1 Outline

In this lecture, we cover

- optimality conditions for convex minimization,
- introduction to gradient descent,
- mean squared error minimization for linear regression,
- convergence of gradient descent for Lipschitz continuous functions.

2 Optimality Conditions for Convex Minimization

2.1 Local Optimality Implies Global Optimality

A feasible solution x^* is **locally optimal** to the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array} \tag{P}$$

if there exists $R > 0$ such that

$$f(x^*) = \min\left\{f(x) : \ x \in C, \ \|x - x^*\| \leq R\right\}.$$

Theorem 3.1. *Any locally optimal solution to a convex optimization problem is (globally) optimal.*

Proof. Suppose for a contradiction that a locally optimal solution x^* to a convex optimization problem $\min_{x \in C} f(x)$ is not globally optimal. Then there exists $y \in C$ such that $f(y) < f(x^*)$. By the local optimality of x^*, there exists $R > 0$ such that $f(x^*) = \min\{f(x) : x \in C, \ \|x - x^*\| \leq R\}$, which implies that $\|y - x^*\| > R$. Let z be defined as

$$z = x^* + \frac{R}{\|y - x^*\|}(y - x^*) = \left(1 - \frac{R}{\|y - x^*\|}\right) x^* + \frac{R}{\|y - x^*\|} y.$$

Since z is a convex combination of x^* and y, it follows that $z \in C$ and

$$f(z) \leq \frac{R}{\|y - x^*\|} f(y) + \left(1 - \frac{R}{\|y - x^*\|}\right) f(x^*) < f(x^*)$$

However, we have $\|z - x^*\| = R$, contradicting the assumption that $f(x^*) = \min\{f(x) : x \in C, \ \|x - x^*\| \leq R\}$. □

For nonconvex problems, a locally optimal solution is not necessarily an optimal solution, illustrated in Figure 3.1.

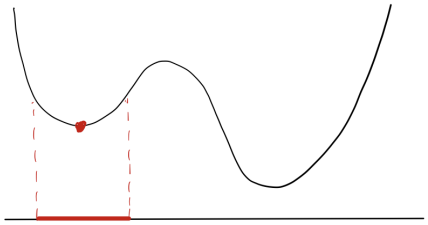

Figure 3.1: Local optimal solution that is not optimal

2.2 First-Order Optimality Condition

Next we establish an optimality condition for convex optimization problems with a differentiable objective.

Theorem 3.2. *For a convex optimization problem of the form* (P) *with f differentiable, $x^* \in C$ is an optimal solution if and only if*

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \textit{for all } x \in C.$$

Figure 3.2 describes the optimality conditions for functions from $\mathbb{R}^2$ to $\mathbb{R}$. Basically, a solution x^* is optimal if we cannot move further from x^* in C in the direction of decreasing f. If $\nabla f(x^*) = 0$, then x^* is optimal.

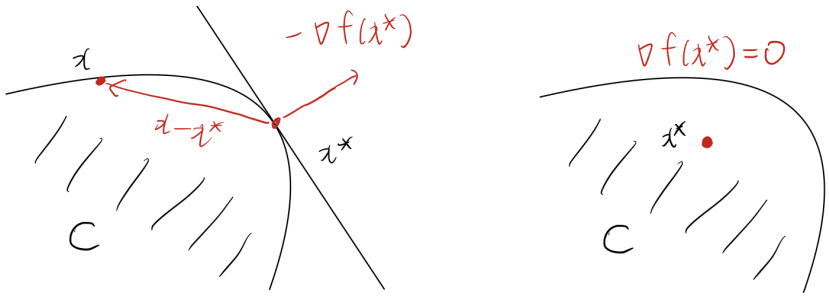

Figure 3.2: Optimality of bi-variate convex functions

By Theorem 3.2, a sufficient condition for optimality is that $\nabla f(x^*) = 0$. This, in fact, is a necessary and sufficient condition for the unconstrained case.

Theorem 3.3. *$x^* \in \mathbb{R}^d$ is optimal to $\min_{x \in \mathbb{R}^d} f(x)$ if and only if*

$$\nabla f(x^*) = 0.$$

Proof. ($\Leftarrow$) If $\nabla f(x^*) = 0$, then it trivially holds that $\nabla f(x^*)^\top (x - x^*) \geq 0$ for $x \in \mathbb{R}^d$. Then x^* is optimal due to Theorem 3.2.

($\Rightarrow$) Let $x = x^* - \alpha \nabla f(x^*)$. Then by Theorem 3.2, we have

$$\nabla f(x^*)^\top (x - x^*) = -\alpha \|\nabla f(x^*)\|_2^2 \geq 0.$$

This in turn implies that $\|\nabla f(x^*)\|_2 = 0$ and thus $\nabla f(x^*) = 0$. □

Figure 3.3 describes the optimality conditions for functions from $\mathbb{R}$ to $\mathbb{R}$.

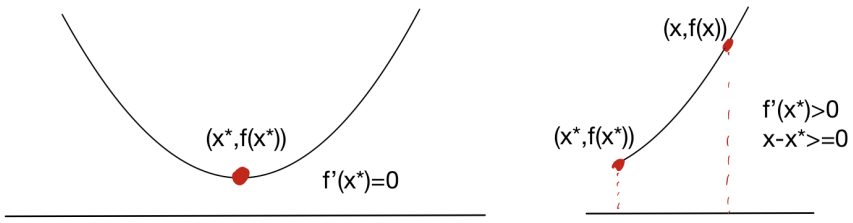

Figure 3.3: Optimality of univariate convex functions

3 Introduction to Gradient Descent

3.1 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 $0 < \eta \leq \epsilon$.

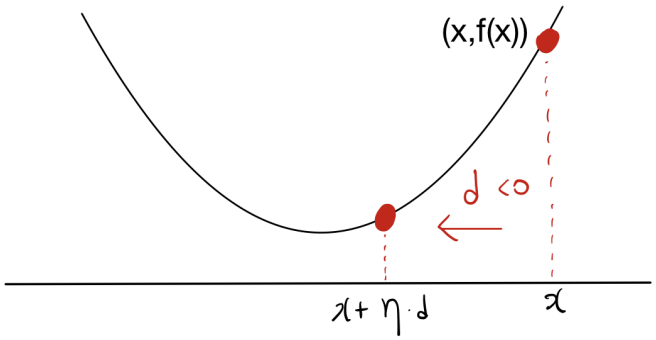

Figure 3.4: Illustration of descent directions

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. Here, η is referred to as a **step size** or as a **learning rate**. Based on descent directions and proper step sizes, we may develop the following algorithm for minimizing a function.

Algorithm 1 Generic descent method

Initialize $x_1 \in \text{dom}(f)$.
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 $t \geq 1$.

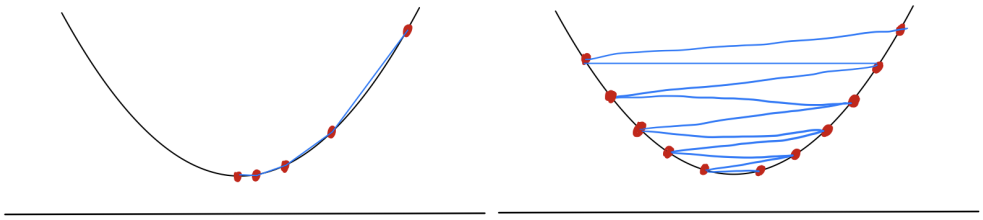

Figure 3.5: Different sequences of step sizes and convergence behavior

We characterize descent directions in terms of the gradient. If f is differentiable, we have

$$\lim_{\eta \to 0+} \frac{f(x+\eta d) - f(x)}{\eta} = d^\top \nabla f(x) \tag{3.1}$$

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 (3.1) that holds for differentiable functions.

Lemma 3.4. *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.

3.2 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

4 Mean Squared Error Minimization for Linear Regression

As discussed in a previous lecture, linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is commonly used for prediction and inference tasks in various fields such as economics, finance, and machine learning.

In linear regression, we have a vector $x \in \mathbb{R}^d$ of predictor variables and one response variable y. The relationship between x and y is modeled using a linear equation as follows.

$$y = \theta_0 + \theta_1^\top x + \epsilon$$

where:

- $\theta_0 \in \mathbb{R}$ is the bias term,
- $\theta_1 \in \mathbb{R}^d$ is the coefficient vector,
- $\epsilon \in \mathbb{R}$ is the error term representing unexplained variation.

Here, the parameters θ_0 and θ_1 are estimated using the method of least squares, which minimizes the average of squared differences between the observed and predicted values of y. Namely, given a set of n data $(x_1, y_1), \ldots, (x_n, y_n)$, we solve

$$\min_{\theta_0, \theta_1} \quad \frac{1}{n} \sum_{i=1}^{n} \left(\underbrace{y_i}_{\text{observed}} - \underbrace{(\theta_0 + \theta_1^\top x_i)}_{\text{predicted}} \right)^2, \tag{3.2}$$

which we refer to as the mean squared error minimization. Then a solution (θ_0, θ_1) gives rise to a model $y = \theta_0 + \theta_1^\top x$ as described in Figure 3.6.

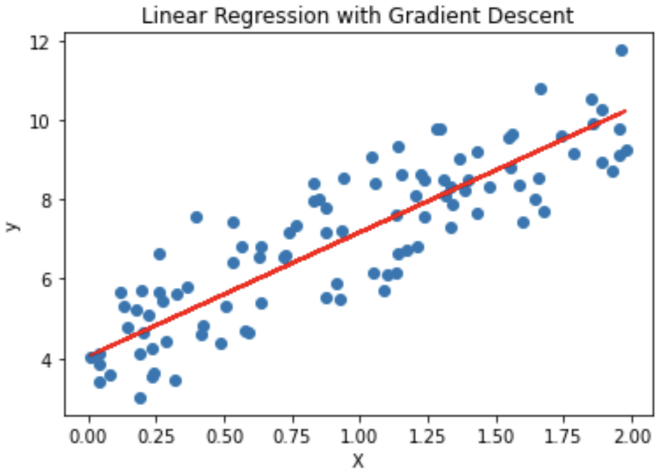

Figure 3.6: Linear regression example

To solve the optimization problem (3.2), we apply gradient descent. Let Y denote the vector whose components are $y_1, \ldots, y_n$, let X denote the matrix whose rows are $x_1^\top, \ldots, x_n^\top$, and let $\mathbf{1}$ denote the vector of all ones. Then it follows that

$$f(\theta_0, \theta_1) := \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - \theta_1^\top x_i \right)^2 = \frac{1}{n} \| Y - \mathbf{1} \cdot \theta_0 - X\theta_1 \|_2^2.$$

Then let us apply gradient descent to solve the optimization problem. Let $f(\theta_0, \theta_1)$ denote the mean squared error as a function of θ_0, θ_1. Then

$$\nabla f(\theta_0, \theta_1) = \left(-\frac{2}{n} \cdot \mathbf{1}^\top \left(Y - \mathbf{1} \cdot \theta_0 - X\theta_1 \right), \quad -\frac{2}{n} \cdot X^\top \left(Y - \mathbf{1} \cdot \theta_0 - X\theta_1 \right) \right).$$

As we run gradient descent, we update (θ_0^t, θ_0^t) where t denotes the tth iteration of gradient descent. Here, the loss of the current solution (θ_0^t, θ_0^t) is simply given by $f(\theta_0^t, \theta_0^t)$. Figure 3.7 shows how the loss decreases as iterations proceed.

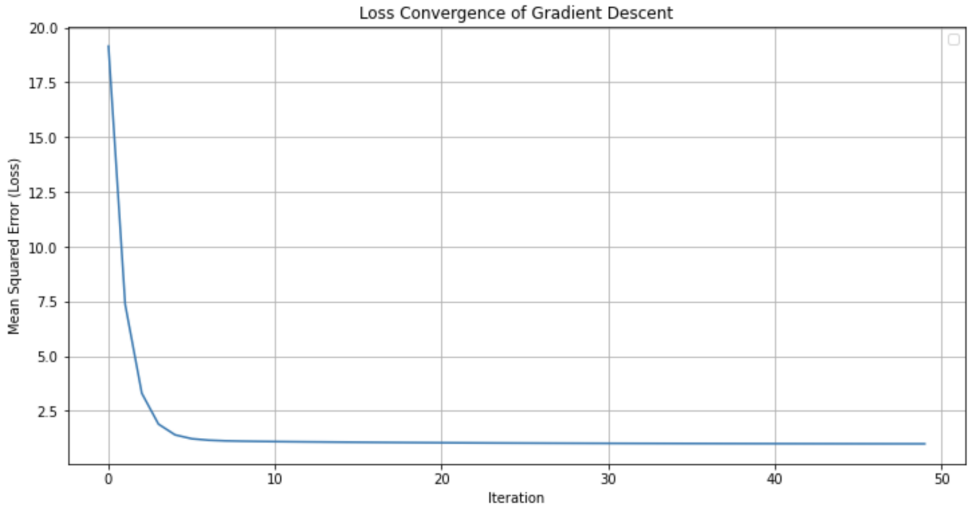

Figure 3.7: Convergence of gradient descent

The last remark of this section is about choosing right learning rates for gradient descent. Figure 3.7 shows the convergence of gradient descent under a constant learning rate of 0.1. In fact, the convergence pattern can vary greatly depending on the learning rate. As the figure suggests,

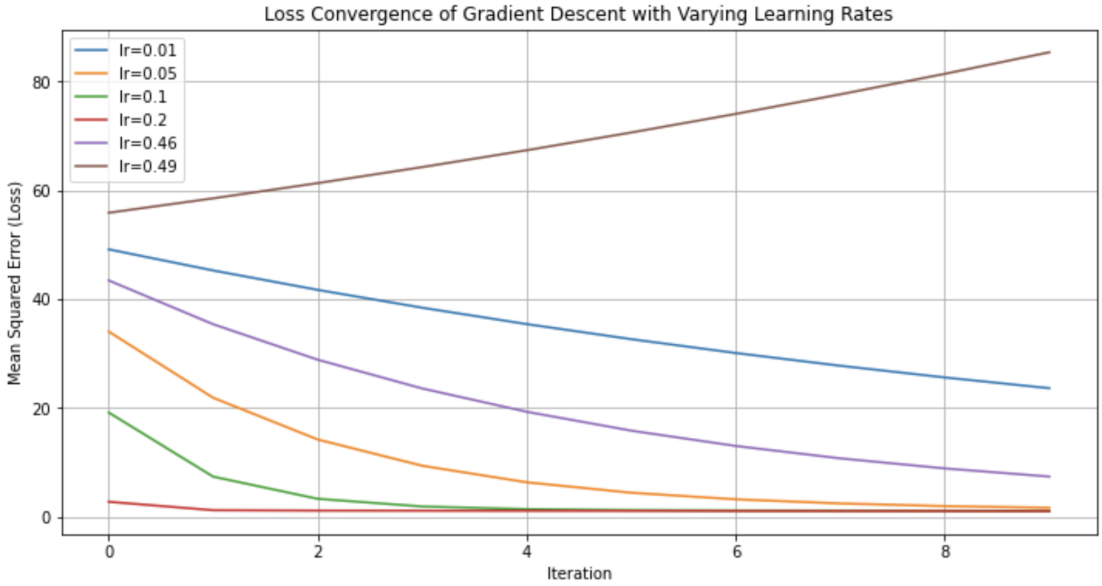

Figure 3.8: Convergence of gradient descent under varying learning rates

choosing right step sizes is crucial for the performance of gradient descent.

5 Convergence of Gradient Descent for Lipschitz Continuous Functions

In this section, we consider the convergence of gradient descent for Lipschitz continuous functions. What is important to guarantee convergence, we need to choose proper step sizes (equivalently, learning rates).

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 3.5. *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 3.5. 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. □