CS295: Convex Optimization

Xiaohui Xie
Department of Computer Science
University of California, Irvine

Course information

- Prerequisites: multivariate calculus and linear algebra
- Textbook: Convex Optimization by Boyd and Vandenberghe
- Course website:
  http://eee.uci.edu/wiki/index.php/CS_295_Convex_Optimization_(Winter_2011)
- Grading based on:
  - final exam (50%)
  - final project (50%)

Mathematical optimization

Mathematical **optimization problem**:

$$\begin{aligned} &\text{minimize} && f_0(\mathbf{x}) \\ &\text{subject to} && f_i(\mathbf{x}) \le \mathbf{b}_i, \quad i = 1, \cdots, m \end{aligned}$$

where

- $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$: optimization variables
- $f_0 : \mathbb{R}^n \to \mathbb{R}$: objective function
- $f_i : \mathbb{R}^n \to \mathbb{R}$: constraint function

Optimal solution $\mathbf{x}^*$ has smallest value of f_0 among all vectors that satisfy the constraints.

Examples

- transportation - product transportation plan
- finance - portfolio management
- machine learning - support vector machines, graphical model structure learning

Transportation problem

We have a product that can be produced in amounts a_i at location i with $i = 1, \cdots, m$. The product must be shipped to n destinations, in quantities b_j to destination j with $j = 1, \cdots, n$. The amount shipped from origin i to destination j is x_{ij}, at a cost of c_{ij} per unit.

Transportation problem

We have a product that can be produced in amounts a_i at location i with $i = 1, \cdots, m$. The product must be shipped to n destinations, in quantities b_j to destination j with $j = 1, \cdots, n$. The amount shipped from origin i to destination j is x_{ij}, at a cost of c_{ij} per unit.

To find the transportation plan that minimizes the total cost, we solve an LP:

$$
\begin{aligned}
\min \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} c_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} = a_i \quad i = 1, \cdots, m \\
& \sum_{i=1}^{m} x_{ij} = b_j \quad j = 1, \cdots, n \\
& x_{ij} \geq 0
\end{aligned}
$$

Markowitz portfolio optimization

Consider a simple portfolio selection problem with n stocks held over a period of time:

- $\mathbf{x} = (x_1, \cdots, x_n)$: the optimization variable with x_i denoting the amount to invest in stock i
- $\mathbf{p} = (p_1, \cdots, p_n)$: a random vector with p_i denoting the reward from stock i. Suppose its mean μ and covariance matrix Σ are known.
- $r = \mathbf{p}^T\mathbf{x}$: the overall return on the portfolio. r is a random variable with mean $\mu^T\mathbf{x}$ and variance $x^T\Sigma x$.

Markowitz portfolio optimization

The Markowitz portfolio optimization problem is the QP

$$\begin{aligned} \min \quad & \mathbf{x}^T \Sigma \mathbf{x} \\ \text{s.t.} \quad & \mu^T \mathbf{x} \geq r_{\min} \\ & 1^T x = B \\ & x_i \geq 0, \quad i = 1, \cdots, n \end{aligned}$$

which find the portfolio that minimizes the return variance subject to three constraints:

- achieving a minimum acceptable mean return $r_{\min}$
- satisfying the total budget B
- no short positions ($x_i \geq 0$)

Support vector machines (SVMs)

Input: a set of training data,

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}, i = 1, \cdots, n\}$$

where y_i is either 1 or -1, indicating the class to which $\mathbf{x}_i$ belongs.

Problem: find the **optimal separating hyperplane** that separates the two classes and maximizes the distance to the closet point from either class.

Support vector machines (SVMs) 2

Define a hyperplane by $w^T x - b = 0$. Suppose the training data are linearly separably. So we can find w and b such that $w^T x_i - b \geq 1$ for all x_i from class 1 and $w^T x_i - b \leq -1$ for all x_i from class -1.

The distance between the two parallel hyperplans, $w^T x_i - b = 1$ and $w^T x_i - b = -1$, is $\frac{2}{\|w\|}$, called **margin**.

To find the optimal separating hyperplane, we choose w and b that maximize the margin:

$$\begin{aligned} \min \quad & \|w\|^2 \\ \text{s.t.} \quad & y_i(w^T x_i - b) \geq 1, \quad i = 1, \cdots, n \end{aligned}$$

Undirected graphical models

Input: a set of training data,

$$\mathcal{D} = \{(\mathbf{x}_i) \mid \mathbf{x}_i \in \mathbb{R}^p \; i = 1, \cdots, n\}$$

Assume the data were sampled from a Gaussian graphical model with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. The inverse covariance matrix, Σ^{-1}, encodes the structure of the graphical model in the sense that the variables i and j are connected only if the (i, j)-entry of Σ^{-1} is nonzero.

Problem: Find the maximum likelihood estimation of Σ^{-1} with a sparsity constraint, $\|\Sigma^{-1}\|_1 \leq \lambda$.

Undirected graphical models 2

Let S be the empirical covariance matrix:

$$S := \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T.$$

Denote $\Theta = \Sigma^{-1}$.

The convex optimization problem:

$$\begin{aligned} \min \quad & -\log\det\Theta + \mathrm{tr}(S\Theta) \\ \text{s.t.} \quad & \|\Theta\|_1 \leq \lambda \\ & \Theta \succ 0 \end{aligned}$$

Solving optimization problems

The optimization problem is in general difficult to solve: taking very long long time, or not always finding the solution

Exceptions: certain classes of problems can be solved efficiently:

- least-square problems
- linear programming problems
- convex optimization problems

Least-squares

$$\text{minimize} \quad \|Ax - b\|_2^2$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k \times n}$.

- analytical solution: $x^* = (A^T A)^{-1} A^T b$ (assuming $k > n$ and $\mathbf{rank}\ A = n$)
- reliable and efficient algorithms available
- computational time proportional to $n^2 k$, and can be further reduced if A has some special structure

Linear programming

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & a_i^T x \leq b_i, \quad i = 1, \cdots, m \end{aligned}$$

where the optimization variable $x \in \mathbb{R}^n$, and $c, a_i, b_i \in \mathbb{R}^n$ are parameters.

- no analytical formula for solution
- reliable and efficient algorithms available (e.g., Dantzig's simplex method, interior-point method)
- computational time proportional to $n^2 m$ if $m \leq n$ (interior-point method); less with structure

Linear programming: example

The Chebyshev approximation problem:

$$\text{minimize} \quad \|Ax - b\|_\infty$$

with $x \in \mathbb{R}^n$, $b \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k\times n}$. The problem is similar to the least-square problem, but with the ℓ_∞-norm replacing the ℓ_2-norm:

$$\|Ax - b\|_\infty = \max_{i=1,\cdots,k} |a_i^T x - b_i|$$

where $a_i \in R^n$ is the ith column of A^T.

Linear programming: example

The Chebyshev approximation problem:

$$\text{minimize} \quad \|Ax - b\|_\infty$$

with $x \in \mathbb{R}^n$, $b \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k\times n}$. The problem is similar to the least-square problem, but with the ℓ_∞-norm replacing the ℓ_2-norm:

$$\|Ax - b\|_\infty = \max_{i=1,\cdots,k} |a_i^T x - b_i|$$

where $a_i \in R^n$ is the ith column of A^T.

An equivalent linear programming:

$$\begin{aligned} \min \quad & t \\ \text{s.t.} \quad & a_i^T x - t \leq b_i, \quad i = 1, \cdots, k \\ & -a_i^T x - t \leq -b_i, \quad i = 1, \cdots, k \end{aligned}$$

Convex optimization problems

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq \mathbf{b}_i, \quad i = 1, \cdots, m \end{aligned}$$

where $x \in \mathbb{R}^n$.

- both objective and constraint functions are convex

  $$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

  for any $0 \leq \theta \leq 1$, and any x and y in the domain of f_0 and f_i for all i.
- includes least-square and linear programming problems as special cases.
- no analytical formula for solution
- reliable and efficient algorithms available

Topics to be covered

- Convex sets and convex functions
- Duality
- Unconstrained optimization
- Equality constrained optimization
- Interior-point methods
- Semidefinite programming

Brief history of optimization

- 1700s: theory for unconstrained optimization (Fermat, Newton, Euler)
- 1797: theory for equality constrained optimization (Lagrange)
- 1947: simplex method for linear programming (Dantzig)
- 1960s: early interior-point methods (Fiacco, McCormick, Dikin, etc)
- 1970s: ellipsoid method and other subgradient methods
- 1980s: polynomial-time interior-point methods for linear programming (Karmarkar)
- 1990s: polynomial-time interior-point methods for nonlinear convex optimization (Nesterorv & Nemirovski)
- 1990-now: many new applications in engineering (control, signal processing, communications, etc); new problem classes (semidefinite and second-order cone programming, robust optimization, convex relaxation, etc)

Convex set

Definition

A set C is called **convex** if

$$\mathbf{x}, \mathbf{y} \in C \implies \theta \mathbf{x} + (1-\theta)\mathbf{y} \in C \quad \forall \theta \in [0, 1]$$

In other words, a set C is convex if the line segment between any two points in C lies in C.

Convex combination

Definition

A **convex combination** of the points $x_1, \cdots, x_k$ is a point of the form

$$\theta_1 x_1 + \cdots + \theta_k x_k,$$

where $\theta_1 + \cdots + \theta_k = 1$ and $\theta_i \geq 0$ for all $i = 1, \cdots, k$.

A set is convex if and only if it contains every convex combinations of the its points.

Convex hull

Definition

The **convex hull** of a set C, denoted **conv** C, is the set of all convex combinations of points in C:

$$\textbf{conv}\ C = \left\{ \sum_{i=1}^{k} \theta_i x_i \mid x_i \in C, \theta_i \geq 0, i = 1, \cdots, k, \sum_{i=1}^{k} \theta_k = 1 \right\}$$

Properties:

- A convex hull is always convex
- **conv** C is the smallest convex set that contains C, i.e., $B \supseteq C$ is convex $\implies \textbf{conv}\ C \subseteq B$

Convex cone

A set C is called a **cone** if $x \in C \implies \theta x \in C, \ \forall \theta \geq 0$.

A set C is a **convex cone** if it is convex and a cone, i.e.,

$$x_1, x_2 \in C \implies \theta_1 x_1 + \theta_2 x_2 \in C, \quad \forall \theta_1, \theta_2 \geq 0$$

The point $\sum_{i=1}^{k} \theta_i x_i$, where $\theta_i \geq 0, \forall i = 1, \cdots, k$, is called a **conic combination** of $x_1, \cdots, x_k$.

The **conic hull** of a set C is the set of all conic combinations of points in C.

Hyperplanes and halfspaces

A **hyperplane** is a set of the form $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = b\}$ where $a \neq 0, b \in \mathbb{R}$.

A (closed) **halfspace** is a set of the form $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} \leq b\}$ where $a \neq 0, b \in \mathbb{R}$.

- **a** is the normal vector
- hyperplanes and halfspaces are convex

Euclidean balls and ellipsoids

Euclidean ball in R^n with center x_c and radius r:

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \leq r\} = \{x_c + ru \mid \|u\|_2 \leq 1\}$$

Euclidean balls and ellipsoids

Euclidean ball in R^n with center x_c and radius r:

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \leq r\} = \{x_c + ru \mid \|u\|_2 \leq 1\}$$

ellipsoid in R^n with center x_c:

$$\mathcal{E} = \left\{x \mid (x - x_c)^T P^{-1}(x - x_c) \leq 1\right\}$$

where $P \in S^n_{++}$ (i.e., symmetric and positive definite)

- the lengths of the semi-axes of $\mathcal{E}$ are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of P.
- An alternative representation of an ellipsoid: with $A = P^{1/2}$

$$\mathcal{E} = \{x_c + Au \mid \|u\|_2 \leq 1\}$$

Euclidean balls and ellipsoids

Euclidean ball in R^n with center x_c and radius r:

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \leq r\} = \{x_c + ru \mid \|u\|_2 \leq 1\}$$

ellipsoid in R^n with center x_c:

$$\mathcal{E} = \left\{x \mid (x - x_c)^T P^{-1}(x - x_c) \leq 1\right\}$$

where $P \in S^n_{++}$ (i.e., symmetric and positive definite)

- the lengths of the semi-axes of $\mathcal{E}$ are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of P.
- An alternative representation of an ellipsoid: with $A = P^{1/2}$

$$\mathcal{E} = \{x_c + Au \mid \|u\|_2 \leq 1\}$$

Euclidean balls and ellipsoids are convex.

Norms

A function $f : R^n \to R$ is called a **norm**, denoted $\|x\|$, if

- nonegative: $f(x) \geq 0$, for all $x \in R^n$
- definite: $f(x) = 0$ only if $x = 0$
- homogeneous: $f(tx) = |t| f(x)$, for all $x \in R^n$ and $t \in R$
- satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$

notation: $\| \cdot \|$ denotes a general norm; $\| \cdot \|_{\mathrm{symb}}$ denotes a specific norm

Distance: $dist(x, y) = \|x - y\|$ between $x, y \in R^n$.

Examples of norms

- ℓ_p-norm on R^n: $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$
  - ℓ_1-norm: $\|x\|_1 = \sum_i |x_i|$
  - ℓ_∞-norm: $\|x\|_\infty = \max_i |x_i|$
- Quadratic norms: For $P \in S^n_{++}$, define the P-quadratic norm as

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

Equivalence of norms

Let $\| \cdot \|_a$ and $\| \cdot \|_b$ be norms on R^n. Then $\exists \alpha, \beta > 0$ such that $\forall x \in R^n$,

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

Norms on any finite-dimensional vector space are equivalent (define the same set of open subsets, the same set of convergent sequences, etc.)

Norm balls and norm cones

norm ball with center x_c and radius r: $\{x \mid \|x - x_c\| \leq r\}$

norm cone: $C = \{(x, t) \mid \|x\| \leq t\} \subseteq \mathbb{R}^{n+1}$

- the second-order cone is the norm cone for the Euclidean norm

norm balls and cones are convex

Polyhedra

A **polyhedron** is defined as the solution set of a finite number of linear equalities and inequalities:

$$\mathcal{P} = \{x \mid Ax \preceq b, Cx = d\}$$

where $A \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{p \times n}$, and $\preceq$ denotes *vector inequality* or *componentwise inequality*.

A polyhedron is the intersection of finite number of halfspaces and hyperplanes.

Simplexes

The **simplex** determined by $k+1$ affinely independent points $v_0, \cdots, v_k \in \mathbb{R}^n$ is

$$C = \textbf{conv}\{v_0, \cdots, v_k\} = \left\{\theta_0 v_0 + \cdots + \theta_k v_k \mid \theta \succeq 0, \mathbf{1}^T\theta = 1\right\}$$

The affine dimension of this simplex is k, so it is often called k-dimensional simplex in $\mathbb{R}^n$.

Some common simplexes: let $e_1, \cdots, e_n$ be the unit vectors in R^n.

- **unit simplex**: $\textbf{conv}\{0, e_1, \cdots, e_n\} = \{x | x \succeq 0, \mathbf{1}^T\theta \leq 1\}$
- **probability simplex**: $\textbf{conv}\{e_1, \cdots, e_n\} = \{x | x \succeq 0, \mathbf{1}^T\theta = 1\}$

Positive semidefinite cone

notation:

- S^n: the set of symmetric $n \times n$ matrices
- $S^n_+ = \{X \in S^n \mid X \succeq 0\}$: symmetric positive semidefinite matrices
- $S^n_{++} = \{X \in S^n \mid X \succ 0\}$ symmetric positive definite matrices

S^n_+ is a convex cone, called positive semidefinte cone. S^n_{++} comprise the cone interior; all singular positive semidefinite matrices reside on the cone boundary.

Example:

$$X = \begin{bmatrix} x & y \\ y & z \end{bmatrix} \in S^2_+ \Longleftrightarrow x \geq 0, z \geq 0, xz \geq y^2$$

Operations that preserve complexity

- intersection
- affine function
- perspective function
- linear-fractional functions

Inner product, Euclidean norm

- Standard inner product on R^n: $\langle x, y \rangle = x^T y = \sum_i x_i y_i$
- Euclidean norm (ℓ_2 norm): $\|x\|_2 = \langle x, x \rangle^{1/2}$
- Cauchy-Schwartz inequality: $\langle x, y \rangle \leq \|x\|_2 \, \|y\|_2$
- Standard inner product on $R^{m \times n}$:

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

- Frobenius norm: $\|X\|_F = \langle X, X \rangle^{1/2}$

Norms and distance

- A function $f : R^n \to R$ with $dom\ f = R^n$ is called a *norm*, written as $f(x) = \|x\|$, if
  - $f(x) \geq 0$, for all $x \in R^n$
  - $f(x) = 0$ only if $x = 0$
  - $f(tx) = |t| f(x)$, for all $x \in R^n$ and $t \in R$
  - $f(x + y) \leq f(x) + f(y)$
- Distance: $dist(x, y) = \|x - y\|$ between $x, y \in R^n$.
- Unit ball: $B = \{x \in R^n | \|x\| \leq 1\}$
  - B is convex
  - B is closed, bounded, and has nonempty interior
  - B is symmetric about the origin, i.e., $x \in B$ iff $-x \in B$.

Examples of norms

- ℓ_p-norm on R^n: $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$
  - ℓ_1-norm: $\|x\|_1 = \sum_i |x_i|$
  - ℓ_∞-norm: $\|x\|_\infty = \max_i |x_i|$
- Quadratic norms: For $P \in S^n_{++}$, define the P-quadratic norm as

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

Equivalence of norms

Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be norms on R^n. Then $\exists \alpha, \beta > 0$ such that $\forall x \in R^n$,

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

In fact, norms on any finite-dimensional vector space are equivalent (define the same set of open subsets, the same set of convergent sequences, etc.)

- Let $\|\cdot\|$ be a norm on R^n. Then $\exists$ a quadratic norm $\|\cdot\|_P$ such that $\forall x \in R^n$,

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

“Minimum Norms” Lemma

Lemma

Suppose X is an n-dimensional normed vector space over $\mathbb{R}$ or ($\mathbb{C}$) with basis $\{x_1, \cdots, x_n\}$. There exists a $c > 0$ such that

$$\|\alpha_1 x_1 + \cdots + \alpha_n x_n\| \geq c(|\alpha_1| + \cdots + |\alpha_n|)$$

for any selection of $\alpha_1, \cdots, \alpha_n$ in the field.

Operator norms

Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be norms on R^m and R^n, respectively. The operator norm of $X \in R^{m\times n}$, induced by $\|\cdot\|_a$ and $\|\cdot\|_b$, is defined to be

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

- Spectral norm (ℓ_2-norm):

$$\|X\|_2 = \|X\|_{2,2} = \sigma_{\max}(X) = (\lambda_{\max}(X^T X))^{1/2}$$

- Max-row-sum norm:

$$\|X\|_\infty = \|X\|_{\infty,\infty} = \max_{i=1,\cdots,m} \sum_{j=1}^{n} |X_{ij}|$$

- Max-column-sum norm:

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

Dual norm

Let $\|\cdot\|$ be a norm on R^n. The associated dual norm, denoted $\|\cdot\|_*$, is defined as

$$\|z\|_* = \sup\ \{z^T x \mid \|x\| \leq 1\}.$$

- $z^T x \leq \|x\|\ \|z\|_*$ for all $x, z \in R^n$
- $\|x\|_{**} = \|x\|$ for all $x \in R^n$
- The dual of the Euclidean norm is the Euclidean norm
- The dual of the ℓ_∞ norm is the ℓ_1 norm
- The dual of the ℓ_p-norm is the ℓ_q-norm, where $1/p + 1/q = 1$
- The dual of the ℓ_2-norm on $R^{m \times n}$ is the nuclear norm,

$$\begin{aligned}\|Z\|_{2*} &= \sup\ \{tr(Z^T X) \mid \|X\|_2 \leq 1\} \\ &= \sigma_1(Z) + \cdots + \sigma_r(Z) = tr(Z^T Z)^{1/2},\end{aligned}$$

where $r = rank\ Z$.

Continuity

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $x \in \textit{dom } f$ if $\forall \epsilon > 0 \; \exists \; \delta > 0$ such that

$$\|y - x\| < \delta \implies \|f(y) - f(x)\| < \epsilon.$$

Continuity can also be described in terms of limits: whenever the sequence (x_i) converges to a point $x \in \textit{dom } f$, the sequence $(f(x_i))$ converges to $f(x)$,

$$\lim_{i \to \infty} f(x_i) = f(\lim_{i \to \infty} x_i).$$

A function f is **continuous** if it is continuous at every point in its domain.

Derivatives

The function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $x \in$ *int dom* f if there exists a matrix $Df(x) \in \mathbb{R}^{m\times n}$ that satisfies

$$\lim_{z\to x} \frac{\|f(z) - f(x) - Df(x)(z-x)\|}{\|z-x\|} = 0,$$

with $z \in$ *dom* $f \backslash \{x\}$. $Df(x)$ is called the **derivative** of f at x. The function f is differentiable if *dom* f is open, and it is differentiable at every point in its domain.

The derivative can be found from partial derivatives:

$$Df(x)_{ij} = \frac{\partial f_i(x)}{\partial x_j},$$

for all $i = 1, \cdots, m$, and $j = 1, \cdots, n$.

Gradient

The gradient of the function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\nabla f(x) = Df(x)^T,$$

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

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

The first-order approximation of f at $x \in \mathit{int\ dom}\ f$ is

$$f(x) + \nabla f(x)^T (z - x).$$

Chain rule

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \textit{int dom } f$, and $g : \mathbb{R}^m \to \mathbb{R}^p$ are differentiable at $f(x) \in \textit{int dom } g$. Define the composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(x) = g(f(x))$. Then h is differentiable at x, with derivative

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

Examples:

- $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$:

$$\nabla(g \circ f)(x) = g'(f(x))\nabla f(x)$$

- $h(x) = f(Ax + b)$, where $A \in \mathbb{R}^{n\times m}$ and $g : \mathbb{R}^m \to \mathbb{R}$:

$$\nabla h(x) = A^T \nabla f(Ax + b)$$

Second derivative

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

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

provided f is twice differentiable at x.

The second-order approximation of f, at or near x, is:

$$\hat{f}(z) = f(x) + \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x).$$

Chain rule for second derivative

Some special cases:

- $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$:

$$\nabla^2 (g \circ f)(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x) \nabla f(x)^T$$

- $h(x) = f(Ax + b)$, where $A \in \mathbb{R}^{n \times m}$ and $g : \mathbb{R}^m \to \mathbb{R}$:

$$\nabla^2 h(x) = A^T \nabla^2 f(Ax + b) A$$

Matrix calculus

Suppose $A \in \mathbb{R}^{n \times n}$. adj(A) denotes the adjugate of A, the transpose of the cofactor matrix of A. The derivative of $\det(A)$,

$$\frac{\mathrm{d}\det(A)}{\mathrm{d}\alpha} = \mathrm{tr}\left(\mathrm{adj}(A)\frac{\mathrm{d}A}{\mathrm{d}\alpha}\right).$$

If A is invertible,

$$\frac{\mathrm{d}\det(A)}{\mathrm{d}\alpha} = \det(A)\,\mathrm{tr}\left(A^{-1}\frac{\mathrm{d}A}{\mathrm{d}\alpha}\right).$$

In particular,

$$\frac{\partial \det(A)}{\partial A_{ij}} = \mathrm{adj}(A)_{ji} = \det(A)(A^{-1})_{ji}$$

$$\frac{\partial \log\det(A)}{\partial A_{ij}} = (A^{-1})_{ji}$$

Derivative of $\det(A)$

Proof.

Denote C the cofactor matrix of A. $\det(A) = \sum_k A_{ik} C_{ik}$.

$$
\begin{aligned}
\frac{\mathrm{d}\det(\mathrm{A})}{\mathrm{d}\alpha} &= \sum_i \sum_j \frac{\partial \det(\mathrm{A})}{\partial \mathrm{A}_{ij}} \frac{\mathrm{d}\mathrm{A}_{ij}}{\mathrm{d}\alpha} \\
&= \sum_i \sum_j \frac{\partial}{\partial \mathrm{A}_{ij}} \sum_k \mathrm{A}_{ik} C_{ik} \frac{\mathrm{d}\mathrm{A}_{ij}}{\mathrm{d}\alpha} \\
&= \sum_i \sum_j C_{ij} \frac{\mathrm{d}\mathrm{A}_{ij}}{\mathrm{d}\alpha} \\
&= \mathrm{tr}\left(\mathrm{adj}(\mathrm{A}) \frac{\mathrm{d}\mathrm{A}}{\mathrm{d}\alpha} \right).
\end{aligned}
$$

□

Example: the gradient of the log det function

Consider the function $f : S^n_{++} \to \mathbb{R}$, given by $f(X) = \log\det X$.
The first-order approximation of f is

$$\log\det(X + \Delta X) = \log\det(X) + \mathrm{tr}(X^{-1}\Delta X),$$

which implies that

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

Example: the gradient of the log det function II

The result can be proved using the formula on the derivative of det(X) function. But here we use a different technique based on the first-order approximation.

$$
\begin{aligned}
\log\det(X + \Delta X) &= \log\det\left(X^{1/2}(I + X^{-1/2}\Delta X X^{-1/2})X^{1/2}\right) \\
&= \log\det X + \log\det(I + X^{-1/2}\Delta X X^{-1/2}) \\
&= \log\det X + \sum_i \log(1 + \lambda_i) \\
&\approx \log\det X + \sum_i \lambda_i \\
&= \log\det X + \mathrm{tr}(X^{-1/2}\Delta X X^{-1/2}) \\
&= \log\det X + \mathrm{tr}(X^{-1}\Delta X),
\end{aligned}
$$

where λ_i is the ith eigenvalue of $X^{-1/2}\Delta X X^{-1/2}$.