Connections Between Mirror Descent, Thompson Sampling and the Information Ratio

Julian Zimmert
DeepMind, London/
University of Copenhagen
zimmert@di.ku.dk

Tor Lattimore
DeepMind, London
lattimore@google.com

Abstract

The information-theoretic analysis by Russo and Van Roy [25] in combination with minimax duality has proved a powerful tool for the analysis of online learning algorithms in full and partial information settings. In most applications there is a tantalising similarity to the classical analysis based on mirror descent. We make a formal connection, showing that the information-theoretic bounds in most applications can be derived from existing techniques for online convex optimisation. Besides this, for k-armed adversarial bandits we provide an efficient algorithm with regret that matches the best information-theoretic upper bound and improve best known regret guarantees for online linear optimisation on ℓ_p-balls and bandits with graph feedback.

1 Introduction

The combination of minimax duality and the information-theoretic machinery by Russo and Van Roy [25] has proved a powerful tool in the analysis of online learning algorithms. This has led to short and insightful analysis for k-armed bandits, linear bandits, convex bandits and partial monitoring, all improving on prior best known results. The downside is that the approach is non-constructive. The application of minimax duality demonstrates the existence of an algorithm with a given bound in the adversarial setting, but provides no way of constructing that algorithm.

The fundamental quantity in the information-theoretic analysis is the 'information ratio' in round t, which informally is

$$\text{information ratio}_t = \frac{(\text{expected regret in round } t)^2}{\text{expected information gain in round } t},$$

where the information gain is either measured using the mutual information [25] or a generalisation based on a Bregman divergence [21]. Proving the information ratio is small corresponds to showing that either the learner is suffering small regret in round t or gaining information, which ultimately leads to a bound on the cumulative regret. The aforementioned generalisation by Lattimore and Szepesvári [21] (restated in the supplementary) lead to a short analysis for k-armed adversarial bandits that is minimax optimal except for small constant factors. The authors speculated that the new idea should lead to improved bounds for a range of online learning problems and suggested a number of applications, including bandits with graph feedback [3] and linear bandits on ℓ_p-balls [11].

We started to follow this plan, successfully improving existing minimax bounds for bandits with graph feedback and online linear optimisation for ℓ_p-balls with full information (the bandit setting remains a mystery). Along the way, however, we noticed a striking connection between the analysis techniques for bounding the information ratio and controlling the stability of online stochastic mirror descent (OSMD), which is a classical algorithm for online convex optimisation. A connection was

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

already hypothesised by Lattimore and Szepesvári [21], who noticed a similarity between the bounds obtained. Notably, why does using the negentropy potential in the information-theoretic analysis lead to almost identical bounds for k-armed bandits as Exp3? Why does this continue to hold with the Tsallis entropy and the INF strategy [6]?

Contribution Our main contribution is a formal connection between the information-theoretic analysis and OSMD. Specifically, we show how tools for analysing OSMD can be applied to a modified version of Thompson sampling that uses the same sampling strategy as OSMD, but replaces the mirror descent update with a Bayesian update. This contribution is valuable for several reasons: (a) it explains the similarity between the information-theoretic and OSMD style analysis, (b) it allows for the transfer of techniques for OSMD to Bayesian regret analysis and (c) it opens the possibility of a constructive transfer of ideas from Bayesian regret analysis to the adversarial framework, as we illustrate in the next contribution.

A curiosity in the Bayesian analysis of adversarial k-armed bandits is that the resulting bound was always a factor of 2 smaller than the corresponding bound for OSMD. This was true in the original analysis [25] and its generalisation [21]. Our new theorem entirely explains the difference, and indeed, allows us to improve the bounds for OSMD. This leads to an efficient algorithm for adversarial k-armed bandits with regret $\mathfrak{R}_n \leq \sqrt{2kn} + O(k)$, matching the information-theoretic upper bound except for small lower-order terms.

Finally, we improve the regret guarantees for two online learning problems. First, for bandits with graph feedback we improve the minimax regret in the 'easy' setting by a $\log(n)$ factor, matching the lower bound up to a factor of $\log^{3/2}(k)$. Second, for online linear optimisation over the ℓ_p-balls we improve existing bounds by arbitrarily large constant factors. At first we had proved these results using the information-theoretic tools and minimax duality, but here we present the unified view and consequentially the analysis also applies to OSMD for which we have efficient algorithms.

Related work The information-theoretic Bayesian regret analysis was introduced by [24, 25, 26]. The focus in these papers is on the analysis of Bayesian algorithms in the stochastic setting, a line of work continued recently by [15]. [10] noticed that the stochastic assumption is not required and that the results continued to hold in a Bayesian adversarial setting where the prior is over arbitrary sequences of losses, rather than over (parametric) distributions as is usual in Bayesian statistics. The idea to use minimax duality to derive minimax regret bounds is due to [1] and has been applied and generalised by a number of authors [10, 17, 21, 9]. Mirror descent was developed by [22] and [23] for optimization. As far as we know its first application to bandits was by [2], which precipitated a flood of papers as summarised in the books by [8, 20]. We work in the partial monitoring framework, which goes back to [27]. Most of the focus since then has been on classifying the growth of the regret on the horizon for finite partial monitoring games [13, 16, 5, 7, 19]. Bandits with graph feedback are a special kind of partial monitoring problem and have been studied extensively [3, 14, 4, and others], with a monograph on the subject by [28]. Online linear optimisation is an enormous subject by itself. We refer the reader to the books by [12, 18].

Notation The reader will find omitted proofs in the supplementary material. Let $[n] = \{1, 2, \ldots, n\}$ and $B_p^d = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$ be the standard ℓ_p-ball. For positive definite A we write $\|x\|_A^2 = x^\top Ax$. Given a topological space X, let $\operatorname{int}(X)$ be its interior and $\Delta(X)$ be the space of probability measures on X with the Borel σ-algebra. We write $X^\circ = \{y \in \mathbb{R}^d : \sup_{x \in X} |\langle x, y\rangle| \leq 1\}$ for the functional analysts polar and $\operatorname{co}(X)$ for the convex hull of X. The domain of a convex function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is $\operatorname{dom}(F) = \{x : F(x) < \infty\}$. For $x, y \in \operatorname{dom}(F)$ the Bregman divergence between x and y with respect to F is $\mathrm{D}_F(x, y) = F(x) - F(y) - \nabla F_{x-y}(y)$ where $\nabla_v F(x)$ is the directional derivative of F at x in the direction v. The diameter of X with respect to F is $\operatorname{diam}_F(X) = \sup_{x,y \in X} F(x) - F(y)$. We abuse notation by writing $\nabla^{-2} F(x) = (\nabla^2 F(x))^{-1}$. For $x, y \in \mathbb{R}^d$ we let $[x, y] = \operatorname{co}(\{x, y\})$ be the convex hull of x and y, which is the set of points on the chord between x and y.

Linear partial monitoring Our results are most easily expressed in a linear version of the partial monitoring framework, which extends the standard adversarial linear bandit framework to general feedback structures. Let $\mathcal{A}$ be the action space and $\mathcal{L}$ the loss space, which are subsets of $\mathbb{R}^d$ with $\mathcal{A}$ compact. The convex hull of $\mathcal{A}$ is $\mathcal{X} = \operatorname{co}(\mathcal{A})$. When $\mathcal{A}$ is finite we let $k = |\mathcal{A}|$. The signal function is a known function $\Phi : \mathcal{A} \times \mathcal{L} \to \Sigma$ for some observation space Σ. An adversary and learner interact over n rounds. First the adversary secretly chooses $(\ell_t)_{t=1}^n$ with $\ell_t \in \mathcal{L}$ for all t. In each round t the learner samples an action $A_t \in \mathcal{A}$ from a distribution depending on observations $A_1, \Phi_1, \dots, A_{t-1}, \Phi_{t-1}$ where $\Phi_s = \Phi(A_s, \ell_s)$ is the observation in round s. The regret of policy π in environment $(\ell_t)_{t=1}^n$ is

$$\mathfrak{R}_n(\pi, (\ell_t)_{t=1}^n) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a, \ell_t \rangle\right],$$

where the expectation is with respect to the randomness in the actions. The regret depends on a policy and the losses. The minimax regret is

$$\mathfrak{R}_n^* = \inf_\pi \sup_{(\ell_t)_{t=1}^n} \mathfrak{R}_n(\pi, (\ell_t)_{t=1}^n),$$

where the infimum is over all policies and the supremum over all loss sequences in $\mathcal{L}^n$. From here on the dependence of $\mathfrak{R}_n$ on the policy and loss sequence is omitted.

Examples The standard k-armed bandit is recovered when $\mathcal{A} = \{e_1, \dots, e_k\}$, $\mathcal{L} = [0,1]^k$ and $\Phi(a, \ell) = \langle a, \ell \rangle \in \Sigma = [0,1]$. For linear bandits the set $\mathcal{A}$ is an arbitrary compact set and $\mathcal{L}$ is typically $\mathcal{A}^\circ$. Bandits with graph feedback have a richer signal function as we explain in Section 4.

Bayesian setting In the Bayesian setting the sequence of losses $(\ell_t)_{t=1}^n$ are sampled from a known prior probability measure ν on $\mathcal{L}^n$ and subsequently the learner interacts with the sampled losses as normal. The optimal action is now a random variable $A^* = \arg\min_{a \in \mathcal{A}} \sum_{t=1}^n \langle a, \ell_t \rangle$ and the Bayesian regret is

$$\mathfrak{BR}_n = \mathbb{E}\left[\sum_{t=1}^n \langle A_t - A^*, \ell_t \rangle\right].$$

Finally, define $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ with $\mathcal{F}_t = \sigma(A_1, \Phi_1, \dots, A_t, \Phi_t)$, $\Delta_t = \langle A_t - A^*, \ell_t \rangle$. A crucial piece of notation is $X_t = \mathbb{E}_{t-1}[A_t] \in \mathcal{X}$, which is the conditional expected action played in round t.

2 Mirror descent, Thompson sampling and the information ratio

We now develop the connection between OSMD and the information-theoretic Bayesian regret analysis. Specifically we show that instances of OSMD can be transformed into an algorithm similar to Thompson sampling (TS) for which the Bayesian regret can be bounded in the same way as the regret of the original algorithm. The similarity to TS is important. Any instance of OSMD with a uniform bound on the adversarial regret enjoys the same bound on the Bayesian regret for any prior without modification. Our result has a different flavour because we prove a bound for a variant of OSMD that replaces the mirror descent update with a Bayesian update.

Algorithm 1: OSMD

Input: $\mathscr{A} = (P, E, F)$ and η
Initialize $X_1 = \arg\min_{a \in \mathcal{X}} F(a)$
for $t = 1, \dots, n$ **do**
  Sample $A_t \sim P_{X_t}$ and observe Φ_t
  Construct: $\hat{\ell}_t = E(X_t, A_t, \Phi_t)$
  Update: $X_{t+1} = f_t(X_t, A_t)$

OSMD is a modular algorithm that depends on defining three components: (1) A sampling scheme that determines how the algorithm explores, (2) a method for estimating the unobserved loss vectors, and (3) a convex 'potential' and learning rate that determines how the algorithm updates its iterates. The following definition makes this more precise.

Definition 1. An instance of OSMD is determined by a tuple $\mathscr{A} = (P, F, E)$ and learning rate $\eta > 0$ such that

(a) The sampling scheme is a collection $P = \{P_x : x \in \mathcal{X}\}$ of probability measures in $\Delta(\mathcal{A})$ such that $\mathbb{E}_{A \sim P_x}[A] = x$ for all $x \in \mathcal{X}$.

(b) The potential is a Legendre function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with $\mathrm{dom}(F) \cap \mathcal{X} \neq \emptyset$ and $\eta > 0$ is the learning rate.

(c) The estimation function is $E : \mathcal{X} \times \mathcal{A} \times \Sigma \to \mathbb{R}^d$, which we assume satisfies $\mathbb{E}_{A \sim P_x}[E(x, A, \Phi(A, \ell))] = \ell$ for all $\ell \in \mathcal{L}$ and $x \in \mathcal{X}$.

The assumptions on the mean of P_x and that E is unbiased are often relaxed in minor ways, but for simplicity we maintain the strict definition. For the remainder we fix $\mathscr{A} = (P, F, E)$ and $\eta > 0$ and abbreviate

$$E_t(x, a) = E(x, a, \Phi(a, \ell_t)) \qquad \text{and} \qquad \hat{\ell}_t = E(X_t, A_t, \Phi_t)\,.$$

You should think of $E_t(x, a)$ as the estimated loss vector when the learner plays action a while sampling from P_x and $\hat{\ell}_t$ as the realisation of this estimate in round t. OSMD starts by initialising X_1 as the minimiser of F constrained to $\mathcal{X}$. Subsequently it samples $A_t \sim P_{X_t}$ and updates

$$X_{t+1} = \operatorname*{arg\,min}_{y \in \mathcal{X}} \eta \langle y, \hat{\ell}_t \rangle + \mathrm{D}_F(y, X_t)\,.$$

A useful notation is to let $(f_t)_{t=1}^n$ and $(g_t)_{t=1}^n$ be sequences of functions from $\mathcal{X} \times \mathcal{A}$ to $\mathbb{R}^d$ with

$$\begin{aligned} f_t(x, a) &= \operatorname*{arg\,min}_{y \in \mathcal{X}} \left(\eta \langle y, E_t(x, a)\rangle + \mathrm{D}_F(y, x)\right) \qquad \text{and} \\ g_t(x, a) &= \operatorname*{arg\,min}_{y \in \operatorname{int}(\operatorname{dom}(F))} \left(\eta \langle y, E_t(x, a)\rangle + \mathrm{D}_F(y, x)\right)\,, \end{aligned}$$

which means that $X_{t+1} = f_t(X_t, A_t)$, while g_t is the same as f_t, but without the constraint to $\mathcal{X}$. The complete algorithm is summarised in Algorithm 1. The next theorem is well known [20, §28].

Theorem 2 (OSMD REGRET BOUND)**.** *The regret of OSMD satisfies*

$$\mathfrak{R}_n \leq \frac{\operatorname{diam}_F(\mathcal{X})}{\eta} + \frac{\eta}{2} \mathbb{E}\left[\sum_{t=1}^n \operatorname{stab}_t(X_t; \eta)\right]\,,$$

where $\operatorname{stab}_t(x; \eta) = \dfrac{2}{\eta} \mathbb{E}_{A \sim P_x}\left[\langle x - f_t(x, A), E_t(x, A)\rangle - \dfrac{\mathrm{D}_F(f_t(x, A), x)}{\eta}\right]$.

The random variable $\operatorname{stab}_t(X_t; \eta)$ measures the stability of the algorithm relative to the learning rate and is usually almost surely bounded. The diameter term depends on how fast the algorithm can move from the starting point to optimal, which is large when the learning rate is small. In this sense the learning rate is tuned to balance the stability of the algorithm and the requirement that (X_t) can tend towards an optimal point. Note that $\operatorname{stab}_t(x)$ depends on P, E, F, η and the loss vector ℓ_t, which means that in the Bayesian setting the stability function is random. The next lemma is also known and is often useful for bounding the stability function.

Lemma 3. *Suppose that F is twice differentiable on* $\operatorname{int}(\operatorname{dom}(F))$*, then*

$$\operatorname{stab}_t(x; \eta) \leq \mathbb{E}_{A \sim P_x}\left[\sup_{z \in [x, f_t(x, A)]} \|E_t(x, A)\|^2_{\nabla^{-2} F(z)}\right]\,.$$

Furthermore, provided that $g_t(x, a)$ exists for all a in the support of P_x, then

$$\operatorname{stab}_t(x; \eta) \leq \mathbb{E}_{A \sim P_x}\left[\sup_{z \in [x, g_t(x, A)]} \|E_t(x, A)\|^2_{\nabla^{-2} F(z)}\right]\,.$$

Bayesian analysis Modified Thompson sampling (MTS) is a variant of TS summarised in Algorithm 2 that depends on a prior distribution ν and a sampling scheme P. The algorithm differs from Algorithm 1 in the computation of X_t. Rather than using the mirror descent update, it uses the Bayesian expected optimal action conditioned on the observations. Expectations in this subsection are with respect to both the prior and the actions, which means that $(\ell_t)_{t=1}^n$ are randomly distributed according to ν and consequently the functions f_t, g_t and stab_t are random. Our main theorem is the following bound on the Bayesian regret of MTS.

Algorithm 2: MTS

Input: Prior ν and P
Initialize $X_1 = \mathbb{E}[A^*]$
for $t = 1, \ldots, n$ **do**
  Sample $A_t \sim P_{X_t}$ and observe Φ_t
  Update: $X_{t+1} = \mathbb{E}_{t-1}[A^*]$

Theorem 4. *MTS satisfies* $\mathfrak{BR}_n \leq \dfrac{\operatorname{diam}_F(\mathcal{X})}{\eta} + \dfrac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \operatorname{stab}_t(X_t;\eta)\right]$.

Remark 5. The stability function depends on $\mathscr{A} = (P, F, E)$ and η while Algorithm 2 only uses P. In this sense Theorem 4 shows that MTS satisfies the given bound for all E, F and η. MTS is the same as TS when sampling from the posterior is the same as sampling from P_{X_t}. A fundamental case where this always holds is when $\mathcal{A} = \{e_1, \ldots, e_d\}$ because each $x \in \mathcal{X}$ is uniquely represented as a linear combination of elements in $\mathcal{A}$ and hence P_x is unique.

Proof of Theorem 4. Beginning with the definition of the per-step regret,

$$
\begin{aligned}
\mathbb{E}_{t-1}[\Delta_t] &= \langle X_t, \mathbb{E}_{t-1}[\ell_t]\rangle - \mathbb{E}_{t-1}[\langle A^*, \ell_t\rangle] \\
&= \langle X_t, \mathbb{E}_{t-1}[\hat{\ell}_t]\rangle - \mathbb{E}_{t-1}\left[\langle A^*, \hat{\ell}_t\rangle\right] && (1)\\
&= \langle X_t, \mathbb{E}_{t-1}[\hat{\ell}_t]\rangle - \mathbb{E}_{t-1}\left[\langle \mathbb{E}_{t-1}[A^* \mid A_t, \Phi_t], \hat{\ell}_t\rangle\right] && (2)\\
&= \mathbb{E}_{t-1}\left[\langle X_t - X_{t+1}, \hat{\ell}_t\rangle\right] && (3)\\
&\leq \mathbb{E}_{t-1}\left[\langle X_t - f_t(X_t, A_t), \hat{\ell}_t\rangle - \frac{1}{\eta}\mathrm{D}_F(f_t(X_t, A_t), X_t) + \frac{1}{\eta}\mathrm{D}_F(X_{t+1}, X_t)\right] && (4)\\
&\leq \mathbb{E}_{t-1}\left[\frac{\eta}{2}\operatorname{stab}_t(X_t;\eta) + \frac{1}{\eta}\mathrm{D}_F(X_{t+1}, X_t)\right]. && (5)
\end{aligned}
$$

Eq. (1) uses that the loss estimators are unbiased. Eq. (2) follows using the tower rule for conditional expectations and the fact that $\hat{\ell}_t$ is a measurable function of X_t, A_t and Φ_t so that

$$\mathbb{E}_{t-1}[\langle A^*, \hat{\ell}_t\rangle] = \mathbb{E}_{t-1}[\mathbb{E}_{t-1}[\langle A^*, \hat{\ell}_t\rangle \mid A_t, \Phi_t]] = \mathbb{E}_{t-1}[\langle \mathbb{E}_{t-1}[A^* \mid A_t, \Phi_t], \hat{\ell}_t\rangle] = \mathbb{E}_{t-1}[\langle X_{t+1}, \hat{\ell}_t\rangle].$$

Eq. (3) uses the definitions of X_{t+1}. Eq. (4) follows from the definition of f_t, which implies that

$$\langle f_t(X_t, A_t), \hat{\ell}_t\rangle + \frac{1}{\eta}\mathrm{D}_F(f_t(X_t, A_t), X_t) \leq \langle X_{t+1}, \hat{\ell}_t\rangle + \frac{1}{\eta}\mathrm{D}_F(X_{t+1}, X_t).$$

Finally, Eq. (5) follows from the definition of stab_t. The proof is completed by summing over the per-step regret, noting that $(X_t)_{t=1}^n$ is a $(\mathcal{F}_t)_t$-adapted martingale and by [21, Theorem 3],

$$\mathbb{E}\left[\sum_{t=1}^n \mathrm{D}_F(X_{t+1}, X_t)\right] \leq \mathbb{E}[F(X_{n+1})] - F(X_1) \leq \operatorname{diam}_F(\mathcal{X}). \qquad \square$$

The stability coefficient The only difference between Theorems 2 and 4 is the trajectory of $(X_t)_{t=1}^n$ and the randomness of the stability function. In most analyses of OSMD the final bound is obtained via a uniform bound on $\operatorname{stab}_t(x;\eta)$ that holds regardless of the losses and in this case the trajectory X_t is irrelevant. This is formalised in the following definition and corollary. Define the stability coefficients by

$$\operatorname{stab}(\mathscr{A};\eta) = \sup_{x\in\mathcal{X}} \max_{t\in[n]} \operatorname{stab}_t(x;\eta) \qquad \text{and} \qquad \operatorname{stab}(\mathscr{A}) = \sup_{\eta>0} \operatorname{stab}(\mathscr{A};\eta).$$

Corollary 6. *The regret of Algorithm 1 for an appropriately tuned learning rate is bounded by*

$$\mathfrak{R}_n \leq \sqrt{2\operatorname{diam}_F(\mathcal{X})\operatorname{stab}(\mathscr{A})n}.$$

The Bayesian regret of Algorithm 2 is bounded by $\mathfrak{BR}_n \leq \sqrt{2\operatorname{diam}_F(\mathcal{X})\operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A}))n}$.

The essential supremum is needed because the stability coefficient depends on the losses $(\ell_t)_{t=1}^n$, which are random in the Bayesian setting. Generally speaking, however, bounds on the stability coefficient are proven in a manner that is independent of the losses.

Remark 7. Often $\operatorname{stab}(\mathscr{A};\eta) \leq a + b\eta$ for constants $a, b \geq 0$ and $\operatorname{stab}(\mathscr{A}) = \infty$. Nevertheless, the same argument shows that the regret of Algorithm 1 is bounded by

$$\mathfrak{R}_n \leq \sqrt{2a\operatorname{diam}_F(\mathcal{X})n} + \frac{b\operatorname{diam}_F(\mathcal{X})}{a},$$

and similarly for the Bayesian regret of Algorithm 2.

Stability and the information ratio The generalised information-theoretic analysis by [21] starts by assuming there exists a constant $\alpha > 0$ such that the following bound on the information ratio holds almost surely:

$$\text{information ratio}_t = \mathbb{E}_{t-1}[\Delta_t]^2 \Big/ \mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] \leq \alpha\,. \tag{6}$$

Then [21, Theorem 3] shows that

$$\mathfrak{BR}_n \leq \sqrt{\alpha n \operatorname{diam}_F(\mathcal{X})}\,. \tag{7}$$

The proof of Theorem 4 directly provides a bound on the information ratio in terms of the stability coefficient. To see this, notice that Eq. (5) holds for all measurable η and let

$$\eta = \sqrt{2\mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] / \operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A}))}\,. \tag{8}$$

Then by Eq. (5) and the definition of $\operatorname{stab}(\mathscr{A})$ it follows that

$$\mathbb{E}_{t-1}[\Delta_t]^2 \Big/ \mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] \leq 2 \operatorname{ess\,sup}(\operatorname{stab}(\mathscr{A})) \;\; a.s.\,.$$

In other words, the usual methods for bounding the stability coefficient in the analysis of OSMD can be used to bound the information ratio in the information-theoretic analysis.

Example 8. To make the abstraction more concrete, consider the k-armed bandit problem where $\mathcal{L} = [0,1]^k$ and $\mathcal{A} = \{e_1, \ldots, e_k\}$. In this case there is a unique sampling scheme defined by $P_x(a) = \langle x, a\rangle$. The standard loss estimation function is to use importance-weighting, which leads to

$$E_t(x, a)_i = \ell_{ti} \mathbb{1}(a = e_i) / x_i\,. \tag{9}$$

A commonly used potential is the unnormalised negentropy $F(x) = \sum_{i=1}^k x_i \log(x_i) - x_i$ that satisfies $\nabla^{-2} F(x) = \operatorname{diag}(x)$. The instance of OSMD resulting from these choices is called Exp3 for which an explicit form for X_t is well known:

$$X_{ti} = \exp\left(-\eta \textstyle\sum_{s=1}^{t-1} \hat{\ell}_{si}\right) \Big/ \left(\textstyle\sum_{j=1}^k \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{sj}\right)\right)\,.$$

A short calculation shows that $g_t(x, a)_i = x_i \exp(-\eta \hat{\ell}_{ti}) \leq x_i$. The stability function is bounded using the second part of Lemma 3 by

$$\begin{aligned}
\operatorname{stab}_t(x; \eta) &\leq \mathbb{E}_{A \sim P_x}\left[\sup_{z \in [x, g_t(x, A)]} \|E_t(x, A)\|^2_{\nabla^{-2} F(z)}\right] \\
&= \mathbb{E}_{A \sim P_x}\left[\sup_{z \in [x, g_t(x,A)]} \sum_{i=1}^k z_{ti} \frac{\mathbb{1}(A = e_i)\ell_{ti}^2}{x_{ti}^2}\right] = \mathbb{E}_{A \sim P_x}\left[\frac{\mathbb{1}(A = e_i)\ell_{ti}^2}{x_{ti}}\right] \leq \sum_{i=1}^k \ell_{ti}^2 \leq k\,.
\end{aligned}$$

Finally, the diameter of the probability simplex $\mathcal{X}$ with respect to the unnormalised negentropy is $\operatorname{diam}_F(\mathcal{X}) = \log(k)$. Applying Theorem 2 shows that the regret of OSMD and Bayesian regret of MTS satisfy

$$\mathfrak{R}_n \leq \sqrt{2nk\log(k)} \quad (\text{OSMD}) \qquad \text{and} \qquad \mathfrak{BR}_n \leq \sqrt{2nk\log(k)} \quad (\text{MTS})\,.$$

Remark 9. Theorems 2 and 4 are vacuous when $\operatorname{diam}_F(\mathcal{X}) = \infty$. The most straightforward resolution is to restrict X_t to a subset of $\mathcal{X}$ on which the diameter is bounded and then control the additive error. This idea also works in the Bayesian setting as described by [21]. We omit a detailed discussion to avoid technicalities.

3 Bandits

The best known bound on the minimax regret for k-armed bandits is $\mathfrak{R}_n \leq \sqrt{2kn}$ by [21]. They let $F(x) = -2\sum_{i=1}^k \sqrt{x_i}$ be the $1/2$-Tsallis entropy and prove that

$$\mathbb{E}_{t-1}[\Delta_t]^2 \Big/ \mathbb{E}_{t-1}[\mathrm{D}_F(X_{t+1}, X_t)] \leq \sqrt{k}\,.$$

By Cauchy-Schwarz $\operatorname{diam}_F(\mathcal{X}) \leq 2\sqrt{k}$ and then Eq. (7) shows that $\mathfrak{BR}_n \leq \sqrt{2nk}$ for all priors ν. Minimax duality is used to conclude that $\mathfrak{R}_n^* \leq \sqrt{2kn}$. Meanwhile, using the importance-weighted estimator in Eq. (9) leads to a bound on the stability coefficient of $\operatorname{stab}(\mathscr{A}) \leq 2\sqrt{k}$ and then Theorem 2 yields a bound of $\mathfrak{R}_n \leq \sqrt{8nk}$.

The discrepancy between these methods is entirely explained by the naive choice of importance-weighted estimator. The approach based on bounding the information ratio is effectively shifting the losses, which can be achieved in the OSMD framework by shifting the importance-weighted estimators (see Fig. 1). This idea reduces the worst-case variance of the importance weighted estimators by a factor of 4.

Lemma 10. *If the loss estimator in Example 8 with* $F(s) = -2\sum_{i=1}^k \sqrt{x_i}$ *is replaced by*

$$E_t(x,a)_i = \frac{(\ell_{ti} - c_{ti})\mathbb{1}(a = e_i)}{x_i} + c_{ti}\,,$$
$$\text{where } c_{ti} = \frac{1}{2}(1 - \mathbb{1}(X_{ti} < \eta^2))\,,$$

then the stability coefficient for any $\eta \leq 1/2$ *is bounded by* $\operatorname{stab}(\mathscr{A};\eta) \leq k^{1/2}/2 + 12k\eta$.

Theorem 11. *The regret of OSMD with the loss estimator of Lemma 10 and appropriate learning rate satisfies:* $\mathfrak{R}_n \leq \sqrt{2kn} + 48k$.

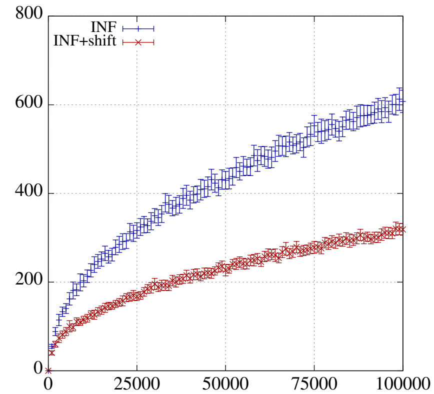

Figure 1: Comparison of INF with and without shifted loss estimators. x-axis is number of time-steps and y-axis the empirical regret estimation. η is tuned to the horizon and all experiments use Bernoulli losses with $\mathbb{E}[\ell_t] = (0.45, 0.55, \ldots, 0.55)^T$ $(k = 5)$. We repeat the experiment 100 times with error bars indicating three standard deviations. The empirical result matches our theoretical improvement of a factor 2.

4 Bandits with graph feedback

In bandits with graph feedback the action set is $\mathcal{A} = \{e_1, \ldots, e_k\}$ and $\mathcal{L} = [0,1]^k$. Let $E \subseteq [k] \times [k]$ be a set of directed edges over vertex set $[k]$ so that $\mathcal{G} = ([k], E)$ is a directed graph. The signal function is $\Phi(e_i, \ell) = \{(j, \ell_j) : j \in \mathcal{N}(i)\}$. The standard bandit framework is recovered when $E = \{(i,i) : i \in [k]\}$ while the full information setup corresponds to $E = [k] \times [k]$. Of course there are settings between and beyond these extremes. The difficulty of the graph feedback problem is determined by the connectivity of the graph. For example, when $E = \emptyset$, the learner has no way to estimate the losses and the regret is linear in the worst case. Like finite partial monitoring, graph feedback problems can be classified into one of four regimes for which:

$$\mathfrak{R}_n^* \in \left\{\mathcal{O}(1),\, \tilde{\Theta}(n^{1/2}),\, \Theta(n^{2/3}),\, \Omega(n)\right\}\,.$$

Our focus is on graph feedback problems that fit in the second category, which is the most challenging to analyse.

Definition 12. $\mathcal{G}$ is called strongly observable if for every vertex $i \in [k]$ at least one of the following holds: (a) $a \in \mathcal{N}(b)$ for all $b \neq a$ or (b) $a \in \mathcal{N}(a)$.

Alon et al. [3] prove the minimax regret for bandits with graph feedback is $\tilde{\Theta}(n^{1/2})$ if and only if $k > 1$ and $\mathcal{G}$ is strongly observable. They also prove the following theorem upper and lower bounding the dependence of the minimax regret on the horizon, the number of actions and a graph functional called the independence number.

Theorem 13 ([3])**.** *Let* $\mathcal{G}_{ind}$ *be the independence number of* $\mathcal{G}$*, which is the cardinality of the largest subset of vertices such that no tow distinct vertices are connected by an edge. Suppose* $k > 1$ *and* $\mathcal{G}$ *is strongly observable. Then* $\mathfrak{R}_n^* = \mathcal{O}(\sqrt{\mathcal{G}_{ind} n}\log(kn))$ *and* $\mathfrak{R}_n^* = \Omega(\sqrt{\mathcal{G}_{ind} n})$.

The logarithmic dependence on n in the proof of Theorem 13 appears quite naturally, which raises the question of whether or not the upper or lower bound is tight. In fact, as n tends to infinity the upper bound in Theorem 13 could be improved to $\mathcal{O}(\sqrt{nk})$ by using a finite-armed algorithm that ignores the feedback except for the played action. Perhaps the independence number is not as fundamental as first thought? The following theorem shows the upper bound can be improved.

Theorem 14. *Let* $\mathscr{A} = (P, E, F)$ *be a triple defining OSMD with* $P_x(a) = \langle a, x\rangle$,

$$F(x) = \frac{1}{\alpha(1-\alpha)}\sum_{i=1}^k x_i^\alpha \qquad \text{where} \quad \alpha = 1 - 1/\log(k)\,.$$

Finally, define the unbiased loss estimation function E by

$$E_t(x,a)_i = \frac{\ell_{ti}\mathbb{1}(a \in \mathcal{N}(i))}{\sum_{b \in \mathcal{N}(i)} x_b} \text{ for } i \notin I_t, \text{ and } E_t(x,a)_i = \frac{(\ell_{ti}-1)\mathbb{1}(a \neq i)}{1-x_i} + 1 \text{ otherwise},$$

where $I_t = \{i \in [k] : i \notin \mathcal{N}(i) \text{ and } X_{ti} > 1/2\}$. Then for any $k \geq 8$ and an appropriately tuned learning rate the regret of OSMD with $\mathscr{A}$ satisfies $\mathfrak{R}_n = \mathcal{O}(\sqrt{\mathcal{G}_{ind} n \log(k)^3})$.

5 Online linear optimisation over ℓ_p-balls

p	Regret	Algorithm
$p = 1$	$\sqrt{n \log(d)}$	Hedge
$p > 1$	$\sqrt{n/(p-1)}$	[12, §11.5]
$p \geq 1$	$\sqrt{d^{2/p-1} n}$	OGD [18]

Table 1: Known results for ℓ_p-balls

We now consider full information online linear optimization on the ℓ_p balls with $p \in [1,2]$, which is modelled in our framework by choosing $\mathcal{A} = B_p^d$ and $\mathcal{L} = B_q^d$ with $1/p + 1/q = 1$ and $\Phi(a,\ell) = \ell$. Table 1 summarises the known results. When $p = 1$ the situation is unambiguous, with matching upper and lower bounds. For $p \in (1,2]$ there exist algorithms for which the regret is dimension free, but with constants that become arbitrarily large as p tends to 1. Known results for online gradient descent (OGD) prove the blowup in terms of p is avoidable, but with a price that is polynomial in the dimension.

Theorem 15. *For any $p \in [1,2]$, let h be the following convex and twice continuously differentiable function:*

$$h(x) = \begin{cases} \frac{d}{2}x^2 & \text{if } |x| \leq d^{\frac{1}{p-2}} \\ \frac{p-2}{p-1} d^{\frac{p-1}{p-2}} |x| + \frac{|x|^p}{p(p-1)} + \frac{2-p}{2p} d^{\frac{p}{p-2}} & \text{otherwise}. \end{cases}$$

Then for OSMD using potential $F(x) = \sum_{i=1}^d h(x_i)$, loss estimator $E(x,a,\sigma) = \sigma$, an arbitrary exploration scheme and appropriately tuned learning rate,

$$\mathfrak{R}_n = \mathcal{O}\left(\sqrt{\min\{1/(p-1), \log(d)\}\, n}\right).$$

Furthermore, the Bayesian regret of TS is bounded by the same quantity.

Remark 16. In the full information setting the loss estimation is independent of the action, which explains the arbitrariness of the exploration scheme. The intuitive justification for the slightly cryptic potential function is provided in the appendix.

6 Discussion

We demonstrated a connection between the information-theoretic analysis and OSMD. For k-armed bandits, we explained the factor of two difference between the regret analysis using information-theoretic and convex-analytic machinery and improved the bound for the latter. For graph bandits we improved the regret by a factor of $\log(n)$. Finally, we designed a new potential for which the regret for online linear optimisation over the ℓ_p-balls improves the previously best known bound by arbitrarily large constant factors.

Open problems The main open problem is whether or not we can 'close the circle' and use the information-theoretic analysis to directly construct OSMD algorithms. Another direction is to try and relax the assumption that the loss is linear. The leading constant in the new bandit analysis now matches the best known information-theoretic bound [21]. There is still a constant lower-order term, which presently seems challenging to eliminate. In bandits with graph feedback one can ask whether the $\log(k)$ dependency can be improved. Lower bounds are still needed for ℓ_p-balls and extending the idea to the bandit setting is an obvious followup. Finally, the best known algorithms for finite partial monitoring also use the information-theoretic machinery. Understanding how to borrow the ideas for OSMD remains a challenge.

References

[1] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[2] J. D. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 263–274. Omnipress, 2008.

[3] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35, Paris, France, 03–06 Jul 2015. PMLR.

[4] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6): 1785–1826, 2017.

[5] A. Antos, G. Bartók, D. Pál, and Cs. Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.

[6] 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.

[7] G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and Cs. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4): 967–997, 2014.

[8] 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.

[9] S. Bubeck and M. Sellke. First-order regret analysis of Thompson sampling. *arXiv preprint arXiv:1902.00681*, 2019.

[10] S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. 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 266–278, Paris, France, 03–06 Jul 2015. PMLR.

[11] 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.

[12] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[13] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31:562–580, 2006.

[14] A. Cohen, T. Hazan, and T. Koren. Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*, pages 811–819, 2016.

[15] S. Dong and B. Van Roy. An information-theoretic analysis for thompson sampling with many actions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4157–4165. Curran Associates, Inc., 2018.

[16] D. Foster and A. Rakhlin. No internal regret via neighborhood watch. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 382–390, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[17] N. Gravin, Y. Peres, and B. Sivan. Towards optimal algorithms for prediction with expert advice. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 528–547. SIAM, 2016.

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

[19] T. Lattimore and Cs. Szepesvári. Cleaning up the neighbourhood: A full classification for adversarial partial monitoring. In *International Conference on Algorithmic Learning Theory*, 2019.

[20] T. Lattimore and Cs. Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

[21] T. Lattimore and Cs. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, 2019.

[22] A. S. Nemirovsky. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979.

[23] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[24] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, NIPS, pages 1583–1591. Curran Associates, Inc., 2014.

[25] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[26] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471, 2016. ISSN 1532-4435.

[27] A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1): 224–243, 1999.

[28] M. Valko. Bandits on graphs and structures, 2016.

A Theorem 3 of [21]

Theorem. Let $(M_t)_{t=1}^{n+1}$ be an $\mathbb{R}^d$-valued martingale adapted to $(\mathcal{F}_t)_{t=1}^{n+1}$ and $M_t \in \mathcal{X} \subset \mathbb{R}^d$ almost surely for all t. Then let F be a convex function with $\operatorname{diam}_F(\mathcal{X}) < \infty$. Suppose there exist constants $\alpha, \beta \geq 0$ such that $\mathbb{E}_t[\Delta_t] \leq \alpha + \sqrt{\beta \mathbb{E}_t[D_F(M_{t+1}, M_t)]}$ almost surely for all t. Then $\mathfrak{BR}_n \leq \alpha n + \sqrt{n\beta \operatorname{diam}_F(\mathcal{X})}$.

B Proof of Lemma 3

The proof is rather standard. In fact, the first part is [20, Theorem 26.13]. For the second part, fix $x \in \mathcal{X}$ and $a \in \mathcal{A}$ and define

$$\Psi(y) = \eta\langle y, E_t(x,a)\rangle + \mathrm{D}_F(y,x)\,.$$

By the assumption that $g_t(x,a) \in \operatorname{int}(\operatorname{dom}(F)) = \operatorname{int}(\operatorname{dom}(\Psi))$ and the definition of $g_t(x,a)$ as the minimizer of Ψ it follows that

$$0 = \nabla\Psi(g_t(x,a)) = \eta E_t(x,a) + \nabla F(g_t(x,a)) - \nabla F(x)\,.$$

Hence

$$\begin{aligned}
\operatorname{stab}_t(x) &= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\langle x - f_t(x,A), E_t(x,A)\rangle - \frac{\mathrm{D}_F(f_t(x,A),x)}{\eta}\right] \\
&= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\frac{1}{\eta}\langle x - f_t(x,A), \nabla F(x) - \nabla F(g_t(x,a))\rangle - \frac{\mathrm{D}_F(f_t(x,A),x)}{\eta}\right] \\
&= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\frac{1}{\eta}\mathrm{D}_F(x, g_t(x,A)) - \frac{1}{\eta}\mathrm{D}_F(f_t(x,a), g_t(x,A))\right] \\
&\leq \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\frac{\mathrm{D}_F(x, g_t(x,A))}{\eta}\right]\,. \qquad (10)
\end{aligned}$$

Let F^* be the Legendre dual of F. Since F is Legendre and twice differentiable on $\operatorname{int}(\operatorname{dom}(F))$ it follows from Taylor's theorem and duality that there exists a $z^* \in [\nabla F(x), \nabla F(x) - \eta E_t(x,a)]$

such that

$$\begin{aligned}\mathrm{D}_F(x, g_t(x,a)) &= \mathrm{D}_{F^*}(\nabla F(g_t(x,a)), \nabla F(x)) \\ &= \mathrm{D}_{F^*}(\nabla F(x) - \eta E_t(x,a), \nabla F(x)) \\ &= \frac{\eta^2}{2}\|E_t(x,a)\|^2_{\nabla^2 F^*(z^*)} \\ &= \frac{\eta^2}{2}\|E_t(x,a)\|^2_{\nabla^{-2} F(\nabla F^*(z^*))} \\ &\le \sup_{z \in [x, g_t(x,a)]} \frac{\eta^2}{2}\|E_t(x,a)\|^2_{\nabla^{-2}F(z)}\,.\end{aligned}$$

Substituting into Eq. (10) completes the result.

Refined bound for the probability simplex For the proofs in the next sections, we require a refined version of Lemma 3. Let 1_k denote the vector with all ones.

Lemma 17. *Assume that $\mathcal{A} = \{e_1, \dots, e_k\}$ and for $c \in \mathbb{R}$ define*

$$\begin{aligned}f_{tc}(x,a) &= \operatorname*{arg\,min}_{y \in \mathcal{X}} \left(\eta\langle y, E_t(x,a) + c1_k\rangle + \mathrm{D}_F(y,x)\right)\,, \\ g_{tc}(x,a) &= \operatorname*{arg\,min}_{y \in \mathrm{int}(\mathrm{dom}(F))} \left(\eta\langle y, E_t(x,a) + c1_k\rangle + \mathrm{D}_F(y,x)\right)\,.\end{aligned}$$

Provided that $g_{tc}(x,a)$ exists for all a in the support of P_x,

$$\mathrm{stab}_t(x;\eta) \le \frac{2}{\eta^2}\mathbb{E}_{A\sim P_x}\left[\mathrm{D}_F(x, g_{tc}(x,A))\right] \le \mathbb{E}_{A\sim P_x}\left[\sup_{z\in[x,g_{tc}(x,A)]}\|E_t(x,A) + c1_k\|^2_{\nabla^{-2}F(z)}\right]\,.$$

Proof. Since $\mathcal{X}$ is the probability simplex $\langle y, c1_k\rangle = c$ for all $y \in \mathcal{X}$. Therefore $f_{tc}(x,a) = f_t(x,a)$ and $\langle x - f_t(x,a), c1_k\rangle = 0$. Hence

$$\begin{aligned}\mathrm{stab}_t(x) &= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\langle x - f_t(x,A), E_t(x,A)\rangle - \frac{\mathrm{D}_F(f_t(x,A),x)}{\eta}\right] \\ &= \frac{2}{\eta}\mathbb{E}_{A\sim P_x}\left[\langle x - f_{tc}(x,A), E_t(x,A) + c1_k\rangle - \frac{\mathrm{D}_F(f_{tc}(x,A),x)}{\eta}\right]\,.\end{aligned}$$

The remaining proof is analogous to the proof of Lemma 3 substituting f_t, g_t by f_{tc}, g_{tc} and the loss $E_t(x,a)$ by $E_t(x,a) + c1_k$. □

C Proof of Corollary 6

Starting with the adversarial regret bound. By Theorem 2,

$$\mathfrak{R}_n \le \frac{\mathrm{diam}_F(\mathcal{X})}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \mathrm{stab}_t(X_t)\right] \le \frac{\mathrm{diam}_F(\mathcal{X})}{\eta} + \frac{\eta n\,\mathrm{stab}(\mathscr{A})}{2}\,.$$

The first part follows by choosing

$$\eta = \sqrt{\frac{2\,\mathrm{diam}_F(\mathcal{X})}{n\,\mathrm{stab}(\mathscr{A})}}\,.$$

The Bayesian case follows from an identical argument and Theorem 4 and the fact that

$$\mathbb{E}\left[\sum_{t=1}^n \mathrm{stab}_t(X_t)\right] \le \mathbb{E}\left[\sum_{t=1}^n \mathrm{stab}(\mathscr{A})\right] \le n\,\mathrm{ess\,sup}(\mathrm{stab}(\mathscr{A}))\,.$$

The result claimed in Remark 7 follows similarly with the same choice of learning rate.

D Proof of Theorem 11

Proof of Lemma 10. We use Lemma 17 with $c = -\frac{1}{2}$. As a reminder, we have

$$E_t(x, a)_i + c = \frac{(\ell_{ti} - c_{ti})\mathbb{1}(a = e_i)}{x_i} + c_{ti} + c\,,\ \text{where } c_{ti} = \frac{1}{2}(1 - \mathbb{1}(X_{ti} < \eta^2).$$

Let $\tilde{\ell}_t = E_t(X_t, A_t) + c1_k$. We start by calculating the Hessian of F. Since $F(a) = -\sum_{i=1}^k 2\sqrt{a_i}$,

$$\nabla F(a) = -1/\sqrt{a} \qquad \text{and} \qquad \nabla^2 F(a) = \operatorname{diag}(a^{-3/2}/2)\,.$$

The next step is to bound $g_{tc}(X_t, A_t)_i^{\frac{3}{2}}$. By definition

$$g_{tc}(X_t, A_t) = \underset{y \in \operatorname{int}(\operatorname{dom}(F))}{\arg\min}\ \eta\langle y, \tilde{\ell}_t\rangle + F(y) - F(X_t) - \langle y - X_t, \nabla F(X_t)\rangle\,,$$

which implies that $\eta\tilde{\ell}_t + \nabla F(g_{tc}(X_t, A_t)) - \nabla F(X_t) = 0$. Substituting the gradient of the potential shows that

$$\eta\tilde{\ell}_{ti} - \frac{1}{\sqrt{g_{tc}(X_t, A_t)_i}} + \frac{1}{\sqrt{X_{ti}}} = 0\,.$$

Solving for $g_{tc}(X_t, A_t)_i$ yields

$$g_{tc}(X_t, A_t)_i^{\frac{3}{2}} = \frac{X_{ti}^{\frac{3}{2}}}{(1 + \tilde{\ell}_t \eta X_{ti}^{\frac{1}{2}})^3}\,. \tag{11}$$

For $\tilde{\ell}_{ti} \geq 0$, Eq. (11) directly implies $g_{tc}(X_t, A_t)_i^{\frac{3}{2}} \leq X_{ti}^{\frac{3}{2}}$. Let $\tilde{\ell}_{ti} < 0$, then we get the following lower bound by definition of $\tilde{\ell}_t$:

$$X_{ti} \geq \eta^2 :\ \tilde{\ell}_{ti} = -\frac{(\ell_{ti} - 1)\mathbb{1}(A_t = e_i)}{2X_{ti}} \geq -\frac{1}{2X_{ti}} \geq -\frac{1}{2\eta X_{ti}^{1/2}},$$

$$X_{ti} < \eta^2 :\ \tilde{\ell}_{ti} = \frac{\ell_{ti}\mathbb{1}(A_t = e_i)}{X_{ti}} - \frac{1}{2} \geq -\frac{1}{2\eta X_{ti}^{1/2}} \geq -\frac{1}{2X_{ti}}.$$

This directly implies $-\tilde{\ell}_{ti}\eta X_{ti}^{1/2} \leq \frac{1}{2}\eta X_{ti}^{-1/2}$ and $1 + \tilde{\eta} X_{ti}^{1/2} \geq \frac{1}{2}$. Going back to Eq. (11), the following bound on $f(x) = x^{-3}$ holds due to convexity for all $x > -1$: $f(1+x) \leq f(1) + xf'(1+x)$. Using all three inequalities provides the bound

$$X_{ti}^{\frac{3}{2}}(1 + \tilde{\ell}_{ti}\eta X_{ti}^{\frac{1}{2}})^{-3} \leq X_{ti}^{\frac{3}{2}}\left(1 - 3(1 + \tilde{\ell}_{ti}\eta X_{ti}^{\frac{1}{2}})^{-4}\tilde{\ell}_{ti}\eta X_{ti}^{\frac{1}{2}}\right) \leq X_{ti}^{\frac{3}{2}} + 24\eta X_{ti}\,.$$

Hence for any $z \in [X_t, g_{tc}(X_t, A_t)]$ we have

$$\nabla^{-2}F(z) \preceq \operatorname{diag}(2X_t^{\frac{3}{2}} + 48\eta X_t \circ \mathbb{1}(\tilde{\ell}_t < 0))\,,$$

where $\mathbb{1}(\tilde{\ell}_t > 0)$ is vector of element wise applied indicator function. Finally we are ready to bound the stability:

$$\mathbb{E}_{A \sim P_{X_t}}\left[\sup_{z \in [X_t, g_{tc}(X_t, A)]} \|E_t(X_t, A) + c1_k\|^2_{\nabla^{-2}F(z)}\right]$$

$$\leq \sum_{i : X_{ti} \geq \eta^2} X_{ti}\frac{(\ell_{ti} - \frac{1}{2})^2}{X_{ti}^2}(2X_{ti}^{\frac{3}{2}} + 48\eta X_{ti}) + \sum_{i : X_{ti} < \eta^2} \frac{1}{2^2}(2X_{ti}^{\frac{3}{2}} + 48\eta X_{ti}) + X_{ti}\frac{\ell_{ti}^2}{X_{ti}^2} 2X_{ti}^{\frac{3}{2}} \tag{12}$$

$$\leq \sum_{i : X_{ti} \geq \eta^2} \frac{X_{ti}^{\frac{1}{2}}}{2} + 12\eta + \sum_{i : X_{ti} < \eta^2} \frac{25\eta^3}{2} + 2\eta \leq \frac{\sqrt{k}}{2} + 12\eta k\,. \tag{13}$$

Eq. (12) follows because for $X_{ti} \geq \eta^2$ the term $E_t(X_t, A)_i + c$ is non zero with probability X_{ti}, while for $X_{ti} < \eta^2$, $E_t(X_t, A)_i + c$ is either non positive and bounded by $-\frac{1}{2}$, or it is positive with probability lower or equal to X_{ti}. Eq. (13) uses the condition $X_{ti} \leq \eta$ in the second sum and the upper bound $\eta \leq 1/2$. □

Proof of Theorem 11. Combine Lemma 10 with Theorem 2, Corollary 6, and Remark 7. □

E Proof of Theorem 14

We make use of the following lemma.

Lemma 18 (Alon et al. 3). *Let $p \in \Delta([k])$. Then*

$$\sum_{i=1}^{k} \frac{p_i}{\sum_{j \in \mathcal{N}(i)} p_j} \leq 4\mathcal{G}_{ind} \log\left(\frac{4k}{\mathcal{G}_{ind} \min_i p_i}\right).$$

Proof of Theorem 14. Starting from Corollary 6 we need to bound the diameter and stability.

$$\mathrm{diam}_F(\mathcal{X}) \leq \frac{k^{1-\alpha}}{\alpha(1-\alpha)} = \frac{k^{\frac{1}{\log(k)}} \log(k)}{1 - \frac{1}{\log(k)}} = \frac{e \log(k)}{1 - \frac{1}{\log(k)}} \leq 2e \log(k),$$

where in the last inequality we used the assumption that $k \geq 8 > e^2$. Moving to the stability term. As a reminder we have

$$E_t(X_t, A_t)_i = \frac{\ell_{ti} \mathbb{1}(A_t \in \mathcal{N}(i))}{\sum_{b \in \mathcal{N}(i)} X_{tb}} \text{ for } i \in I_t \text{ and } E_t(X_t, A_t)_i = \frac{(\ell_{ti} - 1)\mathbb{1}(A_t \neq i)}{1 - X_{ti}} + 1 \text{ otherwise}$$

where $I_t = \{i \in [k] : i \notin \mathcal{N}(i) \text{ and } X_{ti} > 1/2\}$. The set I_t is either empty or contains exactly one element, since the action set it the probability simplex. As a slight abuse of notation, I_t denotes either the (possible empty) set or the unique element within. We use Lemma 17 with

$$c = \mathbb{1}(I_t \neq \emptyset) \frac{(1 - \ell_{tI_t})\mathbb{1}(a \in \mathcal{N}(I_t))}{1 - X_{tI_t}} \geq 0.$$

The Hessian of F is $\nabla F^2(x) = \mathrm{diag}(x^{\alpha - 2})$. The non-negativity of $E_t(X_t, A_t) + c1_k$ ensures that $g_t(X_t, A_t)_i \leq X_{ti}$ almost surely and hence by the definition of the potential $\nabla^{-2}F(z) \preceq \nabla^{-2}F(X_t)$ for all $z \in [X_t, g_t(X_t, A_t)]$,

$$\begin{aligned}
&\mathbb{E}_{A \sim P_{X_t}}\left[\sup_{z \in [X_t, g_{tc}(X_t, A)]} \|E_t(X_t, A) + c1_k\|^2_{\nabla^{-2}F(z)}\right] \\
&= \mathbb{E}_{A \sim P_{X_t}}\left[\|E_t(X_t, A) + c1_k\|^2_{\nabla^{-2}F(X_t)}\right] \\
&= \sum_{i \notin I_t} \mathbb{E}_{A \sim P_{X_t}}\left[(E_t(X_t, A)_i + c)^2 \nabla^{-2}F(X_t)_{ii}\right] + \mathbb{1}(I_t \neq \emptyset)\mathbb{E}_{A \sim P_{X_t}}[\nabla^{-2}F(X_t)_{I_t I_t}] \\
&\leq 2\sum_{i \notin I_t} \mathbb{E}_{A \sim P_{X_t}}\left[E_t(X_t, A)_i^2 X_{ti}^{2-\alpha}\right] + 2\mathbb{E}_{A \sim P_{X_t}}[c^2] \sum_{i \notin I_t} X_{ti}^{2-\alpha} + 1.
\end{aligned}$$

We first bound the c term

$$2\mathbb{E}_{A \sim P_{X_t}}[c^2] \sum_{i \notin I_t} X_{ti}^{2-\alpha} = 2\mathbb{1}(I_t \neq \emptyset) \sum_{i \notin I_t} X_{ti} \left(\frac{1 - \ell_{tI_t}}{\sum_{i \notin I_t} X_{ti}}\right)^2 \sum_{i \notin I_t} X_{ti}^{2-\alpha} \leq 2.$$

Then we bound the contribution of arms i with $i \notin \mathcal{N}(i)$ and $i \notin I_t$, which implies $X_{ti} \leq 1/2$

$$2\mathbb{E}_{A \sim P_x}\left[\sum_{i : i \notin \mathcal{N}(i) \cup I_t} E_t(X_t, A)_i^2 X_{ti}^{2-\alpha}\right] = 2 \sum_{i : i \notin \mathcal{N}(i) \cup I_t} \frac{\ell_{ti}^2 X_{ti}^{2-\alpha}}{1 - X_{ti}} \leq 4.$$

Finally we bound the remaining term

$$2\mathbb{E}_{A \sim P_x}\left[\sum_{i : i \in \mathcal{N}(i)} E_t(X_t, A)_i^2 X_{ti}^{2-\alpha}\right] \leq 2 \sum_{i : i \in \mathcal{N}(i)} \frac{\ell_{ti}^2 X_{ti}^{2-\alpha}}{\sum_{j \in \mathcal{N}(i)} X_{tj}} \leq 2 \max_{a \in \Delta([k])} \sum_{i=1}^{k} \frac{a_i^{2-\alpha}}{\sum_{j \in \mathcal{N}(i)} a_j}.$$

We bound the max using Lemma 18:

$$\begin{aligned}\max_{a\in\Delta([k])}\sum_{i=1}^{k}\frac{a_i^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)}a_j} &= \max_{a\in\Delta([k])}\sum_{i:a_i>\exp(-\log(k)^2)}\frac{a_{ti}^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)}a_j}+\sum_{i:a_i\leq\exp(-\log(k)^2)}\frac{a_i^{2-\alpha}}{\sum_{j\in\mathcal{N}(i)}a_j}\\ &\leq 4\mathcal{G}_{ind}\log\left(\frac{4k\exp(\log(k)^2)}{\mathcal{G}_{ind}}\right)+k\exp(-\log(k)^{-1}\log(k)^2)\\ &= 4\mathcal{G}_{ind}\left(\log\left(\frac{4k}{\mathcal{G}_{ind}}\right)+\log(k)^2\right)+1\,,\end{aligned}$$

where in the final inequality we used Lemma 18 on the sub-graph $\{a : X_{ta} > \exp(-\log(k)^2)$ and noted the fact the independence number of a sub-graph of $\mathcal{G}$ cannot be larger than the independence number of $\mathcal{G}$. Combining everything, we have shown that

$$\mathrm{stab}(\mathscr{A})\leq 8\mathcal{G}_{ind}\left(\log\left(\frac{4k}{\mathcal{G}_{ind}}\right)+\log(k)^2\right)+9.$$

The proof is completed by tuning the learning rate according to Corollary 6. $\square$

F Proof of Theorem 15

Remember that the potential is $F(x)=\sum_{i=1}^{d}h(x_i)$ where

$$h(x)=\begin{cases}\frac{d}{2}x^2 & \text{if } |x|\leq d^{\frac{1}{p-2}}\\ \frac{p-2}{p-1}d^{\frac{p-1}{p-2}}|x|+\frac{|x|^p}{p(p-1)}+\frac{2-p}{2p}d^{\frac{p}{p-2}} & \text{otherwise}\,.\end{cases}$$

Before the proof we provide some intuition for this choice of the potential. By the problem setting for $q=\frac{p}{1-p}$, it holds that $\|\ell_t\|_q,\|X_t\|_p\leq 1$. Assuming we have a 'separable' potential $F(x)=\sum_{i=1}^{d}\tilde{h}(x_i)$, we can write the stability term as

$$\|\ell_t\|^2_{\nabla^{-2}F(z)}=\langle\ell_t\circ\ell_t,(\tilde{h}''(z_i)^{-1})_{i=1,\dots,d}\rangle\leq\|\ell_t\circ\ell_t\|_{q'}\|(\tilde{h}''(z_i)^{-1})_{i=1,\dots,d}\|_{p'}\,.$$

Choosing $q'=\frac{q}{2},p'=\frac{q'}{q'-1}=\frac{p}{2-p}$, the first factor is bounded by 1 and setting $\tilde{h}''(z_i)=|z_i|^{p-2}$ ensures the second factor is bounded by 1. Unfortunately, this leads to the potential $\tilde{h}(x)=\frac{1}{p(p-1)}|x|^p$, whose diameter can be arbitrarily large. To prevent the potential from exploding, we clip $h''(x)$ at d, as shown in Fig. 2. Any upper bound on the second derivative will serve the purpose of decreasing the diameter, however the threshold must be chosen such that the stability doesn't suffer too much. The value d happens to be the lowest value that keeps the stability dimension independent.

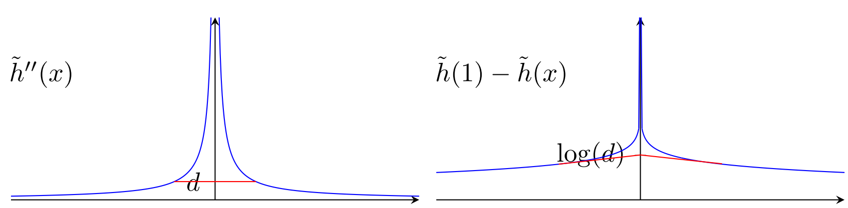

Figure 2: $p=1$: $\tilde{h}''(x)$ and $\tilde{h}(1)-\tilde{h}(x)$ for $p=1$. Red lines indicate h'' and h respectively.

Proof of Theorem 15. By the definition of the loss estimator $\hat{\ell}_t=\ell_t$. As usual, our plan is to bound the stability and diameter and then apply Corollary 6.

Bounding the stability By definition $h''(x) = \min\{|x|^{p-2}, d\}$. Then by Lemma 3 and the assumption that $E_t(x,a) = \ell_t$ for all x and a,

$$\begin{aligned}
\mathrm{stab}_t(x;\eta) &\leq \max_{z\in\mathcal{X}} ||\ell_t||^2_{\nabla F^{-2}(z)} \\
&\leq \max_{z\in\mathcal{X}} \left(\sum_{i:|z_i|\geq d^{\frac{1}{p-2}}} \ell_{ti}^2 |z_i|^{2-p} + \sum_{i:|z_i|<d^{\frac{1}{p-2}}} \frac{1}{d} \right) \\
&\leq \max_{z\in\mathcal{X}} \left(\sum_{i=1}^{d} \ell_{ti}^2 |z_i|^{2-p} + 1 \right) \\
&\leq \max_{z\in\mathcal{X}} \left(\left(\sum_{i=1}^{d} (\ell_{ti}^2)^{\frac{p}{2p-2}} \right)^{\frac{2p-2}{p}} \left(\sum_{i=1}^{d} (|z_i|^{2-p})^{\frac{p}{2-p}} \right)^{\frac{2-p}{p}} + 1 \right) \qquad (14) \\
&= \left(\max_{z\in\mathcal{X}} \|\ell_t\|_q^2 \|z\|_p^{2-p} + 1 \right) \leq 2\,,
\end{aligned}$$

where Eq. (14) follows from Cauchy-Schwarz.

Bounding the diameter First notice that $F(x) \geq 0$ for all $x \in \mathcal{X}$ and $F(0) = 0$. Hence

$$\mathrm{diam}_F(\mathcal{X}) = \max_{x\in\mathcal{X}} F(x)\,.$$

For arbitrary $x \in \mathcal{X}$ define $J = \{i \in [d] | x_i \geq d^{\frac{1}{p-2}}\}$, $I = [d] \setminus J$ and for any $S \subset [d]$ define the vector x_S as the $|S|$-dimensional vector consisting of entries $(x_i)_{i\in S}$. Then it holds

$$F(x) = \frac{d}{2}\|x_I\|_2^2 - \frac{2-p}{p-1} d^{\frac{p-1}{p-2}} \|x_J\|_1 + \frac{\|x_J\|_p^p}{p(p-1)} + \frac{2-p}{2p} d^{\frac{p}{p-2}} |J|.$$

Maximizing this expression over x_J under the constraints of keeping both the set J and $\|x_J\|_p$ constant is setting all but 1 coordinate in x_J to $d^{\frac{1}{p-2}}$ and shifting all other weight towards a single entry. This follows directly from the fact that $\|x\|_p$ is convex, so the minimum of $\|x\|_1$ under constant $\|x\|_p$ is on the boundary. The optimal $y \in \arg\max_{x\in\mathcal{X}} F(x)$ can therefore only have a single coordinate i such that $|y_i| > d^{\frac{1}{p-2}}$, which we assume without loss of generality is $i = 1$.

$$F(y) = h(y_1) + \frac{d}{2}\sum_{i=2}^{d} y_i^2 \leq h(y_1) + \frac{d^2}{2} d^{\frac{2}{p-2}} \leq h(1) + \frac{1}{2}\,.$$

It follows that

$$\begin{aligned}
\mathrm{diam}_F(\mathcal{X}) \leq h(1) + \frac{1}{2} &= \frac{p-2}{p-1} d^{\frac{p-1}{p-2}} + \frac{1}{p(p-1)} + \frac{2-p}{2p} d^{\frac{p}{p-2}} + \frac{1}{2} \\
&= \frac{1 - d^{\frac{p-1}{p-2}}}{p-1} + d^{\frac{p-1}{p-2}} - \frac{1}{p} + \frac{2-p}{2p} d^{\frac{p}{p-2}} + \frac{1}{2} \leq \frac{1 - d^{\frac{p-1}{p-2}}}{p-1} + 1.
\end{aligned}$$

We immediately get the bound $\mathrm{diam}_T(\mathcal{X}) \leq \frac{2}{p-1}$. Let $p \leq \frac{3}{2}$, we substitute $z = \frac{p-1}{2-p}$ and get

$$\mathrm{diam}_F(\mathcal{X}) \leq \frac{1 - d^{-z}}{(2-p)z} + 1 \leq 2\frac{1 - d^{-z}}{z} + 1 \leq 2\log(d) + 1,$$

where we use the fact that for $z \geq 0$ the term $\frac{1-d^{-z}}{z}$ is monotonically decreasing in z with limit $\log(d)$ for $z \to 0$.

We have shown that $\mathrm{diam}_F(\mathcal{X}) \leq \mathcal{O}(\min\{\frac{1}{p-1}, \log(d)\})$ and $\mathrm{stab}(\mathscr{A}) \leq \mathcal{O}(1)$. The proof is completed by tuning the learning rate according to Corollary 6. $\square$