Tight Bounds for Bandit Combinatorial Optimization

Alon Cohen   ALON.COHEN@TECHNION.AC.IL
Tamir Hazan   TAMIR.HAZAN@TECHNION.AC.IL
Technion—Israel Institute of Technology

Tomer Koren   TKOREN@GOOGLE.COM
Google

Abstract

We revisit the study of optimal regret rates in bandit combinatorial optimization—a fundamental framework for sequential decision making under uncertainty that abstracts numerous combinatorial prediction problems. We prove that the attainable regret in this setting grows as $\widetilde{\Theta}(k^{3/2}\sqrt{dT})$ where d is the dimension of the problem and k is a bound over the maximal instantaneous loss, disproving a conjecture of Audibert, Bubeck, and Lugosi (2013) who argued that the optimal rate should be of the form $\widetilde{\Theta}(k\sqrt{dT})$. Our bounds apply to several important instances of the framework, and in particular, imply a tight bound for the well-studied bandit shortest path problem. By that, we also resolve an open problem posed by Cesa-Bianchi and Lugosi (2012).

1. Introduction

We consider the problem of online combinatorial optimization with bandit feedback, also known as *bandit combinatorial optimization*, or more succinctly as *combinatorial bandits*. The problem can be described as the following game between a learner and an environment, that proceeds for T rounds. On each round $t = 1, 2, \ldots, T$, the learner has to pick, possibly at random, an action x_t from a subset $S \subseteq \{0,1\}^d$ of the hypercube in d-dimensions, with the property that each element $x \in S$ has at exactly k non-zero entries, that is $\sum_{i=1}^{d} x_i = k$. Simultaneously, the environment privately chooses a loss vector $\ell_t \in [0,1]^d$. The learner then incurs the loss $\ell_t \cdot x_t \in [0, k]$ and may observe only this loss (but not the vector ℓ_t) as feedback. The goal of the learner throughout the T rounds of the game is to minimize her regret, defined as

$$\sum_{t=1}^{T} \ell_t \cdot x_t \;-\; \min_{x \in S} \sum_{t=1}^{T} \ell_t \cdot x \;.$$

Bandit combinatorial optimization is a fundamental primitive of sequential decision making under uncertainty, and abstracts several major problems in this context (see, e.g., Bubeck and Cesa-Bianchi, 2012). Perhaps the most important and well-studied problem captured by this framework is online network routing, also known as the online shortest path problem (Takimoto and Warmuth, 2003; Kalai and Vempala, 2005). In this setting, a source station s repeatedly sends communication packets to a target station t through a network represented by a connected directed acyclic graph. On each decision round, the environment associates each edge in the network with a loss, and the learner suffers the loss accumulated over the edges in her chosen path. Each packet can be routed differently

and the station has to pick routes so as to minimize the overall amount of time it takes the packets to arrive. In the bandit version of the problem, the only feedback that the source station observes is the roundtrip time of each packet—namely the time it takes the packet to travel to its destination and return to the source.

The network routing problem can be cast in the online combinatorial optimization framework as follows: the set of all s-t paths can be represented as a set $S \subseteq \{0,1\}^d$ where d is the number of edges in the graph, and the non-zero entries in each $x \in S$ indicate the edges that are contained in the path x; then, if $\ell_t \in [0,1]^d$ is the loss vector that associates costs to edges in the network on decision round t, then the cost of path x is given by $\ell_t \cdot x$. The assumption that $\sum_{i=1}^d x_i = k$ for all $x \in S$ means that the length of an s-t path in the network is exactly k (which is also an upper bound on the maximal cost of any s-t path).

The study of bandit combinatorial optimization dates back to the work of Awerbuch and Kleinberg (2004), who considered the online shortest path problem in the bandit setting, henceforth called the *bandit shortest path* problem, in which the learner observes only the loss that she has suffered, and showed an $O(kd^{5/3}T^{2/3})$ bound on the expected regret. Dani et al. (2008) and Abernethy et al. (2008) considered the problem in the wider context of bandit linear optimization and established a regret bound with the optimal $\sqrt{T}$ dependence. Subsequently, Cesa-Bianchi and Lugosi (2012) focused on bandit combinatorial optimization, and showed that a similar bound can be achieved for a large number of problems under this framework, often with computationally efficient algorithms. For the bandit shortest path problem, Cesa-Bianchi and Lugosi (2012) conjectured that the general upper bound is in fact suboptimal and that the correct tight bound is of the form $\widetilde{O}(k\sqrt{dT})$, and could be obtained by a clever adaptation of their algorithm.

More recently, Bubeck et al. (2012); Audibert et al. (2013) showed that an $\widetilde{O}(k^{3/2}\sqrt{dT})$ upper bound holds for any combinatorial bandit problem using a general online optimization algorithm. Additionally, the authors gave a new lower bound of $\Omega(k\sqrt{dT})$ on the expected regret in combinatorial bandits, which leaves a gap of $\sqrt{k}$ between that and their upper bound (ignoring logarithmic factors). They conjectured as well that the lower bound is, in fact, the correct rate and articulated that the upper bound could be improved by non-trivial modifications of the existing algorithmic techniques.

In this paper, we revisit the study of optimal regret rates in bandit combinatorial optimization. Our main contribution is in disproving the conjectures of Cesa-Bianchi and Lugosi (2012) and Audibert et al. (2013) and showing that the expected regret of combinatorial bandits in general, and of the bandit shortest path problem in particular, is in fact $\widetilde{\Theta}(k^{3/2}\sqrt{dT})$. Namely, we show a new lower bound of $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$ for combinatorial bandits that matches the best known upper bound up to logarithmic factors, and also holds (via simple adaptations) in the context of bandit shortest path. Furthermore, we show how this lower bound can be adapted to the setting of online ranking (Helmbold and Warmuth, 2009).

Surprisingly, the construction used in our lower bound is very simple and is based on straightforward adaptations of the one used by Audibert et al. (2013). Furthermore, our analysis is also significantly simpler and shorter than theirs. In a nutshell, the improvement in the bound is obtained via the following observation: when picking its randomized losses for fooling the learner, the environment can choose noisy vectors whose entries are strongly correlated with each other rather than being independent, as is the case in typical lower

bound constructions (and, in particular, as suggested by Audibert et al., 2013).[1] Since the learner never observes individual entries of the loss vectors and can only see a sum of k of them (for a particular choice of the action set S), she cannot exploit this correlation in any way. On the other hand, with correlated noise terms the observed loss value can have a variance that grows *quadratically* with k, rather linearly as is the case with i.i.d. noise, which directly deteriorates the learner's regret by an additional factor of $\sqrt{k}$.

1.1. Related work

Combinatorial bandit optimization is closely related to a somewhat more general online learning scenario known as *bandit linear optimization*, which was first considered by McMahan and Blum (2004) (and subsequently by Dani et al., 2008; Abernethy et al., 2008). In this setting, the decision set S is not restricted to subsets of the hypercube $\{0, 1\}^d$ and may be an arbitrary compact convex set in $\mathbb{R}^d$; instead, the only requirement is that the loss the learner incurs by picking any action in S is bounded (say, by 1 in absolute value) for all possible loss vectors of the environment. State-of-the-art bounds for this problem were obtained by Bubeck et al. (2012) and Hazan and Karnin (2016), the latter using computationally-efficient algorithms.

The general linear optimization setting allows for more general geometries of the sets in which the decisions and the loss vectors reside (e.g., they are typically assumed to be subsets of the Euclidean unit ball), and consequently the bounds obtained in that setting are often not immediately comparable to those in the combinatorial one. In particular, the lower bounds proved by Dani et al. (2008) and more recently by Shamir (2015) hold in the general linear optimization setting (with Euclidean geometry) and do not apply to any natural problem in the combinatorial setting.

A significant amount of work has been devoted to combinatorial optimization in the closely related semi-bandit feedback model (e.g., György et al., 2007; Kale et al., 2010; Audibert et al., 2013; Neu, 2015; Neu and Bartók, 2016), in which after playing an action x_t the learner may observe the individual entries of the loss vector ℓ_t that correspond to active entries of x_t, namely those entries i for which $x_t(i) = 1$. For example, in the context of the online shortest path problem, instead of observing just the overall cost of the chosen path (as is the case in the bandit setting), the player may observe the individual cost of each edge in that path. In the semi-bandit case, however, the regret of bandit combinatorial optimization is by now well understood, and is known to be of the form $\Theta(\sqrt{kdT})$; see Audibert et al. (2013) and the references therein.

For further and more detailed account on related partial information models and their regret analysis, we refer to the recent survey by Bubeck and Cesa-Bianchi (2012).

2. Main results

We now state the main results of this paper. As our results are lower bounds on the learner's regret, we will henceforth focus on oblivious environments, that are required to choose the entire sequence $\ell_1, \dots, \ell_T$ before the game begins and thus do not react adaptively to the

---

1. Note that the correlation discussed here is between different entries of the same loss vector, rather than between different loss vectors at different rounds. In particular, the loss vectors in our lower bound constructions are still chosen i.i.d. so our bounds also apply to the stochastic i.i.d. case.

player's randomized decisions. (A lower bound for such environments also implies a lower bound for more general adaptive environments.) In this setup, we will give bounds on the expected regret, defined as

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t \cdot x_t\right] - \min_{x \in S} \sum_{t=1}^{T} \ell_t \cdot x \ , \tag{1}$$

where the expectations are taken over the random choices of the learner.

Our first result deals with the general combinatorial bandits setting and shows that if the environment is free to choose any action set S, the regret of the learner can be very large. Our lower bound is attained in the multitask bandit problem, in which a learner is simultaneously trying to solve k instances of the n-armed bandit problem (Auer et al., 2002) with $n = d/k$ (which we assume is an integer for simplicity). At every round of the game, the learner plays k actions, one in each of the bandit problems, and observes the sum of the losses that correspond with these k actions. The set S of actions is given as follows:

$$S = \left\{ x \in \{0,1\}^d \ : \ \forall j \in [k] \sum_{i=(j-1)n+1}^{jn} x(i) = 1 \right\} \ . \tag{2}$$

Theorem 1 (multitask MAB)**.** *Assume that $n \geqslant 2$, and let the set of actions $S \subseteq \{0,1\}^d$ be as defined in Eq. (2). Any learning algorithm for the multitask bandit problem must incur at least $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$ expected regret in the worst case.*

The bound in the theorem hides a factor of $\log^{-1/2} T$ which is an artifact of our construction and is likely to be redundant. Nevertheless, up to logarithmic factors the bound is tight and matches the upper bounds of Bubeck et al. (2012) and Hazan and Karnin (2016).

The lower bound of Theorem 1 does not hold for any set S but rather to an instance of the multitask bandit problem, yet, as we show in the following results, is still general enough to imply lower bounds for two important instances of bandit linear optimization. Our next theorem gives a lower bound for the bandit shortest path problem, and shows that even when we limit the action set S to paths in a certain graph, the regret of the learner can still be forced to be large. Formally, given a connected DAG $G = (V, E)$ with d edges and two nodes $s, t \in V$, we define the set of actions $S \subseteq \{0,1\}^d$ as follows:

$$S = \left\{ x \in \{0,1\}^d \ : \ \text{the set } \{e \in E : x(e) = 1\} \text{ forms an } s\text{-}t \text{ path} \right\} \ . \tag{3}$$

Then, we have the following:

Theorem 2 (online shortest paths)**.** *Assume that $k \leqslant d/2$. There exists a graph with d edges such that any s-t path has exactly k edges (see Figure 1), for which the action set S is defined as in Eq. (3). Against this graph any online learning algorithm for the bandit shortest path problem must suffer at least $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$ expected regret in the worst case.*

Again, the theorem implies that the regret rate for bandit shortest path is $\widetilde{\Theta}(k^{3/2}\sqrt{dT})$, contrary to what was conjectured in the literature (Cesa-Bianchi and Lugosi, 2012).

Our last main result shows a lower bound for the online ranking problem. This problem can be cast as finding a maximum matching in the complete bipartite graph $K_{k,n}$, that has

$d = kn$ edges. The set of all of these matchings is represented by the action set $S \subseteq \{0,1\}^d$, and the non-zero entries of every $x \in S$ indicate which edges participate in the matching that corresponds with x. Formally,

$$S = \left\{ x \in \{0,1\}^d \ : \ \forall j \in [k] \sum_{i=(j-1)n+1}^{jn} x(i) = 1, \ \forall l \in [n] \sum_{i=1}^{k} x((i-1)n+l) = 1 \right\} . \quad (4)$$

Theorem 3 (online ranking). *Assume that $k \leqslant n/2$. Consider the problem of online ranking between k and n elements, whose action set S is defined in Eq. (4). Any bandit learning algorithm for this problem must suffer at least $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$ expected regret in the worst case.*

3. Proofs

3.1. Main result

In this section we prove Theorem 1. We show a lower bound of $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$ on the regret of any online learning algorithm applied to an instance of the multitask bandit problem. Surprisingly, the factor $\sqrt{k}$ improvement is obtained via a simple modification of previous constructions (Audibert et al., 2013).

We start by applying Yao's minimax principle, implying that it suffices to show randomized strategy for the environment that forces any deterministic learning algorithm to suffer $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$ regret in expectation.[2] We start by constructing an environment that generates unbounded losses, and subsequently adapt the resulting lower bound to the bounded case.

Set $\epsilon = \sigma\sqrt{kd/(4T)}$ for some $\sigma > 0$ whose value is to be determined later. Before the game begins, in each of the k problems in S, the environment chooses the best arm uniformly at random. We denote the vector indicating this choice by $x^\star \in S$. At every round t, the environment samples $Z_t \sim \mathcal{N}(0, \sigma^2)$. Denote the loss generated by environment on round t as $L'_t(i) = 1/2 - \epsilon \cdot x^\star(i) + Z_t$ for $i = 1, \ldots, d$.

The idea behind this construction is as follows. In order to avoid large losses and minimize her regret, the learner has to identify the best arm in each of the k subproblems, namely to recover $x^\star$. Now suppose that the losses of each coordinate were sampled independently, and each entry in L'_t were to receive an i.i.d. sample of the Gaussian noise. Then the variance of the loss observed by the learner, namely of the random variable $L_t \cdot x$ for any choice of $x \in S$, is of the order of k. On the other hand, because of the correlation between the losses of the different coordinates in the construction above, the variance of the observed loss is of the order of k^2. This allows us to gain and additional $\sqrt{k}$ factor in the lower bound on the regret. Note that, crucially, the learner always observes a sum of k random noise terms and can never peek into the individual terms in the sum (this is due to the bandit feedback and the specific structure of the decision set S); hence, the correlation in the noise cannot be exploited by the learner and the increase in the overall variance comes at no price.

For the construction above, we have the following lemma.

---

2. Consequently, in the definition of the regret and throughout the analysis, expectations are taken with respect to the environment's randomization rather than the learner's who is now deterministic.
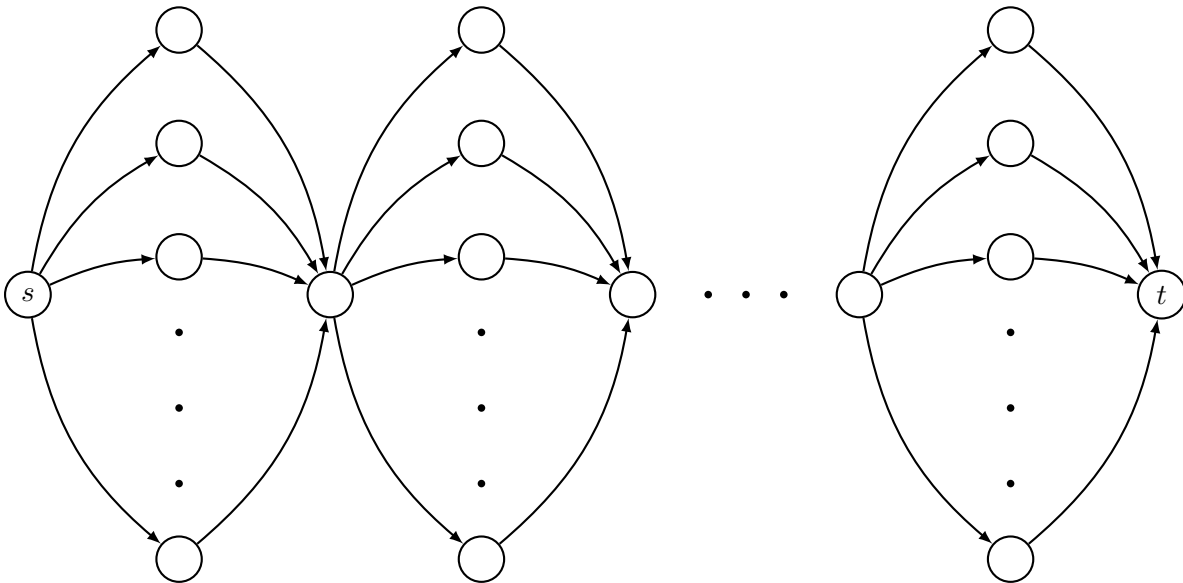
Figure 1: Graph for the lower bound. The graph consists of $k/2$ layers, in each the learner has to choose one of d/k vertices for an s-t path to pass through.

Lemma 4. *Any deterministic player must suffer a regret of at least $\sigma k^{3/2}\sqrt{dT}/8$ in expectation against an environment that plays the losses $L'_1, \ldots, L'_T$.*

To show that Theorem 1 holds we need to show that the learner suffers large regret against an environment that plays losses that are bounded in $[0,1]^d$. While the losses $L'_1, \ldots, L'_T$ are unbounded, for the right choice of σ they are bounded with high probability. We now show that this allows us to obtain a lower bound on the regret against an environment that plays losses $L_1, L_2, \ldots, L_T$, such that $L_t(i) = \mathrm{clip}(L'_t(i))$ for $\mathrm{clip}(a) = \max\{\min\{a, 1\}, 0\}$.

Theorem 5. *Assume that $T \geq kd$ and let $\sigma^2 = 1/(192 + 96 \log T)$. Any deterministic player must suffer an expected regret of at least $\sigma k^{3/2}\sqrt{dT}/16$ against an environment that plays the losses $L_1, \ldots, L_T$.*

The proof of Theorem 1 is now given by setting the value of σ into the bound in Theorem 5.

3.2. Bandit shortest path

In this section we show a lower bound for the bandit shortest path problem, proving Theorem 2. Suppose without loss of generality that k and d are even, and that d is a multiple of k. We show a lower bound on the regret by constructing a graph that simulates the multitask bandit problem with $k/2$ problems of d/k arms each.

This graph is shown in Figure 1. The graph consists of d edges and $d/2 + k/2 + 1$ vertices set in $k/2$ layers. Each layer has an incoming vertex connected to d/k intermediate vertices,

all of them connected to the same outgoing vertex. This outgoing vertex is the incoming vertex of the next layer and so forth. Note that to form an s-t path the learner has to pass through exactly one of the d/k vertices in each layer, and therefore every such path has exactly k edges.

Now, given the losses $L_1, L_2, \ldots, L_T$ generated by the environment of Section 3.1, we shall construct an environment for the shortest path problem such that the regret of the learner would be the same as the one in the proof of Theorem 1. Indeed, recall that the loss at coordinates $(j-1)d/k + 1, \ldots, jd/k$ is associated with the losses of the j'th d/k-armed bandit problem. Then on round t for the j'th layer of the graph, we can set the losses $L_t((j-1)d/k + 1), \ldots, L_t(jd/k)$ to the edges going from the incoming vertex to the intermediate vertices, and a loss of 0 to the edges going from the intermediate vertices to the outgoing vertex.

Therefore, we have a bijection between any s-t path and a set of $k/2$ arms in the aforementioned multitask bandit problem, such that the sum of the losses on the edges of the s-t path and the sum of the losses of these $k/2$ arms are the same. We conclude by invoking Theorem 1 that says that any learner must suffer an expected regret of at least $\widetilde{\Omega}(k^{3/2}\sqrt{dT})$, as claimed.

3.3. Online ranking

In this section we prove Theorem 3 by a similar construction to the one in Section 3.1, for which we present the following random environment.

Set $\epsilon = \sigma\sqrt{kd/(8T)}$. Before the game starts, the environment samples a maximum matching in $K_{k,n}$ unfiromly at random, and denote the vector indicating this choice by $x^\star \in S$, for the set S defined in Eq. (4). At every round t, the environment samples $Z_t \sim \mathcal{N}(0, \sigma^2)$. Denote the loss generated by the environment on round t as $L'_t(i) = 1/2 - \epsilon \cdot x^\star(i) + Z_t$ for all $i = 1, 2, \ldots, d$.

We have the following lemma.

Lemma 6. *Any deterministic player must suffer regret of at least $\sigma k^{3/2}\sqrt{dT}/8$ in expectation against an environment that plays the losses $L'_1, \ldots, L'_T$.*

Now to prove Theorem 3, the result above can be adapted to bounded losses in the same manner as done in Theorem 5.

4. Additional proofs

4.1. Proof of Lemma 4

Proof. Let us denote by $i_1^\star, \ldots, i_k^\star$ the locations of the non-zero coordinates of the random variable $x^\star$, arranged in increasing order. We next introduce the random variables $T_1, \ldots, T_k$, where each T_j is the number of times the learner played an x_t such that $x_t(i_j^\star) = 1$. For each $x \in S$, we introduce the notations $\mathbb{P}_x$ and $\mathbb{E}_x$ indicating probability and expectation with respect to the marginal distributions under which $x^\star = x$. Then,

$$\begin{aligned}
\mathrm{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T} L_t' \cdot x_t - \min_{x \in S} \sum_{t=1}^{T} L_t' \cdot x\right] \\
&\geqslant \mathbb{E}\left[\sum_{t=1}^{T} L_t' \cdot x_t - \sum_{t=1}^{T} L_t' \cdot x^\star\right] \\
&= \frac{1}{n^k} \sum_{x \in S} \mathbb{E}_x \left[\sum_{t=1}^{T} L_t' \cdot x_t - \sum_{t=1}^{T} L_t' \cdot x\right] \\
&= \frac{1}{n^k} \sum_{x \in S} \epsilon \cdot \mathbb{E}_x \left[\sum_{j=1}^{k} (T - T_j)\right] \\
&= \epsilon \left(kT - \sum_{j=1}^{k} \frac{1}{n^k} \sum_{x \in S} \mathbb{E}_x[T_j]\right) ,
\end{aligned} \tag{5}$$

and in order to proceed, we need to upper bound $\mathbb{E}_x[T_j]$ for each j.

For every $x \in S$ and $j \in [k]$ we introduce a new distribution referred to as $\mathbb{P}_{x,-j}$ and $\mathbb{E}_{x,-j}$. It is the same as $\mathbb{P}_x$ except that the loss of coordinate $i_j^\star$ is also $1/2 + Z_t$ (instead of $1/2 + Z_t - \epsilon$). Let λ_t be the loss observed at time t, and $\lambda^{(t)} = (\lambda_1, \ldots, \lambda_t)$ be the losses observed up to and including time t. Then, since the sequence $\lambda^{(T)}$ determines the actions of the learner over the entire game, and by Pinsker's inequality,

$$\begin{aligned}
\mathbb{E}_x[T_j] - \mathbb{E}_{x,-j}[T_j] &\leqslant T \cdot D_{\mathrm{TV}}\left(\mathbb{P}_{x,-j}\left[\lambda^{(T)}\right], \mathbb{P}_x\left[\lambda^{(T)}\right]\right) \\
&\leqslant T\sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathbb{P}_{x,-j}\left[\lambda^{(T)}\right] \,\|\, \mathbb{P}_x\left[\lambda^{(T)}\right]\right)} .
\end{aligned} \tag{6}$$

Moreover, by the chain rule of KL-divergence, $D_{\mathrm{KL}}\left(\mathbb{P}_{x,-j}[\lambda^{(T)}] \,\|\, \mathbb{P}_x[\lambda^{(T)}]\right)$ equals

$$\sum_{t=1}^{T} \mathbb{E}_{\lambda^{(t-1)} \sim \mathbb{P}_{x,-j}} \left[D_{\mathrm{KL}}\left(\mathbb{P}_{x,-j}\left[\lambda_t \middle| \lambda^{(t-1)}\right] \,\|\, \mathbb{P}_x\left[\lambda_t \middle| \lambda^{(t-1)}\right]\right)\right] . \tag{7}$$

Consider a single term in the sum, and recall that $\lambda^{(t-1)}$ determines the action x_t chosen by the learner on round t. If $x_t(i_j^\star) = 0$, the loss observed under $\mathbb{P}_x$ and $\mathbb{P}_{x,-j}$ are the same, and the KL divergence is 0. If $x_t(i_j^\star) = 1$ then the observed losses under $\mathbb{P}_x$ and $\mathbb{P}_{x,-j}$ are both Gaussian whose means are ϵ apart, and the variance of both of them is $\sigma^2 k^2$. Therefore,

$$D_{\mathrm{KL}}\left(\mathbb{P}_{x,-j}\left[\lambda_t \middle| \lambda^{(t-1)}\right] \,\|\, \mathbb{P}_x\left[\lambda_t \middle| \lambda^{(t-1)}\right]\right) \leqslant \frac{\epsilon^2}{2k^2\sigma^2} .$$

Plugging the above back into Eq. (7),

$$D_{\mathrm{KL}}\left(\mathbb{P}_{x,-j}\left[\lambda^{(T)}\right] \,\|\, \mathbb{P}_x\left[\lambda^{(T)}\right]\right) \leqslant \sum_{t=1}^{T} \mathbb{P}_{x,-j}\left[x_t(i_j^\star) = 1\right] \cdot \frac{\epsilon^2}{2k^2\sigma^2} = \frac{\epsilon^2}{2k^2\sigma^2} \mathbb{E}_{x,-j}[T_j] ,$$

and the latter back into Eq. (6), we get $\mathbb{E}_x[T_j] \leqslant \mathbb{E}_{x,-j}[T_j] + \epsilon T/(2k\sigma) \cdot \sqrt{\mathbb{E}_{x,-j}[T_j]}$.

Next, we need the following lemma that we prove in Section 4.2.

Lemma 7. *In the conditions of Lemma 4 and by the construction above, we have*

$$\frac{1}{n^k} \sum_{x \in S} \mathbb{E}_{x,-j}\left[T_j\right] = \frac{T}{n} \ .$$

For every problem $j = 1, \ldots, k$,

$$\frac{1}{n^k} \sum_{x \in S} \mathbb{E}_x[T_j] \leqslant \frac{1}{n^k} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j] + \frac{\epsilon T}{2k\sigma} \cdot \frac{1}{n^k} \sum_{x \in S} \sqrt{\mathbb{E}_{x,-j}[T_j]}$$

$$\leqslant \frac{1}{n^k} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j] + \frac{\epsilon T}{2k\sigma} \sqrt{\frac{1}{n^k} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j]}$$

$$\leqslant \frac{T}{2} + \frac{\epsilon T}{2\sigma} \sqrt{\frac{T}{kd}} \ ,$$

where the second inequality is by Jensen's inequality, and the last inequality is since $n \geqslant 2$ and $d = kn$. Let us now return to Eq. (5). We can lower bound the regret as

$$\mathrm{R}_T \geqslant \epsilon \left(kT - \sum_{j=1}^{k} \left(\frac{T}{2} + \frac{\epsilon T}{2\sigma} \sqrt{\frac{T}{kd}} \right) \right) = \epsilon kT \left(\frac{1}{2} - \frac{\epsilon}{2\sigma} \sqrt{\frac{T}{kd}} \right) \ .$$

For our choice of ϵ, we get that $\epsilon/(2\sigma)\sqrt{T/(kd)}$ is at most $1/4$, and so

$$\mathrm{R}_T \geqslant \sigma \sqrt{\frac{kd}{4T}} \cdot kT \left(\frac{1}{2} - \frac{1}{4} \right) = \frac{\sigma}{8} k^{3/2} \sqrt{dT} \ ,$$

as claimed. ∎

4.2. Proof of Lemma 7

Proof. For any choice $i_1^\star, i_2^\star, \ldots, i_k^\star$, let us denote by $x(i^\star)$ the corresponding $x^\star \in S$. Following Audibert et al. (2013), we consider

$$\sum_{x \in S} \mathbb{E}_{x,-j}\left[T_j\right] = \sum_{i_1^\star, \ldots, i_{j-1}^\star, i_{j+1}^\star, \ldots, i_k^\star} \sum_{i_j^\star} \mathbb{E}_{x(i^\star),-j}\left[T_j\right] \ .$$

Now, keeping $i_1^\star, \ldots, i_{j-1}^\star, i_{j+1}^\star, \ldots, i_k^\star$ fixed the distribution $\mathbb{P}_{x(i^\star),-j}$ is the same for any choice of $i_j^\star$ and therefore, since at every round of the game the learner must choose exactly one arm in the j'th problem, we must have $\sum_{i_j^\star} \mathbb{E}_{x(i^\star),-j}[T_j] = T$.

Putting it all together, we obtain

$$\sum_{i_1^\star, \ldots, i_{j-1}^\star, i_{j+1}^\star, \ldots, i_k^\star} \sum_{i_j^\star} \mathbb{E}_{x(i^\star),-j}\left[T_j\right] = \sum_{i_1^\star, \ldots, i_{j-1}^\star, i_{j+1}^\star, \ldots, i_k^\star} T = n^{k-1}T \ ,$$

and thus

$$\frac{1}{n^k} \sum_{x \in S} \mathbb{E}_{x,-j}\left[T_j\right] = \frac{1}{n^k} n^{k-1}T = \frac{T}{n} \ . \qquad \blacksquare$$

4.3. Proof of Theorem 5

Proof. Let $X_1, X_2, \ldots, X_T$ be the predictions of the learner against an environment that plays $L_1, L_2, \ldots, L_T$, and let $\hat{R}_T$ be the regret attained by the learner,

$$\hat{R}_T = \sum_{t=1}^{T} L_t \cdot X_t - \min_{x \in S} \sum_{t=1}^{T} L_t \cdot x \ .$$

Also define the pretend-regret obtained by playing $X_1, X_2, \ldots, X_T$ against an enivronment that plays $L'_1, L'_2, \ldots, L'_T$ as

$$\hat{R}'_T = \sum_{t=1}^{T} L'_t \cdot X_t - \min_{x \in S} \sum_{t=1}^{T} L'_t \cdot x \ .$$

Now note that if it happens that at every round t, all coordinates of L'_t are between 0 and 1, then $\hat{R}_T = \hat{R}'_T$. Denote this event by E. Then,

$$\mathbb{E}[\hat{R}'_T] \leqslant \mathbb{E}[\hat{R}_T] + kT \cdot \mathbb{P}[E^c] = R_T + kT \cdot \mathbb{P}[E^c] \tag{8}$$

where the inequality is true since the regret is at most kT with probability 1.

It thus remains to upper bound the probability that E does not occur. We will show that $\mathbb{P}[E^c] \leqslant \epsilon/8$, which by combining Eq. (8) and Lemma 4 would yield:

$$R_T \geqslant \frac{\sigma k^{3/2}\sqrt{dT}}{8} - \frac{\sigma k^{3/2}\sqrt{dT}}{16} = \frac{\sigma k^{3/2}\sqrt{dT}}{16} \ ,$$

as required. Now, for E to occur it suffices that $\epsilon \leqslant 1/4$ and that $Z_t \leqslant 1/4$ for every round t. Since

$$\epsilon = \sqrt{\frac{\sigma^2 kd}{4T}} \leqslant \sqrt{\frac{kd}{(192 + 96\log T)T}} \leqslant \sqrt{\frac{1}{192}} \leqslant \frac{1}{4} \ ,$$

by our choice of ϵ and σ and since $T \geqslant kd$ by assumption, we have that the probability $\mathbb{P}[E^c]$ is upper bounded by the probability that $Z_t > 1/4$ at some (at least one) round t. Employing the standard tail bound $\mathbb{P}(Z > x) \leqslant \exp(-x^2/2\sigma^2)$ for the normal distribution and the union bound, the latter is bounded by

$$T \cdot \mathbb{P}[Z_1 > 1/4] \leqslant T \exp\left(-\frac{1}{2\sigma^2}\left(\frac{1}{4}\right)^2\right)$$

$$= T \exp\left(-(6 + 3\log T)\right)$$

$$= e^{-6}\frac{1}{T^2} \ .$$

Therefore, for the probability that E does not occur to be at most $\epsilon/8$ it suffices to have

$$16e^{-6}\frac{1}{T^2} \leqslant \sqrt{\frac{1}{(192 + 96\log T)T}} \ .$$

Rearranging the terms it suffices to have $T^3 \geqslant 0.16 + 0.08\log T$, that holds for any $T \geqslant 1$. ∎

4.4. Proof of Lemma 6

Proof. Let us denote by $i_1^\star, \ldots, i_k^\star$ the locations of the nonzero coordinates of the random variable $x^\star$, arranged in increasing order. We next introduce the random variables $T_1, \ldots, T_k$, where each T_j is the number of times the learner played an x_t such that $x_t(i_j^\star) = 1$. For each $x \in S$, we introduce the notations $\mathbb{P}_x$ and $\mathbb{E}_x$ indicating probability and expectation with respect to the marginal distributions under which $x^\star = x$. Then,

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} L_t' \cdot x_t - \min_{x \in S} \sum_{t=1}^{T} L_t' \cdot x\right] \geqslant \epsilon \left(kT - \sum_{j=1}^{k} \frac{(n-k)!}{n!} \sum_{x \in S} \mathbb{E}_x[T_j]\right) , \qquad (9)$$

and in order to proceed, we need to upper bound $\mathbb{E}_x[T_j]$ for each j.

For every $x \in S$ and $j \in [k]$ we introduce a new distribution, which is the same as $\mathbb{P}_x$ except that the loss of coordinate $i_j^\star$ is also $1/2 + Z_t$. We shall refer to these new laws by $\mathbb{P}_{x,-j}$ and $\mathbb{E}_{x,-j}$. From now on the proof proceeds similarly to that of Lemma 4, with the exception that Lemma 7 is replaced by the following lemma whose proof can be found in Section 4.5.

Lemma 8. *In the conditions of Lemma 6 and by the construction above, we have*

$$\frac{(n-k)!}{n!} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j] \leqslant \frac{T}{n-k+1} .$$

Recall that $k \leqslant n/2$ by assumption, that in particular implies $n - k + 1 \geqslant 2$ as well as $n - k + 1 \geqslant n/2$. Therefore, for all $j = 1, 2, \ldots, k$,

$$\begin{aligned}
\frac{(n-k)!}{n!} \sum_{x \in S} \mathbb{E}_x[T_j] &\leqslant \frac{(n-k)!}{n!} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j] + \frac{\epsilon T}{2k\sigma} \cdot \frac{(n-k)!}{n!} \sum_{x \in S} \sqrt{\mathbb{E}_{x,-j}[T_j]} \\
&\leqslant \frac{(n-k)!}{n!} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j] + \frac{\epsilon T}{2k\sigma} \sqrt{\frac{(n-k)!}{n!} \sum_{x \in S} \mathbb{E}_{x,-j}[T_j]} \\
&\leqslant \frac{T}{2} + \frac{\epsilon T}{2k\sigma} \sqrt{\frac{2T}{n}} .
\end{aligned}$$

Let us now return to Eq. (9). Using the fact that $n = d/k$, we can lower bound the regret as

$$\mathrm{R}_T \geqslant \epsilon kT \left(\frac{1}{2} - \frac{\epsilon}{\sigma}\sqrt{\frac{T}{2kd}}\right) ,$$

which, by our choice of ϵ, allows us to obtain the desired lower bound. ∎

4.5. Proof of Lemma 8

Proof. Recall that we sample $x^\star$ uniformly at random from S, the set defined in Eq. (4), and denote by $U(S)$ the uniform distribution over S. Then, recalling the random variables $i_1^\star, i_2^\star, \ldots, i_k^\star$ we can compute

$$\mathbb{E}_{x^\star \sim U(S)} \mathbb{E}_{x^\star,-j}[T_j] ,$$

by conditioning on $i_1^\star, \ldots, i_{j-1}^\star, i_{j+1}^\star, \ldots, i_k^\star$ and taking the outer expectation only over $i_j^\star$.

Now, there are exactly $n - k + 1$ possible ways to choose $i_j^\star$ in order to complete a maximal matching. In addition, the distribution $\mathbb{P}_{x^\star, -j}$ is the same for any possible choice of $i_j^\star$, and since at every round of the game the learner must choose exactly one position for the j'th element, we must have

$$
\mathbb{E}_{x^\star \sim U(S)} \left[\mathbb{E}_{x^\star, -j} [T_j] \mid i_1^\star = i_1, \ldots, i_{j-1}^\star = i_{j-1}, i_{j+1}^\star = i_{j+1}, \ldots, i_k^\star = i_k \right]
$$

$$
= \frac{1}{n - k + 1} \sum_{x^\star \in S} \mathbf{1}_{\left[i_1^\star = i_1, \ldots, i_{j-1}^\star = i_{j-1}, i_{j+1}^\star = i_{j+1}, \ldots, i_k^\star = i_k \right]} \mathbb{E}_{x^\star, -j} [T_j]
$$

$$
\leqslant \frac{T}{n - k + 1} \ .
$$
■

5. Conclusion and open problems

In this paper, we gave a tight characterization of the optimal regret rate in bandit combinatorial optimization and proved that it grows as $\widetilde{\Theta}(k^{3/2}\sqrt{dT})$, disproving the conjectures of Cesa-Bianchi and Lugosi (2012) and Audibert et al. (2013). Our lower bounds apply to important instances of the framework, including the bandit versions of the online shortest path and the online ranking problems.

An interesting direction for future work is to explore instance-specific bounds, i.e., bounds that depend on the structure of the specific action set S used. What are the geometric and combinatorial properties of the set S that dictate the optimal rate of regret in the induced learning problem? In particular, in the specific context of the bandit shortest path problem, what are the graph-theoretic properties of the network that govern the difficulty of the online problem? Even in extremely simple graphs, such as the two-dimensional directed grid over n^2 nodes (where the s and t nodes are located in two opposite corners), characterizing the optimal rate of regret remains an open problem. We suspect such problems to be non-trivial already in full-information online combinatorial optimization, but expect the bandit setting to be particularly challenging.

For the problem of online ranking, Theorem 3 handles the case of $k \times n$ permutations in which k is smaller than n. However, quantifying the rate of regret in the important case of full permutations (i.e., with $k = n$) remains an open problem. In particular, is the optimal regret $\Theta(n^2\sqrt{T})$ in this setting?

Acknowledgements

References

Jacob Duncan Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory*, 2008.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53. ACM, 2004.

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

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *COLT*, volume 23, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(Oct): 2369–2403, 2007.

Elad Hazan and Zohar Karnin. Volumetric spanners: An efficient exploration basis for learning. *Journal of Machine Learning Research*, 17:1–34, 2016.

David P Helmbold and Manfred K Warmuth. Learning permutations with exponential weights. *Journal of Machine Learning Research*, 10(Jul):1705–1736, 2009.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Satyen Kale, Lev Reyzin, and Robert E Schapire. Non-stochastic bandit slate problems. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2010.

H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *International Conference on Computational Learning Theory*, pages 109–123. Springer, 2004.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of The 28th Conference on Learning Theory*, pages 1360–1375, 2015.

Gergely Neu and Gábor Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *Journal of Machine Learning Research*, 17(154):1–21, 2016.

Ohad Shamir. On the complexity of bandit linear optimization. In *Proceedings of The 28th Conference on Learning Theory*, pages 1523–1551, 2015.

Eiji Takimoto and Manfred K Warmuth. Path kernels and multiplicative updates. *Journal of Machine Learning Research*, 4(Oct):773–818, 2003.