CHAPTER 1. GRADIENT DESCENT

Jinghua Huang Kun Yuan

February 29, 2024

1 Problem formulation

This chapter considers the following unconstrained problem

$$\min_{x\in\mathbb{R}^d} \quad f(x) \tag{1}$$

where $f(x)$ is a differentiable objective function.

Notation. We introduce the following notations:

- Let $x^\star := \arg\min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimial solution to problem (1).
- Let $f^\star := \min_{x\in\mathbb{R}^d}\{f(x)\}$ be the optimal function value.

2 Gradient descent

Given any arbitrary initialization variable x_0, gradient descent iterates as follows

$$x_{k+1} = x_k - \gamma\nabla f(x_k), \quad \forall\, k = 0, 1, 2, \cdots \tag{2}$$

where γ is the learning rate.

3 Convergence analysis

3.1 Smooth and non-convex problem

Assumption 3.1. We assume $f(x)$ is L-smooth, i.e., there exists a constant $L > 0$ such that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|_2^2, \quad \forall\, x, y \in \mathbb{R}^d. \tag{3}$$

Theorem 3.2. Under Assumption 3.1, if $\gamma = 1/L$, gradient descent converges as

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x_k)\|^2 \leq \frac{2L\big(f(x_0) - f^\star\big)}{K+1}. \tag{4}$$

Proof. Since $f(x)$ is L-smooth (Assumption 3.1), it holds that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\overset{(3)}{=} f(x_k) - \gamma\|\nabla f(x_k)\|^2 + \frac{\gamma^2 L}{2}\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - \frac{\gamma}{2}\|\nabla f(x_k)\|^2. \end{aligned} \tag{5}$$

where the last inequality holds when $\gamma \leq 1/L$. The above inequality can be rewritten as

$$\|\nabla f(x_k)\|^2 \leq \frac{2\big(f(x_k) - f(x_{k+1})\big)}{\gamma} \tag{6}$$

Averaging the above inequality over $k = 0, 1, \cdots, K$, we get

$$\frac{1}{K+1}\sum_{k=0}^{K-1}\|\nabla f(x_k)\|^2 \leq \frac{2\big(f(x_0) - f(x_K)\big)}{(K+1)\gamma} \leq \frac{2\big(f(x_0) - f^\star\big)}{(K+1)\gamma} = \frac{2L\big(f(x_0) - f^\star\big)}{K+1} \tag{7}$$

where the last equality holds when $\gamma = 1/L$. □

Iteration complexity implies the number of iterations that an algorithm requires to achieve an ϵ-accurate solution. Given the convergence rate of an algorithm, it is easy to derive its iteration complexity.

Corollary 3.3. Under Assumption 3.1, if $\gamma = 1/L$, the iteration complexity of gradient descent is $O(L/\epsilon)$.

Proof. According to Theorem 3.2, to guarantee gradient descent to converge to an ϵ-accurate solution, it is enough to let

$$\frac{2L\big(f(x_0) - f^\star\big)}{K+1} \leq \epsilon, \tag{8}$$

which implies that $K \geq \frac{2L(f(x_0)-f^\star)}{\epsilon} - 1$. In other words, gradient descent will converge to an ϵ-accurate solution with at most $\frac{2L(f(x_0)-f^\star)}{\epsilon} - 1$ iterations, which implies $K = O(L/\epsilon)$. $\square$

3.2 Smooth and convex problem

3.2.1 Supporting lemmas

Lemma 3.4. If $f(x)$ is L-smooth, i.e. $f(x)$ satisfies Assumption 3.1, it holds that

$$\|\nabla f(x)\|^2 \leq 2L\big(f(x) - f^\star\big), \quad \forall\, x \in \mathbb{R}^d \tag{9}$$

Please note that the above lemma holds without assuming convexity.

Proof. Let $y = x - \frac{1}{L}\nabla f(x)$ in (12), we have

$$\begin{aligned} f^\star &\leq f\big(x - \frac{1}{L}\nabla f(x)\big) \\ &\leq f(x) + \langle \nabla f(x), -\frac{1}{L}\nabla f(x)\rangle + \frac{1}{2L}\|\nabla f(x)\|^2 = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2, \end{aligned} \tag{10}$$

which concludes the proof. $\square$

We also recall the following lemma from Chapter 0.

Lemma 3.5. If $f(x)$ is convex, it holds that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle, \quad \forall\, x, y \in \mathbb{R}^d, \tag{11}$$

3.2.2 Convergence analysis

We make the following assumptions in the smooth and convex scenario.

Assumption 3.6. We assume $f(x)$ is convex and L-smooth, i.e., there exists a constant $L > 0$ such that

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y - x\|_2^2, \quad \forall\, x, y \in \mathbb{R}^d. \tag{12}$$

Theorem 3.7. Under Assumption 3.6, if $\gamma = 1/(2L)$, gradient descent converges as

$$f(x_K) - f^\star \leq \frac{2L\|x_0 - x^\star\|^2}{K+1}. \tag{13}$$

Proof. With GD recursion as in (2), we have

$$\begin{aligned}\|x_{k+1}-x^\star\|^2 &= \|x_k-x^\star-\gamma\nabla f(x_k)\|^2\\ &= \|x_k-x^\star\|^2-2\gamma\langle x_k-x^\star,\nabla f(x_k)\rangle+\gamma^2\|\nabla f(x_k)\|^2\\ &\overset{(a)}{\leq} \|x_k-x^\star\|^2-2\gamma\big(f(x_k)-f^\star\big)+2L\gamma^2(f(x_k)-f^\star)\\ &= \|x_k-x^\star\|^2-\frac{1}{2L}\big(f(x_k)-f^\star\big)\end{aligned}$$

where inequality (a) holds due to Lemmas 3.4 and 3.5, and the last equality holds when $\gamma = 1/(2L)$. We rewrite the above inequality as

$$f(x_k)-f^\star \leq 2L\left(\|x_k-x^\star\|^2-\|x_{k+1}-x^\star\|^2\right). \tag{14}$$

Averaging the above inequality over $k = 0, 1, \cdots, K$, we get

$$\frac{1}{K+1}\sum_{k=0}^{K} f(x_k)-f^\star \leq \frac{2L\|x_0-x^\star\|^2}{K+1}$$

Since $f(x_{k+1}) \leq f(x_k)$ (see (5)), we have $f(x_k) \geq f(x_k)$. Substituting this relation to the above inequality, we prove the final result. □

With similar proof arguments as in Corollary 3.3, we can achieve the iteration complexity of gradient descent in the smooth and convex setting.

Corollary 3.8. Under Assumption 3.6, if $\gamma = 1/(2L)$, the iteration complexity of gradient descent is $O(L/\epsilon)$.

3.3 Smooth and strongly-convex problem

Assumption 3.9 (μ-strongly convex)**.** We assume $f(x)$ is μ–strongly convex, i.e., there exists a constant $\mu > 0$ such that

$$f(y) \geq f(x)+\langle\nabla f(x), y-x\rangle+\frac{\mu}{2}\|y-x\|_2^2, \quad \forall\, x, y \in \mathbb{R}^d$$

The following lemma is from Chpater 0.

Lemma 3.10 (L-smooth μ-strongly convex property)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex, then for $\forall\, x, y \in \mathbb{R}^n$, the following inequality holds:

$$\langle\nabla f(y)-\nabla f(x), y-x\rangle \geq \frac{\mu L}{\mu+L}\|y-x\|_2^2+\frac{1}{\mu+L}\|\nabla f(y)-\nabla f(x)\|_2^2.$$

The following theorem establishes the convergence rate of GD in the smooth and strongly-convex scenario.

Theorem 3.11. Under Assumption 3.6 and Assumption 3.9, if $\gamma = 2/(L+\mu)$, gradient descent converges as

$$\|x_K - x^\star\| \leq \left(\frac{L-\mu}{L+\mu}\right)^{K+1} \|x_0 - x^\star\| \tag{15}$$

Proof. With GD recursion as in (2), we have

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star - \gamma \nabla f(x_k)\|^2 \\
&= \|x_k - x^\star\|^2 - 2\gamma \langle x_k - x^\star, \nabla f(x_k)\rangle + \gamma^2 \|\nabla f(x_k)\|^2 \\
&\overset{(a)}{=} \|x_k - x^\star\|^2 - 2\gamma \langle x_k - x^\star, \nabla f(x_k) - \nabla f(x^\star)\rangle + \gamma^2 \|\nabla f(x_k) - \nabla f(x^\star)\|^2 \\
&\overset{(b)}{\leq} \left(1 - \frac{2\gamma\mu L}{\mu + L}\right) \|x_k - x^\star\|^2 - \left(\frac{2\gamma}{\mu + L} - \gamma^2\right) \|\nabla f(x_k) - \nabla f(x^\star)\|^2 \\
&\overset{(c)}{=} \left(1 - \frac{4\mu L}{(\mu + L)^2}\right) \|x_k - x^\star\|^2 \\
&= \left(\frac{L-\mu}{\mu+L}\right)^2 \|x_k - x^\star\|^2
\end{aligned}$$

where equality (a) holds due to the optimality condition $\nabla f(x^\star) = 0$, inequality (b) holds due to Lemma 3.10, and (c) holds when $\gamma = \frac{2}{\mu+L}$. □

Corollary 3.12. Under Assumption 3.6 and Assumption 3.9, if $\gamma = 2/(L+\mu)$, the iteration complexity of gradient descent is $O((L/\mu)\log(1/\epsilon))$.

Proof. According to Theorem 3.11, we can get the iteration complexity by:

$$\|x_K - x^\star\| \leq \left(\frac{L+\mu}{L-\mu}\right)^{K+1} \|x_0 - x^\star\| \leq \epsilon,$$

from which we have

$$(K+1)\log\left(\frac{L+\mu}{L-\mu}\right) \geq \left(\log\left(\frac{1}{\epsilon}\right) - \log\left(\frac{1}{\|x_0 - x^\star\|}\right)\right). \tag{16}$$

Since $\log((L+\mu)/(L-\mu)) \approx 2\mu/(L-\mu)$ when $L \gg \mu$, we finally achieve the iteration complexity $K = O(\frac{L}{\mu}\log(1/\epsilon))$. □

4 Convergence rate summary

GD	Convergence Rate	Iteration complexity
Smooth non-convex	$\frac{2L\left(f(x_0)-f^\star\right)}{K}$	$O(\frac{L}{\epsilon})$
Smooth convex	$\frac{2L\|x_0-x^\star\|^2}{K}$	$O(\frac{L}{\epsilon})$
Smooth strongly-convex	$\left(\frac{L-\mu}{L+\mu}\right)^K \|x_0-x^\star\|$	$O(\frac{L}{\mu}\log\frac{1}{\epsilon})$

5 Experiments

We demonstrate the optimization process of gradient descent method in three scenarios.

5.1 Gradient descent on non-convex function

We minimize the non-convex and smooth optimization problem:

$$f(x) = \sin \pi x + 2x$$

We set the learning rate to $\gamma = \frac{1}{40}$, GD is as shown in the following figure1:

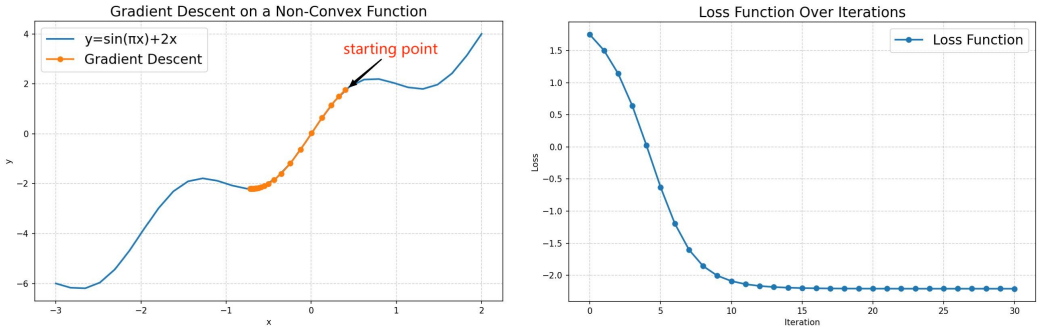

Figure 1: Gradient descent on non-convex function

5.2 Gradient descent on convex function

We minimize the convex and smooth optimization function:

$$f(x) = \begin{cases} x^2 + \frac{1}{2}x, & \text{if } x \leq -\frac{1}{2} \\ x^2 - \frac{1}{3}x, & \text{if } x \geq \frac{1}{3} \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to verify that $f(x)$ is L-smooth with $L = 2$. If we set $\gamma = \frac{1}{5L}$, GD converges as follows:

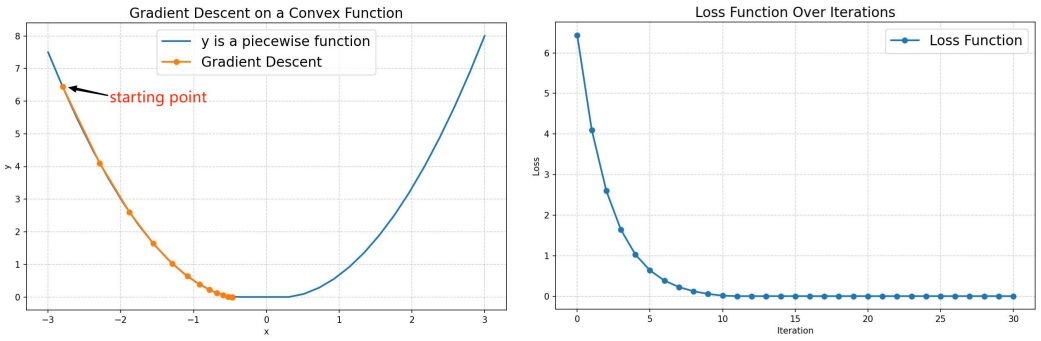

Figure 2: Gradient descent on convex function

If we set $\gamma = 1/L = 1/2$, GD converges very quickly as shown in the left plot in Figure 3. If we set $\gamma = \frac{2.06}{L}$, GD will oscillate a lot as shown in the right plot in Figure 3.

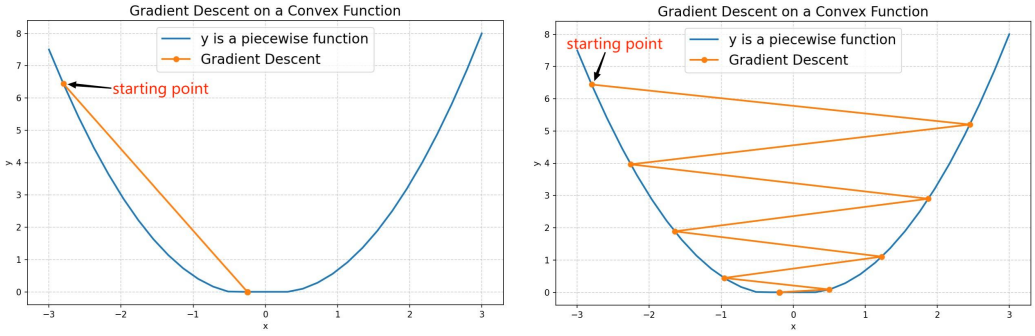

Figure 3: Gradient descent on convex function

The experiments described above demonstrate the influence of the learning rate on convergence performance. In general, both excessively small and excessively large learning rate values can lead to slow convergence. This highlights the sensitivity of gradient descent to the selected learning rate.

5.3 Gradient descent on strongly-convex function

We minimize the smooth and strongly-convex optimization problem:

$$f(x) = x^2 + x$$

We set the learning rate to $\gamma = \frac{8}{9}$, GD is as shown in the following figure4:

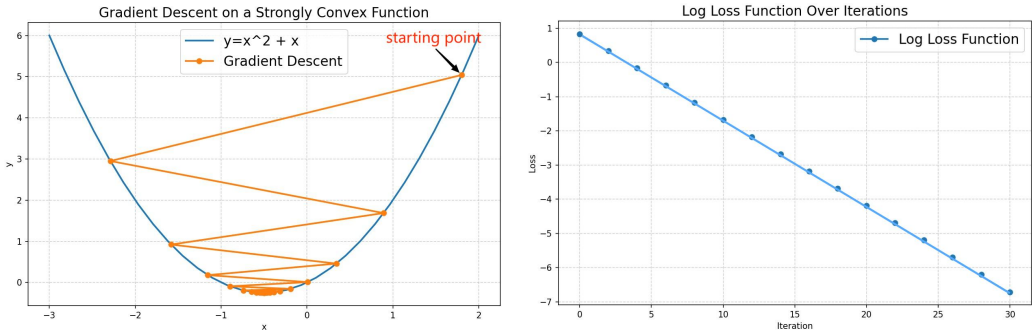

Figure 4: Gradient descent on strongly-convex function

We minimize the strongly-convex and smooth optimization problem:

$$f(x) = \begin{cases} \frac{3}{2}x^2, & \text{if } x \leq 0 \\ \frac{9}{2}x^2, & \text{if } x > 0 \end{cases}$$

If we set the learning rate to $\gamma = \frac{1}{20} < \frac{2}{L+\mu}(L = 9, \mu = 3)$, Double-start GD is as shown in the following figure5:

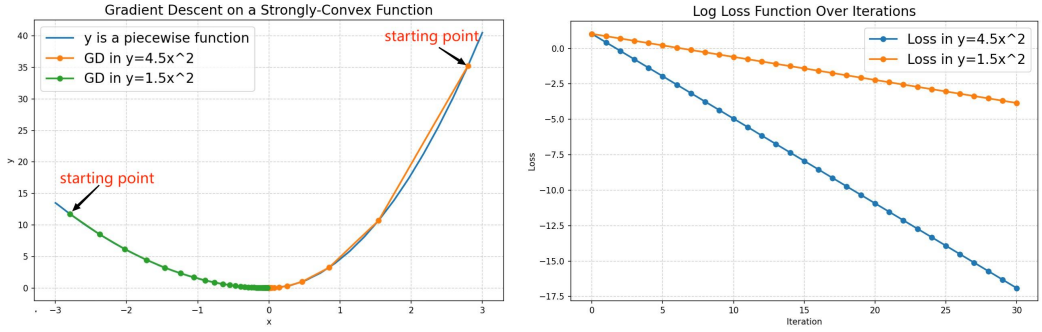

Figure 5: Double-start gradient descent on strongly-convex function

If we consider this piecewise function as two separate functions, We can find that he strong convexity constant μ affects the convergence rate and iteration complexity. In experiments, for the same value of γ, the quadratic function with a larger μ clearly converges faster.This also aligns with the conclusion drawn in our fourth section.

5.4 Gradient descent on bivariate strongly-convex function

We minimize two strongly-convex and smooth optimization problems:

$$f(x, y) = x^2 + y^2$$
$$f_1(x, y) = 10x^2 + y^2$$

It is easy to verify that $f(x, y)$ is L-smooth with $L = 2$ and μ-strongly with $\mu = 2$. We set its learning rate $\gamma = \frac{1}{4} = \frac{1}{L+\mu}$. Similarly, $f_1(x, y)$ is also L-smooth and μ-strongly, its $L = 20$ and $\mu = 2$. We set its learning rate $\gamma_1 = \frac{1}{11} = \frac{2}{L+\mu}$, GD converges as follows:

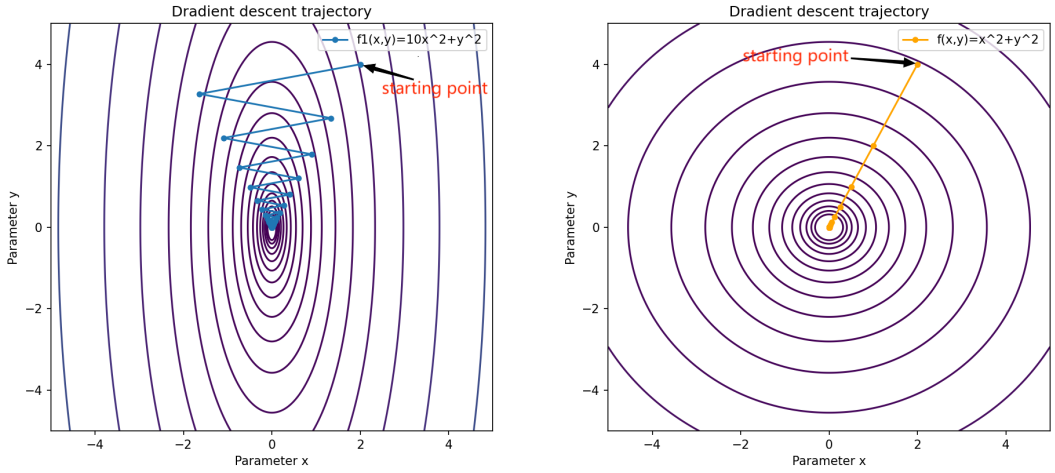

Figure 6: Gradient descent trajectory of $f(x, y)$ and $f_1(x, y)$

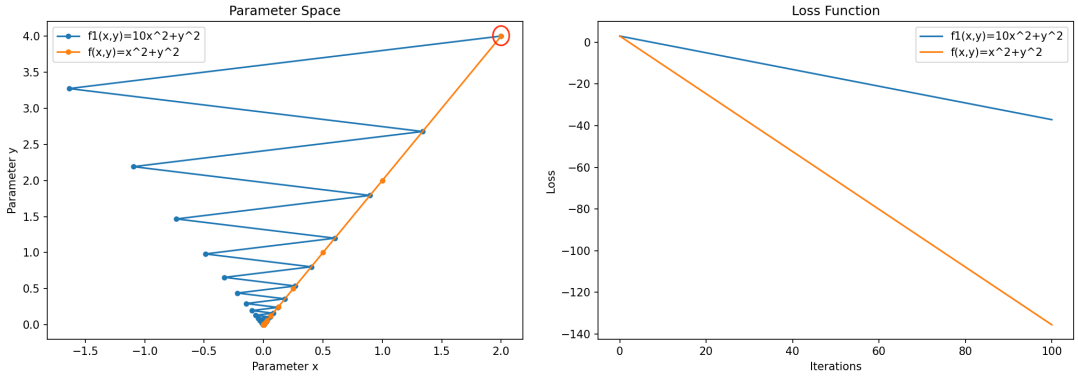

Figure 7: Parameter space trajectories and log loss function for the gradient descent of $f(x, y)$ and $f_1(x, y)$

We observe that the L-smooth constant L of strongly convex function indeed affects the convergence of gradient descent. When both functions share the same μ, if the L of $f(x, y)$ is smaller, with γ set to $C/(L+\mu)$, $f(x, y)$ converges faster and has a lower iteration complexity.