Optimization in 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

Course setup

Grading Policy:

- Homework, 40%
- Final project, 60%

Teaching Assistant:

- 胡子成：51275902019@stu.ecnu.edu.cn
- 贾廷锴：51275902086@stu.ecnu.edu.cn

Website:

- Homepage: `chengchen8.github.io/optml2024.html`
- Homework: 大夏学堂 `elearning.ecnu.edu.cn`

What can I learn in this course?

In practice, libraries are available, algorithms are treated as "black box".

NOT HERE: we look inside the optimization algorithms and try to understand why and how fast they work.

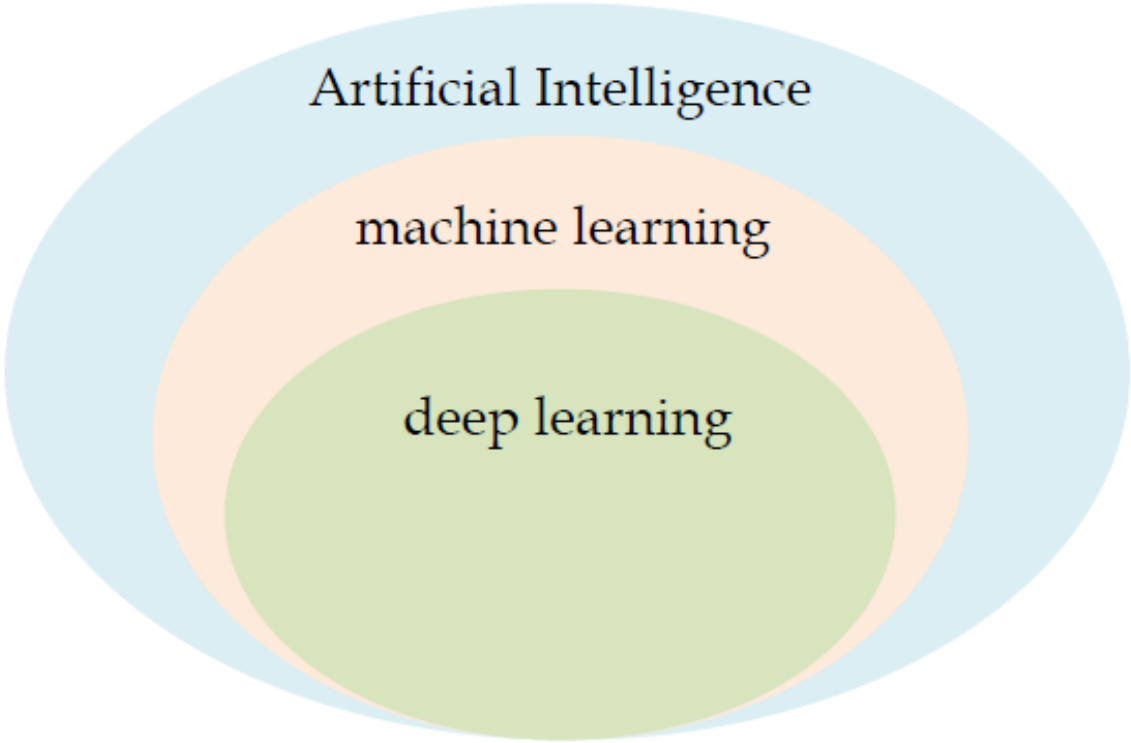
Prerequisite course: calculus, linear algebra, probability, Python/Matlab.

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

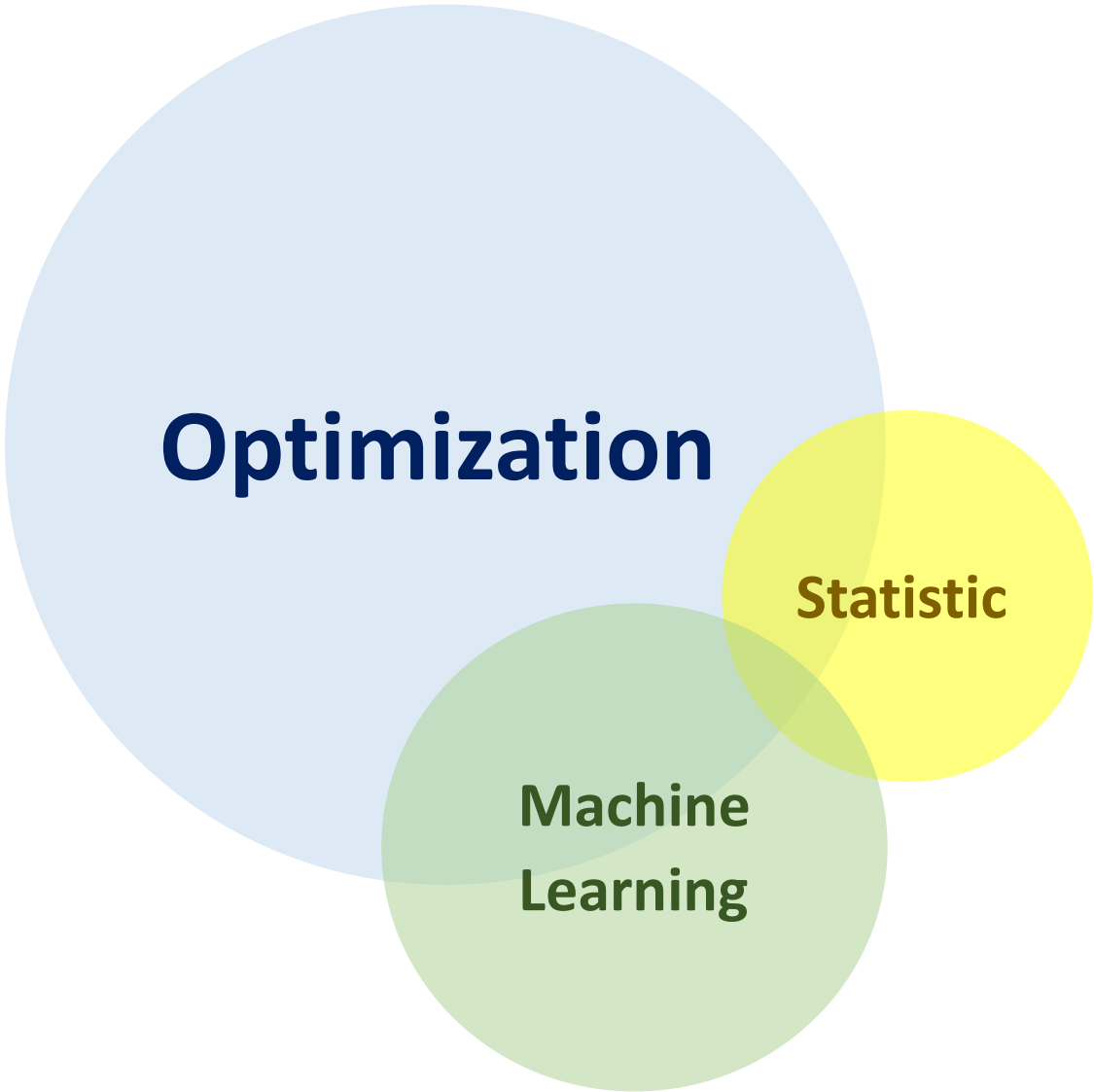
Outline

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. of UW):

Machine Learning = Representation + Evaluation + Optimization

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) for machine learning
- 2015-today: optimization methods for deep learning

Optimization problems

General optimization problem:

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

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

Classifications of optimization problems in machine learning

The description of the feasible set:

- unconstrained vs. constrained

The properties of the objective function:

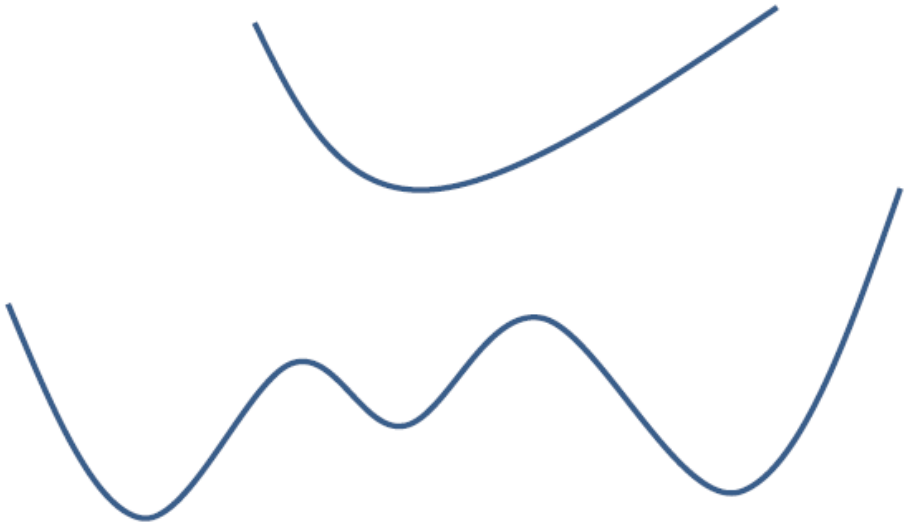
- linear vs. nonlinear
- smooth vs. nonsmooth
- convex vs. nonconvex

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



No-free-lunch theorem for optimization

D. H. Wolpert and W. G. Macready (1997):

- There is no universally better algorithms exist.
- 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.

Empirical risk minimization

The most common optimization problem in machine learning is the empirical risk minimization:

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

where $\mathbf{a}_i$ is the data point, b_i is the corresponding label and $\mathbf{x}$ is the parameter of the model.

$R(x)$ is called the regularization term.

Loss functions

Some traditional loss functions:

- squared loss (least square regression):

$$\ell(\mathbf{x}; \mathbf{a}_i, b_i) = (\mathbf{a}_i^\top \mathbf{x} - b_i)^2$$

- hinge loss (support vector machine):

$$\ell(\mathbf{x}; \mathbf{a}_i, b_i) = \max\{1 - b_i \mathbf{a}_i^\top \mathbf{x}, 0\}$$

- logistic loss (logistic regression):

$$\ell(\mathbf{x}; \mathbf{a}_i, b_i) = \ln(1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x}))$$

Regularization terms

Some popular regularization terms:

- Ridge regularization:

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

- Lasso regularization:

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

Deep learning

For deep learning the loss function ℓ could be highly nonconvex.

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

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

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

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

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

Symmetric eigenvalue decomposition

The eigenvalue decomposition (EVD) of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is

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

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is orthogonal and $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ is a diagonal matrix, i.e., $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ where λ_i are eigenvalues of $\mathbf{A}$.

Usually we order the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We use $\lambda_i(\mathbf{A})$ to denote the i-th largest eigenvalue of $\mathbf{A}$.

Singular value decomposition

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

Usually we order the eigenvalues as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}}$. We use $\sigma_i(\mathbf{A})$ to denote the i-th largest singular value of $\mathbf{A}$.

Singular value decomposition

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

Q-convergence Rates

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 rates

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.