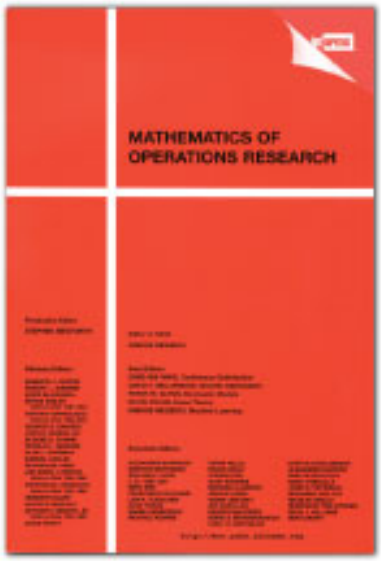

Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Regret in Online Combinatorial Optimization

Jean-Yves Audiberthttp://certis.enpc.fr/~audibert/, Sébastien Bubeckhttp://www.princeton.edu/~sbubeck/, Gábor Lugosihttp://www.econ.upf.edu/~lugosi/

To cite this article:
Jean-Yves Audiberthttp://certis.enpc.fr/~audibert/, Sébastien Bubeckhttp://www.princeton.edu/~sbubeck/, Gábor Lugosihttp://www.econ.upf.edu/~lugosi/ (2014) Regret in Online Combinatorial Optimization. Mathematics of Operations Research 39(1):31-45. http://dx.doi.org/10.1287/moor.2013.0598

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 39, No. 1, February 2014, pp. 31–45
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2013.0598

Regret in Online Combinatorial Optimization

Jean-Yves Audibert
Imagine, University Paris Est; and Sierra, CNRS/ENS/INRIA, Paris, France,
audibert@imagine.enpc.fr, http://certis.enpc.fr/~audibert/

Sébastien Bubeck
Department of Operations Research and Financial Engineering, Princeton University,
sbubeck@princeton.edu, http://www.princeton.edu/~sbubeck/

Gábor Lugosi
ICREA and Pompeu Fabra University, Barcelona, Spain,
gabor.lugosi@upf.edu, http://www.econ.upf.edu/~lugosi/

We address online linear optimization problems when the possible actions of the decision maker are represented by binary vectors. The regret of the decision maker is the difference between her realized loss and the minimal loss she would have achieved by picking, in hindsight, the best possible action. Our goal is to understand the magnitude of the best possible (minimax) regret. We study the problem under three different assumptions for the feedback the decision maker receives: full information, and the partial information models of the so-called "semi-bandit" and "bandit" problems. In the full information case we show that the standard exponentially weighted average forecaster is a provably suboptimal strategy. For the semi-bandit model, by combining the Mirror Descent algorithm and the INF (Implicitely Normalized Forecaster) strategy, we are able to prove the first optimal bounds. Finally, in the bandit case we discuss existing results in light of a new lower bound, and suggest a conjecture on the optimal regret in that case.

Keywords: online optimization; combinatorial optimization; mirror descent; multi-armed bandits, minimax regret
MSC2000 subject classification: Primary: 68T05; secondary: 90C27
OR/MS subject classification: Primary: decision analysis; secondary: sequential
History: Received April 19, 2012; revised October 23, 2012, February 25, 2013. Published online in *Articles in Advance* May 6, 2013.

1. Introduction. In this paper we consider the framework of online linear optimization. The setup may be described as a repeated game between a "decision maker" (or simply "player" or "forecaster") and an "adversary" as follows: at each time instance $t = 1, \dots, n$, the player chooses, possibly in a randomized way, an action from a given finite 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 forecaster is $a_t^T z_t$. The goal of the player is to minimize the expected cumulative loss $\mathbb{E}\sum_{t=1}^n a_t^T z_t$, where the expectation is taken with respect to the player's internal randomization (and eventually the adversary's randomization).

In the basic "full-information" version of this problem, the player observes the adversary's move, z_t, at the end of round t. Another important model for feedback is the so-called *bandit* problem, in which the player only observes the incurred loss $a_t^T z_t$. As a measure of performance we define the regret[1] of the player as

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

In this paper we address a specific example of online linear optimization: we assume that the action set $\mathcal{A}$ is a subset of the d-dimensional hypercube $\{0,1\}^d$ such that $\forall a \in \mathcal{A}$, $\|a\|_1 = m$, and the adversary has a bounded loss per coordinate; that is[2] $\mathcal{Z} = [0,1]^d$. We call this setting *online combinatorial optimization*. As we will see below, this restriction of the general framework contains a rich class of problems. Indeed, in many interesting cases, actions are naturally represented by Boolean vectors.

In addition to the full information and bandit versions of online combinatorial optimization, we also consider another type of feedback which makes sense only in this combinatorial setting. In the *semi-bandit* version, we assume that the player observes only the coordinates of z_t that were played in a_t; that is, the player observes the vector $(a_t(1)z_t(1), \dots, a_t(d)z_t(d))$. All three variants of online combinatorial optimization are sketched in Figure 1.

[1] In the full information version, it is straightforward to obtain upper bounds for the stronger notion of regret $\mathbb{E}\sum_{t=1}^n a_t^T z_t - \mathbb{E}\min_{a\in\mathcal{A}}\sum_{t=1}^n a^T z_t$, which is always at least as large as R_n. However, for partial information games, this requires more work. In this paper we only consider R_n as a measure of the regret.

[2] Note that since all actions have the same size, i.e., $\|a\|_1 = m$, $\forall a \in \mathcal{A}$, one can reduce the case of $\mathcal{Z} = [\alpha, \beta]^d$ to $\mathcal{Z} = [0,1]^d$ via a simple renormalization.

Parameters: set of actions $\mathcal{A} \subset \{0,1\}^d$; number of rounds $n \in \mathbb{N}$.
For each round $t = 1, 2, \ldots, n$;
(1) the player chooses a probability distribution p_t over $\mathcal{A}$ and draws a random action $a_t \in \mathcal{A}$ according to p_t;
(2) simultaneously, the adversary selects a loss vector $z_t \in [0,1]^d$ (without revealing it);
(3) the player incurs the loss $a_t^T z_t$. She observes
—the loss vector z_t in the full information setting,
—the coordinates $z_t(i)a_t(i)$ in the semi-bandit setting,
—the instantaneous loss $a_t^T z_t$ in the bandit setting.
Goal: The player tries to minimize her cumulative loss $\sum_{t=1}^n a_t^T z_t$.

FIGURE 1. Online combinatorial optimization.

More rigorously, online combinatorial optimization is defined as a repeated game between a "player" and an "adversary." At each round $t = 1, \ldots, n$ of the game, the player chooses a probability distribution p_t over the set of actions $\mathcal{A} \subset \{0,1\}^d$ and draws a random action $a_t \in \mathcal{A}$ according to p_t. Simultaneously, the adversary chooses a vector $z_t \in [0,1]^d$. More formally, z_t is a measurable function of the "past" $(p_s, a_s, z_s)_{s=1,\ldots,t-1}$. In the full information case, p_t is a measurable function of $(p_s, a_s, z_s)_{s=1,\ldots,t-1}$. In the semi-bandit case, p_t is a measurable function of $(p_s, a_s, (a_s(i)z_s(i))_{i=1,\ldots,d})_{s=1,\ldots,t-1}$, and in the bandit problem it is a measurable function of $(p_s, a_s, (a_s^T z_s))_{s=1,\ldots,t-1}$.

1.1. Motivating examples. Many problems can be tackled under the online combinatorial optimization framework. We give here three simple examples:

- *m-sets*. In this example we consider the set $\mathcal{A}$ of all $\binom{d}{m}$ Boolean vectors in dimension d with exactly m ones. In other words, at every time step, the player selects m actions out of d possibilities. When $m = 1$, the semi-bandit and bandit versions coincide and correspond to the standard (adversarial) multi-armed bandit problem.
- *Online shortest path problem*. Consider a communication network represented by a graph in which one has to send a sequence of packets from one fixed vertex to another. For each packet one chooses a path through the graph and suffers a certain delay which is the sum of the delays on the edges of the path. Depending on the traffic, the delays on the edges may change, and, at the end of each round, according to the assumed level of feedback, the player observes either the delays of all edges, the delays of each edge on the chosen path, or only the total delay of the chosen path. The player's objective is to minimize the total delay for the sequence of packets.

One can represent the set of valid paths from the starting vertex to the end vertex as a set $\mathcal{A} \subset \{0,1\}^d$, where d is the number of edges. If at time t, $z_t \in [0,1]^d$ is the vector of delays on the edges, then the delay of a path $a \in \mathcal{A}$ is $z_t^T a$. Thus this problem is an instance of online combinatorial optimization in dimension d, where d is the number of edges in the graph. In this paper we assume, for simplicity, that all valid paths have the same length m.

- *Ranking*. Consider the problem of selecting a ranking of m items out of M possible items. For example a website could have a set of M ads, and it has to select a ranked list of m of these ads to appear on the web page. One can rephrase this problem as selecting a matching of size m on the complete bipartite graph $K_{m,M}$ (with $d = m \times M$ edges). In the online learning version of this problem, each day the website chooses one such list, and gains one dollar for each click on the ads. This problem can easily be formulated as an online combinatorial optimization problem.

Our theory applies to many more examples, such as spanning trees (which can be useful in certain communication problems), or m-intervals.

1.2. Previous work.

- *Full information*. The full-information setting is now fairly well understood, and an optimal regret bound (in terms of m, d, n) was obtained by Koolen et al. [26]. Previous papers under full information feedback also include Gentile and Warmuth [14], Kivinen and Warmuth [25], Grove et al. [15], Takimoto and Warmuth [34], Kalai and Vempala [22], Warmuth and Kuzmin [36], Herbster and Warmuth [20], and Hazan et al. [18].
- *Semi-bandit*. The first paper on the adversarial multi-armed bandit problem (i.e., the special case of m-sets with $m = 1$) is by Auer et al. [4] who derived a regret bound of order $\sqrt{dn \log d}$. This result was improved to $\sqrt{dn}$ by Audibert and Bubeck [2, 3]. György et al. [16] consider the online shortest path problem and derive suboptimal regret bounds (in terms of the dependency on m and d). Uchiya et al. [35] (respectively,

TABLE 1. Bounds on the minimax regret (up to constant factors). The new results are set in boldface. In this paper we also show that EXP2 in the full information case has a regret bounded below by $d^{3/2}\sqrt{n}$ (when m is of order d).

	Full information	Semi-bandit	Bandit
Lower bound	$m\sqrt{n \log \frac{d}{m}}$	$\sqrt{mdn}$	$\mathbf{m\sqrt{dn}}$
Upper bound	$m\sqrt{n \log \frac{d}{m}}$	$\mathbf{\sqrt{mdn}}$	$m^{3/2}\sqrt{dn \log \frac{d}{m}}$

Kale et al. [23]) derived optimal regret bounds for the case of m-sets (respectively, for the problem of ranking selection) up to logarithmic factors.

- *Bandit*. McMahan and Blum [27], and Awerbuch and Kleinberg [5] were the first to consider this setting, and obtained suboptimal regret bounds (in terms of n). The first paper with optimal dependency in n was by Dani et al. [12]. The dependency on m and d was then improved in various ways by Abernethy et al. [1], Cesa-Bianchi and Lugosi [11], and Bubeck et al. [9]. We discuss these bounds in detail in §4. In particular, we argue that the optimal regret bound in terms of d (and m) is still an open problem.

We also refer the interested reader to the recent survey by Bubeck and Cesa-Bianchi [8] for an overview of bandit problems in various other settings.

1.3. Contribution and contents of the paper. In this paper we are primarily interested in the optimal *minimax regret* in terms of m, d, and n. More precisely, our aim is to determine the order of magnitude of the following quantity: For a given feedback assumption, write sup for the supremum over all adversaries and inf for the infimum over all allowed strategies for the player under the feedback assumption. (Recall the definition of "adversary" and "player" from the introduction.) Then we are interested in

$$\max_{\mathcal{A} \subset \{0,1\}^d : \forall a \in \mathcal{A},\ \|a\|_1 = m} \inf \sup R_n.$$

Our contribution to the study of this quantity is threefold. First, we unify the algorithms used in Abernethy et al. [1], Koolen et al. [26], Uchiya et al. [35], and Kale et al. [23] under the umbrella of mirror descent. The idea of mirror descent goes back to Nemirovski [28], and Nemirovski and Yudin [29]. A somewhat similar concept was re-discovered in online learning by Herbster and Warmuth [20], Grove et al. [15],and Kivinen and Warmuth [25] under the name of potential-based gradient descent; see Cesa-Bianchi and Lugosi [10, Chapter 11]. Recently, these ideas have been flourishing; see, for instance, Shalev-Schwartz [33], Rakhlin [30], Hazan [17], and Bubeck [7]. Our main theorem (Theorem 2.2) allows one to recover almost all known regret bounds for online combinatorial optimization. This first contribution leads to our second main result, the improvement of the known upper bounds for the semi-bandit game. In particular, we propose a different proof of the minimax regret bound of the order of $\sqrt{nd}$ in the standard d-armed bandit game that is much simpler than the one provided in Audibert and Bubeck [3] (which also improves the constant factor). In addition to these upper bounds we prove two new lower bounds. First we answer a question of Koolen et al. [26] by showing that the exponentially weighted average forecaster is provably suboptimal for online combinatorial optimization. Our second lower bound is a minimax lower bound in the bandit setting which improves known results by an order of magnitude. A summary of known bounds and the new bounds proved in this paper can be found in Table 1.

The paper is organized as follows. In §2 we introduce the two algorithms discussed in this paper. In particular in §2.1 we discuss the popular exponentially weighted average forecaster and we show that it is a provably suboptimal strategy. Then in §2.2 we describe our main algorithm, OSMD (online stochastic mirror descent), and prove a general regret bound in terms of the Bregman divergence of the Fenchel-Legendre dual of the Legendre function defining the strategy. In §3 we derive upper bounds for the regret in the semi-bandit case for OSMD with appropriately chosen Legendre functions. Finally in §4 we prove a new lower bound for the bandit setting, and we formulate a conjecture on the correct order of magnitude of the regret for that problem based on this new result and the regret bounds obtained in Abernethy et al. [1] and Bubeck et al. [9].

2. Algorithms. In this section we discuss two classes of algorithms that have been proposed for online combinatorial optimization.

EXP2:
Parameter: Learning rate η.
Let $p_1 = (1/|\mathcal{A}|, \ldots, 1/|\mathcal{A}|) \in \mathbb{R}^{|\mathcal{A}|}$.
For each round $t = 1, 2, \ldots, n$;
(a) Play $a_t \sim p_t$ and observe
—the loss vector z_t in the full information game,
—the coordinates $z_t(i)\mathbb{1}_{a_t(i)=1}$ in the semi-bandit game,
—the instantaneous loss $a_t^T z_t$ in the bandit game.
(b) Estimate the loss vector z_t by $\tilde{z}_t$. For instance, one may take
—$\tilde{z}_t = z_t$ in the full information game,
—$\tilde{z}_t(i) = (z_t(i)/\sum_{a\in\mathcal{A}:a(i)=1} p_t(a))a_t(i)$ in the semi-bandit game,
—$\tilde{z}_t = P_t^+ a_t a_t^T z_t$, with $P_t = \mathbb{E}_{a\sim p_t}(aa^T)$ in the bandit game.
(c) Update the probabilities, for all $a \in \mathcal{A}$,

$$p_{t+1}(a) = \frac{\exp(-\eta a^T \tilde{z}_t) p_t(a)}{\sum_{b\in\mathcal{A}} \exp(-\eta b^T \tilde{z}_t^T) p_t(b)}.$$

FIGURE 2. The EXP2 strategy. The notation $\mathbb{E}_{a\sim p_t}$ denotes expected value with respect to the random choice of a when it is distributed according to p_t.

2.1. Expanded Exponential weights (EXP2). The simplest approach to online combinatorial optimization is to consider each action of $\mathcal{A}$ as an independent "expert," and then apply a generic regret minimizing strategy. Perhaps the most popular such strategy is the exponentially weighted average forecaster (see, e.g., Cesa-Bianchi and Lugosi [10]). (This strategy is sometimes called Hedge; see Freund and Schapire [13].) We call the resulting strategy for the online combinatorial optimization problem EXP2; see Figure 2. In the full information setting, EXP2 corresponds to "Expanded Hedge," as defined in Koolen et al. [26]. In the semi-bandit case, EXP2 was studied by György et al. [16] while in the bandit case in Dani et al. [12], Cesa-Bianchi and Lugosi [11], and Bubeck et al. [9]. Note that in the bandit case, EXP2 is mixed with an *exploration distribution*; see §4 for more details.

Despite strong interest in this strategy, no optimal regret bound has been derived for it in the combinatorial setting. More precisely, the best bound (which can be derived from a standard argument; see for example Dani et al. [12] or Koolen et al. [26]) is of order $m^{3/2}\sqrt{n\log(d/m)}$. On the other hand, in Koolen et al. [26] the authors showed that by using mirror descent (see next section) with the negative entropy, one obtains a regret bounded by $m\sqrt{n\log(d/m)}$. Furthermore, this latter bound is clearly optimal (up to a numerical constant), as one can see from the standard lower bound in prediction with expert advice (consider the set $\mathcal{A}$ that corresponds to playing m expert problems in parallel with d/m experts in each problem). In Koolen et al. [26] the authors leave as an open question the problem of whether it would be possible to improve the bound for EXP2 to obtain the optimal order of magnitude. The following theorem shows that this is impossible, and that, in fact, EXP2 is a provably suboptimal strategy.

THEOREM 2.1. *Let $n \geq d$. There exists a subset $\mathcal{A} \subset \{0,1\}^d$ such that in the full information setting, the regret of the* EXP2 *strategy (for any learning rate η), satisfies*

$$\sup_{\text{adversary}} R_n \geq 0.01\, d^{3/2}\sqrt{n}.$$

The proof is deferred to the appendix.

2.2. Online Stochastic Mirror Descent. In this section we describe the main algorithm studied in this paper. We call it online stochastic mirror descent (OSMD). Each term in this name refers to a part of the algorithm: *Mirror descent* originates in the work of Nemirovski and Yudin [29]. The idea of mirror descent is to perform a gradient descent, where the update with the gradient is performed in the dual space (defined by some Legendre function F) rather than in the primal (see below for a precise formulation). The *stochastic* part takes its origin from Robbins and Monro [31] and from Kiefer and Wolfowitz [24]. The key idea is that it is enough to observe an unbiased estimate of the gradient rather than the true gradient to perform a gradient descent. Finally the *online* part comes from Zinkevich [37]. Zinkevich derived the online gradient descent (OGD) algorithm, which is a version of gradient descent tailored to online optimization.

To properly describe the OSMD strategy, we recall a few concepts from convex analysis; see Hiriart-Urruty and Lemaréchal [21] for a thorough treatment of this subject. Let $\mathcal{D} \subset \mathbb{R}^d$ be an open convex set, and let $\bar{\mathcal{D}}$ be the closure of $\mathcal{D}$.

Definition 2.1. We call Legendre any continuous function $F\colon \bar{\mathcal{D}} \to \mathbb{R}$ such that
(i) F is strictly convex continuously differentiable on $\mathcal{D}$,
(ii) $\lim_{x \to \bar{\mathcal{D}} \setminus \mathcal{D}} \|\nabla F(x)\| = +\infty$.[3]
The Bregman divergence $D_F\colon \bar{\mathcal{D}} \times \mathcal{D}$ associated to a Legendre function F is defined by

$$D_F(x, y) = F(x) - F(y) - (x - y)^T \nabla F(y).$$

Moreover, we say that $\mathcal{D}^* = \nabla F(\mathcal{D})$ is the dual space of $\mathcal{D}$ under F. We also denote by F^* the Legendre-Fenchel transform of F defined by

$$F^*(u) = \sup_{x \in \bar{\mathcal{D}}} \big(x^T u - F(x)\big).$$

Lemma 2.1. *Let F be a Legendre function. Then $F^{**} = F$ and $\nabla F^* = (\nabla F)^{-1}$ on the set $\mathcal{D}^*$. Moreover, $\forall\, x, y \in \mathcal{D}$,*

$$D_F(x, y) = D_{F^*}\big(\nabla F(y), \nabla F(x)\big). \tag{1}$$

The lemma above is the key to understanding how a Legendre function acts on the space. The gradient ∇F maps $\mathcal{D}$ to the dual space $\mathcal{D}^*$, and ∇F^* is the inverse mapping from the dual space to the original (primal) space. Moreover, (1) shows that the Bregman divergence in the primal space corresponds exactly to the Bregman divergence of the Legendre-Fenchel transform in the dual space. A proof of this result can be found, for example, in Cesa-Bianchi and Lugosi [10, Chapter 11].

We now have all ingredients to describe the OSMD strategy; see Figure 3 for the precise formulation. Note that step (d) is well defined if the following consistency condition is satisfied:

$$\nabla F(x) - \eta \tilde{z}_t \in \mathcal{D}^*, \quad \forall\, x \in \mathrm{Conv}(\mathcal{A}) \cap \mathcal{D}. \tag{2}$$

In the full information setting, algorithms of this type were studied by Abernethy et al. [1], Rakhlin [30], and Hazan [17]. In these papers the authors adopted the presentation suggested by Beck and Teboulle [6], which corresponds to a follow-the-regularized-leader (FTRL)-type strategy. There the focus was on F being strongly convex with respect to some norm. Moreover, in Abernethy et al. [1] the authors also consider the bandit case, and switch to F being a self-concordant barrier for the convex hull of $\mathcal{A}$ (see §4 for more details). Another line of work studied this type of algorithm with F being the negative entropy; see Koolen et al. [26] for the full information case and Uchiya et al. [35], and Kale et al. [23] for specific instances of the semi-bandit case. All these results are unified and described in detail in Bubeck [7]. In this paper we consider a new type of Legendre functions F inspired by Audibert and Bubeck [3]; see §3.

Regarding computational complexity, OSMD is efficient as soon as the polytope $\mathrm{Conv}(\mathcal{A})$ can be described by a polynomial (in d) number of constraints. Indeed in that case steps (a)–(b) can be performed efficiently jointly (one can get an algorithm by looking at the proof of Carathéodory's theorem), and step (d) is a convex program with a polynomial number of constraints. In many interesting examples (such as m-sets, selection of rankings, spanning trees, paths in acyclic graphs) one can describe the convex hull of $\mathcal{A}$ by a polynomial number of constraints; see Schrijver [32]. On the other hand, there also exist important examples where this is not the case (such as paths on general graphs). Also note that for some specific examples it is possible to implement OSMD with improved computational complexity; see Koolen et al. [26].

In this paper we restrict our attention to the combinatorial learning setting in which $\mathcal{A}$ is a subset of $\{0, 1\}^d$, and the loss is linear. However, one should note that this specific form of $\mathcal{A}$ plays no role in the definition of OSMD. Moreover, if the loss is not linear, then one can modify OSMD by performing a gradient update with a gradient of the loss (rather than the loss vector z_t). See Bubeck [7] for more details on this approach.

The following result is at the basis of our improved regret bounds for OSMD in the semi-bandit setting; see §3.

Theorem 2.2. *Suppose that (2) is satisfied and the loss estimates are unbiased in the sense that $\mathbb{E}_{a_t \sim p_t} \tilde{z}_t = z_t$. Then the regret of the* OSMD *strategy satisfies*

$$R_n \le \frac{\sup_{a \in \mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \mathbb{E}\, D_{F^*}(\nabla F(x_t) - \eta \tilde{z}_t, \nabla F(x_t)).$$

[3] By the equivalence of norms in $\mathbb{R}^d$, this definition does not depend on the choice of the norm.

OSMD:
Parameters:
- learning rate $\eta > 0$,
- Legendre function F defined on $\bar{\mathcal{D}} \supset \mathrm{Conv}(\mathcal{A})$.

Let $x_1 \in \arg\min_{x \in \mathrm{Conv}(\mathcal{A})} F(x)$.
For each round $t = 1, 2, \ldots, n$;
(a) Let p_t be a distribution on the set $\mathcal{A}$ such that $x_t = \mathbb{E}_{a \sim p_t} a$.
(b) Draw a random action a_t according to the distribution p_t and observe the feedback.
(c) Based on the observed feedback, estimate the loss vector z_t by $\tilde{z}_t$.
(d) Let $w_{t+1} \in \mathcal{D}$ satisfy

$$\nabla F(w_{t+1}) = \nabla F(x_t) - \eta \tilde{z}_t. \tag{3}$$

(e) Project the weight vector w_{t+1} defined by (3) on the convex hull of $\mathcal{A}$:

$$x_{t+1} \in \arg\min_{x \in \mathrm{Conv}(\mathcal{A})} D_F(x, w_{t+1}). \tag{4}$$

FIGURE 3. Online stochastic mirror descent (OSMD).

PROOF. Let $a \in \mathcal{A}$. Using that a_t and $\tilde{z}_t$ are unbiased estimates of x_t and z_t, we have

$$\mathbb{E} \sum_{t=1}^{n} (a_t - a)^T z_t = \mathbb{E} \sum_{t=1}^{n} (x_t - a)^T \tilde{z}_t.$$

Using (3), and applying the definition of the Bregman divergences, one obtains

$$\begin{aligned} \eta \tilde{z}_t^T (x_t - a) &= (a - x_t)^T (\nabla F(w_{t+1}) - \nabla F(x_t)) \\ &= D_F(a, x_t) + D_F(x_t, w_{t+1}) - D_F(a, w_{t+1}). \end{aligned}$$

By the Pythagorean theorem for Bregman divergences (see, e.g., Cesa-Bianchi and Lugosi [10, Lemma 11.3]), we have $D_F(a, w_{t+1}) \geq D_F(a, x_{t+1}) + D_F(x_{t+1}, w_{t+1})$; hence

$$\eta \tilde{z}_t^T (x_t - a) \leq D_F(a, x_t) + D_F(x_t, w_{t+1}) - D_F(a, x_{t+1}) - D_F(x_{t+1}, w_{t+1}).$$

Summing over t gives

$$\sum_{t=1}^{n} \eta \tilde{z}_t^T (x_t - a) \leq D_F(a, a_1) - D_F(a, a_{n+1}) + \sum_{t=1}^{n} \big(D_F(x_t, w_{t+1}) - D_F(x_{t+1}, w_{t+1}) \big).$$

By the nonnegativity of the Bregman divergences, we get

$$\sum_{t=1}^{n} \eta \tilde{z}_t^T (x_t - a) \leq D_F(a, a_1) + \sum_{t=1}^{n} D_F(x_t, w_{t+1}).$$

From (1), one has $D_F(x_t, w_{t+1}) = D_{F^*}(\nabla F(x_t) - \eta \tilde{z}_t, \nabla F(x_t))$. Moreover, by writing the first-order optimality condition for x_1, one directly obtains $D_F(a, x_1) \leq F(a) - F(x_1)$ which concludes the proof. □

Note that, if F admits a Hessian, denoted $\nabla^2 F$, that is always invertible, then one can prove that, up to a third-order term (in $\tilde{z}_t$), the regret bound can be written as

$$R_n \lessapprox \frac{\sup_{a \in \mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \tilde{z}_t^T \big(\nabla^2 F(x_t) \big)^{-1} \tilde{z}_t. \tag{5}$$

The main technical difficulty is to control the third-order error term in this inequality.

3. Semi-bandit feedback. In this section we consider online combinatorial optimization with semi-bandit feedback. As we already discussed, in the full information case Koolen et al. [26] proved that OSMD with the negative entropy is a minimax optimal strategy. We first prove a regret bound when one uses this strategy with the following estimate for the loss vector:

$$\tilde{z}_t(i) = \frac{z_t(i) a_t(i)}{x_t(i)}. \tag{6}$$

Note that this is a valid estimate since it only makes use of $(z_t(1)a_t(1), \ldots, z_t(d)a_t(d))$. Moreover, it is unbiased with respect to the random draw of a_t from p_t, since by definition, $\mathbb{E}_{a_t \sim p_t} a_t(i) = x_t(i)$. In other words, $\mathbb{E}_{a_t \sim p_t} \tilde{z}_t(i) = z_t(i)$.

THEOREM 3.1. *The regret of OSMD with* $F(x) = \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d x_i$ *(and* $\mathcal{D} = (0, +\infty)^d$*) and any nonnegative unbiased loss estimate* $\tilde{z}_t(i) \geq 0$ *satisfies*

$$R_n \leq \frac{m \log(d/m)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^d x_t(i) \tilde{z}_t(i)^2.$$

In particular, with the estimate (6) *and* $\eta = \sqrt{2((m \log dm)/(nd))}$,

$$R_n \leq \sqrt{2mdn \log \frac{d}{m}}.$$

PROOF. One can easily see that for the negative entropy the dual space is $\mathcal{D}^* = \mathbb{R}^d$. Thus, (2) is verified and OSMD is well defined. Moreover, again by straightforward computations, one can also see that

$$D_{F^*}\big(\nabla F(x), \nabla F(y)\big) = \sum_{i=1}^d y(i) \Theta\big((\nabla F(x) - \nabla F(y))(i)\big), \tag{7}$$

where $\Theta(x) = \exp(x) - 1 - x$. Thus, using Theorem 2.2 and the facts that $\Theta(x) \leq x^2/2$ for $x \leq 0$ and $\sum_{i=1}^d x_t(i) \leq m$, one obtains

$$\begin{aligned} R_n &\leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^n \mathbb{E}\, D_{F^*}\big(\nabla F(x_t) - \eta \tilde{z}_t, \nabla F(x_t)\big) \\ &\leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^d x_t(i) \tilde{z}_t(i)^2 \end{aligned}$$

. The proof of the first inequality is concluded by noting that:

$$F(a) - F(x_1) \leq \sum_{i=1}^d x_1(i) \log \frac{1}{x_1(i)} \leq m \log\left(\sum_{i=1}^d \frac{x_1(i)}{m} \frac{1}{x_1(i)} \right) = m \log \frac{d}{m}.$$

The second inequality follows from

$$\mathbb{E}\, x_t(i) \tilde{z}_t(i)^2 \leq \mathbb{E}\, \frac{a_t(i)}{x_t(i)} = 1. \quad \square$$

Using the standard $\sqrt{dn}$ lower bound for the multi-armed bandit (which corresponds to the case where $\mathcal{A}$ is the canonical basis; see, e.g., Audibert and Bubeck [3, Theorem 30]), one can directly obtain a lower bound of order $\sqrt{mdn}$ for our setting. Thus the upper bound derived in Theorem 3.1 has an extraneous logarithmic factor compared to the lower bound. This phenomenon already appeared in the basic multi-armed bandit setting. In that case, the extra logarithmic factor was removed in Audibert and Bubeck [2] by resorting to a new class of strategies for the expert problem, called INF (implicitely normalized forecaster). Next we generalize this class of algorithms to the combinatorial setting, and thus remove the extra logarithmic factor. First we introduce the notion of a potential and the associated Legendre function.

DEFINITION 3.1. Let $\omega \geq 0$. A function $\psi\colon (-\infty, a) \to \mathbb{R}_+^*$ for some $a \in \mathbb{R} \cup \{+\infty\}$ is called an ω-potential if it is convex, continuously differentiable, and satisfies

$$\begin{aligned} &\lim_{x \to -\infty} \psi(x) = \omega, \qquad &&\lim_{x \to a} \psi(x) = +\infty, \\ &\psi' > 0, &&\int_\omega^{\omega+1} |\psi^{-1}(s)|\, ds < +\infty. \end{aligned}$$

For every potential ψ we associate the function F_ψ defined on $\mathcal{D} = (\omega, +\infty)^d$ by:

$$F_\psi(x) = \sum_{i=1}^d \int_\omega^{x_i} \psi^{-1}(s)\, ds.$$

In this paper we restrict our attention to 0-potentials which we will simply call *potentials*. A nonzero value of ω may be used to derive regret bounds that hold with high probability (instead of pseudo-regret bounds; see footnote 1).

The first order optimality condition for (4) implies that OSMD with F_ψ is a direct generalization of INF with potential ψ, in the sense that the two algorithms coincide when $\mathcal{A}$ is the canonical basis. Note, in particular, that with $\psi(x) = \exp(x)$ we recover the negative entropy for F_ψ. In Audibert and Bubeck [3], the choice of $\psi(x) = (-x)^q$ with $q > 1$ was recommended. We show in Theorem 3.2 that here, again, this choice gives a minimax optimal strategy.

LEMMA 3.1. *Let ψ be a potential. Then $F = F_\psi$ is Legendre and for all $u, v \in \mathcal{D}^* = (-\infty, a)^d$ such that $u_i \le v_i$, $\forall i \in \{1, \dots, d\}$,*

$$D_{F^*}(u, v) \le \tfrac{1}{2} \sum_{i=1}^{d} \psi'(v_i)(u_i - v_i)^2.$$

PROOF. A direct examination shows that $F = F_\psi$ is a Legendre function. Moreover, since $\nabla F^*(u) = (\nabla F)^{-1}(u) = (\psi(u_1), \dots, \psi(u_d))$, we obtain

$$D_{F^*}(u, v) = \sum_{i=1}^{d} \left(\int_{v_i}^{u_i} \psi(s)\, ds - (u_i - v_i)\psi(v_i) \right).$$

From a Taylor expansion, we get

$$D_{F^*}(u, v) \le \sum_{i=1}^{d} \max_{s \in [u_i, v_i]} \tfrac{1}{2} \psi'(s)(u_i - v_i)^2.$$

Since the function ψ is convex, and $u_i \le v_i$, we have

$$\max_{s \in [u_i, v_i]} \psi'(s) \le \psi'(\max(u_i, v_i)) \le \psi'(v_i),$$

which gives the desired result. □

THEOREM 3.2. *Let ψ be a potential. The regret of* OSMD *with $F = F_\psi$ and any nonnegative unbiased loss estimate $\tilde{z}_t$ satisfies*

$$R_n \le \frac{\sup_{a \in \mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{i=1}^{d} \mathbb{E} \frac{\tilde{z}_t(i)^2}{(\psi^{-1})'(x_t(i))}.$$

In particular, with the estimate (6), $\psi(x) = (-x)^{-q}$, $q > 1$, and $\eta = \sqrt{(2/(q-1))(m^{1-2/q}/d^{1-2/q})(1/n)}$,

$$R_n \le q \sqrt{\frac{2}{q-1} mdn}.$$

With $q = 2$ this gives

$$R_n \le 2\sqrt{2mdn}.$$

In the case $m = 1$, the above theorem improves the bound $R_n \le 8\sqrt{nd}$ obtained in Audibert and Bubeck [3, Theorem 11].

PROOF. First note that since $\mathcal{D}^* = (-\infty, a)^d$ and $\tilde{z}_t$ has nonnegative coordinates, OSMD is well defined (that is, (2) is satisfied).

The first inequality follows from Theorem 2.2 and the fact that $\psi'(\psi^{-1}(s)) = 1/(\psi^{-1})'(s)$.

Let $\psi(x) = (-x)^{-q}$. Then $\psi^{-1}(x) = -x^{-1/q}$ and $F(x) = -q/(q-1) \sum_{i=1}^{d} x_i^{1-1/q}$. In particular, note that by Hölder's inequality, since $\sum_{i=1}^{d} x_1(i) = m$,

$$F(a) - F(x_1) \le \frac{q}{q-1} \sum_{i=1}^{d} x_1(i)^{1-1/q} \le \frac{q}{q-1} m^{(q-1)/q} d^{1/q}.$$

Moreover, note that $(\psi^{-1})'(x) = (1/q)x^{-1-1/q}$, and

$$\sum_{i=1}^{d} \mathbb{E} \frac{\tilde{z}_t(i)^2}{(\psi^{-1})'(x_t(i))} \le q \sum_{i=1}^{d} x_t(i)^{1/q} \le q m^{1/q} d^{1-1/q},$$

which concludes the proof. □

4. Bandit feedback. In this section we consider online combinatorial optimization with bandit feedback. This setting is much more challenging than the semi-bandit case, and to obtain sublinear regret bounds all known strategies add an *exploration* component to the algorithm. For example, in EXP2, instead of playing an action at random according to the exponentially weighted average distribution p_t, one draws a random action from p_t with probability $1-\gamma$ and from some fixed "exploration" distribution μ with probability γ. On the other hand, in OSMD, one randomly perturbs x_t to some $\tilde{x}_t$, and then plays at random a point in $\mathcal{A}$ such that on average one plays $\tilde{x}_t$.

In Bubeck et al. [9], the authors study the EXP2 strategy with the exploration distribution μ supported on the contact points between the polytope $\mathrm{Conv}(\mathcal{A})$ and the John ellipsoid of this polytope (i.e., the ellipsoid of minimal volume enclosing the polytope). Using this method they are able to prove the best known upper bound for online combinatorial optimization with bandit feedback. They show that the regret of EXP2 mixed with John's exploration (and with the estimate described in Figure 2) satisfies

$$R_n \le 2m^{3/2}\sqrt{3dn\log\frac{ed}{m}}.$$

Our next theorem shows that no strategy can achieve a regret less than a constant times $m\sqrt{dn}$, leaving a gap of a factor of $\sqrt{m\log(d/m)}$. As we argue below, we conjecture that the lower bound is of the correct order of magnitude. However, improving the upper bound seems to require some substantially new ideas. Note that the following bound gives limitations that no strategy can surpass, on the contrary to Theorem 2.1 which was dedicated to the EXP2 strategy.

THEOREM 4.1. *Let $n \ge d \ge 2m$. There exists a subset $\mathcal{A} \subset \{0,1\}^d$ such that $\|a\|_1 = m$, $\forall\, a \in \mathcal{A}$, under bandit feedback, one has*

$$\inf_{\text{strategies}} \sup_{\text{adversaries}} R_n \ge 0.02\, m\sqrt{dn}, \tag{8}$$

where the infimum and the supremum are taken over the class of strategies for the "player" and for the "adversary" as defined in the introduction.

Note that it should not come as a surprise that EXP2 (with John's exploration) is suboptimal, since even in the full information case the basic EXP2 strategy was provably suboptimal; see Theorem 2.1. We conjecture that the correct order of magnitude for the minimax regret in the bandit case is $m\sqrt{dn}$, as the above lower bound suggests.

A promising approach to resolve this conjecture is to consider again the OSMD approach. However, we believe that in the bandit case, one has to consider Legendre functions with nondiagonal Hessian (on the contrary to the Legendre functions considered so far in this paper). Abernethy et al. [1] propose to use a self-concordant barrier function for the polytope $\mathrm{Conv}(\mathcal{A})$. Then they randomly perturb the point x_t given by OSMD using the eigenstructure of the Hessian. This approach leads to a regret upper bound of order $md\sqrt{\theta n\log n}$ for $\theta > 0$ when $\mathrm{Conv}(\mathcal{A})$ admits a θ-self-concordant barrier function. Unfortunately, even when there exists a $O(1)$-self-concordant barrier, this bound is still larger than the conjectured optimal bound by a factor $\sqrt{d}$. In fact, it was proved in Bubeck et al. [9] that in some cases there exist better choices for the Legendre function and the perturbation than those described in Abernethy et al. [1], even when there is a $O(1)$-self-concordant function for the action set. How to generalize this approach to the polytopes involved in online combinatorial optimization is a challenging open problem.

Acknowledgments. G. Lugosi is supported by the Spanish Ministry of Science and Technology [Grant MTM2009-09063] and PASCAL2 Network of Excellence [EC Grant 216886].

Appendix A. Proof of Theorem 2.1. For the sake of simplicity, we assume that d is a multiple of 4 and that n is even. We consider the following subset of the hypercube:

$$\mathcal{A} = \Bigg\{ a \in \{0,1\}^d \colon \sum_{i=1}^{d/2} a_i = d/4 \text{ and } \big(a_i = 1,\ \forall\, i \in \{d/2+1;\dots,d/2+d/4\}\big) \text{ or } \big(a_i = 1,\ \forall\, i \in \{d/2+d/4+1,\dots,d\}\big) \Bigg\}.$$

That is, choosing a point in $\mathcal{A}$ corresponds to choosing a subset of $d/4$ elements among the first half of the coordinates, and choosing one of the two first disjoint intervals of size $d/4$ in the second half of the coordinates.

We prove that for any parameter η, there exists an adversary such that EXP2 (with parameter η) has a regret of at least $nd/16 \tanh(\eta d/8)$, and that there exists another adversary such that its regret is at least $\min((d\log 2)/12\eta, nd/12)$. As a consequence, we have

$$\begin{aligned}\sup R_n &\geq \max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \min\left(\frac{d\log 2}{12\eta}, \frac{nd}{12}\right)\right)\\ &\geq \min\left(\max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \frac{d\log 2}{12\eta}\right), \frac{nd}{12}\right) \geq \min\left(A, \frac{nd}{12}\right),\end{aligned}$$

with

$$\begin{aligned}A &= \min_{\eta\in[0,+\infty)}\max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \frac{d\log 2}{12\eta}\right)\\ &\geq \min\left(\min_{\eta d\geq 8}\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \min_{\eta d<8}\max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \frac{d\log 2}{12\eta}\right)\right)\\ &\geq \min\left(\frac{nd}{16}\tanh(1), \min_{\eta d<8}\max\left(\frac{nd}{16}\frac{\eta d}{8}\tanh(1), \frac{d\log 2}{12\eta}\right)\right)\\ &\geq \min\left(\frac{nd}{16}\tanh(1), \sqrt{\frac{nd^3\log 2\cdot\tanh(1)}{128\cdot 12}}\right) \geq \min\left(0.04\,nd, 0.01\,d^{3/2}\sqrt{n}\right),\end{aligned}$$

where we used the fact that tanh is concave and increasing on $\mathbb{R}_+$. As $n \geq d$, this implies the stated lower bound.

First we prove the lower bound $nd/16 \tanh(\eta d/8)$. Define the following adversary:

$$z_t(i) = \begin{cases} 1 & \text{if } i \in \{d/2+1; \ldots, d/2+d/4\} \text{ and } t \text{ odd},\\ 1 & \text{if } i \in \{d/2+d/4+1, \ldots, d\} \text{ and } t \text{ even},\\ 0 & \text{otherwise.}\end{cases}$$

This adversary always puts a zero loss on the first half of the coordinates, and alternates between a loss of $d/4$ for choosing the first interval (in the second half of the coordinates) and the second interval. At the beginning of odd rounds, any vertex $a \in \mathcal{A}$ has the same cumulative loss and thus EXP2 picks its expert uniformly at random, which yields an expected cumulative loss equal to $nd/16$. On the other hand, at even rounds the probability distribution to select the vertex $a \in \mathcal{A}$ is always the same. More precisely, the probability of selecting a vertex which contains the interval $\{d/2+d/4+1, \ldots, d\}$ (i.e., the interval with a $d/4$ loss at this round) is exactly $1/(1+\exp(-\eta d/4))$. This adds an expected cumulative loss equal to $nd/8(1/(1+\exp(-\eta d/4)))$. Finally, note that the loss of any fixed vertex is $nd/8$. Thus, we obtain

$$R_n = \frac{nd}{16} + \frac{nd}{8}\frac{1}{1+\exp(-\eta d/4)} - \frac{nd}{8} = \frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right).$$

It remains to show a lower bound proportional to $1/\eta$. To this end, we consider a different adversary defined by

$$z_t(i) = \begin{cases} 1-\varepsilon & \text{if } i \leq d/4,\\ 1 & \text{if } i \in \{d/4+1, \ldots, d/2\},\\ 0 & \text{otherwise,}\end{cases}$$

for some fixed $\varepsilon > 0$.

Note that against this adversary the choice of the interval (in the second half of the components) does not matter. Moreover, by symmetry, the weight of any coordinate in $\{d/4+1, \ldots, d/2\}$ is the same (at any round). Finally, note that this weight is decreasing with t. Thus, we have the following identities (in the big sums i represents the number of components selected in the first $d/4$ components):

$$\begin{aligned}R_n &= \frac{n\varepsilon d}{4}\frac{\sum_{a\in\mathcal{A}: a_{d/2}=1}\exp(-\eta n z_1^T a)}{\sum_{a\in\mathcal{A}}\exp(-\eta n z_1^T a)}\\ &= \frac{n\varepsilon d}{4}\frac{\sum_{i=0}^{d/4-1}\binom{d/4}{i}\binom{d/4-1}{d/4-i-1}\exp(-\eta(nd/4-in\varepsilon))}{\sum_{i=0}^{d/4}\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(-\eta(nd/4-in\varepsilon))}\\ &= \frac{n\varepsilon d}{4}\frac{\sum_{i=0}^{d/4-1}\binom{d/4}{i}\binom{d/4-1}{d/4-i-1}\exp(\eta in\varepsilon)}{\sum_{i=0}^{d/4}\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(\eta in\varepsilon)}\\ &= \frac{n\varepsilon d}{4}\frac{\sum_{i=0}^{d/4-1}(1-(4i)/d)\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(\eta in\varepsilon)}{\sum_{i=0}^{d/4}\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(\eta in\varepsilon)}\end{aligned}$$

where we used $\binom{d/4-1}{d/4-i-1} = (1-(4i)/d)\binom{d/4}{d/4-i}$ in the last equality. Thus, taking $\varepsilon = \min(\log 2/(\eta n), 1)$ yields

$$R_n \geq \min\left(\frac{d\log 2}{4\eta}, \frac{nd}{4}\right)\frac{\sum_{i=0}^{d/4-1}(1-(4i)/d)\binom{d/4}{i}^2 \min(2,\exp(\eta n))^i}{\sum_{i=0}^{d/4}\binom{d/4}{i}^2 \min(2,\exp(\eta n))^i} \geq \min\left(\frac{d\log 2}{12\eta}, \frac{nd}{12}\right),$$

where the last inequality follows from Lemma C.1 in the appendix. This concludes the proof of the lower bound.

Appendix B. Proof of Theorem 4.1. The structure of the proof is similar to that of Audibert and Bubeck [3, Theorem 30], which deals with the simple case where $m = 1$. The main important conceptual difference is contained in Lemma C.2, which is at the heart of this new proof. The main argument follows the line of standard lower bounds for bandit problems; see, e.g., Cesa-Bianchi and Lugosi [10]: The worst-case regret is bounded from below by taking an average over a conveniently chosen class of strategies of the adversary. Then, by Pinsker's inequality, the problem is reduced to computing the Kullback-Leibler divergence of certain distributions. The main technical argument, given in Lemma C.2, is for proving manageable bounds for the relevant Kullback-Leibler divergence.

For the sake of simplifying notation, we assume that d is a multiple of m, and we identify $\{0,1\}^d$ with the set of $m \times (d/m)$ binary matrices $\{0,1\}^{m\times d/m}$. We consider the following set of actions:

$$\mathcal{A} = \left\{a \in \{0,1\}^{m\times(d/m)}\colon \forall i \in \{1,\dots,m\}, \sum_{j=1}^{d/m} a(i,j) = 1\right\}.$$

In other words, the player is playing in parallel m finite games with d/m actions.

From Steps 1 to 3 we restrict our attention to the case of deterministic strategies for the player, and we show how to extend the results to arbitrary strategies in Step 4.

Step 1. *Definitions.*

We denote by $I_{i,t} \in \{1,\dots,m\}$ the random variable such that $a_t(i, I_{i,t}) = 1$. That is, $I_{i,t}$ is the action chosen at time t in the ith game. Moreover, let τ be drawn uniformly at random from $\{1,\dots,n\}$.

In this proof we consider random adversaries indexed by $\mathcal{A}$. More precisely, for $\alpha \in \mathcal{A}$, we define the *α-adversary* as follows: For any $t \in \{1,\dots,n\}$, $z_t(i,j)$ is drawn from a Bernoulli distribution with parameter $\frac{1}{2} - \varepsilon\alpha(i,j)$. In other words, against adversary α, in the ith game, the action j such that $\alpha(i,j) = 1$ has a loss slightly smaller (in expectation) than the other actions. We denote by $\mathbb{E}_\alpha$ integration with respect to the loss generation process of the α-adversary. We write $\mathbb{P}_{i,\alpha}$ for the probability distribution of $\alpha(i, I_{i,\tau})$ when the player faces the α-adversary. Note that we have $\mathbb{P}_{i,\alpha}(1) = \mathbb{E}_\alpha(1/n)\sum_{t=1}^n \mathbb{1}_{\alpha(i,I_{i,t})=1}$; hence, against the α-adversary, we have

$$\bar{R}_n = \mathbb{E}_\alpha \sum_{t=1}^n \sum_{i=1}^m \varepsilon \mathbb{1}_{\alpha(i,I_{i,t})\neq 1} = n\varepsilon \sum_{i=1}^m (1 - \mathbb{P}_{i,\alpha}(1)),$$

which implies (since the maximum is larger than the mean)

$$\max_{\alpha\in\mathcal{A}} \bar{R}_n \geq n\varepsilon \sum_{i=1}^m \left(1 - \frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}} \mathbb{P}_{i,\alpha}(1)\right). \tag{B1}$$

Step 2. *Information inequality.*

Let $\mathbb{P}_{-i,\alpha}$ be the probability distribution of $\alpha(i, I_{i,\tau})$ against the adversary which plays like the α-adversary except that in the ith game, the losses of all coordinates are drawn from a Bernoulli distribution of parameter $1/2$. We call it the $(-i,\alpha)$-adversary and we denote by $\mathbb{E}_{(-i,\alpha)}$ integration with respect to its loss generation process. By Pinsker's inequality,

$$\mathbb{P}_{i,\alpha}(1) \leq \mathbb{P}_{-i,\alpha}(1) + \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbb{P}_{-i,\alpha}, \mathbb{P}_{i,\alpha})},$$

where KL denotes the Kullback-Leibler divergence. Moreover, note that by symmetry of the adversaries $(-i,\alpha)$,

$$\begin{aligned}
\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}} \mathbb{P}_{-i,\alpha}(1) &= \frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}} \mathbb{E}_{(-i,\alpha)}\,\alpha(i, I_{i,\tau}) \\
&= \frac{1}{(d/m)^m}\sum_{\beta\in\mathcal{A}}\frac{1}{d/m}\sum_{\alpha\colon(-i,\alpha)=(-i,\beta)} \mathbb{E}_{(-i,\alpha)}\,\alpha(i, I_{i,\tau}) \\
&= \frac{1}{(d/m)^m}\sum_{\beta\in\mathcal{A}}\frac{1}{d/m}\,\mathbb{E}_{(-i,\beta)}\sum_{\alpha\colon(-i,\alpha)=(-i,\beta)} \alpha(i, I_{i,\tau}) \\
&= \frac{1}{(d/m)^m}\sum_{\beta\in\mathcal{A}}\frac{1}{d/m} \\
&= \frac{m}{d},
\end{aligned} \tag{B2}$$

and thus, thanks to the concavity of the square root,

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{i,\alpha}(1)\le\frac{m}{d}+\sqrt{\frac{1}{2(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})}. \tag{B3}$$

Step 3. *Computation of* $\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})$ *with the chain rule.*

Note that since the forecaster is deterministic, the sequence of observed losses (up to time n) $W_n\in\{0,\dots,m\}^n$ uniquely determines the empirical distribution of plays, and, in particular, the probability distribution of $\alpha(i,I_{i,\tau})$ conditionally to W_n is the same for any adversary. Thus, if we denote by $\mathbb{P}^n_\alpha$ (respectively, $\mathbb{P}^n_{-i,\alpha}$) the probability distribution of W_n when the forecaster plays against the α-adversary (respectively, the $(-i,\alpha)$-adversary), then one can easily prove that $\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})\le\mathrm{KL}(\mathbb{P}^n_{-i,\alpha},\mathbb{P}^n_\alpha)$. Now we use the chain rule for Kullback-Leibler divergence iteratively to introduce the probability distributions $\mathbb{P}^t_\alpha$ of the observed losses W_t up to time t. More precisely, we have,

$$\begin{aligned}&\mathrm{KL}(\mathbb{P}^n_{-i,\alpha},\mathbb{P}^n_\alpha)\\&=\mathrm{KL}(\mathbb{P}^1_{-i,\alpha},\mathbb{P}^1_\alpha)+\sum_{t=2}^n\sum_{w_{t-1}\in\{0,\dots,m\}^{t-1}}\mathbb{P}^{t-1}_{-i,\alpha}(w_{t-1})\mathrm{KL}(\mathbb{P}^t_{-i,\alpha}(\cdot\mid w_{t-1}),\mathbb{P}^t_\alpha(\cdot\mid w_{t-1}))\\&=\mathrm{KL}(\mathcal{B}_\varnothing,\mathcal{B}'_\varnothing)\mathbb{1}_{\alpha(i,I_{i,1})=1}+\sum_{t=2}^n\sum_{w_{t-1}:\alpha(i,I_{i,1})=1}\mathbb{P}^{t-1}_{-i,\alpha}(w_{t-1})\mathrm{KL}(\mathcal{B}_{w_{t-1}},\mathcal{B}'_{w_{t-1}}),\end{aligned}$$

where $\mathcal{B}_{w_{t-1}}$ and $\mathcal{B}'_{w_{t-1}}$ are sums of m Bernoulli distributions with parameters in $\{1/2,1/2-\varepsilon\}$ and such that the number of Bernoullis with parameter $1/2$ in $\mathcal{B}_{w_{t-1}}$ is equal to the number of Bernoullis with parameter $1/2$ in $\mathcal{B}'_{w_{t-1}}$ plus one. Now using Lemma C.2 (see below) we obtain,

$$\mathrm{KL}(\mathcal{B}_{w_{t-1}},\mathcal{B}'_{w_{t-1}})\le\frac{8\varepsilon^2}{(1-4\varepsilon^2)m}.$$

In particular, this gives

$$\mathrm{KL}(\mathbb{P}^n_{-i,\alpha},\mathbb{P}^n_\alpha)\le\frac{8\varepsilon^2}{(1-4\varepsilon^2)m}\mathbb{E}_{-i,\alpha}\sum_{t=1}^n\mathbb{1}_{\alpha(i,I_{i,t})=1}=\frac{8\varepsilon^2 n}{(1-4\varepsilon^2)m}\mathbb{P}_{-i,\alpha}(1).$$

Summing and plugging this into (B3) we obtain (again thanks to (B2)), for $\varepsilon\le1/\sqrt{8}$,

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{i,\alpha}(1)\le\frac{m}{d}+\varepsilon\sqrt{\frac{8n}{d}}.$$

To conclude the proof of (8) for deterministic players one needs to plug this last equation in (B1) along with straightforward computations.

Step 4. *Fubini's theorem to handle nondeterministic players.*

Consider now a randomized player, and let $\mathbb{E}_{\mathrm{rand}}$ denote the expectation with respect to the randomization of the player. Then one has (thanks to Fubini's theorem),

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}\sum_{t=1}^n(a_t^Tz_t-\alpha^Tz)=\mathbb{E}_{\mathrm{rand}}\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}_\alpha\sum_{t=1}^n(a_t^Tz_t-\alpha^Tz).$$

Now note that if we fix the realization of the forecaster's randomization, then the results of the previous steps apply and, in particular, one can lower bound $(1/(d/m)^m)\sum_{\alpha\in\mathcal{A}}\mathbb{E}_\alpha\sum_{t=1}^n(a_t^Tz_t-\alpha^Tz)$ as before (note that α is the optimal action in expectation against the α-adversary).

Appendix C. Technical lemmas.

Lemma C.1. *For any* $k\in\mathbb{N}^*$, *for any* $1\le c\le2$, *we have*

$$\frac{\sum_{i=0}^k(1-i/k)\binom{k}{i}^2c^i}{\sum_{i=0}^k\binom{k}{i}^2c^i}\ge1/3.$$

Proof. Let $f(c)$ denote the expression on the left-hand side of the inequality. Introduce the random variable X, which is equal to $i\in\{0,\dots,k\}$ with probability $\binom{k}{i}^2c^i/\sum_{j=0}^k\binom{k}{j}^2c^j$. We have

$$f'(c)=\frac{1}{c}\mathbb{E}\left[X\frac{1-X}{k}\right]-\frac{1}{c}\mathbb{E}(X)\mathbb{E}\left(\frac{1-X}{k}\right)=-\frac{1}{ck}\operatorname{Var}X\le0.$$

So the function f is decreasing on $[1,2]$, and therefore it suffices to consider $c=2$. The numerator and denominator on the left-hand side differ only by the factor $1-i/k$. A lower bound for the left-hand side can thus be obtained by showing that

the terms for i close to k are not essential to the value of the denominator. To prove this, we may use Stirling's formula which implies that for any $k \geq 2$ and $i \in [1, k-1]$,

$$\left(\frac{k}{i}\right)^i \left(\frac{k}{k-i}\right)^{k-i} \frac{\sqrt{k}}{\sqrt{2\pi i(k-i)}} e^{-1/6} < \binom{k}{i} < \left(\frac{k}{i}\right)^i \left(\frac{k}{k-i}\right)^{k-i} \frac{\sqrt{k}}{\sqrt{2\pi i(k-i)}} e^{1/12},$$

hence

$$\left(\frac{k}{i}\right)^{2i} \left(\frac{k}{k-i}\right)^{2(k-i)} \frac{k e^{-1/3}}{2\pi i(k-i)} < \binom{k}{i}^2 < \left(\frac{k}{i}\right)^{2i} \left(\frac{k}{k-i}\right)^{2(k-i)} \frac{k e^{1/6}}{2\pi i}.$$

Introduce $\lambda = i/k$ and $\chi(\lambda) = 2^\lambda / (\lambda^{2\lambda}(1-\lambda)^{2(1-\lambda)})$. We have

$$[\chi(\lambda)]^k \frac{2e^{-1/3}}{\pi k} < \binom{k}{i}^2 2^i < [\chi(\lambda)]^k \frac{e^{1/6}}{2\pi\lambda}. \tag{C1}$$

Lemma C.1 can be numerically verified for $k \leq 10^6$. We now consider $k > 10^6$. For $\lambda \geq 0.666$, since the function χ can be shown to be decreasing on $[0.666, 1]$, the inequality $\binom{k}{i}^2 2^i < [\chi(0.666)]^k (e^{1/6}/2 \times 0.666 \times \pi)$ holds. We have $\chi(0.657)/\chi(0.666) > 1.0002$. Consequently, for $k > 10^6$, we have $[\chi(0.666)]^k < 0.001 \times [\chi(0.657)]^k / k^2$. So, for $\lambda \geq 0.666$ and $k > 10^6$, we have

$$\begin{aligned}
\binom{k}{i}^2 2^i < 0.001 \times [\chi(0.657)]^k \frac{e^{1/6}}{2\pi \times 0.666 \times k^2} &< [\chi(0.657)]^k \frac{2e^{-1/3}}{1{,}000\,\pi k^2} \\
&= \min_{\lambda \in [0.656,\, 0.657]} [\chi(\lambda)]^k \frac{2e^{-1/3}}{1{,}000\,\pi k^2} \\
&< \frac{1}{1{,}000\,k} \max_{i \in \{1,\ldots,k-1\} \cap [0,\, 0.666k)} \binom{k}{i}^2 2^i,
\end{aligned} \tag{C2}$$

where the last inequality comes from (C1) and the fact that there exists $i \in \{1, \ldots, k-1\}$ such that $i/k \in [0.656, 0.657]$. Inequality (C2) implies that for any $i \in \{1, \ldots, k\}$, we have

$$\sum_{0.666k \leq i \leq k} \binom{k}{i}^2 2^i < \frac{1}{1{,}000} \max_{i \in \{1,\ldots,k-1\} \cap [0,\, 0.666k)} \binom{k}{i}^2 2^i < \frac{1}{1{,}000} \sum_{0 \leq i < 0.666k} \binom{k}{i}^2 2^i.$$

To conclude, introducing $A = \sum_{0 \leq i < 0.666k} \binom{k}{i}^2 2^i$, we have

$$\frac{\sum_{i=0}^k (1 - i/k) \binom{k}{i}^2 2^i}{\sum_{i=0}^k \binom{k}{i} \binom{k}{k-i} 2^i} > \frac{(1 - 0.666)\, A}{A + 0.001\, A} \geq \frac{1}{3}. \quad \square$$

Lemma C.2. *Let l and n be integers with $\frac{1}{2} \leq n/2 \leq l \leq n$. Let $p, p', q, p_1, \ldots, p_n$ be real numbers in $(0,1)$ with $q \in \{p, p'\}$, $p_1 = \cdots = p_l = q$ and $p_{l+1} = \cdots = p_n$. Let $\mathcal{B}$ (resp. $\mathcal{B}'$) be the sum of $n+1$ independent Bernoulli distributions with parameters $p, p_1, \ldots, p_n$ (resp. $p', p_1, \ldots, p_n$). We have*

$$\mathrm{KL}(\mathcal{B}, \mathcal{B}') \leq \frac{2(p' - p)^2}{(1 - p')(n+2)q}.$$

Proof. Let $Z, Z', Z_1, \ldots, Z_n$ be independent Bernoulli distributions with parameters $p, p', p_1, \ldots, p_n$. Define $S = \sum_{i=1}^l Z_i$, $T = \sum_{i=l+1}^n Z_i$, and $V = Z + S$. By a slight and usual abuse of notation, we use KL to denote Kullback-Leibler divergence of both probability distributions and random variables. Then we may write (the inequality is an easy consequence of the chain rule for Kullback-Leibler divergence):

$$\begin{aligned}
\mathrm{KL}(\mathcal{B}, \mathcal{B}') &= \mathrm{KL}\big((Z+S)+T, (Z'+S)+T\big) \\
&\leq \mathrm{KL}\big((Z+S, T), (Z'+S, T)\big) \\
&= \mathrm{KL}(Z+S, Z'+S).
\end{aligned}$$

Let $s_k = \mathbb{P}(S = k)$ for $k = -1, 0, \ldots, l+1$. Using the equalities

$$s_k = \binom{l}{k} q^k (1-q)^{l-k} = \frac{q}{1-q} \frac{l-k+1}{k} \binom{l}{k-1} q^{k-1} (1-q)^{l-k+1} = \frac{q}{1-q} \frac{l-k+1}{k} s_{k-1},$$

which hold for $1 \le k \le l+1$, we obtain

$$\begin{aligned}
\mathrm{KL}(Z+S, Z'+S) &= \sum_{k=0}^{l+1} \mathbb{P}(V=k) \log\left(\frac{\mathbb{P}(Z+S=k)}{\mathbb{P}(Z'+S=k)}\right) \\
&= \sum_{k=0}^{l+1} \mathbb{P}(V=k) \log\left(\frac{ps_{k-1}+(1-p)s_k}{p's_{k-1}+(1-p')s_k}\right) \\
&= \sum_{k=0}^{l+1} \mathbb{P}(V=k) \log\left(\frac{p((1-q)/q)k+(1-p)(l-k+1)}{p'((1-q)/q)k+(1-p')(l-k+1)}\right) \\
&= \mathbb{E}\log\left(\frac{(p-q)V+(1-p)q(l+1)}{(p'-q)V+(1-p')q(l+1)}\right). \qquad \text{(C3)}
\end{aligned}$$

Case 1. $q = p'$.

By Jensen's inequality, using that $\mathbb{E}\,V = p'(l+1)+p-p'$ in this case, we get

$$\begin{aligned}
\mathrm{KL}(Z+S, Z'+S) &\le \log\left(\frac{(p-p')\,\mathbb{E}(V)+(1-p)p'(l+1)}{(1-p')p'(l+1)}\right) \\
&= \log\left(\frac{(p-p')^2+(1-p')p'(l+1)}{(1-p')p'(l+1)}\right) \\
&= \log\left(1+\frac{(p-p')^2}{(1-p')p'(l+1)}\right) \le \frac{(p-p')^2}{(1-p')p'(l+1)}.
\end{aligned}$$

Case 2. $q = p$.

In this case, V is a binomial distribution with parameters $l+1$ and p. From (C3), we have

$$\begin{aligned}
\mathrm{KL}(Z+S, Z'+S) &\le -\mathbb{E}\log\left(\frac{(p'-p)V+(1-p')p(l+1)}{(1-p)p(l+1)}\right) \\
&\le -\mathbb{E}\log\left(1+\frac{(p'-p)(V-\mathbb{E}\,V)}{(1-p)p(l+1)}\right). \qquad \text{(C4)}
\end{aligned}$$

To conclude, we will use the following lemma.

Lemma C.3. *The following inequality holds for any $x \ge x_0$ with $x_0 \in (0,1)$:*

$$-\log(x) \le -(x-1)+\frac{(x-1)^2}{2x_0}.$$

Proof. Introduce $f(x) = -(x-1)+(x-1)^2/(2x_0)+\log(x)$. We have $f'(x) = -1+(x-1)/x_0+1/x$, and $f''(x) = 1/x_0-1/x^2$. From $f'(x_0)=0$, we get that f' is negative on $(x_0,1)$ and positive on $(1,+\infty)$. This leads to f nonnegative on $[x_0,+\infty)$. □

Finally, from Lemma C.3 and (C4), using $x_0 = (1-p')/(1-p)$, we obtain

$$\begin{aligned}
\mathrm{KL}(Z+S, Z'+S) &\le \left(\frac{p'-p}{(1-p)p(l+1)}\right)^2 \frac{\mathbb{E}[(V-\mathbb{E}\,V)^2]}{2x_0} \\
&= \left(\frac{p'-p}{(1-p)p(l+1)}\right)^2 \frac{(l+1)p(1-p)^2}{2(1-p')} \\
&= \frac{(p'-p)^2}{2(1-p')(l+1)p}. \quad \square
\end{aligned}$$

References

[1] Abernethy J, Hazan E, Rakhlin A (2008) Competing in the dark: An efficient algorithm for bandit linear optimization. *Proc. 21st Annual Conf. Learn. Theory (COLT)* (Omnipress, Madison, WI), 263–274.

[2] Audibert J-Y, Bubeck S (2009) Minimax policies for adversarial and stochastic bandits. *Proc. 22nd Annual Conf. Learn. Theory (COLT)* (Omnipress, Madison, WI), 773–818.

[3] Audibert J-Y, Bubeck S (2010) Regret bounds and minimax policies under partial monitoring. *J. Machine Learn. Res.* 11(October): 2635–2686.

[4] Auer P, Cesa-Bianchi N, Freund Y, Schapire R (2003) The nonstochastic multi-armed bandit problem. *SIAM J. Comput.* 32(1):48–77.

[5] Awerbuch B, Kleinberg R (2004) Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. *STOC '04: Proc. Thirty-Sixth Annual ACM Sympos. Theory Comput.* (ACM, New York), 45–53.

[6] Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* 31(3):167–175.

[7] Bubeck S (2011) Introduction to online optimization. *Lecture Notes*, Princeton University.
[8] Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.
[9] Bubeck S, Cesa-Bianchi N, Kakade SM (2012) Towards minimax policies for online linear optimization with bandit feedback. *JMLR Workshop Conf. Proc.* (*COLT*) Vol. 23, 41.1–41.14.
[10] Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, New York).
[11] Cesa-Bianchi N, Lugosi G (2012) Combinatorial bandits. *J. Comput. System Sci.* 78(5):1404–1422.
[12] Dani V, Hayes T, Kakade S (2008) The price of bandit information for online optimization. *Adv. Neural Inform. Processing Systems* (*NIPS*) 20:345–352.
[13] Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55:119–139.
[14] Gentile C, Warmuth M (1998) Linear hinge loss and average margin. *Proc. Adv. Neural Inform. Processing Systems (NIPS)* (MIT Press, Cambridge, MA), 225–231.
[15] Grove A, Littlestone N, Schuurmans D (2001) General convergence results for linear discriminant updates. *Machine Learn.* 43(3):173–210.
[16] György A, Linder T, Lugosi G, Ottucsák G (2007) The on-line shortest path problem under partial monitoring. *J. Machine Learn. Res.* 8(October):2369–2403.
[17] Hazan E (2011) The convex optimization approach to regret minimization. Sra S, Nowozin S, Wright S, eds. *Optimization for Machine Learning* (MIT Press, Cambridge, MA), 287–303.
[18] Hazan E, Kale S, Warmuth M (2010) Learning rotations with little regret. *Proc. 23rd Annual Conf. Learn. Theory (COLT)* (Omnipress, Madison, WI), 144–154.
[19] Helmbold DP, Warmuth M (2009) Learning permutations with exponential weights. *J. Machine Learn. Res.* 10:1705–1736.
[20] Herbster M, Warmuth M (1998) Tracking the best expert. *Machine Learn.* 32:151–178.
[21] Hiriart-Urruty J-B, Lemaréchal C (2001) *Fundamentals of Convex Analysis* (Springer, Berlin).
[22] Kalai A, Vempala S (2005) Efficient algorithms for online decision problems. *J. Comput. System Sci.* 71:291–307.
[23] Kale S, Reyzin L, Schapire R (2010) Nonstochastic bandit slate problems. *Adv. Neural Inform. Processing Systems* (*NIPS*) (Currant Associates, Vancouver, BC), 1054–1062.
[24] Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23:462–466.
[25] Kivinen J, Warmuth M (2001) Relative loss bounds for multidimensional regression problems. *Machine Learn.* 45:301–329.
[26] Koolen W, Warmuth M, Kivinen J (2010) Hedging structured concepts. *Proc.* 23*rd Annual Conf. Learn. Theory* (*COLT*) (Omnipress, Madison, WI), 93–105.
[27] McMahan H, Blum A (2004) Online geometric optimization in the bandit setting against an adaptive adversary. *Proc. 17th Annual Conf. Learn. Theory (COLT)* (Omnipress, Madison, WI), 109–123.
[28] Nemirovski A (1979) Efficient methods for large-scale convex optimization problems. *Ekonomika I Matematicheskie Metody* 15. [In Russian.]
[29] Nemirovski A, Yudin D (1983) *Problem Complexity and Method Efficiency in Optimization* (Wiley Interscience, New York).
[30] Rakhlin A (2009) Lecture notes on online learning, http://www.stat.wharton.upenn.edu/~rakhlin/courses/stat928/stat928_notes.pdf.
[31] Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22:400–407.
[32] Schrijver A (2003) *Combinatorial Optimization* (Springer, Berlin).
[33] Shalev-Shwartz S (2007) Online learning: Theory, algorithms, and applications. Ph.D. thesis, The Hebrew University of Jerusalem.
[34] Takimoto E, Warmuth M (2003) Paths kernels and multiplicative updates. *J. Machine Learn. Res.* 4(October):773–818.
[35] Uchiya T, Nakamura A, Kudo M (2010) Algorithms for adversarial bandit problems with multiple plays. *Proc. 21st Internat. Conf. Algorithmic Learn. Theory (ALT)* (Springer, Berlin), 375–389.
[36] Warmuth M, Kuzmin D (2008) Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *J. Machine Learn. Res.* 9(October):2287–2320.
[37] Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. *Proc. Twentieth Internat. Conf. Machine Learn.* (*ICML*) (AAAI Press, Cambridge, MA), 928–936.