1 Outline

In this lecture, we cover

- projected subgradient method,
- gradient descent for smooth functions,
- adaptive gradient (AdaGrad).

2 Convergence of Gradient Descent for Smooth Functions

Remember that gradient descent for a Lipschitz continuous function has a conergence rate of $O(1/\sqrt{T})$ and uses a constant step size of order $O(1/\sqrt{T})$. In this section, we consider **smooth** convex functions for which gradient descent achieves a faster convergence rate.

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

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

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 $\mathbb{R}$.

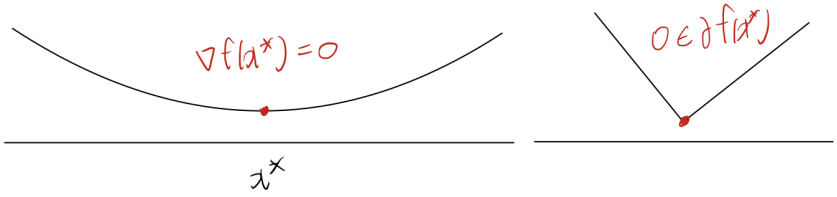

Figure 4.1: Smooth functions vs non-smooth functions

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 when we get realy close 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 $f : \mathbb{R}^d \to \mathbb{R}$.

Lemma 4.1. *If f is β-smooth in the ℓ_2 norm, then*

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

Proof. By the fundamental theorem of calculus and the Cauchy-Schwarz inequality, we obtain the following.

$$\begin{aligned}\left|f(y) - f(x) - \nabla f(x)^\top (y-x)\right| &= \left|\int_0^1 (y-x)^\top \left(\nabla f(x + t(y-x)) - \nabla f(x)\right) dt\right| \\ &\leq \int_0^1 \|y-x\|_2 \left\|\nabla f(x + t(y-x)) - \nabla f(x)\right\|_2 dt \\ &\leq \int_0^1 \beta t \|y-x\|_2^2 dt \\ &= \frac{\beta}{2}\|y-x\|_2^2\end{aligned}$$

where the equality is due to the fundamental theorem of calculus, the first inequality is by the Cauchy-Schwarz inequality, and the second inequality is from the β-smoothness of f. □

Consider a gradient step 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 4.2. *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.

3 Projected Subgradient Method

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 4.3. 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

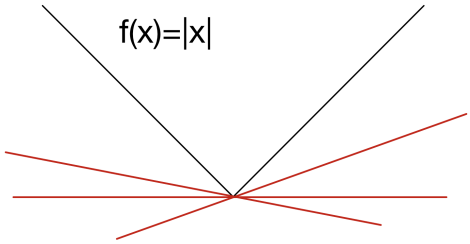

Figure 4.2: Subgradients of $f(x) = |x|$ at $x = 0$

in Figure 4.2, 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 4.4. 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 $i \in [d]$,

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

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 Projected Subgradient Method

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

In Algorithm 1, we have the operation $\text{Proj}_{\mathcal{X}}(x_t - \eta_t g_t)$. Here, the operator $\text{Proj}_{\mathcal{X}}(\cdot)$ is the **projection** operator defined as

$$\text{Proj}_{\mathcal{X}}(x) = \operatorname*{argmin}_{y \in \mathcal{X}} \|x - y\|_2.$$

Lemma 4.5. *Let $u, v \in \mathbb{R}^d$. Then*

$$\|\text{Proj}_{\mathcal{X}}(u) - \text{Proj}_{\mathcal{X}}(v)\|_2 \leq \|u - v\|_2.$$

Proof. Note that

$$\text{Proj}_{\mathcal{X}}(x) = \operatorname*{argmin}_{y \in \mathcal{X}} \frac{1}{2}\|x - y\|_2^2.$$

By the optimality condition, it follows that

$$(\text{Proj}_{\mathcal{X}}(x) - x)^\top (y - \text{Proj}_{\mathcal{X}}(x)) \geq 0$$

for all $y \in \mathcal{X}$. Equivalently,

$$(\text{Proj}_{\mathcal{X}}(x) - x)^\top (\text{Proj}_{\mathcal{X}}(x) - y) \leq 0 \quad \text{for all } y \in \mathcal{X}.$$

Next let us consider two points u, v and their projections onto C, given by $\text{Proj}_C(u)$ and $\text{Proj}_C(v)$, respectively. Then we have

$$\begin{aligned}(\text{Proj}_C(u) - u)^\top (\text{Proj}_C(u) - \text{Proj}_C(v)) \leq 0,\\ (\text{Proj}_C(v) - v)^\top (\text{Proj}_C(v) - \text{Proj}_C(u)) \leq 0.\end{aligned}$$

Adding these two inequalities, we obtain

$$\|\text{Proj}_C(u) - \text{Proj}_C(v)\|_2^2 - (u - v)^\top (\text{Proj}_C(u) - \text{Proj}_C(v)) \leq 0.$$

Then it follows from the Cauchy-Schwarz inequality that $\|\text{Proj}_C(u) - \text{Proj}_C(v)\|_2 \leq \|u - v\|_2$, as required. □

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 4.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex 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*

$$\eta_t = \frac{C}{\sqrt{T}}$$

for each t for some constant C. Then

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{C'}{\sqrt{T}}$$

for some constant C' where x^* is an optimal solution to $\min_{x \in \mathcal{X}} f(x)$.

Proof. Let $\eta = C/\sqrt{T}$. Note that

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

where the first inequality follows from Lemma 4.5 and the second inequality follows from $f(x^*) \geq f(x_t) + g_t^\top (x^* - x_t)$ as g_t is a subgradient at x_t. Then it follows that

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

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 $O(1/\epsilon^2)$.

4 Adaptive Gradient Method

We observed that the performance of gradient descent heavily depends on our choice of learning rates. We have discussed how to choose learning rates to achieve the best theoretical convergence guarantee. For example, for a β-smooth function, our choice was $\eta = 1/\beta$. However, to implement this choice of learning rate, we need to know the smoothness parameter β, which can be difficult in practice.

In this section, we study what is known as **adaptive gradient (AdaGrad)**, which is a variant of gradient descent that automatically adapt to the problem structure such as smoothness and Lipschitz continuity, AdaGrad deploys a schedule of learning rates that do not require knowledge of the smoothness parameter β and the Lipschitz continuity constant L. Moreover, AdaGrad achieves a convergence rate of order $O(1/\sqrt{T})$ for Lipschitz continuous functions and a convergence rate of order $O(1/T)$ for smooth functions.

Algorithm 2 AdaGrad

Initialize $x_1 \in \mathcal{X}$ and $S_0 = 0$.
for $t = 1, \ldots, T$ **do**
  Obtain a subgradient $g_t \in \partial f(x_t)$.
  Set $S_t = S_{t-1} + \|g_t\|_2^2$ and $\eta_t = R/\sqrt{2S_t}$.
  Update $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}(x_t - \eta_t g_t)$.
end for

In Algorithm 2, R denotes an upper bound on the diameter of $\mathcal{X}$, i.e., $R \geq \sup_{u,v\in\mathcal{X}} \|u - v\|_2$. The following lemma holds under AdaGrad, and it will be used to provide convergence guarantees for AdaGrad.

Lemma 4.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. Let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by AdaGrad. Then*

$$\sum_{t=1}^{T} g_t^\top (x_t - x^*) \leq \sqrt{2R^2 \sum_{t=1}^{T} \|g_t\|_2^2}.$$

Proof. Note that

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

Then it follows that

$$\begin{aligned}\sum_{t=1}^{T} g_t^\top (x_t - x^*) &\le \frac{\|x_1 - x^*\|_2^2}{2\eta_1} + \sum_{t=1}^{T} \frac{\|x_t - x^*\|_2^2}{2}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \sum_{t=1}^{T} \frac{\eta_t}{2}\|g_t\|_2^2 \\ &\le \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2}\|g_t\|_2^2 \\ &\le \frac{R}{2}\sqrt{2S_T} + \frac{R}{2\sqrt{2}} \sum_{t=1}^{T} \frac{\|g_t\|_2^2}{\sqrt{S_t}}.\end{aligned}$$

It is known that for any non-negative numbers $a_1, \ldots, a_n$, we have

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \le 2\sqrt{\sum_{i=1}^{n} a_i}.$$

Applying this inequality, it follows that

$$\sum_{t=1}^{T} g_t^\top (x_t - x^*) \le \frac{R}{2}\sqrt{2S_T} + \frac{R}{\sqrt{2}}\sqrt{S_T} = \sqrt{2R^2 \sum_{t=1}^{T} \|g_t\|_2^2},$$

as required. □

Based on Lemma 4.7, we can prove the following theorem which states that AdaGrad guarantees a convergence rate of $O(1/\sqrt{T})$.

Theorem 4.8. *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 $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by AdaGrad. Then*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \le \frac{LR\sqrt{2}}{\sqrt{T}}.$$

Next, we consider smooth functions. To analyze AdaGrad for smooth functions, we need the following lemma on smooth functions.

Lemma 4.9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. If f is β-smooth, then for any $x, y \in \mathbb{R}^d$,*

$$f(y) \ge f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Proof. Since f is β-smooth, we have

$$f(z) \le f(x) + \nabla f(x)^\top (z - x) + \frac{\beta}{2}\|z - x\|_2^2 \quad \text{for any } z \in \mathbb{R}^d.$$

Then taking $z = y + (1/\beta)(\nabla f(x) - \nabla f(y))$,

$$\begin{aligned}f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\le -\nabla f(x)^\top (z - x) + \nabla f(y)^\top (z - y) + \frac{\beta}{2}\|z - y\|_2^2 \\ &= \nabla f(x)^\top (x - y) + (\nabla f(x) - \nabla f(y))^\top (y - z) + \frac{\beta}{2}\|z - y\|_2^2 \\ &= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_2^2.\end{aligned}$$

Then it follows that

$$\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) - f(x) + \nabla f(x)^\top (x - y),$$

as required. □

The following theorem states that AdaGrad guarantees a convergence rate of $O(1/T)$ for smooth functions.

Theorem 4.10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that is β-smooth. Let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by AdaGrad. Then*

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

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

Proof. By Lemma 4.9, we have

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2,$$

which implies that

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

Moreover, by Lemma 4.7, we deduce that

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{1}{T}\sqrt{2R^2 \sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2} - \frac{1}{2\beta T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2.$$

Here,

$$\sqrt{2R^2 \sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2} - \frac{1}{2\beta}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq \max_{z \geq 0}\left\{Rz\sqrt{2} - \frac{1}{2\beta}z^2\right\} = \beta R^2.$$

Therefore, it follows that

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{\beta R^2}{T},$$

as required. □

The convergence guarantees of AdaGrad provided by Theorems 4.8 and 4.10 have the same asymptotic rates as those of gradient descent. Gradient descent, however, need to adjust learning rates based on whether a given function is smooth of Lipschitz continuous. In contrast, the adaptive schedule of learning rates of AdaGrad automatically adapt to the structure of a given function.