Learning objectives.

1. Derive the descent lemma and the useful fixed-step range $0 < \alpha < 2/L$;
2. Prove the classical $O(1/\sqrt{N})$ best-iterate gradient bound for smooth objectives bounded below;
3. Use the μ-Polyak–Łojasiewicz condition to obtain linear $O((1-\mu/L)^N)$ convergence

1 Recap: Optimality conditions

Consider $\min_{x\in\mathbb{R}^n} f(x)$. In Lecture 5, we saw that if gradient descent

$$x_{k+1} = x_k - \alpha\nabla f(x_k) \tag{GD}$$

converges, its limit must be a first-order critical point $\nabla f(x) = 0$.

Question 1. Does (GD) always converge?

1.1 Infimum and coercivity

To avoid failure, we define the *infimum* as the greatest lower bound on the function

$$f_{\inf} \coloneqq \max_{t \in \mathbb{R}} \{t : f(x) \geq t \text{ for all } x \in \mathbb{R}^n\}.$$

We further say that the function is *coercive* if

$$\|x\| \to \infty \quad \Longrightarrow \quad f(x) \to \infty.$$

After the Midterm, we will see $f_{\inf}$ is closely associated with the *Lagranian dual*.

Question 2. Which failure mode does having an infimum prevent?

Question 3. Which failure mode does coercivity prevent?

1.2 First-order optimality

For simplicitly, let f be coercive. If it is further differentiable, then it is bounded from below, so it attains its infimum.

Theorem 1. First-order Weierstrass. Let $f : \mathbb{R}^n \to \mathbb{R}$ be *differentiable*. Then,

$$f(x) = \inf_{x \in \mathbb{R}^n} f(x) \quad \implies \quad \nabla f(x) = 0.$$

If f is further *convex*, then

$$f(x) = \inf_{x \in \mathbb{R}^n} f(x) \quad \iff \quad \nabla f(x) = 0.$$

The set of *first-order critical points* is defined

$$\text{FOC} \coloneqq \{x \in \mathbb{R}^n : \nabla f(x) = 0\}.$$

If f is convex, then any $x \in \text{FOC}$ is globally optimal. Otherwise, must exhaustively enumerate FOC and check every candidate.

Question 4. Why is nonconvex considered "hard"?

1.3 Second-order optimality

We can prune obviously bad points in FOC by considering the next term in the local Taylor expansion:

$$f(x_\star + d) \approx f(x_\star) + \nabla f(x_\star)^T d + \frac{1}{2} d^T \nabla^2 f(x_\star) d.$$

Intuition: If f is *locally convex* about $x_\star$, and its gradient is zero at $x_\star$, then f is locally optimal.

Theorem 2. Second-order Weierstrass. Let $f : \mathbb{R}^n \to \mathbb{R}$ be *twice differentiable* and *coercive*, and *bounded*. Then,

$$f(x) = \inf_{x \in \mathbb{R}^n} f(x) \quad \Longrightarrow \quad \nabla f(x) = 0,\ \nabla^2 f(x) \succeq 0.$$

The set of *second-order critical points* is defined

$$\text{SOC} \coloneqq \{x \in \mathbb{R}^n : \nabla f(x) = 0,\ \nabla^2 f(x) \succeq 0\}$$

GD with random init $x_0 \sim \mathcal{N}(0, I_n)$ almost surely $x_k \to \text{SOC}$ as $k \to \infty$. Faster convergence with noisy gradients.

Example 3. Classify the origin $(x, y) = (0, 0)$ for

$$f_+(x, y) = x^2 + y^4, \qquad f_-(x, y) = x^2 - y^4.$$

2 Main: Convergence of GD

Let f be bounded and coercive.

Question 5. Does $x_{k+1} = x_k - \alpha \nabla f(x_k)$ converge to $\nabla f(x) = 0$?

A differentiable function is said to have *L-Lipschitz gradient* or is said to be *L-smooth* on some set C if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 \qquad \text{for every } x, y \in C.$$

The property is said to hold *globally* if it holds for all $x, y \in \mathbb{R}^n$.

We default to "global" if C is unspecified.

Lemma 4. If f is twice differentiable, then it is L-smooth over every bounded set C.

But the numerical value of L is usually hard to find.

Theorem 5. Let $f_{\text{inf}} = \inf_x f(x)$. If f is L-smooth, then GD

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

with step-size $\alpha = \frac{1}{L}$ achieves

$$\min_{k \in \{0,1,\dots,N\}} \|\nabla f(x_k)\| \leq \sqrt{\frac{2L(f(x_0) - f_{\text{inf}})}{N}}.$$

Put simply, $\|\nabla f(x_{\text{best}})\| = O(1/\sqrt{N})$ in the worst case.

This is super slow, but it is **attained**!

Question 6. Does the theorem require f to be bounded?

Question 7. Does the theorem require f to be coercive?

The theorem does not say that the iterates converge, and it does not say that a critical point is globally optimal. It does not even say that the last iterate is necessarily the best.

3 Preliminary: Descent lemma

The following is the main reason we assume L-smoothness.

Lemma 6. Let C be a convex set. If f is L-smooth on C, then

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2 \text{ for all } x, y \in C.$$

Lemma 7. Let f be L-smooth on $\mathbb{R}^n$. Then,

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x)\|_2^2$$

Question 8. What stepsize guarantees *strict objective decrease*?

Question 9. What stepsize maximizes the guaranteed decrease?

Question 10. Is that the *best* fixed step-size?

4 Proof of Main Theorem

Picking $\alpha = \frac{1}{L}$ in the descent lemma yields the following.

Lemma 8. Optimized Descent Lemma. Let f be L-smooth on $\mathbb{R}^n$. With the step-size $\alpha = \frac{1}{L}$, we have

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

Theorem 9. Let $f_{\text{inf}} = \inf_x f(x)$. If f is L-smooth, then GD

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

with step-size $\alpha = \frac{1}{L}$ achieves

$$\min_{k \in \{0,1,\ldots,N\}} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_{\text{inf}})}{LN}.$$

5 Linear convergence under μ-PL

Sublinear convergence is the general guarantee, but it is very slow. Gradient descent converges much much faster for a function satisfying the μ-Polyak–Łojasiewicz (μ-PL) condition:

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu[f(x) - f_{\text{inf}}] \qquad \text{for every } x.$$

This definition isolates the critical part for geometric / exponential convergence.

Theorem 10. If f is L-smooth and satisfies μ-PL, then gradient descent $x_{k+1} = x_k - \alpha\nabla f(x_k)$ with fixed step-size $\alpha = \frac{1}{L}$ satisfies

$$f(x_k) - f_{\text{inf}} \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_{\text{inf}}].$$

Recall that $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \text{ for all } x, y \in \mathbb{R}^n.$$

It is moreover *μ-strongly convex* if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2 \text{ for all } x, y \in \mathbb{R}^n.$$

Equivalently, if f has a Hessian, then

$$f \text{ is } \mu\text{-strongly convex} \iff \nabla^2 f(x) \succeq \mu I \text{ for all } x \in \mathbb{R}^n.$$

All μ-strongly convex functions globally satisfy μ-PL, but many non-convex functions also *locally* satisfy μ-PL.

Lemma 11. If f is μ-strongly convex, then it also satisfies μ-PL.

6 (*) When is a function L-smooth?

Having L-Lipschitz gradient is stronger than simply having a gradient, but generally weaker than having a Hessian.

Lemma 12. If f is twice differentiable, then it is L-smooth over every bounded set C.

Gradient descent is a monotone method, meaning that its iterates satisfy

$$f(x_0) \geq f(x_1) \geq f(x_2) \geq \cdots \geq f(x_k).$$

As such, we only need L-smoothness over the initial level set

$$\mathcal{L}(x_0) = \{x : f(x) \leq f(x_0)\}.$$

This greatly widens the set of functions that are L-smooth.

Question 11. If f is once differentiable, is it L-smooth over every bounded set C?

Example 13. Are these functions L-smooth? Compute L.

1. $f(x) = x^2$ over $x \in \mathbb{R}$
2. $f(x) = -x^2$ over $x \in \mathbb{R}$
3. $f(x) = x^4$ over $x \in \mathbb{R}$
4. $f(x) = x^4$ over $x \in \mathcal{L}(1)$
5. $f(x) = x^3$ over $x \in \mathcal{L}(1)$
6. $f(x) = \sin x$ over $x \in \mathbb{R}$
7. $f(x, y) = x^2 + 4y^2$ over $x, y \in \mathbb{R}$
8. $f(x, y) = xy$ over $x, y \in \mathbb{R}$

7 Summary

Descent lemma is the critical tool for studying gradient descent:

- L-smooth and bounded below $\Longrightarrow$ first-order stationarity;
- L-smooth and μ-PL $\Longrightarrow$ linear convergence.

All require step $\alpha = \frac{1}{L}$, but L is usuall unknown. Hand tune or adaptive step (next time).

The homework repeats this proof architecture for preconditioned gradient descent,

$$x_{k+1} = x_k - \alpha B \nabla f(x_k), \qquad B = B^T \succ 0,$$

using the paired norms

$$\|v\|_B^2 = v^T B v, \qquad \|v\|_{B^{-1}}^2 = v^T B^{-1} v.$$

The geometry changes, but the pattern remains identical: one-step decrease, telescope, then invoke extra structure if available.