Towards Minimax Policies for Online Linear Optimization with Bandit Feedback

Sébastien Bubeck   SBUBECK@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ, USA

Nicolò Cesa-Bianchi   NICOLO.CESA-BIANCHI@UNIMI.IT
Department of Computer Science
Università degli Studi di Milano,
Milano, Italy

Sham M. Kakade   SKAKADE@MICROSOFT.COM
Microsoft Research New England
Cambridge, MA, USA

Abstract

We address the online linear optimization problem with bandit feedback. Our contribution is twofold. First, we provide an algorithm (based on exponential weights) with a regret of order $\sqrt{dn \log N}$ for any finite action set with N actions, under the assumption that the instantaneous loss is bounded by 1. This shaves off an extraneous $\sqrt{d}$ factor compared to previous works, and gives a regret bound of order $d\sqrt{n \log n}$ for any compact set of actions. Without further assumptions on the action set, this last bound is minimax optimal up to a logarithmic factor. Interestingly, our result also shows that the minimax regret for bandit linear optimization with expert advice in d dimension is the same as for the basic d-armed bandit with expert advice. Our second contribution is to show how to use the Mirror Descent algorithm to obtain computationally efficient strategies with minimax optimal regret bounds in specific examples. More precisely we study two canonical action sets: the hypercube and the Euclidean ball. In the former case, we obtain the first computationally efficient algorithm with a $d\sqrt{n}$ regret, thus improving by a factor $\sqrt{d \log n}$ over the best known result for a computationally efficient algorithm. In the latter case, our approach gives the first algorithm with a $\sqrt{dn \log n}$ regret, again shaving off an extraneous $\sqrt{d}$ compared to previous works.

Keywords: online linear optimization; multi-armed bandits; linear bandits with expert advice; minimax regret; exponential weights; mirror descent.

1. Introduction

In this paper we consider the framework of online linear optimization: at each time instance $t = 1, \ldots, n$, the player chooses, possibly in a randomized way, an action from a given compact action set $\mathcal{A} \subset \mathbb{R}^d$. The action chosen by the player at time t is denoted by $a_t \in \mathcal{A}$. Simultaneously to the player, the adversary chooses a loss vector $z_t \in \mathcal{Z} \subset \mathbb{R}^d$ and the loss incurred by the player is $a_t^\top z_t$. The goal of the player is to minimize the expected cumulative loss $\mathbb{E} \sum_{t=1}^n a_t^\top z_t$ where the expectation is taken with respect to the player's internal randomization (and possibly the adversary's

randomization). In the basic version of this problem, the player observes the adversary's move z_t at the end of round t. We consider here the *bandit* version, where the player only observes the incurred loss $a_t^\top z_t$. As a measure of performance we define the regret of the player as

$$R_n = \mathbb{E} \sum_{t=1}^n a_t^\top z_t - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^n a^\top z_t \ .$$

In this paper we are interested in the dual setting, where the adversary plays on a dual action set, i.e., $\mathcal{A}$ and $\mathcal{Z}$ are such that $|a^\top z| \leq 1, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$.

1.1. Contributions and relation to previous works

In the full information case, the online optimization setting (for convex losses) was introduced by Zinkevich (2003). The specific online linear optimization problem with bandit feedback was first studied by McMahan and Blum (2004) and Awerbuch and Kleinberg (2004). Our first contribution to this problem is to complete the research program started by Dani et al. (2008) and Cesa-Bianchi and Lugosi (2011). In these papers the authors studied the EXP2 (Expanded Exp) algorithm, also called Geometric Hedge, Expanded Hedge, or ComBand. This strategy applies to a finite set of actions; it assigns an exponential weight to each action, and then draws an action at random from the corresponding probability distribution. Using a basic estimation procedure (first used by Auer et al. (2002) for the basic multi-armed bandit problem), one can estimate the loss vector z_t. However, to control the range of the estimates, one has to mix the probability given by EXP2 with an "exploration distribution". Dani et al. (2008) chose this distribution to be uniform over a barycentric spanner for the action set, while in (Cesa-Bianchi and Lugosi, 2011) the distribution was uniform over all actions. Using ideas from convex geometry, we propose a new distribution that allows us to derive a minimax optimal regret bound. More precisely, we show that for any finite action set, EXP2 with the exploration distribution given by John's Theorem (see Theorem 3) attains a regret of order $\sqrt{dn \log N}$ for any set of N actions. This improves by a factor $\sqrt{d}$ over previous works. Moreover this rate is optimal: there exist action sets (such as the hypercube) where the minimax rate is of order $d\sqrt{n}$ —see (Dani et al., 2008). Surprisingly, this result also shows that EXP2 with John's exploration can be used for linear bandits with N experts to obtain a regret of order $\sqrt{dn \log N}$, which is no worse than the minimax regret for the basic d-armed bandit with N experts problem.

While these results show that, without further assumption on the set of action, the regret of EXP2 is optimal, they do not say anything about optimality for a *specific* set of actions. In fact, it was proven by Audibert et al. (2011) that for some pair $(\mathcal{A}, \mathcal{Z})$ the exponential weights is a provably suboptimal strategy (with a gap of order $\sqrt{d}$). To address this issue, another class of algorithms has been studied for online optimization: the Mirror Descent style algorithms of Nemirovski and Yudin (1983) —this class of algorithms was rediscovered in the learning community, see for example Kivinen and Warmuth (2001). In recent years the number of papers using Mirror Descent to solve problems in online optimization has been growing very rapidly. In the full information setting (when one observes z_t), we have a very good understanding of how to use Mirror Descent to obtain optimal regret bounds that adapt to the geometry of the problem —see (Rakhlin, 2009; Hazan, 2011; Bubeck, 2011). In particular, a recent paper suggests that in this basic setting Mirror Descent is "universal", see (Srebro et al., 2011). On the other hand, in the limited feedback scenario the picture is much more scattered. In the particular cases of *semi-bandit* feedback —see (Audibert et al., 2011)— and *two-points bandit* feedback —see (Agarwal et al., 2010), we know how to use

Mirror Descent to obtain optimal regret bounds. However, in both scenarios the feedback is much stronger than in the more fundamental bandit problem. In this latter case, there is only one paper that successfully applies Mirror Descent, namely the seminal work of Abernethy et al. (2008) —see also the follow-up paper Abernethy and Rakhlin (2009). Unfortunately, for a convex and compact set $\mathcal{A}$, this approach (which combines Mirror Descent with a self-concordant barrier for the action set) leads to a regret bound of order $d\sqrt{\theta n \log n}$ for any $\theta > 0$ such that $\mathcal{A}$ admits a θ-self concordant barrier. For example, in the case of the hypercube the best we know is $\theta = O(d)$, which results in the suboptimal $d^{3/2}\sqrt{n \log n}$ regret (compared to $d\sqrt{n}$ for EXP2 with John's ellipsoid). However, note that in this particular case it is not known if EXP2 can be implemented efficiently, while Mirror Descent is polynomial time.

Our second main contribution is to propose an efficient algorithm based on Mirror Descent, with an optimal regret bound for two canonical pairs $(\mathcal{A}, \mathcal{Z})$. Namely, the (hypercube, cross-polytope) pair, which corresponds to an L_∞/L_1 type of constraints, and the (Euclidean ball, Euclidean ball) pair, which corresponds to an L_2/L_2 constraint. In the former case this results in the first computationally efficient algorithm with a regret of order $d\sqrt{n}$, while in the latter case it is the first efficient algorithm with a regret of order $\sqrt{dn \log n}$. Indeed, the approach of Abernethy et al. (2008) only gives $d\sqrt{n \log n}$ for the pair (Euclidean ball, Euclidean ball) since there exists a $O(1)$-self concordant barrier for the Euclidean ball. Note also that this specific example was studied in Abernethy and Rakhlin (2009), we discuss their result in Section 5.

1.2. Outline of the paper

The paper is organized as follows. In Section 2 we introduce the two algorithms discussed in the paper: Expanded Exp (EXP2) and Online Stochastic Mirror Descent (OSMD). In both cases we state a general regret bound. In Section 3 we detail our exploration strategy for EXP2, and show the corresponding regret bound. We also discuss briefly the extension to linear bandits with expert advice. Then in Section 4 (respectively Section 5) we show how to use OSMD to obtain a computationally efficient strategy with optimal regret for the hypercube (respectively for the Euclidean ball, up to a logarithmic factor).

2. Algorithms

We briefly describe here the two algorithmic templates that we shall use in this paper. First, EXP2 is described in Figure 1. The general regret bound for this algorithm is the following. The proof of this result follows a standard argument, see for example [Chapter 7, Bubeck (2011)].

Theorem 1 *Let $\mathcal{A}$ be a finite set of N actions. For the EXP2 strategy, provided that $\eta|a^\top \widetilde{z}_t| \leq 1, \forall a \in \mathcal{A}$, one has*

$$R_n \leq 2\gamma n + \frac{\log N}{\eta} + \eta \, \mathbb{E} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} p_t(a)\big(a^\top \widetilde{z}_t\big)^2 \ .$$

Figure 2 describes OSMD in the bandit setting. Note that step (c) can be written in several equivalent ways, such as a Follow The Regularized Leader equation, or a mirror gradient descent step if F is a Legendre function. When written as a gradient descent step, one usually has to project back on $\mathcal{A}$ (using the Bregman divergence associated to F). Here the projection is implicit in the evaluation of ∇F^*. The following theorem states a general regret bound for OSMD. Recall that the Bregman

---

Algorithm: EXP2 *with exploration μ.*
Parameters: learning rate η; mixing coefficient γ; distribution μ over the action set $\mathcal{A}$.

Let $q_1 = \left(\frac{1}{|\mathcal{A}|}, \ldots, \frac{1}{|\mathcal{A}|}\right) \in \mathbb{R}^{|\mathcal{A}|}$. For each round $t = 1, 2, \ldots, n$;

(a) Let $p_t = (1 - \gamma)q_t + \gamma\mu$, and play $a_t \sim p_t$.

(b) Estimate the loss vector z_t by $\widetilde{z}_t = P_t^+ a_t a_t^\top z_t$, with $P_t = \mathbb{E}_{a \sim p_t}[aa^\top]$.

(c) Update the exponential weights, for all $a \in \mathcal{A}$,

$$q_{t+1}(a) = \frac{\exp(-\eta a^\top \widetilde{z}_t)q_t(a)}{\sum_{b \in \mathcal{A}} \exp(-\eta b^\top \widetilde{z}_t)q_t(b)}.$$

Figure 1: EXP2 strategy for bandit feedback.

Algorithm: OSMD.
Parameters: learning rate $\eta > 0$; regularization function $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with effective domain $\mathcal{A}$, and such that the Legendre-Fenchel dual F^* is differentiable on $\mathbb{R}^d$; perturbation scheme for step (a) below.

Let $a_1 \in \operatorname{argmin}_{a \in \mathcal{A}} F(a)$. For each round $t = 1, 2, \ldots, n$;

(a) Play $\widetilde{a}_t$ at random from some probability distribution p_t over $\mathcal{A}$
($\widetilde{a}_t$ is a randomly perturbated version of a_t, see Section 4 and Section 5 for examples).

(b) Estimate the loss vector z_t by $\widetilde{z}_t = P_t^+ \widetilde{a}_t \widetilde{a}_t^\top z_t$, with $P_t = \mathbb{E}_{a \sim p_t}[aa^\top]$.

(c) Let $a_{t+1} = \nabla F^*\left(-\eta \sum_{s=1}^{t} \widetilde{z}_s\right)$.

Figure 2: Online Stochastic Mirror Descent (OSMD) for bandit feedback.

divergence with respect to F is defined as $D_F(x, y) = F(x) - F(y) - (x - y)^\top \nabla F(y)$, and the Legendre-Fenchel dual of F is defined as $F^*(v) = \sup_{x \in \mathcal{A}} x^\top v - F(x)$. In the following, we write x_1^t to denote $x_1 + \cdots + x_t$.

Theorem 2 *Let $\mathcal{A}$ be a compact set of actions, and F a function with effective domain $\mathcal{A}$, and such that F^* is differentiable on $\mathbb{R}^d$. Then OSMD satisfies (for any norm $\|\cdot\|$)*

$$R_n \leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n} \mathbb{E}D_{F^*}\left(-\eta\widetilde{z}_1^t, -\eta\widetilde{z}_1^{t-1}\right) + \sum_{t=1}^{n} \mathbb{E}\big\|a_t - \mathbb{E}[\widetilde{a}_t \mid a_t]\big\| \cdot \|z_t\|_* .$$

Proof The proof is adapted from Kakade et al. (2010). Using the elementary Young's inequality $x^\top v \le F^*(v) + F(x)$, one obtains $\forall a \in \mathcal{A}$

$$-\eta \sum_{t=1}^{n} a^\top \widetilde{z}_t \le F(a) + F^* \left(-\eta \widetilde{z}_1^n\right)$$

$$= F(a) + F^*(0) + \sum_{t=1}^{n} \left(F^* \left(-\eta \widetilde{z}_1^t\right) - F^* \left(-\eta \widetilde{z}_1^{t-1}\right)\right)$$

$$= F(a) + F^*(0) + \sum_{t=1}^{n} \left(\nabla F^* \left(-\eta \widetilde{z}_1^{t-1}\right)^\top \left(-\eta \widetilde{z}_t\right) + D_{F^*}\left(-\eta \widetilde{z}_1^t, -\eta \widetilde{z}_1^{t-1}\right)\right)$$

$$= F(a) + F^*(0) + \sum_{t=1}^{n} \left(-\eta a_t^\top \widetilde{z}_t + D_{F^*}\left(-\eta \widetilde{z}_1^t, -\eta \widetilde{z}_1^{t-1}\right)\right)$$

since $F^*(0) = -F(a_1)$. This shows that:

$$\sum_{t=1}^{n}(a_t - a)^\top \widetilde{z}_t \le \frac{F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n} D_{F^*}\left(-\eta \widetilde{z}_1^t, -\eta \widetilde{z}_1^{t-1}\right).$$

Taking into account the randomness induced by $\widetilde{a}_t$ and $\widetilde{z}_t$ is then an easy exercise, see for example (Bubeck, 2011, Chapter 7). ∎

This theorem proves to be particularly useful when applied with a Legendre function F —see (Cesa-Bianchi and Lugosi, 2006, Chapter 11) for the definition of a Legendre function. Indeed, in that case F^* is differentiable if F is differentiable, and moreover the corresponding gradient mappings are inverse of each other, which gives a simple way to do computations with the Bregman divergence D_{F^*}.

3. EXP2 **with John's exploration**

We propose here a new exploration distribution μ for the EXP2 strategy, that allows us to derive the first $\sqrt{dn \log N}$ regret bound for online linear optimization with bandit feedback. We use the following result from convex geometry, see (Ball, 1997) for a proof.

Theorem 3 *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set. If the ellipsoid $\mathcal{E}$ of minimal volume enclosing $\mathcal{K}$ is the unit ball in some norm derived from a scalar product $\langle \cdot, \cdot \rangle$, then there exists $M \le d(d+1)/2 + 1$ contact points $u_1, \ldots, u_M$ between $\mathcal{E}$ and $\mathcal{K}$, and $\mu \in \Delta_M$ (the simplex of dimension $M - 1$), such that*

$$x = d \sum_{i=1}^{M} \mu_i \langle x, u_i \rangle u_i, \forall x \in \mathbb{R}^d.$$

To use this theorem, we need to perform a preprocessing of the action set as follows:

1. First, we assume that $\mathcal{A}$ is of full rank (that is such that linear combinations of $\mathcal{A}$ span $\mathbb{R}^d$). If it is not the case, then one can rewrite the elements of $\mathcal{A}$ in some lower dimensional vector space and work there.

2. Find John's ellipsoid for $Conv(\mathcal{A})$ —i.e., the ellipsoid of minimal volume enclosing $Conv(\mathcal{A})$: $\mathcal{E} = \{x \in \mathbb{R}^d : (x - x_0)^\top H^{-1}(x - x_0) \leq 1\}$. The first preprocessing step is to translate everything by x_0. In other words, we assume now that $\mathcal{A}$ is such that $x_0 = 0$. Furthermore, we define the inner product $\langle x, y \rangle = x^\top H y$.

3. We can now assume that we are playing on $\mathcal{A}' = H^{-1}\mathcal{A}$, and the loss of playing $a' \in \mathcal{A}'$ when the adversary plays z is $\langle a', z \rangle$. Indeed: $\langle H^{-1}a, z \rangle = a^\top z$. Moreover, note that John's ellipsoid for $Conv(\mathcal{A}')$ is the unit ball for the inner product $\langle \cdot, \cdot \rangle$ because $\langle H^{-1}x, H^{-1}x \rangle = x^\top H^{-1}x$.

4. Find the contact points $u_1, \ldots, u_M$ and $\mu \in \Delta_M$ that satisfy Theorem 3 for $Conv(\mathcal{A}')$. Note that the contact points are in $\mathcal{A}'$, thus they are valid points to play. We say that μ is John's exploration distribution.

In the following we drop the prime on $\mathcal{A}'$. More precisely, we play on a set $\mathcal{A}$ such that John's ellipsoid for $Conv(\mathcal{A})$ is the unit ball for some inner product $\langle \cdot, \cdot \rangle$, and the loss is given by $\langle a, z \rangle$. Thus, we also need to slightly change the algorithm by replacing the dot products $a^\top \widetilde{z}_t$ and $b^\top \widetilde{z}_t$ in step (c) of Figure 1 by, respectively, $\langle a, \widetilde{z}_t \rangle$ and $\langle b, \widetilde{z}_t \rangle$. We also need to modify the loss estimate in step (b) as follows. Recall that the outer product $u \otimes u$ is defined as the linear mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $u \otimes u(x) = \langle u, x \rangle u$. Note that one can also view $u \otimes u$ as a $d \times d$ matrix, so that the evaluation of $u \otimes u$ is equivalent to a multiplication by the corresponding matrix. Now let:

$$P_t = \sum_{a \in \mathcal{A}} p_t(a) a \otimes a.$$

Note that this matrix is invertible, since $\mathcal{A}$ is of full rank and $p_t(a) > 0, \forall a \in \mathcal{A}$. The estimate for z_t is given by:

$$\widetilde{z}_t = P_t^{-1}(a_t \otimes a_t) z_t. \tag{1}$$

Note that this is a valid estimate since $(a_t \otimes a_t) z_t = \langle a_t, z_t \rangle a_t$ and P_t^{-1} are observed quantities. Moreover, it is also clearly an unbiased estimate. We can now prove the following result.

Theorem 4 EXP2 *with John's exploration and estimate* (1) *satisfies, for* $\frac{\eta d}{\gamma} \leq 1$,

$$R_n \leq 2\gamma n + \frac{\log N}{\eta} + \eta n d.$$

In particular with $\gamma = \eta d$ *and* $\eta = \sqrt{\frac{\log N}{3nd}}$ *we have that*

$$R_n \leq 2\sqrt{3nd \log N}.$$

Proof With the chosen scalar product, it is easy to see that the condition $\eta|a^\top \widetilde{z}_t| \leq 1$ in Theorem 1 rewrites as $\eta|\langle a, \widetilde{z}_t \rangle| \leq 1$, while the third term in the regret bound rewrites as $\mathbb{E} \sum_{a \in \mathcal{A}} p_t(a) \langle a, \widetilde{z}_t \rangle^2$. Thus it remains to control those two quantities. Let us start with the latter:

$$\sum_{a \in \mathcal{A}} p_t(a) \langle a, \widetilde{z}_t \rangle^2 = \sum_{a \in \mathcal{A}} p_t(a) \langle \widetilde{z}_t, (a \otimes a) \widetilde{z}_t \rangle$$
$$= \langle \widetilde{z}_t, P_t \widetilde{z}_t \rangle = \langle a_t, z_t \rangle^2 \langle P_t^{-1} a_t, P_t P_t^{-1} a_t \rangle \leq \langle P_t^{-1} a_t, a_t \rangle.$$

Now we use a spectral decomposition of P_t in an orthonormal basis for $\langle \cdot, \cdot \rangle$ and write $P_t = \sum_{i=1}^{d} \lambda_i v_i \otimes v_i$. In particular, we have $P_t^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda_i} v_i \otimes v_i$ and thus:

$$\mathbb{E}\langle P_t^{-1} a_t, a_t \rangle = \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathbb{E}\langle (v_i \otimes v_i) a_t, a_t \rangle = \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathbb{E}\langle (a_t \otimes a_t) v_i, v_i \rangle = \sum_{i=1}^{d} \frac{1}{\lambda_i} \langle P_t v_i, v_i \rangle = d.$$

This concludes the bound for $\mathbb{E} \sum_{a \in \mathcal{A}} p_t(a) \langle a, \widetilde{z}_t \rangle^2$. We turn now to $\langle a, \widetilde{z}_t \rangle$:

$$\langle a, \widetilde{z}_t \rangle = \langle a_t, z_t \rangle \langle a, P_t^{-1} a_t \rangle \le \langle a, P_t^{-1} a_t \rangle \le \frac{1}{\min_{1 \le i \le d} \lambda_i}$$

where the last inequality follows from the fact that $\langle a, a \rangle \le 1$ for any $a \in \mathcal{A}$, since $\mathcal{A}$ is included in the unit ball. Now to conclude the proof we need to lower bound the smallest eigenvalue of P_t. Using Theorem 3, one can see that $P_t \succeq \frac{\gamma}{d} I_d$, and thus $\lambda_i \ge \frac{\gamma}{d}$ concluding the proof. ∎

Using the discretization argument of Dani et al. (2008), EXP2 with John's exploration can be used to obtain a regret of order $d\sqrt{n \log n}$ for any compact set of action $\mathcal{A}$.

3.1. Computational issues

If $\mathcal{A}$ is given by a finite set of points, then Grötschel et al. (1993) give a polynomial time algorithm for computing a constant factor approximation to the John's ellipsoid (and this approximate basis will provide the same order of regret). However, if $\mathcal{A}$ is specified by the intersection of half spaces, then Nemirovski (2007) shows that obtaining such a constant factor approximation to this ellipsoid is NP-hard in general. Here, it is possible to efficiently compute an ellipsoid where the factor of d in Theorem 3 is replaced by $d^{3/2}$ —see (Grötschel et al., 1993), which leads to a slightly worse dependence on d in the regret bound.

In special cases, we conjecture that the John's ellipsoid may be computed efficiently, as for certain problems, there are efficient implementations of GeometricHedge that lead to optimal rates (such as shortest path problems and other settings where dynamic programming solutions exist).

3.2. Application to bandits with experts

Consider the following model of linear bandits with N experts. At each time step $t = 1, 2, \dots, n$, each expert $k = 1, \dots, N$ suggests an action $a_t(k) \in \mathbb{R}^d$. The goal here is to compete with the best expert, that is at each time step the strategy chooses an expert $k_t \in \{1, \dots, N\}$ and the regret is given by:

$$R_n^{\text{exp}} = \mathbb{E} \sum_{t=1}^{n} a_t(k_t)^\top z_t - \min_{k \in \{1, \dots, N\}} \mathbb{E} \sum_{t=1}^{n} a_t(k)^\top z_t.$$

One can use EXP2 with John's exploration to obtain a regret of order $\sqrt{dn \log N}$ for this problem. Indeed, it suffices at every turn to do the preprocessing step on $\mathcal{A}_t = \{a_t(1), \dots, a_t(N)\}$ and to build the corresponding John's exploration μ_t, the straightforward details are omitted.

For example, at each time t each expert $i = 1, \dots, N$ is associated with a hidden loss estimate $z_t(i) \in \mathcal{Z}$ and an arbitrary "context set" $\mathcal{A}_t \subseteq \mathcal{A}$ is observed. Each expert i then suggests the best action according to the current loss estimate, $a_t(i) = \text{argmin}_{a \in \mathcal{A}_t} z_t(i)^\top a$. This can be viewed as a natural nonstochastic variant of the contextual linear bandit model of Chu et al. (2011). Another

notable special case is the d-armed bandit problem with expert advice, where we can view the suggested actions as the corners of the d-dimensional simplex. Here, the EXP4 algorithm of Auer et al. (2002) achieves a regret of order $\sqrt{dn \ln N}$. Interestingly, the regret achievable in the more general d-dimensional linear optimization setting is no worse than in the seemingly simpler d-armed bandit with expert advice setting.

4. Computationally efficient strategy for the hypercube

In this section we restrict our attention to the action set $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$. Using EXP2 with John's exploration on $\{-1, 1\}^d$ one obtains a regret bound of order $d\sqrt{n}$ for this problem, and as it was shown by Dani et al. (2008) this regret is minimax optimal. However, it is not known if it is possible to sample from the exponential weights distribution in polynomial time for this particular set of actions. In this section we propose to turn to OSMD, and we show that with the appropriate regularizer F and random perturbation $\widetilde{a}_t$ (see step (a) in Figure 2), one can obtain a minimax optimal algorithm with computational complexity linear in d. More precisely we use an entropic regularizer

$$F(x) = \frac{1}{2} \sum_{i=1}^{d} \big((1 + x_i)\log(1 + x_i) + (1 - x_i)\log(1 - x_i)\big) \tag{2}$$

together with the following perturbation of a point a_t in the interior of $\mathcal{A}$:

> With probability γ, play $\widetilde{a}_t$ uniformly at random from the canonical basis (with random sign). With probability $1 - \gamma$, play $\widetilde{a}_t = \xi_t$ where $\xi_t(i)$ is drawn independently from a Rademacher with parameter $\frac{1+a_t(i)}{2}$.

It is easy to check that this perturbation is almost unbiased, indeed one has:

$$\mathbb{E}\,\widetilde{a}_t(i) = (1 - \gamma)\left(\frac{1 + a_t(i)}{2} - \frac{1 - a_t(i)}{2}\right) = (1 - \gamma)a_t(i),$$

and thus:

$$\big\|\mathbb{E}[\widetilde{a}_t \mid a_t] - a_t\big\|_\infty \leq \gamma. \tag{3}$$

We can now prove the following result.

Theorem 5 *Consider the online linear optimization problem with bandit feedback on $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$, and with $\mathcal{Z} = \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$. Then OSMD on $\mathcal{A}$ with regularizer (2) satisfies, for any η and $\gamma \in (0, 1)$ such that $\frac{\eta d}{\gamma} \leq \frac{1}{2}$,*

$$R_n \leq \gamma n + \frac{d\log 2}{\eta} + \eta \sum_{t=1}^{n} \sum_{i=1}^{d} \mathbb{E}\Big[\big(1 - a_t(i)^2\big)\widetilde{z}_t(i)^2\Big]. \tag{4}$$

In particular, with $\gamma = 2d\sqrt{\frac{\log 2}{3n}}$ and $\eta = \sqrt{\frac{\log 2}{3n}}$,

$$R_n \leq 2d\sqrt{3n\log 2}. \tag{5}$$

Remark that the regularizer (2) used here is in the class of Legendre functions with exchangeable Hessian. More precisely, following Audibert et al. (2011), (2) can be written (up to a numerical constant) as

$$F(x) = \sum_{i=1}^{d} \int_{-1}^{x_i} \tanh^{-1}(s) ds .$$

This type of regularizer was first studied (implicitly) by Audibert and Bubeck (2009) and Audibert and Bubeck (2010).

Proof Since F is Legendre on $\mathcal{A}$, F^* is differentiable on $\mathbb{R}^d$ and the gradient mapping of F^* is the inverse of the gradient mapping of F. Therefore, $(\nabla F^*)_i = \tanh$ because $(\nabla F)_i = \tanh^{-1}$. Then, thanks to (3) and Theorem 2, the regret can be bounded as:

$$\gamma n + \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \mathbb{E}\, D_{F^*}\big(-\eta \widetilde{z}_1^t, -\eta \widetilde{z}_1^{t-1}\big).$$

For the first term it is easy to see that $F(a) - F(a_1) \le d \log 2$. For the term involving the Bregman divergence, using elementary computations one obtains

$$D_{F^*}(u, v) = \sum_{i=1}^{d} \left(\log \frac{\cosh(u_i)}{\cosh(v_i)} - \tanh(v_i)(u_i - v_i) \right).$$

To prove (4) we need to show that $D_{F^*}(u, v) \le \sum_{i=1}^{d} \big(1 - \tanh^2(v_i)\big)(u_i - v_i)^2$. In fact, we prove that this inequality is true as soon as $\|u - v\|_\infty \le \frac{1}{2}$. The fact that the property is satisfied for the pair $(u, v) = \big(-\eta \widetilde{z}_1^t, -\eta \widetilde{z}_1^{t-1}\big)$ under consideration is established at the very end of the proof.

Using a basic hyperbolic identity, and the elementary inequalities $\exp(x) \le 1 + x + x^2, \forall x : |x| \le 1$ and $\log(1 + x) \le x$, one obtains

$$\log \left(\frac{\cosh(u_i)}{\cosh(v_i)} \right) - \tanh(v_i)(u_i - v_i)$$

$$= \log \left(\frac{\cosh(v_i)\cosh(u_i - v_i) + \sinh(v_i)\sinh(u_i - v_i)}{\cosh(v_i)} \right) - \tanh(v_i)(u_i - v_i)$$

$$= \log \left(\cosh(u_i - v_i) + \tanh(v_i)\sinh(u_i - v_i) \right) - \tanh(v_i)(u_i - v_i)$$

$$= \log \left(\frac{1 + \tanh(v_i)}{2} \exp(u_i - v_i) + \frac{1 - \tanh(v_i)}{2} \exp(-(u_i - v_i)) \right)$$

$$\quad - \log \exp \left(\tanh(v_i)(u_i - v_i) \right)$$

$$= \log \left(\frac{1 + \tanh(v_i)}{2} \exp \big((1 - \tanh(v_i))(u_i - v_i)\big) \right.$$

$$\quad \left. + \frac{1 - \tanh(v_i)}{2} \exp \big(-(1 + \tanh(v_i))(u_i - v_i)\big) \right)$$

$$\le \log \big(1 + (1 - \tanh^2(v_i))(u_i - v_i)^2\big) \le (1 - \tanh^2(v_i))(u_i - v_i)^2$$

which concludes the proof of (4). Now for the proof of (5) we first compute the matrix P_t:

$$
\begin{aligned}
\mathbb{E}\, \widetilde{a}_t \widetilde{a}_t^\top &= \frac{\gamma}{d} I_d + (1-\gamma) \sum_{i,j=1}^d \mathbb{E}\, \xi_t(i)\xi_t(j)\, e_i e_j^\top \\
&= \frac{\gamma}{d} I_d + (1-\gamma) I_d + (1-\gamma) \sum_{i \neq j} \mathbb{E}\, \xi_t(i)\xi_t(j)\, e_i e_j^\top \\
&= \frac{\gamma}{d} I_d + (1-\gamma) I_d + (1-\gamma) \sum_{i \neq j} a_t(i)a_t(j)\, e_i e_j^\top \\
&= \frac{\gamma}{d} I_d + (1-\gamma) a_t a_t^\top + (1-\gamma) \sum_{i=1}^d \left(1 - a_t(i)^2\right) e_i e_i^\top .
\end{aligned}
$$

To obtain (5) first note that $(1-\gamma) \sum_{i=1}^d \mathbb{E}\big[(1 - a_t(i)^2)\widetilde{z}_t(i)^2\big] \le \mathbb{E}\, \widetilde{z}_t^\top P_t \widetilde{z}_t$. Now we use a spectral decomposition of P_t in an orthonormal basis and write: $P_t = \sum_{i=1}^d \lambda_i v_i v_i^\top$. In particular we have $P_t^{-1} = \sum_{i=1}^d \frac{1}{\lambda_i} v_i v_i^\top$ and thus:

$$
\mathbb{E}\, \widetilde{a}_t^\top P_t^{-1} \widetilde{a}_t = \sum_{i=1}^d \frac{1}{\lambda_i} \mathbb{E}\, \widetilde{a}_t^\top v_i v_i^\top \widetilde{a}_t = \sum_{i=1}^d \frac{1}{\lambda_i} v_i^\top P_t v_i = \sum_{i=1}^d \frac{1}{\lambda_i} \lambda_i v_i^\top v_i = d.
$$

To conclude the proof it remains now to show that $\eta \|\widetilde{z}_t\|_\infty \le \frac{1}{2}$. First note that the smallest eigenvalue of P_t is larger than γ/d, and thus:

$$
\eta |\widetilde{z}_t(i)| = \eta \big| e_i^\top P_t^{-1} \widetilde{a}_t \widetilde{a}_t^\top z_t \big| \le \eta \big| e_i^\top P_t^{-1} \widetilde{a}_t \big| \le \frac{\eta d}{\gamma} \le \frac{1}{2}
$$

where the penultimate inequality follows from $|e_i^\top \widetilde{a}_t| \le 1$ and the last inequality follows from the assumption on η and γ. ∎

5. Improved regret for the Euclidean ball

In this section we restrict our attention to the action set $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\| \le 1\}$, where $\|\cdot\|$ denotes the Euclidean norm. Using EXP2 with John's exploration on a discretization of the Euclidean ball one obtains a regret bound of order $d\sqrt{n \log n}$ for this problem. A similar regret bound can be obtained with a computationally efficient algorithm, using the technique developed by Abernethy et al. (2008). Here we show that in fact one can attain efficiently a regret of order $\sqrt{dn \log n}$ using OSMD with the approriate regularizer F and random perturbation $\widetilde{a}_t$. More precisely here we use $F(x) = -\log(1 - \|x\|) - \|x\|$ (the motivation for this particular regularizer comes from the proof, see below). Moreover we perform the following perturbation of a point a_t in the interior of $\mathcal{A}$:

> Let ξ_t be a Bernoulli of parameter $\|a_t\|$, let I_t be drawn uniformly at random in $\{1, \ldots, d\}$, and let ε_t be Rademacher with parameter $\frac{1}{2}$. If $\xi_t = 1$, then play $\widetilde{a}_t = a_t/\|a_t\|$, else play $\widetilde{a}_t = \varepsilon_t e_{I_t}$.

It is easy to check that this perturbation is unbiased, in the sense that $\mathbb{E}\big[\widetilde{a}_t \mid a_t\big] = a_t$. Here we modify the estimate of step (b) in Figure 2, and instead we use:

$$\widetilde{z}_t = (1 - \xi_t)\frac{d}{1 - \|a_t\|}(z_t^\top \widetilde{a}_t)\widetilde{a}_t. \tag{6}$$

It is easy to check that this estimator satisfies the same key unbiasedness property than the one in step (b) in Figure 2, that is $\mathbb{E}\big[\widetilde{z}_t \mid a_t\big] = z_t$.

Note that the problem studied in this section was also specifically considered in Abernethy and Rakhlin (2009), with an emphasis on high probability bounds. In this paper the authors used the self-concordant barrier $F(x) = -\log(1 - \|x\|^2)$ with a similar perturbation scheme to the one proposed above. They obtain suboptimal rates, but a more careful analysis (precisely slightly modifying Section V.B., step (E)) can actually yield the same rate as the one we obtain. The strength of our approach is that it is in a sense more elementary (e.g., we do not require any results from the Interior Point Methods literature), but on the other hand the result of Abernethy and Rakhlin (2009) holds with high probability (though it is not clear if it possible to get the rate $\sqrt{dn \log n}$ with high probability).

Theorem 6 *Consider the online linear optimization problem with bandit feedback on $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$, and with $\mathcal{Z} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$. Then* OSMD *on $\mathcal{A}' = \{x \in \mathbb{R}^d : \|x\| \leq 1 - \gamma\}$ with the estimate (6), and $F(x) = -\log(1 - \|x\|) - \|x\|$ satisfies, for any η such that $\eta d \leq \frac{1}{2}$,*

$$R_n \leq \gamma n + \frac{\log \gamma^{-1}}{\eta} + \eta \sum_{t=1}^n \mathbb{E}\Big[\big(1 - \|a_t\|\big)\|\widetilde{z}_t\|^2\Big]. \tag{7}$$

In particular, with $\gamma = \frac{1}{\sqrt{n}}$ and $\eta = \sqrt{\frac{\log n}{2nd}}$,

$$R_n \leq 3\sqrt{dn \log n}. \tag{8}$$

Proof First, it is clear that by playing on $\mathcal{A}'$ instead of $\mathcal{A}$, one incurs an extra γn regret. Second, note that F is strictly convex (it is the composition of a convex and nondecreasing function with the Euclidean norm), differentiable, and

$$\nabla F(x) = \frac{x}{1 - \|x\|} . \tag{9}$$

In particular F is Legendre on $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$, and thus F^* is differentiable on $\mathbb{R}^d$. Now the regret with respect to $\mathcal{A}'$ can be bounded as follows, thanks to Theorem 2,

$$\frac{\sup_{a \in \mathcal{A}'} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^n \mathbb{E}\, D_{F^*}\Big(\nabla F(a_t) - \eta\widetilde{z}_t, \nabla F(a_t)\Big).$$

The first term is clearly bounded by $\frac{1}{\eta}\log\frac{1}{\gamma}$ (we use the fact that $a_1 = 0$). For the second term we need to do a few computations (the first one follows from (9) and the fact that F is Legendre):

$$
\begin{aligned}
\nabla F^*(u) &= \frac{u}{1 + \|u\|}, \\
F^*(u) &= -\log(1 + \|u\|) + \|u\|, \\
D_{F^*}(u, v) &= \frac{1}{1 + \|v\|}\left(\|u\| - \|v\| + \|u\| \cdot \|v\| - v^\top u - (1 + \|v\|)\log\left(1 + \frac{\|u\| - \|v\|}{1 + \|v\|}\right)\right).
\end{aligned}
$$

Let $\Theta(u, v)$ such that $D_{F^*}(u, v) = \frac{1}{1+\|v\|}\Theta(u, v)$. First note that

$$\frac{1}{1 + \|\nabla F(a_t)\|} = 1 - \|a_t\| . \tag{10}$$

Thus, in order to prove (7) it remains to show that $\Theta(u, v) \leq \|u - v\|^2$, for $(u, v) = \left(-\eta \widetilde{z}_1^t, -\eta \widetilde{z}_1^{t-1}\right)$. In fact we shall prove that this inequality holds true as soon as $\frac{\|u\|-\|v\|}{1+\|v\|} \geq -\frac{1}{2}$. This is the case for the pair (u, v) under consideration, since by the triangle inequality, equations (6) and (10), and the assumption on η:

$$\frac{\|u\| - \|v\|}{1 + \|v\|} \geq -\frac{\eta\|\widetilde{z}_t\|}{1 + \|v\|} \geq -\eta d \geq -\frac{1}{2}.$$

Now using that $\log(1 + x) \geq x - x^2, \forall x \geq -\frac{1}{2}$, we obtain that for u, v such that $\frac{\|u\|-\|v\|}{1+\|v\|} \geq -\frac{1}{2}$,

$$\begin{aligned}
\Theta(u, v) &\leq & \frac{(\|u\| - \|v\|)^2}{1 + \|v\|} + \|u\| \cdot \|v\| - v^\top u \\
&\leq & (\|u\| - \|v\|)^2 + \|u\| \cdot \|v\| - v^\top u \\
&= & \|u\|^2 + \|v\|^2 - \|u\| \cdot \|v\| - v^\top u \\
&= & \|u - v\|^2 + 2v^\top u - \|u\| \cdot \|v\| - v^\top u \\
&\leq & \|u - v\|^2
\end{aligned}$$

which concludes the proof of (7). Now for the proof of (8) it suffices to note that:

$$\mathbb{E}\left[\left(1 - \|a_t\|\right)\|\widetilde{z}_t\|^2\right] = (1 - \|a_t\|) \sum_{i=1}^{d} \frac{1 - \|a_t\|}{d} \frac{d^2}{(1 - \|a_t\|)^2}(z_t^\top e_i)^2 = d\|z_t\|^2 \leq d$$

along with straightforward computations. ∎

Acknowledgments

References

J. Abernethy and A. Rakhlin. Beating the adaptive bandit with high probability. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

J.-Y. Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

B. Awerbuch and R. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. Cambridge University Press, 1997.

S. Bubeck. Introduction to online optimization. Lecture Notes, 2011.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2011. To appear.

W. Chu, L. Li, L. Reyzin, and R.E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Conference and Workshop Proceedings, 2011.

V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 345–352, 2008.

Martin Grötschel, Lászlo Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer, second corrected edition edition, 1993. ISBN 3-540-56740-2, 0-387-56740-2 (U.S.).

E. Hazan. The convex optimization approach to regret minimization. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2011.

S. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. arXiv:0910.0610v2, 2010.

J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45:301–329, 2001.

H. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *In Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 109–123, 2004.

A. Nemirovski. Advances in convex optimiza- tion: Conic programming. In *Proceedings of the International Congress of Mathematicians, 2006.* EMS-European Mathematical Society Publishing House, 2007.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience, 1983.

A. Rakhlin. Lecture notes on online learning. 2009.

N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.