数学基础

机器学习导论（2023年春季）

赵 鹏

zhaop@lamda.nju.edu.cn

Outline

- Gradient and Derivatives
- Hessian
- Chain Rule

Notational Convention

- $[n] = \{1, \ldots, n\}$
- $\mathbf{x}, \mathbf{y}, \mathbf{v}$: vectors
- A, B: matrices
- $\mathcal{X}, \mathcal{Y}, \mathcal{K}$: domain
- d: dimension
- I: identity matrix
- X, Y: random variables
- $\boldsymbol{p}, \boldsymbol{q}$: probability distributions

Gradient and Derivatives (First Order)

- The gradient and derivative of a scalar function ($f : \mathbb{R} \mapsto \mathbb{R}$) is the same.
- The derivative of vector functions ($f : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R}$) is the transpose of its gradient.

We focus on the "gradient" language (i.e., column vector).

Definition 2 (Gradient). Let $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Let $\mathbf{x} = [x_1, \cdots, x_d]^\top \in \mathcal{X}$. Then, the gradient of f at $\mathbf{x}$ is a vector in $\mathbb{R}^d$ denoted by $\nabla f(\mathbf{x})$ and defined by

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{bmatrix}.$$

Example

Example 1. The gradient of $f(\mathbf{x}) = \|\mathbf{x}\|_2^2 \triangleq \sum_{i=1}^{d} x_i^2$ is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_d \end{bmatrix} = 2\mathbf{x}.$$

Example 2. The gradient of $f(\mathbf{x}) = -\sum_{i=1}^{d} x_i \ln x_i$ is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} -(\ln x_1 + 1) \\ \vdots \\ -(\ln x_d + 1) \end{bmatrix}.$$

Hessian (Second Order)

Definition 3 (Hessian). Let $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Let $\mathbf{x} = [x_1, \cdots, x_d]^\top \in \mathcal{X}$. Then, the Hessian of f at $\mathbf{x}$ is the matrix in $\mathbb{R}^{d\times d}$ denoted by $\nabla^2 f(\mathbf{x})$ and defined by

$$\nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i, x_j}(\mathbf{x})\right]_{1\le i,j\le d}.$$

Example 3. The Hessian of $f(\mathbf{x}) = -\sum_{i=1}^d x_i \ln x_i$ is $\nabla^2 f(\mathbf{x}) = \mathrm{diag}(-\frac{1}{x_1}, \ldots, -\frac{1}{x_d})$.

Example 4. The Hessian of $f(\mathbf{x}) = x_1^3 x_2^2 - 3x_1 x_2^3 + 1$ is $\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 6x_1x_2^2 & 6x_1^2x_2 - 6x_2 \\ 6x_1^2x_2 - 9x_2^2 & 2x_1^3 - 18x_1x_2 \end{bmatrix}$.

Chain Rule

- Consider scalar functions for simplicity.

Chain Rule. For $h(x) = f(g(x))$,

- the gradient of $h(x)$ is $h'(x) = f'(g(x))g'(x)$.
- the Hessian of $h(x)$ is $h''(x) = f''(g(x))(g'(x))^2 + f'(g(x))g''(x)$.

Reference: The Matrix Cookbook

The derivatives of **vectors, matrices, norms, determinants, etc** can be found therein.

2.4.1 First Order

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \tag{69}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \tag{70}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T \tag{71}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \tag{72}$$

$$\frac{\partial \mathbf{X}}{\partial X_{ij}} = \mathbf{J}^{ij} \tag{73}$$

$$\frac{\partial (\mathbf{X}\mathbf{A})_{ij}}{\partial X_{mn}} = \delta_{im}(\mathbf{A})_{nj} = (\mathbf{J}^{mn}\mathbf{A})_{ij} \tag{74}$$

$$\frac{\partial (\mathbf{X}^T\mathbf{A})_{ij}}{\partial X_{mn}} = \delta_{in}(\mathbf{A})_{mj} = (\mathbf{J}^{nm}\mathbf{A})_{ij} \tag{75}$$

https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

2 Derivatives

This section is covering differentiation of a number of expressions with respect to a matrix $\mathbf{X}$. Note that it is always assumed that $\mathbf{X}$ has *no special structure*, i.e. that the elements of $\mathbf{X}$ are independent (e.g. not symmetric, Toeplitz, positive definite). See section 2.8 for differentiation of structured matrices. The basic assumptions can be written in a formula as

$$\frac{\partial X_{kl}}{\partial X_{ij}} = \delta_{ik}\delta_{lj} \tag{32}$$

that is for e.g. vector forms,

$$\left[\frac{\partial \mathbf{x}}{\partial y}\right]_i = \frac{\partial x_i}{\partial y} \qquad \left[\frac{\partial x}{\partial \mathbf{y}}\right]_i = \frac{\partial x}{\partial y_i} \qquad \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right]_{ij} = \frac{\partial x_i}{\partial y_j}$$

The following rules are general and very useful when deriving the differential of an expression ([19]):

$$\begin{aligned}
\partial \mathbf{A} &= 0 \qquad (\mathbf{A} \text{ is a constant}) & (33)\\
\partial(\alpha\mathbf{X}) &= \alpha\partial\mathbf{X} & (34)\\
\partial(\mathbf{X}+\mathbf{Y}) &= \partial\mathbf{X}+\partial\mathbf{Y} & (35)\\
\partial(\mathrm{Tr}(\mathbf{X})) &= \mathrm{Tr}(\partial\mathbf{X}) & (36)\\
\partial(\mathbf{X}\mathbf{Y}) &= (\partial\mathbf{X})\mathbf{Y}+\mathbf{X}(\partial\mathbf{Y}) & (37)\\
\partial(\mathbf{X}\circ\mathbf{Y}) &= (\partial\mathbf{X})\circ\mathbf{Y}+\mathbf{X}\circ(\partial\mathbf{Y}) & (38)\\
\partial(\mathbf{X}\otimes\mathbf{Y}) &= (\partial\mathbf{X})\otimes\mathbf{Y}+\mathbf{X}\otimes(\partial\mathbf{Y}) & (39)\\
\partial(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1} & (40)\\
\partial(\det(\mathbf{X})) &= \mathrm{Tr}(\mathrm{adj}(\mathbf{X})\partial\mathbf{X}) & (41)\\
\partial(\det(\mathbf{X})) &= \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X}) & (42)\\
\partial(\ln(\det(\mathbf{X}))) &= \mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X}) & (43)\\
\partial\mathbf{X}^T &= (\partial\mathbf{X})^T & (44)\\
\partial\mathbf{X}^H &= (\partial\mathbf{X})^H & (45)
\end{aligned}$$