Projection-free Online Learning over Strongly Convex Sets

Yuanyu Wan[1], **Lijun Zhang**[1,2,*]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]Pazhou Lab, Guangzhou 510330, China
{wanyy, zhanglj}@lamda.nju.edu.cn

Abstract

To efficiently solve online problems with complicated constraints, projection-free algorithms including online frank-wolfe (OFW) and its variants have received significant interest recently. However, in the general case, existing efficient projection-free algorithms only achieved the regret bound of $O(T^{3/4})$, which is worse than the regret of projection-based algorithms, where T is the number of decision rounds. In this paper, we study the special case of online learning over strongly convex sets, for which we first prove that OFW can enjoy a better regret bound of $O(T^{2/3})$ for general convex losses. The key idea is to refine the decaying step-size in the original OFW by a simple line search rule. Furthermore, for strongly convex losses, we propose a strongly convex variant of OFW by redefining the surrogate loss function in OFW. We show that it achieves a regret bound of $O(T^{2/3})$ over general convex sets and a better regret bound of $O(\sqrt{T})$ over strongly convex sets.

Introduction

Online convex optimization (OCO) is a powerful framework that has been used to model and solve problems from diverse domains such as online routing (Awerbuch and Kleinberg 2004, 2008) and online portfolio selection (Blum and Kalai 1999; Agarwal et al. 2006; Luo, Wei, and Zheng 2018). In OCO, an online player plays a repeated game of T rounds against an adversary. At each round t, the player chooses a decision $\mathbf{x}_t$ from a convex set $\mathcal{K}$. After that, a convex function $f_t(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ chosen by the adversary is revealed, and incurs a loss $f_t(\mathbf{x}_t)$ to the player. The goal of the player is to choose decisions so that the regret defined as

$$R(T) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

is minimized. Over the past decades, various algorithms such as online gradient descent (OGD) (Zinkevich 2003), online Newton step (Hazan, Agarwal, and Kale 2007) and follow-the-regularized-leader (Shalev-Shwartz 2007; Shalev-Shwartz and Singer 2007) have been proposed to yield optimal regret bounds under different scenarios.

To ensure the feasibility of each decision, a common way in these algorithms is to apply a projection operation to any infeasible decision. However, in many high-dimensional problems with complicated decision sets, the projection step becomes a computational bottleneck (Zhang et al. 2013; Chen et al. 2016; Yang, Lin, and Zhang 2017). For example, in semi-definite programs (Hazan 2008) and multiclass classification (Hazan and Luo 2016), it amounts to computing the singular value decomposition (SVD) of a matrix, when the decision set consists of all matrices with bounded trace norm. To tackle the computational bottleneck, Hazan and Kale (2012) proposed a projection-free algorithm called online Frank-Wolfe (OFW) that replaces the time-consuming projection step with a more efficient linear optimization step. Taking matrix completion as an example again, the linear optimization step can be carried out by computing the top singular vector pair of a matrix, which is significantly faster than computing the SVD. Although OFW can efficiently handle complicated decision sets, it only attains a regret bound of $O(T^{3/4})$ for the general OCO, which is worse than the optimal $O(\sqrt{T})$ regret bound achieved by projection-based algorithms.

More specifically, OFW is an online extension of an offline algorithm called Frank-Wolfe (FW) (Frank and Wolfe 1956; Jaggi 2013) that iteratively performs linear optimization steps to minimize a convex and smooth function. In each round, OFW updates the decision by utilizing a single step of FW to minimize a surrogate loss function, which implies that the approximation error caused by the single step of FW could be the main reason for the regret gap between projection-based algorithms and OFW. Recently, Garber and Hazan (2015) made a quadratic improvement in the convergence rate of FW for strongly convex and smooth offline optimization over strongly convex sets compared to the general case. They used a simple line search rule to choose the step-size of FW, which allows FW to converge faster even if the strong convexity of the decision set is unknown. It is therefore natural to ask whether the faster convergence of FW can be utilized to improve the regret of OFW. In this paper, we give an affirmative answer by improving OFW to achieve an regret bound of $O(T^{2/3})$ over strongly convex sets, which is better than the original $O(T^{3/4})$ regret bound. Inspired by Garber and Hazan (2015), the key idea is to re-

*Lijun Zhang is the corresponding author.

Algorithm	Extra Condition on Loss	Extra Condition on $\mathcal{K}$	Regret Bound
OFW			$O(T^{3/4})$
LLOO-OCO		polyhedral	$O(\sqrt{T})$
LLOO-OCO	strongly convex	polyhedral	$O(\log T)$
Fast OGD		smooth	$O(\sqrt{T})$
Fast OGD	strongly convex	smooth	$O(\log T)$
OSPF	smooth		$O(T^{2/3})$
OFW with Line Search (this work)		strongly convex	$O(T^{2/3})$
SC-OFW (this work)	strongly convex		$O(T^{2/3})$
SC-OFW (this work)	strongly convex	strongly convex	$O(\sqrt{T})$

Table 1: Comparisons of regret bounds for efficient projection-free online algorithms including OFW (Hazan and Kale 2012; Hazan 2016), LLOO-OCO (Garber and Hazan 2016), Fast OGD (Levy and Krause 2019), OSPF (Hazan and Minasyan 2020) and our algorithms.

fine the decaying step-size in the original OFW by a simple line search rule.

Furthermore, we study OCO with strongly convex losses, and propose a strongly convex variant of OFW (SC-OFW). Different from OFW, we introduce a new surrogate loss function from Garber and Hazan (2016) to utilize the strong convexity of losses, which has been used to achieve an $O(\log T)$ regret bound for strongly convex OCO over polyhedral sets. Theoretical analysis reveals that SC-OFW for strongly convex OCO attains a regret bound of $O(T^{2/3})$ over general convex sets[1] and a better regret bound of $O(\sqrt{T})$ over strongly convex sets.

Related Work

In this section, we briefly review the related work about efficient projection-free algorithms for OCO, the time complexity of which is a constant in each round.

OFW (Hazan and Kale 2012; Hazan 2016) is the first projection-free algorithms for OCO, which attains a regret bound of $O(T^{3/4})$ by only performing 1 linear optimization step at each round. Recently, some studies have proposed projection-free online algorithms which are as efficient as OFW and attain better regret bounds, for special cases of OCO. If the decision set is a polytope, Garber and Hazan (2016) proposed a variant of OFW named as LLOO-based online convex optimization (LLOO-OCO), which enjoy an $O(\sqrt{T})$ regret bound for convex losses and an $O(\log T)$ regret bound for strongly convex losses. For OCO over smooth sets, Levy and Krause (2019) proposed a projection-free variant of OGD via devising a fast approximate projection for such sets, and established $O(\sqrt{T})$ and $O(\log T)$ regret bounds for convex and strongly convex losses, respectively. Besides these improvements for OCO over special decision sets, Hazan and Minasyan (2020) proposed online smooth projection free algorithm (OSPF) for OCO with smooth losses, which is a randomized method and achieves an expected regret bound of $O(T^{2/3})$.

[1]When this paper is under review, we notice that a concurrent work (Garber and Kretzu 2020b) also proposed an algorithm similar to SC-OFW and established the $O(T^{2/3})$ regret bound.

Furthermore, OFW has been extended to two practical scenarios. The first scenario is the bandit setting (Flaxman, Kalai, and McMahan 2005; Bubeck et al. 2015), where only the loss value is available to the player. Chen, Zhang, and Karbasi (2019) proposed the first bandit variant of OFW, and established an expected regret bound of $O(T^{4/5})$. Later, two improved bandit variants of OFW were proposed to enjoy better expected regret bounds on the order of $O(T^{3/4})$ for convex losses (Garber and Kretzu 2020a) and $\widetilde{O}(T^{2/3})$ for strongly convex losses (Garber and Kretzu 2020b). The second scenario is the distributed setting (Duchi, Agarwal, and Wainwright 2011; Hosseini, Chapman, and Mesbahi 2013), where many players are distributed over a network and each player can only access to the local loss function. The first projection-free algorithm for distributed OCO was proposed by Zhang et al. (2017), which requires the communication complexity of $O(T)$. Recently, Wan, Tu, and Zhang (2020) further reduced the communication complexity from $O(T)$ to $O(\sqrt{T})$.

Despite these great progresses about projection-free online learning, for the general OCO, OFW is still the best known efficient projection-free algorithm and the regret bound of $O(T^{3/4})$ has remained unchanged. In this paper, we will study OCO over strongly convex sets, and improve the regret bound to $O(T^{2/3})$ and $O(\sqrt{T})$ for convex and strongly convex losses, respectively. The detailed comparisons between efficient projection-free algorithms are summarized in Table 1.

Main Results

In this section, we first introduce necessary preliminaries including common notations, definitions and assumptions. Then, we present an improved regret bound for OFW over strongly convex sets by utilizing the line search. Finally, we introduce our SC-OFW algorithm for strongly convex OCO as well as its theoretical guarantees.

Preliminaries

In this paper, the convex set $\mathcal{K}$ belongs to a finite vector space $\mathbb{E}$. For any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, the inner product is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$. We recall two standard definitions for smooth and strongly convex functions (Boyd and Vandenberghe 2004), respectively.

Definition 1 *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called β-smooth over $\mathcal{K}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Definition 2 *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called α-strongly convex over $\mathcal{K}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\alpha}{2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

When $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is an α-strongly convex function and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} f(\mathbf{x})$, Garber and Hazan (2015) have proved that for any $\mathbf{x} \in \mathcal{K}$

$$\frac{\alpha}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \tag{1}$$

and

$$\|\nabla f(\mathbf{x})\|_2 \geq \sqrt{\frac{\alpha}{2}} \sqrt{f(\mathbf{x}) - f(\mathbf{x}_*)}. \tag{2}$$

Then, we introduce the definition of the strongly convex set, which has been well studied in offline optimization (Levitin and Polyak 1966; Demyanov and Rubinov 1970; Dunn 1979; Garber and Hazan 2015; Rector-Brooks, Wang, and Mozafari 2019).

Definition 3 *A convex set $\mathcal{K} \in \mathbb{E}$ is called α-strongly convex with respect to a norm $\|\cdot\|$ if for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, $\gamma \in [0, 1]$ and $\mathbf{z} \in \mathbb{E}$ such that $\|\mathbf{z}\| = 1$, it holds that*

$$\gamma \mathbf{x} + (1 - \gamma)\mathbf{y} + \gamma(1 - \gamma)\frac{\alpha}{2} \|\mathbf{x} - \mathbf{y}\|^2 \mathbf{z} \in \mathcal{K}.$$

As shown by Garber and Hazan (2015), various balls induced by ℓ_p norms, Schatten norms and group norms are strongly convex, which are commonly used to constrain the decision. For example, the ℓ_p norm ball defined as

$$\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\|_p \leq r\}$$

is $\frac{(p-1)d^{\frac{1}{2} - \frac{1}{p}}}{r}$-strongly convex with respect to the ℓ_2 norm for any $p \in (1, 2]$ (Garber and Hazan, 2015, Corollary 1).

We also introduce two common assumptions in OCO (Shalev-Shwartz 2011; Hazan 2016).

Assumption 1 *The diameter of the convex decision set $\mathcal{K}$ is bounded by D, i.e.,*

$$\|\mathbf{x} - \mathbf{y}\|_2 \leq D$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$.

Assumption 2 *At each round t, the loss function $f_t(\mathbf{x})$ is G-Lipschitz over $\mathcal{K}$, i.e.,*

$$|f_t(\mathbf{x}) - f_t(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|_2$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$.

Algorithm 1 OFW with Line Search

Input: feasible set $\mathcal{K}$, η
Initialization: choose $\mathbf{x}_1 \in \mathcal{K}$
for $t = 1, \cdots, T$ **do**
  Define $F_t(\mathbf{x}) = \eta \sum_{\tau=1}^{t} \langle \nabla f_\tau(\mathbf{x}_\tau), \mathbf{x} \rangle + \|\mathbf{x} - \mathbf{x}_1\|_2^2$
  $\mathbf{v}_t \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \nabla F_t(\mathbf{x}_t), \mathbf{x} \rangle$
  $\sigma_t = \operatorname*{argmin}_{\sigma \in [0,1]} \langle \sigma(\mathbf{v}_t - \mathbf{x}_t), \nabla F_t(\mathbf{x}_t) \rangle + \sigma^2 \|\mathbf{v}_t - \mathbf{x}_t\|_2^2$
  $\mathbf{x}_{t+1} = \mathbf{x}_t + \sigma_t(\mathbf{v}_t - \mathbf{x}_t)$
end for

OFW with Line Search

For the general OCO, OFW (Hazan and Kale 2012; Hazan 2016) arbitrarily chooses $\mathbf{x}_1$ from $\mathcal{K}$, and then iteratively updates its decision as the following linear optimization step

$$\mathbf{v} = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \nabla F_t(\mathbf{x}_t), \mathbf{x} \rangle$$
$$\mathbf{x}_{t+1} = \mathbf{x}_t + \sigma_t(\mathbf{v} - \mathbf{x}_t)$$

where

$$F_t(\mathbf{x}) = \eta \sum_{\tau=1}^{t} \langle \nabla f_\tau(\mathbf{x}_\tau), \mathbf{x} \rangle + \|\mathbf{x} - \mathbf{x}_1\|_2^2 \tag{3}$$

is the surrogate loss function, σ_t and η are two parameters. According to Hazan (2016), OFW with $\eta = O(T^{-3/4})$ and $\sigma_t = O(t^{-1/2})$ attains the regret bound of $O(T^{3/4})$. However, this decaying step-size $\sigma_t = O(t^{-1/2})$ cannot utilize the property of strongly convex sets. Inspired by a recent progress on FW over strongly convex sets (Garber and Hazan 2015), we utilized a line search rule to choose the parameter σ_t as

$$\sigma_t = \operatorname*{argmin}_{\sigma \in [0,1]} \langle \sigma(\mathbf{v}_t - \mathbf{x}_t), \nabla F_t(\mathbf{x}_t) \rangle + \sigma^2 \|\mathbf{v}_t - \mathbf{x}_t\|_2^2.$$

The detailed procedures are summarized in Algorithm 1.

From previous discussions, the approximation error of minimizing $F_t(\mathbf{x})$ by a single step of FW has a significant impact on the regret of OFW. Therefore, we first present the following lemma with respect to the approximation error for Algorithm 1 over strongly convex sets.

Lemma 1 *Let $\mathcal{K}$ be an α_K-strongly convex set with respect to the ℓ_2 norm. Let $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} F_{t-1}(\mathbf{x})$ for any $t \in [T+1]$, where $F_t(\mathbf{x})$ is defined in (3). Then, for any $t \in [T+1]$, Algorithm 1 with $\eta = \frac{D}{2G(T+2)^{2/3}}$ has*

$$F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*) \leq \epsilon_t = \frac{C}{(t+2)^{2/3}}$$

where $C = \max\left(4D^2, \frac{4096}{3\alpha_K^2}\right)$.

We find that the approximation error incurred by a single step of FW is upper bounded by $O(t^{-2/3})$ for Algorithm 1 over strongly convex sets. For a clear comparison, we note that the approximation error for the original OFW over general convex sets has a worse bound of $O(1/\sqrt{t})$ (Hazan, 2016, Lemma 7.3).

Algorithm 2 Strongly Convex Variant of OFW (SC-OFW)

Input: feasible set $\mathcal{K}$, λ
Initialization: choose $\mathbf{x}_1 \in \mathcal{K}$
for $t = 1, \cdots, T$ **do**
  Define $F_t(\mathbf{x}) = \sum_{\tau=1}^{t}(\langle \nabla f_\tau(\mathbf{x}_\tau), \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_\tau\|_2^2)$
  $\mathbf{v}_t \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \nabla F_t(\mathbf{x}_t), \mathbf{x}\rangle$
  $\sigma_t = \operatorname*{argmin}_{\sigma\in[0,1]} \langle \sigma(\mathbf{v}_t - \mathbf{x}_t), \nabla F_t(\mathbf{x}_t)\rangle + \frac{\sigma^2\lambda t}{2}\|\mathbf{v}_t-\mathbf{x}_t\|_2^2$
  $\mathbf{x}_{t+1} = \mathbf{x}_t + \sigma_t(\mathbf{v}_t - \mathbf{x}_t)$
end for

By applying Lemma 1 and further analyzing the property of decisions $\mathbf{x}_1^*, \cdots, \mathbf{x}_T^*$ defined in Lemma 1, we prove that OFW with line search enjoys the following regret bound over strongly convex sets.

Theorem 1 *Let $\mathcal{K}$ be an α_K-strongly convex set with respect to the ℓ_2 norm and $C = \max\left(4D^2, \frac{4096}{3\alpha_K^2}\right)$. For any $\mathbf{x}^* \in \mathcal{K}$, Algorithm 1 with $\eta = \frac{D}{2G(T+2)^{2/3}}$ ensures*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \le \frac{11}{4} G\sqrt{C}(T+2)^{2/3}.$$

We find that the $O(T^{2/3})$ regret bound in Theorem 1 is better than the $O(T^{3/4})$ bound achieved by previous studies for the original OFW over general convex sets (Hazan and Kale 2012; Hazan 2016).

SC-OFW for Strongly Convex Losses

Moreover, we propose a strongly convex variant of OFW (SC-OFW), which is inspired by Garber and Hazan (2016). To be precise, Garber and Hazan (2016) have proposed a variant of OFW for strongly convex OCO over polyhedral sets, which enjoys the $O(\log T)$ regret bound. Compared with the original OFW, their algorithm has two main differences. First, a local linear optimization step for polyhedral sets is developed to replace the linear optimization step utilized in OFW. Second, to handle λ-strongly convex losses, $F_t(\mathbf{x})$ is redefined as

$$\begin{aligned} F_t(\mathbf{x}) = &\sum_{\tau=1}^{t}\left(\langle \nabla f_\tau(\mathbf{x}_\tau), \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_\tau\|_2^2\right) \\ &+ \frac{c\lambda}{2}\|\mathbf{x}-\mathbf{x}_1\|_2^2 \end{aligned} \tag{4}$$

where c is a parameter that depends on properties of polyhedral sets.

Since this paper considers OCO over strongly convex sets, our SC-OFW adopts the linear optimization step utilized in the original OFW, and simplifies $F_t(\mathbf{x})$ in (4) as

$$F_t(\mathbf{x}) = \sum_{\tau=1}^{t}\left(\langle \nabla f_\tau(\mathbf{x}_\tau), \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_\tau\|_2^2\right) \tag{5}$$

by simply setting $c = 0$. The detailed procedures are summarized in Algorithm 2, where the parameter σ_t is selected by the line search as

$$\sigma_t = \operatorname*{argmin}_{\sigma\in[0,1]} \langle \sigma(\mathbf{v}_t - \mathbf{x}_t), \nabla F_t(\mathbf{x}_t)\rangle + \frac{\sigma^2\lambda t}{2}\|\mathbf{v}_t - \mathbf{x}_t\|_2^2.$$

To the best of our knowledge, SC-OFW is the first projection-free algorithm for strongly convex OCO over any decision set.

In the following lemma, we upper bound the approximation error of minimizing the surrogate loss function for SC-OFW over strongly convex sets.

Lemma 2 *Let $\mathcal{K}$ be an α_K-strongly convex set with respect to the ℓ_2 norm. Let $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_{t-1}(\mathbf{x})$ for any $t = 2, \cdots, T+1$, where $F_t(\mathbf{x})$ is defined in (5). Then, for any $t = 2, \cdots, T+1$, Algorithm 2 has*

$$F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*) \le C$$

where $C = \max\left(\frac{4(G+\lambda D)^2}{\lambda}, \frac{288\lambda}{\alpha_K^2}\right)$.

Furthermore, based on Lemma 2, we provide the following theorem with respect to the regret bound of SC-OFW over strongly convex sets.

Theorem 2 *Let $\mathcal{K}$ be an α_K-strongly convex set with respect to the ℓ_2 norm. If $f_t(\mathbf{x})$ is λ-strongly convex for any $t \in [T]$, for any $\mathbf{x}^* \in \mathcal{K}$, Algorithm 2 ensures*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \le C\sqrt{2T} + \frac{C\ln T}{2} + GD$$

where $C = \max\left(\frac{4(G+\lambda D)^2}{\lambda}, \frac{288\lambda}{\alpha_K^2}\right)$.

From Theorem 2, we achieve a regret bound of $O(\sqrt{T})$ for strongly convex losses, which is better than the above $O(T^{2/3})$ bound for general convex losses.

Moreover, we show the regret bound of Algorithm 2 over any general convex set $\mathcal{K}$ in the following theorem.

Theorem 3 *If $f_t(\mathbf{x})$ is λ-strongly convex for any $t \in [T]$, for any $\mathbf{x}^* \in \mathcal{K}$, Algorithm 2 ensures*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \le \frac{3\sqrt{2}CT^{2/3}}{8} + \frac{C\ln T}{8} + GD$$

where $C = \frac{16(G+\lambda D)^2}{\lambda}$.

By comparing Theorem 3 with Theorem 2, we can find that our Algorithm 2 enjoys a better regret bound when the decision set is strongly convex.

Theoretical Analysis

In this section, we prove Theorems 1 and 2. The omitted proofs can be found in the full version (Wan and Zhang 2020).

Proof of Theorem 1

In the beginning, we define $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_{t-1}(\mathbf{x})$ for any $t = 1, \cdots, T+1$, where $F_t(\mathbf{x})$ is defined in (3).

Since each function $f_t(\mathbf{x})$ is convex, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \\
\leq &\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle \\
= &\underbrace{\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_t^* \rangle}_{:=A} + \underbrace{\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}^* \rangle}_{:=B}.
\end{aligned} \tag{6}$$

So, we will upper bound the regret by analyzing A and B.

By applying Lemma 1, we have

$$\begin{aligned}
A = &\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_t^* \rangle \\
\leq &\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2 \|\mathbf{x}_t - \mathbf{x}_t^*\|_2 \\
\leq &\sum_{t=1}^{T} G\sqrt{F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*)} \\
\leq &\sum_{t=1}^{T} \frac{G\sqrt{C}}{(t+2)^{1/3}} \leq \frac{3G\sqrt{C}(T+2)^{2/3}}{2}
\end{aligned} \tag{7}$$

where the second inequality is due to (1) and the fact that $F_{t-1}(\mathbf{x})$ is 2-strongly convex, and the last inequality is due to $\sum_{t=1}^{T} (t+2)^{-1/3} \leq 3(T+2)^{2/3}/2$.

To bound B, we introduce the following lemma.

Lemma 3 *(Lemma 6.6 of Garber and Hazan (2016)) Let $\{f_t(\mathbf{x})\}_{t=1}^T$ be a sequence of loss functions and let $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{\tau=1}^{t} f_\tau(\mathbf{x})$ for any $t \in [T]$. Then, it holds that*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t^*) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq 0.$$

To apply Lemma 3, we define $\tilde{f}_1(\mathbf{x}) = \langle \eta\nabla f_1(\mathbf{x}_1), \mathbf{x}\rangle + \|\mathbf{x} - \mathbf{x}_1\|_2^2$ for $t = 1$ and $\tilde{f}_t(\mathbf{x}) = \langle \eta\nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle$ for any $t \geq 2$. We recall that $F_t(\mathbf{x}) = \sum_{\tau=1}^{t} \tilde{f}_\tau(\mathbf{x})$ and $\mathbf{x}_{t+1}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_t(\mathbf{x})$ for any $t = 1, \cdots, T$. Then, by applying Lemma 3 to $\{\tilde{f}_t(\mathbf{x})\}_{t=1}^T$, we have

$$\sum_{t=1}^{T} \tilde{f}_t(\mathbf{x}_{t+1}^*) - \sum_{t=1}^{T} \tilde{f}_t(\mathbf{x}^*) \leq 0$$

which implies that

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1}^* - \mathbf{x}^* \rangle \\
\leq &(\|\mathbf{x}^* - \mathbf{x}_1\|_2^2 - \|\mathbf{x}_2^* - \mathbf{x}_1\|_2^2)/\eta \\
\leq &D^2/\eta.
\end{aligned} \tag{8}$$

where the last inequality is due to $\|\mathbf{x}_2^* - \mathbf{x}_1\|_2^2 \geq 0$ and Assumption 1.

Moreover, by combining the fact that $F_t(\mathbf{x})$ is 2-strongly convex with (1), we have

$$\begin{aligned}
&\|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\|_2^2 \\
\leq &F_t(\mathbf{x}_t^*) - F_t(\mathbf{x}_{t+1}^*) \\
= &F_{t-1}(\mathbf{x}_t^*) - F_{t-1}(\mathbf{x}_{t+1}^*) + \eta\nabla f_t(\mathbf{x}_t)^\top(\mathbf{x}_t^* - \mathbf{x}_{t+1}^*) \\
\leq &\eta\|\nabla f_t(\mathbf{x}_t)\|_2 \|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\|_2
\end{aligned}$$

which implies that

$$\|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\|_2 \leq \eta\|\nabla f_t(\mathbf{x}_t)\|_2 \leq \eta G. \tag{9}$$

By combining (8), (9) and $\eta = \frac{D}{2G(T+2)^{2/3}}$, we have

$$\begin{aligned}
B = &\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}^* \rangle \\
= &\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1}^* - \mathbf{x}^* \rangle \\
&+ \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t^* - \mathbf{x}_{t+1}^* \rangle \\
\leq &\frac{D^2}{\eta} + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2 \|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\|_2 \\
\leq &\frac{D^2}{\eta} + \eta T G^2 \\
\leq &2DG(T+2)^{2/3} + \frac{DG(T+2)^{1/3}}{2} \\
\leq &G\sqrt{C}(T+2)^{2/3} + \frac{G\sqrt{C}(T+2)^{2/3}}{4}
\end{aligned} \tag{10}$$

where the last inequality is due to $D \leq \sqrt{C}/2$ and $(T+2)^{1/3} \leq (T+2)^{2/3}$ for any $T \geq 1$.

By combining (7) and (10), we complete the proof.

Proof of Theorem 2

Let $\tilde{f}_t(\mathbf{x}) = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2$ for any $t \in [T]$ and $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_{t-1}(\mathbf{x})$ for any $t = 2, \cdots, T+1$.

Since each function $f_t(\mathbf{x})$ is λ-strongly convex, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \\
\leq &\sum_{t=1}^{T} \left(\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{\lambda}{2}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \right) \\
= &\sum_{t=1}^{T} (\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}^*)) \\
= &\underbrace{\sum_{t=1}^{T} (\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_{t+1}^*))}_{:=A} + \underbrace{\sum_{t=1}^{T} (\tilde{f}_t(\mathbf{x}_{t+1}^*) - \tilde{f}_t(\mathbf{x}^*))}_{:B}.
\end{aligned}$$

So, we will derive a regret bound by analyzing A and B.

To bound A, we introduce the following lemma.

Lemma 4 *(Lemma 6.7 of Garber and Hazan (2016)) For any $t \in [T]$, the function $\tilde{f}_t(\mathbf{x}) = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \frac{\lambda}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$ is $(G+\lambda D)$-Lipschitz over $\mathcal{K}$.*

By applying Lemma 4, for any $t = 3, \cdots, T+1$, we have

$$\begin{aligned}
&F_{t-1}(\mathbf{x}_{t-1}^*) - F_{t-1}(\mathbf{x}_t^*) \\
=&F_{t-2}(\mathbf{x}_{t-1}^*) - F_{t-2}(\mathbf{x}_t^*) + \tilde{f}_{t-1}(\mathbf{x}_{t-1}^*) - \tilde{f}_{t-1}(\mathbf{x}_t^*) \\
\leq&(G+\lambda D)\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_2.
\end{aligned}$$

Moreover, since each $F_t(\mathbf{x})$ is $t\lambda$-strongly convex, for any $t = 3, \cdots, T+1$, we have

$$\begin{aligned}
\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_2^2 \leq&\frac{2(F_{t-1}(\mathbf{x}_{t-1}^*) - F_{t-1}(\mathbf{x}_t^*))}{(t-1)\lambda} \\
\leq&\frac{2(G+\lambda D)\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_2}{(t-1)\lambda}.
\end{aligned}$$

Therefore, for any $t = 3, \cdots, T+1$, we have

$$\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_2 \leq \frac{2(G+\lambda D)}{(t-1)\lambda}. \tag{11}$$

By applying Lemmas 2 and 4, we have

$$\begin{aligned}
&\sum_{t=2}^{T}(\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_{t+1}^*)) \\
\leq&\sum_{t=2}^{T}(G+\lambda D)\|\mathbf{x}_t - \mathbf{x}_{t+1}^*\|_2 \\
\leq&(G+\lambda D)\sum_{t=2}^{T}\|\mathbf{x}_t - \mathbf{x}_t^*\|_2 \\
&+ (G+\lambda D)\sum_{t=2}^{T}\|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\|_2 \\
\leq&(G+\lambda D)\sum_{t=2}^{T}\sqrt{\frac{2(F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*))}{(t-1)\lambda}} \\
&+ (G+\lambda D)\sum_{t=2}^{T}\frac{2(G+\lambda D)}{t\lambda} \\
\leq&(G+\lambda D)\sum_{t=2}^{T}\sqrt{\frac{2C}{(t-1)\lambda}} + (G+\lambda D)\sum_{t=2}^{T}\frac{2(G+\lambda D)}{t\lambda} \\
\leq&2(G+\lambda D)\sqrt{\frac{2TC}{\lambda}} + 2(G+\lambda D)^2\frac{\ln T}{\lambda} \\
\leq&C\sqrt{2T} + \frac{C\ln T}{2}
\end{aligned}$$

where the third inequality is due to $\sqrt{(t-1)\lambda}\|\mathbf{x}_t - \mathbf{x}_t^*\|_2 \leq \sqrt{2(F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*))}$ for $t \geq 2$ and (11), and the last inequality is due to $2(G+\lambda D) \leq \sqrt{\lambda C}$.

Due to $\|\nabla f_1(\mathbf{x}_1)\|_2 \leq G$ and $\|\mathbf{x}_1 - \mathbf{x}_2^*\|_2 \leq D$, we have

$$\begin{aligned}
A =&\tilde{f}_1(\mathbf{x}_1) - \tilde{f}_1(\mathbf{x}_2^*) + \sum_{t=2}^{T}(\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_{t+1}^*)) \\
=&\langle \nabla f_1(\mathbf{x}_1), \mathbf{x}_1 - \mathbf{x}_2^* \rangle - \frac{\lambda}{2}\|\mathbf{x}_2^* - \mathbf{x}_1\|_2^2 \\
&+ \sum_{t=2}^{T}(\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_{t+1}^*)) \\
\leq&\|\nabla f_1(\mathbf{x}_1)\|_2\|\mathbf{x}_1 - \mathbf{x}_2^*\|_2 + C\sqrt{2T} + \frac{C\ln T}{2} \\
\leq&GD + C\sqrt{2T} + \frac{C\ln T}{2}.
\end{aligned} \tag{12}$$

Moreover, we note that $F_t(\mathbf{x}) = \sum_{\tau=1}^{t}\tilde{f}_\tau(\mathbf{x})$ and $\mathbf{x}_{t+1}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_t(\mathbf{x})$ for any $t \in [T]$. By applying Lemma 3 to $\{\tilde{f}_t(\mathbf{x})\}_{t=1}^T$, we have

$$B = \sum_{t=1}^{T}\tilde{f}_t(\mathbf{x}_{t+1}^*) - \sum_{t=1}^{T}\tilde{f}_t(\mathbf{x}^*) \leq 0. \tag{13}$$

By combining (12) and (13), we complete the proof.

Proof of Lemma 1

For brevity, we define $h_t = F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*)$ for $t \in [T]$ and $h_t(\mathbf{x}_{t-1}) = F_{t-1}(\mathbf{x}_{t-1}) - F_{t-1}(\mathbf{x}_t^*)$ for $t = 2, \cdots, T$.

For $t = 1$, since $\mathbf{x}_1 = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}\|\mathbf{x} - \mathbf{x}_1\|_2^2$, we have

$$h_1 = F_0(\mathbf{x}_1) - F_0(\mathbf{x}_1^*) = 0 \leq \epsilon_1. \tag{14}$$

For any $T+1 \geq t \geq 2$, assuming $h_{t-1} \leq \epsilon_{t-1}$, we have

$$\begin{aligned}
h_t(\mathbf{x}_{t-1}) =&F_{t-1}(\mathbf{x}_{t-1}) - F_{t-1}(\mathbf{x}_t^*) \\
=&F_{t-2}(\mathbf{x}_{t-1}) - F_{t-2}(\mathbf{x}_t^*) \\
&+ \langle \eta\nabla f_{t-1}(\mathbf{x}_{t-1}), \mathbf{x}_{t-1} - \mathbf{x}_t^* \rangle \\
\leq&F_{t-2}(\mathbf{x}_{t-1}) - F_{t-2}(\mathbf{x}_{t-1}^*) \\
&+ \langle \eta\nabla f_{t-1}(\mathbf{x}_{t-1}), \mathbf{x}_{t-1} - \mathbf{x}_t^* \rangle \\
\leq&\epsilon_{t-1} + \eta\|\nabla f_{t-1}(\mathbf{x}_{t-1})\|_2\|\mathbf{x}_{t-1} - \mathbf{x}_t^*\|_2 \\
\leq&\epsilon_{t-1} + \eta\|\nabla f_{t-1}(\mathbf{x}_{t-1})\|_2\|\mathbf{x}_{t-1}^* - \mathbf{x}_t^*\|_2 \\
&+ \eta\|\nabla f_{t-1}(\mathbf{x}_{t-1})\|_2\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_2 \\
\leq&\epsilon_{t-1} + \eta^2 G^2 + \eta G\sqrt{\epsilon_{t-1}}
\end{aligned} \tag{15}$$

where the first inequality is due to $\mathbf{x}_{t-1}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_{t-2}(\mathbf{x})$ and the last inequality is due to $\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^*\|_2 \leq \sqrt{F_{t-2}(\mathbf{x}_{t-1}) - F_{t-2}(\mathbf{x}_{t-1}^*)}$ and (9).

Then, by substituting $\eta = D/(2G(T+2)^{2/3})$ into (15), we have

$$\begin{aligned}
&h_t(\mathbf{x}_{t-1}) \\
\leq&\epsilon_{t-1} + \frac{D\sqrt{C}}{2(T+2)^{2/3}(t+1)^{1/3}} + \frac{D^2}{4(T+2)^{4/3}} \\
\leq&\epsilon_{t-1} + \frac{\epsilon_{t-1}}{4(t+1)^{1/3}} + \frac{\epsilon_{t-1}}{16(t+1)^{2/3}} \\
\leq&\epsilon_{t-1}(1 + 1/(2(t+1)^{1/3}))
\end{aligned} \tag{16}$$

where the second inequality is due to $T \geq t-1$ and $D \leq \frac{\sqrt{C}}{2}$.

Then, to bound $h_t = F_{t-1}(\mathbf{x}_t) - F_{t-1}(\mathbf{x}_t^*)$ by ϵ_t, we further introduce the following lemma.

Lemma 5 *(Derived from Lemma 1 of Garber and Hazan (2015)) Let* $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ *be a convex and* β_f*-smooth function, where* $\mathcal{K}$ *is* α_K*-strongly convex with respect to the* ℓ_2 *norm. Moreover, let* $\mathbf{x}_{\text{in}} \in \mathcal{K}$ *and* $\mathbf{x}_{\text{out}} = \mathbf{x}_{\text{in}} + \sigma'(\mathbf{v} - \mathbf{x}_{\text{in}})$*, where* $\mathbf{v} \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \nabla f(\mathbf{x}_{\text{in}}), \mathbf{x}\rangle$ *and* $\sigma' = \operatorname{argmin}_{\sigma\in[0,1]} \langle \sigma(\mathbf{v}-\mathbf{x}_{\text{in}}), \nabla f(\mathbf{x}_{\text{in}})\rangle + \frac{\sigma^2\beta_f}{2}\|\mathbf{v}-\mathbf{x}_{\text{in}}\|_2^2$*. For any* $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$*, we have*

$$\begin{aligned}&f(\mathbf{x}_{\text{out}}) - f(\mathbf{x}^*)\\ \le&(f(\mathbf{x}_{\text{in}}) - f(\mathbf{x}^*)) \max\left(\frac{1}{2}, 1 - \frac{\alpha_K\|\nabla f(\mathbf{x}_{\text{in}})\|_2}{8\beta_f}\right).\end{aligned}$$

We note that $F_{t-1}(\mathbf{x})$ is 2-smooth for any $t \in [T+1]$. By applying Lemma 5 with $f(\mathbf{x}) = F_{t-1}(\mathbf{x})$ and $\mathbf{x}_{\text{in}} = \mathbf{x}_{t-1}$, for any $t \in [T+1]$, we have $\mathbf{x}_{\text{out}} = \mathbf{x}_t$ and

$$h_t \le h_t(\mathbf{x}_{t-1}) \max\left(\frac{1}{2}, 1 - \frac{\alpha_K\|\nabla F_{t-1}(\mathbf{x}_{t-1})\|_2}{16}\right). \tag{17}$$

Because of (16), (17) and $1 + \frac{1}{2(t+1)^{1/3}} \le \frac{3}{2}$, if $\frac{1}{2} \le \frac{\alpha_K\|\nabla F_{t-1}(\mathbf{x}_{t-1})\|_2}{16}$, it is easy to verify that

$$\begin{aligned}h_t \le& \frac{3}{4}\epsilon_{t-1} = \frac{3}{4}\frac{C}{(t+1)^{2/3}}\\ =& \frac{C}{(t+2)^{2/3}}\frac{3(t+2)^{2/3}}{4(t+1)^{2/3}}\\ \le& \frac{C}{(t+2)^{2/3}} = \epsilon_t\end{aligned} \tag{18}$$

where the last inequality is due to $\frac{3(t+2)^{2/3}}{4(t+1)^{2/3}} \le 1$ for any $t \ge 2$.

Then, if $\frac{1}{2} > \frac{\alpha_K\|\nabla F_{t-1}(\mathbf{x}_{t-1})\|_2}{16}$, there exist two cases. First, if $h_t(\mathbf{x}_{t-1}) \le \frac{3C}{4(t+1)^{2/3}}$, it is easy to verify that

$$h_t \le h_t(\mathbf{x}_{t-1}) \le \frac{3C}{4(t+1)^{2/3}} \le \epsilon_t \tag{19}$$

where the lase inequality has been proved in (18).

Second, if $h_t(\mathbf{x}_{t-1}) > \frac{3C}{4(t+1)^{2/3}}$, we have

$$\begin{aligned}&h_t\\ \le&h_t(\mathbf{x}_{t-1})\left(1 - \frac{\alpha_K\|\nabla F_{t-1}(\mathbf{x}_{t-1})\|_2}{16}\right)\\ \le&\epsilon_{t-1}\left(1 + \frac{1}{2(t+1)^{1/3}}\right)\left(1 - \frac{\alpha_K\|\nabla F_{t-1}(\mathbf{x}_{t-1})\|_2}{16}\right)\\ \le&\epsilon_{t-1}\left(1 + \frac{1}{2(t+1)^{1/3}}\right)\left(1 - \frac{\alpha_K\sqrt{h_t(\mathbf{x}_{t-1})}}{16}\right)\\ \le&\epsilon_t\frac{(t+2)^{2/3}}{(t+1)^{2/3}}\left(1 + \frac{1}{2(t+1)^{1/3}}\right)\left(1 - \frac{\alpha_K\sqrt{3C}}{32(t+1)^{1/3}}\right)\end{aligned}$$

where the second inequality is due to (16) and the third inequality is due to (2).

Since $(t+2)^{2/3} \le (t+1)^{2/3} + 1$ for any $t \ge 0$, it is easy to verify that

$$\frac{(t+2)^{2/3}}{(t+1)^{2/3}}\left(1 + \frac{1}{2(t+1)^{1/3}}\right) \le 1 + \frac{2}{(t+1)^{1/3}}$$

which further implies that

$$\begin{aligned}h_t \le&\epsilon_t\left(1 + \frac{2}{(t+1)^{1/3}}\right)\left(1 - \frac{\alpha_K\sqrt{3C}}{32(t+1)^{1/3}}\right)\\ \le&\epsilon_t\left(1 + \frac{2}{(t+1)^{1/3}}\right)\left(1 - \frac{2}{(t+1)^{1/3}}\right)\\ \le&\epsilon_t.\end{aligned} \tag{20}$$

where the second inequality is due to $\alpha_K\sqrt{3C} \ge 64$.

By combining (14), (18), (19) and (20), we complete the proof.

Conclusion and Future Work

In this paper, we first prove that the classical OFW algorithm with line search attains an $O(T^{2/3})$ regret bound for OCO over strongly convex sets, which is better than the $O(T^{3/4})$ regret bound for the general OCO. Furthermore, for strongly convex losses, we introduce a strongly convex variant of OFW, and prove that it achieves a regret bound of $O(T^{2/3})$ over general convex sets and a better regret bound of $O(\sqrt{T})$ over strongly convex sets.

An open question is whether the regret of OFW and its strongly convex variant over strongly convex sets can be further improved if the losses are smooth. We note that Hazan and Minasyan (2020) have proposed a projection-free algorithm for OCO over general convex sets, and established an improved regret bound of $O(T^{2/3})$ by taking advantage of the smoothness.

Acknowledgments

This work was partially supported by the NSFC (61921006), NSFC-NRF Joint Research Project (61861146001), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

Agarwal, A.; Hazan, E.; Kale, S.; and Schapire, R. E. 2006. Algorithms for portfolio management based on the Newton method. In *Proceedings of the 23rd International Conference on Machine Learning*, 9–16.

Awerbuch, B.; and Kleinberg, R. 2008. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences* 74(1): 97–114.

Awerbuch, B.; and Kleinberg, R. D. 2004. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 45–53.

Blum, A.; and Kalai, A. 1999. Universal portfolios with and without transaction costs. *Machine Learning* 35(3): 193–205.

Boyd, S.; and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Bubeck, S.; Dekel, O.; Koren, T.; and Peres, Y. 2015. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Conference on Learning Theory*, 266–278.

Chen, J.; Yang, T.; Lin, Q.; Zhang, L.; and Chang, Y. 2016. Optimal stochastic strongly convex optimization with a logarithmic number of projections. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 122–131.

Chen, L.; Zhang, M.; and Karbasi, A. 2019. Projection-free bandit convex optimization. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2047–2056.

Demyanov, V. F.; and Rubinov, A. M. 1970. *Approximate Methods in Optimization Problems*. Elsevier Publishing Company.

Duchi, J. C.; Agarwal, A.; and Wainwright, M. J. 2011. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control* 57(3): 592–606.

Dunn, J. C. 1979. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization* 17(2): 187–211.

Flaxman, A. D.; Kalai, A. T.; and McMahan, H. B. 2005. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 385–394.

Frank, M.; and Wolfe, P. 1956. An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 3(1–2): 95–110.

Garber, D.; and Hazan, E. 2015. Faster rates for the frank-wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on Machine Learning*, 541–549.

Garber, D.; and Hazan, E. 2016. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *SIAM Journal on Optimization* 26(3): 1493–1528.

Garber, D.; and Kretzu, B. 2020a. Improved regret bounds for projection-free bandit convex optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2196–2206.

Garber, D.; and Kretzu, B. 2020b. Revisiting projection-free online learning: the strongly convex case. *ArXiv e-prints* arXiv: 2010.07572.

Hazan, E. 2008. Sparse approximate solutions to semidefinite programs. In *Latin American Symposium on Theoretical Informatics*, 306–316.

Hazan, E. 2016. Introduction to online convex optimization. *Foundations and Trends in Optimization* 2(3–4): 157–325.

Hazan, E.; Agarwal, A.; and Kale, S. 2007. Logarithmic regret algorithms for online convex optimization. *Machine Learning* 69(2): 169–192.

Hazan, E.; and Kale, S. 2012. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning*, 1843–1850.

Hazan, E.; and Luo, H. 2016. Variance-reduced and projection-free stochastic optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, 1263–1271.

Hazan, E.; and Minasyan, E. 2020. Faster projection-free online learning. In *Proceedings of the 33rd Annual Conference on Learning Theory*, 1877–1893.

Hosseini, S.; Chapman, A.; and Mesbahi, M. 2013. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, 1484–1489.

Jaggi, M. 2013. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, 427–435.

Levitin, E. S.; and Polyak, B. T. 1966. Constrained minimization methods. *USSR Computational mathematics and mathematical physics* 6: 1–50.

Levy, K. Y.; and Krause, A. 2019. Projection free online learning over smooth sets. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 1458–1466.

Luo, H.; Wei, C.-Y.; and Zheng, K. 2018. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems 31*, 8235–8245.

Rector-Brooks, J.; Wang, J.-K.; and Mozafari, B. 2019. Revisiting projection-free optimization for strongly convex constraint sets. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 1576–1583.

Shalev-Shwartz, S. 2007. *Online Learning: Theory, Algorithms, and Applications*. Ph.D. thesis, The Hebrew University of Jerusalem.

Shalev-Shwartz, S. 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning* 4(2): 107–194.

Shalev-Shwartz, S.; and Singer, Y. 2007. A primal-dual perspective of online learning algorithm. *Machine Learning* 69(2–3): 115–142.

Wan, Y.; Tu, W.-W.; and Zhang, L. 2020. Projection-free distributed online convex optimization with $O(\sqrt{T})$ communication complexity. In *Proceedings of the 37th International Conference on Machine Learning*, 9818–9828.

Wan, Y.; and Zhang, L. 2020. Projection-free online learning over strongly convex sets. *ArXiv e-prints* arXiv:2010.08177.

Yang, T.; Lin, Q.; and Zhang, L. 2017. A richer theory of convex constrained optimization with reduced projections and improved rates. In *Proceedings of the 34th International Conference on Machine Learning*, 3901–3910.

Zhang, L.; Yang, T.; Jin, R.; and He, X. 2013. $O(\log T)$ projections for stochastic optimization of smooth and strongly convex functions. In *Proceedings of the 30th International Conference on Machine Learning*, 1121–1129.

Zhang, W.; Zhao, P.; Zhu, W.; Hoi, S. C. H.; and Zhang, T. 2017. Projection-free distributed online learning in networks. In *Proceedings of the 34th International Conference on Machine Learning*, 4054–4062.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, 928–936.