1 + \exp(\langle a, x\rangle)} = \frac{\exp(\langle a, x\rangle)\langle a, dx\rangle}{1 + \exp(\langle a, x\rangle)} = \frac{\langle a, dx\rangle}{1 + \exp(-\langle a, x\rangle)} \\
&\boxed{= \sigma(\langle a, x\rangle)\langle a, dx\rangle}.
\end{aligned}
$$

Здесь введено обозначение $\sigma : \mathbb{R} \to \mathbb{R}$ для *сигмоидной функции*:

$$\boxed{\sigma(x) := \frac{1}{1 + \exp(-x)}}.$$

Чтобы найти градиент, приведем $df(x)$ к канонической форме $df(x) = \langle \nabla f(x), dx \rangle$:

$$df(x) = \langle \sigma(\langle a, x\rangle) a, dx \rangle \quad \Rightarrow \quad \boxed{\nabla f(x) = \sigma(\langle a, x\rangle) a}.$$

Таким образом, градиент $\nabla f(x)$ — это вектор, коллинеарный вектору a с коэффициентом $\sigma(\langle a, x\rangle) \in (0, 1)$. В зависимости от точки x меняется лишь длина вектора $\nabla f(x)$, но не его направление.

Теперь найдем вторую производную:

$$
\begin{aligned}
\boxed{d^2 f(x)} &= d(\sigma(\langle a, x\rangle)\langle a, dx_1\rangle) = d(\sigma(\langle a, x\rangle))\langle a, dx_1\rangle = \{ d(\sigma(x)) = \sigma'(x) dx \} = (\sigma'(\langle a, x\rangle) d\langle a, x\rangle)\langle a, dx_1\rangle \\
&= \sigma'(\langle a, x\rangle)\langle a, dx_2\rangle\langle a, dx_1\rangle = \{ \sigma'(x) = \sigma(x)(1 - \sigma(x)) \} \\
&\boxed{= \sigma(\langle a, x\rangle)(1 - \sigma(\langle a, x\rangle))\langle a, dx_2\rangle\langle a, dx_1\rangle}.
\end{aligned}
$$

Чтобы найти гессиан, приведем $d^2 f(x)$ к канонической форме $d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx_2 \rangle$:

$$
\begin{aligned}
d^2 f(x) &= \sigma(\langle a, x\rangle)(1 - \sigma(\langle a, x\rangle))\langle dx_1, a\rangle\langle a, dx_2\rangle \\
&= \langle (\sigma(\langle a, x\rangle)(1 - \sigma(\langle a, x\rangle)) a a^T) dx_1, dx_2 \rangle \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = \sigma(\langle a, x\rangle)(1 - \sigma(\langle a, x\rangle)) a a^T}.
\end{aligned}
$$

Заметим, что гессиан $\nabla^2 f(x)$ — это одноранговая матрица, пропорциональная матрице aa^T с коэффициентом $\sigma(\langle a, x\rangle)(1 - \sigma(\langle a, x\rangle)) \in (0, 0.25)$. Точка x влияет лишь на коэффициент пропорциональности.

Задача 5 (Логарифм определителя)**.** Найти первую и вторую производные $df(X)$ и $d^2 f(X)$, а также градиент $\nabla f(X)$ функции

$$f(X) := \ln(\mathrm{Det}(X)),$$

заданной на множестве $\mathbb{S}^n_{++}$ в пространстве $\mathbb{S}^n$.

Решение. Найдём первую производную:

$$\boxed{df(X)} = d(\ln \mathrm{Det}(X)) = \left\{ d(\ln(x)) = \frac{dx}{x} \right\} = \frac{d(\mathrm{Det}(X))}{\mathrm{Det}(X)} = \frac{\cancel{\mathrm{Det}(X)}\langle X^{-1}, dX\rangle}{\cancel{\mathrm{Det}(X)}} \boxed{= \langle X^{-1}, dX\rangle}.$$

Заметим, что $df(X)$ уже и так записан в канонической форме $df(X) = \langle \nabla f(X), dX\rangle$.[1] Поэтому,

$$\boxed{\nabla f(X) = X^{-1}}.$$

Теперь найдем вторую производную:

$$\boxed{d^2 f(X)} = d\langle X^{-1}, dX_1\rangle = \langle d(X^{-1}), dX_1\rangle = \langle -X^{-1}(dX_2)X^{-1}, dX_1\rangle \boxed{= -\langle X^{-1}(dX_2)X^{-1}, dX_1\rangle}.$$

В итоге получилась билинейная форма от приращений dX_1 и dX_2 в пространстве матриц.

Рассмотрим

$$D^2 f(X)[H, H] = -\langle X^{-1}HX^{-1}, H\rangle.$$

Покажем, что $D^2 f(X)[H, H]$ имеет отрицательный знак для всех $X \in \mathbb{S}^n_{++}$ и $H \in \mathbb{S}^n$, т. е. что функция f является вогнутой функцией. Действительно, раскладывая $X^{-1} = X^{-1/2}X^{-1/2}$, перепишем $D^2 f(X)[H, H]$ в следующем виде:

$$D^2 f(X)[H, H] = -\langle X^{-1/2}HX^{-1/2}, X^{-1/2}HX^{-1/2}\rangle = -\|X^{-1/2}HX^{-1/2}\|_F^2.$$

Отсюда видно, что $D^2 f(X)[H, H]$, действительно, имеет отрицательный знак.

Задача 6. Найти производную $df(X)$ и градиент $\nabla f(X)$ функции

$$f(X) := \|AX - B\|_F, \quad X \in \mathbb{R}^{k\times n},$$

где $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{m\times n}$.

Решение. Вычислим отдельно $d(\|X\|_F)$:

$$\begin{aligned} d(\|X\|_F) &= d(\langle X, X\rangle^{1/2}) = \left\{ d(x^{1/2}) = \frac{1}{2}x^{-1/2}dx \right\} = \frac{1}{2}\langle X, X\rangle^{-1/2} d\langle X, X\rangle \\ &= \frac{1}{\cancel{2}}\|X\|_F^{-1}\cancel{2}\langle X, dX\rangle = \|X\|_F^{-1}\langle X, dX\rangle. \end{aligned}$$

Теперь используем полученную формулу, чтобы найти $df(X)$:

$$\begin{aligned} \boxed{df(X)} &= d(\|AX - B\|_F) = \|AX - B\|_F^{-1}\langle AX - B, d(AX - B)\rangle \\ &\boxed{= \|AX - B\|_F^{-1}\langle AX - B, AdX\rangle}. \end{aligned}$$

Чтобы найти градиент, приведем $df(X)$ к канонической форме $df(X) = \langle \nabla f(X), dX\rangle$:

$$df(X) = \langle \|AX - B\|_F^{-1}A^T(AX - B), dX\rangle \quad \Rightarrow \quad \boxed{\nabla f(X) = \|AX - B\|_F^{-1}A^T(AX - B)}.$$

---

[1] В этом примере мы работаем в пространстве симметричных матриц $\mathbb{S}^n$, поэтому знак транспонирования можно опустить.

Задача 7. Найти производную $df(X)$ и градиент $\nabla f(X)$ функции

$$f(X) := \mathrm{Tr}(AXBX^{-1}), \quad X \in \mathbb{R}^{n\times n},\ \mathrm{Det}(X) \neq 0,$$

где $A, B \in \mathbb{R}^{n\times n}$.

Решение. Для удобства перепишем след через скалярное произведение:

$$f(X) = \langle I_n, AXBX^{-1}\rangle.$$

Найдем первую производную:

$$\begin{aligned}\boxed{df(X)} &= d\langle I_n, AXBX^{-1}\rangle = \langle I_n, d(AXBX^{-1})\rangle = \langle I_n, (d(AXB))X^{-1} + (AXB)d(X^{-1})\rangle \\ &= \langle I_n, (A(dX)B)X^{-1} + (AXB)(-X^{-1}(dX)X^{-1})\rangle \boxed{= \langle I_n, A(dX)BX^{-1} - AXBX^{-1}(dX)X^{-1}\rangle}.\end{aligned}$$

Чтобы найти градиент, приведем $df(X)$ к канонической форме $df(X) = \langle \nabla f(X), dX\rangle$:

$$\begin{aligned}df(X) &= \langle I_n, A(dX)BX^{-1}\rangle - \langle I_n, AXBX^{-1}(dX)X^{-1}\rangle = \\ &= \langle A^TX^{-T}B^T, dX\rangle - \langle X^{-T}B^TX^TA^TX^{-T}, dX\rangle \\ &= \langle A^TX^{-T}B^T - X^{-T}B^TX^TA^TX^{-T}, dX\rangle \\ &\Rightarrow \quad \boxed{\nabla f(X) = A^TX^{-T}B^T - X^{-T}B^TX^TA^TX^{-T}}.\end{aligned}$$

Задача 8. Рассмотрим функцию скалярного аргумента

$$\phi(\alpha) := f(x + \alpha p), \quad \alpha \in \mathbb{R},$$

где $x, p \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ — дважды непрерывно дифференцируемая функция. Найдите первую и вторую производные $\phi'(\alpha)$ и $\phi''(\alpha)$ и выразите их через градиент $\nabla f(\cdot)$ и гессиан $\nabla^2 f(\cdot)$.

Решение. В этой задаче нужно постоянно помнить, что дифференцирование выполняется по α, а x — постоянный вектор.

Найдем первую производную:

$$\begin{aligned}d\phi(\alpha) &= d_\alpha(f(x + \alpha p)) = \{df(x) = \langle \nabla f(x), dx\rangle\} = \langle \nabla f(x + \alpha p), d_\alpha(x + \alpha p)\rangle \\ &= \langle \nabla f(x + \alpha p), (d\alpha)p\rangle = \langle \nabla f(x + \alpha p), p\rangle d\alpha.\end{aligned}$$

Здесь последнее равенство следует из того, что $d\alpha$ — это скаляр. Заметим, что мы представили $d\phi(\alpha)$ в канонической форме $d\phi(\alpha) = \phi'(\alpha)d\alpha$. Значит,

$$\boxed{\phi'(\alpha) = \langle \nabla f(x + \alpha p), p\rangle}.$$

Теперь найдем вторую производную:

$$\begin{aligned}d^2\phi(\alpha) &= d_\alpha(\langle \nabla f(x + \alpha p), p\rangle d\alpha_1) = \langle d_\alpha \nabla f(x + \alpha p), p\rangle d\alpha_1 = \{d\nabla f(x) = \nabla^2 f(x)dx\} \\ &= \langle \nabla^2 f(x + \alpha p)d_\alpha(x + \alpha p), p\rangle d\alpha_1 = \langle \nabla^2 f(x + \alpha p)(d\alpha_2)p, p\rangle d\alpha_1 \\ &= \langle \nabla^2 f(x + \alpha p)p, p\rangle d\alpha_1 d\alpha_2.\end{aligned}$$

Таким образом, из канонической формы $d^2\phi(\alpha) = \phi''(\alpha)\alpha_1\alpha_2$, получаем

$$\boxed{\phi''(\alpha) = \langle \nabla^2 f(x + \alpha p)p, p\rangle}.$$

Задача 9. Рассмотрим функцию скалярного аргумента

$$\phi(\alpha) := \|r(x + \alpha p)\|_2, \quad \alpha \in \mathbb{R}_+, \ r(x + \alpha p) \neq 0,$$

где $x, p \in \mathbb{R}^n$, $r : \mathbb{R}^n \to \mathbb{R}^m$ — дифференцируемое отображение. Найдите производную $\phi'(\alpha)$ и выразите ее через матрицу Якоби $J_r(\cdot)$.

Решение. В этой задаче, как и в предыдущей, нужно постоянно помнить, что дифференцирование выполняется по α, а x — постоянный вектор.

Найдем первую производную:

$$\begin{aligned} d\phi(\alpha) &= d_\alpha(\|r(x + \alpha p)\|_2) = \left\{ d\|x\|_2 = \frac{\langle x, dx \rangle}{\|x\|_2} \right\} = \frac{\langle r(x + \alpha p), d_\alpha r(x + \alpha p) \rangle}{\|r(x + \alpha p)\|_2} = \{dr(x) = J_r(x)dx\} \\ &= \frac{\langle r(x + \alpha p), J_r(x + \alpha p) d_\alpha (x + \alpha p) \rangle}{\|r(x + \alpha p)\|_2} = \frac{\langle r(x + \alpha p), J_r(x + \alpha p)(d\alpha) p \rangle}{\|r(x + \alpha p)\|_2} \\ &= \frac{\langle r(x + \alpha p), J_r(x + \alpha p) p \rangle}{\|r(x + \alpha p)\|_2} d\alpha. \end{aligned}$$

Отсюда

$$\boxed{\phi'(\alpha) = \frac{\langle r(x + \alpha p), J_r(x + \alpha p) p \rangle}{\|r(x + \alpha p)\|_2}}.$$

A Производные: теория

A.1 Определение

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 A.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:

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

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 A.2. 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 A.3. 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 A.4. *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 A.5. 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 A.6. Встречаются разные обозначения производной фунции f в точке x:

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

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 A.7. *Пусть f дифференцируема в точке x. Выберем произвольное направление h. Тогда:*

$$Df(x)[h] = \frac{\partial f(x)}{\partial h} := \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

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*:

$$\frac{\partial f(x)}{\partial x_i} := \lim_{t \to 0} \frac{f(x+te_i) - f(x)}{t} = Df(x)[e_i].$$

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

Пример A.8. Рассмотрим функцию $f(x) = \|x\|_2$. 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:

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

  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:

  $$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right) \in \mathbb{R}^n.$$

  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:

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

  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 A.9 (Дифференциальное исчисление)**.** *Пусть U и V — векторные пространства, X — подмножество U, $x \in X$ — внутренняя точка X. Справедливы следующие свойства:*

(a) *(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 $df(x) = 0$.*

(b) *(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 $h \in U$.*

(c) *(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*

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

(d) *(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 $h \in U$.

(e) *(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*

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

(f) *(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 A.9 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 A.10. *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.)*

A.2 Вторая производная

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 A.11. Если для функции g в некоторой точке x существует производная, то она называется *второй производной* функции f в точке x:

$$D^2f(x)[h_1, h_2] := Dg(x)[h_2].$$

Можно показать, что $D^2f(x)[h_1, h_2]$ 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:

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

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:

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

A.3 Формула Тейлора

Для дважды непрерывно-дифференцируемой функции справедлива формула Тейлора:

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

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

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

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

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

A.4 Подсчет табличных производных

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

Пример A.12 (Линейная функция)**.** Пусть $c \in \mathbb{R}^n$, и пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция $f(x) := \langle c, x \rangle$. 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:

$$f(x+h) - f(x) = \langle c, x+h \rangle - \langle c, x \rangle = \langle c, h \rangle.$$

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

Пример A.13 (Квадратичная форма)**.** Пусть $A \in \mathbb{R}^{n \times n}$, и пусть $f : \mathbb{R}^n \to \mathbb{R}$ — функция $f(x) := \langle Ax, x \rangle$. Зафиксируем произвольную точку $x \in \mathbb{R}^n$ и произвольное приращение аргумента $h \in \mathbb{R}^n$ и вычислим соответствующее приращение функции:

$$f(x+h) - f(x) = \langle A(x+h), x+h \rangle - \langle Ax, x \rangle = \langle (A + A^T)x, h \rangle + \langle Ah, h \rangle.$$

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

$$|\langle Ah, h \rangle| \le \|h\|\|Ah\| \le \|A\|\|h\|^2.$$

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 $Df(x)[h] = \langle (A + A^T)x, h \rangle$.

Пример A.14 (Обратная матрица)**.** Пусть $S := \{X \in \mathbb{R}^{n \times n} : \mathrm{Det}(X) \neq 0\}$ — 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:[2]

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

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 A.15. 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, $dI_n = d\big(XX^{-1}\big) = (dX)X^{-1} + Xd(X^{-1})$. 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 A.16 (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 $\lambda_1(X^{-1}H), \dots, \lambda_n(X^{-1}H)$ — собственные значения матрицы $X^{-1}H$ (пронумерованные в произвольном порядке и, возможно, комплексные). Заметим, что собственными значениями матрицы $I_n + X^{-1}H$ будут $1 + \lambda_1(X^{-1}H), \dots, 1 + \lambda_n(X^{-1}H)$. Поэтому

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

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

[2]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 $|q| < 1$.

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 $|\lambda| \|x\| = \|Ax\| \le \|A\| \|x\|$.) 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 $df(x)[H] = \mathrm{Det}(X)\,\mathrm{Tr}(X^{-1}H) = \mathrm{Det}(X)\langle X^{-T}, H\rangle$.

Замечание A.17. 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 $df(X)[H] = \mathrm{Det}(X)\,\mathrm{Tr}(X^{-1}H)$.