IE 539: Convex Optimization **KAIST, Fall 2023**
Lecture #8: Optimality conditions & Intro to gradient descent September 20, 2023
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- Optimality conditions for convex minimization,
- Normal cones and projection,
- Introduction to gradient descent,

2 Optimality conditions for convex minimization

2.1 Local optimality implies global optimality

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

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

if there exists $R > 0$ such that

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

Theorem 8.1. *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 8.1.

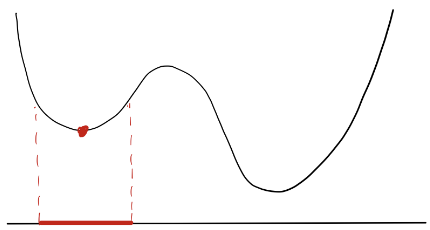

Figure 8.1: Local optimal solution that is not optimal

2.2 First-order optimality condition

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

Theorem 8.2. *For a convex optimization problem of the form* (P) *with f differentiable, $x^* \in C$ is an optimal solution if and only if*

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

We will prove this later in the course, when we discuss the general case allowing nondifferentiable objectives. Figure 8.2 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.

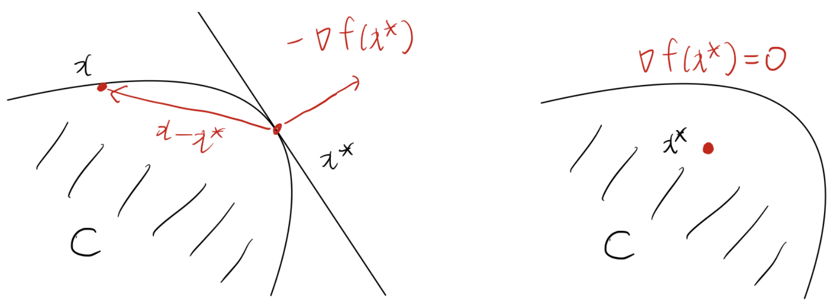

Figure 8.2: Optimality of bi-variate convex functions

By Theorem 8.2, 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 8.3. *$x^* \in \mathbb{R}^d$ 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 8.2.

($\Rightarrow$) Let $x = x^* - \alpha \nabla f(x^*)$. Then by Theorem 8.2, 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 $\nabla f(x^*) = 0$. $\square$

Figure 8.3 describes the optimality conditions for functions from $\mathbb{R}$ to $\mathbb{R}$.

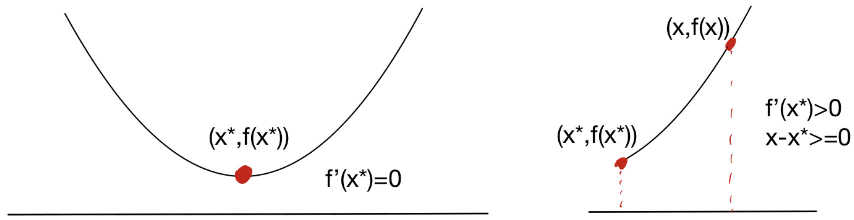

Figure 8.3: Optimality of univariate convex functions

Example 8.4. Consider the following equality-constrained problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

where f is convex and A, b are matrices of appropriate dimensions. Then a solution x^* is optimal if and only if $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all x such that $Ax = b$. Note that the latter condition is equivalent to $\nabla f(x^*)^\top v = 0$ for all v in the null space of A. Since the orthogonal complement of null(A) is the column space of $A^\top$, we have $\nabla f(x^*) = A^\top u$ for some u.

The *normal cone* of C at $x \in C$ is defined as

$$N_C(x) = \{g \in \mathbb{R}^d : \ g^\top (y - x) \leq 0 \text{ for all } y \in C\}.$$

Figure 8.4 shows some examples.

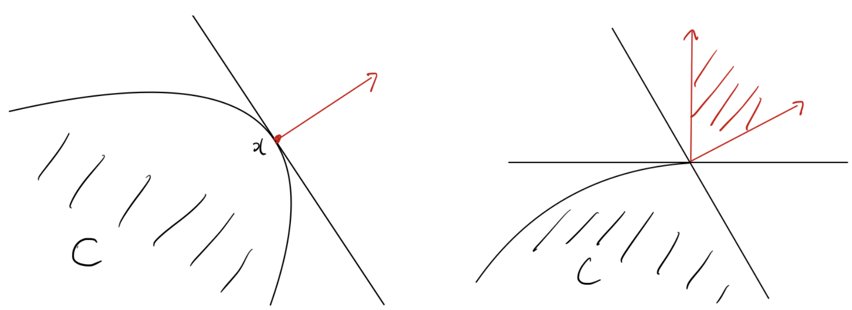

Figure 8.4: Optimality of univariate convex functions

Then the optimality condition in Theorem 8.2 is equivalent to

$$-\nabla f(x^*) \in N_C(x^*) \quad \leftrightarrow \quad 0 \in \nabla f(x^*) + N_C(x^*).$$

Later in the course, we will give a direct proof for this equivalent condition.

2.3 Projection

We consider the problem of projecting a point p onto a convex set C, that is to find a point $x \in C$ minimizing the distance to p.

$$\begin{array}{ll} \text{minimize} & \|x - p\|_2^2 \\ \text{subject to} & x \in C \end{array}$$

Let $\text{Proj}_C(p)$ denote the projection of p to C[1]. By definition, $\text{Proj}_C(p)$ is an optimal solution to the optimization problem. Note that the gradient of $\|x-p\|_2^2$ is

$$2(x-p).$$

As $\text{Proj}_C(p)$ is the optimal solution to the above optimization problem, it follows from Theorem 8.2 that

$$2(\text{Proj}_C(p)-p)^\top(x-\text{Proj}_C(p)) \geq 0 \quad \text{for all } x \in C.$$

Equivalently,

$$\langle \text{Proj}_C(p)-p,\ \text{Proj}_C(p)-x\rangle \leq 0 \quad \text{for all } x \in C.$$

Next let us consider two points u, v and their projections onto C, given by $\text{Proj}_C(u)$ and $\text{Proj}_C(v)$, respectively. Then we have

$$\begin{aligned}\langle \text{Proj}_C(u)-u,\ \text{Proj}_C(u)-\text{Proj}_C(v)\rangle &\leq 0,\\ \langle \text{Proj}_C(v)-v,\ \text{Proj}_C(v)-\text{Proj}_C(u)\rangle &\leq 0.\end{aligned}$$

Adding these two inequalities, we obtain

$$\|\text{Proj}_C(u)-\text{Proj}_C(v)\|_2^2 - \langle u-v,\ \text{Proj}_C(u)-\text{Proj}_C(v)\rangle \leq 0.$$

Then it follows from the Cauchy-Schwarz inequality that

$$\|\text{Proj}_C(u)-\text{Proj}_C(v)\|_2 \leq \|u-v\|_2.$$

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 $0<\eta\leq\epsilon$.

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 $x_1 \in \text{dom}(f)$.
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

[1] In fact, there exists a unique optimal solution to the above optimization problem. Why?

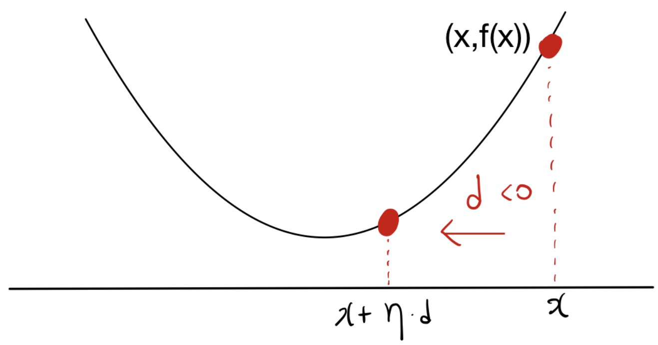

Figure 8.5: Illustration of descent directions

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

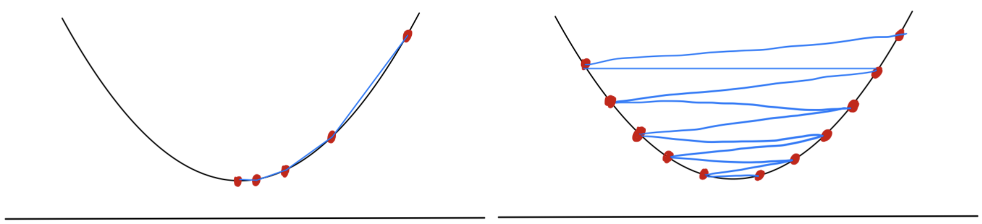

Figure 8.6: Different sequences of step sizes and convergence behavior

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 8.5. *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 8.5, 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 shrink $\eta \leftarrow \beta\eta$.

$$f(x + \eta d_t) < f(x) + \alpha\eta \nabla f(x)^\top d_t.$$

4. We take the final η and set $\eta_t = \eta$.

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

Example 8.6. 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 $\log(1/\epsilon)$.