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

1. explain why the fixed step $\alpha = 1/L$ is a worst-case guarantee and why adaptive steps can perform better;
2. explain why objective decrease alone does not guarantee convergence to a stationary point;
3. state the Armijo condition and prove finite termination and sufficient decrease of backtracking line search;
4. prove convergence to first-order stationarity, and use the PL condition to obtain linear objective convergence.

1 Recap: Fixed-step Gradient Descent

Recall that $f : \mathbb{R}^n \to \mathbb{R}$ is said to be globally L-smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \text{ for all } x, y \in \mathbb{R}^n.$$

If f is twice-differentiable, then the above is equivalently written

$$\left\|\nabla^2 f(x)\right\|_{\text{op}} = \max_i \left|\lambda_i(\nabla^2 f(x))\right| \leq L \text{ for all } x \in \mathbb{R}^n.$$

Immediate consequence of L-smoothness:

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

The fixed-step analysis of Lecture 6 was critically anchored by the following result, which is obtained by substituting $x_+ = x - \alpha \nabla f(x)$ into the above.

Lemma 1. Descent. If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and $x_+ = x - \alpha \nabla f(x)$, then

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

In particular, choosing stepsize $\alpha = \frac{1}{L}$ yields

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

Question 1. What is the significance of the stepsize $\alpha = \frac{1}{L}$?

Theorem 2. 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

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

The μ-Polyak–Łojasiewicz (μ-PL) condition reads:

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

For example, satisfied by μ-strongly convex functions

$$\nabla^2 f(x) \succeq \mu I \text{ for all } x \in \mathbb{R}^n \implies f \text{ satisfies } \mu\text{-PL}.$$

Guarantees linear convergence.

Theorem 3. 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

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

2 What is the best step-size?

For L-smooth f, **fixed step** $\alpha = \frac{1}{L}$ is the "best". (In what sense?)

But in real-world, value of L usually unknown.

Even when L is known, **adaptive step** is usually still much faster:

1. Initialize optimistically large stepsize $\alpha = 1$.
2. If
$$f(x - \alpha \nabla f(x)) < f(x),$$
return $x_+ = x - \alpha \nabla f(x)$.
3. Else if $f(x - \alpha \nabla f(x)) \geq f(x)$, half the stepsize $\alpha \leftarrow \alpha/2$ and go back to **Step 2**.

Question 2. Why is adaptive faster than "optimal" fixed step?

Question 3. Why do many applications still use fixed step?

3 But function decrease is not enough

In practice, the strict decrement rule works well:

$$\text{If } f(x - \alpha \nabla f(x)) < f(x), \text{ accept } \alpha.$$

But cannot guarantee convergence to $\nabla f(x) = 0$.

Question 4. Why are rigorous guarantees *impossible*?

Example 4. Consider gradient descent on $f(x) = x^2$ starting at $x_0 = 1$ with step-sizes

$$\alpha_k = \frac{1}{2(k+2)^2}.$$

1. Does the sequence satisfy $f(x_{k+1}) < f(x_k)$?
2. Does the sequence converge to $f'(x) = 0$?

4 Backtracking line search

Gradient descent choose search direction $v = -\nabla f(x)$ and satisfies the descent lemma

$$f(x + \alpha v) \leq f(x) + \alpha \left(1 - \frac{L\alpha}{2}\right) \nabla f(x)^T v$$

Inspired by this, the **Armijo condition** selects $c_1 \in (0, 1)$ and requires:

Our previous strict descent condition forces $c_1 \to 0^+$. Turns out, finite $c_1 > 0$ allows convergence results essentially as good as $\alpha = \frac{1}{L}$.

1. Initialize optimistically large stepsize $\alpha = \bar{\alpha} > 0$.
2. If
$$f(x - \alpha \nabla f(x)) \leq f(x) + \alpha c_1 \nabla f(x)^T v,$$
return $x_+ = x - \alpha \nabla f(x)$.
3. Else if $f(x - \alpha \nabla f(x)) \geq f(x)$, half the stepsize $\alpha \leftarrow c_2 \alpha$ and go back to **Step 2**.

Typical values are $c_1 = 0.01$ (guarantee 1% linear decrement) and $c_2 = 0.5$ (half stepsize at every trial).

Lemma 5. Suppose that f is L-smooth and v is any descent direction satisfying $\nabla f(x)^T v < 0$. Every trial step satisfying

$$0 < \alpha \leq -\frac{2(1-c_1)\nabla f(x)^T v}{L \|v\|_2^2}$$

passes the Armijo condition.

For gradient descent $v = -\nabla f(x)$, the acceptance condition reduces to

$$0 < \alpha \leq -\frac{2(1-c_1)\nabla f(x)^T v}{L \|v\|_2^2} = \frac{2(1-c_1)}{L}.$$

So if α satisfies the above, then the backtracking terminates.

Lemma 6. Suppose f is L-smooth. At each gradient-descent iteration, backtracking line search terminates after at most

$$1 + \max\left\{0, \left\lceil \log_{1/c_2}\left(\frac{\bar{\alpha}L}{2(1-c_1)}\right)\right\rceil\right\} \text{ trials,}$$

and the resulting step satisfies $\alpha \geq \underline{\alpha}$ where

$$\underline{\alpha} \coloneqq \min\left\{\bar{\alpha}, \frac{2c_2(1-c_1)}{L}\right\} > 0.$$

5 Convergence of BT line search

Theorem 7. Let $f_{\inf} = \inf_x f(x)$. If f is L-smooth, then GD

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

with backtracking line search satisfies

$$\min_{k \in \{0,1,\dots,N\}} \|\nabla f(x_k)\| \leq \sqrt{\frac{f(x_0) - f_{\inf}}{c_1 \underline{\alpha} N}}$$

where $\underline{\alpha} \coloneqq \min\left\{\bar{\alpha}, \frac{2c_2(1-c_1)}{L}\right\}$.

Theorem 8. Let $f_{\text{inf}} = \inf_x f(x)$. If f is L-smooth and satisfies μ-PL,

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

then GD $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ with backtracking line search satisfies

$$f(x_k) - f_{\text{inf}} \leq (1 - 2c_1\underline{\alpha}\mu)^k [f(x_0) - f_{\text{inf}}].$$

where $\underline{\alpha} \coloneqq \min\left\{\bar{\alpha}, \frac{2c_2(1-c_1)}{L}\right\}$.

6 (*) Converge to convex optimality

We proved sublinear convergence to first-order optimality. If f is also convex, then this implies convergence to global optimality. We will now prove that directly.

Theorem 9. If f is μ-strongly convex and L-smooth and $f(x^\star) = f_{\text{inf}}$, then N steps of gradient descent $x_{k+1} = x_k - \alpha \nabla f(x_k)$ with fixed step-size $\alpha = \frac{1}{L}$ satisfies

$$f(x_N) - f^\star \leq \min\left\{\left(1 - \frac{\mu}{L}\right)^N, \frac{1}{N}\right\} \cdot \frac{L}{2} \|x_0 - x^\star\|_2^2 .$$

Note that all gradient descent iterates remain in the sublevel set

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

Therefore, the above can be modified to *local convexity* within a level set. We avoid that complication to keep the discussion simple.

The proof amounts to the following lemma.

Lemma 10. Under the assumptions above, $x_+ = x - \frac{1}{L}\nabla f(x)$ satisfies

$$\|x_+ - x^\star\|_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|x - x^\star\|_2^2 - \frac{2}{L}\left[f(x_+) - f_{\text{inf}}\right]$$

Question 5. Assume the following for all $k \in \{0, 1, \ldots, N-1\}$

$$\|x_{k+1} - x^\star\|_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^\star\|_2^2 - \frac{2}{L} \left[f(x_{k+1}) - f_{\text{inf}}\right].$$

Why is the following true?

$$f(x_N) - f^\star \leq \frac{1}{N} \cdot \frac{L}{2} \|x_0 - x^\star\|_2^2$$

Question 6. Assume the following for all $k \in \{0, 1, \ldots, N-1\}$

$$\|x_{k+1} - x^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^\star\|^2 - \frac{2}{L} \left[f(x_{k+1}) - f_{\text{inf}}\right].$$

Why is the following true?

$$f(x_N) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^N \cdot \frac{L}{2} \|x_0 - x^\star\|^2$$

7 Summary

The fixed step $\alpha = 1/L$ gives a uniform guarantee, but it can be conservative for a particular objective. Backtracking line search instead starts with an optimistic step and reduces it only until the Armijo condition is satisfied:

$$f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - c_1 \alpha_k \|\nabla f(x_k)\|^2$$

For an L-smooth objective, sufficiently small steps always satisfy Armijo. Consequently, backtracking terminates after finitely many trials and the accepted steps satisfy a uniform lower bound $\alpha_k \geq \underline{\alpha} > 0$.

This sufficient decrease gives the same basic convergence pattern as fixed-step gradient descent:

- L-smooth and bounded below $\implies$ $O(1/\sqrt{N})$ convergence of the best gradient norm;
- L-smooth and μ-PL $\implies$ linear convergence of the objective gap.

The proof strategy is the same in both cases: establish a one-step decrease, obtain a uniform lower bound on the step-size, and telescope. Additional structure such as PL or convexity then strengthens the conclusion.