1 Outline

In this lecture, we study

- Properties of smooth functions and strongly convex functions,
- Convergence of gradient descent for smooth functions.

2 Gradient descent for smooth and strongly convex functions

2.1 Smooth functions and strongly convex functions

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth with respect to the ℓ_2 norm form 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 $\mathbb{R}$.

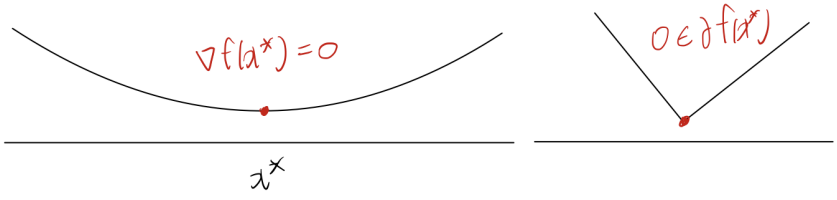

Figure 10.1: Smooth functions vs non-smooth functions

Recall that the subgradient method requires a constant but small step size $\approx O(1/\sqrt{T})$ 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.

Theorem 10.1. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth with respect to the ℓ_2 norm for some $\beta > 0$ if and only if $g(x) = (\beta/2)\|x\|_2^2 - f(x)$ is convex.*

Proof. ($\Rightarrow$) Recall that g is convex if and only if the monotonicity condition is satisfied. Note that $\nabla g(x) = \beta x - \nabla f(x)$ and

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle = \beta \|x - y\|^2 - \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Then by the Cauchy-Schwarz inequality,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\|_2 \cdot \|x - y\|_2 \leq \beta \|x - y\|_2^2.$$

Therefore, we have $\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq 0$, which implies that g is convex.

($\Leftarrow$) Since g is convex, the first-order characterization of convex functions states that

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

which is equivalent to

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

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

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

Then it follows that

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

Similarly, we obtain

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

Adding these two inequalities, it follows that

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

Since $(\nabla f(x) - \nabla f(y))^\top (x - y) \leq \|\nabla f(x) - \nabla f(y)\|_2 \cdot \|x - y\|_2$ by the Cauchy-Schwarz inequality, we obtain

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

Therefore, f is β-smooth. □

By the first-order characterization of convex functions, $g(x) = \frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex if and only if $g(y) \geq g(x) + \nabla g(x)^\top (y - x)$ for any $x, y \in \mathbb{R}^d$. This condition is equivalent to

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

This basically means that a smooth function can be upper bounded by a quadratic function.

In fact, we can prove a stronger statement given as follows.

Lemma 10.2. *If f is β-smooth with respect to the ℓ_2 norm, then*

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

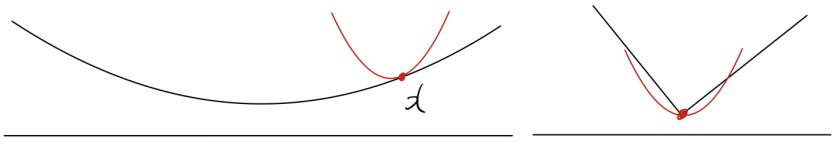

Figure 10.2: Quadratic upper bound on a smooth function

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. □

Example 10.3. Quadratic functions $x^\top Q x/2 + p^\top x + r$.

Recall that a function f is α-strongly convex in the norm $\|\cdot\|_2$ for some $\alpha > 0$ if $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex. Why is strong-convexity useful? Let us first derive the following property of strongly convex functions.

Lemma 10.4. *If f is α-strongly convex with respect to the ℓ_2 norm, then*

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

Proof. By definition, we have that $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex. Then by the first-order characterization of convex functions, we have $g(y) \geq g(x) + \nabla g(x)^\top (y-x)$ for any $x, y \in \mathbb{R}^d$. This is equivalent to

$$\begin{aligned}f(y) &\geq f(x) + (\nabla f(x) - \alpha x)^\top (y-x) - \frac{\alpha}{2}\|x\|_2^2 + \frac{\alpha}{2}\|y\|_2^2 \\ &= f(x) + \nabla f(x)^\top (y-x) + \frac{\alpha}{2}\|y-x\|_2^2,\end{aligned}$$

as required. □

Lemma 10.4 implies that a strongly convex function is lower bounded by a quadratic function. This means that, when a point is far from an optimal solution, the gradient at this point has to be large. Hence, when applying gradient descent or the subgradient method, this leads to a faster convergence.

In Figure 10.4, we compare smoothness and strong convexity. The first figure shows a strongly convex function that is not smooth, obtained by taking the point-wise maximum of two smooth functions. The second figure illustrates a smooth function that is not strongly convex. In particular, the second figure shows a smooth curve around the minimum point while it exhibits a linear growth when being far away from the minimum point.

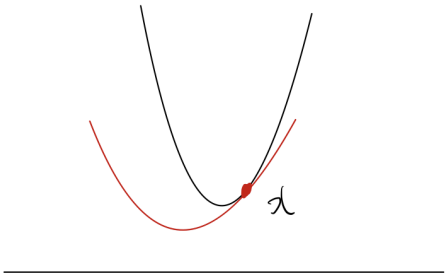

Figure 10.3: Quadratic lower bound on a strongly convex function

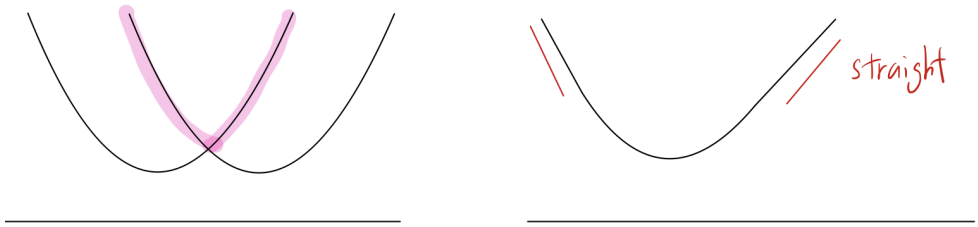

Figure 10.4: Strongly convex but non-smooth function vs smooth function that is not strongly convex

2.2 Convergence result for smooth functions

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

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.