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

Review of Smooth and Strongly Convex

We say a differentiable function f is L-smooth if for all $\mathbf{x}, \mathbf{y}$ we have

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

We say a function f is μ-strongly convex if the function

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

is convex for some $\mu > 0$.

Let f be L-smooth and μ-strongly convex. Its condition number is defined as $\kappa \triangleq \frac{L}{\mu}$ and we have

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

Review of Gradient Descent

Gradient Descent: Start with the initial point $\mathbf{x}_0$ and computes

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

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

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

To achieve ϵ-accuracy, i.e., $\|\mathbf{x}_t - \mathbf{x}^*\|_2 \leq \epsilon$, the necessary number of iterations is

$$\frac{\log(\|\mathbf{x}_0 - \mathbf{x}^*\|_2 / \epsilon)}{\log(\frac{\kappa+1}{\kappa-1})} = \underbrace{O\left(\kappa \log \frac{1}{\epsilon}\right)}_{\text{iteration complexity}} .$$

Convergence of $f(\mathbf{x}_t) - f(\mathbf{x}^*)$

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

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

By smoothness and strong convexity, we know

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \kappa \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2t} (f(\mathbf{x}_0) - f(\mathbf{x}^*)).$$

Convergence of $f(\mathbf{x}_t) - f(\mathbf{x}^*)$

Let f be L-smooth and μ-strongly convex. If $\eta_t = \eta = \frac{1}{L}$, then the outputs of GD satisfies

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\kappa}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}^*)),$$

which means the iteration complexity is also $O\left(\kappa \log \frac{1}{\epsilon}\right)$

Line Search (线搜索)

In practice, one often performs line searches rather than adopting constant stepsizes because:

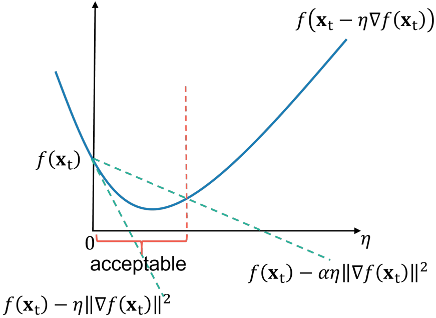
- L may be unknown;
- L may be too high.

Exact line search:

$$\eta_t = \arg\min_{\eta \geq 0} f(\mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)).$$

Exact line search is usually not practical since the subproblem is hard to solve.

Backtracking Line Search (回溯线搜索)

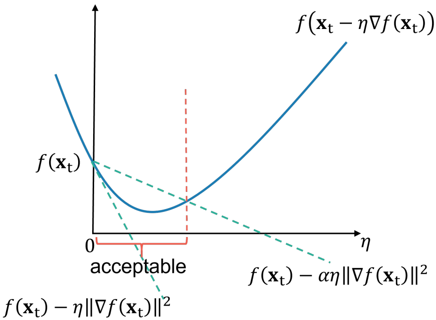


Armijo condition: for $0 < \alpha < 1$,

$$f(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)) < f(\mathbf{x}_t) - \alpha\eta\|\nabla f(\mathbf{x}_t)\|_2^2$$

- $f(\mathbf{x}_t) - \alpha\eta\|\nabla f(\mathbf{x}_t)\|_2^2$ lies above $f(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t))$ for small η
- ensures sufficient decrease of objective values

Backtracking Line Search (回溯线搜索)



1: Initialize $\eta = 1$, $0 < \alpha \leq 1/2$, $0 < \beta < 1$.
2: **while** $f(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)) > f(\mathbf{x}_t) - \alpha\eta\|\nabla f(\mathbf{x}_t)\|_2^2$ **do**
3: $\quad \eta \leftarrow \beta\eta$

Convergence of Backtracking Line Search

Theorem (Boyd, Vandenberghe '04)

Let f be L-smooth and μ-strongly convex. With backtracking line search,

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \min\left\{2\mu\alpha, \frac{2\alpha\beta\mu}{L}\right\}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}^*))$$

Summary

So far we have established linear convergence under strong convexity and smoothness.

Is strong convexity necessary for linear convergence?

Example: Logistic Regression

Suppose we obtain n independent binary samples

$$b_i = \begin{cases} 1 & \text{with prob. } \frac{1}{1+\exp(-\mathbf{a}_i^\top \mathbf{x})} \\ -1 & \text{with prob. } \frac{1}{1+\exp(\mathbf{a}_i^\top \mathbf{x})} \end{cases}$$

where the $\mathbf{a}_i$ and b_i are the feature vector and the label of the i-th data sample respectively, $\mathbf{x}$ is the model parameters.

Example: Logistic Regression

The maximum likelihood estimate (MLE) is given by (after a little manipulation)

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x}))$$

- $\nabla^2 f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(-b_i \mathbf{a}_i^\top \mathbf{x})}{(1+\exp(-b_i \mathbf{a}_i^\top \mathbf{x}))^2} \mathbf{a}_i \mathbf{a}_i^\top \xrightarrow{\mathbf{x} \to \infty} 0$
  $\Rightarrow f$ is 0-strongly convex
- Does it mean we no longer have linear convergence?

Local Strong Convexity

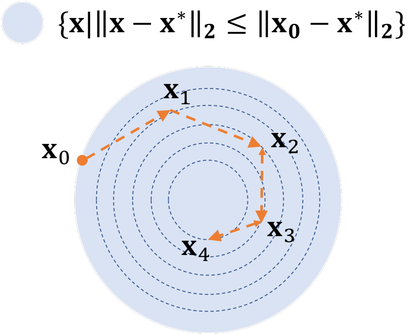
Let f be locally L-smooth and μ-strongly convex such that

$$\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}, \ \forall \mathbf{x} \in \mathcal{B}_0$$

where $\mathcal{B}_0 = \{\mathbf{x} | \, \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2\}$. Then GD obeys

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

Local Strong Convexity



- Suppose $\mathbf{x}_t \in \mathcal{B}_0$. Then follow previous analysis yields $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \leq \frac{\kappa-1}{\kappa+1} \|\mathbf{x}_t - \mathbf{x}^*\|_2$
- This means $\mathbf{x}_{t+1} \in \mathcal{B}_0$, so the above bound continues to hold for the next iteration ...

Local Strong Convexity

The local strong convexity parameter of the logistic regression example is given by

$$\inf_{\{\mathbf{x} \mid \|\mathbf{x}-\mathbf{x}^*\|_2 \le \|\mathbf{x}_0-\mathbf{x}^*\|_2\}} \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\exp(-b_i\mathbf{a}_i^\top\mathbf{x})}{(1+\exp(-b_i\mathbf{a}_i^\top\mathbf{x}))^2}\mathbf{a}_i\mathbf{a}_i^\top\right)$$

which is often stricly bounded away from 0.

Polyak-Lojasiewicz Condition

Recall that an equivalent condition of μ-strongly convex is

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

If we choose $\mathbf{y} = \mathbf{x}^*$, we get the Polyak-Lojasiewicz (PL) condition

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2.$$

where $\mathbf{x}^*$ can be any minimum of f.

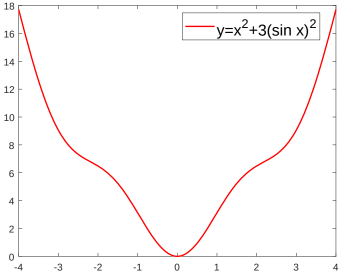
The PL condition guarantees that gradient grows fast as we move away from the optimal objective value.

Polyak-Lojasiewicz (PL) Condition

PL condition:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|_2^2$$

- does NOT imply the function is convex
- does NOT imply the uniqueness of global minima
- guarantees that every stationary point is a global minimum



Convergence under PL condition

Suppose f is L-smooth and satisfies PL condition with parameter μ. If $\eta_t = \eta = \frac{1}{L}$, then GD obeys

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\kappa}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}^*)),$$

which means the iteration complexity is also $O\left(\kappa \log \frac{1}{\epsilon}\right)$

Example: Over-parameterized Linear Regression

Linear regression:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2.$$

Over-parameterization: model dimension $>$ sample size, i.e., $(d > n)$.

- $\nabla^2 f(\mathbf{x}) = \sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top$ is rank-deficient if $d > n$, thus $f(\mathbf{x})$ is not strongly convex
- PL condition is met

Example: Over-parameterized Linear Regression

Suppose $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times d}$ has rank n, and that $\eta_t = \eta = \frac{1}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}$. Then GD obeys

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \frac{\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}^*)).$$

- very mild assumption on $\mathbf{A}$
- while there are many global minima for this over-parametrized problem, GD converges to a global min closest to initialization $\mathbf{x}_0$.

Dropping strong convexity

What happens if we completely drop (local) strong convexity?

We only suppose $f(\mathbf{x})$ is smooth and convex.

Convergence rate for convex and smooth problems

Let f be convex and L-smooth. If $\eta_t = \eta = \frac{1}{L}$, then GD obeys

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{t}$$

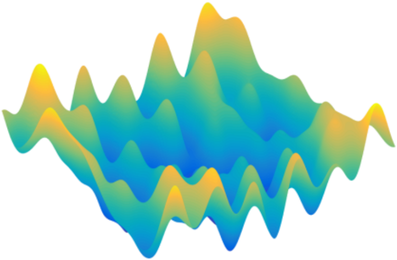
- Without strong convexity, convergence is typically much slower than linear convergence
- attains ϵ-accuracy within $O(\frac{1}{\epsilon})$ iterations (vs $O(\log(\frac{1}{\epsilon}))$ iterations for linear convergence)

Nonconvex problems

Many objective functions in machine learning are nonconvex:

- low-rank matrix completion
- mixture models
- learning deep neural nets
- ...

Challenges



- there may be local minima everywhere
- no algorithm can solve nonconvex problems efficiently in all cases

Typical Convergence Guarantees

We cannot hope for efficient global convergence to global minima in general, but we may have

- convergence to stationary points ,i.e., $\nabla f(\mathbf{x}) = 0$
- convergence to local minima
- local convergence to global minima i.e., when initialized suitably

Making Gradients Small

Suppose we aim to find a stationary point, which means that our goal is merely to find a point $\mathbf{x}$ with

$$\|\nabla f(\mathbf{x})\|_2 \leq \epsilon \text{ (called } \epsilon\text{-approximate stationary point)}$$

ϵ-approximate stationary point does not imply local minima for nonconvex optimization.

Making Gradients Small

Let f be L-smooth and $\eta_t = \eta = \frac{1}{L}$, then GD obeys

$$\min_{0 \le k \le t} \|\nabla f(\mathbf{x}_t)\|_2 \le \sqrt{\frac{2L(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{t}}.$$

- GD finds an ϵ-approximate stationary point in $O(1/\epsilon^2)$ iterations.
- does not imply GD converges to stationary points; it only says that there exists an approximate stationary point in the GD trajectory

Questions