Learning objectives. By the end of class, you should be able to

1. state the first-order necessary condition and use it to solve simple convex problems in closed form;
2. distinguish boundedness below, infimum, and minimum;
3. use continuity and coercivity to certify that a minimum is attained;
4. state the second-order necessary condition and use it to solve simple benignly nonconvex problems in closed form;
5. state the second-order sufficient conditions and recognize the gap between necessity and sufficiency.

1 First-order critical points

This week we begin analyzing gradient descent, the ancestor of many modern optimization algorithms. Consider the fixed-step iteration

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \qquad \alpha > 0.$$

Where does this sequence converge to, assuming that it converges?

Question 1. Suppose that gradient descent converges to a fixed point $x_\star$. What equation must $x_\star$ satisfy?

Solution: Since $x_k \to x_\star$, we also have $x_{k+1} \to x_\star$ and $\nabla f(x_k) \to \nabla f(x_\star)$. Therefore

$$x_\star = x_\star - \alpha \nabla f(x_\star),$$

so $\nabla f(x_\star) = 0$.

A point x is called a *first-order critical point* or a *stationary point* if it satisfies

$$\nabla f(x) = 0.$$

From this viewpoint, gradient descent is an iterative method for solving the nonlinear equation $\nabla f(x) = 0$.

Theorem 1. First-order necessary condition. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and $x_\star$ is a local minimizer, then

$$\nabla f(x_\star) = 0.$$

Key idea is that every *directional derivative* must vanish.

Proof: Fix any direction $d \in \mathbb{R}^n$ and define the line restriction

$$\phi(t) = f(x_\star + td).$$

Since $x_\star$ is a local minimum of f, it follows that $t = 0$ is a local minimum of ϕ. This implies that

$$0 = \phi'(0) = \nabla f(x_\star)^T d.$$

In order for $x_\star$ to be a local minimum of f, the above must hold for every d. Taking $d = \nabla f(x_\star)$ gives

$$0 = \nabla f(x_\star)^T d = \|\nabla f(x_\star)\|^2$$

which can hold if and only if $\nabla f(x_\star) = 0$. $\square$

In general, a critical point can be a local minimum, a local maximum, or a saddle point. Solving for $\nabla f(x) = 0$ generates only *candidates*, not necessarily global minima. For a convex function, however, every critical point is *guaranteed* to be a global minimum.

Lemma 2. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then

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

In fact, the above is "if and only if", but we will only prove "if".

Proof: The chord inequality for convex f reads

$$\begin{aligned} f(y + t(x - y)) &= f(tx + (1 - t)y) \\ &\leq tf(x) + (1 - t)f(y) = f(y) + t(f(x) - f(y)). \end{aligned}$$

Rearranging, f is convex if and only if

$$\frac{f(y + t(x - y)) - f(y)}{t} \leq f(x) - f(y) \text{ for all } x, y \in \mathbb{R}^n. \qquad \text{(A)}$$

Taking the limit

$$\lim_{t \to 0} \frac{f(y + t(x - y)) - f(y)}{t} = \nabla f(y)^T (x - y).$$

□

Theorem 3. Convex first-order sufficient condition. If $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* and differentiable, then

$$\nabla f(x_\star) = 0 \iff f(x_\star) \leq f(x) \text{ for all } x \in \mathbb{R}^n.$$

Proof: ($\Longrightarrow$) If f is convex, then

$$f(x) \geq f(x_\star) + \nabla f(x_\star)^T (x - x_\star) = f(x_\star).$$

($\Longleftarrow$) See previous page. □

2 Solving by first-order optimality

Instead of letting gradient descent converge to a first-order critical point, we can simply solve for $\nabla f(x) = 0$. We loop back to the example from Lect 1, where we claimed that "linear least-squares can be solved in closed-form."

Example 4. Convex unconstrained. Consider

$$f(a,b) = \frac{1}{2}\sum_{i=1}^{m}(ax_i + b - y_i)^2,$$

where the x_i are not all equal. Solve $\nabla f(a,b) = 0$ and show that the first-order critical point is the unique global minimizer.

Solution: Differentiate with respect to the two unknowns a and b:

$$\frac{\partial f}{\partial a} = \sum_{i=1}^{m}(ax_i + b - y_i)x_i, \qquad \frac{\partial f}{\partial b} = \sum_{i=1}^{m}(ax_i + b - y_i).$$

Therefore $\nabla f(a,b) = 0$ is equivalent to

$$a\sum_i x_i^2 + b\sum_i x_i = \sum_i x_i y_i, \qquad a\sum_i x_i + mb = \sum_i y_i.$$

Collect these into the following equations

$$\begin{bmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{bmatrix}.$$

Writing $\bar{x} = m^{-1}\sum_i x_i$ and $\bar{y} = m^{-1}\sum_i y_i$ gives

$$a_\star = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \qquad b_\star = \bar{y} - a_\star \bar{x}.$$

The denominator is strictly positive because the x_i are not all equal, so this critical point is unique.

Finally, we confirm that f is convex, because

$$\nabla^2 f(a,b) = \begin{bmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & m \end{bmatrix}$$

and any $(u,v) \neq (0,0)$,

$$\begin{bmatrix} u & v \end{bmatrix} \nabla^2 f(a,b) \begin{bmatrix} u \\ v \end{bmatrix} = \sum_{i=1}^{m} (u x_i + v)^2 \geq 0.$$

Therefore, every first-order critical point must also be globally optimal.

3 Infimum versus minimum

In previous section, we **assumed** that gradient descent converges to a fixed point. Is it actually true?

Question 2. Why might gradient descent fail to converge?

Solution: Algorithm: Step-size too large.

Problem: Function f is *not differentiable*, e.g. $f(x) = |x|$.

Problem: Objective is *unbounded*, or if first-order critical points are "at infinity". In these two cases, the equation $\nabla f(x) = 0$ has no solution, and gradient descent will iterate forever.

We want to exclude the problem-specific failure modes. Define the *infimum* as the greatest lower bound on the function

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

The objective is bounded (from below) if and only if $f_{\text{inf}} > -\infty$.

Example 5. Unbounded. Let $f(x) = cx$ with $c \neq 0$. What is f_{inf} for this function, and what is a first-order critical point? Derive the gradient-descent iterates and explain their asymptotic behavior.

Solution: Since $f'(x) = c \neq 0$, no first-order critical point. Taking $x \to -\infty$ when $c > 0$, or $x \to +\infty$ when $c < 0$, shows that the function is unbounded below, so $f_{\text{inf}} = -\infty$. Gradient descent satisfies

$$x_{k+1} = x_k - \alpha c, \qquad x_k = x_0 - \alpha c k,$$

and drives $f(x_k) = cx_0 - \alpha c^2 k \to -\infty$.

Even when f is bounded, a minimum exists only if some point attains this value:

$$f(x_\star) = f_{\inf}.$$

This is not generally true, not even for convex functions.

Example 6. Bounded but not attained. Let $f(x) = \exp(-x^2/2)$. Find the infimum, decide whether a minimum exists, and classify every critical point.

Solution: The function is positive for every finite x and approaches zero as $|x| \to \infty$. Therefore $f_{\inf} = 0$, but the infimum is not attained. Also

$$f'(x) = -x\exp(-x^2/2),$$

so $x = 0$ is the only critical point. Since $f''(0) = -1$, it is a strict global maximum.

How to prevent those two failure modes? A useful existence condition is *coercivity*:

$$\|x\| \to \infty \quad \implies \quad f(x) \to +\infty.$$

Coercivity prevents a minimizing sequence from escaping to infinity.

Theorem 7. First-order Weierstrass. If $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable* and *coercive*, then f attains a global minimum at a *first-order critical point.*

Note that "differentiable" just means that the gradient exists everywhere.

Example 8. Are the following functions coercive?

1. $f(x) = x^2$
2. $f(x) = -x^2$
3. $f(x, y) = xy$
4. $f(x, y) = x^2$
5. $f(x) = (x^2 - 1)^2$
6. $f(x, y) = (xy - 1)^2$
7. $f(x, y) = (x - y)^2$
8. $f(x, y) = x^2 + y^4 - xy$

Solution:

1. Coercive. $|x| \to \infty$ implies $f(x) = x^2 \to \infty$.
2. Not coercive. $|x| \to \infty$ implies $f(x) = -x^2 \to -\infty$.
3. Not coercive. Along $(x, y) = (t, 0)$, we have $\|(x, y)\| \to \infty$ but $f(t, 0) = 0$.
4. Not coercive. Along $(x, y) = (0, t)$, we have $\|(x, y)\| \to \infty$ but $f(0, t) = 0$.
5. Coercive. $|x| \to \infty$ implies $f(x) = (x^2 - 1)^2 \to \infty$.
6. Not coercive. Along $(x, y) = (t, 1/t)$, we have $\|(x, y)\| \to \infty$ but $f(t, 1/t) = 0$.
7. Not coercive. Along $(x, y) = (t, t)$, we have $\|(x, y)\| \to \infty$ but $f(t, t) = 0$.
8. Coercive. Since $x^2 + y^4 - xy = (x - y/2)^2 + y^4 - y^2/4$, we have $f(x, y) \to \infty$ whenever $\|(x, y)\| \to \infty$.

4 Second-order optimality conditions

A first-order critical point with $\nabla f(x) = 0$ is not necessarily optimal, not even locally. We can make a better determination 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 9. Second-order necessary condition. If f is twice differentiable and $x_\star$ is a local minimizer, then

$$\nabla f(x_\star) = 0, \qquad \nabla^2 f(x_\star) \succeq 0.$$

Such a point is called a *second-order critical point*. Randomly initialized gradient descent almost surely converges to such a point in its asymptotic limit, though this may take exponential time.

Proof: The first condition follows from first-order necessity. For any direction d, the function $\phi(t) = f(x_\star + td)$ has a local minimum at zero, so

$$0 \le \phi''(0) = d^T \nabla^2 f(x_\star) d.$$

This inequality for every d is exactly $\nabla^2 f(x_\star) \succeq 0$. □

Theorem 10. Second-order sufficient condition. If f is twice continuously differentiable near $x_\star$ and

$$\nabla f(x_\star) = 0, \qquad \nabla^2 f(x_\star) \succ 0,$$

then $x_\star$ is a strict local minimizer.

Proof: Positive definiteness gives a number $m > 0$ such that $d^T \nabla^2 f(x_\star) d \geq m \|d\|^2$. By continuity, the Hessian remains positive definite in a sufficiently small neighborhood. Taylor's theorem then gives

$$f(x_\star + d) - f(x_\star) \geq \frac{m}{4} \|d\|^2 > 0$$

for every sufficiently small nonzero d. □

The necessary condition allows a singular Hessian; the sufficient condition does not. When the Hessian is positive semidefinite but singular, higher-order terms decide the classification.

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

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

What do the gradient and Hessian say in each case?

Solution: Both functions satisfy

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

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.

Example 11. Benignly nonconvex. Let $C = C^T$ with $\lambda_{\min}(C) = -\gamma^2 < 0$, and

$$f(x) = \frac{1}{2}x^T Cx + \frac{1}{4}\|x\|^4.$$

Is this function convex? Nevertheless, solve $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$ and show that any second-order critical point is the global minimizer.

Solution: Let $\lambda_{\min}(C)$ denote the smallest eigenvalue of C. First compute

$$\nabla f(x) = Cx + \|x\|^2 x, \qquad \nabla^2 f(x) = C + \|x\|^2 I + 2xx^T.$$

Given that C has a negative eigenvalue, then $\nabla^2 f(0) = C \not\succeq 0$, so f is *nonconvex*.

We now show that even in this nonconvex case every second-order critical point is a global minimizer. The first-order condition is

$$0 = \nabla f(x) = (C + \|x\|^2 I)x.$$

Indeed, $x = 0$ is a first-order critical point, since $\nabla f(0) = 0$, but $\nabla^2 f(0) = C \not\succeq 0$, so it is not a second-order critical point.

Now, suppose $x \neq 0$. Then, $\nabla f(x) = 0$ is rewritten

$$Cx = -\|x\|^2 x,$$

so x is an eigenvector of C with eigenvalue $-\|x\|^2$. Rescale x so that $\|x\| = \gamma$, then

$$\begin{aligned} z^T[\nabla^2 f(x)]z &= z^T[C + \|x\|^2 I + 2xx^T]z \\ &= z^T Cz + \|x\|^2\|z\|^2 + 2(x^T z)^2 \\ &\geq \lambda_{\min}(C)\|z\|^2 + \|x\|^2\|z\|^2 + 2(x^T z)^2 \\ &= -\gamma^2\|z\|^2 + \gamma^2\|z\|^2 + 2(x^T z)^2 \geq 0, \end{aligned}$$

so indeed $\nabla^2 f(x) \succeq 0$. In conclusion, any x satisfying the following

$$Cx = -\gamma x, \qquad \|x\| = \gamma$$

is a second-order critical point.

Finally, we verify that this point is actually globally optimal. In general, we have the lower-bound

$$f(z) = \frac{1}{2} z^T C z + \frac{1}{4}\|z\|^4 \geq \frac{1}{2}\lambda_{\min}(C)\|z\|^2 + \frac{1}{4}\|z\|^4.$$

But the second-order critical point found above attains this lower-bound with equality:

$$f(x) = \frac{1}{2} x^T C x + \frac{1}{4}\|z\|^4 = -\frac{1}{2}\gamma\|x\|^2 + \frac{1}{4}\|x\|^4.$$

It is therefore globally optimal.

5 Summary

Gradient descent seeks first-order critical points, but several failure modes remain:

- the objective may be unbounded below;
- the infimum may be finite but unattained;
- a first-order critical point may fail to be locally optimal.

With random initialization, its asympotitic limit is a second-order critical point.

- For convex problems, all first-order critical points are globally optimal.
- For benign nonconvex problems, all second-order critical points are globally optimal.
- In general, even second-order critical points may fail to be locally optimal.

Next time: how fast does gradient descent approach first-order criticality $\nabla f(x) \approx 0$?