Refined Lower Bounds for Adversarial Bandits

Sébastien Gerchinovitz, Tor Lattimore

▶ **To cite this version:**

Sébastien Gerchinovitz, Tor Lattimore. Refined Lower Bounds for Adversarial Bandits. 2016. hal-01319572v1

HAL Id: hal-01319572
https://hal.science/hal-01319572v1

Preprint submitted on 20 May 2016 (v1), last revised 25 Feb 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Refined Lower Bounds for Adversarial Bandits

Sébastien Gerchinovitz
Institut de Mathématiques de Toulouse
Université Toulouse 3 Paul Sabatier
Toulouse, 31062, France
sebastien.gerchinovitz@math.univ-toulouse.fr

Tor Lattimore
Department of Computing Science
University of Alberta
Edmonton, Canada
tor.lattimore@gmail.com

Abstract

We provide new lower bounds on the regret that must be suffered by adversarial bandit algorithms. The new results show that recent upper bounds that either (a) hold with high-probability or (b) depend on the total loss of the best arm or (c) depend on the quadratic variation of the losses, are close to tight. Besides this we prove two impossibility results. First, the existence of a single arm that is optimal in every round cannot improve the regret in the worst case. Second, the regret cannot scale with the effective range of the losses. In contrast, both results are possible in the full-information setting.

1 Introduction

We consider the standard K-armed adversarial bandit problem, which is a game played over T rounds between a learner and an adversary. In every round $t \in \{1, \dots, T\}$ the learner chooses a probability distribution $p_t = (p_{i,t})_{1 \leqslant i \leqslant K}$ over $\{1, \dots, K\}$. The adversary then chooses a loss vector $\ell_t = (\ell_{i,t})_{1 \leqslant i \leqslant K} \in [0,1]^K$, which may depend on p_t. Finally the learner samples an action from p_t denoted by $I_t \in \{1, \dots, K\}$ and observes her own loss $\ell_{I_t,t}$. The learner would like to minimise her regret, which is the difference between cumulative loss suffered and the loss suffered by the optimal action in hindsight.

$$R_T(\ell_{1:T}) = \sum_{t=1}^{T} \ell_{I_t,t} - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell_{i,t} \,,$$

where $\ell_{1:T} \in [0,1]^{TK}$ is the set of losses chosen by the adversary. A famous strategy is called Exp3, which satisfies $\mathbb{E}[R_T(\ell_{1:T})] = \mathcal{O}(\sqrt{KT \log(K)})$ where the expectation is taken over the randomness in the algorithm and the choices of the adversary [Auer et al., 2002]. There is also a lower bound showing that for every learner there is an adversary for which the expected regret is $\mathbb{E}[R_T(\ell_{1:T})] = \Omega(\sqrt{KT})$ [Auer et al., 1995]. If the losses are chosen ahead of time, then the adversary is called oblivious, and in this case there exists a learner for which $\mathbb{E}[R_T(\ell_{1:T})] = \mathcal{O}(\sqrt{KT})$ [Audibert and Bubeck, 2009]. One might think that this is the end of the story, but it is not so. While the worst-case expected regret is one quantity of interest, there are many situations where a refined regret guarantee is more informative. Recent research on adversarial bandits has primarily focussed on these issues, especially the questions of obtaining regret guarantees that hold with high probability as well as stronger guarantees when the losses are "nice" in some sense. While there are now a wide range of strategies with upper bounds that depend on various quantities, the literature is missing lower bounds for many cases, some of which we now provide.

We focus on three classes of lower bound, which are described in detail below. The first addresses the optimal regret achievable with high probability, where we show there is little room for improvement over existing strategies. Our other results concern lower bounds that depend on some kind of regularity in the losses ("nice" data). Specifically we prove lower bounds that replace T in the regret bound with the loss of the best action (called first-order bounds) and also with the quadratic variation of the losses (called second-order bounds).

High probability bounds Existing strategies Exp3.P [Auer et al., 2002] and Exp-IX [Neu, 2015] are tuned with a confidence parameter $\delta \in (0,1)$ and satisfy for some universal constant $c > 0$

$$\mathbb{P}\left(R_T(\ell_{1:T}) \geqslant c\sqrt{KT\log(K/\delta)}\right) \leqslant \delta\,. \tag{1}$$

An alternative tuning of Exp-IX or Exp3.P [Bubeck and Cesa-Bianchi, 2012] leads to a single algorithm for which

$$\forall \delta \in (0,1) \qquad \mathbb{P}\left(R_T(\ell_{1:T}) \geqslant c\sqrt{KT}\left(\sqrt{\log(K)} + \frac{\log(1/\delta)}{\sqrt{\log(K)}}\right)\right) \leqslant \delta\,. \tag{2}$$

The difference is that in (1) the algorithm depends on δ while in (2) it does not. The cost of not knowing δ is that the $\log(1/\delta)$ moves outside the square root. In Section 2 we prove two lower bounds showing that there is little room for improvement in either (1) or (2).

First-order bounds An improvement over the worst-case regret bound of $\mathcal{O}(\sqrt{TK})$ is the so-called *improvement for small losses*. Specifically, there exist strategies (eg., FPL-TRIX by Neu [2015] with earlier results by Stoltz [2005], Allenberg et al. [2006], Rakhlin and Sridharan [2013]) such that for all $\ell_{1:T} \in [0,1]^{KT}$

$$\mathbb{E}[R_T(\ell_{1:T})] \leqslant \mathcal{O}\left(\sqrt{L_T^* K\log(K)} + K\log(KT)\right), \text{ with } L_T^* = \min_{1\leqslant i\leqslant K}\sum_{t=1}^{T}\ell_{i,t}\,, \tag{3}$$

where the expectation is with respect to the internal randomisation of the algorithm (the losses are fixed). This result improves on the $\mathcal{O}(\sqrt{KT})$ bounds since $L_T^* \leqslant T$ is always guaranteed and sometimes L_T^* is much smaller than T. In order to evaluate the optimality of this bound, we first rewrite it in terms of the small-loss balls $\mathcal{B}_{\alpha,T}$ defined for all $\alpha \in [0,1]$ and $T \geqslant 1$ by

$$\mathcal{B}_{\alpha,T} \triangleq \left\{\ell_{1:T} \in [0,1]^{KT} \;:\; \frac{L_T^*}{T} \leqslant \alpha\right\}\,. \tag{4}$$

Corollary 1. *The first-order regret bound (3) of Neu [2015] is equivalent to:*

$$\forall \alpha \in [0,1], \qquad \sup_{\ell_{1:T}\in\mathcal{B}_{\alpha,T}} \mathbb{E}[R_T(\ell_{1:T})] \leqslant \mathcal{O}\left(\sqrt{\alpha TK\log(K)} + K\log(KT)\right)\,.$$

The proof is straightforward. Our main contribution in Section 3 is a lower bound of the order of $\sqrt{\alpha TK}$ for all $\alpha \in \Omega(\log(T)/T)$. This minimax lower bound shows that we cannot hope for a better bound than (3) (up to log factors) if we only know the value of L_T^*.

Second-order bounds Another type of improved regret bound was derived by Hazan and Kale [2011b] and involves a second-order quantity called the quadratic variation.

$$Q_T = \sum_{t=1}^{T}\|\ell_t - \mu_T\|_2^2 \leqslant \frac{TK}{4}\,, \tag{5}$$

where $\mu_T = \frac{1}{T}\sum_{t=1}^{T}\ell_t$ is the mean of all loss vectors. (In other words, Q_T/T is the sum of the empirical variances of all the K arms). Hazan and Kale [2011b] addressed the general online linear optimisation setting. In the particular case of adversarial K-armed bandits with an oblivious adversary (as is the case here), they showed that there exists an efficient algorithm such that for some absolute constant $c > 0$ and for all $T \geqslant 2$

$$\forall \ell_{1:T} \in [0,1]^{KT}, \qquad \mathbb{E}[R_T(\ell_{1:T})] \leqslant c\left(K^2\sqrt{Q_T\log T} + K^{1.5}\log^2 T + K^{2.5}\log T\right)\,. \tag{6}$$

As before we can rewrite the regret bound (6) in terms of the small-variation balls $\mathcal{V}_{\alpha,T}$ defined for all $\alpha \in [0,1/4]$ and $T \geqslant 1$ by

$$\mathcal{V}_{\alpha,T} \triangleq \left\{\ell_{1:T} \in [0,1]^{KT} \;:\; \frac{Q_T}{TK} \leqslant \alpha\right\}\,. \tag{7}$$

Corollary 2. *The second-order regret bound (6) of Hazan and Kale [2011b] is equivalent to:*

$$\forall \alpha \in [0, 1/4], \qquad \sup_{\ell_{1:T} \in \mathcal{V}_{\alpha,T}} \mathbb{E}[R_T(\ell_{1:T})] \leqslant c\left(K^2\sqrt{\alpha T K \log T} + K^{3/2}\log^2 T + K^{5/2}\log T\right) .$$

The proof is straightforward because the losses are deterministic and fixed in advance by an oblivious adversary. In Section 4 we provide a lower bound of order $\sqrt{\alpha T K}$ that holds whenever $\alpha = \Omega(\log(T)/T)$. This minimax lower bound shows that we cannot hope for a bound better than (7) by more than a factor of $K^2\sqrt{\log T}$ if we only know the value of Q_T. Closing the gap is left as an open question.

Two impossibility results in the bandit setting We also show in Section 4 that, in contrast to the full-information setting, regret bounds involving the cumulative variance of the algorithm as in [Cesa-Bianchi et al., 2007] cannot be obtained in the bandit setting. More precisely, we prove that two consequences that hold true in the full-information case, namely: (i) a regret bound proportional to the effective range of the losses and (ii) a bounded regret if one arm performs best at all rounds, must fail in the worst case for every bandit algorithm.

Additional notation and key tools Before the theorems we develop some additional notation and describe the generic ideas in the proofs. For $1 \leqslant i \leqslant K$ let $N_i(t)$ be the number of times action i has been chosen after round t. All our lower bounds are derived by analysing the regret incurred by strategies when facing randomised adversaries that choose the losses for all actions from the same joint distribution in every round (sometimes independently for each action and sometimes not). $\mathrm{Ber}(\alpha)$ denotes the Bernoulli distribution with parameter $\alpha \in [0, 1]$. If $\mathbb{P}$ and $\mathbb{Q}$ are measures on the same probability space, then $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ is the KL-divergence between them. For $a < b$ we define $\mathrm{clip}_{[a,b]}(x) = \min\{b, \max\{a, x\}\}$ and for $x, y \in \mathbb{R}$ we let $x \vee y = \max\{x, y\}$. Our main tools throughout the analysis are the following information-theoretic lemmas. The first bounds the KL divergence between the laws of the observed losses/actions for two distributions on the losses.

Lemma 1. *Fix a randomised bandit algorithm and two probability distributions Q_1 and Q_2 on $[0, 1]^K$. Assume the loss vectors $\ell_1, \ldots, \ell_T \in [0, 1]^K$ are drawn i.i.d. from either Q_1 or Q_2, and denote by $\mathbb{Q}_j$ the joint probability distribution on all sources of randomness when Q_j is used (formally, $\mathbb{Q}_j = \mathbb{P}_{\mathrm{int}} \otimes (Q_j^{\otimes T})$, where $\mathbb{P}_{\mathrm{int}}$ is the probability distribution used by the algorithm for its internal randomisation). Let $t \geqslant 1$. Denote by $h_t = (I_s, \ell_{I_s,s})_{1 \leqslant s \leqslant t-1}$ the history available at the beginning of round t, by $\mathbb{Q}_j^{(h_t, I_t)}$ the law of (h_t, I_t) under $\mathbb{Q}_j$, and by $Q_{j,i}$ the ith marginal distribution of Q_j. Then,*

$$\mathrm{KL}\left(\mathbb{Q}_1^{(h_t,I_t)}, \mathbb{Q}_2^{(h_t,I_t)}\right) = \sum_{i=1}^{K} \mathbb{E}_{\mathbb{Q}_1}\left[N_i(t-1)\right] \mathrm{KL}\big(Q_{1,i}, Q_{2,i}\big) .$$

Results of roughly this form are well known and the proof follows immediately from the chain rule for the relative entropy and the independence of the loss vectors across time (see [Auer et al., 2002] or Appendix A). One difference is that the losses need not be independent across the arms, which we heavily exploit in our proofs by using correlated losses. The second key lemma is an alternative to Pinsker's inequality that proves useful when the Kullback-Leibler divergence is larger than 2. It has previously been used for bandit lower bounds (in the stochastic setting) by Bubeck et al. [2013].

Lemma 2 (Lemma 2.6 in Tsybakov 2008)**.** *Let P and Q be two probability distributions on the same measurable space. Then, for every measurable subset A (whose complement we denote by A^c),*

$$P(A) + Q(A^c) \geqslant \frac{1}{2}\exp\big(-\mathrm{KL}(P, Q)\big) .$$

2 Zero-Order High Probability Lower Bounds

We prove two new high-probability lower bounds on the regret of any bandit algorithm. The first shows that no strategy can enjoy smaller regret than $\Omega(\sqrt{KT\log(1/\delta)})$ with probability at least $1 - \delta$. Upper bounds of this form have been shown for various algorithms including Exp.3P [Auer et al., 2002] and Exp3-IX [Neu, 2015]. Although this result is not very surprising, we are not aware of any existing work on this problem and the proof is less straightforward than one might expect. An added

benefit of our result is that the loss sequences producing large regret have two special properties. First, the optimal arm is the same in every round and second the range of the losses in each round is $\mathcal{O}(\sqrt{K\log(1/\delta)/T})$. These properties will be useful in subsequent analysis.

In the second lower bound we show that any algorithm for which $\mathbb{E}[R_T(\ell_{1:T})] = \mathcal{O}(\sqrt{KT})$ must necessarily suffer a high probability regret of at least $\Omega(\sqrt{KT}\log(1/\delta))$ for some sequence $\ell_{1:T}$. The important difference relative to the previous result is that strategies with $\log(1/\delta)$ appearing inside the square root depend on a specific value of δ, which must be known in advance.

Theorem 1. *Suppose $K \geqslant 2$ and $\delta \in (0, 1/4)$ and $T \geqslant 32(K-1)\log(2/\delta)$, then there exists a sequence of losses $\ell_{1:T} \in [0,1]^{KT}$ such that*

$$\mathbb{P}\left(R_T(\ell_{1:T}) \geqslant \frac{1}{27}\sqrt{(K-1)T\log(1/(4\delta))}\right) \geqslant \delta/2\,,$$

where the probability is taken with respect to the randomness in the algorithm. Furthermore $\ell_{1:T}$ can be chosen in such a way that there exists an i such that for all t it holds that $\ell_{i,t} = \min_j \ell_{j,t}$ and $\max_{j,k}\{\ell_{j,t} - \ell_{k,t}\} \leqslant \sqrt{(K-1)\log(1/(4\delta))/T}/(4\sqrt{\log 2})$.

Theorem 2. *Suppose $K \geqslant 2$, $T \geqslant 1$, and there exists a strategy and constant $C > 0$ such that for any $\ell_{1:T} \in [0,1]^{KT}$ it holds that $\mathbb{E}[R_T(\ell_{1:T})] \leqslant C\sqrt{(K-1)T}$. Let $\delta \in (0, 1/4)$ satisfy $\sqrt{(K-1)/T}\log(1/(4\delta)) \leqslant C$ and $T \geqslant 32\log(2/\delta)$. Then there exists $\ell_{1:T} \in [0,1]^{KT}$ for which*

$$\mathbb{P}\left(R_T(\ell_{1:T}) \geqslant \frac{\sqrt{(K-1)T}\log(1/(4\delta))}{203C}\right) \geqslant \delta/2\,,$$

where the probability is taken with respect to the randomness in the algorithm.

Corollary 3. *If $p \in (0,1)$ and $C > 0$, then there does not exist a strategy such that for all T, K, $\ell_{1:T} \in [0,1]^{KT}$ and $\delta \in (0,1)$ the regret is bounded by $\mathbb{P}\left(R_T(\ell_{1:T}) \geqslant C\sqrt{(K-1)T}\log^p(1/\delta)\right) \leqslant \delta$.*

The corollary follows easily by integrating the assumed high-probability bound and applying Theorem 2 for sufficiently large T and small δ. The proof may be found in Appendix E.

Proof of Theorems 1 and 2 Both proofs rely on a carefully selected choice of correlated stochastic losses described below. Let $Z_1, Z_2, \ldots, Z_T$ be a sequence of i.i.d. Gaussian random variables with mean $1/2$ and variance $\sigma^2 = 1/(32\log(2))$. Let $\Delta \in [0, 1/30]$ be a constant that will be chosen differently in each proof and define K random loss sequences $\ell^1_{1:T}, \ldots, \ell^K_{1:T}$ where

$$\ell^j_{i,t} = \begin{cases} \mathrm{clip}_{[0,1]}(Z_t - \Delta) & \text{if } i = 1 \\ \mathrm{clip}_{[0,1]}(Z_t - 2\Delta) & \text{if } i = j \neq 1 \\ \mathrm{clip}_{[0,1]}(Z_t) & \text{otherwise}\,. \end{cases}$$

For $1 \leqslant j \leqslant K$ let $\mathbb{Q}_j$ be the measure on $\ell_{1:T} \in [0,1]^{KT}$ and $I_1, \ldots, I_T$ when $\ell_{i,t} = \ell^j_{i,t}$ for all $1 \leqslant i \leqslant K$ and $1 \leqslant t \leqslant T$. Informally, $\mathbb{Q}_j$ is the measure on the sequence of loss vectors and actions when the learner interacts with the losses sampled from the jth environment defined above.

Lemma 3. *Let $\delta \in (0,1)$ and suppose $\Delta \leqslant 1/30$ and $T \geqslant 32\log(2/\delta)$. Then $\mathbb{Q}_i\left(R_T(\ell^i_{1:T}) \geqslant \Delta T/4\right) \geqslant \mathbb{Q}_i\left(N_i(T) \leqslant T/2\right) - \delta/2$ and $\mathbb{E}_{\mathbb{Q}_i}[R_T(\ell^i_{1:T})] \geqslant 7\Delta\mathbb{E}_{\mathbb{Q}_i}[T - N_i(T)]/8$.*

The proof may be found in Appendix D.

Proof of Theorem 1. First we choose the value of Δ that determines the gaps in the losses by $\Delta = \sqrt{\sigma^2(K-1)\log(1/(4\delta))/(2T)} \leqslant 1/30$. By the pigeonhole principle there exists an $i > 1$ for which $\mathbb{E}_{\mathbb{Q}_1}[N_i(T)] \leqslant T/(K-1)$. Therefore by Lemmas 2 and 1, and the fact that the KL divergence between clipped Gaussian distributions is always smaller than without clipping (see Lemma 7 in Appendix B),

$$\begin{aligned}\mathbb{Q}_1\left(N_1(T) \leqslant T/2\right) + \mathbb{Q}_i\left(N_1(T) > T/2\right) &\geqslant \frac{1}{2}\exp\left(-\mathrm{KL}\left(\mathbb{Q}_1^{(h_T, I_T)}, \mathbb{Q}_i^{(h_T, I_T)}\right)\right) \\ &\geqslant \frac{1}{2}\exp\left(-\frac{\mathbb{E}_{\mathbb{Q}_1}[N_i(T)](2\Delta)^2}{2\sigma^2}\right) \geqslant \frac{1}{2}\exp\left(-\frac{2T\Delta^2}{\sigma^2(K-1)}\right) = 2\delta\,.\end{aligned}$$

But by Lemma 3

$$\max_{k\in\{1,i\}} \mathbb{Q}_k\left(R_T(\ell^k_{1:T}) \geqslant T\Delta/4\right) \geqslant \max\left\{\mathbb{Q}_1\left(N_1(T) \leqslant T/2\right), \mathbb{Q}_i\left(N_i(T) \leqslant T/2\right)\right\} - \delta/2$$
$$\geqslant \frac{1}{2}\left(\mathbb{Q}_1\left(N_1(T) \leqslant T/2\right) + \mathbb{Q}_i\left(N_1(T) > T/2\right)\right) - \delta/2 \geqslant \delta/2\,.$$

Therefore there exists an $i \in \{1,\dots,K\}$ such that

$$\mathbb{Q}_i\left(R_T(\ell^i_{1:T}) \geqslant \sqrt{\frac{\sigma^2 T(K-1)}{32}\log\left(\frac{1}{4\delta}\right)}\right) = \mathbb{Q}_i\left(R_T(\ell^i_{1:T}) \geqslant T\Delta/4\right) \geqslant \delta/2\,.$$

The result is completed by substituting the value of $\sigma^2 = 1/(32\log(2))$ and by noting that $\max_{j,k}\{\ell_{j,t} - \ell_{k,t}\} \leqslant 2\Delta \leqslant \sqrt{(K-1)\log(1/(4\delta))/T}/(4\sqrt{\log 2})$ $\mathbb{Q}_i$-almost surely. □

Proof of Theorem 2. By the assumption on δ we have $\Delta = \frac{7\sigma^2}{16C}\sqrt{\frac{K-1}{T}\log\left(\frac{1}{4\delta}\right)} \leqslant 1/30$. Suppose for all $i > 1$ that

$$\mathbb{E}_{\mathbb{Q}_1}[N_i(T)] > \frac{\sigma^2}{2\Delta^2}\log\left(\frac{1}{4\delta}\right)\,. \tag{8}$$

Then by the assumption in the theorem statement and the second part of Lemma 3 we have

$$C\sqrt{(K-1)T} \geqslant \mathbb{E}_{\mathbb{Q}_1}[R_T(\ell^1_{1:T})] \geqslant \frac{7\Delta}{8}\mathbb{E}_{\mathbb{Q}_1}\left[\sum_{i=2}^K N_i(T)\right] > \frac{7\sigma^2(K-1)}{16\Delta}\log\frac{1}{4\delta} = C\sqrt{(K-1)T}\,,$$

which is a contradiction. Therefore there exists an $i > 1$ for which Eq. (8) does not hold. Then by the same argument as the previous proof it follows that

$$\max_{k\in\{1,i\}} \mathbb{Q}_k\left(R_T(\ell^k_{1:T}) \geqslant \frac{7\sigma^2}{4\cdot 16C}\sqrt{(K-1)T}\log\frac{1}{4\delta}\right) = \max_{k\in\{1,i\}} \mathbb{Q}_k\left(R_T(\ell^k_{1:T}) \geqslant T\Delta/4\right) \geqslant \delta/2\,.$$

The result is completed by substituting the value of $\sigma^2 = 1/(32\log(2))$. □

3 First-Order Lower Bound

First-order upper bounds provide improvement over minimax bounds when the loss of the optimal action is small. Recall from Corollary 1 that first-order bounds can be rewritten in terms of the small-loss balls $\mathcal{B}_{\alpha,T}$ defined in (4). Theorem 3 below provides a new lower bound of order $\sqrt{L^*_T K}$, which matches the best existing upper bounds up to logarithmic factors. As is standard for minimax results this does not imply a lower bound on every loss sequence $\ell_{1:T}$. Instead it shows that we cannot hope for a better bound if we only know the value of L^*_T.

Theorem 3. *Let $K \geqslant 2$, $T \geqslant K \vee 118$, and $\alpha \in [(c\log(32T) \vee (K/2))/T, 1/2]$, where $c = 64/9$. Then for any randomised bandit algorithm $\sup_{\ell_{1:T}\in\mathcal{B}_{\alpha,T}} \mathbb{E}[R_T(\ell_{1:T})] \geqslant \sqrt{\alpha TK}/27$, where the expectation is taken with respect to the internal randomisation of the algorithm.*

Our proof is inspired by that of Auer et al. [2002, Theorem 5.1]. The key difference is that we take Bernoulli distributions with parameter close to α instead of $1/2$. This way the best cumulative loss L^*_T is ensured to be concentrated around αT, and the regret lower bound $\sqrt{\alpha TK} \approx \sqrt{\alpha(1-\alpha)TK}$ can be seen to involve the variance $\alpha(1-\alpha)T$ of the binomial distribution with parameters α and T.

First we state the stochastic construction of the losses and prove a general lemma that allows us to prove Theorem 3 and will also be useful in Section 4 to a derive a lower bound in terms of the quadratic variation. Let $\varepsilon \in [0, 1-\alpha]$ be fixed and define K probability distributions $(\mathbb{Q}_j)_{j=1}^K$ on $[0,1]^{KT}$ such that under $\mathbb{Q}_j$ the following hold:

- All random losses $\ell_{i,t}$ for $1 \leqslant i \leqslant K$ and $1 \leqslant t \leqslant T$ are independent.
- $\ell_{i,t}$ is sampled from a Bernoulli distribution with parameter $\alpha+\varepsilon$ if $i \neq j$, or with parameter α if $i = j$.

Lemma 4. *Let $\alpha \in (0,1)$, $K \geqslant 2$, and $T \geqslant K/(4(1-\alpha))$. Consider the probability distributions $\mathbb{Q}_j$ on $[0,1]^{KT}$ defined above with $\varepsilon = (1/2)\sqrt{\alpha(1-\alpha)K/T}$, and set $\bar{\mathbb{Q}} = \frac{1}{K}\sum_{j=1}^{K}\mathbb{Q}_j$. Then for any randomised bandit algorithm $\mathbb{E}[R_T(\ell_{1:T})] \geqslant \sqrt{\alpha(1-\alpha)TK}/8$, where the expectation is with respect to both the internal randomisation of the algorithm and the random loss sequence $\ell_{1:T}$ which is drawn from $\bar{\mathbb{Q}}$.*

The assumption $T \geqslant K/(4(1-\alpha))$ above ensures that $\varepsilon \leqslant 1-\alpha$, so that the $\mathbb{Q}_j$ are well defined.

Proof of Lemma 4. We lower bound the regret by the pseudo-regret for each distribution $\mathbb{Q}_j$:

$$\mathbb{E}_{\mathbb{Q}_j}\left[\sum_{t=1}^{T}\ell_{I_t,t} - \min_{1\leqslant i\leqslant K}\sum_{t=1}^{T}\ell_{i,t}\right] \geqslant \mathbb{E}_{\mathbb{Q}_j}\left[\sum_{t=1}^{T}\ell_{I_t,t}\right] - \min_{1\leqslant i\leqslant K}\mathbb{E}_{\mathbb{Q}_j}\left[\sum_{t=1}^{T}\ell_{i,t}\right]$$
$$= \sum_{t=1}^{T}\mathbb{E}_{\mathbb{Q}_j}\left[\alpha+\varepsilon-\varepsilon\mathbb{1}_{\{I_t=j\}}\right] - T\alpha = T\varepsilon\left(1 - \frac{1}{T}\sum_{t=1}^{T}\mathbb{Q}_j(I_t=j)\right), \tag{9}$$

where the first equality follows because $\mathbb{E}_{\mathbb{Q}_j}[\ell_{I_t,t}] = \mathbb{E}_{\mathbb{Q}_j}[\mathbb{E}_{\mathbb{Q}_j}[\ell_{I_t,t}|\ell_{1:t-1},I_t]] = \mathbb{E}_{\mathbb{Q}_j}[\alpha+\varepsilon-\varepsilon\mathbb{1}_{\{I_t=j\}}]$ since under $\mathbb{Q}_j$, the conditional distribution of ℓ_t given $(\ell_{1:t-1},I_t)$ is simply $\otimes_{i=1}^{K}\mathcal{B}(\alpha+\varepsilon-\varepsilon\mathbb{1}_{\{i=j\}})$. To bound (9) from below, note that by Pinsker's inequality we have for all $t \in \{1,\dots,T\}$ and $j \in \{1,\dots,K\}$, $\mathbb{Q}_j(I_t=j) \leqslant \mathbb{Q}_0(I_t=j) + (\mathrm{KL}(\mathbb{Q}_0^{I_t},\mathbb{Q}_j^{I_t})/2)^{1/2}$, where $\mathbb{Q}_0 = \mathrm{Ber}(\alpha)^{\otimes KT}$ is the joint probability distribution that makes all the $\ell_{i,t}$ i.i.d. $\mathrm{Ber}(\alpha)$, and $\mathbb{Q}_0^{I_t}$ and $\mathbb{Q}_j^{I_t}$ denote the laws of I_t under $\mathbb{Q}_0$ and $\mathbb{Q}_j$ respectively. Plugging the last inequality above into (9), averaging over $j=1,\dots,K$ and using the concavity of the square root yields

$$\mathbb{E}_{\bar{\mathbb{Q}}}\left[\sum_{t=1}^{T}\ell_{I_t,t} - \min_{1\leqslant i\leqslant K}\sum_{t=1}^{T}\ell_{i,t}\right] \geqslant T\varepsilon\left(1 - \frac{1}{K} - \sqrt{\frac{1}{2T}\sum_{t=1}^{T}\frac{1}{K}\sum_{j=1}^{K}\mathrm{KL}\big(\mathbb{Q}_0^{I_t},\mathbb{Q}_j^{I_t}\big)}\right), \tag{10}$$

where we recall that $\bar{\mathbb{Q}} = \frac{1}{K}\sum_{j=1}^{K}\mathbb{Q}_j$. The rest of the proof is devoted to upper-bounding $\mathrm{KL}(\mathbb{Q}_0^{I_t},\mathbb{Q}_j^{I_t})$. Denote by $h_t = (I_s,\ell_{I_s,s})_{1\leqslant s\leqslant t-1}$ the history available at the beginning of round t. From Lemma 1

$$\mathrm{KL}\Big(\mathbb{Q}_0^{I_t},\mathbb{Q}_j^{I_t}\Big) \leqslant \mathrm{KL}\Big(\mathbb{Q}_0^{(h_t,I_t)},\mathbb{Q}_j^{(h_t,I_t)}\Big) = \mathbb{E}_{\mathbb{Q}_0}\big[N_j(t-1)\big]\,\mathrm{KL}\big(\mathcal{B}(\alpha+\varepsilon),\mathcal{B}(\alpha)\big)$$
$$\leqslant \mathbb{E}_{\mathbb{Q}_0}\big[N_j(t-1)\big]\,\frac{\varepsilon^2}{\alpha(1-\alpha)}, \tag{11}$$

where the last inequality follows by upper bounding the KL divergence by the χ^2 divergence (see Appendix B). Averaging (11) over $j \in \{1,\dots,K\}$ and $t \in \{1,\dots,T\}$ and noting that $\sum_{t=1}^{T}(t-1) \leqslant T^2/2$ we get

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{j=1}^{K}\mathrm{KL}\big(\mathbb{Q}_0^{I_t},\mathbb{Q}_j^{I_t}\big) \leqslant \frac{1}{T}\sum_{t=1}^{T}\frac{(t-1)\varepsilon^2}{K\alpha(1-\alpha)} \leqslant \frac{T\varepsilon^2}{2K\alpha(1-\alpha)}.$$

Plugging the above inequality into (10) and using the definition of $\varepsilon = (1/2)\sqrt{\alpha(1-\alpha)K/T}$ yields

$$\mathbb{E}_{\bar{\mathbb{Q}}}\left[\sum_{t=1}^{T}\ell_{I_t,t} - \min_{1\leqslant i\leqslant K}\sum_{t=1}^{T}\ell_{i,t}\right] \geqslant T\varepsilon\left(1-\frac{1}{K}-\frac{1}{4}\right) \geqslant \frac{1}{8}\sqrt{\alpha(1-\alpha)TK}. \qquad \square$$

Proof of Theorem 3. We show that there exists a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ such that $L_T^* \leqslant \alpha T$ and $\mathbb{E}[R_T(\ell_{1:T})] \geqslant (1/27)\sqrt{\alpha TK}$. Lemma 4 above provides such kind of lower bound, but without the guarantee on L_T^*. For this purpose we will use Lemma 4 with a smaller value of α (namely, $\alpha/2$) and combine it with Bernstein's inequality to prove that $L_T^* \leqslant T\alpha$ with high probability.

Part 1: Applying Lemma 4 with $\alpha/2$ (note that $T \geqslant K \geqslant K/(4(1-\alpha))$ by assumption on T and α) and noting that $\max_j \mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})] \geqslant \mathbb{E}_{\bar{\mathbb{Q}}}[R_T(\ell_{1:T})]$ we get that for some $j \in \{1,\dots,K\}$ the probability distribution $\mathbb{Q}_j$ defined with $\varepsilon = (1/2)\sqrt{(\alpha/2)(1-\alpha/2)K/T}$ satisfies

$$\mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})] \geqslant \frac{1}{8}\sqrt{\frac{\alpha}{2}\left(1-\frac{\alpha}{2}\right)TK} \geqslant \frac{1}{32}\sqrt{6\alpha TK} \tag{12}$$

since $\alpha \leqslant 1/2$ by assumption.

Part 2: Next we prove that $$\mathbb{Q}_j(L_T^* > T\alpha) \leqslant \frac{1}{32T} \,. \tag{13}$$

To this end, first note that $L_T^* \leqslant \sum_{t=1}^T \ell_{j,t}$. Second, note that under $\mathbb{Q}_j$, the $\ell_{j,t}$, $t \geqslant 1$, are i.i.d. $\mathrm{Ber}(\alpha/2)$. We can thus use Bernstein's inequality: applying Theorem 2.10 (and a remark on p.38) of Boucheron et al. [2013] with $X_t = \ell_{j,t} - \alpha/2 \leqslant 1 = b$, with $v = T(\alpha/2)(1-\alpha/2)$, and with $c = b/3 = 1/3$), we get that, for all $\delta \in (0,1)$, with $\mathbb{Q}_j$-probability at least $1-\delta$,

$$\begin{aligned} L_T^* &\leqslant \sum_{t=1}^T \ell_{j,t} \leqslant \frac{T\alpha}{2} + \sqrt{2T\,\frac{\alpha}{2}\left(1-\frac{\alpha}{2}\right)\log\frac{1}{\delta}} + \frac{1}{3}\log\frac{1}{\delta} \\ &\leqslant \frac{T\alpha}{2} + \left(1+\frac{1}{3}\right)\sqrt{T\alpha\log\frac{1}{\delta}} \leqslant \frac{T\alpha}{2} + \frac{T\alpha}{2} = T\alpha \,, \end{aligned} \tag{14}$$

where the second last inequality is true whenever $T\alpha \geqslant \log(1/\delta)$ and that last is true whenever $T\alpha \geqslant (8/3)^2\log(1/\delta) = c\log(1/\delta)$. By assumption on α, these two conditions are satisfied for $\delta = 1/(32T)$, which concludes the proof of (13).

Conclusion: We show by contradiction that there exists a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ such that $L_T^* \leqslant \alpha T$ and

$$\mathbb{E}[R_T(\ell_{1:T})] \geqslant \frac{1}{64}\sqrt{6\alpha TK} \,, \tag{15}$$

where the expectation is with respect to the internal randomisation of the algorithm. Imagine for a second that (15) were false for every loss sequence $\ell_{1:T} \in [0,1]^{KT}$ satisfying $L_T^* \leqslant \alpha T$. Then we would have $\mathbb{1}_{\{L_T^* \leqslant \alpha T\}}\mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})|\ell_{1:T}] \leqslant (1/64)\sqrt{6\alpha TK}$ almost surely (since the internal source of randomness of the bandit algorithm is independent of $\ell_{1:T}$). Therefore by the tower rule for the first expectation on the r.h.s. below, we would get

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})] &= \mathbb{E}_{\mathbb{Q}_j}\left[R_T(\ell_{1:T})\mathbb{1}_{\{L_T^* \leqslant \alpha T\}}\right] + \mathbb{E}_{\mathbb{Q}_j}\left[R_T(\ell_{1:T})\mathbb{1}_{\{L_T^* > \alpha T\}}\right] \\ &\leqslant \frac{1}{64}\sqrt{6\alpha TK} + T\cdot\mathbb{Q}_j(L_T^* > T\alpha) \leqslant \frac{1}{64}\sqrt{6\alpha TK} + \frac{1}{32} < \frac{1}{32}\sqrt{6\alpha TK} \end{aligned} \tag{16}$$

where (16) follows from (13) and by noting that $1/32 < (1/64)\sqrt{6\alpha TK}$ since $\alpha \geqslant K/(2T) > 4/(6T) \geqslant 4/(6TK)$. Comparing (16) and (12) we get a contradiction, which proves that there exists a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ satisfying both $L_T^* \leqslant \alpha T$ and (15). We conclude the proof by noting that $\sqrt{6}/64 \geqslant 1/27$. Finally, the condition $T \geqslant K \vee 118$ is sufficient to make the interval $\left[(c\log(32T) \vee (K/2))/(T), \frac{1}{2}\right]$ non empty. □

4 Second-Order Lower Bounds

We start by giving a lower bound on the regret in terms of the quadratic variation that is close to existing upper bounds except in the dependence on the number of arms. Afterwards we prove that bandit strategies cannot adapt to losses that lie in a small range or the existence of an action that is always optimal.

Lower bound in terms of quadratic variation We prove a lower bound of $\Omega(\sqrt{\alpha TK})$ over any small-variation ball $\mathcal{V}_{\alpha,T}$ (as defined by (7)) for all $\alpha = \Omega(\log(T)/T)$. This minimax lower bound matches the upper bound of Corollary 2 up to a multiplicative factor of $K^2\sqrt{\log(T)}$. Closing this gap is left as an open question, but we conjecture that the upper bound is loose (see also the COLT open problem by Hazan and Kale [2011a]).

Theorem 4. *Let $K \geqslant 2$, $T \geqslant (32K) \vee 601$, and $\alpha \in [(2c_1\log(T) \vee 8K)/T, 1/4]$, where $c_1 = (4/9)^2(3\sqrt{5}+1)^2 \leqslant 12$. Then for any randomised bandit algorithm, $\sup_{\ell_{1:T}\in\mathcal{V}_{\alpha,T}} \mathbb{E}[R_T(\ell_{1:T})] \geqslant \sqrt{\alpha TK}/25$, where the expectation is taken with respect to the internal randomisation of the algorithm.*

The proof is very similar to that of Theorem 3; it also follows from Lemma 4 and Bernstein's inequality. It is postponed to Appendix C.

Impossibility results In the full-information setting (where the entire loss vector is observed after each round) Cesa-Bianchi et al. [2007, Theorem 6] designed a carefully tuned exponential weighting algorithm for which the regret depends on the variation of the algorithm and the range of the losses:

$$\forall \ell_{1:T} \in \mathbb{R}^{KT}, \qquad \mathbb{E}[R_T(\ell_{1:T})] \leqslant 4\sqrt{V_T \log K} + 4E_T \log K + 6E_T \;, \tag{17}$$

where the expectation is taken with respect to the internal randomisation of the algorithm and $E_T = \max_{1 \leqslant t \leqslant T} \max_{1 \leqslant i,j \leqslant K} |\ell_{i,t} - \ell_{j,t}|$ denotes the effective range of the losses and $V_T = \sum_{t=1}^{T} \mathrm{Var}_{I_t \sim p_t}(\ell_{I_t,t})$ denotes the cumulative variance of the algorithm (in each round t the expert's action I_t is drawn at random from the weight vector p_t). The bound in (17) is not closed-form because V_T depends on the algorithm, but has several interesting consequences:

1. If for all t the losses $\ell_{i,t}$ lie in an unknown interval $[a_t, a_t + \rho]$ with a small width $\rho > 0$, then $\mathrm{Var}_{I_t \sim p_t}(\ell_{I_t,t}) \leqslant \rho^2/4$, so that $V_T \leqslant T\rho^2/4$. Hence
$$\mathbb{E}[R_T(\ell_{1:T})] \leqslant 2\rho\sqrt{T \log K} + 4\rho \log K + 6\rho \;.$$
Therefore, though the algorithm by Cesa-Bianchi et al. [2007, Section 4.2] does not use the prior knowledge of a_t or ρ, it is able to incur a regret that scales linearly in the effective range ρ.
2. If all the losses $\ell_{i,t}$ are nonnegative, then by Corollary 3 of [Cesa-Bianchi et al., 2007] the second-order bound (17) implies the first-order bound
$$\mathbb{E}[R_T(\ell_{1:T})] \leqslant 4\sqrt{L_T^* \left(M_T - \frac{L_T^*}{T}\right) \log K + 39 M_T \max\{1, \log K\}} \;, \tag{18}$$
where $M_T = \max_{1 \leqslant t \leqslant T} \max_{1 \leqslant i \leqslant K} \ell_{i,t}$.
3. If there exists an arm i^* that is optimal at every round t (i.e., $\ell_{i^*,t} = \min_i \ell_{i,t}$ for all $t \geqslant 1$), then a translation-invariant algorithm with regret guarantees as in (18) above suffers a bounded regret. This is the case for the algorithm[1] of Cesa-Bianchi et al. [2007, Section 4.1] tuned with $E = 1$ (if all losses lie in $[0, 1]$). Then by the translation invariance of the algorithm all losses $\ell_{i,t}$ appearing in the regret bound can be replaced with the translated losses $\ell_{i,t} - \ell_{i^*,t} \geqslant 0$, so that a bound of the same form as (18) implies a regret bound of $\mathcal{O}(\log K)$.
4. Assume that the loss vectors ℓ_t are i.i.d. with a unique optimal arm in expectation (i.e., there exists i^* such that $\mathbb{E}[\ell_{i^*,1}] < \mathbb{E}[\ell_{i,1}]$ for all $i \neq i^*$). Then using the Hoeffding-Azuma inequality we can show that the algorithm of Cesa-Bianchi et al. [2007, Section 4.2] has with high probability a bounded cumulative variance V_T, and therefore (by (17)) incurs a bounded regret, in the same spirit as in de Rooij et al. [2014], Gaillard et al. [2014].

We already know that point 2 has a counterpart in the bandit setting. If one is prepared to ignore logarithmic terms, then point 4 also has an analogue in the bandit setting due to the existence of logarithmic regret guarantees for stochastic bandits [Lai and Robbins, 1985]. The following corollaries show that in the bandit setting it is not possible to design algorithms to exploit the range of the losses or the existence of an arm that is always optimal. We use Theorem 1 as a general tool but the bounds can be improved to $\sqrt{TK}/30$ by analysing the expected regret directly (similar to Lemma 4).

Corollary 4. *Let $K \geqslant 2$, $T \geqslant 32(K-1)\log(14)$ and $\rho \geqslant 0.22\sqrt{(K-1)/T}$. Then for any randomised bandit algorithm, $\sup_{\ell_1, \ldots, \ell_T \in \mathcal{C}_\rho} \mathbb{E}[R_T(\ell_{1:T})] \geqslant \sqrt{T(K-1)}/504$, where the expectation is with respect to the randomness in the algorithm, and $\mathcal{C}_\rho \triangleq \left\{x \in [0,1]^K : \max_{i,j} |x_i - x_j| \leqslant \rho\right\}$.*

Corollary 5. *Let $K \geqslant 2$ and $T \geqslant 32(K-1)\log(14)$. Then, for any randomised bandit algorithm, there is a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ such that there exists an arm i^* that is optimal at every round t (i.e., $\ell_{i^*,t} = \min_i \ell_{i,t}$ for all $t \geqslant 1$), but $\mathbb{E}[R_T(\ell_{1:T})] \geqslant \sqrt{T(K-1)}/504$, where the expectation is with respect to the randomness in the algorithm.*

Proof of Corollaries 4 and 5. Both results follow from Theorem 1 by choosing $\delta = 0.15$. Therefore there exists an $\ell_{1:T}$ such that $\mathbb{P}\{R_T(\ell_{1:T}) \geqslant \sqrt{(K-1)T\log(1/(4 \cdot 0.15))}/27\} \geqslant 0.15/2$, which implies (since $R_T(\ell_{1:T}) \geqslant 0$ here) that $\mathbb{E}[R_T(\ell_{1:T})] \geqslant \sqrt{(K-1)T}/504$. Finally note that $\ell_{1:T} \in \mathcal{C}_\rho$ since $\rho \geqslant \sqrt{(K-1)\log(1/(4\delta))/T}/(4\sqrt{\log 2})$ and there exists an i such that $\ell_{i,t} \leqslant \ell_{j,t}$ for all j and t. □

[1] The fully automatic algorithm of Cesa-Bianchi et al. [2007, Section 4.2] is not translation-invariant because the estimated ranges E_t only take values on a dyadic grid.

Acknowledgments

The authors would like to thank Aurélien Garivier and Émilie Kaufmann for insightful discussions. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grants ANR-13-BS01-0005 (project SPADRO) and ANR-13-CORD-0020 (project ALICIA).

References

C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of ALT'2006*, pages 229–243. Springer, 2006.

J. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226, 2009.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.

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

S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proceedings of The 26th Conference on Learning Theory*, pages 122–134, 2013.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2/3):321–352, 2007.

T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], second edition, 2006.

S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *J. Mach. Learn. Res.*, 15(Apr):1281–1316, 2014.

P. Gaillard, G. Stoltz, and T. van Erven. A second-order bound with excess losses. In *Proceedings of the 27th Conference on Learning Theory (COLT'14)*, 2014.

E. Hazan and S. Kale. A simple multi-armed bandit algorithm with optimal variation-bounded regret. In *Proceedings of the 24th Conference on Learning Theory*, pages 817–820, 2011a.

E. Hazan and S. Kale. Better algorithms for benign bandits. *J. Mach. Learn. Res.*, 12(Apr):1287–1311, 2011b.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6: 4–22, 1985.

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

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Conference on Learning Theory*, pages 993–1019, 2013.

G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Paris-Sud XI University, 2005.

A. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

A Proof of Lemma 1

The proof is well known (eg., Auer et al. [2002]). We only write it for the convenience of the reader. Recall that $h_t = \big(I_s, \ell_{I_s,s}\big)_{1 \leqslant s \leqslant t-1}$. Next we write $\mathbb{Q}_j^{X|Y}$ the law of X conditionally on Y under $\mathbb{Q}_j$. By the chain rule for the Kullback-Leibler divergence, note that

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{Q}_1^{(h_t,I_t)}, \mathbb{Q}_2^{(h_t,I_t)}\Big) &= \sum_{s=1}^{t-1} \mathbb{E}_{\mathbb{Q}_1}\left[\mathrm{KL}\Big(\mathbb{Q}_1^{I_s|h_s}, \mathbb{Q}_2^{I_s|h_s}\Big) + \mathrm{KL}\Big(\mathbb{Q}_1^{\ell_{I_s,s}|(h_s,I_s)}, \mathbb{Q}_2^{\ell_{I_s,s}|(h_s,I_s)}\Big)\right] \\ &\quad + \mathbb{E}_{\mathbb{Q}_1}\left[\mathrm{KL}\Big(\mathbb{Q}_1^{I_t|h_t}, \mathbb{Q}_2^{I_t|h_t}\Big)\right] . \end{aligned} \tag{19}$$

Note that $\mathbb{Q}_1(I_s = i|h_s) = p_{i,s} = \mathbb{Q}_2(I_s = i|h_s)$ for all s (by definition of a randomised algorithm with weight vector p_s at time s), so that the first and third KL terms equal zero. As for the second one, we can check that $\mathbb{Q}_j^{\ell_{I_s,s}|(h_s,I_s)} = Q_{j,I_s}$ (the I_s-th marginal of Q_j). Combining all these remarks with (19), we get

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{Q}_1^{(h_t,I_t)}, \mathbb{Q}_2^{(h_t,I_t)}\Big) &= \sum_{s=1}^{t-1} \mathbb{E}_{\mathbb{Q}_1}\big[\mathrm{KL}(Q_{1,I_s}, Q_{2,I_s})\big] = \sum_{s=1}^{t-1} \mathbb{E}_{\mathbb{Q}_1}\left[\sum_{i=1}^{K} \mathbb{1}_{\{I_s=i\}} \mathrm{KL}(Q_{1,i}, Q_{2,i})\right] \\ &= \sum_{i=1}^{K} \mathbb{E}_{\mathbb{Q}_1}\big[N_i(t-1)\big] \mathrm{KL}\big(Q_{1,i}, Q_{2,i}\big) ,\end{aligned}$$

which concludes the proof.

B Inequalities from Information Theory

We first recall below a well-known data-processing inequality that follows from conditional Jensen's inequality. A variant of it in the discrete setting can be found, e.g., in Cover and Thomas [2006]. The main message is that transforming the data at hand can only reduce the ability to distinguish between two probability distributions.

Lemma 5 (Contraction of entropy)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$. Denote by $\mathbb{P}^X$ and $\mathbb{Q}^X$ the law of X under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then,*

$$\mathrm{KL}\big(\mathbb{P}^X, \mathbb{Q}^X\big) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q}) .$$

Next we recall an inequality between the Kullback-Leibler divergence and the chi-squared divergence. In the particular case of Bernoulli distributions with parameters $p, q \in [0,1]$, these divergences are given respectively by[2]

$$\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \qquad \text{and} \qquad \chi^2(p,q) = \frac{(p-q)^2}{q(1-q)} .$$

Lemma 6 (Consequence of Lemma 2.7 in Tsybakov 2008)**.** *Let $p, q \in [0,1]$. Then*

$$\mathrm{kl}(p,q) \leqslant \chi^2(p,q) .$$

The final lemma is a straightforward corollary of Lemma 5 and the KL divergence between two Gaussians.

Lemma 7. *For $a \leqslant b$ and define $\mathrm{clip}_{[a,b]}(x) = \max\{a, \min\{x, b\}\}$. Let Z be normally distributed with mean $1/2$ and variance $\sigma^2 > 0$. Define $X = \mathrm{clip}_{[0,1]}(Z)$ and $Y = \mathrm{clip}_{[0,1]}(Z - \varepsilon)$ for $\varepsilon \in \mathbb{R}$. Then $\mathrm{KL}(\mathbb{P}^X, \mathbb{P}^Y) \leqslant \varepsilon^2/(2\sigma^2)$.*

[2] We use the usual conventions: $0 \log 0 = 0/0 = 0$ and $a/0 = +\infty$ for all $a > 0$.

C Proof of Theorem 4

The proof follows the same lines as that of Theorem 3. In the sequel we show that there exists a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ such that $Q_T \leqslant \alpha TK$ and $\mathbb{E}[R_T(\ell_{1:T})] \geqslant (1/25)\sqrt{\alpha TK}$. As in the proof of Theorem 3, we use Lemma 4 with $\alpha/2$ and combine it with Bernstein's inequality to prove that $Q_T \leqslant \alpha TK$ with high probability.

Part 1: Applying Lemma 4 with $\alpha/2$ (note that $T \geqslant 32K \geqslant K/(4(1-\alpha))$ by assumption on T and α) and noting that $\max_j \mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})] \geqslant \mathbb{E}_{\bar{\mathbb{Q}}}[R_T(\ell_{1:T})]$ we get that for some $j \in \{1,\dots,K\}$ the probability distribution $\mathbb{Q}_j$ defined with $\varepsilon = (1/2)\sqrt{(\alpha/2)(1-\alpha/2)K/T}$ satisfies

$$\mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})] \geqslant \frac{1}{8}\sqrt{\frac{\alpha}{2}\left(1-\frac{\alpha}{2}\right)TK} \geqslant \frac{1}{32}\sqrt{7\alpha TK} \tag{20}$$

since $\alpha \leqslant 1/4$ by assumption.

Part 2: Next we prove that

$$\mathbb{Q}_j(Q_T > \alpha TK) \leqslant \frac{1}{32T}\,. \tag{21}$$

To this end recall that $\mu_T = \frac{1}{T}\sum_{t=1}^T \ell_t$ and

$$Q_T = \sum_{t=1}^T \|\ell_t - \mu_T\|_2^2 = \sum_{i=1}^K \underbrace{\sum_{t=1}^T (\ell_{i,t} - \mu_{i,T})^2}_{=:v_{i,T}}\,.$$

Noting that $\ell_{i,t} \in \{0,1\}$ almost surely, we have $v_{i,T} = T\mu_{i,T}(1-\mu_{i,T}) \leqslant T\mu_{i,T} = \sum_{t=1}^T \ell_{i,t}$.

Recall that under $\mathbb{Q}_j$, the $\ell_{i,t}$, $t \geqslant 1$, are i.i.d. $\mathrm{Ber}(\alpha_i^j)$ where $\alpha_i^j = (\alpha/2) + \varepsilon \mathbb{1}_{\{i \neq j\}}$ (we used Lemma 4 with $\alpha/2$). We now apply Bernstein's inequality exactly as after (13): combined with a union bound, it yields that, for all $\delta \in (0,1)$, with $\mathbb{Q}_j$-probability at least $1-\delta$, for all $i \in \{1,\dots,K\}$,

$$\begin{aligned}\sum_{t=1}^T \ell_{i,t} &\leqslant T\alpha_i^j + \sqrt{2T\alpha_i^j\left(1-\alpha_i^j\right)\log\frac{K}{\delta}} + \frac{1}{3}\log\frac{K}{\delta} \\ &\leqslant T\left(\frac{\alpha}{2}+\varepsilon\right) + \sqrt{2T\left(\frac{\alpha}{2}+\varepsilon\right)\log\frac{K}{\delta}} + \frac{1}{3}\log\frac{K}{\delta}\,.\end{aligned} \tag{22}$$

Now note that, by definition of $\varepsilon = (1/2)\sqrt{(\alpha/2)(1-\alpha/2)K/T}$ and by the assumption $T \geqslant 8K/\alpha$,

$$\frac{\alpha}{2} + \varepsilon \leqslant \frac{\alpha}{2} + \frac{1}{2}\sqrt{\frac{\alpha K}{2T}} \leqslant \frac{5\alpha}{8}\,.$$

Substituting the last upper bound in (22) and using the assumption $T\alpha \geqslant 4\log(K/\delta)$ (that we check later) to obtain $\log(K/\delta) \leqslant (1/2)\sqrt{T\alpha\log(K/\delta)}$ we get

$$\sum_{t=1}^T \ell_{i,t} \leqslant \frac{5T\alpha}{8} + \left(\frac{\sqrt{5}}{2}+\frac{1}{6}\right)\sqrt{T\alpha\log\frac{K}{\delta}} \leqslant \frac{5T\alpha}{8} + \frac{3T\alpha}{8} = T\alpha\,, \tag{23}$$

where the last inequality is true whenever $T\alpha \geqslant c_1\log(K/\delta)$ with $c_1 = (4/9)^2(3\sqrt{5}+1)^2$. By the assumption $\alpha \geqslant 2c_1\log(T)/T \geqslant c_1\log(32TK)/T$ (since $T \geqslant 32K$), the condition $T\alpha \geqslant c_1\log(K/\delta)$ is satisfied for $\delta = 1/(32T)$ (as well as the weaker condition $T\alpha \geqslant 4\log(K/\delta)$ mentioned above). We conclude the proof of (21) via $Q_T = \sum_{i=1}^K v_{i,T} \leqslant \sum_{i=1}^K \sum_{t=1}^T \ell_{i,t} \leqslant \alpha TK$ by (23).

Conclusion: We show by contradiction that there exists a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ such that $Q_T \leqslant \alpha TK$ and

$$\mathbb{E}[R_T(\ell_{1:T})] \geqslant \frac{1}{64}\sqrt{7\alpha TK}\,, \tag{24}$$

where the expectation is with respect to the internal randomisation of the algorithm. Imagine for a second that (24) were false for every loss sequence $\ell_{1:T} \in [0,1]^{KT}$ satisfying $Q_T \leqslant \alpha TK$. Then we would have $\mathbb{1}_{\{Q_T \leqslant \alpha TK\}} \mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})|\ell_{1:T}] \leqslant (1/64)\sqrt{7\alpha TK}$ almost surely (since the internal source of randomness of the bandit algorithm is independent of $\ell_{1:T}$). Therefore, using the tower rule for the first expectation on the r.h.s. below, we would get

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}_j}[R_T(\ell_{1:T})] &= \mathbb{E}_{\mathbb{Q}_j}\left[R_T(\ell_{1:T})\mathbb{1}_{\{Q_T \leqslant \alpha TK\}}\right] + \mathbb{E}_{\mathbb{Q}_j}\left[R_T(\ell_{1:T})\mathbb{1}_{\{Q_T > \alpha TK\}}\right] \\
&\leqslant \frac{1}{64}\sqrt{7\alpha TK} + T \cdot \mathbb{Q}_j(Q_T > \alpha TK) \\
&\leqslant \frac{1}{64}\sqrt{7\alpha TK} + \frac{1}{32} < \frac{1}{32}\sqrt{7\alpha TK} \qquad (25)
\end{aligned}$$

where (25) follows from (21) and by noting that $1/32 < (1/64)\sqrt{7\alpha TK}$ since $\alpha \geqslant 8K/T > 4/(7TK)$. Comparing (25) and (20) we get a contradiction, which proves that there exists a loss sequence $\ell_{1:T} \in [0,1]^{KT}$ satisfying both $Q_T \leqslant \alpha TK$ and (24). We conclude the proof by noting that $\sqrt{7}/64 \geqslant 1/25$.

Nota: the assumption $T \geqslant (32K) \vee 601$ is sufficient to make the interval $\left[\frac{2c_1 \log(T) \vee (8K)}{T}, \frac{1}{4}\right]$ non empty.

D Proof of Lemma 3

First we use the definition of the losses to bound

$$R_T(\ell^i_{1:T}) = \sum_{t=1}^{T}(\ell_{I_t,t} - \ell_{i,t}) \geqslant \Delta \sum_{t=1}^{T} \mathbb{1}_{\{Z_t \in [2\Delta, 1-2\Delta] \text{ and } I_t \neq i\}} .$$

Let $W_t = \mathbb{1}_{\{Z_t \in [2\Delta, 1-2\Delta]\}}$, which forms an i.i.d. Bernoulli sequence with

$$\mathbb{Q}_i\left(W_t = 0\right) \leqslant \exp\left(-\frac{(1/2 - 2\Delta)^2}{2\sigma^2}\right) = p \leqslant 1/8\,,$$

where the inequality follows by standard tail bounds on the Gaussian integral [Boucheron et al., 2013, Exercise 2.7]. Therefore by Hoeffding's bound

$$\mathbb{Q}_i\left(\sum_{t=1}^{T} W_t \leqslant \frac{3T}{4}\right) = \mathbb{Q}_i\left(\sum_{t=1}^{T} W_t - T\mathbb{E}_{\mathbb{Q}_i}[W_1] \leqslant \frac{3T}{4} - (1-p)T\right) \leqslant \exp\left(-T/32\right) \leqslant \delta/2\,.$$

The first part of the statement follows from the union bound and because if $N_i(T) \leqslant T/2$ and $\sum_{t=1}^{T} W_t \geqslant 3T/4$, then

$$\Delta \sum_{t=1}^{T} \mathbb{1}_{\{Z_t \in [2\Delta, 1-2\Delta] \text{ and } I_t \neq i\}} \geqslant \Delta T/4\,.$$

For the second part we use below the tower rule (conditioning on I_t and the history h_t up to $t-1$) and the fact that W_t is independent of I_t given h_t but also independent of h_t to get that

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}_i}[R_T(\ell^i_{1:T})] &\geqslant \Delta \sum_{t=1}^{T} \mathbb{Q}_i\left(W_t = 1 \text{ and } I_t \neq i\right) \\
&\geqslant \Delta \sum_{t=1}^{T} \mathbb{Q}_i\left(W_t = 1\right)\mathbb{Q}_i\left(I_t \neq i\right) \geqslant \frac{7\Delta}{8}\mathbb{E}_{\mathbb{Q}_i}[T - N_i(T)]\,,
\end{aligned}$$

which completes the proof.

E Proof of Corollary 3

Suppose on the contrary that such a strategy exists. Then

$$\begin{aligned}\mathbb{E}[R_T(\ell_{1:T})] &\leqslant \int_0^\infty \mathbb{P}\left(R_T(\ell_{1:T}) \geqslant x\right) dx \\ &\leqslant \int_0^\infty \exp\left(-\left(\frac{x}{C\sqrt{(K-1)T}}\right)^{\frac{1}{p}}\right) dx \leqslant C\sqrt{(K-1)T}\,.\end{aligned}$$

By the assumption in the corollary we have

$$\begin{aligned}\delta &\geqslant \mathbb{P}\left(R_T(\ell_{1:T}) \geqslant C\sqrt{(K-1)T}\log^p(1/\delta)\right) \\ &= \mathbb{P}\left(R_T(\ell_{1:T}) \geqslant \frac{\sqrt{(K-1)T}\log(1/(16\delta))}{203C} \cdot \frac{203C^2\log^p(1/\delta)}{\log(1/(16\delta))}\right),\end{aligned}$$

which leads to a contradiction by choosing δ sufficiently small and T sufficiently large and applying Theorem 2 to show that there exists an $\ell_{1:T} \in [0,1]^{KT}$ for which

$$\mathbb{P}\left(R_T(\ell_{1:T}) \geqslant \frac{\sqrt{(K-1)T}\log(1/(16\delta))}{203C}\right) \geqslant 2\delta\,.$$