Optimization and Convex Analysis

1 Optimization

In this class we consider the optimization problem, which can be written 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. This form of optimization is called unconstrained optimization.

Of particular importance is the finite sum structure with regularization, where we have

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) + R(x).$$

This is the form we encounter in training a machine learning model.

A generalization of (1) is constrained optimization problem below

$$\begin{aligned} &\min_x f(x) \\ &\text{subject to } x \in C, \end{aligned} \tag{2}$$

where $C \subset \mathbb{R}^d$ is a closed set on x. In general, we often define the constraint set as follows:

$$C = \{x : g(x) \leq 0\} \tag{3}$$

for a k-dimensional vector valued function $g(x) : \mathbb{R}^d \to \mathbb{R}^k$, which defines k-constraints. For example, for $k = 3$ and $d = 3$, let $x = [x_1, x_2, x_3] \in \mathbb{R}^d$, we can st $g(x) = [-x_1, x_1, x_1^2 + x_2^2 + x_3^2 - 1]$ defines the set $C = \{x : x_1 = 0, x_2^2 + x_3^2 \leq 1\}$.

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

Definition 1 *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 $x \in C$,

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

For general functions that are not convex, it can be very challenge to find a global optimal solution for these problems. However, it is often easier to find a local solution. For convex functions, local solutions are global solutions. Therefore convex problems can be solved efficiently. Significant progress has been made for convex formulations, where both $f(x)$ and $g(x)$ are convex functions. Theoretical analysis in terms of convergence rate can be obtained for convex formulations. Many algorithms developed for convex problems can also be applied to nonconvex problems, although the theory will become more difficult.

In our lectures, we will illustrate some of the most important algorithms for solving these optimization problems in machine learning applications. In particular, we are interested in first order algorithms that are computationally scalable. We will present these algorithms as well as their convergence analysis, with a focus on convex problems.

2 Convexity

A good reference of convex analysis is [1]. We will introduce basic concepts useful for this class. Consider a set $C \subset \mathbb{R}^d$. We say 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 points in C also belongs to C. This can be illustrated in Figure 1 below.

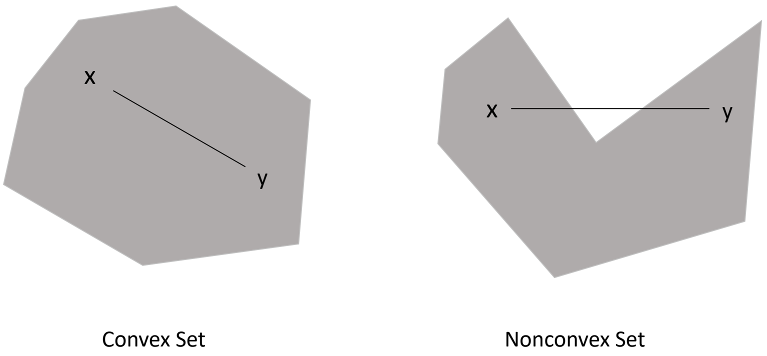


Figure 1: Convex Set versus Nonconvex Set

In this class, 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 point in C that is closest to y:

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

The projection is uniquely defined. If $y \notin C$, then $z = \mathrm{proj}_C(y)$ lies on the boundary of C. See Figure 2.

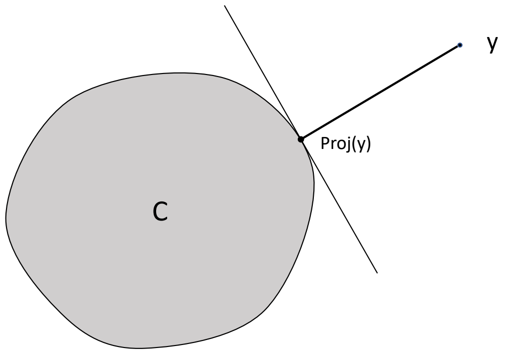


Figure 2: Projection onto Convex Set

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 hyperplane is called a supporting hyperplane, which may not be unique.

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

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

By convention, one often extend a convex function with domain $C \subset \mathbb{R}^d$ to a proper convex function on $\mathbb{R}^d$ by letting $f(x) = +\infty$ for $x \notin C$. The domain of such an extended convex function $f(x)$ is the set $\{x : f(x) < +\infty\}$.

The epigraph of a function $f(x) : C \to \mathbb{R}$ is defined as the set

$$\text{epi } f = \{(x, u) \in C \times \mathbb{R} : f(x) \leq u\}.$$

A function $f(x)$ is convex if and only if its epigraph is a convex set. See Figure 3. A convex function is closed if its epigraph is closed. In our lectures, we will only consider closed convex functions.

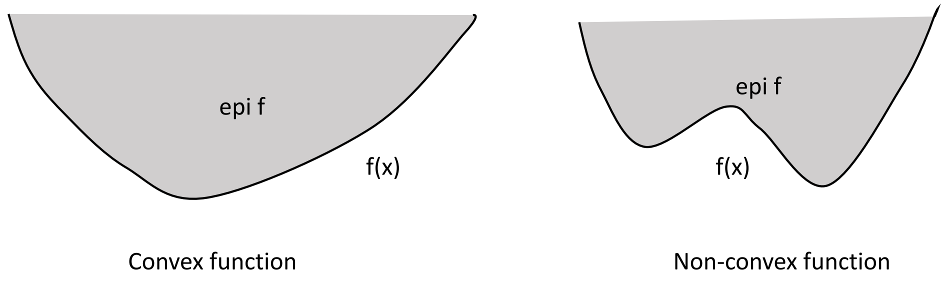


Figure 3: Convex Function

We will present most important properties of convex functions for optimization.

Theorem 1 *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)*.*

Proof Consider any $\epsilon > 0$. Given any $x \in C$, and consider $x' = f(\alpha x + (1-\alpha)\tilde{x}) \in C$. There is a sufficiently small $\alpha > 0$ such that $\|x' - \tilde{x}\| \leq \epsilon$. Local optimality of $\tilde{x}$ implies that

$$f(\tilde{x}) \leq f(x') \leq \alpha f(x) + (1-\alpha) f(\tilde{x}).$$

This implies that $f(\tilde{x}) \leq f(x)$. ■

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

3 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 $u^\top v \leq \|u\|_* \cdot \|v\|$.

The following norms are commonly used in machine learning:

- L_p-norm on $\mathbb{R}^d$, with $1 \leq p \leq \infty$: $\|x\|_p = (\sum_{j=1}^d |x_j|^p)^{1/p}$. Its dual-norm is the L_q-norm, where $1/p + 1/q = 1$. The L_2 norm is also referred to as the Euclidean norm. The L_∞ norm is $\|x\|_\infty = \max_j |x_j|$.
- Given a symmetric positive definite matrix $H \in \mathbb{R}^d \times \mathbb{R}^d$, the H-norm on $\mathbb{R}^d$ is defined as $\|x\|_H = \|H^{1/2} x\|_2$, and its dual norm is $\|\cdot\|_{H^{-1}}$.
- The inner product of two matrices is defined as trace$(A^\top B)$. The trace-norm of a matrix A, $\|A\|_*$, is defined as the sum of its singular values. Its dual norm is the spectral norm of A, defined as its largest singular value. The F-norm (Frobenius norm) of A, defined as the 2-norm of A as a vector, and its dual norm is also the F-norm.

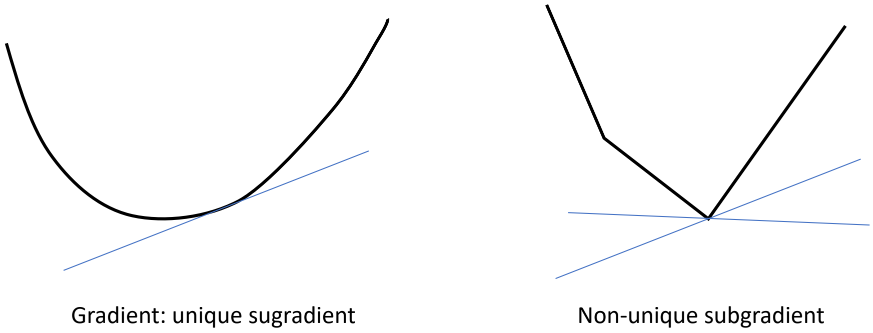
4 Subgradient

If $f(x)$ is differentiable in C, then the gradient $\nabla f(x)$ exists. In such case, $f(x)$ is convex if and only if $\forall x, y$:

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

More generally, 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 $\forall y$:

$$f(y) \geq f(x) + g^\top (y - x). \tag{5}$$



Gradient: unique sugradient

Non-unique subgradient

Figure 4: Subgradient

A subgradient of a convex function defined on $\mathbb{R}^d$ always exists [1], but may not be unique. See Figure 4. A convex function $f(x)$ is differentiable at x if it has a unique subgradient at x.

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

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

For example, consider the convex function $f(x) = |x|$ for $x \in \mathbb{R}$. Its subdifferential at 0 is the set $\partial f(0) = [-1, 1]$.

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

The following result characterizes the solution of convex optimization problem.

Theorem 2 *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 $0 \in \partial f(x_*)$.*

Proof We will only prove a weaker version where $f(x)$ is differentiable on C. For the general case of subgradients, please refer to [1].

If $g_*^\top (y - x_*) \geq 0$, then by convexity condition (5), we have $f(y) - f(x_*) \geq g_*^\top (y - x_*) \geq 0$, which implies that x_* is a solution of (2).

Given any y,

$$\nabla f(x_*)^\top (y - x_*) = \lim_{\alpha \to 0^+} \frac{f(x_* + \alpha(y - x_*)) - f(x_*)}{\alpha}.$$

If x_* is a solution of (2), then the right hand side is non-negative for all $\alpha \in [0, 1]$. Therefore the limit is non-negative, which implies that $\nabla f(x_*)^\top (y - x_*) \geq 0$. ■

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

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

For convex functions, the Lipschitz condition is equivalent to the condition of bounded subgradient: $\|\nabla f(x)\|_2 \leq G$ for all x and subgradient $\nabla f(x)$.

Consider a differentiable function $f(x)$, we say that $f(x)$ is L-smooth for some $L > 0$ if it is gradient Lipschitz: for all x and y

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

The smoothness condition is equivalent to the following inequality:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2. \tag{6}$$

Let $g = \nabla f(x)$ be a subgradient of $f(x)$ at x. Then a convex function satisfies (5). If furthermore, there exists a parameter $\lambda > 0$ such that

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\lambda}{2}\|y - x\|_2^2, \tag{7}$$

then we say $f(x)$ is λ-strongly convex. If $f(x)$ is second order differentiable, then an equivalent condition of λ-strong convexity is $\nabla^2 f(x) \geq \lambda I$.

If $f(x)$ is strongly convex, then the solution of (2) is unique. Moreover, the solution is stable in the following sense

Proposition 1 *If $f(x)$ is λ strongly convex. Then (2) has a unique solution x_*. For an ϵ-approximate solution $\tilde{x}$: $f(\tilde{x}) \leq f(x_*) + \epsilon$, then $\|\tilde{x} - x_*\|_2^2 \leq 2\epsilon/\lambda$.*

Proof Let x_* be a solution of (2), then

$$f(\tilde{x}) \geq f(x_*) + \nabla f(x_*)^\top (\tilde{x} - x_*) + \frac{\lambda}{2}\|\tilde{x} - x_*\|_2^2.$$

We note that Theorem 2 implies that $\nabla f(x_*)^\top (\tilde{x} - x_*) \geq 0$. It follows that

$$f(\tilde{x}) \geq f(x_*) + \frac{\lambda}{2}\|\tilde{x} - x_*\|_2^2 \geq f(\tilde{x}) - \epsilon + \frac{\lambda}{2}\|\tilde{x} - x_*\|_2^2.$$

This implies the desired bound. ■

Due to the stability result, strongly convex function is also easier to solve, with faster convergence rate.

5 Examples of Convex Functions

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

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) + R(w), \tag{8}$$

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

$$f_i(w) = \phi(w^\top X_i, Y_i),$$

and $\phi(u, y)$ is a loss function.

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

- Least squares loss $\phi(u, y) = (u - y)^2$ is smooth, and strongly convex. It is not Lipschitz.
- Logistic loss $\phi(u, y) = \ln(1 + \exp(uy)$ (where $y \in \{\pm 1\}$) is Lipschitz, smooth, but not strongly convex.
- Hinge loss $\phi(u, y) = \max(0, 1 - uy)$ (where $y \in \{\pm 1\}$ is Lipschitz, non-smooth, and not strongly convex.
- Multi-class logistic regression with $y \in \{1, \ldots, k\}$ and $u \in \mathbb{R}^k$, we have $\phi(u, y) = -u_y + \ln \sum_j \exp(u_j)$. It is Lipschitz, smooth, but not strongly convex.

If $\phi(u)$ is a convex function of $u \in \mathbb{R}^k$, and W is a $k \times d$ matrix, and $b \in \mathbb{R}^k$, then $f(u) = \phi(Wx + b)$ is a convex function. The subdifferential of $\phi(Wx + b)$ with respect to x can be obtained using the chain rule as $\{W^\top g : g \in \partial\phi(u)|_{u=Wx+b}\}$.

The commonly used convex regularizers are

- L_2: $R(w) = \frac{\lambda}{2}\|w\|_2^2$, which is smooth and strongly convex.
- L_1: $R(w) = \lambda\|w\|_1$, which is non-smooth and not strongly convex.
- $L_1 - L_2$: $R(w) = \lambda_1\|w\|_1 + \frac{\lambda_2}{2}\|w\|_2^2$, which is non-smooth and strongly convex.
- Trace-norm for matrix w: $R(w) = \lambda\|w\|_*$ (where $\|\cdot\|_*$ is the matrix trace-norm), which is non-smooth and not strongly convex.

References

[1] R. Tyrrell Rockafellar. *Convex analysis.* Princeton University Press, Princeton, NJ, 1970.