1 Outline

In this lecture, we study

- Introduction to gradient descent II,
- Gradient descent for smooth functions

2 Introduction to gradient descent

2.1 Backtracking line search

Before we describe the backtracking line search procedure, we characterize descent directions in terms of the gradient. If f is differentiable, we have

$$\lim_{\eta \to 0+} \frac{f(x+\eta d) - f(x)}{\eta} = d^\top \nabla f(x) \tag{8.1}$$

as the limit exists. Then $\nabla f(x)^\top d$ measures the rate of decrease of f in direction d at x.

Moreover, the following lemma directly follows from (8.1) that holds for differentiable functions.

Lemma 8.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then a nonzero vector $d \in \mathbb{R}^d \setminus \{0\}$ is a descent direction if and only if*

$$\nabla f(x)^\top d < 0.$$

For example, $-\nabla f(x)$ is a descent direction at any x.

Based on the characterization of descent directions in Lemma 8.1, we do backtracking line search described as follows.

1. Fix parameters $0 < \beta < 1$ and $0 < \alpha < 1$.
2. Start with an initial step size $\eta > 0$.
3. Until the following condition is satisfied, we shirink $\eta \leftarrow \beta\eta$.
$$f(x + \eta d_t) < f(x) + \alpha\eta \nabla f(x)^\top d_t.$$
4. We take the final η and set $\eta_t = \eta$.

2.2 Gradient descent method

The *steepest direction* of a differentiable function f at a point x can be defined as

$$\arg\min \left\{ \nabla f(x)^\top d : \ \|d\|_2 = 1 \right\} = \{-\nabla f(x)\}.$$

Basically, the steepest direction, which is the direction opposite to the gradient, is the one with the highest rate of decrease of f at x. Then using $-\nabla f$ for a descent direction at any point of the descent method, we obtain the following algorithm, which is commonly known as gradient descent.

Algorithm 1 Gradient descent method

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

Example 8.2. We consider $f(x) = 2x^2 + 3x : \mathbb{R} \to \mathbb{R}$. We already know that the minimizer of f is given by $x^* = -3/4$, but we apply gradient descent to obtain the same conclusion. Let us take an arbitrary initial point x_1. For now, we use a constant step size, i.e. $\eta_t = \eta$ for any $t \geq 1$.

$$\begin{aligned}
x_{t+1} &= x_t - \eta \nabla f(x_t) \\
&= x_t - \eta(4x_t + 3) \\
&= (1 - 4\eta)x_t - 3\eta \\
&= (1 - 4\eta)((1 - 4\eta)x_{t-1} - 3\eta) - 3\eta \\
&= (1 - 4\eta)^2 x_{t-1} - 3\eta((1 - 4\eta) + 1) \\
&\;\;\vdots \\
&= (1 - 4\eta)^t x_1 - 3\eta \sum_{i=0}^{t-1} (1 - 4\eta)^i \\
&= (1 - 4\eta)^t x_1 - 3\eta \cdot \frac{1 - (1 - 4\eta)^t}{1 - (1 - 4\eta)} \\
&= (1 - 4\eta)^t \left(x_1 + \frac{3}{4} \right) - \frac{3}{4}.
\end{aligned}$$

Hence, as long as $|1 - 4\eta| < 1$, x_t converges to $-3/4$. Note that

$$f(x_{T+1}) - f(x^*) = O((1 - 4\eta)^T).$$

Here, the convergence rate is $(1 - 4\eta)^T$, so the error term exponentially decreases. Therefore, after $T = O(\log(1/\epsilon))$ iterations, we obtain

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

This is often called a "linear convergence". Here, the term "linear" means that the required number of iterations is linear in $\log(1/\epsilon)$.

2.3 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.

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

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. Here, 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 Example 8.2, 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 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 8.3. *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 (Algorithm 1) with $\eta_t = 1/\beta$ for each t. Then*

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

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

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