Lecture 1

- About the course
- History of optimization
- Overview of optimization

About myself

- IEEE/IAPR Fellow
- NSF Distinguished Young Researcher
- Research Areas: Machine Learning, Computer Vision
- Associate Editor of IEEE T. PAMI and IJCV
- Area Chair: CVPR 2014/2016, ICCV 2015, NIPS 2015

Welcome to apply to me!

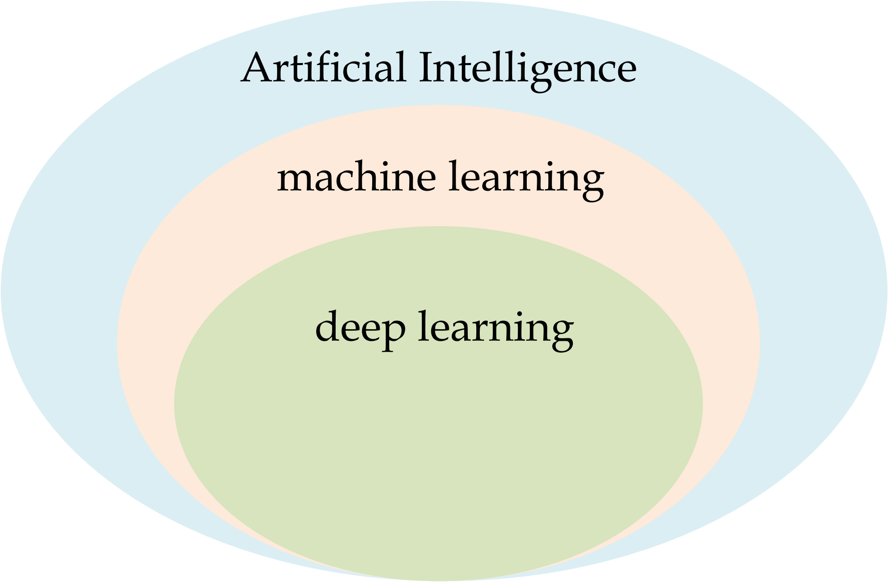
- Good math
- Good coding
- Good English
- **Good attitude**

What is this course?

- Optimization in machine learning
- Convex analysis

Machine Learning is one of the most important branches of Artificial Intelligence. It studies how to empower computers to automatically improve their own abilities by utilizing data.

- Supervised learning
- Unsupervised learning
- Weakly supervised learning
  - Reinforcement learning
  - Semi-supervised learning
  - Multi-instance learning
  - etc.



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

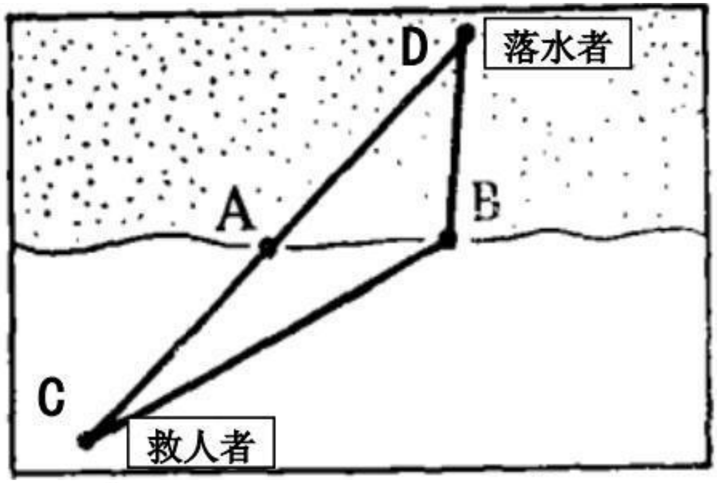
$$\begin{aligned} &\min \operatorname{tr}(\mathbf{X}^T\mathbf{Y}) - \log\det(\mathbf{X}) + \lambda\|\mathbf{X}\|_1, \\ &s.t. \quad \mathbf{X} \in \mathbb{R}^{n\times n}, \mathbf{X}^\top = \mathbf{X}, \mathbf{X} \succeq \mathbf{0}. \end{aligned}$$

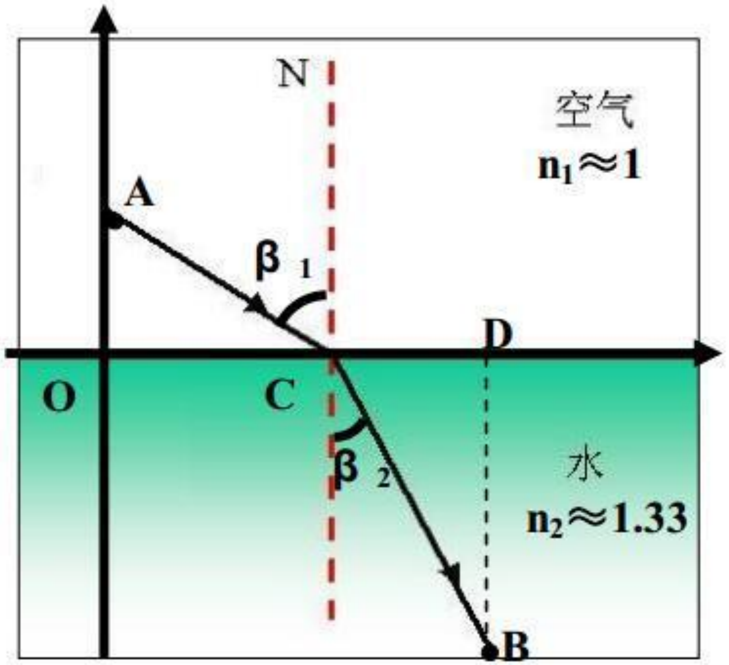
What does optimization concern?

- Pattern Recognition: Speed, Error Rate
- Numerical Algebra: Speed, Numerical Stability
- Optimization: Convergence, Convergence Rate

Why is optimization important?

- Optimization is everywhere!
  - Industrial applications
  - Engineering design
  - Financial markets
  - Daily life
- Nature is also optimizing!
  - A principle as important as symmetry and conservation
  - Principle of least action
  - Minimum optical path (Fermat's principle)





Why is optimization important?

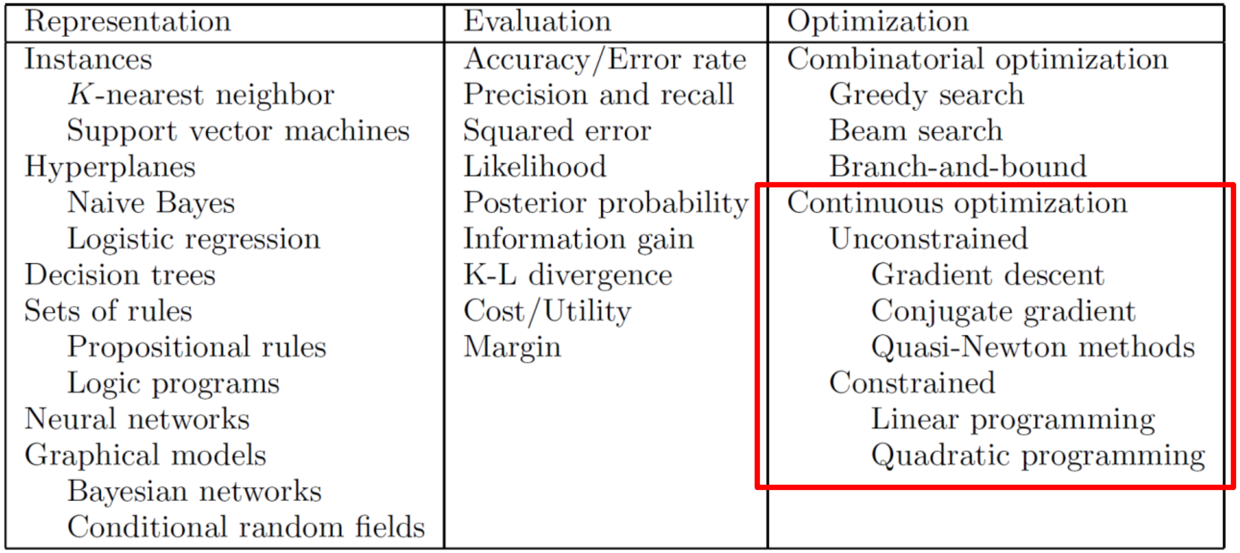
- Pedro Domingos (AAAI Fellow, Prof. @ UW):

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.

Is optimization difficult?

- No!

南京大学数学系张高飞：代数几何>复分析、调和分析、微分方程>几何>动力系统>组合数学>统计>计算数学

最后这一条是专门针对那些悲情人物的。他们连小学的数学也没学好。不要说把上千个数加起来，就是把两个数加起来，对他们来说都是件很吃力的事。然而这一切丝毫没有削弱他们对数学的一片痴情。他们日日夜夜泡在图书馆里。他们翻阅了所有的数学文献，却从未找到一本能读懂的。 但他们仍坚持不懈，为的就是找到一个适合自己的专业。他们的行为感动了上帝。上世纪的某一天，上帝为他们创造了一台机器帮他们计算。这就是计算机。借助计算机，他们可以很快的进行加减乘除的运算。这就是计算数学。

APG:

$$\begin{aligned} \mathbf{y}_k &= \mathbf{x}_k + \frac{t_{k-1}-1}{t_k}(\mathbf{x}_k - \mathbf{x}_{k-1}), \\ \mathbf{x}_{k+1} &= \operatorname{prox}_{\alpha_k g}(\mathbf{y}_k - \alpha_k \nabla f(\mathbf{y}_k)), \\ t_{k+1} &= \frac{\sqrt{4(t_k)^2+1}+1}{2}, \end{aligned}$$

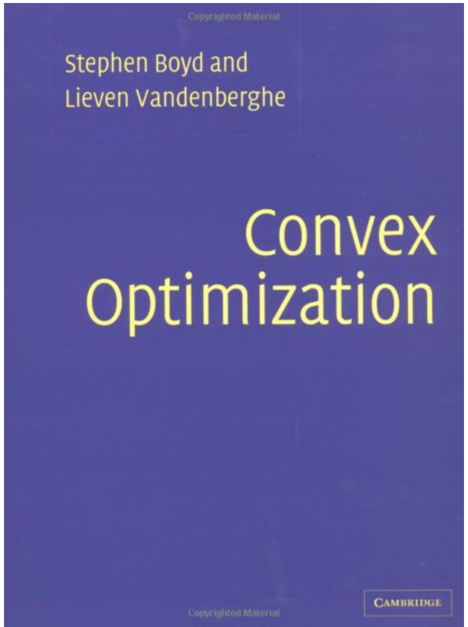
- Yes!
  - constructive algorithm, linear algebra, calculus, functional analysis, differential geometry, probability, … to **prove**

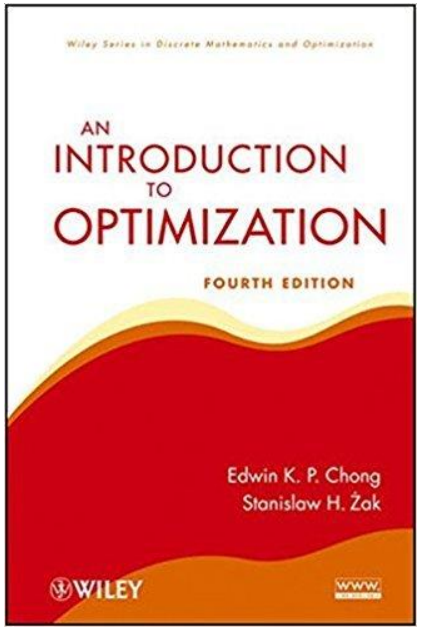
simple but deep

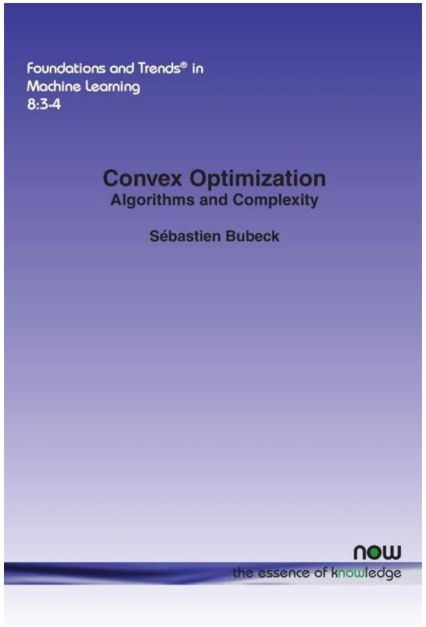
张高飞，如何选择数学方向，http://blog.sina.com.cn/s/blog_649886d60102drfd.html

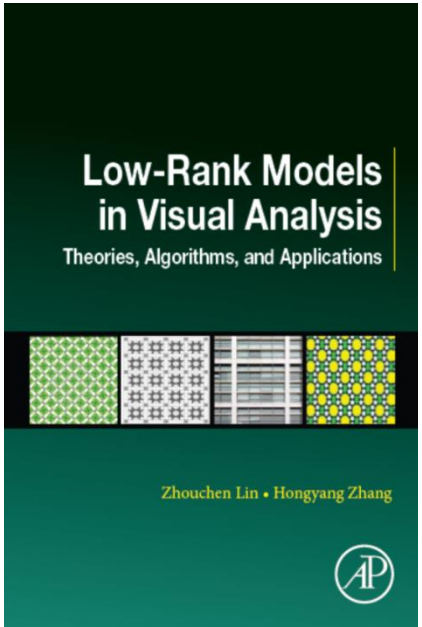
Reference books

- 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



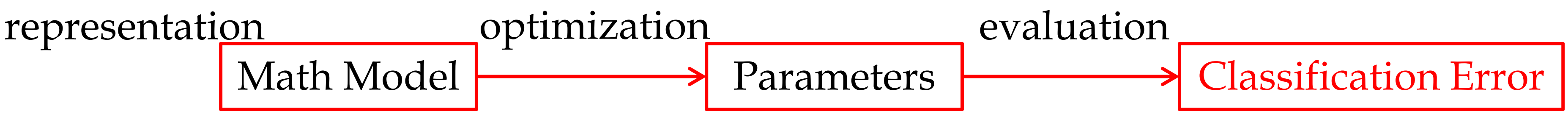





Low-Rank Models
in Visual Analysis
Theories, Algorithms, and Applications
Zhouchen Lin • Hongyang Zhang

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:

Do NOT print the whole lecture notes!
Print weekly instead!

– Constrained optimization

— Problem reformulation techniques
— Distributed/Asynchronous opt.?

History of optimization

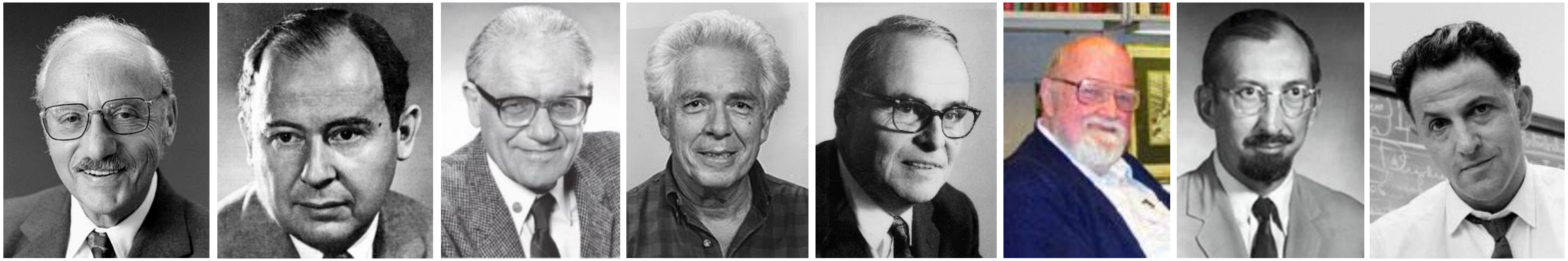
- Optimization is as old as human history.
- The study of optimization problems is as old as science itself.
- Before 1900
  - Euclid: rectangle vs. square
  - Heron: mirror
  - Kepler: marriage problem
  - Snell: law of refraction
  - Fermat: Fermat's principle
  - Newton: shape of minimal resistance
  - Bernoulli: curve of quickest descent
  - Maupertuis: principle of least action

History of optimization

- Before 1900
  - Monge: transportation problem
  - Gauss: method of least squares
  - Ricardo: law of diminishing return
  - Cauchy: iterative method for linear systems (gradient method and steepest descent)

History of optimization

- In twentieth century before 1960
  - (1906) Jensen: convexity & Jensen's inequality
  - (1917) Hancock: first book on optimization "Theory of Minima and Maxima"
  - (1930) Menger: Messenger’s problem
  - (1939) Kantorovich: linear programming
  - (1947) Danzig: simplex method
  - (1947) Neumann: duality
  - (1951) Karush, Kuhn, Tucker: KKT conditions
  - (1954) Ford and Fulkerson: network flow
  - (1957) Bellman: dynamic programming

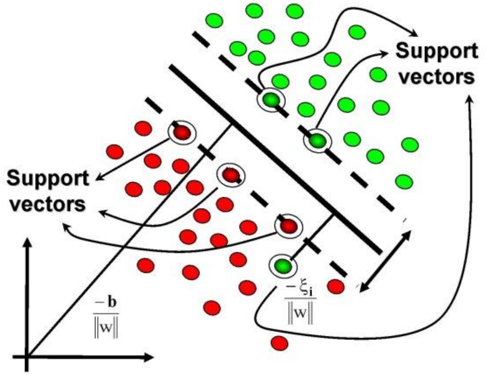


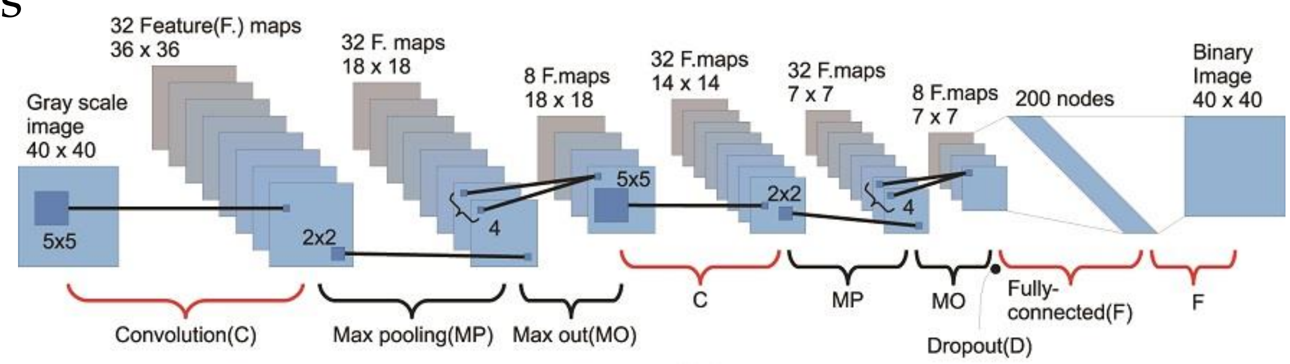
History of optimization

- 1960 - 1990
  - (1960) Zoutendijk: methods of feasible directions
  - (1963) Wilson: sequential quadratic programming
  - (1979) Khachiyan: ellipsoid method
  - (1984) Karmarkar: polynomial time algorithm for LP
  - (1980s) heuristic algorithms
  - (1990s) semidefinite programming

Recent advances (1990-)

- Major theories are done
- Revive and refine of existing techniques
  - Spiral Ascent
- Application driven
  - Support Vector Machines → Quadratic Programing
  - Deep Learning → Stochastic Gradient Descent
  - Big Data → Randomized/Decentralized/Asynchronous Algorithms



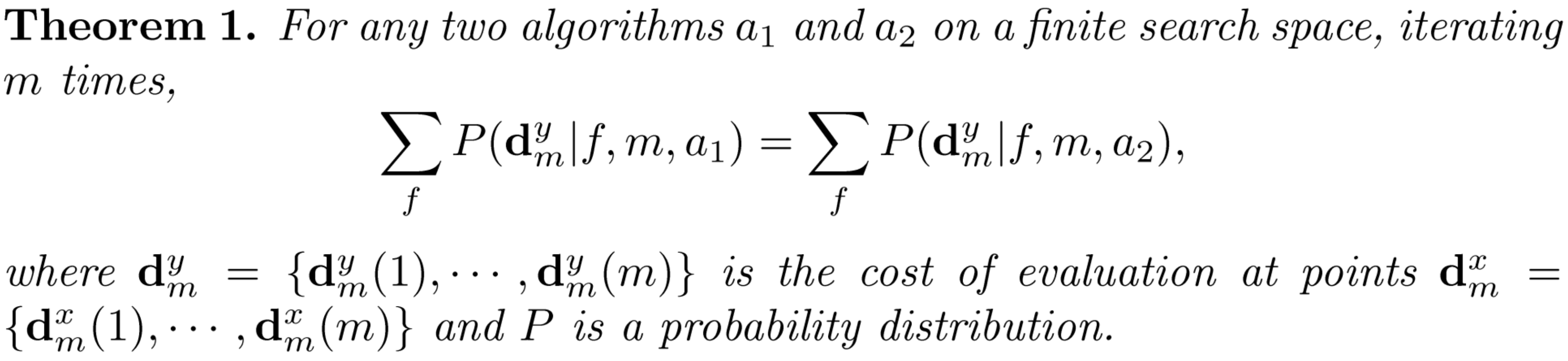


Recent advances (1990-)

- Smooth → Nonsmooth
- Convex → Nonconvex
- One/Two Blocks → Multiple Blocks
- Deterministic → Stochastic
- Synchronous → Asynchronous
- Centralized → Decentralized
- Improved Convergence & Convergence Rate

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 $\mathbb{R}^n$.

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

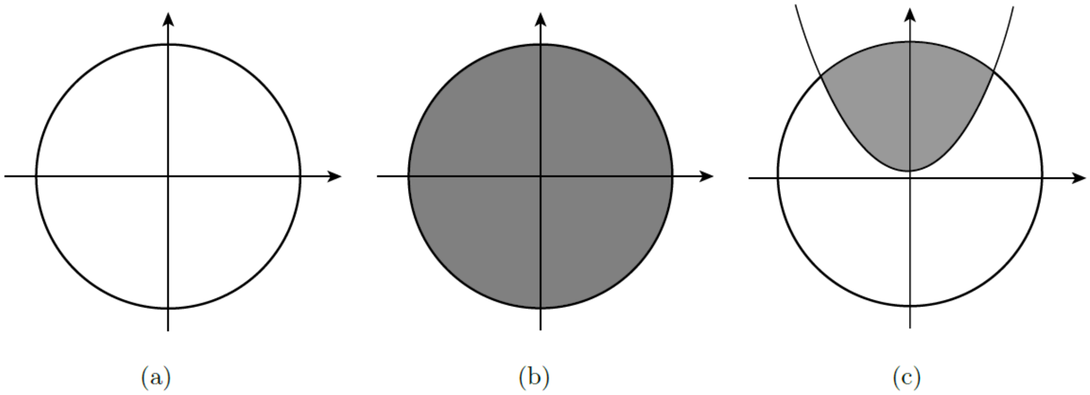
$$s.t. \; (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 = 1\}.$$

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

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

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

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

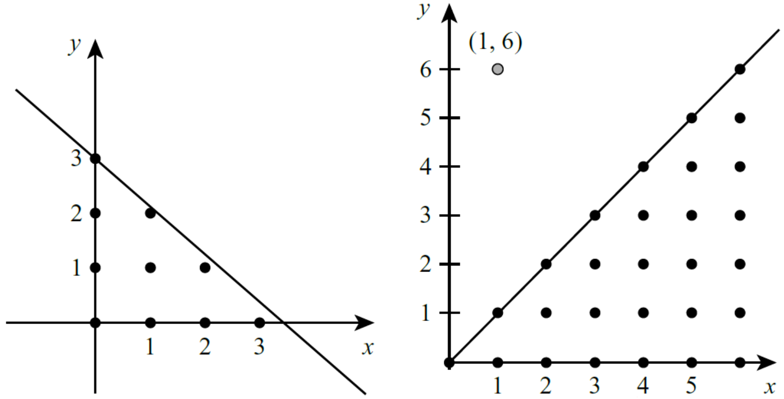


Discrete opt. problems

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

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

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



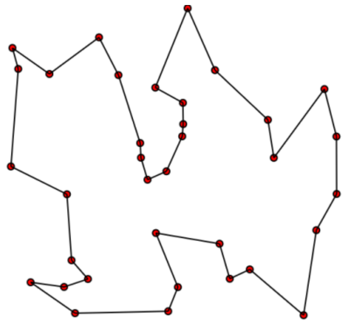
(a)

(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.

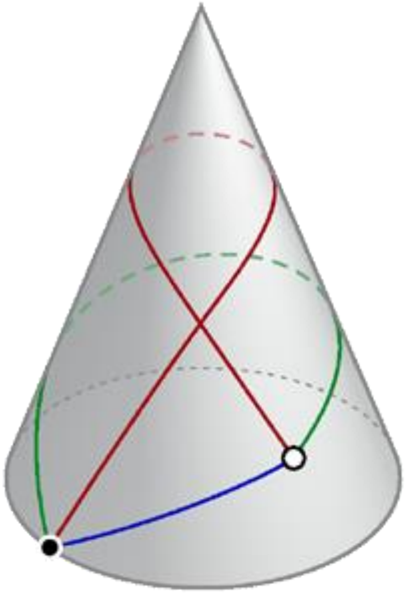


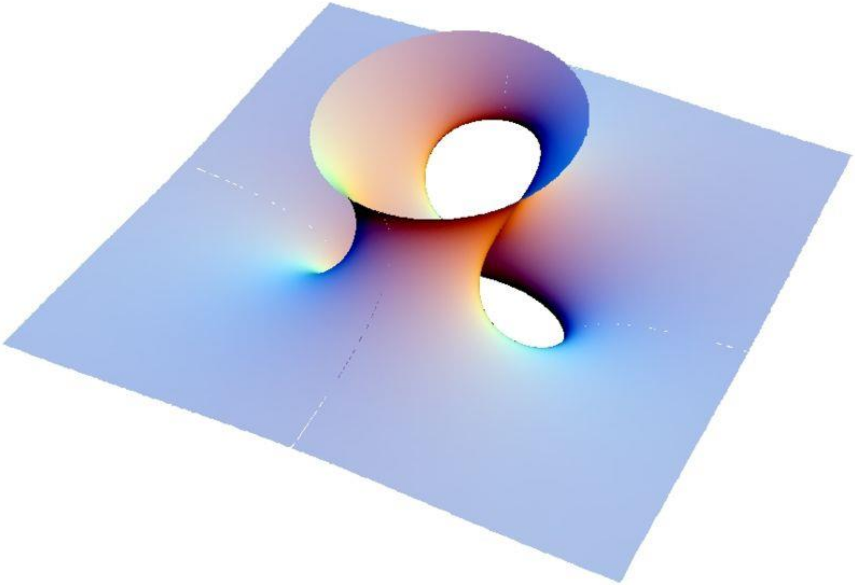
Variational opt. problems

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: $g_j(\mathbf{x}) \leq 0, j = 1, \cdots, J.$

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

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

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

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

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

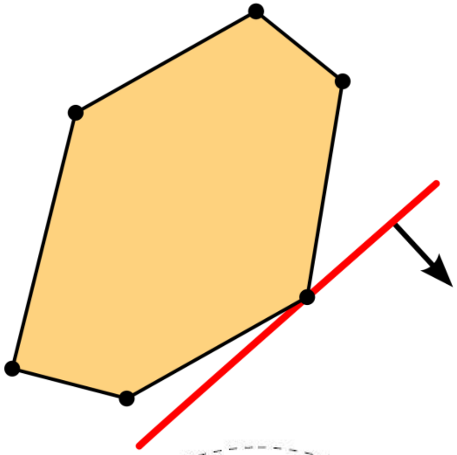
$$s.t.\ (x,y) \in \mathcal{C} = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1, y - x^2 \geq 0\}.$$

Classification based on the objective function

- Convex 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.

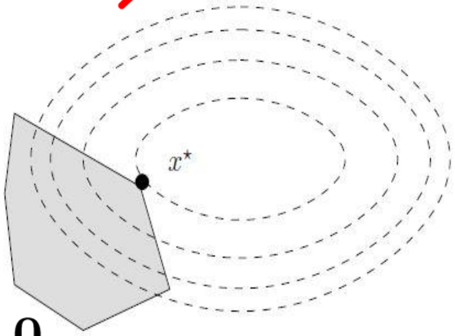


  - Linear programs

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

  - Quadratic programs

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



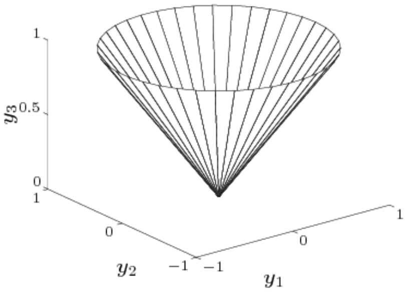
  - Semi-definite programs

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

  - 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

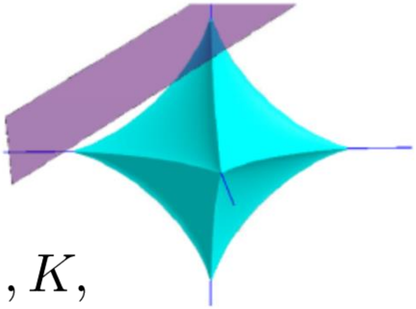
- Non-convex programs

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

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

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

Contact information

- TA: 陈琪, 1701110040@pku.edu.cn, 王一飞 (please send your homework only to 陈琪)
- zlin@pku.edu.cn
- http://www.cis.pku.edu.cn/faculty/vision/zlin/zlin.htm

Homework (1)

1. There is at least one photo of the people in “History of optimization” which is incorrect (not accounting for the ages and poses). Find them, provide your correct photos and give justifications.
2. The no-free-lunch (NFL) theorem is actually quite general. Give more instances that the NFL theorem applies.

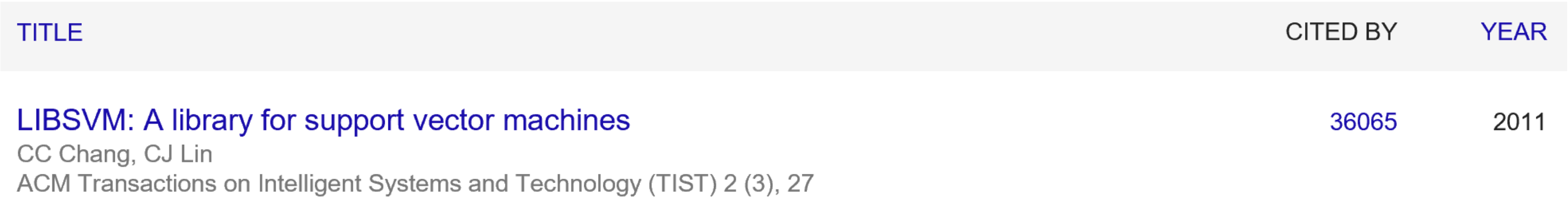
Rethink whether to take this course if your math is not good enough

Homework requirements

1. You MUST use Latex to typeset all your homework, except the computer codes.
2. Submit your homework to QI CHEN (1701110040@pku.edu.cn) by the next class.

Course projects

- History of optimization after 1960
  - Divide by years or subjects
- Program for solving linear SVM
  - LibSVM is the baseline = 90 points



TITLE	CITED BY	YEAR
LIBSVM: A library for support vector machines CC Chang, CJ Lin ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3), 27	36065	2011

I will add more problems and give a reasonable way to evaluate problem 1.

Lecture 2

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

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

- 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

- Sequences

A sequence in $\mathbb{R}^n$ is a set of vectors in $\mathbb{R}^n$ indexed by positive integers, denoted as $\{\mathbf{x}_k\}_{k=1}^{\infty}$ of simply $\{\mathbf{x}_k\}$. Sometimes the index may start from 0 or even negative integers for convenience.

Examples: $x_k = 10^k$, $x_k = 10^{-k}$, $x_k = (-1)^k k$

A sequence $\{x_k\}$ in $\mathbb{R}$ is *increasing* if $x_k < x_{k+1}$ for all k. If $x_k \leq x_{k+1}$, then we say that the sequence is *nondecreasing*. Similarly, we can define *decreasing* and *nonincreasing* sequences. Nonincreasing or nondecreasing sequences are called *monotone* sequences.

Examples: $x_k = 10^k$, $x_k = 10^{-k}$, $x_k = \lfloor k/4 \rfloor$, $x_k = (-1)^k k$

A sequence $\{\mathbf{x}_k\}$ is *bounded* if $\exists R > 0$ such that $\|\mathbf{x}_k\| < R$, $\forall k$.

Examples: $x_k = 10^k$, $x_k = 10^{-k}$, $x_k = \lfloor k/4 \rfloor$, $x_k = (-1)^k k$

Analysis in R^n

- Sequences

A sequence $\{x_k\}$ in $\mathbb{R}$ is *upper bounded* (or *bounded above*) if $\exists u$ such that $x_k \leq u$, $\forall k$. Lower boundedness (or *bounded below*) is defined similarly.

Examples: $x_k = 10^k$, $x_k = 10^{-k}$, $x_k = \lfloor k/4 \rfloor$, $x_k = (-1)^k k$

- Supremum and infimum

Suppose $\mathcal{C} \subseteq \mathbb{R}$. A number a is an *upper bound* on $\mathcal{C}$ if for each $x \in \mathcal{C}$, $x \leq a$. The set of upper bounds on a set C is either empty (in which case we say $\mathcal{C}$ is unbounded above), all of $\mathbb{R}$ (only when $\mathcal{C} = \emptyset$), or a closed infinite interval $[b, \infty)$. The number b is called the *least upper bound* or *supremum* of the set $\mathcal{C}$, and is denoted $\sup \mathcal{C}$. We take $\sup \emptyset = -\infty$, and $\sup \mathcal{C} = \infty$ if $\mathcal{C}$ is unbounded above.
We define lower bound, and infimum, in a similar way.

Examples: $x_k = 2 - 10^{-k}$, $x_k = 1 + 10^{-k}$, $x_k = 1 + (-1)^k \times 10^{-k}$

Analysis in R^n

- Limit

A sequence $\{x_k\}$ in $\mathbb{R}^n$ is *convergent* (or *has a limit*) if $\exists \mathbf{x}^*$, such that for all $\varepsilon > 0$, $\exists K$ such that $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \varepsilon$, $\forall k > K$.

Examples: $x_k = 2 - 10^{-k}$, $x_k = 1 + 10^{-k}$, $x_k = 1 + (-1)^k \times 10^{-k}$

- Accumulation point

Given a sequence $\{\mathbf{x}_k\} \subseteq \mathbb{R}^n$, we call $\mathbf{x}^*$ an accumulation point of $\{\mathbf{x}_k\}$ if for any $\varepsilon > 0$, there exists k_j such that $\|\mathbf{x}_{k_j} - \mathbf{x}^*\| < \varepsilon$.

Examples: $x_k = (-1)^k(1 + 10^{-k})$, $x_k = (k \mod 4) + (-1)^k/k$

Analysis in R^n

- Bolzano-Weierstrass theorem

Theorem 1. *Any bounded sequence in $\mathbb{R}^n$ contains a convergent subsequence.*

Proof. Suppose that a sequence $\{\mathbf{x}_k\}$ is in a hyper-cube $\mathcal{H} = \{\mathbf{x}|l_i \leq x_i \leq u_i, i = 1, \cdots, n\}$. Each time we divide $\mathcal{H}$ in half, then at least one of the halves has infinite number of points in $\{\mathbf{x}_k\}$, denote it as $\mathcal{H}_j$. Then the diameter of $\mathcal{H}_j$ approaches 0 and $\mathbf{x}^* = \bigcap\limits_{j=1}^{\infty} \mathcal{H}_j$ is an accumulation point. Picking any of the points in $\mathcal{H}_{j-1} \backslash \mathcal{H}_j$ gives a subsequence that converges to $\mathbf{x}^*$.

This is written as $|x_n - L| = O(h^p)$ using the big O notation.

Analysis in R^n

- Global and local convergence

We say that an iterative algorithm is *globally convergent* if for any arbitrary starting point the algorithm is guaranteed to generate a sequence of points converging to a point that satisfies the (FONC) for a minimizer. When the algorithm is not globally convergent, it may still generate a sequence that converges to a point satisfying the FONC, provided that the initial point is sufficiently close to the point. In this case we say that the algorithm is *locally convergent*.

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

Q-linear

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

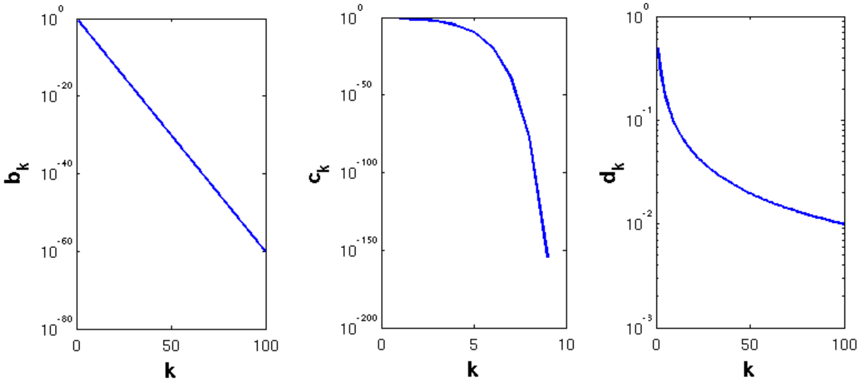
Sublinear: $r = 1$, $C = 1$;

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

Quadratic: $r = 2$;

Cubic: $r = 3$.

r may be non-integers



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