Basics of Convex Optimization

Yanze Song

December 24, 2023

1 Introduction to Convexity

1.1 Convex Sets

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

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

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

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

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

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

1.2 Convex Set Examples

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

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

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 $\mathcal{P} = \{\boldsymbol{x} \,|\, A\boldsymbol{x} = \boldsymbol{b}, B\boldsymbol{x} \leq \boldsymbol{d}\}$ be a polyhedron, then for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{P}, \theta \in [0,1]$, consider $\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}$:

$$A(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) = \theta A\boldsymbol{x} + (1-\theta)A\boldsymbol{y} = \theta \boldsymbol{b} + (1-\theta)\boldsymbol{b} = \boldsymbol{b},$$
$$B(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) = \theta B\boldsymbol{x} + (1-\theta)B\boldsymbol{y} \leq \theta \boldsymbol{d} + (1-\theta)\boldsymbol{d} = \boldsymbol{d}.$$

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

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

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

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

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

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

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 $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ be an affine function, then $\boldsymbol{f}(\mathcal{C})$, i.e. the image of $\boldsymbol{f}$ over $\mathcal{C}$ is also convex.*

Proof. For all $\boldsymbol{y_1}, \boldsymbol{y_2} \in \boldsymbol{f}(\mathcal{C})$, there exist $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{C}$ s.t.

$$\boldsymbol{f}(\boldsymbol{x}_1) = A\boldsymbol{x}_1 + \boldsymbol{b} = \boldsymbol{y_1},\ \boldsymbol{f}(\boldsymbol{x}_2) = A\boldsymbol{x}_2 + \boldsymbol{b} = \boldsymbol{y_2}.$$

We'll then show that $\theta\boldsymbol{y_1} + (1-\theta)\boldsymbol{y_2} \in \boldsymbol{f}(\mathcal{C})$ for all $\theta \in [0,1]$ as well:

$$\theta\boldsymbol{y_1} + (1-\theta)\boldsymbol{y_2} = \theta(A\boldsymbol{x_1} + \boldsymbol{b}) + (1-\theta)(A\boldsymbol{x_2} + \boldsymbol{b}) = A(\theta\boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2) + \boldsymbol{b} = \boldsymbol{f}(\theta\boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2),$$

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

1.4 Convex Functions

Definition 1.4.1 (*Convex Functions*)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is convex if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $\theta \in [0,1]$,*

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \le \theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y}).$$

Definition 1.4.2 (*Strictly Convex Functions*)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is strictly convex if for all $\boldsymbol{x} \neq \boldsymbol{y} \in \mathcal{C}$, $\theta \in (0,1)$,*

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

Definition 1.4.3 (*Concave Functions*)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is concave if $-f$ is convex. It is equivalent to define that f is concave if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $\theta \in [0,1]$,*

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \ge \theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y}).$$

Definition 1.4.4 (*Affine Functions*)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is affine if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\theta \in [0,1]$,*

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

This is equivalent to definition 1.3.1.

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

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

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

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

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

- g preserves scalar multiplication, i.e., $\forall\, \alpha \in \mathbb{R} = [0,1] \cup (1,\infty) \cup (-\infty,0) : g(\alpha\boldsymbol{x}) = \alpha g(\boldsymbol{x})$:
  - $\alpha \in [0,1] \Rightarrow \alpha, 1-\alpha \in [0,1]$:
  $$g(\alpha\boldsymbol{x}) = g(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{0}) = \alpha g(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{0}) = \alpha g(\boldsymbol{x}) \tag{1}$$
  - $\alpha \in (1,\infty) \Rightarrow \frac{1}{\alpha} \in [0,1]$:
  $$g(\alpha\boldsymbol{x}) = \alpha\frac{1}{\alpha}g(\alpha\boldsymbol{x}) \overset{(1)}{=} \alpha g(\frac{1}{\alpha}\alpha\boldsymbol{x}) = \alpha g(\boldsymbol{x}) \tag{2}$$
  - $\alpha \in (-\infty,0) \Rightarrow -\alpha \in (0,\infty)$:
  $$g(\alpha\boldsymbol{x}) = g(-\alpha(-\boldsymbol{x})) = -\alpha g(\boldsymbol{0} - \boldsymbol{x}) \overset{(4)}{=} -\alpha(g(\boldsymbol{0}) - g(\boldsymbol{x})) = \alpha g(\boldsymbol{x}) \tag{3}$$

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

$$g(\boldsymbol{x}+\boldsymbol{y}) = g(\theta\frac{1}{\theta}\boldsymbol{x} + (1-\theta)\frac{1}{1-\theta}\boldsymbol{y}) = \theta g(\frac{1}{\theta}\boldsymbol{x}) + (1-\theta)g(\frac{1}{1-\theta}\boldsymbol{y}) \overset{(1)}{=} g(\boldsymbol{x}) + g(\boldsymbol{y}). \tag{4}$$

□

1.5 Theorems of Convex Functions

Proposition 1.5.1. *Let $f : \mathcal{C} \to \mathbb{R}$ be a function, then*

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

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

Theorem 1.5.1 (*Convex Along All Lines*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a function, then f is convex if and only if for all $\boldsymbol{x} \in \mathcal{C}$, $\boldsymbol{d} \in \mathcal{D} = \{\boldsymbol{d} \in \mathbb{R}^n \,|\, \exists\, t > 0 : \boldsymbol{x} + t\boldsymbol{d} \in \mathcal{C}\}$, i.e. the set of all feasible directions at $\boldsymbol{x}$,*

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

where dom(g) $= \{t \,|\, \boldsymbol{x} + t\boldsymbol{d} \in \mathcal{C}\}$.

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

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

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

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

□

Theorem 1.5.2 (*First Order Convexity Condition*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a differentiable function, then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$,*

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

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

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

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

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

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

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

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

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

Remark 1.5.1. Here're some details for $g'(\theta)$:

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

so by the Chain Rule,

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

Proposition 1.5.2 (*Subgradient*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a convex function, then for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, there exist $\boldsymbol{z}$ s.t.*

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

Such an $\boldsymbol{z}$ is called a subgradient of f at $\boldsymbol{x}$, and the set of all subgradients of f at $\boldsymbol{x}$ is denoted $\partial f(\boldsymbol{x})$. If f is differentiable, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.

Proof. □

Theorem 1.5.3 (*Second Order Convexity Condition*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a twice differentiable function, then f is convex if and only if for all $\boldsymbol{x} \in \mathcal{C}$,*

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

Proof. □

1.6 Optimality of Convex Functions

Theorem 1.6.1 (*Local & Global Optimality*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a convex function, then any locally optimal point is also globally optimal.*

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

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

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

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

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

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

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

□

Theorem 1.6.2 (*First Order Optimality Condition*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be convex and differentiable, then $\boldsymbol{x}^*$ minimizes f over $\mathcal{C}$ if and only if for all $\boldsymbol{x} \in \mathcal{C}$,*

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

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

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

Proof. To prove by contradiction, assume $\boldsymbol{x}^*$ is the minimizer of f over $\mathcal{C}$, and there exists $\boldsymbol{x}_0 \in \mathcal{C}$ s.t. $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle < 0$. Similar to Theorem 1.5.2, define $g(\theta) = f(\theta \boldsymbol{x}_0 + (1-\theta)\boldsymbol{x}^*)$, then

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

which implies that $g(0+\theta) < g(0)$ for some small $\theta > 0$, contradicting with the minimality of $\boldsymbol{x}^*$. Conversely, assume $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0 \Rightarrow f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq f(\boldsymbol{x}^*)$. By Theorem 1.5.2, for all $\boldsymbol{x} \in \mathcal{C}$,

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

showing that $\boldsymbol{x}^*$ is the minimizer of f over $\mathcal{C}$. If $\mathcal{C} = \mathbb{R}^n$, assume $\boldsymbol{x}^*$ minimizes f over $\mathbb{R}^n$, and $\frac{\partial}{\partial x_i} f(\boldsymbol{x}^*) > 0$ for some i w.l.o.g.. We then have $f(\boldsymbol{x}^* - \theta \boldsymbol{e}_i) < f(\boldsymbol{x}^*)$ for some small $\theta > 0$, which contradict with the minimality of $\boldsymbol{x}^*$. The converse can be proved directly using (7). $\square$

1.7 Other types of Convex Functions

Definition 1.7.1 (*Strongly Convex Functions*)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is σ-strongly convex w.r.t. some norm $\|\cdot\|$ if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $\theta \in [0, 1]$,*

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

Definition 1.7.2 (*Exponentially Concave Functions*)**.**

Definition 1.7.3 (*Lipschitz Continuity*)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz continuous w.r.t. some norm $\|\cdot\|$ if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, we have*

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

Proposition 1.7.1 (*Quadratic Lower Bound of Strongly Convex Functions*)**.**

Proof. $\square$

Proposition 1.7.2 (*Addition of Strongly Convex Functions*)**.**

Proof. $\square$

Proposition 1.7.3 (*Convexity of Exponentially Concave Functions*)**.**

Proof. $\square$

Proposition 1.7.4 (*Lipschitzness, Convexity and Bounded Gradient*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be convex, differentiable and L-Lipschitz w.r.t. some norm $\|\cdot\|$, then for all $\boldsymbol{x} \in \mathcal{C}$, we have*

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

Proof. For all $\boldsymbol{x} \in \mathrm{int}(\mathcal{C})$, there exist $\eta > 0$ s.t. $\boldsymbol{x} + \eta \nabla f(\boldsymbol{x}) \in \mathcal{C}$. By Theorem 1.5.2,

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

then by lipschitzness,

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

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

1.8 Convex Function Examples

Example 1.8.1 (*Common Convex and Concave Functions*)**.** The convexity of the following functions can be proved using the above two theorems:

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

1.9 Relationship between Convex Sets and Convex Functions

Definition 1.9.1 (*Graphs, Epigraphs, Level Sets and Sublevel Sets*)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a function, then*

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

Proposition 1.9.1. *Let $f : \mathcal{C} \to \mathbb{R}$ be a function, then f is convex if and only if*

epi(f) is convex.

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

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

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

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

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

Proposition 1.9.2. *Let $f : \mathcal{C} \to \mathbb{R}$ be convex, then for all α,*

C_α is convex.

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

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

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

1.10 Operations that Preserve the Convexity of Functions

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

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

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

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

□

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

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

where dom(g) $= \{\boldsymbol{x} \,|\, A\boldsymbol{x} + \boldsymbol{b} \in \mathcal{C}\}$.

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

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

□

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

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

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

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

□

2 Convex Optimization: Concepts

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

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

Here are some related terminologies:

- $\boldsymbol{x}$: optimization variable
- f_0: objective function
- $\{f_i(\boldsymbol{x}) \leq 0\}_{i=1}^m$: inequality constraints
- $\{h_i(\boldsymbol{x}) = 0\}_{j=1}^p$: equality constraints
- A point $\boldsymbol{x}$ 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 $\inf_{\boldsymbol{x}} \{f_0(\boldsymbol{x}) \,|\, \boldsymbol{x} \in C\}$, which may or may not be attainable.
- A feasible point $\boldsymbol{x}^*$ is *globally optimal*, or *optimal* if $f(\boldsymbol{x}^*) = p^*$. There may be multiple optimal points.
- A feasible point $\boldsymbol{x}^*$ is *locally optimal* if $\exists\, R > 0 : f_0(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} \{f_0(\boldsymbol{x}) \,|\, \boldsymbol{x} \in C \text{ and } ||\boldsymbol{x} - \boldsymbol{x}^*|| \leq R\}$.
- 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:

$$\underset{\boldsymbol{x}}{\text{minimize:}} \;\; f_0(\boldsymbol{x}) + I_C(\boldsymbol{x})$$

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

$$\begin{aligned} I_C \colon \; & \mathbb{R}^n \to \mathbb{R} \\ \boldsymbol{x} \;\; & \mapsto \begin{cases} f_0(\boldsymbol{x}), \text{ if } \boldsymbol{x} \in C \\ \infty, \text{ otherwise.} \end{cases} \end{aligned}$$

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:

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

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

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

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

or equivalently, it has the form:

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

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 $\{h_j(\boldsymbol{x}) = \boldsymbol{a}_j^T \boldsymbol{x} - b_j\}_{j=1}^p$ 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 $\{\boldsymbol{x} \,|\, \boldsymbol{a}_j^T \boldsymbol{x} - b_j = 0\}_{j=1}^p$, 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:

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

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

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

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

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* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as follows:

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

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 $\boldsymbol{x}$ in f_0 through the penalty, there is no need to add additional constraints on $\boldsymbol{x}$, 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 $\boldsymbol{x} \in \mathbb{R}^n$, 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.

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

□

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

$$\operatorname{relint}(S) \coloneqq \{\boldsymbol{x} \in S \,|\, \exists\, r > 0 : (B(\boldsymbol{x}, r) \cap \operatorname{aff}(S)) \subseteq C\},$$

where B is a ball of radius r centered at $\boldsymbol{x}$, i.e., $B(\boldsymbol{x}, r) = \{\boldsymbol{y} \,|\, ||\boldsymbol{y} - \boldsymbol{x}|| \leq r\}$, and $\operatorname{aff}(S)$ 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.,*

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

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

Proof. Skipped for now. $\square$

Here are two immediate results followed from strong duality:

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

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

and complementary slackness:

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

Proof.

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

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

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

We have the following conclusions:

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

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

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

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

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

By assumption, we also have

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

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

4 Convex Optimization: Algorithms

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

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

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

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

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

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

that is, the gradient of f is bounded.

Proof. $\square$

Algorithm 1 Gradient Descent

```
Initialize x_0, ε, and k = 0.
while ||∇f(x_k)|| > ε do                    ▷ Could use other stopping criteria
    Direction: −∇f(x_k).
    Step size: α_k.
    Update: x_{k+1} := x_k − α_k ∇f(x_k).
    k := k + 1.
end while
```

Theorem 4.0.1 (*Convergence Rate of Gradient Descent: Convex Case*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function, and additionally ∇f is Lipschitz continuous with a constant $L > 0$, that is, $\forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n$: $||\nabla f(\boldsymbol{x_1}) - \nabla f(\boldsymbol{x_2})||_2 \leq L||\boldsymbol{x_1} - \boldsymbol{x_2}||_2$. Then gradient descent with a fixed step size $\alpha \leq \frac{1}{L}$ satisfies:*

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

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 $\epsilon := |f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)|$.

Proof. □

Theorem 4.0.2 (*Convergence rate of Gradient descent: Strongly Convex Case*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ 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, $\forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n : f(\boldsymbol{x_2}) \geq f(\boldsymbol{x_1}) + \nabla f(\boldsymbol{x_1})^T(\boldsymbol{x_2} - \boldsymbol{x_1}) + \frac{m}{2}||\boldsymbol{x_2} - \boldsymbol{x_1}||_2^2$. Then gradient descent with a fixed step size $\alpha \leq \frac{2}{m+L}$ satisfies:*

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

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 $\boldsymbol{x}_0$, and $k = 0$.
while $||\nabla f(x_k)|| > \epsilon$ **do** ▷ Could use other stopping criteria
  Direction: $-\nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)$.
  Step size: α_k.
  Update: $\boldsymbol{x}_{k+1} \coloneqq \boldsymbol{x}_k - \alpha_k \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)$.
  $k \coloneqq k + 1$.
end while

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

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize:}} \quad & f_0(\boldsymbol{x}) \\ \text{subject to:} \quad & A\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}. \end{aligned}$$

The idea is to eliminate the equality constraints.

Lastly, we will consider the general convex optimization problem:

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

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*

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

s.t. the following axioms hold:

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

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*

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

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

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*

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

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

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

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

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

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

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 $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}$,*

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

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 $\boldsymbol{x} \in \mathbb{R}^n$,*

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

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

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

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

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

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

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

6 References