JMLR: Workshop and Conference Proceedings 83:1–25, 2018 Algorithmic Learning Theory 2018

Bandit Regret Scaling with the Effective Loss Range

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Dipartimento di Informatica
Università degli Studi di Milano
Milano, Italy

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, Israel

Editor: Mehryar Mohri and Karthik Sridharan

Erratum (updated January 2020)

The results of Sec. 4, except Theorem 10, are incorrect as stated, due to a crucial bug in the proof of Theorem 7 (in particular, the claim in Lemma 14 is not correct). However, the following modification of Theorem 7 is still correct:

Theorem 1 *Assume that in each round t, after choosing I_t the learner is told a number $a_t \geq 0$ such that $\min_i \ell_t(i) \geq a_t$. Then Exp3 performing updates based on loss vectors $\widetilde{\boldsymbol{\ell}}_t = \boldsymbol{\ell}_t - a_t\mathbf{1}$ achieves*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,K} \sum_{t=1}^{T} \ell_t(i) \leq \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}(\ell_t(i) - a_t)^2 .$$

The theorem is based on a simple shifting argument, and is an immediate corollary of Lemma 13 and the argument leading to Eq. (13) in the proof of Theorem 7. Compared to Theorem 7, we: (1) Need to make the different assumption that we are told a lower bound on the losses (e.g., the smallest loss), as opposed to any loss; (2) Define $\widetilde{\boldsymbol{\ell}}_t$ a bit differently; and (3) The regret bound now depends on the variation of the losses through $\sum_i (\ell_t(i) - a_t)^2$ rather than through a Laplacian-based bound. Thus, we still get a regret bound which is always better than the standard Exp3 regret bounds, improves as the per-round losses become more similar, and becomes zero if all the losses are exactly the same (as it should be expected). Finally, Theorem 10, which provides a Laplacian-based lower bound, is still correct but is no longer matched by a similar upper bound. It remains open whether such an upper bound exists for a bandit setting similar to the one considered here.

Thanks to Étienne de Montbrun for finding the problem in Lemma 14.

Abstract

We study how the regret guarantees of nonstochastic multi-armed bandits can be improved, if the effective range of the losses in each round is small (for example, the maximal difference between two losses in a given round). Despite a recent impossibility result, we show how this can be made possible under certain mild additional assumptions, such as availability of rough estimates of the losses, or knowledge of the loss of a single, possibly unspecified arm, at the end of each round. Along the way, we develop a novel technique which might be of independent interest, to convert any multi-armed bandit algorithm with regret depending on the loss range, to an algorithm with regret depending only on the effective range, while attaining better regret bounds than existing approaches.

1. Introduction

In the online learning and bandit literature, a recent and important trend has been the development of algorithms which are capable of exploiting "easy" data, in the sense of improved regret guarantees if the losses presented to the learner have certain favorable patterns. For example, a series of works have studied how the regret can be improved if the losses do not change much across rounds —e.g., (Chiang et al., 2012; Hazan and Kale, 2010, 2011; Karnin and Anava, 2016; Steinhardt and Liang, 2014); being simultaneously competitive w.r.t. both "hard" and "easy" data —e.g., (Seldin and Slivkins, 2014; Sani et al., 2014; Auer and Chiang, 2016; Bubeck and Slivkins, 2012); attain richer feedback on the losses —e.g., (Alon et al., 2014), have some predictable structure (Rakhlin and Sridharan, 2013), and so on. In this paper, we continue this research agenda in a different direction, focusing on improved regret performance in nonstochastic settings with partial feedback where the learner has some knowledge about the variability of the losses *within* each round.

In the full information setting, where the learner sees the entire vector $\ell_t(1), \ldots, \ell_t(K)$ of losses after each round t, it is possible to obtain regret bounds of order $\varepsilon\sqrt{T \log K}$ scaling with the unknown *effective range* $\varepsilon = \max_{t,i,j} |\ell_t(i) - \ell_t(j)|$ of the losses[1] (Cesa-Bianchi et al., 2007, Corollary 1). Unfortunately, the situation in the bandit setting, where the learner only observes the loss of the chosen arm, is quite different. A recent surprising result (Gerchinovitz and Lattimore, 2016, Corollary 4) implies that in the bandit setting, the standard $\Omega\big(\sqrt{KT}\big)$ regret lower bound holds, even when $\varepsilon = \mathcal{O}\big(\sqrt{K/T}\big)$. The proof defines a process where losses are kept ε-close to each other, but where the values oscillate unpredictably between rounds. Since the learner observes the loss of just one arm every round, the closeness between the losses across arms cannot be utilized. Based on this result, one may think that it is impossible to attain improved bounds in the bandit setting which depend on ε, or some other measure of variability of the losses across arms. In this paper, we study situations where partial information about the losses allows one to circumvent the impossibility result in some interesting ways. We analyze two specific settings: one in which the learner can roughly estimate in advance the actual loss value of each arm, and one where she knows the exact loss of some arbitrary and unknown arm.

In order to motivate the first setting, consider a scenario where the learner knows each arm's loss up to a certain precision (which may be different for each arm). For example, in the context of stock prices (Hazan and Kale, 2009; Abernethy et al., 2013) the learner may

1. This is a simplified form of a stronger bound of order $\sqrt{(\log K) \sum_t \varepsilon_t^2}$, where $\varepsilon_t = \max_{i,j} |\ell_t(i) - \ell_t(j)|$.

have a stochastic model providing some estimates of the loss means for the next round, and we would like to take advantage of that. In other cases, the losses of arms may be in $[0, 1]$ most of the time (perhaps with one arm tending to perform better than the others, same as in standard online learning), but there are a few arms which occasionally and predictably get a very high loss. For example, in routing the learner may know in advance that some route is down, and a large loss is incurred if that route is picked. In this scenario, a reasonable algorithm should be able to occasionally avoid picking that route; however, that breaks the regret guarantees of standard expert/bandit algorithms, which typically require each arm to be chosen with some positive probability. In the resulting regret bounds, it is difficult to avoid at least some dependence on the highest loss values.

To formalize these scenarios and considerations, we study a setting where for each arm i at round t, the learner is told that the loss will be in $\big[m_t(i) - \varepsilon_t(i), m_t(i) + \varepsilon_t(i)\big]$ for some $m_t(i), \varepsilon_t(i)$. In this setting, we show a generic reduction, which allows one to convert any algorithm for bounded losses and certain feedbacks (bandit feedback being just a special case), to an algorithm with regret depending only on the effective range of the losses (that is, only on $\varepsilon_t(i)$, independent of $m_t(i)$). Concretely, taking the simple case where the loss of each arm i at each round t is in $\big[m_t(i) - \varepsilon, m_t(i) + \varepsilon\big]$ for some $m_t(i)$ and fixed ε, and assuming the step size is properly chosen, we can get a regret bound of $\widetilde{\mathcal{O}}\big(\varepsilon\sqrt{KT}\big)$ for the bandit feedback, completely independent of $m_t(i)$ and the losses' actual range. Note that this has the desired behavior that as $\varepsilon \to 0$, the regret also converges to zero (in the extreme case where $\varepsilon = 0$, the learner essentially knows the losses in advance, and hence can avoid any regret). With full information feedback (where the entire loss vector is revealed at the end of each round), we can use the same technique to recover the regret bound of $\mathcal{O}\big(\varepsilon\sqrt{T\log K}\big)$. We also provide lower bounds, partially matching our upper bounds and demonstrating their tightness in certain regimes.

We note that the setting we study is a special case of the predictable sequences setting studied by Rakhlin and Sridharan (2013), who propose an algorithm with a regret bound which (in the setting above) scales as $\widetilde{\mathcal{O}}(\varepsilon\sqrt{K^3T})$. Our result has a better dependence on the number of arms K —which we also show to be optimal— and our reduction can be applied to any algorithm, rather than the specific one proposed in their paper. On the flip side, the algorithm proposed there is applicable to the more general setting of bandit linear optimization, and does not require the range parameter ε to be known in advance (see Sec. 3 for a more detailed comparison).

A second setting we study, motivating partial knowledge about the loss vectors, is the following. Consider a system for recommending products to visitors of some website. Say that two products are similar if the typical visitor tends to like them both or dislike them both. Hence, a plausible assumption on the similarity graph over the set of products would be that the likelihood of purchase (or any related index of the visitor's behavior) is a smooth function over this graph. Formally, the loss vectors $\boldsymbol{\ell}_t = \big(\ell_t(1), \ldots, \ell_t(K)\big)$ at each round t are such that $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell} = \sum_{(i,j)\in E_t} \big(\ell_t(i) - \ell_t(j)\big)^2$ is small, where L_t is the Laplacian matrix associated with some graph over the arms with edge set E_t. In this setting, we provide improved bandit regret bounds depending on the spectral properties of the Laplacian. To circumvent the impossibility result of Gerchinovitz and Lattimore (2016) mentioned earlier, we make the reasonable assumption that at the end of each round, the learner is given an "anchor point", corresponding to the loss of some unspecified arm. In our motivating example, the recommender system may assume, for instance, that each visitor has some product that she most likely won't buy. Using a simple modification of the Exp3 algorithm, we show that a regret bound of order (ignoring log factors)

$$\sqrt{\sum_{t=1}^{T} \min_{L_t} \left(1 + \frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}}{\lambda_2(L_t)}\right)}$$

where $L_1, \ldots, L_T$ are Laplacians of simple and connected graphs, and each quantity $\lambda_2(L_t) \in (0, K]$ is the smallest nonzero eigenvalue of L_t (also known as the algebraic connectivity number of the graph represented by L_t). If the learner is told the minimal loss at every round (rather than any loss), this bound can be improved to order of

$$\sqrt{\sum_{t=1}^{T} \min_{L_t} \frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}}{\lambda_2(L_t)}}$$

(again, ignoring log factors) which vanishes, as it should, when $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell} = 0$ for all t; that is, when all arms share the same loss value. Compared to the state-of-the-art bandit regret bound $\sqrt{\sum_t \|\boldsymbol{\ell}_t\|^2}$, we easily show that $\min_{L_t} \boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t / \lambda_2(L_t)$ is never larger than $\|\boldsymbol{\ell}_t\|^2$, and is much smaller than $\|\boldsymbol{\ell}_t\|^2$ when the loss components $\ell_t(1), \ldots, \ell_t(K)$ tend to be close to each other.

We also provide a lower bound, showing that this upper bound is the best possible (up to log factors) in the worst case. Although our basic results pertain to connected graphs, using the range-dependent reductions discussed earlier we show it can be applied to graphs with multiple connected components and anchor points.

The paper is structured as follows: In Sec. 2, we formally define the standard experts/bandit online learning setting, which is the focus of our paper, and devote a few words to the notation we use. In Sec. 3, we discuss the situation where each individual loss is known to lie in a certain range, and provide an algorithm as well as upper and lower bounds on the expected regret. In Sec. 4, we consider the setting of smooth losses (as defined above). Some of the proofs are presented in the appendix.

2. Setting and notation

The standard experts/bandit learning setting (with nonstochastic losses) is phrased as a repeated game between a learner and an adversary, defined over a fixed set of K arms/actions. Before the game begins, the adversary assigns losses for each of K arms and each of T rounds (this is also known as an oblivious adversary, as opposed to a nonoblivious one which sets the losses during the game's progress). The loss of arm i at round t is defined as $\ell_t(i)$, and following standard convention, is assumed w.l.o.g. to lie in $[0, 1]$. At the beginning of each round, the learner chooses an arm $I_t \in \{1, \ldots, K\}$, and receives the associated loss $\ell_t(i)$. With bandit feedback, the learner then observes only her own loss $\ell_t(I_t)$, whereas with full information feedback, the learner gets to observe $\ell_t(i)$ for all i. The learner's goal is to minimize the expected regret (sometimes denoted as pseudo-regret), defined as

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\ldots,K} \sum_{t=1}^{T} \ell_t(i)$$

where the expectation is over the learner's internal randomness. We use $\mathbb{I}\{A\}$ to denote the indicator of the event A, and let log denote the natural logarithm. Given an (undirected) graph over K nodes, its Laplacian L is defined as the $K \times K$ matrix where $L_{i,i}$ equals the degree of node i, $L_{i,j}$ for $i \neq j$ equals -1 if node i is adjacent to node j, and 0 otherwise. We let $\lambda_2(L)$ denote the smallest nonzero eigenvalue of L. This is also known as the algebraic connectivity number, and is larger the more well-connected is the graph. In particular, $\lambda_2(L) = 0$ for disconnected graphs, and $\lambda_2(L) = K$ for the complete graph.

3. Rough estimates of individual losses

We consider a variant of the online learning setting presented in Sec. 2, where at the beginning of every round t, the learner is provided with additional side information in the form of $\big(m_t(i), \varepsilon_t(i)\big)$ for $i = 1, \dots, K$, with the guarantee that $\big|\ell_t(i) - m_t(i)\big| \leq \varepsilon_t(i)$ for all $i = 1, \dots, K$. We then propose an algorithmic reduction, which allows to convert any regret-minimizing algorithm $\mathcal{A}$ (with some generic feedback), to an algorithm with regret depending on $\varepsilon_t(i)$ and independent of $m_t(i)$. We assume that given a loss vector $\boldsymbol{\ell}_t$ and chosen action I_t, the algorithm $\mathcal{A}$ receives as feedback some function $f_t\big(\boldsymbol{\ell}_t, I_t\big)$. For example, if $\mathcal{A}$ is an algorithm for the multi-armed bandits setting, then $f_t\big(\boldsymbol{\ell}_t, I_t\big) = \boldsymbol{\ell}_t(I_t)$, whereas if $\mathcal{A}$ is an algorithm for the experts setting, $f_t\big(\boldsymbol{\ell}_t, I_t\big) = \boldsymbol{\ell}_t$. In our reduction, $\mathcal{A}$ is sequentially fed, at the end of each round t, with $f_t\big(\widetilde{\boldsymbol{\ell}}_t, \widetilde{I}_t\big)$ (where $\widetilde{\boldsymbol{\ell}}_t$ and $\widetilde{I}_t$ are not necessarily the same as the actual loss vector $\boldsymbol{\ell}_t$ and actual chosen arm I_t), and returns a recommended arm $\widetilde{I}_{t+1}$ for the next round, which is used to choose the actual arm I_{t+1}.

To formally describe the reduction, we need a couple of definitions. For all t, let

$$j_t \in \operatorname*{argmin}_{i=1,\dots,K} \big\{m_t(i) - \varepsilon_t(i)\big\}$$

denote the arm with the lowest potential loss, based on the provided side-information (if there are ties, we choose the one with smallest $\varepsilon_t(i)$, and break any remaining ties arbitrarily). At round t define any arm i as "bad" if $m_t(i) - \varepsilon_t(i) > m_t(j_t) + \varepsilon_t(j_t)$ and "good" if $m_t(i) - \varepsilon_t(i) \leq m_t(j_t) + \varepsilon_t(j_t)$. Intuitively, "bad" arms are those which cannot possibly have the smallest loss in round t. For loss vector $\boldsymbol{\ell}_t$, define the transformed loss vector $\widetilde{\boldsymbol{\ell}}_t$ as

$$\widetilde{\ell}_t(i) = \begin{cases} \ell_t(i) - m_t(j_t) + \varepsilon_t(j_t) & \text{if } i \text{ is good} \\ 2\,\varepsilon_t(j_t) & \text{if } i \text{ is bad.} \end{cases}$$

It is easily verified that $\widetilde{\ell}_t(i) \in \big[0, 2(\varepsilon_t(i) + \varepsilon_t(j_t))\big]$ always. Hence, the range of the transformed losses does not depend on $m_t(i)$. The meta-algorithm now does the following at every round:

Crucially, note that in line 2d of the meta-algorithm (Algorithm 1) we assume that $f_t\big(\widetilde{\boldsymbol{\ell}}_t, \widetilde{I}_t\big)$ can be constructed based on $f_t\big(\boldsymbol{\ell}_t, I_t\big)$. For example, this is certainly true in the full information setting (as we are given $\boldsymbol{\ell}_t$, hence can explicitly compute $\widetilde{\boldsymbol{\ell}}_t$). This is also true in the bandit setting: If $\widetilde{I}_t$ is a "good" arm, then $I_t = \widetilde{I}_t$, hence we can construct $\widetilde{\ell}_t(\widetilde{I}_t) = \ell_t(I_t) - m_t(j_t) + \varepsilon_t(j_t)$ based on the feedback $\ell_t(I_t)$ actually given to the meta-algorithm. If $\widetilde{I}_t$ is a "bad" arm, then we can construct $\widetilde{\ell}_t(\widetilde{I}_t) = 2\,\varepsilon_t(j_t)$, since $\varepsilon_t(j_t)$ is given

Algorithm 1 (Meta-Algorithm)

1. Require: Base algorithm $\mathcal{A}$
2. For $t = 1, 2, \ldots$
   (a) Get an arm recommendation $\widetilde{I}_t$ from $\mathcal{A}$
   (b) If $\widetilde{I}_t$ is a good arm, then $I_t = \widetilde{I}_t$, else $I_t = j_t$
   (c) Choose arm I_t and get feedback $f_t\big(\boldsymbol{\ell}_t, I_t\big)$
   (d) Construct feedback $f_t\big(\widetilde{\boldsymbol{\ell}}_t, \widetilde{I}_t\big)$ and feed it to algorithm $\mathcal{A}$

to the meta-algorithm as side-information. In general, this framework can potentially be used for other partial-feedback settings as well.

The following key theorem implies that the expected regret of this meta-algorithm can be upper bounded by the expected regret of $\mathcal{A}$, with respect to the *transformed* losses $\widetilde{\boldsymbol{\ell}}_t$ whose range is independent of $m_t(i)$.

Theorem 2 *Suppose (without loss of generality) that in Algorithm 1 the index $\widetilde{I}_t$ given by $\mathcal{A}$ is chosen at random by sampling from a probability distribution $\widetilde{p}_t(1), \ldots, \widetilde{p}_t(K)$. Let $p_t(1), \ldots, p_t(K)$ be the induced distribution[2] of I_t. Then for any fixed arm $a \in \{1, \ldots, K\}$, it holds that*

$$\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)\ell_t(i) - \sum_{t=1}^{T}\ell_t(a) \;\le\; \sum_{t=1}^{T}\sum_{i=1}^{K} \widetilde{p}_t(i)\widetilde{\ell}_t(i) - \sum_{t=1}^{T}\widetilde{\ell}_t(a)\ . \tag{1}$$

This implies in particular that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(I_t)\right] - \sum_{t=1}^{T}\ell_t(a) \;\le\; \mathbb{E}\left[\sum_{t=1}^{T}\Big(\widetilde{\ell}_t(\widetilde{I}_t) - \widetilde{\ell}_t(a)\Big)\right]$$

where the expectation is over the possible randomness of the algorithm $\mathcal{A}$. Moreover, $\widetilde{\ell}_t(i) \in \big[0, 2(\varepsilon_t(j_t) + \varepsilon_t(i))\big]$ for any good i, and $\widetilde{\ell}_t(i) = 2\,\varepsilon_t(j_t)$ for any bad i.

Proof The proof carefully utilizes the definition of the transformed losses and actions. The main challenge is to prove Eq. (1). The in-expectation bound follows by applying expectations on both sides of the inequality, and noting that conditioned on rounds $1, \ldots, t-1$, the conditional expectation of $\ell_t(I_t)$ equals $\sum_{i=1}^{K} p_t(i)\ell_t(i)$, and the conditional expectation of $\widetilde{\ell}_t(\widetilde{I}_t)$ equals $\sum_{i=1}^{K}\widetilde{p}_t(i)\widetilde{\ell}_t(i)$. Also, the statement on the range of each $\widetilde{\ell}_t(i)$ is immediate from the definition of $\widetilde{\ell}_t(i)$ and Eq. (4) below.

2. By definition of the meta-algorithm, we have $p_t(i) = \widetilde{p}_t(i)$ if $i \neq j_t$ is good, $p_t(i) = 0$ if i is bad, and $p_t(j_t) = \widetilde{p}_t(j_t) + \sum_{i \text{ is bad}} \widetilde{p}_t(i)$.

We now turn to prove Eq. (1). By adding and subtracting terms, it is sufficient to prove that

$$\sum_t \sum_i \big(p_t(i)\ell_t(i) - m_t(j_t) + \varepsilon_t(j_t)\big) - \sum_t \big(\ell_t(a) - m_t(j_t) + \varepsilon_t(j_t)\big) \\ \leq \sum_{t,i} \widetilde{p}_t(i)\widetilde{\ell}_t(i) - \sum_t \widetilde{\ell}_t(a) \ . \tag{2}$$

We will rely on the following facts, which are immediate from the definition of good and bad arms: Any bad arm i must satisfy

$$\ell_t(i) \geq m_t(j_t) + \varepsilon_t(j_t) \tag{3}$$

and any good arm i must satisfy

$$m_t(j_t) - \varepsilon_t(j_t) \leq \ell_t(i) \leq m_t(j_t) + \varepsilon_t(j_t) + 2\varepsilon_t(i) \ . \tag{4}$$

Based on this, we have the following two claims, whose proof is provided in the appendix.

Claim 1. For any fixed arm a, $\widetilde{\ell}_t(a) \leq \ell_t(a) - m_t(j_t) + \varepsilon_t(j_t)$.

Claim 2.

$$\sum_{i=1}^{K} p_t(i)\ell_t(I_t) - m_t(j_t) + \varepsilon_t(j_t) \leq \sum_{i=1}^{K} \widetilde{p}_t(i)\widetilde{\ell}_t(i)$$

where

$$p_t(i) = \begin{cases} \widetilde{p}_t(i) & i \neq j_t \text{ and } i \text{ is good} \\ 0 & i \neq j_t \text{ and } i \text{ is bad} \\ \widetilde{p}_t(j_t) + \sum_{i \text{ is bad}} p(i) & i = j_t. \end{cases}$$

Combining the two claims and summing over t, we get Eq. (2) as required. ∎

Since the range of $\widetilde{\ell}_t$ is independent of m_t, we get a regret bound for our meta-algorithm which depends only on ε_t. This is exemplified in the following two corollaries:

Corollary 3 *With bandit feedback and using Exp3 as the algorithm $\mathcal{A}$ (with step size η), the expected regret of Algorithm 1 is*

$$\mathcal{O}\left(\frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \left(K\varepsilon_t(j_t)^2 + \sum_{i \in G_t} \varepsilon_t(i)^2\right)\right)$$

where $G_t \subseteq \{1, \ldots, K\}$ is the set of "good" arms at round t, and j_t is the arm with lowest potential loss $m_t(i) - \varepsilon_t(i)$ in round t.

The optimal choice of η leads to a regret of order $\sqrt{(\log K)\sum_{t=1}^{T}\left(K\varepsilon_t(j_t)^2 + \sum_{i\in G_t}\varepsilon_t(i)^2\right)}$. This recovers the standard Exp3 bound in the case $m_t(i) = \varepsilon_t(i) = \frac{1}{2}$ (i.e., the standard setting where the losses are only known to be bounded in $[0, 1]$), but can be considerably better if the $\varepsilon_t(i)$ terms are small, or the $m_t(i)$ terms are large. We also note that the $\log K$ factor can in principle be removed, i.e., by using the implicitly normalized forecaster of Audibert and Bubeck (2009) with appropriate parameters. A similar corollary can be obtained in the full information setting, using a standard algorithm such as Hedge (Freund and Schapire, 1995).

Corollary 4 *With full information feedback and using Hedge as the algorithm $\mathcal{A}$ (with step size η), the expected regret of the meta-algorithm is*

$$\mathcal{O}\left(\frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \max_{i=1,\ldots,K} \varepsilon_t(i)^2\right).$$

The optimal choice of η leads to regret of order $\sqrt{(\log K)\sum_{t=1}^{T}\max_i \varepsilon_t(i)^2}$. As in the bandit setting, our reduction can be applied to other algorithms as well, including those with more refined loss-dependent guarantees —e.g., (Steinhardt and Liang, 2014) and references therein.

Finally, we note that Theorem 2 can easily be used to provide high-probability bounds on the actual regret $\sum_{t=1}^{T}\ell_t(I_t) - \sum_{t=1}^{T}\ell_t(a)$, rather than just bounds in expectation, as long as we have a high-probability regret bound for $\mathcal{A}$. This is due to Eq. (1), and can be easily shown using standard martingale arguments.

3.1. Related work

As mentioned in the introduction, a question similar to ours was studied in (Rakhlin and Sridharan, 2013) —see also (Hazan and Kale, 2011) for earlier results— under the name of learning with predictable sequences. Unlike our setting, however, Rakhlin and Sridharan (2013) do not require knowledge of $\varepsilon_t(i)$, and consider a more generic setting corresponding to bandit linear optimization. Assuming the step size is chosen appropriately, they provide algorithms with expected regret bounds scaling as

$$\underbrace{\sqrt{K^2(\log K)\sum_{t=1}^{T}\sum_{i=1}^{K}\varepsilon_t(i)^2}}_{\text{(bandit feedback)}} \qquad \text{and} \qquad \underbrace{\sqrt{(\log K)\sum_{t=1}^{T}\max_i \varepsilon_t(i)^2}}_{\text{(full information feedback)}}$$

Comparing these bounds to Corollaries 3 and 4, we see that we obtain a similar bound in the full information setting, and a significantly better bound in the bandit setting: The dependence on K is improved (by a factor of between $\sqrt{K}$ and K, depending on the $\varepsilon_t(i)$ values), and there is a better dependence on the $\varepsilon_t(i)$ values if $\varepsilon_t(j_t)$ is not too large and the number of "good" arms tends to be small. In fact, as we show in the next subsection, our bound is optimal in certain regimes. Also, our algorithmic approach is based on a reduction, which can be applied in principle to any algorithm and to general families of feedback settings, rather than a specific algorithm. On the flip side, the results in Rakhlin and Sridharan (2013) are tailored to the more general setting of linear online optimization, and do not require knowing $\varepsilon_t(i)$ in advance.

Another related line of work is path-based bounds, where it is assumed that the losses $\ell_t(i)$ tend to vary slowly with t, and $\ell_{t-1}(i)$ can provide a good estimate of $\ell_t(i)$. This can be linked to our setting by taking $m_t(i) = \ell_{t-1}(i)$, and $\varepsilon_t(i)$ be some known upper bound on $|\ell_t(i) - \ell_{t-1}(i)|$. However, implementing this requires the assumption that $\boldsymbol{\ell}_{t-1}$ is revealed at the next round t, which does not fit the bandit setting. Thus, it is difficult to directly compare these results to ours. Most work on this topic has focused on the full information feedback setting —see (Steinhardt and Liang, 2014) and references therein— and the bandit setting was studied for instance in (Hazan and Kale, 2011).

3.2. Lower bound

We now turn to consider the tightness of our results. Since the focus of this paper is to study the variability of the losses across arms, rather than across time, we will consider for simplicity the case where $\varepsilon_t(j)$ are fixed for all $t = 1, \ldots, T$ (hence the t subscript can be dropped).

Theorem 5 *Fix $T, K > 1$ and nonnegative $\varepsilon(1), \ldots, \varepsilon(K)$ such that $\min_{j\,:\,\varepsilon(j)>0} \varepsilon(j)^2 \geq \frac{2}{T}\sum_j \varepsilon(j)^2$. Then there exists fixed parameters $m(j)$ for $j = 1, \ldots, K$ such that the following holds: For any (possibly randomized) learner strategy A, there exists a loss assignment satisfying $|\ell_t(j) - m(j)| \leq \varepsilon(j)$ for all t, j, such that*

$$\mathbb{E}_A\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{j=1,\ldots,K} \sum_{t=1}^{T} \ell_t(j) \geq \begin{cases} c\sqrt{T\sum_{j=1}^{K}\varepsilon(j)^2} & \text{with bandit feedback} \\ c\sqrt{T \max\limits_{j=1,\ldots,K} \varepsilon(j)^2} & \text{with full information feedback} \end{cases}$$

where $c > 0$ is a universal constant.

The theorem implies that the dependencies on $\sum_j \varepsilon(j)^2$ and $\max_j \varepsilon(j)$ in our upper bounds (in the bandit and full information case, respectively) cannot be improved in general. Moreover, when $\varepsilon(j)$ is the same for all j, our bound in the bandit setting is tight up to logarithmic factors.

Remark 6 *The lower bound construction in the bandit setting is such that all arms are potentially "good" in the sense used in Corollary 3, and hence $\sum_{j=1}^{K} \varepsilon(j)^2$ coincides with $\sum_{j\in G_t} \varepsilon(j)^2$ (recall G_t is the set of "good" arms at time t). If one wishes to consider a situation where some arms j are "bad", and obtain a bound dependent on $\sum_{j\in G_t} \varepsilon(j)^2$, one can simply pick some sufficiently large values $m(j)$ for them, and ignore their contribution to the regret in the lower bound analysis.*

The proof (provided in the appendix) is conceptually similar to the standard regret lower bound for nonstochastic multi-armed bandits —see (Bubeck and Cesa-Bianchi, 2012)— where the losses are generated stochastically, with one randomly chosen and hard-to-find arm having a slightly smaller loss in expectation. However, we utilize a more involved stochastic process to generate the losses as well as to choose the better arm, which takes the values of $\varepsilon(i)$ into account.

The lower bound leaves open the possibility of removing the dependence on $K\varepsilon_t(j_t)^2$ in the upper bound. This term is immaterial when $K\varepsilon_t(j_t)^2$ is comparable to, or smaller than $\sum_{i\in G_t} \varepsilon_t(i)^2$ (e.g., if most arms are good, and $\varepsilon_t(i)$ is about the same for all i), but can be significant otherwise. This question is left to future work.

4. Smooth losses

We now turn to consider a different set of assumptions, which utilize smoothness of the losses across arms to get improved regret guarantees, and avoids the lower bound of Gerchinovitz and Lattimore (2016). Specifically, we consider a situation where the learner is given (or can compute) an "anchor point" a_t at the end of each round t, which equals the loss of *some* arm at round t, independent of the learner's randomness at that round. Importantly, the learner need not even know which arm has this loss. For example, it is often reasonable to assume that there is always some arm which attains a minimal loss of 0, or some arm which attains a maximal loss of 1. In that case, instead of estimating losses $\ell_t(i)$ in $[0,1]$, it is enough to estimate losses of the form $\ell_t(i) + (1 - a_t)$, which may lie in a much narrower range if $|\ell_t(i) - a_t|$ tends to be small.

To see why this "anchor point" side-information circumvents the lower bound of Gerchinovitz and Lattimore (2016), we briefly discuss their construction (in a slightly simplified manner): The authors consider a situation where the losses are generated stochastically and independently at each round according to $\ell_t(i) = \mathrm{clip}_{[0,1]}\left(Z_t - \Delta\mathbb{I}\{i = i^*\}\right)$, with Z_t being a standard Gaussian random variable, $\Delta = \Theta(\sqrt{K/T})$, and i^* being some arm chosen uniformly at random. Hence, at every round, the loss of arm i^* is smaller by an amount of order $\Omega(\sqrt{K/T})$ than the loss of all other arms. Getting an expected regret smaller than $\Omega(\sqrt{KT})$ would then imply detecting i^*. However, since the learner observes only a *single* loss every round, the similarity of the losses for different arms at a given round does not help much. In contrast, if the learner had access to the loss a_t of any fixed arm (independent of the learner's randomness), she could easily detect i^* in $\mathcal{O}(K)$ rounds, simply by maintaining a "feasible set" $\mathcal{I}$ of possible arms, picking arms $i \in \mathcal{I}$ at random, and removing it from $\mathcal{I}$ if $\ell_t(i) - a_t$ is positive. This process ends once $\mathcal{I}$ contains a single arm, which must be i^*.

To formalize our setting in a flexible manner, we follow a graph-based approach inspired by Valko et al. (2014). Consider the standard Exp3 bound, which in its tightest form is of order $\sqrt{(\log K)\sum_t \|\boldsymbol{\ell}_t\|^2}$ —see Lemma 13 in the Appendix. We will show how each term $\|\boldsymbol{\ell}_t\|^2$ can be replaced by the smaller term

$$\min_{L_t} \frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)} \;=\; \min_{L_t} \frac{\sum_{(i,j)\in E_t}\left(\ell_t(i) - \ell_t(j)\right)^2}{\lambda_2(L_t)},$$

where the minimum is over the Laplacians L_t of all possible simple and connected graphs over K vertices (with corresponding edge set E_t), and $\lambda_2(L_t)$ is the smallest nonzero eigenvalue of L_t. The smaller is $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t$, the more similar are the losses, on average. Crucially, note that the learner *need not* have explicit knowledge of L_t. The only thing we do expect the learner to know (at the end of each round t) is the "anchor point" a_t as described above. We also note that this setting is quite distinct from the graph bandits setting of (Mannor and Shamir, 2011; Alon et al., 2014), which also assumes a graph structure over the bandits, but this graph encodes what feedback the learner receives, as opposed to encoding similarities between the losses themselves.

We now turn to describe the algorithm and associated regret bound. The algorithm itself is very simple: Run a standard multiarmed bandits algorithm suitable for our setting —for example, Exp3 (Auer et al., 2002)— using the shifted losses $\widetilde{\ell}_t(i) = \ell_t(i) + 1 - a_t$. The associated regret guarantee is formalized in the following theorem (from now on, let $\mathbf{1} = (1, \ldots, 1)$ be the all-ones vector).

Theorem 7 *Assume that in each round t, after choosing I_t the learner is told a number a_t chosen by the oblivious adversary and such that there exists some arm k_t with $\ell_t(k_t) = a_t$.*

Then Exp3 performing updates based on loss vectors $\widetilde{\boldsymbol{\ell}}_t = \boldsymbol{\ell}_t + (1 - a_t)\mathbf{1}$ *achieves*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,K} \sum_{t=1}^{T} \ell_t(i) \le \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \min_{L_t}\left(1 + \frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)}\right)$$

where each L_t *is the Laplacian of an arbitrary simple and connected graph on* $\{1,\dots,K\}$.

The proof (provided in the appendix) is based on Euclidean-norm regret bounds for the Exp3 algorithm, combined with a careful analysis of the associated quantities based on the Laplacian $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t$.

We now show how doubling-trick arguments, combined with Theorem 7, deliver a regret bound of order

$$\sqrt{(\log K)\sum_{t=1}^{T} \min_{L_t}\left(1 + \frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)}\right)}\,.$$

More specifically, we apply the doubling trick to Exp3 in order to obtain a regret bound scaling as $\sqrt{(\log K)\sum_t \|\widetilde{\ell}_t\|^2}$. As all of our results depend on upper bounding $\|\widetilde{\ell}_t\|^2$, the argument applies in particular to the settings of Theorem 7 and Corollary 9.

Theorem 8 *Consider the algorithm that operates by running Exp3 for* T_r *consecutive time steps with parameter* $\eta_r = \sqrt{(2\log K)/2^r}$ *for each* $r = 0, 1, \dots$ *while monitoring the observable random quantity*[3] $Q_s = \sum_{i=1}^{K} p_s(i)\widehat{\ell}_s(i)^2$. *Whenever* $\sum_{t \in T_r} Q_t > 2^r$ *is detected while Exp3 is running with* $\eta = \eta_r$, *the algorithm restarts Exp3 with a new parameter* $\eta = \eta_{r+1}$. *Then the regret of this algorithm satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(k) \le \log_2\big(KT + 1\big) + 4\sqrt{2(\log K)\left(1 + \sum_{t=1}^{T} \|\boldsymbol{\ell}_t\|^2\right)}\,.$$

The bound of Theorem 7 is not fully satisfying as it does not vanish when $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t = 0$ (which, assuming the graph is connected, implies that all losses are the same). The reason is that we need to add 1 to each loss component in order to guarantee that we do not end up with negative components when $\ell_t(k_t)\cdot\mathbf{1}$ is subtracted from $\boldsymbol{\ell}_t$. This is avoided when in each round t, the revealed loss $\ell_t(k_t)$ is the smallest component of $\boldsymbol{\ell}_t$, as formalized in the following corollary.

Corollary 9 *Assume that in each round* t, *after choosing* I_t *the learner is told* $a_t = \min_i \ell_t(i)$. *Then Exp3 performing updates using losses* $\widetilde{\boldsymbol{\ell}}_t = \boldsymbol{\ell}_t - a_t \cdot \mathbf{1}$ *achieves*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,K} \sum_{t=1}^{T} \ell_t(i) \le \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left(\min_{L_t} \frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)}\right)$$

where each L_t *is the Laplacian of an arbitrary simple and connected graph on* $\{1,\dots,K\}$.

3. Here $\widehat{\ell}_s(i) = \frac{\ell_s(i)}{p_s(i)}\mathbb{I}\{I_s = i\}$ are Exp3's standard importance-weighting loss estimates.

We leave the question of getting such a bound, without a_t being the smallest loss, as an open problem.

When η is tuned optimally (e.g., via Theorem 8), the bounds of Theorem 7 and Corollary 9 take, respectively, the form

$$\sqrt{(\log K)\sum_{t=1}^{T}\min_{L_t}\left(1+\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)}\right)} \quad \text{and} \quad \sqrt{(\log K)\sum_{t=1}^{T}\min_{L_t}\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)}}\ . \tag{5}$$

As a sanity check, we now show that the bound in the right-hand side of Eq. (5) is never worse than the $\sqrt{\log(K)\sum_t \|\ell_t\|^2}$ bound for Exp3 mentioned earlier (which does not require an anchor point assumption). To see this, recall that our bounds are achieved for all choices of $L_1,\ldots,L_T$ that correspond to simple and connected graphs. In particular, let L_t be the Laplacian of the K-clique for each t. Then L_t has all nonzero eigenvalues equal to K, and so $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t = K$. As $\lambda_2(L_t)$ is also equal to K, we have that $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t/\lambda_2(L_t) = \|\boldsymbol{\ell}_t\|^2$.

Finally, we show that for fixed graphs $L_t = L$, the bound in the right-hand side of Eq. (5) is tight in the worst-case up to log factors.

Theorem 10 *There exist universal constants c_1, c_2 such that the following holds: For any randomized algorithm, any $C > 0$, any $\lambda \in (0,1]$, and any sufficiently large K and T, there exists a K-node graph with Laplacian L satisfying $\lambda_2(L) \in [c_1\lambda, c_2\lambda]$ and an adversary strategy $\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T$, such that $\boldsymbol{\ell}_t^\top L \boldsymbol{\ell}_t \le C$ for all $t = 1,\ldots,T$, and the expected regret (w.r.t. the algorithm's internal randomization) is at least*

$$\Omega\left(\min\left\{\sqrt{K}, \frac{C}{\sqrt{\lambda_2(L)}}\right\}\sqrt{T}\right)$$

The proof is provided in the appendix. This theorem matches Eq. (5), assuming that $L_t = L$ for all t and that $\lambda_2(L) = \mathcal{O}(1)$. Note that the latter assumption is generally the interesting regime for $\lambda_2(L)$ (for example, $\lambda_2(L) \le 1$ as long as there is *some* node connected by a single edge). The proof is based on considering an "octopus" graph, composed of long threads emanating from one central node, and applying a standard bandit lower bound strategy on the nodes at the ends of the threads.

4.1. Multiple connected components

The previous results of this section need the graph represented by L_t to be connected, in order for the guarantees to be non-vacuous. This is not just an artifact of the analysis: If the graph is not connected, at least some arms can have losses which are arbitrarily different than other arms, and the anchor point side information is not necessarily useful. Indeed, if there are multiple connected components, then $\lambda_2 = 0$ and our bounds become trivial. Nevertheless, we now show it is still possible to get improved regret performance in some cases, as long as the learner is provided with anchor point information on each connected component of the graph.

We assume that at every round t, there is some graph defined over the arms, with edge set E_t. However, here we assume that this graph may have multiple connected components (indexed by s in some set $\mathcal{C}_t$). For each connected component s, with associated Laplacian

$L_t(s)$, we assume the learner has access to an anchor point $m_t(s)$. Unlike the case discussed previously, here the anchor points may be different at different components, so a simple shifting of the losses (as done in Sec. 4) no longer suffices to get a good bound. However, the anchor points still allow us to compute some interval, in which each loss must lie, which in turn can be plugged into the algorithmic reduction presented in Sec. 3. This is formalized in the following lemma, whose proof is provided in the appendix.

Lemma 11 *For any connected component $s \in \mathcal{C}_t$, and any arm i in that component, $\left|\ell_t(i) - m_t(s)\right| \leq \sqrt{\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2\left(L_t(s)\right)}}$.*

Based on this lemma, we know that any arm at any connected component s has values in $\left[m_t(s) - \sqrt{\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2\left(L_t(s)\right)}}, m_t(s) + \sqrt{\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2\left(L_t(s)\right)}}\right]$. Using this and applying Corollary 3, we have the following result.

Theorem 12 *For any fixed arm j, the algorithm described in Corollary 3 satisfies*

$$\mathbb{E}\left[\sum_t \ell_t(I_t) - \sum_t \ell_t(j)\right] \;\leq\; \frac{\log(K)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\left(\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2\left(L_t(s_{\min})\right)} + \sum_{s\in G_t}\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2\left(L_t(s)\right)} N_t(s)\right)$$

where each L_t is the Laplacian of an arbitrary simple graph on $\{1, \ldots, K\}$, $N_t(s)$ is the number of arms in connected component s, and $s_{\min}$ is a connected component s for which $m_t(s) - C_t \Big/ \sqrt{\lambda_2\left(L_t(s)\right)}$ is smallest.

This allows us to get results which depend on the Laplacians $L_t(s)$, even when these subgraphs are disconnected. We note however that this theorem does not recover the results of Sec. 4 when there is only one connected component, as we get $\frac{\log K}{\eta} + \frac{\eta(K+1)}{2}\sum_{t=1}^{T}\frac{\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t}{\lambda_2(L_t)}$, where the $K+1$ factor is spurious. The reason for this looseness is that we go through a coarse upper bound on the magnitude of the losses, and lose the dependence on the Laplacian along the way. This is not just an artifact of the analysis: Recall that the algorithmic reduction proceeds by using transformations of the actual losses, and these transformations may not satisfy the same Laplacian constraints as the original losses. Getting a better algorithm with improved regret performance in this particular setting is left to future work.

References

Jacob Abernethy, Peter L Bartlett, Rafael Frongillo, and Andre Wibisono. How to hedge an option against an adversary: Black-scholes pricing is minimax optimal. In *NIPS*, 2013.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *arXiv preprint arXiv:1409.8428*, 2014.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. *arXiv preprint arXiv:1605.08722*, 2016.

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.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT*, pages 42–1, 2012.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT*, pages 6–1, 2012.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *NIPS*, 2016.

Robert Grone, Russell Merris, and V S_ Sunder. The laplacian spectrum of a graph. *SIAM Journal on Matrix Analysis and Applications*, 11(2):218–238, 1990.

Elad Hazan and Satyen Kale. On stochastic and worst-case models for investing. In *NIPS*, 2009.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2-3):165–188, 2010.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(Apr):1287–1311, 2011.

Zohar S Karnin and Oren Anava. Multi-armed bandits: Competing with optimal sequences. In *NIPS*, 2016.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

Ali Ajdari Rad, Mahdi Jalili, and Martin Hasler. A lower bound for algebraic connectivity based on the connection-graph-stability method. *Linear Algebra and Its Applications*, 1 (435):186–192, 2011.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT*, pages 993–1019, 2013.

Amir Sani, Gergely Neu, and Alessandro Lazaric. Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems*, pages 810–818, 2014.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *ICML*, 2014.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *ICML*, pages 1593–1601, 2014.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomas Kocak. Spectral bandits for smooth graph functions. In *ICML*, 2014.

Appendix A. Proofs

Proofs of Claim 1 and Claim 2 from Sec. 3

To show Claim 1, we consider separately the case where a is a bad arm at round t, and where a a good arm at round t. If a is a bad arm, then $\widetilde{\ell}_t(a) = 2\varepsilon_t(j_t)$, which is at most $\ell_t(a) - m_t(j_t) + \varepsilon_t(j_t)$ by Eq. (3). Otherwise, if a is a good arm at round t, the observation follows by definition of $\widetilde{\ell}_t$.

To show Claim 2, recall that if i is a good arm, then $\widetilde{\ell}_t(i) = \ell_t(i) - m_t(j_t) + \varepsilon_t(j_t)$, and otherwise, we have $\widetilde{\ell}_t(I_t) = 2\varepsilon_t(j_t) \geq \ell_t(j_t) - m_t(j_t) + \varepsilon_t(j_t)$ (since $\ell_t(j_t) \leq m_t(j_t) + \varepsilon_t(j_t)$ by definition). Letting G_t denote the set of good arms at round t, we have:

$$\begin{aligned}
\sum_i \widetilde{p}_t(i)\widetilde{\ell}_t(i) &= \sum_{i\in G_t} \widetilde{p}_t(i)\widetilde{\ell}_t(i) + \sum_{i \text{ bad}} \widetilde{p}_t(i)\widetilde{\ell}_t(i) \\
&= \sum_{i\in G_t} \widetilde{p}_t(i)\big(\ell_t(i) - m_t(j_t) + \varepsilon_t(j_t)\big) + \sum_{i \text{ is bad}} \widetilde{p}_t(i)\, 2\,\varepsilon_t(j_t) \\
&\geq \sum_{i\in G_t} \widetilde{p}_t(i)\big(\ell_t(i) - m_t(j_t) + \varepsilon_t(j_t)\big) + \sum_{i \text{ is bad}} \widetilde{p}_t(i)\big(\ell_t(j_t) - m_t(j_t) + \varepsilon_t(j_t)\big) \\
&= \sum_i p_t(i)\ell_t(I_t) - m_t(j_t) + \varepsilon_t(j_t)\ .
\end{aligned}$$

Proof of Theorem 5

Suppose the learner uses some (possibly randomized) strategy, and let A be a random variable denoting its random coin flips. Our goal is to provide lower bounds on

$$\sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \left(\mathbb{E}_A\left[\sum_{t=1}^T \ell_t(I_t)\right] - \min_{j=1,\ldots,K} \sum_{t=1}^T \ell_t(j) \right)$$

where the expectation is with respect to the learner's (possibly randomized) strategy. Clearly, this is lower bounded by

$$\mathbb{E}_{J,L}\mathbb{E}_A A\left[\sum_{t=1}^T \ell_t(I_t) - \sum_{t=1}^T \ell_t(J)\right]$$

where $\mathbb{E}_{J,L}$ signifies expectation over some distribution over indices J and losses $\{\ell_i(t)\}$. By Fubini's theorem, this equals

$$\mathbb{E}_A\mathbb{E}_{J,L}\left[\sum_{t=1}^T \ell_t(I_t) - \sum_{t=1}^T \ell_t(J)\right] \;\;\geq\;\; \inf_A \mathbb{E}_{\{\ell_i(t)\}_{i,t},j}\left[\sum_{t=1}^T \ell_t(I_t) - \sum_{t=1}^T \ell_t(j)\right]$$

where $\inf_A$ refers an infimum over the learner's random coin flips. Thus, we need to provide some distribution over indices J and losses, so that for any *deterministic* learner,

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(I_t) - \sum_{t=1}^T \ell_t(J)\right] \tag{6}$$

is lower bounded as stated in the theorem. Let

$$p_\varepsilon(j) = \frac{\varepsilon(j)^2}{\sum_{j'} \varepsilon(j')^2} \qquad j = 1, \ldots, K\ .$$

The proof will be composed of two constructions, depending on whether we are in the bandit of full information setting, and whether $\max_j p_\varepsilon(j)$ is larger or smaller than $1/4$.

The case $\max_j p_\varepsilon(j) \le \frac{1}{4}$ with bandit feedback

For this case, we will consider the following distribution: Let J be distributed on $\{1, \ldots, K\}$ according to the probability distribution $p(1), \ldots, p(k)$ (to be specified later). Conditioned on any $J = j$, we define the distribution over losses as follows, independently for each round t and index i:

- If $i \neq j$, then $\ell_t(i)$ equals $\max_r \varepsilon(r) + \varepsilon(i)$ w.p, $\frac{1}{2}$, and $\max_r \varepsilon(r) - \varepsilon(i)$ w.p. $\frac{1}{2}$.
- If $i = j$, then $\ell_t(i)$ equals $\max_r \varepsilon(r) + \varepsilon(i)$ w.p. $\frac{1-\delta(i)}{2}$, and $\max_r \varepsilon(r) - \varepsilon(i)$ w.p. $\frac{1+\delta(i)}{2}$.

Also, let $\mathbb{E}_j, \mathbb{P}_j$ denote expectation and probabilities (over the space of possible losses and indices) conditioned on the event $J = j$. With this construction, we note that $\mathbb{E}_j[\ell_t(j)] = \max_r \varepsilon(r) - \delta(j)\delta(j)$, and $\mathbb{E}_j[\ell_t(i)] = \max_r \varepsilon(r)$ if $i \neq j$. As a result,

$$\mathbb{E}_j[\ell_t(I_t) - \ell_t(j)] = \mathbb{P}_j(I_t \neq j) \cdot \mathbb{E}_j[\ell_t(I_t) - \ell_t(j) | I_t \neq j)] = \mathbb{P}_j(I_t \neq j)\varepsilon(j)\delta(j),$$

and therefore Eq. (6) equals

$$\begin{aligned}\sum_{j=1}^{K} p(j)\mathbb{E}_j\left[\sum_{t=1}^{T}\big(\ell_t(I_t) - \ell_t(j)\big)\right] &= \sum_{j=1}^{K} p(j) \sum_{t=1}^{T} \mathbb{P}_j(I_t \neq j)\varepsilon(j)\delta(j) \\ &= \sum_{j=1}^{K} p(j)\varepsilon(j)\delta(j) \sum_{t=1}^{T}\big(1 - \mathbb{P}_j(I_t = j)\big). \qquad (7)\end{aligned}$$

Let $\mathbb{P}_0$ denote the probability distribution over $\{\ell_t(i)\}_{t,i}$, where for *any* i and t, $\ell_t(i)$ is independent and equals $\max_r \varepsilon(r) \pm \varepsilon(i)$ with equal probability (note that this induces a probability on any event which is a deterministic function of the loss assignments, such as $I_t = j$ for some t, j). By a standard information-theoretic argument (see for instance (Bubeck and Cesa-Bianchi, 2012, proof of Lemma 3.6)), we have that

$$\big|\mathbb{P}_j(I_t = j) - \mathbb{P}_0(I_t = j)\big| \ \le\ \sqrt{\frac{\mathbb{E}_0[T(j)]}{2}\,\mathrm{KL}\left(\tfrac{1}{2}\middle|\middle|\tfrac{1-\delta(j)}{2}\right)}$$

where $T(j)$ is the number of times arm j was chosen by the learner, and $\mathrm{KL}\left(\frac{1}{2}\middle|\middle|\frac{1-\delta(j)}{2}\right) = \frac{1}{2}\log\left(\frac{1}{1-\delta(j)^2}\right)$ is the Kullback-Leibler divergence between Bernoulli distributions with $1/2$ and $(1-\delta(j))/2$. Using the easily-verified fact that $\log(1/(1-z)) \le 2z$ for all $z \in [0, 1/2]$, it follows that

$$\big|\mathbb{P}_j(I_t = j) - \mathbb{P}_0(I_t = j)\big| \ \le\ \sqrt{\frac{\mathbb{E}_0[T(j)]\delta(j)^2}{2}}$$

as long as $\delta(j)^2 \le 1/2$. Plugging this back into Eq. (7), we get the lower bound

$$\sum_{j=1}^{K} p(j)\varepsilon(j)\delta(j) \sum_{t=1}^{T} \left(1 - \mathbb{P}_0(I_t = j) - \sqrt{\frac{\mathbb{E}_0[T(j)]\delta(j)^2}{2}}\right) \tag{8}$$

which is valid as long as $\max_j \delta(j)^2 \le 1/2$. Now, for all $j = 1, \ldots, K$, we pick

$$p(j) = p_\varepsilon(j) \qquad \text{and} \qquad \delta(j) = \mathbb{I}\big\{\varepsilon(j) > 0\big\} \frac{\sqrt{\sum_j \varepsilon(j)^2}}{\varepsilon(j)\sqrt{T}}$$

(assuming that $\max_j \delta(j)^2 \le 1/2$).

Substituting in Eq. (8), and letting $J = \big\{j \in \{1, \ldots, k\} \,:\, \varepsilon(j) > 0\big\}$, we get

$$\begin{aligned}
&\sum_{j \in J} \frac{\varepsilon(j)^2}{\sqrt{T \sum_j \varepsilon(j)^2}} \cdot \sum_{t=1}^{T} \left(1 - \mathbb{P}_0(I_t = j) - \sqrt{\frac{\mathbb{E}_0[T(j)]}{2p_\varepsilon(j)T}}\right) \\
&\quad = \sqrt{T \sum_j \varepsilon(j)^2} \cdot \sum_{j \in J} \frac{p_\varepsilon(j)}{T} \sum_{t=1}^{T} \left(1 - \mathbb{P}_0(I_t = j) - \sqrt{\frac{\mathbb{E}_0[T(j)]}{2p_\varepsilon(j)T}}\right) \\
&\quad \ge \sqrt{T \sum_j \varepsilon(j)^2} \left(1 - \max_j p_\varepsilon(j) - \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j \in J} p_\varepsilon(j) \sqrt{\frac{\mathbb{E}_0[T(j)]}{2p_\varepsilon(j)T}}\right)\right) \\
&\quad \ge \sqrt{T \sum_j \varepsilon(j)^2} \left(1 - \max_j p_\varepsilon(j) - \sqrt{\frac{\sum_{j \in J} \mathbb{E}_0[T(j)]}{2T}}\right) \\
&\quad \ge \sqrt{T \sum_j \varepsilon(j)^2} \left(1 - \max_j p_\varepsilon(j) - \sqrt{\frac{1}{2}}\right)
\end{aligned}$$

where in the second-to-last step we used the fact that

$$\sum_{j \in J} p_\varepsilon(j)\sqrt{a_j} \le \sqrt{\sum_{j \in J} p_\varepsilon(j) a_j}$$

for any non-negative a_j, which follows from Jensen's inequality and the fact that $p_\varepsilon(j)$ represents a probability distribution over the indices in J. Since we assume that $p_\varepsilon(j) \le \frac{1}{4}$, the above is at least $0.04\sqrt{T \sum_j \varepsilon(j)^2}$, so we get overall that

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t) - \sum_{t=1}^{T} \ell_t(J)\right] \;\ge\; 0.04\sqrt{T \sum_j \varepsilon(j)^2}$$

under the assumption that $\max_j p_\varepsilon(j) \le \frac{1}{4}$ and that T is sufficiently large so that

$$\max_j \frac{\mathbb{I}\big\{\varepsilon(j) > 0\big\}}{p_\varepsilon(j)T} \;\le\; \frac{1}{2}\;.$$

Note that the latter condition indeed holds under the theorem's conditions.

The case $\max_j p_\varepsilon(j) \geq \frac{1}{4}$ with bandit feedback or with full information feedback

We now turn to consider either the full information setting, or the bandit setting when $\max_j p_\varepsilon(j) \geq \frac{1}{4}$. In the latter case, we note that $\sum_j \varepsilon(j)^2$ is at most a constant factor larger than $\max_j \varepsilon(j)^2$, so it is sufficient to prove a lower bound of $c\sqrt{T\max_j \varepsilon(j)^2}$ for some universal positive c. In fact, we will prove this lower bound regardless of the values of $\varepsilon(1),\ldots,\varepsilon(K)$, and even in the easier full information case. Therefore, the same construction will give us a lower bound for both the full information setting, and the bandit setting when $\max_j p_\varepsilon(j) \geq \frac{1}{4}$.

To lower bound Eq. (6), we will use the following distribution over losses and J, letting $i_{\max}$ be some arbitrary index in $\arg\max_{i\in\{1,\ldots,K\}} \varepsilon(i)$, and $\delta \in (0, 1/2]$ be some parameter to be chosen later:

- For any $t = 1,\ldots,T$ and $i \neq i_{\max}$, we fix $\ell_t(i) = \varepsilon(i_{\max})$.
- We pick a value z uniformly at random from $\{-1, 1\}$. Then, for all $t = 1,\ldots,T$, we let $\ell_t(i_{\max})$ equal $2\varepsilon(i_{\max})$ with probability $\frac{1-z\delta}{2}$, and 0 with probability $\frac{1+z\delta}{2}$. Also, if $z = 1$, we let $J = 1$, and if $z = -1$, we let $J = 2$.

Clearly, this loss assignment is valid (as $\big|\ell_t(i)-\varepsilon(i_{\max})\big| \leq \varepsilon(i)$ for all t, i). Intuitively, we let all arms but $i_{\max}$ have a fixed loss of 1/2, and randomly choose $i_{\max}$ to be either a "good" arm or a "bad" arm compared to the other arms (with expected value $(1-z\delta)\varepsilon(i_{\max})$, which can be wither $(1+\delta)\varepsilon(i_{\max})$ or $(1-\delta)\varepsilon(i_{\max})$). By letting $\delta = \Theta(1/\sqrt{T})$, we ensure that the algorithm cannot distinguish between these two events, and therefore will "err" and pick $\Omega\big(\varepsilon(i_{\max})/\sqrt{T}\big)$-suboptimal arms with at least constant probability throughout the T rounds, hence incurring $\Omega\big(\varepsilon(i_{\max})\sqrt{T}\big)$ regret.

To make this more formal, let $\mathbb{E}_+, \mathbb{P}_+$ denote expectations and probabilities conditioned on $z = 1$, and $\mathbb{E}_-, \mathbb{P}_-$ denote expectations and probabilities conditioned on $z = -1$. With this notation, Eq. (6) can be written as

$$\begin{aligned}
&\frac{1}{2}\,\mathbb{E}_+\left[\sum_{t=1}^{T}\big(\ell_t(I_t)-\ell_t(1)\big)\right] + \frac{1}{2}\mathbb{E}_-\left[\sum_{t=1}^{T}\big(\ell_t(I_t)-\ell_t(2)\big)\right]\\
&= \frac{1}{2}\sum_{t=1}^{T}\mathbb{P}_+(I_t \neq 1)\delta\varepsilon(i_{\max}) + \frac{1}{2}\sum_{t=1}^{T}\mathbb{P}_-(I_t = 1)\delta\varepsilon(i_{\max})\\
&= \frac{\delta\varepsilon(i_{\max})}{2}\sum_{t=1}^{T}\big(1-\mathbb{P}_+(I_t=1)+\mathbb{P}_-(I_t=1)\big)\\
&\geq \frac{\delta\varepsilon(i_{\max})}{2}\sum_{t=1}^{T}\Big(1-\big|\mathbb{P}_+(I_t=1)-\mathbb{P}_-(I_t=1)\big|\Big)
\end{aligned} \tag{9}$$

Noting that I_t (as a random variable) depends only on the random loss assignments of arm 1, and applying Pinsker's inequality, we have that $\big|\mathbb{P}_+(I_t = 1) - \mathbb{P}_-(I_t = 1)\big| \leq \sqrt{\frac{1}{2}KL(P_t^+||P_t^-)}$, where $\mathrm{KL}(P_t^+||P_t^-)$ is the Kullback-Leibler divergence between the distributions of the losses of arm 1 in rounds $1, 2,\ldots,t-1$, under $z = -1$ and under $z = 1$.

Since the losses are independent across rounds, we can apply the chain rule and get that

$$\big|\mathbb{P}_+(I_t = 1) - \mathbb{P}_-(I_t = 1)\big| \;\leq\; \sqrt{\frac{1}{2}\mathrm{KL}(P_t^+||P_t^-)} \;\leq\; \sqrt{\frac{t-1}{2}\mathrm{KL}\left(\frac{1-\delta}{2}\bigg|\bigg|\frac{1+\delta}{2}\right)}$$

where $\mathrm{KL}\left(\frac{1-\delta}{2}\big|\big|\frac{1+\delta}{2}\right)$ is the Kullback-Leibler divergence between Bernoulli distributions with parameters $\frac{1-\delta}{2}$ and $\frac{1+\delta}{2}$. This in turn equals

$$\sqrt{\frac{t-1}{2}\cdot\delta\log\left(\frac{1+\delta}{1-\delta}\right)} \;\leq\; \sqrt{\frac{3T}{2}\delta^2}$$

where we used the easily-verified fact that $\log\left(\frac{1+\delta}{1-\delta}\right) \leq 3\delta$ for all $\delta \in (0, 1/2]$. Plugging this back into Eq. (9), we get overall that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t) - \ell_t(J))\right] \;\geq\; \frac{\delta\varepsilon(i_{\max})}{2}T\left(1-\delta\sqrt{\frac{3T}{2}}\right)$$

Picking $\delta = 1/2\sqrt{T}$ (which is valid since it is in $(0, 1/2]$ for all T), we get a lower bound of $c\varepsilon(i_{\max})\sqrt{T} = \sqrt{T\max_j \varepsilon(j)^2}$ for some positive c as required.

Proof of Theorem 7

We start by recalling the classical analysis of the Exp3 regret.

Lemma 13 *For losses $\ell_t(i) \in [0,1]$, the regret of the Exp3 algorithm run with parameter $\eta > 0$ satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(I_t)\right] - \min_{k=1,\dots,K}\sum_{t=1}^{T}\ell_t(k) \leq \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{\ell}_t\|^2 \ .$$

Proof The proof is as follows,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i=1}^{K}\frac{w_{t+1}(i)}{W_t} \\
&= \sum_{i=1}^{K}\frac{w_t(i)}{W_t}\,\exp\!\big(-\eta\,\widehat{\ell}_t(i)\big) \\
&= \sum_{i=1}^{K}p_t(i)\exp\!\big(-\eta\,\widehat{\ell}_t(i)\big) \\
&\leq \sum_{i=0}^{K}p_t(i)\left(1-\eta\,\widehat{\ell}_t(i) + \frac{\big(\eta\,\widehat{\ell}_t(i)\big)^2}{2}\right) \\
&\quad \text{(using } e^{-x} \leq 1 - x + x^2/2 \text{ for all } x \geq 0\text{)} \\
&\leq 1 - \eta\sum_{i=1}^{K}p_t(i)\widehat{\ell}_t(i) + \frac{\eta^2}{2}\sum_{i=1}^{K}p_t(i)\widehat{\ell}_t(i)^2 \ .
\end{aligned}$$

Taking logs, upper bounding, and summing over $t = 1, \dots, T$ yields

$$\log \frac{W_{T+1}}{W_1} \le -\eta \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i) + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i)^2 \ .$$

Moreover, for any fixed comparison arm k, we also have

$$\log \frac{W_{T+1}}{W_1} \ge \log \frac{w_{T+1}(k)}{W_1} = -\eta \sum_{t=1}^{T} \widehat{\ell}_t(k) - \log K \ .$$

Putting together,

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i) - \sum_{t=1}^{T} \widehat{\ell}_t(k) \le \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i)^2 \ . \tag{10}$$

Next, note that

$$\mathbb{E}_t\Big[\widehat{\ell}_t(i)\Big] = \ell_t(i) \qquad \text{and} \qquad \mathbb{E}_t\Big[\widehat{\ell}_t(i)^2\Big] = \frac{\ell_t(i)^2}{p_t(i)} \ . \tag{11}$$

This immediately gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(k) \le \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} \ell_t(i)^2$$

concluding the proof. ■

A simple graph is an unweighted, undirected graph, containing no self-loops or multiple edges.

Lemma 14 *Let $G = (V, E)$ be a simple and connected graph with $|V| = K$ nodes, and let L be its Laplacian matrix. Let $L(i,i)$ be the $(K-1) \times (K-1)$ submatrix obtained by deleting the i-th row and the i-th column from L. Then, for any $i = 1, \dots, K$, the eigenvalues of $L(i,i)$ are the non-zero eigenvalues of L.*

Proof L has $K-1$ non-zero eigenvalues because it is connected. Let $\boldsymbol{v} = (v_1, \dots, v_K)$ be an eigenvector of L with eigenvalue $\lambda > 0$. Since $L\mathbf{1} = \mathbf{0}$, we have that

$$\lambda^2 = \boldsymbol{v}^\top L \boldsymbol{v} = \big(\boldsymbol{v} - v_i\mathbf{1}\big)^\top L\big(\boldsymbol{v} - v_i\mathbf{1}\big) = \boldsymbol{v}_i^\top L(i,i)\boldsymbol{v}_i$$

where $\boldsymbol{v}_i = \big(v_1 - v_i, \dots, v_{i-1} - v_i, v_{i+1} - v_i, \dots, v_K - v_i\big)$. Hence $\boldsymbol{v}_i$ is an eigenvector of $L(i,i)$ with eigenvalue λ. ■

We are now ready to prove Theorem 7.

Proof of Theorem 7. Using the invariance of the regret to translation of the losses and the fact that $\ell_t(i) + 1 - a_t \ge 0$ for all t, i, and k_t,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(k) &= \mathbb{E}\left[\sum_{t=1}^{T} \widetilde{\ell}_t(I_t)\right] - \sum_{t=1}^{T} \widetilde{\ell}_t(k) \\
&\le \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\widetilde{\boldsymbol{\ell}}_t\|^2 \ .
\end{aligned} \tag{12}$$

Let $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t = C_t^2$. Since $\boldsymbol{\ell}_t^\top L_t \boldsymbol{\ell}_t = \widetilde{\boldsymbol{\ell}}_t^\top L_t \widetilde{\boldsymbol{\ell}}_T$, and since $\widetilde{\boldsymbol{\ell}}_t$ has component k_t equal to 1, we can upper bound each term $\|\widetilde{\boldsymbol{\ell}}_t\|^2$ by the solution of the convex program

$$
\begin{aligned}
&\max_{\boldsymbol{v}\in\mathbb{R}^K} \|\boldsymbol{v}\|^2 \\
&\text{such that} \quad \boldsymbol{v}^\top L_t \boldsymbol{v} \le C_t^2 \\
&\qquad\qquad\quad \exists i \in \{1,\dots,K\}\ v(i) = 1 \ .
\end{aligned} \tag{13}
$$

Using Lemma 14, the above program is equivalent to

$$
\begin{aligned}
&\max_{\boldsymbol{v}\in\mathbb{R}^{K-1}} \left(1 + \|\boldsymbol{v}\|^2\right) \\
&\quad\text{such that} \qquad \boldsymbol{v}^\top L_t(1,1) \boldsymbol{v} \le C_t^2
\end{aligned}
$$

where $L(1,1)$ is full rank. Hence we can set $\boldsymbol{u} = L_t(1,1)^{1/2}\boldsymbol{v}/C_t$ and obtain the equivalent program

$$
1 + \max_{\boldsymbol{u}\in\mathbb{R}^{K-1}\,:\,\|\boldsymbol{u}\|\le 1} C_t^2\big(\boldsymbol{u}^\top L_t(1,1)^{-1}\boldsymbol{u}\big) = 1 + \frac{C_t^2}{\lambda_2(L_t)} \tag{14}
$$

which gives us the claimed bound. ■

Proof of Theorem 8

Let $\bar{Q}_t = Q_1 + \cdots + Q_s$. The largest r we need is the smallest R such that

$$
\sum_{r=0}^{R} 2^r \ge \bar{Q}_T
$$

and so $R = \big\lfloor \log_2(\bar{Q}_T + 1)\big\rfloor$. Therefore

$$
\sum_{r=0}^{R} 2^{r/2} < 4\sqrt{\bar{Q}_T + 1} \ .
$$

Because of Eq. (10),

$$
\sum_{t=1}^{T}\left(\sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i) - \widehat{\ell}_t(k)\right) \le \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i)^2
$$

and so

$$
\sum_{t\in S_r}\left(\sum_{i=1}^{K} p_t(i)\widehat{\ell}_t(i) - \widehat{\ell}_t(k)\right) \le \frac{\log K}{\eta_r} + \frac{\eta_r}{2}\sum_{t\in S_r} Q_t \le \sqrt{2(\log K)2^r} \ .
$$

Since a regret of at most 1 is incurred whenever Exp3 is restarted, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}p_t(i)\widehat{\ell}_t(i)\right]-\sum_{t=1}^{T}\widehat{\ell}_t(k)\\ &\le \mathbb{E}\Big[\big\lfloor \log_2\big(\bar{Q}_T+1\big)\big\rfloor\Big]+4\,\mathbb{E}\left[\sqrt{2(\log K)\big(\bar{Q}_T+1\big)}\right]\\ &\le \log_2\big(KT+1\big)+4\sqrt{2(\log K)\left(1+\sum_{t=1}^{T}\|\boldsymbol{\ell}_t\|^2\right)}\end{aligned}$$

where in the last step we used Jensen's inequality and Eq. (11).

Proof of Theorem 10

To prove the theorem, let $d=\lceil 1/\lambda\rceil$, and k be any integer such that d divides $k-1$. Finally, define $G_{k,d}$ to be an "octopus" graph composed of $(k-1)/d$ tentacles of equals length d. Formally, for any two nodes $i,j\in\{1,\dots,k\}$ where $j>i$ w.l.o.g., we have $(i,j)\in E$ if and only if

$$((j=k)\ \text{ and }\ (i=1 \bmod d))\quad\text{or}\quad((j\neq k)\ \text{ and }\ (j=i+1)\ \text{ and }\ (i\neq 0 \bmod d))$$

Note that here, node k is the "central" node, from which all tentacles emanate (first tentacle corresponding to nodes $1,2,\dots,d$, second tentacle corresponding to nodes $d{+}1,d{+}2,\dots,2d$ and so on).

The theorem is a straightforward corollary of the following two lemmas.

Lemma 15 *For an octopus graph $G_{k,d}$ with Laplacian $L_{k,d}$, for any $C>0$, and for any randomized algorithm, there exists an adversary strategy $\boldsymbol{\ell}_1,\boldsymbol{\ell}_2,\dots$ such that the expected regret is $\Omega\left(\min\left\{\sqrt{k},Cd\right\}\sqrt{T}\right)$ while $\boldsymbol{\ell}_t^\top L_{k,d}\boldsymbol{\ell}_t\le C$ for each $t=1,\dots,T$.*

Proof (Sketch) In the graph, there are $\Omega(k)$ points at a distance $\Omega(d)$ from the center. These "faraway" points can get loss magnitudes as large as $\frac{1}{2}\pm\min\{1,Cd/\sqrt{k}\}$, while satisfying the budget constraints and having the central point as an anchor with a fixed loss of $1/2$ (to make sure the budget constraint is satisfied, assign losses in increments of roughly $C/\sqrt{k}$ along each tentacle, so points in the faraway half of each points gets losses varying as $\pm\min\{1,Cd/\sqrt{k}\}$). Specifically, all those faraway points will get the same random binary loss every round (equalling $1/2$ in expectation), except for one point whose loss will always be a $\Theta(\sqrt{k/T})$ smaller in expectation. Assuming the hidden constant is small enough, and using the standard proof technique of lower bounds for nonstochastic multi-armed bandits, we have that the algorithm is unlikely to detect the arm with smaller-in-expectation loss, resulting in an expected regret lower bound of

$$\Omega\left(\min\left\{1,\frac{Cd}{\sqrt{k}}\right\}\sqrt{kT}\right)=\Omega\left(\min\left\{\sqrt{k},Cd\right\}\sqrt{T}\right).$$

■

Lemma 16 *For an Octopus graph $G_{k,d}$, $\lambda_2 = \Theta(1/d^2)$ (where $\Theta(\cdot)$ hides universal constants).*

Proof We begin with the lower bound. By (Rad et al., 2011, Theorem 1), λ_2 is lower bounded by $k/C_{\max}$, where $C_{\max} = \max_{e \in E} C_e$, and C_e is the sum, over all pairs of distinct nodes i and j of the length of the shortest path between (i,j) passing through e assuming this path exists. For the graph as defined above, any edge separates at most d nodes from at most $k-1$ other nodes, and the length of the path between any two nodes is at most $2d$. Therefore, $C_{\max} \leq 2kd^2$, so

$$\lambda_2 \geq \frac{k}{2kd^2} = \frac{1}{2d^2} = \Omega\left(\frac{1}{d^2}\right).$$

Turning to the upper bound, by corollary 4.4 in Grone et al. (1990), for any tree graph of diameter D, $\lambda_2 \leq 2\left(1 - \cos(\pi/(D+1))\right)$, and since $\cos(x) \geq 1 - x^2/2$ for all x, this implies

$$\lambda_2 \leq \frac{\pi^2}{(D+1)^2}.$$

An Octopus graph $G_{k,d}$ is a tree with diameter $2d$, hence

$$\lambda_2 \leq \frac{\pi^2}{(2d+1)^2} = \mathcal{O}\left(\frac{1}{d^2}\right)$$

from which the result follows. ■

Proof of Lemma 11

For simplicity, we will drop the s,t subscripts, as they play no role here. The proof follows by an analysis similar to that of Eq. (13), where 1 is replaced by $m_t(s)$ and noting that the 2-norm upper bounds the ∞-norm. Specifically, making the worst-case assumption that the adversary budget $C = \boldsymbol{\ell}^\top L \boldsymbol{\ell}$ is spent solely on the connected component we are concerned with, we need to solve the convex program

$$\begin{aligned} &\max_{\boldsymbol{v} \in \mathbb{R}^K} \|\boldsymbol{v} - m \cdot \mathbf{1}\|_\infty \\ &\quad \text{such that} \qquad \boldsymbol{v}^\top L \boldsymbol{v} \leq C^2 \\ &\qquad\qquad\qquad\quad\ \exists i \in \{1, \ldots, K\}\ v(i) = m\ . \end{aligned}$$

which is equivalent (using the fact that $\boldsymbol{\ell}^\top L \boldsymbol{\ell}$ is invariant to shifting the coordinates of $\boldsymbol{\ell}$) to

$$\begin{aligned} &\max_{\boldsymbol{v} \in \mathbb{R}^K} \|\boldsymbol{v}\|_\infty \\ &\quad \text{such that} \quad \boldsymbol{v}^\top L \boldsymbol{v} \leq C^2 \\ &\qquad\qquad\quad\ \exists i \in \{1, \ldots, K\}\ v(i) = 0\ . \end{aligned}$$

Upper bounding the ∞-norm by the 2-norm, and using Lemma 14, the above program is equivalent to

$$\begin{aligned} &\max_{\boldsymbol{v} \in \mathbb{R}^{K-1}} \|\boldsymbol{v}\|_2 \\ &\quad \text{such that} \quad \boldsymbol{v}^\top L(1,1) \boldsymbol{v} \leq C^2 \end{aligned}$$

where $L(1,1)$ is full rank. Hence we can set $\boldsymbol{u} = L(1,1)^{1/2}\boldsymbol{v}/C$ and obtain the equivalent program

$$\max_{\boldsymbol{u}\in\mathbb{R}^{K-1}\,:\,\|\boldsymbol{u}\|\le 1} C\sqrt{\left(\boldsymbol{u}^\top L(1,1)^{-1}\boldsymbol{u}\right)} = \frac{C}{\sqrt{\lambda_2(L)}}$$

which gives us the claimed bound.