Optimization for Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Outline

Outline

What can I learn in this course?

In this course, you can learn:

1. the mathematical principles behind optimization methods;
2. how to choose suitable optimization algorithms for machine learning problems;
3. implementation of optimization methods.

Prerequisite course: calculus, linear algebra, probability, Python/Matlab.

It would be better if you learned: machine learning, convex optimization.

Course Overview

Grading Policy:

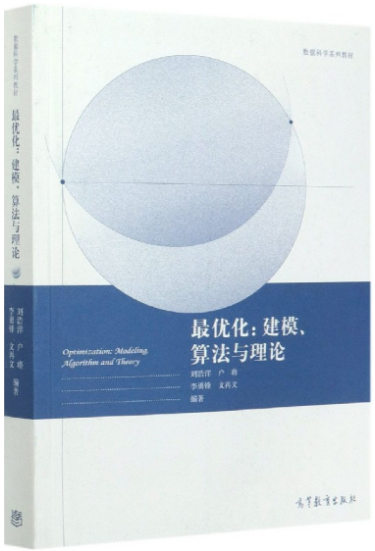
- Attendance, 10%
- Homework, 50%
- Mini-project, 40%

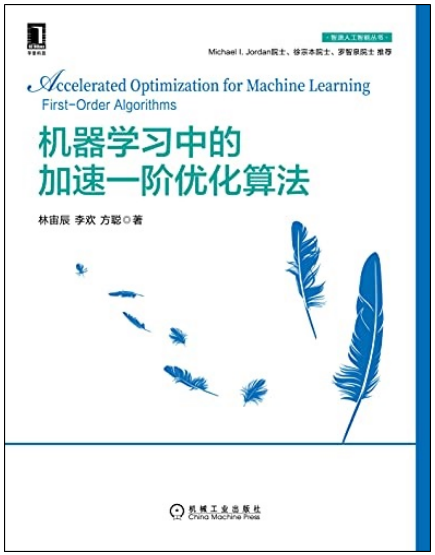
Teaching Assistant:

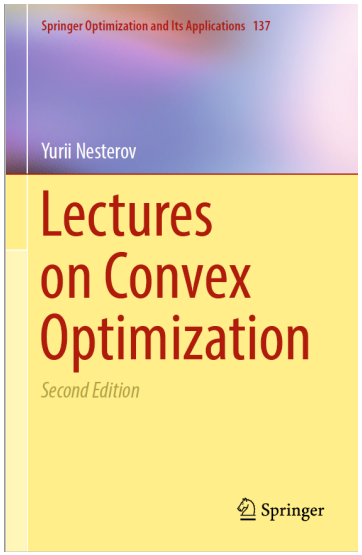
姚子奇：51265902073@stu.ecnu.edu.cn

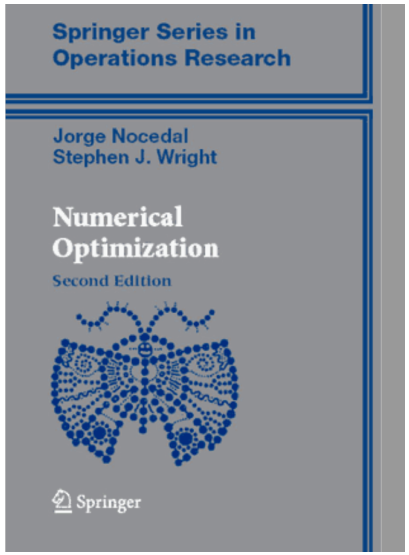
Course Overview

Recommended reading:





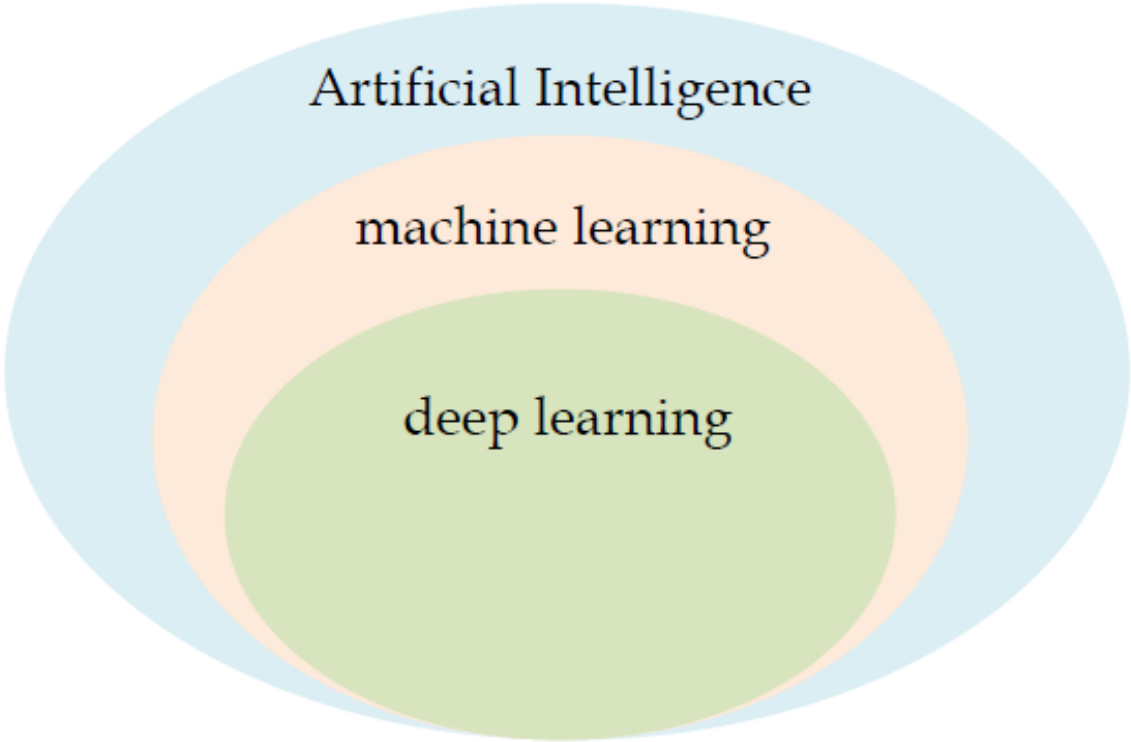




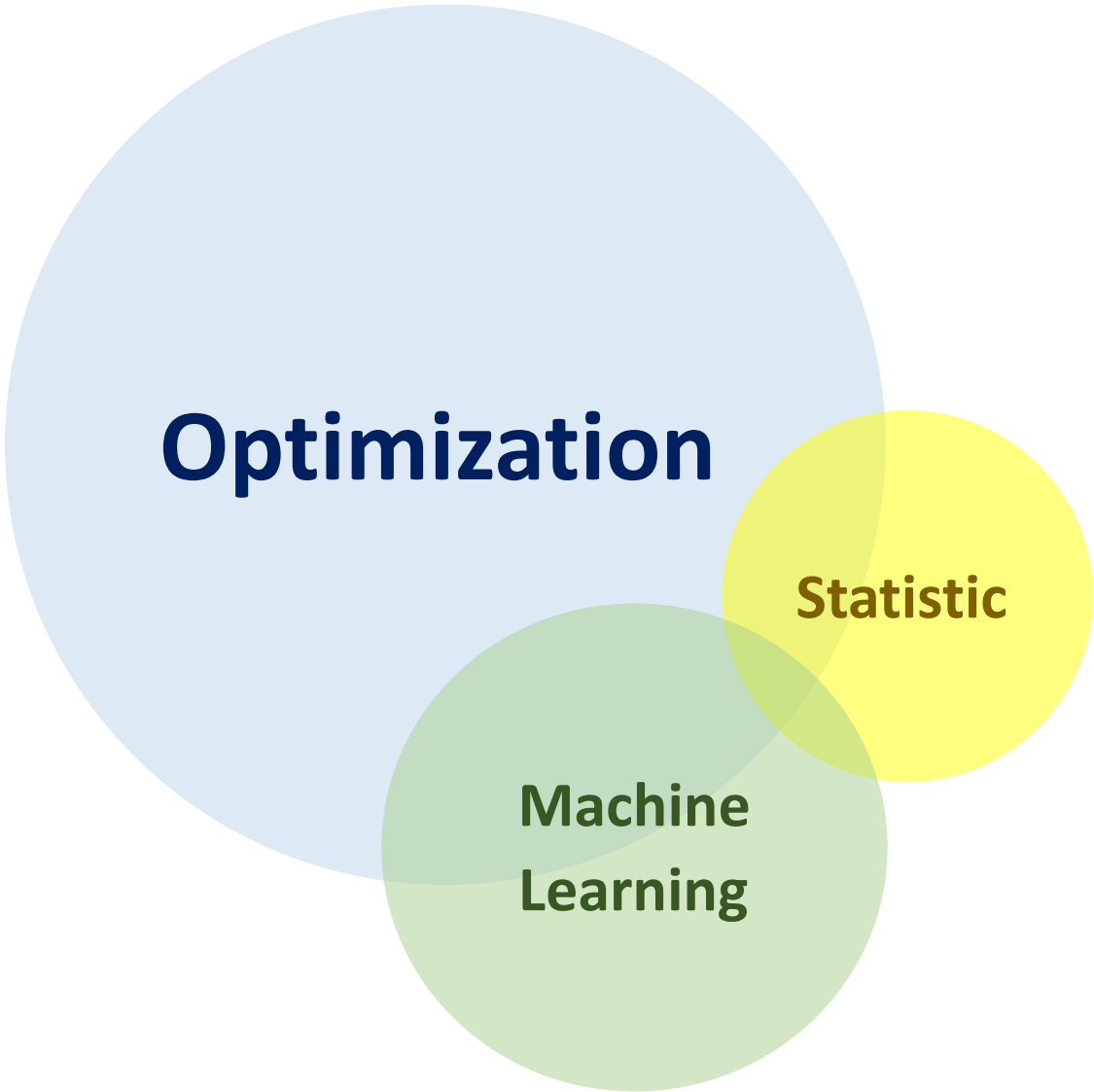
Outline

What is Machine Learning?

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



What is Optimization?



Why is Optimization Important?

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

Machine Learning = Representation + Evaluation + Optimization

Optimization

General optimization problem:

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

- feasible set: $\mathcal{X} \in \mathbb{R}^d$;
- objective function: $f : \mathbb{R}^d \to \mathbb{R}$;
- usually f is continuous in machine learning problems.

History of Optimization

- 1847: Cauchy proposes gradient descent
- 1950s: Linear Programs, soon followed by non-linear, Stochastic Gradient Descent (SGD)
- 1980s: General optimization, convergence theory
- 2005-2015: Large scale optimization (mostly convex), convergence of SGD
- 2015-today: Improved understanding of SGD for deep learning

No-Free-Lunch Theorem for Optimization

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.
- There is no universally better algorithms exist.

The Classification of Optimization Problems

The description of the feasible set:

- unconstrained vs. constrained

The properties of the objective function:

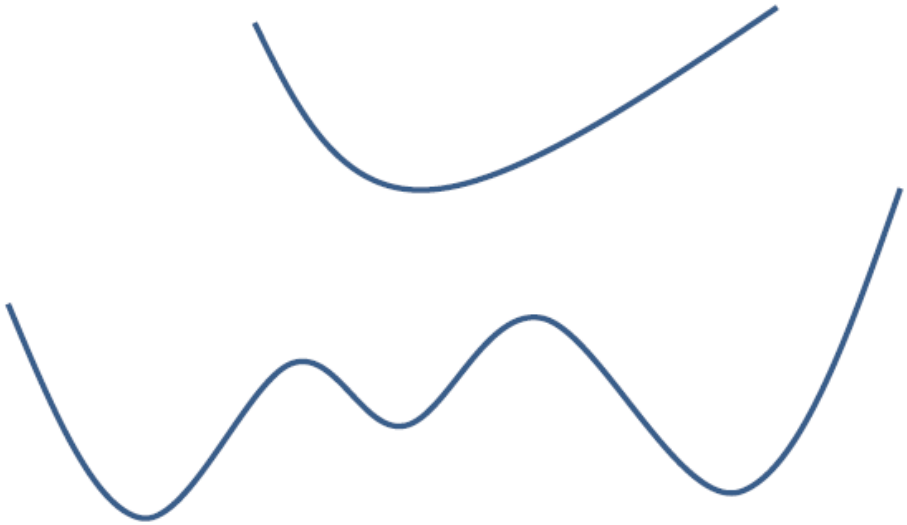
- linear vs. nonlinear
- smooth vs. nonsmooth
- convex vs. nonconvex

The settings in real application:

- deterministic vs. stochastic
- non-distributed vs. distributed

Convex vs. Nonconvex

"In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity." by R. T. Rockfeller



Convex Optimization in Machine Learning

The typical optimization problem in machine learning

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(b_i \mathbf{a}_i^\top \mathbf{x}).$$

We consider the following loss functions.

1. hinge loss (support vector machine):

$$l(z) = \max\{1 - z, 0\}$$

2. logistic loss (logistic regression):

$$l(z) = \ln(1 + \exp(-z))$$

Convex Optimization in Machine Learning

We typically introduce the regularization term

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(b_i \mathbf{a}_i^\top \mathbf{x}) + \lambda R(\mathbf{x}), \quad \text{where } \lambda > 0.$$

Some popular regularization terms:

- ridge regularization

$$R(\mathbf{x}) \triangleq \|\mathbf{x}\|_2^2$$

- Lasso regularization

$$R(\mathbf{x}) \triangleq \|\mathbf{x}\|_1$$

Non-convex Optimization in Machine Learning

We can use more general loss function and formulate

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{x};\mathbf{a}_i,b_i) + \lambda R(\mathbf{x}), \quad \text{where } \lambda > 0.$$

Here the loss function $l(\mathbf{x};\mathbf{a}_i,b_i)$ can be non-convex functions, e.g., neural networks.

Outline

Notations

We use x_i to denote the entry of the n-dimensional vector $\mathbf{x}$ such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

We use a_{ij} to denote the entry of matrix $\mathbf{A}$ with dimension $m \times n$ such that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Vector Norms

A norm of a vector $\mathbf{x} \in \mathbb{R}^n$ written by $\|\mathbf{x}\|$, is informally a measure of the length of the vector. For example, we have the commonly-used Euclidean norm (or ℓ_2 norm),

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Formally, a norm is any function $\mathbb{R}^n \to \mathbb{R}$ that satisfies four properties:

1. For all $\mathbf{x} \in \mathbb{R}^n$, we have $\|\mathbf{x}\| \geq 0$ (non-negativity).
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (definiteness).
3. For all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have $\|t\mathbf{x}\| = |t| \, \|\mathbf{x}\|$ (homogeneity).
4. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

Vector Norms

Some examples for $\mathbf{x} \in \mathbb{R}^n$:

- The ℓ_1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- The ℓ_2-norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
- The ℓ_∞-norm: $\|\mathbf{x}\|_\infty = \max_i |x_i|$

Vector Inner Product

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

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

We have following properties:

- $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|_2^2$
- $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ (Cauchy–Schwarz inequality)

Matrix Norms

General matrix norm is any function $\mathbb{R}^{m\times n} \to \mathbb{R}$ that satisfies:

1. For all $\mathbf{A} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A}\| \geq 0$ (non-negativity).
2. $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$ (definiteness).
3. For all $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, we have $\|t\mathbf{A}\| = |t| \, \|\mathbf{A}\|$ (homogeneity).
4. For all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ (triangle inequality).

Frobenius norm of $m \times n$ matrix $\mathbf{A}$:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}^2}$$

Induced Matrix Norms

Given vector norm $\|\cdot\|$, the corresponding induced matrix norm of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\|\mathbf{A}\| = \sup_{\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| = 1} \|\mathbf{A}\mathbf{x}\| .$$

For example, we define

$$\|\mathbf{A}\|_1 = \sup_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_1 = 1} \|\mathbf{A}\mathbf{x}\|_1$$

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2$$

$$\|\mathbf{A}\|_\infty = \sup_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_\infty = 1} \|\mathbf{A}\mathbf{x}\|_\infty .$$

Singular Value Decomposition

The singular value decomposition (SVD) of $\mathbf{A} \in \mathbb{R}^{m \times n}$ matrix is

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

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ is orthogonal, $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is rectangular diagonal matrix with non-negative real numbers on the diagonal and $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal.

Singular Value Decomposition

The SVD is not unique. It is always possible to choose the decomposition so that the singular values σ_i are in descending order.

The term sometimes refers to the compact SVD, a similar decomposition

$$\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top$$

in which $\mathbf{\Sigma}_r$ is square diagonal of size $r \times r$, where $r \leq \min\{m, n\}$ is the rank of $\mathbf{A}$, and has only the non-zero singular values. In this variant, $\mathbf{U}_r$ is an $m \times r$ column orthogonal matrix and $\mathbf{V}_r$ is an $n \times r$ column orthogonal matrix such that $\mathbf{U}_r^\top \mathbf{U}_r = \mathbf{V}_r^\top \mathbf{V}_r = \mathbf{I}$.

Pseudo-inverse of General Matrices

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 $\mathrm{rank}(\mathbf{A}) = r$. We define the pseudo-inverse of $\mathbf{A}$ as

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

Quadratic Forms

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the scalar $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is called a quadratic form and we have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

We often implicitly assume that the matrices appearing in a quadratic form are symmetric.

Definiteness

1. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive definite (PD) if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$. This is usually denoted by $\mathbf{A} \succ \mathbf{0}$.
2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive semi-definite (PSD) if for all vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$. This is usually denoted by $\mathbf{A} \succeq \mathbf{0}$.

Outline

Topology in Euclidean Space

Open set, closed set, bounded set and compact set:

1. A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called open, if for every $\mathbf{x} \in \mathcal{C}$ there exists $\delta > 0$ such that the ball $\mathcal{B}_\delta(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 \leq \delta\}$ is included in $\mathcal{C}$.
   Example: $\{x | a < x < b\}$, $\{\mathbf{x} | \mathbf{x} > 0\}$, $\{\mathbf{x} | \, \|\mathbf{x} - \mathbf{a}\| < 1\}$.
2. 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.
   Example: $\{x | a \leq x \leq b\}$, $\{\mathbf{x} | \mathbf{x} \geq 0\}$, $\{\mathbf{x} | \, \|\mathbf{x} - \mathbf{a}\| \leq 1\}$.
3. A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called bounded, if there exists $r > 0$ such that $\|\mathbf{x}\|_2 < r$ for all $\mathbf{x} \in \mathcal{C}$.
   Example: $\{x | a \leq x < b\}$, $\{\mathbf{x} | 1 > \mathbf{x} \geq 0\}$, $\{\mathbf{x} | \, \|\mathbf{x} - \mathbf{a}\| < 1\}$.
4. A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called compact, if it is both bounded and closed.
   Example: $\{x | a \leq x \leq b\}$, $\{\mathbf{x} | 1 \geq \mathbf{x} \geq 0\}$, $\{\mathbf{x} | \, \|\mathbf{x} - \mathbf{a}\| \leq 1\}$.

Topology in Euclidean Space

Interior, closure and boundary:

1. The interior of $\mathcal{C} \in \mathbb{R}^n$ is defined as

$$\mathcal{C}^\circ = \{\mathbf{y} : \text{there exist } \varepsilon > 0 \text{ such that } \mathcal{B}_\varepsilon(\mathbf{y}) \subset \mathcal{C}\}$$

2. The closure of $\mathcal{C} \in \mathbb{R}^n$ is defined as

$$\overline{\mathcal{C}} = \mathbb{R}^n \backslash (\mathbb{R}^n \backslash \mathcal{C})^\circ.$$

3. The boundary of $\mathcal{C} \in \mathbb{R}^n$ is defined as $\overline{\mathcal{C}} \backslash \mathcal{C}^\circ$.

Q-Convergence Rate

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

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

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

$$\lim_{k \to +\infty} \frac{z_{k+1}}{z_k^r} = C \quad \text{for some } C \in \mathbb{R}.$$

- linear: $r = 1$, $0 < C < 1$; Q-linear
- sublinear: $r = 1$, $C = 1$;
- superlinear: $r = 1$, $C = 0$;
- quadratic: $r = 2$.

Q-Convergence Rate

Examples:

- $x_k = 1/k^2$
- $x_k = 10^{-k}$
- $x_k = 10^{-2^k}$
- $x_{k+1} = x_k/2 + 2/x_k,\ x_1 = 4$

Convergence Rates

Consider the example

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

It should converge to $x^* = 1$ linearly, however,

$$\lim_{k \to +\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|}$$

does not exist.

R-Convergence Rates

Suppose that $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$. The sequence is said to converge R-linearly to $\mathbf{x}^*$ if there exists a sequence $\{\epsilon_k\}$ such that

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \epsilon_k$$

for all k and $\{\epsilon_k\}$ converges Q-linearly to zero.

The sequence

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

R-linearly converges to one.

Outline

Derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{x} \in (\mathrm{dom}\ f)^\circ$. The derivative at $\mathbf{x}$ is

$$Df(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

This matrix is also called Jacobian matrix.

Gradient

When f is real-valued, i.e., $f : \mathbb{R}^n \to \mathbb{R}$, the gradient of f is:

$$\nabla f(\mathbf{x}) = Df(\mathbf{x})^\top = \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

Gradient of Matrices

Suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$. Then the gradient of f with respect to $\mathbf{X}$ is

$$\nabla f(\mathbf{X}) = \frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Some Basic Results

1. For $\mathbf{X} \in \mathbb{R}^{m \times n}$, we have $\dfrac{\partial (f(\mathbf{X}) + g(\mathbf{X}))}{\partial \mathbf{X}} = \dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + \dfrac{\partial g(\mathbf{X})}{\partial \mathbf{X}}$.
2. For $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $t \in \mathbb{R}$, we have $\dfrac{\partial t f(\mathbf{X})}{\partial \mathbf{X}} = t \dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$.
3. For $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m \times n}$, we have $\dfrac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}$.
4. For $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

   If $\mathbf{A}$ is symmetric, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

We can find more results in the matrix cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

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 is differentiable at $\mathbf{x}$ and

$$Dh(\mathbf{x}) = D(g(f(\mathbf{x}))D(f(\mathbf{x})).$$

Examples:

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

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

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$. Define $h : \mathbb{R}^p \to \mathbb{R}$ as $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then,

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

Gradient of Logistic Regression

What is the gradient of the following function?

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

The Hessian Matrix

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function that takes as input a matrix $\mathbf{x} \in \mathbb{R}^n$ and returns a real value. Then the Hessian matrix with respect to $\mathbf{x}$, written as $\nabla^2 f(\mathbf{x})$, which is defined as

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Taylor's expansion for multivariable function $f : \mathbb{R}^n \to \mathbb{R}$

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

Chain Rules for Second Derivative

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

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

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$. Define $h : \mathbb{R}^p \to \mathbb{R}$ as $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then,

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