最优化理论基础（上）

Hailiang ZHAO @ ZJU-CS

http://hliangzhao.me

2021 年 7 月 10 日

基础数学理论请参考http://hliangzhao.me/math/math.pdf。

向量范数

范数是高维空间中**计量事物自身长度**的一种方式。

定义

向量范数 是一个从向量空间 $\mathbb{R}^n$ 到实数域 $\mathbb{R}$ 的非负函数 $\|\cdot\|$，且满足

- 正定性：$\forall v \in \mathbb{R}^n,\ \|v\| \geq 0$ 且 $\|v\| = 0 \iff v = 0$
- 齐次性：$\forall v \in \mathbb{R}^n, \alpha \in \mathbb{R},\ \|\alpha v\| = |\alpha|\|v\|$
- 三角不等式：$\forall v, w \in \mathbb{R}^n,\ \|v + w\| \leq \|v\| + \|w\|$

l_p 范数 $(p \geq 1)$：$\|v\|_p \triangleq \left(|v_1|^p + ... + |v_n|^p\right)^{\frac{1}{p}}$，默认 $p = 2$

由正定阵 A 诱导的范数：$\|v\|_A \triangleq \sqrt{v^{\mathrm{T}} A v}$

定理

柯西不等式 $\forall a, b \in \mathbb{R}^n,\ |a^{\mathrm{T}} b| \leq \|a\|\|b\|$。

矩阵范数

矩阵的 l_p 范数（$p \geq 1$）：

$$\|A\|_p \triangleq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p \right)^{\frac{1}{p}}$$

$p = 2$ 时，称之为矩阵的 Frobenius 范数，记为 $\|\cdot\|_F$：

$$\|A\|_F \triangleq \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{Tr}\left(A^{\mathrm{T}} A\right)}$$

定理

正交不变性 对于任意的正交矩阵 $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$，

$$\|UAV\|_F^2 = \mathrm{Tr}\left(UAVV^{\mathrm{T}} A^{\mathrm{T}} U^{\mathrm{T}}\right) = \|A\|_F^2$$

这里运用了性质 $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$。

矩阵范数

算子范数（由向量范数诱导的矩阵范数）：
给定 m 维和 n 维空间的向量范数 $\|\cdot\|_{(m)}$ 和 $\|\cdot\|_{(n)}$，

$$\|A\|_{(m,n)} \triangleq \max_{\forall x\in\mathbb{R}^n, \|x\|_{(n)}=1} \|Ax\|_{(m)}$$

显然有结论：

$$\|Ax\|_{(m)} \leq \|A\|_{(m,n)}\|x\|_{(n)}$$

（矩阵范数和给定的向量范数**相容**）

若 $\|\cdot\|_{(m)}$ 和 $\|\cdot\|_{(n)}$ 均取相应空间的 l_p 范数，如 $p=2$，则可得到**矩阵的 2 范数**：

$$\|A\|_2 \triangleq \max_{\forall x\in\mathbb{R}^n, \|x\|_2=1} \|Ax\|_2 = \max_{r=1,\ldots,\mathrm{rank}(A)} \sigma_r(\text{最大奇异值})$$

严格区分矩阵的 2 范数和矩阵的 F 范数（l_2 范数）！

矩阵范数

核范数 (非零奇异值之和):

$$\|A\|_* \triangleq \sum_{r=1}^{\mathrm{rank}(A)} \sigma_r$$

内积用来表征两个矩阵（或其张成的空间）之间的夹角。常用的内积是 Frobenius 内积:

$$\langle A, B\rangle \triangleq \mathrm{Tr}(AB^{\mathrm{T}}) = \sum_{i,j} a_{ij} b_{ij}$$

显然，$\langle A, A\rangle = \|A\|_F^2$。

定理

矩阵范数的柯西不等式 $\forall A, B \in \mathbb{R}^{m\times n}$，$|\langle A, B\rangle| \le \|A\|_F \|B\|_F$。在 A 和 B 线性相关的时候取等。

导数

定义

梯度 若函数 $f:\mathbb{R}^n \to \mathbb{R}$ 在点 x 的一个邻域内有意义且存在 $g \in \mathbb{R}^n$ 使得

$$\lim_{p \to 0} \frac{f(x+p) - f(x) - g^{\mathrm{T}} p}{\|p\|} = 0,$$

则称 f 在点 x 处**可微**。我们称 g 为 f 在点 x 处的梯度，记为 $\nabla f(x)$。若区域 D 上每一个点 x 都有 $\nabla f(x)$ 存在，则称 f 在 D 上可微。

基于该定义，若依次对每一个维度 i 取 $p = \varepsilon e_i$（e_i 为第 i 维的标准基），则可得到

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right]^{\mathrm{T}}$$

二阶导数

定义

海瑟矩阵 若函数 $f:\mathbb{R}^n \to \mathbb{R}$ 在点 x 的二阶偏导数 $\left\{\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right\}_{i,j=1,\ldots,n}$ 都存在，则

$$\nabla^2 f(x) \triangleq \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{i,j=1,\ldots,n} \in \mathbb{R}^{n\times n}$$

称为 f 在点 x 处海瑟矩阵。若区域 D 上每一个点 x 都有 $\nabla^2 f(x)$ 存在，则称 f 在 D 上可微。若 $\nabla^2 f(x)$ 在 D 上还连续，则称 f 在 D 上二阶连续可微，此时 $\nabla^2 f(x)$ 是一个对称阵。

定义

雅可比矩阵 对于向量值函数 $f:\mathbb{R}^n \to \mathbb{R}^m$，可类比**梯度**定义它的雅可比矩阵：

$$J(x) = \left[\frac{\partial f_i(x)}{\partial x_j}\right]_{ij} \in \mathbb{R}^{n\times m}$$

泰勒展开

若 $f:\mathbb{R}^n \to \mathbb{R}$ 是连续可微的，则

$$f(x+p) = f(x) + \nabla f(x+tp)^{\mathrm{T}} p,$$

其中 $0 < t < 1$，$p \in \mathbb{R}^n$。

若 $f:\mathbb{R}^n \to \mathbb{R}$ 是二阶连续可微的，则

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)p dt \qquad (\text{对上式求导})$$

$$f(x+p) = f(x) + \nabla f(x)^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} \nabla^2 f(x+tp)p,$$

其中 $0 < t < 1$，$p \in \mathbb{R}^n$。

梯度 Lipschitz 连续

定义

梯度 Lipschitz 连续 对于可微函数 f，若存在 $L > 0$ 对于任意 $x, y \in \mathbf{dom} f$ 有

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

则称 f 是 L-Lipschitz 连续的。

L-Lipschitz 连续表明 $\nabla f(x)$ 的变化可以被自变量 x 的变化所控制，这在算法的收敛性证明中将发挥重要的作用。

接下来两页的内容常常出现在许多算法的收敛性分析中。

梯度 Lipschitz 连续

定理

二次上界 *L-Lipschitz* 连续的函数 $f:\mathbb{R}^n \to \mathbb{R}$ 有二次上界：

$$f(y) \leq f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \frac{L}{2}\|y-x\|^2, \forall x, y \in \mathbf{dom} f.$$

证明.

通过构造辅助函数 $g(t) = f\big(x + t(y-x)\big), t \in [0,1]$ 来推导

$$f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y-x)$$

的上界。 □

这说明 L-Lipschitz 连续函数被一个二次函数的上界所控制，即 $f(x)$ 的增长速度不超过二次。（$\mathbf{dom} f$ 仅需是凸集即可，不需要是 $\mathbb{R}^n$。）

梯度 Lipschitz 连续

定理

差距的下界 若 L-*Lipschitz* 连续的函数 $f:\mathbb{R}^n \to \mathbb{R}$ 的定义域为 $\mathbb{R}^n$ 且存在一个全局极小点 x^*，则

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^*), \forall x \in \mathbb{R}^n.$$

证明.

根据上一定理的结论可得

$$\begin{aligned} f(x^*) &\leq \inf_{y\in\mathbb{R}^n}\left\{f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \frac{L}{2}\|y-x\|^2\right\} \\ &= f(x) - \frac{1}{2L}\|\nabla f(x)\|^2. \qquad \triangleright -\frac{b}{2a} \end{aligned}$$

□

矩阵变量函数的导数

定义

Frechet **可微** 对于函数 $f : \mathbb{R}^{m\times n} \to \mathbb{R}$，若存在矩阵 $G \in \mathbb{R}^{m\times n}$ 满足

$$\lim_{V\to 0} \frac{f(X+V) - f(X) - \langle G, V\rangle}{\|V\|} = 0,$$

则称 f 在 X 处 Frechet 可微，G 是相应意义下的梯度，记为 $\nabla f(X)$。

同样有：

$$\nabla f(X) = \left[\frac{\partial f}{\partial x_{ij}}\right]_{ij}$$

矩阵变量函数的导数

定义

Gateaux **可微** 对于函数 $f:\mathbb{R}^{m\times n}\to\mathbb{R}$，若存在矩阵 $G\in\mathbb{R}^{m\times n}$ 对任意方向 $V\in\mathbb{R}^{m\times n}$ 和任意 $t\in\mathbb{R}$ 满足

$$\lim_{t\to 0}\frac{f(X+tV)-f(X)-t\langle G,V\rangle}{t}=0,$$

则称 f 在 X 处 Gateaux 可微，G 是相应意义下的梯度，仍然记为 $\nabla f(X)$。

Gateaux 可微其实是方向导数的某种推广。此外，若 f 是 Frechet 可微的，则 f 是 Gateaux 可微的，且二者意义下的梯度相等。

我们主要关注 Gateaux 可微。

利用 Gateaux 可微的定义计算梯度

（利用 $\mathrm{Tr}(A+B)=\mathrm{Tr}(A)+\mathrm{Tr}(B)$、$\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$ 以及 $\|\cdot\|_F$ 的定义来计算）

示例 1：考虑线性函数 $f(X)=\mathrm{Tr}(AX^{\mathrm{T}}B)$，其中 $A\in\mathbb{R}^{p\times n}$，$B\in\mathbb{R}^{m\times p}$，$X\in\mathbb{R}^{m\times n}$，则

$$\begin{aligned}\lim_{t\to 0}\frac{f(X+tV)}{t}&=\frac{\mathrm{Tr}\big(A(X+tV)^{\mathrm{T}}B\big)-\mathrm{Tr}(AX^{\mathrm{T}}B)}{t}\\&=\mathrm{Tr}(AV^{\mathrm{T}}B)=\langle BA,V\rangle.\end{aligned}$$

所以 $\nabla f(X)=BA$。

示例 2：考虑二次函数 $f(X,Y)=\frac{1}{2}\|XY-A\|_F^2$，其中 $(X,Y)\in\mathbb{R}^{m\times p}\times\mathbb{R}^{p\times n}$，则

$$f(X,Y+tV,Y)-f(X,Y)=t\langle V,(XY-A)Y^{\mathrm{T}}\rangle+\mathcal{O}(t^2),$$

所以 $\nabla_Y f(X,Y)=X^{\mathrm{T}}(XY-A)$。同理 $\nabla_X f(X,Y)=(XY-A)Y^{\mathrm{T}}$。

利用 Gateaux 可微的定义计算梯度

示例 3：考虑 *In-det* 函数 $f(X) = \ln(\det(X))$，其中 $X \in \mathcal{S}_{++}^n$，则给定 $X \succ 0$，对任意方向 $V \in \mathcal{S}^n$ 以及 $t \in \mathbb{R}$，有

$$
\begin{aligned}
& f(X+tV) - f(X) \\
&= \ln\Big(\det\big(X^{1/2}(I + tX^{-1/2}VX^{-1/2})X^{1/2}\big)\Big) - \ln\big(\det(X)\big) \\
&= \ln\Big(\det\big(I + tX^{-1/2}VX^{-1/2}\big)\Big). \qquad \triangleright \det(AB) = \det(A)\det(B)
\end{aligned}
$$

对于正定阵 $X^{-1/2}VX^{-1/2}$，因为可以正交对角化[1]，所以不妨设其特征值为 $\lambda_1, ..., \lambda_n$，则

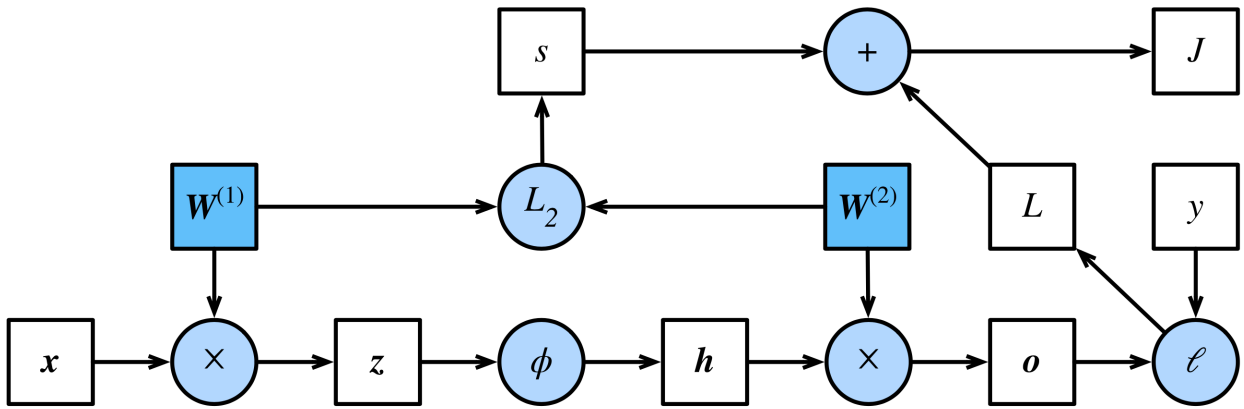
$$
\begin{aligned}
\ln\Big(\det(I + tX^{-1/2}VX^{-1/2})\Big) &= \ln\prod_{i=1}^{n}(1+t\lambda_i) \\
&= \sum_{i=1}^{n} t\lambda_i + \mathcal{O}(t^2) \qquad \triangleright x \to 0 : \ln(1+x) = x + \mathcal{O}(x^2) \\
&= t\mathrm{Tr}(X^{-1/2}VX^{-1/2}) + \mathcal{O}(t^2) \qquad \triangleright \text{迹是特征值之和} \\
&= t\langle (X^{-1})^{\mathrm{T}}, V\rangle + \mathcal{O}(t^2) \qquad \triangleright \nabla f(X) = (X^{-1})^{\mathrm{T}}
\end{aligned}
$$

[1] 更多细节见http://hliangzhao.me/math/math.pdf。

自动微分

依据微积分中的链式法则，沿着从计算图中输出层到输入层的顺序，依次计算并存储目标函数关于计算图中各层的中间变量以及参数的梯度。

第 l 层的误差可由第 $l+1$ 层的误差得到。



正向传播的计算图示例。

细节请参考文档http://hliangzhao.me/math/cheatsheet.pdf的章节 5.2。

广义实值函数

定义

广义实值函数 令 $\bar{\mathbb{R}} \triangleq \mathbb{R} \cup \{\pm\infty\}$ 为广义实数空间，则 $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ 被称为广义实值函数。

定义

适当函数 给定广义实值函数 f 和非空集合 $\mathcal{X}$，若

$$\begin{cases} \exists x \in \mathcal{X}, f(x) < +\infty \\ \forall x \in \mathcal{X}, f(x) > -\infty, \end{cases}$$

则称 f 是关于 $\mathcal{X}$ 的适当函数（**否则求 minimum 没有意义**）。

若无特别指示，接下来及后续文档中提及的 f **均为适当函数**，且默认 $\mathbb{R} = \bar{\mathbb{R}}$。

闭函数

定义

α-**下水平集** 对于函数 $f : \mathbb{R}^n \to \mathbb{R}$，称

$$C_\alpha \triangleq \{x \mid f(x) \leq \alpha\}$$

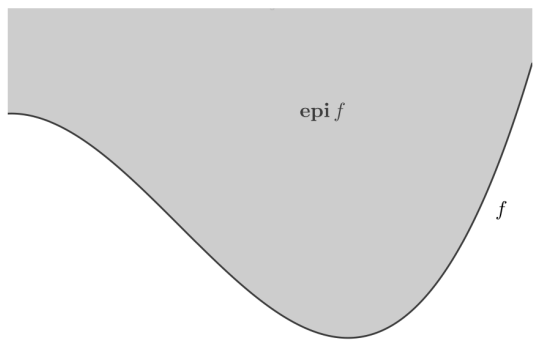
为 f 的 α-下水平集。α-下水平集是不超过某个阈值的自变量的集合。

定义

上方图（用于反推函数 f 的性质）对于函数 $f : \mathbb{R}^n \to \mathbb{R}$，称

$$\mathbf{epi} f \triangleq \{(x, t) \in \mathbb{R}^{n+1} \mid f(x) \leq t\}$$

为 f 的上方图。



闭函数

定义

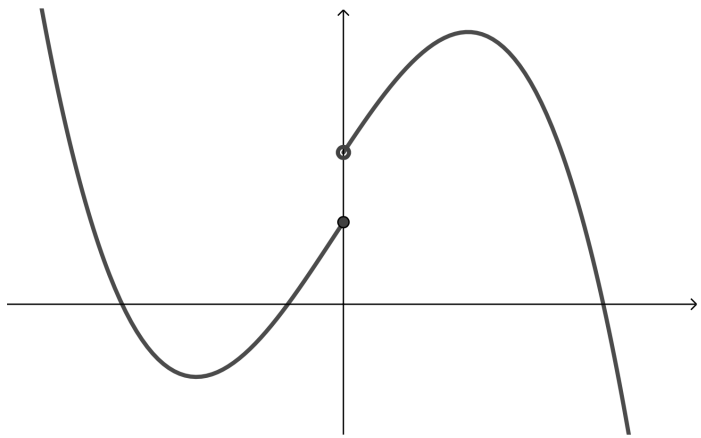
闭函数 对于函数 $f : \mathbb{R}^n \to \mathbb{R}$，若 $\mathbf{epi}f$ 为闭集（集合所有的极限点都是这个集合中的点），则称 f 为闭函数。

定义

下半连续函数 对于函数 $f : \mathbb{R}^n \to \mathbb{R}$，若 $\forall x \in \mathbb{R}^n$，

$$\liminf_{y \to x} f(y) \geq f(x),$$

则称 f 为下半连续函数。



闭函数

定理

闭函数和下半连续函数的等价性 对于函数 $f:\mathbb{R}^n \to \mathbb{R}$，以下命题等价：

1. f 的任意 α-下水平集都是闭集；
2. f 是下半连续的；
3. f 是闭函数。

证明.

$2 \Rightarrow 3$：设 $(x_k, y_k) \in \mathbf{epi} f$ 且 $\lim_{k\to\infty}(x_k, y_k) = (\bar{x}, \bar{y})$，则

$$f(\bar{x}) \le \liminf_{k\to\infty} f(x_k) \le \lim_{k\to\infty} y_k = \bar{y},$$

这说明 $(\bar{x}, \bar{y}) \in \mathbf{epi} f$，所以 $\mathbf{epi} f$ 是闭集。 □

闭函数

定理

闭函数和下半连续函数的等价性 对于函数 $f:\mathbb{R}^n \to \mathbb{R}$，以下命题等价：

1. f 的任意 α-下水平集都是闭集；
2. f 是下半连续的；
3. f 是闭函数。

证明.

$3 \Rightarrow 1$：取 α-下水平集的元素 $x_k \to \bar{x}$，有 $(x_k, \alpha) \in \mathbf{epi} f$ 且 $(x_k, \alpha) \to (\bar{x}, \alpha)$，因为 f 是闭函数，所以必然有 $(\bar{x}, \alpha) \in \mathbf{epi} f$，即 $f(\bar{x}) \leq \alpha$。这说明 f 的任意 α-下水平集都是闭集。 □

闭函数

定理

闭函数和下半连续函数的等价性 对于函数 $f:\mathbb{R}^n \to \mathbb{R}$，以下命题等价：

1. f 的任意 α-下水平集都是闭集；
2. f 是下半连续的；
3. f 是闭函数。

证明.

$1 \Rightarrow 2$：反证法。假设存在序列 $\{x_k\} \to \bar{x}(k \to \infty)$ 但 $f(\bar{x}) > \liminf_{k\to\infty} f(x_k)$，取 t 使得

$$f(\bar{x}) > t > \liminf_{k\to\infty} f(x_k).$$

根据下极限的定义，$\{x_k \mid f(x_k) \le t\}$ 中必然含有无穷多个 x_k，因此 $\{x_k\}$ 中存在子列 $\{x_{k_l}\}$ 使得 $f(x_{x_l}) \le t$ 且 $\lim_{x_{k_l}} = \bar{x}$，则 t-下水平集不是闭集，这与命题 1 矛盾。 □

闭函数

以上关于闭函数的相关命题将为后面的定理证明提供极大的方便。
闭函数间的简单运算会保持原有性质：

1. **加法**：若 f 和 g 均为闭函数，且 $\mathbf{dom} f \cap \mathbf{dom} g \neq \varnothing$，则 $f+g$ 也是闭函数。
2. **仿射映射的复合**：若 f 为闭函数，则 $f(Ax+b)$ 也为闭函数。
3. **取上确界**：若每一个 f_α 均为闭函数，则 $\sup_\alpha f_\alpha(x)$ 也为闭函数。