

Basics of Convex Optimization

Yanze Song

January 19, 2024

1 Introduction to Convexity

1.1 Convex Sets

Definition 1.1.1 (*Lines and Line Segments*). Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be two distinct points, then

- $\{\theta\mathbf{x} + (1 - \theta)\mathbf{y} : \theta \in \mathbb{R}\}$ represents the line passing through \mathbf{x} and \mathbf{y} ,
- $\{\theta\mathbf{x} + (1 - \theta)\mathbf{y} : \theta \in [0, 1]\}$ represents the line segment passing through \mathbf{x} and \mathbf{y} .

Definition 1.1.2 (*Affine Sets*). A set $\mathcal{A} \subseteq \mathbb{R}^n$ is affine if for all $\mathbf{x}, \mathbf{y} \in \mathcal{A}$, $\theta \in \mathbb{R}$,

$$\theta\mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{A}.$$

Definition 1.1.3 (*Convex Sets*). A set $\mathcal{C} \subseteq \mathbb{R}^n$ is convex if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$\theta\mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{C}.$$

1.2 Convex Set Examples

Definition 1.2.1 (*Hyperplanes, Halfspaces & Polyhedra*). Let $\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, $b \in \mathbb{R}$, $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_m(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{d} \in \mathbb{R}^m$, then

- A hyperplane has the form $\{\mathbf{x} : \mathbf{a}^T \mathbf{x} = b\}$,
- A halfspace has the form $\{\mathbf{x} : \mathbf{a}^T \mathbf{x} \leq b\}$,
- A polyhedron has the form $\{\mathbf{x} : A\mathbf{x} = \mathbf{b}, B\mathbf{x} \leq \mathbf{d}\}$.

Proposition 1.2.1. Hyperplanes, halfspaces and polyhedra are all convex. In particular, hyperplanes are affine.

Proof. We'll first prove that polyhedra are convex. Let $\mathcal{P} = \{\mathbf{x} : A\mathbf{x} = \mathbf{b}, B\mathbf{x} \leq \mathbf{d}\}$ be a polyhedron, then for all $\mathbf{x}, \mathbf{y} \in \mathcal{P}$, $\theta \in [0, 1]$, consider $\theta\mathbf{x} + (1 - \theta)\mathbf{y}$:

$$A(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) = \theta A\mathbf{x} + (1 - \theta)A\mathbf{y} = \theta\mathbf{b} + (1 - \theta)\mathbf{b} = \mathbf{b},$$

$$B(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) = \theta B\mathbf{x} + (1 - \theta)B\mathbf{y} \leq \theta\mathbf{d} + (1 - \theta)\mathbf{d} = \mathbf{d}.$$

We proved that $\theta\mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{P}$, and therefore \mathcal{P} is convex. It's trivial to see that hyperplanes are affine. \square

1.3 Operations that Preserve the Convexity of Sets

Proposition 1.3.1. Let $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathbb{R}^n$ be convex, then $\mathcal{C}_1 \cap \mathcal{C}_2$ is also convex.

Proof. For all $\mathbf{x}, \mathbf{y} \in \mathcal{C}_1 \cap \mathcal{C}_2$, $\theta \in [0, 1]$, consider $\theta\mathbf{x} + (1 - \theta)\mathbf{y}$:

$$\mathbf{x}, \mathbf{y} \in \mathcal{C}_1 \Rightarrow \theta\mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{C}_1, \mathbf{x}, \mathbf{y} \in \mathcal{C}_2 \Rightarrow \theta\mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{C}_2,$$

that is, we have $\theta\mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{C}_1 \cap \mathcal{C}_2$, and this completes the proof. \square

Definition 1.3.1 (*Affine Functions*). Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a function, then \mathbf{f} is affine if there exists $A \in \mathcal{M}_{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$ s.t. for all $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

In particular, scalar and linear (equivalently, matrix) transformations are affine transformations.

Proposition 1.3.2. Let $\mathcal{C} \subseteq \mathbb{R}^n$ be convex, and $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be an affine function, then $\mathbf{f}(\mathcal{C})$, i.e. the image of \mathbf{f} over \mathcal{C} is also convex.

Proof. For all $\mathbf{y}_1, \mathbf{y}_2 \in \mathbf{f}(\mathcal{C})$, there exist $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ s.t.

$$\mathbf{f}(\mathbf{x}_1) = A\mathbf{x}_1 + \mathbf{b} = \mathbf{y}_1, \mathbf{f}(\mathbf{x}_2) = A\mathbf{x}_2 + \mathbf{b} = \mathbf{y}_2.$$

We'll then show that $\theta\mathbf{y}_1 + (1 - \theta)\mathbf{y}_2 \in \mathbf{f}(\mathcal{C})$ for all $\theta \in [0, 1]$ as well:

$$\theta\mathbf{y}_1 + (1 - \theta)\mathbf{y}_2 = \theta(A\mathbf{x}_1 + \mathbf{b}) + (1 - \theta)(A\mathbf{x}_2 + \mathbf{b}) = A(\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2) + \mathbf{b} = \mathbf{f}(\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2),$$

where $\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \in \mathcal{C}$, and this completes the proof. \square

1.4 Assumptions

Remark 1.4.1. Unless otherwise specified, we assume \mathcal{C} is a **open** convex set throughout this note to resolve the differentiability issue. Or we can equivalently assume \mathbf{f} is differentiable over an open set that contains \mathcal{C} .

1.5 Convex Functions

Definition 1.5.1 (*Convex Functions*). A function $f : \mathcal{C} \rightarrow \mathbb{R}$ is convex if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}).$$

Definition 1.5.2 (*Strictly Convex Functions*). A function $f : \mathcal{C} \rightarrow \mathbb{R}$ is strictly convex if for all $\mathbf{x} \neq \mathbf{y} \in \mathcal{C}$, $\theta \in (0, 1)$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}).$$

Definition 1.5.3 (*Concave Functions*). A function $f : \mathcal{C} \rightarrow \mathbb{R}$ is concave if $-f$ is convex. It is equivalent to define that f is concave if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) \geq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}).$$

Definition 1.5.4 (*Affine Functions*). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is affine if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) = \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}).$$

This is equivalent to Definition 1.3.1.

Proof. We're to prove for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) = \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}) \Leftrightarrow f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} + b, \forall \mathbf{x} \in \mathbb{R}^n$$

for some $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$. We'll first prove \Leftarrow . For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\theta \in [0, 1]$, we have

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) = \mathbf{a}^T(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) + b = \theta(\mathbf{a}^T \mathbf{x} + b) + (1 - \theta)(\mathbf{a}^T \mathbf{y} + b) = \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}).$$

Conversely, we define $g : \mathbb{R}^n \rightarrow \mathbb{R}$ s.t. $g(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{0})$, then it's equivalent to prove that g is linear. Note that g satisfies the LHS as well, and $g(\mathbf{0}) = 0$.

- g preserves scalar multiplication, i.e., $\forall \alpha \in \mathbb{R} = [0, 1] \cup (1, \infty) \cup (-\infty, 0) : g(\alpha\mathbf{x}) = \alpha g(\mathbf{x})$:

$$- \alpha \in [0, 1] \Rightarrow \alpha, 1 - \alpha \in [0, 1]:$$

$$g(\alpha\mathbf{x}) = g(\alpha\mathbf{x} + (1 - \alpha)\mathbf{0}) = \alpha g(\mathbf{x}) + (1 - \alpha)g(\mathbf{0}) = \alpha g(\mathbf{x}) \quad (1)$$

– $\alpha \in (1, \infty) \Rightarrow \frac{1}{\alpha} \in [0, 1]$:

$$g(\alpha \mathbf{x}) = \alpha \frac{1}{\alpha} g(\alpha \mathbf{x}) \stackrel{(1)}{=} \alpha g\left(\frac{1}{\alpha} \alpha \mathbf{x}\right) = \alpha g(\mathbf{x}) \quad (2)$$

– $\alpha \in (-\infty, 0) \Rightarrow -\alpha \in (0, \infty)$:

$$g(\alpha \mathbf{x}) = g(-\alpha(-\mathbf{x})) = -\alpha g(\mathbf{0} - \mathbf{x}) \stackrel{(4)}{=} -\alpha(g(\mathbf{0}) - g(\mathbf{x})) = \alpha g(\mathbf{x}) \quad (3)$$

• g preserves addition, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : g(\mathbf{x} + \mathbf{y}) = g(\mathbf{x}) + g(\mathbf{y})$: Let $\theta \in (0, 1)$, then

$$g(\mathbf{x} + \mathbf{y}) = g\left(\theta \frac{1}{\theta} \mathbf{x} + (1 - \theta) \frac{1}{1 - \theta} \mathbf{y}\right) = \theta g\left(\frac{1}{\theta} \mathbf{x}\right) + (1 - \theta) g\left(\frac{1}{1 - \theta} \mathbf{y}\right) \stackrel{(1)}{=} g(\mathbf{x}) + g(\mathbf{y}). \quad (4)$$

□

1.6 Theorems of Convex Functions

Proposition 1.6.1. *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a function, then*

$$f \text{ is affine} \Leftrightarrow f \text{ is both convex and concave.}$$

Proof. This holds if $\text{dom}(f) = \mathbb{R}^n$, so it directly applies to \mathcal{C} . □

Proposition 1.6.2 (*Convex Along All Lines*). *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a function, then f is convex if and only if for all $\mathbf{x} \in \mathcal{C}$, $\mathbf{d} \in \mathcal{D} = \{\mathbf{d} \in \mathbb{R}^n : \exists t > 0 : \mathbf{x} + t\mathbf{d} \in \mathcal{C}\}$, i.e. the set of all feasible directions at \mathbf{x} ,*

$$g(t) = f(\mathbf{x} + t\mathbf{d}) \text{ is convex,}$$

where $\text{dom}(g) = \{t : \mathbf{x} + t\mathbf{d} \in \mathcal{C}\}$.

Proof. It's trivial to show that $\text{dom}(g)$ is convex. Assume f is convex, then for all $t_1, t_2 \in \text{dom}(g)$, $\theta \in [0, 1]$,

$$\begin{aligned} g(\theta t_1 + (1 - \theta)t_2) &= f(\mathbf{x} + (\theta t_1 + (1 - \theta)t_2)\mathbf{d}) = f(\theta(\mathbf{x} + t_1\mathbf{d}) + (1 - \theta)(\mathbf{x} + t_2\mathbf{d})) \\ &= \theta f(\mathbf{x} + t_1\mathbf{d}) + (1 - \theta)f(\mathbf{x} + t_2\mathbf{d}) = \theta g(t_1) + (1 - \theta)g(t_2). \end{aligned}$$

Conversely, assume g is convex, then for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$\begin{aligned} f(\theta \mathbf{y} + (1 - \theta)\mathbf{x}) &= f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) = g(\theta) \\ &= g(\theta \cdot 1 + (1 - \theta) \cdot 0) = \theta g(1) + (1 - \theta)g(0) = \theta f(\mathbf{y}) + (1 - \theta)f(\mathbf{x}). \end{aligned}$$

□

Theorem 1.6.1 (*First Order Convexity Condition*). *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a differentiable function, then f is convex if and only if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

Proof. First we assume f is convex, then for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in (0, 1]$,

$$f(\theta \mathbf{y} + (1 - \theta)\mathbf{x}) \leq \theta f(\mathbf{y}) + (1 - \theta)f(\mathbf{x}) \Rightarrow \frac{f(\theta \mathbf{y} + (1 - \theta)\mathbf{x}) - f(\mathbf{x})}{\theta} + f(\mathbf{x}) \leq f(\mathbf{y}).$$

Let $g(\theta) = f(\theta \mathbf{y} + (1 - \theta)\mathbf{x})$, so in particular $f(\mathbf{x}) = g(0)$, then it becomes for all $\theta \in (0, 1]$,

$$\frac{g(0 + \theta) - g(0)}{\theta} + f(\mathbf{x}) \leq f(\mathbf{y}).$$

With $\theta \rightarrow 0^+$ we have $g'(0) + f(\mathbf{x}) \leq f(\mathbf{y})$, where $g'(0) = \langle \nabla f(\theta \mathbf{y} + (1 - \theta)\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle|_{\theta=0} = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$. Conversely, for all $\mathbf{x} \neq \mathbf{y} \in \mathcal{C}$ (the case $\mathbf{x} = \mathbf{y}$ is trivial), $\theta \in [0, 1]$, let $\mathbf{z} = \theta \mathbf{x} + (1 - \theta)\mathbf{y} \in \mathcal{C}$, then

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle, \quad (5)$$

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle, \quad (6)$$

and $\theta \cdot (5) + (1 - \theta) \cdot (6)$ completes the proof. \square

Remark 1.6.1. Here're some details for $g'(\theta)$:

$$g(\theta) = f(\underbrace{\theta \mathbf{y} + (1 - \theta)\mathbf{x}}_{\mathbf{u}(\theta)}) = f(\underbrace{\theta y_1 + (1 - \theta)x_1}_{u_1(\theta)}, \dots, \underbrace{\theta y_n + (1 - \theta)x_n}_{u_n(\theta)}),$$

so by the Chain Rule,

$$\begin{aligned} \frac{dg}{d\theta} &= \frac{d}{d\theta} f(u_1(\theta), \dots, u_n(\theta)) = \sum_{i=1}^n \frac{\partial f}{\partial u_i} \frac{du_i}{d\theta} = \sum_{i=1}^n \frac{\partial f}{\partial u_i} (y_i - x_i) \\ &= \langle \nabla f(\mathbf{u}), \mathbf{y} - \mathbf{x} \rangle = \langle \nabla f(\theta \mathbf{y} + (1 - \theta)\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle. \end{aligned}$$

Theorem 1.6.2 (*Second Order Convexity Condition*). Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a twice differentiable function, then f is convex if and only if for all $\mathbf{x} \in \mathcal{C}$,

$$\nabla^2 f(\mathbf{x}) \succcurlyeq \mathbf{0}.$$

Proof. For all $\mathbf{x} \in \mathcal{C}$, we have for all $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{x} + t\mathbf{d} \in \mathcal{C}$ for some small $t > 0$. Then by the Taylor's Theorem,

$$f(\mathbf{x} + t\mathbf{d}) = f(\mathbf{x}) + t\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle + \frac{1}{2}t^2 \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} + o(t^2)$$

\square

1.7 Optimality of Convex Functions

Theorem 1.7.1 (*Local & Global Optimality*). Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a convex function, then any locally optimal point is also globally optimal.

Proof. Let \mathbf{x}^* be a local optimum, then there exists $R > 0$ s.t. for all $\mathbf{x} \in \{\mathbf{x} \in \mathcal{C} : \|\mathbf{x} - \mathbf{x}^*\| \leq R\}$,

$$f(\mathbf{x}^*) \leq f(\mathbf{x}).$$

To prove by contradiction, assume there exists $\mathbf{x}_0 \in \mathcal{C} \setminus \{\mathbf{x}^*\}$ s.t.

$$f(\mathbf{x}_0) < f(\mathbf{x}^*),$$

and it's clear that $\|\mathbf{x}_0 - \mathbf{x}^*\| > R$. Now consider the point $\mathbf{x}_R = \theta \mathbf{x}^* + (1 - \theta)\mathbf{x}_0 \in \mathcal{C}$ where $\theta = 1 - \frac{R}{\|\mathbf{x}_0 - \mathbf{x}^*\|} \in (0, 1)$, note that $\|\mathbf{x}_R - \mathbf{x}^*\| = R$, then

$$f(\theta \mathbf{x}^* + (1 - \theta)\mathbf{x}_0) = f(\mathbf{x}_R) \geq f(\mathbf{x}^*) > \theta f(\mathbf{x}^*) + (1 - \theta)f(\mathbf{x}_0),$$

which contradicts with the convexity of f , and this completes the proof. \square

Theorem 1.7.2 (*First Order Optimality Condition*). Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be convex and differentiable, then \mathbf{x}^* minimizes f over \mathcal{C} if and only if for all $\mathbf{x} \in \mathcal{C}$,

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0. \quad (7)$$

In particular if $\mathcal{C} = \mathbb{R}^n$, then \mathbf{x}^* minimizes f over \mathbb{R}^n if and only if

$$\nabla f(\mathbf{x}^*) = \mathbf{0}.$$

Proof. To prove by contradiction, assume \mathbf{x}^* minimizes f over \mathcal{C} , and there exists $\mathbf{x}_0 \in \mathcal{C}$ s.t. $\langle \nabla f(\mathbf{x}^*), \mathbf{x}_0 - \mathbf{x}^* \rangle < 0$. Similar to Theorem 1.6.1, define $g(\theta) = f(\theta\mathbf{x}_0 + (1 - \theta)\mathbf{x}^*)$, then

$$\lim_{\theta \rightarrow 0^+} \frac{g(0 + \theta) - g(0)}{\theta} = \langle \nabla f(\mathbf{x}^*), \mathbf{x}_0 - \mathbf{x}^* \rangle < 0,$$

which implies that $g(0 + \theta) < g(0)$ for some small $\theta > 0$, contradicting with the minimality of \mathbf{x}^* . Conversely, assume $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \Rightarrow f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq f(\mathbf{x}^*)$. By Theorem 1.6.1, for all $\mathbf{x} \in \mathcal{C}$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq f(\mathbf{x}^*),$$

showing that \mathbf{x}^* minimizes f over \mathcal{C} . If $\mathcal{C} = \mathbb{R}^n$, assume \mathbf{x}^* minimizes f over \mathbb{R}^n , and $\frac{\partial}{\partial x_i} f(\mathbf{x}^*) \neq 0$ for some i . We then have $f(\mathbf{x}^* + \theta \mathbf{e}_i) < f(\mathbf{x}^*)$ for some small $\theta \neq 0$, which contradict with the minimality of \mathbf{x}^* . The converse can be proved directly using the above part. \square

1.8 Other types of Convex Functions

1.8.1 Strong Convexity

Definition 1.8.1 (*Strongly Convex Functions*). A function $f : \mathcal{C} \rightarrow \mathbb{R}$ is σ -strongly convex w.r.t. some norm $\|\cdot\|$ for some $\sigma > 0$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}) - \frac{\sigma}{2}\theta(1 - \theta)\|\mathbf{x} - \mathbf{y}\|^2.$$

Proposition 1.8.1 (*Quadratic Lower Bound of Strongly Convex Functions*). Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be differentiable, then f is σ -strongly convex if and only if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Proof. It's very similar to Proof 1.6. Let f be σ -strongly convex, then for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in (0, 1]$,

$$\begin{aligned} f(\theta\mathbf{y} + (1 - \theta)\mathbf{x}) &\leq \theta f(\mathbf{y}) + (1 - \theta)f(\mathbf{x}) - \frac{\sigma}{2}\theta(1 - \theta)\|\mathbf{y} - \mathbf{x}\|^2 \\ \Rightarrow f(\mathbf{x}) + \frac{f(\theta\mathbf{y} + (1 - \theta)\mathbf{x}) - f(\mathbf{x})}{\theta} &+ \frac{\sigma}{2}(1 - \theta)\|\mathbf{y} - \mathbf{x}\|^2 \leq f(\mathbf{y}), \end{aligned}$$

then with $\theta \rightarrow 0^+$ we completes the proof for \Rightarrow . Conversely, for all $\mathbf{x} \neq \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$, let $\mathbf{z} = \theta\mathbf{x} + (1 - \theta)\mathbf{y}$, then

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{z}\|^2, \quad (8)$$

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{z}\|^2 \quad (9)$$

and $\theta \cdot (8) + (1 - \theta) \cdot (9)$ we get $f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}) - \sigma\theta(1 - \theta)\|\mathbf{x} - \mathbf{y}\|^2$, which implies the one in Definition 1.8.1. \square

Proposition 1.8.2 (*Sum of Strongly Convex Functions*). Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be σ -strongly convex, $g : \mathcal{C} \rightarrow \mathbb{R}$ be convex, both of which are differentiable, then

$$h = f + g \text{ is also } \sigma\text{-strongly convex.}$$

Proof. For $\mathbf{x}, \mathbf{y} \in \mathcal{C}$,

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2, \\ g(\mathbf{y}) &\geq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \end{aligned}$$

and after adding we have

$$\underbrace{f(\mathbf{y}) + g(\mathbf{y})}_{h(\mathbf{y})} \geq \underbrace{f(\mathbf{x}) + g(\mathbf{x})}_{h(\mathbf{x})} + \underbrace{\langle \nabla f(\mathbf{x}) + \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle}_{\nabla h(\mathbf{x})} + \frac{\sigma}{2} \|\mathbf{y} - \mathbf{x}\|^2,$$

so $h = f + g$ is σ -strongly convex. \square

Proposition 1.8.3. *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be differentiable and σ -strongly convex w.r.t. some norm $\|\cdot\|$, and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$, then for all $\mathbf{x} \in \mathcal{C}$, we have*

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}^*\|^2.$$

Proof. By strong convexity,

$$f(\mathbf{w}) - f(\mathbf{w}^*) \geq \langle \nabla f(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle + \frac{\sigma}{2} \|\mathbf{w} - \mathbf{w}^*\|^2,$$

where $\langle \nabla f(\mathbf{w}^*), f(\mathbf{w}) - \mathbf{w}^* \rangle \geq 0$ by convexity, and this completes the proof. This gives a bit more information than $f(\mathbf{w}) - f(\mathbf{w}^*) \geq 0$. \square

1.8.2 Lipschitz Continuity

Definition 1.8.2 (*Lipschitz Continuous Functions*). *A function $f : \mathcal{X} \rightarrow \mathbb{R}$ is L -Lipschitz continuous w.r.t. some norm $\|\cdot\|$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

Proposition 1.8.4 (*Lipschitzness, Convexity and Bounded Gradient*). *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be convex and differentiable, then f is L -Lipschitz w.r.t. some norm $\|\cdot\|$ if and only if for all $\mathbf{x} \in \mathcal{C}$,*

$$\|\nabla f(\mathbf{x})\| \leq L.$$

Proof. For all $\mathbf{x} \in \operatorname{int}(\mathcal{C})$, there exist $\eta > 0$ s.t. $\mathbf{x} + \eta \nabla f(\mathbf{x}) \in \mathcal{C}$. By Theorem 1.6.1,

$$f(\mathbf{x} + \eta \nabla f(\mathbf{x})) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x} + \eta \nabla f(\mathbf{x}) - \mathbf{x} \rangle \Rightarrow |f(\mathbf{x} + \eta \nabla f(\mathbf{x})) - f(\mathbf{x})| \geq \|\eta \nabla f(\mathbf{x})\|^2,$$

and then by lipschitzness,

$$L \|\mathbf{x} + \eta \nabla f(\mathbf{x}) - \mathbf{x}\| \geq |f(\mathbf{x} + \eta \nabla f(\mathbf{x})) - f(\mathbf{x})| \geq \|\eta \nabla f(\mathbf{x})\|^2,$$

which yields $\|\nabla f(\mathbf{x})\| \leq L$. \square

1.9 Convex Function Examples

Example 1.9.1 (*Common Convex and Concave Functions*). The convexity of the followings can be proved using the above:

- $e^x, x \log x, -\log x$ are convex
- x^α is convex on $\mathbb{R}_{>0}$ for $\alpha \geq 1$ or $\alpha \leq 0$
- Every norm $\|\mathbf{x}\|$ on \mathbb{R}^n is convex
- Geometric mean $f(\mathbf{x}) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave on $\mathbb{R}_{\geq 0}^n$

1.10 Operations that Preserve the Convexity of Functions

Proposition 1.10.1 (*Non-negative Weighted Sums*). *Let $f_1, f_2 : \mathcal{C} \rightarrow \mathbb{R}$ be convex, $\omega_1, \omega_2 \geq 0$, then*

$$f = \omega_1 f_1 + \omega_2 f_2 \text{ is convex.}$$

Proof. For all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$\begin{aligned} f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) &= (\omega_1 f_1 + \omega_2 f_2)(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) \\ &= \omega_1 f_1(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) + \omega_2 f_2(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) \\ &\leq \theta \omega_1 f_1(\mathbf{x}) + (1 - \theta) \omega_1 f_1(\mathbf{y}) + \theta \omega_2 f_2(\mathbf{x}) + (1 - \theta) \omega_2 f_2(\mathbf{y}) \\ &= \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y}). \end{aligned}$$

□

Proposition 1.10.2 (*Composition with an Affine Map*). *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be convex, $A \in \mathcal{M}_{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, then*

$$g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}) \text{ is convex,}$$

where $\text{dom}(g) = \{\mathbf{x} : A\mathbf{x} + \mathbf{b} \in \mathcal{C}\}$.

Proof. For all $\mathbf{x}, \mathbf{y} \in \text{dom}(g)$, $\theta \in [0, 1]$,

$$\begin{aligned} g(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) &= f(A(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) + \mathbf{b}) \\ &= f(\theta(A\mathbf{x} + \mathbf{b}) + (1 - \theta)(A\mathbf{y} + \mathbf{b})) \\ &\leq \theta f(A\mathbf{x} + \mathbf{b}) + (1 - \theta) f(A\mathbf{y} + \mathbf{b}) \\ &= \theta g(\mathbf{x}) + (1 - \theta) g(\mathbf{y}). \end{aligned}$$

□

Proposition 1.10.3 (*Pointwise Maximum*). *Let $f_1, f_2 : \mathcal{C} \rightarrow \mathbb{R}$ be convex, then*

$$f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\} \text{ is convex.}$$

Proof. For all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$,

$$\begin{aligned} f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) &= \max\{f_1(\theta \mathbf{x} + (1 - \theta) \mathbf{y}), f_2(\theta \mathbf{x} + (1 - \theta) \mathbf{y})\} \\ &\leq \max\{\theta f_1(\mathbf{x}) + (1 - \theta) f_1(\mathbf{y}), \theta f_2(\mathbf{x}) + (1 - \theta) f_2(\mathbf{y})\} \\ &\leq \theta \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\} + (1 - \theta) \max\{f_1(\mathbf{y}), f_2(\mathbf{y})\} \\ &= \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y}) \end{aligned}$$

□

1.11 Relationship between Convex Sets and Convex Functions

Definition 1.11.1 (*Graphs, Epigraphs, Level Sets and Sublevel Sets*). *Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a function, then*

- The graph of f is defined as $\{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in \mathcal{X}\} \subseteq \mathbb{R}^{n+1}$,
- The epigraph of f is defined as $\{(\mathbf{x}, t) : \mathbf{x} \in \mathcal{X}, t \geq f(\mathbf{x})\} \subseteq \mathbb{R}^{n+1}$, denoted by $\text{epi}(f)$,
- The α -level set of f is defined as $\{\mathbf{x} : \mathbf{x} \in \mathcal{X}, f(\mathbf{x}) = \alpha\}$,
- The α -sublevel set of f is defined as $\{\mathbf{x} : \mathbf{x} \in \mathcal{X}, f(\mathbf{x}) \leq \alpha\}$, denoted by C_α .

Proposition 1.11.1. *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a function, then f is convex if and only if*

$$\text{epi}(f) \text{ is convex.}$$

Proof. Assume f is convex, then for all $(\mathbf{x}, t_1), (\mathbf{y}, t_2) \in \text{epi}(f)$, $\theta \in [0, 1]$, consider $\theta(\mathbf{x}, f(\mathbf{x})) + (1 - \theta)(\mathbf{y}, f(\mathbf{y})) = (\theta\mathbf{x} + (1 - \theta)\mathbf{y}, \theta t_1 + (1 - \theta)t_2)$,

$$\theta t_1 + (1 - \theta)t_2 \geq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}) \geq f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}),$$

and hence $\theta(\mathbf{x}, f(\mathbf{x})) + (1 - \theta)(\mathbf{y}, f(\mathbf{y})) \in \text{epi}(f)$. Conversely, assume $\text{epi}(f)$ is convex, then for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\theta \in [0, 1]$, consider $(\mathbf{x}, f(\mathbf{x})), (\mathbf{y}, f(\mathbf{y})) \in \text{epi}(f)$,

$$\theta(\mathbf{x}, f(\mathbf{x})) + (1 - \theta)(\mathbf{y}, f(\mathbf{y})) = (\theta\mathbf{x} + (1 - \theta)\mathbf{y}, \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y})) \in \text{epi}(f),$$

and hence $\theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}) \geq f(\theta\mathbf{x} + (1 - \theta)\mathbf{y})$. □

Proposition 1.11.2. *Let $f : \mathcal{C} \rightarrow \mathbb{R}$ be convex, then for all α ,*

$$C_\alpha \text{ is convex.}$$

Proof. For all $\mathbf{x}, \mathbf{y} \in C_\alpha$, $\theta \in [0, 1]$,

$$f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}) \leq \alpha,$$

and hence $\theta\mathbf{x} + (1 - \theta)\mathbf{y} \in C_\alpha$. □

2 Convex Optimization: Concepts

Definition 2.0.1 (*Mathematical Optimization*). Let $\{f_i\}_{i=0}^m$ and $\{h_j\}_{j=1}^p : \mathbb{R}^n \rightarrow \mathbb{R}$ be functions. We use the following notation to represent the standard/canonical form of a *mathematical optimization* problem:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize:}} && f_0(\mathbf{x}) \\ & \text{subject to:} && f_i(\mathbf{x}) \leq 0, \quad \forall i \in \{1, 2, \dots, m\} \\ & && h_j(\mathbf{x}) = 0, \quad \forall j \in \{1, 2, \dots, p\}. \end{aligned}$$

Here are some related terminologies:

- \mathbf{x} : optimization variable
- f_0 : objective function
- $\{f_i(\mathbf{x}) \leq 0\}_{i=1}^m$: inequality constraints
- $\{h_j(\mathbf{x}) = 0\}_{j=1}^p$: equality constraints
- A point \mathbf{x} is *feasible* if it satisfies all constraints, and *infeasible* otherwise.
- The *feasible set* $C \subseteq \mathbb{R}^n$ is the set of all feasible points.
- The problem is *feasible* if $C \neq \emptyset$, and *infeasible* otherwise.
- The *optimal value* p^* is defined as $\inf_{\mathbf{x}} \{f_0(\mathbf{x}) \mid \mathbf{x} \in C\}$, which may or may not be attainable.
- A feasible point \mathbf{x}^* is *globally optimal*, or *optimal* if $f_0(\mathbf{x}^*) = p^*$. There may be multiple optimal points.
- A feasible point \mathbf{x}^* is *locally optimal* if $\exists R > 0 : f_0(\mathbf{x}^*) = \min_{\mathbf{x}} \{f_0(\mathbf{x}) \mid \mathbf{x} \in C \text{ and } \|\mathbf{x} - \mathbf{x}^*\| \leq R\}$.
- The problem is *unbounded below* if $p^* = -\infty$.

Here are some other equivalent forms to represent an optimization problem:

Definition 2.0.2 (*Indicator Function Form*). With respect to the problem above, the *indicator function form* looks like:

$$\underset{\mathbf{x}}{\text{minimize:}} \quad f_0(\mathbf{x}) + I_C(\mathbf{x})$$

where the *indicator function* I_C is defined as follows:

$$I_C: \mathbb{R}^n \rightarrow \mathbb{R}$$

$$\mathbf{x} \mapsto \begin{cases} f_0(\mathbf{x}), & \text{if } \mathbf{x} \in C \\ \infty, & \text{otherwise.} \end{cases}$$

Remark 2.0.1. The Indicator function form relaxes the problem, while sacrificing its convex property, i.e., it is no longer a convex optimization problem.

Definition 2.0.3 (*Epigraph Form*). With respect to the problem above, the *epigraph form* looks like:

$$\underset{(\mathbf{x}, t)}{\text{minimize:}} \quad t$$

$$\text{subject to: } f_i(\mathbf{x}) \leq 0, \quad \forall i \in \{1, 2, \dots, m\}$$

$$h_j(\mathbf{x}) = 0, \quad \forall j \in \{1, 2, \dots, p\}$$

$$f_0(\mathbf{x}) \leq t.$$

Remark 2.0.2. The optimization variable changes from \mathbf{x} to (\mathbf{x}, t) , so rigorously, all constraint functions should be (slightly) modified correspondingly. But we will skip these for simplicity.

Definition 2.0.4 (*Convex Optimization*). Let $\{f_i\}_{i=0}^m : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex functions, $\{\mathbf{a}_j\}_{j=1}^p \in \mathbb{R}^n$, and $\{b_j\}_{j=1}^p \in \mathbb{R}$, then a *convex optimization* problem has the form:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize:}} && f_0(\mathbf{x}) \\ & \text{subject to:} && f_i(\mathbf{x}) \leq 0, \quad \forall i \in \{1, 2, \dots, m\} \\ & && \mathbf{a}_j^T \mathbf{x} - b_j = 0, \quad \forall j \in \{1, 2, \dots, p\}, \end{aligned}$$

or equivalently, it has the form:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize:}} && f_0(\mathbf{x}) \\ & \text{subject to:} && f_i(\mathbf{x}) \leq 0, \quad \forall i \in \{1, 2, \dots, m\} \\ & && A\mathbf{x} - \mathbf{b} = \mathbf{0}. \end{aligned}$$

Remark 2.0.3. Convex optimization has three more requirements:

- The objective function f_0 must be convex,
- The inequality constraint functions $\{f_i\}_{i=1}^m$ must be convex,
- The equality constraint functions $\{h_j(\mathbf{x}) = \mathbf{a}_j^T \mathbf{x} - b_j\}_{j=1}^p$ must be affine.

The resulting feasible set from the form above is convex because:

- Any sublevel set of a convex function $\{f_i\}_{i=1}^m$ is convex,
- Hyperplanes are affine $\{\mathbf{x} \mid \mathbf{a}_j^T \mathbf{x} - b_j = 0\}_{j=1}^p$, and therefore convex,
- The intersection of convex sets is convex.

Remark 2.0.4. A concave maximization problem can be transformed into an equivalent convex minimization problem.

Remark 2.0.5. We may encounter a case where the constraint functions are not convex, but the feasible set is still convex. Here we do **not** consider it a convex optimization problem. We must strictly follow the definition.

3 Convex Optimization: Duality

Consider a general optimization problem (not necessarily convex) in the canonical form:

$$\begin{aligned} \min_{\mathbf{x}}: & f_0(\mathbf{x}) \\ \text{s.t.}: & f_i(\mathbf{x}) \leq 0, \quad \forall i \in \{1, 2, \dots, m\} \\ & h_j(\mathbf{x}) = 0, \quad \forall j \in \{1, 2, \dots, p\}. \end{aligned}$$

Denote the optimal value by p^* , and the optimal point by \mathbf{x}^* . Then we define the following associated functions:

Definition 3.0.1 (*Lagrangian Function*). The associated *Lagrangian function* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ is defined as follows:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}),$$

where $\{\lambda_i\}_{i=1}^m$ and $\{\nu_j\}_{j=1}^p$ are called the *Lagrange multipliers*.

Remark 3.0.1. This involves the idea of *relaxation*: we are more interested in a *nearby* problem which is easier to solve. The way we acquire a nearby problem is to move the constraints to the objective function, and penalize the violations of the constraints using the multipliers. A solution of a nearby problem provides information about the original problem.

Definition 3.0.2 (*Dual Function*). As motivated, the associated *dual function* $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ is defined as follows:

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are called the *dual variables*.

Remark 3.0.2. The motivation to define g is intuitive: since we have already considered the feasibility of \mathbf{x} in f_0 through the penalty, there is no need to add additional constraints on \mathbf{x} , i.e., we can relax the problem. However, it can be the case that L being minimal is due to negative penalty + not-optimized f_0 , which makes it only a nearby problem.

Remark 3.0.3. Regardless of the concavity of the original problem, g is always concave. To see this, if we traverse all $\mathbf{x} \in \mathbb{R}^n$, we will have a set of an infinite number of affine functions of $(\boldsymbol{\lambda}, \boldsymbol{\nu})^T$. The pointwise infimum function over such a set is concave.

Theorem 3.0.1 (*Weak Duality*). *With respect to an optimization problem, we have*

$$\forall \boldsymbol{\lambda} \geq \mathbf{0} : \forall \boldsymbol{\nu} \in \mathbb{R}^p : g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*.$$

*This property is called the weak duality, and it holds for **any** optimization problem.*

Proof.

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &:= \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= \inf_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*), \text{ since } \mathbf{x}^* \text{ is feasible} \\ &= p^*. \end{aligned}$$

□

Remark 3.0.4. Weak duality says that under $\boldsymbol{\lambda} \geq \mathbf{0}$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a lower bound for p^* . A natural question is then raised: what is $\max_{\boldsymbol{\lambda} \geq \mathbf{0}} g(\boldsymbol{\lambda}, \boldsymbol{\nu})$, i.e. the largest lower bound? Can it be equal to p^* ? In that case, we say the *strong duality* holds. We are interested in these questions, because it will be our best approximation of p^* from the dual perspective.

Definition 3.0.3 (*Dual Problem*). As motivated, we are to consider the following optimization problem:

$$\begin{aligned} \max_{(\boldsymbol{\lambda}, \boldsymbol{\nu})} & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{s.t.} & \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned}$$

which is called the associated *dual problem*, and the original one is called the *primal problem*. Denote the optimal value by d^* , and the optimal point by $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)^T$. The *duality gap* is defined as $p^* - d^*$.

Remark 3.0.5. Regardless of the convexity of the primal problem, the dual problem is always convex. To see this, we argued that g is always concave, and maximizing a concave function is equivalent to minimizing a convex function. In addition, the inequality constrained function is convex, and hence the problem is convex.

Remark 3.0.6. Under strong duality, we can solve p^* from the dual perspective, which is always convex. It turns out that most (but not all) convex optimization problems have strong duality. There are many results establishing conditions (called *constraint qualifications*) on the problem, under which the strong duality holds. We will see one below.

Definition 3.0.4 (*Relative Interior*). Let $S \subseteq \mathbb{R}^n$ be a set, then its *relative interior* is defined as

$$\text{relint}(S) := \{\mathbf{x} \in S \mid \exists r > 0 : (B(\mathbf{x}, r) \cap \text{aff}(S)) \subseteq C\},$$

where B is a ball of radius r centered at \mathbf{x} , i.e., $B(\mathbf{x}, r) = \{\mathbf{y} \mid \|\mathbf{y} - \mathbf{x}\| \leq r\}$, and $\text{aff}(S)$ is the affine hull of S , i.e. the smallest affine set that contains S .

Theorem 3.0.2 (*Slater's Condition*). *Given a convex optimization problem, strong duality holds if there exists a strictly feasible point in the relative interior of C , i.e.,*

$$\exists \mathbf{x} \in \text{relint}(C) : f_i(\mathbf{x}) < 0, \forall i \in \{1, 2, \dots, m\}, \text{ and } A\mathbf{x} - \mathbf{b} = \mathbf{0}.$$

In particular, when the inequality constraint functions are all affine, the feasibility does not have to be strict.

Proof. Skipped for now. □

Here are two immediate results followed from strong duality:

Proposition 3.0.1 (*Stationarity & Complementary Slackness*). *Assume strong duality holds, then we have stationarity:*

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = \nabla f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0},$$

and complementary slackness:

$$\forall i \in \{1, 2, \dots, m\} : \lambda_i^* f_i(\mathbf{x}^*) = 0.$$

Proof.

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &= \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &\leq L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &= f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*), \end{aligned}$$

which means that it should be equality everywhere. Therefore, \mathbf{x}^* is a minimizer of $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ over \mathbb{R}^n , and we have stationarity by the First-Order Optimality Condition. Due to the feasibility of \mathbf{x}^* and $\boldsymbol{\lambda}^*$, we have $\sum_{j=1}^p \nu_j^* h_j(\mathbf{x}^*) = 0$ and therefore $\sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) = 0$. The fact that $\lambda_i^* f_i(\mathbf{x}^*)$ is non-positive forces it to be zero, and then we have complementary slackness. \square

Theorem 3.0.3 (KKT Conditions). *The KKT conditions are as follows:*

- $\forall i \in \{1, 2, \dots, m\} : f_i(\mathbf{x}^*) \leq 0$ and $A\mathbf{x}^* - \mathbf{b} = \mathbf{0}$ primal feasibility
- $\boldsymbol{\lambda} \geq \mathbf{0}$ dual feasibility
- $\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$ stationarity
- $\forall i \in \{1, 2, \dots, m\} : \lambda_i^* f_i(\mathbf{x}^*) = 0$ complementary slackness

We have the following conclusions:

- For any optimization problem, if $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ satisfies the KKT conditions, then \mathbf{x}^* and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are primal and dual optimal. sufficiency
- Provided that the strong duality holds, if \mathbf{x}^* and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are primal and dual optimal, then $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ satisfies the KKT conditions. necessity

Putting up together, assume we have strong duality (e.g. convex problem + Slater's condition),

$$(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \text{ satisfies the KKT conditions} \Leftrightarrow \mathbf{x}^* \text{ and } (\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \text{ are primal and dual optimal.}$$

Proof. Necessity is trivial to prove (we in fact proved it from the above proposition). Regarding sufficiency, we assume $(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ satisfies the KKT conditions. By weak duality, we have

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}^*).$$

By assumption, we also have

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}) \\ &= f_0(\mathbf{x}), \end{aligned}$$

and hence we have $f_0(\mathbf{x}) \leq f_0(\mathbf{x}^*)$. It must be the case that $f_0(\mathbf{x}) = f_0(\mathbf{x}^*)$, or otherwise it will contradict with the optimality of \mathbf{x}^* . This proves that \mathbf{x} is primal optimal. In addition, we also have $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) = f_0(\mathbf{x}^*) = p^*$, and clearly, $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ has reached its maximum, which makes $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ dual optimal. \square

4 Convex Optimization: Algorithms

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex and differentiable function. We will first consider the unconstrained problem:

$$\min_{\mathbf{x}} : f(\mathbf{x}).$$

By the First-Order Optimality Condition, it is equivalent to solve

$$\nabla f(\mathbf{x}) = \mathbf{0},$$

which is a root-finding problem, where *fixed point iteration* can be found useful. Several algorithms in this section are instances of the fixed point iteration. Before moving on, we need some definitions first:

Proposition 4.0.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. With respect to some norm, f is L -Lipschitz continuous implies

$$\forall \mathbf{x} \in \mathbb{R}^n : \|\nabla f(\mathbf{x})\| \leq L,$$

that is, the gradient of f is bounded.

Proof.

□

Algorithm 1 Gradient Descent

Initialize \mathbf{x}_0 , ϵ , and $k = 0$.

while $\|\nabla f(\mathbf{x}_k)\| > \epsilon$ **do**

▷ Could use other stopping criteria

 Direction: $-\nabla f(\mathbf{x}_k)$.

 Step size: α_k .

 Update: $\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$.

$k := k + 1$.

end while

Theorem 4.0.1 (*Convergence Rate of Gradient Descent: Convex Case*). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex and differentiable function, and additionally ∇f is Lipschitz continuous with a constant $L > 0$, that is, $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n : \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 \leq L \|\mathbf{x}_1 - \mathbf{x}_2\|_2$. Then gradient descent with a fixed step size $\alpha \leq \frac{1}{L}$ satisfies:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\alpha k}.$$

This means that gradient descent is guaranteed to converge with rate $\mathcal{O}(\frac{1}{k})$, or reaching a sub-optimal tolerance level ϵ requires $\mathcal{O}(\frac{1}{\epsilon})$ iterations, where $\epsilon := |f(\mathbf{x}_k) - f(\mathbf{x}^*)|$.

Proof.

□

Theorem 4.0.2 (Convergence rate of Gradient descent: Strongly Convex Case). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex and differentiable function, where ∇f is Lipschitz continuous with a constant $L > 0$, and additionally f is strongly convex with a parameter m , that is, $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n : f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^T(\mathbf{x}_2 - \mathbf{x}_1) + \frac{m}{2} \|\mathbf{x}_2 - \mathbf{x}_1\|_2^2$. Then gradient descent with a fixed step size $\alpha \leq \frac{2}{m+L}$ satisfies:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\gamma^k L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2}, \text{ where } \gamma \in (0, 1).$$

This means that gradient descent is guaranteed to converge with rate $\mathcal{O}(\gamma^k)$, or reaching a sub-optimal tolerance level ϵ requires $\mathcal{O}(\frac{1}{\log(\frac{1}{\epsilon})})$ iterations.

Proof.

□

Algorithm 2 Newton's Method

Initialize \mathbf{x}_0 , and $k = 0$.

while $\|\nabla f(\mathbf{x}_k)\| > \epsilon$ **do**

▷ Could use other stopping criteria

Direction: $-\nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$.

Step size: α_k .

Update: $\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha_k \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$.

$k := k + 1$.

end while

Now, we will then consider the equality-constrained problem:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize:}} & f_0(\mathbf{x}) \\ \text{subject to:} & A\mathbf{x} - \mathbf{b} = \mathbf{0}. \end{array}$$

The idea is to eliminate the equality constraints.

Lastly, we will consider the general convex optimization problem:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize:}} && f_0(\mathbf{x}) \\ & \text{subject to:} && f_i(\mathbf{x}) \leq 0, \quad \forall i \in \{1, 2, \dots, m\} \\ & && h_j(\mathbf{x}) = 0, \quad \forall j \in \{1, 2, \dots, p\}. \end{aligned}$$

There are multiple algorithms to solve the unconstrained problem, and here we will focus on one algorithm called *Barrier methods*, a.k.a. *interior point methods*, (*IPM*). We will use *barrier functions* such that high cost will be added due to infeasibility.

Our first choice of the barrier function is the indicator function:

5 Other Useful Concepts

5.1 Inner Product

Definition 5.1.1 (*Inner Product*). An inner product on a vector space \mathcal{V} is a function

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$$

s.t. the following axioms hold:

- $\forall \mathbf{u}, \mathbf{v} \in \mathcal{V} : \langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle,$
- $\forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V} : \langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle,$
- $\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}, c \in \mathbb{R} : \langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle,$
- $\forall \mathbf{u} \in \mathcal{V} : \langle \mathbf{u}, \mathbf{u} \rangle \geq 0, \text{ and } \langle \mathbf{u}, \mathbf{u} \rangle = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}.$

A vector space \mathcal{V} together with an inner product $\langle \cdot, \cdot \rangle$ is called an inner product space, denoted by $(\mathcal{V}, \langle \cdot, \cdot \rangle)$.

Definition 5.1.2 (*Euclidean Inner Product*). The Euclidean inner product is an inner product on \mathbb{R}^n with $\langle \cdot, \cdot \rangle$ defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Sometimes it's referred to as the dot product, denoted by $\mathbf{u} \cdot \mathbf{v}$.

5.2 Norm

Definition 5.2.1 (*Norm*). W.r.t. the inner product space $(\mathcal{V}, \langle \cdot, \cdot \rangle)$, the induced norm is the function

$$\|\cdot\| : \mathcal{V} \rightarrow \mathbb{R}_{\geq 0}$$

where $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ for all $\mathbf{u} \in \mathcal{V}$.

Lemma 5.2.1 (*Properties of Norms*). Let $\|\cdot\|$ be a norm on \mathcal{V} ,

- $\forall \mathbf{u} \in \mathcal{V} : \|\mathbf{u}\| \geq 0, \text{ and } \|\mathbf{u}\| = 0 \Leftrightarrow \mathbf{u} = \mathbf{0},$
- $\forall \mathbf{u} \in \mathcal{V}, c \in \mathbb{R} : \|c\mathbf{u}\| = |c| \cdot \|\mathbf{u}\|.$

Theorem 5.2.1 (*Cauchy-Schwarz Inequality*). Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be an inner product space, then for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|.$$

The equality holds when they're linearly dependent.

Theorem 5.2.2 (*Triangle Inequality*). Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be an inner product space, then for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

Definition 5.2.2 (ℓ_p -Norm). On \mathbb{R}^n , the ℓ_p -norm $\|\cdot\|_p$ for some $p \geq 1$ is a norm s.t. for all $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{\frac{1}{p}}.$$

Remark 5.2.1. Here're some common ℓ_p -norms on \mathbb{R}^n :

- $p = 1$: $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$

- $p = 2$: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{\mathbf{x}^T \mathbf{x}}$, which is called the Euclidean norm
- $p = \infty$: $\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_n|\}$

Definition 5.2.3 (*Dual Norm*). Let $\|\cdot\|$ be a norm on \mathbb{R}^n , then its associated dual norm $\|\cdot\|_*$ is defined as

$$\|\mathbf{u}\|_* = \sup\{\mathbf{u}^T \mathbf{v} : \|\mathbf{v}\| \leq 1, \mathbf{v} \in \mathbb{R}^n\}$$

for all $\mathbf{u} \in \mathbb{R}^n$.

6 References