Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 2: Optimization and Convex Analysis

Optimization

In this class we consider the following optimization problem as

$$\min_x f(x), \tag{1}$$

where f is a certain function, and $x \in \mathbb{R}^d$ is the parameter to be optimized.

Constrained Optimization

A generalization is constrained optimization problem below

$$\min_x f(x) \quad \text{subject to } x \in C, \tag{2}$$

where $C \subset \mathbb{R}^d$ is a closed set on x.

Local and Global Solutions

In general, we are interested in optimization algorithms to solve (1) and (2). The solution can be local and global, defined as follows.

Definition

A point $\tilde{x} \in C$ is a local solution of (2) if there exists $\epsilon > 0$ such that for all $x \in C$, $\|x - \tilde{x}\| \leq \epsilon$,

$$f(\tilde{x}) \leq f(x).$$

A point $\tilde{x} \in C$ is a global solution of (2) if for all $x \in C$,

$$f(\tilde{x}) \leq f(x).$$

Convexity

Consider a closed set $C \subset \mathbb{R}^d$, the set is convex if for all $x, y \in C$, and $\forall \alpha \in [0, 1]$,

$$\alpha x + (1 - \alpha) y \in C.$$

Geometrically, this means that the line-segment connecting any two point in C also belongs to C.

Projection

In this course, we are mainly interested in closed convex sets. Given a closed convex set C, and any point y, we can define the projection of y onto C as the closest point to y in C:

$$\mathrm{proj}_C(y) = \arg\min_{x \in C} \|y - x\|_2^2.$$

The projection is uniquely defined.

Separation

If $y \notin C$, then $z = \mathrm{proj}_C(y)$ lies on the boundary of C. The hyperplane $\{x : (y - z)^\top (x - z) = 0\}$ separates y and C in that they lie on different sides of the hyperplane.

Given any z on the boundary of C, we can find a hyperplane passing C such that C is on one side of the hyperplane. This is called a supporting hyperplane, which may not be unique.

Convex Function

A function $f(x) : C \to \mathbb{R}$, defined on a convex set C, is convex if for all $x, y \in C$,

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

The epigraph of a function $f(x) : C \to \mathbb{R}$ is defined as the set $\{(x, u) \in C \times \mathbb{R} : f(x) \leq u\}$. A function $f(x)$ is convex if and only if its epigraph is a convex set.

Convex Optimization

Theorem

Consider a convex function $f(x)$ defined on a convex set C. If $\tilde{x}$ is a local solution of (2), then it is a global solution of (2).

Constructing Convex Set from Convex Functions

We can define convex sets using convex functions.
Given any $g(x) : \mathbb{R}^d \to \mathbb{R}^k$, such that each component $g_j(x)$ is convex, then the set $\{x : g(x) \leq 0\}$ is convex.
We note $\{x : g(x) \leq 0\}$ is the intersection of $\{x : g_j(x) \leq 0\}$ for $j = 1, \ldots, k$.

Properties

In general, the intersection of convex sets is a convex set, and a weighted sum of convex sets is a convex set. The sup over a family of convex functions is convex, and a positively weighted sum of convex functions is convex.
A function $f(x)$ on $\mathbb{R}^d$ is called concave if $-f(x)$ is convex. Linear functions are both convex and concave.

Norm

A norm $\|\cdot\|$ on $\mathbb{R}^d$ is a function that satisfies the following conditions: $\|u+v\| \leq \|u\| + \|v\|$, $\|\rho u\| = |\rho|\|u\|$ for all $\rho \in \mathbb{R}$, and $\|u\| = 0$ if and only if $u = 0$.

Any norm is a convex function.

Given a norm $\|\cdot\|$ on $\mathbb{R}^D$, one can define its dual norm $\|\cdot\|_*$ on $\mathbb{R}^d$ as follows:

$$\|u\|_* = \sup_{\|v\|=1} u^\top v.$$

This inequality implies that $u^\top v \leq \|u\|_* \cdot \|v\|$.

Subgradient

If a funtion $f(x)$ is differentiable, then $f(x)$ is convex if and only if $\forall x, y$:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x). \tag{3}$$

For a convex function $f(x)$, we may define a generalization of gradient called *subgradient as follows*. A vector $g \in \mathbb{R}^d$ is a subgradient of $f(x)$ at x if $\forall y$:

$$f(y) \geq f(x) + g^\top (y - x). \tag{4}$$

A subgradient of a convex function defined on $\mathbb{R}^d$ always exists, but may not be unique. A convex function $f(x)$ is differentiable at x if it has a unique subgradient at x.

Subdifferential

The set of subgradients at x is called subdifferential of $f(x)$, defined as:

$$\partial f(x) = \{g \in \mathbb{R}^d : f(y) \geq f(x) + g^\top (y - x) \ \forall y\}.$$

A convex function $f(x)$ is called *non-smooth* if its subgradient is not unique.

Optimal Solution

The following result characterizes the solution of convex optimization problem.

Theorem

A point $x_* \in C$ is a solution of (2) if and only if there exists a subgradient $g_* \in \partial f(x_*)$, such that $\forall y \in C$:

$$g_*^\top (y - x_*) \geq 0.$$

In particular, x_* is the solution for the unconstrained problem (1) if $0 \in \partial f(x_*)$.

We say that a function $f : C \to \mathbb{R}$ is G-Lipschitz if for all $x, y \in C$:

$$|f(x) - f(y)| \leq G\|x - y\|_2.$$

The smoothness condition is equivalent to the following inequality:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2. \tag{5}$$

we say $f(x)$ is λ-strongly convex

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\lambda}{2}\|y - x\|_2^2, \tag{6}$$

If $f(x)$ is strongly convex, then the solution of (2) is unique.

Examples

In machine learning, we encounter an optimization problem of the form

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) + R(w), \tag{7}$$

where w is the model parameter, $f_i(w)$ is the loss at (X_i, Y_I).

Loss Functions

The following are common loss functions that are convex in u:

- Least squares loss $\phi(u, y) = (u - y)^2$
- Logistic loss $\phi(u, y) = \ln(1 + \exp(uy)$ (where $y \in \{\pm 1\}$)
- Hinge loss $\phi(u, y) = \max(0, 1 - uy)$ (where $y \in \{\pm 1\}$
- Multi-class logistic regression with $y \in \{1, \ldots, k\}$ and $u \in \mathbb{R}^k$, we have $\phi(u, y) = -u_y + \ln \sum_j \exp(u_j)$.

Regularizer

The commonly used convex regularizers are

- L_2: $R(w) = \frac{\lambda}{2}\|w\|_2^2$
- L_1: $R(w) = \lambda\|w\|_1$
- $L_1 - L_2$: $R(w) = \lambda_1\|w\|_1 + \frac{\lambda_2}{2}\|w\|_2^2$
- Trace-norm for matrix w: $R(w) = \lambda\|w\|_*$ (where $\|\cdot\|_*$ is the matrix trace-norm)