Improved Regret Guarantees for Online Smooth Convex Optimization with Bandit Feedback

Ankan Saha
University of Chicago

Ambuj Tewari
University of Texas at Austin

Abstract

The study of online convex optimization in the bandit setting was initiated by Kleinberg (2004) and Flaxman et al. (2005). Such a setting models a decision maker that has to make decisions in the face of adversarially chosen convex loss functions. Moreover, the only information the decision maker receives are the losses. The identities of the loss functions themselves are not revealed. In this setting, we reduce the gap between the best known lower and upper bounds for the class of smooth convex functions, i.e. convex functions with a Lipschitz continuous gradient. Building upon existing work on self-concordant regularizers and one-point gradient estimation, we give the first algorithm whose expected regret is $O(T^{2/3})$, ignoring constant and logarithmic factors.

1 INTRODUCTION

The problem of sequential decision making is of utmost importance in disciplines as varied as Artificial Intelligence, Control Theory, Economics, Operations Research, and Statistics. In this paper, we are concerned with a situation where a decision maker (or learner) has to make decisions in the presence of an adversary. After the learner has chosen its action at the current time step, the adversary responds with a loss function. We assume that the set of actions available to the learner is a convex set and the adversary also chooses convex loss functions. To add to the difficulties of the learner, it is further assumed that the actual loss function chosen by the adversary is not revealed to the learner. Instead, the learner simply observes the value of the loss function at the point it chose from its convex action set.

One would ideally like to design learning algorithms that minimize the cumulative sum, over a total of T time steps, of the losses incurred by the learner. But that is already asking for too much, even in the full information setting where the learner observes the adversary's loss functions. So, we have to lower our unrealistic expectations and search for learning algorithms that minimize *regret*. Regret is the difference between (i) the cumulative loss of the learner, and (ii) the minimum possible loss had the adversary's sequence of loss functions been known in advance and the learner could choose the best fixed decision or action in response to it.

This formalization of sequential decision making is known as *online convex optimization* (or OCO in short) and got off to a start in a remarkably clear and elegant paper of Zinkevich (2003). He considered the full information version and showed that a simple gradient descent strategy for the learner incurs $O(\sqrt{T})$ regret. The constants hidden in the big-Oh notation are known and depend on properties such as the Lipschitz constant of the loss functions and the diameter of the learner's action set. Surprisingly, it was soon shown by Kleinberg (2004) and Flaxman et al. (2005) that one could design algorithms with $o(T)$ regret even in the "bandit" setting, where only evaluations of the loss functions, not the loss functions themselves, are revealed. The algorithm of Flaxman et al. (2005) is particularly striking in its simplicity and elegance as it uses point evaluations of convex functions to approximately estimate the gradient. Finally these estimates are fed to Zinkevich's algorithm and a clever analysis shows that we get a non-trivial regret guarantee even in the bandit setting.

Striking as these first results in bandit OCO were, they all gave rates whose dependence on T was worse than $\sqrt{T}$. This immediately leads to the question of *the price of bandit information*, to use an appealing phrase borrowed from Dani et al. (2007). What does the learner pay, in terms of the regret it incurs, for not

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

having the ability to observe the adversary's loss functions? Does the asymptotic behavior of regret, as T goes to infinity, change? If not, then do the constants in the $O(\sqrt{T})$ full information guarantee change?

If we specialize to the case of online *linear* optimization, i.e. a setting where the adversary plays linear loss functions, then, again with the element of surprise somewhat characteristic of the work in this area, there is no "price of bandit information" to be paid if we only care about the dependence on T. Specifically, a series of papers each building on its predecessors have shown the following two results: (i) it is possible to design an algorithm that has $O(\sqrt{T})$ regret against an adaptive adversary with high probability Dani et al. (2007), and (ii) if the convex set admits an efficiently computable self-concordant barrier then the algorithm of Abernethy et al. (2008) is efficient and achieves $O(\sqrt{T\log(T)})$ expected regret against an oblivious adversary.

1.1 Our Contributions

Our first contribution is conceptual. We point out that, as in first-order convex optimization and full information online convex optimization, the assumptions made on the convex functions played by the adversary are very important. The class of Lipschitz convex functions contains several important subclasses: linear, smooth, strongly convex etc. It is known that the optimal regret can depend on the particular subclass chosen. For example, the full-information algorithm of Hazan et al. (2007) achieves $O(\log(T))$ regret against adversaries that play strongly convex functions. This is much better than the Zinkevich guarantee of $O(\sqrt{T})$. On the other hand, it is known that against an adversary that plays smooth functions, no guarantee better than $O(\sqrt{T})$ can be given. We argue that research in bandit OCO needs to similarly chart out the regret guarantees for the entire territory of important subclasses of convex functions. Some results are already known. For example, Agarwal et al. (2010b) shows how to achieve $O^\star(T^{2/3})$ regret against strongly convex functions. But many questions, including lower bounds, remain open.

Our second contribution is algorithmic. We give an algorithm that achieves $O^\star(T^{2/3})$ regret[1] when the adversary plays smooth functions. This is an improvement on the previous regret guarantees of $O(T^{3/4})$ due to Kleinberg (2004). To do this, we use the self-concordance based algorithm of Abernethy et al. (2008); Abernethy and Rakhlin (2009) coupled with the single-point gradient estimation idea of Flaxman et al. (2005). To the best of our knowledge this is the first $O^\star(T^{2/3})$ algorithm for the class of *non-linear* convex smooth functions.

[1] Our informal $O^\star(\cdot)$ notation hides constant and logarithmic factors in T. We will give precise statements later.

2 PRELIMINARIES

Lower bold case letters (*e.g.*, $\mathbf{w}, \mathbf{x}$ etc.) denote vectors, w_i denotes the i-th component of $\mathbf{w}$, ∂K refers to the boundary of the set K, $\mathbb{S}^d$ refers to the surface of the unit sphere in d dimensions while $\mathbb{B}^d$ refers to the unit ball in d dimensions. The diameter of a closed convex set, K is given by $\mathcal{D} = \max\{\|\mathbf{x}-\mathbf{y}\| : \mathbf{x}, \mathbf{y} \in K\}$. Unless specified otherwise, $\|\cdot\|$ refers to the Euclidean norm $\|\mathbf{w}\| := \left(\sum_i w_i^2\right)^{1/2}$, and $\langle\cdot,\cdot\rangle$ denotes the Euclidean dot product $\langle \mathbf{x}, \mathbf{w}\rangle = \sum_i x_i w_i$.

2.1 Strong Convexity and Smoothness

The following notions of strong convexity and smoothness are extensively used in the sequel. We remark that the concepts of strong convexity and smoothness can be defined w.r.t. an arbitrary norm $\|\cdot\|$. However, for simplicity, we will only work with the standard Euclidean norm in a finite dimensional space $\mathbb{R}^d$.

Definition 1. (Strong Convexity) *Suppose $K \subseteq \mathbb{R}^d$. A convex function $f : K \to \mathbb{R}$ is said to be strongly convex with respect to $\|\cdot\|$ on K if there exists a constant $\rho > 0$ such that $f - \frac{\rho}{2}\|\cdot\|^2$ is convex on K. ρ is called the modulus of strong convexity of f, and for brevity we will call f ρ-strongly convex.*

Definition 2. (Smoothness) *Suppose $K \subseteq \mathbb{R}^d$. Let $f : K \to \mathbb{R}$ be differentiable on K. Then f is said to be smooth (or, alternatively, have Lipschitz continuous gradient (l.c.g)) with respect to $\|\cdot\|$ if there exists a constant $H \geq 0$ such that, for all $\mathbf{w}, \mathbf{w}' \in K$,*

$$\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}')\| \leq H\|\mathbf{w} - \mathbf{w}'\| \ . \tag{1}$$

For brevity, we will call f H-smooth.

We note the standard fact that if f is H-smooth then it satisfies a certain second order upper bound at any point in its domain.

Lemma 1. *If a function f is H-smooth then, for all $\mathbf{w}, \mathbf{w}' \in K$,*

$$f(\mathbf{w}') \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{H}{2}\|\mathbf{w}' - \mathbf{w}\|^2 \ . \tag{2}$$

2.2 Self-Concordant Barriers and Local Norms

The notion of a self-concordant barrier plays a central role in modern convex optimization, especially in the

theory of interior point methods (see, e.g., Nemirovski and Todd (2008)). Let K be a closed convex set. A function $R : K \to \mathbb{R}$ is a self-concordant barrier, if: (i) R tends to infinity near the boundary of K, (ii) both R and $\nabla^2 R$ are Lipschitz continuous w.r.t. the local norm defined by R,

$$\|\mathbf{x}\|_{R,\mathbf{w}} = \sqrt{\langle \mathbf{x}, \nabla^2 R(\mathbf{w})\mathbf{x}\rangle} \tag{3}$$

The formal definition is as follows.

Definition 3. (Self-concordant barrier) [Nemirovski and Todd (2008) Definition 2.1] *Let $K \subseteq \mathbb{R}^d$ be a closed convex set. A function $R : \text{int}(K) \to \mathbb{R}$ is called a ν-self-concordant barrier for K, if*

1. *R is three times continuously differentiable with $R(\mathbf{w}_k) \to \infty$ if $\mathbf{w}_k \to \partial K$, and*
2. *R satisfies, for all $\mathbf{w} \in \text{int}(K), \mathbf{x} \in \mathbb{R}^d$,*

$$\left|\nabla^3 R(\mathbf{w})[\mathbf{x},\mathbf{x},\mathbf{x}]\right| \leq 2 \cdot \|\mathbf{x}\|^3_{R,\mathbf{w}} \ ,$$
$$|\langle \nabla R(\mathbf{w}), \mathbf{x}\rangle| \leq \sqrt{\nu} \cdot \|\mathbf{x}\|_{R,\mathbf{w}} \ .$$

Note that the above definition automatically implies that R and $\nabla^2 R$ are Lipschitz continuous w.r.t the local norm with constants $\sqrt{\nu}$ and 2 respectively (Nemirovski and Todd, 2008).

The following fact will be very useful. If R is a self-concordant barrier for K, then, for any $\mathbf{w} \in \text{int}(K)$, the **Dikin Ellipsoid** centered at $\mathbf{w}$,

$$\{\mathbf{w}' \ : \ \|\mathbf{w}' - \mathbf{w}\|_{R,\mathbf{w}} \leq 1\}$$

is entirely contained in K.

2.3 One-point Gradient Estimates

We will use the following idea of Flaxman et al. (2005). Suppose we have a bounded differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ whose gradient at some point $\mathbf{x}$ needs to be estimated from a single random point evaluation. Then, choose $\mathbf{u}$ uniformly at random from the surface of the unit sphere $\mathbb{S}^d$ and use the estimate

$$\frac{d}{\delta} \cdot f(\mathbf{x} + \delta\mathbf{u}) \cdot \mathbf{u} \ .$$

The usefulness of this is captured by the following lemma proved in Flaxman et al. (2005).

Lemma 2. *Define $\hat{f}(\mathbf{x}) = \mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}[f(\mathbf{x} + \delta\mathbf{v})]$. Then, we have,*

$$\nabla \hat{f}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\in\mathbb{S}^d}\left[\frac{d}{\delta} \cdot f(\mathbf{x} + \delta\mathbf{u}) \cdot \mathbf{u}\right] \ .$$

Note that there is a bias-variance trade-off here. The bias of the estimator vanishes as δ becomes small but then the variance becomes large. On the other hand, the variance vanishes as δ increases but then the bias grows. In our application of this lemma, we will have to carefully balance the two.

2.4 Bandit Online Convex Optimization and Regret

We consider the following repeated game of T rounds played between a player/learner/decision maker and an adversary. The set of arms/actions/decisions is a closed bounded convex set $K \subseteq \mathbb{R}^d$. In this paper, an (oblivious) adversary is simply a sequence $f_1, f_2, \ldots, f_T$ of functions chosen from some function class $\mathcal{F} \subseteq \mathcal{F}_{\text{cvx}}$ where $\mathcal{F}_{\text{cvx}}$ is the class of all differentiable convex functions on K. At round t of the game

- Player plays a (possibly random) $\mathbf{y}_t \in K$,
- The adversary responds with $f_t \in \mathcal{F}$,
- Player gets to see and suffers loss $f_t(\mathbf{y}_t)$.

This is referred to as bandit online convex optimization (bandit OCO in short). A related and much better understood setting is that of full information OCO where the identity of f_t is also revealed at time step t.

We define the **regret** as:

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

Note that this a random variable as the player might be using a randomized algorithm. The expected regret is simply

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{y}_t)\right] - \min_{\mathbf{x}_\star \in K} \sum_{t=1}^{T} f_t(\mathbf{x}_\star)$$

where the expectation is over the randomness in the player algorithm. The second term is deterministic as it is solely a function of the adversary sequence $f_1, \ldots, f_T$. Given an interesting subclass $\mathcal{F} \subseteq \mathcal{F}_{\text{cvx}}$, our goal is to design algorithms for the player such the expected regret is small *no matter what the adversary sequence is* (as long as all the functions are chosen from $\mathcal{F}$).

3 SUBCLASSES OF CONVEX FUNCTIONS

Whether we are interested in first-order (i.e. gradient based) methods for convex optimization, stochastic convex optimization, or in (full information) online

convex optimization, the type of assumptions made on the underlying class $\mathcal{F}$ of convex functions makes a big difference. The rates of convergence for optimization accuracy or for cumulative regret depend on properties, such as the degree of differentiability, measure of curvature, etc., of the convex functions under consideration.

Recall that $\mathcal{F}_{\text{cvx}}$ denotes the class of differentiable convex functions on K. The following four main subclasses of $\mathcal{F}_{\text{cvx}}$ are often isolated for study:

1. Lipschitz:

$$\mathcal{F}_{\text{lip}}(L) = \{f \in \mathcal{F}_{\text{cvx}} \ : \ \forall \mathbf{w} \in K, \ \|\nabla f(\mathbf{w})\| \leq L\}$$

2. Smooth:

$$\mathcal{F}_{\text{smth}}(H) = \{f \in \mathcal{F}_{\text{cvx}} \ : \ f \text{ is } H\text{-smooth}\}$$

3. Lipschitz and Strongly Convex:

$$\mathcal{S}_{\text{lip}}(L, \rho) = \{f \in \mathcal{F}_{\text{lip}}(L) : f \text{ is } \rho\text{-strongly convex}\}$$

4. Smooth and Strongly Convex:

$$\mathcal{S}_{\text{smth}}(H, \rho) = \{f \in \mathcal{F}_{\text{smth}}(H) : f \text{ is } \rho\text{-strongly convex}\}$$

We will omit the various constants in the definitions of the subclasses above if they are obvious from context.

Let us first consider first-order convex optimization, i.e. we are optimizing a single function $f \in \mathcal{F}$ but we can only access the function by asking for its gradient at arbitrary points in the domain. For a total budget of T queries, the best optimization accuracy

$$f(\mathbf{w}_T) - \min_{\mathbf{w} \in K} f(\mathbf{w})$$

will typically be some decreasing function of T. Here $\mathbf{w}_T$ is the iterate of the optimization algorithm after T first-order queries. For $\mathcal{F}_{\text{lip}}$ and $\mathcal{F}_{\text{smth}}$, optimal first-order methods achieve accuracies of $\Theta(1/\sqrt{T})$ and $\Theta(1/T^2)$ respectively. With strong convexity, the rates become much better. For example, a first-order method can achieve an accuracy of $\exp(-\Theta(T))$ after T queries.

For full information online convex optimization, the optimal rate at which the (cumulative) regret scales in T is known to be $\Theta(\sqrt{T})$ for $\mathcal{F}_{\text{lip}}$ and $\Theta(\log(T))$ for $\mathcal{S}_{\text{lip}}$. Morover, adding smoothness assumptions does not help here, in contrast to first-order convex optimization. Thus, the optimal rates remain $\Theta(\sqrt{T})$ and $\Theta(\log(T))$ for $\mathcal{F}_{\text{smth}}$ and $\mathcal{S}_{\text{smth}}$.

We can ask the analogous question for bandit OCO: what are the optimal regret rates for the four classes $\mathcal{F}_{\text{lip}}, \mathcal{F}_{\text{smth}}, \mathcal{S}_{\text{lip}}, \mathcal{S}_{\text{smth}}$ mentioned above?

The algorithms of Flaxman et al. (2005) and Kleinberg (2004) achieve $O(T^{3/4})$ expected regret for the classes $\mathcal{F}_{\text{lip}}$ and $\mathcal{F}_{\text{smth}}$ respectively. So, Kleinberg's algorithm achieves the same $O(T^{3/4})$ regret under stronger smoothness assumptions. Should we expect a better rate under the smoothness assumption? We show below that the answer is "yes". Our Algorithm 1 achieves $O^\star(T^{2/3})$ expected regret against $\mathcal{F}_{\text{smth}}$ thus improving upon Kleinberg's algorithm. Note that the full information lower bound of $\Omega(\sqrt{T})$ for the class $\mathcal{F}_{\text{smth}}$ is trivially a lower bound in the bandit setting too.

We note that the subclass $\mathcal{S}_{\text{lip}}$ has already been considered before in the bandit OCO setting by Agarwal et al. (2010b). They show that an algorithm combining the ideas of Hazan et al. (2007) and Flaxman et al. (2005) achieves $O^\star(T^{2/3})$ expected regret against $\mathcal{S}_{\text{lip}}$. Unfortunately, no lower bound better that the trivial $\Omega(\log(T))$ full information lower bound has so far appeared in the literature.

4 ALGORITHM FOR SMOOTH FUNCTIONS

In this section, we present our algorithm along with its formal regret guarantee. All proofs will be given later.

Like most bandit algorithms, Algorithm 1 works by estimating the "missing information" and then passing on this estimate to a full information algorithm. In this case, the "missing information" at any time step is the gradient of the current loss function at the current point. The underlying full information algorithm is an algorithm from Abernethy et al. (2008); Abernethy and Rakhlin (2009) that we will refer to as the AHR algorithm hereafter.

Algorithm 1: Bandit OCO Algorithm for Smooth Functions

Parameters: $\eta > 0, \delta \in [0, 1]$, R – a ν-self-concordant barrier for K

Pick $\mathbf{x}_1 \in K$
for $t = 1$ to T **do**
  $A_t \leftarrow \sqrt{(\nabla^2 R(\mathbf{x}_t))^{-1}}$
  Draw $\mathbf{u}_t \sim \mathbb{S}^d$ uniformly at random
  $\mathbf{y}_t \leftarrow \mathbf{x}_t + \delta A_t \mathbf{u}_t$
  Play $\mathbf{y}_t$ and receive $f_t(\mathbf{y}_t) \in \mathbb{R}$
  $\mathbf{g}_t \leftarrow \frac{d}{\delta} \cdot f_t(\mathbf{y}_t) \cdot A_t^{-1} \mathbf{u}_t$
  $\mathbf{x}_{t+1} \leftarrow \text{argmin}_{\mathbf{x} \in K} \ \eta \sum_{s=1}^{t} \langle \mathbf{g}_s, \mathbf{x} \rangle + R(\mathbf{x})$
end for

It is easy to verify that we have a bandit algorithm since the only feedback used is the number $f_t(\mathbf{y}_t)$.

Note that $\mathbf{y}_t$, the player's move at time t, lies in the Dikin ellipsoid centered at $\mathbf{x}_t$, the point suggested by the full information algorithm:

$$\|\mathbf{y}_t - \mathbf{x}_t\|^2_{R,\mathbf{x}_t} = \left\langle \delta A_t \mathbf{u}_t, (A_t)^{-2} \delta A_t \mathbf{u}_t \right\rangle = \delta^2 \|\mathbf{u}_t\|_2^2 \leq 1 \, .$$

Above, we used the fact that $A_t^{-2} = \nabla R(\mathbf{x}_t)$. By the Dikin ellipsoid property, $\mathbf{y}_t \in K$ and we have a valid algorithms that plays points in the set K on every round.

The gradient estimate $\mathbf{g}_t$ is then passed on to the full information AHR algorithm. In the analysis, we show that $\mathbf{g}_t$ is neither too big (in norm) nor too bad an estimate of the gradient.

Theorem 3. *Let the set K have diameter $\mathcal{D}$. Suppose we run Algorithm 1 against an arbitrary sequence of functions f_t all drawn from $\mathcal{F}_{\text{smth}}(H)$ and bounded by C. Then, for appropriate choices of the parameters η, δ, the expected regret is bounded as:*

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{y}_t)\right] &- \min_{\mathbf{x}^\star \in K} \sum_{t=1}^T f_t(\mathbf{x}^\star) \\
&\leq 3 \left(H\nu \log T\right)^{1/3} \left(Cd\mathcal{D}\right)^{2/3} T^{2/3} \\
&\quad + \left(\frac{2C}{\mathcal{D}} + \mathcal{D}H\right)\sqrt{T} \\
&= O\left(T^{2/3} (\log(T))^{1/3}\right)
\end{aligned}$$

5 PROOFS

In the proofs, $\mathcal{H}_{<t}$ stands for the history of the algorithm in question up to (but not including) time t. The conditional expectation w.r.t. all the randomness used by the algorithm previous to step t is denoted by $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}_{<t}]$.

5.1 Proofs of Theorem 3

Let $\mathbf{x}_\star$ be an arbitrary point in K. Let $\tilde{\mathbf{x}}$ be the point closest to $\mathbf{x}_\star$ that is also at least $1/\sqrt{T}$ distance away from the boundary ∂K. For such an $\tilde{\mathbf{x}}$, $R(\tilde{\mathbf{x}}) \leq 2\nu \log(T)$ (see Abernethy and Rakhlin (2009)) and since H-smooth functions bounded by C on a set of diameter $\mathcal{D}$ necessarily have a Lipschitz constant bounded by

$$L' = \frac{2C}{\mathcal{D}} + \mathcal{D}H$$

we also have $|f_t(\mathbf{x}_\star) - f_t(\tilde{\mathbf{x}})| \leq L'/\sqrt{T}$. Hence the regret is bounded as:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{y}_t)\right] - \sum_{t=1}^T f_t(\mathbf{x}_\star) &\leq \mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{y}_t)\right] - \sum_{t=1}^T f_t(\tilde{\mathbf{x}}) \\
&\quad + T \cdot \frac{L'}{\sqrt{T}} \, . \qquad (4)
\end{aligned}$$

Thus, we focus on the sum on the RHS above. We can write it as:

$$\begin{aligned}
(A) \qquad & \mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{y}_t) - \hat{f}_t(\mathbf{y}_t)\right] \\
(B) \qquad & + \mathbb{E}\left[\sum_{t=1}^T \hat{f}_t(\mathbf{y}_t) - \hat{f}_t(\mathbf{x}_t)\right] \\
(C) \qquad & - \mathbb{E}\left[\sum_{t=1}^T f_t(\tilde{\mathbf{x}}) - \hat{f}_t(\tilde{\mathbf{x}})\right] \\
(D) \qquad & + \mathbb{E}\left[\sum_{t=1}^T \hat{f}_t(\mathbf{x}_t) - \hat{f}_t(\tilde{\mathbf{x}})\right]
\end{aligned}$$

where we define $\hat{f}_t$ as the following smoothed version of the adversary's function f_t:

$$\hat{f}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \in \mathbb{B}^d}\left[f_t(\mathbf{x} + \delta A_t \mathbf{v})\right] \, .$$

Note that $\hat{f}_t$ is a random function measurable w.r.t. $\mathcal{H}_{<t}$ as it depends on the random matrix A_t. Also note that it is well-defined since $\mathbf{x} + \delta A_t \mathbf{v} \in K$ by the Dikin ellipsoid property. Hence only evaluate f_t on K in the above definition.

We now bound the sums (A) – (D). The first three are easy to bound. The last term (D) is the main term and involves an appeal to the analysis of the underlying full information AHR algorithm.

Bounding (A) This term is simply non-positive since, by Jensen's inequality, we have

$$\begin{aligned}
\hat{f}_t(\mathbf{y}_t) &= \mathbb{E}_{\mathbf{v}_t \in \mathbb{S}^d}\left[f_t(\mathbf{y}_t + \delta A_t \mathbf{v})\right] \\
&\geq f_t(\mathbb{E}_{\mathbf{v}_t \in \mathbb{S}^d}\left[\mathbf{y}_t + \delta A_t \mathbf{v}_t\right]) = f_t(\mathbf{y}_t)
\end{aligned}$$

Bounding (B) We have,

$$\begin{aligned}
(B) &= \sum_{t=1}^T \mathbb{E}\left[\hat{f}_t(\mathbf{x}_t + \delta A_t \mathbf{u}_t) - \hat{f}_t(\mathbf{x}_t)\right] \\
&\leq \sum_{t=1}^T \mathbb{E}\left[\left\langle \nabla \hat{f}_t(\mathbf{x}_t), \delta A_t \mathbf{u}_t \right\rangle + \frac{H}{2}\left\|\delta A_t \mathbf{u}_t\right\|^2\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\delta \left\langle \nabla \hat{f}_t(\mathbf{x}_t), A_t \mathbb{E}_t\left[\mathbf{u}_t\right] \right\rangle\right] + \frac{H\delta^2}{2}\mathbb{E}\left[\left\|A_t \mathbf{u}_t\right\|^2\right] \\
&\leq \frac{HT\delta^2\mathcal{D}^2}{2} \qquad (5)
\end{aligned}$$

where $\mathcal{D}$ is the diameter of the set K. Here the first equality follows from the definition of $\mathbf{y}_t$ from algorithm 1 while the second inequality follows by using (2), since f_t is H-smooth. The second equality is a bit

subtle and follows by observing that the only randomness is due to $\mathbf{u}_t$ as the definition of A_t from algorithm 1 clearly makes it deterministic once the history $\mathcal{H}_{<t}$ is fixed. Expectation of $\mathbf{u}_t$ over the surface of the unit ball gives 0. Finally (5) follows, since the Dikin ellipsoid property ensures that $\mathbf{y}_t = \mathbf{x}_t + A_t\mathbf{u}_t$ lies in the set K, thus bounding $\|A_t\mathbf{u}_t\|$ by the diameter of K.

Bounding (C) We have,

$$(C) \le \frac{HT\delta^2\mathcal{D}^2}{2} \tag{6}$$

because

$$\begin{aligned}
-f_t(\tilde{\mathbf{x}}) &= -\hat{f}_t(\tilde{\mathbf{x}}) + \hat{f}_t(\tilde{\mathbf{x}}) - f_t(\tilde{\mathbf{x}}) \\
&= -\hat{f}_t(\tilde{\mathbf{x}}) + \mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}\left[f_t(\tilde{\mathbf{x}} + \delta A_t\mathbf{v}) - f_t(\tilde{\mathbf{x}})\right] \\
&\le -\hat{f}_t(\tilde{\mathbf{x}}) + \mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}\left[\langle\nabla f_t(\tilde{\mathbf{x}}), \delta A_t\mathbf{v}\rangle + \frac{H}{2}\|\delta A_t\mathbf{v}\|^2\right] \\
&\le -\hat{f}_t(\tilde{\mathbf{x}}) + \frac{H\delta^2\mathcal{D}^2}{2} \ .
\end{aligned}$$

The second equality follows from the definition of $\hat{f}$ while the subsequent inequality follows from (2). The last inequality follows because $\mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}[\mathbf{v}] = \mathbf{0}$ and $\|A_t\mathbf{v}\| \le \mathcal{D}$.

Bounding (D) The analysis will depend on the following guarantee for the AHR full information algorithm. This result is essentially derived in Abernethy and Rakhlin (2009).

Lemma 4. *Let K be a convex set and R be a self-concordant barrier on K. Then, for any random sequence $h_1, \ldots, h_T$ of convex functions where h_t is measurable w.r.t. $\mathcal{H}_{<t}$, if we run the AHR algorithm*

$$\mathbf{x}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{x}\in K} \sum_{s=1}^{t} \eta\langle\mathbf{g}_s, \mathbf{x}\rangle + R(\mathbf{x})$$

with gradient estimates $\mathbf{g}_t$'s such that $\mathbb{E}_t[\mathbf{g}_t] = \nabla h_t(\mathbf{x}_t)$, we have, for any $\tilde{\mathbf{x}} \in K$,

$$\sum_{t=1}^{T}\mathbb{E}\left[h_t(\mathbf{x}_t) - h_t(\tilde{\mathbf{x}})\right] \le \eta\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{g}_t\|_{t,\star}^2\right] + \frac{R(\tilde{\mathbf{x}})}{\eta} \ , \tag{7}$$

where $\|\cdot\|_{t,\star}$ is the norm dual to the local norm $\|\cdot\|_t = \|\cdot\|_{R,\mathbf{x}_t}$, i.e.

$$\|\mathbf{g}_t\|_{t,\star}^2 = \left\langle\mathbf{g}_t, \left(\nabla^2 R(\mathbf{x}_t)\right)^{-1}\mathbf{g}_t\right\rangle \ . \tag{8}$$

Before we can use this lemma with $h_t = \hat{f}_t$, we need to make sure that the $\mathbf{g}_t$'s used by Algorithm 1 are indeed unbiased estimates of the gradients.

Lemma 5. *Let $\hat{f}_t$'s be defined as above. Then, we have,*

$$\mathbb{E}\left[\mathbf{g}_t \mid \mathcal{H}_{<t}\right] = \nabla\hat{f}_t(\mathbf{x}_t)$$

for $\mathbf{g}_t$ as defined in Algorithm 1.

Proof. Condition on $\mathcal{H}_{<t}$. Then $\mathbf{u}_t$ is an independent random variable distributed uniformly on the unit sphere and hence we have,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{g}_t|\mathcal{H}_{<t}\right] &= \mathbb{E}_{\mathbf{u}\in\mathbb{S}^d}\left[\frac{d}{\delta}\cdot f_t(\mathbf{x}_t + \delta A_t\mathbf{u})\cdot A_t^{-1}\mathbf{u}\right] \\
&= A_t^{-1}\mathbb{E}_{\mathbf{u}\in\mathbb{S}^d}\left[\frac{d}{\delta}\cdot f_t(\mathbf{x}_t + \delta A_t\mathbf{u})\cdot \mathbf{u}\right] \\
&= A_t^{-1}\mathbb{E}_{\mathbf{u}\in\mathbb{S}^d}\left[\frac{d}{\delta}\cdot F_t(A_t^{-1}\mathbf{x}_t + \delta\mathbf{u})\cdot \mathbf{u}\right] \\
&= A_t^{-1}\nabla\hat{F}_t(A_t^{-1}\mathbf{x}_t) \\
&= A_t^{-1}A_t\nabla\hat{f}_t(\mathbf{x}_t) = \nabla\hat{f}(\mathbf{x}_t) \ ,
\end{aligned}$$

where the third equality holds simply by defining $F_t(\mathbf{x}) = f_t(A_t\mathbf{x})$. The fourth equality holds by Lemma 2 where

$$\hat{F}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}\left[F_t(\mathbf{x} + \delta\mathbf{v})\right] = \mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}\left[f_t(A_t\mathbf{x} + \delta A_t\mathbf{v})\right] \ .$$

The fifth equality holds because differentiating the above equality gives

$$\nabla\hat{F}_t(\mathbf{x}) = A_t\mathbb{E}_{\mathbf{v}\in\mathbb{B}^d}\left[\nabla f_t(A_t\mathbf{x} + \delta A_t\mathbf{v}\right] = A_t\nabla\hat{f}_t(A_t\mathbf{x}) \ .$$

□

To use the bound in (7), we need to bound the norm in (8). The next lemma does this.

Lemma 6. *Given $\|\mathbf{g}_t\|_t$, as defined in (8), we have*

$$\|\mathbf{g}_t\|_t^2 \le \left(\frac{Cd}{\delta}\right)^2 \tag{9}$$

Proof. Using the definition of $\mathbf{g}_t$ from algorithm 1 we have

$$\begin{aligned}
\|\mathbf{g}_t\|^2 &= \frac{d^2}{\delta^2}(f_t(\mathbf{y}_t)^2\left\|A_t^{-1}\mathbf{u}_t\right\|_t^2 \\
&= \frac{d^2}{\delta^2}(f_t(\mathbf{y}_t)^2\left\langle A_t^{-1}\mathbf{u}_t, \left(\nabla^2 R(\mathbf{x}_t)\right)^{-1}A_t^{-1}\mathbf{u}_t\right\rangle \\
&= \frac{d^2}{\delta^2}(f_t(\mathbf{y}_t)^2\mathbf{u}_t^\top A_t^{-1}A_t^2A_t^{-1}\mathbf{u}_t \\
&= \frac{d^2}{\delta^2}(f_t(\mathbf{y}_t)^2 \qquad (\because \|\mathbf{u}_t\| = 1) \\
&\le \left(\frac{Cd}{\delta}\right)^2
\end{aligned}$$

□

Now using (7) with $h_t = \hat{f}_t$, we get

$$\sum_{t=1}^{T} \mathbb{E}\left[\hat{f}_t(\mathbf{x}_t) - \hat{f}_t(\tilde{\mathbf{x}})\right] \leq \eta \sum_{t=1}^{T} \left(\frac{Cd}{\delta}\right)^2 + \frac{R(\tilde{\mathbf{x}})}{\eta}$$
$$= \eta T \left(\frac{Cd}{\delta}\right)^2 + \frac{2\nu \log(T)}{\eta}$$

where we used the fact that $R(\tilde{\mathbf{x}}) \leq 2\nu \log(T)$. Minimizing over η, we have that

$$\sum_{t=1}^{T} \mathbb{E}\left[\hat{f}_t(\mathbf{x}_t) - \hat{f}_t(\tilde{\mathbf{x}})\right] \leq \frac{2Cd}{\delta}\sqrt{2\nu T \log(T)} \qquad (10)$$

Putting it together Plugging in (5), (6), and (10) into (4), we get, for an arbitrary $\mathbf{x}_\star \in K$,

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{y}_t)\right] - \sum_{t=1}^{T} f_t(\mathbf{x}_\star)$$
$$\leq HT\delta^2\mathcal{D}^2 + \frac{2Cd}{\delta}\sqrt{2\nu T \log(T)} + L'\sqrt{T}\,. \qquad (11)$$

Optimizing now over δ gives Theorem 3.

Note that for linear functions, $H = 0$. Plugging that in (11) and setting $\delta = 1$ helps us retrieve the previous rates of $O(\sqrt{T \log T})$ by Abernethy et al. (2008).

6 DISCUSSION

In this paper we reduced the gap between lower and upper bounds for bandit OCO against smooth convex functions. The trivial full information lower bound is $\Omega(\sqrt{T})$ and the best known upper bound was previously $O(T^{3/4})$. We improve the upper bound to $O^\star(T^{2/3})$ for the class $\mathcal{F}_{\text{smth}}(H)$ when the convex functions played by the adversary are smooth.

One of the main open questions in bandit OCO is reducing the gap between the $\Omega(\sqrt{T})$ and $O(T^{3/4})$ lower and upper bounds on regret for the most general class of convex Lipschitz functions, $\mathcal{F}_{\text{lip}}(L)$. A recent result by Agarwal et al. (2010a) is that we can obtain regret guarantees of $O(\sqrt{T})$ if multi-point feedback is available, i.e. the player can ask for the function value at multiple points. This shows that if function evaluations at multiple points are available, the regret guarantee has similar dependence on T as in the full information setting. Perhaps we can hope that there is again no "price of bandit information" to be paid in the bandit OCO setting against convex Lipschitz functions.

References

Jacob Abernethy and Alexander Rakhlin. Beating the adaptive bandit with high probability. In *COLT*, 2009.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pages 263–274, 2008.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, 2010a.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback, 2010b. longer version available at http://www.cs.berkeley.edu/~alekh/bandit-colt.pdf.

Varsha Dani, Thomas P. Hayes, and Sham Kakade. The price of bandit information for online optimization. In *NIPS*, 2007.

Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, pages 385–394, 2005.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS*, 2004.

Arkadi S. Nemirovski and Michael J. Todd. Interior-point methods for optimization. *Acta Numerica*, 17: 191–234, 2008.

M. Zinkevich. Online convex programming and generalised infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.