Statistical Optimization: Lecture 7

Gradient Descent: Theoretical Analysis

Zijian Guo

Zhejiang University
Center for Data Science

April 1, 2026

Gradient descent

We consider the optimization problem

$$\min_{\theta \in \mathbb{R}^d} f(\theta).$$

For a small step $v \in \mathbb{R}^d$, the first-order Taylor expansion gives

$$f(\theta + v) \approx f(\theta) + \nabla f(\theta)^\top v.$$

So to decrease f, we want to choose v such that

$$\nabla f(\theta)^\top v < 0.$$

Gradient descent (GD) starts from an initial point $\theta_0 \in \mathbb{R}^d$, and updates by

$$\theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t), \qquad t \geq 0.$$

Here we choose a fixed stepsize $\gamma > 0$ for theoretical analysis. One can also consider a varying stepsize sequence $(\gamma_t)_{t \geq 0}$.

Objective

In this lecture, we study whether gradient descent converges and how fast it converges under three assumptions:

- convexity + Lipschitz;
- convexity + smoothness;
- strong convexity+ smoothness.

We will analyze two notions of convergence:

- **parameter convergence:** how θ_t approaches $\theta^\star$,
- **function-value convergence:** how $f(\theta_t)$ approaches $f(\theta^\star)$.

Outline

Vanilla analysis

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ **steps**

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ **steps**

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ **steps**

Vanilla analysis

Let $(\theta_t)_{t\geq 0}$ be generated by gradient descent

$$\theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t),$$

and let $\theta^\star$ be a minimizer of a convex function f. Define

$$g_t := \nabla f(\theta_t).$$

Then for every integer $T \geq 1$, we have

$$\sum_{t=0}^{T-1} \big(f(\theta_t) - f(\theta^\star)\big) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2\gamma} \|\theta_0 - \theta^\star\|^2.$$

Proof of vanilla analysis

Proof. Since f is convex, we have

$$f(\theta^\star) \geq f(\theta_t) + \nabla f(\theta_t)^\top (\theta^\star - \theta_t).$$

Rearranging and use $g_t := \nabla f(\theta_t)$, gives

$$f(\theta_t) - f(\theta^\star) \leq g_t^\top (\theta_t - \theta^\star).$$

So it suffices to bound

$$g_t^\top (\theta_t - \theta^\star).$$

Proof of vanilla analysis

From the gradient descent update, $\theta_{t+1} = \theta_t - \gamma g_t$, hence

$$g_t = \frac{\theta_t - \theta_{t+1}}{\gamma}.$$

Therefore,

$$g_t^\top (\theta_t - \theta^\star) = \frac{1}{\gamma}(\theta_t - \theta_{t+1})^\top (\theta_t - \theta^\star).$$

Now apply the identity

$$2v^\top w = \|v\|^2 + \|w\|^2 - \|v - w\|^2$$

with

$$v = \theta_t - \theta_{t+1}, \qquad w = \theta_t - \theta^\star.$$

We obtain

$$g_t^\top (\theta_t - \theta^\star) = \frac{1}{2\gamma}\Big(\|\theta_t - \theta_{t+1}\|^2 + \|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2\Big).$$

Proof of vanilla analysis

Since $\theta_t - \theta_{t+1} = \gamma g_t$, this becomes

$$g_t^\top(\theta_t - \theta^\star) = \frac{\gamma}{2}\|g_t\|^2 + \frac{1}{2\gamma}\Big(\|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2\Big).$$

Summing over $t = 0, \ldots, T-1$, the right side second sum telescopes, so we get

$$\sum_{t=0}^{T-1} g_t^\top(\theta_t - \theta^\star) = \frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma}\Big(\|\theta_0 - \theta^\star\|^2 - \|\theta_T - \theta^\star\|^2\Big)$$

and hence

$$\sum_{t=0}^{T-1} g_t^\top(\theta_t - \theta^\star) \le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma}\|\theta_0 - \theta^\star\|^2.$$

Proof of vanilla analysis

Finally, using

$$f(\theta_t) - f(\theta^\star) \leq g_t^\top (\theta_t - \theta^\star)$$

for each t, summing from $t = 0$ to $T - 1$ gives

$$\sum_{t=0}^{T-1} \big(f(\theta_t) - f(\theta^\star)\big) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2\gamma} \|\theta_0 - \theta^\star\|^2.$$

This completes the proof. □

What the vanilla analysis tells us

- Question: Is this result strong? In general, the answer is no.
- Limitations of the bound:
  - Dependence on $\|\theta_0 - \theta^\star\|$ is expected (farther start $\rightarrow$ longer time).
  - The dependence on the squared gradients $\sum_{t=0}^{T-1} \|g_t\|^2$ is more of an issue, and if we cannot control them, we cannot say much.
- Interpretation:
  - The result provides a guarantee but is not yet very informative.
  - Stronger assumptions (e.g., smoothness) will be needed for sharper convergence rates.

Outline

Vanilla analysis

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Lipschitz convex functions

Definition. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called *B-Lipschitz* if

$$|f(\theta) - f(\eta)| \leq B\|\theta - \eta\|, \qquad \forall \theta, \eta \in \mathbb{R}^d.$$

If f is differentiable, this is equivalent to

$$\|\nabla f(\theta)\| \leq B, \qquad \forall \theta \in \mathbb{R}^d.$$

Remark. This is a rather restrictive assumption. For example, $f(\theta) = \theta^2$ is not Lipschitz on $\mathbb{R}$, since $\nabla f(\theta) = 2\theta$ is unbounded.

Average error bound with Lipschitz convex functions

Theorem (average error bound with Lipschitz convex functions)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, and let $\theta^\star$ be a global minimizer. Assume that

$$\|\theta_0 - \theta^\star\| \le R \qquad \text{and} \qquad \|\nabla f(\theta)\| \le B \quad \text{for all } \theta.$$

Choose the stepsize

$$\gamma := \frac{R}{B\sqrt{T}}.$$

Then gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(f(\theta_t) - f(\theta^\star)\big) \le \frac{RB}{\sqrt{T}}.$$

Proof of Theorem

Proof. From the vanilla analysis,

$$\sum_{t=0}^{T-1}\big(f(\theta_t)-f(\theta^\star)\big)\le\frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2+\frac{1}{2\gamma}\|\theta_0-\theta^\star\|^2.$$

Using the assumptions

$$\|\theta_0-\theta^\star\|\le R,\qquad \|g_t\|=\|\nabla f(\theta_t)\|\le B,$$

we obtain

$$\sum_{t=0}^{T-1}\big(f(\theta_t)-f(\theta^\star)\big)\le\frac{\gamma}{2}B^2T+\frac{R^2}{2\gamma}.$$

Now define

$$q(\gamma):=\frac{\gamma}{2}B^2T+\frac{R^2}{2\gamma}.$$

Proof of Theorem

To get the best bound, we minimize $q(\gamma)$ over $\gamma > 0$. Setting $q'(\gamma) = 0$ gives

$$\gamma = \frac{R}{B\sqrt{T}}.$$

Substituting this choice back yields

$$\sum_{t=0}^{T-1} \big(f(\theta_t) - f(\theta^\star)\big) \leq RB\sqrt{T}.$$

Dividing both sides by T, we get

$$\frac{1}{T}\sum_{t=0}^{T-1} \big(f(\theta_t) - f(\theta^\star)\big) \leq \frac{RB}{\sqrt{T}}.$$

□

Theorem result analysis

- To guarantee

$$\min_{t=0,\dots,T-1} \big(f(\theta_t) - f(\theta^\star)\big) \leq \varepsilon,$$

  gradient descent requires at least

$$T \geq \frac{R^2 B^2}{\varepsilon^2}$$

  iterations. The required number of iterations can be extremely large.
- On the positive side, the iteration complexity does not depend on the dimension d of the space, which is crucial for high-dimensional optimization problems.
- Although R (initial distance to optimum) and B (gradient bound) may depend on d, in many practical cases this dependence is mild.

Outline

Vanilla analysis

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Smooth convex functions

We now introduce a condition that gives an *upper bound* on the function value by a quadratic approximation.

Definition Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be differentiable, let $\Theta \subseteq \mathrm{dom}(f)$ be convex, and let $L > 0$. We say that f is *L-smooth over* Θ if

$$f(\eta) \leq f(\theta) + \nabla f(\theta)^\top (\eta - \theta) + \frac{L}{2}\|\eta - \theta\|^2, \qquad \forall \theta, \eta \in \Theta.$$

If $\Theta = \mathrm{dom}(f)$, we simply say that f is *L-smooth*.

Smooth convex functions

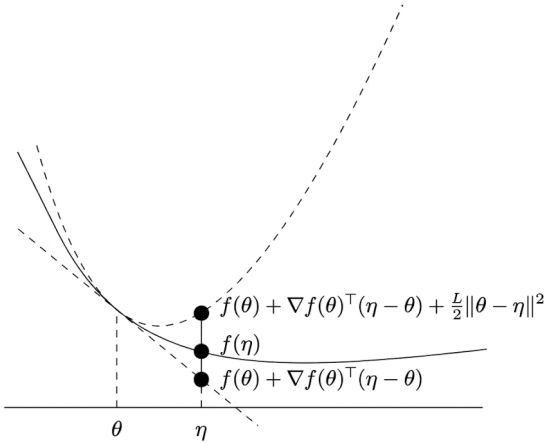

Figure: A smooth convex function

Sufficient decrease under smoothness

Lemma (Sufficient decrease lemma)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and L-smooth. With stepsize

$$\gamma = \frac{1}{L},$$

gradient descent satisfies

$$f(\theta_{t+1}) \leq f(\theta_t) - \frac{1}{2L}\|\nabla f(\theta_t)\|^2, \qquad t \geq 0.$$

Remark. This does NOT require convexity; smoothness alone is enough.

Proof of Lemma

Proof. From L-smoothness and the update rule

$$\theta_{t+1} = \theta_t - \frac{1}{L}\nabla f(\theta_t),$$

we obtain

$$\begin{aligned} f(\theta_{t+1}) &\leq f(\theta_t) + \nabla f(\theta_t)^\top(\theta_{t+1} - \theta_t) + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \\ &= f(\theta_t) - \frac{1}{L}\|\nabla f(\theta_t)\|^2 + \frac{1}{2L}\|\nabla f(\theta_t)\|^2 \\ &= f(\theta_t) - \frac{1}{2L}\|\nabla f(\theta_t)\|^2. \end{aligned}$$

□

Convergence with smooth convex functions

Theorem (Convergence with smooth convex functions)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and let $\theta^\star$ be a global minimizer. Assume that f is smooth with parameter L. Choose the stepsize

$$\gamma = \frac{1}{L}.$$

Then gradient descent yields

$$f(\theta_T) - f(\theta^\star) \leq \frac{L}{2T}\|\theta_0 - \theta^\star\|^2, \qquad T \geq 1.$$

Corollary. If we write $R^2 := \|\theta_0 - \theta^\star\|^2$, then it suffices to take $T \geq \frac{LR^2}{2\varepsilon}$ to guarantee $f(\theta_T) - f(\theta^\star) \leq \varepsilon$.

Proof of Theorem

Proof. First by sufficient decrease lemma,

$$f(\theta_{t+1}) \leq f(\theta_t) - \frac{1}{2L}\|\nabla f(\theta_t)\|^2.$$

Summing over $t = 0, \ldots, T-1$, we get

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\theta_t)\|^2 \leq \sum_{t=0}^{T-1}\big(f(\theta_t) - f(\theta_{t+1})\big) = f(\theta_0) - f(\theta_T). \tag{7.1}$$

Besides, using $\gamma = \frac{1}{L}$ in the vanilla analysis, we get

$$\sum_{t=0}^{T-1}\big(f(\theta_t) - f(\theta^\star)\big) \leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\theta_t)\|^2 + \frac{L}{2}\|\theta_0 - \theta^\star\|^2.$$

Proof of Theorem

Using (7.1) to bound the gradient sum in the vanilla analysis, we obtain

$$\sum_{t=0}^{T-1}\big(f(\theta_t)-f(\theta^\star)\big)\le f(\theta_0)-f(\theta_T)+\frac{L}{2}\|\theta_0-\theta^\star\|^2.$$

Rearranging gives

$$\sum_{t=1}^{T}\big(f(\theta_t)-f(\theta^\star)\big)\le \frac{L}{2}\|\theta_0-\theta^\star\|^2.$$

Since $f(\theta_{t+1})\le f(\theta_t)$, we have

$$f(\theta_T)-f(\theta^\star)\le \frac{1}{T}\sum_{t=1}^{T}\big(f(\theta_t)-f(\theta^\star)\big)\le \frac{L}{2T}\|\theta_0-\theta^\star\|^2.$$

□

Equivalence of smoothness

Lemma (equivalence of smoothness)

Suppose that $\mathrm{dom}(f)$ *is open and convex, and that* $f : \mathrm{dom}(f) \to \mathbb{R}$ *is differentiable. Let* $L > 0$*. Then the following are equivalent:*

(i) *f is L-smooth.*

(ii) *The function*

$$g(\theta) := \frac{L}{2}\theta^\top\theta - f(\theta)$$

is convex on $\mathrm{dom}(g) := \mathrm{dom}(f)$.

Proof of lemma (i) $\rightarrow$ (ii)

Proof of (i) $\rightarrow$ (ii). Let $\theta, \eta \in \mathrm{dom}(f)$. Then

$$\begin{aligned} g(\eta) - g(\theta) - \nabla g(\theta)^\top(\eta - \theta) &= \frac{L}{2}\|\eta\|^2 - f(\eta) - \left(\frac{L}{2}\|\theta\|^2 - f(\theta)\right) - (L\theta - \nabla f(\theta))^\top(\eta - \theta) \\ &= \frac{L}{2}\left(\|\eta\|^2 - \|\theta\|^2 - 2\theta^\top(\eta - \theta)\right) - \left(f(\eta) - f(\theta) - \nabla f(\theta)^\top(\eta - \theta)\right) \\ &= \frac{L}{2}\|\eta - \theta\|^2 - \left(f(\eta) - f(\theta) - \nabla f(\theta)^\top(\eta - \theta)\right). \end{aligned}$$

By L-smoothness of f,

$$f(\eta) - f(\theta) - \nabla f(\theta)^\top(\eta - \theta) \leq \frac{L}{2}\|\eta - \theta\|^2,$$

hence

$$g(\eta) - g(\theta) - \nabla g(\theta)^\top(\eta - \theta) \geq 0 \rightarrow g(\eta) \geq g(\theta) + \nabla g(\theta)^\top(\eta - \theta), \qquad \forall \theta, \eta,$$

so g is convex. □

Proof of lemma (ii) $\rightarrow$ (i)

Proof of (ii) $\rightarrow$ (i). Assume that g is convex. Then for all $\theta, \eta \in \mathrm{dom}(f)$,

$$g(\eta) \geq g(\theta) + \nabla g(\theta)^\top (\eta - \theta).$$

That is,

$$\frac{L}{2}\|\eta\|^2 - f(\eta) \geq \frac{L}{2}\|\theta\|^2 - f(\theta) + \big(L\theta - \nabla f(\theta)\big)^\top (\eta - \theta).$$

Rearranging gives

$$f(\eta) \leq f(\theta) + \nabla f(\theta)^\top (\eta - \theta) + \frac{L}{2}\|\eta - \theta\|^2.$$

Thus f is L-smooth.

□

Quadratic functions are smooth

Lemma (Quadratic functions are smooth)

Let

$$f(\theta) = \theta^\top Q\theta + b^\top \theta + c,$$

where $Q \in \mathbb{R}^{d\times d}$ is symmetric, $b \in \mathbb{R}^d$, and $c \in \mathbb{R}$. Then f is smooth with parameter

$$L = 2\|Q\|,$$

where $\|Q\|$ is the spectral norm of Q.

Proof of lemma

Proof. Since

$$\nabla f(\theta) = 2Q\theta + b,$$

we have

$$\begin{aligned} f(\eta) - f(\theta) - \nabla f(\theta)^\top (\eta - \theta) &= (\eta^\top Q\eta - \theta^\top Q\theta) - (2Q\theta + b)^\top (\eta - \theta) \\ &= (\eta - \theta)^\top Q(\eta - \theta) \\ &\leq \|Q\| \, \|\eta - \theta\|^2. \end{aligned}$$

Therefore,

$$f(\eta) \leq f(\theta) + \nabla f(\theta)^\top (\eta - \theta) + \frac{L}{2}\|\eta - \theta\|^2$$

with $L = 2\|Q\|$.

□

Equivalence of L-smooth

Lemma (equivalence of L-smooth)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Then the following are equivalent:

(i) *f is smooth with parameter L.*

(ii) $\|\nabla f(\theta) - \nabla f(\eta)\| \leq L\|\theta - \eta\|, \qquad \forall \theta, \eta \in \mathbb{R}^d.$

Proof of Lemma (i) $\rightarrow$ (ii)

Proof of (i) $\rightarrow$ (ii).

Step 1. Assume that f is L-smooth. For any fixed θ, define

$$f_\theta(z) := f(z) - \nabla f(\theta)^\top z,$$

and similarly, for any fixed η, define

$$f_\eta(z) := f(z) - \nabla f(\eta)^\top z.$$

It is easy to check that f_θ and f_η are still convex and L-smooth, and

$$\nabla f_\theta(\theta) = \nabla f(\theta) - \nabla f(\theta) = 0,$$

so $\theta \in \arg\min_z f_\theta(z)$. Similarly,

$$\eta \in \arg\min_z f_\eta(z).$$

Proof of Lemma (i) $\rightarrow$ (ii)

Step 2. For any L-smooth function g and any u, v, we have

$$g(v) \leq g(u) + \nabla g(u)^\top (v - u) + \frac{L}{2}\|v - u\|^2.$$

By setting $v = u - \frac{1}{L}\nabla g(u)$ in L-smooth condition, we get

$$g\Big(u - \frac{1}{L}\nabla g(u)\Big) \leq g(u) - \frac{1}{2L}\|\nabla g(u)\|^2.$$

Apply this to $g = f_\theta$ at $u = \eta$. Since θ minimizes f_θ,

$$f_\theta(\theta) \leq f_\theta\Big(\eta - \frac{1}{L}\nabla f_\theta(\eta)\Big) \leq f_\theta(\eta) - \frac{1}{2L}\|\nabla f_\theta(\eta)\|^2.$$

Hence

$$f_\theta(\eta) - f_\theta(\theta) \geq \frac{1}{2L}\|\nabla f_\theta(\eta)\|^2.$$

Proof of Lemma (i) $\rightarrow$ (ii)

Step 3. Now

$$f_\theta(\eta) - f_\theta(\theta) = f(\eta) - f(\theta) - \nabla f(\theta)^\top(\eta - \theta),$$

and

$$\nabla f_\theta(\eta) = \nabla f(\eta) - \nabla f(\theta).$$

Therefore,

$$f(\eta) - f(\theta) - \nabla f(\theta)^\top(\eta - \theta) \geq \frac{1}{2L}\|\nabla f(\eta) - \nabla f(\theta)\|^2.$$

Swapping θ and η, we also get

$$f(\theta) - f(\eta) - \nabla f(\eta)^\top(\theta - \eta) \geq \frac{1}{2L}\|\nabla f(\theta) - \nabla f(\eta)\|^2.$$

Adding the two inequalities yields

$$(\nabla f(\theta) - \nabla f(\eta))^\top(\theta - \eta) \geq \frac{1}{L}\|\nabla f(\theta) - \nabla f(\eta)\|^2.$$

Proof of Lemma (i) $\rightarrow$ (ii)

Step 4. By Cauchy–Schwarz,

$$\|\nabla f(\theta) - \nabla f(\eta)\|^2 \leq L(\nabla f(\theta) - \nabla f(\eta))^\top(\theta - \eta) \leq L\|\nabla f(\theta) - \nabla f(\eta)\|\,\|\theta - \eta\|.$$

If $\nabla f(\theta) \neq \nabla f(\eta)$, divide both sides by $\|\nabla f(\theta) - \nabla f(\eta)\|$ to obtain

$$\|\nabla f(\theta) - \nabla f(\eta)\| \leq L\|\theta - \eta\|.$$

This proves (ii). $\square$

Proof of lemma (ii) $\to$ (i)

Proof of (ii) $\to$ (i). Define $g(t) := f\big(t(\theta-\eta)+\eta\big)$ for $t \in [0,1]$.

$$
\begin{aligned}
f(\theta) - f(\eta) - \nabla f(\eta)^\top(\theta-\eta) &= g(1) - g(0) - \nabla f(\eta)^\top(\theta-\eta) \\
&= \int_0^1 g'(t)\,dt - \nabla f(\eta)^\top(\theta-\eta) \\
&= \int_0^1 \big(\nabla f(t(\theta-\eta)+\eta) - \nabla f(\eta)\big)^\top(\theta-\eta)\,dt \\
&\le \int_0^1 \|\nabla f(t(\theta-\eta)+\eta) - \nabla f(\eta)\|\,\|\theta-\eta\|\,dt \\
&\le \int_0^1 Lt\|\theta-\eta\|^2\,dt \\
&= \frac{L}{2}\|\theta-\eta\|^2.
\end{aligned}
$$

Proof of lemma (ii) $\rightarrow$ (i)

Hence

$$f(\theta) - f(\eta) - \nabla f(\eta)^\top (\theta - \eta) \leq \frac{L}{2} \|\theta - \eta\|^2.$$

Rearranging gives

$$f(\theta) \leq f(\eta) + \nabla f(\eta)^\top (\theta - \eta) + \frac{L}{2} \|\theta - \eta\|^2.$$

This is exactly the definition of L-smoothness.

□

Properties of smoothness

Lemma (Properties of smoothness)

(i) *If $f_1, \ldots, f_m$ are smooth with parameters $L_1, \ldots, L_m$ and $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$, then*

$$f := \sum_{i=1}^{m} \lambda_i f_i$$

is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$ on $\operatorname{dom}(f) = \bigcap_{i=1}^m \operatorname{dom}(f_i)$.

(ii) *If $f : \operatorname{dom}(f) \to \mathbb{R}$ is smooth with parameter L and $g(\theta) = A\theta + b$ is affine, then $f \circ g$ is smooth with parameter $L\|A\|^2$ on*

$$\operatorname{dom}(f \circ g) = \{\theta \in \mathbb{R}^m : \ g(\theta) \in \operatorname{dom}(f)\}.$$

Proof of Lemma (i)

Proof of (i). For each i, since f_i is L_i-smooth,

$$f_i(\eta) \le f_i(\theta) + \nabla f_i(\theta)^\top (\eta - \theta) + \frac{L_i}{2}\|\eta - \theta\|^2.$$

Multiply by $\lambda_i \ge 0$ and sum over $i = 1, \dots, m$. Then

$$f(\eta) \le f(\theta) + \nabla f(\theta)^\top (\eta - \theta) + \frac{1}{2}\Big(\sum_{i=1}^{m} \lambda_i L_i\Big)\|\eta - \theta\|^2.$$

Hence f is smooth with parameter $\sum_{i=1}^{m} \lambda_i L_i$. □

Proof of Lemma (ii)

Proof of (ii). Let

$$h(\theta) := f(A\theta + b).$$

For any θ, η, define

$$u := A\theta + b, \qquad v := A\eta + b.$$

By L-smoothness of f,

$$f(v) \le f(u) + \nabla f(u)^\top (v - u) + \frac{L}{2}\|v - u\|^2.$$

Since

$$v - u = A(\eta - \theta),$$

we get

$$f(A\eta + b) \le f(A\theta + b) + \nabla f(A\theta + b)^\top A(\eta - \theta) + \frac{L}{2}\|A(\eta - \theta)\|^2.$$

Proof of Lemma (ii)

Also,

$$\nabla h(\theta) = A^\top \nabla f(A\theta + b),$$

and

$$\|A(\eta - \theta)\| \leq \|A\| \, \|\eta - \theta\|.$$

Therefore,

$$h(\eta) \leq h(\theta) + \nabla h(\theta)^\top (\eta - \theta) + \frac{L\|A\|^2}{2} \|\eta - \theta\|^2.$$

So h is smooth with parameter $L\|A\|^2$.

□

Outline

Vanilla analysis

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Theorem (Convergence with smooth and strongly convex functions)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Assume that f is L-smooth and strongly convex with parameter $\mu > 0$. Choose

$$\gamma = \frac{1}{L}.$$

Then gradient descent with arbitrary initial point θ_0 satisfies:

(i) ***Geometric decay of squared distance:***

$$\|\theta_{t+1} - \theta^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|\theta_t - \theta^\star\|^2, \qquad t \geq 0.$$

(ii) ***Exponential decay of function error:***

$$f(\theta_T) - f(\theta^\star) \leq \frac{L}{2} \left(1 - \frac{\mu}{L}\right)^T \|\theta_0 - \theta^\star\|^2, \qquad T \geq 1.$$

Proof of Theorem (i)

Proof. We start from the vanilla analysis that

$$\nabla f(\theta_t)^\top(\theta_t - \theta^\star) = \frac{\gamma}{2}\|\nabla f(\theta_t)\|^2 + \frac{1}{2\gamma}\Big(\|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2\Big). \tag{7.2}$$

By strong convexity, with $\theta = \theta_t$ and $\eta = \theta^\star$,

$$f(\theta^\star) \geq f(\theta_t) + \nabla f(\theta_t)^\top(\theta^\star - \theta_t) + \frac{\mu}{2}\|\theta_t - \theta^\star\|^2.$$

Rearranging gives

$$\nabla f(\theta_t)^\top(\theta_t - \theta^\star) \geq f(\theta_t) - f(\theta^\star) + \frac{\mu}{2}\|\theta_t - \theta^\star\|^2.$$

Proof of Theorem (i)

Combining this with (7.2), we obtain

$$\frac{1}{2\gamma}\Big(\|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2\Big) \geq f(\theta_t) - f(\theta^\star) + \frac{\mu}{2}\|\theta_t - \theta^\star\|^2 - \frac{\gamma}{2}\|\nabla f(\theta_t)\|^2.$$

Rearranging,

$$\|\theta_{t+1} - \theta^\star\|^2 \leq 2\gamma\big(f(\theta^\star) - f(\theta_t)\big) + \gamma^2\|\nabla f(\theta_t)\|^2 + (1 - \mu\gamma)\|\theta_t - \theta^\star\|^2. \tag{7.3}$$

By sufficient decrease lemma, we know

$$f(\theta^\star) - f(\theta_t) \leq f(\theta_{t+1}) - f(\theta_t) \leq -\frac{1}{2L}\|\nabla f(\theta_t)\|^2.$$

Proof of Theorem (i)

Using $\gamma = \frac{1}{L}$, combining above inequality with (7.3) yields

$$\|\theta_{t+1} - \theta^\star\|^2 \leq (1 - \mu\gamma)\|\theta_t - \theta^\star\|^2 = \left(1 - \frac{\mu}{L}\right)\|\theta_t - \theta^\star\|^2.$$

Iterating this bound gives

$$\|\theta_T - \theta^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|\theta_0 - \theta^\star\|^2.$$

This proves part (i).

Proof of Theorem (ii)

To prove part (ii), use smoothness at $\theta^\star$:

$$f(\theta_T) \leq f(\theta^\star) + \nabla f(\theta^\star)^\top (\theta_T - \theta^\star) + \frac{L}{2}\|\theta_T - \theta^\star\|^2.$$

Since $\theta^\star$ is a minimizer, we have

$$\nabla f(\theta^\star) = 0.$$

Therefore,

$$f(\theta_T) - f(\theta^\star) \leq \frac{L}{2}\|\theta_T - \theta^\star\|^2.$$

Using part (i), we obtain

$$f(\theta_T) - f(\theta^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\theta_0 - \theta^\star\|^2.$$

□

Summary

Gradient descent alone does not guarantee a sharp rate. The vanilla analysis only yields the cumulative bound

$$\sum_{t=0}^{T-1}\big(f(\theta_t)-f(\theta^\star)\big)\le\frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\theta_t)\|^2+\frac{1}{2\gamma}\|\theta_0-\theta^\star\|^2.$$

To obtain explicit convergence rates, we need additional assumptions on f.

	Lipschitz convex functions	smooth convex functions	smooth & strongly convex functions
gradient descent	$\mathcal{O}(1/\varepsilon^2)$	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(\log(1/\varepsilon))$