Chapter 1
Basics: Convex Optimization

Abstract This chapter provides a concise introduction to foundational concepts in convex optimization, including convex sets and functions, Fenchel conjugates, Lagrangian duality, and the Karush-Kuhn-Tucker (KKT) conditions. Definitions are accompanied by illustrative examples to build intuition and support practical understanding. While convex optimization is a rich and expansive subject that merits its own dedicated volume, our focus is intentionally selective. We present only the essential tools and results that the author considers most relevant for understanding and analyzing optimization problems encountered in later chapters. The goal is to equip readers with a practical yet rigorous foundation, enabling them to appreciate the theoretical underpinnings of algorithm design and analysis in subsequent chapters.

Convex Optimization is the foundation of foundations!

Contents

1.1 Notations and Definitions

This book uses the following notations.

- Let us denote by $\|\cdot\|_2$ the Euclidean norm, and by $\|\cdot\|$ a general norm.
- For a differentiable function f, let $\nabla f(\mathbf{x})$ denote its gradient at $\mathbf{x}$, and $\partial f(\mathbf{x})$ denote its subdifferential set at $\mathbf{x}$.
- Let $\partial_1 f(\mathbf{w}, \mathbf{u})$ and $\partial_2 f(\mathbf{w}, \mathbf{u})$ denote the partial subgradients of f with respect to the first variable $\mathbf{w}$ and the second variable $\mathbf{u}$, respectively.
- Define the d-dimensional probability simplex as

$$\Delta_d = \left\{ \mathbf{x} \in \mathbb{R}^d : x_i \geq 0 \ \forall i, \ \sum_{i=1}^{d} x_i = 1 \right\}.$$

- Let $\mathbb{I}(\cdot)$ denote the standard indicator function, which returns 1 if the input condition is true and 0 otherwise. Let $\mathbb{I}_{0-\infty}(\cdot)$ denote the zero-infinity indicator function, which returns 0 if the input condition is true and ∞ otherwise.
- Denote by $\mathbf{1}$ a vector of all ones. Let $\mathbf{e}_i$ denote the standard basis vector with a 1 in the i-th coordinate and 0 in all other entries.
- Let $\mathbf{x} \sim \mathbb{P}$ denote a random variable that follows a distribution $\mathbb{P}$.
- $[n]$ denotes the set of all integers from 1 to n, i.e., $[n] = \{1, \ldots, n\}$.
- We use $\langle \mathbf{x}, \mathbf{y} \rangle$ interchangeable with $\mathbf{x}^\top \mathbf{y}$ to denote the inner product of two vectors.
- $\log(x)$ is in the base of natural constant e.
- w.r.t is short for with respect to.
- s.t. is short for subject to.

Definition 1.1 (Dual Norm) Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, then its dual norm $\|\cdot\|_* : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\|\mathbf{y}\|_* := \sup\{\mathbf{x}^\top \mathbf{y} : \|\mathbf{x}\| \leq 1\}.$$

Examples

Example 1.1. *$\|\cdot\|_2$ is the dual norm of itself as $\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$.*

Example 1.2. *$\|\cdot\|_\infty$ and $\|\cdot\|_1$ are dual norms of each other as $\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty$.*

Example 1.3. *Let $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A \mathbf{x}}$, where $A \succ 0$ is a positive definite matrix. Then $\|\mathbf{y}\|_* = \sqrt{\mathbf{y}^\top A^{-1} \mathbf{y}}$. This is because that $\mathbf{x}^\top \mathbf{y} = \mathbf{x}^\top A^{1/2} A^{-1/2} \mathbf{y} \leq \|A^{1/2}\mathbf{x}\|_2 \|A^{-1/2}\mathbf{y}\|_2 \leq \|A^{-1/2}\mathbf{y}\|_2$.*

Definition 1.2 (Convex set) A set C is convex if the line segment between any two points in C lies in C, i.e. $\forall \mathbf{x}_1, \mathbf{x}_2 \in C, \forall \theta \in [0, 1]$,

$$\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \in C.$$

Definition 1.3 (Convex function) A function $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if its domain $\mathrm{dom}(f)$ is convex and

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}), \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f), \theta \in [0,1].$$

It is strictly convex if strict inequality holds whenever $\mathbf{x} \neq \mathbf{y}$ and $\theta \in (0,1)$.

This inequality implies that the graph of a convex function lies below the straight line connecting any two points on the graph—like a bowl: if you place a chopstick across its edges, it will stay above the surface of the bowl.

Lemma 1.1 (First-order condition) *Suppose f is differentiable (i.e., its gradient ∇f exists at each point in* dom f*). Then f is convex if and only if* dom f *is convex and*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \tag{1.1}$$

holds for all $\mathbf{x}, \mathbf{y} \in$ dom f.

Proof. We first prove for one-dimensional convex function $\phi(\cdot) : \mathbb{R} \to \mathbb{R}$, we have

$$\phi(t) \geq \phi(s) + \phi'(s)(t-s). \tag{1.2}$$

According to the definition of convexity, we have

$$\phi(t) \geq \phi(s) + \frac{\phi(s + \alpha(t-s)) - \phi(s)}{\alpha}.$$

Taking the limit $\alpha \to 0$ yields (1.2).

($\Rightarrow$) Assume f is convex and differentiable on the open convex set dom f. Fix $\mathbf{x} \in$ dom f and any $\mathbf{y} \in$ dom f. Define $\phi : [0,1] \to \mathbb{R}$ by

$$\phi(t) = f\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big).$$

Since f is convex and the map $t \mapsto \mathbf{x} + t(\mathbf{y} - \mathbf{x})$ is affine, ϕ is a convex function on $[0,1]$. For a convex (one-dimensional) differentiable function, we have proved that

$$\phi(1) \geq \phi(0) + \phi'(0)(1-0).$$

By the chain rule,

$$\phi'(0) = \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Thus

$$f(\mathbf{y}) = \phi(1) \geq \phi(0) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

($\Leftarrow$) Assume dom f is convex and for all $\mathbf{x}, \mathbf{y} \in$ dom f,

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

Take any $\mathbf{x}, \mathbf{y} \in$ dom f and $\theta \in [0,1]$, and set $\mathbf{z} = \theta \mathbf{x} + (1-\theta)\mathbf{y} \in$ dom f. Apply the assumption with $(\mathbf{x}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$:

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}), \qquad f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}).$$

Multiply the first by θ and the second by $(1-\theta)$ and add:

$$\theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top \big(\theta(\mathbf{x} - \mathbf{z}) + (1-\theta)(\mathbf{y} - \mathbf{z})\big).$$

Since $\theta(\mathbf{x} - \mathbf{z}) + (1-\theta)(\mathbf{y} - \mathbf{z}) = 0$, we get

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

i.e., $f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$. Hence f is convex. □

Definition 1.4 (Subgradient) For a non-differentiable convex function f, let the subgradient of f at $\mathbf{x}$ be denoted by $\partial f(\mathbf{x})$, which consists of all vectors $\mathbf{v}$ satisfying:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{v}^\top (\mathbf{y} - \mathbf{x}), \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$$

Without causing any confusion, we often write

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \partial f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f),$$

where $\partial f(\mathbf{x})$ refers to some specific element of the subgradient set.

Examples

Example 1.4. $f(x) = [x]_+ = \max(0, x)$. *At* $x = 0$ *it has a subgradient* $\partial f(0) = \{\xi \in [0,1]\}$, $\partial f(x) = 1, \forall x > 0$, *and* $\partial f(x) = 0, \forall x < 0$.

Definition 1.5 (Strongly Convex Function) A function $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ is called μ-strongly convex with respect to a norm $\|\cdot\|$ if there exists a constant $\mu > 0$ such that for any $\mathbf{x}, \mathbf{y}$ and $\mathbf{v} \in \partial f(\mathbf{x})$ we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{v}^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Examples

Example 1.5. *The function* $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ *is 1-strongly convex with respect to the Euclidean norm* $\|\cdot\|_2$. *This follows directly from the identity:*

$$\frac{1}{2}\|\mathbf{y}\|_2^2 = \frac{1}{2}\|\mathbf{x}\|_2^2 + \mathbf{x}^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2,$$

which satisfies the definition of strong convexity with parameter 1.

Definition 1.6 (Smooth function) A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is called L-smooth with respect to a norm $\|\cdot\|$ if it is differentiable and its gradient is L-Lipchitz continuous, i.e., there exists a positive real constant L such that, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$, or equivalently,

$$|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{1.3}$$

Definition 1.7 (Bregman Divergence) Let $\varphi : \Omega \to \mathbb{R}$ be be a continuously-differentiable, strictly convex function defined on a convex set Ω, the Bregman divergence induced by $\varphi(\cdot)$ is defined as

$$D_\varphi(\mathbf{x}, \mathbf{y}) := \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - \nabla\varphi(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}).$$

Examples:

Example 1.6 (Euclidean distance). $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ *induces the Euclidean distance:*

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 - \mathbf{y}^\top(\mathbf{x} - \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2. \tag{1.4}$$

Example 1.7 (Kullback–Leibler (KL) divergence). $\varphi(\mathbf{x}) = \sum_{i=1}^d x_i \log x_i$ *for* $\mathbf{x} \in \Delta_d$ *induces the Kullback–Leibler (KL) divergence:*

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d y_i \log y_i - \sum_{i=1}^d (\log y_i + 1)(x_i - y_i) = \sum_{i=1}^d x_i \log \frac{x_i}{y_i}. \tag{1.5}$$

Example 1.8 (Itakura–Saito distance). $\varphi(\mathbf{x}) = -\sum_{i=1}^d \log x_i$ *for* $\mathbf{x} > 0$ *induces the Itakura–Saito distance:*

$$D_\varphi(\mathbf{x}, \mathbf{y}) = -\sum_{i=1}^d \log x_i + \sum_{i=1}^d \log y_i + \sum_{i=1}^d \frac{1}{y_i}(x_i - y_i) = \sum_{i=1}^d \frac{x_i}{y_i} - \sum_{i=1}^d \log \frac{x_i}{y_i} - 1. \tag{1.6}$$

1.2 Verification of Convexity

In practice, directly applying the definition of convexity or verifying the first-order condition of convexity can be challenging when proving that a function is convex. The following rules offer practical tools to simplify the verification process.

Second-order Condition for Twice Differentiable Functions

If a function $f(\mathbf{x})$ is twice differentiable, then it is convex if and only if

$$\nabla^2 f(\mathbf{x}) \succeq 0, \quad \forall \mathbf{x},$$

i.e., its Hessian is positive semidefinite everywhere.

Examples

We can use the above rule to verify the convexity of the following functions.

Example 1.9 (Log-Sum-Exp Function).

$$\ell(\mathbf{y}) = \log\left(\sum_{i=1}^{K} \exp(y_i)\right), \quad \mathbf{y} \in \mathbb{R}^K.$$

Its Hessian matrix is given by

$$H = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^T,$$

where $\mathbf{p}$ *is the vector of softmax probabilities with components* $p_i = \frac{\exp(y_i)}{\sum_{k=1}^{K} \exp(y_k)}$. *It is positive semidefinite as* $\mathbf{v}^\top H\mathbf{v} = \sum_{i=1}^{K} p_i v_i^2 - (\sum_{i=1}^{K} p_i v_i)^2 \geq 0$ *due to Cauchy-Schwarz inequality.*

Example 1.10 (Negative entropy).

$$\varphi(\mathbf{p}) = \sum_{i=1}^{n} p_i \log p_i$$

where $\mathbf{p} \in \Delta_n = \{\mathbf{q} : \sum_{i=1}^{n} q_i = 1, q_i \geq 0, \forall i\}$ *is a probability vector. Its Hessian matrix is*

$$H = diag(1/\mathbf{p})$$

is positive definite.

Operations that Preserve Convexity

The following operations preserve convexity:

- **Affine Composition:** If f is convex, then $f(A\mathbf{x} + \mathbf{b})$ is convex for any matrix A and vector $\mathbf{b}$.
- **Non-Negative Weighted Sums:** If f_i is convex for all i, and $\alpha_i \geq 0$, then

  $$f(\mathbf{x}) = \sum_i \alpha_i f_i(\mathbf{x})$$

  is convex.
- **Pointwise Maximum:** If $g(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for all $\mathbf{y}$, then

  $$f(\mathbf{x}) = \max_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})$$

is convex.

- **Function Composition:** The composition $h(\mathbf{x}) = f(g(\mathbf{x}))$ is convex if one of the following holds:
  - f is convex and non-decreasing, and $g(\mathbf{x})$ is convex.
  - f is convex and non-increasing, and $g(\mathbf{x})$ is concave.

  To quickly verify this, we compute the Hessian matrix assuming that both f and g are twice-differentiable:

$$\nabla^2 h(\mathbf{x}) = f'(g(\mathbf{x}))\nabla^2 g(\mathbf{x}) + f''(g(\mathbf{x}))\nabla g(\mathbf{x})\nabla g(\mathbf{x})^\top,$$

  which is positive semi-definite under either of the above two conditions.

1.3 Fenchel Conjugate

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Its **Fenchel conjugate** (also called the convex conjugate) is defined as:

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathrm{dom}(f)} \left\{ \mathbf{x}^\top \mathbf{y} - f(\mathbf{x}) \right\},$$

where the domain of the conjugate function consists of $\mathbf{y} \in \mathbb{R}^d$ for which the supremum is finite. From the definition of conjugate function, we immediately obtain the inequality

$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{x}^\top \mathbf{y}, \forall \mathbf{x}, \mathbf{y}.$$

This is called Fenchel's inequality. If f is proper, convex, and lower semicontinuous, then the conjugate of the conjugate of a convex function is the original function, i.e., $(f^*)^* = f$.

Definition 1.8 (Legendre function) Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous, convex function with $\mathrm{int}(\mathrm{dom}\, f) \neq \emptyset$. The function f is called a *Legendre function* if it satisfies:

(i) f is differentiable on $\mathrm{int}(\mathrm{dom}\, f)$, and for any sequence $\{\mathbf{x}_k\} \subset \mathrm{int}(\mathrm{dom}\, f)$ with $\mathbf{x}_k$ converging to a boundary point of $\mathrm{dom}\, f$, we have $\|\nabla f(\mathbf{x}_k)\| \to \infty$.

(ii) f is strictly convex on every convex subset of $\mathrm{dom}(\partial f)$.

If f is Legendre function, its Fenchel conjugate reduces to the Legendre transform, defined by

$$f^*(\mathbf{y}) = \mathbf{x}(\mathbf{y})^\top \mathbf{y} - f(\mathbf{x}(\mathbf{y})),$$

where $\mathbf{x}(\mathbf{y}) = \arg\min_{\mathbf{x}} (\mathbf{x}^\top \mathbf{y} - f(\mathbf{x}))$ is the unique solution to the first-order optimality condition $\nabla f(\mathbf{x}) = \mathbf{y}$.

Examples

Example 1.11 (**Conjugate of the Quadratic Function.**). *Let* $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. *Then:*

$$f^*(\mathbf{y}) = \sup_{\mathbf{x}} \left\{ \mathbf{x}^\top \mathbf{y} - \frac{1}{2}\|\mathbf{x}\|_2^2 \right\} = \frac{1}{2}\|\mathbf{y}\|_2^2.$$

Example 1.12 (**Conjugate of the Squared Hinge.**). *Let* $f(x) = \max(x, 0)^2$. *Then:*

$$f^*(y) = \sup_x xy - \max(x, 0)^2 = \begin{cases} \frac{y^2}{4}, & y \geq 0 \\ \infty, & y < 0 \end{cases}.$$

The Legendre transform is not defined in this case since f *is not strictly convex.*

Example 1.13. *Log-sum-exp and negative entropy are conjugates of each other. Please refer to the Example 1.16.*

1.4 Convex Optimization

A standard optimization problem is defined by:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & f_0(\mathbf{x}) \\ s.t. \quad & f_i(\mathbf{x}) \leq 0, i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \ldots, n. \end{aligned} \tag{1.7}$$

Definition 1.9 A standard optimization problem (1.7) is a convex optimization problem if $f_i(\mathbf{x})$ is convex for $i = 0, \ldots, m$ and $h_j(\mathbf{x}) = \mathbf{a}_j^\top \mathbf{x} + b_j$ is an affine function for $j = 1, \ldots, n$.

The problem (1.7) is feasible if there exists at least one point such that all constraints are satisfied, and infeasible otherwise. The set of all feasible points is called the feasible set, denoted by

$$\mathcal{X} = \{\mathbf{x} : f_i(\mathbf{x}) \leq 0, i = 1, \ldots, m, h_i(\mathbf{x}) = 0, j = 1, \ldots, n\}.$$

The Optimal value and optimal solutions

The optimal value of (1.7) is defined as

$$f_* = \inf\{f_0(\mathbf{x}) | \mathbf{x} \in \mathcal{X}\}.$$

where inf returns the greatest value that is less than or equal to all possible objective values at feasible points if such a value exists. For example inf $e^{-x} = 0$. If the problem is infeasible, we let $f_* = \infty$.

A solution $\mathbf{x}_*$ is an optimal solution if it is feasible, i.e., satisfying all constraints, and $f_0(\mathbf{x}_*) = f_*$. Hence, we may have a set of optimal solutions:

$$\mathcal{X}_* = \arg\min\{f_0(\mathbf{x})|\mathbf{x} \in \mathcal{X}\} = \{\mathbf{x} : \mathbf{x} \in \mathcal{X}, f_0(\mathbf{x}) = f_*\}.$$

The optimal solution is unique if the objective is strongly convex.

1.4.1 Local Minima and Global Minima

A solution $\mathbf{x}$ is called a local minima if there is an $R > 0$ such that

$$f_0(\mathbf{x}) = \inf\{f_0(\mathbf{y})|\mathbf{y} \in \mathcal{X}, \|\mathbf{y} - \mathbf{x}\|_2 \leq R\}. \tag{1.8}$$

Theorem 1.1 *For a convex optimization problem, a local minima* $\mathbf{x}$ *is also a global minima.*

Proof. Suppose $\mathbf{x}$ is not a global minima. It means that there exists a feasible $\mathbf{z}$ such that $f_0(\mathbf{z}) < f_0(\mathbf{x})$. Then $\|\mathbf{z} - \mathbf{x}\|_2 > R$ because $\mathbf{x}$ is an optimal solution in the local region $\Omega = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 \leq R\}$.

Let us derive a contradiction. Let $\mathbf{y} = \mathbf{x} + \theta(\mathbf{z} - \mathbf{x})$, where $\theta = \frac{R}{\|\mathbf{x}-\mathbf{z}\|_2}$ such that $\|\mathbf{y} - \mathbf{x}\|_2 \leq \theta\|\mathbf{z} - \mathbf{x}\|_2 \leq R$. Then $f_0(\mathbf{y}) \leq \theta f_0(\mathbf{z}) + (1 - \theta) f_0(\mathbf{x}) < f_0(\mathbf{x})$, which contradicts to the fact that $\mathbf{x}$ is an optimal solution in the region $\Omega = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 \leq R\}$. Hence such an $\mathbf{z}$ does not exist. □

1.4.2 Optimality Conditions

Let us consider a differential objective function f_0.

Theorem 1.2 *For a convex optimization problem (1.7) with non-empty* $\mathcal{X}_*$, $\mathbf{x}$ *is optimal if and only if* $\mathbf{x} \in \mathcal{X}$ *and*

$$\nabla f_0(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geq 0, \forall \mathbf{y} \in \mathcal{X}. \tag{1.9}$$

For non-differential function, the above condition is replaced by $\exists \mathbf{v} \in \partial f_0(\mathbf{x})$ such that $\mathbf{v}^\top (\mathbf{y} - \mathbf{x}) \geq 0, \forall \mathbf{y} \in \mathcal{X}$.

Proof. To prove the sufficient condition, we use the convexity of f_0. For any $\mathbf{y} \in \mathcal{X}$, we have

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \nabla f_0(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geq f_0(\mathbf{x}).$$

Hence $\mathbf{x}$ is an optimal solution. Let us prove the necessary condition. If (1.9) does not hold for an $\mathbf{y}$, i.e., $\nabla f_0(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) < 0$, let us consider $\mathbf{z}(t)) = t\mathbf{y} + (1-t)\mathbf{x}$, which is feasible. Thence $\nabla_t f_0(\mathbf{z}(t))|_{t=0} = \nabla f_0(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) < 0$, which means there exists a small $t > 0$ such that $f_0(\mathbf{z}(t)) \leq f_0(\mathbf{z}(0)) = f_0(\mathbf{x})$, which is impossible as $\mathbf{x}$ is an optimal solution. □

When the problem is unconstrained such that $\mathcal{X} = \mathbb{R}^d$, then the optimality condition (1.9) implies that $\mathbf{x}$ is optimal if and only if $\nabla f_0(\mathbf{x}) = 0$.

Lemma 1.2 *For a convex optimization problem (1.7), if f_0 is strongly convex, then $\mathcal{X}_*$ contains only a single element if it is not empty.*

Proof. Assume $\mathcal{X}_*$ contains two different solutions $\mathbf{x}_1 \neq \mathbf{x}_2$ such that $f_0(\mathbf{x}_1) = f_0(\mathbf{x}_2)$. We will derive a contradiction. Since f_0 is strongly convex, we have

$$f_0(\mathbf{x}_1) \geq f_0(\mathbf{x}_2) + \partial f_0(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2) + \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2.$$

Due to the optimality condition, $\partial f_0(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2) \geq 0$, hence $f_0(\mathbf{x}_1) \geq f_0(\mathbf{x}_2) + \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 > f_0(\mathbf{x}_2)$, which contradicts to the fact $f_0(\mathbf{x}_1) = f_0(\mathbf{x}_2)$.

□

1.4.3 Karush–Kuhn–Tucker (KKT) Conditions

Constrained optimization problems such as (1.7) are often challenging to analyze and solve directly. The Karush-Kuhn-Tucker (KKT) conditions, derived from Lagrangian duality theory, offer first-order necessary conditions for optimality. These conditions can simplify the original problem, sometimes enabling a transformation into a more tractable form or even leading to a closed-form solution.

The Lagrangian function and the Lagrangian dual function

For the constrained optimization (1.7), the Lagrangian function is defined as:

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

where $\lambda_1, \ldots, \lambda_m, \nu_1, \ldots, \nu_n$ are called the Lagrangian multipliers.

The Lagrangian dual function is defined as:

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

Based on this, we define the Lagrangian dual problem:

$$g_* = \sup_{\lambda \geq 0} g(\lambda, \nu).$$

Regarding the original optimal value f_* and the dual optimal value g_*, we have the following weak duality.

Lemma 1.3 *We always have $g_* \leq f_*$.*

Proof. Let $\mathbf{x}_*$ be an optimal solution to (1.7). For any $\lambda \geq 0, \nu$, we have

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

where the last inequality uses the fact $h_j(\mathbf{x}_*) = 0$, $f_i(\mathbf{x}_*) \leq 0$, and $\lambda \geq 0$. The conclusion follows. □

KKT conditions

An interesting scenario is the strong duality where $g_* = f_*$. In such case, we can derive two conditions.

Lemma 1.4 *Suppose that the primal and dual optimal values are attained and equal. Let $\mathbf{x}_*$ be an optimal primal solution and λ_*, ν_* be optimal dual solutions. Assume that f, g_i, h_j are continuously differentiable, then the following conditions hold:*

$$\nabla f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} \nabla f_i(\mathbf{x}_*) + \sum_{j=1}^{n} \nu_{*,j} \nabla h_j(\mathbf{x}_*) = 0, \tag{1.10}$$

$$\lambda_{*,i} f_i(\mathbf{x}_*) = 0, i = 1, \ldots, m, \tag{1.11}$$

where the second condition is called the complementary slackness.

Proof. First, we have

$$\begin{aligned} g_* &= \sup_{\lambda \geq 0} g(\lambda, \nu) = g(\lambda_*, \nu_*) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda_*, \nu_*) \\ &= \inf_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_{*,j} h_j(\mathbf{x}) \\ &\leq f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}_*) + \sum_{j=1}^{n} \nu_{*,j} h_j(\mathbf{x}_*) \\ &\leq f_0(\mathbf{x}_*) = f_*. \end{aligned}$$

Since $g_* = f_*$, the inequalities will become equalities. The first equality is

$$\inf_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_{*,j} h_j(\mathbf{x}) = f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}_*) + \sum_{j=1}^{n} \nu_{*,j} h_j(\mathbf{x}_*),$$

which implies that $\mathbf{x}_*$ optimizes $L(\mathbf{x}, \lambda_*, \nu_*)$. Hence, by the first-order optimality condition, we have $\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda_*, \nu_*) = 0$, which is (1.10). The second equality is

$$f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}_*) + \sum_{j=1}^{n} \nu_{*,j} h_j(\mathbf{x}_*) = f_0(\mathbf{x}_*),$$

which implies $\lambda_{*,i} f_i(\mathbf{x}_*) = 0, \forall, i$ because $\lambda_{*,i} f_i(\mathbf{x}_*) \leq 0, \forall, i$ and they cannot be larger than zero; otherwise the equality will not hold. □

KKT conditions

Assume that f, g_i, h_j are continuously differentiable. Let $\mathbf{x}_*$ be an optimal primal solution and λ_*, ν_* be optimal dual solutions. The KKT conditions are:

$$\text{(Stationarity)} \quad \nabla f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} \nabla f_i(\mathbf{x}_*) + \sum_{j=1}^{n} \nu_{*,j} \nabla h_j(\mathbf{x}_*) = 0,$$

$$\text{(Primal feasibility)} \quad f_i(\mathbf{x}_*) \leq 0, \quad h_j(\mathbf{x}_*) = 0, \forall i, j,$$

$$\text{(Dual feasibility)} \quad \lambda_{*,i} \geq 0, \forall i,$$

$$\text{(Complementary slackness)} \quad \lambda_{*,i} f_i(\mathbf{x}_*) = 0, \forall i.$$

Slater's condition

How to ensure the strong duality holds? Constraint qualifications have been developed as sufficient conditions of strong duality. One simple constraint qualification is Slater's condition for a convex optimization problem: There exists an $\mathbf{x} \in \text{relint}(D)$ (where relint denotes the relative interior of the convex set $D := \cap_{i=0}^{m} \text{dom}(f_i)$ such that

$$f_i(\mathbf{x}) < 0, \forall i, \quad \text{and} \quad \mathbf{a}_j^\top \mathbf{x} + b_j = 0, \forall j.$$

An important theorem of Lagrangian duality is that the strong duality holds when **the primal problem is convex** and Slater's condition holds. This suggests a tangible approach to compute $\mathbf{x}_*$ or transform the original problem into a simplified one. First, we solve the dual problem to obtain an optimal dual solution (λ_*, ν_*):

$$(\lambda_*, \nu_*) = \arg\max_{\lambda \geq 0, \nu} g(\lambda, \nu). \tag{1.12}$$

Then we use the stationarity condition of KKT conditions to derive a close form of $\mathbf{x}_*$. In addition, we have

$$\min_{\mathbf{x}}\{f_0(\mathbf{x}),\ s.t.\ \mathbf{x}\in\mathcal{X}\} = \max_{\lambda\geq 0,\nu} g(\lambda,\nu).$$

Examples

Example 1.14 (Dual of Distributionally Robust optimization (DRO)). *The following problem often arises in robust machine learning:*

$$f(\ell_1,\ldots,\ell_n) = \max_{\mathbf{p}\in\Delta_n} \sum_{i=1}^n p_i\ell_i - \tau\sum_{i=1}^n q_i\phi(p_i/q_i),$$

where $\tau\geq 0, \mathbf{q}\in\Delta_n$ and $\phi(t):\mathbb{R}^+\to\mathbb{R}$ is a proper closed convex function and has a minimum value zero that is attained at $t=1$. Let us derive its dual problem. We write the above problem as a standard convex optimization problem:

$$\begin{aligned}&\min_{\mathbf{p}} -\sum_{i=1}^n p_i\ell_i + \tau\sum_i q_i\phi(p_i/q_i)\\ &s.t.\ \sum_{i=1}^n p_i = 1.\end{aligned}$$

where the constraint $p_i\geq 0$ is enforced by the domain of $\phi(t)$.
We define the Lagrangian function:

$$L(\mathbf{p},\nu) = -\sum_{i=1}^n p_i\ell_i + \tau\sum_{i=1}^n q_i\phi(p_i/q_i) + \nu\left(\sum_{i=1}^n p_i - 1\right).$$

Let us define

$$\phi^*(s) = \max_{t\geq 0} ts - \phi(t). \tag{1.13}$$

By minimizing over $\mathbf{p}\geq 0$, we have

$$\begin{aligned}g(\nu) &= \min_{\mathbf{p}\geq 0} -\sum_{i=1}^n p_i(\ell_i-\nu) + \tau\sum_{i=1}^n q_i\phi(p_i/q_i) - \nu\\ &= -\{\max_{\mathbf{p}\geq 0}\sum_{i=1}^n p_i(\ell_i-\nu) - \tau\sum_{i=1}^n q_i\phi(p_i/q_i)\} - \nu.\end{aligned}$$

With a variable change $\tilde{p} = p/q$, we have

$$g(\nu) = -\max_{\tilde{p}\geq 0} \sum_{i=1}^{n} q_i\{\tilde{p}_i(\ell_i - \nu) - \tau\phi(\tilde{p}_i)\} - \nu$$
$$= -\sum_{i=1}^{n} q_i\{\max_{\tilde{p}_i\geq 0} \tilde{p}_i(\ell_i - \nu) - \tau\phi(\tilde{p}_i)\} - \nu = -\sum_{i=1}^{n} \tau q_i \phi^*\left(\frac{\ell_i - \nu}{\tau}\right) - \nu.$$

Since the Slater's condition holds ($p_i = 1/n$ satisfies), we have

$$\min_{\mathbf{p}\in\Delta} -\sum_{i=1}^{n} p_i\ell_i + \tau\sum_{i=1}^{n} q_i\phi(p_i/q_i)$$
$$= \max_{\nu} g(\nu) = -\left\{\min_{\nu} \sum_{i=1}^{n} \tau q_i \phi^*\left(\frac{\ell_i - \nu}{\tau}\right) + \nu\right\}.$$

Hence,

$$\max_{\mathbf{p}\in\Delta} \sum_{i=1}^{n} p_i\ell_i - \tau\sum_{i=1}^{n} q_i\phi(p_i/q_i) = \min_{\nu} \sum_{i=1}^{n} \tau q_i \phi^*\left(\frac{\ell_i - \nu}{\tau}\right) + \nu. \quad (1.14)$$

Example 1.15 (**Conjugate of ϕ functions.**). *We can derive ϕ^* for three cases below (exercise):*

- $\phi(t) = (t-1)^2$*:*

$$\phi^*(y) = \max_{t\geq 0} yt - (t-1)^2 = \begin{cases} \frac{1}{4}y^2 + y & \text{if } y \geq -2 \\ -1 & \text{o.w.} \end{cases}$$

- $\phi(t) = t\log t - t + 1$ *and*

$$\phi^*(y) = \max_{t\geq 0} yt - (t\log t - t + 1) = \exp(y) - 1.$$

- $\phi(t) = \mathbb{I}_{0-\infty}(t \leq 1/\alpha)$ *for* $\alpha \in (0, 1]$*:*

$$\phi^*(y) = \max_{t\geq 0} yt - \mathbb{I}_{0-\infty}(t \leq 1/\alpha) = \frac{[y]_+}{\alpha}.$$

Example 1.16 (**KKT conditions of DRO with a KL divergence**). *Let us consider a special case of Example 1.14 with $\phi(t) = t\log t - t + 1$:*

$$f(\ell_1, \ldots, \ell_n) = \max_{\mathbf{p}\in\Delta_n} \sum_{i=1}^{n} p_i\ell_i - \tau\sum_{i=1}^{n} p_i \log\frac{p_i}{q_i}. \quad (1.15)$$

We can derive the following KKT conditions:

$$(\ell_i - \nu_*) - \tau(\log \frac{p_i^*}{q_i} + 1) = 0, \forall i \Rightarrow p_i^* = q_i \exp\left(\frac{\ell_i - \nu_* - \tau}{\tau}\right),$$

$$\sum_{i=1}^{n} p_i^* = 1.$$

As a result, we can derive

$$p_i^* = \frac{q_i \exp(\frac{\ell_i}{\tau})}{\sum_{i=1}^n q_i \exp(\frac{\ell_i}{\tau})} \tag{1.16}$$

$$f(\ell_1, \ldots, \ell_n) = \tau \log\left(\sum_{i=1}^{n} q_i \exp\left(\frac{\ell_i}{\tau}\right)\right). \tag{1.17}$$

1.5 Basic Lemmas

Below, we present some basic lemmas that are useful for the presentation and analysis in later chapters.

Lemma 1.5 *For a L-smooth convex function w.r.t. $\|\cdot\|_2$, the following conditions are equivalent:*

(a) $0 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$;
(b) $\frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$;
(c) $\frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x} - \mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\|_2^2$;
(d) $\frac{\alpha(1-\alpha)}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) - f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha(1-\alpha)\frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$.

Proof. Let us prove (a). Since $\frac{df(\mathbf{x}+\gamma\mathbf{p})}{d\gamma} = \nabla f(\mathbf{x}+\gamma\mathbf{p})^\top\mathbf{p}$, according to *Taylor Theory*

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + \int_0^1 \nabla f(\mathbf{x} + \gamma\mathbf{p})^\top\mathbf{p} d\gamma$$

Let $\mathbf{y} = \mathbf{x} + \mathbf{p}$:

$$
\begin{aligned}
&f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
&\quad = \int_0^1 \nabla f(\mathbf{x} + \gamma(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) d\gamma - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
&\quad = \int_0^1 \nabla f(\mathbf{x} + \gamma(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) d\gamma \\
&\quad \leq \int_0^1 \|\nabla f(\mathbf{x} + \gamma(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_2 \|\mathbf{p}\|_2 d\gamma \\
&\quad \leq \int_0^1 L\|\gamma \mathbf{p}\|_2 \|\mathbf{p}\|_2 d\gamma = \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.
\end{aligned}
$$

Let us prove (b). Define $\phi(\mathbf{z}) = f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \mathbf{z}$. We can conclude that $\mathbf{z}^* = \mathbf{x}$ (by the first-order optimality) and that $\phi(\mathbf{z})$ is also convex & L-smooth if f is convex & L-smooth.

$$
\begin{aligned}
\phi(\mathbf{x}) = \min_{\mathbf{z}} \phi(\mathbf{z}) &\leq \min_{\mathbf{z}} \left\{ \phi(\mathbf{y}) + \nabla \phi(\mathbf{y})^\top (\mathbf{z} - \mathbf{y}) + \frac{L}{2}\|\mathbf{z} - \mathbf{y}\|_2^2 \right\} \\
&\overset{r = \mathbf{z} - \mathbf{y}}{=} \min_r \left\{ \phi(\mathbf{y}) + \nabla \phi(\mathbf{y})^\top r + \frac{L}{2}\|r\|_2^2 \right\} \\
&\overset{\text{solve } r}{=} \phi(\mathbf{y}) - \frac{\|\phi(\mathbf{y})\|_2^2}{L} + \frac{\|\nabla \phi(\mathbf{y})\|_2^2}{2L} = \phi(\mathbf{y}) - \frac{\|\nabla \phi(\mathbf{y})\|_2^2}{2L}.
\end{aligned}
$$

Then, we have $2L(\phi(\mathbf{y}) - \phi(\mathbf{x})) \geq \|\nabla \phi(\mathbf{y})\|_2^2$, which prove the result by plugging in $\phi(\mathbf{z}) = f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \mathbf{z}$ and $\nabla \phi(\mathbf{z}) = \nabla f(\mathbf{z}) - \nabla f(\mathbf{x})$.

Let us prove (c). According to part (b) we have

$$
\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).
$$

Similarly,

$$
\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}).
$$

Summing up the above two inequalities leads to

$$
\frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}).
$$

Let us prove (d). Let $\mathbf{x}_\alpha = \alpha \mathbf{x} + (1 - \alpha)\mathbf{y}$. From (a) and (b), we have

$$\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_\alpha)\|_2^2 \le f(\mathbf{x}) - (f(\mathbf{x}_\alpha) + \nabla f(\mathbf{x}_\alpha)^\top (1-\alpha)(\mathbf{x}-\mathbf{y}))$$
$$\le \frac{L}{2}\|(1-\alpha)(\mathbf{x}-\mathbf{y})\|_2^2,$$
$$\frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}_\alpha)\|_2^2 \le f(\mathbf{y}) - (f(\mathbf{x}_\alpha) + \nabla f(\mathbf{x}_\alpha)^\top \alpha(\mathbf{y}-\mathbf{x}))$$
$$\le \frac{L}{2}\|\alpha(\mathbf{y}-\mathbf{x})\|_2^2.$$

Multiplying the first by α and the second by $1-\alpha$, we can prove part (d), where the lower bound is as

$$\frac{\alpha}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_\alpha)\|_2^2 + \frac{1-\alpha}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}_\alpha)\|_2^2 \ge \frac{\alpha(1-\alpha)}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$$

by applying the Young's inequality $\|\mathbf{a}-\mathbf{b}\|_2^2 \le (1+\beta)\|\mathbf{a}-\mathbf{c}\|_2^2 + (1+\frac{1}{\beta})\|\mathbf{b}-\mathbf{c}\|_2^2$ with $\beta = \alpha/(1-\alpha)$. □

Lemma 1.6 *If f is differentiable and μ-strongly convex w.r.t $\|\cdot\|_2$, the following conditions are equivalent:*

(a) $f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) \ge \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$;
(b) $f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) \le \frac{1}{2\mu}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2^2$;
(c) $\mu\|\mathbf{x}-\mathbf{y}\|_2^2 \le (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \le \frac{1}{\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$;
(d) $\frac{\alpha(1-\alpha)\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2 \le \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) - f(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}) \le \alpha(1-\alpha)\frac{1}{2\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$.

From (a) we can derive an useful inequality for strongly convex optimization $\mathbf{x}_* = \arg\min_{\mathbf{x}} f(\mathbf{x})$, i.e., for any $\mathbf{x}$, we have

$$f(\mathbf{x}) \ge f(\mathbf{x}_*) + \nabla f(\mathbf{x}_*)^\top(\mathbf{x}-\mathbf{x}_*) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{x}_*\|_2^2 \ge f(\mathbf{x}_*) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{x}_*\|_2^2. \quad (1.18)$$

Proof of (b). Define $\phi(\mathbf{z}) = f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \mathbf{z}$. We can conclude that $\mathbf{z}^* = \mathbf{x}$ (by the first-order optimality) and that $\phi(\mathbf{z})$ is also convex & μ-strongly convex since f is convex & μ-strongly convex.

$$\phi(\mathbf{x}) = \min_{\mathbf{z}} \phi(\mathbf{z}) \ge \min_{\mathbf{z}} \left\{\phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^\top(\mathbf{z}-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{z}-\mathbf{y}\|_2^2\right\}$$
$$\overset{r=\mathbf{z}-\mathbf{y}}{=} \min_r \left\{\phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^\top r + \frac{\mu}{2}\|r\|_2^2\right\}$$
$$\overset{\text{solve } r}{=} \phi(\mathbf{y}) - \frac{\|\phi(\mathbf{y})\|_2^2}{2\mu}.$$

Then, we have $2\mu(\phi(\mathbf{y}) - \phi(\mathbf{x})) \le \|\phi(\mathbf{y})\|_2^2$, which prove the result by plugging in $\phi(\mathbf{z}) = f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \mathbf{z}$ and $\nabla\phi(\mathbf{z}) = \nabla f(\mathbf{z}) - \nabla f(\mathbf{x})$.

part (b), (c), (d) can be proved similarly as the previous lemma. □

Lemma 1.7 *If $r(\cdot)$ is μ-strongly convex w.r.t $\|\cdot\|_2$ and*

$$\text{prox}_{\eta r}(\mathbf{z}_1) := \arg\min_{\mathbf{w}} r(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}_1\|_2^2, \tag{1.19}$$

$$\text{prox}_{\eta r}(\mathbf{z}_2) := \arg\min_{\mathbf{w}} r(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w}-\mathbf{z}_2\|_2^2\,, \tag{1.20}$$

then we have $\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2 \le \frac{1}{1+\mu\eta}\|\mathbf{z}_1-\mathbf{z}_2\|_2$.

Proof. First, we can see that when $r = 0$, the conclusion trivially holds. Next, we prove it when r is present.

By the optimality of $\text{prox}_{\eta r}(\mathbf{z}_1)$ and $\text{prox}_{\eta r}(\mathbf{z}_1)$ we have

$$\begin{aligned}\mathbf{u} :=& \frac{\mathbf{z}_1 - \text{prox}_{\eta r}(\mathbf{z}_1)}{\eta} \in \partial r(\text{prox}_{\eta r}(\mathbf{z}_1))\\ \mathbf{v} :=& \frac{\mathbf{z}_2 - \text{prox}_{\eta r}(\mathbf{z}_2)}{\eta} \in \partial r(\text{prox}_{\eta r}(\mathbf{z}_2)).\end{aligned}$$

Since $r(\mathbf{x})$ is μ-strongly convex, we have

$$\begin{aligned} r(\text{prox}_{\eta r}(\mathbf{z}_1)) \ge\ & r(\text{prox}_{\eta r}(\mathbf{z}_2)) + \mathbf{v}^\top(\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2))\\ &+ \frac{\mu}{2}\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2^2\\ r(\text{prox}_{\eta r}(\mathbf{z}_2)) \ge\ & r(\text{prox}_{\eta r}(\mathbf{z}_1)) + \mathbf{u}^\top(\text{prox}_{\eta r}(\mathbf{z}_2) - \text{prox}_{\eta r}(\mathbf{z}_1))\\ &+ \frac{\mu}{2}\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2^2. \end{aligned}$$

Adding them together, we have

$$\begin{aligned}&\mu\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2^2 \le (\mathbf{u}-\mathbf{v})^\top(\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2))\\ &= \frac{1}{\eta}(\mathbf{z}_1 - \mathbf{z}_2 + \text{prox}_{\eta r}(\mathbf{z}_2) - \text{prox}_{\eta r}(\mathbf{z}_1))^\top(\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)),\end{aligned}$$

which implies

$$\begin{aligned}&(\mu + \frac{1}{\eta})\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2^2 \le \frac{1}{\eta}(\mathbf{z}_1-\mathbf{z}_2)^\top(\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2))\\ &\le \frac{1}{\eta}\|\mathbf{z}_1-\mathbf{z}_2)\|_2\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2.\end{aligned}$$

Thus $\|\text{prox}_{\eta r}(\mathbf{z}_1) - \text{prox}_{\eta r}(\mathbf{z}_2)\|_2 \le \frac{1}{\mu\eta+1}\|\mathbf{z}_1-\mathbf{z}_2\|_2$. □

Lemma 1.8 *For a proper closed convex function f, the following holds:*

(i) if f is G-Lipchitz continuous w.r.t $\|\cdot\|_2$, then $dom(f^*)$ is bounded and for any $\mathbf{y} \in dom(f^*)$, we have $\|\mathbf{y}\|_2 \le G$;

(ii) if $\mathbf{x}_* \in \arg\max_{\mathbf{x}}\{\mathbf{x}^\top \mathbf{y}_* - f(\mathbf{x})\}$*, then* $\mathbf{y}_* \in \arg\max_{\mathbf{y}}\{\mathbf{y}^\top \mathbf{x}_* - f^*(\mathbf{y})\}$*. Equivalently,* $\mathbf{y}_* \in \partial f(\mathbf{x}_*)$ *(or* $\mathbf{x}_* \in \partial f^*(\mathbf{y}_*)$*);*

(iii) if f *is further a Legendre function, then* $f(\mathbf{x}_*) + f^*(\mathbf{y}_*) = \mathbf{x}_*^\top \mathbf{y}_*$ *if and only if* $\mathbf{y}_* = \nabla f(\mathbf{x}_*)$*, and* $\nabla f^* = (\nabla f)^{-1}$.

Proof. Let us prove (i). For any $\mathbf{y}$ with $\|\mathbf{y}\|_2 > G$, let $\mathbf{u} = \mathbf{y}/\|\mathbf{y}\|_2$ and take $\mathbf{x} = t\mathbf{u}$. By Lipschitz continuity, $f(t\mathbf{u}) \le f(0) + Gt$, hence

$$\mathbf{y}^\top t\mathbf{u} - f(t\mathbf{u}) \ge t(\|\mathbf{y}\|_2 - G) - f(0) \to +\infty,$$

so $f^*(\mathbf{y}) = +\infty$ and thus $\mathbf{y} \notin \mathrm{dom}(f^*)$.

Next, we prove (ii). Since $\mathbf{x}_*$ attains the supremum in the definition of $f^*(\mathbf{y}_*)$, we have $\mathbf{y}_* \in \partial f(\mathbf{x}_*)$ according to the optimality condition and $f^*(\mathbf{y}_*) = \mathbf{x}_*^\top \mathbf{y}_* - f(\mathbf{x}_*)$. Using $f^{**} = f$, we obtain

$$f(\mathbf{x}_*) = \sup_{\mathbf{y}}\{\mathbf{y}^\top \mathbf{x}_* - f^*(\mathbf{y})\} = \mathbf{x}_*^\top \mathbf{y}_* - f^*(\mathbf{y}_*),$$

and the above equality shows that $\mathbf{y}_*$ attains the supremum. Hence, $\mathbf{y}_* \in \arg\max_{\mathbf{y}}\{\mathbf{y}^\top \mathbf{x}_* - f^*(\mathbf{y})\}$, and $\mathbf{x}_* \in \partial f^*(\mathbf{y}_*)$.

Lastly, we prove (iii). By definition, $f^*(\mathbf{y}_*)$ is the supremum of the concave function $F(\mathbf{x}) = \mathbf{y}_*^\top \mathbf{x} - f(\mathbf{x})$. If this supremum is attained at $\mathbf{x}_* \in \mathbb{R}^d$, then $\nabla F(\mathbf{x}_*) = 0$, which is to say $\mathbf{y}_* = \nabla f(\mathbf{x}_*)$. On the other hand, if $\mathbf{y}_* = \nabla f(\mathbf{x}_*)$, then $\mathbf{x}_*$ is a maximizer of $F(\mathbf{x})$, and therefore $f^*(\mathbf{y}_*) = \mathbf{y}_*^\top \mathbf{x}_* - f(\mathbf{x}_*)$. Using the this result twice,

$$\begin{aligned} \mathbf{y} = \nabla f(\mathbf{x}) \quad &\text{if and only if} \quad f(\mathbf{x}) + f^*(\mathbf{y}) = \mathbf{x}^\top \mathbf{y} \\ \mathbf{x} = \nabla f^*(\mathbf{y}) \quad &\text{if and only if} \quad f^*(\mathbf{y}) + f^{**}(\mathbf{x}) = \mathbf{x}^\top \mathbf{y}. \end{aligned}$$

Since $f^** = f$, then $\mathbf{x} = \nabla f^{-1}(\mathbf{y}) = \nabla f^*(\mathbf{y})$. Hence $(\nabla f)^{-1} = \nabla f^*$. □

Lemma 1.9 *If* f *is* μ*-strongly convex w.r.t* $\|\cdot\|_2$*, then its Fenchel conjugate is* $1/\mu$*-smooth. Similarly if* f *is* L*-smooth and convex w.r.t* $\|\cdot\|_2$*, then its Fenchel conjugate is* $1/L$*-strongly convex.*

Proof. Let $f^*(\mathbf{y}) = \max_{\mathbf{x}} \mathbf{x}^\top \mathbf{y} - f(\mathbf{x})$ be the Fenchel conjugate of f.

Suppose f is μ-strongly convex. let $\mathbf{x}(\mathbf{y}) = \arg\max_{\mathbf{x}} \mathbf{x}^\top \mathbf{y} - f(\mathbf{x})$. Then $\nabla f^*(\mathbf{y}) = \mathbf{x}(\mathbf{y})$ due to the Danskin Theorem. Similar to the previous lemma, we can prove that

$$\|\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)\|_2 \le \frac{1}{\mu}\|\mathbf{y}_1 - \mathbf{y}_2\|_2,$$

which proves the Lipchitz continuity of $\nabla f^*(\mathbf{y})$ and hence the smoothness of f^*.

Suppose f is L-smooth and convex. Let us prove f^* is $1/L$-strongly convex. Let us consider $\mathbf{y}_1, \mathbf{y}_2$. Let $\mathbf{x}_1 \in \arg\max_{\mathbf{x}} \mathbf{x}^\top \mathbf{y}_1 - f(\mathbf{x})$ and $\mathcal{X}_2 = \arg\max_{\mathbf{x}} \mathbf{x}^\top \mathbf{y}_2 - f(\mathbf{x})$. Then $\nabla f(\mathbf{x}_1) = \mathbf{y}_1$. For any $\mathbf{x}_2 \in \mathcal{X}_2$, we have $\nabla f(\mathbf{x}_2) = \mathbf{y}_2$. Given that

$$f^*(\mathbf{y}_1) = \mathbf{x}_1^\top \mathbf{y}_1 - f(\mathbf{x}_1), \quad f^*(\mathbf{y}_2) = \mathbf{x}_2^\top \mathbf{y}_2 - f(\mathbf{x}_2),$$

then

$$
\begin{aligned}
&f^*(\mathbf{y}_1) - f^*(\mathbf{y}_2) - \mathbf{x}_2^\top(\mathbf{y}_1 - \mathbf{y}_2) \\
&= \mathbf{x}_1^\top \mathbf{y}_1 - f(\mathbf{x}_1) - (\mathbf{x}_2^\top \mathbf{y}_2 - f(\mathbf{x}_2)) - \mathbf{x}_2^\top(\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)) \\
&= f(\mathbf{x}_2) - f(\mathbf{x}_1) + \mathbf{x}_1^\top \nabla f(\mathbf{x}_1) - \mathbf{x}_2^\top \nabla f(\mathbf{x}_2) - \mathbf{x}_2^\top(\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)) \\
&= f(\mathbf{x}_2) - f(\mathbf{x}_1) + (\mathbf{x}_1 - \mathbf{x}_2)^\top \nabla f(\mathbf{x}_1) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2^2 = \frac{1}{2L}\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2,
\end{aligned}
$$

where the last inequality is due to part (b) of Lemma 1.5. Hence, we can conclude the proof by noting that $\partial f^*(\mathbf{y}_2) = \mathrm{conv}(\mathcal{X}_2)$ due to the generalized Danskin theorem. □

Lemma 1.10 *For $\mathbf{p} \in \Delta_n$, the negative entropy function $R(\mathbf{p}) = \sum_{i=1}^n p_i \log p_i$ is 1-strongly convex w.r.t to the ℓ_1 norm $\|\cdot\|_1$.*

Proof. For any $\mathbf{x}, \mathbf{y} \in \Delta_n$, let $f(t) = R(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$. By the second-order Taylor expansion, for some $t \in (0, 1)$, we have

$$
\begin{aligned}
R(\mathbf{x}) = f(1) &= f(0) + f'(0) + \frac{1}{2}f''(t) \\
&= R(\mathbf{y}) + \nabla R(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top \nabla^2 R(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y}).
\end{aligned}
$$

Hence it suffices to prove that $\mathbf{v}^\top \nabla^2 R(\mathbf{p})\mathbf{v} \geq \|\mathbf{v}\|_1^2$ for any $\mathbf{p} \in \Delta_n$. This can be seen from the following:

$$
\begin{aligned}
\mathbf{v}^\top \nabla^2 R(\mathbf{p})\mathbf{v} = \sum_{i=1}^d v_i^2 p_i^{-1} &= \left[\sum_i v_i^2 p_i^{-1}\right]\left[\sum_i p_i\right] \geq \left[\sum_i (p_i^{-1/2}|v_i|)p_i^{1/2}\right]^2 \\
&= \left[\sum_i |v_i|\right]^2,
\end{aligned}
$$

where the inequality follows by Cauchy inequality.

□

1.6 History and Notes

This chapter has selectively introduced core concepts from convex optimization that are most pertinent to the algorithms and applications discussed in later chapters. While the treatment here is necessarily concise, readers seeking a more comprehensive foundation are encouraged to consult several classic references.

The text by Rockafellar (1970a) provides one of the most comprehensive and authoritative treatments of convex analysis. The textbook by Boyd and Vandenberghe (2004) is an excellent introduction to convex optimization well suited for engineers.

It covers convex sets, convex functions, duality, and optimality conditions in detail, and emphasizes geometric intuition and practical modeling. Many of the definitions and examples in this chapter are inspired by this text. Bertsekas (2009) offers deep insights into convex analysis, duality theory, and constrained optimization from a classical perspective.

The KKT condition is named after three mathematicians, William Karush, Harold W. Kuhn and Albert W. Tucker. It was known due to Kuhn and Tucker, who first published the conditions in 1951 (Kuhn and Tucker, 2014). Later scholars discovered that the necessary conditions for this problem had been stated by Karush in his master's thesis in 1939 (Karush, 1939). The Danskin Theorem originates from the work of Danskin (1967), while its generalized form for subdifferentiable is attributed to Bertsekas (2005).

Nesterov's *Introductory Lectures on Convex Programming* (Nesterov, 2004) provides a more mathematically rigorous treatment, including several key lemmas on smooth and strongly convex functions (Lemma 1.5 and Lemma 1.6) that are presented in this chapter. It is particularly useful for readers interested in complexity analysis and the theoretical underpinnings of first-order methods. The proof of Lemma 1.10 is due to Nemirovski et al. (2009).