

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

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

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

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

Proof: Substitute $x_+ = x - \alpha \nabla f(x)$ into the quadratic upper-bound

$$f(x_+) \leq f(x) + \nabla f(x)^T (x_+ - x) + \frac{L}{2} \|x_+ - x\|^2$$

and group terms. □

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

Solution: It maximizes the worst-case decrement factor $\alpha \left(1 - \frac{L\alpha}{2}\right)$. Indeed,

$$\begin{aligned} \frac{d}{d\alpha} \left[\alpha \left(1 - \frac{L\alpha}{2}\right) \right] &= 1 - \frac{L\alpha}{2} + \alpha \left(-\frac{L}{2}\right) = 1 - L\alpha \\ \frac{d^2}{d\alpha^2} \left[\alpha \left(1 - \frac{L\alpha}{2}\right) \right] &= -L \end{aligned}$$

so the α -maximization problem is concave and attains its maximum at the zero-derivative point.

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

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

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 $N - 1$:

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

On the other hand

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

Rearranging obtains the desired claim.

□

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

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

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

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

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

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

□

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?

Solution: Actually $\alpha = \frac{1}{L}$ is the best step that minimizes *the descent-lemma worst-case upper-bound*. For the actual objective, larger steps can be much better. Fundamental gap is **worst-case vs average-case**. There *exists* a worst-case L -smooth function for which $\alpha = \frac{1}{L}$ is provably the best. But our L -smooth function probably isn't the worst case.

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

Solution: If both $f(x)$ and $\nabla f(x)$ cost the same, then better to do another step than to polish a single step. “Perfect is enemy of good.”

If $\nabla f(x)$ is much more expensive than $f(x)$, then better to reduce gradient calls. “Make most out of every gradient eval.”

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*?

Solution: Allows stagnation with arbitrarily small step-size $\alpha \rightarrow 0$.

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 $f'(x) = 0$?

Solution: Since $f'(x) = 2x$,

$$x_{k+1} = \left(1 - \frac{1}{(k+2)^2}\right) x_k.$$

So indeed $0 < x_{k+1} < x_k$ and $f(x_{k+1}) < f(x_k)$. However,

$$x_N = \prod_{j=2}^{N+1} \frac{(j-1)(j+1)}{j^2} = \frac{N+2}{2(N+1)} \longrightarrow \frac{1}{2}.$$

The iterates converge to $x = 1/2$, but this point is not stationary because $|f'(x_N)| = 2|x_N| \longrightarrow 1$.

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:

Solution:

$$f(x + \alpha v) \leq f(x) + c_1 \alpha \nabla f(x)^T v.$$

Our previous strict descent condition forces $c_1 \rightarrow 0^+$. 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.

Proof: Apply the quadratic upper-bound with $x_+ = x + \alpha v$:

$$f(x + \alpha v) \leq f(x) + \nabla f(x)^T(\alpha v) + \frac{L}{2} \|\alpha v\|_2^2.$$

The assumed stepsize bound gives

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

Substitute this inequality into the descent lemma:

$$\begin{aligned} f(x + \alpha v) &\leq f(x) + \alpha \nabla f(x)^T v - \alpha(1 - c_1)\nabla f(x)^T v \\ &= f(x) + c_1 \alpha \nabla f(x)^T v. \end{aligned}$$

This is exactly 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

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

Proof: If $\bar{\alpha} \leq \frac{2(1-c_1)}{L}$, then the first trial is guaranteed to be accepted. Otherwise, at acceptance, the previous trial α/c_2 was rejected, so

$$\frac{\alpha}{c_2} > \frac{2(1-c_1)}{L} \iff \alpha > \frac{2c_2(1-c_1)}{L}.$$

Now, at acceptance, we have performed at most j trials so that

$$c_2^j \bar{\alpha} \leq \frac{2(1-c_1)}{L}.$$

This at the j -th trial where

$$j = \left\lceil \log_{1/c_2} \left(\frac{\bar{\alpha}L}{2(1-c_1)} \right) \right\rceil.$$

Finally, add the initial trial with $j = 0$. □

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

Proof: Recall that backtracking line search terminates with a step-size of $\alpha_k \geq \underline{\alpha}$. The associated Armijo condition reads

$$f(x_{k+1}) \leq f(x_k) - c_1 \underline{\alpha} \|\nabla f(x_k)\|_2^2.$$

Rearranging yields

$$c_1 \underline{\alpha} \|\nabla f(x_k)\|_2^2 \leq f(x_k) - f(x_{k+1}).$$

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

$$\begin{aligned} c_1 \underline{\alpha} \sum_{k=0}^{N-1} \|\nabla f(x_k)\|_2^2 &\leq [f(x_0) - f(x_1)] + [f(x_1) - f(x_2)] \\ &\quad + \dots + [f(x_{N-1}) - f(x_N)] \\ &= f(x_0) - f(x_N) \leq f(x_0) - f_{\inf}. \end{aligned}$$

On the other hand

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

Rearranging obtains the desired claim. $\square$

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

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

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

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

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

Proof: Recall that backtracking line search terminates with a step-size of $\alpha_k \geq \underline{\alpha}$. The associated Armijo condition reads

$$f(x_{k+1}) \leq f(x_k) - c_1 \underline{\alpha} \|\nabla f(x_k)\|_2^2.$$

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

$$\begin{aligned} f(x_{k+1}) - f_{\inf} &\leq f(x_k) - f_{\inf} - c_1 \underline{\alpha} \|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) - f_{\inf} - 2c_1 \underline{\alpha} \mu (f(x_k) - f_{\inf}) \\ &= (1 - 2c_1 \underline{\alpha} \mu) (f(x_k) - f_{\inf}). \end{aligned}$$

□

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 $f(x^\star) = f_{\inf}$, 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

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

Proof: The update equations give

$$\begin{aligned} x_+ - x^\star &= x - x^\star - \frac{1}{L} \nabla f(x) \\ \|x_+ - x^\star\|^2 &= \|x - x^\star\|^2 - \frac{2}{L} \nabla f(x)^T (x - x^\star) + \frac{1}{L^2} \|\nabla f(x)\|^2. \end{aligned} \tag{A}$$

The first-order characterization of μ -strong convexity gives

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

which rearranges to

$$-\frac{2}{L} \nabla f(x)^T (x - x^*) \leq \frac{2}{L} [f_{\inf} - f(x)] - \frac{\mu}{L} \|x - x^*\|^2. \quad (\text{B})$$

Meanwhile, the optimized descent lemma reads

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

which rearranges into

$$\frac{1}{L^2} \|\nabla f(x)\|_2^2 \leq \frac{2}{L} [f(x) - f(x_+)]. \quad (\text{C})$$

Adding (B) and (C), and then substituting into (A) yields the claim. $\square$

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

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

Why is the following true?

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

Solution: Upper-bound $1 - \frac{\mu}{L} \leq 1$ and rearrange to yield

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

Summing gives

$$\begin{aligned} \frac{2}{L} \sum_{k=0}^{N-1} [f(x_{k+1}) - f^*] &\leq \sum_{k=0}^{N-1} \left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \right) \\ &= \|x_0 - x^*\|_2^2 \end{aligned}$$

Since $f(x_0) \geq f(x_1) \geq \dots \geq f(x_N)$, we further have

$$\frac{2}{L} \sum_{k=0}^{N-1} [f(x_{k+1}) - f^*] \geq \frac{2}{L} \sum_{k=0}^{N-1} [f(x_N) - f^*] = \frac{2N}{L} [f(x_N) - f^*].$$

Rearrange to finish the proof.

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

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

Why is the following true?

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

Solution: Upper-bound $-\frac{2}{L} [f(x_{k+1}) - f_{\inf}] \leq 0$ and rearrange to yield

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^*\|^2$$

Iterating yields

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

Finally, we use the L -smooth quadratic upper-bound

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2$$

so

$$\begin{aligned} f(x_N) &\leq f(x^*) + \nabla f(x^*)^T (x_N - x^*) + \frac{L}{2} \|x_N - x^*\|^2 \\ &= f(x^*) + 0 + \frac{L}{2} \|x_N - x^*\|^2 \\ &\leq f(x^*) + 0 + \frac{L}{2} \left(1 - \frac{\mu}{L}\right)^N \|x_0 - x^*\|^2 \end{aligned}$$

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.