3341.454: Mathematical and Numerical Optimization **SNU, Spring 2026**
Lecture #9: Introduction to gradient descent April 1, 2026
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- introduction to gradient descent,
- convergence analysis of gradient descent for smooth functions, and
- properties of smooth functions.

2 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

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

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 $\mathbb{R}$.

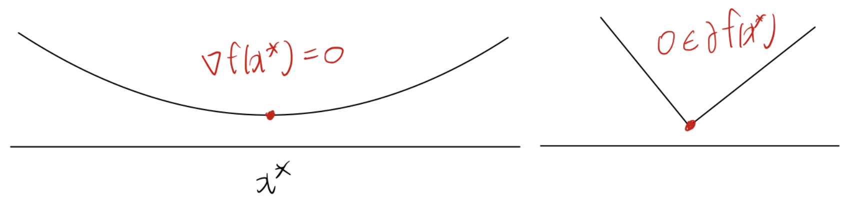

Figure 9.1: Smooth functions vs non-smooth functions

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 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

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

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

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

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 9.1. *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

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

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

Proof. Note that

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

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

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

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

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

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

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

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.

3 Smooth functions

3.1 Quadratic upper bounds on smooth functions

In the last lecture, we discussed the convergence rate of gradient descent on smooth functions. The key property of smooth functions based on which we developed the analysis is the following quadratic upper bound inequality for smooth functions. Namely, for a β-smooth convex function $f : \mathbb{R}^d \to \mathbb{R}$, we have

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

for any $x, y \in \mathbb{R}^d$. Let us verify this while we discuss some important properties of smooth functions.

Theorem 9.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then f is β-smooth in 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

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

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. $\square$

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

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

which verifies the quadratic upper bound property of smooth functions.

In fact, the inequality holds regardless of whether f is convex or not. We prove the following stronger statement.

Lemma 9.3. *If f is β-smooth in the ℓ_2 norm, then*

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

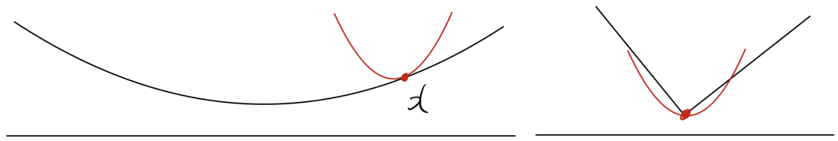

Figure 9.2: Quadratic upper bound on a smooth function

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. $\square$

3.2 Optimality gap for smooth functions

Another interesting result is that when f is smooth, we can measure the gap between the optimal value and $f(x)$ for any given solution x. More precisely, we prove the following result.

Theorem 9.4. *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth in the ℓ_2 norm and convex, then*

$$\frac{1}{2\beta}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{\beta}{2}\|x - x^*\|_2^2 \quad \forall x \in \mathbb{R}^d$$

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

Proof. Let us prove the upper bound on $f(x) - f(x^*)$ first. As f is β-smooth, we have

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

which implies the upper bound as $\nabla f(x^*) = 0$. For the lower bound, note that for any $y \in \mathbb{R}^d$,

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

Here, we can take $y = x - (1/\beta)\nabla f(x)$, which makes the right-most side

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

Then it follows that

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

as required. $\square$

Based on Theorem 9.4, we can prove the following property of smooth functions.

Lemma 9.5. *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth in the ℓ_2 norm and convex, then*

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

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

Proof. Given $x, y \in \mathbb{R}^d$, we take the following two functions.

$$\begin{aligned} g(z) &= f(z) - \nabla f(x)^\top z, \\ h(z) &= f(z) - \nabla f(y)^\top z. \end{aligned}$$

As $\nabla g(z) = \nabla f(z) - \nabla f(x)$ and $\nabla h(z) = \nabla f(z) - \nabla f(y)$, it follows that x and y minimize g and h, respectively. Moreover, g and h are both β-smooth. Note that

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

Similarly, we have

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

Adding these two inequalities, we obtain

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

as required. □