Методы оптимизации. Семинар 1. Матрично-векторное дифференцирование.

Корнилов Никита Максимович

МФТИ ФИВТ

3 сентября 2024г

Матрицы и векторы

Мы будем работать с векторами и матрицами:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, \quad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{pmatrix}.$$

- Складываем матрицы только одинаковых размерностей!
- Умножить матрицу A справа на вектор x можно только если $A \in \mathbb{R}^{n \times m}$ и $x \in \mathbb{R}^m$!!! Матрица умножается слева на строку $x^\top$.
- Перемножать матрицы A, B разных размерностей можно только если $A \in \mathbb{R}^{n \times m}$ и $B \in \mathbb{R}^{m \times k}$!!!
- В общем случае, переставлять матрицы при умножении нельзя:

$$AB \neq BA!!!$$

Норма

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

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

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

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

- Все нормы в $\mathbb{R}^n$ эквивалентны, к примеру,

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2.$$

Свойства опперации

- Циклическое свойство следа для матриц A, B, C, D

$$\mathrm{Tr}(ABCD) = \mathrm{Tr}(DABC) = \mathrm{Tr}(CDAB) = \mathrm{Tr}(BCDA).$$

Перестановка матрицы в скалярном произведении, x, y — векторы:

$$\langle Ax, y \rangle = \langle x, A^\top y \rangle \quad \langle AB, C \rangle = \langle B, A^\top C \rangle = \langle A, CB^\top \rangle$$

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

$$A \in \mathbb{S}^n \quad \iff \quad A = A^\top,$$

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

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

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

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

Скалярное произведение

- Скалярное произведение в $\mathbb{R}^n$ считается по формуле

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

$$\|x\|_2^2 = \langle x, x \rangle.$$

  В случае матриц из $\mathbb{R}^{n \times m}$ скалярное произведение в стандартном базисе определено как

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

- Поэлементное умножение одинаковых по размерностям матриц обозначается как $\odot$

$$(A \odot B)_{ij} = A_{ij} * B_{ij}.$$

Собственные числа

- Собственное значение $\lambda \in \mathbb{C}$:

$$\exists x \neq 0 : \quad Ax = \lambda x \Longleftrightarrow \det(A - \lambda I) = 0.$$

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

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

где $\lambda_i(A)$ — i-ое по модулю собственное число.

- Для симметричных матриц существует ОНБ из действительных собственных векторов:

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

Матричные нормы

Индуцированная векторной нормой $\|\cdot\|_p$ матричная норма для $A \in \mathbb{R}^{n\times m}$ определятся как

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

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

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

Матричные нормы

Свойства:

- Все нормы выше удовлетворяют свойству субмультипликативности

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

- $A \in \mathbb{R}^{n\times n}, \quad |\lambda_{max}(A)| \leq \|A\|,$
- Для любой ортогональной S и нормы Фробениуса верно

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

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

Производная

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 $f : \mathbb{R}^n \to \mathbb{R}^m$, направлений изменения не два, а бесконечно много. Производные по направлению отвечают за изменения функции вдоль одного направления $h \in \mathbb{R}^n$:

Definition

Производной по направлению h функции f в точке x называется

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

Дифференцируемость

Definition

Функция f дифференцируема во внутренней точке x, если существует линейный оператор $L : \mathbb{R}^n \to \mathbb{R}^m$, такой что

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

L называется производной f в точке x и обозначается как $f'(x)$.

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

Definition

Дифференциалом $df(x)[h] \in \mathbb{R}^m$ в точке x дифференцируемой функции f и с приращением h называется вектор $f'(x)[h]$. Часто направление h обозначают как dx, а $f'(x)[h]$ как $df(x)$ или $df(x)[h], df(x)[dx]$.

Дифференцируемость

Definition

Если функция f дифференцируема в x, то произвольная 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.

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

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

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 $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ линейную функцию $df(X)[dX]$ можно представить в виде

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

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

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

Квадратичная функция

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 $f : \mathbb{R}^n \to \mathbb{R}^m$ линейный оператор $df(x)[dx]$ можно представить в виде

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

Матрица $J_x(x)$ 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 (16) и выделить градиенты функций, гессиан или матрицу Якоби.

Дифференциальное исчисление: правила

Правила преобразования
$d(\alpha X) = \alpha dX$ $d(AXB) = AdXB$ $d(X+Y) = dX + dY$ $d(X^T) = (dX)^T$
$d(XY) = (dX)Y + X(dY)$ $d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
$d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$
$d(g(f(x))) = g'(f)df(x)$ $J_{g(f)} = J_g J_f \iff \frac{\partial g}{\partial x} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$
$df(x,y) = J_x dx + J_y dy$

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

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, а не только для стандартного.

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.

Квадратичная функция

Example

Найдите первый $df(x)$ дифференциал функции

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

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

Вторая производная

Пусть f дифференцируема в каждой точке x. Рассмотрим дифференциал функции f при фиксированном приращении h_1 как функцию от x:

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

Definition (Вторая производная)

Если в некоторой точке x функция g дифференцируема, то второй дифференциал имеет вид

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

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

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

Example

Найдите второй $d^2 f(x)$ дифференциал функции

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

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

Example

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

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

Example

Найдите первый и второй дифференциал $df(x), d^2f(x)$, 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 $s(x) = \text{softmax}(x)$

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

Фробениусова норма

Example

Найти градиент $\nabla f(X)$ и дифференциал $df(X)$ 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 $df(X)$ 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.

Example

Найдите первый и второй дифференциалы $df(X)$ и $d^2f(X)$, а также градиент $\nabla f(X)$ функции $f(X)$

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

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

Практика

Example

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

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