MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxxx 20xx, pp. xxx–xxx
ISSN 0364-765X | EISSN 1526-5471 |xx|0000|0xxx

DOI 10.1287/moor.xxxx.xxxx

Regret in Online Combinatorial Optimization

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

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

Gábor Lugosi
ICREA and Pompeu Fabra University, Barcelona, Spain
email: 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 best 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. Combining the Mirror Descent algorithm and the INF (Implicitely Normalized Forecaster) strategy, we are able to prove optimal bounds for the semi-bandit case. We also recover the optimal bounds for the full information setting. 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. Finally we also prove that the standard exponentially weighted average forecaster is provably suboptimal in the setting of online combinatorial optimization.

Key words: online optimization; combinatorial optimization; mirror descent; multi-armed bandits, minimax regret

MSC2000 Subject Classification: Primary: 99999, 88888; Secondary: 88888, 77777 (*See the MSC2000 codes at* http://www.ams.org/msc/)

OR/MS subject classification: Primary: xxxx, yyyy; Secondary: zzzz (*See the OR/MS classification at* http://mor.pubs.informs.org/ORSubject.pdf)

History: Received: Xxxx xx, xxxx; Revised: Yyyyyy yy, yyyy and Zzzzzz zz, zzzz.

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 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

[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.

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 $a_t \in \mathcal{A}$ with the help of an external randomization;

(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 his cumulative loss $\sum_{t=1}^n a_t^T z_t$.

Figure 1: Online Combinatorial Optimization.

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 $(a_t(1)z_t(1), \ldots, a_t(d)z_t(d))$. All three variants of online combinatorial optimization are sketched in Figure 1.

1.1 Motivating examples. A great number of specific 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 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, so that, 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.
- **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 webpage. 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 communication problems), or m-intervals.

1.2 Previous work.

[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.

- **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, Warmuth, and Kivinen [25]. Previous papers under full information feedback include Kivinen and Warmuth [24], Grove, Littlestone, and Schuurmans [13], Takimoto and Warmuth [33], Kalai and Vempala [21], Warmuth and Kuzmin [35], Herbster and Warmuth [18], and Hazan, Kale, and Warmuth [17].
- **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, Cesa-Bianchi, Freund, and Schapire [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, Linder, Lugosi, and Ottucsák[14] consider the online shortest path problem and derive suboptimal regret bounds (in terms of the dependency on m and d). Uchiya, Nakamura, and Kudo [34] (respectively Kale, Reyzin, and Schapire [22]) 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 [26], 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, Hayes, and Kakade [11]. The dependency on m and d was then improved in various ways by Abernethy, Hazan, and Rakhlin [1], Cesa-Bianchi and Lugosi [10], and Bubeck, Cesa-Bianchi and Kakade [8]. We discuss these bounds in detail in Section 5. In particular, we argue that the optimal regret bound in terms of d (and m) is still an open problem.

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. Then we are interested in

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

We prove upper and lower bounds for the minimax regret under the different feedback assumptions. The upper bounds are obtained by constructing prediction strategies. We also discuss the computational complexity of these strategies.

Our contribution is threefold. First, we unify the algorithms used in Abernethy, Hazan, and Rakhlin [1], Koolen, Warmuth, and Kivinen[25], Uchiya, Nakamura, and Kudo [34], and Kale, Reyzin, and Schapire[22] under the umbrella of mirror descent. The idea of mirror descent goes back to Nemirovski [27], Nemirovski and Yudin[28]. A somewhat similar concept was re-discovered in online learning by Herbster and Warmuth [19], Grove, Littlestone, and Schuurmans [13], Kivinen and Warmuth [24] under the name of potential-based gradient descent, see [9, Chapter 11]. Recently, these ideas have been flourishing, see for instance Shalev-Schwartz [32], Rakhlin [29], Hazan [16], 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 we prove several lower bounds. First, we establish lower bounds for the minimax regret under the three feedback assumptions (the main difficulty being the bandit case). Moreover, we also answer a question of Koolen, Warmuth, and Kivinen[25] by showing that the exponentially weighted average forecaster is provably suboptimal for online combinatorial optimization. A summary of the bounds proved in this paper can be found in Tables 1 and 2.

The paper is organized as follows. In Section 2 we introduce the three algorithms discussed in this paper. In particular, in Section 2.2 we describe OSMD (Online Stochastic Mirror Descent) and prove a general regret bound in terms of the Bregman divergence of the Fenchel-Legendre dual of the regularizer. Then in Sections 3 and 4, we derive upper bounds for the regret in the full information case and in the semi-bandit case for OSMD with appropriately chosen regularizers. In Section 5 we discuss the regret bounds obtained in [1, 8] for the bandit version, and formulate a conjecture on the correct order of magnitude of the regret for that problem. We end the paper with the lower bounds in Section 6.

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

Table 1: Bounds on the minimax regret proved in this paper (up to constant factors). The new results are set in boldface.

	Full Information	Semi-Bandit	Bandit
EXP2	$m^{3/2}\sqrt{n \log \frac{d}{m}}$	$\mathbf{m}\sqrt{\mathbf{dn} \log \frac{\mathbf{d}}{\mathbf{m}}}$	$m^{3/2}\sqrt{dn \log \frac{d}{m}}$
FPL	$m\sqrt{dn}$	-	-
OSMD	$m\sqrt{n \log \frac{d}{m}}$	$\mathbf{\sqrt{mdn}}$	$md^{3/2}\sqrt{n \log n}$

Table 2: Upper bounds on R_n for specific algorithms. The new results are in boldface. We also show that the bound for EXP2 in the full information setting is not improvable.

2. Algorithms. Three classes of algorithms have been proposed for online combinatorial optimization. In this section we review them, discuss their computational complexity, and prove general regret bounds that will be useful to derive bounds under specific feedback assumptions.

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., [9]). (This strategy is sometimes called Hedge, see Freund and Schapire [12].) 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, Warmuth, and Kivinen [25]. In the semi-bandit case, EXP2 was studied by György, Linder, Lugosi, and Ottucsák [14] while in the bandit case in Dani, Hayes, and Kakade [11], Cesa-Bianchi and Lugosi [10], and Bubeck, Cesa-Bianchi and Kakade [8]. Note that in the bandit case, EXP2 is mixed with an *exploration distribution*, see Section 5 for more details.

The following theorem shows the regret bound for EXP2 that one may obtain by a standard argument, as, for example, in [10].

THEOREM 2.1 *If the loss estimate is unbiased in the sense that, for each $t = 1, \dots, n$, $\mathbb{E}_{a_t \sim p_t} \widetilde{z}_t = z_t$, then the regret of the* EXP2 *strategy satisfies*

$$R_n \leq \frac{\log(|\mathcal{A}|)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\big[p_t(a)(a^T \widetilde{z}_t)^2 \max\big(1, \exp(-\eta a^T \widetilde{z}_t)\big)\big].$$

In its straightforward implementation, when one calculates $p_t(a)$ separately for each action $a \in \mathcal{A}$, EXP2 is clearly inefficient, unless $\mathcal{A}$ has only polynomially many elements. However, for some specific (non-trivial) sets $\mathcal{A}$, efficient implementation (i.e., of complexity polynomial in d) of EXP2 is possible. We refer to Koolen, Warmuth, and Kivinen [25] and Cesa-Bianchi and Lugosi [10] for examples.

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 [28]. 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 in Robbins and Monro [30], Kiefer and Wolfowitz [23].

EXP2*:*

Parameter: Learning rate η.

Let $p_1 = \left(\frac{1}{|\mathcal{A}|}, \ldots, \frac{1}{|\mathcal{A}|}\right) \in \mathbb{R}^{|\mathcal{A}|}$.

For each round $t = 1, 2, \ldots, n$;

(a) 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 $\widetilde{z}_t$. For instance, one may take

- $\widetilde{z}_t = z_t$ in the full information game,
- $\widetilde{z}_t(i) = \frac{z_t(i)}{\sum_{a \in \mathcal{A}: a(i)=1} p_t(a)} a_t(i)$ in the semi-bandit game,
- $\widetilde{z}_t = P_t^+ a_t a_t^T z_t$, with $P_t = \mathbb{E}_{a \sim p_t}(aa^T)$ in the bandit game.

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

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

Figure 2: The EXP2 strategy.

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 [36]. In this latter paper the Online Gradient Descent (OGD) algorithm was derived, which is a version of gradient descent tailored to online optimization.

In the full information setting, algorithms of this type were studied by Abernethy, Hazan, and Rakhlin [1], Rakhlin [29], and Hazan [16]. 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 [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 Section 5 for more details). Another line of work studied this type of algorithms with F being the negative entropy, see Koolen, Warmuth, and Kivinen [25] for the full information case and Uchiya, Nakamura, and Kudo [34], Kale, Reyzin, and Schapire [22] for specific instances of the semi-bandit case. All these results are unified and described in details in Bubeck [7]. In this paper we consider a new type of Legendre functions F inspired by Audibert and Bubeck [3], see Section 4.

To properly describe the OSMD strategy, we recall a few concepts from convex analysis, see Hiriart-Urruty and Lemaréchal [20] for a thorough treatment of this subject. Let $\mathcal{D} \subset \mathbb{R}^d$ be an open convex set, and $\overline{\mathcal{D}}$ the closure of $\mathcal{D}$.

DEFINITION 2.1 *We call Legendre any continuous function $F : \overline{\mathcal{D}} \to \mathbb{R}$ such that*

(i) F is strictly convex continuously differentiable on $\mathcal{D}$,

(ii) $\lim_{x \to \overline{\mathcal{D}} \setminus \mathcal{D}} ||\nabla F(x)|| = +\infty$.[3]

The Bregman divergence $D_F : \overline{\mathcal{D}} \times \mathcal{D}$ 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

$$F^*(u) = \sup_{x \in \overline{\mathcal{D}}} \left(x^T u - f(x)\right) .$$

[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 $\overline{\mathcal{D}} \supset Conv(\mathcal{A})$.

Let $x_1 \in \operatorname{argmin}_{x \in Conv(\mathcal{A})} F(x)$.

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 $\widetilde{z}_t$.

(d) Let $w_{t+1} \in \mathcal{D}$ satisfy

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

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

$$x_{t+1} \in \operatorname*{argmin}_{x \in Conv(\mathcal{A})} D_F(x, w_{t+1}). \tag{4}$$

Figure 3: Online Stochastic Mirror Descent (OSMD).

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

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

The lemma above is the key to understanding how a Legendre function acts on the space. ∇F maps $\mathcal{D}$ to the dual space $\mathcal{D}^*$, and ∇F^* is the inverse mapping to go 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 [Chapter 11, [9]].

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:

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

Regarding computational complexity, OSMD is efficient as soon as the polytope $Conv(\mathcal{A})$ can be described by a polynomial 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 [31]. 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, Warmuth, and Kivinen [25].

The following result is at the basis of all our upper bounds for the regret of OSMD.

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

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

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

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

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

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

By the Pythagorean theorem for Bregman divergences (see, e.g., Lemma 11.3 of [9]), we have $D_F(a, w_{t+1}) \geq D_F(a, x_{t+1}) + D_F(x_{t+1}, w_{t+1})$, hence

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

Summing over t then gives

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

By the nonnegativity of the Bregman divergences, we get

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

From (1), one has $D_F(x_t, w_{t+1}) = D_{F^*}\big(\nabla F(x_t) - \eta\nabla\ell(x_t, z_t), \nabla F(x_t)\big)$. 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 an Hessian, denoted $\nabla^2 F$, that is always invertible, then one can prove that, up to a third-order term $\big(\text{in } \widetilde{z}_t\big)$, the regret bound can be written as

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

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

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.

2.3 Follow the Perturbated Leader (FPL). The last strategy that we consider, called Follow the Perturbated Leader (FPL), was proposed by Hannan [15], see also Kalai and Vempala [21]. The idea is simple: it is clear that following the leader, that is, choosing at time t,

$$\operatorname*{argmin}_{a\in\mathcal{A}}\sum_{s=1}^{t-1} z_s^T a$$

is a strategy that can be hazardous. In FPL, this choice is *regularized* by adding a small amount of noise. More precisely, let $\xi_1, \ldots, \xi_n$ be an i.i.d. sequence of random variables uniformly distributed on $[0, 1/\eta]^d$. Then FPL picks the action

$$\operatorname*{argmin}_{a\in\mathcal{A}}\left(\xi_t + \sum_{s=1}^{t-1} z_s\right)^T a.$$

This strategy may be generalized to the partial information models (semi-bandit and bandit) by replacing the losses in the definition by their unbiased estimates. Unfortunately however, such a forecaster does not seem to have a good performance. Moreover, as we will see below, even for the full-information case, the known bounds for FPL are suboptimal by a factor $\sqrt{d}$. Nonetheless, FPL has the advantage that it is computationally efficient as soon as there exist efficient algorithms for the offline problem (that is, the problem of linear optimization over $\mathcal{A}$). This is an important property, and an interesting open problem is to decide whether there exists a strategy with this property and optimal regret bounds.

The following regret bound was proved by Kalai and Vempala [21]. We slightly improve the constant (by following the same proof technique as in [21]). We perform the analysis in a restrictive framework, namely we only consider oblivious adversaries (i.e., the loss sequence (z_t) is fixed and z_t cannot depend on the past moves $a_1, \ldots, a_{t-1}$ of the player). The details of the proof are given in the Appendix.

THEOREM 2.3 *For any oblivious adversary, the regret of the FPL strategy satisfies*

$$R_n \leq \frac{m}{2\eta} + \eta m d n \ .$$

In particular, with $\eta = \sqrt{\frac{1}{2dn}}$*, one obtains*

$$R_n \leq m\sqrt{2dn} \ .$$

3. Full Information. In this section we consider online combinatorial optimization with full information feedback. First we analyze the regret of the EXP2 strategy in the full information setting.

THEOREM 3.1 *The regret of the* EXP2 *strategy satisfies*

$$R_n \leq \frac{m \log \frac{ed}{m}}{\eta} + \frac{\eta}{2} n m^2.$$

In particular, with $\eta = \sqrt{\frac{2\log\left(\frac{ed}{m}\right)}{nm}}$*,*

$$R_n \leq m^{3/2}\sqrt{2n \log\left(\frac{ed}{m}\right)}.$$

PROOF. Apply Theorem 2.1 by noting that $\log|\mathcal{A}| \leq \log\binom{d}{m} \leq m \log\left(\frac{ed}{m}\right)$, and $z_t^T a \in [0, m]$. □

Perhaps surprisingly, there is a gap between this upper bound and the minimax lower bound proved below in Theorem 6.2. It is natural to ask whether one can improve the analysis of EXP2. This question was posed by Koolen, Warmuth, and Kivinen [25]. In Theorem 6.1 we give a negative answer, that is, we show that the upper bound of Theorem 3.1 cannot be improved substantially. As we also show that the minimax lower bound can be achieved (up to a constant factor), this proves that the popular EXP2 strategy is suboptimal.

The key to obtaining optimal regret bounds in online combinatorial optimization is to use the OSMD strategy, which gives the flexibility to adapt to the geometry of the action set $\mathcal{A}$. The following theorem shows that the negative entropy is a good choice of the Legendre function.

THEOREM 3.2 *(Koolen, Warmuth, Kivinen [25].) The regret of* OSMD *with* $F(x) = \sum_{i=1}^d x_i \log x_i - x_i$ *(and* $\mathcal{D} = (0, +\infty)^d$*) satisfies*

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

In particular, with $\eta = \sqrt{\frac{2\log\left(\frac{d}{m}\right)}{nm}}$*,*

$$R_n \leq m\sqrt{2n \log\left(\frac{d}{m}\right)}.$$

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

$$D_{F^*}\Big(\nabla F(x), \nabla F(y)\Big) = \sum_{i=1}^d y(i)\ \Theta\Big((\nabla F(x) - \nabla F(y))(i)\Big)\ , \tag{6}$$

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

$$\begin{aligned} R_n &\leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^n \mathbb{E} D_{F^*}\left(\nabla F(x_t) - \eta\widetilde{z}_t, \nabla F(x_t)\right) \\ &\leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^n x_t(i) z_t(i)^2 \qquad (7) \\ &\leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(x_1)}{\eta} + \frac{\eta}{2} nm. \end{aligned}$$

The proof is concluded by noting that:

$$F(a) - F(x_1) \leq \sum_{i=1}^d x_1(i) \log \frac{1}{x_1(i)} \leq m \log\left(\sum_{i=1}^d \frac{x_1(i)}{m}\frac{1}{x_1(i)}\right) = m \log\frac{d}{m}. \qquad (8)$$

□

4. Semi-bandit feedback. In this section we consider online combinatorial optimization with semi-bandit feedback. As we saw in the full information case, the key to obtaining optimal regret bounds is the OSMD strategy. First we analyze the behavior of OSMD with the negative entropy (which is an optimal strategy under full information feedback), and with the following estimate for the loss vector:

$$\widetilde{z}_t(i) = \frac{z_t(i) a_t(i)}{x_t(i)}. \qquad (9)$$

Note that this is a valid estimate since it makes only 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, $\mathbb{E}_{a_t\sim p_t}\widetilde{z}_t(i) = z_t(i)$.

THEOREM 4.1 *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 non-negative unbiased loss estimate $\widetilde{z}_t(i) \geq 0$ satisfies*

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

In particular, with the estimate (9) *and $\eta = \sqrt{2\frac{m\log dm}{nd}}$,*

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

PROOF. The first inequality directly follows from (7) and (8). The second inequality follows from

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

□

As the lower bound of Theorem 6.2 shows, this upper bound has an extra logarithmic factor. This phenomenon already appeared in the basic multi-armed bandit setting (when $\mathcal{A}$ corresponds to the canonical basis). 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 4.1 *Let $\omega \geq 0$. A function $\psi : (-\infty, a) \to \mathbb{R}_+^*$ for some $a \in \mathbb{R}\cup\{+\infty\}$ is called an ω-potential if it is convex, continuously differentiable, and satisfies*

$$\lim_{x\to-\infty}\psi(x) = \omega \qquad \lim_{x\to a}\psi(x) = +\infty$$

$$\psi' > 0 \qquad \int_\omega^{\omega+1} |\psi^{-1}(s)| ds < +\infty.$$

To a 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 non-zero 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 [3], the choice of $\psi(x) = (-x)^q$ with $q > 1$ was recommended. We show below that here, again, this choice gives a minimax optimal strategy.

LEMMA 4.1 *Let ψ be a potential. Then F_ψ is Legendre and for all $u, v \in \mathcal{D}^* = (-\infty, a)^d$ such that $u_i \leq v_i, \forall i \in \{1, \ldots, d\}$,*

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

PROOF. It is easy to check that F is a Legendre function. Moreover, since $\nabla F^*(u) = (\nabla F)^{-1}(u) = \big(\psi(u_1), \ldots, \psi(u_d)\big)$, we obtain

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

From a Taylor expansion, we get

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

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

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

which gives the desired result. □

THEOREM 4.2 *Let ψ be a potential. The regret of* OSMD *with F_ψ and any non-negative unbiased loss estimate $\widetilde{z}_t$ satisfies*

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

In particular, with the estimate (9), *$\psi(x) = (-x)^{-q}$, $q > 1$, and $\eta = \sqrt{\frac{2}{q-1} \frac{m^{1-2/q}}{d^{1-2/q}}}$,*

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

With $q = 2$ this gives

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

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

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

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

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

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

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

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

which concludes the proof. □

For sake of completeness, we end this section with a regret bound for the EXP2 strategy in the semi-bandit setting.

THEOREM 4.3 *The regret of* EXP2 *satisfies*

$$R_n \leq \frac{m\log\frac{ed}{m}}{\eta} + \frac{\eta n m d}{2}.$$

In particular, for $\eta = \sqrt{\frac{2\log\frac{ed}{m}}{n}}$, *we have*

$$R_n \leq m\sqrt{2nd\log\frac{ed}{m}}.$$

PROOF. The proof follows from Theorem 2.1 and the following: let $x_t = \mathbb{E}_{a\sim p_t} a = \sum_{a\in\mathcal{A}} p_t(a)a$. In particular, we have $\widetilde{z}_t(i) = \frac{z_t(i)a_t(i)}{x_t(i)}$, and:

$$\begin{aligned}
\mathbb{E}_{a_t\sim p_t}\sum_{a\in\mathcal{A}} p_t(a)(a^T\widetilde{z}_t)^2 &= \mathbb{E}_{a_t\sim p_t, a_t'\sim p_t}\sum_{i,j}\frac{z_t(i)a_t(i)a_t'(i)}{x_t(i)}\frac{z_t(j)a_t(j)a_t'(j)}{x_t(j)} \\
&\leq \mathbb{E}_{a_t,a_t'}\sum_{i,j} a_t(i)\frac{a_t(j)}{x_t(j)}\frac{a_t'(i)}{x_t(i)} \\
&= m\mathbb{E}_{a_t}\sum_j \frac{a_t(j)}{x_t(j)} \\
&= md.
\end{aligned}$$

□

5. 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 $\widetilde{x}_t$, and then plays at random a point in $\mathcal{A}$ such that on average one plays $\widetilde{x}_t$.

In Bubeck, Cesa-Bianchi and Kakade [8], the authors study the EXP2 strategy with the exploration distribution μ supported on the contact points between the polytope $Conv(\mathcal{A})$ and the John's 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 the John's exploration (and with the estimate described in Figure 2) satisfies

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

This regret is off by a factor $\sqrt{m\log\frac{d}{m}}$ from the minimax lower bound described in Section 6. However this may not come as a surprise, since even in the full information case the EXP2 strategy is provably

suboptimal, see Section 6. We conjecture that the correct order of magnitude for the minimax regret in the bandit case is $m\sqrt{dn}$, as Theorem 6.2 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 non-diagonal Hessian (on the contrary to the Legendre functions considered so far in this paper). Abernethy, Hazan, and Rakhlin [1] propose to use a self-concordant barrier function for the polytope $Conv(\mathcal{A})$. 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 $Conv(\mathcal{A})$ 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 [8] that in some cases there exist better choices for the Legendre function and the perturbation than those described in [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 an interesting open problem.

6. Lower Bounds. In this section we offer various lower bounds. We start this with a result that shows that the EXP2 strategy is suboptimal for online combinatorial optimization, answering a question of Koolen, Warmuth, and Kivinen [25]. Then we turn to *minimax* lower bounds that show limitations that no strategy can surpass.

THEOREM 6.1 *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*

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

PROOF. For the sake of simplicity, we assume here that d is a multiple of 4 and that n is even. We consider the following subset of the hypercube:

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

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 $\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right)$, and that there exists another adversary such that its regret is at least $\min\left(\frac{d\log 2}{12\eta}, \frac{nd}{12}\right)$. 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

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

As $n \geq d$, this implies the stated lower bound.

First we prove the lower bound $\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right)$. Define the following adversary:

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

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 $\frac{1}{1+\exp(-\eta d/4)}$. This adds an expected cumulative loss equal to $\frac{nd}{8}\frac{1}{1+\exp(-\eta d/4)}$. 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

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

where the value of $\varepsilon > 0$ is specified below.

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):

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

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

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

where the last inequality follows from Lemma B.1 below. This concludes the proof of the lower bound. □

The next theorem is one of the main results of this paper. It gives lower bounds for the minimax regret under all three feedback assumptions. Note that the lower bounds for the full information and semi-bandit cases follow easily from standard lower bounds. Our main contribution here is the lower bound for bandit online combinatorial optimization.

THEOREM 6.2 *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 full information feedback,*

$$\inf_{strategies}\sup_{adversaries} R_n \ge 0.03m\sqrt{n\log\frac{d}{m}}, \tag{10}$$

under semi-bandit feedback,

$$\inf_{strategies} \sup_{adversaries} R_n \geq 0.04\sqrt{mdn}, \tag{11}$$

and under bandit feedback,

$$\inf_{strategies} \sup_{adversaries} R_n \geq 0.02m\sqrt{dn}. \tag{12}$$

Before the proof, it is interesting to note that the lower bound (10) is maximized when m is a constant multiple of d. For such sets, the bound under the semi-bandit assumption are not larger than for the full information case. Indeed, the matching upper bounds show the, perhaps surprising, fact that for rich classes, the full information and the semi-bandit problems have essentially the same difficulty.

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

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

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

The bounds (10) and (11) follow directly from Audibert and Bubeck [3, Theorem 30] (which gives the bound in the case $m = 1$). Indeed, full information and semi-bandit feedback, the player faces m independent games. On the other hand, in the bandit case the situation is more delicate. We focus now on this latter setting and divide the proofs in four steps. From step 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.

First step: 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 i^{th} game. Moreover, let τ be drawn uniformly at random in $\{1,\dots,n\}$.

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 i^{th} 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 law of $\alpha(i, I_{i,\tau})$ when the player faces the α-adversary. Note that we have $\mathbb{P}_{i,\alpha}(1) = \mathbb{E}_\alpha \frac{1}{n}\sum_{t=1}^n \mathbb{1}_{\alpha(i,I_{i,t})=1}$, hence, against the α-adversary, we have

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

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

$$\max_{\alpha\in\mathcal{A}} \overline{R}_n \geq n\varepsilon \sum_{i=1}^m \left(1 - \frac{1}{(d/m)^m} \sum_{\alpha\in\mathcal{A}} \mathbb{P}_{i,\alpha}(1)\right). \tag{13}$$

Second step: information inequality.

Let $\mathbb{P}_{-i,\alpha}$ be the law of $\alpha(i, I_{i,\tau})$ against the adversary which plays like the α-adversary except that in the i^{th} 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,

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

Moreover, note that by symmetry of the adversaries $(-i,\alpha)$,

$$\begin{aligned}\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{-i,\alpha}(1) &= \frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}_{(-i,\alpha)}\alpha(i,I_{i,\tau})\\ &= \frac{1}{(d/m)^m}\sum_{\beta\in\mathcal{A}}\frac{1}{d/m}\sum_{\alpha:(-i,\alpha)=(-i,\beta)}\mathbb{E}_{(-i,\alpha)}\alpha(i,I_{i,\tau})\\ &= \frac{1}{(d/m)^m}\sum_{\beta\in\mathcal{A}}\frac{1}{d/m}\mathbb{E}_{(-i,\beta)}\sum_{\alpha:(-i,\alpha)=(-i,\beta)}\alpha(i,I_{i,\tau})\\ &= \frac{1}{(d/m)^m}\sum_{\beta\in\mathcal{A}}\frac{1}{d/m}\\ &= \frac{m}{d}, \end{aligned} \tag{14}$$

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

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{i,\alpha}(1) \le \frac{m}{d} + \sqrt{\frac{1}{2(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})}. \tag{15}$$

Third step: 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 law 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 law 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 laws $\mathbb{P}^t_\alpha$ of the observed losses W_t up to time t. More precisely, we have,

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

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 B.2 (see below) we obtain,

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

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 (15) we obtain (again thanks to (14)), for $\varepsilon \le \frac{1}{\sqrt{8}}$,

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

To conclude the proof of (12) for deterministic players one needs to plug this last equation in (13) along with straightforward computations.

Fourth step: Fubini's theorem to handle non-deterministic players.

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

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

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 $\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}_\alpha\sum_{t=1}^n(a_t^Tz_t-\alpha^Tz)$ as before (note that α is the optimal action in expectation against the α-adversary). □

Appendix A. Proof of Theorem 2.3. The first step of the proof is a simple lemma, see Kalai and Vempala [21].

LEMMA A.1 *Let $\ell:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ be a loss function and*

$$a_t^*=\operatorname*{argmin}_{a\in\mathcal{A}}\sum_{s=1}^t\ell(a,z_t).$$

Then one has

$$\sum_{t=1}^n\ell(a_t^*,z_t)\le\sum_{t=1}^n\ell(a_n^*,z_t).$$

We can now prove the theorem.

PROOF. Let

$$a_t^*=\operatorname*{argmin}_{a\in\mathcal{A}}\left(\xi_1+\sum_{s=1}^t z_s\right)^T a.$$

Using Lemma A.1 with

$$\ell(a,z_t)=\begin{cases}(\xi_1+z_1)^Ta_1 & \text{if } t=1,\\ z_t^Ta & \text{if } t>1,\end{cases}$$

one obtains that for any $u\in\mathcal{A}$,

$$\xi_1^Ta_1^*+\sum_{t=1}^n z_t^Ta_t^*\le\xi_1^Tu+\sum_{t=1}^n z_t^Tu.$$

In particular, we get

$$\mathbb{E}\sum_{t=1}^n z_t^T(a_t^*-u)\le\frac{m}{2\eta}.$$

Now let

$$\widetilde{a}_t=\operatorname*{argmin}_{a\in\mathcal{A}}\left(\xi_t+\sum_{s=1}^t z_s\right)^T a.$$

Since the adversary is oblivious, $\widetilde{a}_t$ has the same distribution as a_t^*. In particular, we have $\mathbb{E}z_t^Ta_t^*=\mathbb{E}z_t^T\widetilde{a}_t$, which implies

$$\mathbb{E}\sum_{t=1}^n z_t^T(\widetilde{a}_t-u)\le\frac{m}{2\eta}.$$

To conclude, it suffices to show that $\mathbb{E}z_t^T(a_t-\widetilde{a}_t)\le\eta md$. Let

$$h(\xi)=z_t^T\left\{\operatorname*{argmin}_{a\in\mathcal{A}}\left(\xi+\sum_{s=1}^{t-1}z_s\right)^T a\right\}.$$

Then one has

$$\begin{aligned}\mathbb{E}z_t^T(a_t-\widetilde{a}_t) &= \mathbb{E}h(\xi_t)-\mathbb{E}h(\xi_t+z_t)\\ &= \eta^d\int_{\xi\in[0,1/\eta]^d}h(\xi)d\xi-\eta^d\int_{\xi\in z_t+[0,1/\eta]^d}h(\xi)d\xi\\ &\le m\eta^d\int_{\xi\in[0,1/\eta]^d\setminus\{z_t+[0,1/\eta]^d\}}\\ &= m\mathbb{P}\left(\exists i\in\{1,\dots,d\}:\xi_1(i)\le z_t(i)\right)\\ &\le \eta md.\end{aligned}$$

□

Appendix B. Technical lemmas.

LEMMA B.1 *For any $k \in \mathbb{N}^*$, for any $1 \le c \le 2$, we have*

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

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 $\binom{k}{i}^2 c^i / \sum_{j=0}^k \binom{k}{j}^2 c^j$. We have $f'(c) = \frac{1}{c}\mathbb{E}[X(1-X/k)] - \frac{1}{c}\mathbb{E}(X)\mathbb{E}(1-X/k) = -\frac{1}{ck}\mathbb{V}\mathrm{ar}\, X \le 0$. So the function f is decreasing on $[1,2]$, and therefore it suffices to consider $c = 2$. Numerator and denominator of 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 $k \ge 2$ and $i \in [1, k-1]$,

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

hence

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

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

$$[\chi(\lambda)]^k \frac{2e^{-1/3}}{\pi k} < \binom{k}{i}^2 2^i < [\chi(\lambda)]^k \frac{e^{1/6}}{2\pi\lambda}. \tag{16}$$

Lemma B.1 can be numerically verified for $k \le 10^6$. We now consider $k > 10^6$. For $\lambda \ge 0.666$, since the function χ can be shown to be decreasing on $[0.666, 1]$, the inequality $\binom{k}{i}^2 2^i < [\chi(0.666)]^k \frac{e^{1/6}}{2\times 0.666\times\pi}$ 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 $\lambda \ge 0.666$ and $k > 10^6$, we have

$$\begin{aligned}\binom{k}{i}^2 2^i < 0.001 \times [\chi(0.657)]^k \frac{e^{1/6}}{2\pi \times 0.666 \times k^2} &< [\chi(0.657)]^k \frac{2e^{-1/3}}{1000\pi k^2} \\ &= \min_{\lambda\in[0.656,0.657]} [\chi(\lambda)]^k \frac{2e^{-1/3}}{1000\pi k^2} \\ &< \frac{1}{1000k} \max_{i\in\{1,\dots,k-1\}\cap[0,0.666k)} \binom{k}{i}^2 2^i\ , \end{aligned} \tag{17}$$

where the last inequality comes from (16) and the fact that there exists $i \in \{1,\dots,k-1\}$ such that $i/k \in [0.656, 0.657]$. Inequality (17) implies that for any $i \in \{1,\dots,k\}$, we have

$$\sum_{\frac{5}{6}k\le i\le k} \binom{k}{i}^2 2^i < \frac{1}{1000} \max_{i\in\{1,\dots,k-1\}\cap[0,0.666k)} \binom{k}{i}^2 2^i < \frac{1}{1000} \sum_{0\le i<0.666k} \binom{k}{i}^2 2^i.$$

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

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

□

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

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

PROOF. Let $Z, Z', Z_1, \ldots, Z_n$ be independent Bernoulli distributions with parameters $p, p', p_1, \ldots, p_n$. Define $S = \sum_{i=1}^{\ell} Z_i$, $T = \sum_{i=\ell+1}^{n} Z_i$ and $V = Z + S$. By a slight abuse of notation, merging in the same notation the distribution and the random variable, we have

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

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

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

which holds for $1 \le k \le \ell+1$, we obtain

$$\begin{aligned}\mathrm{KL}(Z+S, Z'+S) &= \sum_{k=0}^{\ell+1} \mathbb{P}(V=k) \log\left(\frac{\mathbb{P}(Z+S=k)}{\mathbb{P}(Z'+S=k)}\right)\\ &= \sum_{k=0}^{\ell+1} \mathbb{P}(V=k) \log\left(\frac{p s_{k-1} + (1-p) s_k}{p' s_{k-1} + (1-p') s_k}\right)\\ &= \sum_{k=0}^{\ell+1} \mathbb{P}(V=k) \log\left(\frac{p\frac{1-q}{q} k + (1-p)(\ell-k+1)}{p'\frac{1-q}{q} k + (1-p')(\ell-k+1)}\right)\\ &= \mathbb{E}\log\left(\frac{(p-q)V + (1-p)q(\ell+1)}{(p'-q)V + (1-p')q(\ell+1)}\right). \qquad (18)\end{aligned}$$

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

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

Second case: $q = p$.
In this case, V is a binomial distribution with parameters $\ell+1$ and p. From (18), we have

$$\begin{aligned}\mathrm{KL}(Z+S, Z'+S) &\le -\mathbb{E}\log\left(\frac{(p'-p)V + (1-p')p(\ell+1)}{(1-p)p(\ell+1)}\right)\\ &\le -\mathbb{E}\log\left(1 + \frac{(p'-p)(V-\mathbb{E}V)}{(1-p)p(\ell+1)}\right). \qquad (19)\end{aligned}$$

To conclude, we will use the following lemma.

LEMMA B.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 $f(x) = -(x-1) + \frac{(x-1)^2}{2x_0} + \log(x)$. We have $f'(x) = -1 + \frac{x-1}{x_0} + \frac{1}{x}$, and $f''(x) = \frac{1}{x_0} - \frac{1}{x^2}$. 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 B.3 and (19), using $x_0 = \frac{1-p'}{1-p}$, we obtain

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

□

Acknowledgments. G. Lugosi is supported by the Spanish Ministry of Science and Technology grant MTM2009-09063 and PASCAL2 Network of Excellence under EC grant no. 216886.

References

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

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

[3] _____, *Regret bounds and minimax policies under partial monitoring*, Journal of Machine Learning Research **11** (2010), 2635–2686.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, *The non-stochastic multi-armed bandit problem*, SIAM Journal on Computing **32** (2003), no. 1, 48–77.

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

[6] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

[7] S. Bubeck, *Introduction to online optimization*, Lecture Notes, 2011.

[8] S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade, *Towards minimax policies for online linear optimization with bandit feedback*, Arxiv preprint arXiv:1202.3079 (2012).

[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*, Cambridge University Press, 2006.

[10] _____, *Combinatorial bandits*, Journal of Computer and System Sciences (2011), To appear.

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

[12] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences **55** (1997), 119–139.

[13] A. Grove, N. Littlestone, and D. Schuurmans, *General convergence results for linear discriminant updates*, Machine Learning **43** (2001), 173–210.

[14] A. György, T. Linder, G. Lugosi, and G. Ottucsák, *The on-line shortest path problem under partial monitoring*, Journal of Machine Learning Research **8** (2007), 2369–2403.

[15] J. Hannan, *Approximation to Bayes risk in repeated play*, Contributions to the theory of games **3** (1957), 97–139.

[16] E. Hazan, *The convex optimization approach to regret minimization*, Optimization for Machine Learning (S. Sra, S. Nowozin, and S. Wright, eds.), MIT press, 2011, pp. 287–303.

[17] E. Hazan, S. Kale, and M. Warmuth, *Learning rotations with little regret*, Proceedings of the 23rd Annual Conference on Learning Theory (COLT), 2010.

[18] D. P. Helmbold and M. Warmuth, *Learning permutations with exponential weights*, Journal of Machine Learning Research **10** (2009), 1705–1736.

[19] M. Herbster and M. Warmuth, *Tracking the best expert*, Machine Learning **32** (1998), 151–178.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of convex analysis*, Springer, 2001.

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

[22] S. Kale, L. Reyzin, and R. Schapire, *Non-stochastic bandit slate problems*, Advances in Neural Information Processing Systems (NIPS), 2010, pp. 1054–1062.

[23] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, Annals of Mathematical Statistics **23** (1952), 462–466.

[24] J. Kivinen and M. Warmuth, *Relative loss bounds for multidimensional regression problems*, Machine Learning **45** (2001), 301–329.

[25] W. Koolen, M. Warmuth, and J. Kivinen, *Hedging structured concepts*, Proceedings of the 23rd Annual Conference on Learning Theory (COLT), 2010, pp. 93–105.

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

[27] A. Nemirovski, *Efficient methods for large-scale convex optimization problems*, Ekonomika i Matematicheskie Metody **15** (1979), (In Russian).

[28] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley Interscience, 1983.

[29] A. Rakhlin, *Lecture notes on online learning*, 2009.

[30] H. Robbins and S. Monro, *A stochastic approximation method*, Annals of Mathematical Statistics **22** (1951), 400–407.

[31] A. Schrijver, *Combinatorial optimization*, Springer, 2003.

[32] S. Shalev-Shwartz, *Online learning: Theory, algorithms, and applications*, Ph.D. thesis, The Hebrew University of Jerusalem, 2007.

[33] E. Takimoto and M. Warmuth, *Paths kernels and multiplicative updates*, Journal of Machine Learning Research **4** (2003), 773–818.

[34] T. Uchiya, A. Nakamura, and M. Kudo, *Algorithms for adversarial bandit problems with multiple plays*, Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT), 2010.

[35] M. Warmuth and D. Kuzmin, *Randomized online pca algorithms with regret bounds that are logarithmic in the dimension*, Journal of Machine Learning Research **9** (2008), 2287–2320.

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