Методы оптимизации (ФКН ВШЭ, 2017). Семинар 3. Производные и условия оптимальности.

24 января 2017 г.

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.

Правила преобразования

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

Таблица стандартных производных

$$
\begin{aligned}
d(c^T x) &= c^T dx \\
d(x^T A x) &= x^T(A + A^T)dx \\
d(x^T A x) &= 2x^T A dx \qquad (\textbf{если } A = A^T) \\
d(\mathrm{Tr}(X)) &= \mathrm{Tr}(dX) \\
d(\mathrm{Det}(X)) &= \mathrm{Det}(X)\,\mathrm{Tr}(X^{-1}dX) \\
d(X^{-1}) &= -X^{-1}(dX)X^{-1}
\end{aligned}
$$

Здесь A, B — фиксированные матрицы; α — фиксированный скаляр; c — фиксированный вектор; 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 $\phi(x) := \ln(x^T A x)$, где $A \in \mathbb{S}^n_{++}$. В данном случае

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

Подставляем формально в $dg(y)$ вместо y выражение для $f(x) = x^T A x$, а вместо dy выражение

для $df(x) = 2x^T A dx$:

$$d\phi(x) = \frac{2x^T A dx}{x^T A x} \qquad \text{(В нотации с «}D\text{»-большим:} \quad D\phi(x)[\Delta x] = \frac{2x^T A \Delta x}{x^T A x}\text{).}$$

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	$df(x) = [\nabla f(x)]^T dx$ ($\nabla f(x)$: вектор; dx: вектор)	$df(x) = J_f(x)dx$ ($J_f(x)$: матрица; dx: вектор)	–
Матрица	$df(X) = \mathrm{Tr}([\nabla f(X)]^T dX)$ ($\nabla f(X)$: 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

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

$$g(x) = \frac{2x^T A dx_1}{x^T A x}$$

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

$$d^2\phi(x) = d\left(\frac{2x^T A dx_1}{x^T A x}\right) = \frac{d(2x^T A dx_1)(x^T Ax) - (2x^T A dx_1)d(x^T Ax)}{(x^T Ax)^2}$$
$$= \frac{(2(dx_2)^T A dx_1)(x^T Ax) - (2x^T A dx_1)(2x^T A dx_2)}{(x^T Ax)^2} = (dx_1)^T \left(2A - 4\frac{Axx^T A}{(x^T Ax)^2}\right) dx_2.$$

(В нотации с D-большим: $D^2\phi(x)[\Delta x_1, \Delta x_2] = (\Delta x_1)^T \left(2A - 4\frac{Axx^T A}{(x^T Ax)^2}\right) \Delta x_2$.)

Для второй производной каноническая форма для скалярной функции векторного аргумента

$$d^2 f(x) = (dx_1)^T [\nabla^2 f(x)] dx_2.$$

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

Задачи

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

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

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

Решение. Найдем первую производную:

$$\boxed{df(x)} = d\left(\frac{1}{2}x^T Ax - b^T x + c\right) = \frac{1}{2}d(x^T Ax) - d(b^T x) = \frac{1}{2}2x^T Adx - b^T dx \boxed{= (Ax - b)^T dx}.$$

Заметим, что $df(x)$ уже записан в канонической форме $df(x) = [\nabla f(x)]^T dx$, поэтому

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

Теперь найдём вторую производную:

$$\boxed{d^2 f(x)} = d\left((Ax - b)^T dx_1\right) = (d(Ax - b))^T dx_1 = (d(Ax))^T dx_1 = (Adx_2)^T dx_1 \boxed{= (dx_2)^T Adx_1}.$$

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

$$d^2 f(x) = (dx_1)^T Adx_2 \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = A}.$$

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

$$f(x) := \frac{1}{2}\|Ax - b\|_2^2, \quad x \in \mathbb{R}^n,$$

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

Решение. Найдем первую производную:

$$\boxed{df(x)} = d\left(\frac{1}{2}\|Ax - b\|_2^2\right) = \{d(\|x\|_2^2) = d(x^T x) = 2x^T dx\} = \frac{1}{2}2(Ax - b)^T d(Ax - b) \boxed{= (Ax - b)^T Adx}.$$

Чтобы найти градиент, приведем $df(x)$ к канонической форме $df(x) = [\nabla f(x)]^T dx$:

$$df(x) = (A^T(Ax - b))^T dx \quad \Rightarrow \quad \boxed{\nabla f(x) = A^T(Ax - b)}.$$

Теперь найдём вторую производную:

$$\boxed{d^2 f(x)} = d((Ax - b)^T Adx_1) = (d(Ax - b))^T Adx_1 = (Adx_2)^T Adx_1 \boxed{= (dx_2)^T A^T Adx_1}.$$

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

$$d^2 f(x) = (dx_1)^T(A^T A)dx_2 \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = A^T A}.$$

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

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

Решение. Найдем первую производную:

$$\boxed{df(x)} = d\left(\frac{1}{3}\|x\|_2^3\right) = \frac{1}{3}d((x^Tx)^{3/2}) = \frac{1}{\cancel{3}}\frac{\cancel{3}}{2}(x^Tx)^{1/2}d(x^Tx) = \frac{1}{\cancel{2}}\|x\|_2(\cancel{2}x^Tdx)\boxed{= \|x\|_2(x^Tdx)}.$$

Чтобы найти градиент, приведем $df(x)$ к канонической форме $df(x) = [\nabla f(x)]^T dx$:

$$df(x) = (\|x\|_2 x)^T dx \quad \Rightarrow \quad \boxed{\nabla f(x) = \|x\|_2 x}.$$

Теперь найдем вторую производную:

$$\begin{aligned}\boxed{d^2f(x)} &= d(\|x\|_2(x^Tdx_1)) = \underbrace{d(\|x\|_2)}_{=d((x^Tx)^{1/2})}(x^Tdx_1) + \|x\|_2 d(x^Tdx_1) \\ &= \left(\frac{1}{\cancel{2}}(x^Tx)^{-1/2}(\cancel{2}x^Tdx_2)\right)(x^Tdx_1) + \|x\|_2((dx_2)^Tdx_1) \\ &\boxed{= \|x\|_2^{-1}(x^Tdx_2)(x^Tdx_1) + \|x\|_2((dx_2)^Tdx_1)}.\end{aligned}$$

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

$$\begin{aligned}d^2f(x) &= \|x\|_2^{-1}((dx_1)^Tx)(x^Tdx_2) + \|x\|_2((dx_1)^Tdx_2) \\ &= (dx_1)^T(\|x\|_2^{-1}xx^T + \|x\|_2 I_n)dx_2 \quad \Rightarrow \quad \boxed{\nabla^2 f(x) = \|x\|_2^{-1}xx^T + \|x\|_2 I_n}.\end{aligned}$$

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 4 (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:

$$\boxed{df(x)} = d(\|x\|_2) = d((x^Tx)^{1/2}) = \frac{1}{2}(x^Tx)^{-1/2}d(x^Tx) = \frac{1}{\cancel{2}}\|x\|_2^{-1}(\cancel{2}x^Tdx)\boxed{= \|x\|_2^{-1}(x^Tdx)}.$$

Чтобы найти градиент, приведем $df(x)$ к канонической форме $df(x) = [\nabla f(x)]^T dx$:

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

Теперь найдем вторую производную:

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

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

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

Задача 5 (Логистическая функция)**.** Найти первую и вторую производные $df(x)$ и $d^2f(x)$, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

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

где $a \in \mathbb{R}^n$.

Решение. Найдем первую производную:

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

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 $df(x) = [\nabla f(x)]^T dx$:

$$df(x) = (\sigma(a^T x)a)^T dx \quad \Rightarrow \quad \boxed{\nabla f(x) = \sigma(a^T x)a}.$$

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

Теперь найдем вторую производную:

$$\begin{aligned}\boxed{d^2f(x)} &= d(\sigma(a^T x)(a^T dx_1)) = d(\sigma(a^T x))(a^T dx_1) = \{d(\sigma(x)) = \sigma'(x)dx\} = (\sigma'(a^T x)d(a^T x))(a^T dx_1) \\ &= \sigma'(a^T x)(a^T dx_2)(a^T dx_1) = \{\sigma'(x) = \sigma(x)(1 - \sigma(x))\} \\ &\boxed{= \sigma(a^T x)(1 - \sigma(a^T x))(a^T dx_2)(a^T dx_1)}.\end{aligned}$$

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

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

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

Задача 6 (Логарифм определителя)**.** Найти первую и вторую производные $df(X)$ и $d^2f(X)$, а также градиент $\nabla f(X)$ функции

$$f(X) := \ln(\mathrm{Det}(X)), \quad X \in \mathbb{R}^{n\times n},\ \mathrm{Det}(X) > 0.$$

Решение. Найдём первую производную:

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

Чтобы найти градиент, приведем $df(X)$ к канонической форме $df(X) = \mathrm{Tr}([\nabla f(X)]^T dX)$:

$$df(X) = \mathrm{Tr}((X^{-T})^T dX) \quad \Rightarrow \quad \boxed{\nabla f(X) = X^{-T}}.$$

Теперь найдём вторую производную:

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

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

Задача 7. 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 $d(\|X\|_F)$:

$$\begin{aligned} d(\|X\|_F) &= d((\mathrm{Tr}(X^TX))^{1/2}) = \left\{ d(x^{1/2}) = \frac{1}{2}x^{-1/2}dx \right\} = \frac{1}{2}(\mathrm{Tr}(X^TX))^{-1/2} d(\mathrm{Tr}(X^TX)) \\ &= \frac{1}{2}\|X\|_F^{-1}\,\mathrm{Tr}(d(X^TX)) = \frac{1}{2}\|X\|_F^{-1}\,\mathrm{Tr}((dX)^TX + X^TdX) \\ &= \frac{1}{2}\|X\|_F^{-1}(\mathrm{Tr}((dX)^TX) + \mathrm{Tr}(X^TdX)) = \{\mathrm{Tr}(A) = \mathrm{Tr}(A^T)\} \\ &= \frac{1}{\cancel{2}}\|X\|_F^{-1}\cancel{2}\,\mathrm{Tr}(X^TdX) = \|X\|_F^{-1}\,\mathrm{Tr}(X^TdX). \end{aligned}$$

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

$$\begin{aligned} \boxed{df(X)} &= d(\|AX - B\|_F) = \|AX - B\|_F^{-1}\,\mathrm{Tr}((AX - B)^T d(AX - B)) \\ &\boxed{= \|AX - B\|_F^{-1}\,\mathrm{Tr}((AX - B)^T A dX)}. \end{aligned}$$

Чтобы найти градиент, приведем $df(X)$ к канонической форме $df(X) = \mathrm{Tr}([\nabla f(X)]^T dX)$:

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

Задача 8. 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. Найдем первую производную:

$$\boxed{df(X)} = d(\mathrm{Tr}(AXBX^{-1})) = \mathrm{Tr}(d(AXBX^{-1})) = \mathrm{Tr}((d(AXB))X^{-1} + (AXB)d(X^{-1}))$$
$$= \mathrm{Tr}((A(dX)B)X^{-1} + (AXB)(-X^{-1}(dX)X^{-1}))\boxed{= \mathrm{Tr}(A(dX)BX^{-1} - AXBX^{-1}(dX)X^{-1})}.$$

Чтобы найти градиент, приведем $df(X)$ к канонической форме $df(X) = \mathrm{Tr}([\nabla f(X)]^T dX)$:

$$df(X) = \mathrm{Tr}(A(dX)BX^{-1}) - \mathrm{Tr}(AXBX^{-1}(dX)X^{-1}) = \{\mathrm{Tr}(AB) = \mathrm{Tr}(BA)\}$$
$$= \mathrm{Tr}(BX^{-1}AdX) - \mathrm{Tr}(X^{-1}AXBX^{-1}dX) = \mathrm{Tr}((BX^{-1}A - X^{-1}AXBX^{-1})dX)$$
$$= \mathrm{Tr}((A^T X^{-T} B^T - X^{-T} B^T X^T A^T X^{-T})^T dX)$$
$$\Rightarrow \quad \boxed{\nabla f(X) = A^T X^{-T} B^T - X^{-T} B^T X^T A^T X^{-T}}.$$

Задача 9. 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:

$$d\phi(\alpha) = d_\alpha(f(x + \alpha p)) = \{df(x) = \nabla f(x)^T dx\} = \nabla f(x + \alpha p)^T d_\alpha(x + \alpha p)$$
$$= \nabla f(x + \alpha p)^T((d\alpha)p) = (\nabla f(x + \alpha p)^T p)d\alpha.$$

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,

$$\boxed{\phi'(\alpha) = \nabla f(x + \alpha p)^T p}.$$

Теперь найдем вторую производную:

$$d^2\phi(\alpha) = d_\alpha((\nabla f(x + \alpha p)^T p)d\alpha_1) = ((d_\alpha(\nabla f(x + \alpha p)))^T p)d\alpha_1 = \{d(\nabla f(x)) = \nabla^2 f(x)dx\}$$
$$= ((\nabla^2 f(x + \alpha p)d_\alpha(x + \alpha p))^T p)d\alpha_1 = ((\nabla^2 f(x + \alpha p)(d\alpha_2)p)^T p)d\alpha_1$$
$$= (p^T \nabla^2 f(x + \alpha p)p)d\alpha_1 d\alpha_2.$$

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

$$\boxed{\phi''(\alpha) = p^T[\nabla^2 f(x + \alpha p)]p}.$$

Задача 10. 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:

$$d\phi(\alpha) = d_\alpha(\|r(x+\alpha p)\|_2) = \left\{ d\|x\|_2 = \frac{x^T dx}{\|x\|_2} \right\} = \frac{r(x+\alpha p)^T d_\alpha(r(x+\alpha p))}{\|r(x+\alpha p)\|_2} = \{dr(x) = J_r(x)dx\}$$
$$= \frac{r(x+\alpha p)^T (J_r(x+\alpha p) d_\alpha(x+\alpha p))}{\|r(x+\alpha p)\|_2} = \frac{r(x+\alpha p)^T (J_r(x+\alpha p)((d\alpha)p))}{\|r(x+\alpha p)\|_2}$$
$$= \frac{r(x+\alpha p)^T J_r(x+\alpha p)p}{\|r(x+\alpha p)\|_2} d\alpha.$$

Отсюда

$$\boxed{\phi'(\alpha) = \frac{r(x+\alpha p)^T J_r(x+\alpha p)p}{\|r(x+\alpha p)\|_2}}.$$

Задача 11 (Регрессия наименьших квадратов)**.** Рассмотрим следующую задачу оптимизации:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2,$$

где $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\mathrm{Rank}(A) = n$. Найдите множество ее решений и оптимальное значение целевой функции.

Решение. Прежде всего, перейдем от исходной негладкой задачи к эквивалентной ей гладкой:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2. \tag{1}$$

Заметим, что при таком переходе меняется лишь оптимальное значение целевой функции, а множество оптимальных решений остается неизменным.

Чтобы найти решения (1), воспользуемся условием оптимальности первого порядка:

$$\nabla(\|Ax - b\|_2^2) = 2A^T(Ax - b) = 0.$$

Таким образом, все решения задачи (1), содержатся среди решений следующей системы линейных уравнений:

$$A^T Ax = A^T b. \tag{2}$$

Уравнения (2) называются *нормальными уравнениями*. Заметим, что матрица $A^T A$ является обратимой, поскольку она имеет полный ранг: согласно условию, $\mathrm{Rank}(A) = n$, и из линейной алгебры известно, что $\mathrm{Rank}(A^T A) = \mathrm{Rank}(A)$. Отсюда заключаем, что нормальные уравнения (2) имеют единственное решение

$$\boxed{\hat{x} = (A^T A)^{-1} A^T b}.$$

Заметим, что пока еще нельзя утверждать, что $\hat{x}$ является решением задачи (1). Все, что мы показали — это что $\hat{x}$ является стационарной точкой целевой функции. Как известно, стационарная точка может быть, а может и не быть глобальным минимумом. Напомним, что условие оптимальности первого порядка говорит о том, что, если (хотя бы один) глобальный минимум существует, то он обязательно является точкой стационарности. Таким образом, если задача (1) вообще не имеет решений, то условие оптимальности первого порядка здесь применять абсолютно бессмысленно, даже несмотря на то, что точки стационарности существуют. (Например, одномерная функция x^3 не имеет глобального минимума на $\mathbb{R}$, поскольку стремится к $-\infty$ при $x \to -\infty$, однако имеет точку стационарности $x = 0$.) Итак, откуда мы знаем, что задача (1) разрешима, и найденная точка $\hat{x}$ является ее решением? На самом деле, причина здесь кроется в том, что задача (1) обладает весьма важным (и нетривиальным) свойством *выпуклости*, а для выпуклых задач понятия глобального минимума и

стационарной точки эквивалентны. Однако, поскольку мы пока еще не изучали выпуклость, то приведем здесь прямое доказательство этого факта конкретно для задачи (1).

Покажем, что если некоторая точка $\hat{x}$ является стационарной (т. е. удовлетворяет системе (2), то $\hat{x}$ обязательно является глобальным минимумом в задаче (1). Пусть $x \in \mathbb{R}^n$. Тогда

$$\|Ax - b\|_2^2 = \|A(x - \hat{x}) + (A\hat{x} - b)\|_2^2 = \|A(x - \hat{x})\|_2^2 + \|A\hat{x} - b\|_2^2 + 2(x - \hat{x})^T A^T (A\hat{x} - b).$$

Заметим, что последнее слагаемое в правой части равно нулю, поскольку $\hat{x}$ удовлетворяет системе (2). Таким образом, для любого $x \in \mathbb{R}^n$ верно следующее разложение:

$$\|Ax - b\|_2^2 = \|A\hat{x} - b\|_2^2 + \|A(x - \hat{x})\|_2^2.$$

Поскольку второе слагаемое всегда неотрицательное, то это означает (по определению), что точка $\hat{x}$ является глобальным минимумом в задаче (1).

Таким образом, найденная точка $\hat{x}$ является единственным решением исходной задачи. Оптимальное значение целевой функции при этом равно

$$\boxed{\text{Opt}} := \|A\hat{x} - b\|_2 \boxed{= \|A(A^T A)^{-1} A^T b - b\|_2}.$$

Матрица $(A^T A)^{-1} A^T$ называется *псевдообратной* (по Муру-Пенроузу) и обозначается A^+. Это прямое обобщение понятия обратной матрицы для неквадратных матриц. Если A квадратная и обратимая, то псевдообратная матрица A^+ совпадает с обратной матрицей A^{-1}.

Матрица $A\left(A^T A\right)^{-1} A^T$ называется *проекционной матрицей* и проектирует заданный вектор b на линейную оболочку, натянутую на столбцы матрицы A.

Отметим, что задача (1) имеет решения всегда, даже если $\text{Rank}(A) < n$. Действительно, как было показано выше, задача (1) разрешима тогда и только тогда, когда разрешимы нормальные уравнения (2). Нормальные уравнения (2) разрешимы тогда и только тогда, когда $A^T b \in \text{Im}(A^T A)$, где $\text{Im}(A^T A)$ — образ матрицы $A^T A$. Из линейной алгебры известно, что $\text{Im}(A^T A) = \text{Im}(A^T)$. Значит, условие $A^T b \in \text{Im}(A^T A)$ всегда выполняется. Итак, задача (1) всегда разрешима, и в случае $\text{Rank}(A) < n$ имеет бесконечное множество решений. (Одно из возможных решений записывается через псевдообратную матрицу.)