IE 539: Convex Optimization **KAIST, Fall 2024**
Lecture #9: Convergence of gradient descent October 7, 2024
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- Taylor approximation interpretation,
- Convergence of gradient descent.

2 Introduction to gradient descent

2.1 Gradient descent method

In the last lecture, we learned that the following provides a general template for the gradient descent algorithm.

Algorithm 1 Gradient descent method

```
Initialize $x_1 \in \mathrm{dom}(f)$.
for $t = 1, \dots, T$ do
    $x_{t+1} = x_t - \eta_t \nabla f(x_t)$ for a step size $\eta_t > 0$.
end for
```

We also discussed applying gradient descent to find a minimizer of function $f(x) = 2x^2+3x : \mathbb{R} \to \mathbb{R}$. We observed that the minimizer of f is given by $x^* = -3/4$ and that gradient descent with a constant step size $\eta_t = \eta$ proceeds with the following update rule.

$$x_{t+1} = x_t - \eta \nabla f(x_t) = x_t - \eta(4x_t + 3) = (1-4\eta)x_t - 3\eta$$

Then it follows that

$$x_{t+1} = (1-4\eta)^t \left(x_1 + \frac{3}{4}\right) - \frac{3}{4}.$$

Since $x^* = -3/4$, this implies that

$$|x_{T+1} - x^*| = O\left((1-4\eta)^T\right)$$

where x_{T+1} is the solution obtained after running gradient descent for T iterations. We call $x_1, \dots, x_{T+1}$ **iterates**.

- If η is chosen to satisfy $|1-4\eta| < 1$, then $(1-4\eta)^T$ converges to zero as $T \to \infty$.
- Note that
$$f(x_{T+1}) - f(x^*) = 2(x_{T+1} - x^*)^2 + (4x^* + 3)(x_{T+1} - x^*).$$
When $|1-4\eta| < 1$, $(1-4\eta)^T$ converges slower than $(1-4\eta)^{2T}$. Hence, we have
$$|f(x_{T+1}) - f(x^*)| \leq O\left((1-4\eta)^T\right).$$
- Therefore, after $T = O(\log(1/\epsilon))$ iterations, we have
$$|f(x_{T+1}) - f(x^*)| \leq \epsilon.$$

2.2 Taylor approximation interpretation

Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and a point $x_t \in \mathrm{dom}(f)$, the first-order Taylor approximation of f at x_t is given by

$$f(x) \approx f(x_t) + \nabla f(x_t)^\top (x - x_t).$$

If f is convex, then by the first-order characterization of convexity, we know that the first-order Taylor approximation is a lower bound on f. Moreover, as it provides an approximation of f, we can try to minimize $f(x) \approx f(x_t) + \nabla f(x_t)^\top (x - x_t)$ instead of f. However, the first-order Taylor approximation is a linear function, which means that

$$\min_x \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) \right\} = -\infty.$$

Instead of minimizing the first-order Taylor approximation directly, we add a proximity term as follows.

$$f(x) \approx f(x_t) + \nabla f(x_t)^\top (x - x_t) + \underbrace{\frac{1}{2\eta_t} \|x - x_t\|_2^2}_{\text{proximity term}}.$$

When minimizing the second approximation, we cannot choose a solution that is too far from x_t, for otherwise, the corresponding value will be high due to the proximity term. Again, we minimize the approximation instead of f and take a unique minimizer x_{t+1} as follows.

$$x_{t+1} \in \operatorname*{argmin}_x \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right\}.$$

This gives us an iterative algorithm for minimizing f. Again, the proximity term prevents the solution from being too far away from the initial point x_t. The larger η_t is, the closer the solution x_{t+1} is to the starting point x_t. In fact, there is a closed-form expression for x_{t+1}. Recall that the approximation with the proximity term is differentiable, and therefore, it follows from the optimality condition that

$$\nabla f(x_t) + \frac{1}{\eta_t}(x_{t+1} - x_t) = 0.$$

This is equivalent to

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

which is precisely the gradient descent iteration.

3 Convergence of gradient descent

In this section, we cover some convergence results for the gradient descent method. Here, the term "convergence" simply means convergence to an optimal solution or the optimal value. When we talk about convergence results, we often care about the rate of convergence, which measures how quickly a given algorithm converges. We discussed above that it is crucial to choose proper step sizes to achieve convergence. In the previous example with $f(x) = 2x^2 + 3x$, we used a constant step size η satisfying $|1 - 4\eta| < 1$ to guarantee convergence, and if $|1 - 4\eta|$ were greater than 1, gradient descent would not converge. Moreover, the convergence rate was $O(c^t)$ where $c = |1 - 4\eta| < 1$, so gradient descent converges exponentially fast. Based on this, we said that to achieve an ϵ-optimal solution, meaning that the difference between its value and the optimal value is at most ϵ, we need only $O(\log(1/\epsilon))$ iterations. Basically,

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

will be the main subjects of this section.

3.1 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 9.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz continuous in the ℓ_2-norm and convex, and let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by gradient descent with step size*

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

for each t. Then

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

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.

Proof of Theorem 9.1. Let $\eta = \|x_1 - x^*\|_2/L\sqrt{T}$. Then $\eta_t = \eta$ for each $t \geq 1$. Note that

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

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

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

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

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

where the second inequality is because $\|x_{T+1} - x^*\|_2 \geq 0$ and $\|\nabla f(x_t)\|_2 \leq L$. 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. □

3.2 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 9.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and convex, and let $\{x_t : t = 1, \dots, T+1\}$ 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 $\min_{x \in \mathbb{R}^d} f(x)$.

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

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

3.3 Lipschitz continuous and strongly convex functions

We say that a function is strongly convex in the ℓ_2-norm 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 we have

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

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

Theorem 9.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz continuous and α-strongly convex in the ℓ_2-norm, and let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by gradient descent with step size*

$$\eta_t = \frac{2}{\alpha(t+1)}$$

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

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

3.4 Smooth and strongly convex functions

If function f is β-smooth and α-strongly convex in the ℓ_2-norm, 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 9.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and α-strongly convex in the ℓ_2-norm, and let $\{x_t : t = 1, \ldots, T+1\}$ be the sequence of iterates generated by gradient descent with sep size*

$$\eta_t = \frac{2}{\alpha + \beta}$$

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 a linear rate of convergence, and after $T = O(\log(1/\epsilon))$ iterations, we have

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