| **EN.553.741: Machine Learning** | **Spring 2023** |
|---|---|
| Review on Optimization 1: Feb. 23 & Mar. 2 | |
| *Lecturer: James Schmidt* | *TA: Ao Sun* |

## 1.1 Unconstrained Optimization

In unconstrained optimization problem, we minimize an objective function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ depending on a variable $x \in \mathbb{R}^d$ without restrictions on $x$. The mathematical formulation is

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x). \tag{1.1}$$

We are only interested in those functions $f$ which are bounded below, otherwise the minimum of $f$ is $-\infty$.

**Definition 1.1.** *A point $x^\star \in \mathbb{R}^d$ is called a **global minimizer** if $f(x^\star) \le f(x)$ for all $x \in \mathbb{R}^d$. A point $x^\star$ is a **local minimizer** if there is a neighborhood $U$ of $x^\star$ such that $f(x^\star) \le f(x)$ for all $x \in U$.*

Even if a function is bounded below, the global minimizer may not exist, e.g., the function $f(x) = e^x$ for $x \in \mathbb{R}$ is bounded below, since it is nonnegative. However, the infimum of $f$ is 0 which cannot be attained by any $x \in \mathbb{R}$. The global minimizer is usually difficult to find, because our knowledge of $f$ is usually only local. Most algorithms in practice only aim at finding local minimizer.

**Definition 1.2.** *A point $x^\star$ is a **strict local minimizer** if there is a neighborhood $U$ of $x^\star$ such that $f(x^\star) < f(x)$ for all $x \in U \setminus \{x^\star\}$. A point $x^\star$ is an **isolated local minimizer** if there is a neighborhood $U$ of $x^\star$ such that $x^\star$ is the* only *minimizer in $U$.*

The strict local minimizer are not necessarily isolated, e.g., the function $f$ defined by

$$f(x) := \begin{cases} x^4 \cos(1/x) + 2x^4 & \text{if } x \neq 0; \\ 0 & \text{if } x = 0 \end{cases}$$

is of class $C^2(\mathbb{R})$ and has a strict local minimizer at $x^\star = 0$. And there are strict local minimizers at many points of $x_j$, these points are labeled such that $x_j \downarrow 0$ as $j \to \infty$. However, it is easy to see that all isolated local minimizers are strict.

If the function $f$ is sufficiently smooth, we can exploit its gradient and Hessian to tell whether or not $x^\star$ is a local minimizer.

**Theorem 1.3.** *Let $f \in C^1(\mathbb{R}^d)$ and $p \in \mathbb{R}^d$. Then we have*

$$f(x+p) = f(x) + \nabla f(x+tp)^\top p \tag{1.2}$$

*for some $t \in (0,1)$. Moreover, if $f \in C^2(\mathbb{R}^d)$, we have that*

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)p\,, \tag{1.3}$$

*and that*

$$f(x+p) = f(x) + \nabla f(x)^\top p + \frac{1}{2} p^\top \nabla^2 f(x+tp) p, \tag{1.4}$$

*for some $t \in (0,1)$.*

Necessary conditions for optimality are derived by assuming that $x^\star$ is a local minimizer and then proving facts about $\nabla f(x^\star)$ and $\nabla^2 f(x^\star)$.

**Theorem 1.4** (First-order Necessary Conditions)**.** *If $x^\star$ is a local minimizer and $f$ is continuously differentiable in an open neighborhood of $x^\star$, then $\nabla f(x^\star) = 0$.*

**Theorem 1.5** (Second-order Necessary Conditions)**.** *If $x^\star$ is a local minimizer of $f$, and $\nabla^2 f$ exists and is continuous in an open neighborhood of $x^\star$, then $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive semidefinite.*

The necessary conditions in Theorem 1.5 are not sufficient. For example, let $f(x) = x^3$, then $f'(0) = 0$ and $f''(0) = 0$. But $f$ is not a local minimizer at $x = 0$. We now describe sufficient conditions under which $x^\star$ is a local minimizer.

**Theorem 1.6.** *Suppose that $\nabla^2 f$ exists and is continuous in an open neighborhood of $x^\star$, and that $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is* positive definite*. Then $x^\star$ is a strict local minimizer of $f$.*

Note too that the second-order sufficient conditions are not necessary: A point $x^\star$ may be a strict local minimizer, and may fail to satisfy the sufficient conditions. A simple example is given by the function $f(x) = x^4$, for which the point $x^\star = 0$ is a strict local minimizer at which the Hessian vanishes.

All algorithms for unconstrained optimization problem require the user to supply a initial point, which is usually denoted by $x_0$. Beginning at $x_0$, the optimization algorithms generate a sequence of iterates $(x_k)_{k\in\mathbb{N}_0}$ that terminate when

- no more progress can be made;
- a solution has been approximated with sufficient accuracy.

## 1.2 Convex Analysis

Our aim is to introduce basic concepts in convex analysis. The space $V$ is a normed vector space equipped with norm $\|\cdot\|$ unless otherwise specified.

**Definition 1.7.** *A subset $\Omega$ of $V$ is called* ***convex*** *if*

$$\alpha x + (1-\alpha)y \in \Omega$$

*for all $x, y \in \Omega$ and $\alpha \in [0,1]$.*

The empty set is by convention considered to be convex. The following proposition gives some operations that preserve convexity of a set.

**Proposition 1.8.**

1. *Let $I$ be an arbitrary index set. If $(A_i)_{i\in I}$ is a sequence of convex sets, then $\bigcap_{i\in I} A_i$ is still a convex set.*
2. *The vector same $A_1 + A_2 = \{x + y : x \in A_1, y \in A_2\}$ is a convex set provided $A_1, A_2$ are both convex.*
3. *The scalar multiple of a convex set $\alpha A = \{\alpha x : x \in A\}$ is convex. Moreover we have the set equation*
$$\alpha_1 A + \alpha_2 A = (\alpha_1 + \alpha_2)A$$
*where $\alpha_1, \alpha_2 \in [0, +\infty)$.*
4. *The closure and the interior of a convex set are convex.*
5. *The image and the inverse image of a convex set under an affine function are convex.*

If $a$ is a nonzero vector in $V = \mathbb{R}^d$ and $b \in \mathbb{R}$, then we have the following definition. A **hyperplane** is a set specified by a set of the form $\{x : a^\top x = b\}$. A **halfspace** is a set specified by a set of the form $\{x : a^\top x \le b\}$. It is clearly closed and convex. A set is said to be **polyhedral** if it is nonempty and it is the intersection of a finite number of halfspaces. A polyhedral set is convex and closed, being the intersection of halfspaces. A set $A$ is said to be a **cone** if for all

$x \in A$ and $\alpha > 0$, we have $\alpha x \in A$. A cone may not be convex and may not contain the origin, although the origin always lies in the closure of a non-empty cone. A **polyhedral cone** is a set of the form

$$\bigcap_{i=1}^{n} \{x : a_i^\top x \leq 0\}$$

where $a_1, ..., a_n \in \mathbb{R}^d$. A subspace is a special case of a polyhedral cone, which is in turn a special case of a polyhedral set.

**Definition 1.9.** *Let $\Omega$ be a convex subset of $V$. A function $f : \Omega \to (-\infty, +\infty)$ is called **convex** if*

$$f(\alpha x + (1-\alpha)y)) \leq \alpha f(x) + (1-\alpha)f(y)$$

*for all $x, y \in \Omega$ and $\alpha \in [0,1]$. A function $f$ is called strictly convex if the inequality above is strict. A function $f$ is **concave** if $-f$ is convex.*

Common examples of convex functions are affine functions and $\|\cdot\|$, etc. If $f : \Omega \to \mathbb{R}$ is a function and $\lambda \in \mathbb{R}$ is a scalar, the sets $\{x \in \Omega : f(x) \leq \lambda\}$ and $\{x \in \Omega : f(x) < \lambda\}$ are called the **level sets** of $f$. If $x, y \in \Omega$ with $f(x) \leq \lambda$ and $f(y) \leq \lambda$ for some convex function $f$, then for every $\alpha \in [0,1]$, we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \leq \lambda,$$

by the convexity of $f$. So the level sets for convex functions are convex. However the converse is not true as illustrated in the following example.

**Example 1.10.** *$f(x) = \sqrt{|x|}$ is not a convex function, and $\{x : f(x) \leq \lambda\} = \{x : |x| \leq \lambda^2\}$ is convex for sure.*

In the context of optimization and duality, we will encounter operations resulting functions taking infinite values. For example, the the supremum of a infinite collection of functions. We thus are motivated to consider *extended real-valued* functions that can take the values $\pm\infty$ at some points. Such functions can be charaterized using the notion of epigraph discussed below.

**Definition 1.11.** *The **epigraph** of a function $f : \Omega \to [-\infty, +\infty]$, where $\Omega$ is a subset of $V$, is defined to be a subset of $V \times \mathbb{R}$ given by*

$$\operatorname{epi}(f) = \{(x, y) : x \in \Omega, y \in \mathbb{R}, f(x) \leq y\}$$

*the **effective domain** of $f$ is defined to be the set*

$$\operatorname{dom}(f) = \{x \in \Omega : f(x) < +\infty\} \tag{1.5}$$

*We say that $f$ is **proper** if $f(x) < +\infty$ for at least one $x \in \Omega$,* i.e. , *$dom(f) \neq \emptyset$, and $f(x) > -\infty$ for all $x \in \Omega$,* i.e. , epi$(f)$ *does not contain a vertical line, we call $f$ is **improper** otherwise.*

In order to avoid $\infty - \infty$ in the definition for an improper convex function, it is convenient to define convex extended real-valued function by its epigraph

**Definition 1.12.** *Let $\Omega$ be a convex subset of $V$. We say that an extended real-valued function $f : \Omega \to [-\infty, +\infty]$ is convex if* epi$(f)$ *is a convex subset of $V \times \mathbb{R}$.*

It can be easily shown that, according to Definition 1.12, convexity of $f$ implies that its effective domain dom$(f)$ and its level sets

$$\{x \in \Omega : f(x) \leq \lambda\}, \quad \{x \in \Omega : f(x) < \lambda\}$$

are convex sets for all scalars $\lambda \in \mathbb{R}$.

**Exercise 1.13.** *Prove the assertions above.*

Definition 1.12 is consistent with the earlier definition of convexity for real-valued functions.

**Exercise 1.14.** *If $f(x) < \infty$ for all $x \in \Omega$, or $f(x) > -\infty$ for all $x \in \Omega$, then*

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y)$$

*for all $x, y \in \Omega$ and $\alpha \in [0,1]$. In particular, Definition 1.12 is consistent with Definition 1.9.*

By passing to epigraphs, we can use results about sets to infer corresponding results about functions. The reverse is also possible, via function $\delta_\Omega : V \to (-\infty, +\infty]$ of a set $\Omega \subset V$, defined by

$$\delta_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ \infty & \text{otherwise.} \end{cases} \tag{1.6}$$

A set $A$ is convex if and only if $\delta_A$ is a convex function, and $A$ is nonempty if and only if $\delta_A$ is proper. If the epigraph of a function $f : \Omega \to [-\infty, +\infty]$ is a closed set, we say that $f$ is a **closed function**. Closedness is related to the classical notion of lower semi-continuity.

**Definition 1.15.** *An extended real-valued function $f$ is called **lower semicontinuous** at a $x \in \Omega$ if*

$$f(x) \le \liminf_{k\to\infty} f(x_k)$$

*for every sequence $\{x_k\}_{k=1}^\infty \subset \Omega$ with $x_k \to x$. We say that $f$ is **upper semicontinuous** if $-f$ is lower semicontinuous.*

The following proposition connects closedness, lower semicontinuity, and closedness of the level sets of a function.

**Proposition 1.16.** *For an extended real-valued function $f : V \to [-\infty, +\infty]$, the following are equivalent*

1. *The level set $\{x : f(x) \le \lambda\}$ is closed for every $\lambda \in \mathbb{R}$.*
2. *$f$ is lower semicontinuous.*
3. $\operatorname{epi}(f)$ *is a closed set.*

**Corollary 1.17.** *The level set $\{x : f(x) < \lambda\}$ is open for every $\lambda \in \mathbb{R}$ if and only if $f$ is upper semicontinuous.*

In practice, we prefer to use the closeness notion, rather than lower semicontinuity. One reason is that contrary to closeness, lower semicontinuity is a domain dependent property. For example $\delta_{(0,1)}$ is neither closed or lower semicontinuous on $V = \mathbb{R}$; but if its domain is restricted to $(0,1)$ it becomes lower semicontinuous.

The local minimum of a convex function is also a global minimum.

**Theorem 1.18.** *If $\Omega$ is a convex subset of $V$ and $f : V \to (-\infty, \infty]$ is a convex function, then a local minimum of $f$ over $\Omega$ is also a global minimum. If $f$ is strictly convex, then there exists at most one global minimum of $f$ over $\Omega$.*