1 Outline

In this lecture, we study

- Convergence of gradient descent,
- Subgradients,
- Optimality conditions for non-differentiable functions,
- Subgradient method

2 Convergence of gradient descent

Recall that

- choosing proper step sizes,
- analyzing convergence rate, and
- analyzing a required number of iterations

are important when we develop the gradient descent method. In this section, we will see how various structural assumptions on the objective function lead to certain choices of step sizes and the resulting convergence rates. We state the corresponding convergence results without proofs for now, but we will get to prove them later in the course.

2.1 Smooth functions

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is *smooth* if there exists some $\beta > 0$ such that

$$\|\nabla f(x) - \nabla f(y)||_2 \leq \beta \|x - y\|_2$$

holds for any $x, y \in \mathbb{R}^d$. More precisely, we say that f is β-smooth in the norm $\|\cdot\|_2$. Recall that a convex function f satisfies

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

If f is β-smooth, then

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2.$$

Theorem 9.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth, and let $\{x_t : t = 1, \ldots, T + 1\}$ be the sequence of iterates generated by gradient descent with step size $\eta_t = 1/\beta$ for each t. Then*

$$f(x_{T+1}) - f(x^*) \leq \frac{\beta \|x_1 - x^*\|_2^2}{2T}$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Here, x_1 and x^* are some fixed vectors, which means that $\|x_1 - x^*\|_2$ is a constant. Moreover, the smoothness parameter β is also a constant. Hence, the convergence rate is $O(1/T)$. Therefore, after $T = O(1/\epsilon)$ iterations, we have

$$f(x_{T+1}) - f(x^*) \leq \epsilon.$$

2.2 Smooth and strongly convex functions

We say that a function is strongly convex if there exists some $\alpha > 0$ such that

$$f(x) - \frac{\alpha}{2}\|x\|_2^2$$

is convex. More precisely, we say that f is α-strongly convex in the norm $\|\cdot\|_2$. If f is α-strongly convex, then

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|_2^2.$$

If f is β-smooth and α-strongly convex, then it follows that

$$\frac{\alpha}{2}\|y - x\|_2^2 \leq (f(y) - f(x)) - \nabla f(x)^\top (y - x) \leq \frac{\beta}{2}\|y - x\|_2^2.$$

Here, we call $\kappa = \beta/\alpha$ the *condition number* of f. In fact, when f is both smooth and strongly convex, it leads to a drastic improvement in the convergence rate. The convergence rate depends on the condition number κ.

Theorem 9.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and α-strongly convex, and let $\{x_t : t = 1, \ldots, T+1\}$ be the sequence of iterates generated by gradient descent with sep size $\eta_t = 2/(\alpha + \beta)$ for each t. Then*

$$f(x_{T+1}) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4T}{\kappa + 1}\right) \|x_1 - x^*\|_2^2$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Note that $\exp(-4/(\kappa+1)) < 1$, and therefore, the convergence rate is $O(c^T)$ where $c = \exp(-4/(\kappa+1)) < 1$. Hence, we achieve linear convergence, and after $T = O(\log(1/\epsilon))$ iterations, we have

$$f(x_{T+1}) - f(x^*) \leq \epsilon.$$

2.3 Lipschitz continuous functions

We say that a differentiable function is Lipschitz continuous if there exists some $L > 0$ such that

$$|f(x) - f(y)| \leq L\|x - y\|_2$$

for any $x, y \in \mathbb{R}^d$. More precisely, we say that f is L-Lipschitz continuous in the norm $\|\cdot\|_2$. This is equivalent to

$$\|\nabla f(x)\|_2 \leq L$$

for any $x \in \mathbb{R}^d$.

Theorem 9.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz continuous, and let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by gradient descent with step size $\eta_t = \|x_1 - x^*\|_2/L\sqrt{T}$ for each t. Then*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{L\|x_1 - x^*\|_2}{\sqrt{T}}$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Here, we take the average of the points $x_1, \ldots, x_T$. Hence, the convergence rate is $O(1/\sqrt{T})$. This means that after $O(1/\epsilon^2)$ iterations, we have

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon.$$

In fact, Lipschitz continuity extends to non-differentiable functions, and gradient descent guarantees the same convergence rate for any non-differentiable functions as long as they are Lipschitz continuous.

2.4 Lipschitz continuous and strongly convex functions

If we assume strong convexity, then we deduce a faster convergence.

Theorem 9.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz continuous and α-strongly convex, and let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by gradient descent with step size $\eta_t = 2/\alpha(t+1)$ for each t. Then*

$$f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t\right) - f(x^*) \leq \frac{2L^2}{\alpha(T+1)}$$

where x^* is an optimal solution to $\min_{x\in\mathbb{R}^d} f(x)$.

Here, we take an weighted average of the points $x_1, \ldots, x_T$. The converge rate is $O(1/T)$, and after $O(1/\epsilon)$ iterations, we have

$$f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t\right) - f(x^*) \leq \epsilon.$$

3 Subgradients

The first-order characterization of convex functions states that a differentiable function f is convex if and only if $\mathrm{dom}(f)$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

for all $x, y \in \mathrm{dom}(f)$. For a function that is not necessarily differentiable, we can define the notion of *subgradients* as well as *subdifferentials*.

Definition 9.5. Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and a fixed point $x \in \mathrm{dom}(f)$, the *subdifferential* of f at x is defined as

$$\partial f(x) = \left\{ g : \ f(y) \geq f(x) + g^\top (y - x) \ \ \forall y \in \mathrm{dom}(f) \right\}.$$

Here, any $g \in \partial f(x)$ is called a *subgradient* of f at x.

Conversely, the subdifferential is the set of subgradients. If function f is differentiable at x, then we have $\partial f(x) = \{\nabla f(x)\}$, and therefore, the subdifferential reduces to the gradient. In contrast, a non-differentiable function may have more than one subgradient. Moreover, note that for any subgradient g at x, $f(x) + g^\top (y - x)$ provies a lower approximation of the function f.

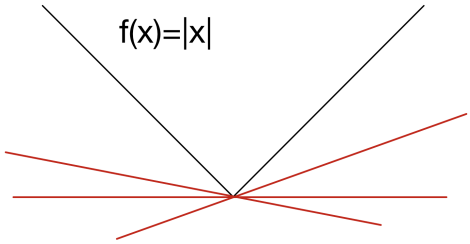

Figure 9.1: Subgradients of $f(x) = |x|$ at $x = 0$

Recall that for a differentiable univariate function f, the gradient of f at some point x is the slope of the line tangent to f at x. We have a similar geometric intuition for subgradients. Consider the the absolute value function $f(x) = |x|$ over $x \in \mathbb{R}$, which is not differentiable at $x = 0$. As depicted in Figure 9.1, there are multiple lines that are below $f(x) = |x|$ and go through $x = 0$. In fact, the subdifferential of f can be computed as follows.

$$\partial f(x) = \begin{cases} \{-1\} = \{\text{sign}(x)\}, & \text{for } x < 0 \\ [-1, 1], & \text{for } x = 0 \\ \{+1\} = \{\text{sign}(x)\}, & \text{for } x > 0 \end{cases}$$
$$= \begin{cases} \{\text{sign}(x)\}, & \text{for } x \neq 0 \\ [-1, 1], & \text{for } x = 0. \end{cases}$$

Let us consider a few more examples.

Example 9.6. Let $f(x) = \|x\|_1 : \mathbb{R}^d \to \mathbb{R}$. Then the subdifferential of f at any point $x = (x_1, \ldots, x_d)^\top$ is the set of vectors $g = (g_1, \ldots, g_d)^\top$ such that for each $i \in [d]$,

$$g_i = \begin{cases} \text{sign}(x_i), & \text{if } x_i \neq 0 \\ [-1, 1], & \text{if } x_i = 0. \end{cases}.$$

Example 9.7. Let $f_1, \ldots, f_k$ be convex functions, and let f is defined as the pointwise maximum of $f_1, \ldots, f_k$. Given a point x, if $f(x) = f_i(x)$ for some $i \in [k]$, then any subgradient of f_i is a subgradient of f.

Example 9.8. Given a convex set $C \subseteq \mathbb{R}^d$, the indicator function $I_C(x)$ at a point $x \in \mathbb{R}^d$ is defined as

$$I_C(x) = \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{if } x \notin C \end{cases}.$$

For a point $x \in C$, what is the subdifferential of the indicator function at x? Note that

$$I_C(x) = \left\{ g \in \mathbb{R}^d : \ 0 \geq 0 + g^\top (y - x) \ \ \forall y \in C \right\} = N_C(x).$$

Therefore, the subdifferential is precisely the normal cone of C at x.

4 Optimality conditions for non-differentiable convex functions

Now we consider the convex minimization problem with a general convex objective function that is not necessarily differentiable.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array} \quad = \quad \begin{array}{ll} \text{minimize} & f(x) + I_C(x) \\ \text{subject to} & x \in \mathbb{R}^d. \end{array}$$

The left is the constrained version, and the right formulation shows its unconstrained version with the indicator function. We discussed optimality conditions for convex minimization problems with a differentiable objective. In this section, we state and prove optimality conditions for the general case, in which the objective can be non-differentiable.

Theorem 9.9. *For a convex optimization problem* $\min_{x\in C} f(x)$*,* $x^* \in C$ *is an optimal solution if and only if there exists* $s \in \partial f(x^*)$ *such that*

$$s^\top (x - x^*) \geq 0 \quad \textit{for all } x \in C.$$

An immediate corollary of Theorem 9.9 is the following optimality condition for unconstrained problems.

Corollary 9.10. *For a convex optimization problem* $\min_{x\in\mathbb{R}^d} f(x)$*,* $x^* \in \mathbb{R}^d$ *is an optimal solution if and only if* $0 \in \partial f(x^*)$*.*

Corollary 9.10 can be applied to the unconstrained formulation of constrained convex minimization. We argued that for the differentiable case, the optimality condition is $-\nabla f(x^*) \in N_C(x^*)$ and $0 \in \{\nabla f(x^*)\} + N_C(x^*)$. As $\partial I_C(x)$ is equivalent to the normal cone $N_C(x)$, we obtain the following as a corollary of Corollary 9.10.

Corollary 9.11. *For a convex optimization problem* $\min_{x\in C} f(x)$*,* $x^* \in C$ *is an optimal solution if and only if*

$$0 \in \partial f(x^*) + N_C(x^*).$$

In this section, we will prove Theorem 9.9 which states the optimality condition for convex minimization. A tool that we need is the separating hyperplane theorem, which is an important result in convex analysis on its own. We state the separating hyperplane theorem without proof.

Theorem 9.12 (Separating hyperplane theorem)**.** *Let* $C, D \subseteq \mathbb{R}^d$ *be disjoint convex sets, i.e.,* $C \cap D = \emptyset$*, then there exists* $a \in \mathbb{R}^d \setminus \{0\}$ *and* $b \in \mathbb{R}$ *such that*

$$a^\top x \geq b, \quad \textit{for all } x \in C$$
$$a^\top x \leq b, \quad \textit{for all } x \in D$$

Let us prove Theorem 9.9 using Theorem 9.12.

Proof of Theorem 9.9. ($\Leftarrow$) Assume that there exists $s \in \partial f(x^*)$ such that $s^\top (x - x^*) \geq 0$ holds for all $x \in C$. Then it follows from the definition of subgradients that

$$f(x) - f(x^*) \geq s^\top (x - x^*) \geq 0 \quad \text{for all } x \in C.$$

This implies that $f(x) \geq f(x^*)$ for all $x \in C$, so x^* is optimal.

($\Rightarrow$) Let us consider the following two sets.

$$C = \{(x - x^*, t) : \ f(x) - f(x^*) \leq t\},$$
$$D = \{(x - x^*, t) : \ x \in C, \ t < 0\}.$$

Since $f(x) - f(x^*) \geq 0$ for any $x \in C$, these two sets are disjoint. Then by Theorem 9.12, there exists $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$ such that $(a, b) \neq (0, 0)$ and

$$a^\top (x - x^*) + bt \geq c, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t \tag{9.1}$$
$$a^\top (x - x^*) + bt \leq c, \quad \forall x \in C, \ t < 0. \tag{9.2}$$

In (9.2), t can be arbitrarily small, so $b \geq 0$. Suppose that $b = 0$, in which case (9.1) becomes

$$a^\top (x - x^*)0 \geq c, \quad \forall x \in \mathbb{R}^d,\ f(x) - f(x^*) \leq t.$$

Here, $x - x^*$ can be $\lambda \cdot a$ where λ is an arbitrarily small number. This implies that $a = 0$. However, this contradicts the condition that $(a, b) \neq (0, 0)$. Therefore, $b > 0$. Then, without loss of generality, we may assume that $b = 1$. Then taking $x = x^*$ and $t = 0$ in (9.1), we obtain $0 \geq c$. Moreover, taking $x = x^*$ and a number that is arbitrarily close to 0 for t, it follows that $0 \leq c$. Hence, $c = 0$. Then (9.1) and (9.2) become

$$a^\top (x - x^*) + t \geq 0, \quad \forall x \in \mathbb{R}^d,\ f(x) - f(x^*) \leq t \tag{9.3}$$

$$a^\top (x - x^*) + t \leq 0, \quad \forall x \in C,\ t < 0. \tag{9.4}$$

Here, we take $t = f(x) - f(x^*)$ in (9.3). Then (9.3) becomes

$$f(x) \geq f(x^*) - a^\top (x - x^*),$$

which implies that $-a \in \partial f(x^*)$. Moreover, we take a number that is arbitrarily close to 0 for t in (9.4). Then it becomes $a^\top (x - x^*) \leq 0$, which is equivalent to $-a^\top (x - x^*) \geq 0$. Hence, $-a$ is the desired vector. □

5 Subgradient method

We discussed the gradient descent method for minimizing a differentiable convex function. For non-differentiable convex functions, we can consider subgradients and use the subgradient method described as follows.

Algorithm 1 Subgradient method

Initialize $x_1 \in \mathrm{dom}(f)$.
for $t = 1, \ldots, T$ **do**
  Obtain a subgradient $g_t \in \partial f(x_t)$.
  $x_{t+1} = x_t - \eta_t g_t$ for a step size $\eta_t > 0$.
end for

We will show that the subgradient method given by Algorithm 1 converges if the subgradients of f are bounded. Recall that for the differentiable case, the ℓ_2 norm of f's gradient is bounded if and only if f is Lipschitz continuous.

Theorem 9.13. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function such that $\|g\|_2 \leq L$ for any $g \in \partial f(x)$ for every $x \in \mathbb{R}^d$. Let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by the subgradient method with step size $\eta_t = \|x_1 - x^*\|_2 / L\sqrt{T}$ for each t. Then*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{L\|x_1 - x^*\|_2}{\sqrt{T}}$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Proof. Let $\eta = \|x_1 - x^*\|_2 / L\sqrt{T}$. Then $\eta_t = \eta$ for each $t \geq 1$. Note that

$$\begin{aligned}
\|x_{t+1} - x^*\|_2^2 &= \|x_t - \eta g_t - x^*\|_2^2 \\
&= \|x_t - x^*\|_2^2 - 2\eta g_t^\top (x_t - x^*) + \eta^2 \|g_t\|_2^2 \\
&\leq \|x_t - x^*\|_2^2 - 2\eta(f(x_t) - f(x^*)) + \eta^2 \|g_t\|_2^2
\end{aligned}$$

where the inequality follows from $f(x^*) \geq f(x_t) + g_t^\top (x^* - x_t)$. Then it follows that

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta} \left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) + \frac{\eta}{2} \|g_t\|_2^2.$$

Summing this over $t = 1, \ldots, T$ and dividing the resulting one by T, we obtain

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*) &\leq \frac{1}{2\eta T} \left(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2 \right) + \frac{\eta}{2T} \sum_{t=1}^{T} \|g_t\|_2^2 \\
&\leq \frac{\|x_1 - x^*\|_2^2}{2\eta T} + \frac{\eta}{2} L^2 \\
&= \frac{L \|x_1 - x^*\|_2}{\sqrt{T}}
\end{aligned}$$

where the second inequality is because $\|x_{T+1} - x^*\|_2 \geq 0$ and $\|g_t\|_2 \leq L$. Lastly, as f is convex,

$$f\left(\frac{1}{T} \sum_{t=1}^{T} x_t \right) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{L \|x_1 - x^*\|_2}{\sqrt{T}},$$

as required. □

Here, the step size η has the order of $O(1/\sqrt{T})$ when we run the subgradient method for T iterations. Then the convergence rate is $O(1/\sqrt{T})$, and the number of required iterations to bound the error by ϵ is $O(1/\epsilon^2)$.

The important property of the subgradient method is that it is "dimension-free" in the sense that the algorithm and the convergence rate do not depend on the ambient dimension d. In many applications, we have a moderate tolerance for the error ϵ while the dimension d is huge. For such applications, the fact that the subgradient method is dimension-free has a huge advantage.