1 Outline

In this lecture, we study

- introduction to gradient descent,
- convergence analysis of gradient descent for Lipschitz continuous functions, and
- subgradients and subgradient method.

2 Introduction to gradient descent

2.1 Generic descent method

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. Given a point $x \in \mathbb{R}^d$, we say that a nonzero vector $d \in \mathbb{R}^d \setminus \{0\}$ is a *descent direction* of f at x if there exists some $\epsilon > 0$ such that

$$f(x + \eta d) < f(x)$$

for any $0 < \eta \leq \epsilon$.

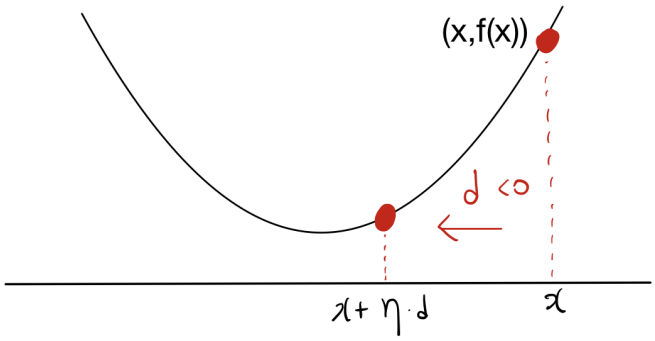

Figure 8.1: Illustration of descent directions

Hence, moving towards a descent direction d can decrease the function value, but how much we move along the direction, captured by η, is important. We often call η a *step size*. Based on descent directions and proper step sizes, we may develop the following algorithm for minimizing a function.

Algorithm 1 Generic descent method

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

Whether the descent method, given by Algorithm 1, converges or not depends on how we choose the step sizes η_t for $t \geq 1$.

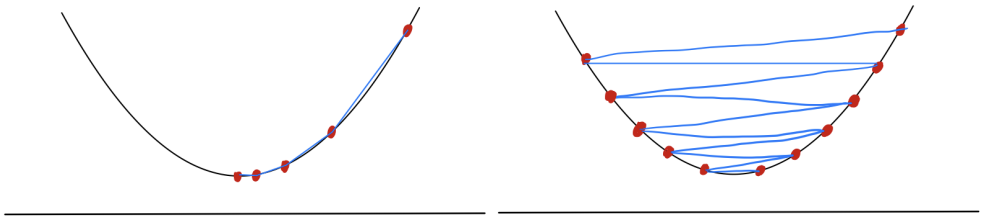

Figure 8.2: Different sequences of step sizes and convergence behavior

Exact line search We choose the step size η_t as

$$\eta_t = \operatorname*{argmin}_{\eta \geq 0} f(x_t + \eta d_t).$$

Here, choosing the step size this way requires solving an optimization problem, which is often an expensive procedure.

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 change in f along 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*

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

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

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

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 2 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 $x_t \in \text{dom}(f)$, 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 8.3. *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 8.3. 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 $t = 1, \ldots, T$ 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 Subgradients

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 8.4. 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 in Figure 8.3, 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.

$$\begin{aligned}\partial f(x) &= \begin{cases}\{-1\} = \{\mathrm{sign}(x)\}, & \text{for } x < 0 \\ [-1, 1], & \text{for } x = 0 \\ \{+1\} = \{\mathrm{sign}(x)\}, & \text{for } x > 0\end{cases} \\ &= \begin{cases}\{\mathrm{sign}(x)\}, & \text{for } x \neq 0 \\ [-1, 1], & \text{for } x = 0.\end{cases}\end{aligned}$$

Let us consider a few more examples.

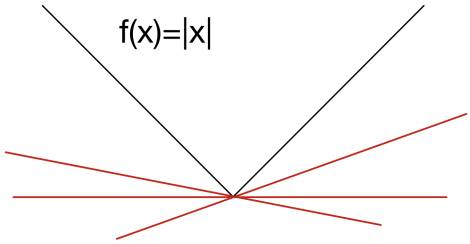

Figure 8.3: Subgradients of $f(x) = |x|$ at $x = 0$

Example 8.5. 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 $i \in [d]$,

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

Example 8.6. Let $f_1, \ldots, f_k$ be convex functions, and let f be defined as the pointwise maximum of $f_1, \ldots, f_k$. Given a point x, if $f(x) = f_i(x)$ for some $i \in [k]$, then any subgradient of f_i is a subgradient of f.

Example 8.7. Given a convex set $C \subseteq \mathbb{R}^d$, the indicator function $I_C(x)$ at a point $x \in \mathbb{R}^d$ is defined as

$$I_C(x) = \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{if } x \notin C \end{cases}.$$

For a point $x \in C$, what is the subdifferential of the indicator function at x? Note that

$$\partial I_C(x) = \left\{ g \in \mathbb{R}^d : \ 0 \geq 0 + g^\top (y - x) \ \ \forall y \in C \right\} = N_C(x)$$

where $N_C(x)$ denotes the normal cone of C at x. Therefore, the subdifferential of the indicator function for C at x is precisely the normal cone of C at x.

3.3 Subgradient method

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 3 Subgradient method

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

We will show that the subgradient method given by Algorithm 3 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 8.8. *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 $\{x_t : t = 1, \dots, T\}$ be the sequence of iterates generated by the subgradient method with step size*

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

for each t. Then

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

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

Proof. Note that

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

where the inequality follows from $f(x^*) \ge f(x_t) + g_t^\top (x^* - x_t)$ as g_t is a subgradient at x_t. The rest of the proof is the same as the argument used for the differentiable case. □

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

The important property of the subgradient method is that it is "dimension-free" in the sense that the algorithm and the convergence rate do not depend on the ambient dimension d. In many applications, we have a moderate tolerance for the error ϵ while the dimension d is huge. For such applications, the fact that the subgradient method is dimension-free has a huge advantage.