Basics of Convex Optimization

Yanze Song

November 29, 2023

1 Intro 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*

$$\{\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \mid \theta \in \mathbb{R}\} \text{ and } \{\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \mid \theta \in [0,1]\}$$

represent the line and the line segment passing through $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively.

Definition 1.1.2 (*Affine Sets*)**.** *A set $\mathcal{A} \subseteq \mathbb{R}^n$ is affine if the line through any two points in $\mathcal{A}$ lies in $\mathcal{A}$. That is, $\mathcal{A}$ 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 the line segment through any two points in $\mathcal{C}$ lies in $\mathcal{C}$. That is, $\mathcal{C}$ 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*)**.** *Let $\boldsymbol{a} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, $b \in \mathbb{R}$, then a hyperplane has the form*

$$\{\boldsymbol{x} \,|\, \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x} = b\},$$

i.e. the solution set of a linear equation, in particular, a line in $\mathbb{R}^2$ and a plane in $\mathbb{R}^3$.

Definition 1.2.2 (*Closed Halfspaces*)**.** *Let $\boldsymbol{a} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, $b \in \mathbb{R}$, then a halfspace has the form*

$$\{\boldsymbol{x} \,|\, \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x} \leq b\},$$

i.e. the solution set of a linear inequality. A hyperplane divides a space into two halfspaces.

Definition 1.2.3 (*Polyhedra*)**.** *Let $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_m(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{d} \in \mathbb{R}^m$, then a polyhedron has the form*

$$\{\boldsymbol{x} \,|\, A\boldsymbol{x} = \boldsymbol{b}, B\boldsymbol{x} \leq \boldsymbol{d}\},$$

i.e. the solution set of a finite number of linear equations and inequalities, or geometically, the intersection of hyperplanes and halfspaces.

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

Proof. We will show the convexity of polyhedra first, and then the convexity of the preceding two sets immediately follows. Hyperplanes will implicitly be proven to be affine.

Let $P = \{\boldsymbol{x} \,|\, A\boldsymbol{x} = \boldsymbol{b},\, B\boldsymbol{x} \leq \boldsymbol{d}\}$ be a polyhedron, and take arbitrary $\boldsymbol{x_1}, \boldsymbol{x_2} \in P$, $\theta \in [0,1]$.
Then consider $\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}$:

$$A(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) = \theta A\boldsymbol{x_1} + (1-\theta)A\boldsymbol{x_2} = \theta\boldsymbol{b} + (1-\theta)\boldsymbol{b} = \boldsymbol{b},$$
$$B(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) = \theta B\boldsymbol{x_1} + (1-\theta)B\boldsymbol{x_2} \leq \theta\boldsymbol{d} + (1-\theta)\boldsymbol{d} = \boldsymbol{d}.$$

We proved that $\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in P$, and therefore P is convex. $\square$

Definition 1.2.4 (*Affine Functions*)**.** *Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, then $\boldsymbol{f}$ is affine if there exists an $A \in \mathcal{M}_{m\times n}$ 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 1.2.2. *Let $S \subseteq \mathbb{R}^n$ be a convex set, and $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ be an affine function, then $\boldsymbol{f}(S)$, i.e. the image of $\boldsymbol{f}$ over S is also convex.*

Proof. Take arbitrary $\boldsymbol{y_1}, \boldsymbol{y_2} \in \boldsymbol{f}(S)$, $\theta \in [0, 1]$, then we have:

$$\exists\, \boldsymbol{s_1}, \boldsymbol{s_2} \in S : \boldsymbol{f}(\boldsymbol{s_1}) = \boldsymbol{y_1}, \boldsymbol{f}(\boldsymbol{s_2}) = \boldsymbol{y_2}.$$

Then we will show that $\theta\boldsymbol{y_1} + (1-\theta)\boldsymbol{y_2} \in \boldsymbol{f}(S)$ as well:

$$\begin{aligned}
\theta\boldsymbol{y_1} + (1-\theta)\boldsymbol{y_2} &= \theta\boldsymbol{f}(\boldsymbol{s_1}) + (1-\theta)\boldsymbol{f}(\boldsymbol{s_2}) \\
&= \theta(A\boldsymbol{s_1} + \boldsymbol{b}) + (1-\theta)(A\boldsymbol{s_2} + \boldsymbol{b}), \text{ since } \boldsymbol{f} \text{ is affine} \\
&= A(\theta\boldsymbol{s_1} + (1-\theta)\boldsymbol{s_2}) + \boldsymbol{b} \\
&= \boldsymbol{f}(\theta\boldsymbol{s_1} + (1-\theta)\boldsymbol{s_2}), \text{ where } \theta\boldsymbol{s_1} + (1-\theta)\boldsymbol{s_2} \in S \text{ since } S \text{ is convex.}
\end{aligned}$$

Therefore,

$$\exists\, \boldsymbol{s} \in S : \boldsymbol{f}(\boldsymbol{s}) = \theta\boldsymbol{y_1} + (1-\theta)\boldsymbol{y_2} \implies \theta\boldsymbol{y_1} + (1-\theta)\boldsymbol{y_2} \in \boldsymbol{f}(S),$$

and this completes the proof. □

Proposition 1.2.3. *Let $S_1, S_2 \subseteq \mathbb{R}^n$ be convex sets, then $S_1 \cap S_2$ is also convex.*

Proof. Take arbitrary $\boldsymbol{x_1}, \boldsymbol{x_2} \in S_1 \cap S_2$, $\theta \in [0, 1]$, and hence $\boldsymbol{x_1}, \boldsymbol{x_2} \in S_1, S_2$. Since S_1, S_2 are convex,

$$\begin{aligned}
\boldsymbol{x_1} \in S_1,\, \boldsymbol{x_2} \in S_1 &\implies \theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in S_1, \\
\boldsymbol{x_1} \in S_2,\, \boldsymbol{x_2} \in S_2 &\implies \theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in S_2.
\end{aligned}$$

That is,

$$\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in S_1 \cap S_2,$$

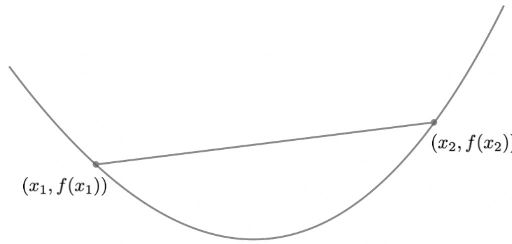
and this completes the proof. □

1.3 Convex Functions

Definition 1.3.1 (*Convex Functions*)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is convex if the line segment between any two points on f lies above or on the corresponding part of f. That is, f is convex if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $\theta \in [0, 1]$,*

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

Example 1.3.1. A convex function in $\mathbb{R}^2$:



Definition 1.3.2 (*Strictly Convex Functions*)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is strictly convex if the line segment between any two distinct points on f lies exactly above the corresponding part of f. That is, f is strictly convex if for all $\boldsymbol{x} \neq \boldsymbol{y} \in \mathcal{C}$, $\theta \in (0, 1)$,*

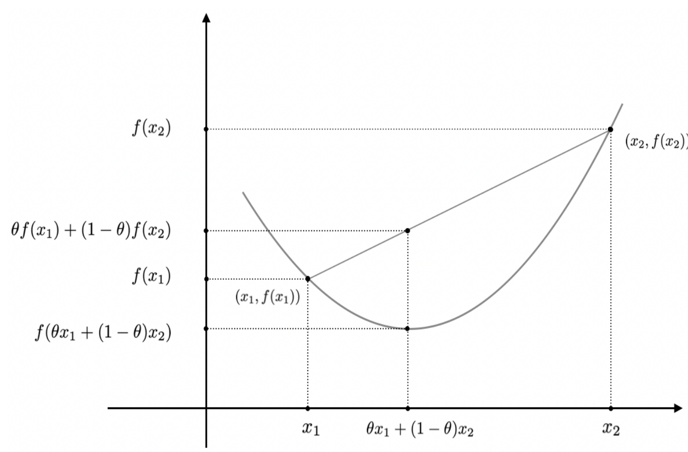
$$f(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y}) < \theta f(\boldsymbol{x}) + (1 - \theta) f(\boldsymbol{y}).$$

Definition 1.3.3 (*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 $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $\theta \in [0, 1]$,*

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

Similar definition applies to strictly concave functions as well.

Remark 1.3.1. The following graph my help understand the definition of convex functions:



Proposition 1.3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, then f is affine $\Leftrightarrow$ f is both convex and concave.*

Proof. Take arbitrary $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n$, $\theta \in [0,1]$. We can show that $f(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) = \theta f(\boldsymbol{x_1}) + (1-\theta)f(\boldsymbol{x_2})$ since f is affine, and this proves $\Rightarrow$. Conversely, since f is convex and concave, we have $f(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) = \theta f(\boldsymbol{x_1}) + (1-\theta)f(\boldsymbol{x_2})$. Define g such that $g(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{0})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. It is equivalent to show that g is linear:

- Closed under scalar multiplication:
  - Case 1 ($\alpha \in [0,1]$):

$$\begin{aligned} g(\alpha\boldsymbol{x}) &= f(\alpha\boldsymbol{x}) - f(\boldsymbol{0}) \\ &= f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{0}) - f(\boldsymbol{0}) \\ &= \alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{0}) - f(\boldsymbol{0}) \\ &= \alpha f(\boldsymbol{x}) - \alpha f(\boldsymbol{0}) \\ &= \alpha g(\boldsymbol{x}) \end{aligned}$$

  - Case 2 ($\alpha \in (1,\infty)$):

$$\begin{aligned} g(\alpha\boldsymbol{x}) &= \alpha \cdot \frac{1}{\alpha} g(\alpha\boldsymbol{x}) \\ &= \alpha g(\frac{1}{\alpha} \cdot \alpha\boldsymbol{x}), \text{ since } \frac{1}{\alpha} \in (0,1) \\ &= \alpha g(\boldsymbol{x}) \end{aligned}$$

- Closed under addition:

$$\begin{aligned} g(\boldsymbol{x_1} + \boldsymbol{x_2}) &= g(\frac{1}{2} \cdot 2\boldsymbol{x_1} + \frac{1}{2} \cdot 2\boldsymbol{x_2}) \\ &= \frac{1}{2} g(2\boldsymbol{x_1}) + \frac{1}{2} g(2\boldsymbol{x_2}) \\ &= \frac{1}{2} \cdot 2g(\boldsymbol{x_1}) + \frac{1}{2} \cdot 2g(\boldsymbol{x_2}) \\ &= g(\boldsymbol{x_1}) + g(\boldsymbol{x_2}) \end{aligned}$$

- Closed under scalar multiplication:
  - Case 3 ($\alpha \in (-\infty, 0)$):

$$\begin{aligned} & g(-\beta\boldsymbol{x}) = \beta g(-\boldsymbol{x}) = \beta g(\boldsymbol{x} - 2\boldsymbol{x}) = \beta g(\boldsymbol{x}) + g(-2 \cdot \beta\boldsymbol{x}), \text{ for some } \beta \in (0,\infty) \\ \implies & g(-\beta\boldsymbol{x}) = \beta g(\boldsymbol{x}) + g(-2 \cdot \beta\boldsymbol{x}) \\ \implies & -\beta g(\boldsymbol{x}) = g(-2 \cdot \beta\boldsymbol{x}) - g(-\beta\boldsymbol{x}) \\ \implies & -\beta g(\boldsymbol{x}) = g(-\beta\boldsymbol{x}) \end{aligned}$$

  Letting $\alpha = -\beta$ yields $\forall\, \alpha \in (-\infty, 0) : g(\alpha\boldsymbol{x}) = \alpha g(\boldsymbol{x})$, which completes the proof.

$\square$

Theorem 1.3.1 (*First-Order Convexity Condition*)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a differentiable function, then f is convex if and only if*

$$f \textit{is convex} \Leftrightarrow \forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n : f(\boldsymbol{x_2}) \geq f(\boldsymbol{x_1}) + \nabla f(\boldsymbol{x_1})^T (\boldsymbol{x_2} - \boldsymbol{x_1}).$$

Proof. To simplify, we assume f lies in $\mathbb{R}^2$. Let f be convex, so for any $x_1, x_2 \in \mathbb{R}$, and $\theta \in [0, 1]$, we have

$$\begin{aligned}
& f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2) \\
\implies & f(x_2 + \theta(x_1 - x_2)) \leq f(x_2) + \theta(f(x_1) - f(x_2)) \\
\implies & \frac{f(x_2 + \theta(x_1 - x_2)) - f(x_2)}{\theta} + f(x_2) \leq f(x_1) \\
\implies & f(x_2) + \frac{g(\theta) - g(0)}{\theta} \leq f(x_1), \text{ where } g(\theta) = f(x_2 + \theta(x_1 - x_2)) \\
\implies & f(x_2) + g'(0) \leq f(x_1), \text{ by letting } \theta \to 0.
\end{aligned}$$

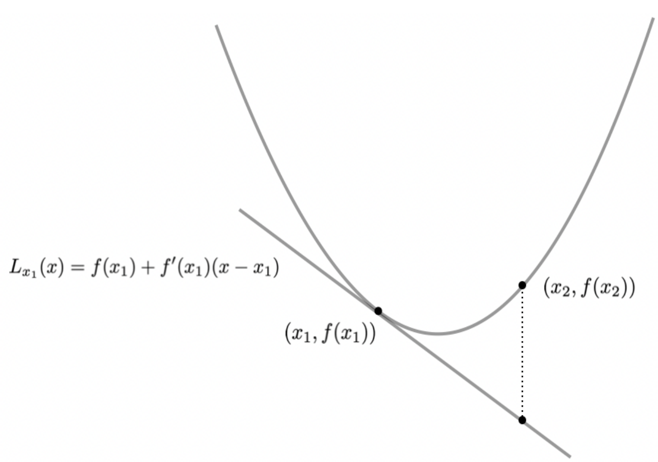
In addition, $g'(\theta) = f'(x_2 + \theta(x_1 - x_2))(x_1 - x_2) \implies g'(0) = f'(x_2)(x_1 - x_2)$. Plugging in we have $f(x_2) + f'(x_2)(x_1 - x_2) \leq f(x_1)$.

Conversely, let the RHS holds. Take arbitrary $x, y \in \mathbb{R}$, and $t \in [0, 1]$. Let $z = tx + (1-t)y$. Then we have

$$\begin{aligned}
f(z) + f'(z)(x - z) \leq f(x) &\implies t f(z) + t f'(z)(x - z) \leq t f(x) \\
f(z) + f'(z)(y - z) \leq f(y) &\implies (1-t) f(z) + (1-t) f'(z)(y - z) \leq (1-t) f(y)
\end{aligned}$$

Adding up we have $f(z) + f'(z)(tx + (1-t)y - z) = f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$, which completes the proof. $\square$

Remark 1.3.2. This theorem says that any linearization of f is a global under-estimator of f.



Theorem 1.3.2 (*Second-Order Convexity Condition*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, then*

$$f \text{ is convex} \Leftrightarrow \forall \boldsymbol{x} \in \mathbb{R}^n : \nabla^2 f(\boldsymbol{x}) \succeq 0.$$

Proof. To simplify, we assume f lies in $\mathbb{R}^2$. We will use the above theorem to prove this one. Let f be convex, and $x_1, x_2 \in \mathbb{R}$ be two distinct points, so we have

$$f(x_2) \geq f(x_1) + f'(x_1)(x_2 - x_1) \text{ and } f(x_1) \geq f(x_2) + f'(x_2)(x_1 - x_2),$$

and hence we have

$$f'(x_1)(x_2 - x_1) \leq f(x_2) - f(x_1) \leq f'(x_2)(x_2 - x_1).$$

Dividing LHS and RHS by $(x_2 - x_1)^2$ yields

$$\forall x_1, x_2\,(x_1 \neq x_2) \in \mathbb{R} : \frac{f'(x_2) - f'(x_1)}{x_2 - x_1} \geq 0,$$

and letting $x_2 \to x_1$ yields

$$\forall x_1 \in \mathbb{R} : f''(x_1) \geq 0.$$

Conversely, assume $f''(x) \geq 0$ for all $x \in \mathbb{R}$. By Taylor's theorem, for all distinct $x_1, x_2 \in \mathbb{R}$, we have

$$f(x_2) = f(x_1) + f'(x_1)(x_2 - x_1) + \frac{f''(\xi)(x_2 - x_1)^2}{2}, \text{ for some } \xi \in (x_1, x_2),$$

assuming $x_1 < x_2$ without loss of generality. By assumption, we then have

$$f(x_2) \geq f(x_1) + f'(x_1)(x_2 - x_1).$$

The case where $x_1 = x_2$ is trivial to prove. $\square$

Remark 1.3.3. In $\mathbb{R}^2$, the theorem says that f is convex $\Leftrightarrow$ f is concave up, or affine.

Example 1.3.2 (*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
- Linear and affine functions are both convex and concave
- $logx$ is concave
- Geometric mean $f(\boldsymbol{x}) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave on $\mathbb{R}^n_{\geq 0}$

Remark 1.3.4. Before moving on to the propositions relating convex functions and convex sets, we will look at some definitions first.

Definition 1.3.4 (*Graph & Epigraph*)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, then the *graph* of f is defined as

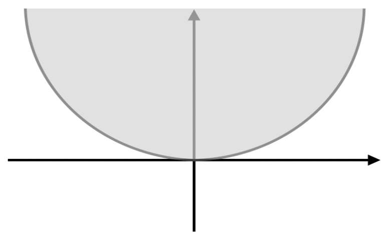
$$\{(\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^{n+1},$$

and the *epigraph* of f is defined as

$$\{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathbb{R}^n, t \geq f(\boldsymbol{x})\} \subseteq \mathbb{R}^{n+1}.$$

Visually, the *epigraph* of f is the area above the graph of f.

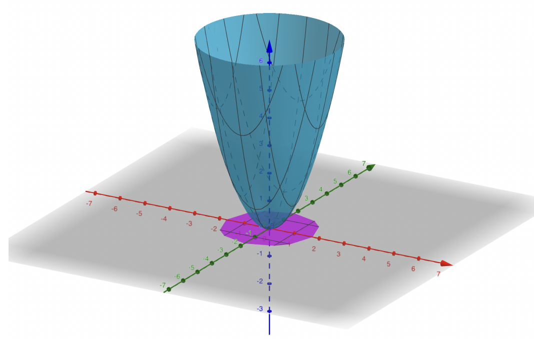
e.g. (Epigraph).



Definition 1.3.5 (*Sublevel Sets*)**.** The α*-sublevel set* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$C_\alpha \coloneqq \{\boldsymbol{x} \mid f(\boldsymbol{x}) \leq \alpha\}.$$

e.g. (Sublevel set).



The area in purple is the 3-sublevel set of $f(x, y) = x^2 + y^2$

Proposition 1.3.2. *A function is convex $\Leftrightarrow$ its epigraph is convex.*

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, and $E = \{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathbb{R}^n, t \geq f(\boldsymbol{x})\}$ be its epigraph.

Let f be convex. Take $p_1 = (\boldsymbol{x_1}, t_1), p_2 = (\boldsymbol{x_2}, t_2) \in E$, for some arbitrary $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n, t_1 \geq f(\boldsymbol{x_1})$, and $t_2 \geq f(\boldsymbol{x_2})$. We want to show

$$\theta \boldsymbol{p_1} + (1-\theta)\boldsymbol{p_2} = (\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}, \theta t_1 + (1-\theta)t_2) \in E$$

for some arbitrary $\theta \in [0,1]$ as well. Trivially, $\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in \mathbb{R}$. Regarding the last component, we have

$$\theta t_1 + (1-\theta)t_2 \geq \theta f(\boldsymbol{x_1}) + (1-\theta) f(\boldsymbol{x_2}) \geq f(\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}),$$

so $\theta \boldsymbol{p_1} + (1-\theta)\boldsymbol{p_2} \in E \implies E$ is convex.

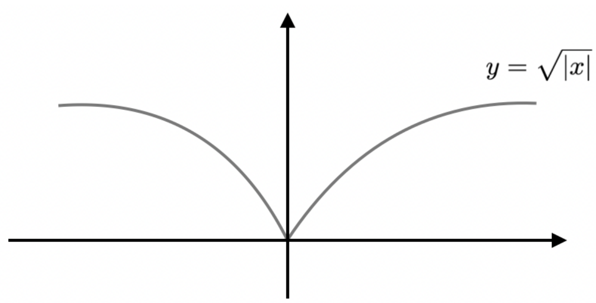
Conversely, let E be convex. Take $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n$, and $\theta \in [0,1]$ arbitrarily, and then consider $p_1 = (\boldsymbol{x_1}, f(\boldsymbol{x_1})), p_2 = (\boldsymbol{x_2}, f(\boldsymbol{x_2})) \in E$. Since E is convex, we have $\theta \boldsymbol{p_1} + (1-\theta)\boldsymbol{p_2} \in E$, in particular, $\theta f(\boldsymbol{x_1}) + (1-\theta) f(\boldsymbol{x_2}) \geq f(\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2})$, and therefore f is convex. $\square$

Proposition 1.3.3. *Any sublevel set of a convex function is convex.*

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Take $\alpha \in \mathbb{R}$, $\boldsymbol{x_1}, \boldsymbol{x_2} \in C_\alpha$, and $\theta \in [0,1]$ arbitrarily. Then we have

$$f(\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) \leq \theta f(\boldsymbol{x_1}) + (1-\theta) f(\boldsymbol{x_2}) \leq \theta\alpha + (1-\theta)\alpha = \alpha,$$

and hence $\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in C_\alpha \implies C_\alpha$ is convex. Note that the converse is not true: consider the following function: $\square$



Here are three common operations that preserve the convexity of functions:

Proposition 1.3.4 (*Non-negative Weighted Sums*)**.** *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex functions, and $\omega_1, \omega_2 \geq 0$, then $f \coloneqq \omega_1 f_1 + \omega_2 f_2$ is also convex.*

Proof. Take $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n$ and $\theta \in [0, 1]$ arbitrarily, then

$$\begin{aligned} f(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) &= (\omega_1 f_1 + \omega_2 f_2)(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) \\ &= \omega_1 f_1(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) + \omega_2 f_2(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) \\ &\leq \omega_1(\theta f_1(\boldsymbol{x_1}) + (1-\theta) f_1(\boldsymbol{x_2})) + \omega_2(\theta f_2(\boldsymbol{x_1}) + (1-\theta) f_2(\boldsymbol{x_2})) \\ &= \theta(\omega_1 f_1(\boldsymbol{x_1}) + \omega_2 f_2(\boldsymbol{x_1})) + (1-\theta)(\omega_1 f_1(\boldsymbol{x_2}) + \omega_2 f_2(\boldsymbol{x_2})) \\ &= \theta f(\boldsymbol{x_1}) + (1-\theta) f(\boldsymbol{x_2}), \end{aligned}$$

which completes the proof. □

Proposition 1.3.5 (*Composition with an Affine Map*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, $A \in M_{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, then g, where $g(\boldsymbol{x}) = f(A\boldsymbol{x} + \boldsymbol{b})$, is also convex.*

Proof. Take $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n$ and $\theta \in [0, 1]$ arbitrarily, then

$$\begin{aligned} g(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) &= f(A(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) + \boldsymbol{b}) \\ &= f(\theta A\boldsymbol{x_1} + \theta\boldsymbol{b} + (1-\theta)A\boldsymbol{x_2} + (1-\theta)\boldsymbol{b}) \\ &= f(\theta(A\boldsymbol{x_1} + \boldsymbol{b}) + (1-\theta)(A\boldsymbol{x_2} + \boldsymbol{b})) \\ &\leq \theta f(A\boldsymbol{x_1} + \boldsymbol{b}) + (1-\theta) f(A\boldsymbol{x_2} + \boldsymbol{b}) \\ &= \theta g(\boldsymbol{x_1}) + (1-\theta) g(\boldsymbol{x_2}), \end{aligned}$$

which completes the proof. □

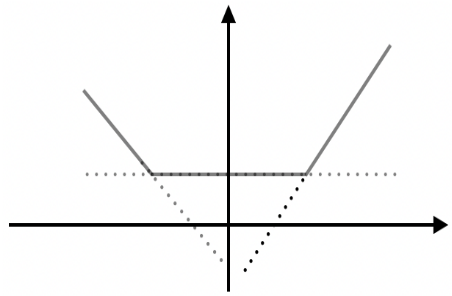
Proposition 1.3.6 (*Pointwise Maximum*)**.** *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex functions, then f, where $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$, is also convex.*

Proof. Take $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n$ and $\theta \in [0, 1]$ arbitrarily, then

$$\begin{aligned} f(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}) &= \max\{f_1(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2}), f_2(\theta\boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2})\} \\ &\leq \max\{\theta f_1(\boldsymbol{x_1}) + (1-\theta) f_1(\boldsymbol{x_2}), \theta f_2(\boldsymbol{x_1}) + (1-\theta) f_2(\boldsymbol{x_2})\} \\ &\leq \max\{\theta f_1(\boldsymbol{x_1}), \theta f_2(\boldsymbol{x_1})\} + \max\{(1-\theta) f_1(\boldsymbol{x_2}), (1-\theta) f_2(\boldsymbol{x_2})\} \\ &= \theta\max\{f_1(\boldsymbol{x_1}), f_2(\boldsymbol{x_1})\} + (1-\theta)\max\{f_1(\boldsymbol{x_2}), f_2(\boldsymbol{x_2})\} \\ &= \theta f(\boldsymbol{x_1}) + (1-\theta) f(\boldsymbol{x_2}), \end{aligned}$$

which completes the proof. □

Remark 1.3.5. The first and the third proposition can be extended to any arbitrary m functions using induction. More generally, the first one can be extended to infinite sum (integral), and the third can be extend to pointwise supremum. The graph below may help understand the convexity of the pointwise maximum functions visually:



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:

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

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.

Theorem 2.0.1 (*Local & Global Optimality*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex objective function, and C be a convex feasible set, then any locally optimal point is also globally optimal.*

Proof. To simplify, we assume f lies in $\mathbb{R}^2$. Let x^* be a local optimum, and hence we have

$$\exists\, R > 0 : f(x^*) = \min_x \{f(x) \,|\, x \in C \text{ and } |x - x^*| \le R\}.$$

We want to prove by contradiction, so we assume

$$\exists\, x_0 \in C : f(x_0) < f(x^*).$$

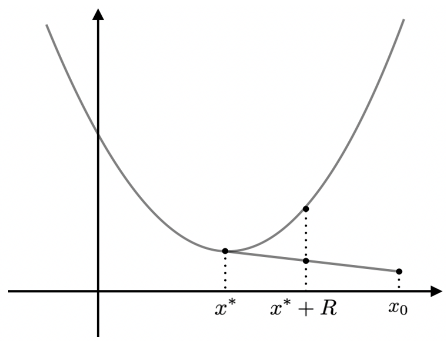
Without loss of generality, we assume x_0 is to the right of x^* and $R < |x_0 - x^*|$. Since x^* is locally optimal and C is convex, we then have

$$\exists\, \theta_0 \in (0, 1) : \theta_0 x^* + (1 - \theta_0)x_0 = x^* + R \in C, \text{ and}$$

$$f(\theta_0 x^* + (1 - \theta_0)x_0) = f(x^* + R) \ge f(x^*) > \theta_0 f(x^*) + (1 - \theta_0)f(x_0),$$

where the strict inequality is because the slope of the line segment connecting $(x^*, f(x^*))$ and $(x_0, f(x_0))$ is negative. This contradicts with the convexity of f, which completes the proof. Similar proof applies in $\mathbb{R}^{n>2}$. $\square$

Remark 2.0.6. The following graph may help visualize the proof:



Theorem 2.0.2 (*First-Order Optimality Condition*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable objective function, and C be a convex feasible set, then*

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in C} f(\boldsymbol{x}) \;\Leftrightarrow\; \forall\, \boldsymbol{x} \in C : \nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \geq 0.$$

In particular, if the problem is unconstrained, i.e. $C = \mathbb{R}^n$, then

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in C} f(\boldsymbol{x}) \;\Leftrightarrow\; \nabla f_0(\boldsymbol{x}^*) = \mathbf{0}.$$

Proof. To simplify, we assume f lies in $\mathbb{R}^2$, and then we will prove $\Rightarrow$ using contradiction. Assume x^* is a minimizer of f_0 over C but not over $\mathbb{R}$, and

$$\exists\, x_0 \in C : f^{'}(x^*)(x_0 - x^*) < 0.$$

Without loss of generality, we assume x_0 is to the right of x^*, i.e., $x_0 - x^* > 0$, so we then have $f^{'}(x^*) < 0$. Therefore, $\exists\, x^{**} \in C : x^{**} > x^*$ and $f(x^{**}) < f(x^*)$, which contradicts with the optimality of x^*. Similar proof applies in $\mathbb{R}^{n>2}$.

Conversely, assume the RHS holds, then we have

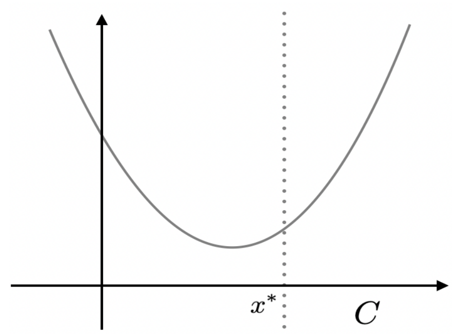
$$\forall\, \boldsymbol{x} \in C : f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \geq f(\boldsymbol{x}^*).$$

By the *First-Order Convexity Condition*, we have

$$\forall\, \boldsymbol{x} \in C : f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*).$$

To sum up, we have $\forall\, \boldsymbol{x} \in C : f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$, which completes the proof. We will skip the proof of the special case, which is useful for finding local optima. $\square$

Remark 2.0.7. The following graph may help visualize the proof:



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 $\max\limits_{\boldsymbol{\lambda} \geq \mathbf{0}} g(\boldsymbol{\lambda}, \boldsymbol{\nu})$, 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. □

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:

Definition 4.0.1 (*Strong Convexity*)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function, then we say f is *strongly convex* with parameter m if

$$\forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^n : f(\boldsymbol{x_2}) \geq f(\boldsymbol{x_1}) + \nabla f(\boldsymbol{x_1})^T(\boldsymbol{x_2} - \boldsymbol{x_1}) + \frac{m}{2}||\boldsymbol{x_2} - \boldsymbol{x_1}||_2^2.$$

Remark 4.0.1. It says that f is strongly convex if there is a quadratic function bounding below.

Definition 4.0.2 (*Lipschitz Continuity*)**.** Let $f : X \to Y$ be a function, then we say f is *L-Lipschitz continuous* if

$$\exists L \geq 0 : \forall x_1, x_2 \in X : d_Y(f(x_1), f(x_2)) \leq L d_X(x_1, x_2),$$

where d_X and d_Y denote the distance metrics on X and Y.

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

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** $\triangleright$ 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:

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

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 References