Mathematical Background

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Outline

- Norms
- Analysis
- Functions
- Derivatives
- Linear Algebra

Inner product

- Inner product on $\mathbf{R}^n$

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

- Euclidean norm, or l_2-norm

$$\|x\|_2 = (x^\top x)^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}, x \in \mathbf{R}^n$$

- Cauchy-Schwartz inequality

$$|x^\top y| \le \|x\|_2 \|y\|_2, x, y \in \mathbf{R}^n$$

- Angle between nonzero vectors $x, y \in \mathbf{R}^n$

$$\angle(x, y) = \cos^{-1}\left(\frac{x^\top y}{\|x\|_2 \|y\|_2}\right), x, y \in \mathbf{R}^n$$

Inner product

- Inner product on $\mathbf{R}^{m\times n}$, $X, Y \in \mathbf{R}^{m\times n}$

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

Here $\mathrm{tr}(\,)$ denotes trace of a matrix.

- Frobenius norm of a matrix $X \in \mathbf{R}^{m\times n}$

$$\|X\|_F = \left(\mathrm{tr}(X^\top X)\right)^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)^{1/2}$$

- Inner product on $\mathbf{S}^n$

$$\langle X, Y\rangle = \mathrm{tr}(XY) = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}\,Y_{ij} = \sum_{i=1}^{n} X_{ii}\,Y_{ii} + 2\sum_{i<j} X_{ij}\,Y_{ij}$$

Norms

- A function $f: \mathbf{R}^n \to \mathbf{R}$ with $\mathrm{dom}\, f = \mathbf{R}^n$ is called a norm if
  - f is nonnegative: $f(x) \geq 0$ for all $x \in \mathbf{R}^n$
  - f is definite: $f(x) = 0$ only if $x = 0$
  - f is homogeneous: $f(tx) = |t| f(x)$, for all $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$
  - f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$, for all $x, y \in \mathbf{R}^n$
- Distance
  - Between vectors x and y as the length of their difference, i.e.,

$$\mathrm{dist}(x, y) = \|x - y\|$$

Norms

Unit ball

- The set of all vectors with norm less than or equal to one,

$$\mathcal{B} = \{x \in \mathbf{R}^n \mid \|x\| \leq 1\}$$

  is called the unit ball of the norm $\|\cdot\|$.

- The unit ball satisfies the following properties:
  - ✓ $\mathcal{B}$ is symmetric about the origin, i.e., $x \in \mathcal{B}$ if and only if $-x \in \mathcal{B}$
  - ✓ $\mathcal{B}$ is convex
  - ✓ $\mathcal{B}$ is closed, bounded, and has nonempty interior

- Conversely, if $C \subseteq \mathbf{R}^n$ is any set satisfying these three conditions, the it is the unit ball of a norm:

$$\|x\| = (\sup\{t \geq 0 | tx \in C\})^{-1}$$

Norms

- Some common norms on $\mathbf{R}^n$
  - Sum-absolute-value, or l_1-norm

$$\|x\|_1 = |x_1| + \cdots + |x_n|, x \in \mathbf{R}^n$$

  - Chebyshev or l_∞-norm

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$$

  - l_p-norm

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$$

  - For $P \in \mathbf{S}_{++}^n$, P-quadratic norm is

$$\|x\|_P = (x^\top P x)^{1/2} = \|P^{1/2}x\|_2$$

Norms

- Some common norms on $\mathbf{R}^{m\times n}$
  - Sum-absolute-value norm

$$\|X\|_{\mathrm{sav}} = \sum_{i=1}^{m}\sum_{j=1}^{n}|X_{ij}|$$

  - Maximum-absolute-value norm

$$\|X\|_{\mathrm{mav}} = \max\{|X_{ij}| \mid i = 1,\ldots,m, j = 1,\ldots,n\}$$

Norms

Equivalence of norms

- Suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbf{R}^n$, there exist positive constants α and β, for all $x \in \mathbf{R}^n$

$$\alpha\|x\|_a \leq \|x\|_b \leq \beta\|x\|_a$$

- If $\|\cdot\|$ is any norm on $\mathbf{R}^n$, then there exists a quadratic norm $\|\cdot\|_P$ for which

$$\|x\|_P \leq \|x\| \leq \sqrt{n}\,\|x\|_P$$

holds for all x.

Norms

- Operator norms
  - Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. Operator norm of $X \in \mathbf{R}^{m\times n}$ induced by $\|\cdot\|_a$ and $\|\cdot\|_b$ is

$$\|X\|_{a,b} = \sup\{\|Xu\|_a \mid \|u\|_b \le 1\}$$

  - When $\|\cdot\|_a$ and $\|\cdot\|_b$ are Euclidean norms, the operator norm of X is its maximum singular value, and is denoted $\|X\|_2$

$$\|X\|_2 = \sigma_{\max}(X) = \left(\lambda_{\max}(X^\top X)\right)^{1/2}$$

    - ✓ Spectral norm or ℓ_2-norm

Norms

Operator norms

- The norm induced by the ℓ_∞-norm on $\mathbf{R}^m$ and $\mathbf{R}^n$, denoted $\|X\|_\infty$, is the max-row-sum norm,

$$\|X\|_\infty = \sup\{\|Xu\|_\infty \mid \|u\|_\infty \leq 1\} = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |X_{ij}|$$

- The norm induced by the ℓ_1-norm on $\mathbf{R}^m$ and $\mathbf{R}^n$, denoted $\|X\|_1$, is the max-column-sum norm,

$$\|X\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |X_{ij}|$$

Norms

Dual norm

- Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$.
- The associated dual norm, denoted $\|\cdot\|_*$, is defined as
$$\|z\|_* = \sup\{z^\top x \mid \|x\| \le 1\}$$
- We have the inequality
$$z^\top x \le \|x\| \|z\|_*$$
- The dual of Euclidean norm
$$\sup\{z^\top x \mid \|x\|_2 \le 1\} = \|z\|_2$$
- The dual of the ℓ_∞-norm
$$\sup\{z^\top x \mid \|x\|_\infty \le 1\} = \|z\|_1$$

Norms

Dual Norm

- The dual of ℓ_p-norm is the ℓ_q-norm such that

$$\frac{1}{p}+\frac{1}{q}=1$$

- The dual of the ℓ_2-norm on $\mathbf{R}^{m\times n}$ is the nuclear norm

$$\|Z\|_{2*}=\sup\{\mathrm{tr}(Z^\top X)\mid \|X\|_2\leq 1\}$$

$$=\sigma_1(Z)+\cdots+\sigma_r(Z)=\mathrm{tr}(Z^\top Z)^{1/2}$$

Outline

- Norms
- Analysis
- Functions
- Derivatives
- Linear Algebra

Analysis

Interior and Open Set

- An element $x \in C \subseteq \mathbf{R}^n$ is called an interior point of C if there exists an $\epsilon > 0$ for which
$$\{y \mid \|y - x\|_2 \leq \epsilon\} \subseteq C$$
i.e., there exists a ball centered at x that lies entirely in C.

- The set of all points interior to C is called the interior of C and is denoted $\operatorname{int} C$.

- A set C is open if $\operatorname{int} C = C$

Analysis

Closed Set and Boundary

- A set $C \subseteq \mathbf{R}^n$ is closed if its complement is open

$$\mathbf{R}^n \setminus C = \{x \in \mathbf{R}^n | x \notin C\}$$

- The closure of a set C is defined as

$$\mathrm{cl}\, C = \mathbf{R}^n \setminus \mathrm{int}(\mathbf{R}^{\mathbf{n}} \setminus C)$$

- The boundary of the set C is defined as

$$\mathrm{bd}\, C = \mathrm{cl}\, C \setminus \mathrm{int}\, C$$

  ✓ C is closed if it contains its boundary. It is open if it contains no boundary points.

Analysis

Supremum and infimum

- The least upper bound or supremum of the set C is denoted $\sup C$.

- The greatest lower bound or infimum of the set C is denoted $\inf C$.

Outline

- Norms
- Analysis
- Functions
- Derivatives
- Linear Algebra

Functions

- Notation

$$f : A \to B$$

  - $\mathrm{dom}\, f \subseteq A$

- An example $f : \mathbf{S}^n \to \mathbf{R}$

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

  - $\mathrm{dom}\, f \subseteq \mathbf{S}^n_{++}$

Functions

Continuity

- A function $f: \mathbf{R}^n \to \mathbf{R}^m$ is continuous at $x \in \operatorname{dom} f$ if for all $\epsilon > 0$ there exists a δ with $y \in \operatorname{dom} f$, such that
  $$\|y - x\|_2 \le \delta \Rightarrow \|f(y) - f(x)\|_2 \le \epsilon$$

Closed functions

- A function $f: \mathbf{R}^n \to \mathbf{R}$ is closed if, for each $\alpha \in \mathbf{R}$, the sublevel set
  $$\{x \in \operatorname{dom} f \mid f(x) \le \alpha\}$$
  is closed. This is equivalent to
  $$\operatorname{epi} f = \{(x, t) \in \mathbf{R}^{n+1} | x \in \operatorname{dom} f, f(x) \le t\}$$

Outline

- Norms
- Analysis
- Functions
- Derivatives
- Linear Algebra

Derivatives

Definition

- Suppose $f: \mathbf{R}^n \to \mathbf{R}^m$ and $x \in \mathrm{int\,dom}\, f$. The function f is differentiable at x if there exists a matrix $Df(x) \in \mathbf{R}^{m \times n}$ that satisfies

$$\lim_{z \in \mathrm{dom}\, f,\, z \neq x,\, z \to x} \frac{\|f(z) - f(x) - Df(x)(z - x)\|_2}{\|z - x\|_2} = 0$$

  in which case we refer to $Df(x)$ as the derivative (or Jacobian) of f at x.

Derivatives

- Definition
  - The affine function of z given by

$$f(x) + Df(x)(z - x)$$

is called the first-order approximation of f at (or near) x.

$$Df(x)_{ij} = \frac{\partial f_i(x)}{\partial x_j}, i = 1, \cdots, m, j = 1, \cdots, n$$

Derivatives

Gradient

- When f is real-valued (i.e., $f: \mathbf{R}^n \to \mathbf{R}$) the derivative $Df(x)$ is a $1 \times n$ matrix (it is a row vector). Its transpose is called the gradient of the function:

$$\nabla f(x) = Df(x)^\top$$

  which is a column vector (in $\mathbf{R}^n$). Its components are the partial derivatives of f:

$$\nabla f(x)_i = \frac{\partial f(x)}{\partial x_i}, i = 1, \cdots, n$$

- The first-order approximation of f at a point $x \in \operatorname{int} \operatorname{dom} f$ can be expressed as (the affine function of z)

$$f(x) + \nabla f(x)^\top (z - x)$$

Derivatives

- Examples

$$f(x) = \frac{1}{2}x^\top Px + q^\top x + r$$

$$\nabla f(x) = Px + q$$

$$f(X) = \log\det X\,, \mathrm{dom}\, f = \mathbf{S}_{++}^n$$

$$\nabla f(X) = X^{-1}$$

Derivatives

Chain rule

- Suppose $f:\mathbf{R}^n \to \mathbf{R}^m$ is differentiable at $x \in \mathrm{int}$ $\mathrm{dom}\, f$ and $g:\mathbf{R}^m \to \mathbf{R}^p$ is differentiable at $f(x) \in \mathrm{int}$ $\mathrm{dom}\, g$.

  Define the composition $h:\mathbf{R}^n \to \mathbf{R}^p$ by $h(z) = g(f(z))$. Then h is differentiable at x, with derivate

$$Dh(x) = Dg(f(x))Df(x)$$

- Suppose $f:\mathbf{R}^n \to \mathbf{R}$, $g:\mathbf{R} \to \mathbf{R}$, and $h(x) = g(f(x))$

$$\nabla h(x) = g'\big(f(x)\big)\nabla f(x)$$

Derivatives

□ Composition of Affine Function

$$g(x) = f(Ax + b)$$

$$\nabla g(x) = A^\top \nabla f(Ax + b)$$

$$f: \mathbf{R}^n \to \mathbf{R}, \qquad g: \mathbf{R} \to \mathbf{R}$$

$$g(t) = f(x + tv), \qquad x, v \in \mathbf{R}^n$$

$$g'(t) = v^\top \nabla f(x + tv)$$

Example 1

- Consider the function $f: \mathbf{R}^n \to \mathbf{R}$

$$f(x) = \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i)$$

  - where $a_1, \dots, a_m \in \mathbf{R}^n$

- $f = g(Ax + b)$

$$g(y) = \log \sum_{i=1}^{m} \exp(y_i)$$

$$\nabla g(y) = \frac{1}{\sum_{i=1}^{m} \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix}$$

Example 1

- Consider the function $f: \mathbf{R}^n \rightarrow \mathbf{R}$

$$f(x) = \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i)$$

  - where $a_1, \ldots, a_m \in \mathbf{R}^n$

- $f = g(Ax + b)$

$$\nabla f(x) = A^\top \nabla g(Ax + b) = \frac{1}{1^\top z} A^\top z$$

$$z = \begin{bmatrix} \exp a_1^\top x + b_1 \\ \vdots \\ \exp a_m^\top x + b_m \end{bmatrix}$$

Example 2

- Consider the function

$$f(x) = \log\det(F_0 + x_1 F_1 + \cdots + x_n F_n)$$

  - where $F_0, \ldots, F_n \in S^p$

- $f(x) = g(F_0 + x_1 F_1 + \cdots + x_n F_n)$

$$g(X) = \log\det X$$

$$\frac{\partial f(x)}{\partial x_i} = \mathrm{tr}(F_i \nabla \log\det(F)) = \mathrm{tr}(F^{-1} F_i)$$

$$\nabla f(x) = \begin{bmatrix} \mathrm{tr}(F^{-1} F_1) \\ \vdots \\ \mathrm{tr}(F^{-1} F_n) \end{bmatrix}$$

Second Derivative

Definition

- Suppose $f: \mathbf{R}^n \to \mathbf{R}$. The second derivative or Hessian matrix of f at $x \in \operatorname{int} \operatorname{dom} f$, denoted $\nabla^2 f(x)$, is given by

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, i = 1, \cdots, n, j = 1, \cdots, n.$$

Second-order Approximation

$$f(x) + \nabla f(x)^\top (z - x) + \frac{1}{2}(z - x)^\top \nabla^2 f(x)(z - x)$$

Derivatives

Examples

$$f(x) = \frac{1}{2}x^\top Px + q^\top x + r$$

$$\nabla f(x) = Px + q$$

$$\nabla^2 f(x) = P$$

$$f(X) = \log\det X, \mathrm{dom}\, f = \mathbf{S}_{++}^n$$

$$\nabla f(X) = X^{-1}$$

$$f(X) + \mathrm{tr}\left(X^{-1}(Z - X)\right) - \frac{1}{2}\mathrm{tr}\left(X^{-1}(Z - X)X^{-1}(Z - X)\right)$$

Second Derivative

- Chain rule
  - Suppose $f: \mathbf{R}^n \to \mathbf{R}$, $g: \mathbf{R} \to \mathbf{R}$, and $h(x) = g(f(x))$.

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^\top$$

  - Composition with affine function:

$$g(x) = f(Ax + b)$$

$$\nabla^2 g(x) = A^\top \nabla^2 f(Ax + b)A$$

Example 1

- Consider the function $f: \mathbf{R}^n \to \mathbf{R}$

$$f(x) = \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i)$$

  - where $a_1, \ldots, a_m \in \mathbf{R}^n$

- $f = g(Ax + b)$

$$g(y) = \log \sum_{i=1}^{m} \exp(y_i)$$

$$\nabla g(y) = \frac{1}{\sum_{i=1}^{m} \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix}$$

$$\nabla^2 g(y) = \operatorname{diag}(\nabla g(y)) - \nabla g(y) \nabla g(y)^\top$$

Example 1

- Consider the function $f: \mathbf{R}^n \to \mathbf{R}$

$$f(x) = \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i)$$

  - where $a_1, \dots, a_m \in \mathbf{R}^n$

- $f = g(Ax + b)$

$$\begin{aligned} \nabla^2 f(x) &= A^\top \nabla g^2 (Ax + b) A \\ &= A^\top \left(\frac{1}{\mathbf{1}^\top z} \operatorname{diag}(z) - \frac{1}{(\mathbf{1}^\top z)^2} z z^\top \right) A \end{aligned}$$

  - $z_i = \exp\left(a_i^\top x + b_i\right), i = 1, \dots, m$

Outline

- Norms
- Analysis
- Functions
- Derivatives
- Linear Algebra

Linear algebra

Range and nullspace

- Let $A \in \mathbf{R}^{m\times n}$, the range of A, denoted $\mathcal{R}(A)$, is the set of all vectors in $\mathbf{R}^m$ that can be written as linear combinations of the columns of A:

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbf{R}^n\} \subseteq \mathbf{R}^m$$

- The nullspace (or kernel) of A, denoted $\mathcal{N}(A)$, is the set of all vectors x mapped into zero by A:

$$\mathcal{N}(A) = \{x \mid Ax = 0\} \subseteq \mathbf{R}^n$$

- if $\mathcal{V}$ is a subspace of $\mathbf{R}^n$, its orthogonal complement, denoted $\mathcal{V}^\perp$, is defined as:

$$\mathcal{V}^\perp = \{x \mid z^\top x = 0 \text{ for all } z \in \mathcal{V}\}$$

Linear algebra

Range and nullspace

- Let $A \in \mathbf{R}^{m\times n}$, the range of A, denoted $\mathcal{R}(A)$, is the set of all vectors in $\mathbf{R}^m$ that can be written as linear combinations of the columns of A:

  $$\mathcal{R}(A) = \{A \quad \mathbf{R}^m$$

- The nullspa denoted $\mathcal{N}(A)$, is th mapped into zero by A:

  $$\mathcal{N}(A) = \mathcal{R}(A^\top)^\perp$$

  $$\mathcal{N}(A) = \{x | Ax = 0\} \subseteq \mathbf{R}^n$$

- if $\mathcal{V}$ is a subspace of $\mathbf{R}^n$, its orthogonal complement, denoted $\mathcal{V}^\perp$, is defined as:

  $$\mathcal{V}^\perp = \{x | z^T x = 0 \text{ for all } z \in \mathcal{V}\}$$

Linear algebra

Symmetric eigenvalue decomposition

- Suppose $A \in \boldsymbol{S}^n$, i.e., A is a real symmetric $n \times n$ matrix. Then A can be factored as

$$A = Q \Lambda Q^\top$$

where $Q \in \mathbf{R}^{n\times n}$ is orthogonal, i.e., satisfies $Q^\top Q = I$, and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$.

- The determinant and trace can be expressed in terms of the eigenvalue.

$$\det A = \prod_{i=1}^n \lambda_i, \mathrm{tr}\, A = \sum_{i=1}^n \lambda_i$$

Linear algebra

□ Norms

$$\|A\|_2 = \max_{i=1,\ldots,n} |\lambda_i| = \max(\lambda_1, -\lambda_n)$$

$$\|A\|_F = \left(\sum_{i=1}^{n} \lambda_i^2 \right)^{1/2}$$

Linear algebra

□ Positive definite Matrix

- A matrix $A \in \boldsymbol{S}^n$ is called positive definite, if for all $x \neq 0, x^\top A x > 0$, denoted as $A \succ 0$.
- If $-A$ is positive definite, we say A is negative definite, denoted as $A \prec 0$.
- We use $\boldsymbol{S}_{++}^n$ to denote the set of positive definite matrices in $\boldsymbol{S}^n$.
- We use $\boldsymbol{S}_{+}^n$ to denote the set of positive semidefinite matrices in $\boldsymbol{S}^n$.

Linear algebra

Singular value decomposition (SVD)

- Suppose $A \in \mathbf{R}^{m\times n}$ with $\operatorname{rank} A = r$. Then A can be factored as

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

where $U \in \mathbf{R}^{m\times r}$ satisfies $U^\top U = I$, $V \in \mathbf{R}^{n\times r}$ satisfies $V^\top V = I$, and $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_r)$ with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$$

- The singular value decomposition can be written

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

Linear algebra

Norms

$$\|A\|_2 = \sigma_1$$

$$\|A\|_F = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2}$$

Linear algebra

Pseudo-inverse

- Let $A = U\Sigma V^\top$ be the singular value decomposition of $A \in \mathbf{R}^{m\times n}$, with $\operatorname{rank} A = r$. The pseudo-inverse or Moore-Penrose inverse of A is

$$A^\dagger = V\Sigma^{-1}U^\top \in \mathbf{R}^{n\times m}$$

Schur complement

- $A \in \mathbf{S}^k$, and a matrix $X \in \mathbf{S}^n$ partitioned as

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

- If $\det A \neq 0$, the matrix

$$S = C - B^\top A^{-1} B$$

is called the Schur complement of A in X.

Application of Schur complement

- PD Matrices
  - $X \succ 0$ if and only if $A \succ 0$ and $S \succ 0$
  - If $A \succ 0$, then $X \succcurlyeq 0$ if and only if $S \succcurlyeq 0$

- PSD Matrices

$$X \succcurlyeq 0 \Leftrightarrow A \succcurlyeq 0, \left(I - AA^{\dagger}\right)B = 0, C - B^{\top}A^{\dagger}B \succcurlyeq 0$$

Summary

- Norms of vectors
  - l_1-norm, l_2-norm, l_∞-norm, P-quadratic norm
- Norms of Matrices
  - Frobenius norm, spectral norm, nuclear norm
- Gradients of Common Functions
  - The Matrix Cookbook
- Eigendecompostion vs SVD
- PSD matrices