# Online Learning
# Notes

Sih-Ray Mao

-

Last modified: 2025-10-08

**Disclaimer.** This is a self-study notes on online learning, mainly following the text *A Modern Introduction to Online Learning* by Francesco Orabona [Ora25]. This notes was prepared as part of my (ongoing) independent study to develop a strong foundation in convex optimization and online learning for my graduate study. The goal is to summarize key concepts, proofs, and algorithms in a clear and structured manner, while adding personal insights and connections relevant to theoretical machine learning. Please note that this notes is for personal study and may not be fully polished. They are not a substitute for the original sources, which should be consulted for rigorous detail. Any error or omissions are on my own.

# Contents

# 1 Introduction

## 1.1 The Online Learning Protocol

Consider the following repeated game. Let $\mathcal{V} \subseteq \mathbb{R}^d$ and let $T$ be the total number of rounds. At the $t$-th round, where $1 \leq t \leq T$, the game proceeds as

(i) the learner announces an action $x_t$ from $\mathcal{V}$;

(ii) then, the adversary announces a function $\ell_t : \mathcal{X} \to \mathbb{R}$; and

(iii) finally, the learner suffers loss $\ell_t(x_t)$.

The set $\mathcal{V}$ is called the **action set** or the **feasible set**, and the function $\ell_t$ is called the **loss function** at the $t$-th round. Note that the loss functions $\ell_t$'s need to be identical.

We do not impose any assumption on the adversary in the above repeated game. This is very different from the usual assumption in machine learning settings, where data are assumed to be generated identically and independently (i.i.d.) from an unknown distribution. We relax the i.i.d. assumption in online learning. Moreover, the adversary, as the name suggests, may even be adversarial. The adversary has the power to select the loss function after observing the learner's actions from previous rounds. In other words, we analyze algorithms under the worst-case assumption in online learning.

**Definition 1.1.1** (Regret)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be the feasible set and let $\ell_1, \dots, \ell_T$ be the loss functions. The **(standard/static) regret** of an online learning algorithm $\mathcal{A}$ with respect to the competitor $u \in \mathcal{V}$ is defined as*

$$\operatorname{Regret}_T^{\mathcal{A}}(u) \coloneqq \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(u).$$

*The **(standard/static) regret** of an online learning algorithm $\mathcal{A}$ is defined as*

$$\operatorname{Regret}_T^{\mathcal{A}} \coloneqq \sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(x).$$

*An online learning algorithm $\mathcal{A}$ is said to be **no-regret** if $\operatorname{Regret}_T^{\mathcal{A}}$ is sublinear, that is, if $\operatorname{Regret}_T^{\mathcal{A}} = o(T)$.*

Namely, the regret of an online learning algorithm $\mathcal{A}$ is defined as the difference between the cumulative loss incurred by $\mathcal{A}$ and that of the optimal fixed choice in hindsight. When the context is clear, we will denote by $\operatorname{Regret}_T$ the regret of an online learning algorithm $\mathcal{A}$.

Considering that the loss functions are not assumed to be identical, we might wonder why measuring the performance of an online learning algorithm in a dynamic environment with static regret makes sense. As we will see, this is because we can use a technique called "online-to-batch conversion" to convert an online learning algorithm into an optimization algorithm working in the batch setting. Moreover, we can ensure that the resulting algorithm has satisfactory convergence guarantee as long as the online learning algorithm has sublinear regret. On the other hand, we can also measure the performance of an online learning algorithm using "dynamic regret" or "tracking regret."

## 1.2 Follow-the-Leader

The simplest online learning algorithm outputs a minimizer of the cumulative loss in each round. This algorithm is known as Follow-the-Leader.

**Algorithm 1** Follow-the-Leader

**Inputs:** A nonempty set $\mathcal{V} \subseteq \mathbb{R}^d$ and an initial point $x_1 \in \mathcal{V}$.

1: **for** $t = 1, \dots, T$ **do**

2: The learner outputs $x_t \in \mathcal{V}$.

3: The adversary chooses a loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.

4: The learner updates by setting

$$x_{t+1} \leftarrow \operatorname*{arg\,min}_{x \in \mathcal{V}} \sum_{\tau=1}^{t} \ell_\tau(x).$$

5: **end for**

**Lemma 1.2.1** (Be-the-Leader lemma, Hannan [Han57]). *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be nonempty and let $\ell_1, \dots, \ell_T : \mathcal{V} \to \mathbb{R}$ be an arbitrary sequence of loss functions. Assume that $\arg\min_{x \in \mathcal{V}} \sum_{\tau=1}^{t} \ell_\tau(x)$ is nonempty and set $x_{t+1} \in \arg\min_{x \in \mathcal{V}} \sum_{\tau=1}^{t} \ell_\tau(x)$ for $t = 1, \dots, T$. Then, we have*

$$\sum_{t=1}^{T} \ell_t(x_{t+1}) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(x) \le 0. \tag{1.1}$$

*Proof.* The proof is by induction on $T$. Since $x_2 \in \arg\min_{x \in \mathcal{V}} \ell_1(x)$, we clearly have $\ell_1(x_2) - \min_{x \in \mathcal{V}} \ell_1(x) \le 0$. Thus, Equation (1.1) holds when $T = 1$. Now, suppose that Equation (1.1) holds for $T - 1$, that is, we have

$$\sum_{t=1}^{T-1} \ell_t(x_{t+1}) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T-1} \ell_t(x) \le 0. \tag{1.2}$$

We can write

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(x_{t+1}) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(x) &= \sum_{t=1}^{T} \ell_t(x_{t+1}) - \sum_{t=1}^{T} \ell_t(x_{T+1}) \\
&= \sum_{t=1}^{T-1} \ell_t(x_{t+1}) - \sum_{t=1}^{T-1} \ell_t(x_{T+1}) \\
&\le \sum_{t=1}^{T-1} \ell_t(x_{t+1}) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T-1} \ell_t(x) \\
&\le 0,
\end{aligned}$$

where the last inequality follows from the induction hypothesis (Equation (1.2)). Therefore, we conclude that Equation (1.1) holds for all $T$. □

In words, Lemma 1.2.1 tells us that "being the leader" guarantees a nonpositive regret. The Be-the-Leader algorithm is as follows.

**Algorithm 2** Be-the-Leader

**Inputs:** A nonempty set $\mathcal{V} \subseteq \mathbb{R}^d$.

1: **for** $t = 1, \dots, T$ **do**
2: The learner outputs

$$x_t \leftarrow \arg\min_{x \in \mathcal{V}} \sum_{\tau=1}^{t} \ell_\tau(x).$$

3: The adversary chooses a loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
4: **end for**

Note that Be-the-Leader is not a valid online learning algorithm. This is because at the $t$-th iteration, the prediction $x_t$ is chosen with the knowledge of the loss function $\ell_t$, which is clearly infeasible.

Let us look at our first concrete example of online learning.

**Example 1.2.2** (Guessing game). *Consider the following online learning problem. Let $[0, 1]$ be the feasible set. At the $t$-th round, where $t = 1, \dots, T$,*

*(i) the learner chooses a number $x_t \in \mathcal{V}$;*

*(ii) the adversary reveals the true number $y_t \in \mathcal{V}$ and the loss function $\ell_t(x) = (x - y_t)^2$; and*

*(iii) the learner suffers loss $\ell_t(x_t) = (x_t - y_t)^2$.*

*Then, the regret of* Follow-the-Leader *is bounded by*

$$\mathrm{Regret}_T \le 4 + 4 \log T.$$

*Proof.* The cumulative loss function over the first $t$ rounds is given by $\sum_{\tau=1}^{t} (x - y_\tau)^2$. Note that the sample average of the $y_\tau$'s is a minimizer of the cumulative loss. Thus, Follow-the-Leader iterates as

$$x_{t+1} = \arg\min_{x \in [0,1]} \sum_{\tau=1}^{t} (x - y_\tau)^2 = \frac{1}{t} \sum_{\tau=1}^{t} y_\tau, \tag{1.3}$$

for $t = 1, \dots, T$. We can write

$$\begin{aligned}
\text{Regret}_T &= \sum_{t=1}^{T} (x_t - y_t)^2 - \min_{x \in [0,1]} \sum_{t=1}^{T} (x - y_t)^2 \\
&\leq \sum_{t=1}^{T} (x_t - y_t)^2 - \sum_{t=1}^{T} (x_{t+1} - y_t)^2 \\
&\leq \sum_{t=1}^{T} |x_t + x_{t+1} - 2y_t| \cdot |x_t - x_{t+1}| \\
&\leq 2 \sum_{t=1}^{T} |x_t - x_{t+1}| \\
&= 2 \sum_{t=1}^{T} \left| \frac{1}{t-1} \sum_{\tau=1}^{t-1} y_\tau - \frac{1}{t} \sum_{\tau=1}^{t} y_\tau \right| \\
&= 2 \sum_{t=1}^{T} \left| \frac{1}{t(t-1)} \sum_{\tau=1}^{t-1} y_\tau - \frac{1}{t} y_t \right| \\
&\leq 2 \sum_{t=1}^{T} \left( \frac{1}{t} |y_t| + \frac{1}{t(t-1)} \sum_{\tau=1}^{t-1} |y_\tau| \right) \\
&\leq \sum_{t=1}^{T} \frac{4}{t}, \qquad (1.4)
\end{aligned}$$

where the first inequality follows from the Be-the-Leader lemma (Lemma 1.2.1), the third inequality follows from the fact that $x_t, x_{t+1}$, and $y_t$ are contained in $[0, 1]$, the second equality follows from Equation (1.3), the second last inequality follows from the triangle inequality, and the last inequality follows from the fact that the $y_\tau$'s lie in $[0, 1]$. Observe that the sum $\sum_{t=2}^{T} 1/t$ is upper bounded by the integral $\int_2^T 1/t \, \mathrm{d}t$. Thus, we have

$$\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \int_2^T \frac{1}{t} \, \mathrm{d}t = 1 + \log T - \log 2 \leq 1 + \log T. \qquad (1.5)$$

Combining Equation (1.4) and Equation (1.5) together yields

$$\text{Regret}_T \leq \sum_{t=1}^{T} \frac{4}{t} \leq 4 + 4 \log T.$$

□

As Example 1.2.2 shows, Follow-the-Leader achieves sublinear regret in the guessing game. Does Follow-the-Leader always achieve sublinear regret? The answer is *no* in general. Consider the following counterexample.

**Example 1.2.3** (Failure of Follow-the-Leader)**.** *Let $\mathcal{V} = [-1, 1]$ be the feasible set and let the loss functions $\ell_t$'s be defined as*

$$\ell_t(x) = \begin{cases} -0.5x, & \text{if } t = 1; \\ x, & \text{if } t = 2, 4, \dots; \\ -x, & \text{if } t = 3, 5, \dots. \end{cases}$$

*Then,* Follow-the-Leader *fails to achieve a sublinear regret in this setting.*

*Proof.* Observe that we have

$$\sum_{\tau=1}^{t} \ell_\tau(x) = \begin{cases} -0.5x, & \text{if } t \text{ is odd}; \\ 0.5x, & \text{if } t \text{ is even}. \end{cases}$$

Then, by the updating rule of Follow-the-Leader (Algorithm 1), we have $x_t = 1$ when $t$ is even and $x_t = -1$ when $t$ is odd (for $t \geq 2$). Thus, the cumulative loss of Follow-the-Leader is given by

$$\sum_{t=1}^{T} \ell_t(x_t) = T - 1 - \frac{x_1}{2}.$$

On the other hand, the cumulative loss of the fixed strategy of choosing $x_t = 0$ for all $t$ is 0. Thus, we have

$$\text{Regret}_T = \sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(x) \geq T - 1 - \frac{x_1}{2} \geq T - \frac{3}{2}.$$

Therefore, Follow-the-Leader fails to achieve a sublinear regret bound in this setting. □

Let us take a closer look at the behavior of Follow-the-Leader in Example 1.2.3. In this example, Follow-the-Leader jumps back and forth between $-1$ and 1. Intuitively, Follow-the-Leader fails to achieve a sublinear regret because it is "unstable." More precisely, from the Be-the-Leader lemma (Lemma 1.2.1), we can write

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(x) \leq \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x_{t+1}) = \sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(x_{t+1})). \tag{1.6}$$

Equation (1.6) tells us that the regret of Follow-the-Leader is small if $x_t$ is close to $x_{t+1}$. In other words, the regret of Follow-the-Leader is determined by the stability of the algorithm.

# 2 Online Subgradient Descent

In this chapter, we introduce the Online Subgradient Descent algorithm, an algorithm for online learning problems with *convex subdifferentiable* losses. Let us first look at the simpler case when the loss functions are assumed to be *convex differentiable*.

## 2.1 Online Gradient Descent

Before we introduce the Online Gradient Descent algorithm for convex loss functions, we need some elementary results from convex analysis. We begin by introducing the notion of *convexity*.

**Definition 2.1.1** (Convex set)**.** *A set $\mathcal{V} \subseteq \mathbb{R}^d$ is said to be* ***convex*** *if and only if*

$$\lambda x + (1-\lambda)y \in \mathcal{V}, \quad \forall x, y \in \mathcal{V}, \lambda \in [0,1].$$

In words, a set $\mathcal{V}$ is convex if and only if for any two points $x, y \in \mathcal{V}$, the line segment between $x$ and $y$ lies entirely in $\mathcal{V}$. For example, the unit open ball $B = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$ in $\mathbb{R}^d$ is convex. The real line $\mathbb{R}$ and the Euclidean space $\mathbb{R}^d$ are clearly convex. The empty set is also convex since the condition that $\lambda x + (1-\lambda)y \in \varnothing$ for all $x, y \in \varnothing$ and $\lambda \in [0,1]$ holds vacuously.

**Proposition 2.1.2.** *Let $\{\mathcal{V}_\alpha\}_{\alpha\in\Lambda}$ be an arbitrary collection of convex sets, where $\Lambda$ is an index set. Then, the intersection $\bigcap_{\alpha\in\Lambda} \mathcal{V}_\alpha$ is convex.*

*Proof.* For brevity, we write $\mathcal{V} \coloneqq \bigcap_{\alpha\in\Lambda} \mathcal{V}_\alpha$. Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. Since each $\mathcal{V}_\alpha$ is convex, we have $\lambda x + (1-\lambda)y \in \mathcal{V}_\alpha$ for each $\alpha \in \Lambda$. Thus, we have $\lambda x + (1-\lambda)y \in \mathcal{V}$. Therefore, $\mathcal{V}$ is convex. □

**Proposition 2.1.3.** *Let $\mathcal{V}_1, \ldots, \mathcal{V}_m$ be convex sets in $\mathbb{R}^d$. Then, the product $\bigtimes_{i=1}^m \mathcal{V}_i$ is convex if and only if each $\mathcal{V}_i$ is convex.*

*Proof.* For brevity, we write $\mathcal{V} \coloneqq \bigtimes_{i=1}^m \mathcal{V}_i$. Let $x = (x_1, \ldots, x_m), y = (y_1, \ldots, y_m) \in \mathcal{V}$, where $x_i, y_i \in \mathcal{V}_i$ for $i = 1, \ldots, m$, and let $\lambda \in [0,1]$. The product $\mathcal{V}$ is convex if and only if $\lambda x + (1-\lambda)y \in \mathcal{V}$, if and only if $\lambda x_i + (1-\lambda)y_i \in \mathcal{V}_i$ for $i = 1, \ldots, m$, if and only if $\mathcal{V}_i$ is convex for $i = 1, \ldots, m$. □

The extended real number $[-\infty, +\infty]$ is obtained by adjoining $-\infty$ and $+\infty$ with the set of real numbers, that is, $[-\infty, +\infty] = \mathbb{R} \cup \{-\infty, +\infty\}$. Arithmetic operations of the extended real numbers are defined as follows. For each $x \in \mathbb{R}$, we define $x + (+\infty) \coloneqq +\infty$ and $x + (-\infty) \coloneqq -\infty$. The expressions $(+\infty) + (-\infty)$ and $(-\infty) + (+\infty)$ are undefined, unless stated otherwise. For $\alpha > 0$, we define $\alpha \cdot (+\infty) \coloneqq +\infty$ and $\alpha \cdot (-\infty) \coloneqq -\infty$. For $\alpha < 0$, we define $\alpha \cdot (+\infty) \coloneqq -\infty$ and $\alpha \cdot (-\infty) \coloneqq +\infty$. For each $x \in \mathbb{R}$, we define $x - (+\infty) \coloneqq x + (-1)(+\infty) = x + (-\infty) = -\infty$ and $x - (-\infty) \coloneqq x + (-1)(-\infty) = x + (+\infty) = +\infty$.

**Definition 2.1.4** (Domain)**.** *An* ***extended real-valued function*** *on $\mathbb{R}^d$ is a function that takes value in $[-\infty, +\infty]$. The* ***(effective) domain*** *of an extended real-valued function $f : \mathbb{R}^d \to [-\infty, +\infty]$, denoted by $\operatorname{dom}(f)$, is defined as*

$$\operatorname{dom}(f) \coloneqq \{x \in \mathbb{R}^d : f(x) < +\infty\}.$$

**Example 2.1.5** (Indicator function)**.** *The* ***indicator function*** *of $\mathcal{V} \subseteq \mathbb{R}^d$, denoted by $\iota_\mathcal{V} : \mathbb{R}^d \to (-\infty, +\infty]$, is defined as*

$$\iota_\mathcal{V}(x) \coloneqq \begin{cases} 0, & \text{if } x \in \mathcal{V}; \\ +\infty, & \text{otherwise.} \end{cases}$$

Let $\mathcal{V} \subseteq \mathbb{R}^d$. Clearly, the indicator function $\iota_\mathcal{V}$ is an extended real-valued function. From the definition of $\iota_\mathcal{V}$ (Example 2.1.5), it follows immediately that the domain of $\iota_\mathcal{V}$ is nonempty if and only if $\mathcal{V}$ is nonempty.

The indicator functions are useful for solving optimization problems. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function. Suppose we want to solve the following constrained minimization problem:

$$\min_{x\in\mathcal{V}} f(x).$$

We can convert this constrained optimization problem into an unconstrained one by introducing the indicator function $\iota_\mathcal{V}$. The above constrained optimization problem is equivalent to the following unconstrained minimization problem:

$$\min_{x\in\mathbb{R}^d} \{f(x) + \iota_\mathcal{V}(x)\}.$$

Indeed, for $x \notin \mathcal{V}$, we have $f(x) + \iota_\mathcal{V}(x) = +\infty$. Thus, the minimum of the function $f(x) + \iota_\mathcal{V}(x)$ must be attained when $x \in \mathcal{V}$. Therefore, the two minimization problems are equivalent.

**Definition 2.1.6** (Epigraph)**.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$ be an extended real-valued function. The* ***epigraph*** *of $f$, denoted by* $\operatorname{epi}(f)$*, is defined as*

$$\operatorname{epi}(f) \coloneqq \{(x, t) \in \operatorname{dom}(f) \times \mathbb{R} : f(x) \leq t\}.$$

Geometrically speaking, the epigraph of an extended real-valued function $f$ is the set above the graph of $f$.

**Definition 2.1.7** (Convex function)**.** *An extended real-valued function $f : \mathbb{R}^d \to [-\infty, +\infty]$ is said to be* ***convex*** *if and only if* $\operatorname{epi}(f)$ *is convex.*

For example, consider the indicator function (Example 2.1.5) of a subset $\mathcal{V} \subseteq \mathbb{R}^d$. The epigraph of $\iota_{\mathcal{V}}$ is $\operatorname{epi}(\iota_{\mathcal{V}}) = \mathcal{V} \times \mathbb{R}$. Then, from Proposition 2.1.3, we know that $\operatorname{epi}(\iota_{\mathcal{V}})$ is convex if and only if $\mathcal{V}$ is convex. Therefore, we conclude that the indicator function $\iota_{\mathcal{V}}$ is convex if and only if $\mathcal{V}$ is convex.

**Theorem 2.1.8** (Zeroth-order condition)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be an extended real-valued function. Then, $f$ is convex if and only if* $\operatorname{dom}(f)$ *is convex and*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \operatorname{dom}(f), \lambda \in [0, 1]. \tag{2.1}$$

*Proof.* Suppose that $f : \mathbb{R}^d \to (-\infty, +\infty]$ is convex. Since $\operatorname{epi}(f)$ is convex, from Proposition 2.1.3, we know that $\operatorname{dom}(f)$ is convex. Let $x, y \in \operatorname{dom}(f)$. Since $(x, f(x)), (y, f(y)) \in \operatorname{epi}(f)$ and $\operatorname{epi}(f)$ is convex, for all $\lambda \in [0, 1]$, we have

$$\lambda \cdot (x, f(x)) + (1-\lambda) \cdot (y, f(y)) = (\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)) \in \operatorname{epi}(f).$$

It follows that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \operatorname{dom}(f), \lambda \in [0, 1].$$

Conversely, suppose that $\operatorname{dom}(f)$ is convex and Equation (2.1) holds. Let $(x, s), (y, t) \in \operatorname{epi}(f)$. Then, for all $\lambda \in [0, 1]$, we can write

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda s + (1-\lambda) t \tag{2.2}$$

Equation (2.2) implies that $(\lambda x + (1-\lambda)y, \lambda s + (1-\lambda)t) \in \operatorname{epi}(f)$ for all $\lambda \in [0, 1]$. Thus, we have $\lambda \cdot (x, s) + (1-\lambda) \cdot (y, t) \in \operatorname{epi}(f)$ for all $\lambda \in [0, 1]$. Therefore, $\operatorname{epi}(f)$ is convex and thus $f$ is convex. □

**Definition 2.1.9** (Convex function)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex and let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be an extended real-valued function. We say that $f$ is* ***convex*** *on $\mathcal{V}$ if and only if*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \mathcal{V}, \lambda \in [0, 1].$$

Namely, a function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is convex on a convex set $\mathcal{V} \subseteq \mathbb{R}^d$ if and only if the zeroth-order condition (Theorem 2.1.8) holds for all $x, y \in \mathcal{V}$ and $\lambda \in [0, 1]$.

**Example 2.1.10** (Norm)**.** *Any norm is convex.*

*Proof.* Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. The domain of $\|\cdot\|$ is the whole space $\mathbb{R}^d$ and is clearly convex. Recall that $\|\cdot\|$ is absolute homogeneous, that is, we have $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^d$. Moreover, $\|\cdot\|$ satisfies the triangle inequality, that is, $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^d$. Then, for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$, we have

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda \|x\| + (1-\lambda)\|y\|.$$

From the zeroth-order condition (Theorem 2.1.8), we know that $\|\cdot\|$ is convex. □

**Example 2.1.11** (Affine function)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is* ***affine*** *if and only if*

$$f(\lambda x + (1-\lambda)y) = \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \mathbb{R}^d, \lambda \in \mathbb{R}. \tag{2.3}$$

*Any affine function is convex.*

*Proof.* Clearly, Equation (2.3) implies the zeroth-order condition (Theorem 2.1.8). We conclude that any affine function is convex. □

**Proposition 2.1.12.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a convex set. The following statements hold.*

*(a) If $f : \mathbb{R}^d \to (-\infty, +\infty]$ is a convex function on $\mathcal{V}$ and $\alpha > 0$ is a constant, then $\alpha f$ is convex on $\mathcal{V}$.*

*(b) If $f, g : \mathbb{R}^d \to (-\infty, +\infty]$ are convex functions on $\mathcal{V}$, then $f + g$ is convex on $\mathcal{V}$.*

*(c) Let $f_1, \dots, f_n : \mathbb{R}^d \to (-\infty, +\infty]$ be convex functions on $\mathcal{V}$. Define $f : \mathbb{R}^d \to (-\infty, +\infty]$ as*

$$f(x) \coloneqq \max_{1 \le i \le n} f_i(x), \quad \forall x \in \mathbb{R}^d.$$

*Then, $f$ is convex on $\mathcal{V}$.*

*Proof of (a).* Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. Since $f$ is convex on $\mathcal{V}$ and $\alpha > 0$, we have

$$(\alpha f)(\lambda x + (1-\lambda)y) = \alpha \cdot f(\lambda x + (1-\lambda)y) \le \alpha\lambda f(x) + \alpha(1-\lambda)f(y) = \lambda(\alpha f)(x) + (1-\lambda)(\alpha f)(y).$$

Thus, $\alpha f$ is convex on $\mathcal{V}$. □

*Proof of (b).* Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. Since $f$ and $g$ are convex functions on $\mathcal{V}$, we have

$$\begin{aligned}(f+g)(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &\le \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y) \\ &= \lambda(f+g)(x) + (1-\lambda)(f+g)(y).\end{aligned}$$

Thus, $f + g$ is convex on $\mathcal{V}$. □

*Proof of (c).* Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. We can write

$$\begin{aligned}f(\lambda x + (1-\lambda)y) &= \max_{1 \le i \le n} f_i(\lambda x + (1-\lambda)y) \\ &\le \max_{1 \le i \le n} \{\lambda f_i(x) + (1-\lambda) f_i(y)\} \\ &\le \lambda \cdot \max_{1 \le i \le n} f_i(x) + (1-\lambda) \cdot \max_{1 \le i \le n} f_i(y) \\ &= \lambda f(x) + (1-\lambda) f(y),\end{aligned}$$

where the first inequality follows from the convexity of the $f_i$'s. Thus, $f$ is convex on $\mathcal{V}$. □

**Theorem 2.1.13** (First-order condition)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a convex function and let $x \in \operatorname{int}(\operatorname{dom}(f))$. If $f$ is (Gâteaux) differentiable at $x$, then we have*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle, \tag{2.4}$$

*for all $y \in \mathbb{R}^d$.*

*Proof.* If $y \notin \operatorname{dom}(f)$, then we must have $f(y) = +\infty$. Clearly, Equation (2.4) holds if $f(y) = +\infty$. It remains to consider the case when $y \in \operatorname{dom}(f)$. Let $y \in \operatorname{dom}(f)$ and let $\lambda \in (0,1)$. By the zeroth-order condition (Theorem 2.1.8), we can write

$$f(x + \lambda(y-x)) - f(x) = f(\lambda y + (1-\lambda)x) - f(x) \le \lambda(f(y) - f(x)). \tag{2.5}$$

Since $f$ is (Gâteaux) differentiable at $x$, by dividing $\lambda$ from both sides of Equation (2.5) and letting $\lambda \downarrow 0$, we get

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

Rearranging, we obtain the desired inequality. This completes the proof. □

We can interpret the first-order condition (Theorem 2.1.13) as follows. Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a (Gâteaux) differentiable convex function. For each $x \in \operatorname{int}(\operatorname{dom}(f))$, let $\varphi_x : \mathbb{R}^d \to \mathbb{R}$ be defined as

$$\varphi_x(y) \coloneqq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall y \in \mathbb{R}^d.$$

Namely, $\varphi_x$ is the first-order Taylor approximation of $f$ at $x$. Recall from Example 2.1.11 that a function $g : \mathbb{R}^d \to \mathbb{R}$ is affine if and only if $g(\lambda x + (1-\lambda)y) = \lambda g(x) + (1-\lambda)g(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$. Let $y, z \in \mathbb{R}^d$ and let $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}\varphi_x(\lambda y + (1-\lambda)z) &= f(x) + \langle \nabla f(x), \lambda y + (1-\lambda)z - x \rangle \\ &= \lambda(f(x) + \langle \nabla f(x), y - x \rangle) + (1-\lambda)(f(x) + \langle \nabla f, z - x \rangle) \\ &= \lambda \varphi_x(y) + (1-\lambda)\varphi_x(z).\end{aligned}$$

Thus, $\varphi_x$ is an affine function. Therefore, for a convex (Gâteaux) differentiable function $f$, we have an "affine lower bound" at each point in the interior of the domain.

Recall the "Fermat's theorem" from basic real analysis.

**Theorem 2.1.14** (Fermat's theorem)**.** *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ attains an extremum at $x^\star \in \mathbb{R}^d$. If $f$ is (Gâteaux) differentiable at $x^\star$, then $\nabla f(x^\star) = 0$.*

**Theorem 2.1.15** (Optimality condition)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set, and let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a convex (Gâteaux) differentiable function on $\mathcal{V}$. Then, we have $x^\star \in \arg\min_{x\in\mathcal{V}} f(x)$ if and only if*

$$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0, \quad \forall x \in \mathcal{V}. \tag{2.6}$$

*Proof.* Let $x^\star \in \arg\min_{x\in\mathcal{V}} f(x)$. The proof is by contradiction. Suppose the contrary that there exists some $\tilde{x} \in \mathcal{V}$ such that

$$\langle \nabla f(x^\star), \tilde{x} - x^\star \rangle < 0.$$

Conversely, suppose that Equation (2.6) holds for some $x^\star \in \mathcal{V}$. Since $f$ is convex and (Gâteaux) differentiable at $x^\star$, by the first-order condition (Theorem 2.1.13), we have

$$f(x) \geq f(x^\star) + \langle \nabla f(x^\star), x - x^\star \rangle \geq f(x^\star), \quad \forall x \in \mathcal{V},$$

which shows that $x^\star \in \arg\min_{x\in\mathcal{V}} f(x)$. □

**Theorem 2.1.16** (Jensen's inequality)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a (measurable) convex function and let $\xi \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}[\xi]$ exists and $\xi \in \mathrm{dom}(f)$ with probability 1. Then, we have*

$$f(\mathbb{E}[\xi]) \leq \mathbb{E}[f(\xi)].$$

**Definition 2.1.17** (Euclidean projection)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed set. The **Euclidean projection** onto $\mathcal{V}$, denoted by $\mathrm{proj}_{\mathcal{V}} : \mathbb{R}^d \to \mathcal{V}$, is defined as*

$$\mathrm{proj}_{\mathcal{V}}(x) \in \operatorname*{arg\,min}_{y\in\mathcal{V}} \|y - x\|_2, \quad \forall x \in \mathbb{R}^d.$$

The Euclidean projection may not be well-defined for arbitrary closed set. However, if $\mathcal{V}$ is a nonempty closed convex set, then the Euclidean projection onto $\mathcal{V}$ is unique. This follows from the fact that a strongly convex function attains a unique minimizer over a closed convex set. This fact is proved in Theorem 3.1.14. For now, we assume that Euclidean projections are well-defined.

Online Gradient Descent takes as input a nonempty closed convex set $\mathcal{V} \subseteq \mathbb{R}^d$, an initial feasible point $x_1 \in \mathcal{V}$, and a sequence of step sizes $\eta_1, \ldots, \eta_T$. The Online Gradient Descent algorithm is as follows.

**Algorithm 3** Online Gradient Descent

**Inputs:** A nonempty closed convex set $\mathcal{V} \subseteq \mathbb{R}^d$, an initial point $x_1 \in \mathcal{V}$, and step sizes $\eta_1, \ldots, \eta_T > 0$.

1: **for** $t = 1, \ldots, T$ **do**
2: The learner outputs $x_t \in \mathcal{V}$.
3: The adversary chooses a convex differentiable loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
4: The learner calls the first-order oracle for $\nabla \ell_t(x_t)$.
5: The learner updates by setting

$$x_{t+1} \leftarrow \mathrm{proj}_{\mathcal{V}}(x_t - \eta_t \nabla \ell_t(x_t)).$$

6: **end for**

**Lemma 2.1.18.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. Then, we have*

$$\|\mathrm{proj}_{\mathcal{V}}(x) - y\|_2 \leq \|x - y\|_2, \quad \forall x \in \mathbb{R}^d, y \in \mathcal{V}.$$

*Proof.* Observe that

$$\mathrm{proj}_{\mathcal{V}}(x) = \operatorname*{arg\,min}_{y\in\mathcal{V}} \|y - x\|_2 = \operatorname*{arg\,min}_{y\in\mathcal{V}} \frac{1}{2}\|y - x\|_2^2.$$

Since $\mathcal{V}$ is a closed and convex set and $\|y - x\|_2^2/2$ is differentiable and convex, by the optimality condition (Theorem 2.1.15), we have

$$\langle \mathrm{proj}_{\mathcal{V}}(x) - x, y - \mathrm{proj}_{\mathcal{V}}(x) \rangle \geq 0, \quad \forall y \in \mathcal{V}. \tag{2.7}$$

Then, we can write

$$\begin{aligned}
\|x - y\|_2^2 &= \|x - \mathrm{proj}_{\mathcal{X}}(x) + \mathrm{proj}_{\mathcal{X}}(x) - y\|_2^2 \\
&= \|x - \mathrm{proj}_{\mathcal{X}}(x)\|_2^2 + \|\mathrm{proj}_{\mathcal{X}}(x) - y\|_2^2 + 2\langle x - \mathrm{proj}_{\mathcal{X}}(x), \mathrm{proj}_{\mathcal{X}}(x) - y \rangle \\
&\geq \|x - \mathrm{proj}_{\mathcal{X}}(x)\|_2^2 + \|\mathrm{proj}_{\mathcal{X}}(x) - y\|_2^2 \\
&\geq \|\mathrm{proj}_{\mathcal{X}}(x) - y\|_2^2,
\end{aligned}$$

where the first inequality follows from Equation (2.7). The lemma follows by taking square root on both sides of the above inequality. □

In words, Lemma 2.1.18 formally verifies our intuition that the distance between $x \in \mathbb{R}^d$ and $y \in \mathcal{V}$ must decrease after projecting $x$ onto $\mathcal{V}$.

**Lemma 2.1.19.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set, and let $\ell_t : \mathcal{V} \to \mathbb{R}$ be a convex differentiable function. Then, with the updating rule of* Online Gradient Descent*, we have*

$$\eta_t(\ell_t(x_t) - \ell_t(u)) \leq \eta_t \langle \nabla \ell_t(x_t), x_t - u \rangle \leq \frac{1}{2}\|x_t - u\|_2^2 - \frac{1}{2}\|x_{t+1} - u\|_2^2 + \frac{\eta_t^2}{2}\|\nabla \ell_t(x_t)\|_2^2, \quad \forall u \in \mathcal{V}.$$

*Proof.* Let $u \in \mathcal{V}$. Since $\ell_t : \mathcal{V} \to \mathbb{R}$ is convex and differentiable, by the first-order condition (Theorem 2.1.13), we can write

$$\ell_t(u) \geq \ell_t(x_t) + \langle \nabla \ell_t(x_t), u - x_t \rangle .$$

Multiply both sides of the above inequality by $\eta_t > 0$ and rearrange the resulting expression, we have

$$\eta_t(\ell_t(x_t) - \ell_t(u)) \leq \eta_t \langle \nabla \ell_t(x_t), x_t - u \rangle . \tag{2.8}$$

Since $x_{t+1} = \text{proj}_{\mathcal{V}}(x_t - \eta_t \nabla \ell_t(x_t))$, from Lemma 2.1.18, we know that

$$\|x_{t+1} - u\|_2 \leq \|x_t - \eta_t \nabla \ell_t(x_t) - u\|_2. \tag{2.9}$$

By squaring both sides of Equation (2.9), we get

$$\|x_{t+1} - u\|_2^2 \leq \|x_t - \eta_t \nabla \ell_t(x_t) - u\|_2^2 = \|x_t - u\|_2^2 + \eta_t^2 \|\nabla \ell_t(x_t)\|_2^2 - 2\eta_t \langle \nabla \ell_t(x_t), x_t - u \rangle .$$

Rearranging, we get

$$\eta_t \langle \nabla \ell_t(x_t), x_t - u \rangle \leq \frac{1}{2}\|x_t - u\|_2^2 - \frac{1}{2}\|x_{t+1} - u\|_2^2 + \frac{\eta_t^2}{2}\|\nabla \ell_t(x_t)\|_2^2. \tag{2.10}$$

The lemma follows by combining Equation (2.9) and Equation (2.10). □

The difference $\ell_t(x_t) - \ell_t(u)$ is called the **instantaneous regret** or the **per-round regret** compared to the competitor $u \in \mathcal{V}$. The regret of an online learning algorithm is simply the sum of the instantaneous regret over all rounds. Lemma 2.1.19 gives us a bound on the instantaneous regret of Online Gradient Descent. With this bound at hand, the regret guarantee of Online Gradient Descent can be proved easily as follows.

**Theorem 2.1.20** (Zinkevich [Zin03])**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set with diameter $D > 0$, i.e., $\sup_{x,y \in \mathcal{X}} \|x - y\|_2 = D$. Let $\ell_1, \ldots, \ell_T : \mathcal{V} \to \mathbb{R}$ be an arbitrary sequence of convex differentiable functions. Let $x_1 \in \mathcal{V}$ and let $\eta_1, \ldots, \eta_T$ be positive constants such that $\eta_{t+1} \leq \eta_t$ for all $t$. Then, we have*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta_T}, \quad \forall u \in \mathcal{V}. \tag{2.11}$$

*Moreover, if $\eta_t \coloneqq \eta > 0$ for $t = 1, \ldots, T$, we have*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla \ell_t(x_t)\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta}, \quad \forall u \in \mathcal{V}. \tag{2.12}$$

*Proof.* Let $u \in \mathcal{V}$. From Lemma 2.1.19, we have

$$\eta_t(\ell_t(x_t) - \ell_t(u)) \leq \frac{1}{2}\|x_t - u\|_2^2 - \frac{1}{2}\|x_{t+1} - u\|_2^2 + \frac{\eta_t^2}{2}\|\nabla \ell_t(x_t)\|_2^2, \tag{2.13}$$

for $t = 1, \ldots, T$. By dividing $\eta_t > 0$ from both sides of Equation (2.13) and summing up the resulting expression over $t = 1, \ldots, T$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\leq \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2\right) + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2 \\
&= \frac{\|x_1 - u\|_2^2}{2\eta_1} + \sum_{t=1}^{T-1}\left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right)\|x_{t+1} - u\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2 \\
&\leq \frac{D^2}{2\eta_1} + D^2\sum_{t=1}^{T-1}\left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right) + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta_T} \\
&= \frac{D^2}{2\eta_1} + D^2\left(\frac{1}{2\eta_T} - \frac{1}{2\eta_1}\right) + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta_T} \\
&= \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta_T},
\end{aligned}$$

where the second inequality follows from the bounded domain assumption. This proves Equation (2.11).

Now, suppose we have $\eta_t \coloneqq \eta$ for some $\eta > 0$ for $t = 1, \dots, T$, then Equation (2.13) becomes

$$\eta(\ell_t(x_t) - \ell_t(u)) \leq \frac{1}{2}\|x_t - u\|_2^2 - \frac{1}{2}\|x_{t+1} - u\|_2^2 + \frac{\eta^2}{2}\|\nabla \ell_t(x_t)\|_2^2. \tag{2.14}$$

By dividing $\eta > 0$ from both sides of Equation (2.14) and summing up the resulting expression over $t = 1, \dots, T$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\leq \frac{1}{2\eta}\sum_{t=1}^{T}\left(\|x_t - u\|_2^2 - \|x_{t+1} - u\|_2^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla \ell_t(x_t)\|_2^2 \\ &= \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla \ell_t(x_t)\|_2^2 - \frac{\|x_{T+1} - u\|_2^2}{2\eta}.\end{aligned}$$

This proves Equation (2.12). □

**Remark 2.1.21.** Note that the assumption that $\mathcal{V}$ is bounded is not used to prove the regret bound of Online Gradient Descent with fixed learning rate (Equation (2.12)). On the contrary, the proof of the regret bound for time-varying learning rates (Equation (2.11)) relies heavily on the assumption that $\mathcal{V}$ is bounded.

Observe that the last term in both of the regret bounds given in Theorem 2.1.20 are nonpositive. Thus, we usually discard them and write

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\nabla \ell_t(x_t)\|_2^2, \quad \forall u \in \mathcal{V},$$

and

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla \ell_t(x_t)\|_2^2, \quad \forall u \in \mathcal{V}.$$

How should we choose the step size sequence $\{\eta_t\}_{t=1}^T$ for Online Gradient Descent? Let us consider the case for fixed step size. One simple strategy is to find the optimal fixed step size $\eta > 0$ for a fixed time horizon $T$. Theorem 2.1.20 gives us a clue to decide the step sizes. If we further assume that $\|\nabla \ell_t(x_t)\|_2 \leq L$ for some constant $L > 0$ for $t = 1, \dots, T$ (i.e., bounded gradients), then Online Gradient Descent achieves

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{D^2}{2\eta} + \frac{\eta}{2}TL^2, \quad \forall u \in \mathcal{V}. \tag{2.15}$$

Let $\eta^\star$ denote the optimal fixed step size. We find the optimal fixed step size $\eta^\star$ by optimizing over $\eta$ on the RHS of the above inequality. Let $\varphi$ be defined as $\varphi(\eta) \coloneqq D^2/(2\eta) + \eta T L^2/2$ for $\eta > 0$. Differentiate with respect to $\eta$, we get

$$\varphi'(\eta) = \frac{-D^2}{2\eta^2} + \frac{TL^2}{2}, \quad \varphi''(\eta) = \frac{D^2}{\eta^3}.$$

Since $\varphi''(\eta) > 0$ for all $\eta$, we know that $\varphi$ is a convex function. We know that $\eta^\star$ must satisfy $\varphi'(\eta^\star) = 0$. Thus, we have $\eta^\star = D/(L\sqrt{T})$. Substitute $\eta^\star$ into Equation (2.15) gives

$$\text{Regret}_T = \sum_{t=1}^{T}\ell_t(x_t) - \min_{x \in \mathcal{V}}\sum_{t=1}^{T}\ell_t(x) \leq DL\sqrt{T}.$$

In conclusion, Online Gradient Descent gives a $O(\sqrt{T})$ regret bound. This regret bound was first obtained by Zinkevich [Zin03].

## 2.2 Online Subgradient Descent

In order to run Online Gradient Descent, the loss functions must be convex and differentiable on some open set containing the feasible set. This assumption is quite strong, and often not met in practice. For example, the *hinge loss*

$$\ell_{\text{hinge}}(w) \coloneqq \max\{1 - y\langle x, w\rangle, 0\}, \quad \forall w \in \mathbb{R}^d,$$

where $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, is clearly not differentiable. In this section, we extend Online Gradient Descent to convex losses that are not necessarily differentiable. This extension is fairly natural and straightforward. It turns out that we can simply replace the gradients in Online Gradient Descent with *subgradients*. We begin by introducing more notions from convex analysis.

**Definition 2.2.1** (Proper function)**.** *An extended real-valued function $f : \mathbb{R}^d \to [-\infty, +\infty]$ is* ***proper*** *if $\operatorname{dom}(f)$ is nonempty and $f(x) > -\infty$ for all $x \in \mathbb{R}^d$. A function is* ***improper*** *if it is not proper.*

For example, the indicator function (Example 2.1.5) $\iota_{\mathcal{V}} : \mathbb{R}^d \to (-\infty, +\infty]$ is proper if and only if $\mathcal{V}$ is nonempty.

Let's see why properness is an important assumption in optimization. Consider an extended real-valued function $f_1 : \mathbb{R}^d \to [-\infty, +\infty]$ defined as $f_1 \equiv +\infty$. We have $\operatorname{epi}(f_1) = \varnothing$. Thus, $\operatorname{epi}(f_1)$ is convex and $f_1$ is a convex function. Consider another extended real-valued function $f_2 : \mathbb{R}^d \to [-\infty, +\infty]$ defined as $f_2 \equiv -\infty$. We have $\operatorname{epi}(f_2) = \mathbb{R}^{d+1}$. Thus, $\operatorname{epi}(f_2)$ is convex and $f_2$ is a convex function. Although $f_1$ and $f_2$ are convex functions, optimizing over them is meaningless. Note that a function is improper if it is either identically $+\infty$ or identically $-\infty$. Therefore, properness is introduced exactly to exclude these cases where the function is technically convex but meaningless for optimization.

**Proposition 2.2.2.** *The following statements hold.*

*(a) If $f : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and $\alpha > 0$ is some constant, then $\alpha f$ is proper.*

*(b) If $f, g : \mathbb{R}^d \to (-\infty, +\infty]$ are proper functions such that $\operatorname{dom}(f) \cap \operatorname{dom}(g)$ is nonempty, then $f + g$ is proper.*

*Proof of (a).* Note that we have $\operatorname{dom}(\alpha f) = \operatorname{dom}(f)$. Since $\operatorname{dom}(f)$ is nonempty, it follows that $\operatorname{dom}(\alpha f)$ is also nonempty. Since $f(x) > -\infty$ for all $x \in \mathbb{R}^d$ and $\alpha > 0$, we clearly have $(\alpha f)(x) > -\infty$ for all $x \in \mathbb{R}^d$. Therefore, $\alpha f$ is proper. □

*Proof of (b).* Note that we have $\operatorname{dom}(f) = \operatorname{dom}(g) \subseteq \operatorname{dom}(f + g)$. Since $\operatorname{dom}(f) \cap \operatorname{dom}(g)$ is assumed to be nonempty, it follows that $\operatorname{dom}(f + g)$ is nonempty. Since $f(x) > -\infty$ and $g(x) > -\infty$ for all $x \in \mathbb{R}^d$, we clearly have $(f + g)(x) = f(x) + g(x) > -\infty$ for all $x \in \mathbb{R}^d$. Therefore, $f + g$ is proper if $\operatorname{dom}(f) \cap \operatorname{dom}(g)$ is nonempty. □

Note that the sum of proper functions may not be proper. For example, let $\mathcal{V}_1, \mathcal{V}_2 \subseteq \mathbb{R}^d$ be nonempty sets such that $\mathcal{V}_1 \cap \mathcal{V}_2 = \varnothing$. Since $\mathcal{V}_1$ and $\mathcal{V}_2$ are nonempty, we know that the indicator functions $\iota_{\mathcal{V}_1}$ and $\iota_{\mathcal{V}_2}$ are proper. However, we have $\iota_{\mathcal{V}_1}(x) + \iota_{\mathcal{V}_2}(x) = +\infty$ for all $x \in \mathbb{R}^d$. Thus, $\operatorname{dom}(\iota_{\mathcal{V}_1} + \iota_{\mathcal{V}_2}) = \varnothing$ and $\iota_{\mathcal{V}_1} + \iota_{\mathcal{V}_2}$ is improper.

**Definition 2.2.3** (Subgradient, subdifferential)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper function and let $x \in \mathbb{R}^d$. A vector $g_x \in \mathbb{R}^d$ is called a* ***subgradient*** *of $f$ at $x$ if and only if*

$$f(y) \geq f(x) + \langle g_x, y - x \rangle, \quad \forall y \in \mathbb{R}^d.$$

*If such $g_x$ exists, then $f$ is said to be* ***subdifferentiable*** *at $x$. The set of all subgradients of $f$ at $x$ is called the* ***subdifferential*** *of $f$ at $x$, and is denoted by $\partial f(x)$. The* ***domain*** *of $\partial f$, denoted by $\operatorname{dom}(\partial f)$, is defined as*

$$\operatorname{dom}(\partial f) \coloneqq \{x \in \mathbb{R}^d : \partial f(x) \neq \varnothing\}.$$

In other words, a vector $g_x \in \mathbb{R}^d$ is a subgradient of $f : \mathbb{R}^d \to (-\infty, +\infty]$ at $x$ if and only if the function $\varphi_x(y) \coloneqq f(x) + \langle g_x, y - x \rangle$ is an affine lower bound of $f$. Note that convexity is not assumed in the definition of subgradients.

**Example 2.2.4** (Normal cone)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty convex set, let $\iota_{\mathcal{V}} : \mathbb{R}^d \to (-\infty, +\infty]$ be the indicator function of $\mathcal{V}$, and let $x \in \mathcal{V}$. If $g \in \partial \iota_{\mathcal{V}}(x)$, then we have*

$$\langle g, y - x \rangle \leq 0, \quad \forall y \in \mathcal{V}.$$

*The set $N_{\mathcal{V}}(x) \coloneqq \{g \in \mathbb{R}^d : \langle g, y - x \rangle \leq 0\}$ is called the* ***normal cone*** *to $\mathcal{V}$ at $x$.*

*Proof.* Let $g \in \partial \iota_{\mathcal{V}}(x)$. Then, for all $y \in \mathbb{R}^d$, we have $\iota_{\mathcal{V}}(y) \geq \iota_{\mathcal{V}}(x) + \langle g, y - x \rangle$. In particular, we have $0 \geq \langle g, y - x \rangle$ for all $y \in \mathcal{V}$. □

A normal is indeed a *cone*. A subset $\mathcal{K} \subseteq \mathbb{R}^d$ is a **cone** if we have $\lambda x \in \mathcal{K}$ for all $x \in \mathcal{K}$ and $\lambda > 0$. Namely, a cone is a set that is closed under positive scalar multiplication. Clearly, the normal cone $N_{\mathcal{V}}(x)$ is closed under positive scalar multiplication.

Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty convex set. Note that we have $N_{\mathcal{V}}(x) = \{0\}$ if $x \in \operatorname{int}(\mathcal{V})$. Let $x \in \operatorname{int}(\mathcal{V})$ and let $g \in N_{\mathcal{V}}(x)$, where $g \neq 0$. Since $N_{\mathcal{V}}(x)$ is a cone, we know that $g' \coloneqq g/\|g\|_2 \in N_{\mathcal{V}}(x)$. Since $x \in \operatorname{int}(\mathcal{V})$, there exists some $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq \mathcal{V}$, where $B_\varepsilon(x) \coloneqq \{y \in \mathbb{R}^d : \|y - x\|_2 < \varepsilon\}$ is the open ball centered at $x$ with radius $\varepsilon$. Since $\|g'\|_2 = 1$, we have $x + \varepsilon g' \in \mathcal{V}$. Then, we can write

$$\langle g', (x + \varepsilon g') - x \rangle \leq 0.$$

The above inequality implies that $g = 0$, which is a contradiction. We conclude that $N_{\mathcal{V}}(x) = \{0\}$ for all $x \in \operatorname{int}(\mathcal{V})$.

**Proposition 2.2.5.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper function. Then,* $\mathrm{dom}(\partial f) \subseteq \mathrm{dom}(f)$.

*Proof.* For $x \notin \mathrm{dom}(f)$, we have $f(x) = +\infty$. Thus, for any $y \in \mathrm{dom}(f)$ and any $g_x \in \mathbb{R}^d$, we would have

$$f(y) < f(x) + \langle g_x, y - x \rangle = +\infty.$$

This shows that $\partial f(x) = \varnothing$. It follows that $\mathrm{dom}(\partial f) \subseteq \mathrm{dom}(f)$. □

**Proposition 2.2.6.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper function and let $x \in \mathbb{R}^d$. The subdifferential $\partial f(x)$ is a convex set in $\mathbb{R}^d$.*

*Proof.* If $\partial f(x) = \varnothing$, then $\partial f(x)$ is clearly a convex set. Suppose that $\partial f(x)$ is nonempty. Let $g_1, g_2 \in \partial f(x)$ and let $\lambda \in [0, 1]$. We have

$$f(y) \geq f(x) + \langle g_1, y - x \rangle, \quad \forall y \in \mathbb{R}^d, \tag{2.16}$$

and

$$f(y) \geq f(x) + \langle g_2, y - x \rangle, \quad \forall y \in \mathbb{R}^d. \tag{2.17}$$

Summing up the resulting expressions after multiplying Equation (2.16) by $\lambda$ and Equation (2.17) by $1 - \lambda$ yields

$$f(y) \geq f(x) + \langle \lambda g_1 + (1 - \lambda) g_2, y - x \rangle, \quad \forall y \in \mathbb{R}^d,$$

which shows that $\lambda g_1 + (1 - \lambda) g_2 \in \partial f(x)$. Therefore, $\partial f(x)$ is convex. □

**Theorem 2.2.7.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex and let $f : \mathbb{R}^d \to \mathbb{R}$. If $\partial f(x)$ is nonempty for all $x \in \mathcal{V}$, then $f$ is convex on $\mathcal{V}$.*

*Proof.* Suppose that $\partial f(x)$ is nonempty for all $x \in \mathcal{V}$. Let $x, y \in \mathcal{V}$, let $\lambda \in [0, 1]$, and let $z \coloneqq \lambda x + (1 - \lambda) y$. Since $\mathcal{V}$ is convex, we know that $z \in \mathcal{V}$ and thus $\partial f(z)$ is nonempty by assumption. Let $g_z \in \partial f(z)$. We can write

$$f(x) \geq f(z) + \langle g_z, x - z \rangle = f(z) + (1 - \lambda) \langle g_z, x - y \rangle, \tag{2.18}$$

and

$$f(y) \geq f(z) + \langle g_z, y - z \rangle = f(z) + \lambda \langle g_z, y - x \rangle. \tag{2.19}$$

Summing up the resulting expressions after multiplying Equation (2.18) by $\lambda$ and Equation (2.19) by $1 - \lambda$ yields

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(z) = f(\lambda x + (1 - \lambda) y).$$

This shows that $f$ is convex on $\mathcal{V}$. □

**Theorem 2.2.8** (Rockafellar [Roc15], Theorem 23.4)**.** *A proper convex function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is subdifferentiable on* $\mathrm{int}(\mathrm{dom}(f))$.

**Theorem 2.2.9.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper convex function and let $x \in \mathrm{int}(\mathrm{dom}(f))$. If $f$ is (Gâteaux) differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$. Conversely, if $\partial f(x) = \{g_x\}$ for some $g_x \in \mathbb{R}^d$, then $f$ is (Gâteaux) differentiable at $x$ and $g_x = \nabla f(x)$.*

*Proof.* Suppose that $f$ is differentiable at $x \in \mathrm{int}(\mathrm{dom}(f))$. Since $f$ is convex and differentiable at $x$, by the first-order condition (Theorem 2.1.13), we can write

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall y \in \mathbb{R}^d.$$

This shows that $\nabla f(x) \in \partial f(x)$. Next, we prove that $\partial f(x)$ is singleton. Let $g_x \in \partial f(x)$. We can write

$$f(x + \lambda y) - f(x) \geq \lambda \langle g_x, y \rangle, \quad \forall \lambda \in \mathbb{R}_{++}, y \in \mathbb{R}^d. \tag{2.20}$$

Since $f$ is differentiable at $x$, by dividing $\lambda > 0$ from both sides of Equation (2.20) and letting $\lambda \downarrow 0$, we get

$$\langle \nabla f(x) - g_x, y \rangle \geq 0, \quad \forall y \in \mathbb{R}^d. \tag{2.21}$$

Substitute $y = g_x - f(x)$ into Equation (2.21), we get $\|\nabla f(x) - g_x\|_2^2 \leq 0$. It follows that $g_x = \nabla f(x)$. Therefore, the subgradient of $f$ at $x$ is unique and we conclude that $\partial f(x) = \{\nabla f(x)\}$. □

**Proposition 2.2.10.** *Let $f_1, \ldots, f_m : \mathbb{R}^d \to (-\infty, +\infty]$ be proper functions, and let $f \coloneqq f_1 + \cdots + f_m$. Then, we have*

$$\partial f_1(x) + \cdots + \partial f_m(x) \subseteq \partial f(x), \quad \forall x \in \mathbb{R}^d.$$

*Proof.* Let $x \in \mathbb{R}^d$ be fixed, and let $g_i \in \partial f_i(x)$ for $i = 1, \ldots, m$. For each $i$, we have

$$f_i(y) \geq f_i(x) + \langle g_i, y - x \rangle, \quad \forall y \in \mathbb{R}^d. \tag{2.22}$$

Summing up Equation (2.22) over $i = 1, \ldots, m$ yields

$$f(y) = \sum_{i=1}^m f_i(y) \geq \sum_{i=1}^m f_i(x) + \sum_{i=1}^m \langle g_i, y - x \rangle = f(x) + \left\langle \sum_{i=1}^m g_i, y - x \right\rangle, \quad \forall y \in \mathbb{R}^d. \tag{2.23}$$

Equation (2.23) implies that $\sum_{i=1}^m g_i \in \partial f(x)$. Since the $g_i$'s are arbitrary, we conclude that $\partial f_1(x) + \cdots + \partial f_m(x) \subseteq \partial f(x)$ for all $x \in \mathbb{R}^d$. □

**Proposition 2.2.11.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper. Define $h(x) \coloneqq f(Ax + b)$, where $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$. Then, we have*

$$A^\top \partial f(Ax + b) \subseteq \partial h(x), \quad \forall x \in \mathbb{R}^m,$$

*where $A^\top \partial f(Ax + b) = \{A^\top g : g \in \partial f(Ax + b)\}$.*

*Proof.* Let $x \in \mathbb{R}^m$ and let $g \in \partial f(Ax + b)$. Since $g \in \partial f(Ax + b)$, we have

$$h(y) = f(Ay + b) \geq f(Ax + b) + \langle g, A(y - x) \rangle = f(Ax + b) = \left\langle A^\top g, y - x \right\rangle = h(x) + \left\langle A^\top g, y - x \right\rangle,$$

for all $y \in \mathbb{R}^m$. Since $g \in \partial f(Ax + b)$ is arbitrary, we conclude that $A^\top \partial f(Ax + b) \subseteq \partial h(x)$ for all $x \in \mathbb{R}^m$. □

**Definition 2.2.12** (Closed function)**.** *An extended real-valued function $f : \mathbb{R}^d \to [-\infty, +\infty]$ is **closed** if and only if $\{x \in \mathbb{R}^d : f(x) \leq \alpha\}$ is closed for all $\alpha \in \mathbb{R}$.*

In words, an extended real-valued function $f$ is closed if and only if each *sublevel set* is closed. Let us see why closedness is an important assumption in optimization. Consider an extended real-valued function $f : \mathbb{R} \to [-\infty, +\infty]$ defined as

$$f(x) \coloneqq \begin{cases} x, & \text{if } x > 0; \\ +\infty, & \text{otherwise.} \end{cases}$$

We have $\operatorname{dom}(f) = \mathbb{R}_{++}$, so $\operatorname{dom}(f)$ is convex and $f$ is proper. From the zeroth-order condition (Theorem 2.1.8), we can see that $f$ is convex. Thus, $f$ is a proper convex function. However, $f$ is not closed. This is because the sublevel set $\{x \in \mathbb{R} : f(x) \leq 1\} = (0, 1]$ is not closed. Although we have $\inf_{x \in \mathbb{R}^d} f(x) = 0$, $f$ does not attain its minimum over $\mathbb{R}$. By assuming closedness, the minimizers (provided they exist) of the objective function must lie in the feasible set. Therefore, in optimization theory, the objective function is usually assumed to be closed so that the minimizers are indeed contained in the feasible set.

**Example 2.2.13** (Indicator function)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$. The indicator function $\iota_{\mathcal{V}} : \mathbb{R}^d \to (-\infty, +\infty]$ is closed if and only if $\mathcal{V}$ is closed.*

*Proof.* Recall from Example 2.1.5 that the indicator function $\iota_{\mathcal{V}}$ is defined as $\iota_{\mathcal{V}}(x) = 0$ for all $x \in \mathcal{V}$ and $\iota_{\mathcal{V}}(x) = +\infty$ for all $x \notin \mathcal{V}$. Denote by $\operatorname{lev}_{\leq \alpha}(\iota_{\mathcal{V}})$ the sublevel set of $\iota_{\mathcal{V}}$ with parameter $\alpha \in \mathbb{R}$, that is, $\operatorname{lev}_{\leq \alpha}(\iota_{\mathcal{V}}) \coloneqq \{x \in \mathbb{R}^d : \iota_{\mathcal{V}}(x) \leq \alpha\}$. We have

$$\operatorname{lev}_{\leq \alpha}(\iota_{\mathcal{V}}) = \begin{cases} \varnothing, & \text{if } \alpha < 0; \\ \mathcal{V}, & \text{otherwise.} \end{cases}$$

Since the empty set is closed, we conclude that $\iota_{\mathcal{V}}$ is closed if and only if $\mathcal{V}$ is closed. □

**Theorem 2.2.14** (Fermat's rule)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper function. Then, we have $x^\star \in \arg\min_{x \in \mathbb{R}^d} f(x)$ if and only if $0 \in \partial f(x^\star)$.*

*Proof.* We have $x^\star \in \arg\min_{x \in \mathbb{R}^d} f(x)$ if and only if $f(x) \geq f(x^\star)$ for all $x \in \mathbb{R}^d$, if and only if

$$f(x) \geq f(x^\star) + \langle 0, x - x^\star \rangle, \quad \forall x \in \mathbb{R}^d,$$

if and only if $0 \in \partial f(x^\star)$. □

Theorem 2.2.14 is called "Fermat's rule" due to its resemblance to the Fermat's theorem (Theorem 2.1.14). The Fermat's rule is a strict generalization of the Fermat's theorem since it does not require $f$ to be differentiable.

**Remark 2.2.15.** Although $f$ is not assumed to be closed in the Fermat's rule (Theorem 2.2.14), we actually need $f$ to be closed for a meaningful result. This is because if $f$ is not closed, then the set $\arg\min_{x \in \mathbb{R}^d} f(x)$ could be empty. In this case, Fermat's rule holds vacuously.

**Definition 2.2.16** (Lipschitz continuity). *Let $\mathcal{V} \subseteq \mathbb{R}^d$. A function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is $L$-**Lipschitz continuous** with respect to $\|\cdot\|$ over $\mathcal{V}$ if and only if*

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{V},$$

*for some constant $L > 0$. The constant $L$ is called the **Lipschitz parameter**.*

**Theorem 2.2.17.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper convex function. Then, $f$ is $L$-Lipschitz continuous with respect to $\|\cdot\|$ over $\operatorname{int}(\operatorname{dom}(f))$ if and only if*

$$\|g_x\|_* \leq L, \quad \forall g_x \in \partial f(x), x \in \operatorname{int}(\operatorname{dom}(f)). \tag{2.24}$$

*Proof.* Suppose that Equation (2.24) holds. Since $f$ is proper and convex, from Theorem 2.2.8, we know that $f$ is subdifferentiable on $\operatorname{int}(\operatorname{dom}(f))$. Let $x, y \in \operatorname{int}(\operatorname{dom}(f))$ and let $g_x \in \partial f(x)$. Then, we can write

$$f(x) - f(y) \leq \langle g_x, x - y\rangle \leq \|g_x\|_*\|x - y\| \leq L\|x - y\|, \tag{2.25}$$

where the second inequality follows from the Hölder's inequality, and the last inequality follows from Equation (2.24). Similarly, let $g_y \in \partial f(y)$, we have

$$f(y) - f(x) \leq \langle g_y, y - x\rangle \leq \|g_y\|_*\|y - x\| \leq L\|y - x\|. \tag{2.26}$$

Equation (2.25) and Equation (2.26) together imply that

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

Therefore, $f$ is $L$-Lipschitz continuous with respect to $\|\cdot\|$ over $\operatorname{int}(\operatorname{dom}(f))$. □

We are now ready to introduce the Online Subgradient Descent algorithm. As mentioned earlier, Online Subgradient Descent is basically the same as Online Gradient Descent. The only difference is that we use subgradients instead of gradients in Online Subgradient Descent. The Online Subgradient Descent algorithm is as follows.

**Algorithm 4** Online Subgradient Descent

**Inputs:** A nonempty closed convex set $\mathcal{V} \subseteq \mathbb{R}^d$, an initial point $x_1 \in \mathcal{V}$, and step sizes $\eta_1, \ldots, \eta_T > 0$.

1: **for** $t = 1, \ldots, T$ **do**
2: The learner outputs $x_t \in \mathcal{V}$.
3: The adversary chooses a convex subdifferentiable loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
4: The learner calls the first-order oracle for $g_t \in \partial \ell_t(x_t)$.
5: The learner updates by setting

$$x_{t+1} \leftarrow \operatorname{proj}_{\mathcal{V}}(x_t - \eta_t g_t).$$

6: **end for**

**Lemma 2.2.18.** *Let $\mathcal{V}$ be a nonempty closed convex subset of $\mathbb{R}^d$, and let $\ell_t : \mathcal{V} \to \mathbb{R}$ be a convex subdifferentiable function. Then, with the updating rule of* Online Subgradient Descent, *we have*

$$\eta_t(\ell_t(x_t) - \ell_t(u)) \leq \eta_t \langle g_t, x_t - u\rangle \leq \frac{1}{2}\|x_t - u\|_2^2 - \frac{1}{2}\|x_{t+1} - u\|_2^2 + \frac{\eta_t^2}{2}\|g_t\|_2^2, \quad \forall u \in \mathcal{V}.$$

*Proof.* Omitted. The proof is basically the same as Lemma 2.1.19. □

**Theorem 2.2.19** (Zinkevich [Zin03]). *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex subset with diameter $D > 0$, i.e., $\sup_{x,y\in\mathcal{V}} \|x - y\|_2 = D$. Let $\ell_1, \ldots, \ell_T : \mathcal{V} \to \mathbb{R}$ be an arbitrary sequence of convex subdifferentiable functions. Let $x_1 \in \mathcal{V}$ and let $\eta_1, \ldots, \eta_T$ be positive constants such that $\eta_{t+1} \leq \eta_t$ for all $t$. Then, we have*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{D^2}{2\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{2}\|g_t\|_2^2, \quad \forall u \in \mathcal{V}. \tag{2.27}$$

*Moreover, if $\eta_t := \eta > 0$ for $t = 1, \ldots, T$, we have*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \leq \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2, \quad \forall u \in \mathcal{V}. \tag{2.28}$$

*Proof.* Omitted. The proof is basically the same as Theorem 2.1.20. □

In both of the regret analysis of Online Gradient Descent (Theorem 2.1.20) and Online Subgradient Descent (Theorem 2.2.19), we rely heavily on the following inequality:

$$\ell_t(x_t) - \ell_t(u) \leq \langle g_t, x_t - u \rangle, \quad \forall u \in \mathcal{V},$$

where $g_t \in \partial \ell_t(x_t)$. Summing up the above inequality over $t = 1, \ldots, T$ gives

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \leq \sum_{t=1}^{T} \langle g_t, x_t - u \rangle, \quad \forall u \in \mathcal{V}.$$

Thus, to upper bound the regret of an online learning algorithm with convex subdifferentiable losses $\ell_1, \ldots, \ell_T : \mathcal{V} \to \mathbb{R}$, it is enough to bound the regret with respect to the "linearized losses" defined as $\tilde{\ell}_t(x) = \langle g_t, x \rangle$, where $g_t \in \partial \ell_t(x_t)$ for $t = 1, \ldots, T$.

## 2.3 Questions

**Question 2.1.** In Theorem 2.1.20, we are able to obtain a regret bound for adaptive step size sequence if we assume the feasible set is bounded, and a regret bound without the bounded domain assumption if the step sizes are fixed. Is it possible to obtain a regret bound for Online Gradient Descent with adaptive step sizes without the bounded domain assumption?

# 3 Beyond $\sqrt{T}$ Regret

## 3.1 Strong Convexity

**Definition 3.1.1** (Strong convexity)**.** *A proper function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is said to be $\mu$-**strongly convex** with respect to $\|\cdot\|$ if we have*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2, \quad \forall x, y \in \mathrm{dom}(f), \lambda \in [0,1], \tag{3.1}$$

*for some constant $\mu > 0$. The constant $\mu$ is called the **strong convexity parameter**. We say that $f$ is **strongly convex** if there exists some $\mu > 0$ and some norm $\|\cdot\|$ such that* Equation (3.1) *holds.*

In view of the zeroth-order condition for convex functions (Theorem 2.1.8), we can see that a convex function is 0-strongly convex and that a strongly convex function is convex.

**Definition 3.1.2** (Strong convexity)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex and let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be an extended real-valued function. We say that $f$ is $\mu$-**strongly convex** with respect to $\|\cdot\|$ on $\mathcal{V}$ if and only if*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2, \quad \forall x, y \in \mathcal{V}, \lambda \in [0,1], \tag{3.2}$$

*for some constant $\mu > 0$. We say that $f$ is **strongly convex** on $\mathcal{V}$ if there exists some $\mu > 0$ and some norm $\|\cdot\|$ such that Equation (3.2) holds.*

Recall from Proposition 2.1.12 that the class of convex functions is closed under operations such as positive scalar multiplication and addition. The class of strongly convex functions (with respect to the same norm $\|\cdot\|$) also enjoys this property.

**Proposition 3.1.3.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex. Suppose that $f : \mathbb{R}^d \to (-\infty, +\infty]$ is $\mu_1$-strongly convex and $g : \mathbb{R}^d \to (-\infty, +\infty]$ is $\mu_2$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$.*

*(a) The sum $f + g$ is $(\mu_1 + \mu_2)$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$.*

*(b) If $\alpha > 0$, then $\alpha f$ is $\alpha\mu_1$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$.*

*(c) If $\mu_1'$ is some constant such that $0 < \mu_1' < \mu$, then $f$ is $\mu_1'$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$.*

*Proof of (a).* Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. We can write

$$\begin{aligned}(f+g)(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &\leq \lambda(f(x)+g(x)) + (1-\lambda)(f(y)+g(y)) - \frac{\mu_1+\mu_2}{2}\lambda(1-\lambda)\|x-y\|^2 \\ &= \lambda(f+g)(x) + (1-\lambda)(f+g)(y) - \frac{\mu_1+\mu_2}{2}\lambda(1-\lambda)\|x-y\|^2.\end{aligned}$$

Since $x, y \in \mathcal{V}$ and $\lambda \in [0,1]$ are arbitrary, we conclude that $f + g$ is $(\mu_1 + \mu_2)$-strongly convex on $\mathcal{V}$. □

*Proof of (b).* Let $\alpha > 0$, let $x, y \in \mathcal{V}$, and let $\lambda \in [0,1]$. We can write

$$\begin{aligned}(\alpha f)(\lambda x + (1-\lambda)y) &= \alpha f(\lambda x + (1-\lambda)y) \\ &\leq \lambda\alpha f(x) + (1-\lambda)\alpha f(y) - \frac{\alpha\mu}{2}\lambda(1-\lambda)\|x-y\|^2 \\ &= \lambda(\alpha f)(x) + (1-\lambda)(\alpha f)(y) - \frac{\alpha\mu}{2}\lambda(1-\lambda)\|x-y\|^2.\end{aligned}$$

This shows that $\alpha f$ is $\alpha\mu_1$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. □

*Proof of (c).* Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. We can write

$$\begin{aligned}f(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu_1}{2}\lambda(1-\lambda)\|x-y\|^2 \\ &\leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu_1'}{2}\lambda(1-\lambda)\|x-y\|^2.\end{aligned}$$

This shows that $f$ is also a $\mu_1'$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. □

**Theorem 3.1.4** (First-order condition)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and let $\mathcal{V} \subseteq \mathrm{dom}(\partial f)$ be a convex set. Then, $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$ if and only if*

$$f(y) \geq f(x) + \langle g_x, y - x\rangle + \frac{\mu}{2}\|y-x\|^2, \quad \forall x, y \in \mathcal{V}, g_x \in \partial f(x). \tag{3.3}$$

*Proof.* Suppose that $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$. Let $x, y \in \mathcal{V}$, let $\lambda \in (0,1)$, and let $g_x \in \partial f(x)$. We can write

$$\begin{aligned}(1-\lambda)\langle g_x, y-x\rangle &= \langle g_x, (1-\lambda)(y-x)\rangle \\ &= \langle g_x, \lambda x + (1-\lambda)y - x\rangle \\ &\leq f(\lambda x + (1-\lambda)y) - f(x) \\ &\leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2 - f(x) \\ &= (1-\lambda)(f(y)-f(x)) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2, \end{aligned} \tag{3.4}$$

where the first inequality follows from the fact that $g_x \in \partial f(x)$, and the second inequality follows from the strong convexity of $f$. By dividing $1-\lambda$ from both sides Equation (3.4) and letting $\lambda \uparrow 1$, we get

$$\langle g_x, y-x\rangle \leq f(y) - f(x) - \frac{\mu}{2}\|x-y\|^2.$$

Rearranging, we obtain

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

Conversely, suppose that $f$ satisfies Equation (3.3). Let $x, y \in \mathcal{V}$ and let $\lambda \in [0,1]$. Since $\mathcal{V}$ is convex, we know that $z \coloneqq \lambda x + (1-\lambda)y \in \mathcal{V}$. Let $g_z \in \partial f(z)$. By Equation (3.3), we can write

$$f(x) \geq f(z) + \langle g_z, x-z\rangle + \frac{\mu}{2}\|x-z\|^2, \tag{3.5}$$

and

$$f(y) \geq f(z) + \langle g_z, y-z\rangle + \frac{\mu}{2}\|y-z\|^2. \tag{3.6}$$

By summing up the resulting expressions after multiplying Equation (3.5) by $\lambda$ and Equation (3.6) by $1-\lambda$, we get

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) &\geq f(z) + \langle g_z, \lambda(x-z) + (1-\lambda)(y-z)\rangle + \frac{\lambda\mu}{2}\|x-z\|^2 + \frac{(1-\lambda)\mu}{2}\|y-z\|^2 \\ &= f(z) + \frac{\mu}{2}\lambda(1-\lambda)^2\|x-y\|^2 + \frac{\mu}{2}(1-\lambda)\lambda^2\|x-y\|^2 \\ &= f(z) + \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2.\end{aligned}$$

Rearranging, we get

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2.$$

Therefore, $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$. □

In words, Theorem 3.1.4 tells us that we can construct a quadratic lower bound for a strongly convex function $f : \mathbb{R}^d \to (-\infty, +\infty]$ at each $x \in \operatorname{dom}(\partial f)$.

We can view Theorem 3.1.4 as an extension of the first-order condition for convex functions (Theorem 2.1.13). Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be convex and differentiable at $x \in \operatorname{int}(\operatorname{dom}(f))$. From Theorem 2.2.9, we know that $\partial f(x) = \{\nabla f(x)\}$. Since a convex function is 0-strongly convex, from Theorem 3.1.4, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle, \quad \forall y \in \mathbb{R}^d.$$

**Theorem 3.1.5.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper. Then, $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ if and only if*

$$\langle g_x - g_y, x-y\rangle \geq \mu\|x-y\|^2, \quad \forall x, y \in \operatorname{dom}(\partial f), g_x \in \partial f(x), g_y \in \partial f(y). \tag{3.7}$$

*Proof.* Suppose that $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$. Let $x, y \in \operatorname{dom}(\partial f)$, let $g_x \in \partial f(x)$, and let $g_y \in \partial f(y)$. From the first-order condition (Theorem 3.1.4), we can write

$$f(x) \geq f(y) + \langle g_y, x-y\rangle + \frac{\mu}{2}\|x-y\|^2, \tag{3.8}$$

and

$$f(y) \geq f(x) + \langle g_x, y-x\rangle + \frac{\mu}{2}\|y-x\|^2. \tag{3.9}$$

Summing up Equation (3.8) and Equation (3.9) yields

$$\langle g_x - g_y, x-y\rangle \geq \mu\|x-y\|^2.$$

□

**Corollary 3.1.6.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper. Then, $f$ is convex if and only if*

$$\langle g_x - g_y, x - y \rangle \geq 0, \quad \forall x, y \in \operatorname{dom}(\partial f), g_x \in \partial f(x), g_y \in \partial f(y).$$

*Proof.* We know that $f$ is convex if and only if $f$ is 0-strongly convex with respect to some norm $\|\cdot\|$. The statement then follows from Theorem 3.1.5. □

Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a convex differentiable function. From Corollary 3.1.6, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^d. \tag{3.10}$$

Equation (3.10) shows that the gradient mapping $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ of a convex differentiable function is a *monotone operator*, which is defined as follows.

**Definition 3.1.7** (Monotone operator)**.** *An operator $T : \mathbb{R}^d \to \mathbb{R}^d$ is said to be **monotone** if and only if*

$$\langle T(x) - T(y), x - y \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^d,$$

*and **strictly monotone** if and only if*

$$\langle T(x) - T(y), x - y \rangle > 0, \quad \forall x, y \in \mathbb{R}^d, x \neq y.$$

**Theorem 3.1.8** (Second-order condition)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex and let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a twice-differentiable function. Then, $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$ if and only if*

$$\left\langle \nabla^2 f(x) h, h \right\rangle \geq \mu \|h\|^2, \quad \forall x \in \mathcal{V}, h \in \mathbb{R}^d.$$

**Example 3.1.9** (Squared norm)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}_+$ be defined as*

$$f(x) \coloneqq \frac{1}{2}\|x\|_2^2, \quad \forall x \in \mathbb{R}^d.$$

*Then, $f$ is 1-strongly convex with respect to $\|\cdot\|_2$ on $\mathbb{R}^d$.*

*Proof.* Differentiate $f(x)$ with respect to $x$, we have $\nabla f(x) = x$ and $\nabla^2 f(x) = I_d$ for all $x \in \mathbb{R}^d$, where $I_d$ denotes the $d \times d$ identity matrix. Thus, we have

$$\left\langle \nabla^2 f(x) h, h \right\rangle = \langle I_d h, h \rangle \geq \|h\|_2^2, \quad \forall x, h \in \mathbb{R}^d.$$

From the second-order condition (Theorem 3.1.8), we know that $f$ is 1-strongly convex with respect to the Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^d$. □

Next, we prove that a strongly convex function has a unique minimizer over a nonempty closed convex set. In order to do so, we digress a bit to recall some notions from real analysis and introduce some additional concepts from convex analysis. Recall from basic real analysis that a set $\mathcal{V}$ is said to be **compact** if and only every open covering of $\mathcal{V}$ has a finite subcovering. Moreover, from the Heine–Borel theorem, we know that a subset $\mathcal{V}$ in $\mathbb{R}^d$ is compact if and only if it is bounded and closed. For example, the closed interval $[0, 1]$ is compact in $\mathbb{R}$.

Given a sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbb{R}$, recall that the **limit inferior** of $\{x_n\}_{n=1}^{\infty}$ is defined as

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \left( \inf_{k \geq n} x_k \right),$$

and the **limit superior** of $\{x_n\}_{n=1}^{\infty}$ is defined as

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \left( \sup_{k \geq n} x_k \right).$$

It follows immediately from the definitions of limit inferior and limit superior that

$$\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n,$$

for any sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbb{R}$.

**Definition 3.1.10** (Lower semicontinuity)**.** *A function $f : \mathbb{R}^d \to [-\infty, +\infty]$ is **lower semicontinuous (lsc)** at $x \in \mathbb{R}^d$ if for any sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbb{R}^d$ such that $x_n \to x$, we have*

$$f(x) \leq \liminf_{n\to\infty} f(x_n).$$

*A function is **lower semicontinuous** if it is lower semicontinuous throughout $\mathbb{R}^d$. A function $f$ is **upper semicontinuous (usc)** at $x \in \mathbb{R}^d$ if $-f$ is lower semicontinuous at $x$. A function $f$ is **upper semicontinuous** if $-f$ is lower semicontinuous.*

**Lemma 3.1.11** (Bauschke and Combettes [BC], Lemma 1.24)**.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$. The following statements are equivalent.*

*(a) $f$ is lower semicontinuous.*

*(b) The epigraph $\operatorname{epi}(f)$ is closed in $\mathbb{R}^d \times \mathbb{R}$.*

*(c) The sublevel set $\operatorname{lev}_{\leq t}(f)$ is closed in $\mathbb{R}^d$ for each $t \in \mathbb{R}$.*

*Proof.* ((a) $\implies$ (b)) Suppose that $f$ is lower semicontinuous. Let $\{(x_n, t_n)\}_{n=1}^{\infty}$ be a sequence in $\operatorname{epi}(f)$ such that $(x_n, t_n) \to (x, t)$ for some $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$. Since $f$ is lower semicontinuous at $x$ and $x_n \to x$, we have

$$f(x) \leq \liminf_{n\to\infty} f(x_n) \leq \liminf_{n\to\infty} t_n = t,$$

which implies that $(x, t) \in \operatorname{epi}(f)$. Thus, $\operatorname{epi}(f)$ is closed in $\mathbb{R}^d \times \mathbb{R}$. □

**Theorem 3.1.12** (Weierstrass theorem)**.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$ be lower semicontinuous and let $\mathcal{V} \subseteq \mathbb{R}^d$ be compact. Suppose that $\mathcal{V} \cap \operatorname{dom}(f)$ is nonempty. Then, $f$ achieves its infimum in $\mathcal{V}$.*

*Proof.* See Theorem 1.28 in Bauschke and Combettes [BC]. □

**Corollary 3.1.13** (Weierstrass theorem)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be compact and let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous. Suppose that $\mathcal{V} \cap \operatorname{dom}(f)$ is nonempty. Then, $f$ attains its extremum in $\mathcal{V}$.*

*Proof.* Since $f$ is continuous, we know that $f$ is both lower semicontinuous and upper semicontinuous. Since $f$ is lower semicontinuous, it follows from Theorem 3.1.12 that $f$ attains its infimum in $\mathcal{V}$. On the other hand, since $f$ is upper semicontinuous, we know that $-f$ is lower semicontinuous. Again, from Theorem 3.1.12, we know that $-f$ attains its infimum in $\mathcal{V}$. Since we have $\inf_{x\in\mathcal{V}} -f(x) = \sup_{x\in\mathcal{V}} f(x)$, we deduce that $f$ attains its supremum in $\mathcal{V}$. Therefore, $f$ attains its extremum in $\mathcal{V}$. □

We now return to our main goal of proving that a strongly convex function attains its infimum over a nonempty closed convex set.

**Theorem 3.1.14.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, closed, and $\mu$-strongly convex with respect to $\|\cdot\|$ over a nonempty closed convex set $\mathcal{V} \subseteq \operatorname{dom}(f)$. Assume that $\operatorname{dom}(\partial f)$ is nonempty. Then, $f$ has a unique minimizer over $\mathcal{V}$.*

*Proof.* Since $f$ is closed, from Lemma 3.1.11, we know that $f$ is lower semicontinuous. Then, it follows from the Weierstrass theorem (Theorem 3.1.12) that $f$ attains its minimum in $\mathcal{V}$. The uniqueness of the minimizer follows from the strong convexity of $f$. □

An immediate consequence of Theorem 3.1.14 is that the Euclidean projection (Definition 2.1.17) is indeed well-defined. From Example 3.1.9, we know that $h(y) \coloneqq \|y - x\|_2^2/2$ is 1-strongly convex with respect to $\|\cdot\|_2$. It then follows from Theorem 3.1.14 that for a nonempty closed convex set $\mathcal{V} \subseteq \mathbb{R}^d$, the Euclidean projection

$$\operatorname{proj}_{\mathcal{V}}(x) = \operatorname*{arg\,min}_{y\in\mathcal{V}} \|y - x\|_2 = \operatorname*{arg\,min}_{y\in\mathcal{V}} \frac{1}{2}\|y - x\|_2^2$$

is well-defined.

## 3.2 Online Subgradient Descent with Strongly Convex Loss Functions

**Theorem 3.2.1** (Bartlett, Hazan, and Rakhlin [BHR07])**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set, and let $\ell_t : \mathcal{V} \to \mathbb{R}$ be subdifferentiable and $\mu_t$-strongly convex with respect to $\|\cdot\|_2$ on $\mathcal{V}$, where $\mu_t > 0$, for $t = 1, \ldots, T$. Run* Online Subgradient Descent *with step sizes $\eta_1, \ldots, \eta_T$ given by*

$$\eta_t = \frac{1}{\sum_{i=1}^t \mu_i}, \quad \forall 1 \leq t \leq T.$$

*Then, we have*

$$\sum_{t=1}^T (\ell_t(x_t) - \ell_t(u)) \leq \frac{1}{2} \sum_{t=1}^T \frac{\|g_t\|_2^2}{\sum_{i=1}^t \mu_i}, \quad \forall u \in \mathcal{V},$$

*where $g_t \in \partial \ell_t(x_t)$ for $t = 1, \ldots, T$.*

*Proof.* Let $u \in \mathcal{V}$ be fixed, and let $g_t \in \partial \ell_t(x_t)$ for $t = 1, \dots, T$. We can write

$$\begin{aligned}
\ell_t(x_t) - \ell_t(u) &\le \langle g_t, x_t - u \rangle - \frac{\mu_t}{2}\|x_t - u\|_2^2 \\
&\le \frac{1}{2\eta_t}\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2 + \frac{\eta_t}{2}\|g_t\|_2^2 - \frac{\mu_t}{2}\|x_t - u\|_2^2 \\
&= \left(\frac{1}{2\eta_t} - \frac{\mu_t}{2}\right)\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2 + \frac{\eta_t}{2}\|g_t\|_2^2 \\
&= \begin{cases} -\frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2 + \frac{\eta_t}{2}\|g_t\|_2^2, & \text{if } t = 1; \\ \frac{1}{2\eta_{t-1}}\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2 + \frac{\eta_t}{2}\|g_t\|_2^2, & \text{otherwise,} \end{cases}
\end{aligned} \tag{3.11}$$

where the first inequality follows from Theorem 3.1.4, the second inequality follows from Lemma 2.2.18, and the last equality follows from the definition of the $\eta_t$'s. Summing up inequality (3.11) over $t = 1, \dots, T$, we have

$$\begin{aligned}
\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\le -\frac{1}{2\eta_1}\|x_2 - u\|_2^2 + \sum_{t=2}^{T}\left(\frac{1}{2\eta_{t-1}}\|x_t - u\|_2^2 - \frac{1}{2\eta_t}\|x_{t+1} - u\|_2^2\right) + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_2^2 \\
&= -\frac{1}{2\eta_1}\|x_2 - u\|_2^2 + \frac{1}{2\eta_1}\|x_2 - u\|_2^2 - \frac{1}{2\eta_T}\|x_{T+1} - u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_2^2 \\
&\le -\frac{1}{2\eta_T}\|x_{T+1} - u\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_2^2 \\
&\le \frac{1}{2}\sum_{t=1}^{T}\frac{\|g_t\|_2^2}{\sum_{i=1}^{t}\mu_i}.
\end{aligned}$$

□

**Corollary 3.2.2** (Bartlett, Hazan, and Rakhlin [BHR07])**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set, and let $\ell_t : \mathcal{V} \to \mathbb{R}$ be subdifferentiable, $\mu$-strongly convex and $L$-Lipschitz with respect to $\|\cdot\|_2$ on $\mathcal{V}$, where $\mu > 0$ and $L > 0$, for $t = 1, \dots, T$. Run* Online Subgradient Descent *with step sizes $\eta_1, \dots, \eta_T$ given by*

$$\eta_t = \frac{1}{t\mu}, \quad \forall 1 \le t \le T.$$

*Then, we have*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \frac{L^2}{2\mu}(1 + \log T), \quad \forall u \in \mathcal{V}.$$

*Proof.* From Theorem 3.2.1, we can write

$$\begin{aligned}
\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\le \frac{1}{2}\sum_{t=1}^{T}\frac{\|g_t\|_2^2}{\sum_{i=1}^{t}\mu_i} \\
&\le \frac{1}{2}\sum_{t=1}^{T}\frac{L}{t\mu} \\
&\le \frac{L}{2\mu}(1 + \log T),
\end{aligned}$$

where the second inequality follows from Theorem 2.2.17, and the last inequality follows by bounding the sum $\sum_{t=2}^{T} 1/t$ with the integral $\int_1^T 1/t \, \mathrm{d}t$. □

## 3.3 Smoothness

**Definition 3.3.1** (Dual norm)**.** *The* ***dual norm*** $\|\cdot\|_*$ *of a norm $\|\cdot\|$ is defined as*

$$\|x\|_* = \sup_{y \in \mathbb{R}^d : \|y\| \le 1} \langle x, y \rangle, \quad \forall x \in \mathbb{R}^d.$$

It follows immediately from the definition of dual norm (Definition 3.3.1) that we have

$$\langle x, y \rangle \le \|x\|_* \|y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{3.12}$$

Inequality (3.12) is known as the **Hölder's inequality** or the **dual norm inequality**, and is a strict generalization of the Cauchy–Schwarz inequality.

**Example 3.3.2** ($\ell_2$-norm)**.** *The dual norm of the $\ell_2$-norm $\|\cdot\|_2$ is itself.*

*Proof.* Let $\|\cdot\|_*$ denote the dual norm of the $\ell_2$-norm. By the Cauchy–Schwarz inequality, we can write

$$\|x\|_* = \sup_{y\in\mathbb{R}^d:\|y\|_2\le 1} \langle x, y\rangle \le \|x\|_2, \quad \forall x \in \mathbb{R}^d.$$

On the other hand, we have

$$\|x\|_* = \sup_{y\in\mathbb{R}^d:\|y\|_2\le 1} \langle x, y\rangle \ge \left\langle x, \frac{x}{\|x\|_2}\right\rangle = \|x\|_2, \quad \forall x \in \mathbb{R}^d.$$

Therefore, we conclude that $\|x\|_* = \|x\|_2$ for all $x \in \mathbb{R}^d$. □

Since the dual norm of the $\ell_2$-norm is itself, we say that the $\ell_2$-norm is **self-dual**. In general, the dual norm of the $\ell_p$-norm is the $\ell_q$-norm, where $p, q \in [1,\infty]$ such that $1/p + 1/q = 1$ (i.e., $p$ and $q$ are *Hölder conjugates* of each other). The proof is given in Exercise 4.4.

**Example 3.3.3** ($\ell_1$-norm)**.** *The dual norm of the $\ell_1$-norm is the $\ell_\infty$-norm.*

Recall from linear algebra that a matrix $A \in \mathbb{R}^{d\times d}$ is said to be *positive-definite* if we have $x^\top Ax > 0$ for all $x \in \mathbb{R}^d \setminus \{0\}$. From the spectral theorem, we know that every positive-definite matrix is diagonalizable and thus invertible. Thus, there exists some invertible matrix $Q \in \mathbb{R}^{d\times d}$ such that $A = Q^{-1}\Lambda Q$, where $\Lambda \in \mathbb{R}^{d\times d}$ is a diagonal matrix with the eigenvalues of $A$ as its diagonal entries. Without loss of generality, we may assume that $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, where $\lambda_1 \ge \cdots \ge \lambda_d$. We define the square root of $A$, denoted by $A^{1/2}$, as the matrix $A^{1/2} \coloneqq Q^{-1}\Lambda^{1/2}Q$, where $\Lambda^{1/2} \coloneqq \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})$. The matrix $A^{1/2}$ is indeed the "square root" of $A$ since

$$A^{1/2}A^{1/2} = (Q^{-1}\Lambda^{1/2}Q)(Q^{-1}\Lambda^{1/2}Q) = Q^{-1}\Lambda^{1/2}\Lambda^{1/2}Q = Q^{-1}\Lambda Q = A.$$

The inverse of $A^{1/2}$, denoted by $A^{-1/2}$, is the matrix defined as $A^{1/2} \coloneqq Q^{-1}\Lambda^{-1/2}Q$, where $\Lambda^{1/2} \coloneqq \operatorname{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_d})$.

**Example 3.3.4** (Matrix norm)**.** *Let $A \in \mathbb{R}^{d\times d}$ be a positive-definite matrix. Then, the mapping $\|\cdot\|_A : \mathbb{R}^d \to \mathbb{R}_+$ defined as*

$$\|x\|_A = \sqrt{x^\top Ax}, \quad \forall x \in \mathbb{R}^d,$$

*is a norm. Moreover, the dual norm of $\|\cdot\|_A$ is $\|\cdot\|_{A^{-1}}$.*

*Proof.* First, we prove that $\|\cdot\|_A$ is a norm. Since $A$ is positive-definite, we know that $x^\top Ax > 0$ for all $x \in \mathbb{R}^d \setminus \{0\}$ and $x^\top Ax = 0$ if and only if $x = 0$. Thus, $\|\cdot\|_A$ is positive-definite. Let $\alpha \in \mathbb{R}$. We have

$$\|\alpha x\|_A = \sqrt{(\alpha x)^\top A(\alpha x)} = \sqrt{\alpha^2(x^\top Ax)} = |\alpha| \cdot \sqrt{x^\top Ax} = |\alpha| \cdot \|x\|_A, \quad \forall x \in \mathbb{R}^d.$$

Thus, $\|\cdot\|_A$ is absolutely homogeneous. Let $x, y \in \mathbb{R}^d$. We can write

$$\begin{aligned}
\|x+y\|_A^2 &= (x+y)^\top A(x+y) \\
&= x^\top Ax + 2x^\top Ay + y^\top Ay \\
&\le \|x\|_A^2 + 2\|x\|_A\|y\|_A + \|y\|_A^2 \\
&= (\|x\|_A + \|y\|_A)^2,
\end{aligned}$$

where the inequality follows from the Cauchy–Schwarz inequality. Take square root on both sides, we get $\|x+y\|_A \le \|x\|_A + \|y\|_A$ for all $x, y \in \mathbb{R}^d$. This shows that $\|\cdot\|_A$ satisfies the triangle inequality. Therefore, $\|\cdot\|_A$ is indeed a norm.

Next, we prove that the dual norm of $\|\cdot\|_A$ is $\|\cdot\|_{A^{-1}}$. Denote by $\|\cdot\|_*$ the dual norm of $\|\cdot\|_A$. For all $x \in \mathbb{R}^d$, we can write

$$\begin{aligned}
\|x\|_* &= \sup_{y\in\mathbb{R}^d:\|y\|_A\le 1} \langle x, y\rangle = \sup_{y\in\mathbb{R}^d:y^\top Ay\le 1} x^\top y \\
&= \sup_{z\in\mathbb{R}^d:z^\top z\le 1} x^\top A^{-1/2}z \\
&= \sup_{z\in\mathbb{R}^d:\|z\|_2\le 1} (A^{-1/2}x)^\top z \\
&= \|A^{-1/2}x\|_2 \\
&= \sqrt{x^\top A^{-1}x} \\
&= \|x\|_{A^{-1}},
\end{aligned}$$

where the second equality follows from a change of variable $z = A^{1/2}y$, and the fourth equality follows from the fact that the $\ell_2$-norm is self-dual (Example 3.3.2). Since $x \in \mathbb{R}^d$ is arbitrary, we conclude that the dual norm of $\|\cdot\|_A$ is $\|\cdot\|_{A^{-1}}$. □

**Definition 3.3.5** (Smoothness)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{V}$. We say that $f$ is **$L$-smooth** with respect to $\|\cdot\|$ on $\mathcal{V}$ if we have*

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in \mathcal{V},$$

*for some constant $L > 0$. The constant $L > 0$ is called the **smoothness parameter**.*

In other words, a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-smooth on $\mathcal{V}$ if the gradient mapping $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is $L$-Lipschitz with respect to $\|\cdot\|$ on $\mathcal{V}$.

**Theorem 3.3.6.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth with respect to $\|\cdot\|$ on $\mathcal{V}$. Then, we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2, \quad \forall x, y \in \mathcal{V}.$$

*Proof.* Let $x, y \in \mathcal{V}$. We have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\|_* \|x - y\| \leq L\|x - y\|^2,$$

where the first inequality follows from Hölder's inequality, and the second inequality follows from the $L$-smoothness of $f$. □

**Theorem 3.3.7.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex and let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth with respect to $\|\cdot\|$ on $\mathcal{V}$. Then, we have*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{V}. \tag{3.13}$$

*Proof.* Let $x, y \in \mathcal{V}$ be fixed. Consider the continuous function $\varphi : [0, 1] \to \mathbb{R}$ defined as

$$\varphi(t) \coloneqq f(x + t(y - x)), \quad \forall t \in [0, 1].$$

Observe that we have $\varphi(0) = f(x)$ and $\varphi(1) = f(y)$. By the Fundamental Theorem of Calculus, we can write

$$\begin{aligned} f(y) &= \varphi(1) \\ &= \varphi(0) + \int_0^1 \varphi'(t)\,\mathrm{d}t \\ &= f(x) + \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle\,\mathrm{d}t \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle\,\mathrm{d}t. \end{aligned} \tag{3.14}$$

We can further bound Equation (3.14) by

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle\,\mathrm{d}t \right| \\ &\leq \int_0^1 |\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle|\,\mathrm{d}t \\ &= \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_* \|y - x\|\,\mathrm{d}t \\ &\leq \int_0^1 tL\|y - x\|^2\,\mathrm{d}t \\ &= \frac{L}{2}\|y - x\|^2, \end{aligned}$$

where the second equality follows from Hölder's inequality, and the second inequality follows from the $L$-smoothness of $f$. This proves Equation (3.13). □

We can rewrite Equation (3.13) as

$$f(x) + \langle \nabla f(x), y - x \rangle - \frac{L}{2}\|y - x\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2,$$

for all $x, y \in \mathcal{V}$. In words, Theorem 3.3.7 tells us that we have a quadratic lower and upper bound of $f$ at each point of $\mathcal{V}$.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function that is $\mu$-strongly convex with respect to $\|\cdot\|$ and let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex. Recall from the first-order condition (Theorem 3.1.4) that we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{V}.$$

If we further assume that $f$ is $L$-smooth, then we can write

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathcal{V}. \tag{3.15}$$

From Equation (3.15), it is easy to see that $\mu \leq L$. Thus, we must have $\kappa_f \coloneqq L/\mu \geq 1$. The quantity $\kappa_f$ is called the **condition number** of $f$.

# 4 Online Mirror Descent

Online Mirror Descent is a generalization of Online Subgradient Descent. Instead of using the $\ell_2$-norm to measure the distance between two points, Online Mirror Descent uses a more general mathematical object to measure distance. This allows us to better capture the geometry of the feasible set, see Section 4.7.

## 4.1 Reinterpreting the Online Subgradient Descent Algorithm

Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set, let $\ell_1, \dots, \ell_T : \mathbb{R}^d \to \mathbb{R}$ be a sequence of subdifferentiable loss functions, and let $\eta_1, \dots, \eta_T$ be the step sizes. Recall that Online Subgradient Descent iterates as

$$x_{t+1} \leftarrow \text{proj}_{\mathcal{V}}(x_t - \eta_t g_t), \quad \text{for } t = 1, \dots, T,$$

where $\text{proj}_{\mathcal{V}} : \mathbb{R}^d \to \mathcal{V}$ is the Euclidean projection onto $\mathcal{V}$ and $g_t \in \partial \ell_t(x_t)$. Namely, we solve the following constrained optimization problem in each iteration $t = 1, \dots, T$:

$$x_{t+1} = \arg\min_{x \in \mathcal{V}} \|x - x_t + \eta_t g_t\|_2^2.$$

We can write

$$\begin{aligned}
\arg\min_{x \in \mathcal{V}} \|x - x_t + \eta_t g_t\|_2^2 &= \arg\min_{x \in \mathcal{V}} \left\{ \|x - x_t\|_2^2 + \eta_t^2 \|g_t\|_2^2 + 2\eta_t \langle g_t, x - x_t \rangle \right\} \\
&= \arg\min_{x \in \mathcal{V}} \left\{ \|x - x_t\|_2^2 + 2\eta_t \langle g_t, x - x_t \rangle \right\} \\
&= \arg\min_{x \in \mathcal{V}} \left\{ \langle g_t, x - x_t \rangle + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right\} \\
&= \arg\min_{x \in \mathcal{V}} \left\{ \ell_t(x_t) + \langle g_t, x - x_t \rangle + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right\}. \qquad (4.1)
\end{aligned}$$

Thus, updating rule of Online Subgradient Descent is equivalent to Equation 4.1. The optimization problem (4.1) can be interpreted as minimizing the affine function $\tilde{\ell}_t(x) \coloneqq \ell_t(x_t) + \langle g_t, x - x_t \rangle$ with the constraint that $x$ lies in the "proximity" of $x_t$. Since $g_t \in \partial \ell_t(x_t)$, we know that $\ell_t(x) \geq \tilde{\ell}_t(x)$. Therefore, the updating rule of Online Subgradient Descent is equivalent to minimizing a lower bound of the loss function $\ell_t$ with the constraint that $x$ is close to $x_t$.

As we will shortly see, the updating rule (4.1) is basically the one we use in Online Mirror Descent. The only difference is that we measure the "locality" with a different mathematical objected called *Bregman divergence* in Online Mirror Descent. In (4.1), we use the squared $\ell_2$-norm to measure the distance between $x$ and $x_t$. As will see in the next section, the squared $\ell_2$-norm is a special case of Bregman divergence.

## 4.2 Bregman Divergence

**Definition 4.2.1** (Strictly convex function)**.** *A function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is **strictly convex** if and only if*

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \text{dom}(f), x \neq y, \lambda \in (0,1). \qquad (4.2)$$

*A function is said to be **strictly convex** on a convex set $\mathcal{V} \subseteq \mathbb{R}^d$ if* Equation (4.2) *holds for all $x, y \in \mathcal{V}$ with $x \neq y$.*

Note that a strongly convex function (with respect to any norm) is also strictly convex. Indeed, let $\mathcal{V} \subseteq \mathbb{R}^d$ be convex and let $f : \mathcal{V} \to (-\infty, +\infty]$ be $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2} \lambda (1-\lambda) \|x - y\|^2 < \lambda f(x) + (1-\lambda) f(y),$$

for all $x, y \in \mathcal{V}$ with $x \neq y$ and $\lambda \in (0,1)$. Thus, $f$ is strictly convex on $\mathcal{V}$.

Recall from Theorem 2.1.13 that for a convex differentiable function $f : \mathbb{R}^d \to (-\infty, +\infty]$, we have an affine lower bound at each point in the interior of the domain. For strictly convex differentiable functions, the lower bounds become strict.

**Theorem 4.2.2** (First-order condition)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a strictly convex function and let $x \in \text{int}(\text{dom}(f))$. If $f$ is (Gâteaux) differentiable at $x$, then we have*

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle, \qquad (4.3)$$

*for all $y \in \mathbb{R}^d$ such that $y \neq x$.*

*Proof.* If $y \notin \operatorname{dom}(f)$, then we must have $f(y) = +\infty$. Clearly, Equation (4.3) holds if $f(y) = +\infty$. It remains to consider the case when $y \in \operatorname{dom}(f)$ such that $y \neq x$. □

**Theorem 4.2.3.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a strictly convex (Gâteaux) differentiable function. Then, we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0, \quad \forall x, y \in \operatorname{int}(\operatorname{dom}(f)), x \neq y.$$

*Proof.* Let $x, y \in \operatorname{int}(\operatorname{dom}(f))$. By the first-order condition (Theorem 4.2.2), we can write

$$f(x) > f(y) + \langle \nabla f(y), x - y \rangle , \tag{4.4}$$

and

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle . \tag{4.5}$$

Summing up Equation (4.4) and Equation (4.5) yields

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0.$$

□

In other words, the gradient mapping of a strictly convex differentiable function is strictly monotone (Definition 3.1.7). Note that a strictly monotone operator is injective. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be strictly monotone. The condition $\langle T(x) - T(y), x - y \rangle > 0$ for all $x, y \in \mathbb{R}^d$ such that $x \neq y$ implies that $T(x) \neq T(y)$ for $x \neq y$, and thus $T$ is one-to-one.

**Definition 4.2.4** (Bregman divergence, Bregman [Bre67])**.** *Let $h : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is differentiable on $\operatorname{int}(\mathcal{X})$. The **Bregman divergence** with respect to $h$, denoted by $D_h : \mathcal{X} \times \operatorname{int}(\mathcal{X}) \to \mathbb{R}$, is defined as*

$$D_h(x \,\|\, y) \coloneqq h(x) - h(y) - \langle \nabla h(y), x - y \rangle , \quad \forall x \in \mathcal{X}, y \in \operatorname{int}(\mathcal{X}).$$

*The function $h$ is called a **distance-generating function (DGF)**.*

Let $h : \mathcal{X} \to \mathbb{R}$ be a strictly convex differentiable function. It follows immediately from Theorem 4.2.2 that $D_h(x \,\|\, y) > 0$ for all $x \in \mathcal{X}$ and $y \in \operatorname{int}(\mathcal{X})$. We also have $D_h(x \,\|\, y) = 0$ if and only if $x = y$. Thus, the Bregman divergence $D_h$ is a nonnegative function. We can interpret $D_h(x \,\|\, y)$ as a "similarity measure" between $x$ and $y$. However, note that $D_h$ is not symmetric, that is, $D_h(x \,\|\, y) \neq D_h(y \,\|\, x)$ in general. Therefore, the Bregman divergence $D_h$ is not a distance function.

If the distance-generating function $h : \mathcal{X} \to \mathbb{R}$ is $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$ and differentiable on $\operatorname{int}(\mathcal{X})$, then by Theorem 3.1.4, we have

$$D_h(x \,\|\, y) \geq \frac{\mu}{2}\|x - y\|^2, \quad \forall x \in \mathcal{X}, y \in \operatorname{int}(\mathcal{X}).$$

**Example 4.2.5** (Squared $\ell_2$-norm)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}_+$ be defined as*

$$h(x) \coloneqq \frac{1}{2}\|x\|_2^2, \quad \forall x \in \mathbb{R}^d.$$

*Then, the corresponding Bregman divergence $D_h$ is given by*

$$D_h(x \,\|\, y) \coloneqq \frac{1}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

*Proof.* Recall from Example 3.1.9 that $h(x)$ is 1-strongly convex. Thus, $h$ is strictly convex. The associated Bregman divergence $D_h$ is defined as

$$\begin{aligned} D_h(x \,\|\, y) &= h(x) - h(y) - \langle \nabla h(y), x - y \rangle \\ &= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y \rangle \\ &= \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2 - \langle y, x \rangle \\ &= \frac{1}{2}\|x - y\|_2^2, \end{aligned}$$

for all $x, y \in \mathbb{R}^d$. □

As mentioned earlier, Bregman divergences are not necessarily symmetric. However, as Example 4.2.5 shows, we have $D_h(x \,\|\, y) = D_h(y \,\|\, x)$ if $h(x) = \|x\|_2^2/2$.

**Example 4.2.6** (Negative Shannon entropy)**.** *Let $h : \mathbb{R}^d_+ \to \mathbb{R}$ be defined as*

$$h(x) \coloneqq \sum_{i=1}^d x_i \log(x_i) - \sum_{i=1}^d x_i, \quad \forall x \in \mathbb{R}^d_+,$$

*where we define $0 \log(0) \coloneqq 0$. Then, the corresponding Bregman divergence $D_h$ is given by*

$$D_h(x \,\|\, y) \coloneqq \sum_{i=1}^d x_i \log\left(\frac{x_i}{y_i}\right) - \left(\sum_{i=1}^d x_i - \sum_{i=1}^d y_i\right), \quad \forall x, y \in \mathbb{R}^d_+.$$

*Proof.* First, we show that the negative Shannon entropy is strictly convex. Differentiate with respect to $x$, we have

$$\nabla h(x) = \begin{pmatrix} \log(x_1) \\ \log(x_2) \\ \vdots \\ \log(x_d) \end{pmatrix}, \quad \nabla^2 h(x) = \begin{pmatrix} 1/x_1 & 0 & \dots & 0 \\ 0 & 1/x_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/x_d \end{pmatrix}, \quad \forall x \in \mathbb{R}^d_+.$$

Since all the eigenvalues of $\nabla^2 h(x)$ are positive, we know that $\nabla^2 h(x)$ is positive-definite. Thus, by Theorem 3.1.8, we know that $h$ is strongly convex and hence strictly convex. Next, we find the corresponding Bregman divergence. The corresponding Bregman divergence $D_h$ is given by

$$\begin{aligned} D_h(x \,\|\, y) &= h(x) - h(y) - \langle \nabla h(y), x - y \rangle \\ &= \sum_{i=1}^d x_i \log(x_i) - \sum_{i=1}^d x_i - \sum_{i=1}^d y_i \log(y_i) + \sum_{i=1}^d y_i - \sum_{i=1}^d \log(y_i)(x_i - y_i) \\ &= \sum_{i=1}^d x_i \log(x_i) - \sum_{i=1}^d x_i \log(y_i) - \sum_{i=1}^d x_i + \sum_{i=1}^d y_i \\ &= \sum_{i=1}^d x_i \log\left(\frac{x_i}{y_i}\right) - \left(\sum_{i=1}^d x_i - \sum_{i=1}^d y_i\right), \end{aligned}$$

for all $x, y \in \mathbb{R}^d_+$. □

**Remark 4.2.7.** In online learning literature, the negative Shannon entropy refer to slightly different functions based on the context. When the domain is the nonnegative orthant $\mathbb{R}^d_+$, the negative Shannon entropy function refers to the one defined in Example 4.2.6. When the domain is the probability simplex $\Delta_d = \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\}$, then the negative Shannon entropy refers to the function $h' : \Delta_d \to \mathbb{R}$ defined as

$$h'(x) \coloneqq \sum_{i=1}^d x_i \log(x_i), \quad \forall x \in \Delta_d.$$

Note that we always have $\sum_{i=1}^d x_i = 1$ for $x \in \Delta_d$. Thus, optimizing $h$ is equivalent to optimizing $h'$ over $\Delta_d$. We can view $h'$ as a normalized version of $h$.

**Lemma 4.2.8** (Three-point equality, Chen and Teboulle [CT93])**.** *Let $h : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is differentiable on $\operatorname{int}(\mathcal{X})$, and let $D_h$ be the Bregman divergence associated with $h$. Then, for any three points $x, y \in \operatorname{int}(\mathcal{X})$ and $z \in \mathcal{X}$, we have the following identity:*

$$D_h(z \,\|\, x) + D_h(x \,\|\, y) = D_h(z \,\|\, y) + \langle \nabla h(y) - \nabla h(x), z - x \rangle .$$

*Proof.* Let $x, y \in \operatorname{int}(\mathcal{X})$ and let $z \in \mathcal{X}$. We can write

$$\begin{aligned} D_h(z \,\|\, x) + D_h(x \,\|\, y) &= h(z) - h(x) - \langle \nabla h(x), z - x \rangle + h(x) - h(y) - \langle \nabla h(y), x - y \rangle \\ &= h(z) - h(y) - \langle \nabla h(x), z - x \rangle - \langle \nabla h(y), x - y \rangle \\ &= (h(z) - h(y) - \langle \nabla h(y), z - y \rangle) - \langle \nabla h(x), z - x \rangle + \langle \nabla h(y), z - x \rangle \\ &= D_h(z \,\|\, y) + \langle \nabla h(y) - \nabla h(x), z - x \rangle . \end{aligned}$$

□

In fact, the three-point equality (Lemma 4.2.8) can be viewed as a generalization of the well-known Pythagorean theorem. If $h(x) \coloneqq \frac{1}{2}\|x\|_2^2$, then by the three-point equality, we have

$$\frac{1}{2}\|z - x\|_2^2 + \frac{1}{2}\|x - y\|_2^2 = \frac{1}{2}\|z - y\|_2^2 + \langle y - x, z - x \rangle , \quad \forall x, y, z \in \mathbb{R}^d.$$

Indeed, by the Pythagorean theorem, we can write

$$\frac{1}{2}\|z-y\|_2^2 = \frac{1}{2}\|(z-x)+(x-y)\|_2^2 = \frac{1}{2}\|z-x\|_2^2 + \frac{1}{2}\|x-y\|_2^2 + \langle z-x, x-y\rangle,$$

for all $x, y, z \in \mathbb{R}^d$.

## 4.3 Online Mirror Descent

The Online Mirror Descent algorithm is as follows.

**Algorithm 5** Online Mirror Descent

**Inputs:** A nonempty closed convex set $\mathcal{V} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$, a strictly convex differentiable function $h : \mathcal{X} \to \mathbb{R}$, an initial point $x_1 \in \mathcal{V} \cap \operatorname{int}(\mathcal{X})$, and step sizes $\eta_1, \ldots, \eta_T > 0$.

1: **for** $t = 1, \ldots, T$ **do**
2: The learner outputs $x_t \in \mathcal{V}$.
3: The adversary chooses a convex subdifferentiable loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
4: The learner calls the first-order oracle for $g_t \in \partial \ell_t(x_t)$.
5: The learner updates by setting

$$x_{t+1} \leftarrow \operatorname*{arg\,min}_{x \in \mathcal{V}} \left\{ \langle g_t, x\rangle + \frac{1}{\eta_t} D_h(x \,\|\, x_t) \right\}.$$

6: **end for**

**Remark 4.3.1.** Note that the iterates of Online Mirror Descent may not be well-defined. The function $\langle g_t, x\rangle + D_h(x \,\|\, x_t)/\eta_t$ may not attain a minimum over the feasible set $\mathcal{V}$. Moreover, the minimizers may not be unique.

We can recover Online Subgradient Descent from Online Mirror Descent. Recall from Example 4.2.5 that if $h(x) = \frac{1}{2}\|x\|_2^2$, then the associated Bregman divergence is given by

$$D_h(x \,\|\, y) = \frac{1}{2}\|x-y\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

Thus, if we choose $h(x) = \frac{1}{2}\|x\|_2^2$ as the distance-generating function, then Online Mirror Descent iterates as

$$x_{t+1} \leftarrow \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ \langle g_t, x\rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\}, \quad \text{for } t = 1, \ldots, T.$$

From Equation (4.1), we know that this is exactly the updating rule of Online Subgradient Descent.

Before we prove the regret guarantee of Online Mirror Descent, we show that Online Mirror Descent is indeed a well-defined algorithm if the distance-generating function is strongly convex. Let $h : \mathcal{X} \to \mathbb{R}$ be $\mu$-strongly convex with respect to $\|\cdot\|$ on a closed convex set $\mathcal{V} \subseteq \mathcal{X}$ and differentiable on some open set containing $\mathcal{V}$, and let $x_t \in \mathcal{V}$ be fixed. Note that the function $D_h(\cdot \,\|\, x_t)$ is also $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. Indeed, for all $x, y \in \mathcal{V}$ and $\lambda \in [0,1]$, we can write

$$\begin{aligned}
D_h(\lambda x + (1-\lambda)y \,\|\, x_t) &= h(\lambda x + (1-\lambda)y) - h(x_t) - \langle \nabla h(x_t), \lambda x + (1-\lambda)y - x_t\rangle \\
&\le \lambda h(x) + (1-\lambda)h(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2 - h(x_t) - \langle \nabla h(x_t), \lambda x + (1-\lambda)y - x_t\rangle \\
&= \lambda(h(x) - h(x_t) - \langle \nabla h(x_t), x - x_t\rangle) + (1-\lambda)(h(y) - h(x_t) - \langle \nabla h(x_t), y - x_t\rangle) \\
&\quad - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2 \\
&= \lambda D_h(x \,\|\, x_t) + (1-\lambda) D_h(y \,\|\, x_t) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2.
\end{aligned}$$

Let $\eta_t > 0$ be a constant and let $g_t \in \mathbb{R}^d$ be fixed. Since $D_h(\cdot \,\|\, x_t)$ is a $\mu$-strongly convex function with respect to $\|\cdot\|$ on $\mathcal{V}$, from part (b) of Proposition 3.1.3, we know that $D_h(\cdot \,\|\, x_t)/\eta_t$ is $(\mu/\eta_t)$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. Moreover, since $\langle g_t, x\rangle$ is convex in $x$ (hence 0-strongly convex), it follows from part (b) of Proposition 3.1.3 that $\langle g_t, x\rangle + D_h(x \,\|\, x_t)/\eta_t$ is $(\mu/\eta_t)$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. Then, from Theorem 3.1.14, we know that

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in \mathcal{V}} \left\{ \langle g_t, x\rangle + \frac{1}{\eta_t} D_h(x \,\|\, x_t) \right\}$$

exists and is unique. Therefore, Online Mirror Descent is a well-defined algorithm.

Next, we prove the regret guarantee of Online Mirror Descent. Similar to what we have done in the regret analysis of Online Subgradient Descent, we first prove a lemma upper bounding the instantaneous regret of Online Mirror Descent.

**Lemma 4.3.2.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{V} \subseteq \operatorname{int}(\mathcal{X})$ be a nonempty closed convex set. Let $h : \mathcal{X} \to \mathbb{R}$ be proper, closed, differentiable, and $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. Then, with the updating rule of* Online Mirror Descent*, we have*

$$\begin{aligned}\eta_t(\ell_t(x_t) - \ell_t(u)) &\le \eta_t \langle g_t, x_t - u\rangle && (4.6)\\ &\le D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1}) - D_h(x_{t+1} \,\|\, x_t) + \langle \eta_t g_t, x_t - x_{t+1}\rangle && (4.7)\\ &\le D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1}) + \frac{\eta_t^2}{2\mu}\|g_t\|_*^2, && (4.8)\end{aligned}$$

*for all $u \in \mathcal{V}$.*

*Proof.* Let $u \in \mathcal{V}$. Since $g_t \in \partial \ell_t(x_t)$, we can write

$$\ell_t(u) \ge \ell_t(x_t) + \langle g_t, u - x_t\rangle . \tag{4.9}$$

Equation (4.6) follows from multiplying both sides of Equation (4.9) by $\eta_t > 0$ and rearranging.

Since Online Mirror Descent iterates as $x_{t+1} \leftarrow \arg\min_{x\in\mathcal{V}} \langle g_t, x\rangle + D_h(x \,\|\, x_t)/\eta_t$, by the optimality condition (Theorem 2.1.15), we have

$$\left\langle g_t + \frac{1}{\eta_t}(\nabla h(x_{t+1}) - \nabla h(x_t)), u - x_{t+1}\right\rangle \ge 0. \tag{4.10}$$

By multiplying both sides of Equation (4.7) by $\eta_t > 0$ and rearranging the resulting expression, we get

$$\langle \eta_t g_t, x_{t+1} - u\rangle \le \langle \nabla h(x_{t+1}) - \nabla h(x_t), u - x_{t+1}\rangle . \tag{4.11}$$

Then, we can write

$$\begin{aligned}\eta_t \langle g_t, x_t - u\rangle &= \langle \eta_t g_t, x_{t+1} - u\rangle + \langle \eta_t g_t, x_t - x_{t+1}\rangle\\ &\le \langle \nabla h(x_{t+1}) - \nabla h(x_t), u - x_{t+1}\rangle + \langle \eta_t g_t, x_t - x_{t+1}\rangle\\ &= D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1}) - D_h(x_{t+1} \,\|\, x_t) + \langle \eta_t g_t, x_t - x_{t+1}\rangle ,\end{aligned}$$

where the first inequality follows from Equation (4.11), and the last equality follows from the three-point identity (Lemma 4.2.8). This proves Equation (4.7).

Finally, we prove Equation (4.8). We can write

$$\begin{aligned}-D_h(x_{t+1} \,\|\, x_t) + \langle \eta_t g_t, x_t - x_{t+1}\rangle &\le -\frac{\mu}{2}\|x_{t+1} - x_t\|^2 + \langle \eta_t g_t, x_t - x_{t+1}\rangle\\ &\le -\frac{\mu}{2}\|x_{t+1} - x_t\|^2 + \eta_t\|g_t\|_*\|x_t - x_{t+1}\|\\ &\le \frac{\eta_t^2}{2\mu}\|g_t\|_*^2,\end{aligned}$$

where the first inequality follows from the strong convexity of $h$, the second inequality follows from Hölder's inequality, and the last inequality follows by optimizing a quadratic concave function. This completes the proof. □

**Theorem 4.3.3.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{V} \subseteq \operatorname{int}(\mathcal{X})$ be a nonempty closed convex set. Let $h : \mathcal{X} \to \mathbb{R}$ be proper, closed, differentiable, and $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathcal{V}$. Then,* Online Mirror Descent *with learning rates $\eta_1, \dots, \eta_T$ achieves*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \max_{1\le t\le T}\frac{D_h(u \,\|\, x_t)}{\eta_T} + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2, \quad \forall u \in \mathcal{V}. \tag{4.12}$$

*Moreover, if $\eta_t \coloneqq \eta$ for $t = 1, \dots, T$, then we have*

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \frac{D_h(u \,\|\, x_1)}{\eta} + \frac{\eta}{2\mu}\sum_{t=1}^{T}\|g_t\|_*^2, \quad \forall u \in \mathcal{V}. \tag{4.13}$$

*Proof.* Let $u \in \mathcal{V}$ be fixed, and let $D^2 \coloneqq \max_{1 \le t \le T} D_h(u \,\|\, x_t)$. By dividing $\eta_t > 0$ from both sides of Equation (4.8), we get

$$\ell_t(x_t) - \ell_t(u) \le \frac{D_h(u \,\|\, x_t)}{\eta_t} - \frac{D_h(u \,\|\, x_{t+1})}{\eta_t} + \frac{\eta_t}{2\mu}\|g_t\|_*^2. \tag{4.14}$$

Summing up Equation (4.14) over $t = 1, \dots, T$ gives

$$\begin{aligned}
\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\le \sum_{t=1}^{T}\left(\frac{D_h(u \,\|\, x_t)}{\eta_t} - \frac{D_h(u \,\|\, x_{t+1})}{\eta_t}\right) + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2 \\
&= \frac{D_h(u \,\|\, x_1)}{\eta_1} - \frac{D_h(u \,\|\, x_{T+1})}{\eta_T} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)D_h(x \,\|\, x_{t+1}) + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2 \\
&\le \frac{D_h(u \,\|\, x_1)}{\eta_1} + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)D_h(x \,\|\, x_{t+1}) + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2 \\
&\le \frac{D^2}{\eta_1} + D^2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2 \\
&= \frac{D^2}{\eta_T} + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2,
\end{aligned}$$

where the second inequality follows from the fact that $D_h(u \,\|\, x_{T+1}) \ge 0$, the third inequality follows from the bounded domain assumption, and the last equality follows by telescoping a sum. This proves Equation (4.12).

Now, consider the case when we have fixed learning rates $\eta_t = \eta > 0$ for $t = 1, \dots, T$. In the case, Equation (4.14) becomes

$$\ell_t(x_t) - \ell_t(u) \le \frac{1}{\eta}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})) + \frac{\eta}{2\mu}\|g_t\|_*^2. \tag{4.15}$$

Summing up Equation (4.15) over $t = 1, \dots, T$ yields

$$\begin{aligned}
\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) &\le \frac{1}{\eta}\sum_{t=1}^{T}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})) + \frac{\eta}{2\mu}\sum_{t=1}^{T}\|g_t\|_*^2 \\
&= \frac{D_h(u \,\|\, x_1)}{\eta} - \frac{D_h(u \,\|\, x_{T+1})}{\eta} + \frac{\eta}{2\mu}\sum_{t=1}^{T}\|g_t\|_*^2 \\
&\le \frac{D_h(u \,\|\, x_1)}{\eta} + \frac{\eta}{2\mu}\sum_{t=1}^{T}\|g_t\|_*^2,
\end{aligned}$$

where the last inequality follows from the fact that $D_h(u \,\|\, x_{T+1}) \ge 0$. This proves Equation (4.13). □

As mentioned earlier, Online Subgradient Descent is a special case of Online Mirror Descent when the distance-generating function is defined as $h(x) \coloneqq \frac{1}{2}\|x\|_2^2$. We can recover the regret guarantee of Online Subgradient Descent (Theorem 2.2.19) from Theorem 4.3.3. Let $\eta_1, \dots, \eta_T$ be a sequence of learning rate such that $\eta_{t+1} \le \eta_t$ for all $t$. Since $h$ is 1-strongly convex with respect to $\|\cdot\|_2$ (Example 3.1.9), by Theorem 4.3.3, we have

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \max_{1 \le t \le T}\frac{D_h(u \,\|\, x_t)}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2 \le \frac{D^2}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|g_t\|_2^2,$$

for all $u \in \mathcal{V}$, where $D > 0$ is the diameter of the feasible set $\mathcal{V}$. This recovers Equation (2.27). If $\eta_t \coloneqq \eta$, by Theorem 4.3.3, we have

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \frac{D_h(u \,\|\, x_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_*^2 = \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2,$$

for all $u \in \mathcal{V}$, which recovers Equation (2.28).

Let us find the optimal fixed step size for Online Mirror Descent. Assume that $D_h(u \,\|\, x_1) \le D^2$ for some $D > 0$ (i.e., bounded domain) and that $\|g_t\|_* \le L$ for some $L > 0$ (i.e., bounded subgradients). Then, from Theorem 4.3.3, we know that Online Mirror Descent achieve

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(u)) \le \frac{D^2}{\eta} + \frac{\eta T L^2}{2\mu}, \quad \forall u \in \mathcal{V}. \tag{4.16}$$

We can find the optimal step size $\eta^\star > 0$ by optimizing over $\eta$ on the RHS of Equation (4.16). Let $\varphi(\eta) \coloneqq D^2/\eta + \eta T L^2/(2\mu)$. Differentiate $\varphi(\eta)$ with respect to $\eta$, we get $\varphi'(\eta) = -D^2/\eta^2 + TL^2/(2\mu)$ and $\varphi''(\eta) = 2D^2/\eta^3 > 0$. Since $\varphi''(\eta) > 0$, we know that $\varphi(\eta)$ is convex. Thus, $\eta^\star$ must satisfy the condition $\varphi'(\eta^\star) = 0$. Solve for $\eta^\star$, we get $\eta^\star = \sqrt{2\mu D^2/(TL^2)}$. Substitute $\eta^\star$ into Equation (4.16), we get

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le \sqrt{\frac{2T}{\mu}} DL.$$

In conclusion, Online Mirror Descent gives a $O(\sqrt{T})$ regret bound.

## 4.4 The Mirror Interpretation

By now, we have introduced the Online Mirror Descent algorithm and proved its regret guarantee. In this section, we will see what "mirror" means in Online Mirror Descent. We begin by introducing the *Fenchel conjugate*.

**Definition 4.4.1** (Fenchel conjugate)**.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$. The **Fenchel conjugate** of $f$, denoted by $f^* : \mathbb{R}^d \to [-\infty, +\infty]$, is defined as*

$$f^*(\theta) \coloneqq \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - f(x)\}, \quad \forall \theta \in \mathbb{R}^d.$$

*The **biconjugate** of $f$ is defined as $f^{**} \coloneqq (f^*)^*$.*

The Fenchel conjugate is also referred to as the **Legendre–Fenchel transformation**, the **Legendre transformation** and the **convex conjugate**. First, let us see some examples of Fenchel conjugate.

**Example 4.4.2** (Inner product)**.** *Let $z \in \mathbb{R}^d \setminus \{0\}$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as*

$$f(x) \coloneqq \langle z, x \rangle, \quad \forall x \in \mathbb{R}^d.$$

*Then, the Fenchel conjugate of $f$ is given by*

$$f^*(\theta) = \begin{cases} 0, & \text{if } \theta = z; \\ +\infty, & \text{otherwise.} \end{cases}$$

*Proof.* We have

$$f^*(\theta) = \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - \langle z, x \rangle\} = \sup_{x \in \mathbb{R}^d} \langle \theta - z, x \rangle, \quad \forall \theta \in \mathbb{R}^d.$$

If $\theta = z$, then we have $\langle \theta - z, x \rangle = 0$ for all $x \in \mathbb{R}^d$. It follows that $f^*(\theta) = 0$ if $\theta = z$. On the other hand, if $\theta \neq z$, then the inner product $\langle \theta - z, x \rangle$ could be made arbitrarily large by setting $x = \alpha(\theta - z)$ for some large $\alpha > 0$. Thus, we have $f^*(\theta) = +\infty$ if $\theta \neq z$. □

**Example 4.4.3** (Squared norm)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}_+$ be defined as*

$$f(x) \coloneqq \frac{1}{2}\|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

*Then, the Fenchel conjugate of $f$ is given by*

$$f^*(\theta) = \frac{1}{2}\|\theta\|_*^2, \quad \forall \theta \in \mathbb{R}^d.$$

*Proof.* Let $\theta \in \mathbb{R}^d$. By Hölder's inequality, we can write

$$\langle \theta, x \rangle - \frac{1}{2}\|x\|^2 \le \|\theta\|_* \|x\| - \frac{1}{2}\|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

Note that $\|\theta\|_*\|x\| - \|x\|^2/2$ is a quadratic concave function of $\|x\|$, and its maximum is attained when $\|x\| = \|\theta\|_*$. Thus, we have

$$f^*(\theta) = \sup_{x \in \mathbb{R}^d} \left\{ \langle \theta, x \rangle - \frac{1}{2}\|x\|^2 \right\} \le \frac{1}{2}\|\theta\|_*^2. \tag{4.17}$$

On the other hand, let $x^\star \in \mathbb{R}^d$ such that $\langle \theta, x^\star \rangle = \|\theta\|_*\|x^\star\|$ and $\|x^\star\| = \|\theta\|_*$. Then, we have

$$f^*(\theta) = \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - f(x)\} \ge \langle \theta, x^\star \rangle - \frac{1}{2}\|x^\star\|^2 = \|\theta\|_*\|x^\star\| - \frac{1}{2}\|x^\star\|^2 = \frac{1}{2}\|\theta\|_*^2. \tag{4.18}$$

Equation (4.17) and Equation (4.18) together implies that $f^*(\theta) = \|\theta\|_*^2/2$ for all $\theta \in \mathbb{R}^d$. □

**Example 4.4.4** (Exponential function)**.** *Let Fenchel conjugate of the exponential function* $\exp : \mathbb{R} \to \mathbb{R}_{++}$ *is given by*

$$\exp^*(\theta) = \begin{cases} \theta \log(\theta) - \theta, & \text{if } \theta > 0; \\ 0, & \text{if } \theta = 0; \\ +\infty, & \text{otherwise.} \end{cases}$$

*Proof.* Consider the following cases.

- **Case 1:** $\theta = 0$**.** If $\theta = 0$, then we have $\exp^*(\theta) = \sup_{x \in \mathbb{R}} -\exp(x) = 0$.
- **Case 2:** $\theta > 0$**.** Let $\theta > 0$ be fixed. Consider the function $\varphi_\theta : \mathbb{R} \to \mathbb{R}$ defined as

  $$\varphi_\theta(x) \coloneqq \theta x - \exp(x), \quad \forall x \in \mathbb{R}.$$

  Differentiate $\varphi_\theta(x)$ with respect to $x$, we get $\varphi_\theta'(x) = \theta - \exp(x)$ and $\varphi_\theta''(x) = -\exp(x)$. Since $\varphi_\theta''(x) < 0$, we know that $\varphi_\theta(x)$ is concave. Denote by $x^\star$ the maximizer of $\varphi_\theta(x)$. We have $\varphi_\theta'(x^\star) = 0$ and thus $x^\star = \log(\theta)$. Thus, if $\theta > 0$, we have

  $$\exp^*(\theta) = \sup_{x \in \mathbb{R}} \{\theta x - \exp(x)\} = \sup_{x \in \mathbb{R}} \varphi_\theta(x) = \varphi_\theta(x^\star) = \theta \log(\theta) - \theta.$$

- **Case 3:** $\theta < 0$**.** If $\theta < 0$, then $\theta x - \exp(x) \to +\infty$ as $x \to +\infty$. Thus, we have $\exp^*(\theta) = +\infty$ if $\theta < 0$.

This completes the proof. □

**Example 4.4.5** (Young's inequality)**.** *Let* $p > 1$ *be fixed. Consider the function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *defined as*

$$f(x) = \frac{1}{p} x^p, \quad \forall x \in \mathbb{R}_+.$$

*Then, the Fenchel conjugate of* $f$ *is given by*

$$f^*(\theta) = \begin{cases} 0, & \text{if } \theta < 0; \\ \frac{1}{q} \theta^q, & \text{otherwise,} \end{cases}$$

*for all* $\theta \in \mathbb{R}$*, where* $q > 1$ *is the constant such that* $1/p + 1/q = 1$.

**Lemma 4.4.6** (Bauschke and Combettes [BC], Proposition 13.9)**.** *Let* $f : \mathbb{R}^d \to [-\infty, +\infty]$*. Then, we have* $-\infty \in f^*(\mathbb{R}^d)$ *if and only if* $f \equiv +\infty$*, if and only if* $f^* \equiv -\infty$*.*

*Proof.* We have $-\infty \in f^*(\mathbb{R}^d)$ if and only if there exists some $\theta \in \mathbb{R}^d$ such that

$$-\infty = f^*(\theta) \geq \langle \theta, x \rangle - f(x), \quad \forall x \in \mathbb{R}^d. \tag{4.19}$$

Note that Equation (4.19) holds if and only if $f(x) = +\infty$ for all $x \in \mathbb{R}^d$. Clearly, we have $f \equiv +\infty$ if and only if $f^* \equiv -\infty$. □

**Lemma 4.4.7** (Bauschke and Combettes [BC], Proposition 13.9)**.** *Let* $f : \mathbb{R}^d \to [-\infty, +\infty]$*. If* $f^*$ *is proper, then* $f$ *is proper.*

*Proof.* The proof is by contradiction. Let $f : \mathbb{R}^d \to [-\infty, +\infty]$ such that $f^*$ is proper. Suppose the contrary that $f$ is not proper. Then, we either have $f \equiv +\infty$ or $f(y) = -\infty$ for some $y \in \mathbb{R}^d$. We consider these two cases separately.

- **Case 1:** $f \equiv +\infty$**.** If $f \equiv +\infty$, from Lemma 4.4.6, we know that $f^* \equiv -\infty$, which contradicts the assumption that $f^*$ is proper.
- **Case 2:** $f(y) = -\infty$**.** If $f(y) = -\infty$ for some $y \in \mathbb{R}^d$, then we have $\langle \theta, y \rangle - f(y) = +\infty$ for all $\theta \in \mathbb{R}^d$. It follows that $f^*(\theta) = +\infty$ for all $\theta \in \mathbb{R}^d$, which contradicts the assumption that $f^*$ is proper.

Since both cases lead to a contradiction, we conclude that $f$ must be proper if $f^*$ is proper. □

The converse statement of Lemma 4.4.7 is not true in general. For example, consider the function $f : \mathbb{R}^d \to (-\infty, +\infty]$ defined as $f(x) = \|x\|^2$ for all $x \in \mathbb{R}^d$, where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$. The Fenchel conjugate of $f$ is given by

$$f^*(\theta) = \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle + \|x\|^2\}, \quad \forall \theta \in \mathbb{R}^d.$$

For any fixed $\theta \in \mathbb{R}^d$, the function $\langle \theta, x \rangle + \|x\|^2 \to +\infty$ as $\|x\| \to +\infty$. Therefore, we have $f^*(\theta) = +\infty$ for all $\theta \in \mathbb{R}^d$ and thus $f^*$ is improper.

**Lemma 4.4.8** (Bauschke and Combettes [BC], Proposition 13.14)**.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$. Then, we have $f^{**}(x) \leq f(x)$ for all $x \in \mathbb{R}^d$.*

*Proof.* It follows immediately from the definition of Fenchel conjugate that $f^*(\theta) \geq \langle \theta, x \rangle - f(x)$ for all $x \in \mathbb{R}^d$. Then, we can write

$$f^{**}(x) = \sup_{\theta \in \mathbb{R}^d} \{\langle x, \theta \rangle - f^*(\theta)\} \leq \sup_{\theta \in \mathbb{R}^d} f(x) = f(x),$$

for all $x \in \mathbb{R}^d$. □

**Lemma 4.4.9.** *Let $f_1 : \mathbb{R}^d \to [-\infty, +\infty]$ and $f_2 : \mathbb{R}^d \to [-\infty, +\infty]$ be functions such that $f_1(x) \leq f_2(x)$ for all $x \in \mathbb{R}^d$. Then, we have $f_1^*(\theta) \geq f_2^*(\theta)$ for all $\theta \in \mathbb{R}^d$.*

*Proof.* Let $\theta \in \mathbb{R}^d$. Since $f_1(x) \leq f_2(x)$ for all $x \in \mathbb{R}^d$, we can write

$$f_1^*(\theta) = \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - f_1(x)\} \geq \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - f_2(x)\} = f_2^*(\theta).$$

Therefore, $f_1^*(\theta) \geq f_2^*(\theta)$ for all $\theta \in \mathbb{R}^d$. □

**Lemma 4.4.10.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$, let $a > 0, b \in \mathbb{R}$, and let $g \in \mathbb{R}^d$. Define $h : \mathbb{R}^d \to [-\infty, +\infty]$ by*

$$h(x) \coloneqq af(x) + \langle g, x \rangle + b, \quad \forall x \in \mathbb{R}^d.$$

*Then, we have*

$$h^*(\theta) = af^*\left(\frac{\theta - g}{a}\right) - b, \quad \forall \theta \in \mathbb{R}^d.$$

*Proof.* Let $\theta \in \mathbb{R}^d$. We have

$$\begin{aligned} h^*(\theta) &= \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - h(x)\} = \sup_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - af(x) - \langle g, x \rangle - b\} \\ &= \sup_{x \in \mathbb{R}^d} \{\langle \theta - g, x \rangle - af(x)\} - b \\ &= a \cdot \sup_{x \in \mathbb{R}^d} \left\{ \left\langle \frac{\theta - g}{a}, x \right\rangle - f(x) \right\} - b \\ &= af^*\left(\frac{\theta - g}{a}\right) - b. \end{aligned}$$

□

**Theorem 4.4.11** (Bauschke and Combettes [BC], Proposition 13.11)**.** *Let $f : \mathbb{R}^d \to [-\infty, +\infty]$. Then, $f^*$ is closed and convex.*

**Theorem 4.4.12** (Fenchel–Young inequality)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ is proper. Then, we have*

$$\langle \theta, x \rangle \leq f^*(\theta) + f(x), \quad \forall \theta, x \in \mathbb{R}^d.$$

*Proof.* Let $\theta, x \in \mathbb{R}^d$. Since $f$ is proper, we know that $f$ is not identically $+\infty$. It follows from Lemma 4.4.6 that $f^*(\theta) > -\infty$. From the definition of Fenchel conjugate, we immediately have $f^*(\theta) \geq \langle \theta, x \rangle - f(x)$. Rearranging, we obtain the desired inequality. □

Let us take a closer look at why properness is needed in Fenchel–Young inequality (Theorem 4.4.12). Consider the function $f : \mathbb{R}^d \to [-\infty, +\infty]$ defined as $f \equiv +\infty$. From Lemma 4.4.6, we know that the Fenchel conjugate of $f$ is given by $f^* \equiv -\infty$. Then, for all $x, \theta \in \mathbb{R}^d$, we have

$$f(x) + f^*(\theta) = (+\infty) + (-\infty),$$

which is an undefined expression. Thus, Fenchel–Young inequality may not hold for improper functions. Properness is assumed in the Fenchel–Young inequality to avoid undefined additions in the extended reals such as $(+\infty) + (-\infty)$ and $(-\infty) + (+\infty)$.

**Theorem 4.4.13** (Bauschke and Combettes [BC], Proposition 16.9)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and let $\theta, x \in \mathbb{R}^d$. The following statements hold.*

*(a) We have $\theta \in \partial f(x)$ if and only if $f(x) + f^*(\theta) = \langle x, \theta \rangle$.*

*(b) If $f(x) + f^*(\theta) = \langle x, \theta \rangle$, then $x \in \partial f^*(\theta)$.*

*Proof of (a).* Since $f$ is proper, by the Fenchel–Young inequality (Theorem 4.4.12), we already have $\langle x, \theta \rangle \leq f(x) + f^*(\theta)$. It remains to prove that $\theta \in \partial f(x)$ if and only if $f(x) + f^*(\theta) \leq \langle x, \theta \rangle$. We have $\theta \in \partial f(x)$ if and only if

$$\langle x, \theta \rangle \geq f(x) + \langle \theta, y \rangle - f(y), \quad \forall y \in \mathbb{R}^d. \tag{4.20}$$

Equation (4.20) holds if and only if

$$\langle x, \theta \rangle \geq f(x) + \sup_{y \in \mathbb{R}^d} \{\langle \theta, y \rangle - f(y)\} = f(x) + f^*(\theta).$$

Therefore, we have $\theta \in \partial f(x)$ if and only if $f(x) + f^*(\theta) = \langle x, \theta \rangle$. □

*Proof of (b).* Suppose that $f(x) + f^*(\theta) = \langle x, \theta \rangle$. We can write

$$\langle x, u \rangle - f^*(u) + f^*(\theta) \leq f^{**}(x) + f^*(\theta) \leq f(x) + f^*(\theta) = \langle x, \theta \rangle, \quad \forall u \in \mathbb{R}^d, \tag{4.21}$$

where the first inequality follows from the definition of the biconjugate, and the second inequality follows from Lemma 4.4.8. Rearranging Equation (4.21) gives

$$f^*(u) \geq f^*(\theta) + \langle x, u - \theta \rangle, \quad \forall u \in \mathbb{R}^d.$$

This shows that $x \in \partial f^*(\theta)$. □

**Theorem 4.4.14** (Fenchel–Moreau theorem)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper. Then, $f$ is closed and convex if and only if $f = f^{**}$. In this case, $f^*$ is also proper.*

*Proof.* For the first statement, see Theorem 13.32 in Bauschke and Combettes [BC].

Now, assume that $f$ is closed and convex. From above, we know that $f = f^{**}$. Since $f$ is proper, it follows immediately that $f^{**}$ is also proper. Then, from Lemma 4.4.7, we know that $f^*$ is also proper. □

**Corollary 4.4.15.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper closed convex function and let $x, \theta \in \mathbb{R}^d$. Then, we have $\theta \in \partial f(x)$ if and only if $f(x) + f^*(\theta) = \langle x, \theta \rangle$, if and only if $x \in \partial f^*(\theta)$.*

*Proof.* From Theorem 4.4.13, we already know that $\theta \in \partial f(x)$ if and only if $f(x) + f^*(\theta) = \langle x, \theta \rangle$. It remains to prove the second statement.

Since $f$ is a proper closed convex function, from the Fenchel–Moreau theorem (Theorem 4.4.14), we know that $f^*$ is proper. Then, we can write

$$\langle \theta, x \rangle \leq f^*(\theta) + f^{**}(x) \leq f^*(\theta) + f(x), \tag{4.22}$$

where the first inequality follows from the Fenchel–Young inequality (Theorem 4.4.12), and the second inequality follows from Lemma 4.4.8. From Equation (4.22), we can see that $f(x) + f^*(\theta) = \langle x, \theta \rangle$ if and only if $f^*(\theta) + f^{**}(x) = \langle \theta, x \rangle$. Since $f^*$ is proper, from part (a) of Theorem 4.4.13, we know that $f^*(\theta) + f^{**}(x) = \langle \theta, x \rangle$ if and only if $x \in \partial f^*(\theta)$. Therefore, we conclude that $f(x) + f^*(\theta) = \langle x, \theta \rangle$ if and only if $x \in \partial f^*(\theta)$. □

We can view the subdifferential $\partial f$ as a set-valued mapping from $\operatorname{dom}(\partial f)$ to $\mathbb{R}^d$. Then, for a proper closed convex function $f$, Corollary 4.4.15 tells us that the inverse of $\partial f$ is simply $\partial f^*$, i.e., $(\partial f)^{-1} = \partial f^*$.

**Theorem 4.4.16.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper closed convex function. Suppose that $\operatorname{dom}(\partial f)$ is nonempty. Then, $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathbb{R}^d$ if and only if $f^*$ is differentiable and $1/\mu$-smooth with respect to $\|\cdot\|_*$ on $\mathbb{R}^d$.*

*Proof.* Suppose that $f$ is $\mu$-strongly convex with respect to $\|\cdot\|$ on $\mathbb{R}^d$. First, we show that $f^*$ is (Gâteaux) differentiable. Let $\theta \in \mathbb{R}^d$. Since $f(x) - \langle \theta, x \rangle$ is a proper closed $\mu$-strongly convex function and $\operatorname{dom}(\partial f)$ is nonempty, from Theorem 3.1.14, we know that $x^\star = \arg\min_{x \in \mathbb{R}^d} f(x) - \langle \theta, x \rangle$ exists and is unique. Note that

$$x^\star = \arg\min_{x \in \mathbb{R}^d} \{f(x) - \langle \theta, x \rangle\} = \arg\max_{x \in \mathbb{R}^d} \{\langle \theta, x \rangle - f(x)\}.$$

Thus, we can write $f(x^\star) + f^*(\theta) = \langle x^\star, \theta \rangle$. It follows from part (b) of Theorem 4.4.13 that $x^\star \in \partial f^*(\theta)$. Suppose that $x' \in \partial f^*(\theta)$. Since $f$ is a proper closed convex function, from Corollary 4.4.15, we know that $x' \in \partial f^*(\theta)$ if and only if $f(x') + f^*(\theta) = \langle x', \theta \rangle$. Thus, we have $x' \in \arg\max_{x \in \mathbb{R}^d} \langle \theta, x \rangle - f(x)$. By the uniqueness of $x^\star$, we deduce that $x' = x^\star$. Now, since $\partial f^*(\theta) = \{x^\star\}$ is singleton, from Theorem 2.2.9, we know that $f^*$ is (Gâteaux) differentiable at $\theta$. Since $\theta \in \mathbb{R}^d$ is arbitrary, we deduce that $f^*$ is (Gâteaux) differentiable on $\mathbb{R}^d$. Next, we show that $f^*$ is $1/\mu$-smooth with respect to $\|\cdot\|_*$ on $\mathbb{R}^d$. Let $\theta_1, \theta_2 \in \mathbb{R}^d$ and let $x_1 = \nabla f^*(\theta_1), x_2 = \nabla f^*(\theta_2)$. Since $f$ is a proper closed convex function and $x_1 = \nabla f^*(\theta_1)$, it follows from

Corollary 4.4.15 that $\theta_1 \in \partial f(x_1)$. Likewise, we have $\theta_2 \in \partial f^*(x_2)$. Then, by the first-order condition for strong convexity (Theorem 3.1.4), we can write

$$f(x_2) \geq f(x_1) + \langle \theta_1, x_2 - x_1 \rangle + \frac{\mu}{2}\|x_2 - x_1\|^2, \tag{4.23}$$

and

$$f(x_1) \geq f(x_2) + \langle \theta_2, x_1 - x_2 \rangle + \frac{\mu}{2}\|x_1 - x_2\|^2. \tag{4.24}$$

Summing up Equation (4.23) and Equation (4.24) gives

$$\begin{aligned}\mu\|\nabla f^*(\theta_1) - \nabla f^*(\theta_2)\|^2 = \mu\|x_1 - x_2\|^2 &\leq \langle \theta_1 - \theta_2, x_1 - x_2 \rangle \\ &\leq \|\theta_1 - \theta_2\|_*\|x_1 - x_2\| && \text{(Hölder's ineq.)} \\ &= \|\theta_1 - \theta_2\|_*\|\nabla f^*(x_1) - \nabla f^*(x_2)\|.\end{aligned}$$

Rearranging, we get

$$\|\nabla f^*(\theta_1) - \nabla f^*(\theta_2)\| \leq \frac{1}{\mu}\|\theta_1 - \theta_2\|.$$

Therefore, $f^*$ is $1/\mu$-smooth with respect to $\|\cdot\|_*$ on $\mathbb{R}^d$. □

We are now ready to see the "mirror" interpretation of Online Mirror Descent.

**Theorem 4.4.17.** *Let $h : \mathcal{X} \to \mathbb{R}$ be proper, closed, $\mu$-strongly convex, and differentiable. Let $\mathcal{V} \subseteq \mathcal{X}$ be a nonempty closed convex set and let $x_t \in \mathcal{V}$. Given $g_t \in \mathbb{R}^d$, define*

$$x_{t+1} \coloneqq \operatorname*{arg\,min}_{x \in \mathcal{V}} \left\{ \langle g_t, x \rangle + \frac{1}{\eta_t} D_h(x \,\|\, x_t) \right\}.$$

*Then, we have*

$$x_{t+1} = \nabla h_{\mathcal{V}}^*(\nabla h(x_t) - \eta_t g_t),$$

*where $h_{\mathcal{V}} \coloneqq h + \iota_{\mathcal{V}}$ is the restriction of $h$ on $\mathcal{V}$.*

*Proof.* We can write

$$\begin{aligned}\operatorname*{arg\,min}_{x \in \mathcal{V}} \left\{ \langle g_t, x \rangle + \frac{1}{\eta_t} D_h(x \,\|\, x_t) \right\} &= \operatorname*{arg\,min}_{x \in \mathcal{V}} \{ \langle \eta_t g_t, x \rangle + D_h(x \,\|\, x_t) \} \\ &= \operatorname*{arg\,min}_{x \in \mathcal{V}} \{ \langle \eta_t g_t, x \rangle + h(x) - h(x_t) - \langle \nabla h(x_t), x - x_t \rangle \} \\ &= \operatorname*{arg\,min}_{x \in \mathcal{V}} \{ \langle \eta_t g_t - \nabla h(x_t), x \rangle + h(x) \} \\ &= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \{ \langle \eta_t g_t - \nabla h(x_t), x \rangle + h(x) + \iota_{\mathcal{V}}(x) \} \\ &= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \{ \langle \eta_t g_t - \nabla h(x_t), x \rangle + h_{\mathcal{V}}(x) \}.\end{aligned}$$

Note that $\langle \eta_t g_t - \nabla h(x_t), x \rangle + h_{\mathcal{V}}(x)$ is a proper function. By Fermat's rule (Theorem 2.2.14), we have $0 \in \eta_t g_t - \nabla h(x_t) + \partial h_{\mathcal{V}}(x_{t+1})$. Rearranging, we get $\nabla h(x_t) - \eta_t g_t \in \partial h_{\mathcal{V}}(x_{t+1})$. Since $h_{\mathcal{V}}$ is a proper closed convex function, it follows from Corollary 4.4.15 that $x_{t+1} \in \partial h_{\mathcal{V}}^*(\nabla h(x_t) - \eta_t g_t)$ and from Theorem 4.4.16 that $h_{\mathcal{V}}^*$ is differentiable. Since $h_{\mathcal{V}}^*$ is differentiable and $x_{t+1} \in \partial h_{\mathcal{V}}^*(\nabla h(x_t) - \eta_t g_t)$, from Theorem 2.2.9, we conclude that $x_{t+1} = \nabla h_{\mathcal{V}}^*(\nabla h(x_t) - \eta_t g_t)$. □

Consider Online Mirror Descent with a closed convex feasible set $\mathcal{V} \subseteq \mathcal{X}$ and a proper closed $\mu$-strongly convex differentiable distant-generating function $h : \mathcal{X} \to \mathbb{R}$. Theorem 4.4.17 tells us that the predictions of Online Mirror Descent are obtained as follows.

(i) First, map the last prediction $x_t$ into the "dual space" by the gradient mapping $\nabla h : \mathcal{X} \to \mathbb{R}^d$.

(ii) Next, perform a subgradient descent step, i.e., set $\tilde{x}_{t+1} \coloneqq \nabla h(x_t) - \eta_t g_t$.

(iii) Finally, map $\tilde{x}_{t+1}$ back to the feasible set $\mathcal{V}$ by the "mirror map" $\nabla h_{\mathcal{V}}^* : \mathbb{R}^d \to \mathcal{V}$ and set $x_{t+1} \coloneqq \nabla h_{\mathcal{V}}^*(\tilde{x}_{t+1})$.

## 4.5 A Two-Step Update

In the preceding section (Section 4.4), we have analyzed Online Mirror Descent with a dual perspective. In this section, we will see yet another interpretation of Online Mirror Descent. We will show that the one-step updating rule of Online Mirror Descent

$$x_{t+1} = \arg\min_{x \in \mathcal{V}} \left\{ \langle g_t, x \rangle + \frac{1}{\eta_t} D_h(x \,\|\, x_t) \right\},$$

is equivalent the following two-step updating rule:

$$\tilde{x}_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ \langle g_t, x \rangle + \frac{1}{\eta_t} D_h(x \,\|\, x_t) \right\}, \tag{4.25}$$

$$x_{t+1} = \arg\min_{x \in \mathcal{V}} D_h(x \,\|\, \tilde{x}_{t+1}). \tag{4.26}$$

We begin by defining *Bregman projection*.

**Definition 4.5.1** (Bregman projection)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set. Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, closed, strictly convex and differentiable on* $\operatorname{int}(\operatorname{dom}(h))$*. The* ***Bregman projection*** *onto $\mathcal{V}$ with respect to $h$ is defined as*

$$\operatorname{proj}_{\mathcal{V},h}(x) \coloneqq \arg\min_{y \in \mathcal{V}} D_h(y \,\|\, x), \quad \forall x \in \mathbb{R}^d.$$

If $h(x) \coloneqq \|x\|_2^2/2$ for all $x \in \mathbb{R}^d$, then from Example 4.2.5, we know that $\operatorname{proj}_{\mathcal{V},h}(x) = \arg\min_{y \in \mathcal{V}} \|y - x\|_2^2/2$ for all $x \in \mathbb{R}^d$. Namely, the Bregman projection onto $\mathcal{V}$ with respect to $h$ is simply the Euclidean projection (Definition 2.1.17) onto $\mathcal{V}$. This shows that Bregman projection is a strict generalization of Euclidean projection.

**Theorem 4.5.2.** *Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, closed, strictly convex function that is differentiable on* $\operatorname{int}(\operatorname{dom}(h))$*. Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a nonempty closed convex set such that $\mathcal{V} \cap \operatorname{dom}(h)$ is nonempty. Assume that $\tilde{y} \in \arg\min_{z \in \mathbb{R}^d} h(z)$ exists and that $\tilde{y} \in \operatorname{int}(\operatorname{dom}(h))$. Denote by $y' \in \arg\min_{z \in \mathcal{V}} D_h(z, \tilde{y})$. Then, the following statements hold:*

*(a) $y \in \arg\min_{z \in \mathcal{V}} h(z)$ exists and is unique.*

*(b) We have $y = y'$.*

*Proof of (a).* The existence of $y$ follows from Theorem 3.1.12. The uniqueness of $y$ follows from the strict convexity of $h$. □

*Proof of (b).* From part (a), we know that $y \in \arg\min_{z \in \mathcal{V}} h(z)$ is unique. Thus, to prove that $y = y'$, it suffices to show that $h(y) = h(y')$. Since $y \in \arg\min_{z \in \mathcal{V}} h(z)$ and $y' \in \mathcal{V}$, it follows immediately that $h(y) \leq h(y')$.

On the other hand, since $y' \in \arg\min_{z \in \mathcal{V}} D_h(z \,\|\, \tilde{y})$ and $y \in \mathcal{V}$, we can write

$$\begin{aligned} 0 &\leq D_h(y \,\|\, \tilde{y}) - D_h(y' \,\|\, \tilde{y}) \\ &= h(y) - h(\tilde{y}) - \langle \nabla h(\tilde{y}), y - \tilde{y} \rangle - h(y') + h(\tilde{y}) + \langle \nabla h(\tilde{y}), y' - \tilde{y} \rangle \\ &= h(y) - h(y') + \langle \nabla h(\tilde{y}), y' - y \rangle . \end{aligned}$$

Rearrange the above inequality, we obtain

$$h(y') + \langle \nabla h(\tilde{y}), y - y' \rangle \leq h(y). \tag{4.27}$$

Since $\tilde{y} = \arg\min_{z \in \mathbb{R}^d} h(z)$ and $h$ is differentiable at $\tilde{y}$, by Theorem 2.2.9 and the Fermat's rule (Theorem 2.2.14), we know that $\nabla h(\tilde{y}) = 0$. Thus, inequality (4.27) becomes $h(y') \leq h(y)$. Therefore, we have $h(y) = h(y')$. It follows from the uniqueness of $y$ that $y = y'$. □

## 4.6 Bounding Online Mirror Descent with Local Norms

As before, we assume that the iterates $\{x_t\}_{t=1}^T$ of Online Mirror Descent is well-defined in the following discussion. Recall from Example 3.3.4 that if $A \in \mathbb{R}^{d \times d}$ is a positive-definite matrix, then $\|\cdot\|_A$ defines a norm on $\mathbb{R}^d$ and the dual norm of $\|\cdot\|_A$ is $\|\cdot\|_{A^{-1}}$.

**Lemma 4.6.1.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed convex set and let $h : \mathcal{X} \to \mathbb{R}$ be a proper closed convex closed function that is twice-differentiable with positive-definite Hessian on $\operatorname{int}(\mathcal{X})$. Then, we have*

$$\eta_t(\ell_t(x_t) - \ell_t(u)) \leq \eta_t \langle g_t, x_t - u \rangle \tag{4.28}$$

$$\leq D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1}) + \frac{\eta_t^2}{2} \min\left\{ \|g_t\|^2_{(\nabla^2 h(z_t))^{-1}}, \|g_t\|^2_{(\nabla^2 h(z_t'))^{-1}} \right\}, \tag{4.29}$$

*for all $u \in \mathcal{X}$, where $z_t$ is a point on the line segment between $x_t$ and $x_{t+1}$, and $z_t'$ is a point on the line segment between $x_t$ and $\tilde{x}_{t+1}$.*

*Proof.* From Lemma 4.3.2, we already have

$$\begin{aligned}\eta_t(\ell_t(x_t) - \ell_t(u)) &\leq \eta_t \langle g_t, x_t - u \rangle \\ &\leq D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1}) - D_h(x_{t+1} \,\|\, x_t) + \langle \eta_t g_t, x_t - x_{t+1} \rangle\end{aligned}$$

This proves Equation (4.28), and it remains to prove that

$$\langle \eta_t g_t, x_t - x_{t+1} \rangle - D_h(x_{t+1} \,\|\, x_t) \leq \frac{\eta_t^2}{2} \min\left\{ \|g_t\|^2_{(\nabla^2 h(z_t))^{-1}}, \|g_t\|^2_{(\nabla^2 h(z_t'))^{-1}} \right\}. \tag{4.30}$$

By Taylor's theorem, there exists some $z_t$ lying on the line segment between $x_t$ and $x_{t+1}$ such that

$$D_h(x_{t+1} \,\|\, x_t) = h(x_{t+1}) - h(x_t) - \langle \nabla h(x_t), x_{t+1} - x_t \rangle = \frac{1}{2}(x_{t+1} - x_t)^\top \nabla^2 h(z_t)(x_{t+1} - x_t). \tag{4.31}$$

Since $\nabla^2 h(z_t)$ is assumed to be positive-definite, we know that $\nabla^2 h(z_t)$ define a norm on $\mathbb{R}^d$. Thus, we can write

$$\frac{1}{2}(x_{t+1} - x_t)^\top \nabla^2 h(z_t)(x_{t+1} - x_t) = \frac{1}{2}\|x_{t+1} - x_t\|^2_{\nabla^2 h(z_t)}. \tag{4.32}$$

Then, by Equation (4.31) and Equation (4.32), we can write

$$\begin{aligned}\langle \eta_t g_t, x_t - x_{t+1} \rangle - D_h(x_{t+1} \,\|\, x_t) &= \langle \eta_t g_t, x_t - x_{t+1} \rangle - \frac{1}{2}\|x_{t+1} - x_t\|^2_{\nabla^2 h(z_t)} \\ &\leq \sup_{x \in \mathbb{R}^d} \left\{ \langle \eta_t g_t, x \rangle - \frac{1}{2}\|x\|^2_{\nabla^2 h(z_t)} \right\} \\ &= \frac{1}{2}\|\eta_t g_t\|^2_{(\nabla^2 h(z_t))^{-1}} \\ &= \frac{\eta_t^2}{2}\|g_t\|^2_{(\nabla^2 h(z_t))^{-1}},\end{aligned} \tag{4.33}$$

where the second equality follows Example 4.4.3. Now, recall that from the two-step updating rule of Online Mirror Descent (Equation (4.25)) that $\tilde{x}_{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \langle \eta_t g_t, x \rangle + D_h(x \,\|\, x_t)$. Thus, we can write

$$\langle \eta_t g_t, x_t - x_{t+1} \rangle - D_h(x_{t+1} \,\|\, x_t) \leq \langle \eta_t g_t, x_t - \tilde{x}_{t+1} \rangle - D_h(\tilde{x}_{t+1} \,\|\, x_t). \tag{4.34}$$

Again, by Taylor's theorem, we know that there exists some $z_t'$ lying on the line segment between $x_t$ and $\tilde{x}_{t+1}$ such that

$$D_h(\tilde{x}_{t+1} \,\|\, x_t) = \frac{1}{2}(\tilde{x}_{t+1} - x_t)^\top \nabla^2 h(z_t')(\tilde{x}_{t+1} - x_t). \tag{4.35}$$

Putting Equation (4.34) and Equation (4.35) together, we can write

$$\begin{aligned}\langle \eta_t g_t, x_t - x_{t+1} \rangle - D_h(x_{t+1} \,\|\, x_t) &\leq \langle \eta_t g_t, x_t - \tilde{x}_{t+1} \rangle - D_h(\tilde{x}_{t+1} \,\|\, x_t) \\ &= \langle \eta_t g_t, x_t - \tilde{x}_{t+1} \rangle - \frac{1}{2}\|\tilde{x}_{t+1} - x_t\|^2_{\nabla^2 h(z_t')} \\ &\leq \sup_{x \in \mathbb{R}^d} \left\{ \langle \eta_t g_t, x \rangle - \frac{1}{2}\|x\|^2_{\nabla^2 h(z_t')} \right\} \\ &= \frac{\eta_t^2}{2}\|g_t\|^2_{(\nabla^2 h(z_t'))^{-1}},\end{aligned} \tag{4.36}$$

where the first equality follows from the positive-definiteness of $\nabla^2 h(z_t')$, and the last equality follows from Example 4.4.3. Finally, Equation (4.33) and Equation (4.36) together proves Equation (4.30). This completes the proof. □

## 4.7 Entropic Mirror Descent

In this section, we introduce the Entropic Mirror Descent algorithm. Let $\Delta_d$ denote the $(d-1)$-dimensional probability simplex in $\mathbb{R}^d$, that is, $\Delta_d = \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\}$, and let $h : \mathbb{R}^d_+ \to \mathbb{R}$ denote the (normalized) negative Shannon entropy on $\Delta_d$, that is,

$$h(x) = \sum_{i=1}^{d} x_i \log(x_i), \quad \forall x \in \Delta_d.$$

The Entropic Mirror Descent algorithm is as follows.

**Algorithm 6** Entropic Mirror Descent

**Inputs:** The probability simplex $\Delta_d$, the (normalized) negative Shannon entropy $h : \Delta_d \to \mathbb{R}$, and a fixed step size $\eta > 0$.

1: Set $x_1 \leftarrow (1/d, \ldots, 1/d)^\top$.
2: **for** $t = 1, \ldots, T$ **do**
3: The learner outputs $x_t \in \Delta_d$.
4: The adversary chooses a convex subdifferentiable loss function $\ell_t : \Delta_d \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
5: The learner calls the first-order oracle for $g_t \in \partial \ell_t(x_t)$.
6: The learner updates by setting

$$x_{t+1} \leftarrow \arg\min_{x \in \Delta_d} \left\{ \langle g_t, x \rangle + \frac{1}{\eta} D_h(x \,\|\, x_t) \right\}.$$

7: **end for**

**Lemma 4.7.1.** *The probability simplex $\Delta_d = \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\}$ is a closed convex set of $\mathbb{R}^d$.*

*Proof.* First, we prove that $\Delta_d$ is convex. Let $x, y \in \Delta_d$ and let $\lambda \in [0, 1]$. Since $x_i \geq 0$ and $y_i \geq 0$ for $i = 1, \ldots, d$, we clearly have $\lambda x_i + (1-\lambda) y_i \geq 0$ for all $i$. Moreover, we have

$$\sum_{i=1}^{d} \lambda x_i + (1-\lambda) y_i = \lambda \sum_{i=1}^{d} x_i + (1-\lambda) \sum_{i=1}^{d} y_i = \lambda + (1-\lambda) = 1.$$

Thus, $\lambda x + (1-\lambda) y \in \Delta_d$ and $\Delta_d$ is convex. □

**Lemma 4.7.2.** *The (normalized) negative Shannon entropy $h : \Delta_d \to \mathbb{R}$ defined as*

$$h(x) = \sum_{i=1}^{d} x_i \log(x_i), \quad \forall x \in \Delta_d,$$

*is 1-strongly convex with respect to the $\ell_1$-norm.*

**Theorem 4.7.3.** Entropic Mirror Descent *achieves*

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_\infty^2, \quad \forall u \in \Delta_d.$$

*Proof.* Let $u \in \Delta_d$. From Lemma 4.7.1, we know that $\Delta_d$ is closed and convex. From Lemma 4.7.2, we know that the negative Shannon entropy is 1-strongly convex with respect to $\|\cdot\|_1$. From Example 3.3.3, we know that the dual norm of the $\ell_1$-norm is the $\ell_\infty$-norm. Thus, from Theorem 4.3.3, we have

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \leq \frac{D_h(u \,\|\, x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_\infty^2. \tag{4.37}$$

We can write

$$D_h(u \,\|\, x_1) = \sum_{i=1}^{d} u_i \log(d u_i) = \log(d) + \sum_{i=1}^{d} u_i \log(u_i) \leq \log(d), \tag{4.38}$$

where the first equality follows from Example 4.2.6, the second equality follows from the fact $u \in \Delta_d$, and the last equality follows from the fact that $u_i \log(u_i) \leq 0$ for $i = 1, \ldots, d$. Equation (4.37) and Equation (4.38) together yields

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_\infty^2.$$

□

Let us find the optimal fixed step size for Entropic Mirror Descent. Let $\eta_\infty^\star$ denote the optimal step size. Assume that $\|g_t\|_\infty \le L_\infty$ for $t = 1, \dots, T$ for some constant $L_\infty > 0$. From Theorem 4.7.3, we have

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le \frac{\log(d)}{\eta} + \frac{\eta T L_\infty^2}{2}, \quad \forall u \in \Delta_d. \tag{4.39}$$

Let $\varphi(\eta) \coloneqq \log(d)/\eta + \eta T L_\infty^2/2$. Differentiate $\varphi(\eta)$ with respect to $\eta$, we get $\varphi'(\eta) = -\log(d)/\eta^2 + TL_\infty^2/2$ and $\varphi''(\eta) = 2\log(d)/\eta^3 > 0$. Since $\varphi''(\eta) > 0$, we know that $\varphi(\eta)$ is a convex function. Thus, $\eta_\infty^\star$ must satisfy the condition $\varphi'(\eta_\infty^\star) = 0$. Solve for $\eta_\infty^\star$, we get $\eta_\infty^\star = \sqrt{2\log(d)/(TL_\infty^2)}$. Substitute $\eta_\infty^\star$ into Equation (4.39), we get

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le \sqrt{2\log(d) T L_\infty^2}, \quad \forall u \in \Delta_d. \tag{4.40}$$

We could also run Online Subgradient Descent in this setting. Assume that $\|g_t\|_\infty \le L_\infty$ for $t = 1, \dots, T$ for some constant $L_\infty > 0$. Then, we have

$$\|g_t\|_2^2 = \sum_{i=1}^{d} g_{t,i}^2 \le \sum_{i=1}^{d} \|g_t\|_\infty^2 = dL_\infty^2, \quad \forall 1 \le t \le T.$$

Set $x_1 = (1/d, \dots, 1/d)^\top \in \Delta_d$. Then, we have $\|u - x_1\|_2 \le 2$ for all $u \in \Delta_d$. From Theorem 2.2.19, we have

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le \frac{\|u - x_1\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_2^2 \le \frac{2}{\eta} + \frac{\eta d T L_\infty^2}{2}, \tag{4.41}$$

for all $u \in \Delta_d$. Next, we find the optimal learning rate $\eta_2^\star$ for Online Subgradient Descent under this setting. Let $\varphi(\eta) \coloneqq 2/\eta + \eta d T L_\infty^2/2$. Differentiate $\varphi(\eta)$ with respect to $\eta$, we get $\varphi'(\eta) = -2/\eta^2 + dTL_\infty^2/2$ and $\varphi''(\eta) = 4/\eta^3 > 0$. Since $\varphi''(\eta) > 0$, we know that $\varphi(\eta)$ is convex. Thus, $\eta_2^\star$ must satisfy the condition $\varphi'(\eta_2^\star) = 0$. Solve for $\eta_2^\star$, we get $\eta_2^\star = \sqrt{4/TdL_\infty^2}$. Substitute $\eta_2^\star$ into Equation (4.41), we get

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le \sqrt{4dTL_\infty^2}, \quad \forall u \in \Delta_d. \tag{4.42}$$

Let us compare the regret bounds of Entropic Mirror Descent and Online Subgradient Descent. Both algorithm achieves a $O(\sqrt{T})$ regret bound in this setting. However, by comparing Equation (4.40) and Equation (4.42), we can see that Entropic Mirror Descent has a milder dependency on the dimension than Online Subgradient Descent. In fact, Entropic Mirror Descent essentially has no dependence on the dimension since $\sqrt{\log(d)}$ grows very slow.

The above example illustrates the importance of choosing the "right" algorithm. By running Entropic Mirror Descent instead of Online Subgradient Descent, the dependency on the dimension goes from $\sqrt{d}$ to $\sqrt{\log(d)}$.

## 4.8 Exponentiated Gradient

In this section, we introduce the Exponentiated Gradient algorithm. As we will shortly see, Exponentiated Gradient is identical to Entropic Mirror Descent. The Exponentiated Gradient algorithm is as follows.

**Algorithm 7** Exponentiated Gradient

**Inputs:** The probability simplex $\Delta_d$, the (normalized) negative Shannon entropy $h : \Delta_d \to \mathbb{R}$, and a fixed constant step size $\eta > 0$.

1: Set $x_1 \leftarrow (1/d, \dots, 1/d)^\top$.
2: **for** $t = 1, \dots, T$ **do**
3: The learner outputs $x_t \in \Delta_d$.
4: The adversary chooses a subdifferentiable loss function $\ell_t : \Delta_d \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
5: The learner calls the first-order oracle for $g_t \in \partial \ell_t(x_t)$.
6: The learner updates by setting

$$x_{t+1,j} \leftarrow \frac{x_{t,j} \exp(-\eta g_{t,j})}{\sum_{i=1}^{d} x_{t,i} \exp(-\eta g_{t,i})}, \quad \forall 1 \le j \le d.$$

7: **end for**

To see why Exponentiated Gradient is identical to Entropic Mirror Descent, we will analyze Exponentiated Gradient with a dual perspective. Let us find the Fenchel conjugate of the (normalized) negative Shannon entropy on $\Delta_d$. The Fenchel conjugate of the (normalized) negative Shannon entropy function $h(x) = \sum_{i=1}^d x_i \log(x_i)$ is defined as

$$h^*(\theta) \coloneqq \sup_{x \in \Delta_d} \{\langle \theta, x \rangle - h(x)\} = \sup_{x \in \Delta_d} \left\{ \sum_{i=1}^d \theta_i x_i - \sum_{i=1}^d x_i \log(x_i) \right\}. \tag{4.43}$$

Note that we can rephrase the constrained maximization problem in Equation (4.43) as an unconstrained minimization problem in $\mathbb{R}^{d-1}$ by

$$\begin{aligned} h^*(\theta) &= \inf_{x \in \Delta_d} \left\{ \sum_{i=1}^d x_i \log(x_i) - \sum_{i=1}^d \theta_i x_i \right\} \\ &= \inf_{x \in \mathbb{R}^{d-1}} \left\{ \sum_{i=1}^{d-1} x_i \log(x_i) + \left(1 - \sum_{i=1}^{d-1} x_i\right) \log\left(1 - \sum_{i=1}^{d-1} x_i\right) - \sum_{i=1}^{d-1} \theta_i x_i - \theta_d \left(1 - \sum_{i=1}^{d-1} x_i\right) \right\}. \end{aligned} \tag{4.44}$$

Let $\varphi : \mathbb{R}^{d-1} \to \mathbb{R}$ be the function inside the infimum in Equation (4.44). We have

$$\frac{\partial}{\partial x_i} \varphi(x) = \log(x_i) - \log\left(1 - \sum_{j=1}^{d-1} x_j\right) - \theta_i + \theta_d, \quad \forall 1 \le i \le d-1. \tag{4.45}$$

Denote by $x^\star$ the minimizer of $\varphi$ over $\mathbb{R}^{d-1}$. We know that $x^\star$ satisfies $\nabla \varphi(x^\star) = 0$. By Equation (4.45), we get

$$x_i^\star = \exp(\theta_i - \theta_d)\left(1 - \sum_{j=1}^{d-1} x_j^\star\right), \quad \forall 1 \le i \le d-1. \tag{4.46}$$

Summing Equation (4.46) over $i = 1, \ldots, d-1$ yields

$$\sum_{i=1}^{d-1} x_i^\star = \left(\sum_{i=1}^{d-1} \exp(\theta_i - \theta_d)\right)\left(1 - \sum_{i=1}^{d-1} x_i^\star\right). \tag{4.47}$$

By adding $1 - \sum_{i=1}^{d-1} x_i^\star$ on both sides of Equation (4.47), we get

$$1 - \sum_{i=1}^{d-1} x_i^\star = \frac{1}{1 + \sum_{i=1}^{d-1} \exp(\theta_i - \theta_d)}. \tag{4.48}$$

Substitute Equation (4.48) into Equation (4.46), we obtain

$$x_i^\star = \frac{\exp(\theta_i - \theta_d)}{1 + \sum_{j=1}^{d-1} \exp(\theta_j - \theta_d)} = \frac{\exp(\theta_i)}{\sum_{j=1}^{d} \exp(\theta_j)}, \quad \forall 1 \le i \le d. \tag{4.49}$$

## 4.9 Learning with Expert Advice

**Learning with Expert Advice (LEA)** is a game between the learner and the adversary. The game is as follows. Given a fixed set of $d$ experts, in each round $t = 1, \ldots, T$,

(i) the learner chooses an expert $I_t$ among the $d$ experts;

(ii) the adversary reveals the losses $g_t = (g_{t,1}, \ldots, g_{t,d})$ of *each* expert; and

(iii) the learner suffers loss $g_{t,I_t}$.

The learner's goal is to minimize his losses compared to the cumulative loss of the best expert in hindsight. The regret of the learner is given by

$$\text{Regret}_T(e_i) = \sum_{t=1}^T \langle g_t, x_t \rangle - \sum_{t=1}^T \langle g_t, e_i \rangle, \quad \forall 1 \le i \le d,$$

where each $x_t \in \mathbb{R}^d$ is defined as

$$x_{t,i} \coloneqq \begin{cases} 1, & \text{if } i = I_t; \\ 0, & \text{otherwise,} \end{cases}$$

and $e_i$ is the $i$-th vector of the canonical ordered basis of $\mathbb{R}^d$.

Before we try to solve the Learning with Expert Advice problem, we should consider whether it is solvable or not, in the sense of achieving a sublinear regret. It turns out that this problem is *not* solvable in the adversarial setting. Consider the following setting. Assume that $d = 2$ and that the losses are the zero-one losses. In each round, after observing the expert we have chosen, the adversary assign a loss of 1 to our chosen expert and assign a loss of 0 to the other expert. Then, after a total of $T$ rounds, our cumulative loss would be exactly $T$. Note that the cumulative loss of the best expert is at most $T/2$. Thus, we have

$$\text{Regret}_T = \sum_{t=1}^{T} \langle g_t, x_t \rangle - \min_{i \in \{1,2\}} \sum_{t=1}^{T} \langle g_t, e_i \rangle \geq T - \frac{T}{2} = \frac{T}{2}.$$

In conclusion, we cannot achieve sublinear regret in the adversarial setting.

The reason that sublinear regret is not achievable in the above example is that the adversary have too much power. One possible way to reduce the adversary's power is by considering randomized algorithms. Let our action set be $\Delta_d$, the $(d-1)$-dimensional probability simplex in $\mathbb{R}^d$, that is,

$$\Delta_d = \{x \in \mathbb{R}^d : x_i \geq 0, \|x\|_1 = 1\}.$$

In each round $t = 1, \ldots, T$, the learner constructs a random variable $i_t$ such that

$$\mathbb{P}(i_t = i) = x_{t,i}, \quad \forall 1 \leq i \leq d.$$

Then, the learner chooses the corresponding expert according to the outcome of $i_t$. Note that we can write the expected loss at round $t$ as

$$\mathbb{E} g_{t,i_t} = \sum_{i=1}^{d} g_{t,i} \cdot \mathbb{P}(i_t = i) = \sum_{i=1}^{d} g_{t,i} \cdot x_{t,i} = \langle g_t, x_t \rangle .$$

Under the randomized setting, the regret is given by

$$\sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, u \rangle , \quad \forall u \in \Delta_d,$$

and the expected regret is

$$\mathbb{E}\left[ \sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, u \rangle \right], \quad \forall u \in \Delta_d.$$

Let us further assume that $\|g_t\|_\infty < L_\infty$ for $t = 1, \ldots, T$. Then, from Equation (4.40), we know that Entropic Mirror Descent with fixed step size $\eta = \sqrt{2\log(d)/(TL_\infty^2)}$ achieves

$$\mathbb{E}\text{Regret}_T \leq \sqrt{2T \log d L_\infty^2}.$$

Therefore, Learning with Expert Advice is solvable in the randomized setting, and Entropic Mirror Descent achieves a sublinear regret.

Later, we will study the **Multi-Armed Bandit (MAB)** problem. Multi-Armed Bandit is similar to Learning with Expert Advice. In both problem settings, we are choosing among $d$ fixed choices in each round. However, in Multi-Armed Bandit, the adversary only reveals the loss of the chosen arm, while the loss of each expert is revealed in Learning with Expert Advice. Naturally, one would expect that the Multi-Armed Bandit problem is harder to solve due to the fact that there are less information available to the learner.

## 4.10 Legendre Functions

**Definition 4.10.1** (Essential smoothness)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper closed convex function. Then, $f$ is said to be **essentially smooth** if and only if*

*(a)* $\text{int}(\text{dom}(f))$ *is nonempty;*

*(b) $f$ is differentiable on* $\text{int}(\text{dom}(f))$*; and*

*(c) $\|\nabla f(x_n)\|_2 \to \infty$ for any sequence $\{x_n\}_{n=1}^{\infty}$ in* $\text{int}(\text{dom}(f))$ *such that $x_n \to x$ for some $x \in \text{bdry}(\text{dom}(f))$.*

**Definition 4.10.2** (Legendre function)**.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper closed convex function. We say that $f$ is a **Legendre (type) function** if $f$ is essentially smooth and strictly convex.*

**Theorem 4.10.3.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a Legendre function. The following statements hold.*

*(a) The gradient mapping $\nabla f : \text{int}(\text{dom}(f)) \to \text{int}(\text{dom}(f^*))$ is a bijection with inverse $(\nabla f)^{-1} = \nabla f^*$.*

*(b) For all $x, y \in \text{int}(\text{dom}(f))$, we have $D_f(x \,\|\, y) = D_{f^*}(\nabla f(y) \,\|\, \nabla f(x))$.*

*(c) The Fenchel conjugate $f^*$ is Legendre.*

## 4.11 Questions

**Question 4.1.** In Online Mirror Descent (Algorithm 5), why do we separate the feasible set $\mathcal{V}$ and the domain $\mathcal{X}$ of the distant-generating function?

**Question 4.2.** Is there any advantage of assuming that the distant-generating function in Online Mirror Descent is Legendre?

**Question 4.3.** The regret bound of Online Mirror Descent with time-varying learning rates depends on the quantity $\max_{1 \leq t \leq T} D_h(u \,\|\, x_t)$ (Theorem 4.3.3). Is there an efficient method to find or upper bound this quantity?

# 5 Follow-the-Regularized-Leader

The simplest strategy one can adopt in an online convex optimization game is the Follow-the-Leader algorithm. We have seen that Follow-the-Leader fails to achieve sublinear regret in Example 1.2.3. The problem is that Follow-the-Leader is not stable. We can stabilize it by considering "regularization." This leads us to the Follow-the-Regularized-Leader algorithm.

## 5.1 The Follow-the-Regularized-Leader Algorithm

The Follow-the-Regularized-Leader algorithm is as follows.

**Algorithm 8** Follow-the-Regularized-Leader

**Inputs:** A sequence of regularizers $\psi_1, \dots, \psi_T : \mathcal{X} \to \mathbb{R}$, and a nonempty closed set $\mathcal{V} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$.

1: **for** $t = 1, \dots, T$ **do**

2: The learner outputs

$$x_t \in \operatorname*{arg\,min}_{x \in \mathcal{V}} \left\{ \sum_{\tau=1}^{t-1} \ell_\tau(x) + \psi_t(x) \right\}.$$

3: The adversary chooses a loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.

4: **end for**

Note that we do not impose any assumption on the regularizers and the loss functions in Algorithm 8. Thus, the $x_t$'s might not be feasible or well-defined. As we will see, by assuming that the regularized loss is a closed strongly convex function and that the feasible is a closed convex set, we can guarantee that the predictions are well-defined (Lemma 5.2.1).

Let us prove an identity that holds for Follow-the-Regularized-Leader with arbitrary regularizers and loss functions.

**Lemma 5.1.1.** *Let $\ell_1, \dots, \ell_T : \mathcal{X} \to \mathbb{R}$ be an arbitrary sequence of loss functions, let $\psi_1, \dots, \psi_{T+1} : \mathcal{X} \to \mathbb{R}$ be an arbitrary sequence of regularizers, and let $\mathcal{V} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed set. Define $F_t(x) := \psi_t(x) + \sum_{\tau=1}^{t-1} \ell_\tau(x)$ for all $t$. Assume that $\operatorname{arg\,min}_{x \in \mathcal{V}} F_t(x)$ is nonempty and set $x_t \in \operatorname{arg\,min}_{x \in \mathcal{V}} F_t(x)$ for all $t$. Then, for any $u \in \mathbb{R}^d$, we have*

$$\begin{aligned} \sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) = \psi_{T+1}(u) - \min_{x \in \mathcal{V}} \psi_1(x) + \sum_{t=1}^{T} (F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t)) \\ + F_{T+1}(x_{T+1}) - F_{T+1}(u). \end{aligned} \tag{5.1}$$

*Proof.* Observe that the sum $\sum_{t=1}^{T} \ell_t(x_t)$ appears on both sides of Equation (5.1). Thus, it suffices to show that

$$-\sum_{t=1}^{T} \ell_t(u) = \psi_{T+1}(u) - \min_{x \in \mathcal{V}} \psi_1(x) + \sum_{t=1}^{T} (F_t(x_t) - F_{t+1}(x_{t+1})) + F_{T+1}(x_{T+1}) - F_{T+1}(u), \tag{5.2}$$

for all $u \in \mathbb{R}^d$. Let $u \in \mathbb{R}^d$. By assumption, we have $F_1(x_1) = \min_{x \in \mathcal{V}} F_1(x) = \min_{x \in \mathcal{V}} \psi_1(x)$. We can write

$$\begin{aligned} -\sum_{t=1}^{T} \ell_t(u) &= \psi_{T+1}(u) - F_{T+1}(u) \\ &= \psi_{T+1}(u) - \min_{x \in \mathcal{V}} \psi_1(x) + F_1(x_1) - F_{T+1}(u) \\ &= \psi_{T+1}(u) - \min_{x \in \mathcal{V}} \psi_1(x) + F_1(x_1) - F_{T+1}(x_{T+1}) + F_{T+1}(x_{T+1}) - F_{T+1}(u) \\ &= \psi_{T+1}(u) - \min_{x \in \mathcal{V}} \psi_1(x) + \sum_{t=1}^{T} (F_t(x_t) - F_{t+1}(x_{t+1})) + F_{T+1}(x_{T+1}) - F_{T+1}(u). \end{aligned}$$

This proves Equation (5.2) and the lemma follows. □

Again, we emphasize that Lemma 5.1.1 is very general. The loss functions and the regularizers are not assumed to be convex.

## 5.2 Follow-the-Regularized-Leader with Strong Convexity

**Lemma 5.2.1.** *Let $\mathcal{V} \subseteq \mathcal{X}$ be a nonempty closed convex set, and let $\ell_1, \dots, \ell_t : \mathcal{X} \to \mathbb{R}$ be a sequence of loss functions. Define $F_t(x) \coloneqq \psi_t(x) + \sum_{\tau=1}^{t-1} \ell_\tau(x)$. If $F_t$ is closed, subdifferentiable, and strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$, then $x_t \in \arg\min_{x \in \mathcal{V}} F_t(x)$ exists and is unique. Moreover, if $\partial \ell_t(x_t)$ is nonempty and $F_t + \ell_t$ is closed, subdifferentiable, and $\mu_t$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$, then we have*

$$F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t) \le \langle g_t, x_t - x_{t+1} \rangle - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \tag{5.3}$$

$$\le \frac{\|g_t\|_*^2}{2\mu_t} + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}), \tag{5.4}$$

*for all $x_{t+1} \in \mathcal{V}$ and $g_t \in \partial \ell_t(x_t)$.*

*Proof.* It follows immediately from Theorem 3.1.14 that $x_t = \arg\min_{x \in \mathcal{V}} F_t(x)$ is well-defined.

By the Fermat's rule (Theorem 2.2.14), we have $0 \in \partial(F_t + \iota_{\mathcal{V}})(x_t)$. Let $g_t \in \partial \ell_t(x_t)$. By Proposition 2.2.10, we have $g_t \in \partial(F_t + \iota_{\mathcal{V}})(x_t) + \partial \ell_t(x_t) \subseteq \partial(F_t + \ell_t + \iota_{\mathcal{V}})(x_t)$. Since $\mathcal{V}$ is convex, we know that $\iota_{\mathcal{V}}$ is convex (hence 0-strongly convex). Then, from part (a) of Proposition 3.1.3, we know that the sum $F_t + \ell_t + \iota_{\mathcal{V}}$ is $\mu_t$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$. Note that we have the following identity:

$$F_{t+1}(x) = F_t(x) + \ell_t(x) + \psi_{t+1}(x) - \psi_t(x), \quad \forall x \in \mathcal{V}. \tag{5.5}$$

Let $x_{t+1} \in \mathcal{V}$. We can write

$$\begin{aligned}
F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t) &= F_t(x_t) - (F_t(x_{t+1}) + \ell_t(x_{t+1}) + \psi_{t+1}(x_{t+1}) - \psi_t(x_{t+1})) + \ell_t(x_t) \\
&= (F_t + \ell_t)(x_t) - (F_t + \ell_t)(x_{t+1}) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\
&= (F_t + \ell_t + \iota_{\mathcal{V}})(x_t) - (F_t + \ell_t + \iota_{\mathcal{V}})(x_{t+1}) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\
&\le \langle g_t, x_t - x_{t+1} \rangle - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}),
\end{aligned}$$

where the first equality follows from the identity (5.5), and the last inequality follows from Theorem 3.1.4 by the strong convexity of $F_t + \ell_t + \iota_{\mathcal{V}}$. This proves Equation (5.3). We can write

$$\langle g_t, x_t - x_{t+1} \rangle - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 \le \|g_t\|_* \|x_t - x_{t+1}\| - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 \le \frac{\|g_t\|_*^2}{2\mu_t},$$

where the first inequality follows from Hölder's inequality, and the last inequality follows by optimizing a concave quadratic function. Equation (5.4) then follows. □

**Corollary 5.2.2.** *Let $\mathcal{V} \subseteq \mathcal{X}$ be a nonempty closed convex set, and let $\psi : \mathcal{X} \to \mathbb{R}$ be closed and $\mu$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$. Define the sequence of regularizers $\psi_1, \dots, \psi_{T+1}$ by*

$$\psi_t(x) \coloneqq \frac{1}{\eta_t}\left(\psi(x) - \min_{z \in \mathcal{V}} \psi(z)\right), \quad \forall x \in \mathcal{X},$$

*and $\psi_{T+1} = \psi_T$, where $\eta_1, \dots, \eta_T$ is a sequence of nonincreasing constants. Assume that $\ell_1, \dots, \ell_T : \mathcal{X} \to \mathbb{R}$ is a sequence of closed convex loss functions such that $\partial \ell_t(x_t)$ is nonempty for $t = 1, \dots, T$. Then,* Follow-the-Regularized-Leader *guarantees*

$$\sum_{t=1}^{T} (\ell_t(x_t) - \ell_t(u)) \le \frac{\psi(u) - \min_{z \in \mathcal{V}} \psi(z)}{\eta_{T-1}} + \frac{1}{2\mu}\sum_{t=1}^{T} \eta_t \|g_t\|_*^2, \quad \forall u \in \mathcal{V},$$

*where $g_t \in \partial \ell_t(x_t)$ for $t = 1, \dots, T$.*

*Proof.* Since $\psi$ is $\mu$-strongly convex, from part (b) of Proposition 3.1.3, we know that $\psi_t$ is $\mu/\eta_t$-strongly convex for $t = 1, \dots, T$. Since the loss functions $\ell_\tau$'s are convex, it follows from part (a) of Proposition 3.1.3 that $F_t = \psi_t + \sum_{\tau=1}^{t-1} \ell_\tau$ and $F_t + \ell_t = \psi_t + \sum_{\tau=1}^{t} \ell_\tau$ are both $\mu/\eta_t$-strongly convex. From Lemma 5.2.1, we can write

$$F_t(x_t) - F_t(x_{t+1}) + \ell_t(x_t) \le \frac{\eta_t}{2\mu}\|g_t\|_*^2 + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \le \frac{\eta_t}{2\mu}\|g_t\|_*^2, \tag{5.6}$$

for $t = 1, \dots, T$, where the last inequality follows from the fact that $\psi_t \le \psi_{t+1}$ for $t = 1, \dots, T$. Finally, by

Lemma 5.1.1, we can write

$$\begin{aligned}\sum_{t=1}^{T}(\ell_t(x_t)-\ell_t(u)) &= \psi_{T+1}(u)-\min_{z\in\mathcal{V}}\psi_1(z)+\sum_{t=1}^{T}(F_t(x_t)-F_{t+1}(x_{t+1})+\ell_t(x_t))+F_{T+1}(x_{T+1})-F_{T+1}(u)\\ &\le \psi_{T+1}(u)-\min_{z\in\mathcal{V}}\psi_1(z)+\frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2+F_{T+1}(x_{T+1})-F_{T+1}(u)\\ &\le \psi_{T+1}(u)-\min_{z\in\mathcal{V}}\psi_1(z)+\frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2\\ &= \frac{\psi(u)-\min_{z\in\mathcal{V}}\psi(z)}{\eta_{T-1}}+\frac{1}{2\mu}\sum_{t=1}^{T}\eta_t\|g_t\|_*^2,\end{aligned}$$

for all $u\in\mathcal{V}$, where the first inequality follows by summing Equation (5.6) over $t=1,\dots,T$, the second inequality follows from the fact that $x_{T+1}\in\arg\min_{z\in\mathcal{V}}F_{T+1}(z)$, and the last inequality follows from the definition of $\psi_{T+1}$. □

**Theorem 5.2.3.** *Let $\mathcal{V}\subseteq\mathbb{R}^d$ be a nonempty closed convex set. Suppose that $\ell_1,\dots,\ell_T:\mathcal{V}\to\mathbb{R}$ is a sequence of loss functions such that $\ell_t$ is closed, subdifferentiable, and $\mu_t$-strongly convex with respect to $\|\cdot\|$, for $t=1,\dots,T$, then $x_t\in\arg\min_{x\in\mathcal{V}}\sum_{\tau=1}^{t-1}\ell_\tau(x)$ exists and is unique. Moreover, we have*

$$\sum_{t=1}^{T}(\ell_t(x_t)-\ell_t(u))\le\frac{1}{2}\sum_{t=1}^{T}\frac{\|g_t\|_*^2}{\sum_{\tau=1}^{t}\mu_\tau},\quad\forall u\in\mathcal{V},g_t\in\partial\ell_t(x_t).$$

*Proof.* Define $F_t(x)\coloneqq\sum_{\tau=1}^{t-1}\ell_\tau(x)$ for $t=1,\dots,T$. Since each $\ell_\tau$ is $\mu_\tau$-strongly convex with respect to $\|\cdot\|$, from part (a) of Proposition 3.1.3, we know that $F_t$ is $\sum_{\tau=1}^{t-1}\mu_\tau$-strongly convex with respect to $\|\cdot\|$. From Theorem 3.1.14, we know that $x_t=\arg\min_{x\in\mathcal{V}}F_t(x)$ is well-defined.

From Lemma 5.1.1, we can write

$$\sum_{t=1}^{T}(\ell_t(x_t)-\ell_t(u))=\sum_{t=1}^{T}(F_t(x_t)-F_{t+1}(x_{t+1})+\ell_t(x_t))+F_{T+1}(x_{T+1})-F_{T+1}(u),\quad\forall u\in\mathcal{V}.\tag{5.7}$$

Let $g_t\in\partial\ell_t(x_t)$ for $t=1,\dots,T$. From Lemma 5.2.1, we have

$$F_t(x_t)-F_{t+1}(x_{t+1})+\ell_t(x_t)\le\frac{\|g_t\|_*^2}{2\sum_{\tau=1}^{t}\mu_\tau}.\tag{5.8}$$

□

In other words, Theorem 5.2.3 gives us the regret bound of Follow-the-Leader over strongly convex losses.

**Definition 5.2.4** (Proximal regularizer)**.** *A sequence of regularizers $\psi_1,\dots,\psi_T$ is **proximal** if we have*

$$x_t\in\arg\min_{x\in\mathcal{V}}\{\psi_{t+1}(x)-\psi_t(x)\},\quad\forall 1\le t\le T.$$

**Example 5.2.5** (Squared norm regularizers)**.** *Let $\mathcal{V}\subseteq\mathbb{R}^d$ be a closed convex set. The sequence of regularizers $\psi_1,\dots,\psi_T:\mathcal{V}\to\mathbb{R}$ given by*

$$\psi_t(x)=\frac{1}{2}\sum_{\tau=1}^{t-1}\|x_\tau-x\|_2^2,\quad\forall x\in\mathcal{V},$$

*satisfies the proximal condition.*

*Proof.* We have $\psi_{t+1}(x)-\psi_t(x)=\|x_t-x\|_2^2/2$ for all $x\in\mathcal{V}$. It is clear that $x_t\in\arg\min_{x\in\mathcal{V}}\psi_{t+1}(x)-\psi_t(x)$. Therefore, the regularizers are proximal. □

**Lemma 5.2.6.** *Let $\mathcal{V}\subseteq\mathcal{X}$ be a nonempty closed convex set, and let $\ell_1,\dots,\ell_t:\mathcal{X}\to\mathbb{R}$ be a sequence of loss functions. Define $F_t(x)\coloneqq\psi_t(x)+\sum_{\tau=1}^{t-1}\ell_\tau(x)$. If $F_t$ is closed, subdifferentiable, and $\mu_t$-strongly convex with respect to $\|\cdot\|$ over $\mathcal{V}$, then $x_t\in\arg\min_{x\in\mathcal{V}}F_t(x)$ exists and is unique. Moreover, if $x_t$ satisfies the proximal condition $x_t\in\arg\min_{x\in\mathcal{V}}\psi_{t+1}(x)-\psi_t(x)$ and $\partial\ell_t(x_t)$ is nonempty, then we have*

$$F_t(x_t)-F_{t+1}(x_{t+1})+\ell_t(x_t)\le\langle g_t,x_t-x_{t+1}\rangle-\frac{\mu_t}{2}\|x_t-x_{t+1}\|^2+\psi_t(x_t)-\psi_{t+1}(x_t)\tag{5.9}$$

$$\le\frac{\|g_t\|_*^2}{2\mu_t}+\psi_t(x_t)-\psi_{t+1}(x_t),\tag{5.10}$$

*for all $x_{t+1}\in\mathcal{V}$ and $g_t\in\partial\ell_t(x_t)$.*

*Proof.* It follows immediately from Theorem 3.1.14 that $x_t = \arg\min_{x \in \mathcal{V}} F_t(x)$ is well-defined.

By the Fermat's rule (Theorem 2.2.14), we have $0 \in \partial(F_t + \iota_{\mathcal{V}})(x_t)$. Let $g_t \in \partial \ell_t(x_t)$. By Proposition 2.2.10, we have $g_t \in \partial(F_t + \iota_{\mathcal{V}})(x_t) + \partial \ell_t(x_t) \subseteq \partial(F_t + \ell_t + \iota_{\mathcal{V}})(x_t)$. Let $x_{t+1} \in \mathcal{V}$. We can write

$$\begin{aligned}
F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t) &= F_t(x_t) - (F_t(x_{t+1}) + \ell_t(x_{t+1}) + \psi_{t+1}(x_{t+1}) - \psi_t(x_{t+1})) + \ell_t(x_t) \\
&= (F_t + \ell_t)(x_t) - (F_t + \ell_t)(x_{t+1}) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\
&= (F_t + \ell_t + \iota_{\mathcal{V}})(x_t) - (F_t + \ell_t + \iota_{\mathcal{V}})(x_{t+1}) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\
&\leq \langle g_t, x_t - x_{t+1} \rangle - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\
&\leq \langle g_t, x_t - x_{t+1} \rangle - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 + \psi_t(x_t) - \psi_{t+1}(x_t),
\end{aligned}$$

where the first inequality follows from Theorem 3.1.4 by the strong convexity of $F_t + \ell_t + \iota_{\mathcal{V}}$, and the last inequality follows from the proximal condition. This proves Equation (5.9). We can write

$$\langle g_t, x_t - x_{t+1} \rangle - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 \leq \|g_t\|_* \|x_t - x_{t+1}\| - \frac{\mu_t}{2}\|x_t - x_{t+1}\|^2 \leq \frac{\|g_t\|_*^2}{2\mu_t},$$

where the first inequality follows from Hölder's inequality, and the last inequality follows by optimizing a concave quadratic function. Equation (5.10) then follows. □

## 5.3 Follow-the-Regularized-Leader with Linearized Losses

In this section, we consider Follow-the-Regularized-Leader with linearized losses.

**Algorithm 9** Follow-the-Regularized-Leader with Linearized Losses

**Inputs:** A sequence of regularizers $\psi_1, \ldots, \psi_T : \mathcal{X} \to \mathbb{R}$, and a nonempty closed convex set $\mathcal{V} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$.

1: **for** $t = 1, \ldots, T$ **do**
2: The leaner outputs

$$x_t \in \arg\min_{x \in \mathcal{V}} \left\{ \left\langle \sum_{\tau=1}^{t-1} g_\tau, x \right\rangle + \psi_t(x) \right\}.$$

3: The adversary chooses a loss function $\ell_t : \mathcal{V} \to \mathbb{R}$ and reveals the loss $\ell_t(x_t)$.
4: Set $g_t \in \partial \ell_t(x_t)$.
5: **end for**

**Theorem 5.3.1.** *Let $\psi_t : \mathcal{X} \to \mathbb{R}$ be a proper closed $\mu$-strongly convex function, let $\mathcal{V} \subseteq \mathcal{X}$ be a nonempty closed convex set, and let $\ell_1, \ldots, \ell_T : \mathcal{X} \to \mathbb{R}$ be given by $\ell_t(x) \coloneqq \langle g_t, x \rangle$, where $g_t \in \mathbb{R}^d$ is fixed, for $t = 1, \ldots, T$. Denote by $\psi_{\mathcal{V},t}$ the restriction of $\psi_t$ on $\mathcal{V}$, that is, $\psi_{\mathcal{V},t}(x) \coloneqq \psi_t(x) + \iota_{\mathcal{V}}(x)$. Define*

$$x_t \coloneqq \arg\min_{x \in \mathcal{V}} \left\{ \left\langle \sum_{\tau=1}^{t-1} g_\tau, x \right\rangle + \psi_t(x) \right\}.$$

*Then, we have*

$$x_t = \nabla \psi_{\mathcal{V},t}^* \left( -\sum_{\tau=1}^{t-1} g_\tau \right).$$

*Proof.* First, note that $\psi_{\mathcal{V},t}$ is a proper closed $\mu$-strongly convex function since $\psi_t$ is a proper closed $\mu$-strongly convex function and $\mathcal{V}$ is a closed convex set. We can write

$$x_t = \arg\min_{x \in \mathcal{V}} \left\{ \left\langle \sum_{\tau=1}^{t-1} g_\tau, x \right\rangle + \psi_t(x) \right\} = \arg\min_{x \in \mathbb{R}^d} \left\{ \left\langle \sum_{\tau=1}^{t-1} g_\tau, x \right\rangle + \psi_{\mathcal{V},t}(x) \right\}.$$

By the Fermat's rule (Theorem 2.2.14), we have $0 \in \sum_{\tau=1}^{t-1} g_\tau + \partial \psi_{\mathcal{V},t}(x_t)$. Rearranging, we get $-\sum_{\tau=1}^{t-1} g_\tau \in \partial \psi_{\mathcal{V},t}(x_t)$. Since $\psi_{\mathcal{V},t}$ is a proper closed $\mu$-strongly convex function, it follows from Corollary 4.4.15 that $x_t \in \partial \psi_{\mathcal{V},t}^*(-\sum_{\tau=1}^{t-1} g_\tau)$ and from Theorem 4.4.16 that $\psi_{\mathcal{V},t}^*$ is differentiable. Since $\psi_{\mathcal{V},t}^*$ is differentiable and $x_t \in \partial \psi_{\mathcal{V},t}^*(-\sum_{\tau=1}^{t-1} g_\tau)$, from Theorem 2.2.9, we know that $x_t = \nabla \psi_{\mathcal{V},t}^*(-\sum_{\tau=1}^{t-1} g_\tau)$. □

Let $\psi_1, \ldots, \psi_T : \mathcal{X} \to \mathbb{R}$ be the regularizers, let $\ell_1, \ldots, \ell_T : \mathcal{X} \to \mathbb{R}$ be given by $\ell_t(x) = \langle g_t, x \rangle$ for $t = 1, \ldots, T$, and let $\mathcal{V} \subseteq \mathcal{X}$ be a closed convex set. Denote by $\tilde{x}_t \coloneqq -\sum_{\tau=1}^{t-1} g_\tau$. Under this setting, Theorem 5.3.1 tells us that the predictions of Follow-the-Regularized-Leader are obtained as follows.

(i) Perform a subgradient descent step in the dual space, i.e., $\tilde{x}_{t+1} \coloneqq \tilde{x}_t - g_t$.

(ii) Then, map $\tilde{x}_{t+1}$ back to the feasible set by the gradient mapping $\nabla\psi^*_{\mathcal{V},t} : \mathbb{R}^d \to \mathcal{V}$ and set $x_{t+1} \coloneqq \nabla\psi^*_{\mathcal{V},t}(\tilde{x}_{t+1})$.

Note that the regularizer $\psi$ is not assumed to be differentiable in Theorem 5.3.1, while the distant-generating function $h$ in Online Mirror Descent is assumed to be differentiable in Theorem 4.4.17.

**Lemma 5.3.2.** *Let $\psi_1, \dots, \psi_{T+1} : \mathcal{X} \to \mathbb{R}$ be an arbitrary sequence of regularizers, let $\ell_1, \dots, \ell_T : \mathcal{X} \to \mathbb{R}$ be a sequence of loss functions given by $\ell_t(x) = \langle g_t, x\rangle$ for $t = 1, \dots, T$, and let $\mathcal{V} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty closed set. Define $F_t(x) \coloneqq \psi_t(x) + \sum_{\tau=1}^{t-1} \ell_\tau(x)$ for all $t$. Assume that the set $\arg\min_{x\in\mathcal{V}} F_t(x)$ is nonempty and set $x_t \in \arg\min_{x\in\mathcal{V}} F_t(x)$ for all $t$. Then, for any $u \in \mathbb{R}^d$, we have*

$$\sum_{t=1}^{T} \langle g_t, x_t - u\rangle = \psi_{\mathcal{V},T+1}(u) + \psi^*_{\mathcal{V},1}(0) + \sum_{t=1}^{T}\left(\psi^*_{\mathcal{V},t+1}\left(-\sum_{\tau=1}^{t} g_\tau\right) - \psi^*_{\mathcal{V},t}\left(-\sum_{\tau=1}^{t-1} g_\tau\right) + \langle g_t, x_t\rangle\right)$$
$$- \psi^*_{\mathcal{V},T+1}\left(-\sum_{\tau=1}^{T} g_\tau\right) - \psi_{\mathcal{V},T+1}(u) - \left\langle \sum_{\tau=1}^{T} g_\tau, u\right\rangle,$$

*where $\psi^*_{\mathcal{V},t}$ is the Fenchel conjugate of $\psi_{\mathcal{V},t} \coloneqq \psi_t + \iota_{\mathcal{V}}$.*

*Proof.* Since the loss functions are linear and $x_t \in \arg\min_{x\in\mathcal{V}} F_t(x)$, we have

$$\begin{aligned} F_t(x_t) &= \min_{x\in\mathcal{V}}\left\{\sum_{\tau=1}^{t-1}\ell_\tau(x) + \psi_t(x)\right\} = \min_{x\in\mathcal{V}}\left\{\left\langle\sum_{\tau=1}^{t-1} g_\tau, x\right\rangle + \psi_t(x)\right\} \\ &= \min_{x\in\mathbb{R}^d}\left\{\left\langle\sum_{\tau=1}^{t-1} g_\tau, x\right\rangle + \psi_{\mathcal{V},t}(x)\right\} \\ &= -\max_{x\in\mathbb{R}^d}\left\{\left\langle-\sum_{\tau=1}^{t-1} g_\tau, x\right\rangle - \psi_{\mathcal{V},t}(x)\right\} \\ &= -\psi^*_{\mathcal{V},t}\left(-\sum_{\tau=1}^{t-1} g_\tau\right). \end{aligned} \tag{5.11}$$

By Equation (5.11), we can write

$$F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t) = -\psi^*_{\mathcal{V},t}\left(-\sum_{\tau=1}^{t-1} g_\tau\right) + \psi^*_{\mathcal{V},t+1}\left(-\sum_{\tau=1}^{t} g_\tau\right) + \langle g_t, x_t\rangle, \tag{5.12}$$

and

$$\min_{x\in\mathcal{V}} \psi_1(x) = \min_{x\in\mathcal{V}} F_1(x) = F_1(x_1) = -\psi^*_{\mathcal{V},1}(0). \tag{5.13}$$

Now, substitute Equation (5.12) and Equation (5.13) into Lemma 5.1.1, we obtain the desired equality. □

**Example 5.3.3** (Equivalence of Online Mirror Descent and Follow-the-Regularized-Leader)**.** *Let $\mathcal{V} = \mathbb{R}^d$ and let $\eta > 0$. Define $\psi(x) \coloneqq \|x\|_2^2/2$ and $\psi_t \coloneqq \psi/\eta$ for $t = 1, \dots, T$. Suppose that the loss functions $\ell_1, \dots, \ell_T$ are given by $\ell_t = \langle g_t, x\rangle$ for $t = 1, \dots, T$. Then,* Online Mirror Descent *with feasible set $\mathcal{V}$, regularizer $\psi$, fixed step sizes $\eta$, and initial point $x_1 = \arg\min_{x\in\mathcal{V}} \psi(x)$ is equivalent to* Follow-the-Regularized-Leader *with feasible set $\mathcal{V}$ and regularizers $\psi_1, \dots, \psi_T$.*

# References


[1] Peter Bartlett, Elad Hazan, and Alexander Rakhlin. "Adaptive Online Gradient Descent". In: *Advances in neural information processing systems* 20 (2007).

[2] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

[3] Lev M Bregman. "The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming". In: *USSR computational mathematics and mathematical physics* 7.3 (1967), pp. 200–217.

[4] Gong Chen and Marc Teboulle. "Convergence Analysis of a Proximal-like Minimization Algorithm using Bregman Functions". In: *SIAM Journal on Optimization* 3.3 (1993), pp. 538–543.

[5] James Hannan. "Approximation to Bayes Risk in Repeated Play". In: *Contributions to the Theory of Games* 3.2 (1957), pp. 97–139.

[6] Francesco Orabona. *A Modern Introduction to Online Learning*. 2025. arXiv: `1912.13213 [cs.LG]`.

[7] Ralph Tyrell Rockafellar. "Convex analysis:(pms-28)". In: (2015).

[8] Martin Zinkevich. "Online Convex Programming and Generalized Infinitesimal Gradient Ascent". In: *Proceedings of the 20th international conference on machine learning (icml-03)*. 2003, pp. 928–936.