Convex Analysis and Optimization Methods

Zhouchen Lin (林宙辰)
Dept. of Machine Intelligence, School of EECS
Peking University

Chapter 1: Overview

- About me and the course
- Overview of optimization

About me

- IEEE/IAPR Fellow
- NSF Distinguished Young Researcher
- Research Areas: Machine Learning, Computer Vision
- Associate Editor of IEEE T. PAMI and IJCV

Welcome to apply to me!

4G

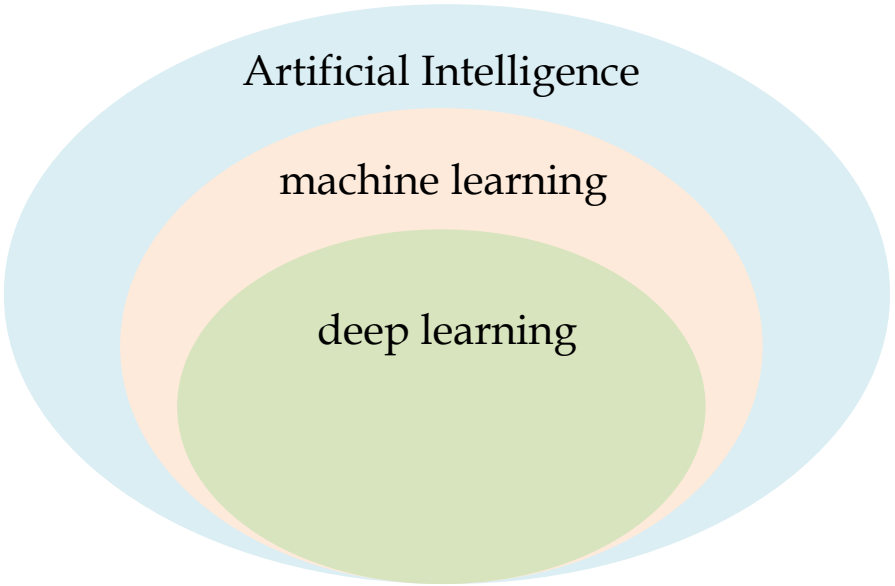
- Good math
- Good coding
- Good English
- **Good attitude**

- Standing Committee of CCF Computer Vision Professional Committee

What is this course?

- Optimization in machine learning
- Convex analysis

Machine Learning studies how to empower computers to automatically improve their own abilities by utilizing data.



What is optimization?

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t. \quad \mathbf{x} \in \mathcal{C}.$$

$$\min_{\mathbf{x} \in \mathcal{C}} \; f(\mathbf{x}).$$

$$\min \|\mathbf{W}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_1.$$

Exemplar optimization problems in ML

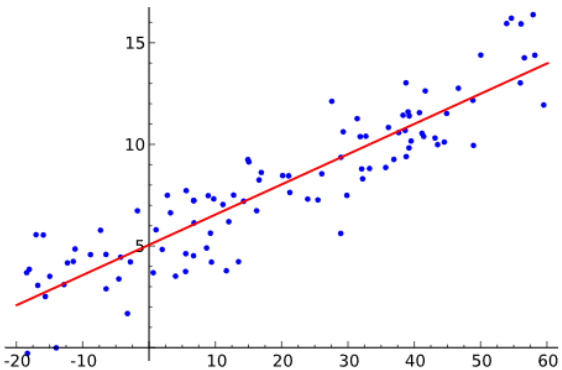
- Typical formulation

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} f_i(\mathbf{x}) + \lambda\mathcal{R}(\mathbf{x}),$$

where the functions $f_1, \ldots, f_m, \mathcal{R}$ are convex, $\mathcal{R}(\mathbf{x})$ is a regularizer and $\lambda \geq 0$ is a is a fixed parameter.

- Ridge regression

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{W}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|^2.$$



- LASSO

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{W}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

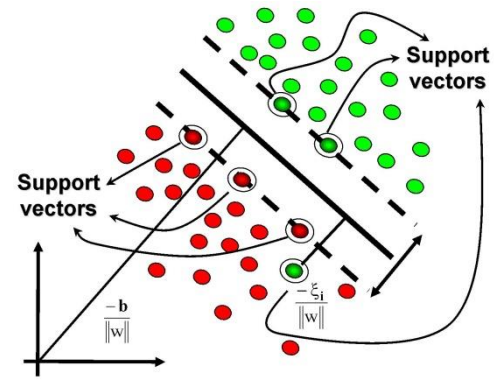
- Sparse inverse covariance estimation

$$\min \operatorname{tr}(\mathbf{X}^T\mathbf{Y}) - \log\det(\mathbf{X}) + \lambda\|\mathbf{X}\|_1,$$
$$s.t. \quad \mathbf{X} \in \mathbb{R}^{n\times n}, \mathbf{X}^\top = \mathbf{X}, \mathbf{X} \succeq \mathbf{0}.$$

Exemplar optimization problems in ML

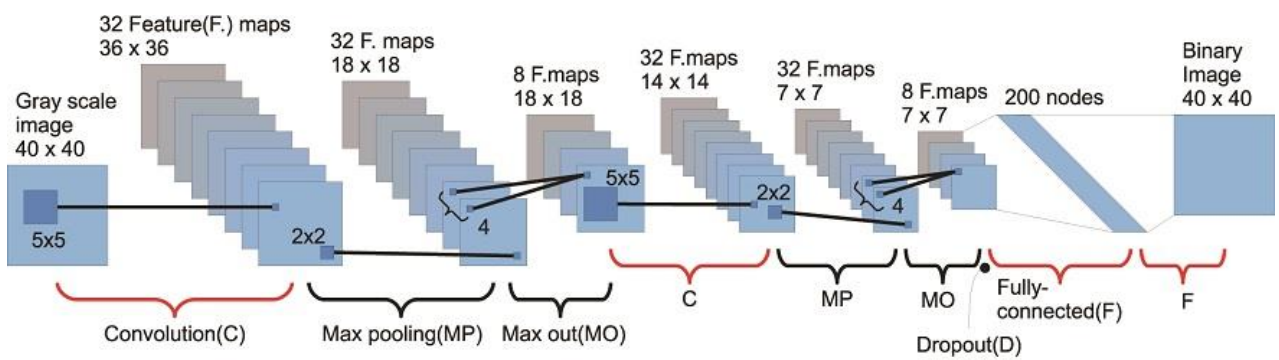
- Support Vector Machine

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2, \quad s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, \cdots, N.$$



- Deep Neural Network

$$\min_{\{\mathbf{W}^i\}} \ell\left(\phi(\mathbf{W}^{n-1}\phi(\cdots\phi(\mathbf{W}^2\phi(\mathbf{W}^1\mathbf{X}^1))\cdots)), \mathbf{L}\right).$$



See Claudio Gambella, Bissan Ghaddar, and Joe Naoum-Sawaya, *Optimization Models for Machine Learning: A Survey*, 2019.

Why is optimization important?

- Pedro Domingos (AAAI Fellow, Prof. @ UW):

Machine Learning

= Representation + Evaluation

+ Optimization

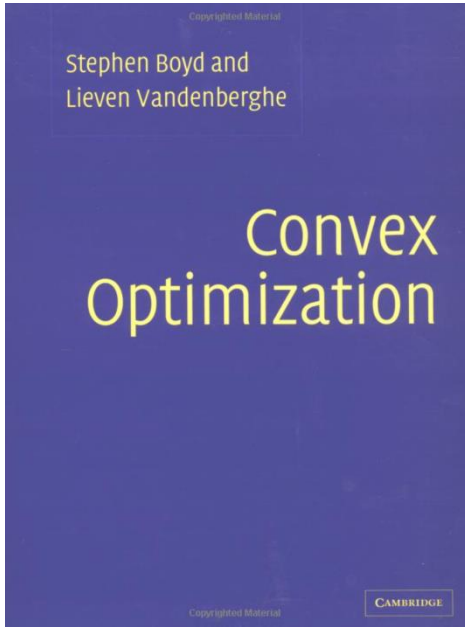
Table 1: The three components of learning algorithms.

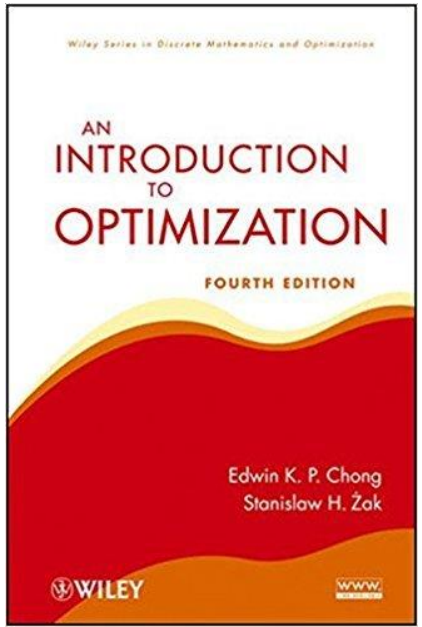
Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

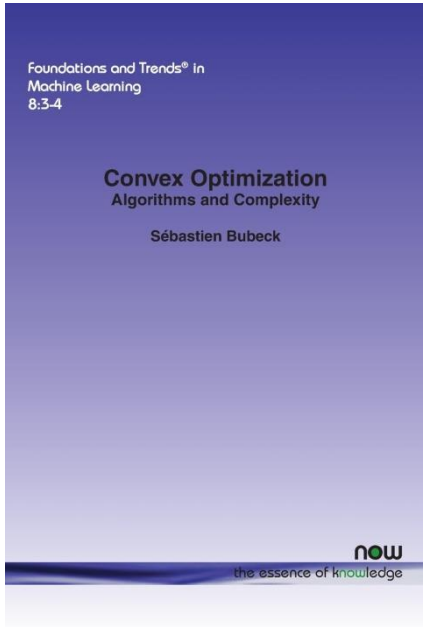
P. Domingos. A few useful things to know about machine learning. Communications of the ACM, 55(10):78-87, 2012.

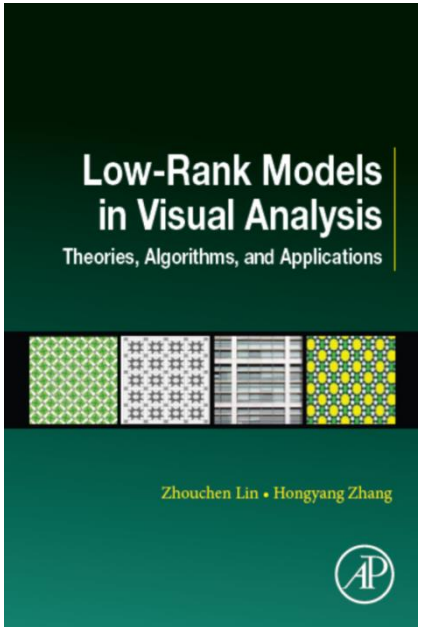
Reference books

- S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004
- Edwin K. P. Chong and Stanislaw H. Zak, An Introduction to Optimization, 4th Edition, John Wiley & Sons, Inc., 2013
- Sébastien Bubeck, Convex Optimization: Algorithms and Complexity, Now Publishers, 2015
- Zhouchen Lin and Hongyang Zhang, Low-Rank Models in Visual Analysis, Academic Press, 2017









Lecture notes

- The theories and algorithms are tailored for machine learning (but the examples and exercises are not yet done)
- First order methods: good balance between speed and numerical accuracy

- Contents:
  - Overview of optimization
  - Background mathematics
  - Convex sets
  - Convex functions
  - Optimality conditions
  - Unconstrained optimization
  - Constrained optimization
  - Acceleration techniques
  - Stochastic gradient

Reference books

- Elad Hazan, Optimization for Machine Learning, draft.

lecture notes:

Optimization for Machine Learning

version 0.54

Elad Hazan

Teaching assistant

- 陈程：jackchen1990@gmail.com

No-free-lunch Theorem for Opt.

- D. H. Wolpert and W. G. Macready (1997):
  - If algorithm *A* performs better than algorithm *B* for some optimization functions, then *B* will outperform *A* for other functions.
  - If averaged over all possible function space, both algorithms *A* and *B* will perform on average equally well.
  - There is no universally better algorithms exist.

Theorem 1. *For any two algorithms a_1 and a_2 on a finite search space, iterating m times,*

$$\sum_f P(\mathbf{d}_m^y | f, m, a_1) = \sum_f P(\mathbf{d}_m^y | f, m, a_2),$$

where $\mathbf{d}_m^y = \{\mathbf{d}_m^y(1), \cdots, \mathbf{d}_m^y(m)\}$ is the cost of evaluation at points $\mathbf{d}_m^x = \{\mathbf{d}_m^x(1), \cdots, \mathbf{d}_m^x(m)\}$ and P is a probability distribution.

D. H. Wolpert and W. G. Macready, No free lunch theorems for optimization, IEEE T. Evolutionary Computation, 1, 67-82 (1997).

Classification of opt. problems

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t. \quad \mathbf{x} \in \mathcal{C}.$$

- Classifications take into account:
  - the nature of the solution set $\mathcal{C}$.
  - the description (definition) of the solution set $\mathcal{C}$.
  - the properties of the objective function f.

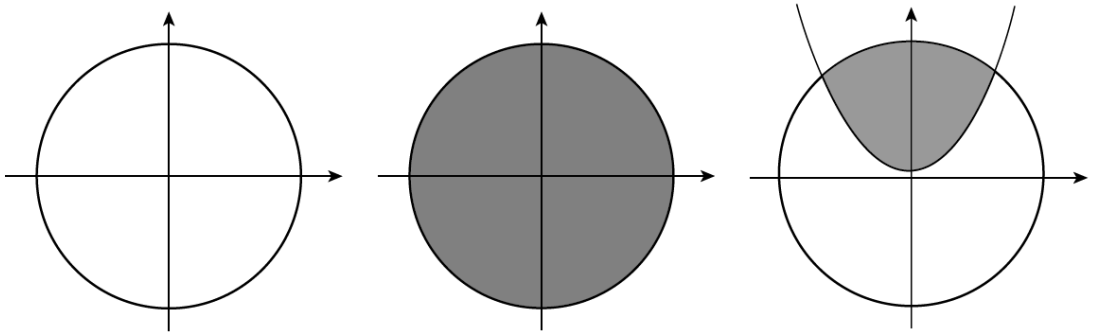
Continuous opt. problems

An optimization problem is called *continuous* when the solution set $\mathcal{C}$ is a continuous subset of $\mathbb{R}^n$.

$$\min_{x,y} f(x,y) = 3x + 4y + 1,$$
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 = 1\}.$$

$$\min_{x,y} f(x,y) = 3x + 4y + 1,$$
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1\}.$$

$$\min_{x,y} f(x,y) = 3x + 4y + 1,$$
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1, y - x^2 \geq 0\}.$$



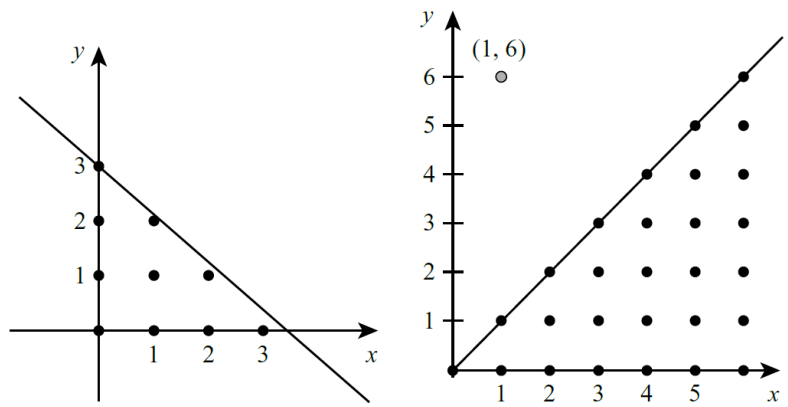
(a) (b) (c)

Discrete opt. problems

An optimization problem is called *discrete* when the solution set $\mathcal{C}$ is a discrete subset of $\mathbb{R}^n$.

$$\min_{x,y} f(x,y) = x + y,$$
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{Z}^2 | 6x + 7y = 21, x \geq 0, y \geq 0\}.$$

$$\min_{x,y} f(x,y) = (x-1)^2 + (y-6)^2,$$
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{Z}^2 | y \leq x, x \geq 0, y \geq 0\}.$$

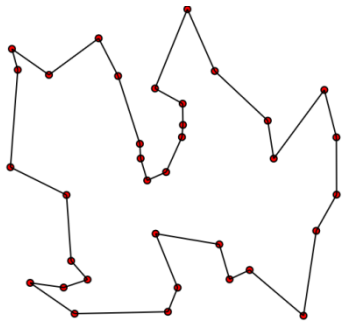


(a) (b)

Combinatorial opt. problems

An optimization problem is said to be *combinatorial* when its solution set $\mathcal{C}$ is finite. Usually, the elements of $\mathcal{C}$ are not explicitly determined. Instead, they are indirectly specified through combinatorial relations.

The *Traveling Salesman Problem.* Given n towns, find the minimum length route that starts in a given town, goes through each one of the other towns, and ends in the starting town.

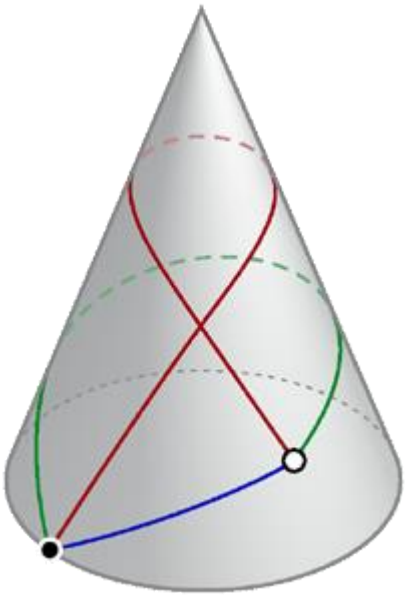


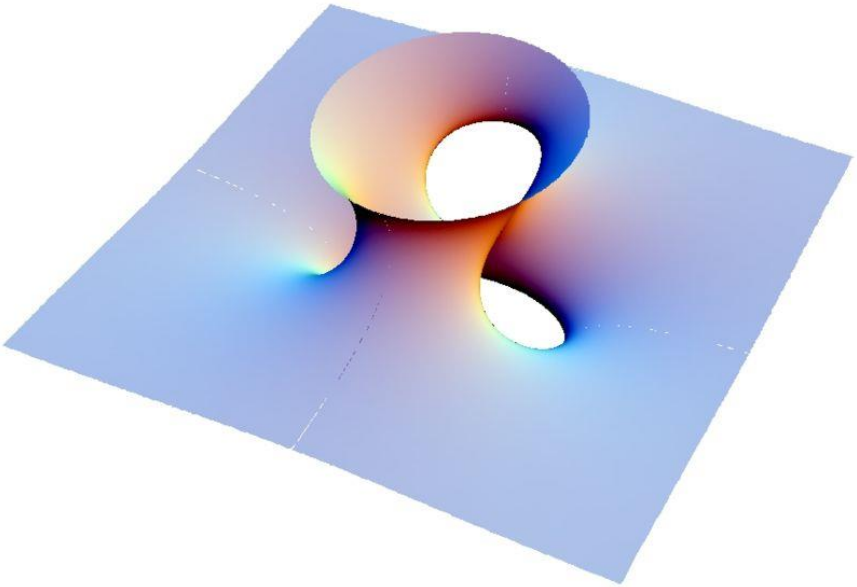
Variational opt. problems

An optimization problem is called a *variational problem* when its solution set $\mathcal{C}$ is an infinite dimensional subset of a space of functions.

The *Geodesic Problem.* Find the path of minimum length joining two points p_1 and p_2 of a given surface.

The *Minimal Surface Problem.* Find the surface of minimum area for a given boundary curve.





Classification based on constraints

- Unconstrained: $\mathcal{C} = \mathbb{R}^n$.

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{W}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

- Constrained:
  - Equality constrained: $h_i(\mathbf{x}) = 0, i = 1, \cdots, m;$
  - Inequality constrained: $g_j(\mathbf{x}) \leq 0, j = 1, \cdots, J.$

$$\min_{x,y} f(x,y) = 3x + 4y + 1,$$

$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 = 1\}.$$

$$\min_{x,y} f(x,y) = 3x + 4y + 1,$$

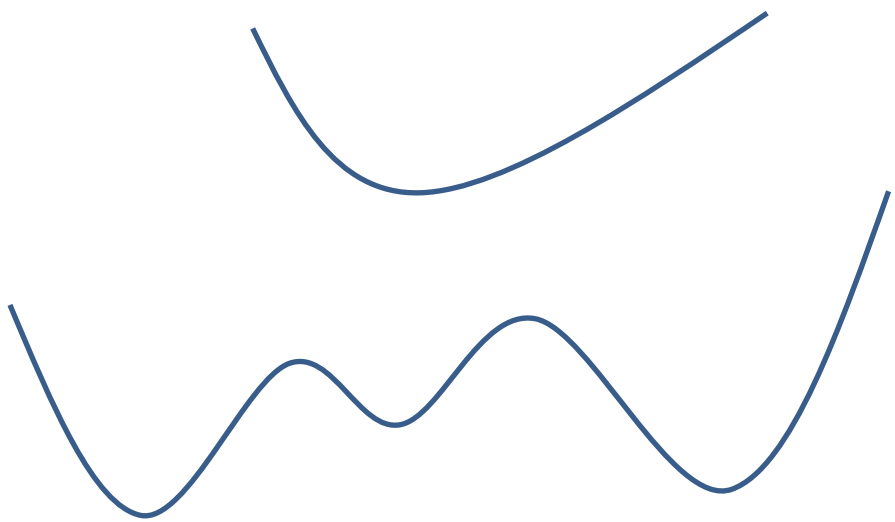
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1\}.$$

$$\min_{x,y} f(x,y) = 3x + 4y + 1,$$

$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1, y - x^2 \geq 0\}.$$

Classification based on the objective function

- Convex vs. nonconvex programs
  - R. T. Rockfeller: *In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.*



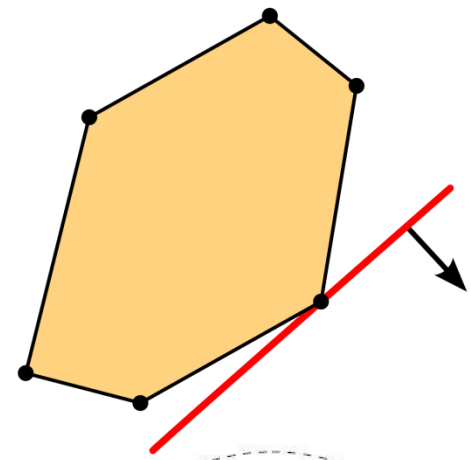
R. T. Rockafellar. Lagrange multipliers and optimality. SIAM Review, 35:183-283, 1993.

Classification based on the objective function

- Convex programs

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t. \quad \mathbf{x} \in \mathcal{C},$$

where f is a convex function and $\mathcal{C}$ is a convex set.

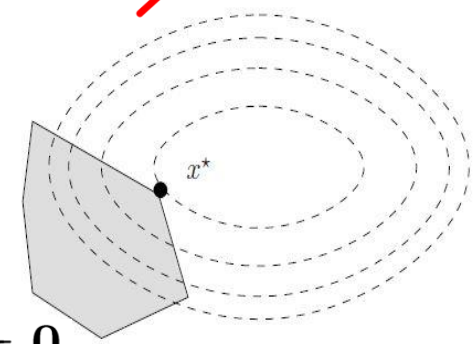


  - Linear programs

$$\min_{\mathbf{x}} \; \mathbf{c}^T\mathbf{x}, \quad s.t. \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}.$$

  - Quadratic programs

$$\min_{\mathbf{x}} \; \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}, \quad s.t. \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}.$$



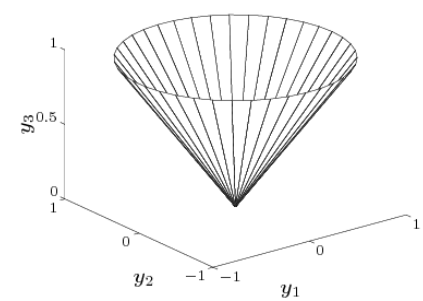
  - Semi-definite programs

$$\min_{\mathbf{X}} \; \langle \mathbf{C}, \mathbf{X} \rangle, \quad s.t. \quad \langle \mathbf{X}, \mathbf{A}_i \rangle \leq b_i, i = 1, \cdots, K, \mathbf{X} \succeq \mathbf{0}.$$

  - Second-order cone programs

$$\min_{\mathbf{x}} \; \mathbf{c}^T\mathbf{x},$$

$$s.t. \; \|\mathbf{A}_i\mathbf{x} + \mathbf{b}_i\|_2 \leq \mathbf{c}_i^T\mathbf{x} + d_i, i = 1, \cdots, K, \mathbf{F}\mathbf{x} = \mathbf{g}.$$



Classification based on the objective function

- Convex programs in machine learning
  - Linear regression $$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{W}\mathbf{x}-\mathbf{y}\|_2^2.$$
  - Ridge regression $$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{W}\mathbf{x}-\mathbf{y}\|_2^2+\lambda\|\mathbf{x}\|^2.$$
  - LASSO regression $$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{W}\mathbf{x}-\mathbf{y}\|_2^2+\lambda\|\mathbf{x}\|_1.$$
  - Logistic regression
  - Softmax regression $$\min_{\boldsymbol{\beta}} \sum_{\alpha=1}^{m} \left\{ n_\alpha \ln[1+\exp(\tilde{\mathbf{x}}_\alpha^T\boldsymbol{\beta})] - y_\alpha \tilde{\mathbf{x}}_\alpha^T\boldsymbol{\beta} \right\}.$$
  - Support vector machine

$$\min_{\boldsymbol{\theta}} -\left[\sum_{i=1}^{m} y_i \log h_{\boldsymbol{\theta}}(\mathbf{x}_i) + (1-y_i)\log(1-h_{\boldsymbol{\theta}}(\mathbf{x}_i))\right],$$

where $h_{\boldsymbol{\theta}}(\mathbf{x}) = 1/(1+\exp(-\boldsymbol{\theta}^T\mathbf{x}))$.

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2, \quad s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i+b) \geq 1, i=1,\cdots,N.$$

Classification based on the objective function

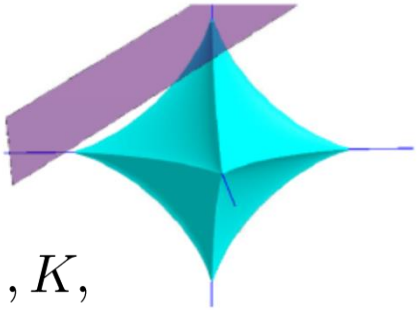
- Non-convex programs

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t. \quad \mathbf{x} \in \mathcal{C},$$

where either f is not a convex function or $\mathcal{C}$ is not a convex set.

  - Sparse/Low-rank models

$$\min_{\mathbf{x}} \; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_p^p.$$



  - Polynomial Programming:

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \quad s.t. \quad h_i(\mathbf{x}) \leq 0, i = 1, \cdots, K,$$

where f and h_i are all polynomials.

  - Fractional Programming

$$\min_{\mathbf{x}} \; f(\mathbf{x})/g(\mathbf{x}), \quad s.t. \quad \mathbf{x} \in \mathcal{C} = \{\mathbf{x} | h_i(\mathbf{x}) \leq 0, i = 1, \cdots, K\},$$

where $g(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{C}$.

Other forms of optimization

- Multi-objective optimization

$$\min_{\mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(x_1, x_2, \cdots, x_n) \\ f_2(x_1, x_2, \cdots, x_n) \\ \vdots \\ f_l(x_1, x_2, \cdots, x_n) \end{pmatrix}$$

$$s.t. \quad \mathbf{x} \in \mathcal{C}$$

where $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^l$ and $\mathcal{C} \subseteq \mathbb{R}^n$.

- Multi-level optimization

$$\begin{aligned} &\min_{\mathbf{x}} \; F(\mathbf{x}, \mathbf{y}) \\ &s.t. \; \mathbf{x} \in \mathcal{C}_1, \\ &\qquad \mathbf{y} \in \operatorname*{argmin}_{\mathbf{y} \in \mathcal{C}_2} f(\mathbf{x}, \mathbf{y}) \end{aligned}$$

Classification, AutoML

Chapter 2: Mathematics

- Background mathematics
  - Topology in $\mathbb{R}^n$
  - Analysis in $\mathbb{R}^n$
  - Linear algebra

Topology in R^n

- Open set

A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called open, if for every $\mathbf{x} \in \mathcal{C}$ there exists $\varepsilon > 0$ such that the ball $B_\varepsilon(\mathbf{x}) = \{\mathbf{y} | \|\mathbf{y} - \mathbf{x}\|_2 \leq \varepsilon\}$ is included in $\mathcal{C}$.

Examples: $\{x | a < x < b\}$, $\{\mathbf{x} | \|\mathbf{x} - \mathbf{a}\| < 1\}$, $\{\mathbf{x} | \mathbf{x} > \mathbf{0}\}$, $\mathbb{S}^n_{++} = \{\mathbf{X} | \mathbf{X} \succ \mathbf{0}\}$

- Closed set

A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called closed, if its complement $\mathcal{C}^c = \mathbb{R}^n \backslash \mathcal{C}$ is open.

Examples: $\{x | a \leq x \leq b\}$, $\{\mathbf{x} | \|\mathbf{x} - \mathbf{a}\| \leq 1\}$, $\{\mathbf{x} | \mathbf{x} \geq \mathbf{0}\}$, $\mathbb{S}^n_{+} = \{\mathbf{X} | \mathbf{X} \succeq \mathbf{0}\}$

- Bounded set

A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called bounded, if $\exists R > 0$ such that $\|\mathbf{x}\| < R$, $\forall \mathbf{x} \in \mathcal{C}$.

Examples: $\{x | a \leq x < b\}$, $\{\mathbf{x} | \|\mathbf{x} - \mathbf{a}\| < 1\}$, $\{\mathbf{x} | \mathbf{1} > \mathbf{x} \geq \mathbf{0}\}$, $\{\mathbf{X} | \mathbf{I} \succeq \mathbf{X} \succ \mathbf{0}\}$

- Compact set

A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called compact, if it is both bounded and closed.

Examples: $\{x | a \leq x \leq b\}$, $\{\mathbf{x} | \|\mathbf{x} - \mathbf{a}\| \leq 1\}$, $\{\mathbf{x} | \mathbf{1} \geq \mathbf{x} \geq \mathbf{0}\}$, $\{\mathbf{X} | \mathbf{I} \succeq \mathbf{X} \succeq \mathbf{0}\}$

Topology in R^n

- Interior

The interior of $\mathcal{C} \subseteq \mathbb{R}^n$ is defined as $\mathcal{C}^\circ = \{\mathbf{y} | \exists \varepsilon > 0 \text{ such that } B_\varepsilon(\mathbf{y}) \subset \mathcal{C}\}$.

Examples: $(\{x|a \leq x \leq b\})^\circ = \{x|a < x < b\}$, $(\{\mathbf{x}|\|\mathbf{x}-\mathbf{a}\| \leq 1\})^\circ = \{\mathbf{x}|\|\mathbf{x}-\mathbf{a}\| < 1\}$, $(\{\mathbf{x}|\mathbf{1} \geq \mathbf{x} \geq \mathbf{0}\})^\circ = \{\mathbf{x}|\mathbf{1} > \mathbf{x} > \mathbf{0}\}$, $(\{\mathbf{X}|\mathbf{I} \succeq \mathbf{X} \succeq \mathbf{0}\})^\circ = \{\mathbf{X}|\mathbf{I} \succ \mathbf{X} \succ \mathbf{0}\}$

- Closure

The closure of $\mathcal{C} \subset \mathbb{R}^n$ is defined as $\bar{\mathcal{C}} = \mathbb{R}^n \backslash (\mathbb{R}^n \backslash \mathcal{C})^\circ = ((\mathcal{C}^c)^\circ)^c$.

Examples: $\overline{\{x|a \leq x < b\}} = \{x|a \leq x \leq b\}$, $\overline{\{\mathbf{x}|\|\mathbf{x}-\mathbf{a}\| < 1\}} = \{\mathbf{x}|\|\mathbf{x}-\mathbf{a}\| \leq 1\}$, $\overline{\{\mathbf{x}|\mathbf{1} \geq \mathbf{x} > \mathbf{0}\}} = \{\mathbf{x}|\mathbf{1} \geq \mathbf{x} \geq \mathbf{0}\}$, $\overline{\{\mathbf{X}|\mathbf{I} \succ \mathbf{X} \succ \mathbf{0}\}} = \{\mathbf{X}|\mathbf{I} \succeq \mathbf{X} \succeq \mathbf{0}\}$

- Boundary

The boundary of $\mathcal{C} \subseteq \mathbb{R}^n$ is defined as $\partial\mathcal{C} = \bar{\mathcal{C}} \backslash \mathcal{C}^\circ$.

Examples: $\partial(\{x|a \leq x < b\}) = \{a, b\}$, $\partial(\{\mathbf{x}|\|\mathbf{x}-\mathbf{a}\| < 1\}) = \{\mathbf{x}|\|\mathbf{x}-\mathbf{a}\| = 1\}$, $\partial(\{\mathbf{x}|\mathbf{1} \geq \mathbf{x} > \mathbf{0}\}) = \{\mathbf{x}|\exists i \text{ such that } x_i = 0 \text{ or } 1\}$, $\partial(\{\mathbf{X}|\mathbf{I} \succ \mathbf{X} \succ \mathbf{0}\}) = \{\mathbf{X}|0 \leq \lambda_i(\mathbf{X}) \leq 1, \forall i, \text{ and } \exists j \text{ such that } \lambda_j(\mathbf{X}) = 0 \text{ or } 1\}$

Analysis in R^n

- Convergence rate

Assume $\mathbf{x}_k \to \mathbf{x}^*$. We define the sequence of errors to be

$$\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}^*.$$

We say that the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$ with rate r and rate constant C if

$$\lim_{k\to\infty} \frac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|^r} = C, \quad (C < \infty).$$

Linear: $r = 1$, $0 < C < 1$; $\boxed{\text{Q-linear}}$

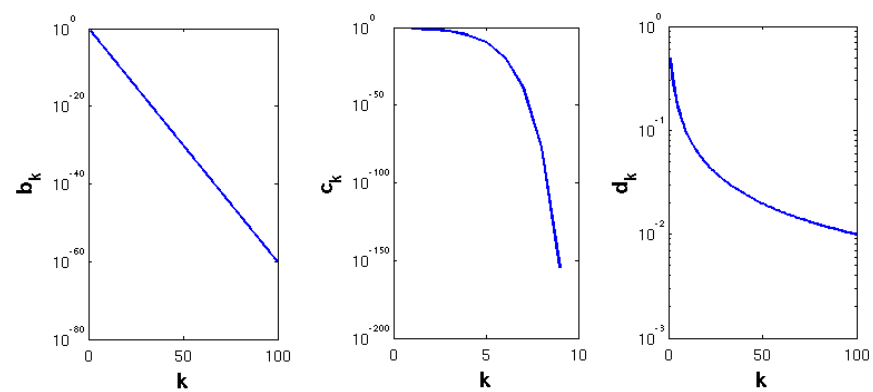
Sublinear: $r = 1$, $C = 1$;

Superlinear: $r = 1$, $C = 0$;

Quadratic: $r = 2$;

Cubic: $r = 3$. r may be non-integers

Can $r = 1$ and $C > 1$ happen?



Examples: $x_k = 10^{-k}$, $x_k = 0.99^k$, $x_k = 10^{-2^k}$, $x_{k+1} = x_k/2 + 2/x_k$ $(x_1 = 4)$,

Analysis in R^n

- Convergence rate

Estimating the order r:

$$r \approx \frac{\log \dfrac{x_{k+1} - x_k}{x_k - x_{k-1}}}{\log \dfrac{x_k - x_{k-1}}{x_{k-1} - x_{k-2}}}.$$

Assume $\mathbf{x}_k \to \mathbf{x}^*$. We say that the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$ R-linearly if

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le e_k$$

R-linear

and $\{e_k\}$ converges to 0 Q-linearly.

Remedies the issue when $\lim\limits_{k\to\infty} \dfrac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|^r}$ does not exist.

Example: $x_k = \begin{cases} 1 + 2^{-k}, & k \text{ even}, \\ 1, & k \text{ odd}. \end{cases}$

Analysis in R^n

- Derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$, and $\mathbf{x} \in (\operatorname{dom} f)^\circ$. If there exists a matrix $\mathbf{J}$ such that

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

for all choice of sequence $\{\mathbf{z}\} \subset \operatorname{dom} f$, then f is said to be differentiable at $\mathbf{x}$ and denote $Df(\mathbf{x}) = \mathbf{J}$. Let $\mathbf{z} = \mathbf{x} + t\mathbf{e}_i$ and let $t \to 0$. Then

$$\begin{aligned}
\lim_{\mathbf{z} \in \operatorname{dom} f, \mathbf{z} \neq \mathbf{x}, \mathbf{z} \to \mathbf{x}} \frac{\|f(\mathbf{z}) - f(\mathbf{x}) - \mathbf{J}(\mathbf{z} - \mathbf{x})\|_2}{\|\mathbf{z} - \mathbf{x}\|_2} &= \lim_{t \to 0} \frac{\|f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x}) - t\mathbf{J}\mathbf{e}_i\|_2}{|t|} \\
&= \lim_{t \to 0} \left\| \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} - \mathbf{J}\mathbf{e}_i \right\|_2 \\
&= \left\| \frac{\partial f(\mathbf{x})}{\partial x_i} - \mathbf{J}\mathbf{e}_i \right\|_2 .
\end{aligned}$$

Jacobian

Therefore, the i-th column of $\mathbf{J}$ is $\frac{\partial f(\mathbf{x})}{\partial x_i}$. Thus $\mathbf{J} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j} \right)$.

Analysis in R^n

- Gradient

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

$$\nabla f(\mathbf{x}) = Df(\mathbf{x})^T.$$

Its components are the partial derivatives of f:

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

The first-order approximation of f at a point $\mathbf{x} \in (\mathrm{dom}\, f)^\circ$ can be expressed as

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

Examples: $f(\mathbf{x}) = (1/2)\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^\top\mathbf{x} + r$, $f(\mathbf{X}) = \log\det\mathbf{X}$, with $\mathrm{dom}\, f = \mathbf{S}^n_{++}$

What if $\mathbf{P}$ is not symmetric?

Analysis in R^n

- Gradient

Suppose $f = f(\mathbf{X})$ is a scalar function, $\mathbf{X}$ is a vector or a matrix. Define the gradient of f w.r.t $\mathbf{X}$ as:

$$\frac{\partial f}{\partial \mathbf{X}} \triangleq \left(\frac{\partial f(\mathbf{X})}{\partial X_{ij}} \right). \tag{1}$$

Example: $f(\mathbf{X}) = \|\mathbf{X}\|_F^2$.

Analysis in R^n

- Gradient of $f(\mathbf{X}) = \log\det\mathbf{X}$, with $\operatorname{dom} f = \mathbb{S}^n_{++}$

Let $\mathbf{Z} \in \mathbb{S}^n_{++}$ be close to $\mathbf{X}$, $\Delta\mathbf{X} = \mathbf{Z} - \mathbf{X}$, and λ_i be the ith eigenvalue of $\mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2}$. We have

$$
\begin{aligned}
\log\det\mathbf{Z} &= \log\det\left(\mathbf{X} + \Delta\mathbf{X}\right) \\
&= \log\det\left(\mathbf{X}^{1/2}(\mathbf{I} + \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}\right) \\
&= \log\det\mathbf{X} + \log\det\left(\mathbf{I} + \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2}\right) \\
&= \log\det\mathbf{X} + \sum_{i=1}^{n}\log\left(1 + \lambda_i\right) \approx \log\det\mathbf{X} + \sum_{i=1}^{n}\lambda_i \\
&= \log\det\mathbf{X} + \operatorname{tr}\left(\mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2}\right) \\
&= \log\det\mathbf{X} + \operatorname{tr}\left(\mathbf{X}^{-1}\Delta\mathbf{X}\right) \\
&= \log\det\mathbf{X} + \operatorname{tr}\left(\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})\right).
\end{aligned}
$$

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

Analysis in R^n

- Chain rule

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

$$Dh(\mathbf{x}) = Dg(f(\mathbf{x}))Df(\mathbf{x}).$$

You may check the dimensions of the matrices in order to determine the order of multiplication.

Examples: suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(\mathbf{x}) = g(f(\mathbf{x}))$, then

$$\nabla h(\mathbf{x}) = g'(f(\mathbf{x}))\nabla f(\mathbf{x}).$$

Composition with affine function: Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, $\mathbf{A} \in \mathbb{R}^{n\times p}$, and $\mathbf{b} \in \mathbb{R}^n$. Define $g : \mathbb{R}^p \to \mathbb{R}^m$ as $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$, with $\operatorname{dom} g = \{\mathbf{x} | \mathbf{A}\mathbf{x} + \mathbf{b} \in \operatorname{dom} f\}$. The derivative of g is $Dg(\mathbf{x}) = Df(\mathbf{A}\mathbf{x} + \mathbf{b})\mathbf{A}$.

When f is real-valued (*i.e.*, $m = 1$),

$$\nabla g(\mathbf{x}) = \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

Analysis in R^n

- Gradient of $\quad f(\mathbf{x}) = \log \sum_{i=1}^{m} \exp\left(\mathbf{a}_i^\top \mathbf{x} + b_i\right)$

It is the composition of the affine function $\mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rows $\mathbf{a}_1^T, ..., \mathbf{a}_m^T$, and the function $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(\mathbf{y}) = \log\left(\sum_{i=1}^m \exp y_i\right)$. Simple differentiation shows that

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

so by the composition formula we have

$$\nabla f(\mathbf{x}) = \frac{1}{\mathbf{1}^T \mathbf{z}} \mathbf{A}^T \mathbf{z}$$

where $z_i = \exp\left(\mathbf{a}_i^T \mathbf{x} + b_i\right), i = 1, \ldots, m.$

Analysis in R^n

- Second derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in (\operatorname{dom} f)^\circ$. Then $Df(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$. Suppose $Df(\mathbf{x})$ is also differentiable, then $D^2 f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and

$$D^2 f(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \frac{\partial (Df(\mathbf{x}))^T}{\partial \mathbf{x}^T} = \left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\right) = \mathbf{H}.$$

Hessian

The second-order approximation of f at a point $\mathbf{x} \in (\operatorname{dom} f)^\circ$ can be expressed as

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

Examples: $f(\mathbf{x}) = (1/2)\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^\top\mathbf{x} + r$, $f(\mathbf{X}) = \log\det \mathbf{X}$, with $\operatorname{dom} f = \mathbf{S}^n_{++}$

Analysis in R^n

- Second derivative of $f(\mathbf{X}) = \log\det\mathbf{X}$, with $\operatorname{dom} f = \mathbb{S}^n_{++}$

For $\mathbf{Z} \in \mathbb{S}^n_{++}$ near $\mathbf{X} \in \mathbb{S}^n_{++}$, and $\Delta\mathbf{X} = \mathbf{Z} - \mathbf{X}$, we have

$$\begin{aligned}
\mathbf{Z}^{-1} = (\mathbf{X} + \Delta\mathbf{X})^{-1} &= \left(\mathbf{X}^{1/2}(\mathbf{I} + \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}\right)^{-1} \\
&= \mathbf{X}^{-1/2}(\mathbf{I} + \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})^{-1}\mathbf{X}^{-1/2} \\
&\approx \mathbf{X}^{-1/2}(\mathbf{I} - \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})\mathbf{X}^{-1/2} \\
&= \mathbf{X}^{-1} - \mathbf{X}^{-1}\Delta\mathbf{X}\mathbf{X}^{-1}.
\end{aligned}$$

The Hessian of f at $\mathbf{X}$ is a fourth-order tensor $\mathcal{T}$. The above shows that $\mathcal{T}(\Delta\mathbf{X}) = -\mathbf{X}^{-1}\Delta\mathbf{X}\mathbf{X}^{-1}$. The (i,j,k,l)-th entry of $\mathcal{T}$ is

$$\begin{aligned}
\mathcal{T}_{ijkl} = \mathbf{e}_i^T[-\mathbf{X}^{-1}(\mathbf{e}_k\mathbf{e}_l^T)\mathbf{X}^{-1}]\mathbf{e}_j &= -(\mathbf{e}_i^T\mathbf{X}^{-1}\mathbf{e}_k)(\mathbf{e}_l^T\mathbf{X}^{-1}\mathbf{e}_j) \\
&= -(\mathbf{X}^{-1})_{ik}(\mathbf{X}^{-1})_{lj}.
\end{aligned}$$

The second-order approximation of f near $\mathbf{X}$ is:

$$f(\mathbf{Z}) \approx f(\mathbf{X}) + \operatorname{tr}(\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})) - (1/2)\operatorname{tr}(\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})).$$

Analysis in R^n

- Chain rules for second derivative

Composition with scalar function. Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(\mathbf{x}) = g(f(\mathbf{x}))$. Simply working out the partial derivatives yields

$$\nabla^2 h(\mathbf{x}) = g'(f(\mathbf{x}))\nabla^2 f(\mathbf{x}) + g''(f(\mathbf{x}))\nabla f(\mathbf{x})\nabla f(\mathbf{x})^T.$$

Composition with affine function. Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times m}$, and $\mathbf{b} \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then we have

$$\nabla^2 g(\mathbf{x}) = \mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{x} + \mathbf{b})\mathbf{A}.$$

Analysis in R^n

- Second derivative of $f(\mathbf{x}) = \log \sum_{i=1}^{m} \exp\left(\mathbf{a}_i^\top \mathbf{x} + b_i\right)$

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

$$\nabla^2 g(\mathbf{y}) = \operatorname{diag}\left(\nabla g(\mathbf{y})\right) - \nabla g(\mathbf{y}) \nabla g(\mathbf{y})^T.$$

By the composition formula we have

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top \left(\frac{1}{\mathbf{1}^T \mathbf{z}} \operatorname{diag}\left(\mathbf{z}\right) - \frac{1}{(\mathbf{1}^T \mathbf{z})^2} \mathbf{z}\mathbf{z}^T \right) \mathbf{A},$$

where $z_i = \exp(\mathbf{a}_i^\top \mathbf{x} + b_i), i = 1, ..., m.$

Linear Algebra

- Symmetric eigenvalue decomposition (EVD)

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

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top, \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is *orthogonal*, *i.e.*, satisfies $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. The (real) numbers λ_i are the *eigenvalues* of $\mathbf{A}$, and are the roots of the *characteristic polynomial* $\det(s\mathbf{I} - \mathbf{A})$. The columns of $\mathbf{Q}$ form an orthonormal set of *eigenvectors* of $\mathbf{A}$. The factorization (1) is called the *spectral decomposition* or (symmetric) *eigenvalue decomposition* of $\mathbf{A}$.

We order the eigenvalues as $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_n$. We use the notation $\lambda_i(\mathbf{A})$ to refer to the ith largest eigenvalue of $\mathbf{A} \in \mathbb{S}^n$. We usually write the largest or maximum eigenvalue as $\lambda_1(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$, and the least or minimum eigenvalue as $\lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$.

$$\det \mathbf{A} = \prod_{i=1}^n \lambda_i, \ \mathrm{tr}\, \mathbf{A} = \sum_{i=1}^n \lambda_i, \ \|\mathbf{A}\|_2 = \max\{\lambda_1, -\lambda_n\}, \ \|\mathbf{A}\|_F = \left(\sum_{i=1}^n \lambda_i^2\right)^{1/2}.$$

Linear Algebra

- Definiteness

$$\lambda_{\max}(\mathbf{A}) = \sup_{x \neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}, \quad \lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{x} \neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}.$$

A matrix $\mathbf{A} \in \mathbb{S}^n$ is called *positive definite* if for all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$. We denote this as $\mathbf{A} \succ \mathbf{0}$. By the inequality above, we see that $\mathbf{A} \succ \mathbf{0}$ if and only all its eigenvalues are positive, *i.e.*, $\lambda_{\min}(\mathbf{A}) > 0$. If $-\mathbf{A}$ is positive definite, we say $\mathbf{A}$ is *negative definite*, which we write as $\mathbf{A} \prec \mathbf{0}$. We use $\mathbb{S}^n_{++}$ to denote the set of positive definite matrices in $\mathbb{S}^n$.

If $\mathbf{A}$ satisfies $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$, we say that $\mathbf{A}$ is *positive semidefinite* or *nonnegative definite*. If $-\mathbf{A}$ is in nonnegative definite, *i.e.*, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq 0$ for all $\mathbf{x}$, we say that $\mathbf{A}$ is *negative semidefinite* or *nonpositive definite*. We use $\mathbb{S}^n_{+}$ to denote the set of nonnegative definite matrices in $\mathbb{S}^n$.

For $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$, we use $\mathbf{A} \prec \mathbf{B}$ to mean $\mathbf{B} - \mathbf{A} \succ \mathbf{0}$, and so on.

Linear Algebra

- Symmetric squareroot

Let $\mathbf{A} \in \mathbb{S}_+^n$, with eigenvalue decomposition $\mathbf{A} = \mathbf{Q}\,\mathrm{diag}(\lambda_1, ..., \lambda_n)\mathbf{Q}^\top$. We define the (symmetric) squareroot of $\mathbf{A}$ as

$$\mathbf{A}^{1/2} = \mathbf{Q}\,\mathrm{diag}(\lambda_1^{1/2}, ..., \lambda_n^{1/2})\mathbf{Q}^\top.$$

The squareroot $\mathbf{A}^{1/2}$ is the unique symmetric positive semidefinite solution of the equation $\mathbf{X}^2 = \mathbf{A}$.

Linear Algebra

- Singular value decomposition (SVD)

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

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{1}$$

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

$$\sigma_{\max}(\mathbf{A}) = \sup_{\mathbf{x},\mathbf{y} \neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{y}}{\|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2} = \sup_{\mathbf{y} \neq 0} \frac{\|\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{y}\|_2}.$$

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

The factorization (1) is called the *singular value decomposition* (SVD) of $\mathbf{A}$. The columns of $\mathbf{U}$ are called *left singular vectors* of $\mathbf{A}$, the columns of $\mathbf{V}$ are *right singular vectors*, and the numbers σ_i are the *singular values*. The singular value decomposition can be written

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Linear Algebra

- Norms
  - Inner product

The *standard inner product* on $\mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

In particular, the inner product between matrices is

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{tr}(\mathbf{X}^T \mathbf{Y}).$$

Linear Algebra

- Norms
  - Norms and distances

A function $f : \mathbb{R}^n \to \mathbb{R}$ with $\operatorname{dom} f = \mathbb{R}^n$ is called a *norm* if

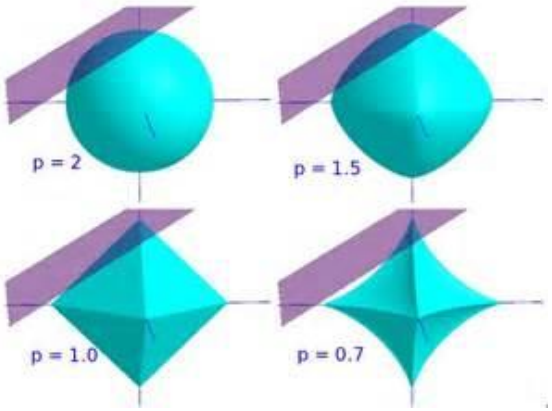
- f is nonnegative: $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$;
- f is definite: $f(\mathbf{x}) = 0$ only if $\mathbf{x} = \mathbf{0}$;
- f is homogeneous: $f(t\mathbf{x}) = |t| f(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$;
- f satisfies the triangle inequality: $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

We use the notation $f(\mathbf{x}) = \|\mathbf{x}\|$.

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

Examples: Euclidean norm, ℓ_p-norm, Mahalanobis norm, Frobenius norm, matrix 1-norm, matrix ∞-norm, matrix 2-norm

Linear Algebra



- Norms
  - Unit balls

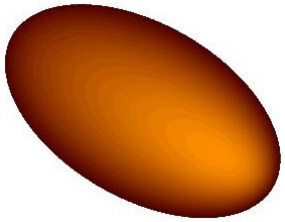
$$\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n | \, \|\mathbf{x}\| \leq 1\},$$

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.*, $\mathbf{x} \in \mathcal{B}$ if and only if $-\mathbf{x} \in \mathcal{B}$;
- $\mathcal{B}$ is convex;
- $\mathcal{B}$ is closed, bounded, and has nonempty interior.

Conversely, if $C \subseteq \mathbb{R}^n$ is any set satisfying these three conditions, then it is the unit ball of a norm, which is given by

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



Examples: unit balls of Euclidean norm, ℓ_p-norm, Mahalanobis norm

Linear Algebra

- Norms
  - Operator/induced norms

Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. We define the *operator norm* of $\mathbf{X} \in \mathbb{R}^{m \times n}$, induced by the norms $\|\cdot\|_a$ and $\|\cdot\|_b$, as

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

Examples: $\|\mathbf{X}\|_2$, $\|\mathbf{X}\|_\infty$, $\|\mathbf{X}\|_1$

Linear Algebra

- Norms
  - Nuclear norm

The nuclear norm $\|\mathbf{A}\|_*$ of a matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A}).$$

It is usually used for approximating the rank of a matrix.

$$||\mathbf{X}||_* = \max_{\mathbf{U}^T\mathbf{U}=\mathbf{I},\mathbf{V}^T\mathbf{V}=\mathbf{I}} \mathrm{tr}(\mathbf{U}^T\mathbf{X}\mathbf{V}).$$

Proposition 1. *Equivalence between matrix norms:*

1. $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \|\mathbf{A}\|_*.$
2. $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1\|\mathbf{A}\|_\infty}.$

Linear Algebra

- Norms
  - (p,q)-norm

The (p, q)-norm of matrices is widely used in sparse representation:

$$\|\mathbf{A}\|_{p,q} = \left(\sum_{i=1}^{n} \|\mathbf{A}_i\|_p^q \right)^{\frac{1}{q}},$$

where $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_n)$ and $p, q \geq 1$.
Note that we use the same notation as the induced norm.

Linear Algebra

- Norms
  - Dual norm

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

$$\|\mathbf{z}\|^* = \sup\{\mathbf{z}^\top \mathbf{x} | \, \|\mathbf{x}\| \leq 1\}.$$

$$\forall \mathbf{x}, \exists \mathbf{x}^*, \text{ such that } \langle \mathbf{x}, \mathbf{x}^* \rangle = \|\mathbf{x}\| \|\mathbf{x}^*\|^*.$$

Generalized Cauchy-Schwartz inequality:

$$\mathbf{z}^\top \mathbf{x} \leq \|\mathbf{x}\| \, \|\mathbf{z}\|^* .$$

$$\|\mathbf{x}\|^{**} = \|\mathbf{x}\|$$

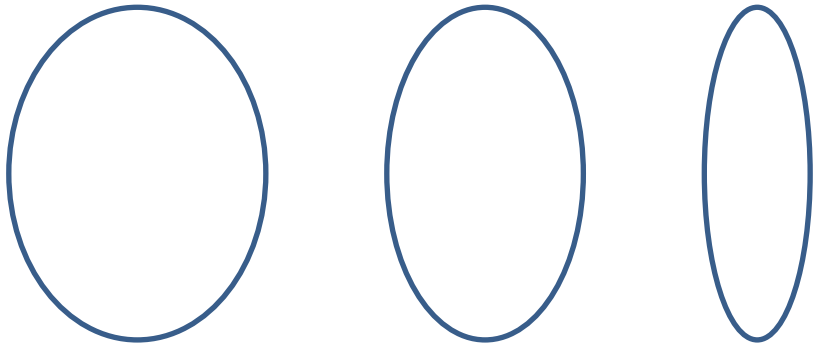
Examples: Euclidean norm, ℓ_p norm, matrix 2-norm.

Linear Algebra

- Condition number

The *condition number* of a nonsingular $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $\operatorname{cond}(\mathbf{A})$ or $\kappa(\mathbf{A})$, is defined as

$$\operatorname{cond}(\mathbf{A}) = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2 = \sigma_{\max}(\mathbf{A})/\sigma_{\min}(\mathbf{A}).$$



Linear Algebra

- Pseudo-inverse

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

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

Linear Algebra

- Adjoint operator

Given a linear operator $\mathcal{A} : \mathbb{R}^m \to \mathbb{R}^n$, its *adjoint operator* is defined as the linear operator $\mathcal{A}^*$ that satisfies:

$$\langle \mathcal{A}^*(\mathbf{x}), \mathbf{y} \rangle = \langle \mathbf{x}, \mathcal{A}(\mathbf{y}) \rangle, \quad \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m.$$

Examples: $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, $\mathcal{A}(\mathbf{X}) = \mathbf{x}$ is the linear operator that extracts entries from $\mathbf{X}$.

Linear Algebra

- von Neumann trace theorem

Theorem 1. *Suppose* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$. *Then*

$$|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).$$

In particular, when both $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ p.s.d. matrices, the inequality becomes

$$\sum_{i=1}^{n} \sigma_i(\mathbf{A})\sigma_{n-i+1}(\mathbf{B}) \leq \mathrm{tr}(\mathbf{AB}) \leq \sum_{i=1}^{n} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).$$

Examples: Prove the openess of $\mathbb{S}^n_{++} = \{\mathbf{X} | \mathbf{X} \succ \mathbf{0}\}$.

$$||\mathbf{X}||_* = \max_{\mathbf{U}^T\mathbf{U}=\mathbf{I}, \mathbf{V}^T\mathbf{V}=\mathbf{I}} \mathrm{tr}(\mathbf{U}^T\mathbf{X}\mathbf{V}).$$