1 Outline

In this lecture, we study

- Convergence of gradient descent for strongly convex functions,
- Subgradient and subdifferential,
- Subgradient method,
- Convergence rate of gradient descent for smooth functions.

2 Convergence of gradient descent

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 non-differentiable function, we can define the notion of *subgradients* as well as *subdifferentials*.

Definition 10.1. Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and a 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.

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 10.1, 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.

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

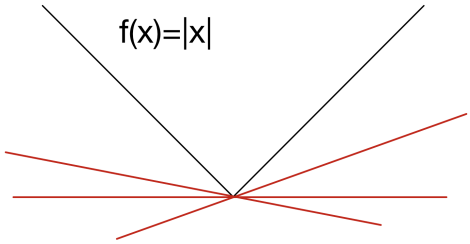

Figure 10.1: Subgradients of $f(x) = |x|$ at $x = 0$

Let us consider a few more examples.

Example 10.2. 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 10.3. Let $f_1, \ldots, f_k$ be convex functions, and let f be 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 10.4. 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

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

where $N_C(x)$ denotes the normal cone of C at x. Therefore, the subdifferential of the indicator function for C at x is precisely the normal cone of C at x.

4 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 $x_1 \in \text{dom}(f)$.
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 10.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function such that $\|g\|_2 \le L$ for any $g \in \partial f(x)$ for every $x \in \mathbb{R}^d$. Let $\{x_t : t = 1, \dots, T\}$ be the sequence of iterates generated by the subgradient method with step size*

$$\eta_t = \frac{\|x_1 - x^*\|_2}{L\sqrt{T}}$$

for each t. Then

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

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

Proof. Note that

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

where the inequality follows from $f(x^*) \ge f(x_t) + g_t^\top (x^* - x_t)$ as g_t is a subgradient at x_t. The rest of the proof is the same as the argument used for the differentiable case. □

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.

5 Convergence of gradient descent for smooth functions

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth with respect to the ℓ_2 norm for some $\beta > 0$ if

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

holds for any $x, y \in \mathbb{R}^d$. Smooth functions have the *self-tuning* property! By the optimality condition (for unconstrained problmes), we have $\nabla f(x^*) = 0$ for any optimal solution x^*. Then the smoothness assumption implies that the gradient gets close to 0 as we approach an optimal solution. This is in contrast to a non-differentiable function, e.g., $f(x) = |x|$ over $\mathbb{R}$.

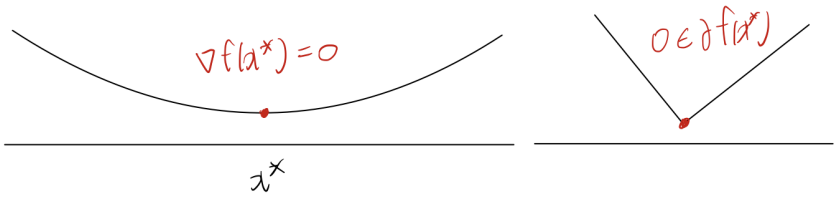

Figure 10.2: Smooth functions vs non-smooth functions

Recall that the gradient descent method for Lipschitz continuous functions requires a constant but small step size $O(1/\sqrt{T})$ where T is the total number of iterations. This is partly because the subgradient does not get smaller even we converge to an optimal solution. In contrast, for smooth functions, we can take large step sizes, because the gradient gets reduced as we converge to an optimal solution. This is referred to as the self-tuning property.

Next we prove the convergence result for smooth function. The first thing we observe is that a gradient step for a smooth function can always guarantee a strict improvement. To explain this, take a differentiable and β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$. Then a gradient step is given by

$$x_{t+1} = x_t - \eta_t \nabla f(x_t).$$

Note that

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2 \\ &= f(x_t) + \left(-\eta_t + \frac{\eta_t^2 \beta}{2}\right) \|\nabla f(x_t)\|_2^2 \\ &\leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2 \end{aligned}$$

where the first inequality follows from the β-smoothness of f and the second inequality is because the term inside the parenthesis is a quadratic function in η_t which can be maximized at $\eta_t = 1/\beta$. Therefore, when $\eta_t = 1/\beta$, we obtain

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2,$$

which implies that $f(x_{t+1})$ is strictly better than $f(x_t)$ when x_t is not an optimal solution. Based on this observation, we can prove the following convergence result for smooth functions.

Theorem 10.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth in the ℓ_2-norm and convex, and let $\{x_t : t = 1, \ldots, T+1\}$ be the sequence of iterates generated by gradient descent with step size*

$$\eta_t = \frac{1}{\beta}$$

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

Proof. Note that

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2 \\ &\leq f(x^*) - \nabla f(x_t)^\top (x^* - x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2 \\ &= f(x^*) + \frac{\beta}{2}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \end{aligned}$$

where the second inequality is because $f(x_t) + \nabla f(x_t)^\top (x - x_t)$ is a lower bound on f and the equality follows because $x_{t+1} = x_t - (1/\beta)\nabla f(x_t)$. This implies that

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

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

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2T}\left(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2\right) \leq \frac{\beta}{2T}\|x_1 - x^*\|_2^2.$$

Recall that each gradient step for smooth functions leads to an improvement, i.e., $f(x_{t+1}) \leq f(x_t)$. Therefore,

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

as required. □

The important takeaway is that we took a constant step size $1/\beta$, which does not depend on the number of iterations T. This is due to the self-tuning property of smooth functions. Although we do not shrink the step size, the change between the current iterate x_t and the next iterate x_{t+1} gets reduced as we approach an optimal solution.

As discussed before, the term $\|x_1 - x^*\|_2$ and the smoothness parameter β are all fixed constants. Hence, the convergence rate is $O(1/T)$. Therefore, after $T = O(1/\epsilon)$ iterations, we have

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

Note that the convergence results for smooth functions improves over $O(1/\sqrt{T})$ and $O(1/\epsilon^2)$ for the subgradient method. Moreover, let us compare the last steps of their analyses. For the subgradient method, we had

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\eta T} + \frac{\eta}{2}L^2,$$

whereas the last step for smooth functions was that

$$f(x_{T+1}) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\eta T}.$$

These are almost the same, but for smooth functions, we did not have the additional term $\eta L^2/2$ on the right-hand side. Moreover, we used the fact that each gradient step improves the objective.