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?

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.

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".

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

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

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.

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?

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

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

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

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.

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

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

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.

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

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

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. $f(x, y) = x^2 + y^4 - xy$

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.

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.

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?

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.

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 $\nabla f(x) \approx 0$?