DSCC 435 Optimization for Machine Learning — Lecture 2

Convexity and Complexity

Lecturer: Jiaming Liang — *September 5, 2023*

1 Convexity

1.1 Convex set

Definition 1. *A set $S \subseteq \mathbb{R}^n$ is called convex if for any $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0,1]$ it holds that $\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in S$. For a set $S \subseteq \mathbb{R}^n$, the convex hull of S, denoted by $\mathrm{Conv}(S)$, is the intersection of all convex sets containing S.*

Clearly, $\mathrm{Conv}(S)$ is by itself a convex set, and it is the smallest convex set containing S (w.r.t. inclusion).

Affine sets can be defined similarly if we replace $\lambda \in [0,1]$ by $\lambda \in \mathbb{R}$. A hyperplane is a subset of $\mathbb{R}^n$ given by

$$H_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{a}, \mathbf{x} \rangle = b\},$$

where $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. It is an easy exercise to show that hyperplanes are affine sets.

Examples: Evidently, affine sets are always convex. Open and closed balls are always convex regardless of the choice of norm. For given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the closed line segment between $\mathbf{x}$ and $\mathbf{y}$ is a subset of $\mathbb{R}^n$ denoted by $[\mathbf{x}, \mathbf{y}]$ and defined as

$$[\mathbf{x}, \mathbf{y}] = \{\alpha \mathbf{x} + (1-\alpha)\mathbf{y} : \alpha \in [0,1]\}$$

The open line segment $(\mathbf{x}, \mathbf{y})$ is similarly defined as

$$(\mathbf{x}, \mathbf{y}) = \{\alpha \mathbf{x} + (1-\alpha)\mathbf{y} : \alpha \in (0,1)\}$$

when $\mathbf{x} \neq \mathbf{y}$ and is the empty set $\emptyset$ when $\mathbf{x} = \mathbf{y}$. Closed and open line segments are convex sets. Another example of convex sets are half-spaces, which are sets of the form

$$H^-_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{a}, \mathbf{x} \rangle \leq b\},$$

where $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

1.2 Convex function

Definition 2. *A proper extended real-valued function f is convex if and only if*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \quad \textit{for all} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \lambda \in [0,1].$$

The epigraph of an extended real-valued function $f : \mathbb{R}^n \to [-\infty, \infty]$ is defined by $\mathrm{epi}(f) = \{(\mathbf{x}, t) : f(\mathbf{x}) \le t, \mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}\}$

Theorem 1. *An extended real-valued function $f : \mathbb{R}^n \to [-\infty, \infty]$ is called convex if* $\mathrm{epi}(f)$ *is a convex set.*

Definition 3. *A continuously differentiable function f is convex on $\mathbb{R}^n$ if for any $x, y \in \mathbb{R}^n$ we have*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle.$$

Definition 4. *A twice continuously differentiable function f is convex on $\mathbb{R}^n$ if and only if for any $x \in \mathbb{R}^n$ we have*

$$\nabla^2 f(x) \succeq 0.$$

Example

- Every linear function $f(x) = \alpha + \langle a, x \rangle$ is convex.
- Let matrix A be symmetric and positive semidefinite. Then the quadratic function

$$f(x) = \alpha + \langle a, x \rangle + \frac{1}{2}\langle Ax, x \rangle$$

  is convex (since $\nabla^2 f(x) = A \succeq 0$).

Operations preserving convexity

- $A : \mathbb{R}^n \to \mathbb{R}^m$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^n \to (-\infty, \infty]$ is a convex function, then

$$g(x) = f(Ax + b)$$

  is convex.
- Let $f_1, f_2, \ldots, f_m : \mathbb{R}^n \to (-\infty, \infty]$ be extended real-valued convex functions, and let $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}_+$. Then the function $\sum_{i=1}^m \alpha_i f_i$ is convex.
- Let $f_i : \mathbb{R}^n \to (-\infty, \infty], i \in I$, be extended real-valued convex functions, where I is a given index set. Then the function

$$f(\mathbf{x}) = \max_{i \in I} f_i(\mathbf{x})$$

  is convex.

Example Let $C \subseteq \mathbb{R}^n$ be a nonempty set, and consider the function

$$\varphi_C(\mathbf{x}) = \frac{1}{2}\left(\|\mathbf{x}\|^2 - d_C^2(\mathbf{x})\right)$$

where

$$d_C(\mathbf{x}) = \min_{\mathbf{y}\in C} \|\mathbf{x}-\mathbf{y}\|.$$

We want to show that $\varphi_C(\mathbf{x})$ is convex. Note that

$$d_C^2(\mathbf{x}) = \min_{\mathbf{y}\in C} \|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 - \max_{\mathbf{y}\in C}\left[2\langle \mathbf{y},\mathbf{x}\rangle - \|\mathbf{y}\|^2\right]$$

Hence,

$$\varphi_C(\mathbf{x}) = \max_{\mathbf{y}\in C}\left[\langle \mathbf{y},\mathbf{x}\rangle - \frac{1}{2}\|\mathbf{y}\|^2\right].$$

Therefore, since φ_C is a maximization of affine—and hence convex—functions, by the maximization rule above, it is necessarily convex. Also note that φ_C is convex regarless of whether C is convex or not.

2 Complexity

Definition 5. *We say $f(x) = \mathcal{O}(g(x))$ if there exists scalars $M > 0$ and $x_0 \in \mathbb{R}$ such that*

$$|f(x)| \leq Mg(x) \quad \text{for all} \quad x \geq x_0.$$

We say $f(x) = \Omega(g(x))$ if there exists scalars $M > 0$ and $x_0 \in \mathbb{R}$ such that

$$f(x) \geq Mg(x) \quad \text{for all} \quad x \geq x_0.$$

We say $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$.

Convergence rate

Sublinear rate. This rate is described in terms of a power function of the iteration counter. For example, suppose that for some method we can prove the rate of convergence $r_k \leq \frac{c}{\sqrt{k}}$. In this case, the upper complexity bound justified by this scheme for the corresponding problem class is $\left(\frac{c}{\epsilon}\right)^2$.

Example Gradient method:

$$f(x_k) - f_* \leq \frac{Ld_0^2}{k}$$

Accelerated gradient method:

$$f(x_k) - f_* \leq \frac{Ld_0^2}{k^2}$$

Subgradient method:

$$f(x_k) - f_* \leq \frac{Md_0}{\sqrt{k}}$$

Linear rate. This rate is given in terms of an exponential function of the iteratic counter. For example,

$$r_{k+1} \leq (1-q)r_k$$

$$r_k \leq c(1-q)^k \leq ce^{-qk}, \quad 0 < q \leq 1.$$

Note that the corresponding complexity bound is $\frac{1}{q}\left(\ln c + \ln \frac{1}{\epsilon}\right)$.

Eample Gradient method:

$$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{\mu d_0^2}{\varepsilon}\right)\right)$$

Accelerated gradient method:

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\mu d_0^2}{\varepsilon}\right)\right)$$

Quadratic rate. This rate has a double exponential dependence in the iteration counter. For example,

$$r_{k+1} \leq cr_k^2$$

The corresponding complexity estimate depends on the double logarithm of the desired accuracy: $\ln\ln\frac{1}{\epsilon}$.

Example Suppose that the initial starting point x_0 is close enough to x^* :

$$\|x_0 - x^*\| \leq \bar{r} = \frac{2\mu}{3M}.$$

Then $\|x_k - x^*\| \leq \bar{r}$ for all k and the Newton's Method converges quadratically:

$$\|x_{k+1} - x^*\| \leq \frac{M\|x_k - x^*\|^2}{2\left(\mu - M\|x_k - x^*\|\right)}.$$

Note that following the definition of $\bar{r}$ and $\|x_k - x^*\| \leq \bar{r}$, we have

$$\|x_{k+1} - x^*\| \leq \frac{M\|x_k - x^*\|^2}{2\left(\mu - M\bar{r}\right)} = \frac{3M\|x_k - x^*\|^2}{2\mu} = \frac{\|x_k - x^*\|^2}{\bar{r}}.$$