# 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

1. Gradient descent

2. Quadratic minimization

3. Smoothness and strongly convex

# Outline

1. Gradient descent

2. Quadratic minimization

3. Smoothness and strongly convex

# Differentiable unconstrained optimization

Suppose the objective function (or loss function) $f$ is differentiable. The unconstrained optimization problem is:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$$

# Optimality condition (最优性条件)

Suppose $f$ is differentiable and convex. A point $\mathbf{x}^*$ is optimal if and only if

$$\nabla f(\mathbf{x}^*) = 0.$$

Strict convex function has unique optimal solution.

# Iterative descent methods

Start with a point $\mathbf{x}_0$ and construct a sequence $\{\mathbf{x}_t\}$ s.t.,

$$f(\mathbf{x}_{t+1}) < f(\mathbf{x}_t), \quad t = 0, 1, \ldots$$

We call $\mathbf{d}$ is a descent direction at $\mathbf{x}$ if

$$f'(\mathbf{x}; \mathbf{d}) \triangleq \underbrace{\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}}_{\text{directional derivative}} = \nabla f(\mathbf{x})^\top \mathbf{d} < 0.$$

# Iterative descent methods

- Start with a point $\mathbf{x}_0$;
- In each iteration, search in descent direction

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{d}_t$$

where $\mathbf{d}_t$ is the descent direction at $\mathbf{x}_t$, $\eta_t$ is the stepsize.

# How to find a descent direction?

By Cauchy-Schwarz inequality,

$$\min_{\|\mathbf{d}\|_2 \leq 1} f'(\mathbf{x}; \mathbf{d}) = \min_{\|\mathbf{d}\|_2 \leq 1} \nabla f(\mathbf{x})^\top \mathbf{d} = -\|\nabla f(\mathbf{x})\|_2$$

$f'(\mathbf{x}; \mathbf{d})$ achieve minimum when $\mathbf{d} = -\nabla f(\mathbf{x}) / \|\nabla f(\mathbf{x})\|_2$.

# Gradient descent (梯度下降法)

One of the most important descent methods: gradient descent

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$$

- descent direction: $\mathbf{d}_t = -\nabla f(\mathbf{x}_t)$
- traced to Augustin Louis Cauchy '1847
- First-order Taylor approximation: $f(\mathbf{x}) \approx f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle$

# Outline

1. Gradient descent

2. Quadratic minimization

3. Smoothness and strongly convex

# Quadratic minimization

We begin with the quadratic objective function:

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{b}^\top \mathbf{x},$$

for some $d \times d$ symmetric matrix $\mathbf{Q} \succ 0$.

- The gradient is $\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x} - \mathbf{b}$.
- The unique optimal solution is $\mathbf{x}^* = \mathbf{Q}^{-1}\mathbf{b}$.
- $\lambda_1(\mathbf{Q})\mathbf{I} \succeq \mathbf{Q} \succeq \lambda_d(\mathbf{Q})\mathbf{I}$, where $\lambda_1(\mathbf{Q})$ and $\lambda_d(\mathbf{Q})$ are largest and smallest eigenvalues of $\mathbf{Q}$ respectively.

# How to find a good stepsize?

According to the GD update rule,

$$\mathbf{x}_{t+1} - \mathbf{x}^* = \mathbf{x}_t - \mathbf{x}^* - \eta_t \nabla f(\mathbf{x}_t) = (\mathbf{I} - \eta_t \mathbf{Q})(\mathbf{x}_t - \mathbf{x}^*)$$

$$\Rightarrow \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \leq \|\mathbf{I} - \eta_t \mathbf{Q}\|_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2$$

We observe that

$$\|\mathbf{I} - \eta_t \mathbf{Q}\|_2 = \underbrace{\max\{|1 - \eta_t \lambda_1(\mathbf{Q})|, |1 - \eta_t \lambda_d(\mathbf{Q})|\}}_{\text{optimal choice is } \eta_t = \frac{2}{\lambda_1(\mathbf{Q}) + \lambda_d(\mathbf{Q})}}$$

$$= \frac{\lambda_1(\mathbf{Q}) - \lambda_d(\mathbf{Q})}{\lambda_1(\mathbf{Q}) + \lambda_d(\mathbf{Q})}$$

# Convergence for constant stepsize

If $\eta_t = \eta = \frac{2}{\lambda_1(\mathbf{Q})+\lambda_d(\mathbf{Q})}$, then

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \leq \left(\frac{\lambda_1(\mathbf{Q}) - \lambda_d(\mathbf{Q})}{\lambda_1(\mathbf{Q}) + \lambda_d(\mathbf{Q})}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2.$$

- linear convergence
- the convergence rate relys on the condition number $\kappa = \frac{\lambda_1(\mathbf{Q})}{\lambda_d(\mathbf{Q})}$

# Outline

1. Gradient descent

2. Quadratic minimization

3. Smoothness and strongly convex

# Generalization

Let's now generalize quadratic minimization to a broader class of problems

$$\min_{\mathbf{x}} f(\mathbf{x})$$

where

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}.$$

# Smoothness (光滑性)

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is $G$-Lipschitz continuous if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq G \left\| \mathbf{x} - \mathbf{y} \right\|_2 .$$

We say a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-smooth if it has $L$-Lipschitz continuous gradient. That is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$\left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \right\|_2 \leq L \left\| \mathbf{x} - \mathbf{y} \right\|_2 .$$

# Smoothness

Which of following functions are smooth?

- $f(x) = x^4$;
- $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ with $\mathbf{Q} \succeq 0$;

# Equivalent first-order characterizations of smoothness

Let $f : \mathbb{R}^d \leftarrow \mathbb{R}$ be a convex and differentiable function. Then the following properties are equivalent characterizations of $L$-smoothness of $f$:

1. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \; \forall\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d;$
2. $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq L\|\mathbf{x} - \mathbf{y}\|_2^2, \; \forall\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d;$
3. $f(\mathbf{y}) \leq \underbrace{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle}_{\text{first-order Taylor expansion}} + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2, \; \forall\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d;$
4. $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2, \; \forall\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d;$
5. $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2, \; \forall\; \mathbf{x}, \mathbf{y} \in \mathbb{R}^d;$

Which characterizations do not hold if $f$ is not convex?

# Equivalent first-order characterizations of smoothness (cont)

Let $f : \mathbb{R}^d \leftarrow \mathbb{R}$ be a convex and differentiable function. Then the following properties are equivalent characterizations of $L$-smoothness of $f$:

6. $\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \leq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) + \frac{L}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$;
7. $\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) + \frac{\lambda(1-\lambda)}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$.

# Equivalent second-order characterization of smoothness

We say a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-smooth if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2 .$$

**Second-Order Characterization:**
Let $f : \mathbb{R}^d \leftarrow \mathbb{R}$ be a twice differentiable function. Then the following property is an equivalent characterization of $L$-smoothness of $f$:

$$-L\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}.$$

# Strongly convexity (强凸性)

We say $f$ is $\mu$-strongly convex if the function

$$g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|_2^2$$

is convex for some $\mu > 0$.

# Equivalent first-order characterizations of strong convexity

Let $f : \mathbb{R}^d \leftarrow \mathbb{R}$ be a convex and differentiable function. Then the following properties are equivalent characterizations of $\mu$-strong convexity of $f$:

1. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \geq \mu\|\mathbf{x} - \mathbf{y}\|_2$, $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;
2. $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|_2^2$, $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;
3. $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;
4. $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$, $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;
5. $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \frac{1}{\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$, $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;

Strongly convex functions are strictly convex.

# Equivalent second-order characterization of strongly convexity

**Second-Order Characterization:**

Let $f : \mathbb{R}^d \leftarrow \mathbb{R}$ be a twice differentiable function. Then the following property is an equivalent characterization of $\mu$-strongly convex of $f$:

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}.$$

# Strongly convex and smooth functions

Let $f$ be $L$-smooth and $\mu$-strongly convex. Then we have

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}.$$

Let $\kappa \triangleq \frac{L}{\mu}$ be the condition number.

# Convergence rate of strongly convex and smooth problems

**Theorem.** Let $f$ be $L$-smooth and $\mu$-strongly convex. If $\eta = \frac{2}{\mu+L}$, then

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \le \left(\frac{\kappa-1}{\kappa+1}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2 .$$

**Proof 1:** Use fundamental theorem of calculus

$$\nabla f(\mathbf{x}_t) - \underbrace{\nabla f(\mathbf{x}^*)}_{=0} = \Big(\int_0^1 \nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_t - \mathbf{x}^*))\,\mathrm{d}\tau\Big)(\mathbf{x}_t - \mathbf{x}^*).$$

Then we have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 &= \|\mathbf{x}_t - \mathbf{x}^* - \eta\nabla f(\mathbf{x}_t)\|_2 \\
&= \left\|(\mathbf{I} - \eta\int_0^1 \nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_t - \mathbf{x}^*))\,\mathrm{d}\tau)(\mathbf{x}_t - \mathbf{x}^*)\right\|_2 \\
&\le \sup_{\tau\in[0,1]} \left\|\mathbf{I} - \eta\nabla^2 f(\mathbf{x}^* + \tau(\mathbf{x}_t - \mathbf{x}^*))\right\|_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2 \\
&\le \frac{L+\mu}{L-\mu}\|\mathbf{x}_t - \mathbf{x}^*\|_2 .
\end{aligned}$$

# Convergence rate of strongly convex and smooth problems

**Theorem.** Let $f$ be $L$-smooth and $\mu$-strongly convex. If $\eta = \frac{2}{\mu+L}$, then

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2 .$$

**Proof 2:** Use the following inequality

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{\mu + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2,$$

we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_{t+1} - \mathbf{x}_t + \mathbf{x}_t - \mathbf{x}^*\|_2^2 \\
&= \eta^2 \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^*)\|_2^2 - 2\eta \langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^* \rangle + \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \\
&\leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right) \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta \underbrace{\left(\eta - \frac{2}{\mu + L}\right)}_{=0} \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^*)\|_2^2
\end{aligned}$$

# Convergence rate of strongly convex and smooth problems

Let $f$ be $L$-smooth and $\mu$-strongly convex. If $\eta_t = \eta = \frac{2}{\mu+L}$, then

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2 .$$

Iteration complexity: To achieve $\epsilon$-accuracy, we require $\frac{\log(\|\mathbf{x}_0 - \mathbf{x}^*\|_2/\epsilon)}{\log(\frac{\kappa+1}{\kappa-1})}$ number of iterations.

Dimension-free: The iteration complexity is independent of problem size $d$ if $\kappa$ does not depend on $d$.

# Summary

- Gradient descent
- Smoothness and strongly convex
  - First-order characterizations
  - Second-order characterizations
- Convergence rate of GD for strongly convex and smooth problems