1.2 预备知识 — 向量范数与矩阵范数

符号:

$$x := (x_1, x_2, \ldots, x_n)^\top = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \langle x, y \rangle := \sum_{i=1}^{n} x_i y_i = x^\top y.$$

定义 1.2.1 称映射$\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ 为$\mathbb{R}^n$上的范数, 当且仅当它具有下列性质(范数三公理):

(i) 对任意的$x \in \mathbb{R}^n$, 有$\|x\| \geq 0$, 且$\|x\| = 0$当且仅当$x = 0$.

(ii) 对任意的$x \in \mathbb{R}^n$和任意的$\alpha \in \mathbb{R}$, 有$\|\alpha x\| = |\alpha|\|x\|$.

(iii) 对任意的$x, y \in \mathbb{R}^n$, 有$\|x + y\| \leq \|x\| + \|y\|$.

1.2 预备知识 — 向量范数与矩阵范数

对任意的$x \in \mathbb{R}^n$, 常用的向量范数有

(1) l_1-范数: $\|x\|_1 = \sum_{i=1}^n |x_i|$.

(2) l_2-范数: $\|x\|_2 = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$.

(3) l_∞-范数: $\|x\|_\infty = \max\{|x_i| \mid i \in \{1, 2, \ldots, n\}\}$.

一般地, 对于$p \in [1, \infty)$, l_p-范数定义为:

(4) $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$.

由上述各种范数的定义，容易验证: 对任意的$x \in \mathbb{R}^n$有

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \le n\|x\|_\infty.$$

1.2 预备知识 — 向量范数与矩阵范数

命题 1.2.1 假设$\|\cdot\|_\alpha$和$\|\cdot\|_\beta$是定义在$\mathbb{R}^n$上的任意两种范数. 那么总存在两个正数ξ_1和ξ_2, 使得对任意的$x \in \mathbb{R}^n$, 有$\xi_1\|x\|_\alpha \le \|x\|_\beta \le \xi_2\|x\|_\alpha$.

向量的椭球范数被定义为 $\|x\|_A := \sqrt{x^\top Ax}, \ \forall x \in \mathbb{R}^n$.

类似于向量范数, 可以定义矩阵范数.

定义 1.2.2 称映射$\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ 为$\mathbb{R}^{n\times n}$上的范数, 当且仅当它具有下列性质:

(i) 对任意的$A \in \mathbb{R}^{n\times n}$, 有$\|A\| \ge 0$, 且$\|A\| = 0$当且仅当$A = 0$.

(ii) 对任意的$A \in \mathbb{R}^{n\times n}$和任意的$\alpha \in \mathbb{R}$, 有$\|\alpha A\| = |\alpha|\|A\|$.

(iii) 对任意的$A, B \in \mathbb{R}^{n\times n}$, 有$\|A + B\| \le \|A\| + \|B\|$.

1.2 预备知识 — 向量范数与矩阵范数

对任意的$A=(a_{ij})_{n\times n}\in\mathbb{R}^{n\times n}$, 最常用的矩阵范数是Frobenius范数:

$$\|A\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2} = \sqrt{\mathrm{Tr}(A^\top A)},$$

其中$\mathrm{Tr}(A^\top A)$表示矩阵$A^\top A$的迹, 即$A^\top A$的所有主对角线元素之和, 也等于$A^\top A$的所有特征值之和. 另一个常用的矩阵范数是由向量所诱导的矩阵范数, 也称为算子范数, 其定义为

$$\|A\|_\alpha := \max_{x\in\mathbb{R}^n\setminus\{0\}}\frac{\|Ax\|_\alpha}{\|x\|_\alpha},\quad \forall A\in\mathbb{R}^{n\times n},$$

其中, $\|\cdot\|$是某种向量范数. 特别地, 对任意的$A\in\mathbb{R}^{n\times n}$,

(1) (列范数) $\|A\|_1 = \max\left\{\sum_{i=1}^n |a_{ij}| \mid j\in\{1,2,\ldots,n\}\right\}$;

(2) (行范数) $\|A\|_\infty = \max\left\{\sum_{j=1}^n |a_{ij}| \mid i\in\{1,2,\ldots,n\}\right\}$;

(3) (谱范数) $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$, 其中$\lambda_{\max}(A^\top A)$表示矩阵$A^\top A$的最大特征值.

相容性条件: $\|AB\|\le\|A\|\|B\|,\quad \|Ax\|\le\|A\|\|x\|.$

1.2 预备知识 — 向量范数与矩阵范数

下面介绍几个常用的不等式.

Cauchy-Schwartz不等式: 对任意的$x, y \in \mathbb{R}^n$, 有

$$|x^\top y| \le \|x\|_2 \|y\|_2,$$

其中, 等号成立当且仅当x和y线性相关.

广义Cauchy-Schwartz不等式: 设$A \in \mathbb{R}^{n\times n}$是正定矩阵. 对任意的$x, y \in \mathbb{R}^n$, 有

$$|x^\top y| \le \|x\|_A \|y\|_{A^{-1}},$$

其中, 等号成立当且仅当x和$A^{-1}y$线性相关.

Young不等式: 假设p和q是大于1的实数且满足$1/p + 1/q = 1$, 如果a和b是实数, 那么

$$|ab| \le \frac{|a|^p}{p} + \frac{|b|^q}{q},$$

其中, 等号成立当且仅当$|a|^p = |b|^q$.

1.2 预备知识 — 向量范数与矩阵范数

Hölder不等式: 对任意的$x, y \in \mathbb{R}^n$, 有

$$|x^\top y| \leq \|x\|_p \|y\|_q = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q},$$

其中, p和q是大于1的实数且满足$1/p + 1/q = 1$.

Minkowski不等式: 对任意的$x, y \in \mathbb{R}^n$, 有

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p},$$

其中, $p \in [1, \infty)$.

1.2 预备知识 — 函数的可微性

给定函数$f: \mathscr{F} \subseteq \mathbb{R}^n \to \mathbb{R}$,

- 如果f在每一点$x \in \mathscr{F}$处连续, 那么称f在$\mathscr{F}$上连续;
- 如果$\mathscr{F}$为开集且在每一点$x \in \mathscr{F}$处, 偏导数$\frac{\partial f(x)}{\partial x_i}$ $(i \in \{1, \ldots, n\})$都存在且连续, 那么称f在$\mathscr{F}$上连续可微;
- 如果在每一点$x \in \mathscr{F}$处, 二阶偏导数$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ $(i, j \in \{1, 2, \ldots, n\})$都存在且连续, 那么称$f$在$\mathscr{F}$上二次连续可微.

多变量实值函数的梯度. 假设函数$f: \mathscr{F} \subseteq \mathbb{R}^n \to \mathbb{R}$一阶连续可微, 那么函数$f$在点$x$处的一阶导数为

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^\top,$$

称其为函数f在点x处的梯度.

多变量实值函数的Hesse阵. 如果函数f是二次连续可微, 那么函数f在点x处的二阶导数为

$$\nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{n\times n} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix},$$

称其为函数f在点x处的Hesse阵.

例如, 对于给定对称矩阵$A \in \mathbb{R}^{n\times n}$, 向量$b \in \mathbb{R}^n$和实数$c$, 则二次函数$f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$ 的梯度为$\nabla f(x) = Ax + b$; Hesse阵为$\nabla^2 f(x) = A$.

1.2 预备知识 — 函数的可微性

多变量向量值函数的雅可比. 给定多变量向量值函数$F: \mathscr{F} \subseteq \mathbb{R}^n \to \mathbb{R}^m$, 如果$F$在点$x \in \mathscr{F}$处连续可微, 那么函数$F$在点$x$处的一阶导数为

$$F'(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1} & \frac{\partial F_m}{\partial x_2} & \cdots & \frac{\partial F_m}{\partial x_n} \end{pmatrix},$$

称其为函数F在点x处的Jacobi矩阵.

例如, 对于给定矩阵$A \in \mathbb{R}^{m\times n}$, 则多变量向量值函数$F(x) = Ax$的Jacobi矩阵为$F'(x) = A$.

1.2 预备知识 — 函数的可微性

例 求函数$f(x)=x_1^2+2x_2^2-2x_1x_2-4x_1+3$的梯度与Hesse阵。

解：（法1）$\dfrac{\partial f(x)}{\partial x_1}=2x_1-2x_2-4;\ \dfrac{\partial f}{\partial x_2}=4x_2-2x_1;$

所以$\nabla f(x)=(2x_1-2x_2-4,4x_2-2x_1)^T.$

$$\frac{\partial^2 f(x)}{\partial x_1^2}=2;\ \frac{\partial^2 f(x)}{\partial x_1\partial x_2}=-2;\ \frac{\partial^2 f(x)}{\partial x_2\partial x_1}=-2;\ \frac{\partial^2 f(x)}{\partial x_2\partial x_1}=4;$$

所以$\partial^2 f(x)=\begin{pmatrix}2 & -2\\ -2 & 4\end{pmatrix}.$

（法2）$f(x)=\dfrac{1}{2}x^T\begin{pmatrix}2 & -2\\ -2 & 4\end{pmatrix}x+\begin{pmatrix}-4\\ 0\end{pmatrix}^T x+3;$

$$\nabla f(x)=Ax+b=\begin{pmatrix}2 & -2\\ -2 & 4\end{pmatrix}x+\begin{pmatrix}-4\\ 0\end{pmatrix}=\begin{pmatrix}2x_1-2x_2-4\\ -2x_1+4x_2\end{pmatrix}.$$

$$\partial^2 f(x)=A=\begin{pmatrix}2 & -2\\ -2 & 4\end{pmatrix}.$$

1.2 预备知识 — 函数的可微性

例 求函数$f(x)=(x_2-x_1^2)^2+(1+x_1)^2$的梯度与Hesse阵。

解：$\dfrac{\partial f(x)}{\partial x_1}=2(x_2-x_1^2)(-2x_1)+2(1+x_1)$; $\dfrac{\partial f(x)}{\partial x_2}=2(x_2-x_1^2)$;

所以$\nabla f(x)=\begin{pmatrix} 2(x_2-x_1^2)(-2x_1)+2(1+x_1) \\ 2(x_2-x_1^2) \end{pmatrix}$;

$$\frac{\partial^2 f(x)}{\partial x_1^2}=12x_1^2-4x_2+2;\ \frac{\partial^2 f(x)}{\partial x_1\partial x_2}=-4x_1;\ \frac{\partial^2 f(x)}{\partial x_2^2}=2;$$

$$\frac{\partial^2 f}{\partial x^2}=\begin{pmatrix} 12x_1^2-4x_2+2 & -4x_1 \\ -4x_1 & 2 \end{pmatrix}.$$

例 求向量值函数$F(x)=(e^{x_1x_2}+3\sin x_2, 1+x_2^2\cos x_1)^T$的Jacobi阵。

解：$F^{'}(x)=\begin{pmatrix} \dfrac{\partial F_1(x)}{\partial x_1} & \dfrac{\partial F_1(x)}{\partial x_2} \\ \dfrac{\partial F_2(x)}{\partial x_1} & \dfrac{\partial F_2(x)}{\partial x_2} \end{pmatrix}=\begin{pmatrix} x_2e^{x_1x_2} & x_1e^{x_1x_2}+3\cos x_2 \\ -x_2^2\sin x_1 & 2x_2\cos x_1 \end{pmatrix}.$

1.2 预备知识 — 函数的可微性

定理 1.2.1 对于函数$f:\ \mathscr{F}\subseteq\mathbb{R}^n\to\mathbb{R}$, 令$x, x^*\in\mathscr{F}$且$x\neq x^*$. 如果函数$f$在$\mathscr{F}$上连续可微, 那么

(i) 存在$\alpha\in(0,1)$使得

$$f(x)=f(x^*)+\nabla f(\xi)^\top(x-x^*),$$

其中, $\xi=x^*+\alpha(x-x^*)$;

(ii) 函数f在x^*处有一阶Taylor展开式为:

$$f(x)=f(x^*)+\nabla f(x^*)^\top(x-x^*)+o(\|x-x^*\|),$$

其中, $o(\|x-x^*\|)$表示: 当$\|x-x^*\|\to0$时, $o(\|x-x^*\|)$是$\|x-x^*\|$的高阶无穷小量.

如果函数f在$\mathscr{F}$上二次连续可微, 那么

(iii) 存在$\beta\in(0,1)$使得

$$f(x)=f(x^*)+\nabla f(x^*)^\top(x-x^*)+\frac{1}{2}(x-x^*)^\top\nabla^2f(\xi)(x-x^*),$$

其中, $\xi=x^*+\beta(x-x^*)$;

(iv) 函数f在x^*处有二阶Taylor展开式为:

$$f(x)=f(x^*)+\nabla f(x^*)^\top(x-x^*)+\frac{1}{2}(x-x^*)^\top\nabla^2f(x^*)(x-x^*)$$
$$+o(\|x-x^*\|^2).$$

1.2 预备知识 — 函数的可微性

证明　构造一元辅助函数$\phi(t) = f(x^* + t(x - x^*))$, 那么$\phi(0) = f(x^*)$, $\phi(1) = f(x)$. 由函数f连续可微(或二次连续可微)知: 函数$\phi(t)$关于t也连续可微(或二次连续可微). 于是有

$$
\begin{aligned}
\phi'(t) &= \sum_{i=1}^{n} \frac{\partial f(x^* + t(x - x^*))}{\partial x_i}(x_i - x_i^*) = \nabla f(x^* + t(x - x^*))^\top (x - x^*), \\
\phi''(t) &= \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \frac{\partial^2 f(x^* + t(x - x^*))}{\partial x_i \partial x_j}(x_j - x_j^*) \right] (x_i - x_i^*) \\
&= (x - x^*)^\top \nabla^2 f(x^* + t(x - x^*))(x - x^*),
\end{aligned}
$$

进而, $\phi'(0) = \nabla f(x^*)^\top (x - x^*)$且$\phi''(0) = (x - x^*)^\top \nabla^2 f(x^*)(x - x^*)$. 因此, 由一元函数的中值定理,

$$
\phi(1) = \phi(0) + \phi'(\alpha), \quad \phi(1) = \phi(0) + \phi'(0) + \frac{1}{2}\phi''(\beta),
$$

可分别得到(i)和(iii)中的结论.

1.2 预备知识 — 函数的可微性

另外, 记

$$y := \frac{x - x^*}{\|x - x^*\|} \quad \text{和} \quad \gamma := \|x - x^*\|.$$

构造函数$\psi(\gamma) := f(x^* + \gamma y)$, 那么

$$\psi(\gamma) = f(x), \quad \psi(0) = f(x^*),$$
$$\psi'(0)\gamma = \nabla f(x^*)^\top (x - x^*),$$
$$\psi''(0)\gamma^2 = (x - x^*)^\top \nabla^2 f(x^*)(x - x^*).$$

因此, 由一元函数的Taylor展开式

$$\psi(\gamma) = \psi(0) + \psi'(0)\gamma + o(\gamma), \quad \psi(\gamma) = \psi(0) + \psi'(0)\gamma + \frac{1}{2}\psi''(0)\gamma + o(\gamma^2),$$

可分别得到(ii)和(iv)中的结论. □

1.2 预备知识 — 小结与作业

小结：

本节介绍了向量和矩阵范数以及函数的可微性，前者重点在于定义及相互关系；后者重点在于梯度和海塞矩阵的计算，以及中值定理及泰勒展开公式。

作业：

1.5 试证明: 对任意的$x \in \mathbb{R}^n$, 有

(1) $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$;

(2) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$;

(3) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.

1.10 计算下列函数的梯度和Hesse矩阵:

(1) $f(x_1, x_2, x_3) = x_1 e^{x_1+x_2+x_3}$;

(2) $f(x_1, x_2) = x_1 x_2 + ln(x_1^2 + x_1 x_2 + x_2^2)$;