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: smooth and strongly convex

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

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

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

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

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: 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: 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: 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

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

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

Review: smooth and strongly convex

Let $f : \mathbb{R}^d \leftarrow \mathbb{R}$ be a convex and differentiable function. Then the following properties are equivalent characterizations of μ-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. $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \frac{1}{\mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2, \ \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;

Review: 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

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

Proof 1 (last week): use fundamental theorem of calculus

Proof 2: Use the following inequality

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

Review: 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

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

Proof 1 (last week): use fundamental theorem of calculus

Proof 2: Use the following inequality

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

Review: 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

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

Proof 1 (last week): use fundamental theorem of calculus

Proof 2: Use the following inequality

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

Convergence property

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

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

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

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.

Line search (线搜索)

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

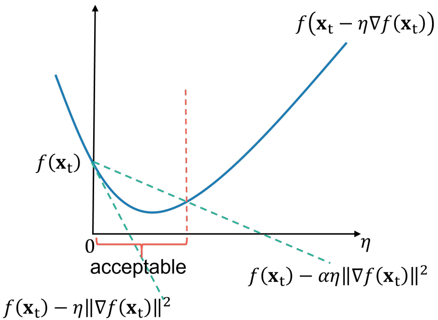
- L may be unknown;
- L may be too high.

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)

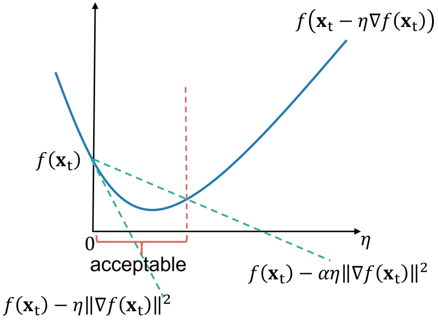


Armijo condition: for $0 < \alpha < 1$,

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

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

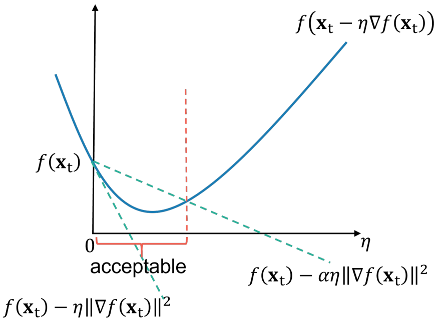


Armijo condition: for $0 < \alpha < 1$,

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

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

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

where the $\mathbf{a}_i$ and y_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 estimation (MLE) is given by (after a little manipulation)

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

- $\nabla^2 f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-y_i \mathbf{a}_i^\top \mathbf{x})}{(1+\exp(-y_i \mathbf{a}_i^\top \mathbf{x}))^2} \mathbf{a}_i \mathbf{a}_i^\top \xrightarrow{\mathbf{x}\to\infty} 0$

  $\Rightarrow f$ is 0-strongly convex
- Does it mean we no longer have linear convergence?

Example: logistic regression

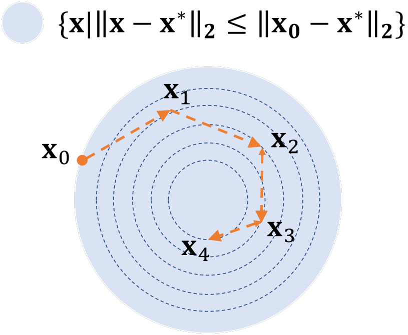
The maximum likelihood estimation (MLE) is given by (after a little manipulation)

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

- $\nabla^2 f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-y_i \mathbf{a}_i^\top \mathbf{x})}{(1+\exp(-y_i \mathbf{a}_i^\top \mathbf{x}))^2} \mathbf{a}_i \mathbf{a}_i^\top \xrightarrow{\mathbf{x}\to\infty} 0$

  $\Rightarrow f$ is 0-strongly convex
- Does it mean we no longer have linear convergence?

Local strong convexity



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

Local strong convexity

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

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

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

which is often strictly bounded away from 0.

Polyak-Lojasiewicz condition

Recall that an equivalent condition of μ-strongly convex is

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

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

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

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 condition

Recall that an equivalent condition of μ-strongly convex is

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

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 condition

Recall that an equivalent condition of μ-strongly convex is

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

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.

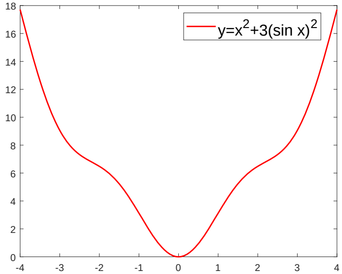
The PL condition guarantees that gradient grows fast as we move away from the optimal objective value.

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

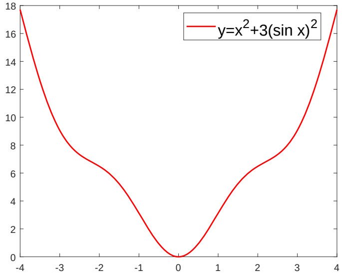


Polyak-Lojasiewicz (PL) condition

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

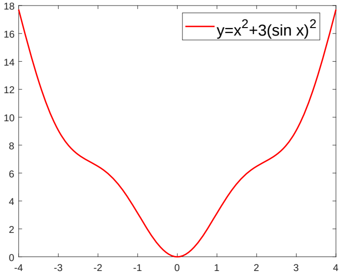


Polyak-Lojasiewicz (PL) condition

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-parametrization: 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-parameterized 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.

Dropping strong convexity

Without strong convexity, it may often be better to focus on objective improvement (rather than improvement on estimation error)

Example: consider $f(x) = 1/x$ $(x > 0)$. GD iterates $\{x_t\}$ might never converge to $x^* = \infty$. In comparison, $f(x_t)$ might approach $f(x^*) = 0$ rapidly.

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)

Summary

	stepsize	convergence rate	iteration complexity
strongly convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\epsilon})$
locally strongly convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\epsilon})$
PL condition & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\epsilon})$
convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\epsilon}\right)$

Table: Convergence Property of GD