Методы оптимизации. Семинар 1. Матрично-векторное дифференцирование.

Корнилов Никита Максимович

МФТИ ФИВТ

4 сентября 2025г

Матрицы и векторы

Мы будем работать с векторами и матрицами:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d, \quad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{pmatrix} \in \mathbb{R}^{n\times m}.$$

- Складывать можно только матрицы одинаковых размерностей $A, B \in \mathbb{R}^{n\times m} : A + B \in \mathbb{R}^{n\times m}$! Вектора тоже $x, y \in \mathbb{R}^d : x + y \in \mathbb{R}^d$!
- Перемножать матрицы A, B разных размерностей можно только если $A \in \mathbb{R}^{n\times m}$ и $B \in \mathbb{R}^{m\times k} : AB \in \mathbb{R}^{n\times k}$!!!
- В общем случае, переставлять квадратные матрицы $A, B \in \mathbb{R}^{n\times n}$ при умножении нельзя: $AB \neq BA$!!!
- Умножить матрицу $A \in \mathbb{R}^{n\times m}$ справа на вектор x можно только если $x \in \mathbb{R}^m : Ax \in \mathbb{R}^n$!!!
- Слева матрицу $A \in \mathbb{R}^{n\times m}$ можно умножить на строку $y^\top$, где $y \in \mathbb{R}^n$:$y^\top A \in \mathbb{R}^{1\times m}$.

След матрицы

След квадратной матрицы $A \in \mathbb{R}^{n \times n}$ обозначается как $Tr : \mathbb{R}^{n \times n} \to \mathbb{R}$ и считается по формуле:

$$Tr(A) = \sum_{i=1}^{n} a_{ii}.$$

Свойства следа:

1. $Tr(A^\top) = Tr(A), \quad A \in \mathbb{R}^{n \times n},$
2. $Tr(A + B) = Tr(A) + Tr(B), \quad A, B \in \mathbb{R}^{n \times n},$
3. $Tr(cA) = cTr(A), \quad A \in \mathbb{R}^{n \times n}, c \in \mathbb{R},$
4. Циклическое свойство:
   $Tr(A_1 A_2 \dots A_{k-1} A_k) = Tr(A_k A_1 A_2 \dots A_{k-1}), \quad A_1, \dots, A_k \in \mathbb{R}^{n \times n}.$

Скалярное произведение

- Стандартное скалярное произведение $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ считается по формуле

$$\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i = x^\top y, \quad x, y \in \mathbb{R}^d.$$

Для матриц скалярное произведение $\langle \cdot, \cdot \rangle : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \to \mathbb{R}$ определено как

$$\langle X, Y \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} Y_{ij} = \langle Y, X \rangle = \mathrm{Tr}(X^\top Y), \quad X, Y \in \mathbb{R}^{n \times m}.$$

Перестановка матрицы A в скалярном произведении:

$$\langle Ax, y \rangle = \langle x, A^\top y \rangle \quad \langle AX, Y \rangle = \langle X, A^\top Y \rangle.$$

- Поэлементное умножение одинаковых по размерностям матриц обозначается как $\odot : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m}$:

$$(X \odot Y)_{ij} = X_{ij} * Y_{ij}, \quad X, Y \in \mathbb{R}^{n \times m}.$$

Симметричные и определенные матрицы

- Множество симметричных матриц $\mathbb{S}^n$:

$$A \in \mathbb{S}^n \quad \Longleftrightarrow \quad A = A^\top.$$

- Множество положительно определённых $\mathbb{S}^n_{++}$:

$$A \in \mathbb{S}^n_{++} \quad \Longleftrightarrow \quad A \in \mathbb{S}^n; \quad \forall x \neq 0 \in \mathbb{R}^n : \quad x^\top A x > 0.$$

  Критерий Сильвестра: все угловые миноры имеют положительный определитель.

- Множество положительно полуопределённых $\mathbb{S}^n_{+}$:

$$A \in \mathbb{S}^n_{+} \quad \Longleftrightarrow \quad A \in \mathbb{S}^n; \quad \forall x \in \mathbb{R}^n : \quad x^\top A x \geq 0.$$

  Критерий Арнольда: все *главные* миноры имеют неотрицательный определитель. Главным минором называется определитель подматрицы, симметричной относительно главной диагонали.

Собственные числа

Для квадратичной матрицы $A \in \mathbb{R}^{n \times n}$ существует n (возможно комплексных) собственных чисел $\{\lambda_i(A)\}_{i=1}^n$, где $\lambda_i(A)$ — i-ое по модулю собственное число.

- Собственное значение $\lambda \in \mathbb{C}$ и собственный вектор $x \neq 0 \in \mathbb{C}^d$:

$$Ax = \lambda x \iff \det(A - \lambda I) = 0.$$

- Определитель и след матрицы A можно выразить через её собственные значения

$$\det(A) = \prod_{i=1}^n \lambda_i(A), \quad \mathrm{Tr}(A) = \sum_{i=1}^n \lambda_i(A).$$

- Для любой симметричной матрицы $A \in \mathbb{S}^n$ существует действительный ортонормированный базис из собственных векторов $S \in \mathbb{R}^{n \times n}$ и действительных собс. чисел:

$$A = S\Sigma S^\top, \quad S^\top S = I, \quad \Sigma - \text{диагональная с собс. числами}.$$

Векторные нормы

Норма $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ — функция, удовлетворяющая свойствам: Неотрицательность, Умножение на скаляр, Неравенство треугольника.

- p-норма $\|\cdot\|_p$ на $\mathbb{R}^d$ для $p \in [1, +\infty]$:

$$\|x\|_p := \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}}, \quad \|x\|_\infty = \max_{i\in[1,d]} |x_i|.$$

Частные случаи: $\|x\|_2 = \sqrt{\sum\limits_{i=1}^{d} |x_i|^2} = \sqrt{\langle x, x\rangle}, \quad \|x\|_1 = \sum\limits_{i=1}^{d} |x_i|.$

- Сопряженная норма $\|\cdot\|_*$ для любой нормы $\|\cdot\|$: $\|y\|_* := \sup_{\|x\|\le 1}\langle x, y\rangle$. Для p-нормы сопряженная это q-норма, $\frac{1}{p} + \frac{1}{q} = 1$.
- Скалярное произведение $\langle\cdot,\cdot\rangle_A$ и квадратичная норма $\|\cdot\|_A$, определенная пол. опр. матрицей $A \in \mathbb{S}^d_+$:

$$\langle x, y\rangle_A := x^\top A y, \quad \|x\|_A = \sqrt{\langle x, x\rangle_A}.$$

Свойства норм

- Неравенства Гёльдера и Коши-Буняковского:
Для векторов $x, y \in \mathbb{R}^d$ и чисел $p \in [1, +\infty]$, $\frac{1}{q} + \frac{1}{p} = 1$ выполняется неравенство

$$|\langle x, y\rangle| \leq \|x\|_p \|y\|_q.$$

В частности неравенство КБШ: $|\langle x, y\rangle| \leq \|x\|_2 \|y\|_2$.

- Все нормы в $\mathbb{R}^d$ эквивалентны, например, для $p < p' \in [1, +\infty]$:

$$\|x\|_{p'} \leq \|x\|_p \leq d^{\frac{1}{p} - \frac{1}{p'}} \|x\|_{p'}, \quad \forall x \in \mathbb{R}^d.$$

В частности, имеем:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2, \quad \forall x \in \mathbb{R}^d.$$

Матричные нормы

Норма $\|\cdot\|_p : \mathbb{R}^{n\times m} \to \mathbb{R}$ — функция, удовлетворяющая свойствам: Неотрицательность, Умножение на скаляр, Неравенство треугольника.

- Матричная норма $\|\cdot\|_p$ для $A \in \mathbb{R}^{n\times m}$, индуцированная векторной нормой $\|\cdot\|_p$, определятся как

$$\|A\|_p := \sup_{\|x\|_p=1} \|Ax\|_p.$$

  Можно привести замкнутые формы для классических норм
  - $\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^m |a_{ij}|$,
  - $\|A\|_1 = \max_{1\le j\le m} \sum_{i=1}^n |a_{ij}|$,
  - $\|A\|_2 = \sup\limits_{\langle x,x\rangle=1} \sqrt{\langle Ax, Ax\rangle} = \sqrt{\lambda_{\max}(A^\top A)}$.
- Норма Фробениуса для матрицы $A \in \mathbb{R}^{n\times m}$ определяется как

$$\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^m A_{ij}^2 := \mathrm{Tr}(A^\top A) = \sum_{i=1}^m \lambda_i(A^\top A).$$

SVD разложение матрицы

Числа $\sigma_i(A) := \sqrt{\lambda_i(A^\top A)} \geq 0, i \in \overline{1, m}$ называются сингулярными числами $A \in \mathbb{R}^{n \times m}$. В случае симметричной матрицы $A \in \mathbb{S}^m$: $\sigma_i(A) = |\lambda_i(A)|, i \in \overline{1, m}$.

Lemma

Матрица $A \in \mathbb{R}^{n \times m}$ представима в виде SVD-разложения:

$$A = U \Sigma V^\top,$$

где $U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{m \times m}$ — ортогональные, $\Sigma \in \mathbb{R}^{n \times m}$ — диагональная матрица, составленная из сингулярных чисел $\sigma_i(A) := \sqrt{\lambda_i(A^\top A)}, i \in \overline{1, m}$, расположенных в порядке убывания.

V - базис из собственных векторов $A^\top A$, U - базис из собственных векторов матрицы $AA^\top$.

Свойства матричных норм

- Индуцированные и Фробениусова нормы удовлетворяют свойству субмультипликативности (в общем случае - нет)

$$\|AB\| \leq \|A\|\|B\|, \quad A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times k}.$$

- Для любой $A \in \mathbb{R}^{n\times n}$ и индуцированной или Фробениусовой нормы $\|\cdot\|$ верно $|\lambda_{max}(A)| \leq \|A\|$.
- Для любой ортогональной S и нормы Фробениуса верно

$$\|SA\|_F = \|A\|_F.$$

- $\|A\|_2^2 \leq \|A\|_\infty \|A\|_1$.
- $\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2$.

Дифференцируемость по Фреше

Работаем в множествах U, V линейных нормированных полных (банаховых) пространств.

Definition

Функция $f : U \to V$ дифференцируема по Фреше во внутренней точке $x \in \mathrm{int} U$, если существует линейный оператор $df(x) : U \to V$, т.ч.

$$f(x+h) = f(x) + L[h] + o(\|h\|), \|h\| \to 0.$$

$df(x)$ называется производной f в точке x.

Если точка x не является внутренней, то понятие дифференцируемости не определено.

Definition

Приращение дифференцируемой функции f в точке x с приращением h называется дифференциалом $df(x)[h] \in V$. Часто направление h обозначают как dx, а дифференциал как $df(x)[dx]$.

Производная по направлению

В одномерном случае $f : \mathbb{R} \to \mathbb{R}$, показателем скорости изменения f в точке x вдоль числовой прямой является производная:

$$f'(x) := \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}.$$

В многомерном случае $f : U \to V$, направлений изменения не два, а бесконечно много. Производные по направлению отвечают за изменения функции вдоль одного направления $h \in U$:

Definition

Производной по направлению $h \in U$ функции $f : U \to V$ во внутреней точке $x \in \operatorname{int} U$ называется

$$\frac{\partial f}{\partial h}(x) := \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}. \tag{1}$$

Если для любого $h \in U$ определена производная по направлению $\frac{\partial f}{\partial h}(x)$, то функция f дифференцируема по Гато в точке x.

Связь определений

Lemma

Если функция f дифференцируема по Фреше в x, то производная по направлению $\frac{\partial f}{\partial h}(x)$ линейна по h и равна дифференциалу $df(x)[h]$.

В классическом матанализе, из дифференцируемости следует существование производных по всем направлением. Однако обратное неверно. Достаточным условием будет непрерывность всех частных производных.

Градиент по вектору

- В случае $f : \mathbb{R}^n \to \mathbb{R}$ дифференциал $df(x)[dx] \in \mathbb{R}$ можно представить в виде

$$df(x)[dx] = \langle \nabla f(x), dx \rangle, \quad \text{где вектор } \nabla f(x) \in \mathbb{R}^n \text{ зависит от } x.$$

Вектор $\nabla f(x)$ называется **градиентом** функции. Взяв $h = e_i = (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0) \in \mathbb{R}^n$, получим формулу градиента в стандартном базисе

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)^\top \in \mathbb{R}^n, \tag{2}$$

где $\frac{\partial f}{\partial x_i}(x) := \lim_{t \to 0} \frac{f(x+te_i) - f(x)}{t}$ — частные производные по i-ой координате.

Градиент по матрице

- В случае $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ дифференциал $df(X)[dX] \in \mathbb{R}$ можно представить в виде

$$df(X)[dX] = \langle \nabla f(X), dX \rangle,$$

где матрица $\nabla f(X) \in \mathbb{R}^{n\times m}$ зависит от X.
Матрица $\nabla f(X)$ также называется **градиентом** функции.
Аналогично взяв $h = e_{ij}$, получим формулу градиента в стандартном базисе

$$\nabla f(X) = \left(\frac{\partial f}{\partial x_{ij}}(X) \right)_{i,j} \in \mathbb{R}^{n\times m}. \tag{3}$$

Example

Найдите градиент $\nabla f(x)$ функции

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,$$

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

Матрица Якоби

- В случае $f : \mathbb{R}^n \to \mathbb{R}^m$ дифференциал $df(x)[dx] \in \mathbb{R}^m$ можно представить в виде

$$df(x)[dx] = J_f(x)dx, \quad \text{где матрица } J_f(x) \in \mathbb{R}^{m\times n} \text{ зависит от } x.$$

Матрица $J_f(x)$ называется матрицей Якоби.
Аналогично взяв $h = e_i$, получим формулу матрицы Якоби в стандартном базисе

$$J_f(x) \equiv \frac{\partial f}{\partial x} := \left(\frac{\partial f_i}{\partial x_j}(x)\right)_{i,j} \in \mathbb{R}^{m\times n}. \tag{4}$$

Заметим, что $\nabla f(x) = J_x^\top$.

Таблица канонических видов

Вход \ Выход	Скаляр	Вектор
Скаляр	$df(x) = f'(x)dx$ $f'(x)$ скаляр, dx скаляр.	-
Вектор	$df(x) = \langle \nabla f(x), dx \rangle$ $f(x)$ вектор, dx вектор	$df(x) = J_x dx$ J_x матрица, dx вектор
Матрица	$df(X) = \langle \nabla f(X), dX \rangle$ $\nabla f(X)$ мат, dX мат	-

Стоит отметить, что данная таблица верна и для произвольных скалярных произведений, а не только для стандартного.

1. **Прямой подход**
   Идея: выразить функцию $f(x)$ через скалярную зависимость от каждой координаты x_i и напрямую искать частную производную $\frac{\partial f(x)}{\partial x_i}$.
2. **Дифференциальный подход**
   Идея: Используя правила вычисления дифференциалов, получить канонический вид из Таблицы (19) и выделить градиенты функций, гессиан или матрицу Якоби.

Дифференциальное исчисление: правила

Правила преобразования
$d(\alpha X) = \alpha dX$ $d(AXB) = AdXB$ $d(X+Y) = dX + dY$ $d(X^T) = (dX)^T$
$d(XY) = (dX)Y + X(dY)$ $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
$d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$
$d(g(f(x))) = dg(f)[df(x)]$ $J_{g(f)} = J_g J_f \iff \frac{\partial g}{\partial x} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$
$df(x,y) = J_x dx + J_y dy$

Стоит отметить, что данная таблица верна и для произвольных скалярных произведений и гельдоровых ЛНП.

Таблица стандартных производных
$dA = 0$ $d\langle A, X\rangle = \langle A, dX\rangle$ $d\langle Ax, x\rangle = \langle (A + A^\top)x, dx\rangle$ $d\,\mathrm{Tr}(X) = \mathrm{Tr}(dX)$
$d(\det(X)) = \det(X)\,\mathrm{Tr}(X^{-1}dX)$
$d(X^{-1}) = -X^{-1}(dX)X^{-1}$

Стоит отметить, что данная таблица верна и для произвольных скалярных произведений и гельдоровых ЛНП.

Hint. Для запоминания формулы $d(X^{-1})$

$$
\begin{aligned}
&I = XX^{-1},\\
&dI = 0 = d(XX^{-1}) = (dX)X^{-1} + Xd(X^{-1}),\\
&d(X^{-1}) = -X^{-1}(dX)X^{-1}.
\end{aligned}
$$

Однако это не является доказательством **существования** дифференциала.

Example

Найдите первый дифференциал и градиент функции $\nabla f(x)$ для

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,$$

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

Вторая производная

Пусть $f : U \to V$ дифференцируема в каждой точке $x \in U$. Рассмотрим дифференциал функции $df(x)[h_1]$ при фиксированном приращении $h_1 \in U$ как функцию от x:

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

Definition (Вторая производная)

Если в некоторой точке $x \in U$ функция $g : U \to V$ дифференцируема, то второй дифференциал $d^2 f(x)[h_1, h_2] : U \times U \to V$ имеет вид

$$d^2 f(x)[h_1, h_2] := d(df[h_1])(x)[h_2]. \tag{5}$$

Можно показать, что $d^2 f(x)[h_1, h_2]$ билинейная функция по h_1, h_2. По аналогии определяется третий дифференциал $d^3 f(x)[h_1, h_2, h_3]$, четвёртый и так далее.

Гессиан

В случае $f : \mathbb{R}^n \to \mathbb{R}$ второй дифференциал, как и любую билинейную функцию, можно представить с помощью матрицы

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

Матрица $\nabla^2 f(x)$ называется **гессианом** функции. В стандартном базисе гессиан имеет вид

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{ij} \in \mathbb{R}^{n \times n}.$$

Напомним, что для дважды непрерывно дифференцируемой функции гессиан - симметричная матрица. В общем случае, удобно считать гессиан как

$$\nabla^2 f(x) = (J_{\nabla f})^\top.$$

Квадратичная функция

Example

Найдите второй дифференциал и гессиан функции $\nabla^2 f(x)$ для

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,$$

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

Example

Найдите первый и второй дифференциалы, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \ln\langle Ax, x\rangle$$

где $x \in \mathbb{R}^n, A \in \mathbb{S}^n_{++}$.

Example

Найдите первый и второй дифференциалы, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \frac{1}{3}\|x\|_2^3, \quad x \in \mathbb{R}^n \setminus \{0\}.$$

Example

Найдите первый и второй дифференциалы, а также градиент $\nabla f(x)$ и гессиан $\nabla^2 f(x)$ функции

$$f(x) = \ln(1 + \exp(\langle a, x \rangle)),$$

где $a \in \mathbb{R}^n$.

Softmax

Example

Найдите матрицу Якоби функции $s(x) = \mathrm{softmax}(x)$

$$\mathrm{softmax}(x) := \left(\frac{\exp(x_1)}{\sum_{i=1}^{n}\exp(x_i)}, \ldots, \frac{\exp(x_n)}{\sum_{i=1}^{n}\exp(x_i)}\right)^{\top}.$$

Example

Найти градиент $\nabla f(X)$ и дифференциал функции $f(X)$

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

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

Практика

Example

Найдите первый дифференциал и градиент $\nabla f(X)$ функции $f(X)$

$$f(X) = \det(AX^{-1}B),$$

где A, X, B – такие матрицы с нужными размерностями, что $AX^{-1}B$ обратима.

Логарифм определителя

Example

Найдите первый и второй дифференциалы, а также градиент $\nabla f(X)$ функции $f(X)$

$$f(X) = \ln(\det(X))$$

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

Example

Найти градиент $\nabla f(X)$ и дифференциал функции $f(X)$

$$f(X) = \mathrm{Tr}(AX^\top X).$$