Mirror Descent and the Information Ratio

Tor Lattimore LATTIMORE@DEEPMIND.COM
DeepMind, London
András György AGYORGY@DEEPMIND.COM
DeepMind, London

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We establish a connection between the stability of mirror descent and the information ratio by Russo and Van Roy (2014). Our analysis shows that mirror descent with suitable loss estimators and exploratory distributions enjoys the same bound on the adversarial regret as the bounds on the Bayesian regret for information-directed sampling. Along the way, we develop the theory for information-directed sampling and provide an efficient algorithm for adversarial bandits for which the regret upper bound matches exactly the best known information-theoretic upper bound.

Keywords: Bandits, partial monitoring, mirror descent, information theory.

1. Introduction

The combination of minimax duality and the information-theoretic machinery developed by Russo and Van Roy (2014) has yielded a series of elegant arguments bounding the minimax regret for a variety of regret minimisation problems. The downside is that the application of minimax duality makes the approach non-constructive: the *existence* of certain policies is established without identifying what those policies are. Our main contribution is to show that the information-theoretic machinery can be translated in a natural way to the language of online linear optimisation, yielding explicit policies. Unfortunately, these policies are not guaranteed to be efficient – they must solve a convex optimisation problem that may be infinite dimensional. Nevertheless, our approach provides a clear path towards algorithm design and/or improved bounds, as we illustrate with an application to finite-armed bandits.

To maximise generality, our results are stated using a linear variant of the partial monitoring framework, which is flexible enough to model most classical setups (Lattimore and Szepesvári, 2020b, Chapter 37). Readers who are not familiar with partial monitoring should not be put off. Our analysis does not depend on subtle concepts specific to finite partial monitoring, like the cell decomposition or observability. Examples are given in Table 1.

A linear partial monitoring game is defined by an action space $\mathcal{A} \subset \mathbb{R}^d$, a signal space Σ, a latent space $\mathcal{Z}$ and two functions: a signal function $\Phi : \mathcal{A} \times \mathcal{Z} \to \Sigma$ and a loss function $\ell : \mathcal{Z} \to \mathbb{R}^d$. Both the signal and loss functions are known to the learner. What is special about partial monitoring is that the learner never directly observes the realised losses, instead receiving signals that are correlated with the losses in a way that depends on the loss and signal functions. At the start of the game, an adversary secretly chooses a sequence $(z_t)_{t=1}^n$ with $z_t \in \mathcal{Z}$. A policy is a mapping from action/signal sequences to distributions over actions. The learner interacts with the environment over n rounds. In each round t, the learner uses their policy to find a distribution P_t over the actions based on the history $(A_s)_{s=1}^{t-1}$ and $(\sigma_s)_{s=1}^{t-1}$, where A_s is the action chosen in round s and $\sigma_s = \Phi_{A_s}(z_s)$ is the signal. The learner then samples A_t from P_t, observes the corresponding signal σ_s, and suffers loss $\langle A_t, \ell(z_t)\rangle$.

Regret The regret of a policy π is defined as $\mathfrak{R}_n(\pi, (z_t)) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a, \ell(z_t)\rangle\right]$, where the expectation integrates over the randomness in the actions chosen by the learner (to simplify notation, we write (z_t) instead of $(z_t)_{t=1}^n$). The arguments π and (z_t) are omitted when they are obvious from the context. The quantity of interest is generally the minimax adversarial regret, defined as $\mathfrak{R}_n^\star = \inf_\pi \sup_{(z_t)} \mathfrak{R}_n(\pi, (z_t))$, where the infimum is taken over all policies of the learner and the supremum is over all possible choices of the adversary. Given a finitely supported distribution μ on $\mathcal{Z}^n$, the Bayesian regret of policy π is

$$\mathfrak{BR}_n(\pi, \mu) = \int_{\mathcal{Z}^n} \mathfrak{R}_n(\pi, (z_t)) \, \mathrm{d}\mu((z_t)) \, .$$

A recently popular method for controlling the adversarial regret non-constructively appeals to minimax duality to show that

$$\mathfrak{R}_n^\star = \sup_\mu \inf_\pi \mathfrak{BR}_n(\pi, \mu) \, , \tag{1}$$

where the supremum is over all finitely supported priors (Bubeck et al., 2015; Bubeck and Eldan, 2018; Lattimore, 2020). The Bayesian regret is then bounded uniformly over all priors using the information-theoretic argument of Russo and Van Roy (2014). A limitation of this approach is that the application of minimax duality is non-constructive. It yields a bound on the minimax regret but gives no hint towards an algorithm.

Contributions Our main contribution is to make a connection between the information-theoretic machinery by Russo and Van Roy (2014) and online linear optimisation. Specifically, we show that the stability of mirror descent (MD) and follow the regularised leader (FTRL), the two most popular algorithms for the latter problem, is upper bounded by a function of the information ratio introduced by Russo and Van Roy (2014). The new machinery partially resolves two open problems related to zeroth-order bandit convex optimisation (Bubeck et al., 2017) and hint towards the existence of improved algorithms for this problem. Our results also provide an effortless proof of the main theorem of Lattimore and Szepesvári (2020a) for regret minimisation in partial monitoring. Along the way, we further generalise the information-theoretic machinery to derive adaptive bounds and to make it more suitable for analysing games for which the minimax regret is not $\Theta(n^{1/2})$. A concrete consequence is an efficient algorithm for d-armed adversarial bandits for which $\mathfrak{R}_n \leq \sqrt{2dn}$, improving on the best known result for an efficient algorithm that is $\mathfrak{R}_n \leq \sqrt{2dn} + 48d$ by Zimmert and Lattimore (2019). A modest improvement that nevertheless illustrates the applicability of the approach.

Related work Mirror descent has its origins in classical convex optimisation (Nemirovsky, 1979), while follow the regularised leader goes back to the work of Gordon (1999). As far as we know, the first application to bandits was by Abernethy et al. (2008). The information-theoretic analysis for bandit problems was developed in two influential papers by Russo and Van Roy (2014, 2016). These focussed on the Bayesian setting, with no connections made to the adversarial framework. Bubeck et al. (2015) used minimax duality to argue that the minimax (adversarial) regret is equal to the worst-case Bayesian regret and used this to derive the first proof that the minimax regret for convex bandits in one dimension is $O(\sqrt{n}\log(n))$. The same plan has been used for convex bandits for larger dimensions (Bubeck and Eldan, 2018; Lattimore, 2020) and finite partial monitoring (Lattimore and Szepesvári, 2019), the latter of which establishes Eq. (1) in the present setup. None of

these works yields an efficient algorithm, but these have now been found for both settings (Bubeck et al., 2017; Lattimore and Szepesvári, 2020a), in both cases based on mirror descent. Connections between the information ratio and mirror descent were investigated by Zimmert and Lattimore (2019), who showed that bounds on the stability of mirror descent imply bounds on the information ratio with somewhat restrictive assumptions. These results hinted at a deeper connection, but the analysis is somehow in the wrong direction, since the adversarial regret is a stronger notion than the Bayesian regret. The policy we propose in Section 5 is similar to the exploration by optimisation algorithm suggested by Lattimore and Szepesvári (2020a). The difference is that now the bias of the loss estimators is incorporated into the optimisation problem in a more natural way, without which the connection to the information ratio is not apparent.

2. Notation and conventions

Recall that a proper convex function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is Legendre if it is lower semi-continuous, essentially smooth and essentially strictly convex (Rockafellar, 2015, §26). Throughout, let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a Legendre function and $\mathcal{D} \subset \operatorname{conv}(\mathcal{A})$ be a compact, convex set with non-empty relative interior, where $\operatorname{conv}(\mathcal{A})$ is the convex hull of $\mathcal{A}$. We make the following assumptions:

(a) (finite action set): $1 < |\mathcal{A}| < \infty$.

(b) (bounded losses): $\sup_{a \in \mathcal{A}} \sup_{z \in \mathcal{Z}} |\langle a, \ell(z) \rangle| < \infty$.

(c) (domain of potential): $\mathcal{D} \subset \operatorname{dom}(F) \triangleq \{x \in \mathbb{R}^d : F(x) < \infty\}$.

(d) (bounded potential): $\operatorname{diam}(\mathcal{D}) = \sup_{x,y \in \mathcal{D}} F(x) - F(y) < \infty$.

The restriction to finite action sets avoids delicate measure-theoretic technicalities. Note, since $\mathcal{D}$ is compact, (d) is automatic when F is continuous on $\mathcal{D}$ with the subspace topology, which holds for all potentials considered in the literature that satisfy (c).

Basic notation Precedence is given to the expectation operator: $\mathbb{E}[X]^\alpha$ denotes $(\mathbb{E}[X])^\alpha$ for random variables X and reals α. The relative interior of a subset A of a topological vector space is $\operatorname{relint}(A)$. The standard basis vectors in $\mathbb{R}^d$ are $e_1, \ldots, e_d$. Let $\mathcal{P}$ be the space of probability distributions over $\mathcal{A}$ and $\mathcal{P}_+ = \{p \in \mathcal{P} : p(a) > 0\,, \forall a \in \mathcal{A}\}$ and $\mathcal{P}_\epsilon = \{p \in \mathcal{P} : p(a) \geq \epsilon\,, \forall a \in \mathcal{A}\}$. Occasionally elements $p \in \mathcal{P}$ are identified with vectors in $\mathbb{R}^{|\mathcal{A}|}$. The Fenchel–Legendre dual of F is the convex function defined by $F^\star(u) = \sup_{x \in \mathbb{R}^d} \langle u, x \rangle - F(x)$. Given $p, q \in \operatorname{dom}(F)$ and $x, y \in \operatorname{dom}(F^\star)$, Bregman divergences with respect to F and $F^\star$ are

$$\mathrm{D}(p, q) = F(p) - F(q) - \nabla_{p-q} F(q) \qquad \mathrm{D}_\star(x, y) = F^\star(x) - F^\star(y) - \nabla_{x-y} F^\star(y)\,,$$

where $\nabla_\phi F$ is the directional derivative of F in direction ϕ. The use of directional derivatives is necessary because F and $F^\star$ are not differentiable on the boundary. The assumption that F is Legendre ensures that duality holds so that $(\nabla F)^{-1} = \nabla F^\star$ and whenever $p, q \in \operatorname{int}(\operatorname{dom}(F))$, then

$$\mathrm{D}(p, q) = \mathrm{D}_\star(\nabla F(q), \nabla F(p))\,. \tag{2}$$

Note that if $q \in \text{int}(\text{dom}(F))$, then $p \mapsto \text{D}(p, q)$ is convex, since F is differentiable on its interior by definition. The space of finitely supported probability distributions on $\mathcal{Z} \times \mathcal{D}$ is denoted by $\mathcal{V}$. Finally, let

$$\epsilon_{\mathcal{D}} = \max_{a \in \mathcal{A}} \min_{b \in \mathcal{D}} \max_{z \in \mathcal{Z}} \langle b - a, \ell(z) \rangle \,,$$

which is always non-negative and vanishes in the typical case that $\mathcal{D} = \text{conv}(\mathcal{A})$.

NAME	$\mathcal{A}$	$\mathcal{Z}$	Σ	$\ell(z)$	$\Phi_a(z)$
full information	$\{e_1, \ldots, e_d\}$	$[0,1]^d$	$[0,1]^d$	z	z
d-armed bandits	$\{e_1, \ldots, e_d\}$	$[0,1]^d$	$[0,1]$	z	z_a
semi-bandits	$\subset \{a \in \{0,1\}^d : \|a\|_1 \leq m\}$	$[0,1]^d$	$[0,1]^*$	z	$(z_a : a = 1)$
linear bandits	arbitrary	$\subset \mathbb{R}^d$	$[0,1]$	z	$\langle a, z \rangle$
graph feedback (†)	$\{e_1, \ldots, e_d\}$	$[0,1]^d$	$[0,1]^*$	z	$(z_b : b \in N_a)$
convex bandit (‡)	$\{e_1, \ldots, e_d\}$	$\subset [0,1]^d$	$[0,1]$	z	z_a

† A bandit with graph feedback problem depends on a directed graph over the actions represented by a collection of sets $(N_a)_{a=1}^d$ with N_a the set of edges originating from action a. When playing action a the learner observes the losses for actions $b \in N_a$.

‡ The set of actions for convex bandits is generally a convex set $\mathcal{K} \subset \mathbb{R}^p$. We linearise by finding a net $\{x_1, \ldots, x_d\} \subset \mathcal{K}$, where d is large enough that all actions are well-approximated. Then $\mathcal{Z} = \{(f(x_1), \ldots, f(x_d)) : f \in [0,1]^{\mathcal{K}} \text{ is convex}\}$.

Table 1: Examples

3. A generalised information ratio

The information ratio was introduced by Russo and Van Roy (2014) for the analysis of an algorithm called information-directed sampling, which explicitly optimises the exploration/exploitation dilemma in a Bayesian framework. This beautiful idea led to a number of short proofs bounding the Bayesian regret for a variety of set-ups (Russo and Van Roy, 2014; Bubeck et al., 2015; Russo and Van Roy, 2016; Dong and Van Roy, 2018; Dong et al., 2019; Lattimore and Szepesvári, 2019; Lattimore, 2020). We introduce a generalisation of the concept and explore the properties of information-directed sampling. At the end of the section we outline the differences between the generalised information ratio and the original.

Definition 1 *Recall that $\mathcal{V}$ is the space of finitely supported probability distributions on $\mathcal{Z} \times \mathcal{D}$. A partial monitoring game has a (generalised) information ratio of $(\alpha, \beta, \lambda) \in \mathbb{R}^3$ if for any $\nu \in \mathcal{V}$, there exists a distribution $p \in \mathcal{P}$ such that when $(Z, A^\star, A)$ has law $\nu \otimes p$, then*

$$\mathbb{E}[\langle A - A^\star, \ell(Z) \rangle] \leq \alpha + \beta^{1-1/\lambda} \mathbb{E}[\text{D}(\mathbb{E}[A^\star | \Phi_A(Z), A], \mathbb{E}[A^\star])]^{1/\lambda} \,.$$

The distributions $p \in \mathcal{P}$ realising the display are called exploratory distributions. As the following theorem shows, the Bayesian regret can be bounded in terms of the generalised information ratio.

Theorem 2 *Suppose a partial monitoring game has an information ratio of (α, β, λ) with $\alpha, \beta \geq 0$ and $\lambda \geq 1$. Then, for any finitely supported distribution μ on $\mathcal{Z}^n$, there exists a policy π such that*

$$\mathfrak{BR}_n(\pi, \mu) \leq n(\epsilon_{\mathcal{D}} + \alpha) + (\beta n)^{1-1/\lambda} \, \text{diam}(\mathcal{D})^{1/\lambda} \,.$$

Proof Let $(Z_t)_{t=1}^n$ be the sequence of outcomes sampled from the prior μ and $\mathbb{E}_t[\cdot]$ be the conditional expectation given the observation history $(A_s)_{s=1}^t$, $(\sigma_s)_{s=1}^t$ and abbreviate $\ell_t = \ell(Z_t)$. Let $A^\star = \arg\min_{a \in \mathcal{D}} \sum_{t=1}^n \langle a, \ell_t \rangle$ and $A_t^\star = \mathbb{E}_{t-1}[A^\star]$ be its expectation given the information available at the start of round t. Consider the policy π that samples A_t from any distribution $P_t \in \mathcal{P}$ for which

$$\mathbb{E}_{t-1}[\langle A_t - A^\star, \ell_t \rangle] \leq \alpha + \beta^{1-1/\lambda} \mathbb{E}_{t-1}[\mathrm{D}(A_{t+1}^\star, A_t^\star)]^{1/\lambda}\,, \tag{3}$$

the existence of which is guaranteed by the assumptions of the theorem. Note, that here we have used the fact that A_t and $(Z_t, A^\star)$ are conditionally independent given $(A_s)_{s=1}^{t-1}$ and $(\sigma_s)_{s=1}^{t-1}$. The Bayesian regret of this policy is bounded by

$$\begin{aligned}
\mathfrak{BR}_n(\pi, \mu) &= \mathbb{E}\left[\max_{a \in \mathcal{A}} \sum_{t=1}^n \langle A_t - a, \ell_t \rangle\right] \leq n\epsilon_{\mathcal{D}} + \mathbb{E}\left[\sum_{t=1}^n \langle A_t - A^\star, \ell_t \rangle\right] \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\sum_{t=1}^n \beta^{1-1/\lambda} \mathbb{E}_{t-1}\left[\mathrm{D}(A_{t+1}^\star, A_t^\star)\right]^{1/\lambda}\right] \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + (\beta n)^{1-1/\lambda} \mathbb{E}\left[\sum_{t=1}^n \mathrm{D}(A_{t+1}^\star, A_t^\star)\right]^{1/\lambda} \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + (\beta n)^{1-1/\lambda} \operatorname{diam}(\mathcal{D})^{1/\lambda}\,,
\end{aligned}$$

where the second inequality follows from Eq. (3), the third from Jensen's inequality and the concavity of $x \mapsto x^{1/\lambda}$. The fourth inequality follows by telescoping the Bregman divergences (Lattimore and Szepesvári, 2019, Theorem 3). ■

Remark 3 *The information ratio defined by Russo and Van Roy (2014) assumed that F is the negentropy, $\alpha = 0$ and $\lambda = 2$. Lattimore and Szepesvári (2019) showed that alternative potential functions sometimes lead to improved bounds and introduced the parameter α. What is new in Theorem 1 is that $\lambda \neq 2$ is permitted. In typical applications, $\alpha = 0$, when λ determines the dependence on the horizon and β the leading constant in the regret. Non-zero values of α generally arise as a consequence of discretising infinite-action games, with the level of discretisation chosen in a horizon-dependent manner to ensure that αn is negligible.*

Theorem 2 suggests the following tantalising question:

Open problem 4 *Assume $\mathcal{D} = \operatorname{conv}(\mathcal{A})$ with F being the negentropy potential. Let $\lambda \geq 1$ be the smallest value such that the game has an (α, β, λ) information ratio with $\alpha = 0$ and some β. Is it true that the minimax regret is $\mathfrak{R}_n^\star = \Theta(n^{1-1/\lambda})$?*

The result is known to be true for finite partial monitoring games, where $\mathcal{A}$ and $\mathcal{Z}$ are finite. Note, the result is not true for any potential: the quadratic potential does not lead to $O(n^{1/2})$ regret for finite-armed bandits.

Remark 5 *The assumption that the prior μ is finitely supported is needed because in Theorem 1 we only assumed the existence of a good exploratory distribution for distributions $\nu \in \mathcal{V}$. Those concerned mostly with the Bayesian setting usually define the information ratio for a richer class of distributions than $\mathcal{V}$ and correspondingly Theorem 2 would apply to more priors. The reason for considering finitely supported priors is for the connection to the stability term in Theorem 25, where (a) the coarse $\mathcal{V}$ is sufficient and (b) richer classes cause measure-theoretic challenges.*

Historical note The original information ratio was defined in a more classically Bayesian way. A stochastic partial monitoring game is determined by a probability measure φ on $\mathcal{Z}$ (with some σ-algebra) and then $(z_t)_{t=1}^n$ is sampled from the product measure $\otimes_{t=1}^n \varphi = \varphi^n$. Of course φ is unknown to the learner. The Bayesian perspective on this problem is to take a prior distribution on the unknown φ. Let Θ be some parameter space and $\theta \mapsto \varphi_\theta$ be a probability kernel from Θ to $\mathcal{Z}$. The Bayesian statistician chooses a prior probability measure ξ on Θ and from this one can construct the mixture measure on $\mathcal{Z}^n$ by $\mu = \int \varphi_\theta \, \mathrm{d}\xi(\theta)$. Note that μ is not a product measure anymore, but nevertheless has a special structure that is not enforced in Theorem 2 nor exploited in Definition 1. Russo and Van Roy (2014) only used the standard Shannon entropy as a potential function, which means the expected Bregman divergence is the conditional mutual information between the observation and the optimal action. More formally, they defined the information ratio in round t as

$$\mathrm{IR}_t = \frac{\mathbb{E}_{t-1}\left[\langle A_t - A^*, \ell(Z_t)\rangle\right]^2}{I_{t-1}(A^\star; A_t, \Phi(A_t, Z_t))} \,,$$

where $\mathbb{E}_{t-1} = \mathbb{E}[\cdot | A_1, \sigma_1, \ldots, A_t, \sigma_t]$ and I_t is the mutual information with respect to conditional probability measure $\mathbb{P}(\cdot | A_1, \sigma_1, \ldots, A_t, \sigma_t)$. There is another small difference, which is that $A^\star$ was taken to be the optimal action in expectation with respect to φ_θ where $\theta \sim \xi$ is now included in the probability space. Hence, the regret at the end of the day was the Bayesian pseudo-regret. Bubeck et al. (2015) noticed that the analysis used by Russo and Van Roy (2014) did not rely on the specific construction of the mixture distribution or the definition of the optimal action. By relaxing this assumption they were able to combine the information-theoretic approach to prove minimax bounds for adversarial problems. Our work builds on this by considering the possibility of using alternative potential functions, introducing the λ parameter and making a connection to online convex optimisation.

4. Information-directed sampling

We now explain the name 'information ratio' and explore some properties of the information-directed sampling (IDS) algorithm introduced by Russo and Van Roy (2014) in the context of our generalisation. Let $\nu \in \mathcal{V}$ and define the λ-information ratio $\Psi_{\nu,\lambda} : \mathcal{P} \to \mathbb{R} \cup \{\infty\}$ by

$$\Psi_{\nu,\lambda}(p) = \frac{\max(0, \mathbb{E}[\langle A - A^\star, \ell(Z)\rangle])^\lambda}{\mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star | \Phi_A(Z), A], \mathbb{E}[A^\star])]} \quad \text{with} \quad (Z, A^\star, A) \sim \nu \otimes p$$

and where we adopt the convention that $0/0 = 0$. The minimax λ-information ratio is $\Psi_{\star,\lambda} = \sup_{\nu \in \mathcal{V}} \min_{p \in \mathcal{P}} \Psi_{\nu,\lambda}(p)$. By rearranging the definitions, a game has a generalised information ratio of (α, β, λ) with $\alpha = 0$ if and only if $\Psi_{\star,\lambda} \leq \beta^{\lambda-1}$.

We now introduce the natural generalisation of IDS that minimises the λ-information ratio. More precisely, suppose that $(Z_t)_{t=1}^n$ are sampled from a known finitely supported prior μ on $\mathcal{Z}^n$ and $A^\star = \arg\min_{a \in \mathcal{D}} \sum_{t=1}^n \langle a, \ell(Z_t)\rangle$. The λ-IDS algorithm samples A_t from the distribution $P_t \in \mathcal{P}$ minimising the λ-information ratio:

$$P_t = \operatorname*{arg\,min}_{p \in \mathcal{P}} \Psi_{\nu_t,\lambda}(p) \,,$$

where ν_t is the law of $(Z_t, A^\star)$ under the posterior at round t, which is $\mathbb{P}(\cdot \mid A_1, \sigma_1, \ldots, A_{t-1}, \sigma_{t-1})$. By following the corresponding theorem by Russo and Van Roy (2014), it is easy to show that $p \mapsto \Psi_{\nu,\lambda}(p)$ is convex for $\lambda \geq 2$ and that there exists a minimising p supported on at most two actions.

One thing that is not so nice about the generalisation is that the algorithm now depends on both the potential function *and* λ. Fortunately, if one is prepared to sacrifice a small constant factor in the regret, then optimising $\Psi_{\nu_t,2}$ is a good surrogate for $\Psi_{\nu_t,\lambda}$ for all $\lambda \geq 2$ as the following theorem shows.

Theorem 6 *For any finitely supported prior μ on $\mathcal{Z}^n$ and $\lambda \geq 2$, the Bayesian regret of 2-IDS is bounded by $\mathfrak{BR}_n \leq n\epsilon_{\mathcal{D}} + 2^{1-2/\lambda} n^{1-1/\lambda} (\Psi_{\star,\lambda} \operatorname{diam}(\mathcal{D}))^{1/\lambda}$.*

Proof Let $\nu \in \mathcal{V}$ and $p = \arg\min_{p \in \mathcal{P}} \Psi_{\nu,2}(p)$. By Lemma 23 in the appendix, for any $\lambda \geq 2$,

$$\Psi_{\nu,\lambda}(p) \leq 2^{\lambda-2} \Psi_{\star,\lambda} \,.$$

Combining the above with the argument (and notation) in Theorem 2,

$$\begin{aligned}
\mathfrak{BR}_n &\leq n\epsilon_{\mathcal{D}} + \mathbb{E}\left[\sum_{t=1}^n \langle A_t - A^\star, \ell(Z_t)\rangle\right] \\
&\leq n\epsilon_{\mathcal{D}} + \mathbb{E}\left[\sum_{t=1}^n \Psi_{\nu_t,\lambda}(P_t)^{1/\lambda} \mathbb{E}[\mathrm{D}(A^\star_{t+1}, A^\star_t)]^{1/\lambda}\right] \\
&\leq n\epsilon_{\mathcal{D}} + 2^{1-2/\lambda} n^{1-1/\lambda} \Psi_{\star,\lambda}^{1/\lambda} \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}[\mathrm{D}(A^\star_{t+1}, A^\star_t)]\right]^{1/\lambda} \\
&\leq n\epsilon_{\mathcal{D}} + 2^{1-2/\lambda} n^{1-1/\lambda} (\Psi_{\star,\lambda} \operatorname{diam}(\mathcal{D}))^{1/\lambda} \,.
\end{aligned}$$

■

Remark 7 *Kirschner et al. (2020) showed using an ad-hoc argument that a 'frequentist' version of information-directed sampling with the squared regret in the information ratio works for globally observable games where $\Theta(n^{2/3})$ regret is expected.*

5. Exploration by optimisation

The policy introduced in this section uses MD or FTRL and solves an optimisation problem to find exploratory distributions and loss estimators in a way that essentially minimises the regret bound. A similar algorithm has been seen before with a less clean form and in the context of finite partial monitoring (Lattimore and Szepesvári, 2020a). The modification compared to that algorithm is essential for our main result in the next section.

Optimisation problem Let $\mathcal{G}$ be the space of functions from $\mathcal{A} \times \Sigma$ to $\mathbb{R}^d$. Functions in $\mathcal{G}$ will be used to estimate the losses and are called estimation functions. An estimation function $g \in \mathcal{G}$ is called unbiased if for all $z \in \mathcal{Z}$ and actions $b, c \in \mathcal{A}$,

$$\left\langle b - c, \ell(z) - \sum_{a \in \mathcal{A}} g(a, \Phi_a(z)) \right\rangle = 0 \,.$$

An unbiased loss estimation function $g \in \mathcal{G}$ can be combined with importance-weighting to estimate relative differences in losses. Specifically, given any $p \in \mathcal{P}_+$ and $A \sim p$,

$$\mathbb{E}\left[\left\langle b - c, \underbrace{\frac{g(A, \Phi_A(z))}{p(A)}}_{\text{loss estimate}}\right\rangle\right] = \langle b - c, \ell(z)\rangle \qquad \text{for all } z \in \mathcal{Z}\,.$$

We now define the objective for an optimisation problem that plays a central role in everything that follows. Given $q \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}(F))$ and $\eta > 0$, define a function $\Lambda_{q,\eta} : \mathcal{Z} \times \mathcal{D} \times \mathcal{P}_+ \times \mathcal{G} \to \mathbb{R}$ by

$$\Lambda_{q,\eta}(z, a^\star, p, g) = \sum_{a \in \mathcal{A}} p(a) \langle a - a^\star, \ell(z)\rangle + \left\langle a^\star - q, \sum_{a \in \mathcal{A}} g(a, \Phi_a(z))\right\rangle + \frac{1}{\eta}\sum_{a \in \mathcal{A}} p(a) \mathcal{S}_q\left(\frac{\eta g(a, \Phi_a(z))}{p(a)}\right),$$

where $\mathcal{S}_q(x) = \mathrm{D}_\star(\nabla F(q) - x, \nabla F(q))$, which is the 'stability' term that appears in regret bounds of MD/FTRL (see Theorem 25). Note that $x \mapsto \mathcal{S}_q(x)$ is convex for $q \in \operatorname{dom}(\nabla F)$, and since sums of convex functions are convex and the perspective of a convex function is convex, the function $(p, g) \mapsto \Lambda_{q,\eta}(z, a^\star, p, g)$ is convex. To give a little more intuition for $\Lambda_{q,\eta}$, notice that

$$\Lambda_{q,\eta}(z, a^\star, p, g) = \sum_{a \in \mathcal{A}} p(a) \langle a - q, \ell(z)\rangle + \left\langle a^\star - q, \sum_{a \in \mathcal{A}} g(a, \Phi_a(z)) - \ell(z)\right\rangle + \frac{1}{\eta}\sum_{a \in \mathcal{A}} p(a) \mathcal{S}_q\left(\frac{\eta g(a, \Phi_a(z))}{p(a)}\right).$$

The first term measures the loss due to sampling an action from p with mean $\sum_{a \in \mathcal{A}} p(a)a$ rather than a distribution with mean q as recommended by MD/FTRL. The second term vanishes when g is unbiased and otherwise provides some measure of the bias. The last term measures the stability of the online learning algorithm. Define $\Lambda^\star_{q,\eta}$ and $\Lambda^\star_\eta$ by

$$\Lambda^\star_{q,\eta} = \inf_{\substack{p \in \mathcal{P}_+ \\ g \in \mathcal{G}}} \sup_{\substack{z \in \mathcal{Z} \\ a^\star \in \mathcal{D}}} \Lambda_{q,\eta}(z, a^\star, p, g) \qquad \Lambda^\star_\eta = \sup_{q \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}(F))} \Lambda^\star_{q,\eta}\,. \tag{4}$$

Theorem 8 *The regret of the policy defined by Algorithm 1 (using either MD or FTRL) when run with precision $\epsilon > 0$ and learning rate $\eta > 0$ is bounded by $\mathfrak{R}_n \leq \frac{\operatorname{diam}(\mathcal{D})}{\eta} + n(\epsilon_{\mathcal{D}} + \epsilon + \Lambda^\star_\eta)$.*

Proof Let $\ell_t = \ell(z_t)$ and $a^\star = \arg\min_{a \in \mathcal{D}} \sum_{t=1}^n \langle a, \ell_t\rangle$ be the optimal action in $\mathcal{D}$ in hindsight. Decomposing the regret relative to $a^\star$ and applying the standard regret bound for MD/FTRL in the

INPUT Learning rate η and precision ϵ
INITIALISE $Q_1 = \arg\min_{q \in \mathcal{D}} F(q)$
for $t = 1$ to n:

Find $P_t \in \mathcal{P}_+$ and $G_t \in \mathcal{G}$ such that $\sup_{z \in \mathcal{Z}, a^\star \in \mathcal{D}} \Lambda_{Q_t,\eta}(z, a^\star, P_t, G_t) \leq \Lambda^\star_\eta + \epsilon$

Sample action $A_t \sim P_t$ and observe signal $\sigma_t = \Phi_{A_t}(z_t)$

Compute loss estimate $\hat{\ell}_t = G_t(A_t, \sigma_t)$ and

$$\underbrace{Q_{t+1} = \arg\min_{q \in \mathcal{D}} \langle q, \hat{\ell}_t \rangle + \frac{1}{\eta} \mathrm{D}(q, Q_t)}_{\text{MD}} \qquad \textit{or} \qquad \underbrace{Q_{t+1} = \arg\min_{q \in \mathcal{D}} \sum_{s=1}^{t} \langle q, \hat{\ell}_s \rangle + \frac{F(q)}{\eta}}_{\text{FTRL}}$$

Algorithm 1: Exploration by optimisation

full information case (Theorem 25) yields

$$\begin{aligned}
\mathfrak{R}_n &\leq n\epsilon_{\mathcal{D}} + \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a \in \mathcal{A}} P_t(a) \langle a - a^\star, \ell_t \rangle\right] \\
&= n\epsilon_{\mathcal{D}} + \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a \in \mathcal{A}} P_t(a) \langle a - a^\star, \ell_t \rangle + \langle a^\star - Q_t, \hat{\ell}_t \rangle + \langle Q_t - a^\star, \hat{\ell}_t \rangle\right] \\
&\leq n\epsilon_{\mathcal{D}} + \frac{\operatorname{diam}(\mathcal{D})}{\eta} + \sum_{t=1}^{n} \underbrace{\mathbb{E}\left[\sum_{a \in \mathcal{A}} P_t(a) \langle a - a^\star, \ell_t \rangle + \langle a^\star - Q_t, \hat{\ell}_t \rangle + \frac{1}{\eta} \mathcal{S}_{Q_t}(\eta \hat{\ell}_t)\right]}_{(\mathrm{A})_t} .
\end{aligned}$$

Using the fact that $p \in \mathcal{P}_+$ and the definition of expectation yields

$$\begin{aligned}
\mathbb{E}[(\mathrm{A})_t] &= \mathbb{E}\left[\sum_{a \in \mathcal{A}} P_t(a) \langle a - a^\star, \ell_t \rangle + \left\langle a^\star - Q_t, \sum_{a \in \mathcal{A}} G_t(a, \Phi_a(z_t)) \right\rangle \right. \\
&\qquad \left. + \frac{1}{\eta} \sum_{a \in \mathcal{A}} P_t(a) \mathcal{S}_{Q_t}\left(\frac{\eta G_t(a, \Phi_a(z_t))}{P_t(a)}\right)\right] \\
&= \mathbb{E}\left[\Lambda_{Q_t,\eta}(z_t, a^\star, P_t, G_t)\right] \leq \Lambda^\star_\eta + \epsilon ,
\end{aligned}$$

where the last inequality follows from the definition of P_t and G_t in Algorithm 1. ■

6. Stability and the information ratio

The next theorem makes a connection between the information ratio and the value of the optimisation problems defined in Eq. (4). The result shows that the bound on the Bayesian regret in Theorem 2 holds for Algorithm 1 with the adversarial regret.

Theorem 9 *Suppose a partial monitoring game has an information ratio of* (α, β, λ) *with* $\lambda > 1$. *Then,* $\Lambda^\star_\eta \leq \alpha + \beta\,(1 - 1/\lambda)\,(\eta/\lambda)^{1/(\lambda-1)}$.

Combining Theorem 9 with Theorem 8 and tuning the learning rate immediately yields the following corollary, showing that the regret of Algorithm 1 matches the bound given in Theorem 2.

Corollary 10 *Under the same conditions as Theorem 9, the regret of Algorithm 1 with precision* $\epsilon > 0$ *and* $\eta = \lambda\,(\mathrm{diam}(\mathcal{D})/(\beta n))^{1-1/\lambda}$ *is bounded by* $\mathfrak{R}_n \leq (\epsilon + \epsilon_{\mathcal{D}} + \alpha)n + \mathrm{diam}(\mathcal{D})^{\frac{1}{\lambda}}(\beta n)^{1-\frac{1}{\lambda}}$.

In the language of λ-information ratios, this corresponds to a regret bound of

$$\mathfrak{R}_n \leq (\epsilon + \epsilon_{\mathcal{D}})n + (\Psi_{\star,\lambda}\,\mathrm{diam}(\mathcal{D}))^{\frac{1}{\lambda}} n^{1-\frac{1}{\lambda}}\,,$$

which holds for any $\lambda > 1$.

Proof sketch of Theorem 9 The complete proof is given in Appendix A. Here we outline the main idea, omitting measure-theoretic and topological concerns that make the rigorous proof more involved. Let $q \in \mathcal{D} \cap \mathrm{int}(\mathrm{dom}(F))$ and abbreviate $\Lambda = \Lambda_{q,\eta}$. Assume that

$$\inf_{\substack{p \in \mathcal{P}_+ \\ g \in \mathcal{G}}} \sup_{\substack{a^\star \in \mathcal{D} \\ z \in \mathcal{Z}}} \Lambda(z, a^\star, p, g) = \sup_{\nu \in \mathcal{V}} \inf_{\substack{p \in \mathcal{P}_+ \\ g \in \mathcal{G}}} \int_{\mathcal{Z} \times \mathcal{D}} \Lambda(z, a^\star, p, g)\,\mathrm{d}\nu(z, a^\star)\,. \tag{5}$$

One might expect such a result from minimax theory, since $(p, g) \mapsto \Lambda(z, a^\star, p, g)$ is convex. There is, however, much delicacy. Topologies need to be chosen in such a way that certain compactness and (semi-)continuity conditions are satisfied. At the moment we are not sure whether or not Eq. (5) holds in general. In the rigorous proof we restrict the domains and make a limiting argument, but for simplicity we take Eq. (5) as given for now. The next step is to bound the right-hand side of Eq. (5). Let $\nu \in \mathcal{V}$ and $p \in \mathcal{P}_+$ be arbitrary and $(Z, A^\star, A)$ have law $\nu \otimes p$, which means that

$$\int_{Z \times \mathcal{D}} \Lambda(z, a^\star, p, g)\,\mathrm{d}\nu(z, a^\star) = \mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle + \left\langle A^\star - q, \frac{g(A, \sigma)}{p(A)}\right\rangle + \frac{1}{\eta} S_q\left(\frac{\eta g(A, \sigma)}{p(A)}\right)\right].$$

The estimation function $g \in \mathcal{G}$ that minimises the right-hand side above is found by differentiating. Precisely, given any action $a \in \mathcal{A}$ and signal $\sigma \in \{\Phi_a(z) : z \in \mathcal{Z},\, \nu(\{z\} \times \mathcal{D}) > 0\}$, let

$$g(a, \sigma) = \frac{p(a)}{\eta}\left(\nabla F(q) - \nabla F(\mathbb{E}[A^\star | \Phi_a(Z) = \sigma])\right)\,,$$

and otherwise let $g(a, \sigma) = \mathbf{0}$. Next, let $\sigma = \Phi_A(Z)$ and $A^\star_{\text{pr}} = \mathbb{E}[A^\star]$ and $A^\star_{\text{po}} = \mathbb{E}[A^\star | A, \Phi_A(Z)]$. Then, using the definitions, non-negativity of the Bregman divergences and duality (Eq. (2)),

$$\begin{aligned}
&\mathbb{E}\left[\left\langle A^\star - q, \frac{g(A,\sigma)}{p(A)}\right\rangle + \frac{1}{\eta}\mathcal{S}_q\left(\frac{\eta g(A,\sigma)}{p(A)}\right)\right] \\
&\quad= \frac{1}{\eta}\mathbb{E}\left[\langle A^\star, \nabla F(q) - \nabla F(A^\star_{\text{po}})\rangle + F^\star\left(\nabla F(A^\star_{\text{po}})\right) - F^\star(\nabla F(q))\right] \\
&\quad= -\frac{1}{\eta}\mathbb{E}\left[F^\star(\nabla F(q)) - F^\star(\nabla F(A^\star_{\text{pr}})) - \langle A^\star, \nabla F(q) - \nabla F(A^\star_{\text{pr}})\rangle\right] \\
&\qquad - \frac{1}{\eta}\mathbb{E}\left[F^\star(\nabla F(A^\star_{\text{pr}})) - F^\star(\nabla F(A^\star_{\text{po}})) - \langle A^\star, \nabla F(A^\star_{\text{pr}}) - \nabla F(A^\star_{\text{po}})\rangle\right] \\
&\quad= -\frac{1}{\eta}\mathbb{E}\left[F^\star(\nabla F(q)) - F^\star(\nabla F(A^\star_{\text{pr}})) - \langle A^\star_{\text{pr}}, \nabla F(q) - \nabla F(A^\star_{\text{pr}})\rangle\right] \\
&\qquad - \frac{1}{\eta}\mathbb{E}\left[F^\star(\nabla F(A^\star_{\text{pr}})) - F^\star(\nabla F(A^\star_{\text{po}})) - \langle A^\star_{\text{po}}, \nabla F(A^\star_{\text{pr}}) - \nabla F(A^\star_{\text{po}})\rangle\right] \\
&\quad= -\frac{1}{\eta}\mathbb{E}\left[\mathrm{D}_\star(\nabla F(q), \nabla F(A^\star_{\text{pr}})) + \mathrm{D}_\star(\nabla F(A^\star_{\text{pr}}), \nabla F(A^\star_{\text{po}}))\right] \\
&\quad\le -\frac{1}{\eta}\mathbb{E}\left[\mathrm{D}_\star(\nabla F(A^\star_{\text{pr}}), \nabla F(A^\star_{\text{po}}))\right] \\
&\quad= -\frac{1}{\eta}\mathbb{E}\left[\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})\right] \ , \qquad (6)
\end{aligned}$$

where the first equality follows from the definitions of $\mathcal{S}_q$ and g. Note, g was chosen so as to minimise this expression. The second equality follows by adding and subtracting terms. The third is true by the definition of $A^\star_{\text{pr}}$ and $A^\star_{\text{po}}$ and the fourth is the definition of the Bregman divergence. The inequality is true since Bregman divergences are always non-negative. The final equality follows from duality (Eq. (2)), though careful readers will notice that we not have guaranteed that $\nabla F(A^\star_{\text{pr}})$ and $\nabla F(A^\star_{\text{po}})$ exist, an issue that is handled carefully in the rigorous proof. By the definition of the information ratio there exists a $p \in \mathcal{P}$ such that

$$\mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle\right] \le \alpha + \beta^{1-1/\lambda}\mathbb{E}\left[\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})\right]^{1/\lambda} \ .$$

Combining this with Eq. (6), the definition of Λ and elementary optimisation shows that

$$\begin{aligned}
\mathbb{E}\left[\Lambda(Z, A^\star, p, g)\right] &\le \mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle - \frac{1}{\eta}\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})\right] \\
&\le \alpha + \beta^{1-1/\lambda}\mathbb{E}[\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})]^{1/\lambda} - \frac{1}{\eta}\mathbb{E}[\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})] \\
&\le \alpha + \beta\left(1 - \frac{1}{\lambda}\right)\left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda-1}} \ ,
\end{aligned}$$

which concludes the sketch. ■

7. Computation

Given $q = Q_t \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}(F))$, Algorithm 1 needs to compute $p \in \mathcal{P}_+$ and $g \in \mathcal{G}$ such that

$$\sup_{z \in \mathcal{Z}, a^\star \in \mathcal{D}} \Lambda_{q,\eta}(z, a^\star, p, g) \leq \Lambda^\star_\eta + \epsilon \,.$$

While this is a convex optimisation problem, $\mathcal{G}$ is often infinite-dimensional and the supremum need not have an explicit form. A fundamental case where things work out is finite partial monitoring games ($\mathcal{Z}$ and $\mathcal{A}$ are finite). Then all relevant quantities are finite and standard convex optimisation libraries can be used to implement Algorithm 1 efficiently. Corollary 10 shows that Algorithm 1 has regret bounded by the same quantities as the information-theoretic bounds given by Lattimore and Szepesvári (2019).

8. Discussion

Thompson sampling In general, Theorem 2 only establishes the existence of loss estimators and exploratory distributions for which the stability term is well controlled. There is one fundamental case where something can be said about which exploratory distributions yield a bound on the stability term. Thompson sampling is the policy that always samples A_t from the conditional law of $A^\star$. We show in Theorem 18 that if Thompson sampling witnesses a bound on the information ratio and $\mathcal{A}$ is the standard basis vectors and F is a Tsallis entropy, then mirror descent with $P_t = Q_t$ has well-controlled stability when $\mathcal{D} = \mathcal{P}_{\eta^{4/3}}$ is a simplex without the corners.

Finite-armed bandits The finite-armed adversarial bandit problem is modelled as a linear partial monitoring game by $\mathcal{A} = \{e_1, \ldots, e_d\}$, $\mathcal{Z} = [0,1]^d$, $\Sigma = [0,1]$ and $\ell(z) = z$ and $\Phi_a(z) = z_a$. Audibert et al. (2014) configured mirror descent with the standard importance-weighted loss estimators and the $1/2$-Tsallis entropy potential: $F(q) = -2\sum_{i=1}^d \sqrt{q_i}$. With these choices they were able to show that the resulting algorithm has a regret bounded by $\mathfrak{R}_n \leq \sqrt{8dn}$, which matches the lower bound up to constant factors (Auer et al., 1995). By modifying the loss estimates, Zimmert and Lattimore (2019) improved the bound to $\mathfrak{R}_n \leq \sqrt{2dn} + 48d$.

Meanwhile, on the information-theoretic side, Lattimore and Szepesvári (2019) used entropy inequalities to show that the same potential has an information ratio bounded by $\Psi_{\star,2} \leq \sqrt{d}$. Combining this with the elementary bound on the diameter $\operatorname{diam}(\mathcal{D}) \leq 2\sqrt{d}$ and Corollary 10 shows that Algorithm 1 has $\mathfrak{R}_n \leq n\epsilon + \sqrt{2dn}$ for arbitrarily small ϵ. The fact that Σ is infinite means that the optimisation problem in Algorithm 1 is infinite-dimensional. Nevertheless, armed with the knowledge that certain loss estimation functions exist, the challenge of finding them is less daunting. As a starting point, we guessed that *(a)* the estimation function could be unbiased and *(b)* that mirror descent with $P_t = Q_t$ would suffice. The latter guess is partially supported by our arguments about Thompson sampling above.

Theorem 11 *The regret of Algorithm 1 using the $1/2$-Tsallis entropy and $P_t = Q_t$ and*

$$G_t(a, \sigma)_b = \mathbf{1}_{a=b}\left(\sigma - 1/2 + \frac{\eta}{8}\left(1 + \frac{1}{Q_{t,b} + \sqrt{Q_{t,b}}}\right)\right) - \frac{\eta Q_{t,a}}{8(Q_{t,b} + \sqrt{Q_{t,b}})}$$

is bounded by $\sqrt{2nd}$.

Convex bandits Although we do not yet have an efficient approximation of Algorithm 1 for convex bandits, the analysis here does provide some insights to that problem. Notably, our results combined with the bound on the information ratio by Lattimore (2020) show there exist loss estimation functions and exploratory distributions such that mirror descent has regret at most $\mathfrak{R}_n \leq O(d^{2.5}\sqrt{n}\log(n))$. This closes a question raised by Bubeck et al. (2017) by showing that restarting is not necessary, at least for *some* loss estimators.

Adaptivity Our results naturally extend to data-dependent regret analysis, such as first order bounds. In Appendix B we introduce an adaptive notion of the information ratio and prove analogues of both the information-theoretic Bayesian regret analysis and the duality in Theorem 9.

Infinite action spaces In principle, infinite actions spaces can be handled using the same arguments. But delicate measure-theoretic issues arise in the application of Sion's theorem and some technical assumptions may be necessary. We leave this as a fun challenge for someone with an inclination to technical measure-theoretic details.

Is this *really* mirror descent? Algorithm 1 generally does not play a distribution with exactly the same mean as recommended by mirror descent and the loss estimates are found by solving an optimisation problem that minimises a bound. Do they deserve to be called loss estimates? Or are we manipulating the flexibility of mirror descent to prove a poor man's minimax theorem. A first observation is that the distribution P_t and loss estimates only depend on Q_t, so in this sense at least some aspect of mirror descent is being used. Second, the loss estimation functions that are solutions to the optimisation problem do satisfy certain properties that we expect of sensible estimators. For simplicity, let us take $\mathcal{A} = \{e_1, \dots, e_d\}$ and $\mathcal{D} = \operatorname{conv}(\mathcal{A})$, which is convenient since elements of $\mathcal{D}$ uniquely represent probability distributions over the actions. We will also assume that $\Lambda^\star_\eta = O(\eta)$, which is typical for games with $O(n^{1/2})$ regret, like bandits. Let $q \in \operatorname{int}(\operatorname{dom}(F))$ and $\eta > 0$ be fixed and suppose that (p, g) optimise the saddle-point problem in Eq. (4). Let $\bar{\ell}(z) = \sum_{a \in \mathcal{A}} g(a, \Phi_a(z))$ be the expected value of the importance-weighted loss estimator using estimation function g. There is no hope to argue that $\bar{\ell}(z)$ is close to $\ell(z)$. What is true is that $\bar{\ell}(z)$ is often close to $\ell(z)$ up to a constant shift, which reflects the fact that the learner only needs to estimate relative losses between actions. Let $\Delta(z) = \bar{\ell}(z) - \ell(z) - \langle q, \ell(z) - \bar{\ell}(z)\rangle \mathbf{1}$. If $\Delta(z)$ is small (in some sense), then $\bar{\ell}(z)$ is close to $\ell(z)$ up to shifts. Furthermore,

$$\sup_{z \in \mathcal{Z}} \langle p - q, \ell(z)\rangle + \frac{1}{2}\langle q, |\Delta(z)|\rangle \leq \sup_{z \in \mathcal{Z}} \langle p - q, \ell(z)\rangle + \sup_{a^\star \in \mathcal{D}} \langle q - a^\star, \ell(z) - \bar{\ell}(z)\rangle = O(\eta)\,,$$

where the second relation follows from the positivity of the Bregman divergence and the assumption that $\Lambda^\star_{q,\eta} = O(\eta)$. Hence, if (p, g) minimises the saddle-point problem, then the expected bias of the loss estimators relative to the distribution of mirror descent cannot be large relative to the regret gain by playing p rather than q. The dropped stability term ensures that the magnitude of the loss estimators is not too extreme. Indeed, when F is the negentropy, then the Taylor series expansion of the stability is the second moment of the loss estimates. Overall, the loss estimation functions do (generally) estimate the real losses approximately up to shifts while maintaining relatively small variance.

References

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

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*, pages 229–243, Berlin, Heidelberg, 2006. Springer-Verlag.

J-Y Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Math. Oper. Res.*, 39(1):31–45, 2014.

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.

S. Bubeck and R. Eldan. Exploratory distributions for convex functions. *Mathematical Statistics and Learning*, 1(1):73–100, 2018.

S. Bubeck and M. Sellke. First-order bayesian regret analysis of thompson sampling. In *Algorithmic Learning Theory*, pages 196–233, 2020.

S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Conference on Learning Theory*, pages 266–278, Paris, France, 2015. JMLR.org.

S. Bubeck, Y-T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

S. Dong and B. Van Roy. An information-theoretic analysis for Thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2018. Curran Associates Inc.

S. Dong, T. Ma, and B. Van Roy. On the performance of thompson sampling on logistic bandits. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1158–1160, Phoenix, USA, 25–28 Jun 2019. PMLR.

G. J. Gordon. Regret bounds for prediction problems. In *Proceedings of the 12th Conference on Learning Theory*, pages 29–40, 1999.

J. Kirschner, T. Lattimore, and A. Krause. Information directed sampling for linear partial monitoring. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2328–2369. PMLR, 2020.

T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *arXiv preprint arXiv:2006.00475*, 2020.

T. Lattimore and Cs. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the 32nd Conference on Learning Theory*, pages 2111–2139, Phoenix, USA, 2019. PMLR.

T. Lattimore and Cs. Szepesvári. Exploration by optimisation in partial monitoring. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2488–2515, 2020a.

T. Lattimore and Cs. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020b.

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

R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591. Curran Associates, Inc., 2014.

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.

M. Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

C-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, pages 1263–1291. JMLR.org, 06–09 Jul 2018.

J. Zimmert and T. Lattimore. Connections between mirror descent, thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems*, pages 11973–11982. Curran Associates, Inc., 2019.

J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In *AISTATS*, pages 467–475, 2019.

Appendix A. Proof of Theorem 9

We start with a technical lemma lower bounding $\Lambda_{q,\eta}$.

Lemma 12 *Let $q \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}(F))$ and $\eta > 0$. Then there exists a constant $C_{q,\eta}$ such that*

$$\Lambda_{q,\eta}(z, a^\star, p, g) \geq C_{q,\eta}$$

for all $z \in \mathcal{Z}$, $a^\star \in \mathcal{D}$, $p \in \mathcal{P}_+$ and $g \in \mathcal{G}$.

Proof By the Fenchel–Young inequality,

$$\begin{aligned}
&\left\langle a^\star - q, \frac{g(a, \Phi_a(z))}{p(a)} \right\rangle + \frac{1}{\eta} \mathcal{S}_q \left(\frac{\eta g(a, \Phi_a(z))}{p(a)} \right) \\
&= \frac{1}{\eta} \left\langle a^\star, \frac{\eta g(a, \Phi_a(z))}{p(a)} \right\rangle + \frac{1}{\eta} F^\star \left(\nabla F(q) - \frac{\eta g(a, \Phi_a(z))}{p(a)} \right) - \frac{1}{\eta} F^\star(\nabla F(q)) \\
&\geq \frac{\langle a^\star, \nabla F(q) \rangle - F(a^\star) - F^\star(\nabla F(q))}{\eta} \\
&\geq -\frac{\|a^\star\| \|\nabla F(q)\| + F(a^\star) + F^\star(\nabla F(q))}{\eta} \,.
\end{aligned}$$

Hence, using the definition of $\Lambda_{q,\eta}$,

$$\begin{aligned}
\Lambda_{q,\eta}(z, a^\star, p, g) &\geq \sum_{a \in \mathcal{A}} p(a) \langle a - a^\star, \ell(z) \rangle - \frac{\|a^\star\| \|\nabla F(q)\| + F(a^\star) + F^\star(\nabla F(q))}{\eta} \\
&\geq -\frac{\|a^\star\| \|\nabla F(q)\| + F(a^\star) + F^\star(\nabla F(q))}{\eta} - 2 \sup_{a \in \mathcal{A}} \sup_{z \in \mathcal{Z}} |\langle a, \ell(z) \rangle| \,.
\end{aligned}$$

The right-hand side is lower bounded by a constant that depends only on q and η since $a^\star \in \mathcal{D}$, F has finite diameter on $\mathcal{D}$ and the losses are bounded by assumption. ■

Proof of Theorem 9 The core ingredients of the proof are an application of Sion's minimax theorem to exchange the $\inf$ and $\sup$ in the definition of $\Lambda^\star_{q,\eta}$ and an algebraic calculation to introduce the information ratio. The argument is complicated by the fact that ∇F need not exist on $\mathcal{D} \setminus \operatorname{relint}(\mathcal{D})$.

Step 1: Notation and setup Let $q \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}(F))$ and $\eta > 0$ be fixed and abbreviate $\Lambda(\cdot) \equiv \Lambda_{q,\eta}(\cdot)$. Substituting the definition of $\mathcal{S}_q$ gives

$$\mathcal{S}_q(x) = F^\star(\nabla F(q) - x) - F^\star(\nabla F(q)) + \langle q, x \rangle \,. \tag{7}$$

Let $y \in \operatorname{relint}(\mathcal{D})$ be arbitrary, which exists by assumption. For $\epsilon > 0$ define

$$\mathcal{D}_\epsilon = \{(1 - \epsilon)x + \epsilon y : x \in \mathcal{D}\} \subset \operatorname{relint}(\mathcal{D}) \,.$$

For the remainder we assume that ϵ is small enough that $q \in \mathcal{D}_\epsilon$. By convexity of F and the assumption that $\mathcal{D} \subset \operatorname{dom}(F)$ and that F is Legendre,

$$\sup_{x \in \mathcal{D}_\epsilon} \|\nabla F(x)\|_\infty < \infty \,.$$

Let $\mathcal{V}_\epsilon \subset \mathcal{V}$ be the space of finitely supported probability distributions on $\mathcal{Z} \times \mathcal{D}_\epsilon$ and $\mathcal{G}_\epsilon \subset \mathcal{G}$ be the set of estimation functions g with $\max_{a \in \mathcal{A}} \sup_{\sigma \in \Sigma} \|g(a, \sigma)\|_\infty \leq C_\epsilon$ where

$$C_\epsilon = \frac{1}{\eta} \sup_{q' \in \mathcal{D}_\epsilon} \|\nabla F(q) - \nabla F(q')\|_\infty \,.$$

Next, let $\mathcal{X}_\epsilon \subset \mathcal{P}_\epsilon \times \mathcal{G}_\epsilon$ be given by

$$\mathcal{X}_\epsilon = \left\{ (p, g) \in \mathcal{P}_\epsilon \times \mathcal{G}_\epsilon : -\frac{\eta g(a, \sigma)}{p(a)} \in \operatorname{conv}(\nabla F(\mathcal{D}_\epsilon)) - \nabla F(q) \triangleq \mathcal{K}_\epsilon \text{ for all } a \in \mathcal{A}, \sigma \in \Sigma \right\} .$$

Since $\mathcal{D}_\epsilon \subset \operatorname{int}(\operatorname{dom} F)$ is compact and ∇F is continuous on $\operatorname{int}(\operatorname{dom} F)$, it follows that $\nabla F(\mathcal{D}_\epsilon)$ is compact. And since $\nabla F(\mathcal{D}_\epsilon) \subset \mathbb{R}^d$ is finite-dimension, its convex hull is also compact and hence so too is $\mathcal{K}_\epsilon$. Note also that because F is Legendre, $\nabla F(\operatorname{int}(\operatorname{dom} F)) = \operatorname{int}(\operatorname{dom}(F^\star))$, which is convex. Therefore $\operatorname{conv}(\nabla F(\mathcal{D}_\epsilon)) \subset \operatorname{int}(\operatorname{dom}(F^\star))$. Convexity of $\mathcal{X}_\epsilon$ follows from convexity of $\mathcal{K}_\epsilon$ and the fact that $\{(x, y) \in \mathbb{R}^d \times (0, \infty) : x/y \in \mathcal{K}\}$ is convex for convex $\mathcal{K} \subset \mathbb{R}^d$.

Step 2: Exchanging inf and sup We will now use Sion's theorem to exchange the $\inf$ and the $\sup$ in the definition of Λ and show that

$$\inf_{\substack{p \in \mathcal{P} \\ g \in \mathcal{G}}} \sup_{\substack{a^\star \in \mathcal{D} \\ z \in \mathcal{Z}}} \Lambda(z, a^\star, p, g) \leq \liminf_{\epsilon \to 0} \sup_{\nu \in \mathcal{V}_\epsilon} \inf_{(p,g) \in \mathcal{X}_\epsilon} \int_{\mathcal{Z} \times \mathcal{D}} \Lambda(z, a^\star, p, g) \, \mathrm{d}\nu(z, a^\star) \,. \tag{8}$$

The analysis in this step depends on some topological tomfoolery and can be skipped by eager readers. When $(p, g) \in \mathcal{X}_\epsilon$, then using Eq. (7),

$$\sup_{\substack{a^\star \in \mathcal{D}_\epsilon \\ z \in \mathcal{Z}}} \mathcal{S}_q \left(\frac{\eta g(a, \Phi_a(z))}{p(a)} \right) \leq \sup_{u \in \operatorname{conv}(\nabla F(\mathcal{D}_\epsilon))} F^\star(u) - F^\star(\nabla F(q)) + \eta \sup_{v \in \mathcal{K}_\epsilon} \|q\| \|v\| < \infty \,,$$

where the final inequality follows because $\operatorname{conv}(\nabla F(\mathcal{D}_\epsilon))$ is a compact subset of $\operatorname{int}(\operatorname{dom}(F^\star))$ and $\mathcal{K}_\epsilon$ is compact. Hence, using the boundedness of the losses and the definition of Λ,

$$\sup_{z \in \mathcal{Z}, a^\star \in \mathcal{D}_\epsilon} \Lambda(z, a^\star, p, g) < \infty \,. \tag{9}$$

Combining with Lemma 12 shows that $(z, a^\star) \mapsto \Lambda(z, a^\star, p, g)$ is bounded when $(p, g) \in \mathcal{X}_\epsilon$. By choosing the discrete topology on $\mathcal{Z} \times \mathcal{D}_\epsilon$, the mapping $(z, a^\star) \mapsto \Lambda(z, a^\star, p, g)$ is automatically continuous. Let $\mathcal{V}_\epsilon$ have the weak* topology and $(p, g) \in \mathcal{X}_\epsilon$. Then $\nu \mapsto \int_{\mathcal{Z} \times \mathcal{D}_\epsilon} \Lambda(z, a^\star, p, g) \, \mathrm{d}\nu(z, a^\star)$ is continuous by the definition of the weak* topology and Eq. (9). The same mapping is clearly linear. Of course $\mathcal{P}_\epsilon$ has the usual topology as a subset of $\mathbb{R}^{|\mathcal{A}|}$. Let $\mathcal{G}_\epsilon$ have the product topology, which is the initial topology of the collection of maps $(g \mapsto g(a, \sigma))_{a \in \mathcal{A}, \sigma \in \Sigma}$. Continuity of $(p, g) \mapsto \Lambda(z, a^\star, p, g)$ follows from the definition of the product topology and the same mapping is convex via the perspective construction as noted in Section 5. We also claim that $\mathcal{X}_\epsilon$ is compact. By Tychonoff's theorem, $\mathcal{P}_\epsilon \times \mathcal{G}_\epsilon$ is compact. Hence, it suffices to show that $\mathcal{X}_\epsilon \subset \mathcal{P}_\epsilon \times \mathcal{G}_\epsilon$ is closed, or equivalently, that its complement (in $\mathcal{P}_\epsilon \times \mathcal{G}_\epsilon$) is open.

$$\begin{aligned}
\mathcal{X}_\epsilon &= \left\{ (p, g) \in \mathcal{P}_\epsilon \times \mathcal{G}_\epsilon : -\frac{\eta g(a, \sigma)}{p(a)} \in \mathcal{K}_\epsilon \text{ for all } a \in \mathcal{A}, \sigma \in \Sigma \right\} \\
&= \bigcap_{a \in \mathcal{A}, \sigma \in \Sigma} \left\{ (p, g) \in \mathcal{P}_\epsilon \times \mathcal{G}_\epsilon : -\frac{\eta g(a, \sigma)}{p(a)} \in \mathcal{K}_\epsilon \right\} .
\end{aligned}$$

Since the pre-image of a closed set by a continuous function is closed, the right-hand side is an intersection of closed sets. Therefore $\mathcal{X}_\epsilon$ is closed and hence compact. To summarise, we have shown that for $(p, g) \in \mathcal{X}_\epsilon$, $(z, a^\star) \mapsto \Lambda(z, a^\star, p, g)$ is continuous and linear and for $(z, a^\star) \in \mathcal{Z} \times \mathcal{D}_\epsilon$, $(p, g) \mapsto \Lambda(z, a^\star, p, g)$ is continuous, convex and $\mathcal{X}_\epsilon$ is comapct. Since both $\mathcal{V}_\epsilon$ and $\mathcal{X}_\epsilon$ are

convex, by Sion's minimax theorem (Sion, 1958),

$$\inf_{(p,g)\in\mathcal{X}_\epsilon} \sup_{\substack{a^\star\in\mathcal{D}_\epsilon \\ z\in\mathcal{Z}}} \Lambda(z,a^\star,p,g) = \inf_{(p,g)\in\mathcal{X}_\epsilon} \sup_{\nu\in\mathcal{V}_\epsilon} \int_{\mathcal{Z}\times\mathcal{D}_\epsilon} \Lambda(z,a^\star,p,g)\,\mathrm{d}\nu(z,a^\star)$$
$$= \sup_{\nu\in\mathcal{V}_\epsilon} \inf_{(p,g)\in\mathcal{X}_\epsilon} \int_{\mathcal{Z}\times\mathcal{D}_\epsilon} \Lambda(z,a^\star,p,g)\,\mathrm{d}\nu(z,a^\star)\,.$$

Combining this with linearity of the map $a \mapsto \Lambda(z,a,p,g)$ and Lemma 12 shows that

$$\begin{aligned}
\inf_{\substack{p\in\mathcal{P}\\ g\in\mathcal{G}}} \sup_{\substack{a^\star\in\mathcal{D}\\ z\in\mathcal{Z}}} \Lambda(z,a^\star,p,g) &\le \inf_{(p,g)\in\mathcal{X}_\epsilon} \sup_{\substack{a^\star\in\mathcal{D}\\ z\in\mathcal{Z}}} \Lambda(z,a^\star,p,g) \\
&= \inf_{(p,g)\in\mathcal{X}_\epsilon} \sup_{\substack{a^\star\in\mathcal{D}\\ z\in\mathcal{Z}}} \frac{\Lambda(z,\epsilon y + (1-\epsilon)a^\star,p,g) - \epsilon\Lambda(z,y,p,g)}{1-\epsilon} \\
&\le \inf_{(p,g)\in\mathcal{X}_\epsilon} \sup_{\substack{a^\star\in\mathcal{D}_\epsilon\\ z\in\mathcal{Z}}} \frac{\Lambda(z,a^\star,p,g) - \epsilon C_{q,\eta}}{1-\epsilon} \\
&= \sup_{\nu\in\mathcal{V}_\epsilon} \inf_{(p,g)\in\mathcal{X}_\epsilon} \int_{\mathcal{Z}\times\mathcal{D}} \frac{\Lambda(z,a^\star,p,g)}{1-\epsilon}\,\mathrm{d}\nu(z,a^\star) - \frac{\epsilon C_{q,\eta}}{1-\epsilon}
\end{aligned}$$

Taking the limit as ϵ tends to zero establishes Eq. (8).

Step 3: Introducing the information ratio Fix $\epsilon > 0$ and $\nu \in \mathcal{V}_\epsilon$ and $p \in \mathcal{P}_\epsilon$ and let $(Z, A^\star, A)$ have law $\nu \otimes p$, which means that

$$\begin{aligned}
\int_{\mathcal{Z}\times\mathcal{D}} \Lambda(z,a^\star,p,g)\,\mathrm{d}\nu(z,a^\star) &= \mathbb{E}[\Lambda(Z,A^\star,p,g)] \\
&= \mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle + \left\langle A^\star - q, \frac{g(A,\sigma)}{p(A)}\right\rangle + \frac{1}{\eta}\mathcal{S}_q\left(\frac{\eta g(A,\sigma)}{p(A)}\right)\right]\,.
\end{aligned}$$

The first term will be bounded using Theorem 24 and the assumptions on the information ratio. The second term is bounded by explicitly minimising the second term. Given any action $a \in \mathcal{A}$ and signal $\sigma \in \{\Phi_a(z) : z \in \mathcal{Z},\, \nu(\{z\}\times\mathcal{D}_\epsilon) > 0\}$, let

$$g(a,\sigma) = \frac{p(a)}{\eta}\left(\nabla F(q) - \nabla F(\mathbb{E}[A^\star|\Phi_a(Z) = \sigma])\right)\,,$$

and otherwise let $g(a,\sigma) = 0$. Since $A^\star \in \mathcal{D}_\epsilon$, it holds that $\mathbb{E}[A^\star|\Phi_a(Z) = \sigma] \in \mathcal{D}_\epsilon$. Therefore $\max_{a\in\mathcal{A}}, \sup_{\sigma\in\Sigma} \|g(a,\sigma)\|_\infty \le C_\epsilon$, which implies that $g \in \mathcal{G}_\epsilon$ and hence $(p,g) \in \mathcal{X}_\epsilon$. Next, let $\sigma = \Phi_A(Z)$ and $A^\star_{\text{pr}} = \mathbb{E}[A^\star]$ and $A^\star_{\text{po}} = \mathbb{E}[A^\star|A,\Phi_A(Z)]$. Then, exactly as in Eq. (6),

$$\mathbb{E}\left[\left\langle A^\star - q, \frac{g(A,\sigma)}{p(A)}\right\rangle + \frac{1}{\eta}\mathcal{S}_q\left(\frac{\eta g(A,\sigma)}{p(A)}\right)\right] \le -\frac{1}{\eta}\mathbb{E}\left[\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})\right]\,. \tag{10}$$

By Theorem 24, $p \in \mathcal{P}_\epsilon$ can be chosen so that

$$\mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle\right] \le \epsilon + \alpha + \beta^{1-1/\lambda}\mathbb{E}\left[\mathrm{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})\right]^{1/\lambda}\,.$$

Combining this with Eq. (10), the definition of Λ and elementary optimisation shows that

$$\begin{aligned}\mathbb{E}\left[\Lambda(Z, A^\star, p, g)\right] &\leq \mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle - \frac{1}{\eta}\operatorname{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})\right] \\ &\leq |\mathcal{A}|\epsilon + \alpha + \beta^{1-1/\lambda}\mathbb{E}[\operatorname{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})]^{1/\lambda} - \frac{1}{\eta}\mathbb{E}[\operatorname{D}(A^\star_{\text{po}}, A^\star_{\text{pr}})] \\ &\leq |\mathcal{A}|\epsilon + \alpha + \beta\left(1 - \frac{1}{\lambda}\right)\left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda-1}} .\end{aligned}$$

All together we have shown that for any $\epsilon > 0$ and $\nu \in \mathcal{V}_\epsilon$ there exists a $(p, g) \in \mathcal{X}_\epsilon$ such that

$$\int_{\mathcal{Z}\times\mathcal{D}} \Lambda(z, a^\star, p, g)\,\mathrm{d}\nu(z, a^\star) \leq |\mathcal{A}|\epsilon + \alpha + \beta\left(1 - \frac{1}{\lambda}\right)\left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda-1}} .$$

The claim of the theorem now follows from Eq. (8). ■

Appendix B. Adaptivity

Data-dependent analysis of bandit algorithms based on exponential weights or FTRL has a long history (Allenberg et al., 2006, for example). Recently, Bubeck and Sellke (2020) developed a data-dependent version of the information-theoretic analysis that was specified towards proving first-order bounds for combinatorial semi-bandits. Here we generalise this concept by introducing an adaptive generalised information ratio and extending the results of earlier sections by showing the existence of a corresponding FTRL strategy.

Definition 13 *Let $\alpha \in \mathbb{R}$ and $\beta : \mathcal{Z} \times \mathcal{A} \to [0, \infty)$ and $\lambda > 1$. A partial monitoring game has an (α, β, λ) adaptive information ratio if for all $\nu \in \mathcal{V}$ there exists a $p \in \mathcal{P}$ such that when $(Z, A^\star, A)$ has law $\nu \otimes p$, then*

$$\mathbb{E}[\langle A - A^\star, \ell(Z)\rangle] \leq \alpha + \mathbb{E}[\beta(Z, A)]^{1-1/\lambda}\mathbb{E}[\operatorname{D}(\mathbb{E}[A^\star|\Phi_A(Z), A], \mathbb{E}[A^\star])]^{1/\lambda} .$$

The next theorem is a straightforward generalisation of Theorem 2. That theorem is recovered exactly when β is a constant function.

Theorem 14 *Suppose a partial monitoring game has a (α, β, λ) adaptive information ratio, then for any prior $\nu \in \mathcal{V}$, there exists a policy such that*

$$\mathfrak{BR}_n \leq n(\epsilon_{\mathcal{D}} + \alpha) + \operatorname{diam}(\mathcal{D})^{1/\lambda}\mathbb{E}\left[\sum_{t=1}^n \beta(Z_t, A_t)\right]^{1-1/\lambda} ,$$

where $(Z_t)_{t=1}^n$ is sampled from ν.

Proof Using the same notation and argument as in Theorem 2,

$$\begin{aligned}\mathfrak{BR}_n &\leq n(\epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}_{t-1}[\beta(Z_t, A_t)]^{1-1/\lambda}\mathbb{E}_{t-1}[\operatorname{D}(A^\star_{t+1}, A^\star_t)]^{1/\lambda}\right] \\ &\leq n(\epsilon_{\mathcal{D}} + \alpha) + \operatorname{diam}(\mathcal{D})^{1/\lambda}\mathbb{E}\left[\sum_{t=1}^n \beta(Z_t, A_t)\right]^{1-1/\lambda} .\end{aligned}$$

■

The next theorem generalises Theorem 9.

Theorem 15 *Suppose a partial monitoring game has an (α, β, λ) adaptive information ratio and β is bounded. Then, for any $\eta > 0$ and $q \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}(F))$,*

$$\inf_{\substack{p \in \mathcal{P}_+ \\ g \in \mathcal{G}}} \sup_{\substack{a^\star \in \mathcal{D} \\ z \in \mathcal{Z}}} \left[\Lambda_{q,\eta}(z, a^\star, p, g) - \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda - 1}} \sum_{a \in \mathcal{A}} p(a) \beta(z, a) \right] \leq \alpha \,.$$

Proof Given $\nu \in \mathcal{V}$ and $p \in \mathcal{P}_+$, let $\mathbb{E}_{\nu,p}$ be the expectation with respect to measure $\nu \otimes p$ on $\mathcal{Z} \times \mathcal{D} \times \mathcal{A}$ and let

$$\bar{D}_{\nu,p} = \mathbb{E}_{\nu,p}[\mathrm{D}(\mathbb{E}[A^\star | \Phi_A(Z), A], \mathbb{E}[A^\star])] \qquad \bar{\beta}_{\nu,p} = \mathbb{E}_{\nu,p}[\beta(Z, A)] \,.$$

Notice that the term added inside the saddle point problem in the theorem statement is linear in p and bounded by assumption. Hence, the application of minimax theorem in the proof of Theorem 9 goes through in the same manner, which shows that

$$\begin{aligned}
&\inf_{p \in \mathcal{P}_+, g \in \mathcal{G}} \sup_{a^\star \in \mathcal{D}, z \in \mathcal{Z}} \Lambda_{q,\eta}(z, a^\star, p, g) - \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda - 1}} \sum_{a \in \mathcal{A}} p(a) \beta(z, a) \\
&\leq \sup_{\nu \in \mathcal{V}} \inf_{p \in \mathcal{P}_+} \left(\mathbb{E}_{\nu,p}\left[\langle A - A^\star, \ell(Z) \rangle\right] - \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda - 1}} \bar{\beta}_{\nu,p} - \frac{\bar{D}_{\nu,p}}{\eta} \right) \\
&\leq \sup_{\nu \in \mathcal{V}} \left(\alpha + \bar{\beta}_{\nu,p(\nu)}^{1 - 1/\lambda} \bar{D}_{\nu,p(\nu)}^{1/\lambda} - \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda - 1}} \bar{\beta}_{\nu,p(\nu)} - \frac{\bar{D}_{\nu,p(\nu)}}{\eta} \right) \\
&\leq \alpha \,,
\end{aligned}$$

where the last inequality follows from elementary optimisation and $p : \mathcal{V} \to \mathcal{P}_+$ is a mapping guaranteed by the adaptive information ratio for which

$$\mathbb{E}_{\nu,p(\nu)}[\langle A - A^\star, \ell(Z) \rangle] \leq \alpha + \bar{\beta}_{\nu,p(\nu)}^{1 - 1/\lambda} \bar{D}_{\nu,p(\nu)}^{1/\lambda} \,. \qquad \blacksquare$$

Algorithm 1 can be made adaptive by optimising P_t and G_t so that

$$\sup_{a^\star \in \mathcal{D}, z \in \mathcal{Z}} \Lambda_{Q_t,\eta}(z, a^\star, P_t, G_t) - \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda - 1}} \sum_{a \in \mathcal{A}} P_t(a) \beta(z, a) \leq \alpha + \epsilon \,.$$

By repeating the analysis in the proof of Theorem 8, it follows that

$$\mathfrak{R}_n \leq n(\epsilon + \epsilon_{\mathcal{D}} + \alpha) + \frac{\operatorname{diam}(\mathcal{D})}{\eta} + \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta}{\lambda}\right)^{\frac{1}{\lambda - 1}} \mathbb{E}\left[\sum_{t=1}^{n} \beta(z_t, A_t)\right] \,. \tag{11}$$

There are two problems. First, the expectation in the right-hand side depends on the law of the actions of the algorithm, which depend on η. Hence, it is not straightforward to optimise the learning rate. Second, even if $(z, a) \mapsto \beta(z, a)$ can be written as a function of z only, the quantity in the expectation is generally not known to the learner in advance. Both problems are resolved by tuning the learning rate online and making an additional assumption on β.

Online tuning Adaptively tuning the learning rate is possible if $(z, a) \mapsto \beta(z, a)$ can be written as a function of the signal $\Phi_a(z)$ and a. For the remainder of the section we assume this is true and abuse notation by writing $\beta(\sigma, a)$. Let

$$\eta_t = \lambda^{-1/\lambda}(\lambda - 1)^{1-1/\lambda} \left(\frac{\operatorname{diam}(\mathcal{D})}{\beta_0 + \sum_{s=1}^{t-1} \beta(\sigma_s, A_s)} \right)^{1-1/\lambda}, \tag{12}$$

where $\beta_0 = \sup_{\sigma \in \Sigma} \max_{a \in \mathcal{A}} \beta(\sigma, a)$. Consider the policy that chooses $P_t \in \mathcal{P}_+$ and $G_t \in \mathcal{G}$ such that

$$\sup_{\substack{z \in \mathcal{Z} \\ a^\star \in \mathcal{D}}} \Lambda_{\eta_t, Q_t}(z, a^\star, P_t, G_t) - \left(1 - \frac{1}{\lambda}\right) \left(\frac{\eta_t}{\lambda}\right)^{\frac{1}{\lambda - 1}} \sum_{a \in \mathcal{A}} P_t(a) \beta(\Phi_a(z), a) \leq \epsilon + \alpha \,, \tag{13}$$

where η_t is defined in Eq. (12) and with $\hat{\ell}_s = G_s(A_s, \sigma_s)$,

$$Q_t = \operatorname*{arg\,min}_{q \in \mathcal{D}} \sum_{s=1}^{t-1} \langle q, \hat{\ell}_s \rangle + \frac{F(q)}{\eta_t} \,.$$

Remark 16 *Mirror descent can behave badly when the learning rate is non-constant, so only the FTRL version of the algorithm is used here.*

Theorem 17 *The regret of the policy choosing P_t and G_t satisfying Eq. (13) is bounded by*

$$\mathfrak{R}_n \leq n(\epsilon + \epsilon_{\mathcal{D}} + \alpha) + \left(\frac{\lambda}{\lambda - 1}\right)^{1 - \frac{1}{\lambda}} \operatorname{diam}(\mathcal{D})^{\frac{1}{\lambda}} \mathbb{E}\left[\left(\beta_0 + \sum_{t=1}^{n-1} \beta(\sigma_t, A_t) \right)^{1 - \frac{1}{\lambda}} \right] .$$

Proof Repeat the analysis in Theorem 8 to show that

$$\mathfrak{R}_n \leq n(\epsilon + \epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\frac{\operatorname{diam}(\mathcal{D})}{\eta_n} + \left(1 - \frac{1}{\lambda}\right) \sum_{t=1}^{n} \left(\frac{\eta_t}{\lambda}\right)^{1-1/\lambda} \beta(\sigma_t, A_t) \right] .$$

Then combine the definition of η_t with Lemma 22 in the appendix. ■

The order of the expectation and $x \mapsto x^{1-1/\lambda}$ has been reversed in Theorem 17 relative to Theorem 14, which except for the marginally larger leading constant and the presence of β_0 is actually an improvement. A similar improvement is possible in Theorem 14. Let $(\eta_t)_{t=1}^n$ be the sequence of

learning rates as defined in Eq. (12). Then, using the notation in the proof of Theorem 14,

$$\begin{aligned}
\mathfrak{BR}_n &\leq n(\epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_{t-1}[\beta(\sigma_t, A_t)]^{1-1/\lambda} \mathbb{E}_{t-1}[\mathrm{D}(A^\star_{t+1}, A^\star_t)]^{1/\lambda}\right] \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\sum_{t=1}^{n} \frac{\mathbb{E}_{t-1}[\mathrm{D}(A^\star_{t+1}, A^\star_t)]}{\eta_t} + (1-\lambda)\left(\frac{\eta_t}{\lambda}\right)^{\frac{1}{\lambda-1}} \beta(\sigma_t, A_t)\right] \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\sum_{t=1}^{n} \frac{F(A^\star_{t+1}) - F(A^\star_t)}{\eta_t} + (1-\lambda)\left(\frac{\eta_t}{\lambda}\right)^{\frac{1}{\lambda-1}} \beta(\sigma_t, A_t)\right] \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + \mathbb{E}\left[\frac{\mathrm{diam}(\mathcal{D})}{\eta_n} + (1-\lambda)\sum_{t=1}^{n}\left(\frac{\eta_t}{\lambda}\right)^{\frac{1}{\lambda-1}} \beta(\sigma_t, A_t)\right] \\
&\leq n(\epsilon_{\mathcal{D}} + \alpha) + \left(\frac{\lambda}{\lambda-1}\right)^{1-1/\lambda} \mathrm{diam}(\mathcal{D})^{1/\lambda} \mathbb{E}\left[\left(\beta_0 + \sum_{t=1}^{n} \beta(\sigma_t, A_t)\right)^{1-1/\lambda}\right],
\end{aligned}$$

where the second inequality holds for any sequence of positive learning rates by elementary optimisation. The third inequality by Fatou's lemma as in (Lattimore and Szepesvári, 2019, Theorem 3). The fourth inequality by telescoping the weighted potential and the fact that the learning rates is non-increasing. The final inequality follows from the definition of the learning rate and standard bounding.

Application To make things concrete, let us give an application to d-armed bandits (see Table 1). The following argument is due to Bubeck and Sellke (2020). Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be the logarithmic barrier, which is defined on the positive orthant by $F(p) = -\sum_{a=1}^{d} \log(p_a)$ and its associated with Bregman divergence

$$\mathrm{D}(p, q) = -\sum_{a=1}^{d} \log\left(\frac{p_a}{q_a}\right) + \langle 1/q, p - q\rangle .$$

Let $\epsilon \in (0, 1/d)$ and $\mathcal{D} = \mathcal{P}_\epsilon$, for which $\epsilon_{\mathcal{D}} \leq d\epsilon$. A simple calculation shows that $\mathrm{diam}(\mathcal{D}) \leq d\log(1/\epsilon)$. Let $\beta(z, a) = z_a^2 = \Phi_a(z)^2$. The results by Bubeck and Sellke (2020) show that whenever $(Z, A^\star)$ has law $\nu \in \mathcal{V}$, then with A sampled independently from $(Z, A^\star)$ with law $\mathbb{E}[A^\star] \in \mathcal{P}$,

$$\mathbb{E}[\langle A - A^\star, \ell(Z)\rangle] \leq \sqrt{\mathbb{E}[\beta(Z, A)]\mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star|\Phi_A(Z), A], \mathbb{E}[A^\star])]} .$$

Hence, by Theorem 14, the Bayesian regret for any prior can be bounded by

$$\begin{aligned}
\mathfrak{BR}_n &\leq nd\epsilon + \sqrt{d\mathbb{E}\left[\sum_{t=1}^{n} \ell_{A_t}(Z_t)^2\right]\log(1/\epsilon)} \\
&\leq nd\epsilon + \sqrt{d\left(\mathfrak{BR}_n + \mathbb{E}\left[\sum_{t=1}^{n} \ell_{A^\star}(Z_t)\right]\right)\log(1/\epsilon)}
\end{aligned}$$

Solving the quadratic shows that

$$\mathfrak{BR}_n \leq nd\epsilon + d\log(1/\epsilon) + \sqrt{d\left(1 + \mathbb{E}\left[\sum_{t=1}^{n} \ell_{A^\star}(Z_t)\right]\right)\log(1/\epsilon)}\,.$$

Theorem 17 shows that a suitable instantiation of FTRL achieves about the same bound, a result which is already known (Lattimore and Szepesvári, 2020b).

Appendix C. Thompson sampling

Theorem 9 provides a bound on $\Lambda^\star_\eta$ in terms of the information ratio, but does not provide much information about which policy and estimation functions yield the bound. A fundamental case where more information can be extracted is when $\mathcal{A} = \{e_1, \ldots, e_d\}$ and a bound on the information ratio is witnessed by Thompson sampling, as is often the case. The next theorem relies on a class of potential functions that are widely used in finite-armed bandits (Wei and Luo, 2018; Zimmert and Seldin, 2019, for example). Given $s \in \mathbb{R}$, the s-Tsallis entropy is

$$F(p) = \sum_{a=1}^{d} \frac{p_a^s - sp_a - (1-s)}{s(s-1)}\,.$$

The limits as $s \to 1$ and $s \to 0$ correspond to the negentropy and logarithmic barrier, respectively.

Theorem 18 *Suppose that F is the s-Tsallis entropy with $s \in [0,1]$ and $\mathcal{A} = \{e_1, \ldots, e_d\}$ and $\mathcal{D} = \mathcal{P}_{\eta^{4/3}}$. Assume that for any $(Z, A^\star)$ with law $\nu \in \mathcal{V}$ and independent A with law $p = \mathbb{E}[A^\star]$,*

$$\mathbb{E}\left[\left|\mathbb{E}[\langle A, \ell(Z)\rangle] - \mathbb{E}[\langle A^\star, \ell(Z)\rangle]\right|\right] \leq \sqrt{\beta \mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star|\Phi_A(Z), A], \mathbb{E}[A^\star])]}\,,$$

where $\beta \geq 0$ is a constant. Then,

$$\inf_{g \in \mathcal{G}} \sup_{\substack{a^\star \in \mathcal{D} \\ z \in \mathcal{Z}}} \Lambda_{q,\eta}(z, a^\star, p, g) \leq (1 + O(\eta^{2/3}))\frac{\beta\eta}{4}\,,$$

where the Big-O hides a constant depending only on β.

Note, the presence of the absolute values in the conditions of Theorem 18 is slightly stronger than the definition of the information ratio in Theorem 1. As far as we are aware, all known bounds on the information ratio hold for this stronger definition.

Corollary 19 *Under the same assumptions as Theorem 18, There exist estimation functions such that MD/FTRL with $\mathcal{D} = \mathcal{P}_{\eta^{4/3}}$ and $P_t = Q_t$ and $\eta = 2\sqrt{\operatorname{diam}(\mathcal{D})/(n\beta)}$ satisfies $\mathfrak{R}_n = \sqrt{(1 + o(1))\beta n \operatorname{diam}(\mathcal{D})}$.*

Proof Combine Theorems 8 and 18 yields the following corollary and note that $\epsilon_{\mathcal{D}} \leq d\eta^{4/3}$, which contributes negligibly for large n. ∎

Let us start with a simple lemma that, like the theorem, assumes that F is the s-Tsallis entropy for $s \in [0,1]$.

Lemma 20 *Suppose that* $\epsilon \in [-1,1]^d$ *and* $q \in \mathcal{P}$ *and* $r \in [0,1]^d$*, then* $\langle q - r, \epsilon \rangle - \mathrm{D}(r,q) \leq \frac{e}{2}\langle q, \epsilon^2 \rangle$.

Proof It suffices to prove the result when $d = 1$. Let $f_s(p) = -(p^s - sp - (1-s))/(s(s-1))$, which has $f_s''(p) = p^{s-2}$. A tedious calculation shows that the value of r maximising the left-hand side satisfies $r \leq eq$. By Taylor's theorem and the fact that $p \mapsto f_s''(p)$ is decreasing,

$$\epsilon(q-r) - \mathrm{D}(r,q) \leq \frac{\epsilon^2}{2f''(\max(q,r))} = \frac{\epsilon^2}{2}(\max(q,r))^{2-s} \leq \frac{eq\epsilon^2}{2}\,. \qquad \blacksquare$$

Proof of Theorem 18 Let $\epsilon > 0$ be sufficiently small and $(Z, A^\star)$ have law $\nu \in \mathcal{V}_\epsilon$ and $r = \mathbb{E}[A^\star]$. It suffices to show that when A has law q, then

$$\begin{aligned}\inf_{g \in \mathcal{G}_\epsilon} \mathbb{E}[\Lambda(Z, A^\star, p, g)] &= \mathbb{E}\left[\langle A - A^\star, \ell(Z)\rangle - \frac{1}{\eta}\mathrm{D}(\mathbb{E}[A^\star|\Phi_A(Z), A], r) - \frac{1}{\eta}\mathrm{D}(r,q)\right] \\ &\leq (1 + O(\eta^{1/2}))\frac{\eta\beta}{4}\,.\end{aligned}$$

Let $\mathfrak{I}_a = \mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star|\Phi_a(Z)], \mathbb{E}[A^\star])]$ and $\Delta_a = |\mathbb{E}[\ell_a(Z)] - \mathbb{E}[\langle A^\star, \ell(Z)\rangle]|$. Suppose first that $\langle q, \Delta\rangle \leq \langle q, \mathfrak{I}\rangle/\eta$. Then, by the positivity of the Bregman divergence,

$$\mathbb{E}[\Lambda(Z, A^\star, q, g)] \leq \langle q, \Delta\rangle - \frac{\langle q, \mathfrak{I}\rangle}{\eta} - \frac{1}{\eta}\mathrm{D}(r,q) \leq 0\,.$$

On the other hand, if $\langle q, \Delta\rangle > \langle q, \mathfrak{I}\rangle/\eta$, then

$$\begin{aligned}
&\inf_{g \in \mathcal{G}_\epsilon} \mathbb{E}[\Lambda(Z, A^\star, q, g)] \leq \langle q, \Delta\rangle - \frac{\langle q, \mathfrak{I}\rangle}{\eta} - \frac{1}{\eta}\mathrm{D}(r,q) \\
&= \langle r, \Delta\rangle - \frac{\langle q, \mathfrak{I}\rangle}{\eta} + \langle q - r, \Delta\rangle - \frac{1}{\eta}\mathrm{D}(r,q) \\
&\leq \sqrt{\beta\langle r, \mathfrak{I}\rangle} - \frac{\langle q, \mathfrak{I}\rangle}{\eta} + \langle q - r, \Delta\rangle - \frac{1}{\eta}\mathrm{D}(r,q) \\
&\leq \sqrt{\beta\langle q, \mathfrak{I}\rangle} - \frac{\langle q, \mathfrak{I}\rangle}{\eta} + \frac{|\langle r - q, \mathfrak{I}\rangle|\sqrt{\beta}}{2\sqrt{\langle q, \mathfrak{I}\rangle}} + \langle q - r, \Delta\rangle - \frac{1}{\eta}\mathrm{D}(r,q) \\
&\overset{(\star)}{\leq} \sqrt{\beta\langle q, \mathfrak{I}\rangle} - \frac{\langle q, \mathfrak{I}\rangle}{\eta} + \eta\left\langle q, \left(\frac{\mathfrak{I}\sqrt{\beta}}{\sqrt{q, \mathfrak{I}}} + 2\Delta\right)^2\right\rangle - \frac{1}{2\eta}\mathrm{D}(r,q) \\
&\leq \sqrt{\beta\langle q, \mathfrak{I}\rangle} - \frac{\langle q, \mathfrak{I}\rangle}{\eta} + \frac{2\eta\beta}{\langle q, \mathfrak{I}\rangle}\langle q, \mathfrak{I}^2\rangle + 8\eta\langle q, \Delta^2\rangle - \frac{1}{2\eta}\mathrm{D}(r,q) \\
&\leq \sqrt{\beta\langle q, \mathfrak{I}\rangle} - \frac{\langle q, \mathfrak{I}\rangle}{\eta} + \frac{2\beta\langle q, \mathfrak{I}\rangle}{\eta^{1/3}} + 8\eta\langle q, \Delta\rangle - \frac{1}{2\eta}\mathrm{D}(r,q) \\
&\leq \sqrt{\beta\langle q, \mathfrak{I}\rangle} - \left(1 - 2\beta\eta^{2/3} - 8\eta\right)\frac{\langle q, \mathfrak{I}\rangle}{\eta} + 8\eta\left(\langle q, \Delta\rangle - \frac{\langle q, \mathfrak{I}\rangle}{\eta} - \frac{1}{\eta}\mathrm{D}(r,q)\right)\,,
\end{aligned}$$

where the first inequality follows from Eq. (10), the second by assumption and the third since $(x+\delta)^{1/2} \leq x^{1/2} + \frac{1}{2}|\delta|x^{-1/2}$. The fifth inequality follows from the fact that $(x+y)^2 \leq 2x^2 + 2y^2$

and the sixth since $\langle q, \mathfrak{I}^2\rangle \leq \langle q, \mathfrak{I}\rangle^2/\eta^{4/3}$ and $\Delta \leq \mathbf{1}$. The last inequality follows from naive simplification and re-arranging and by taking η suitably small. The inequality marked with a $(\star)$ follows from Theorem 20, which is justified because

$$\frac{\eta \mathfrak{I}_a \sqrt{\beta}}{\sqrt{\langle q, \mathfrak{I}\rangle}} + 2\eta\Delta_a \leq \frac{\eta\sqrt{\beta\langle q, \mathfrak{I}\rangle}}{q_a} + 2\eta\Delta_a \leq \frac{\eta\sqrt{\beta\eta}}{q_a} + 2\eta\Delta_a \leq 1\,,$$

which holds for all sufficiently small η since $q_a \geq \eta^{4/3}$. Rearranging shows that

$$\mathbb{E}[\Lambda(Z, A^\star, q, g)] \leq \frac{1}{1-8\eta}\left(\sqrt{\beta\langle q, \mathfrak{I}\rangle} - \left(1 - 2\beta\eta^{2/3} - 8\eta\right)\frac{\langle q, \mathfrak{I}\rangle}{\eta}\right) = (1 + O(\eta^{2/3}))\frac{\eta\beta}{4}\,. \quad \blacksquare$$

Appendix D. Proof of Theorem 11

Since G_t is unbiased, the regret of mirror descent with $P_t = Q_t$ and this estimation function is bounded by

$$\begin{aligned}
\mathfrak{R}_n &\leq \frac{\operatorname{diam}(\mathcal{D})}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \mathcal{S}_{Q_t}\left(\frac{\eta g(A_t, \Phi_{A_t}(z_t))}{Q_{tA_t}}\right)\right] \\
&\leq \frac{\operatorname{diam}(\mathcal{D})}{\eta} + \frac{n}{\eta}\sup_{\substack{q\in\operatorname{relint}(\mathcal{D})\\ z\in\mathcal{Z}}}\sum_{a=1}^d q_a \mathcal{S}_q\left(\frac{\eta g(a, \Phi_a(z))}{q_a}\right)\,. \qquad (14) \\
&\leq \frac{\operatorname{diam}(\mathcal{D})}{\eta} + \frac{n\eta\sqrt{d}}{4} \\
&= \sqrt{2nd}\,,
\end{aligned}$$

where the final inequality follows by bounding $\operatorname{diam}(\mathcal{D}) \leq 2\sqrt{d}$ and choosing $\eta = \sqrt{8/n}$ and the second inequality follows from the following lemma. Note that when $n \leq 4$, then $\mathfrak{R}_n \leq \sqrt{2dn}$ is immediate. Hence we may assume that $\eta \leq \sqrt{2}$.

Lemma 21 *Suppose that $\eta \leq \sqrt{2}$. Then stability term in the right-hand side of Eq. (14) is bounded by*

$$\frac{1}{\eta}\sup_{\substack{q\in\operatorname{relint}(\mathcal{D})\\ z\in\mathcal{Z}}}\sum_{a=1}^d q_a \mathcal{S}_q\left(\frac{\eta g(a, \Phi_a(z))}{q_a}\right) \leq \frac{\eta\sqrt{d}}{4}\,.$$

Proof Let $z \in \mathcal{Z}$ and $q \in \operatorname{relint}(\mathcal{D})$ be arbitrary. Then,

$$\begin{aligned}
\frac{1}{\eta}\sum_{a=1}^d p(a)\mathcal{S}_q\left(\frac{\eta g(a, \Phi_a(z))}{q_a}\right) &= \eta\sum_{a=1}^d q_a \sum_{b=1}^d \frac{q_b\left(\frac{g(a,\Phi_a(z))_b}{q_a}\right)^2}{\sqrt{\frac{1}{q_b}} + \frac{\eta g(a,\Phi_a(z))_b}{q_a}} \\
&= \eta\sum_{b=1}^d \sqrt{q_b}\underbrace{\left(\sum_{a=1}^d \frac{q_a\sqrt{q_b}\left(\frac{g(a,\Phi_a(z))_b}{q_a}\right)^2}{\sqrt{\frac{1}{q_b}} + \frac{\eta g(a,\Phi_a(z))_b}{q_a}}\right)}_{(\mathrm{A})_b} \leq \frac{\eta}{4}\sum_{b=1}^d \sqrt{q_b} \leq \frac{\eta\sqrt{d}}{4}\,,
\end{aligned}$$

where the first inequality follows from the messy calculation below and the second inequality follows from Cauchy–Schwarz. For the messy calculation:

$$\begin{aligned}
(\mathrm{A})_b &= \sum_{a=1}^{d} \frac{q_a \sqrt{q_b} \left(\frac{g(a,\Phi_a(z))_b}{q_a}\right)^2}{\sqrt{\frac{1}{q_b} + \frac{\eta g(a,\Phi_a(z))_b}{q_a}}} \\
&= \frac{1}{8}\left(\frac{\left(\eta + 4(2z_b - 1)\sqrt{q_b}\right)^2}{\eta^2 + 4\eta(2z_b-1)\sqrt{q_b} + 8q_b} + \frac{\eta^2\left(1-\sqrt{q_b}\right)}{8\left(\sqrt{q_b}+1\right) - \eta^2}\right) \\
&= \frac{1}{8}\left(2 - \frac{\eta^2}{\eta^2 + 4\eta(2z_b-1)\sqrt{q_b} + 8q_b} + \frac{\eta^2\left(1-\sqrt{q_b}\right)}{8\left(\sqrt{q_b}+1\right) - \eta^2}\right) \\
&\leq \frac{1}{8}\left(2 - \frac{\eta^2}{\eta^2 + 4\eta\sqrt{q_b} + 8q_b} + \frac{\eta^2\left(1-\sqrt{q_b}\right)}{8\left(\sqrt{q_b}+1\right) - \eta^2}\right) \\
&\leq \frac{1}{4}\,,
\end{aligned}$$

where the final inequality follows since $\eta \leq \sqrt{2}$. ■

Appendix E. Technical inequalities

Here we collect some technical results.

Lemma 22 *Let $\lambda > 1$ and $(\beta_t)_{t=0}^n$ be a sequence of positive reals with $\beta_0 \geq \beta_t$ for all $1 \leq t \leq n$. Then,*

$$\sum_{t=1}^{n} \beta_t \left(\sum_{s=0}^{t-1} \beta_s\right)^{1/\lambda - 1} \leq \lambda \left(\sum_{t=1}^{n} \beta_t\right)^{1/\lambda}.$$

Proof Let $B(t) = \int_0^t \beta_{\lceil s \rceil} \,\mathrm{d}s$. Then,

$$\sum_{t=1}^{n} \beta_t \left(\sum_{s=0}^{t-1} \beta_s\right)^{1/\lambda - 1} \leq \int_0^n B'(t) B(t)^{1/\lambda - 1} \,\mathrm{d}t = \lambda B(n)^{1/\lambda} = \lambda \left(\sum_{t=1}^{n} \beta_t\right)^{1/\lambda}. \qquad ■$$

Lemma 23 *Let $\nu \in \mathcal{V}$ and $p = \arg\min_{p \in \mathcal{P}} \Psi_{\nu,2}(p)$. Then for all $\lambda \geq 2$, $\Psi_{\nu,\lambda}(p) \leq 2^{\lambda-2} \min_{q \in \mathcal{P}} \Psi_{\nu,\lambda}(q)$.*

Proof Let $(Z, A) \sim \nu$ and define regret and information vectors $\Delta, \mathfrak{I} \in \mathbb{R}^{\mathcal{A}}$ by $\Delta_a = \mathbb{E}[\langle a - A^\star, \ell(Z)\rangle]$ and $\mathfrak{I}_a = \mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star | \Phi_a(Z)], \mathbb{E}[A^\star])]$. The result is immediate if $\langle p, \Delta\rangle \leq 0$, so assume for the remainder that $\langle p, \Delta\rangle > 0$. By the first-order optimality conditions

$$0 \leq \langle \nabla \Psi_{\nu,2}(p), q - p\rangle = \frac{2\langle q - p, \Delta\rangle\langle p, \Delta\rangle}{\langle p, \mathfrak{I}\rangle} - \frac{\langle q - p, \mathfrak{I}\rangle \langle p, \Delta\rangle^2}{\langle p, \mathfrak{I}\rangle^2}\,.$$

Rearranging shows that

$$\langle p, \Delta \rangle \left(1 + \frac{\langle q, \mathfrak{I} \rangle}{\langle p, \mathfrak{I} \rangle}\right) \leq 2\langle q, \Delta \rangle \,. \tag{15}$$

Since the information gain is non-negative, it follows that $\langle p, \Delta \rangle \leq 2\langle q, \Delta \rangle$. Therefore,

$$\Psi_{\nu,\lambda}(p) = \frac{\langle p, \Delta \rangle^\lambda}{\langle p, \mathfrak{I} \rangle} \leq \frac{2^{\lambda-2}\langle p, \Delta \rangle^2 \langle q, \Delta \rangle^{\lambda-2}}{\langle p, \mathfrak{I} \rangle} \leq \frac{2^{\lambda-2}\langle q, \Delta \rangle^2}{\langle q, \mathfrak{I} \rangle} = 2^{\lambda-2} \min_{q \in \mathcal{P}} \Psi_{\nu,\lambda}(q) \,,$$

where the first inequality follows form Eq. (15) and the second since p minimises $\Psi_{\nu,2}$. ■

The next simple lemma is used to show that the exploratory distribution can be chosen to assign non-zero probability to all actions with arbitrarily small loss.

Lemma 24 *Suppose a partial monitoring game has an information ratio of (α, β, λ) with $\lambda \geq 1$. Then for any $\nu \in \mathcal{V}$ and $\epsilon \in (0, 1)$, there exists a $q \in \mathcal{P}_\epsilon$ such that when $(Z, A^\star, A)$ is sampled from the product measure $\nu \otimes q$, then*

$$\mathbb{E}[\langle A - A^\star, \ell(Z) \rangle] \leq |\mathcal{A}|\epsilon + \alpha + \beta^{1-1/\lambda} \mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star | \Phi_A(Z), A], \mathbb{E}[A^\star])]^{1/\lambda} \,.$$

Proof Let $p \in \mathcal{P}$ be the distribution guaranteed by the definition of the information ratio and $q = (1 - \epsilon)p + \epsilon \mathbf{1}$. Then $q \in \mathcal{P}_\epsilon$, and

$$\begin{aligned}
\sum_{a \in \mathcal{A}} q(a) \mathbb{E}[\langle a - A^\star, \ell(Z) \rangle] &= (1 - \epsilon) \sum_{a \in \mathcal{A}} p(a) \mathbb{E}[\langle a - A^\star \rangle, \ell(Z)] + \epsilon \sum_{a \in \mathcal{A}} \mathbb{E}[\langle a - A^\star, \ell(Z) \rangle] \\
&\leq |\mathcal{A}|\epsilon + (1 - \epsilon) \left[\alpha + \beta^{1-1/\lambda} \left(\sum_{a=1}^{k} p(a) \mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star | \Phi_a(Z)], \mathbb{E}[A^\star])] \right)^{1/\lambda} \right] \\
&\leq |\mathcal{A}|\epsilon + \alpha + \beta^{1-1/\lambda} \left(\sum_{a=1}^{k} q(a) \mathbb{E}[\mathrm{D}(\mathbb{E}[A^\star | \Phi_a(Z)], \mathbb{E}[A^\star])] \right)^{1/\lambda} ,
\end{aligned}$$

where in the first inequality we used the assumption that $\langle a, \ell(z) \rangle \in [0, 1]$ for all $a \in \mathcal{A}$ and $z \in \mathcal{Z}$. The second follows by the non-negativity of the Bregman divergence and the fact that $(1 - \epsilon) \leq (1 - \epsilon)^{1/\lambda}$ since $\lambda \geq 1$ and $\epsilon \in (0, 1)$. ■

Appendix F. Mirror descent and FTRL

Given a sequence of loss estimates $(\hat{\ell}_t)_{t=1}^n$ with $\hat{\ell}_t \in \mathbb{R}^d$ and a sequence of non-increasing and strictly positive learning rates $(\eta_t)_{t=1}^n$, MD produces a sequence $(q_t)_{t=1}^n$ with $q_t \in \mathcal{D}$ defined inductively by

$$q_1 = \arg\min_{q \in \mathcal{D}} F(q) \qquad\qquad q_{t+1} = \arg\min_{q \in \mathcal{D}} \langle q, \hat{\ell}_t \rangle + \frac{\mathrm{D}(q, q_t)}{\eta_t} \,. \tag{16}$$

Follow the regularised leader also produces a sequence $(q_t)_{t=1}^n$ with $q_t \in \mathcal{D}$ defined by

$$q_t = \operatorname*{arg\,min}_{q \in \mathcal{D}} \sum_{s=1}^{t-1} \langle q, \hat{\ell}_s \rangle + \frac{F(q)}{\eta_t} \,. \tag{17}$$

The next theorem bounds the regret of MD and FTRL with respect to the estimated losses. There are many sources for results like this, though this exact version is the dual of what is normally seen, which avoids the need for assumptions on the loss estimates.

Theorem 25 *Suppose that one of the following is true:*

(a) $(q_t)_{t=1}^n$ *are chosen according to Eq.* (16) *and* $\eta_t = \eta$ *is constant; or*

(b) $(q_t)_{t=1}^n$ *is chosen according to Eq.* (17)*.*

Then, $\max_{a^\star \in \mathcal{D}} \sum_{t=1}^n \langle q_t - a^\star, \hat{\ell}_t \rangle \leq \dfrac{\operatorname{diam}(\mathcal{D})}{\eta_n} + \sum_{t=1}^n \dfrac{\mathcal{S}_{q_t}(\eta_t \hat{\ell}_t)}{\eta_t}$.

Proof By (Lattimore and Szepesvári, 2020b, Theorem 28.4,Exercise 28.12),

$$\max_{a^\star \in \mathcal{D}} \sum_{t=1}^n \langle q_t - a^\star, \hat{\ell}_t \rangle \leq \frac{\operatorname{diam}(\mathcal{D})}{\eta_n} + \sum_{t=1}^n \langle q_t - q_{t+1}, \hat{\ell}_t \rangle - \frac{1}{\eta_t} \operatorname{D}(q_{t+1}, q_t) \,,$$

The result follows from Lemma 26 below and the fact that $q_t \in \operatorname{int}(\operatorname{dom}(F))$, which holds since F is Legendre. ■

Lemma 26 *Suppose that* $p, q \in \mathcal{D}$ *with* $q \in \operatorname{int}(\operatorname{dom}(F))$*. Then, for any* $\ell \in \mathbb{R}^d$*,*

$$\langle q - p, \ell \rangle - \frac{1}{\eta} \operatorname{D}(p, q) \leq \frac{1}{\eta} \operatorname{D}_\star(\nabla F(q) - \ell, \nabla F(q))$$

Proof If $\nabla F(q) - \ell \in \operatorname{int}(\operatorname{dom}(F^\star))$, then the p maximising the left-hand side is $\nabla F^\star(\nabla F(q) - \eta\ell)$ and the result follows from elementary calculations and duality. On the other hand, the result is immediate if $\nabla F(q) - \ell \notin \operatorname{dom}(F^\star)$, since then the right-hand side is infinite. Suppose for the remainder that $\nabla F(q) - \ell \in \operatorname{dom}(F^\star)$. Since $\nabla F(q) \in \operatorname{int}(\operatorname{dom}(F^\star))$, $\nabla F(q) - (1-\epsilon)\ell \in \operatorname{int}(\operatorname{dom}(F^\star))$ for any $\epsilon > 0$. Therefore,

$$\begin{aligned}
\langle q - p, \ell \rangle - \frac{1}{\eta} \operatorname{D}(p, q) &\leq \epsilon \|p - q\| \|\ell\| + \langle q - p, (1-\epsilon)\ell \rangle - \frac{1}{\eta} \operatorname{D}(p, q) \\
&\leq \epsilon \|p - q\| \|\ell\| + \frac{1}{\eta} \operatorname{D}_\star(\nabla F(q) - (1-\epsilon)\ell, \nabla F(q)) \\
&\leq \epsilon \|p - q\| \|\ell\| + \frac{1-\epsilon}{\eta} \operatorname{D}_\star(\nabla F(q) - \ell, \nabla F(q)) \,.
\end{aligned}$$

Since $\mathcal{D}$ is compact, the result follows by taking the limit as ϵ tends to 0. ■