On First-Order Bounds, Variance and Gap-Dependent Bounds for Adversarial Bandits

Roman Pogodin
Gatsby Computational Neuroscience Unit
University College London, London, UK
roman.pogodin.17@ucl.ac.uk

Tor Lattimore
DeepMind
London, UK
tor.lattimore@gmail.com

Abstract

We make three contributions to the theory of k-armed adversarial bandits. First, we prove a first-order bound for a modified variant of the INF strategy by Audibert and Bubeck [2009], without sacrificing worst case optimality or modifying the loss estimators. Second, we provide a variance analysis for algorithms based on follow the regularised leader, showing that without adaptation the variance of the regret is typically $\Omega(n^2)$ where n is the horizon. Finally, we study bounds that depend on the degree of separation of the arms, generalising the results by Cowan and Katehakis [2015] from the stochastic setting to the adversarial and improving the result of Seldin and Slivkins [2014] by a factor of $\log(n)/\log\log(n)$.

1 INTRODUCTION

The k-armed adversarial bandit is a sequential game played over n rounds. At the start of the game the adversary secretly chooses a sequence of losses $(\ell_t)_{t=1}^n$ with $\ell_t \in [0,1]^k$. In each round t the learner chooses a distribution P_t over the actions $[k] = \{1, 2, \ldots, k\}$. An action $A_t \in [k]$ is sampled from P_t and the learner observes the loss ℓ_{tA_t}. Like prior work we focus on controlling the regret, which is

$$\hat{R}_n = \max_{i \in [k]} \sum_{t=1}^{n} (\ell_{tA_t} - \ell_{ti}) \ .$$

This quantity is a random variable, so the standard objective is to bound $\hat{R}_n$ with high probability or its expectation: $R_n = \mathbb{E}[\hat{R}_n]$.

We make three contributions, with the common objective of furthering our understanding of the application of follow the regularised leader (FTRL) to adversarial bandit problems. Our first contribution is a modification of the INF policy by Audibert and Bubeck [2009] in order to prove first-order bounds (i.e. in terms of the loss of the best action) without sacrificing minimax optimality. Then we turn our attention to the variance of algorithms based on FTRL. Here we prove that using the standard importance-weighted estimators and a large class of potentials leads to a variance of $\Omega(n^2)$, which is the worst possible for bounded losses. Finally, we investigate the asymptotic performance of algorithms when there is a linear separation between the losses of the arms. We improve the result by Seldin and Slivkins [2014] by a factor of $\log(n)/\log\log(n)$ and generalise known results in the stochastic setting by Cowan and Katehakis [2015] to the adversarial one by constructing an algorithm for which the regret grows arbitrarily slowly almost surely.

Related work The literature on adversarial bandits is enormous. See the books by Bubeck and Cesa-Bianchi [2012] and Lattimore and Szepesvári [2019] for a comprehensive account. The common thread in the three components of our analysis is adaptivity for algorithms based on follow the regularised leader. The INF policy that underlies much of our analysis was introduced by Audibert and Bubeck [2009]. The connection to mirror descent and follow the regularised leader came later [Audibert and Bubeck, 2010, Bubeck and Cesa-Bianchi, 2012], which greatly simplified the analysis. The principle justification for introducing this algorithm was to prove bounds on the minimax regret. Remarkably, it was recently shown that by introducing a non-adaptive decaying learning rate, the algorithm retains minimax optimality while simultaneously achieving a near-optimal logarithmic regret in the stochastic setting [Zimmert and Seldin, 2019]. Despite its simplicity, the algorithm improves on the state-of-the-art for this problem Bubeck and Slivkins [2012], Seldin and Slivkins [2014], Seldin and Lugosi [2017]. See also

the extension to the combinatorial semibandit setting [Zimmert et al., 2019]. First-order bounds for bandits were first given by Allenberg et al. [2006], who analysed a modification of Exp3 [Auer et al., 1995]. As far as we know, previous algorithms with first order bounds have not been minimax optimal ($R_n = O(\sqrt{kn})$): the recent work by Neu [2015b] achieved $O(\sqrt{kn(\log(k)+1)})$ expected regret, and [Wei and Luo, 2018] had a $O(\sqrt{kn\log n})$ bound. Both papers used the idea of an adaptive learning rate similar to our analysis. In the setting of gains rather than losses Audibert and Bubeck [2010] have shown that by introducing biased estimators it is possible to prove a bound of $O(\sqrt{kG^*})$ where G^* is the maximum gain. Although it is not obvious, we suspect the same idea could be applied in our setting. We find it interesting nevertheless that the same affect is possible without modifying the loss estimators. The aforementioned work also assumes knowledge of G^*. Possibly our adaptive learning rates could be used to make this algorithm anytime without a doubling trick.

Although it is well known that straightforward applications of follow the regularised leader or mirror descent with importance-weighted estimators leads to poor concentration of the regret, we suspect the severity of the situation is not widely appreciated. As far as we know, the quadratic variance of Exp3 was only derived recently [Lattimore and Szepesvári, 2019, §11]. There are, however, a number of works modifying the importance-weighted estimators to prove high probability bounds Auer et al. [1995], Abernethy and Rakhlin [2009], Neu [2015a] with matching lower bounds by Gerchinovitz and Lattimore [2016]. Finally, we note there are many kinds of adaptivity beyond first-order bounds. For example sparsity and variance [Bubeck et al., 2018, Hazan and Kale, 2011, and others].

2 NOTATION

Given a vector $x \in \mathbb{R}^d$ let $\operatorname{diag}(x) \in \mathbb{R}^{d\times d}$ be the diagonal matrix with x along the diagonal. The interior of a topological space X is $\operatorname{interior}(X)$ and its boundary is ∂X. The standard basis vectors are $e_1, \ldots, e_d$. The $(d-1)$-dimensional probability simplex is $\Delta^{d-1} = \{x \in [0,1]^d : \|x\|_1 = 1\}$. A convex function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ has domain $\operatorname{dom}(F) = \{x \in \mathbb{R}^d : F(x) \neq \infty\}$. The Bregman divergence with respect to a differentiable F is a function $D_F : \operatorname{dom}(F) \times \operatorname{dom}(F) \to [0,\infty]$ defined by $D_F(x,y) = F(x) - F(y) - \langle \nabla F(y), x - y\rangle$. The Fenchel dual of F is $F^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ defined by $F^*(u) = \sup_{x\in\mathbb{R}^d} \langle x, u\rangle - F(x)$.

There are k arms and the horizon is n, which may or may not be known. The losses are $(\ell_t)_{t=1}^n$ with $\ell_t \in [0,1]^k$. We let $L_t = \sum_{s=1}^t \ell_s$. The importance-weighted estimator of ℓ_t is $\hat{\ell}_t$ defined by $\hat{\ell}_{ti} = \mathbb{1}\{A_t = i\}\,\ell_{ti}/P_{ti}$. All algorithms proposed here ensure that $P_{ti} > 0$ for all t and i, so this quantity is always well defined. Let $\hat{L}_t = \sum_{s=1}^t \hat{\ell}_s$. Expectations are with respect to the randomness in the actions $(A_t)_{t=1}^n$. Of course the learner can only choose P_t based on information available at the start of round t. Let $\mathcal{F}_t = \sigma(A_1, \ldots, A_t)$. Then P_t is $\mathcal{F}_{t-1}$-measurable. Let $A_{ti} = \mathbb{1}\{A_t = i\}$ and $T_i(t) = \sum_{s=1}^t A_{si}$ be the number of times arm i is played in the first t rounds. Our standing assumption is that the first arm is optimal. All our algorithms are symmetric, so this is purely for notational convenience.

Potential	**Definition**	**Alg.**
Negentropy	$\frac{1}{\eta}\sum_{i=1}^k p_i(\log(p_i) - 1)$	Exp3
1/2-Tsallis	$-\frac{2}{\eta}\sum_{i=1}^k \sqrt{p_i}$	INF
Log barrier	$-\frac{1}{\eta}\sum_{i=1}^k \log(p_i)$	

Table 1: Common potential functions

Assumption 2.1. $L_{t1} = \min_{i\in[k]} L_{ti}$.

3 FOLLOW THE REGULARISED LEADER

Follow the regularized leader (FTRL) is a popular tool for online optimization [Shalev-Shwartz, 2007, Hazan, 2016]. The basic algorithm depends on a sequence of potential functions $(F_t)_{t=1}^\infty$ where $F_t : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex and $\operatorname{dom}(F_t) \cap \Delta^{k-1} \neq \emptyset$. In each round the algorithm chooses the distribution

$$P_t = \operatorname*{arg\,min}_{p\in\Delta^{k-1}} \langle p, \hat{L}_{t-1}\rangle + F_t(p)\,,$$

which we assume exists. The action $A_t \in [k]$ is sampled from P_t. In many applications $F_t = F$ is chosen in a time independent way, with examples given in Table 1. This has the disadvantage that F must be chosen in advance in a way that depends on the horizon, which may be unknown. This weakness can be overcome by choosing $F_t = F/\eta_t$ where $(\eta_t)_{t=1}^\infty$ is a sequence of learning rates, which may be chosen in advance or adaptively in a data-dependent way.

A modification that will prove useful is to let $(\mathcal{A}_t)_{t=1}^\infty$ be a sequence of subsets of Δ^{k-1} and define

$$P_t = \operatorname*{arg\,min}_{p\in\mathcal{A}_t} \langle p, \hat{L}_{t-1}\rangle + F_t(p)\,.$$

The restriction to a subset of Δ^{k-1} can be useful to control the gradients of $F_t(P_t)$, which is sometimes crucial. The following theorem provides a generic bound for FTRL with changing potentials and constraint sets.

The result is reminiscent of many previous bounds for FTRL, but a reference for this result seems elusive. Most related is the generic analysis by Joulani et al. [2017], which also provides the most comprehensive literature summary.

Theorem 3.1. *Assume $\mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{A}_{n+1} \subseteq \Delta^{k-1}$ and $(F_t)_{t=1}^{n+1}$ is a sequence of convex functions with $\mathrm{dom}(F_t) \cap \mathcal{A}_t \neq \emptyset$ for all t. Define*

$$d_t = \max_{y \in \mathcal{A}_{t+1}} \min_{x \in \mathcal{A}_t} \|x - y\|_1 \,, \quad g_t = \sup_{x \in \mathcal{A}_t} \|\nabla F_t(x)\|_\infty$$
$$\text{and} \qquad v_n = \sum_{t=1}^n d_t(g_t + (t-1)) \,.$$

Then the regret of FTRL is bounded by

$$\begin{aligned} R_n \leq v_n &+ \mathbb{E}\left[\sum_{t=1}^n \left\langle P_t - P_{t+1}, \hat{\ell}_t \right\rangle - D_{F_t}(P_{t+1}, P_t)\right] \\ &+ \mathbb{E}\left[\min_{p \in \mathcal{A}_{n+1}} (F_{n+1}(p) + n\|p - e_1\|_1) - F_1(P_1)\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^n (F_t(P_{t+1}) - F_{t+1}(P_{t+1}))\right] . \end{aligned}$$

Proof. Let $p \in \mathcal{A}_{n+1}$. Using the fact that $\hat{\ell}_t$ is unbiased,

$$\begin{aligned} R_n &= \mathbb{E}\left[\sum_{t=1}^n \langle P_t - e_1, \hat{\ell}_t \rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n \langle P_t - p, \hat{\ell}_t \rangle\right] + \mathbb{E}\left[\sum_{t=1}^n \langle p - e_1, \ell_t \rangle\right] . \end{aligned}$$

The second sum is the approximation error, and by Holder's inequality,

$$\sum_{t=1}^n \langle p - e_1, \ell_t \rangle \leq \|p - e_1\|_1 \sum_{t=1}^n \|\ell_t\|_\infty \leq n\|p - e_1\|_1 \,.$$

Therefore,

$$\begin{aligned} R_n \leq \mathbb{E}\left[\sum_{t=1}^n \langle P_t - P_{t+1}, \hat{\ell}_t \rangle + \sum_{t=1}^n \langle P_{t+1} - p, \hat{\ell}_t \rangle\right] \\ + n\|p - e_1\|_1 \,. \end{aligned}$$

Let $\Phi_t(q) = F_t(q) + \sum_{s=1}^{t-1} \langle q, \hat{\ell}_s \rangle$, which is chosen so that $P_t = \arg\min_{q \in \mathcal{A}_t} \Phi_t(q)$. Then the second sum in the above display equals

$$\begin{aligned} &\sum_{t=1}^n (\Phi_{t+1}(P_{t+1}) - \Phi_t(P_{t+1}) - F_{t+1}(P_{t+1}) + F_t(P_{t+1})) \\ &\qquad - \Phi_{n+1}(p) + F_{n+1}(p) \\ &= \sum_{t=1}^n (\Phi_t(P_t) - \Phi_t(P_{t+1})) \\ &\qquad + \Phi_{n+1}(P_{n+1}) - \Phi_1(P_1) - \Phi_{n+1}(p) \\ &\qquad + F_{n+1}(p) + \sum_{t=1}^n (F_t(P_{t+1}) - F_{t+1}(P_{t+1})) \end{aligned}$$

We can rewrite the Φ-differences as

$$\begin{aligned} &\Phi_t(P_t) - \Phi_t(P_{t+1}) \\ &\qquad = -D_{\Phi_t}(P_{t+1}, P_t) - \langle \nabla \Phi_t(P_t), P_{t+1} - P_t \rangle \,. \end{aligned}$$

Let $\delta_t = P_{t+1} - \arg\min_{q \in \mathcal{A}_t} \|q - P_{t+1}\|_1$. Then due to first-order optimality condition for P_t on $\mathcal{A}_t$,

$$\mathbb{E}\left[\langle \nabla \Phi_t(P_t), (P_{t+1} - \delta_t) - P_t \rangle\right] \geq 0,$$

therefore

$$\begin{aligned} \mathbb{E}\left[\langle \nabla \Phi_t(P_t), P_{t+1} - P_t \rangle\right] &\geq \mathbb{E}\left[\langle \nabla \Phi_t(P_t), \delta_t \rangle\right] \\ &\geq \mathbb{E}\left[\langle \nabla F_t(P_t), \delta_t \rangle + \sum_{s=1}^{t-1} \langle \hat{\ell}_s, \delta_t \rangle\right] \\ &\geq -\mathbb{E}\left[\|\delta_t\|_1 \left(\|\nabla F_t(P_t)\|_\infty + \sum_{s=1}^{t-1} \|\ell_s\|_\infty\right)\right] \\ &\geq -d_t g_t - d_t \sum_{s=1}^{t-1} \|\ell_s\|_\infty \geq -d_t(g_t + (t-1)) \,, \end{aligned}$$

where we used Holder's inequality, the definitions of d_t and g_t, non-negativity of $\hat{\ell}_s$ and that $\mathbb{E}\,\hat{\ell}_s = \ell_s \in [0,1]$. It follows that

$$\Phi_t(P_t) - \Phi_t(P_{t+1}) \leq d_t(g_t + k(t-1)) - D_{\Phi_t}(P_{t+1}, P_t) \,.$$

Since $p \in \mathcal{A}_{n+1}$ and P_{n+1} is the minimiser of Φ_{n+1} in $\mathcal{A}_{n+1}$, we have

$$\Phi_{n+1}(P_{n+1}) - \Phi_{n+1}(p) \leq 0 \,.$$

Finally, noting that $\Phi_1 = F_1$ and $D_{\Phi_t}(P_{t+1}, P_t) = D_{F_t}(P_{t+1}, P_t)$ we obtain

$$\begin{aligned} R_n \leq n\|p - e_1\|_1 &+ \sum_{t=1}^n d_t(g_t + (t-1)) \\ &+ \mathbb{E}\left[\sum_{t=1}^n \langle P_t - P_{t+1}, \hat{\ell}_t \rangle - D_{F_t}(P_{t+1}, P_t)\right] \\ &+ \mathbb{E}\left[F_{n+1}(p) - F_1(P_1)\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^n (F_t(P_{t+1}) - F_{t+1}(P_{t+1}))\right] , \end{aligned}$$

from which the statement follows. □

4 FIRST ORDER BOUNDS

We now introduce the modification of the INF strategy, which takes inspiration from Wei and Luo [2018], Zimmert and Seldin [2019], Zimmert et al. [2019]. The new algorithm plays on the 'chopped' simplex, with the magnitude of the cut dependent on the round,

$$\mathcal{A}_t = \Delta^{k-1} \cap [1/t, 1]^k \,. \tag{1}$$

Then for a convex potential $f_t(p)$ with $\mathrm{dom}(f_t)^k \cap \Delta^{k-1} \neq \emptyset$ define a potential

$$F_t(p) = \frac{1}{\eta_t} \sum_{i=1}^{k} f_t(p_i) \,, \tag{2}$$

where the learning rate η_t is given by

$$\eta_t = \frac{\eta_0}{\sqrt{1 + \sum_{s=1}^{t-1} \hat{\ell}_{sA_s}^2 (\nabla^2(f_s)(P_{sA_s}))^{-1}}} \,, \tag{3}$$

where η_0 is positive constant to be tuned later.

The Hessian of the potential plays a fundamental role in the regret, simplifying the derivation of a generic first-order bound:

Theorem 4.1. *Suppose that $\nabla^2 f_t$ is decreasing on $(0,1)$ and there exist $B, C \geq 0$ such that*

$$\frac{1}{p^2 \nabla^2 f_t(p)} \leq B, \quad \mathbb{E}\left[\frac{1}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})}\right] \leq C \,,$$

for all $p \in (0,1)$ and $t \in [n]$. Assume additionally that there exist a non-negative constant h_1 and a non-negative function $h_2(n)$ such that

$$v_n + \min_{p \in \mathcal{A}_{n+1}} (F_{n+1}(p) + \|p - e_1\|_1 n) - F_1(P_1) + \sum_{t=1}^{n} (F_t(P_{t+1}) - F_{t+1}(P_{t+1})) \leq \frac{h_1}{\eta_{n+1}} + h_2(n) \,,$$

almost surely. Then the expected regret of FTRL with $\eta_0 = \sqrt{h_1}/2^{1/4}$ simultaneously satisfies

$$R_n \leq \frac{\sqrt{h_1}}{2^{5/4}} B + h_2(n) + 2\sqrt{2} B h_1 + 2^{7/4}\sqrt{h_1} \times \sqrt{1 + \frac{BL_{n1}}{2} + \frac{Bh_2(n)}{2} + B^2\left(\frac{\sqrt{h_1}}{2^{9/4}} + \frac{h_1}{\sqrt{2}}\right)} \,,$$

$$R_n \leq \frac{\sqrt{h_1}}{2^{5/4}} B + h_2(n) + 2^{7/4}\sqrt{h_1}\sqrt{1 + \frac{Cn}{2}} \,.$$

Remark 4.2. h_1 and $h_2(n)$ reflect the approximation error, non-stationarity of the potential f_t and how sensitive it is to the changes in $\mathcal{A}_t$. In a simple case with $\mathcal{A}_t = \mathcal{A}$, $f_t = f$ for all t, this is a standard bound for the sum of the potential differences. h_1 can be a function of n when the horizon n is known, as we choose the learning rate based on it.

As an application of this general first-order result, we derive a worst-case optimal bound for a carefully chosen mixture of the INF regularizer and the log-barrier:

Corollary 4.3. *For $\eta_0 = k^{1/4}\sqrt{\frac{13}{3\sqrt{2}} + \frac{3}{\sqrt{2}q}}$ and*

$$f_t(p) = -2\sqrt{p} - \frac{\log p}{\sqrt{k}\log^{1+q}\max\{3,t\}}$$

and any $q > 0$ and $n \geq 3$, the regret grows with n as

$$R_n = O\left(\sqrt{kL_{n1}\log^{1+q}(n) + k^2\log^{2(1+q)} n} + k\log^{1+q}(n)\right) \,,$$

with some constants proportional to $1/q$.

Corollary 4.4. *For $q = 1$, $\eta_0 = k^{1/4}\sqrt{22/(3\sqrt{2})}$,*

$$R_n \leq 19k^2 + 22k\log^2(n) + 2k\log(n) + 6.5\log(n) \times \sqrt{kL_{n1} + 19k^3 + 2k^2\log(n) + 11.2k^2\log^2(n)} \,.$$

In the worst-case scenario the regret satisfies

$$\limsup_{n\to\infty} \frac{R_n}{\sqrt{kn}} \leq 9.2.$$

Corollary 4.5. *If the horizon $n \geq 3$ is known in advance, using $\mathcal{A}_t = \Delta^{k-1} \cap [1/n, 1]^k$, $\eta_0 = k^{1/4}\sqrt{3}/2^{1/4}$ and*

$$f_t(p) = -2\sqrt{p} - \frac{\log p}{\sqrt{k}\log n}$$

results in

$$R_n \leq k + 9.1k\log n + 4.2\sqrt{kL_{n1}\log(n) + 2\sqrt{k} + 6k^2\log^2(n)} \,,$$

$$\limsup_{n\to\infty} \frac{R_n}{\sqrt{kn}} \leq 5.9 \,.$$

The proof of the last corollary simply repeats previous statements, also using the stationarity of the constraint set and $f_t(p)$. See the supplementary material for more details.

Remark 4.6. Theorem 4.1 with known n reproduces the result of [Wei and Luo, 2018] (note that they used a slightly different algorithm and the learning rate schedule): for the log-barrier potential $f_t(p) = -\log p$ we have $B = C = 1$ and $h_1(n) \propto k\log n$, such that the worst-case regret is $R_n = O(\sqrt{kn\log n})$.

The proof of Theorem 4.1 follows from Theorem 3.1 and the following lemmas:

Lemma 4.7. *For a potential of the form Eq. (2) with $\nabla^2 f_t(p)$ that is monotonically decreasing on $p \in (0,1)$,*

$$\sum_{t=1}^{n} \left\langle P_t - P_{t+1}, \hat{\ell}_t \right\rangle - D_{F_t}(P_{t+1}, P_t) \leq \sum_{t=1}^{n} \frac{\eta_t}{2} \frac{\ell_{tA_t}^2}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} .$$

Proof. Let $t \in [n]$ and suppose that $P_{t+1,A_t} > P_{tA_t}$. Then using the fact that the loss estimators and Bregman divergence are non-negative,

$$\begin{aligned}\langle P_t - P_{t+1}, \hat{\ell}_t \rangle - D_{F_t}(P_{t+1}, P_t) &\leq \langle P_t - P_{t+1}, \hat{\ell}_t \rangle \\ &= (P_{tA_t} - P_{t+1,A_t})\hat{\ell}_{tA_t} \leq 0 .\end{aligned}$$

Now suppose that $P_{t+1,A_t} \leq P_{tA_t}$. By [Lattimore and Szepesvári, 2019, Theorem 26.5],

$$\langle P_t - P_{t+1}, \hat{\ell}_t \rangle - D_{F_t}(P_{t+1}, P_t) \leq \frac{1}{2} \|\hat{\ell}_t\|^2_{(\nabla^2 F_t(z))^{-1}} ,$$

where $z = \alpha P_t + (1-\alpha) P_{t+1}$ for some $\alpha \in [0,1]$. By definition $\nabla^2 F_t(z) = \operatorname{diag}(\nabla^2 f_t(z))/\eta_t$ and since $\hat{\ell}_{ti} = 0$ for $i \neq A_t$,

$$\frac{1}{2} \|\hat{\ell}_t\|^2_{\nabla^2 F_t(z)^{-1}} = \frac{\eta_t \hat{\ell}_{tA_t}^2}{2\nabla^2 f_t(z_{A_t})} \leq \frac{\eta_t \hat{\ell}_{tA_t}^2}{2\nabla^2 f_t(P_{tA_t})} ,$$

where we used the fact that $z_{A_t} \leq P_{tA_t}$ and that $\nabla^2 f_t(p)$ is decreasing. The result follows by substituting the definition of $\hat{\ell}_{tA_t}$ and summing over $t \in [n]$. □

Lemma 4.8. *Let $(x_t)_{t=1}^n$ be a sequence with $x_t \in [0, B]$ for all t. Then*

$$\sum_{t=1}^{n} \frac{x_t}{\sqrt{1 + \sum_{s=1}^{t-1} x_s}} \leq 4 \sqrt{1 + \frac{1}{2} \sum_{t=1}^{n} x_t} + B .$$

The proof follows from a comparison to an integral and is given in the supplementary material.

Proof of Theorem 4.1. Using the result of Theorem 3.1, Lemma 4.7 and the assumption on the difference in the potentials, we have

$$R_n \leq h_2(n) + \mathbb{E}\left[\frac{h_1}{\eta_{n+1}} + \sum_{t=1}^{n} \frac{\eta_t}{2} \frac{\ell_{tA_t}^2}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \right] .$$

As $\ell_{tA_t} \leq 1$, we can apply Lemma 4.8 with

$$x_t = \frac{\ell_{tA_t}^2}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \leq B .$$

It follows that $\eta_t = \eta_0 / \sqrt{1 + \sum_{s=1}^{t-1} x_s}$, and thus

$$\sum_{t=1}^{n} \frac{\eta_t}{2} \frac{\ell_{tA_t}^2}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \leq 2\eta_0 \sqrt{1 + \frac{1}{2} \sum_{t=1}^{n} \frac{\ell_{tA_t}^2}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})}} + \frac{\eta_0}{2} B .$$

The first term in the last line is proportional to $1/\eta_{n+1}$, therefore using the definition of η_t, Jensen's inequality and $\ell_{tA_t}^2 \leq \ell_{tA_t}$, the regret can be bounded as

$$R_n \leq h_2(n) + \frac{\eta_0}{2} B + \left(\frac{\sqrt{2} h_1}{\eta_0} + 2\eta_0 \right) \times \sqrt{1 + \frac{1}{2} \mathbb{E}\left[\sum_{t=1}^{n} \frac{\ell_{tA_t}}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \right]} .$$

The first bound in the theorem follows from

$$\mathbb{E}\left[\sum_{t=1}^{n} \frac{\ell_{tA_t}}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \right] \leq B \mathbb{E}\left[\sum_{t=1}^{n} (\ell_{tA_t} - \ell_{t1} + \ell_{t1}) \right] = B R_n + B L_{n1}$$

and then from choosing $\eta_0 = \sqrt{h_1}/2^{1/4}$ and solving the resulting quadratic equation with respect to R_n.

For the second bound, we use $l_{tA_t} \leq 1$ and the definition of C, such that

$$\mathbb{E}\left[\sum_{t=1}^{n} \frac{\ell_{tA_t}}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \right] \leq \frac{Cn}{2} .$$

□

To prove the corollaries, we need to bound h_1, $h_2(n)$, B, and C:

Lemma 4.9. *The Hessian of the hybrid potential in Corollary 4.3 is monotonically decreasing, and for $n \geq 3$*

$$\frac{1}{p^2 \nabla^2 f_t(p)} \leq \sqrt{k} \log^{1+q} n, \quad \mathbb{E}\left[\frac{1}{P_{tA_t}^2 \nabla^2 f_t(P_{tA_t})} \right] \leq 2\sqrt{k} ,$$

Proof. For $p \in \operatorname{interior}(\Delta^{k-1})$,

$$\nabla^2 f_t(p) = \frac{1}{2p^{3/2}} + \frac{1}{p^2 \sqrt{k} \log^{1+q} \max\{3, t\}}$$

is a decreasing function of p. It follows that for $n \geq 3$

$$\frac{1}{p^2 \nabla^2 f_t(p)} \leq \sqrt{k} \log^{1+q} n .$$

Moreover,

$$\sup_{t,P_t\in\mathcal{A}_t}\mathbb{E}\left[\frac{1}{P_{tA_t}^2\nabla^2 f_t(P_{tA_t})}\right] \le \sup_{t,P_t\in\Delta^{k-1}}\mathbb{E}\left[\frac{2}{\sqrt{P_{tA_t}}}\right] = 2\sqrt{k}\,. \qquad \square$$

Lemma 4.10. *Under the conditions of Corollary 4.3,*

$$v_n \le \frac{\sqrt{k}}{\eta_{n+1}}\left(\frac{4}{3}+\frac{2}{q}\right) + \frac{5.5k}{\eta_0}\sqrt{1+9k^{3/2}\log^{1+q}(9k^{3/2})} + \frac{3.7\sqrt{k}}{\eta_0}\sqrt{1+3\sqrt{k}\log^{1+q}3} + 2k\log n\,.$$

Proof. Due to the chopped simplex and the factorised potential, we have (recall the definition in Theorem 3.1 and bound the second sum with the integral, as shown in the supplementary material)

$$\begin{aligned} v_n &= \sum_{t=1}^{n} d_t(g_t+(t-1)) \\ &\le \sum_{t=1}^{n}\frac{2k}{t^2}\left(\frac{1}{\eta_t}\sup_{p\in[1/t,1]}|\nabla f_t(p)|+(t-1)\right) \\ &\le \sum_{t=1}^{n}\frac{2k}{t^2}\left(\frac{1}{\eta_t}\sup_{p\in[1/t,1]}|\nabla f_t(p)|\right)+2k\log n\,. \end{aligned}$$

For $p\in[1/t,1]$ the gradient is bounded as

$$\begin{aligned} |\nabla f_t(p)| &\le \frac{1}{\sqrt{p}} + \frac{1}{p\sqrt{k}\log^{1+q}\max\{3,t\}} \\ &\le \sqrt{t} + \frac{t}{\sqrt{k}\log^{1+q}\max\{3,t\}}\,. \end{aligned}$$

Therefore, the corresponding sum in v_n converges. By a straightforward calculation (see the supplementary material), the Hessian is bounded as in Lemma 4.9),

$$v_n \le \frac{\sqrt{k}}{\eta_{n+1}}\left(\frac{4}{3}+\frac{2}{q}\right) + \frac{5.5k}{\eta_0}\sqrt{1+9k^{3/2}\log^{1+q}(9k^{3/2})} + \frac{3.7\sqrt{k}}{\eta_0}\sqrt{1+3\sqrt{k}\log^{1+q}3} + 2k\log n\,. \qquad \square$$

Lemma 4.11. *Under the conditions of Corollary 4.3,*

$$\begin{aligned} &\min_{p\in\mathcal{A}_{n+1}}(F_{n+1}(p)+\|p-e_1\|_1 n) - F_1(P_1) \\ &\quad+\sum_{t=1}^{n}(F_t(P_{t+1})-F_{t+1}(P_{t+1})) \\ &\le \frac{\sqrt{k}}{\eta_{n+1}}\left(3+\frac{1}{q}\right)+k+\frac{\sqrt{k}}{3\eta_0}\sqrt{1+3\sqrt{k}\log^{1+q}3}\,. \end{aligned}$$

Proof. The potential is a mixture of the INF and the log-barrier parts, $F_t(p) = -\frac{2}{\eta_t}\sum_i\sqrt{p_i} - \frac{\alpha_t}{\eta_t}\sum_i\log p_i$ with $\alpha_t = 1/(\sqrt{k}\log^{1+q}\max\{3,t\})$.

To control the contribution of the INF term, first notice that the INF part of $F_{n+1}(p)$ is negative. Moreover,

$$\left(-\frac{2}{\eta_t}+\frac{2}{\eta_{t+1}}\right)\sum_{i=1}^{k}\sqrt{P_{t+1,i}} \le 2\sqrt{k}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\,.$$

Summing with the INF part of $-F_1(P_1)$ and telescoping shows that it contributes at most $2\sqrt{k}/\eta_{n+1}$ to the sum.

For log-barrier, suppose $\alpha_t/\eta_t \le \alpha_{t+1}/\eta_{t+1}$. Then

$$\left(-\frac{\alpha_t}{\eta_t}+\frac{\alpha_{t+1}}{\eta_{t+1}}\right)\sum_{i=1}^{k}\log(P_{t+1,i}) \le 0\,.$$

Now suppose that $\alpha_t/\eta_t > \alpha_{t+1}/\eta_{t+1}$. For $t\ge 3$, as $P_{t+1}\in\mathcal{A}_{t+1}$,

$$\begin{aligned} &\left(-\frac{\alpha_t}{\eta_t}+\frac{\alpha_{t+1}}{\eta_{t+1}}\right)\sum_{i=1}^{k}\log(P_{t+1,i}) \\ &\quad\le \left(\frac{\alpha_t}{\eta_t}-\frac{\alpha_{t+1}}{\eta_{t+1}}\right)k\log(t+1) \\ &\quad\le \frac{1}{\eta_t}\left(\frac{\log(t+1)}{\log^{1+q}(t)}-\frac{1}{\log^q(t)}\right)\sqrt{k} \\ &\quad\le \frac{\sqrt{k}}{\eta_t t\log^{1+q}t}\,. \end{aligned}$$

Summing over t and noting that due to $\alpha_1=\alpha_2=\alpha_3$ the potential is unchanged,

$$\begin{aligned} &\sum_{t=1}^{n}\left(-\frac{\alpha_t}{\eta_t}+\frac{\alpha_{t+1}}{\eta_{t+1}}\right)\sum_{i=1}^{k}\log w_{t+1,i} \\ &\quad\le \sum_{t=3}^{n}\frac{\sqrt{k}}{\eta_t t\log^{1+q}t} \le \frac{\sqrt{k}}{\eta_{n+1}q}+\frac{\sqrt{k}}{3\eta_3}\,, \end{aligned}$$

where the last inequality (shown in the supplementary material) essentially compares the sum to the integral of $1/(t\log^{1+q}t)$ and uses that $1/\log^q t\le 1$ for $t\ge 3$. We can further bound η_3 as

$$\frac{1}{\eta_3} \le \frac{1}{\eta_0}\sqrt{1+3\sqrt{k}\log^{1+q}3}$$

by using the fact that the Hessian is bounded (see the proof of Lemma 4.9).

Finally, the log-barrier part of $-F_1(P_1)$ is negative. The log-barrier part of $F_{n+1}(p)$ is bounded by $\sqrt{k}/\eta_{n+1}$ as $p\in\mathcal{A}_{n+1}$. Thus,

$$\begin{aligned} &\min_{p\in\mathcal{A}_{n+1}}F_{n+1}(p)+n\|p-e_1\|_1 \\ &\quad\le \frac{\sqrt{k}}{\eta_{n+1}}+\min_{p\in\mathcal{A}_{n+1}}n\|p-e_1\|_1 = \frac{\sqrt{k}}{\eta_{n+1}}+\frac{kn}{n+1}\,. \end{aligned}$$

Combining the three bounds and using that $kn/(n+1) \leq k$ concludes the proof. □

Proof of Corollary 4.3. From Lemma 4.9, Lemma 4.10 and Lemma 4.11, we find

$$B = \sqrt{k}\log^{1+q} n\,,\; C = 2\sqrt{k}\,,\; h_1 = \sqrt{k}\left(\frac{13}{3} + \frac{3}{q}\right)\,,$$
$$h_2(n) = 2k\log n + \frac{5.5.k}{\eta_0}\sqrt{1 + 9k^{3/2}\log^{1+q}(9k^{3/2})} + \frac{4.1}{\eta_0}\sqrt{k}\sqrt{1 + 3\sqrt{k}\log^{1+q} 3} + k\,.$$

□

Now applying Theorem 4.1 with $\eta_0 = k^{1/4}\sqrt{\frac{13}{3\sqrt{2}} + \frac{3}{\sqrt{2}q}}$ completes the proof. Note that in the big-O notation, we only kept the leading terms that grow with n.

Proof of Corollary 4.4. Starting from the end of the previous proof, choosing $q = 1$ and upper-bounding the numerical coefficients, we obtain the corollary. □

5 VARIANCE OF THE REGRET

The expected regret is just one measure of the performance of an algorithm. Algorithms with small expected regret may suffer from a large variance. Since the adversarial model is often motivated on the grounds of providing robustness, it would be unfortunate if proposed algorithms suffered from high variance. Recently, however, it was shown that the variance of Exp3 without exploration is quadratic in the horizon [Lattimore and Szepesvári, 2019, §11], and a similar result holds for Thompson sampling in a Bayesian setting [Bubeck and Sellke, 2019]. Here we generalise these arguments to prove quadratic variance of the regret for a class of algorithms based on FTRL with importance-weighted loss estimators. This is the worst possible result for bandits with bounded losses. The class of policies covered by our theorem includes INF and Exp3, but not FTRL with the log barrier. To keep things simple we restrict ourselves to algorithms of the form

$$P_t = \operatorname*{arg\,min}_{p \in \Delta^{k-1}} \langle p, \hat{L}_{t-1}\rangle + \frac{1}{\eta_n}\sum_{i=1}^{k} f(p_i)\,,$$

where f is convex and $(\eta_n)_{n=1}^{\infty}$ is a sequence of learning rates. Note that this corresponds to a sequence of algorithms, each with a fixed learning rate.

Assumption 5.1. The number of actions is $k = 2$ and f is Legendre with $(0,1) \subseteq \operatorname{dom}(f)$ and $0 \in \partial\operatorname{dom}(f)$.

The assumption on the potential is satisfied by all standard potentials for bandits on the probability simplex, including those in Table 1. It allows us to write P_t in a simple form. Let $g(p) = f(p) + f(1-p)$, which is convex and Legendre with $\operatorname{dom}(g) = (0,1)$. Given $x \geq 0$,

$$\operatorname*{arg\,min}_{p \in [0,1]} (px + g(p)) = \nabla g^*(-x)\,,$$

where we used the fact that for Legendre functions the gradient is invertible and $(\nabla g)^{-1} = \nabla g^*$. That g is Legendre with $\operatorname{dom}(g) = (0,1)$ also ensures that ∇g^* is nondecreasing and $\lim_{x\to-\infty}\nabla g^*(x) = 0$ and $\lim_{x\to\infty}\nabla g^*(x) = 1$. By symmetry, we also have $\nabla g^*(0) = 1/2$. The point is that by the definition of FTRL, $P_{t1} = \nabla g^*(\eta_n(\hat{L}_{t-1,2} - \hat{L}_{t-1,1}))$.

Theorem 5.2. *Assume* $\limsup_{n\to\infty} n\nabla g^*(-an\eta_n) < \infty$ *for all* $a > 0$. *Then for all sufficiently large* n *there exists a bandit for which* $\mathbb{P}(\hat{R}_n \geq n/4) \geq c$, *where* $c > 0$ *is a constant that depends on the algorithm, but not the horizon.*

Corollary 5.3. *Under the same conditions as Theorem 5.2 the variance of the regret is* $\operatorname{Var}[\hat{R}_n] = \Omega(n^2)$.

Examples Suppose $\eta_n = an^{-1/2}$ for some $a > 0$. Then the conditions of the theorem are satisfied when f is the negentropy. In this case ∇g^* is the sigmoid function and the corresponding algorithm is just Exp3. When $f(p) = -2\sqrt{p}$ and $x \leq 0$, then

$$\nabla g^*(x) = \frac{1}{2}\left(1 - \sqrt{1 + \frac{4\left(2\sqrt{1+x^2} - 2 - x^2\right)}{x^4}}\right)\,,$$

which satisfies $\limsup_{n\to\infty}\nabla g^*(-a\sqrt{n})n = 1/a^2$. In this sense 1/2-Tsallis entropy with $\eta_n = \Theta(n^{-1/2})$ just barely satisfies the conditions. The consequence is that the minimax optimal INF policy proposed by Audibert and Bubeck [2009] has quadratic variance. The log barrier does not satisfy the conditions and we speculate it is more stable.

Proof of Theorem 5.2. Assume for simplicity that 4 is a factor of n. Let $\alpha_n \in [0, 1/2]$ be a constant to be tuned subsequently and consider a bandit defined by

$$\ell_{t1} = \begin{cases} \alpha_n & \text{if } t \leq n/2 \\ 0 & \text{otherwise}\,. \end{cases} \qquad \ell_{t2} = \begin{cases} 0 & \text{if } t \leq n/2 \\ 1 & \text{otherwise}\,. \end{cases}$$

Clearly the first arm is optimal. Let $c_1 > 0$ be a constant such that for all sufficiently large n it holds that $\nabla g^*(-n\eta_n) \leq c_1/n$, which is guaranteed to exist by the assumptions in the theorem. Then define events $F_t = \cap_{s=n/2+1}^{t}\{A_s = 2, P_{s1} \leq c_1/n\}$. On the event F_n the random regret satisfies

$$\hat{R}_n \geq \frac{n}{2} - \frac{\alpha_n n}{2} \geq \frac{n}{4}\,. \tag{4}$$

The theorem follows by proving that $\mathbb{P}(F_n) \geq c$ for all sufficiently large n and constant $c > 0$. The idea is to show that the estimated loss for the optimal arm after the first $n/2$ rounds is large enough that the algorithm never plays the optimal arm in the second half of the game with constant probability.

First half dynamics The choice of α_n determines the dynamics of the interaction between the algorithm and environment in the first $n/2$ rounds. Before the main proof we establish some facts about this. Let $\alpha \in [0, 1/2]$ and define $(p_s(\alpha))_{s=0}^n$ inductively by $p_0(\alpha) = 1/2$ and

$$p_{s+1}(\alpha) = \nabla g^* \left(-\eta_n \sum_{u=0}^{s} \frac{\alpha}{p_u(\alpha)} \right),$$

which is chosen so that $P_{t+1,1} = p_s(\alpha)$ whenever $t + 1 \leq n/2$ and $T_1(t) = s$. Here we used the fact that $\hat{L}_{t2} = 0$ for $t \leq n/2$, which follows from the definition of the bandit. Let $Q_s(\alpha) = \sum_{u=0}^{s-1} \alpha / p_u(\alpha)$. Clearly $Q_2(1/2) > 0$ and $Q_s(0) = 0$ for all s. Furthermore, $Q_s(\alpha)$ is increasing in both α and s and continuous in α. Therefore there exists an $\alpha_\circ \in (0, 1/2)$ such that $Q_2(1/2) \geq Q_3(\alpha_\circ)$. Now suppose that $Q_s(1/2) \geq Q_{s+1}(\alpha_\circ)$. Using the fact that ∇g^* is increasing,

$$\begin{aligned} Q_{s+1}(1/2) &= Q_s(1/2) + \frac{1}{2} \nabla g^* \left(-\eta_n Q_s(1/2) \right)^{-1} \\ &\geq Q_{s+1}(\alpha_\circ) + \alpha_\circ \nabla g^* \left(-\eta_n Q_{s+1}(\alpha_\circ) \right)^{-1} = Q_{s+2}(\alpha_\circ), \end{aligned}$$

which by induction means that $Q_s(1/2) \geq Q_{s+1}(\alpha_\circ)$ for all $s \geq 2$. Notice that $\hat{L}_{t1} = Q_s(\alpha_n)$ when $T_1(t-1) = s$.

Second half dynamics Define threshold λ_n by

$$\lambda_n = n + n^2/(2(n - c_1)) \leq 2n,$$

where the latter inequality holds for all sufficiently large n. Let E be the event $E = \{\hat{L}_{n/2,1} \geq \lambda_n\}$. We claim that $\mathbb{P}(F_n \mid E) \geq \exp(-c_1/2)$. Suppose that $t > n/2$ and $E \cap F_t$ occurs. Then

$$\hat{L}_{t2} = \sum_{s=n/2+1}^{t} \frac{1}{P_{s2}} \leq \sum_{s=n/2+1}^{t} \frac{1}{1 - c_1/n} \leq \frac{n^2}{2(n - c_1)},$$

where the first inequality follows from the definition of F_t. Therefore, since $\hat{L}_{t1} \geq \hat{L}_{n/2,1} \geq \lambda_n$,

$$\begin{aligned} P_{t+1,1} &= \nabla g^*(\eta_n(\hat{L}_{t2} - \hat{L}_{t1})) \\ &\leq \nabla g^* \left(\eta_n \left(\frac{n^2}{2(n - c_1)} - \lambda_n \right) \right) \leq \frac{c_1}{n}. \end{aligned} \quad (5)$$

Hence $\mathbb{P}(F_{t+1} \mid F_t, E) \geq 1 - c_1/n$. Noting that Eq. (5) implies that $P_{n/2+1,1} \leq c_1/n$ shows that $E \subseteq F_{n/2+1}$ and hence by induction

$$\mathbb{P}(F_n \mid E) \geq \left(1 - \frac{c_1}{n}\right)^{n/2} \geq \exp(-c_1/2). \quad (6)$$

Lower bounding $\mathbb{P}(E)$ By Eqs. (4) and (6) it suffices to prove that $\mathbb{P}(E)$ is larger than a constant for sufficiently large n. Let $s = \min\{u : Q_u(1/2) \geq \lambda_n\}$, which by our assumptions on ∇g^* for sufficiently large n is at least $s > 2$ and at most $s \leq n/2$. Then $Q_s(\alpha_\circ) \leq Q_{s-1}(1/2) < \lambda_n \leq Q_s(1/2)$. By the intermediate value theorem and the continuity of $\alpha \mapsto Q_s(\alpha)$ we may choose $\alpha_n \in (\alpha_\circ, 1/2]$ such that $Q_s(\alpha_n) = \lambda_n$. Now introduce a sequence of independent geometric random variables $(G_u)_{u=0}^s$ with $G_u \in \{1, 2, \ldots\}$ and $\mathbb{E}[G_u] = 1/p_u(\alpha)$. Then by construction,

$$\mathbb{P}(T_1(n/2) \geq s) = \mathbb{P}\left(\sum_{u=0}^{s-1} G_u \leq \frac{n}{2} \right). \quad (7)$$

You should think of G_u as the number of rounds before the algorithm plays action 1 for the uth time. Let

$$\kappa = \min \left\{ m : \sum_{u=0}^{s-m-1} \frac{1}{p_u(\alpha_n)} \leq \frac{n}{8} \right\}.$$

Then either $\sum_{u=0}^{s-\kappa-1} 1/p_u(\alpha_n) \leq n/16$ in which case $1/p_{s-\kappa}(\alpha_n) \geq n/16$ or $\sum_{u=0}^{s-\kappa-1} 1/p_u(\alpha_n) \geq n/16$. Then there exists a constant $c_2 \geq 0$ such that for sufficiently large n,

$$\begin{aligned} p_{s-\kappa}(\alpha_n) &= \nabla g^*(-\eta_n Q_{s-\kappa-1}(\alpha_n)) \\ &= \nabla g^* \left(-\eta_n \sum_{u=0}^{s-\kappa-1} \frac{\alpha_n}{p_u(\alpha_n)} \right) \\ &\leq \nabla g^* \left(-\frac{\alpha_\circ n \eta_n}{16} \right) \leq \frac{c_2}{n}. \end{aligned}$$

Combining the two cases and choosing $c_2 \geq 16$ guarantees that $p_{s-\kappa}(\alpha_n) \leq c_2/n$ for sufficiently large n. Using the fact that $s \mapsto p_s(\alpha_n)$ is decreasing,

$$2n \geq \lambda_n = \sum_{u=0}^{s-1} \frac{\alpha_n}{p_u(\alpha_n)} \geq \sum_{u=s-\kappa}^{s-1} \frac{\alpha_\circ n}{c_2} = \frac{\kappa \alpha_\circ n}{c_2}.$$

Rearranging shows that κ is less than a constant that is independent of n. By Markov's inequality

$$\begin{aligned} &\mathbb{P}\left(\sum_{u=0}^{s-\kappa-1} G_u \geq \frac{n}{4} \right) \\ &\qquad \leq \mathbb{P}\left(\sum_{u=0}^{s-\kappa-1} G_u \geq 2 \sum_{u=0}^{s-\kappa-1} \frac{1}{p_u(\alpha_n)} \right) \leq \frac{1}{2}. \end{aligned}$$

Hence

$$\mathbb{P}\left(\sum_{u=0}^{s-\kappa-1} G_u < \frac{n}{4} \right) \geq \frac{1}{2}. \quad (8)$$

Furthermore,

$$\frac{\alpha_\circ}{p_{s-1}(\alpha_n)} \leq \frac{\alpha_n}{p_{s-1}(\alpha_n)} \leq \sum_{u=0}^{s-1} \frac{\alpha_n}{p_u(\alpha_n)} = Q_s(\alpha_n) = \lambda_n \leq 2n\,.$$

Therefore, using again that $s \mapsto p_s(\alpha)$ is decreasing,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{u=s-\kappa}^{s-1} G_u \leq \frac{n}{4}\right) \\
&\geq \binom{n/4}{\kappa} p_{s-1}(\alpha_n)^\kappa \left(1 - p_{s-\kappa}(\alpha_n)\right)^{n/4-\kappa} \\
&\geq \binom{n/4}{\kappa} \left(\frac{\alpha_\circ}{2n}\right)^\kappa \left(1 - \frac{c_2}{n}\right)^{n/4-\kappa}\,,
\end{aligned}$$

which for sufficiently large n is larger than a strictly positive constant and the result follows by combining the above with Eqs. (7) and (8). □

Remark 5.4. We believe the result continues to hold for adaptive learning rates under the assumption that $\limsup_{t\to\infty} t\nabla g^*(-a t\eta_t) < \infty$ for all $a > 0$. The proof becomes significantly more delicate, however.

6 LINEARLY SEPARABLE BANDITS

In this section we consider the case where the adversary chooses an infinite sequence of loss vectors $(\ell_t)_{t=1}^\infty$. The main objective is to prove logarithmic (or better) regret under the following assumption.

Assumption 6.1. There is a linear separation between the optimal and suboptimal arms:

$$\Delta_i = \liminf_{n\to\infty} (L_{ni} - L_{n1})/n > 0 \quad \text{for all } i > 1\,.$$

Note that if $(\ell_t)_{t=1}^\infty$ are independent and identically distributed random vectors, then the above holds almost surely whenever there is a unique optimal arm. We provide two results in this setting. The first generalises a known result from stochastic bandits that there exist algorithms for which the asymptotic random regret grows arbitrarily slowly almost surely [Cowan and Katehakis, 2015].

Theorem 6.2. *For any nondecreasing function $f : \mathbb{N} \to \mathbb{N}$ with $\lim_{n\to\infty} f(n) = \infty$ there exists an algorithm such that $\limsup_{n\to\infty} \hat{R}_n / f(n) < \infty$ almost surely.*

The algorithm realising the bound in Theorem 6.2 explores uniformly at random on a set E for which $\limsup_{n\to\infty} |E \cap [n]|/f(n) \leq 1$ almost surely. The reader is warned that the constants hidden by the asymptotics are potentially quite enormous.

Of course this result says nothing about the expected regret, which must be logarithmic for consistent algorithms [Lai and Robbins, 1985]. The following theorem improves on a result by Seldin and Slivkins [2014] by a factor of $\log(n)/\log\log(n)$.

Theorem 6.3. *There exists an algorithm such that for any adversarial bandit $R_n = O(\sqrt{kn})$. Furthermore, under Assumption 6.1 it holds that*

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)^2 \log\log(n)} < \infty\,.$$

The algorithm is INF with enough forced exploration that the loss estimators are guaranteed to be sufficiently accurate to detect a linear separation. The proofs of Theorems 6.2 and 6.3 use standard concentration results and are given in the supplementary material.

7 OPEN QUESTIONS

Despite the relatively long history and extensive research, many open questions exist about k-armed adversarial bandits. Perhaps the most exciting question is the existence/nature of a genuinely instance-optimal algorithm. The work by Zimmert and Seldin [2019] suggests the possibility of an algorithm for which $R_n = O(\sqrt{kn})$ and $R_n = O(\sum_{i:\Delta_i>0} \log(n)/\Delta_i)$, where $\Delta_i = \frac{1}{n}\sum_{t=1}^n (\ell_{ti} - \ell_{t1})$ is the empirical gap between the arms. In fact, one could hope for a little more. For stochastic Bernoulli bandits with means $(\theta_i)_{i=1}^k$, the KL-UCB algorithm by Cappé et al. [2013] satisfies $R_n = O(\sum_{i:\Delta_i>0} \Delta_i \log(n)/d(\theta_i, \theta^*))$ where $d(\theta_i, \theta_1)$ is the relative entropy between Bernoulli distributions with bias θ_i and θ_1 respectively. We are not aware of a lower bound proving that such a result is not possible for adversarial bandits with $\theta_i = \frac{1}{n}\sum_{t=1}^n \ell_{ti}$. At present it is not clear whether or not our modified algorithm from Corollary 4.3 retains the logarithmic regret in the stochastic setting, both because we use an adaptive learning rate and a hybrid potential. Finally, it is known that sub-exponential tail bounds are incompatible with logarithmic regret in the stochastic setting [Audibert et al., 2009], but by appropriately tuning the confidence intervals it is straightforward to prove the variance is linear in n, which is optimal. Missing is an adaptation of INF that enjoys (a) minimax regret, (b) logarithmic regret in the stochastic setting and (c) linear variance.

Acknowledgements

This work was supported by the Gatsby Charitable Foundation. The authors thank Haipeng Luo for spotting an error in the earlier version of the manuscript.

References

J. Abernethy and A. Rakhlin. Beating the adaptive bandit with high probability. In *2009 Information Theory and Applications Workshop*, pages 280–289, 2009. 2

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 the 17th International Conference on Algorithmic Learning Theory*, ALT, pages 229–243, Berlin, Heidelberg, 2006. Springer-Verlag. 2

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226, 2009. 1, 7

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of Conference on Learning Theory (COLT)*, 2010. 1, 2

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. 9

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. 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. 2

S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated, 2012. 1

S. Bubeck and M. Sellke. First-order regret analysis of thompson sampling. *arXiv preprint arXiv:1902.00681*, 2019. 7

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT*, pages 42.1–42.23, 2012. 1

S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In F. Janoos, M. Mohri, and K. Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 111–127. PMLR, 07–09 Apr 2018. 2

O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013. 9

W. Cowan and M. N. Katehakis. Asymptotic behavior of minimal-exploration allocation policies: Almost sure, arbitrarily slow growing regret. *arXiv preprint arXiv:1505.02865*, 2015. 1, 9

S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, NIPS, pages 1198–1206. Curran Associates, Inc., 2016. 2

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4): 157–325, 2016. 2

E. Hazan and S. Kale. A simple multi-armed bandit algorithm with optimal variation-bounded regret. In S. M. Kakade and U. von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 817–820. PMLR, 2011. 2

P. Joulani, A. György, and Cs. Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, and variational bounds. In *ALT*, 2017. 3

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. 9

T. Lattimore and Cs. Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019. 1, 2, 5, 7

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 3168–3176. Curran Associates, Inc., 2015a. 2

G. Neu. First-order regret bounds for combinatorial semi-bandits. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1360–1375, Paris, France, 03–06 Jul 2015b. PMLR. 2

Y. Seldin and G. Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1743–1759, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR. 1

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1287–1295, Bejing, China, 22–24 Jun 2014. PMLR. 1, 9

S. Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, The Hebrew University of Jerusalem, 2007. 2

C-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 06–09 Jul 2018. 2, 4

J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 467–475. PMLR, 16–18 Apr 2019. 1, 4, 9

J. Zimmert, H. Luo, and C-Y. Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7683–7692. PMLR, 09–15 Jun 2019. 2, 4