

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?

Solution: No. Reasons for failure:

1. Step-size too large.
2. Function is *not differentiable*, e.g. $f(x) = |x|$.
3. Unbounded function value $f(x_k) \rightarrow -\infty$, e.g. $f(x) = x$.
4. Iterate blow-up $\|x_k\| \rightarrow \infty$. This occurs when critical point is at “infinity”, e.g. $f(x) = \exp(x)$.

1.1 Infimum and coercivity

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

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

We further say that the function is *coercive* if

$$\|x\| \rightarrow \infty \implies f(x) \rightarrow \infty.$$

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

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

Solution: It prevents **unbounded function value**.

Indeed, the objective is bounded (from below) if and only if $f_{\inf} > -\infty$. So gradient descent objective value cannot diverge $f(x_k) \rightarrow -\infty$ as $k \rightarrow \infty$.

Question 3. Which failure mode does coercivity prevent?

Solution: It prevents **iterate blowup**. Otherwise, the infimum might not be attained. For example, $f(x) = \exp(x)$ has $f_{\inf} = 0$ but there exists no optimum x^* for which $f(x^*) = 0$. Gradient descent objective value converges as $f(x_k) \rightarrow 0$, but its iterate diverges $x_k \rightarrow -\infty$. If f is coercive then this cannot happen.

1.2 First-order optimality

For simplicity, 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be *differentiable*. Then,

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

If f is further *convex*, then

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

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

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

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”?

Solution: Impossible in practice to truly exhaust FOC by GD. (But you can actually do a great job with random initialization.)

Possibly easy to find a candidate, but hard to certify that it is the best.

Often FOC contains many obviously bad solutions.

1.3 Second-order optimality

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

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

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

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

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

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

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

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

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

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

Solution: Both functions satisfy

$$\nabla f_{\pm}(0, 0) = 0, \quad \nabla^2 f_{\pm}(0, 0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \succeq 0.$$

However, $f_+(x, y) \geq 0$ with equality only at the origin, so the origin is a strict global minimum. For f_- , values are positive along $y = 0$ and negative along $x = 0$, so the origin is a saddle point. Second-order criticality alone does not decide either case.

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 $\nabla f(x) = 0$?

Solution: Not necessarily. We still need f to be *differentiable*, and we still need to choose *good step-size* α .

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

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

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 $f_{\inf} = \inf_x f(x)$. 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

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

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?

Solution: Technically no, although $f_{\inf} = -\infty$ renders the theorem vacuous.

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

Solution: No, GD can converge to critical point at infinity. Theorem allows $x_k \rightarrow \infty$ so long as $f(x_k) \rightarrow f_{\inf}$.

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

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

Proof: Let $v = y - x$ and define $\phi(t) = f(x + tv)$. Then

$$\phi(1) - \phi(0) = \int_0^1 \phi'(t) dt$$

rewrites into

$$f(y) - f(x) = \int_0^1 \nabla f(x + tv)^T v dt.$$

Add and subtract $\nabla f(x)^T v$:

$$\begin{aligned} f(y) - f(x) &= \nabla f(x)^T v \\ &\quad + \int_0^1 [\nabla f(x + tv) - \nabla f(x)]^T v dt. \end{aligned}$$

By Cauchy–Schwarz and Lipschitz continuity,

$$[\nabla f(x + tv) - \nabla f(x)]^T v \leq Lt \|v\|_2^2.$$

Integrating Lt from zero to one gives the result. □

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

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

Proof: Substitute $y = x - \alpha \nabla f(x)$ into quadratic upper bound. Then,

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T(-\alpha \nabla f(x)) + \frac{L}{2} \|\alpha \nabla f(x)\|_2^2 \\ &= f(x) - \alpha \|\nabla f(x)\|_2^2 + \alpha^2 \frac{L}{2} \|\nabla f(x)\|_2^2. \end{aligned}$$

Finally, group terms. □

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

Solution: $\nabla f(x) \neq 0, 0 < \alpha < \frac{2}{L} \implies f(x - \alpha \nabla f(x)) < f(x)$.

Question 9. What stepsize maximizes the guaranteed decrease?

Solution: Define the line search function

$$\phi(\alpha) = \alpha \left(1 - \frac{L\alpha}{2}\right).$$

This function is concave, because

$$\phi'(\alpha) = 1 - L\alpha, \quad \phi''(\alpha) = -L.$$

Its global maximum is attained at first-order optimality $\phi'(\alpha) = 0$, with step-size $\alpha = \frac{1}{L}$.

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

Solution: Not necessarily. It gives the **largest decrement for the worst-case choice** of L -smooth function f . For your specific f , a different step-size may result in faster convergence.

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 $f_{\inf} = \inf_x f(x)$. 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

$$\min_{k \in \{0, 1, \dots, N\}} \|\nabla f(x_k)\|_2^2 \leq \frac{2(f(x_0) - f_{\inf})}{LN}.$$

Proof: Rearrange optimized descent lemma yields

$$\frac{1}{2L} \|\nabla f(x_k)\|_2^2 \leq f(x_k) - f(x_{k+1}).$$

Sum from $k = 0$ to $N - 1$:

$$\begin{aligned} \frac{1}{2L} \sum_{k=0}^{N-1} \|\nabla f(x_k)\|_2^2 &\leq \sum_{k=0}^{N-1} [f(x_k) - f(x_{k+1})] \\ &= [f(x_0) - f(x_1)] + [f(x_1) - f(x_2)] \\ &\quad + \dots + [f(x_{N-1}) - f(x_N)] \\ &= f(x_0) - f(x_N) \\ &\leq f(x_0) - f_{\inf}. \end{aligned}$$

On the other hand

$$\begin{aligned}\|\nabla f(x_k)\|_2^2 &\geq \|\nabla f(x_{\text{best}})\|_2^2 \\ \frac{1}{2L} \sum_{k=0}^{N-1} \|\nabla f(x_k)\|_2^2 &\geq \frac{1}{2L} \sum_{k=0}^{N-1} \|\nabla f(x_{\text{best}})\|_2^2 \\ &= \frac{N}{2L} \|\nabla f(x_{\text{best}})\|_2^2.\end{aligned}$$

Rearranging obtains the desired claim. □

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:

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

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

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

Proof: Recall the optimized descent lemma

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2.$$

Subtract $f_{\min}$ and substitute μ -PL

$$\begin{aligned} f(x_{k+1}) - f_{\inf} &\leq f(x_k) - f_{\inf} - \frac{1}{2L} \|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) - f_{\inf} - \frac{\mu}{L} (f(x_k) - f_{\inf}) \\ &= \left(1 - \frac{\mu}{L}\right) (f(x_k) - f_{\inf}). \end{aligned}$$

□

Recall that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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.

Proof: If f is μ -strongly convex, then it must attain its infimum $f_{\inf} = f(x^*)$. At this point, we have

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|^2.$$

Reorder, apply Cauchy-Schwarz, and complete the square

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|^2 \\ &\leq \|\nabla f(x)\| \|x - x^*\| - \frac{\mu}{2}\|x^* - x\|^2 \\ &= \frac{1}{2\mu} \|\nabla f(x)\|^2 - \frac{\mu}{2} \left(\|x^* - x\| - \frac{\|\nabla f(x)\|}{\mu} \right)^2 \\ &\leq \frac{1}{2\mu} \|\nabla f(x)\|^2. \end{aligned}$$

□

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 .

Proof: Set $\overline{C} = \text{Conv}(C)$ as the smallest convex set that encloses C . If C is bounded, then $\overline{C}$ is bounded. Conversely, if f is L -smooth over $\overline{C}$, then it is also L -smooth over C .

We now show that f is L -smooth over $\overline{C}$. For any $x, y \in \overline{C}$, write $v = x - y$. Since $\overline{C}$ is convex, we have $x + tv \in \overline{C}$ for $0 \leq t \leq 1$ and

$$\begin{aligned} \nabla f(x) - \nabla f(y) &= \int_0^1 \nabla^2 f(y + tv) v \, dt \\ \|\nabla f(x) - \nabla f(y)\| &= \left\| \int_0^1 \nabla^2 f(y + tv) v \, dt \right\| \\ &\leq \int_0^1 \|\nabla^2 f(y + tv) v\| \, dt \\ &\leq \int_0^1 \|\nabla^2 f(y + tv)\|_{\text{op}} \|v\| \, dt \\ &\leq L \int_0^1 \|v\| \, dt = L \|v\| = L \|x - y\|. \end{aligned}$$

□

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 ?

Solution: Not necessarily. Consider $f(x) = |x|^{3/2}$. Its minimizer is $x = 0$, so $0 \in \mathcal{L}(x_0)$. Then

$$f'(x) = \frac{3}{2}\text{sign}(x)\sqrt{|x|}, \quad f'(0) = 0$$

and for $x > 0$, we have

$$\frac{|f'(x) - f'(0)|}{|x - 0|} = \frac{|\frac{3}{2}\sqrt{x} - 0|}{|x - 0|} = \frac{3}{2\sqrt{x}} \rightarrow \infty.$$

So f is differentiable, and yet no finite L can make the gradient Lipschitz on $\mathcal{L}(x_0)$.

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 $x, y \in \mathbb{R}$

Solution:

1. Yes, $L = 2$. We have $|f''(x)| = |2| \leq 2$ for every $x \in \mathbb{R}$.
2. Yes, $L = 2$. We have $|f''(x)| = |-2| \leq 2$ for every $x \in \mathbb{R}$.
3. No, since $f''(x) = 12x^2$ is unbounded over $\mathbb{R}$.
4. Yes, $L = 12$. Here $\mathcal{L}(1) = [-1, 1]$, and $|f''(x)| = 12x^2 \leq 12$ on this set.
5. No. Here $\mathcal{L}(1) = (-\infty, 1]$, and $|f''(x)| = |6x|$ is unbounded on this set.
6. Yes, $L = 1$. We have $|f''(x)| = |-\sin x| \leq 1$ for every $x \in \mathbb{R}$.
7. Yes, $L = 8$. Since

$$\nabla^2 f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix},$$

we have $\|\nabla^2 f(x, y)\|_{\text{op}} = 8$ for every $(x, y) \in \mathbb{R}^2$.

8. Yes, $L = 1$. Since

$$\nabla^2 f(x, y) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

whose eigenvalues are ± 1 , we have $\|\nabla^2 f(x, y)\|_{\text{op}} = 1$.

7 Summary

Descent lemma is the critical tool for studying gradient descent:

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

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

The homework repeats this proof architecture for preconditioned gradient descent,

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

using the paired norms

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

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