Structured Projection-free Online Convex Optimization with Multi-point Bandit Feedback

Yuhao Ding and Javad Lavaei

***Abstract*— We consider structured online convex optimization (OCO) with bandit feedback, where either the loss function is smooth or the constraint set is strongly convex. Projection-free methods are among the most popular and computationally efficient algorithms for solving this problem, mainly due to their ability to handle convex constraints appearing in machine learning for which computing projections is often impractical in high-dimensional settings. Despite the improved regret bound results for the full-information setting where the gradients of the functions are readily available, it remains unclear whether simple projection-free zero-order algorithms become more efficient for structured OCO problems in the case when multiple function values can be sampled at each time instance. In this paper, we develop some simple projection-free algorithms and prove that they indeed achieve the same improved regret bounds as the full-information case under various additional problem structures. This implies that leveraging the structural properties of the problem compensates for the lack of access to the gradients. Experiments on the online matrix completion reveal several attractive advantages of the proposed algorithms, including their simplicity, easy implementation, and effectiveness, as they outperform other competing algorithms.**

I. INTRODUCTION

Recent years have witnessed a large number of projection-based online algorithms, including online gradient descent (OGD) [34], online Newton step (ONS) [16], follow-theregularized- leader (FTRL) [30] and follow-the-perturbed-leader (FPL) [23], which yield optimal regret bounds under different scenarios. However, a significant barrier to the direct application of projection-based online algorithms in many machine learning applications lies in the computational bottleneck of performing a projection over complicated constraint sets in high-dimensional settings. Practitioners have long been aware of this computational bottleneck of the projection-based online algorithms in matrix completion [32] and collaborative filtering [24]. Indeed, the projection amounts to computing the singular value decomposition (SVD) of a matrix, whose cost increases dramatically as the dimension grows. This difficulty motivates the development of the so-called projection-free online algorithm [17], which is indeed an online variant of the classical Frank-Wolfe algorithm [9], [27], [21]. This approach replaces the projection step with a linear optimization over the constraint set, which is proven to be efficient. The algorithm is sometimes by orders of magnitude faster than projection-based online algorithms when the constraint sets arise from a combinatorial structure, e.g., paths/matches/spanning trees in graphs or matroids, or from a low-rank matrix structure, as demonstrated by encouraging empirical results in low-rank matrix approximation [22], [29], multitask learning [7], structural support vector machine [25], semidefinite programming [26] and image processing [14]. From a theoretical point of view, for an online optimization problem with the time horizon T, the online Frank-Wolfe (OFW) algorithm and its variants are efficient in the sense that they achieve the sublinear regret bound of $O(T^{3/4})$ when the loss function is convex [17] and improved regret bound of $O(T^{2/3})$ or even $O(\sqrt{T})$ under additional assumptions on the smoothness of the loss function or the strong convexity of the constraint set [20], [33], [12].

The above-mentioned online algorithms require a gradient oracle for all loss functions f_t such that the learner can update the next decision x_{t+1} by using the gradient information $\nabla f_t(x_t)$. Unfortunately, for many problems in optimization, machine learning and statistics [28], [13], [2], [1], [6], the only information that a learner can observe at iteration t is the loss f_t because the direct calculation of the gradient may be computationally infeasible, expensive, or impossible. We refer to online convex optimization setting with zero-order information or bandit feedback [3], [19]. This setting necessities an accurate estimation of the gradient information, making the development of efficient learning algorithms significantly more challenging. However, if the function f_t can be evaluated at two points or multiple points, some of the difficulties inherent in optimization using only a single function evaluation can be alleviated [1], [6]. Such multi-point settings are useful for online optimization problems in which the adversary is oblivious (the loss functions $f_1, f_2, \ldots$ are chosen beforehand and do not depend on the decisions of the learner) and multiple function values can be sampled for a given function f_t. Applications of such bandit problems include online auctions and advertisement selection for search engines.

There are several attempts to adapt the OGD and OFW algorithms to the zero-order online optimization setting and we refer the reader to Section 2 for more details. Recently, it has been shown in [11], [12] that the state-of-arts projection-free regret bounds for the full information setting (where the exact gradient is available) and single-point bandit feedback setting match up to a factor in $\log(T)$, provided that the functions are either convex or strongly convex. Both of those algorithms are, however, double-loop algorithms and thus relatively complicated to implement. More importantly, it is well-known that projection-free algorithms can benefit from additional structures existing in real-world applications for the full information setting, such as the smoothness of the

Y. Ding and J. Lavaei are with the department of Industrial Engineering and Operations Research, University of California, Berkeley, USA. {yuhao_ding, lavaei}@berkeley.edu. This work was supported by grants from ARO, ONR, AFOSR and NSF.

loss function [20] or the strongly convexity of the constraint set [33]. Based on this observation, it is natural to ask the important question: *Can the simplest projection-free online optimization algorithm with bandit feedback achieve the same regret bounds for the full-information setting under additional problem structures if multi-point bandit feedback is available?*

In this work, we provide an affirmative answer to the above question, and develop some improved regret bounds for two algorithms which are based on OFW [17] and smoothed FPL [20] with a multiple-point estimator [1], [31]. When the constraint set is strongly convex, the first algorithm and its variant achieve the expected regret bounds $\tilde{O}(T^{2/3})$ and $\tilde{O}(\sqrt{T})$ for convex and strongly convex loss functions, respectively. When the loss functions are smooth, the second algorithm attains the expected regret bound $O(T^{2/3})$. These bounds match the existing bounds for the full-information setting. We also develop some regret bounds that hold with high probability. Experiments on the online matrix completion demonstrate the benefit of using a two-point estimator and levering additional problem structures, and show achieving the same regret bound as if the gradient was readily available.

Notation. We use bold lower-case letters to denote vectors, as in x, and calligraphic upper case letters to denote sets, as in $\mathcal{X}$. The sets of real numbers and natural numbers are shown as $\mathbb{R}$ and $\mathbb{N}$, respectively. We denote by $\mathcal{S}^d$ and $\mathcal{B}^d$ the unit sphere and ball in $\mathbb{R}^d$, and $\mathbf{u} \sim \mathcal{S}^d$ and $\mathbf{u} \sim \mathcal{B}^d$ mean that $\mathbf{u}$ is a random vector sampled uniformly from $\mathcal{S}^d$ and $\mathcal{B}^d$, respectively. For a scalar $r > 0$, we denote $r\mathcal{B}^d$ as the ball with radius r. The notion $[T]$ refers to $\{1, 2, \ldots, T\}$ for some integer $T > 0$. The notions $\mathbb{E}_\xi[\cdot]$ and $\mathbb{E}[\cdot]$ refer to the expectation over the random variable ξ and over all of the randomness. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, let $\nabla f(\mathrm{x})$ denote the gradient of f at x. For a scalar a, let $|a|$ denote its absolute value. For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\|\mathbf{x}\|$ denote the ℓ_2-norm and $\langle \mathbf{x}, \mathbf{y} \rangle$ be the inner product. We also define the set $\mathcal{X}_\delta = \{(1-\delta)\mathbf{x} \mid \mathbf{x} \in \mathcal{X}\}$ given a constant $\delta \in [0, 1]$. The notion $\mathcal{A}_i$ refers to Algorithm i. Lastly, given the number of prediction rounds T, the notation $a = O(b(T))$ means $a \leq C \cdot b(T)$ for some constant $C > 0$ that is independent of T. Similarly, $a = \tilde{O}(b(T))$ indicates that the previous inequality may depend on the function $\log(T)$, where $C > 0$ is also independent of T.

II. PRELIMINARIES

According to [15], in online convex optimization with a bandit feedback setting, an online learner repeatedly chooses a decision x_t from a convex and compact set $\mathcal{K} \in \mathbb{R}^d$ on round $t \in [T]$, and observes the associated loss $f_t(\mathbf{x}_t)$, where T is known in advance and $f_t : \mathcal{K} \to \mathbb{R}$ is a convex and Lipschitz adversarial function. Besides the loss oracle f_t, the learner does not gain any additional knowledge of f_t. The goal is to find an algorithm $\mathcal{A}$ for generating the sequence $\{\mathbf{x}_t\}_{t \geq 1}$ such that the expected regret bound $\mathbb{E}[\mathcal{R}_T(\mathcal{A})]$ on T is minimized where $\mathcal{R}_T(\mathcal{A})$ is defined as

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

For the multi-point bandit setting [1], the player queries each loss function at k randomized points $\mathbf{y}_{t,1}, \ldots, \mathbf{y}_{t,k}$, rather than at a single point. In this model, the regret is defined as

$$\mathcal{R}_T(\mathcal{A}) = \frac{1}{k} \sum_{t=1}^{T} \sum_{i=1}^{k} f_t(\mathbf{y}_{t,i}) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)$$

Similarly to [5], we assume that the adversary is oblivious, meaning that the loss functions f_t's are chosen beforehand and do not depend on the decisions of the learner. Throughout this paper, we use some standard assumptions given below.

Assumption 1. *The constrained action set $\mathcal{K} \subset \mathbb{R}^d$ is convex and compact.*

Assumption 2. *For every iteration $t \in \mathbb{N}$, the loss function $f_t : \mathcal{K} \to \mathbb{R}$ is convex and differentiable.*

Because of the compactness of the constraint set, let R denote an upper bound on the norm of all points in the set, i.e., $\|\mathrm{x}\| \leq R$ for all $\mathrm{x} \in \mathcal{K}$, and let L denote an upper bound on the gradient norms over the set, i.e., $\|\nabla f_t(\mathrm{x})\| \leq L$ for all $\mathrm{x} \in \mathcal{K}$.

A. Linear optimization oracle

Consider the following well-known linear optimization oracle and linear value oracle over the constraint set $\mathcal{K}$:

$$\mathrm{LP}_{\mathcal{K}}(\mathrm{g}) = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{K}} \mathrm{g}^\top \mathrm{x}, \qquad \text{for all } \mathrm{g} \in \mathbb{R}^d,$$
$$\mathrm{VAL}_{\mathcal{K}}(\mathrm{g}) = \max_{\mathbf{x} \in \mathcal{K}} \mathrm{g}^\top \mathbf{x}, \qquad \text{for all } \mathrm{g} \in \mathbb{R}^d.$$

The focus of this work is on the bandit setting where performing a projection on the constraint set $\mathcal{K}$, as a quadratic optimization over $\mathcal{K}$, has a significantly higher computational cost than solving a linear optimization over $\mathcal{K}$. This scenario covers a wide range of application problems, for which projection-based bandit algorithms (e.g.,[8], [1], [18], [4]) are inferior to projection-free bandit algorithms (e.g., [5], [11]). Moreover, it follows from the compactness of $\mathcal{K}$ and the continuity of linear functions that $\mathrm{LP}_{\mathcal{K}}(\mathrm{g})$ exists for all $\mathrm{g} \in \mathbb{R}^d$ as stated below.

Lemma 1 (Lemma 2.4 in [20])**.** *The linear value oracle $\mathrm{VAL}_{\mathcal{K}} : \mathbb{R}^d \to \mathbb{R}$ is well defined and satisfies the properties $\mathrm{VAL}_{\mathcal{K}}(\mathrm{g}) = \mathrm{g}^\top \mathrm{LP}_{\mathcal{K}}(\mathrm{g})$ and $\nabla \mathrm{VAL}_{\mathcal{K}}(\mathrm{g}) = \mathrm{LP}_{\mathcal{K}}(\mathrm{g})$. Moreover, $\mathrm{VAL}_{\mathcal{K}}$ is R-Lipschitz, namely $|\mathrm{VAL}_{\mathcal{K}}(\mathbf{g}_1) - \mathrm{VAL}_{\mathcal{K}}(\mathbf{g}_2)| \leq R\|\mathbf{g}_1 - \mathbf{g}_2\|$ for all $\mathbf{g}_1, \mathbf{g}_2 \in \mathbb{R}^d$.*

B. Gradient estimates from bandit feedback

In this subsection, we introduce a gradient estimator based on a multi-point estimator. To this end, recall the one-point estimator [8], where the gradient of f_t at some point x is estimated from a single random point evaluation. Let $\mathbf{v} \sim \mathcal{B}^d$ be a uniform random vector in the unit ball, and define the δ-smoothed loss function $\hat{f}_{t,\delta}(\mathrm{x}) = \mathbb{E}_{\mathrm{v} \sim \mathcal{B}^d}(f_t(\mathrm{x} + \delta \mathrm{v}))$. Since f_t is convex and L-Lipschitz, its δ-smoothed version $\hat{f}_{t,\delta}$ is also convex and differentiable. Then, a one-point estimator

can be used that queries the objective function $f_t(\mathrm{x})$ only once:

$$\mathbf{g}_t^1 = \frac{d}{\delta} f_t\left(\mathbf{x} + \delta \mathbf{u}_t\right) \mathbf{u}_t, \quad \text{for some } \mathbf{u}_t \sim \mathcal{S}^d, \tag{1}$$

where g_t^1 denotes an estimate of $\nabla f_t(x)$.

Lemma 2 (Lemma 3.4 in [15])**.** *The one-point estimator defined in* (1) *is unbiased for the δ-smoothed function $\hat{f}_{t,\delta}$ in the sense that $\nabla \hat{f}_{t,\delta}(\mathrm{x}) = \mathbb{E}_{\mathbf{u}_t \sim \mathcal{S}^d}\left[\mathbf{g}_t^1\right]$. Furthermore, if the loss function satisfies $|f_t(x)| \leq M$ for all $\mathrm{x} \in \mathcal{K}$ for some constant M, then the estimated gradient vector satisfies the bounded norm inequality $\left\|\mathbf{g}_t^1\right\| \leq \frac{dM}{\delta}$.*

On the other hand, $\hat{f}_{t,\delta}$ satisfactorily approximates f_t when δ is small since $\left|\hat{f}_{t,\delta}\left(\mathrm{x}_t\right) - f_t\left(\mathrm{x}_t\right)\right| = O(\delta)$ [15, Lemma 2.6]. Thus, it suffices to evaluate f_t only at $\mathbf{x} + \delta\mathbf{u}$ in order to approximate the gradient of f_t at $\mathbf{x}$. A two-point estimator is proposed in [1], in which the player estimates the gradient by querying each loss function at two points. This requires using two loss function values $f_t\left(\mathbf{x}_t + \delta \mathbf{u}_t\right)$ and $f_t\left(\mathbf{x}_t - \delta \mathbf{u}_t\right)$ to construct a gradient estimator:

$$\mathrm{g}_t^2 = \frac{d}{2\delta}\left(f_t\left(\mathrm{x}_t + \delta \mathbf{u}_t\right) - f_t\left(\mathrm{x}_t - \delta \mathbf{u}_t\right)\right) \mathbf{u}_t, \tag{2}$$

for some $\mathbf{u}_t \sim \mathcal{S}^d$, where g_t^2 denotes an estimate of $\nabla f_t(x)$. The intuition is readily seen in the one-dimensional ($d = 1$) case, where the expectation of (2) equals $\frac{1}{2\delta}\left(f_t\left(\mathbf{x}_t + \delta\right) - f_t\left(\mathbf{x}_t - \delta\right)\right)$, which approximates the derivative of f_t at x_t if δ is small enough. Moreover, this estimator is different from the one-point estimator $\mathrm{g}_t = \frac{d}{\delta} f_t\left(\mathrm{x}_t + \delta \mathbf{u}_t\right) \mathbf{u}_t$ since its norm does not grow unboundedly as δ tends to zero.

Lemma 3 ([1])**.** *The two-point estimator defined in* (2) *is unbiased for the δ-smoothed function $\hat{f}_{t,\delta}$ in the sense that $\mathbb{E}_{\mathbf{u}_t \sim \mathcal{S}^d}\left[\mathbf{g}_t^2\right] = \nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)$. Moreover, it satisfies the bounded norm inequality $\left\|\mathrm{g}_t^2\right\| \leq Ld$.*

We make some remarks on the advantage of the two-point estimator over the one-point estimator for projection-free online convex optimization with bandit feedback. First, the two-point estimator may seem unnecessary when the loss functions are convex or strongly convex since a nontrivial combination of the one-point estimator and an online Frank-Wolfe algorithm achieves the same regret bounds as if the gradient is precisely known [11], [12]. However, under additional problem structures where the complexity of the problem reduces, the one-point estimator becomes a significant barrier to a further improvement, since the function approximation error $O(\delta)$ together with the norm of the one-point estimator being $O\left(\delta^{-1}\right)$ dominates the regret bound; see [11, Lemma 6] and [12, Lemma 8]. In contrast, the norm of the two-point estimator is bounded, which can be leveraged to further improve the regret bound.

Furthermore, if the player is allowed to query each function f_t at $(d+1)$ points, a deterministic gradient estimator can be constructed as follows:

$$\tilde{g}_t^d = \frac{1}{\delta}\sum_{i=1}^{d}\left(f_t\left(x_t + \delta e_i\right) - f_t\left(x_t\right)\right) e_i \tag{3}$$

where $\mathbf{e}_i$'s are the standard unit basis vectors and $\tilde{g}_t^d$ denotes an estimate of $\nabla f_t(x)$. With an additional smoothness assumption on the loss function, it is shown in [1] that the deterministic gradient estimator in (3) satisfactorily approximates the true gradient of the original loss function f_t.

Definition 1. *A function f is ℓ-smooth over $\mathcal{K}$ if $\|\nabla f(\mathrm{x}) - \nabla f\left(\mathrm{x}'\right)\| \leq \ell \|\mathrm{x} - \mathrm{x}'\|$ for all $\mathrm{x}, \mathrm{x}' \in \mathcal{K}$.*

Lemma 4 ([1])**.** *If the loss function f_t is ℓ-smooth, the $(d+1)$-point estimator defined in* (3) *satisfies the bounded norm inequality $\left\|\mathrm{g}_t^d\right\| \leq Ld$ and the bounded estimation error inequality $\|\mathrm{g}_t^d - \nabla f_t\left(\mathrm{x}_t\right)\| \leq \frac{\sqrt{d}\ell\delta}{2}$.*

III. Improved Regret Bounds for Strongly Convex Set

In this section, we present our algorithms along with their formal regret guarantee for online convex optimization with two-point bandit feedback, in the case when the constraint set is strongly convex. We first introduce the definition of strongly convex sets [10].

Definition 2. *Given a strictly positive number α, a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ is α-strongly convex with respect to the norm $\|\cdot\|$ if*

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

$\forall \mathrm{x}, \mathrm{x}' \in \mathcal{K}$, $\forall \gamma \in [0,1]$ and $\forall \mathbf{u} \in \mathcal{S}^d$.

Remark 1. *While strong convexity is a crucial property for improving the convergence rate/regret bound of projection-free algorithms, this assumption is not conservative and many sets used to constrain the decisions in real-world problems are strongly convex. Indeed, it is shown in [10] that various balls induced by ℓ_p norms, Schatten norms and group norms are strongly convex, where $\|\cdot\|$ refers to Frobenius norm for the later two cases. Moreover, the per-iteration cost is not a concern since the linear optimization over most of these sets is straightforward and admits a closed-form solution. For the brevity of the presentation, we will only consider ℓ_2 norm in the rest of the paper, but generalization to an arbitrary $\|\cdot\|$ is straightforward.*

A. Bandit feedback and convex loss function

For online convex optimization with full information about the gradient and general convex loss function, online Frank-Wolfe (OFW) [33] chooses an arbitrary point $\mathbf{x}_1$ from $\mathcal{K}$, and then iteratively updates its decision via the formulas:

$$\mathbf{v}_t = \mathrm{LP}_{\mathcal{K}}\left(-\nabla F_t\left(\mathbf{x}_t\right)\right)$$
$$\mathbf{x}_{t+1} = \mathbf{x}_t + \sigma_t\left(\mathbf{v}_t - \mathbf{x}_t\right)$$

where

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

is the surrogate loss function, σ_t is chosen by the line search as

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

and η is a parameter. According to [33], OFW with an appropriate choice of σ_t and η attains the regret bound of $O\left(T^{2/3}\right)$ over a strongly convex set $\mathcal{K}$. However, in the bandit setting, only the value function is available and the gradient needs to be estimated. By using the two-point estimator (2) whose bounded norm does not depend on the smoothing parameter δ, we prove that OFW with a two-point bandit feedback enjoys the following regret bound over strongly convex sets.

Theorem 1. *Let $\mathcal{K}$ be an α-strongly convex set with respect to the ℓ_2 norm, $r\mathcal{B}^d \subseteq \mathcal{K} \subseteq R\mathcal{B}^d$ and $C = \max\left(16R^2, \frac{4096}{3\alpha^2}\right)$. For every $\mathrm{x}^* \in \mathcal{K}$, Algorithm $\mathcal{A}_1$ with $\eta = \frac{R}{Ld(T+2)^{2/3}}$ and $\delta = T^{-1}$ leads to the regret bound*

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_T(\mathcal{A}_1)\right] \leq & 3L + LR/r + RLd(T+2)^{1/3} \\ & + 4RLd(T+2)^{2/3} + \frac{3}{2}L\sqrt{C}(T+2)^{2/3}.\end{aligned}$$

Algorithm 1 2-point bandit algorithm with online Frank-Wolfe (OFW)

Inputs: Horizon T, feasible set $\mathcal{K}$, η, δ
Outputs: $\mathbf{y}_{1,1}, \mathbf{y}_{2,1}, \ldots, \mathbf{y}_{1,T}, \mathbf{y}_{2,T}$
Initialization: $\mathbf{x}_1 \in \mathcal{K}_\delta$
for $t = 1, 2, \ldots,$T **do**
  $\mathbf{y}_{1,t} = \mathbf{x}_t + \delta \mathbf{u}_t$ and $\mathbf{y}_{2,t} = \mathbf{x}_t - \delta \mathbf{u}_t$ where $\mathbf{u}_t \sim \mathcal{S}^d$
  Play $\mathbf{y}_{1,t}$, $\mathbf{y}_{2,t}$ and observe $f_t(\mathbf{y}_{1,t})$, $f_t(\mathbf{y}_{2,t})$
  $\mathbf{g}_t = \frac{d}{2\delta}\left(f_t(\mathbf{y}_{1,t}) - f_t(\mathbf{y}_{2,t})\right)\mathbf{u}_t$
  Define $F_t(\mathbf{x}) = \eta \sum_{\tau=1}^{t-1} \langle \mathbf{g}_t, \mathbf{x} \rangle + \|\mathbf{x} - \mathbf{x}_1\|^2$
  $\mathbf{v}_t = \mathrm{LP}_{\mathcal{K}_\delta}\left(-\nabla F_t(\mathbf{x}_t)\right)$
  $\sigma_t = \operatorname*{argmin}_{\sigma \in [0,1]} \langle \sigma(\mathbf{v}_t - \mathbf{x}_t), \nabla F_t(\mathbf{x}_t) \rangle + \sigma^2 \|\mathbf{v}_t - \mathbf{x}_t\|^2$
  $\mathbf{x}_{t+1} = \mathbf{x}_t + \sigma_t(\mathbf{v}_t - \mathbf{x}_t)$
end for

Before proving the above theorem, we first present some useful lemmas below.

Lemma 5 ([30]). *] Let $\{\mathbf{w}_t\}_{t=1}^T$ be a sequence of vectors in $\mathcal{K}_\delta$ such that $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in \mathcal{K}_\delta} \sum_{\tau=1}^{t-1} f_\tau(\mathbf{w}) + R(\mathbf{w})$, where $R(\mathbf{w})$ is an arbitrary strongly convex function. Then, for every $\mathbf{z} \in \mathcal{K}_\delta$, it holds that*

$$\begin{aligned}\sum_{t=1}^{T}\left(f_t(\mathbf{w}_t) - f_t(\mathbf{z})\right) \leq & \sum_{t=1}^{T}\left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})\right) \\ & + R(\mathbf{z}) - R(\mathbf{w}_1).\end{aligned}$$

Lemma 6 ([33]). *Suppose that $\mathcal{K}_\delta$ is an α-strongly convex set with respect to the ℓ_2 norm. Define $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}_\delta} F_t(\mathbf{x})$ for all $t \in \{1, 2, \ldots, T+1\}$, where $F_t(\mathbf{x})$ is defined in line 8 of Algorithm $\mathcal{A}_1$. Then, Algorithm $\mathcal{A}_1$ with $\eta = \frac{R}{Ld(T+2)^{2/3}}$ leads to the inequality*

$$F_t(\mathbf{x}_t) - F_t(\mathbf{x}_t^*) \leq \epsilon_t$$

where $\epsilon_t = \frac{C}{(t+2)^{2/3}}$ and $C = \max\left(16R^2, \frac{4096}{3\alpha^2}\right)$.

Lemma 7. *Let $\mathcal{K}$ be an α-strongly convex set with respect to the ℓ_2 norm. For every $\mathrm{x}^* \in \mathcal{K}$, Algorithm $\mathcal{A}_1$ leads to the following relations:*

$$\begin{aligned}\mathcal{R}_T(\mathcal{A}_1) = & \sum_{t=1}^{T}\frac{1}{2}\left(f_t(\mathbf{y}_{1,t}) + f_t(\mathbf{y}_{2,t})\right) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \\ \leq & \sum_{t=1}^{T}\left(\hat{f}_{t,\delta}(\mathbf{x}_t) - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)\right) + 3\delta LT + \delta LRT/r.\end{aligned}$$

where $\tilde{\mathrm{x}}^*$ denotes the projection of x^* onto $\mathcal{K}_\delta$.

Proof. It holds that

$$\begin{aligned}\mathcal{R}_T(\mathcal{A}_1) = & \sum_{t=1}^{T}\frac{1}{2}\left(f_t(\mathbf{y}_{1,t}) + f_t(\mathbf{y}_{2,t})\right) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \\ = & \sum_{t=1}^{T}\frac{1}{2}\left(f_t(\mathbf{y}_{1,t}) + f_t(\mathbf{y}_{2,t})\right) - \sum_{t=1}^{T} f_t(\mathbf{x}_t) \\ & + \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\tilde{\mathbf{x}}^*) + \sum_{t=1}^{T} f_t(\tilde{\mathbf{x}}^*) - \sum_{t=1}^{T} f_t(\mathbf{x}^*).\end{aligned}$$

Since f_t is L-Lipschitz, one can write

$$\begin{aligned}& \sum_{t=1}^{T}\frac{1}{2}\left(f_t(\mathbf{y}_{1,t}) + f_t(\mathbf{y}_{2,t})\right) - \sum_{t=1}^{T} f_t(\mathbf{x}_t) \\ & = \sum_{t=1}^{T}\left[\frac{1}{2} f_t(\mathbf{x}_t + \delta\mathbf{u}_t) - f_t(\mathbf{x}_t)\right] \\ & \quad + \sum_{t=1}^{T}\left[\frac{1}{2} f_t(\mathbf{x}_t - \delta\mathbf{u}_t) - f_t(\mathbf{x}_t)\right] \\ & \leq \sum_{t=1}^{T} L\|\delta\mathbf{u}_t\| \leq \delta LT.\end{aligned}$$

Also, we have

$$\begin{aligned}& \sum_{t=1}^{T} f_t(\tilde{\mathbf{x}}^*) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \leq \sum_{t=1}^{T} L\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\| \\ & = \sum_{t=1}^{T} L\|(1 - \delta/\tau)x^* - x^*\| \leq \delta LRT/r.\end{aligned}$$

We split $f_t(\mathbf{x}_t) - f_t(\tilde{\mathbf{x}}^*)$ into $f_t(\mathbf{x}_t) - \hat{f}_{t,\delta}(\mathbf{x}_t) + \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*) - f_t(\tilde{\mathbf{x}}^*) + \hat{f}_{t,\delta}(\mathbf{x}_t) - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)$, and thus obtain

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\tilde{\mathbf{x}}^*) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \hat{f}_{t,\delta}(\mathbf{x}_t) + \sum_{t=1}^{T} \left(\hat{f}_{t,\delta}(\overline{\mathbf{x}}^*) - f_t(\tilde{\mathbf{x}}^*)\right) + \sum_{t=1}^{T} \left(\hat{f}_{t,\delta}(\mathbf{x}_t) - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)\right) \leq 2\delta LT + \sum_{t=1}^{T} \left(\hat{f}_{t,\delta}(\mathbf{x}_t) - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)\right).$$

This completes the proof. □

Now, we are ready to prove the Theorem 1.

Proof of Theorem 1. In light of Lemma 7, it suffices to bound $\sum_{t=1}^{T} \left(\mathbb{E}\left[\hat{f}_{t,\delta}(\mathbf{x}_t)\right] - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)\right)$. Let $\mathbf{x}_t^* = \operatorname{argmin}_{x \in \mathcal{K}_\delta} F_t(x)$ for every $t \in \{2, \ldots, T+1\}$. Since $\hat{f}_{t,\delta}(\mathbf{x})$ is convex, we have

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\hat{f}_{t,\delta}(\mathbf{x}_t)\right] - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)\right) \leq \sum_{t=1}^{T} \mathbb{E}\left[\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}^* + \mathbf{x}_t^* - \mathbf{x}_t^*)\right]. \quad (4)$$

Since $\left\|\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)\right\| \leq L$ and $F_t(\mathbf{x})$ is 2-strongly convex for all $t \in \{1, 2, \ldots, T\}$, it follows from Lemma 6 that

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left[\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_t^*)\right] &\leq L \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^*\|\right] \\ &\leq L \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{F_t(\mathbf{x}_t) - F_t(\mathbf{x}_t^*)}\right] \leq L\sqrt{C} \sum_{t=1}^{T} \frac{1}{(t+2)^{1/3}} \\ &\leq \frac{3}{2} L\sqrt{C}(T+2)^{2/3}. \end{aligned} \quad (5)$$

Since F_t is 2-strongly convex and $F_t(\mathbf{x}_t^*) \leq F_t\left(\mathbf{x}_{t+1}^*\right)$, we have

$$\begin{aligned}\left\|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\right\|^2 &\leq F_{t+1}(\mathbf{x}_t^*) - F_{t+1}\left(\mathbf{x}_{t+1}^*\right) \\ &= F_t(\mathbf{x}_t^*) - F_t\left(\mathbf{x}_{t+1}^*\right) + \eta \mathbf{g}_t^\top \left(\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\right) \\ &\leq \eta \|\mathbf{g}_t\| \left\|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\right\|. \end{aligned} \quad (6)$$

Thus, it holds that $\left\|\mathrm{x}_t^* - \mathrm{x}_{t+1}^*\right\| \leq \eta \|\mathrm{g}_t\|$. Let $\mathcal{F}_t$ be the σ-field generated by $\mathbf{x}_1, \mathbf{g}_1, \mathbf{x}_2, \mathbf{g}_2, \ldots, \mathbf{x}_{t-1}, \mathbf{g}_{t-1}, \mathbf{x}_t$. Note that $\mathbf{x}_t^*$ is a function of $\mathrm{g}_1, \ldots, \mathrm{g}_{t-1}$ and thus measurable with respect to $\mathcal{F}_t$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right] \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_t \mid \mathcal{F}_t\right]^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right] \\ &= \mathbb{E}\left[\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right]. \end{aligned}$$

Then, Lemma 5 yields that

$$\sum_{t=1}^{T} \mathbb{E}\left[\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{g}_t^\top \left(\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\right)\right] + \frac{1}{\eta} \|\tilde{\mathbf{x}}^* - \mathbf{x}_1\|^2 \leq \eta \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{g}_t\|^2\right] + \frac{4R^2}{\eta} \leq \eta (Ld)^2 T + \frac{4R^2}{\eta}. \quad (7)$$

Combining (4), (5) and (7) concludes that

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\hat{f}_{t,\delta}(\mathbf{x}_t)\right] - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)\right) \leq \eta (Ld)^2 T + \frac{4R^2}{\eta} + \frac{3}{2} L\sqrt{C}(T+2)^{2/3}.$$

In light of Lemma 7, the above inequality leads to

$$\mathbb{E}\left[\mathcal{R}_T(\mathcal{A}_1)\right] \leq 3\delta LT + \delta LRT/r + \eta (Ld)^2 T + \frac{4R^2}{\eta} + \frac{3}{2} L\sqrt{C}(T+2)^{2/3}.$$

Substituting $\delta = T^{-1}$ and $\eta = \frac{R}{Ld(T+2)^{2/3}}$ produces the desired bound. □

Note that the regret bound $O\left(T^{2/3}\right)$ in Theorem 1 is better than the regret bound $O\left(T^{3/4}\right)$ which is achieved by an algorithm requiring double loops over general convex sets [11]. In addition, since the approximate function error $O(\delta)$ together with the norm of the one-point estimator being $O\left(\delta^{-1}\right)$ dominates the regret bound, simply replacing the one-point estimator in [11] with the two-point estimator will not help improve the regret bound.

While bounding the expected regret is an important problem, the regret may still have a high variance. In order to ensure that Algorithm $\mathcal{A}_1$ enjoys a small regret, it is necessary to prove a bound that holds with high probability. To this end, we use the Hoeffding-Azuma inequality to derive a high probability guarantee for the convex functions over strongly convex set.

Theorem 2. *Consider Algorithm $\mathcal{A}_1$ with $\delta = T^{-1}$ and $\eta = \frac{R}{Ld(T+2)^{2/3}}$. For every $\mathrm{x} \in \mathcal{K}$ and $\xi > 0$, the inequality*

$$\mathcal{R}_T(\mathcal{A}_1) \leq LR/r + RLd(T+2)^{1/3} + 4RLd(T+2)^{2/3} + \frac{3}{2} L\sqrt{C}(T+2)^{2/3} + 2(L+Ld)R\sqrt{2T\log(1/\xi)} + 3L$$

holds with probability at least $1 - \xi$.

Proof. For each point $\tilde{\mathbf{x}}^* \in \mathcal{K}_\delta$, define

$$Z_t = \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*) - \mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*).$$

Since $\mathbb{E}\left[\nabla \hat{f}_{t,\delta}(\mathbf{x})^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right] = \mathbb{E}\left[\mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right]$ for every $\mathbf{x}$ that is independent of $\mathbf{u}_t$, we have $\mathbb{E}[Z_t] = 0$. In addition,

$$\begin{aligned}&\left\|\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*) - \mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\right\| \\ &\leq \left\|\nabla \hat{f}_{t,\delta}(\mathbf{x}_t) - \mathbf{g}_t\right\| \|(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\| \\ &\leq \left(\left\|\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)\right\| + \|\mathbf{g}_t\|\right) \|(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)\| \\ &\leq 2(L + Ld)R. \end{aligned}$$

Thus, the sequence $\{Z_t\}_{t=1}^T$ is a bounded martingale difference sequence. Using the Hoeffding-Azuma inequality, we obtain

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t > \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{2TB^2}\right)$$

where $B = 2(L + Ld)R$. Consider $\epsilon = B\sqrt{2T\log(1/\xi)}$, which leads to $\xi = \exp\left(-\frac{\epsilon^2}{2TB^2}\right)$. Hence, the following inequality holds with probability at least $1 - \xi$:

$$\sum_{t=1}^{T} \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top \left(\mathbf{x}_t^* - \sum_{t=1}^{T} \tilde{\mathbf{x}}^*\right) \leq \sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*) + 2(L + Ld)R\sqrt{2T\log(1/\xi)}.$$

Then, it results from Lemma 5 that

$$\sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*) \leq \sum_{t=1}^{T} \mathbf{g}_t^\top \left(\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\right) + \frac{1}{\eta}\|\tilde{\mathbf{x}}^* - \mathbf{x}_1\|^2 \leq \eta \sum_{t=1}^{T} \|\mathbf{g}_t\|^2 + \frac{4R^2}{\eta} \leq \eta(Ld)^2 T + \frac{4R^2}{\eta}.$$

The second inequality holds because of $\left\|\mathrm{x}_t^* - \mathrm{x}_{t+1}^*\right\| \leq \eta\|\mathrm{g}_t\|$ (see (6)) and the third inequality holds because of $\|\mathbf{g}_t\| \leq Ld$. Since $F_t(\mathbf{x})$ is 2-strongly convex for all $t = 1, \ldots, T$, using Lemma 6, we have that

$$\sum_{t=1}^{T} \mathbb{E}\left[\nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_t^*)\right] \leq L \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^*\|\right] \leq L \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{F_t(\mathbf{x}_t) - F_t(\mathbf{x}_t^*)}\right] \leq L\sqrt{C} \sum_{t=1}^{T} \frac{1}{(t+2)^{1/3}} \leq \frac{3}{2} L\sqrt{C}(T+2)^{2/3}.$$

Then, by combining the above results with Lemma 7, we obtain the desired high probability bound. □

B. Bandit feedback and strongly convex loss function

In this subsection, we propose a variant of OFW with a two-point bandit feedback for strongly convex functions, which matches the best known regret bound of $O(\sqrt{T})$ associated with the full information case for which the gradients are known completely [33]. Due to the space restriction, the proofs are moved to the appendix.

Definition 3. *Given a strictly positive number λ, a function f is λ-strongly convex over $\mathcal{K}$ if for all $\mathrm{x}, \mathrm{x}' \in \mathcal{K}$, it holds that*

$$f(\mathrm{x}') \geq f(\mathrm{x}) + (\mathrm{x}' - \mathrm{x})^\top \nabla f(\mathrm{x}) + \frac{\lambda}{2}\|\mathrm{x}' - \mathrm{x}\|^2.$$

In order to handle strongly convex losses, $F_t(x)$ is redefined as $\tilde{F}_t(\mathbf{x}) = \sum_{\tau=1}^{t-1}\left(\langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_\tau\|^2\right)$. The detailed procedures under this setting are summarized in Algorithm $\mathcal{A}_2$.

Algorithm 2 Strongly Convex Variant of 2-point bandit algorithm with online Frank-Wolfe (OFW)

Inputs: Horizon T, feasible set $\mathcal{K}$, η, δ
Outputs: $\mathbf{y}_{1,1}, \mathbf{y}_{2,1}, \ldots, \mathbf{y}_{1,T}, \mathbf{y}_{2,T}$
Initialization: $x_1 \in \mathcal{K}_\delta$
for $t = 1, 2, \ldots,$T **do**
  $\mathbf{y}_{1,t} = \mathbf{x}_t + \delta\mathbf{u}_t$ and $\mathbf{y}_{2,t} = \mathbf{x}_t - \delta\mathbf{u}_t$ where $\mathbf{u}_t \sim \mathcal{S}^d$
  Play $\mathbf{y}_{1,t}$, $\mathbf{y}_{2,t}$ and observe $f_t(\mathbf{y}_{1,t})$, $f_t(\mathbf{y}_{2,t})$
  $\mathbf{g}_t = \frac{d}{2\delta}\left(f_t(\mathbf{y}_{1,t}) - f_t(\mathbf{y}_{2,t})\right)\mathbf{u}_t$
  Define $\tilde{F}_t(\mathbf{x}) = \sum_{\tau=1}^{t-1}\left(\langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_\tau\|^2\right)$
  $\mathbf{v}_t \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}\left\langle \nabla\tilde{F}_t(\mathbf{x}_t), \mathbf{x}\right\rangle$
  $\sigma_t = \operatorname{argmin}_{\sigma\in[0,1]}\left\langle \sigma(\mathbf{v}_t - \mathbf{x}_t), \nabla\tilde{F}_t(\mathbf{x}_t)\right\rangle + \frac{\sigma^2\lambda t}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2$
  $\mathbf{x}_{t+1} = \mathbf{x}_t + \sigma_t(\mathbf{v}_t - \mathbf{x}_t)$
end for

Theorem 3. *Let $\{f_t(x)\}_{t=1}^T$ be $\lambda-$strongly convex functions, $\mathcal{K}$ be an α-strongly convex set with respect to the ℓ_2 norm, and $r\mathcal{B}^d \subseteq \mathcal{K} \subseteq R\mathcal{B}^d$. Define $\tilde{C} = \max\left(\frac{4(Ld+\lambda D)}{\lambda}, \frac{288\lambda}{\alpha^2}\right)$. for every $\mathrm{x}^* \in \mathcal{K}$, Algorithm $\mathcal{A}_2$ with $\delta = T^{-1}$ ensures that*

$$\mathbb{E}\left[\mathcal{R}_T(\mathcal{A}_2)\right] \leq 2L\sqrt{\frac{2\tilde{C}T}{\lambda}} + \frac{2(Ld + 2\lambda R)^2}{\lambda}\log(T) + 3L + LR/r.$$

Note that the regret bound $O(\sqrt{T})$ in Theorem 3 is better than the existing bound $O\left(T^{2/3}\right)$ achieved by an algorithm requiring double loops for strongly convex loss functions over general convex sets [12]. Furthermore, this regret bound $O(\sqrt{T})$ matches the regret bound attained with the full information feedback for strongly convex loss functions over general convex sets [33].

As before, we use the Hoeffding-Azuma inequality to derive a high probability guarantee for the case of strongly convex functions over a strongly convex set.

Theorem 4. *Consider Algorithm $\mathcal{A}_2$ with $\delta = T^{-1}$. For every $\mathrm{x} \in \mathcal{K}$ and $\xi > 0$, the inequality*

$$\mathcal{R}_T(\mathcal{A}_2) \leq LR/r + 2L\sqrt{\frac{2CT}{\lambda}} + \frac{2(Ld + 2\lambda R)^2}{\lambda}\log(T) + 2(L + Ld)R\sqrt{2T\log(1/\xi)} + 3L.$$

holds with probability at least $1 - \xi$.

Compared with Theorem 2, the high-probability regret bound is improved to $O(\sqrt{T})$ by utilizing the strong convexity of the loss functions.

IV. Improved Regret Bounds for Smooth Loss Functions

In this section, we present our algorithm along with its formal regret guarantee for online convex optimization with $(d+1)$ points bandit feedback under the assumption that the loss function is smooth. A projection-free algorithm has been recently proposed in [20] for online convex optimization with

smooth loss functions over general convex sets, which has led to the improved regret bound $O\left(T^{2/3}\right)$ by leveraging the smoothness. That algorithm is not based on the online Frank-Wolfe method, but rather a version of the Follow-the-Perturbed-Leader (FPT) method [23]. That algorithm is based on an online primal-dual methodology, which uses the smoothness of the loss function to control the error-propagation caused by the random estimation of the mean (as the output of the FPT). However, in the bandit setting, a new challenge arises when the gradient information is unavailable. In order to leverage the smoothness in FPL, an accurate estimate of the gradient with a small estimation error is required. Due to the space restriction, the proofs of the results of this section are moved to the appendix.

Theorem 5. *Let $\{f_t(x)\}_{t=1}^T$ be ℓ-smooth convex functions. For every $\mathrm{x} \in \mathcal{K}$, Algorithm $\mathcal{A}_3$ with $\sigma = 2/L\sqrt{d}T^{-2/3}, \delta = T^{-1}$ and $k = T^{1/3}$ yields that*

$$\begin{aligned}&\mathcal{R}_T\left(\mathcal{A}_3\right)\\ =&\sum_{t=1}^{T}\frac{1}{1+d}\left(f_t\left(\mathrm{x}_t\right)+\sum_{i=1}^{d}f_t\left(\mathrm{x}_t+\delta\mathbf{e}_i\right)\right)-\sum_{t=1}^{T}f_t\left(\mathrm{x}^*\right)\\ \leq& L+LR/r+RL\sqrt{d}T^{2/3}+4\ell R^2T^{2/3}+\frac{dR\ell}{2}T^{1/3}.\end{aligned}$$

Note that the regret bound $O\left(T^{2/3}\right)$ in Theorem 5 is better than the bound $O\left(T^{3/4}\right)$ which is achieved by an algorithm requiring double loops for arbitrary convex loss functions [11]. Furthermore, this regret bound $O\left(T^{2/3}\right)$ matches the regret bound attained via a full information feedback in the smooth losses setting [20].

Algorithm 3 Online smooth projection-free algorithm with multi-point bandit feedback

Inputs: Horizon T, feasible set $\mathcal{K}$, η, δ
Outputs: $\{\mathbf{y}_{i,1}\}_{i=1}^d, \{\mathbf{y}_{i,2}\}_{i=1}^d, \ldots, \{\mathbf{y}_{i,T}\}_{i=1}^d$
Initialization: $m = 0$, $x_0 \in \mathcal{K}_\delta$
for $t = 1, 2, \ldots,$T **do**
  if t mod $k \neq 0$ **then**
    $\mathbf{x}_t = \mathbf{x}_{t-1}$.
    $\mathbf{y}_{i,t} = \mathbf{x}_t + \delta\mathbf{e}_i$ and observe $f_t(\mathbf{y}_{i,t})$ for $i = 1, \ldots, d$.
    $\mathbf{g}_t = \frac{1}{\delta}\sum_{i=1}^d \left(f_t(\mathbf{y}_{i,t}) - f_t(\mathbf{x}_{i,t})\right)\mathbf{e}_i$.
  else
    sample $\mathbf{v}_{t-k+j} \sim \mathcal{B}^d$ uniformly for $j = 1, \ldots, k$.
    $\mathbf{x}_t^j = \mathrm{LP}_{\mathcal{K}_\delta}(\sum_{\tau=1}^{t-1}\mathbf{g}_\tau + \frac{1}{\sigma}\mathbf{v}_t^j)$ for $j = 1, \ldots, k$ and $\mathbf{x}_t = \frac{1}{k}\sum_{j=1}^k \mathbf{x}_t^j$.
    play $\mathbf{y}_{i,t} = \mathbf{x}_t + \delta\mathbf{e}_i$ and observe $f_t(\mathbf{y}_{i,t})$ for $i = 1, \ldots, d$.
    $\mathbf{g}_t = \frac{1}{\delta}\sum_{i=1}^d \left(f_t(\mathbf{y}_{i,t}) - f_t(\mathbf{x}_{i,t})\right)\mathbf{e}_i$.
  end if
end for

Even though expected regret is a widely accepted metric for online randomized algorithms, it is essential to understand whether the expectation bound holds only due to a balance of large and small chunks of regret or the given result holds most of the time. To address this question, we use the Hoeffding-Azuma inequality to derive a high probability guarantee for the case of convex, smooth functions over a general convex set.

Theorem 6. *Consider Algorithm $\mathcal{A}_3$ with $\sigma = 2/L\sqrt{d}T^{-2/3}, \delta = T^{-1}$ and $k = T^{1/3}$. For every $\mathbf{x} \in \mathcal{K}$ and $\xi > 0$, the inequality*

$$\begin{aligned}&\mathcal{R}_T\left(\mathcal{A}_3\right) \leq L + LR/r + RL\sqrt{d}T^{2/3} + \frac{dR\ell}{2}T^{1/3}\\ &+ 2LR\sqrt{2\log\left(2/\xi\right)}T^{2/3} + 8\ell R^2\log\left(4T^{1/3}/\xi\right)T^{1/3}.\end{aligned}$$

holds with probability at least $1-\xi$.

V. Numerical Examples

We conduct numerical experiments to evaluate the performance of the proposed algorithms using the task of online matrix completion. Let $\{\mathbf{M}_t\}_{t=1}^T$ be symmetric positive semi-definite matrices, where $\mathbf{M}_t = \mathbf{N}_t^\top\mathbf{N}_t$ such that each entry of $\mathbf{N}_t \in \mathbb{R}^{k\times n}$ obeys the standard normal distribution. At each iteration, half of the entries of $\mathbf{M}_t$ are observed. We set $n = 20$ and $k = 18$. We denote the set of entries of $\mathbf{M}_t$ observed at the t-th iteration by $\mathbf{O}_t$. It is desirable to minimize $f_t\left(\mathbf{X}_t\right) = \frac{1}{2}\sum_{(i,j)\in\mathbf{O}_t}\left(\mathbf{X}_t[i,j] - \mathbf{M}_t[i,j]\right)^2$ subject to $\|\mathbf{X}_t\|_* \leq k$, where $\mathbf{X}_t$ denotes the optimization variable of the same dimension as $\mathbf{M}_t$, $\mathbf{X}_t[i,j]$ denotes the (i,j) entry of $\mathbf{X}_t$ and $\|\cdot\|_*$ denotes the nuclear norm. The nuclear norm constraint is a standard convex relaxation of the rank constraint $\mathrm{rank}(\mathbf{X}) \leq k$. In this example, the loss function f_t is smooth and the constraint set is strongly convex [10]. Thus, Algorithms $\mathcal{A}_1$ and $\mathcal{A}_3$ are applicable. The linear optimization step in Line 9 of Algorithm $\mathcal{A}_1$ (or the linear optimization step in Line 11 of Algorithm $\mathcal{A}_3$) has the closed-form solution $k\mathbf{v}_{\max}\mathbf{v}_{\max}^\top$, where $\mathbf{v}_{\max}$ denotes the unit eigenvector of the largest eigenvalue of $-\nabla F_t\left(\mathbf{X}_t\right)$ (or $-\sum_{\tau=1}^{t-1}\mathbf{g}_\tau + \frac{1}{\sigma}\mathbf{v}_{t-k+j}$) (see [15]). We compare our proposed Algorithms $\mathcal{A}_1$ and $\mathcal{A}_3$ with the baseline approaches: (1) OFW: OFW with the *full information* feedback over strongly convex sets [33], (2) OSPF: Online smooth projection free algorithm with the *full information* feedback [20], (3) PFBCO: OFW with the *bandit* feedback over arbitrary convex sets [5] (we replace the one-point estimator in the original algorithm with the two-pointer estimator to demonstrate that the improvement in our algorithm is indeed due to utilizing the structural properties of the problem). For all of these algorithms, we report the average loss defined as $\sum_{t=1}^T f_t\left(\mathbf{X}_t\right)/T$. All of the experiments are conducted in Python3 on a workstation with a 3.1 GHz Intel Core i 5 and 8GB memory, equipped with macOS 10.14.6.

We can observe from Fig. 1 that the average losses of Algorithms $\mathcal{A}_1$ and $\mathcal{A}_3$ almost overlap with the average losses of OFW and OSPF, respectively, when the problem possesses structural properties, such as strongly convexity of the feasible set or the smoothness of the loss function. This demonstrates the power of bandit feedback on utilizing the problem structures in the simple projection-free algorithm even if the exact gradient information is unavailable. In

addition, Algorithms $\mathcal{A}_1$ and $\mathcal{A}_3$ have a similar average loss, which supports the results in Theorem 1 and Theorem 5 stating that Algorithms $\mathcal{A}_1$ and $\mathcal{A}_3$ enjoy regret bounds of the same order. Finally, the improvement of our methods over the previous method with the simple algorithm [5] can be seen from the comparison of the average losses of Algorithm $\mathcal{A}_1$ and PFBCO.

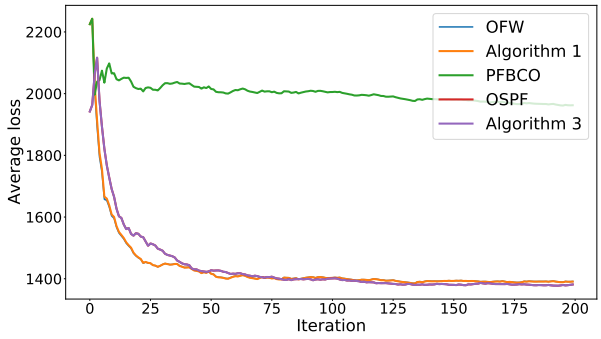

Fig. 1. We show the average loss versus the number of iterations for 5 algorithms on the example of online matrix completion. The curve of OFW and OSPF almost overlap with Algorithms $\mathcal{A}_1$ and $\mathcal{A}_3$, respectively.

VI. CONCLUSION

In this paper, we developed efficient projection-free algorithms for structured online convex optimization with multi-point bandit feedback, leading to improved regret bounds. More specifically, we developed a projection-free algorithm with two-point bandit feedback achieving the regret bound $O(T^{2/3})$ when loss functions are convex and constraint sets are strongly convex. This regret bound can be further improved to $O(\sqrt{T})$ if the loss functions are strongly convex. In addition, we developed a projection-free algorithm with multi-point bandit feedback achieving the regret bound $O(T^{2/3})$ when loss functions are smooth and constraint sets are convex. These bounds match that for the full-information setting, demonstrating again the power of zero-order optimization. These algorithms are thus an effective alternative to the existing best-known projection-free bandit algorithms if two-point bandit feedback is available. The results of this work are applicable to practical problems where multiple function values can be sampled at each time instance.

REFERENCES

[1] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40, 2010.

[2] Alekh Agarwal, Dean P Foster, Daniel Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

[3] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[4] S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: \sqrtt regret in one dimension. In *COLT*, pages 266–278, 2015.

[5] L. Chen, M. Zhang, and A. Karbasi. Projection-free bandit convex optimization. In *AISTATS*, pages 2047–2056. PMLR, 2019.

[6] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[7] M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *AISTATS*, pages 327–336, 2012.

[8] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, pages 385–394, 2005.

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

[10] D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *ICML*, pages 541–549. PMLR, 2015.

[11] D. Garber and B. Kretzu. Improved regret bounds for projection-free bandit convex optimization. In *AISTATS*, pages 2196–2206. PMLR, 2020.

[12] D. Garber and B. Kretzu. Revisiting projection-free online learning: the strongly convex case. *ArXiv Preprint: 2010.07572*, 2020.

[13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[14] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick. Large-scale image classification with trace-norm regularization. In *CVPR*, pages 3386–3393. IEEE, 2012.

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

[16] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[17] E. Hazan and S. Kale. Projection-free online learning. In *ICML*, pages 1843–1850. PMLR, 2012.

[18] E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *NeurIPS*, pages 784–792, 2014.

[19] E. Hazan and Y. Li. An optimal algorithm for bandit convex optimization. *ArXiv Preprint: 1603.04350*, 2016.

[20] E. Hazan and E. Minasyan. Faster projection-free online learning. In *COLT*, pages 1877–1893. PMLR, 2020.

[21] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435. PMLR, 2013.

[22] M. Jaggi and M. Sulovskỳ. A simple algorithm for nuclear norm regularized problems. In *ICML*, pages 471–478. PMLR, 2010.

[23] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[24] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[25] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural svms. In *ICML*, pages 53–61. PMLR, 2013.

[26] S. Laue. A hybrid algorithm for convex semidefinite optimization. In *ICML*, pages 1083–1090, 2012.

[27] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[28] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[29] A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *AISTATS*, pages 636–642, 2011.

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

[31] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

[32] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *ICML*, pages 720–727, 2003.

[33] Y. Wan and L. Zhang. Projection-free online learning over strongly convex sets. *ArXiv Preprint: 2010.08177*, 2020.

[34] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936. PMLR, 2003.

APPENDIX

A. Proof of Theorems in Section III-B

First, we state two lemmas below.

Lemma 8 ([30]). *Let $\{\mathbf{w}_t\}_{t=1}^T$ be a sequence of vectors in $\mathcal{K}_\delta$ defined as $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}\in\mathcal{K}_\delta} \sum_{\tau=1}^{t-1} f_\tau(\mathbf{w})$. Then, for every $\mathbf{z} \in \mathcal{K}_\delta$ we have*

$$\sum_{t=1}^{T} \left(f_t\left(\mathbf{w}_t\right) - f_t(\mathbf{z})\right) \leq \sum_{t=1}^{T} \left(f_t\left(\mathbf{w}_t\right) - f_t\left(\mathbf{w}_{t+1}\right)\right).$$

Lemma 9 ([33]). *Assume that $\mathcal{K}_\delta$ is an α-strongly convex set with respect to the ℓ_2 norm. Let $\mathbf{x}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \tilde{F}_t(\mathbf{x})$ for all $t \in \{1,\ldots,T\}$, where $\tilde{F}_t(\mathbf{x})$ is defined in line 8 of Algorithm $\mathcal{A}_2$. Then, Algorithm $\mathcal{A}_2$ gives rise to*

$$\tilde{F}_t\left(\mathbf{x}_t\right) - \tilde{F}_t\left(\mathbf{x}_t^*\right) \leq \tilde{C}, \quad \forall t \in \{1,\ldots,T\},$$

where $\tilde{C} = \max\left(\frac{4(Ld+2\lambda R)}{\lambda}, \frac{288\lambda}{\alpha^2}\right)$.

Proof of Theorem 3. In light of Lemma 7, it suffices to bound $\sum_{t=1}^T \left(\mathbb{E}\left[\hat{f}_{t,\delta}\left(\mathbf{x}_t\right)\right] - \hat{f}_{t,\delta}\left(\tilde{\mathbf{x}}^*\right)\right)$. Let $\mathbf{x}_t^* = \operatorname{argmin}_{x\in\mathcal{K}_\delta} \tilde{F}_t(x)$ for all $t \in \{2,\ldots,T+1\}$. Since $\hat{f}_{t,\delta}(\mathbf{x})$ is λ-strongly convex, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\mathrm{E}\left[\hat{f}_{t,\delta}\left(\mathbf{x}_t\right)\right] - \hat{f}_{t,\delta}\left(\tilde{\mathbf{x}}^*\right)\right) \\
&\leq \sum_{t=1}^{T} \mathbf{E}\left[\nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t - \tilde{\mathbf{x}}^*\right) - \frac{\lambda}{2}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\right\|^2\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t - \mathbf{x}_t^*\right) + \nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right)\right] \\
&- \sum_{t=1}^{T} \mathbb{E}\left[\frac{\lambda}{2}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\right\|^2\right]. \qquad (8)
\end{aligned}$$

Since $\left\|\nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)\right\| \leq L$ and $F_{t+1}(\mathbf{x})$ is $t\lambda$-strongly convex, for all $t \in \{1,2,\ldots,T\}$, it follows from Lemma 8 that

$$\begin{aligned}
&\sum_{t=1}^{T} \mathrm{E}\left[\nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t - \mathbf{x}_t^*\right)\right] \\
&\leq L \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{\frac{2\left(\tilde{F}_t\left(\mathbf{x}_t\right) - \tilde{F}_t\left(\mathbf{x}_t^*\right)\right)}{t\lambda}}\right] \leq 2L\sqrt{\frac{2\tilde{C}T}{\lambda}}. \qquad (9)
\end{aligned}$$

Let $\tilde{g}_t(\mathbf{x}) = \langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{\lambda}{2}\left\|\mathbf{x} - \mathbf{x}_t\right\|^2$ for all $t \in \{1,2,\ldots,T\}$. Since $\tilde{g}_t(\mathbf{x})$ is $(\|\mathbf{g}_t\| + 2\lambda R)$-Lipschitz over $\mathcal{K}$, we have

$$\tilde{g}_t\left(\mathbf{x}_t^*\right) - \tilde{g}_t\left(\mathbf{x}_{t+1}^*\right) \leq \left(\|\mathbf{g}_t\| + 2\lambda R\right)\left\|\mathbf{x}_t^* - \mathbf{x}_{t+1}^*\right\|. \qquad (10)$$

Moreover, the inequality $\tilde{F}_t\left(\mathbf{x}_t^*\right) \leq \tilde{F}_t\left(\mathbf{x}_{t+1}^*\right)$ leads to

$$\begin{aligned}
\left\|\mathrm{x}_t^* - \mathrm{x}_{t+1}^*\right\|^2 &\leq \frac{2}{\lambda t}\left(F_{t+1}\left(\mathrm{x}_t^*\right) - F_{t+1}\left(\mathrm{x}_{t+1}^*\right)\right) \\
&= \frac{2}{\lambda t}\left(\tilde{F}_t\left(\mathrm{x}_t^*\right) - \tilde{F}_t\left(\mathrm{x}_{t+1}^*\right) + \tilde{g}_t\left(\mathrm{x}_t^*\right) - \tilde{g}_t\left(\mathrm{x}_{t+1}^*\right)\right) \\
&\leq \frac{2\left(\|\mathrm{g}_t\| + 2\lambda R\right)}{\lambda t}\left\|\mathrm{x}_t^* - \mathrm{x}_{t+1}^*\right\|.
\end{aligned}$$

This together with (10) yields that

$$\begin{aligned}
&\sum_{t=1}^{T} \tilde{g}_t\left(\mathbf{x}_t^*\right) - \tilde{g}_t\left(\mathbf{x}_{t+1}^*\right) \leq \sum_{t=1}^{T} \frac{2\left(\|\mathbf{g}_t\| + 2\lambda R\right)^2}{\lambda t} \\
&\leq 2(Ld + 2\lambda R)^2 \sum_{t=1}^{T} \frac{1}{\lambda t} = \frac{2(Ld+2\lambda R)^2}{\lambda}\log(T). \qquad (11)
\end{aligned}$$

Let $\mathcal{F}_t$ be the σ-field generated by $\mathbf{x}_1, \mathbf{g}_1, \mathbf{x}_2, \mathbf{g}_2, \ldots, \mathbf{x}_{t-1}, \mathbf{g}_{t-1}, \mathbf{x}_t$. Note that $\mathbf{x}_t^*$ is a function of $\mathrm{g}_1, \ldots, \mathrm{g}_{t-1}$ and thus measurable with respect to $\mathcal{F}_t$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{g}_t^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_t^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right)\right] \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_t \mid \mathcal{F}_t\right]^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right)\right] \\
&= \mathbb{E}\left[\nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right)\right].
\end{aligned}$$

Then, from the definition of $\tilde{g}_t(\mathbf{x})$ and the fact that $\frac{\lambda}{2}\left\|\mathbf{x}_t^* - \mathbf{x}_t\right\|^2 \geq 0$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{E}\left[\nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right) - \frac{\lambda}{2}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\right\|^2\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{g}_t^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right) - \frac{\lambda}{2}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\right\|^2\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\tilde{g}_t\left(\mathbf{x}_t^*\right) - \tilde{g}_t\left(\tilde{\mathbf{x}}^*\right)\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\tilde{g}_t\left(\mathbf{x}_t^*\right) - \tilde{g}_t\left(\mathbf{x}_{t+1}^*\right)\right] \\
&\leq \frac{2(Ld+2\lambda R)^2}{\lambda}\log(T). \qquad (12)
\end{aligned}$$

The second inequality is due to Lemma 8 and the last inequality is due to (11). Combining (8),(9) and (12) gives rise to

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\mathbb{E}\left[\hat{f}_{t,\delta}\left(\mathbf{x}_t\right)\right] - \hat{f}_{t,\delta}\left(\tilde{\mathbf{x}}^*\right)\right) \\
&\leq \frac{2(Ld+2\lambda R)^2}{\lambda}\log(T) + 2L\sqrt{\frac{2\tilde{C}T}{\lambda}}
\end{aligned}$$

Combining the last result with Lemma 7, we obtain the bound

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T\right] \leq &\frac{2(Ld+2\lambda R)^2}{\lambda}\log(T) + 2L\sqrt{\frac{2\tilde{C}T}{\lambda}} \\
&+ 3\delta LT + \delta LRT/r.
\end{aligned}$$

Substituting $\delta = T^{-1}$, we obtain the desire bound. □

Proof of Theorem 4. Define $\mathbf{x}_t^* = \operatorname{argmin}_{x\in\mathcal{K}_\delta} \tilde{F}_t(x)$. Due to the proof of Theorem 2, we know that for every point $\tilde{\mathbf{x}}^* \in \mathcal{K}_\delta$, the points

$$Z_t = \nabla \hat{f}_{t,\delta}\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right) - \mathbf{g}_t^\top \left(\mathbf{x}_t^* - \tilde{\mathbf{x}}^*\right)$$

construct a bounded martingale difference sequence. Using the Hoeffding-Azuma inequality, we have that

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t > \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{2TB^2}\right)$$

where $B = 2(L + Ld)R$. Define $\epsilon = B\sqrt{2T\log(1/\xi)}$, leading to $\xi = \exp\left(-\frac{\epsilon^2}{2TB^2}\right)$. Hence, the inequality

$$\sum_{t=1}^{T} \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top \left(\mathbf{x}_t^* - \sum_{t=1}^{T} \tilde{\mathbf{x}}^*\right) \leq \sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{x}_t^* - \overline{\mathbf{x}}^*) + 2(L + Ld)R\sqrt{2T\log(1/\xi)}$$

holds with probability at least $1-\xi$. Define $\tilde{g}_t(\mathbf{x}) = \langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$ for every $t \in \{1, 2, \ldots, T\}$. Using the fact that $\frac{\lambda}{2}\|\mathbf{x}_t^* - \mathbf{x}_t\|^2 \geq 0$, one can write

$$\sum_{t=1}^{T} \mathbf{g}_t^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*) - \frac{\lambda}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\|^2 \leq \sum_{t=1}^{T} \tilde{g}_t(\mathbf{x}_t^*) - \bar{g}_t(\overline{\mathbf{x}}^*)$$
$$\leq \sum_{t=1}^{T} \tilde{g}_t(\mathbf{x}_t^*) - \tilde{g}_t\left(\mathbf{x}_{t+1}^*\right) \leq \frac{2(Ld + 2\lambda R)^2}{\lambda}\log(T). \quad (13)$$

The second inequality is due to Lemma 8 and the last inequality is due to (11). Since $F_{t+1}(\mathbf{x})$ is $t\lambda$-strongly convex, it follows from Lemma 9 that

$$\sum_{t=1}^{T} \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_t^*) \leq L\sum_{t=1}^{T}\sqrt{\frac{2\left(\tilde{F}_t(\mathbf{x}_t) - \tilde{F}_t(\mathbf{x}_t^*)\right)}{t\lambda}} \leq 2L\sqrt{\frac{2\tilde{C}T}{\lambda}}.$$

Then, by combining the above equation with (13) and the fact that each $\hat{f}_{t,\delta}(\mathbf{x})$ is λ-strongly convex, we have

$$\sum_{t=1}^{T} \hat{f}_{t,\delta}(\mathbf{x}_t) - \hat{f}_{t,\delta}(\tilde{\mathbf{x}}^*)$$
$$\leq \sum_{t=1}^{T} \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}^*) - \frac{\lambda}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\|^2$$
$$= \sum_{t=1}^{T} \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_t^*) + \nabla \hat{f}_{t,\delta}(\mathbf{x}_t)^\top (\mathbf{x}_t^* - \tilde{\mathbf{x}}^*)$$
$$- \frac{\lambda}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}^*\|^2$$
$$\leq 2L\sqrt{\frac{2\tilde{C}T}{\lambda}} + \frac{2(Ld + 2\lambda R)^2}{\lambda}\log(T)$$
$$+ 2(L + Ld)R\sqrt{2T\log(1/\xi)}.$$

Then, by combining the above results with Lemma 7, we obtain the required high probability bound. □

B. Proof of Theorems in Section IV

Algorithm 4 Sampled follow-the-perturbed-leader algorithm with multi-point bandit feedback

Inputs: Horizon T, feasible set $\mathcal{K}$, η, δ
Outputs: $\{\mathbf{y}_{i,1}\}_{i=1}^d, \{\mathbf{y}_{i,2}\}_{i=1}^d, \ldots, \{\mathbf{y}_{i,T}\}_{i=1}^d$
Initialization: $x_0 \in \mathcal{K}_\delta$
for $t = 1, 2, \ldots,$T **do**
  sample $\mathbf{v}_t^j \sim \mathcal{B}^d$ uniformly for $j = 1, \ldots, k$.
  $\mathbf{x}_t^j = \mathrm{LP}_{\mathcal{K}_\delta}(\sum_{\tau=1}^{t-1} \mathbf{g}_\tau + \frac{1}{\sigma}\mathbf{v}_t^j)$ for $j = 1, \ldots, k$ and $\mathbf{x}_t = \frac{1}{k}\sum_{j=1}^{k} \mathbf{x}_t^j$.
  play $\mathbf{y}_{i,t} = \mathbf{x}_t + \delta\mathbf{e}_i$ and observe $f_t(\mathbf{y}_{i,t})$ for $i = 1, \ldots, d$.
  $\mathbf{g}_t = \frac{1}{\delta}\sum_{i=1}^{d}(f_t(\mathbf{y}_{i,t}) - f_t(\mathbf{x}_{i,t}))\,\mathbf{e}_i$.
end for

Before proving the regret bound for Algorithm $\mathcal{A}_3$, we first consider an algorithm that mimics the expected FPL by replacing the computationally expensive expectations with empirical averages of independent and identically distributed (i.i.d.) samples. This is provided in detail in Algorithm $\mathcal{A}_4$. The difference between Algorithm $\mathcal{A}_3$ and Algorithm $\mathcal{A}_4$ lies in whether the blocking technique is used to group several game iterations into one and thereby changing the decision less often. We will first prove a regret bound for Algorithm $\mathcal{A}_4$ and show that Algorithm $\mathcal{A}_3$ can actually be reduced from Algorithm $\mathcal{A}_4$. The following two lemmas are useful in studying Algorithm $\mathcal{A}_4$.

Lemma 10 ([20]). *Given an L-Lipschitz function $g : \mathbb{R}^d \to \mathbb{R}$, the function $\hat{g}(\mathbf{y}) = \mathbb{E}_{\mathbf{v}\sim\mathbb{B}}\left[g\left(\mathbf{y} + \frac{1}{\sigma}\cdot\mathbf{v}\right)\right]$ is σdL-smooth.*

Lemma 11 ([20]). *Given an ℓ-smooth function $f_t : \mathcal{K} \to \mathbb{R}$, let $\hat{\mathbf{x}}_t = \mathbb{E}_{\xi_t}[\mathbf{x}_t]$, where $\mathbf{x}_t$ is generated by Algorithm $\mathcal{A}_4$ and $\xi_t = \left\{\mathrm{v}_t^1, \ldots, \mathrm{v}_t^k\right\}$ comprises the randomness used at iteration t. Then,*

$$\mathbb{E}_{\xi_{1:T}}[\langle \nabla_t, \mathbf{x}_t - \hat{\mathbf{x}}_t\rangle] \leq \frac{4\ell R^2}{k}.$$

Lemma 12. *For every $\mathrm{x}^* \in \mathcal{K}$, Algorithm $\mathcal{A}_4$ yields that*

$$\mathbb{E}[\mathcal{R}_T(\mathcal{A}_4)]$$
$$= \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{1+d}\left(f_t(\mathbf{x}_t) + \sum_{i=1}^{d} f_t(\mathbf{x}_t + \delta\mathbf{e}_i)\right) - \sum_{t=1}^{T} f_t(\mathbf{x}^*)\right]$$
$$\leq \frac{2R}{\sigma} + \frac{\sigma dRL^2}{2}T + \frac{\delta\sigma d^{3/2}RL\ell}{4}T^2$$
$$+ \frac{4\ell R^2}{k} + \delta LT + \delta LRT/r.$$

Proof. Let $\tilde{\mathbf{x}}^*$ be the projection of $\mathbf{x}^*$ onto $\mathcal{K}_\delta$. It holds that

$$\begin{aligned}\mathcal{R}_T =& \sum_{t=1}^{T} \frac{1}{1+d}\left(f_t\left(\mathrm{x}_t\right)+\sum_{i=1}^{d} f_t\left(\mathrm{x}_t+\delta \mathrm{e}_i\right)\right)-\sum_{t=1}^{T} f_t\left(\mathrm{x}^*\right) \\ =& \sum_{t=1}^{T} \frac{1}{1+d}\left(f_t\left(\mathrm{x}_t\right)+\sum_{i=1}^{d} f_t\left(\mathrm{x}_t+\delta \mathrm{e}_i\right)\right)-\sum_{t=1}^{T} f_t\left(\mathrm{x}_t\right) \\ &+\sum_{t=1}^{T} f_t\left(\mathrm{x}_t\right)-\sum_{t=1}^{T} f_t\left(\tilde{\mathrm{x}}^*\right)+\sum_{t=1}^{T} f_t\left(\tilde{\mathrm{x}}^*\right)-\sum_{t=1}^{T} f_t\left(\mathrm{x}^*\right).\end{aligned}$$

Since f_t is L-Lipschitz, one can write

$$\begin{aligned}&\sum_{t=1}^{T} \frac{1}{1+d}\left(f_t\left(\mathrm{x}_t\right)+\sum_{i=1}^{d} f_t\left(\mathrm{x}_t+\delta \mathbf{e}_i\right)\right)-\sum_{t=1}^{T} f_t\left(\mathrm{x}_t\right) \\ =& \sum_{t=1}^{T} \frac{1}{1+d} \sum_{i=1}^{d}\left(f_t\left(\mathrm{x}_t+\delta \mathrm{e}_i\right)-f_t\left(\mathrm{x}_t\right)\right) \\ \leq& \sum_{t=1}^{T} L\left\|\delta \mathrm{e}_i\right\| \leq \delta L T.\end{aligned}$$

Furthermore,

$$\begin{aligned}&\sum_{t=1}^{T} f_t\left(\tilde{\mathbf{x}}^*\right)-\sum_{t=1}^{T} f_t\left(\mathbf{x}^*\right) \leq \sum_{t=1}^{T} L\left\|\tilde{\mathbf{x}}^*-\mathbf{x}^*\right\| \\ &=\sum_{t=1}^{T} L\left\|(1-\delta / r) x^*-x^*\right\| \leq \delta L R T / r.\end{aligned}$$

Thus, it only remains to bound $\sum_{t=1}^{T} f_t\left(\mathrm{x}_t\right)-\sum_{t=1}^{T} f_t\left(\tilde{\mathrm{x}}^*\right)$. The proof of its regret bound is based on duality for the optimization problem

$$\min_{\mathbf{x} \in \mathcal{K}_\delta}\left\{\sum_{t=1}^{T} f_t(\mathbf{x})+h_\sigma(\mathbf{x})\right\}$$

which resembles the loss suffered by the best-in-hindsight fixed action. The dual objective, that is to be maximized, can be obtained using Lagrange multipliers and is given by

$$\mathcal{D}\left(\lambda_1, \ldots, \lambda_T\right)=-h_\sigma^*\left(-\mu_{1: T}\right)-\sum_{t=1}^{T} f_t^*\left(\mu_t\right).$$

The term $h_\sigma(\cdot)$ serves as a regularization and is defined implicitly through its Fenchel conjugate $h_\sigma^*(\mathbf{y}) = \mathbb{E}_{\mathbf{v} \sim \mathbb{B}}\left[\mathrm{VAL}_{\mathcal{K}_\delta}\left(\mathbf{y}+\frac{1}{\sigma} \mathbf{v}\right)\right]$, a stochastic smoothing of the value oracle that is $\delta d R$-smooth according to Lemma 10 and the fact that $\mathrm{VAL}_{\mathcal{K}_\delta}$ is R-Lipschitz.

We select $\mu_t=\nabla f_t\left(\mathrm{x}_t\right)$ for all $t \in\{1,2, \ldots, T\}$, where x_t is generated by Algorithm $\mathcal{A}_4$, and use the shorthand notation $\nabla_t=\nabla f_t\left(\mathrm{x}_t\right)$. Denote the incremental difference as $\Delta_t = \mathcal{D}\left(\nabla_1, \ldots, \nabla_t, \mathbf{0}, \ldots, \mathbf{0}\right)-\mathcal{D}\left(\nabla_1, \ldots, \nabla_{t-1}, \mathbf{0}, \ldots, \mathbf{0}\right)$ and notice that the dual can be written as $\mathcal{D}\left(\nabla_1, \ldots, \nabla_T\right) = \sum_{t=1}^{T} \Delta_t+\mathcal{D}(\mathbf{0}, \ldots, 0)$. For each $t \in\{1, \ldots, T\}$, we have

$$\begin{aligned}\Delta_t =& -\left[h_\sigma^*\left(-\sum_{i=1}^{t} \nabla_i\right)-h_\sigma^*\left(-\sum_{i=1}^{t-1} \nabla_i\right)\right] \\ &-f_t^*\left(\nabla_t\right)+f_t^*(\mathbf{0}) \\ \geq& \left\langle \nabla_t, \nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \nabla_i\right)\right\rangle-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2 \\ &-f_t^*\left(\nabla_t\right)+f_t^*(\mathbf{0}) \\ =& \left\langle \nabla_t, \nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \mathbf{g}_i\right)\right\rangle \\ &+\left\langle \nabla_t, \nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \nabla_i\right)-\nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \mathbf{g}_i\right)\right\rangle \\ &-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2-f_t^*\left(\nabla_t\right)+f_t^*(\mathbf{0}) \\ \geq& \left\langle \nabla_t, \nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \mathbf{g}_i\right)\right\rangle \\ &-L\left\|\nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \nabla_i\right)-\nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \mathbf{g}_i\right)\right\| \\ &-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2-f_t^*\left(\nabla_t\right)+f_t^*(\mathbf{0}) \\ \geq& \left\langle \nabla_t, \nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \mathbf{g}_i\right)\right\rangle \\ &-\sigma L d R\left\|\sum_{i=1}^{t-1} \nabla_i-\sum_{i=1}^{t-1} \mathbf{g}_i\right\|-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2 \\ &-f_t^*\left(\nabla_t\right)+f_t^*(\mathbf{0}) \\ =& \left\langle \nabla_t, \hat{\mathbf{x}}_t\right\rangle-f_t^*\left(\nabla_t\right)-\sigma L d R\left\|\sum_{i=1}^{t-1} \nabla_i-\sum_{i=1}^{t-1} \mathbf{g}_i\right\| \\ &-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2+f_t^*(\mathbf{0}) \\ =& \left\langle \nabla_t, \mathbf{x}_t\right\rangle+\left\langle \nabla_t, \hat{\mathbf{x}}_t-\mathbf{x}_t\right\rangle-f_t^*\left(\nabla_t\right) \\ &-\sigma L d R\left\|\sum_{i=1}^{t-1} \nabla_i-\sum_{i=1}^{t-1} \mathbf{g}_i\right\|-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2+f_t^*(\mathbf{0}) \\ =& f_t\left(\mathbf{x}_t\right)+\left\langle \nabla_t, \hat{\mathbf{x}}_t-\mathbf{x}_t\right\rangle-\frac{\sigma d R}{2}\left\|\nabla_t\right\|^2+\hat{f}_{t, \delta}^*(\mathbf{0}) \\ &-\sigma L d R\left\|\sum_{i=1}^{t-1} \nabla_i-\sum_{i=1}^{t-1} \mathbf{g}_i\right\|. \end{aligned} \tag{14}$$

The first inequality holds because $h_\sigma^*(\cdot)$ is $\sigma d R$-smooth given that $\mathrm{LP}_{\mathcal{K}_\delta}(\cdot)$ is R-Lipschitz as shown in Lemma 10. The third inequality holds because $\nabla h_\sigma^*(\cdot)$ is $\sigma d R$-Lipschitz. The third equality holds because $\hat{\mathbf{x}}_t$ can alternatively be expressed as $\nabla h_\sigma^*\left(-\sum_{i=1}^{t-1} \mathbf{g}_i\right)$, where $\hat{\mathbf{x}}_t = \mathbb{E}\left[\mathbf{x}_t\right]$. The last equality holds because of the Fenchel dual identity $f_t\left(\mathrm{x}_t\right)=\left\langle \nabla f_t\left(\mathrm{x}_t\right), \mathrm{x}_t\right\rangle-f_t^*\left(\nabla f_t\left(\mathrm{x}_t\right)\right)$. Note that by definition $\mathcal{D}(\mathbf{0}, \ldots, \mathbf{0})=-h_\sigma^*(\mathbf{0})-\sum_{t=1}^{T} f_t^*(\mathbf{0})$, which gives the identity $\sum_{t=1}^{T} \Delta_t-\sum_{t=1}^{T} f_t^*(\mathbf{0})=\mathcal{D}\left(\nabla_1, \ldots, \nabla_T\right)+h_\sigma^*(\mathbf{0})$.

Now, it results from (14) that

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) \le \mathcal{D}(\nabla_1, \dots, \nabla_T) + h_\sigma^*(\mathbf{0}) + \frac{\sigma d R}{2} \sum_{t=1}^{T} \|\nabla_t\|^2 + \sum_{t=1}^{T} \sigma L d R \left\| \sum_{i=1}^{t-1} \nabla_i - \sum_{i=1}^{t-1} \mathbf{g}_i \right\| + \sum_{t=1}^{T} \langle \nabla_t, \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle$$
$$\le h_\sigma(\tilde{\mathbf{x}}^*) + \sum_{t=1}^{T} f_t(\tilde{\mathbf{x}}^*) + h_\sigma^*(\mathbf{0}) + \frac{\sigma d R}{2} \sum_{t=1}^{T} \|\nabla_t\|^2 + \sum_{t=1}^{T} \sigma t L d R \|\nabla_t - \mathbf{g}_t\| + \sum_{t=1}^{T} \langle \nabla_t, \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle$$

where the second inequality holds because of the weak duality theorem. On the other hands, by definition $\mathrm{VAL}_{\mathcal{K}_\delta}(\mathbf{0}) = 0$ and the Lipschitzness of $\mathrm{VAL}_{\mathcal{K}_\delta}(\cdot)$, we have $\left|\mathrm{VAL}_{\mathcal{K}_\delta}\left(\frac{1}{\sigma}\mathbf{v}\right)\right| \le R\|\mathbf{v}\|/\sigma \le R/\sigma$ for every $\mathbf{v} \in \mathcal{B}^d$, so $h_\sigma^*(\mathbf{0}) \le R/\sigma$. Moreover, for every $\mathbf{y} \in \mathcal{K}_\delta$ and $\mathbf{v} \sim \mathcal{B}^d$, we have

$$\begin{aligned}
&\langle \tilde{\mathbf{x}}^*, \mathbf{y} \rangle - \max_{\mathbf{x}' \in \mathcal{K}} \left\langle \mathbf{x}', \mathbf{y} + \frac{1}{\sigma} \cdot \mathbf{v} \right\rangle \\
\le &\langle \tilde{\mathbf{x}}^*, \mathbf{y} \rangle - \left\langle \tilde{\mathbf{x}}^*, \mathbf{y} + \frac{1}{\sigma} \cdot \mathbf{v} \right\rangle \\
= &\left\langle \tilde{\mathbf{x}}^*, -\frac{1}{\sigma} \cdot \mathbf{v} \right\rangle \le \|\tilde{\mathbf{x}}^*\| \|\mathbf{v}\|/\sigma \le R/\sigma.
\end{aligned}$$

Thus, $h_\sigma(\tilde{\mathbf{x}}^*) = \sup_{\mathbf{y} \in \mathcal{K}_\delta} \langle \tilde{\mathbf{x}}^*, \mathbf{y} \rangle - h_\delta^*(\mathbf{y}) = \mathbb{E}_{\mathbf{v} \sim \mathbb{B}} \left[\langle \tilde{\mathbf{x}}^*, \mathbf{y} \rangle - \max_{\mathbf{x}' \in \mathcal{K}} \left\langle \mathbf{x}', \mathbf{y} + \frac{1}{\delta} \cdot \mathbf{v} \right\rangle \right] \le R/\sigma$. Then, since $\|\nabla_t\| \le L$ and $\|\nabla_t - \mathrm{g}_t\| \le \frac{\sqrt{d}\ell\delta}{2}$ because of Lemma 4, we have

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathrm{x}_t) \le &\sum_{t=1}^{T} f_t(\tilde{\mathrm{x}}^*) + \frac{2R}{\sigma} + \frac{\sigma d R L^2}{2} T \\
&+ \frac{\delta \sigma d^{3/2} R L \ell}{4} T^2 + \sum_{t=1}^{T} \langle \nabla_t, \mathrm{x}_t - \hat{\mathrm{x}}_t \rangle . \quad (15)
\end{aligned}$$

Finally, by taking the expectation on the both sides and using Lemma 11, we obtain the desired result. □

Now, we derive a high probability guarantee for the regret bound of Algorithm $\mathcal{A}_4$.

Lemma 13. *for every $\xi > 0$, with probability at least $1 - \xi$, Algorithm $\mathcal{A}_4$ guarantees that*

$$\mathcal{R}_T(\mathcal{A}_1) \le \delta L R T / r + \frac{\sigma d R L^2}{2} T + \frac{8 \ell R^2 T}{k} \log(4T/\xi) + \frac{\delta \sigma d^{3/2} R L \ell}{4} T^2 + 2LR\sqrt{2T \log(2/\xi)} + \delta L T + \frac{2R}{\sigma}.$$

Proof. Let $\hat{\mathbf{x}}_t = \mathbb{E}_{\xi_t}[\mathbf{x}_t]$, where $\mathbf{x}_t$ is generated by Algorithm $\mathcal{A}_4$, where $\xi_t = \left\{ \mathrm{v}_t^1, \dots, \mathrm{v}_t^k \right\}$ comprises the randomness used at iteration t. Since the losses are smooth, we have

$$\begin{aligned}
\langle \nabla_t, \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle &= \langle \nabla f_t(\mathbf{x}_t) - \nabla f_t(\hat{\mathbf{x}}_t), \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle \\
&\quad + \langle \nabla f_t(\hat{\mathbf{x}}_t), \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle \\
&\le \ell \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2 + \langle \nabla f_t(\hat{\mathbf{x}}_t), \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle .
\end{aligned}$$

By defining $Z_j = \frac{1}{k}\left(\hat{\mathbf{x}}_t - \mathbf{x}_t^j\right), j = 1, \dots, k$, it holds that $\mathbb{E}_{\mathbf{v}_t^j}\left[Z_j \mid \mathbf{v}_t^1, \dots, \mathrm{v}_t^{j-1} \right] = 0$ given the definition of $\hat{\mathrm{x}}_t$ and i.i.d. uniform samples $\mathrm{v}_t^j \sim \mathcal{B}^d$. In addition, $\left\| \frac{1}{k}\left(\hat{\mathrm{x}}_t - \mathrm{x}_t^j\right) \right\| \le \frac{2R}{k}$. Thus, the sequence $\{Z_j\}_{j=1}^k$ is a bounded martingale difference sequence. Using the Hoeffding-Azuma inequality, we have that

$$\mathbb{P}\left(\|\hat{\mathbf{x}}_t - \mathbf{x}_t\| > \epsilon_1\right) = \mathbb{P}\left(\left\| \sum_{t=1}^{k} Z_t \right\| > \epsilon_1 \right) \le 2\exp\left(\frac{-\epsilon_1^2}{2kB^2} \right)$$

where $B_1 = 2R/k$. Define $\epsilon_1 = B_1\sqrt{2k\log(4T/\xi)}$, leading to $\xi/(2T) = 2\exp\left(-\frac{\epsilon_1^2}{2kB^2}\right)$. Hence, by the union bound, the inequality

$$\sum_{t=1}^{T} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \le \frac{8R^2T}{k} \log(4T/\xi) \quad (16)$$

holds with probability at least $1 - \xi/2$. Define $Q_t = \langle \nabla f_t(\hat{\mathrm{x}}_t), \mathrm{x}_t - \hat{\mathrm{x}}_t \rangle$ for all $t \in \{1, 2, \dots, T\}$. Since $\hat{\mathrm{x}}_t = \mathbb{E}_{\xi_t}[\mathrm{x}_t \mid \xi_1, \dots, \xi_{t-1}]$ and $\nabla f_t(\hat{\mathrm{x}}_t)\ \hat{\mathrm{x}}_t$ is independent of $\{\xi_1, \dots, \xi_{t-1}\}$, it must hold that $\mathbb{E}_{\xi_t}[Q_t \mid \xi_1, \dots, \xi_{t-1}] = 0$. In addition, since $\|Q_t\| \le 2LR$, the sequence $\{Q_t\}_{t=1}^T$ is a bounded martingale difference sequence. Using the Hoeffding-Azuma inequality, we have that

$$\mathbb{P}\left(\sum_{t=1}^{T} \langle \nabla f_t(\hat{\mathrm{x}}_t), \mathrm{x}_t - \hat{\mathrm{x}}_t \rangle > \epsilon_2 \right) = \mathbb{P}\left(\sum_{t=1}^{T} Q_t > \epsilon_2 \right) \le \exp\left(\frac{-\epsilon_2^2}{2TB_2^2} \right)$$

where $B_2 = 2LR$. Define $\epsilon_2 = B_2\sqrt{2T\log(2/\xi)}$, leading to $\xi/2 = \exp\left(\frac{-\epsilon_2^2}{2TB_2^2}\right)$. Hence, the inequality

$$\sum_{t=1}^{T} \langle \nabla f_t(\hat{\mathbf{x}}_t), \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle \le 2LR\sqrt{2T\log(2/\xi)} \quad (17)$$

holds with probability at least $1 - \xi/2$. Combining (16) and (17) implies the satisfaction of the following inequality with probability at least $1 - \xi$:

$$\sum_{t=1}^{T} \langle \nabla_t, \mathbf{x}_t - \hat{\mathbf{x}}_t \rangle \le 2LR\sqrt{2T\log(2/\xi)} + \frac{8\ell R^2 T}{k} \log(4T/\xi).$$

This together with (15) gives rise to the desired high probability regret bound. □

Proof of Theorem 5. The reduction from Theorem 4 to Theorem 3 follows from a simple blocking technique, i.e. grouping several rounds into one as detailed in Algorithm $\mathcal{A}_3$. Let $T = nk$, where n and k are assumed to be integers for simplicity and denote $f_i' = \sum_{t=(i-1)\times k+1}^{i \times k} f_t, \forall i \in \{1, \dots, n\}$. Since f_i' contains k losses from the original problem, the player is allowed k linear optimizations to handle a single loss f_i'. Hence, we use Algorithm $\mathcal{A}_4$ for n iterations with k samples at each iteration to obtain actions $\mathrm{x}_1', \dots, \mathrm{x}_n'$ and play $\mathrm{x}_t = \mathrm{x}_i'$

for all $(i-1)\times k+1 \leq t \leq i\times k$ in the original setting $-$ call this algorithm $\mathcal{A}_4'$. The corresponding constants of the constructed game are $R'=R, L'=L\times k$ and $\ell'=\ell\times k$. Note that Algorithm $\mathcal{A}_4'$ is equivalent to Algorithm $\mathcal{A}_3$. Hence, it results from Lemma 12 that the expected regret bound of Algorithm $\mathcal{A}_3$ for smooth convex functions is given by

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_T\left(\mathcal{A}_3\right)\right] \leq \delta(L\cdot k)n + \delta(L\cdot k)Rn/r + \frac{4(\ell\cdot k)R^2n}{k} + \\
&\sigma dR(L\cdot k)^2\cdot n/2 + \frac{\delta\sigma d^{3/2}R(L\cdot k)(\ell\cdot k)}{4}n^2 + 2R/\sigma \\
&= RL\sqrt{nd}k + \frac{4(\ell\cdot k)R^2n}{k} + \frac{\delta dR\ell k}{2}n^{3/2} \\
&+ \delta LT + \delta LRT/r
\end{aligned}$$

with the choice of $\sigma = 2/L\sqrt{d}\sqrt{n}k$. By choosing $n = T^{2/3}, k = T^{1/3}$ and $\delta = T^{-1}$, we attain the expected regret bound $O\left(T^{2/3}\right)$ for Algorithm 3 with one linear optimization per iteration. □

Proof of Theorem 6. Similar to the proof of Theorem 5, it follows from Lemma 13 that the high probability regret bound for Algorithm $\mathcal{A}_3$ for smooth convex functions can be written as

$$\begin{aligned}
&\mathcal{R}_T\left(\mathcal{A}_3\right) \leq \delta(L\cdot k)n + \delta(L\cdot k)Rn/r + 2R/\sigma \\
&+ \sigma dR(L\cdot k)^2\cdot n/2 + \frac{\delta\sigma d^{3/2}R(L\cdot k)(\ell\cdot k)}{4}n^2 \\
&+ 2(L\cdot k)R\sqrt{2n\log\left(2/\xi\right)} + \frac{8(\ell\cdot k)R^2n}{k}\log\left(4n/\xi\right) \\
&= \delta LT + \delta LRT/r + RL\sqrt{nd}k + \frac{\delta dR\ell k}{2}n^{3/2} + \\
&2LkR\sqrt{2n\log\left(2/\xi\right)} + 8\ell R^2n\log\left(4n/\xi\right)
\end{aligned}$$

with the choice of $\sigma = 2/L\sqrt{d}\sqrt{n}k$. By choosing $n = T^{2/3}, k = T^{1/3}$ and $\delta = T^{-1}$, we attain the high probability regret bound $\tilde{O}\left(T^{2/3}\right)$ for Algorithm $\mathcal{A}_3$ with one linear optimization per iteration. □