Minimax Regret for Partial Monitoring: Infinite Outcomes and Rustichini's Regret

Tor Lattimore   LATTIMORE@DEEPMIND.COM
DeepMind, London

Abstract

We show that a version of the generalised information ratio of Lattimore and György (2021) determines the asymptotic minimax regret for all finite-action partial monitoring games provided that *(a)* the standard definition of regret is used but the latent space where the adversary plays is potentially infinite; or *(b)* the regret introduced by Rustichini (1999) is used and the latent space is finite. Our results are complemented by a number of examples. For any $p \in [1/2, 1]$ there exists an infinite partial monitoring game for which the minimax regret over n rounds is n^p up to subpolynomial factors and there exist finite games for which the minimax Rustichini regret is $n^{4/7}$ up to subpolynomial factors.

Keywords: Partial monitoring; bandits; information ratio.

1. Introduction

The minimax asymptotic regret for partial monitoring is only understood for finite games and where the notion of regret compares the learner's accumulated losses to the loss of the best action in hindsight. For many partial monitoring games the learner is at a severe disadvantage compared to the adversary because the signal function may be relatively uninformative. The notion of regret introduced by Rustichini (1999) serves as a better measure of performance for these games, as it measures the learner's performance compared to a baseline that is achievable, while the standard notion of regret may be linear. Our main contribution is twofold. We show that the generalised information ratio of Lattimore and György (2021) determines the asymptotic minimax regret for all finite-action games where either *(a)* the standard regret is used but the latent space where the adversary plays is possibly infinite; or *(b)* the definition of the regret proposed by Rustichini (1999) is used and the latent space is finite. To state our main theorem informally, for $\lambda > 1$ let

$$\Psi_\star^\lambda = \sup_{\text{prior}} \min_{\text{policy}} \frac{[\text{expected regret}]^\lambda}{\text{information gain}}$$

be the minimax information ratio, where the supremum is over priors on the adversary's decision for one round and the infimum is over distributions on actions. The information gain and regret will of course be specified later, with the latter bounded in $[0, 1]$, which means that $\lambda \mapsto \Psi_\star^\lambda$ is non-increasing. Let $\lambda_\star = \inf\{\lambda > 1 : \Psi_\star^\lambda < \infty\}$. We prove that

$$\limsup_{n \to \infty} \frac{\log(\mathfrak{R}_n^\star)}{\log(n)} = 1 - \frac{1}{\lambda_\star} \,,$$

where $\mathfrak{R}_n^\star$ is the minimax regret over n rounds. Because of the logarithm and the limit, our result says nothing about the leading constant, or even subpolynomial factors. In exchange, the result

holds in extraordinary generality. Only finiteness of the action set is needed. An interesting consequence of our proofs is that the lower bound is always attained by a stationary stochastic adversary. As illustrated by our construction in Section 5, this is a powerful result because it allows you to prove upper bounds for the adversarial setting by constructing algorithms for the stochastic setting. Besides showing that the information ratio characterises the minimax regret, we also prove the existence of finite games for which the minimax Rustichini regret is $n^{4/7}$ up to subpolynomial factors and infinite games for which the standard regret is n^p up to subpolynomial factors for any $p \in [1/2, 1]$.

Setting Let $[k] = \{1, \ldots, k\}$. A k-action partial monitoring game is determined by a latent space $\mathcal{Z}$, a loss function $\mathcal{L} : [k] \times \mathcal{Z} \to [0, 1]$ and a signal function $\mathcal{S} : [k] \times \mathcal{Z} \to \Sigma$ where Σ is called the signal set. Let $\mathcal{K} = \mathscr{P}(\mathcal{Z})$ be the space of finitely supported distributions on $\mathcal{Z}$ and extend the second argument of $\mathcal{L}$ linearly to $\mathcal{K}$ by $\mathcal{L}(a, x) = \int_{\mathcal{Z}} \mathcal{L}(a, z) \, \mathrm{d}x(z)$ for $x \in \mathcal{K}$. The condition that distributions in $\mathcal{K}$ have finite support is a useful trick to avoid measure-theoretic challenges. The learner and adversary interact over n rounds. At the beginning the adversary secretly chooses a sequence of distributions $(x_t)_{t=1}^n$ in $\mathcal{K}$ and then $(z_t)_{t=1}^n$ are sampled from the product measure $\otimes_{t=1}^n x_t$ but not observed. The learner sequentially takes actions $(a_t)_{t=1}^n$ in $[k]$ and observes signals $(\sigma_t)_{t=1}^n$ with $\sigma_t = \mathcal{S}(a_t, z_t)$. To be clear, a_t is chosen based on $a_1, \sigma_1, \ldots, a_{t-1}, \sigma_{t-1}$, possibly with randomisation. The loss suffered by the learner in round t is $\mathcal{L}(a_t, z_t)$, which is not observed. The regret of a learner $\mathscr{L}$ over n rounds is

$$\mathfrak{R}_n(\mathscr{L}, (x_t)_{t=1}^n) = \mathbb{E}\left[\sum_{t=1}^n \mathcal{V}(a_t, x_t)\right] - n\mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^n x_t\right),$$

where $\mathcal{V} : [k] \times \mathcal{K} \to \mathbb{R}$ and $\mathcal{V}_\star : \mathcal{K} \to \mathbb{R}$ are functions for which we will consider two possible definitions and the expectation is with respect to the randomness in the learner's actions and the outcomes (z_t). The standard choice of $\mathcal{V}$ and $\mathcal{V}_\star$ is

$$\mathcal{V}(a, x) = \mathcal{L}(a, x) \qquad\qquad \mathcal{V}_\star(x) = \min_{a \in [k]} \mathcal{L}(a, x). \qquad (1)$$

With these definitions the regret compares the cumulative expected loss of the learner to the cumulative loss of the best action in hindsight. As we explain in a moment, the regret definition arising from the choice in Eq. (1) can be too demanding for games where the signal function provides limited information to the learner. Before that, we need a little more notation. The minimax regret is

$$\mathfrak{R}_n^\star = \inf_{\mathscr{L}} \sup_{(x_t)_{t=1}^n} \mathfrak{R}_n(\mathscr{L}, (x_t)_{t=1}^n),$$

which of course depends on the structure of the game $(\mathcal{L}, \mathcal{S})$ and the definition of $\mathcal{V}$. If $\mathcal{Z}$ is finite and $\mathcal{V}$ is defined by Eq. (1), then the classification of finite partial monitoring (Bartók et al., 2014) states that the minimax regret falls into one of four categories: $\mathfrak{R}_n^\star \in \{0, \Theta(n^{1/2}), \Theta(n^{2/3}), \Theta(n)\}$. Partial monitoring games where $\mathfrak{R}_n^\star = \Theta(n)$ are called hopeless. These are the games where the signal function does not provide enough information for the learner to identify an optimal action and where an alternative definition of $\mathcal{V}$ is more natural.

Example 1 (MATCHING PENNIES IN THE DARK, Perchet 2011a) *Let $k = 2$ and $\mathcal{Z} = [2]$ and $\mathcal{L}(a, z) = \mathbf{1}_{a=z}$ and $\mathcal{S}(a, z) = \perp$ for all $a \in [k]$ and $z \in \mathcal{Z}$. Learning is hopeless in this game*

because the learner's observation does not depend on z. There is no way to differentiate between an adversary that plays a Dirac on $z_t = 1$ in every round and one that plays $z_t = 2$ in every round. Since the optimal action is different in each case, the learner suffers linear minimax regret.

In games like matching pennies in the dark, the expectation that the learner compete with the optimal action in hindsight is too strict. Rustichini (1999) proposed to compare the learner's performance to the Nash equilibrium of the game where the learner knows the average distribution of the signal for each action, which can be estimated by playing all actions. When $x \in \mathcal{K}$ and $a \in [k]$, let $\mathcal{S}(a, x)$ be the law of $\mathcal{S}(a, z)$ when $z \sim x$ and define an equivalence relation R on $\mathcal{K} \times \mathcal{K}$ by $x\mathrm{R}y$ if $\mathcal{S}(a, x) = \mathcal{S}(a, y)$ for all $a \in [k]$. That is, $x\mathrm{R}y$ is true if the signal distribution for all actions is the same for both x and y. The Rustichini notion of regret uses

$$\mathcal{V}(\pi, x) = \sup_{y : y\mathrm{R}x} \mathcal{L}(\pi, y) \qquad\qquad \mathcal{V}_\star(x) = \min_{\pi \in \Delta_k} \mathcal{V}(\pi, x), \qquad (2)$$

where $\Delta_k = \{\pi \in [0,1]^k : \|\pi\|_1 = 1\}$ and $\mathcal{L}(\pi, x) = \sum_{a \in [k]} \pi(a)\mathcal{L}(a, x)$. Note that the minimising policy in Eq. (2) may not be an extreme point of the simplex. Indeed, in the game of Example 1, $x\mathrm{R}y$ for all x and y. An easy calculation shows that $\mathcal{V}_\star(x) = 1/2$ and the policy π such that $\mathcal{V}(\pi, x) = \mathcal{V}_\star(x)$ randomises uniformly between the two actions. The standard and Rustichini regret coincide when $x\mathrm{R}y$ if and only if $x = y$. Otherwise the situation is quite subtle, as we explain at length in Appendix F, a section that is best read after the main body.

Important remark   When we say a result holds in the *standard setting* we mean that $\mathcal{V}$ and $\mathcal{V}_\star$ are defined as in Eq. (1) and $\mathcal{Z}$ may be finite or infinite, while in the *Rustichini setting* $\mathcal{V}$ and $\mathcal{V}_\star$ are defined as in Eq. (2) and $\mathcal{Z}$ is finite. Occasionally we use the notation $\mathrm{const}_{\mathcal{LS}}$ to denote a constant that depends only on $\mathcal{L}$ and $\mathcal{S}$ and may vary from one equation to the next. Without the indices, const refers to a universal constant.

Related work   Partial monitoring and the non-standard definition of the regret used in part of this work was proposed by Rustichini (1999), who proved a Hannan-type consistency result using an argument based on Blackwell approachability (Blackwell, 1956). More recent work on partial monitoring has focussed on finite games using the classical definition of the regret, where the learner competes with the best action in hindsight. A long line of work covering various cases (Piccolboni and Schindelhauer, 2001; Cesa-Bianchi et al., 2006; Foster and Rakhlin, 2012; Antos et al., 2013; Lattimore and Szepesvári, 2019a) led to a complete classification of finite partial monitoring games (Bartók et al., 2014). This shows that the minimax regret for all such games is either 0, $\Theta(n^{1/2})$, $\Theta(n^{2/3})$ or $\Theta(n)$, with the leading constants and rate depending on the structure of the game. More recent work has developed simple algorithms and improved dependence on constants (Lattimore and Szepesvári, 2019b, 2020). As far as we know, there has been no general study of infinite partial monitoring games, though of course many specific instances fall into this category (bandits with a continuum of possible losses, for example).

Besides Rustichini (1999), few authors have directly considered partial monitoring with Rustichini's regret. Lugosi et al. (2008) improve on the results of Rustichini (1999) by providing rates and high probability bounds, which in the general case are $O(n^{4/5})$ with certain special cases giving improved bounds. The setting can be viewed as an instance of the even more general problem of Blackwell approachability with partial monitoring studied by Perchet (2011b); Mannor and Shamir (2011); Mannor et al. (2014) and Kwon and Perchet (2017). When applied to our setting, the

last of these papers prove that the minimax regret is $\mathfrak{R}_n^\star = O(n^{2/3})$. There are games for which $\mathfrak{R}_n^\star = \Omega(n^{2/3})$, so this rate cannot be improved in general. Note this means there is no concept of a hopeless game when the Rustichini regret is used (essentially this is why it is used). An interesting question is whether or not our ideas can be lifted to approachability with partial monitoring.

Our approach for proving upper bounds is based on the duality between information directed sampling (Russo and Van Roy, 2014) and its connection to quantities involved in bounding the regret of mirror descent (Zimmert and Lattimore, 2019; Lattimore and György, 2021). The principle novelty here is new lower bounds in terms of the information ratio. While one might guess that the relative entropy might be the notion of information gain most suitable for this, it turns out that defining the information gain as an expected Bregman divergence between a posterior and prior with respect to the logarithmic barrier (rather than negentropy) is easier to work with. Correspondingly, our algorithm is based on mirror descent with the logarithmic barrier as a potential function.

A recent paper by Foster et al. (2021) investigates a related idea and also prove upper and lower bounds in terms of an information-ratio-like quantity. Their results are stronger in some dimensions and weaker in others. For example, while we consider adversarial partial monitoring problems, they focus on stochastic problems with bandit feedback (but can handle contextual problems). On the other hand, our work essentially relies on a (small) finite action set and only characterises the minimax regret in terms of the horizon, while their methods are suitable for large action sets and yield bounds that demonstrate dependencies on other game-dependent quantities. One might hope to one day get the best of all these worlds, though at the moment it seems some real new idea is needed for such a result.

Notation   The vectors of all ones and zeros are $\mathbf{1}$ and $\mathbf{0}$ respectively. The dimension of these quantities should understood from the context. Let $\mathscr{P}(A)$ denote the set of finitely supported probability distributions on a set A and $\mathscr{P}_m(A) \subset \mathscr{P}(A)$ the set of those distributions supported on at most m elements. The support of a distribution $\mu \in \mathscr{P}(A)$ is $\mathrm{spt}(\mu) = \{z \in A : \mu(z) > 0\}$. The relative interior of a set $A \subset \mathbb{R}^d$ is $\mathrm{ri}(A)$ and its convex hull is $\mathrm{conv}(A)$. When A is convex, its exterior points are denoted by $\mathrm{ext}(A)$. The indicator function of A is $\mathbf{1}_A$. When A is a polytope, $\mathrm{faces}(A)$ is the set of faces of A, including itself and the empty set (formal definition in Appendix A). The relative entropy between probability measures ν and μ on the same measurable space is $\mathrm{KL}(\nu, \mu)$ and the total variation distance is $\|\nu - \mu\|_{\mathrm{TV}}$. The ℓ_1 distance between a point $x \in \mathbb{R}^d$ and a nonempty closed set $F \subset \mathbb{R}^d$ is $\mathrm{dist}_1(x, F) = \min_{y \in F} \|x - y\|_1$. A game is called finite if $\mathcal{Z}$ and k are finite. For finite games the loss and signal functions are often represented by matrices where the adversary plays in the columns and the learner plays in the rows: $\mathcal{L}_{a,z} = \mathcal{L}(a, z)$ and $\mathcal{S}_{a,z} = \mathcal{S}(a, z)$. Given a convex function $F : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, the Bregman divergence with respect to F is $\mathrm{D}(x, y) = F(x) - F(y) - \langle \nabla F(y), x - y \rangle$, which is defined provided that $x \in \mathrm{dom}(F)$ and $y \in \mathrm{dom}(\nabla F)$. The Fenchel conjugate of F is $F^\star(x) = \sup_{u \in \mathbb{R}^m} \langle u, x \rangle - F(x)$.

Properties of $\mathcal{V}$ The functions $\mathcal{V}$ and $\mathcal{V}_\star$ play a central role in our analysis.

Lemma 1 *The following hold in both the standard and Rustichini settings:*
  (a) $\mathcal{V}_\star$ is concave and piecewise linear.
  (b) $x \mapsto \mathcal{V}(\pi, x)$ is concave and piecewise linear for all $\pi \in \Delta_k$ (linear in the standard setting).
  (c) $\pi \mapsto \mathcal{V}(\pi, x)$ is convex and piecewise linear for all $x \in \mathcal{K}$ (linear in the standard setting).

The proof is elementary in the standard setting. In the Rustichini setting it follows from classical linear programming arguments. Details are in Appendix B. This lemma is one place where

the finiteness of $\mathcal{Z}$ is used in our results on the Rustichini notion of regret. Let m be the smallest number of pieces needed to define $\mathcal{V}_\star$ and let $\mathcal{V}^m : \mathcal{K} \to \mathbb{R}^m$ be linear so that $\mathcal{V}_\star(x) = \min_{p \in \Delta_m} \langle p, \mathcal{V}^m(x) \rangle$. In the standard setting $m \leq k$ is immediate while in the Rustichini setting m can be larger depending on the structure of the signal function but is finite whenever $\mathcal{Z}$ is finite. Note that $\max_{x \in \mathcal{K}} \|\mathcal{V}^m(x)\|_\infty < \infty$ is obvious for both settings. The constant m is used to denote the number of linear pieces needed to represent $\mathcal{V}_\star$ for the remainder of the article.

2. Mirror descent, duality and the information ratio

We are now in a position to formally define the information ratio and state our main theorem. Given a prior $\mu \in \mathscr{P}(\mathcal{K})$, let

$$
\mathrm{I}(\pi, \mu) = \sum_{a=1}^{k} \pi(a) \sum_{x \in \mathrm{spt}(\mu)} \mathrm{KL}\left(\mathcal{S}(a, \mu), \mathcal{S}(a, x)\right),
$$

which will be used as a measure of information gain. At a very high level, $\mathrm{I}(\pi, \mu)$ is a measure of the variation of signal distributions over different $x \in \mathrm{spt}(\mu)$ when an action a is sampled from π. This definition differs from the version of information gain based on the mutual information introduced by Russo and Van Roy (2014) and used in most followup works. Instead, it corresponds to using the logarithmic barrier in the generalised machinery by Lattimore and György (2021). Very briefly, the mutual information is defined as an expected Bregman divergence between posterior and prior with respect to the negentropy potential. Our information gain can be written as an expected Bregman divergence with respect to the logarithmic barrier (Appendix E). We also need to define an instantaneous regret estimate:

$$
\Delta(\pi, \mu) = \int_{\mathcal{K}} (\mathcal{V}(\pi, x) - \mathcal{V}_\star(x)) \, \mathrm{d}\mu(x). \tag{3}
$$

We will constantly abuse notation by writing $\Delta(\pi, x)$ as an abbreviation for $\Delta(\pi, \mu)$ with μ a Dirac on x. Finally, let $\Delta_\star(\mu) = \min_{\pi \in \Delta_k} \Delta(\pi, \mu)$, which exists by continuity and compactness arguments (Lemma 14).

Fact 2 $\pi \mapsto \Delta(\pi, \mu)$ *is convex,* $\mu \mapsto \Delta(\pi, \mu)$ *is linear and* $\pi \mapsto \mathrm{I}(\pi, \mu)$ *is linear.*

The λ-information ratio is a function $\Psi^\lambda : \Delta_k \times \mathscr{P}(\mathcal{K})$ is given by

$$
\Psi^\lambda(\pi, \mu) = \frac{\Delta(\pi, \mu)^\lambda}{\mathrm{I}(\pi, \mu)},
$$

where $0/0 = 0$. Let $\Psi^\lambda_\star(\mu) = \min_{\pi \in \Delta_k} \Psi^\lambda(\pi, \mu)$. The minimax information ratio is $\Psi^\lambda_\star = \sup_{\mu \in \mathscr{P}_m(\mathcal{K})} \Psi^\lambda_\star(\mu)$. Note the restriction to measures $\mu \in \mathscr{P}_m(\mathcal{K})$ with support size at most m, which is the number of pieces needed to represent piecewise linear function $\mathcal{V}_\star$. Since $\Delta(\pi, \mu) \in [0, 1]$, the map $\lambda \mapsto \Psi^\lambda_\star$ is nonincreasing. Let $\lambda_\star = \inf\{\lambda > 1 : \Psi^\lambda_\star < \infty\}$, where the infimum of the empty set is ∞.

Theorem 3 (MAIN THEOREM) *In both the standard and Rustichini settings:*

$$
\limsup_{n \to \infty} \frac{\log(\mathfrak{R}^\star_n)}{\log(n)} = 1 - \frac{1}{\lambda_\star}.
$$

Note that Theorem 3 does not say that the regret does not depend on whether or not the standard or Rustichini regret is used because $\mathcal{V}$ appears in the definition of the regret function Δ and hence also the information ratio. We prove Theorem 3 by establishing upper and lower bounds in terms of the information ratio, starting with the upper bound.

Algorithm and upper bound   Our algorithm is a modification of the exploration-by-optimisation algorithm (Lattimore and György, 2021), which itself is a version of mirror descent with loss estimators and exploration distribution computed by solving an optimisation problem. The mirror descent component plays on the domain $\mathcal{D} = \Delta_m \cap [1/n, 1]^m$ and using as a potential function the logarithmic barrier $F : \mathbb{R}^m \to \mathbb{R}$ defined on the positive orthant by $F(q) = -\sum_{i=1}^m \log(q_i)$. Let $F^\star$ be the Fenchel conjugate of F and let D and $\mathrm{D}_\star$ be the Bregman divergences with respect to F and $F^\star$. Let $\mathcal{E}$ be the space of all functions from $[k] \times \Sigma \to \mathbb{R}^m$. Given $q \in \mathcal{D}$ and $\eta > 0$, define $\Lambda_{q\eta} : \mathrm{ri}(\Delta_k) \times \mathcal{E} \to \mathbb{R}$ by $\Lambda_{q\eta}(\pi, \hat{v}) =$

$$
\sup_{\substack{p \in \mathcal{D} \\ x \in \mathcal{K}}} \left(\mathcal{V}(\pi, x) - \langle p, \mathcal{V}^m(x) \rangle + \mathbb{E} \left[\frac{\langle p - q, \hat{v}(a, \sigma) \rangle}{\pi(a)} + \frac{1}{\eta} \mathrm{D}_\star \left(\nabla F(q) - \frac{\eta \hat{v}(a, \sigma)}{\pi(a)}, \nabla F(q) \right) \right] \right),
$$

where the expectation is with respect to $(z, a) \sim x \otimes \pi$ and the signal is $\sigma = \mathcal{S}(a, z)$. Then let $\Lambda_{q\eta}^\star = \inf_{\pi \in \mathrm{ri}(\Delta_k), \hat{v} \in \mathcal{E}} \Lambda_{q\eta}(\pi, \hat{v})$. At a high level, $\Lambda_{q\eta}$ is a quantity that appears in the regret bound of mirror descent. Our algorithm will minimise this bound.

---

input: learning rate $\eta > 0$ and precision $\epsilon > 0$ and horizon n

$q_1 = \mathbf{1}/m \in \Delta_m$

for $t = 1$ **to** n:

    find $\pi_t \in \mathrm{ri}(\Delta_k)$ and $\hat{v}_t \in \mathcal{E}$ such that $\Lambda_{q_t \eta}(\pi_t, \hat{v}_t) \leq \Lambda_{q_t \eta}^\star + \epsilon$

    sample $a_t \sim \pi_t$ and observe σ_t

    update: $q_{t+1} = \arg\min_{q \in \mathcal{D}} \dfrac{\langle q, \hat{v}(a_t, \sigma_t) \rangle}{\pi_t(a_t)} + \dfrac{1}{\eta} \mathrm{D}(q, q_t)$

Algorithm 1: Exploration by optimisation

In the standard setting Algorithm 1 is the same as the original. In the Rustichini setting the difference is that our algorithm plays mirror descent on Δ_m rather than Δ_k and $\pi \mapsto \mathcal{V}(\pi, x)$ is convex, rather than linear. The upper bound in Theorem 3 follows almost immediately from Theorem 4 below, the proof of which is essentially a corollary of Theorems 8 and 9 in the paper by Lattimore and György (2021). The differences and additional steps are explained in Appendix E.

Theorem 4   *Provided that $\eta > 0$, $n \geq m$ and $\epsilon \geq 0$, then for any $\lambda > 1$ the regret of Algorithm 1 is bounded by*

$$
\mathfrak{R}_n \leq n\epsilon + 2 \max_{x \in \mathcal{K}} \|\mathcal{V}^m(x)\|_\infty + \frac{m \log(n/m)}{\eta} + n \left(\frac{\lambda - 1}{\lambda} \right) \left(\frac{\eta \Psi_\star^\lambda}{\lambda} \right)^{\frac{1}{\lambda - 1}}.
$$

The upper bound of Theorem 3 is a corollary of Theorem 4.

Proof [THEOREM 3, UPPER BOUND] If $\lambda_\star = \infty$, then the result is obvious. Otherwise, suppose that $\lambda_\star < \lambda < \infty$. Then $\Psi_\star^\lambda < \infty$ and by Theorem 4, if Algorithm 1 is run with learning rate

$$\eta = \lambda \left(\frac{m \log(n/m)}{n} \right)^{1-1/\lambda} (\Psi_\star^\lambda)^{-1/\lambda} \tag{4}$$

and $\epsilon = 1/n$, then

$$\mathfrak{R}_n \leq 1 + 2 \max_{x \in \mathcal{K}} \|\mathcal{V}^m(x)\|_\infty + \frac{m \log(n/m)}{\eta} + n \left(\frac{\lambda - 1}{\lambda} \right) \left(\frac{\eta \Psi_\star^\lambda}{\lambda} \right)^{\frac{1}{\lambda - 1}}$$

$$= 1 + 2 \max_{x \in \mathcal{K}} \|\mathcal{V}^m(x)\|_\infty + \left(m \Psi_\star^\lambda \log(n/m) \right)^{1/\lambda} n^{1 - \frac{1}{\lambda}} .$$

Since this bound is independent of the adversary's decisions, $\limsup_{n \to \infty} \log(\mathfrak{R}_n^\star) / \log(n) \leq 1 - \frac{1}{\lambda}$. The theorem follows by taking the limit as λ tends to $\lambda_\star$ from above. ∎

Computation complexity The computation complexity of Algorithm 1 depends on the structure of the game. For finite $\mathcal{Z}$, there are two challenges. First, some pre-computation is needed to find $\mathcal{V}^m$. The optimisation to find π_t and $\hat{v}_t$ is a convex program and can also be solved to reasonable precision in polynomial time. Updating q_t is naively a convex program as well, but can be solved analytically except for a line search. The optimal tuning of η given in Eq. (4) is probably hard to find but there is no problem tuning the learning rate online as explained by Lattimore and György (2021). All up, for medium-sized Rustichini games Algorithm 1 is a reasonably practical option. For infinite games the optimisation problem to find π_t and $\hat{v}_t$ remains convex but is infinite-dimensional. In this case we view the arguments as more of a tool for understanding the regret and an inspiration for algorithm design.

3. Lower bounds

More or less all lower bounds for bandits and online learning are established via information-theoretic techniques. Our lower bound (below) is proven similarly but because of the limited structure of the problem only the most general machinery can be used. The first lemma is used to relate the instantaneous regret to the minimax information ratio in a linear way. A similar result when $\lambda = 2$ was given by Kirschner et al. (2021). The proof is in Appendix B.

Lemma 5 *For all $\mu \in \mathscr{P}(\mathcal{K})$ and $\pi \in \Delta_k$, $\Psi_\star^\lambda(\mu) \leq \frac{2^\lambda \Delta_\star(\mu)^{\lambda-1} \Delta(\pi, \mu)}{\mathrm{I}(\pi, \mu)}$.*

Next we need a result showing that for any $\mu \in \mathscr{P}_m(\mathcal{K})$, if a collection of policies on index set $\mathrm{spt}(\mu)$ are sufficiently close to each other, then at least one of the policies has regret comparable to $\Delta_\star(\mu)$.

Lemma 6 *There exists a game-dependent constant $\varphi > 0$ such that the following holds. Let $\mu \in \mathscr{P}_m(\mathcal{K})$ and suppose that $(\pi_x)_{x \in \mathrm{spt}(\mu)}$ is a set of policies with $\|\pi_x - \pi\|_\infty \leq \varphi$ for some $\pi \in \Delta_k$. Then $\max_{x \in \mathrm{spt}(\mu)} \Delta(\pi_x, x) \geq \mathrm{const}_{\mathcal{LS}} \Delta_\star(\mu)$.*

Proof [STANDARD SETTING] We prove Lemma 6 in the standard setting with $\varphi = 1/(4k)$. Let $\mu \in \mathscr{P}_m(\mathcal{K})$ and π and $(\pi_x)_{x \in \text{spt}(\mu)}$ be a collection of policies satisfying the conditions of the lemma. Let $\mathcal{A} = \{a \in [k] : \pi(a) \geq 1/(2k)\}$, which satisfies $\sum_{a \in \mathcal{A}} \pi(a) \geq 1/2$. Define a policy $\rho(a) = \pi(a) \mathbf{1}_{\mathcal{A}}(a) / \sum_{a \in \mathcal{A}} \pi(a)$. Then,

$$\Delta_\star(\mu) \leq \Delta(\rho, \mu) = \frac{\int_{\mathcal{K}} \sum_{a \in \mathcal{A}} \pi(a) \Delta(a, x) \, \mathrm{d}\mu(x)}{\sum_{a \in \mathcal{A}} \pi(a)} \leq 2 \int_{\mathcal{K}} \sum_{a \in \mathcal{A}} \pi(a) \Delta(a, x) \, \mathrm{d}\mu(x)$$

$$\leq 2 \max_{x \in \text{spt}(\mu)} \sum_{a \in \mathcal{A}} \pi(a) \Delta(a, x) \leq 4 \max_{x \in \text{spt}(\mu)} \sum_{a \in \mathcal{A}} \pi_x(a) \Delta(a, x) \leq 4 \max_{x \in \text{spt}(\mu)} \Delta(\pi_x, x) . \qquad \blacksquare$$

The proof above relied on the linearity of $\pi \mapsto \Delta(\pi, x)$, which does not hold in the Rustichini setting. There may be an elegant linear programming proof of Lemma 6 for the Rustichini setting, but we offer a relatively involved geometric argument in Appendix C. Now we have all the tools needed to prove the lower bound in our main theorem.

Proof [THEOREM 3, LOWER BOUND] Our proof works by showing that for every learner there exists a stochastic adversary that samples $(z_t)_{t=1}^n$ independently from some carefully chosen $x \in \mathcal{K}$ for which the regret is large. We start by introducing some notation (step 1). In step 2 we make the necessary relative entropy calculations and connect the information gain to the relative entropy between various stochastic environments. The regret is then decomposed in two ways, depending on whether or not the algorithm explores a lot or a little (step 3). Finally we balance the constants in the proof to obtain the result.

Step 1: Notation and setup The result is obvious if $\lambda_\star = 1$. For the remainder assume that $\lambda_\star > 1$. Fix an arbitrary policy for the learner and let $1 < \lambda < \lambda_\star$ and $\mu \in \mathscr{P}_m(\mathcal{K})$. Given $x \in \mathcal{K}$, let $\mathbb{P}_x$ be the measure on $(a_t)_{t=1}^n$, $(\pi_t)_{t=1}^n$ and $(\sigma_t)_{t=1}^n$ when the learner interacts with the (stochastic) environment where $x_t = x$ for all t. Let $\mathbb{P} = \mathbb{P}_y$ with $y = \int_{\mathcal{K}} x \, \mathrm{d}\mu(x)$. Expectations with respect to these measures are denoted by $\mathbb{E}_{\mathbb{P}}$ and $\mathbb{E}_{\mathbb{P}_x}$. The regret when the learner faces the stochastic environment with $(z_t)_{t=1}^n$ sampled from $\otimes_{t=1}^n x$ is denoted by

$$\mathfrak{R}_n(x) = \mathbb{E}_{\mathbb{P}_x} \left[\sum_{t=1}^n \mathcal{V}(\pi_t, x) \right] - n \mathcal{V}_\star(x) .$$

Let $\mathcal{F}_t = \sigma(\pi_1, a_1, \sigma_1, \ldots, \pi_t, a_t, \sigma_t)$ be the σ-algebra generated by the history after t observations. Given an $(\mathcal{F}_t)_{t=0}^n$-adapted stopping time τ, let $\mathbb{P}^\tau$ and $\mathbb{P}_x^\tau$ denote the measures $\mathbb{P}$ and $\mathbb{P}_x$ restricted to $\mathcal{F}_\tau$.

Step 2: Relative entropy Let τ be a stopping time with respect to the filtration $(\mathcal{F}_t)_{t=0}^n$ with $\tau \in \{0, \ldots, n\}$ almost surely. Then by the chain rule for relative entropy,

$$\text{KL}(\mathbb{P}^\tau, \mathbb{P}_x^\tau) = \mathbb{E}_{\mathbb{P}} \left[\sum_{t=1}^\tau \text{KL}(\mathcal{S}(a_t, \mu), \mathcal{S}(a_t, x)) \right] \leq \mathbb{E}_{\mathbb{P}} \left[\sum_{t=1}^\tau \sum_{y \in \text{spt}(\mu)} \text{KL}(\mathcal{S}(a_t, \mu), \mathcal{S}(a_t, y)) \right]$$

$$= \mathbb{E}_{\mathbb{P}} \left[\sum_{t=1}^\tau \mathrm{I}(\pi_t, \mu) \right] . \tag{5}$$

We also need a Pinsker-like inequality, which we prove in Appendix B.

Lemma 7 *Suppose that $x \in \mathrm{spt}(\mu)$ and $\epsilon > 0$ and $t \in \{0, \ldots, n\}$. Then,*

$$\|\mathbb{P}_x^t - \mathbb{P}^t\|_{\mathrm{TV}} \leq \sqrt{2\epsilon} + \mathbb{P}\left(\sum_{s=1}^{t} \mathrm{I}(\pi_s, \mu) \geq \epsilon\right).$$

Step 3: Regret decomposition Let φ be the constant from Lemma 6 and $\delta = \varphi/(1 + 1/m)$ and $\epsilon = (\delta/m)^2/2$. Suppose that

$$\mathbb{P}\left(\sum_{t=1}^{n} \mathrm{I}(\pi_t, \mu) \geq \epsilon\right) \leq \delta. \tag{6}$$

By Lemma 7, for all $t \in [n]$, $\|\mathbb{E}_{\mathbb{P}_x}[\pi_t] - \mathbb{E}_{\mathbb{P}}[\pi_t]\|_{\infty} \leq \sqrt{2\epsilon} + \delta = \varphi$. Hence, by a triangle inequality,

$$\left\|\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\mathbb{P}_x}[\pi_t] - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\mathbb{P}}[\pi_t]\right\|_{\infty} \leq \frac{1}{n}\sum_{t=1}^{n}\|\mathbb{E}_{\mathbb{P}_x}[\pi_t] - \mathbb{E}_{\mathbb{P}}[\pi_t]\|_{\infty} \leq \varphi.$$

Lemma 6 now shows there exists an $x \in \mathrm{spt}(\mu)$ such that

$$\Delta\left(\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\mathbb{P}_x}[\pi_t], x\right) \geq \mathrm{const}_{\mathcal{LS}}\,\Delta_{\star}(\mu).$$

Combining this with the convexity of $\pi \mapsto \Delta(\pi, x)$,

$$\mathfrak{R}_n^{\star} \geq \max_{x \in \mathrm{spt}(\mu)} \mathfrak{R}_n(x) = \max_{x \in \mathrm{spt}(\mu)} \sum_{t=1}^{n}\mathbb{E}_{\mathbb{P}_x}\left[\Delta(\pi_t, x)\right]$$

$$\geq n \max_{x \in \mathrm{spt}(\mu)} \Delta\left(\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\mathbb{P}_x}[\pi_t], x\right) \geq \mathrm{const}_{\mathcal{LS}}\, n\Delta_{\star}(\mu).$$

Suppose for the remainder of this step that Eq. (6) does not hold. Let $\tau = \max\{t \in [n] : \sum_{s=1}^{t}\mathrm{I}(\pi_s, \mu) \leq \epsilon\}$, which is a stopping time because $\mathrm{I}(\pi_{t+1}, \mu)$ is $\mathcal{F}_t$-measurable. By Lemma 5,

$$\mathbb{P}\left(\tau < n \text{ and } \sum_{t=1}^{\tau+1}\Delta(\pi_t, \mu) \geq \frac{2^{-\lambda}\epsilon\Psi_{\star}^{\lambda}(\mu)}{\Delta_{\star}(\mu)^{\lambda-1}}\right) \geq \mathbb{P}(\tau < n) > \delta.$$

Hence, by a union bound, there exists an x such that

$$\mathbb{P}\left(\tau < n \text{ and } \sum_{t=1}^{\tau+1}\Delta(\pi_t, x) \geq \frac{2^{-\lambda}\epsilon\Psi_{\star}^{\lambda}(\mu)}{\Delta_{\star}(\mu)^{\lambda-1}}\right) > \frac{\delta}{m}.$$

Since $\mathrm{KL}(\mathbb{P}^{\tau}, \mathbb{P}_x^{\tau}) \leq \epsilon$ by the definition of τ and Eq. (5) and $\Delta(\pi_t, x)$ is $\mathcal{F}_{t-1}$-measurable, by Pinsker's inequality,

$$\mathbb{P}_x\left(\tau < n \text{ and } \sum_{t=1}^{\tau+1}\Delta(\pi_t, x) \geq \frac{2^{-\lambda}\epsilon\Psi_{\star}^{\lambda}(\mu)}{\Delta_{\star}(\mu)^{\lambda-1}}\right) > \frac{\delta}{m} - \sqrt{\epsilon/2} = \frac{\delta}{2m}.$$

Therefore,

$$\mathfrak{R}_n^{\star} \geq \mathfrak{R}_n(x) = \mathbb{E}_{\mathbb{P}_x}\left[\sum_{t=1}^{n}\Delta(\pi_t, x)\right] > \frac{2^{-\lambda}\epsilon\Psi_{\star}^{\lambda}(\mu)}{\Delta_{\star}(\mu)^{\lambda-1}} \cdot \frac{\delta}{2m} = \frac{2^{-\lambda-2}\varphi^3}{m^3\left(1 + \frac{1}{m}\right)^3} \cdot \frac{\Psi_{\star}^{\lambda}(\mu)}{\Delta_{\star}(\mu)^{\lambda-1}}.$$

Step 4: Balancing the terms The results in the previous steps show that

$$\mathfrak{R}_n^\star \geq \text{const}_{\mathcal{LS}} \min \left(n\Delta_\star(\mu), \frac{\Psi_\star^\lambda(\mu)}{\Delta_\star(\mu)^{\lambda-1}} \right) .$$

Since this holds for any horizon n, let us choose $n = \lfloor \Psi_\star^\lambda(\mu)/\Delta_\star(\mu)^\lambda \rfloor$. Naive bounding and rearranging shows that for suitably large $\Psi_\star^\lambda(\mu)$,

$$\mathfrak{R}_n(x) \geq \text{const}_{\mathcal{LS}}\, n^{1-\frac{1}{\lambda}} \left(\Psi_\star^\lambda(\mu) \right)^{\frac{1}{\lambda}} .$$

Since $\lambda < \lambda_\star$, $\Psi_\star^\lambda = \infty$ and so there exists a sequence $(\mu_t)_{t=1}^\infty$ in $\mathscr{P}_m(\mathcal{K})$ with $\lim_{t\to\infty} \Psi_\star^\lambda(\mu_t) = \infty$. Because $\Delta_\star(\mu) \in [0,1]$ for all μ, the horizon as defined above is always at least $\Omega(\Psi_\star^\lambda(\mu))$. Taking logarithms and a limit shows that $\limsup_{n\to\infty} \log(\mathfrak{R}_n^\star)/\log(n) \geq 1 - 1/\lambda$ for all $\lambda < \lambda_\star$. Taking the limit as λ tends to $\lambda_\star$ from below completes the proof. ∎

4. Cell Structure for Rustichini's Regret

Before the proof of Lemma 6 for the Rustichini setting we need to explore the combinatorial and geometric structure of Rustichini's regret. These concepts are needed anyway to understand why Rustichini games do not admit the same classification as finite partial monitoring games and to understand our example in Section 5. Assume for this section that we are in the Rustichini setting so that $\mathcal{V}$ and $\mathcal{V}_\star$ are defined as in Eq. (2) and $\mathcal{Z}$ is finite and $\mathcal{K}$ is the probability simplex over $\mathcal{Z}$.

The idea is to appropriately generalise the cell decomposition used in the standard finite setting, which was introduced by Piccolboni and Schindelhauer (2001) and played a fundamental role in all subsequent work. A morally similar structure is used by Kwon and Perchet (2017), but the setting is different and matching the ideas does not seem straightforward. The most important takeaway is that policies can be optimal on multiple cells, which is not true in the standard setting.

Recall that $\mathcal{V}_\star(x) = \min_{\alpha \in [m]} \mathcal{V}^m(x)_\alpha$ with $\mathcal{V}^m : \mathcal{K} \to \mathbb{R}^m$ linear. For $\alpha \in [m]$, the cell of α is the convex polytope $P_\alpha = \{x \in \mathcal{K} : \mathcal{V}_\star(x) = \mathcal{V}^m(x)_\alpha\}$. The set $\mathcal{F} = \cup_{\alpha \in [m]} \text{faces}(P_\alpha)$ is a polyhedral complex (see Appendix A) and corresponds to the cell decomposition in standard finite partial monitoring games. A policy $\pi \in \Delta_k$ is optimal at $x \in \mathcal{K}$ if $\Delta(\pi, x) = 0$ and the set of such policies is denoted by $\mathcal{N}(x) = \{\pi \in \Delta_k : \Delta(\pi, x) = 0\}$. Note that $\mathcal{N}(x)$ is nonempty by the definition of Δ. The following lemma shows that $\mathcal{N}(x)$ only depends on the face F such that $x \in \text{ri}(F)$. The proof is given in Appendix B.

Lemma 8 *Suppose that $F \in \mathcal{F}$ and $x, y \in \text{ri}(F)$. Then $\mathcal{N}(x) = \mathcal{N}(y)$.*

Based on Lemma 8, it makes sense to define $\mathcal{N}(F) = \mathcal{N}(x)$ where $x \in \text{ri}(F)$ is any point in the relative interior of F. Note that because $x \mapsto \Delta(\pi, x)$ is continuous, it follows immediately that $\Delta(\pi, x) = 0$ for all $x \in F$ and $\pi \in \mathcal{N}(F)$. Continuity of Δ also implies that $\mathcal{N}(F)$ is closed for all faces $F \in \mathcal{F}$. Further, $\pi \mapsto \Delta(\pi, x)$ is convex and piecewise linear for any x, so $\mathcal{N}(F)$ is a convex polytope. An essential difference between the cell decomposition in the standard setting and the Rustichini setting is that $\mathcal{N}(P_\alpha)$ and $\mathcal{N}(P_\beta)$ are not necessary disjoint for distinct $\alpha, \beta \in [m]$. The importance of this difference will become apparent in the next section.

5. Exotic Rustichini Games

Theorem 3 shows that the minimax regret of all finite Rustichini games is characterised by the generalised information ratio. There has been some speculation about whether or not Rustichini games have a minimax regret in $\{0, \Theta(n^{1/2}), \Theta(n^{2/3})\}$, like standard finite partial monitoring games except for the absence of hopeless games. The theorem of this section shows that this is not the case. The fundamental difference between Rustichini games and the standard setting is that policies can be optimal on multiple cells. In partial monitoring the regret is determined by the ability of the learner to estimate in which cell the adversary is playing, while in Rustichini games there is a richer combinatorial structure. The example in the analysis below illustrates some of the challenges.

Theorem 9 *There exists a finite Rustichini game such that* $\limsup_{n\to\infty} \log(\mathfrak{R}_n^\star)/\log(n) = 4/7$.

Proof [SKETCH] The game we use to exhibit the desired minimax regret is written below in matrix form. The construction is not exactly trivial and the important features of the game cannot be absorbed by staring at the matrices.

$$
\mathcal{L} = \begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 1/4 & 0 & 1/4 & 1 & 1/4 & 0 & 1/4 \\
1 & 1/4 & 0 & 1/4 & 1 & 1/4 & 0 & 1/4 \\
1 & 1/4 & 0 & 1/4 & 1 & 1/4 & 0 & 1/4 \\
2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
2 & 2 & 2 & 2 & 2 & 2 & 2 & 2
\end{bmatrix}
\qquad
\mathcal{S} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0
\end{bmatrix} .
$$

Before diving into details, let us outline the intuitive role of the actions. For this game, the set of optimal policies $\mathcal{N}(x)$ for a given $x \in \mathcal{K}$ are the convex combinations of the deterministic policies that are optimal at x. Note that is not true in general. At every point x one of the actions in $\{1, 2, 3, 4\}$ are optimal. Actions $\{5, 6, 7\}$ are optimal on a low dimensional subset while actions $\{8, 9\}$ are never optimal. Their only role is exploration and the regret when playing them is always large.

Loss structure The domain $\mathcal{K}$ of the adversary is the probability simplex over the 8 outcomes, which is 7-dimensional and not easy to represent. Fortunately, the regret function can be represented via a low dimensional parameterisation. A little notation is needed for this. Let $u(x) = x_4 - x_2$ and $v(x) = x_8 - x_6$. A tedious calculation shows that

$$
\Delta(\pi, x) \overset{\times}{=} \pi(\{8, 9\}) +
\begin{cases}
|u(x)|\pi(\{4, 5, 6, 7\}) + |v(x)|\pi(\{2, 5, 6, 7\}) & \text{if } u(x) \geq 0 \text{ and } v(x) \geq 0 \\
|u(x)|\pi(\{4, 5, 6, 7\}) + |v(x)|\pi(\{1, 5, 6, 7\}) & \text{if } u(x) \geq 0 \text{ and } v(x) \leq 0 \\
|u(x)|\pi(\{3, 5, 6, 7\}) + |v(x)|\pi(\{2, 5, 6, 7\}) & \text{if } u(x) \leq 0 \text{ and } v(x) \geq 0 \\
|u(x)|\pi(\{3, 5, 6, 7\}) + |v(x)|\pi(\{1, 5, 6, 7\}) & \text{if } u(x) \leq 0 \text{ and } v(x) \leq 0 .
\end{cases}
$$

where $f \overset{\times}{=} g$ indicates that $f(x) \leq \text{const } g(x) \leq \text{const } f(x)$ for all x. Notice this means that the deterministic policy that plays action $a = 1$ is optimal if $v(x) \geq 0$, while $a = 2$ is optimal if $v(x) \leq 0$ and $a = 3$ is optimal if $u(x) \geq 0$ and $a = 4$ is optimal if $u(x) \leq 0$. Therefore the learner can identify an optimal policy by learning *either* the sign of $u(x)$ or $v(x)$.

11

Signal structure   Actions $\{1, 2, 3, 4\}$ are entirely uninformative. Actions $\{5, 6, 7\}$ can be used to estimate $u(x)$. By playing all actions the learner can estimate both $u(x)$ and $v(x)$.

Bounding the regret   By the lower bound construction in Theorem 3 it suffices to consider only stochastic adversaries where $x_t = x$ for some unknown $x \in \mathcal{K}$ for all t. The learner wants $\mathfrak{R}_n = O(n^{4/7})$ and the adversary wants $\mathfrak{R}_n = \Omega(n^{4/7})$. Clearly the learner can play all actions $\Theta(n^{4/7})$ times and use the information gained to estimate $u(x)$ and $v(x)$ to accuracy $O(n^{-2/7})$. If the adversary chose x such that $|u(x)| + |v(x)| = \Omega(n^{-2/7})$, then the learner can identify the sign of either $u(x)$ or $v(x)$ and subsequently play optimally. On the other hand, if $|u(x)| + |v(x)| = O(n^{-2/7})$, then the learner can play actions $\{5, 6, 7\}$ $O(n^{6/7})$ times and learn $u(x)$ to accuracy $O(n^{-3/7})$. Subsequently if $|u(x)| = \Omega(n^{-3/7})$, then the learner has enough information to know the sign of $u(x)$ and play optimally. On the other hand, if $|u(x)| = O(n^{-3/7})$, then the learner plays either action in $\{3, 4\}$ and suffers regret $O(n^{4/7})$ as required. The adversary can force this by randomising $v(x) \in \{n^{-2/7}, -n^{-2/7}\}$ and $u(x) \in \{n^{-3/7}, -n^{-3/7}\}$. Note the power of our lower bound construction in Theorem 3. Except for subpolynomial factors we can prove a minimax adversarial regret upper bound by analysing what is essentially explore-then-commit in the stochastic setting. ∎

6. Discussion

For infinite partial monitoring games with the standard regret it was not known what rates the minimax regret could have. We show in Appendix D that for all $\lambda \in (2, \infty)$ there exists a game with $\lambda_\star = \lambda$. The Rustichini regret and the standard regret usually coincide for games where $\mathcal{Z}$ is finite and for which the standard minimax regret is not linear. The cases where they do not are a subset of the so-called degenerate games. The details are explained in Appendix F.

Classification of finite Rustichini games   We believe that for finite Rustichini games

$$\lim_{n \to \infty} \frac{\log(\mathfrak{R}_n^\star)}{\log(n)} \in \{0\} \cup \{1/2\} \cup \left\{ \frac{2^i}{2^{i+1} - 1} : i \in 1, 2, 3, \dots \right\}$$

Games with this regret can be constructed by generalising the example in Section 5 to higher dimensions. Proving that this kind of construction is essentially the only case remains an interesting challenge.

Non-existence of very easy games   We now know that in the standard setting there exist games for which $\lambda_\star \in \{1\} \cup [2, \infty]$. These are the only possibilities. Games for which $\lambda_\star \in (1, 2)$ do not exist. To see why, note that if there exists an action a that is optimal irrespective of what the adversary plays, then the minimax regret vanishes and $\lambda_\star = 1$. On the other hand, if no action is optimal for all choices of the adversary, then because there are only finitely many actions there exists a subset of the outcome space such that the finite partial monitoring game with the adversary restricted to this subset is not trivial. On this restricted game the classification theorem for finite partial monitoring shows that $\lambda_\star \geq 2$ and since the unrestricted game is certainly not easier it follows that $\lambda_\star \in \{0\} \cup [2, \infty]$ for all partial monitoring games.

Infinite action sets Our analysis heavily relies on the action set being finite. One roundabout demonstration that the ideas cannot be generalised to infinite action sets without heavy modification is that for infinite action sets the adversarial and stochastic regret are no longer the same (Cesa-Bianchi et al., 2021, for example).

Other questions At the moment we do not know how to obtain high probability bounds in terms of the information ratio, which would be at least one route towards handling non-oblivious adversaries. Our analysis shows that the minimax information ratio with respect to the logarithmic barrier characterises the regret. An interesting question is whether or not the same result holds when the logarithmic barrier is replaced by the negentropy, when the information gain would correspond to the usual mutual information. The obstruction in our current analysis is essentially Eq. (5), which does not hold anymore. Even if the minimax information ratio with the negentropy does not yield a classification theorem, some version of exponential weights may still have good results. Devraj et al. (2021) hints that this may sometimes be possible. Understanding this may be important because the logarithmic barrier has a serious limitation that the diameter of its potential is linear in the dimension, which makes it poorly suited for high-dimensional structured problems.

References

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

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.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

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

N. Cesa-Bianchi, T. Cesari, R. Colomboni, F. Fusco, and S. Leonard. Bilateral trade: A regret minimization perspective. *arXiv preprint arXiv:2109.12974*, 2021.

A. Devraj, B. Van Roy, and K. Xu. A bit better? quantifying information for bandit learning. *arXiv preprint arXiv:2102.09488*, 2021.

D. Foster and A. Rakhlin. No internal regret via neighborhood watch. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 382–390, La Palma, Canary Islands, 2012. JMLR.org.

D. Foster, S. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

J. Kirschner, T. Lattimore, C. Vernade, and Cs. Szepesvári. Asymptotically optimal information-directed sampling. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2777–2821. PMLR, 2021.

J. Kwon and V. Perchet. Online learning and blackwell approachability with partial monitoring: optimal convergence rates. In *Artificial Intelligence and Statistics*, pages 604–613. PMLR, 2017.

T. Lattimore and A. György. Mirror descent and the information ratio. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2021.

T. Lattimore and Cs. Szepesvári. Cleaning up the neighbourhood: A full classification for adversarial partial monitoring. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, 2019a.

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, 2019b. PMLR.

T. Lattimore and Cs. Szepesvári. Exploration by optimisation in partial monitoring. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2020.

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

G. Lugosi, S. Mannor, and G. Stoltz. Strategies for prediction under imperfect monitoring. *Mathematics of Operations Research*, 33(3):513–528, 2008.

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

S. Mannor, V. Perchet, and G. Stoltz. Set-valued approachability and online learning with partial monitoring. *The Journal of Machine Learning Research*, 15(1):3247–3295, 2014.

V. Perchet. Internal regret with partial monitoring: Calibration-based optimal algorithms. *Journal of Machine Learning Research*, 12(6), 2011a.

V. Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677, 2011b.

A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Computational Learning Theory*, pages 208–223. Springer, 2001.

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.

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

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.

Appendix A. Polyhedral complexes and triangulations

Here we collect some definitions and basic facts about polyhedral complexes and the existence of triangulations. Throughout the section, $\mathcal{K}$ is a compact convex subset of $\mathbb{R}^d$. Given $x, y \in \mathbb{R}^d$, let $[x, y] = \{(1 - \alpha)x + \alpha y : \alpha \in [0, 1]\}$ denote the closed chord connecting x and y and $(x, y) = \{(1 - \alpha)x + \alpha y : \alpha \in (0, 1)\}$ the open chord. Given a polytope P, a set $F \subset P$ is a face of P if for all $x, y \in P$ with $(x, y) \cap F \neq \emptyset$, $[x, y] \subset F$. Note that P and the empty set are faces of P. Remember that the set of all faces is denoted by $\mathrm{faces}(P)$.

Definition 10 *A convex polyhedral complex on $\mathcal{K}$ is a finite set of closed convex polytopes $\mathcal{P}$ such that:*
  (a) $\mathrm{faces}(P) \subset \mathcal{P}$ *for all $P \in \mathcal{P}$; and*
  (b) $P \cap Q \in \mathcal{P}$ *for all $P, Q \in \mathcal{P}$; and*
  (c) $\mathcal{K} = \cup_{P \in \mathcal{P}} P$.

Fact 11 *Given any $x \in \mathcal{K}$ and convex polyhedral complex $\mathcal{P}$ on $\mathcal{K}$, there exists a unique $P \in \mathcal{P}$ such that $x \in \mathrm{ri}(P)$.*

A polyhedral complex $\mathcal{P}$ is called a simplicial complex if P is a simplex for all $P \in \mathcal{P}$.

Fact 12 *Given any continuous piecewise linear function $f : \mathcal{K} \to \mathbb{R}^s$, there exists a simplicial complex $\mathcal{P}$ on $\mathcal{K}$ such that the restriction $f_{|P}$ is linear for all $P \in \mathcal{P}$.*

Appendix B. Technical lemmas

Proof [LEMMA 1] We start by proving the claims in the standard setting, which means that that $\mathcal{V}$ and $\mathcal{V}_\star$ are defined by Eq. (1). Since $x \mapsto \mathcal{L}(\pi, x) = \int_{\mathcal{Z}} \mathcal{L}(\pi, z) \, dx(z)$ is linear, it follows that $\mathcal{V}_\star$ is the minimum of k linear functions and hence is concave and piecewise linear with at most k pieces. Parts (b) and (c) in the standard setting are trivial, since $\mathcal{V}$ is bilinear. Moving on to the Rustichini setting. Since $\mathcal{Z}$ is finite, $\mathcal{K}$ corresponds to a probability simplex over $\mathcal{Z}$. Since $\mathcal{Z}$ is finite we assume without loss of generality that Σ is finite and $\Sigma = [|\Sigma|]$. Notice that $\mathcal{V}_\star$ can be written as a linear program and its dual is

$$\mathcal{V}_\star(x) = \min_{\pi \geq \mathbf{0}} \min_\lambda \langle \lambda, Sx \rangle \quad \text{subject to } \pi^\top \mathcal{L} - S^\top \lambda \leq \mathbf{0} \text{ and } \langle \pi, \mathbf{1} \rangle = 1 \,.$$

where $S \in \{0, 1\}^{a|\Sigma| \times d}$ is the matrix with $S_{a\sigma, z} = \mathbf{1}(\mathcal{S}(a, z) = \sigma)$. And again, $\mathcal{V}_\star$ can be written as the minimum of finitely many linear functions. The results for $\mathcal{V}$ follow along the same lines. Let $\pi \in \Delta_k$ be fixed and consider the mapping $x \mapsto \mathcal{V}(\pi, x)$, which by duality satisfies

$$\mathcal{V}(\pi, x) = \min_\lambda \langle \lambda, Sx \rangle \quad \text{subject to } \pi^\top \mathcal{L} - S^\top \lambda \leq \mathbf{0} \text{ and } \langle \pi, \mathbf{1} \rangle = 1 \,.$$

As above, this is the minimum of finitely many linear functions and hence concave and piecewise linear. For the last part, let $x \in \mathcal{K}$ be fixed, then

$$\mathcal{V}(\pi, x) = \max_{y : y \mathrm{R} x} \mathcal{L}(\pi, y) \,.$$

15

Since $\{y \in \mathcal{K} : y\mathrm{R}x\}$ is a polytope, the maximum is achieved on an extreme point of this polytope of which there are finitely many. Hence $\pi \mapsto \mathcal{V}(\pi, x)$ is the maximum of finitely many linear functions and hence is convex and piecewise linear. ∎

Proof [LEMMA 5] Let ρ be a policy such that $\Delta_\star(\mu) = \Delta(\rho, \mu)$, which exists by Lemma 14. Then,

$$\Psi_\star^\lambda(\mu) \leq \min_{p \in [0,1]} \frac{(p\Delta(\pi, \mu) + (1-p)\Delta(\rho, \mu))^\lambda}{p\mathrm{I}(\pi, \mu)} \leq \frac{2^\lambda \Delta_\star(\mu)^{\lambda-1}\Delta(\pi, \mu)}{\mathrm{I}(\pi, \mu)},$$

where the first inequality follows from convexity of $\pi \mapsto \Delta(\pi, \mu)$ and because $\mathrm{I}(p\pi + (1-p)\rho, \mu) \geq p\mathrm{I}(\pi, \mu)$. The second inequality follows by choosing $p = \Delta(\rho, \mu)/\Delta(\pi, \mu)$, which is in $[0, 1]$ by the definition of ρ. ∎

Proof [LEMMA 7] Let $A \in \mathcal{F}_t$ and $\tau = \min\{s \leq t : \sum_{u=1}^{s+1} \mathrm{I}(\pi_u, \mu) \geq \epsilon\}$ with the minimum of the empty set defined to be $t + 1$. Because $\mathrm{I}(\pi_{u+1}, \mu)$ is $\mathcal{F}_u$-measurable, τ is a stopping time adapted to $(\mathcal{F}_s)_{s=1}^t$. Therefore,

$$\mathbb{P}_x^t(A) \leq \mathbb{P}_x^t(\tau < t) + \mathbb{P}_x^t(A \cap \{\tau \geq t\})$$

$$\leq \mathbb{P}^t(\tau < t) + 2\sqrt{\frac{\epsilon}{2}} + \mathbb{P}^t(A \cap \{\tau \geq t\})$$

$$\leq \mathbb{P}^t(\tau < t) + 2\sqrt{\frac{\epsilon}{2}} + \mathbb{P}^t(A)$$

$$= \mathbb{P}^t\left(\sum_{s=1}^t \mathrm{I}(\pi_s, \mu) \geq \epsilon\right) + \sqrt{2\epsilon} + \mathbb{P}^t(A),$$

where in the second inequality we used Pinsker's inequality and Eq. (5) and the fact that both $\{\tau < t\} \in \mathcal{F}_\tau$ and $A \cap \{\tau \geq t\} \in \mathcal{F}_\tau$. The other direction follows along the same lines. ∎

Proof [LEMMA 8] Let $\pi \in \mathcal{N}(y)$. Since $x, y \in \mathrm{ri}(F)$, there exists an $\epsilon > 0$ such that $y + \epsilon(y - x) \in F$. Since $\mathcal{V}_\star$ is linear on F, $z \mapsto \Delta(\pi, z)$ is concave on F. Therefore,

$$0 \leq \Delta(\pi, y + \epsilon(y - x)) \leq \epsilon(\Delta(\pi, y) - \Delta(\pi, x)) = -\epsilon\Delta(\pi, x),$$

which implies that $\Delta(\pi, x) = 0$. Therefore $\mathcal{N}(x) \subset \mathcal{N}(y)$. A symmetric argument shows that $\mathcal{N}(y) \subset \mathcal{N}(x)$ and hence $\mathcal{N}(x) = \mathcal{N}(y)$. ∎

Lemma 13 *Suppose that $(C_i)_{i=1}^k$ is a finite collection of compact sets in $\mathbb{R}^d$ with $\bigcap_{i=1}^k C_i = \emptyset$. Then there exists an $\epsilon > 0$ such that*

$$\inf_{x \in \mathbb{R}^d} \max_{i \in [k]} \mathrm{dist}_1(x, C_i) > \epsilon.$$

Proof Suppose not. There exists a sequence $(x_t)_{t=1}^\infty$ such that $\lim_{n \to \infty} \max_{i \in [k]} \mathrm{dist}_1(x_t, C_i) = 0$ for all $i \in [k]$. Clearly the sequence (x_t) is bounded and hence has a convergent subsequence. Let x be an accumulation point (x_t). Then, since C_i is closed, $d(x, C_i) = 0$ for all i and hence $x \in \bigcap_{i \in [k]} C_i$, which is a contradiction. ∎

Lemma 14 *For all $\lambda > 1$ and $\mu \in \mathscr{P}(\mathcal{K})$ there exists a policy $\pi \in \Delta_k$ such that $\Psi^\lambda(\pi, \mu) = \Psi^\lambda_\star(\mu)$.*

Proof There is nothing to prove if $\Psi^\lambda_\star(\mu) = \infty$, so assume this is not the case. If there exists a π such that $\Delta(\pi, \mu) = 0$. Then $\Psi^\lambda_\star(\mu) = \Psi^\lambda(\pi, \mu) = 0$. Suppose now that $\Delta(\pi, \mu) > 0$ for all π. Since $\pi \mapsto \Delta(\pi, \mu)$ is continuous and Δ_k is compact, $\min_{\pi \in \Delta_k} \Delta(\pi, \mu) > 0$. Let $\pi_1, \pi_2, \ldots$ be a sequence of policies with $\lim_{t \to \infty} \Psi^\lambda(\pi, \mu) = \Psi^\lambda_\star(\mu)$. By compactness this sequence has a convergent subsequence converging to some policy $\pi_\star$. Clearly there exists an $a \in \mathrm{spt}(\pi)$ with $\mathrm{KL}(\mathcal{S}(a, \mu), \mathcal{S}(a, x)) > 0$. Therefore $\pi \mapsto \Psi^\lambda(\pi, \mu)$ is continuous in a neighbourhood of $\pi_\star$ and the result follows. $\blacksquare$

Lemma 15 *Suppose that $F \subset \mathbb{R}^d$ is a polytope. Then $x \mapsto \mathrm{dist}_1(x, F)$ is piecewise linear.*

Proof Suppose that $F = \{x : Ax \le b\}$ for suitable $A \in \mathbb{R}^{s \times d}$ and $b \in \mathbb{R}^s$ and let $S \in \mathbb{R}^{2^d \times d}$ be the matrix with rows formed from all possible elements of $\{-1, 1\}^d$. Then,

$$
\begin{aligned}
\mathrm{dist}_1(x, F) &= \min_{\substack{y \in \mathbb{R}^d \\ z \in \mathbb{R}}} \left\{ z : Ay \le b, S(y - x) \le z\mathbf{1} \right\} \\
&= \max_{\substack{\mathbf{0} \le \lambda \in \mathbb{R}^s \\ \mathbf{0} \le \varphi \in \mathbb{R}^{2^d}}} \left\{ \langle \varphi, Sx \rangle - \langle \lambda, b \rangle : A^\top \lambda = S^\top \varphi, \langle \varphi, \mathbf{1} \rangle = 1 \right\}, \qquad \text{(duality)}
\end{aligned}
$$

which can be written as the maximum of finitely many linear functions and hence is convex and piecewise linear. $\blacksquare$

Appendix C. Proof of Lemma 6 [RUSTICHINI SETTING]

The mission in this section is to prove Lemma 6 in the Rustichini setting. The notation in Section 4 will be used, so that section should be read first.

Lemma 16 *There exists a game-dependent constant Lip_Δ such that for all nonempty $F \in \mathcal{F}$,*

$$
\max_{\pi \in \mathcal{N}(F)} \Delta(\pi, x) \le \mathrm{Lip}_\Delta \, \mathrm{dist}_1(x, F) .
$$

Proof Let $F \in \mathcal{F}$ be nonempty and recall from Section 4 that $\mathcal{N}(F)$ is a polytope. By Lemma 1, $\pi \mapsto \Delta(\pi, x)$ is convex for all $x \in \mathcal{K}$ and $x \mapsto \Delta(\pi, x)$ is piecewise linear for all $\pi \in \Delta_k$. Since $\mathcal{K}$ is compact, there can be only finitely many pieces. Convex functions on polytopes are maximised on their extreme points and so

$$
\max_{\pi \in \mathcal{N}(F)} \Delta(\pi, x) = \max_{\pi \in \mathrm{ext}(\mathcal{N}(F))} \Delta(\pi, x) .
$$

Since $x \mapsto \Delta(\pi, x)$ is piecewise linear, for any $\pi \in \Delta_k$ there exists a constant $c_\pi \ge 0$ such that $x \mapsto \Delta(\pi, x)$ is c_π-Lipschitz with respect to $\| \cdot \|_1$. Let $y \in F$ be such that $\mathrm{dist}_1(x, F) = \|x - y\|_1$, which exists because F is closed. Therefore,

$$
\max_{\pi \in \mathrm{ext}(\mathcal{N}(F))} \Delta(\pi, x) \le \max_{\pi \in \mathrm{ext}(\mathcal{N}(F))} \Delta(\pi, y) + c_\pi \|x - y\|_1 = \mathrm{dist}_1(x, F) \max_{\pi \in \mathrm{ext}(\mathcal{N}(F))} c_\pi ,
$$

where we used the fact that for $\pi \in \text{ext}(\mathcal{N}(F))$ and $y \in F$, $\Delta(\pi, y) = 0$. The result follows with $\text{Lip}_\Delta = \max_{F \in \mathcal{F}: F \neq \emptyset} \max_{\pi \in \text{ext}(\mathcal{N}(F))} c_\pi$, which is finite because $\mathcal{F}$ is finite and $\text{ext}(\mathcal{N}(F))$ is finite for all $F \in \mathcal{F}$. ∎

Lemma 17 *Let $\epsilon > 0$, $\mathcal{G} \subset \mathcal{F}$ and $\Pi = \{\pi : \text{dist}_1(\pi, \mathcal{N}(F)) \geq \epsilon \text{ for all } F \in \mathcal{G}\}$ and define $f, g : \mathcal{K} \to \mathbb{R}$ by*

$$f(x) = \min_{\pi \in \Pi} \Delta(\pi, x) \qquad\qquad g(x) = \text{dist}_1\left(x, \cup_{F \in \mathcal{F} \backslash \mathcal{G}} F\right).$$

Then there exists a strictly positive constant const_ϵ depending on the game and ϵ such that

$$f(x) \geq \text{const}_\epsilon \, g(x).$$

Proof If $\mathcal{G}$ is empty, then $g(x) = 0$ for all x and the result is immediate because f is non-negative. Assume for the remainder that $\mathcal{G}$ is nonempty. Note that Π is can be written as a union of convex polytopes. Therefore,

$$f(x) = \min_{\pi \in \Pi} \max_{y: y \mathbf{R} x} \mathcal{L}(\pi, y) - \mathcal{V}_\star(x),$$

is the difference between two concave piecewise linear functions and hence is piecewise linear. The function g is also piecewise linear because

$$g(x) = \text{dist}_1(x, \cup_{F \in \mathcal{F} \backslash \mathcal{G}} F) = \min_{F \in \mathcal{F} \backslash \mathcal{G}} \text{dist}_1(x, F),$$

which is the minimum of finitely many piecewise linear functions (Lemma 15) and hence piecewise linear. Suppose that $F \in \mathcal{G}$. By definition, if $x \in \text{ri}(F)$, then $\Delta(\pi, x) = 0$ if and only if $\pi \in \mathcal{N}(F)$. Hence, because $\text{dist}_1(\pi, \mathcal{N}(F)) \geq \epsilon$ for all $\pi \in \Pi$, $f(x) > 0$ for all $x \in \text{ri}(F)$. Let $\mathcal{T}$ be a simplicial complex of $\mathcal{K}$ such that the restrictions $f_{|T}$ and $g_{|T}$ are linear for all $T \in \mathcal{T}$, which exists by Fact 12. Given $T \in \mathcal{T}$ and $t \in \text{ext}(T)$, let $F \in \mathcal{F}$ be such that $t \in \text{ri}(F)$, which is unique by Fact 11. Suppose that $g(t) > 0$. Then $F \in \mathcal{G}$, since otherwise $g(t) = \text{dist}_1(t, \cup_{F \in \mathcal{F} \backslash \mathcal{G}} F) = 0$. Therefore $f(t) > 0$. Hence,

$$g(t) > 0 \implies f(t) > 0 \text{ for all } t \in \text{ext}(T). \tag{7}$$

Let $x \in T$ and $p \in \mathscr{P}(\text{ext}(T))$ be a distribution such that $x = \sum_{t \in \text{ext}(T)} p(t) t$. Combining Eq. (7) with the facts that both f and g are linear on T and non-negative,

$$g(x) = \sum_{t \in \text{ext}(T): g(t) > 0} p(t) g(t) \qquad\qquad (g_{|T} \text{ linear, } g \geq 0)$$

$$= \sum_{t \in \text{ext}(T): g(t) > 0} p(t) f(t) \frac{g(t)}{f(t)} \qquad\qquad (\text{by Eq. (7)})$$

$$\leq \left(\max_{t \in \text{ext}(T): g(t) > 0} \frac{g(t)}{f(t)}\right) \sum_{t \in \text{ext}(T): g(t) > 0} p(t) f(t) \qquad\qquad (f, g \geq 0)$$

$$\leq \left(\max_{t \in \text{ext}(T): g(t) > 0} \frac{g(t)}{f(t)}\right) \sum_{t \in \text{ext}(T)} p(t) f(t) \qquad\qquad (f \geq 0)$$

$$= \left(\max_{t \in \text{ext}(T): g(t) > 0} \frac{g(t)}{f(t)}\right) f(x). \qquad\qquad (f_{|T} \text{ linear})$$

Rearranging shows that

$$f(x) \geq \left(\min_{t \in \text{ext}(T): g(t) > 0} \frac{f(t)}{g(t)} \right) g(x) .$$

The minimum is always positive for any T by Eq. (7). The result follows since there are only finitely many $T \in \mathcal{T}$. ∎

Proof [LEMMA 6, RUSTICHINI SETTING] We proceed in a few steps.

Step 1: Separation arguments Let $\epsilon > 0$ be a constant such that for any finite index set $\mathcal{I}$ and set $(\mathcal{F}_\alpha)_{\alpha \in \mathcal{I}}$ of subsets of $\mathcal{F}$ such that whenever

$$\bigcap_{\alpha \in \mathcal{I}} \bigcup_{F \in \mathcal{F}_\alpha} \mathcal{N}(F) = \emptyset ,$$

it holds that

$$\min_{\pi \in \Delta_k} \max_{\alpha \in \mathcal{I}} \text{dist}_1 \left(\pi, \bigcup_{F \in \mathcal{F}_\alpha} \mathcal{N}(F) \right) > \epsilon .$$

Such a constant exists by Lemma 13 and because there are only finitely many subsets of $\mathcal{F}$ and because $\mathcal{N}(F)$ is closed for all $F \in \mathcal{F}$ and the union of finitely many closed sets is closed.

Step 2: Separation arguments The claim of the lemma is trivial if $\Delta_\star(\mu) = 0$, so for the remainder assume that $\Delta_\star(\mu) > 0$. For $x \in \text{spt}(\mu)$ let $\mathcal{F}_x = \{ F \in \mathcal{F} : \text{dist}_1(x, F) < \Delta_\star(\mu) / \text{Lip}_\Delta \}$, where Lip_Δ is the constant from Lemma 16. Suppose that

$$\bigcap_{x \in \text{spt}(\mu)} \bigcup_{F \in \mathcal{F}_x} \mathcal{N}(F) \neq \emptyset .$$

Then there exists a policy π such that for all $x \in \text{spt}(\mu)$, $\pi \in \cup_{F \in \mathcal{F}_x} \mathcal{N}(F)$. By Lemma 16, for this policy it holds for all $x \in \text{spt}(\mu)$ that

$$\Delta(\pi, x) \leq \text{Lip}_\Delta \max_{F \in \mathcal{F}_x} \text{dist}_1(x, F) < \Delta_\star(\mu) ,$$

which contradicts the definition of $\Delta_\star(\mu)$ as the minimum of $\pi \mapsto \Delta(\pi, \mu)$. Therefore,

$$\bigcap_{x \in \text{spt}(\mu)} \bigcup_{F \in \mathcal{F}_x} \mathcal{N}(F) = \emptyset .$$

By the definition of ϵ in the first step,

$$\min_{\pi \in \Delta_k} \max_{x \in \text{spt}(\mu)} \text{dist}_1 \left(\pi, \bigcup_{F \in \mathcal{F}_x} \mathcal{N}(F) \right) > \epsilon .$$

Step 3: Perturbations Let $\varphi = \epsilon/(2k)$ and $(\pi_x)_{x \in \mathrm{spt}(\mu)}$ be a collection of policies and π be a policy such that $\|\pi - \pi_x\|_\infty \leq \varphi$. Then $\|\pi - \pi_x\|_1 \leq \epsilon/2$ for all $x \in \mathrm{spt}(\mu)$. By the previous step there exists an $x \in \mathrm{spt}(\mu)$ such that

$$\mathrm{dist}_1 \left(\pi, \bigcup_{F \in \mathcal{F}_x} \mathcal{N}(F) \right) > \epsilon \,.$$

Therefore, by the triangle inequality,

$$\mathrm{dist}_1 \left(\pi_x, \bigcup_{F \in \mathcal{F}_x} \mathcal{N}(F) \right) > \epsilon/2 \,.$$

By Lemma 17,

$$\begin{aligned}
\Delta(\pi_x, x) &\geq \mathrm{const}_{\epsilon/2, \mathcal{F}_x} \, \mathrm{dist}_1 \left(x, \bigcup_{F \in \mathcal{F} \setminus \mathcal{F}_x} F \right) \\
&= \mathrm{const}_{\epsilon/2, \mathcal{F}_x} \min_{F \in \mathcal{F} \setminus \mathcal{F}_x} \mathrm{dist}_1(x, F) \\
&\geq \frac{\mathrm{const}_{\epsilon/2, \mathcal{F}_x}}{\mathrm{Lip}_\Delta} \Delta_\star(\mu) \,,
\end{aligned}$$

where in the final inequality we used the definition of $\mathcal{F}_x$. The result follows because there are only finitely many possibles values of $\mathcal{F}_x$. ∎

Appendix D. Exotic standard games

Many classical partial monitoring problems have infinite latent spaces. Even the most elementary prediction with expert advice problem, for example, has infinite latent spaces because the losses are allowed to take real values. Theorem 3 shows that the information ratio characterises the regret for games like this. More interestingly, the information ratio also characterises the regret for more exotic games. Here we show that in contrast to finite partial monitoring games, when the latent space is infinite, then the minimax regret can have a continuum of rates.

Theorem 18 *For any $q \in (1/2, 1)$ there exists a partial monitoring game for which*

$$\lim_{n \to \infty} \frac{\log(\mathfrak{R}_n^\star)}{\log(n)} = q \,.$$

Note that for $q \in \{0, 1/2, 1\}$ the above is already known even for finite partial monitoring games.

Proof The construction we employ is based on a lower bound for finite partial monitoring by Lattimore and Szepesvári (2019a) showing that the size of the signal set can play a role in the minimax regret for some games.

Step 1: Constructing the game Let $p = 2 - 2q \in (0, 1)$. For $\alpha \in \mathbb{N}$ define a game $\mathcal{G}_\alpha$ with $k = 2$ actions and $2\alpha - 1$ outcomes with loss and signal matrices given by

$$\mathcal{L}_\alpha = \begin{bmatrix} 0 & \alpha^{-p} & 0 & \cdots & \alpha^{-p} & 0 \\ \alpha^{-p} & 0 & \alpha^{-p} & \cdots & 0 & \alpha^{-p} \end{bmatrix} \quad \mathbb{S}_\alpha = \begin{bmatrix} 1 & 2 & 2 & 3 & 3 & \cdots & \alpha & \alpha \\ 1 & 1 & 2 & 2 & 3 & \cdots & \alpha - 1 & \alpha \end{bmatrix}.$$

This is the same game used by Lattimore and Szepesvári (2019a) except that the losses have been scaled by a factor of α^{-p}. Let $g_\alpha(a, \sigma) = \alpha^{-p}(a + 2(\sigma - 1))^{3-2a}$, which has the property that for all $i \in [2\alpha - 1]$,

$$\sum_{a \in [k]} g_\alpha(a, \mathbb{S}_\alpha(a, i)) = \mathcal{L}_\alpha(1, i) - \mathcal{L}_\alpha(2, i) = \alpha^{-p}(-1)^i.$$

In the language of finite partial monitoring, g_α is an estimation function that can be used to estimate the differences in losses between actions using only the observed signals. Note that $\|g_\alpha\|_\infty \leq 2\alpha^{1-p}$. Next, let $\mathcal{G}$ be the infinite game where the adversary chooses an interleaving of outcomes from games $(\mathcal{G}_\alpha)_{\alpha=1}^\infty$ and the learner gets to observe the signal corresponding to its action and the game/outcome chosen by the adversary and also the identity of the game. Formally,

$$\mathcal{Z} = \cup_{\alpha \in \mathbb{N}}\{(\alpha, i) : i \in [2\alpha - 1]\}$$
$$\Sigma = \cup_{\alpha \in \mathbb{N}}\{(\alpha, i) : i \in [\alpha]\}$$
$$\mathcal{L}(a, (\alpha, i)) = \mathcal{L}_\alpha(a, i)$$
$$\mathbb{S}(a, (\alpha, i)) = (\alpha, \mathbb{S}_\alpha(a, i)).$$

Step 2: Lower bound Let $\mathfrak{R}_n^\star(\mathcal{G}_\alpha)$ denote the minimax regret for the game $\mathcal{G}_\alpha$. Lattimore and Szepesvári (2019a) showed that for any α,

$$\mathfrak{R}_n^\star(\mathcal{G}_\alpha) \geq \text{const} \min(\alpha^{1-p}\sqrt{n}, n\alpha^{-p}),$$

where $\text{const} > 0$ is a universal constant. Obviously the adversary can choose to limit its choices to any particular game and hence

$$\mathfrak{R}_n^\star(\mathcal{G}) \geq \text{const} \max_{\alpha \in \mathbb{N}} \min(\alpha^{1-p}\sqrt{n}, n\alpha^{-p})).$$

By optimising $\alpha = \lceil\sqrt{n}\rceil$ and passing to the limit we conclude that

$$\lim_{n\to\infty} \frac{\log(\mathfrak{R}_n^\star(\mathcal{G}))}{\log(n)} \geq 1 - p/2 = q.$$

Step 3: Upper bound To obtain the upper bound we control the generalised information ratio. Let $g(a, (\alpha, \sigma)) = g_\alpha(a, \sigma)$ where g_α is the function defined in the first step. By the definitions, we have the following:

    (a) $|g(a, (\alpha, \sigma))| \leq 2\alpha^{1-p}$ for all actions $a \in [k]$ and $(\alpha, \sigma) \in \Sigma$.

    (b) $\sum_{a \in [k]} g(a, \mathbb{S}(a, z)) = \mathcal{L}(2, z) - \mathcal{L}(1, z)$ for all $z \in \mathcal{Z}$.

Because there are only two actions it suffices to bound the information ratio for priors $\mu \in \mathscr{P}_2(\mathcal{K})$ so let $x, y \in \mathcal{K}$ be arbitrary and $\mu(x) = s$ and $\mu(y) = 1 - s$ for some $s \in [0, 1]$. We need to show that

$$\Psi_\star^{2/p}(\mu) = \min_{\pi \in \Delta_k} \frac{\Delta(\pi, \mu)^{2/p}}{\mathrm{I}(\pi, \mu)} \leq \text{const},$$

21

where const is a universal constant that is independent of μ. Let $\pi \in \Delta_k$ be the distribution with $\pi(1) = s$ and $\pi(2) = 1 - s$ and $\mathcal{Z}_{<\beta} = \{(\alpha, z) \in \mathcal{Z} : \alpha < \beta\}$. Then, letting $u = sx + (1 - s)y$, which satisfies $\mathcal{S}(a, u) = \mathcal{S}(a, \mu)$,

$$
\begin{aligned}
\Delta(\pi, \mu) &= s\left(\mathcal{L}(\pi, x) - \mathcal{L}(1, x)\right) + (1 - s)\left(\mathcal{L}(\pi, y) - \mathcal{L}(2, y)\right) \\
&= s(1 - s)\left(\mathcal{L}(2, x) - \mathcal{L}(1, x)\right) + s(1 - s)\left(\mathcal{L}(1, y) - \mathcal{L}(2, y)\right) \\
&= s(1 - s) \int_{\mathcal{Z}} \left(\mathcal{L}(2, z) - \mathcal{L}(1, z)\right) x(\mathrm{d}z) + s(1 - s) \int_{\mathcal{Z}} \left(\mathcal{L}(1, y) - \mathcal{L}(2, y)\right) y(\mathrm{d}z) \\
&\leq \frac{\beta^{-p}}{2} + s(1 - s) \int_{\mathcal{Z}_{<\beta}} \left(\mathcal{L}(2, z) - \mathcal{L}(1, z)\right) x(\mathrm{d}z) + s(1 - s) \int_{\mathcal{Z}_{<\beta}} \left(\mathcal{L}(1, y) - \mathcal{L}(2, y)\right) y(\mathrm{d}z) \\
&= \frac{\beta^{-p}}{2} + s(1 - s) \int_{\mathcal{Z}_{<\beta}} \sum_{a \in [k]} g(a, \mathcal{S}(a, z)) x(\mathrm{d}z) - s(1 - s) \int_{\mathcal{Z}_{<\beta}} \sum_{a \in [k]} g(a, \mathcal{S}(a, z)) y(\mathrm{d}z) \\
&= \frac{\beta^{-p}}{2} + s(1 - s) \int_{\mathcal{Z}_{<\beta}} \sum_{a \in [k]} g(a, \mathcal{S}(a, z)) \left((x - u) + (u - y)\right)(\mathrm{d}z) \\
&\leq \frac{\beta^{-p}}{2} + \beta^{1-p} s(1 - s) \sum_{a \in [k]} \left(\sqrt{\frac{\mathrm{KL}(\mathcal{S}(a, \mu), \mathcal{S}(a, x))}{2}} + \sqrt{\frac{\mathrm{KL}(\mathcal{S}(a, \mu), \mathcal{S}(a, y))}{2}}\right) \\
&\leq \frac{\beta^{-p}}{2} + \beta^{1-p} s(1 - s) \sum_{a \in [k]} \sqrt{\mathrm{KL}(\mathcal{S}(a, \mu), \mathcal{S}(a, x)) + \mathrm{KL}(\mathcal{S}(a, \mu), \mathcal{S}(a, y))} \\
&= \frac{\beta^{-p}}{2} + \beta^{1-p} s(1 - s) \sum_{a \in [k]} \sqrt{\mathrm{I}(a, \mu)} \\
&\leq \frac{\beta^{-p}}{2} + \beta^{1-p} \sqrt{\mathrm{I}(\pi, \mu)} \,.
\end{aligned}
$$

By choosing $\beta = \left\lceil \sqrt{1/\mathrm{I}(\pi, \mu)} \right\rceil$, it follows that

$$
\Delta(\pi, \mu) \leq \mathrm{const}\, \mathrm{I}(\pi, \mu)^{p/2} \,.
$$

Therefore, $\Psi_\star^{2/p}(\mu) \leq \mathrm{const}$ and

$$
\mathfrak{R}_n^\star \leq \mathrm{const}\, n^{1-p/2} \log(n)^{p/2} \,.
$$

The result follows by passing to the limit. ∎

Appendix E. Proof of Theorem 4

Our proof combines and mirrors the proofs of Theorems 8 and 9 of Lattimore and György (2021). There are only two novel aspects. One is that we use convexity of $\pi \mapsto \mathcal{V}(\pi, x)$ and concavity of $x \mapsto \mathcal{V}(\pi, x)$, which in the previous analysis were always linear. The second (marginally) novel part is to relate our definition of the information gain to the Bregman divergences used by Lattimore and György (2021), which is really just a moderately tricky application of Bayes' law. Recall that $F(p) = -\sum_{\alpha=1}^m \log(p_\alpha)$ is the logarithmic barrier on $\mathcal{D} = \Delta_m \cap [1/n, 1]^m$.

Step 1: Bounding the regret by the stability   We start by bounding the regret in terms of $\Lambda_{q\eta}^\star$. Using the definition of the regret and $\Lambda_{q\eta}^\star$ and letting $\epsilon_{\mathcal{D}} = \frac{1}{n} \max_{x \in \mathcal{K}} \|\mathcal{V}^m(x)\|_\infty$,

$$
\begin{aligned}
\mathfrak{R}_n &= \mathbb{E}\left[\sum_{t=1}^n \mathcal{V}(\pi_t, x_t) - n \mathcal{V}_\star\left(\frac{1}{n}\sum_{t=1}^n x_t\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^n \mathcal{V}(\pi_t, x_t) - n \min_{p \in \Delta_m}\left\langle p, \mathcal{V}^m\left(\frac{1}{n}\sum_{t=1}^n x_t\right)\right\rangle\right] \\
&= \max_{p \in \Delta_m} \mathbb{E}\left[\sum_{t=1}^n \mathcal{V}(\pi_t, x_t) - \langle p, \mathcal{V}^m(x_t)\rangle\right] \\
&\le \max_{p \in \mathcal{D}} \mathbb{E}\left[\sum_{t=1}^n \mathcal{V}(\pi_t, x_t) - \langle p, \mathcal{V}^m(x_t)\rangle\right] + 2n\epsilon_{\mathcal{D}}\,.
\end{aligned}
$$

By the standard analysis of mirror descent (Lattimore and Szepesvári, 2020, Theorem 28.4) with $\mathrm{diam}(\mathcal{D}) = \max_{x,y \in \mathcal{D}} F(x) - F(y)$ and $p \in \mathcal{D}$, then

$$
\begin{aligned}
\sum_{t=1}^n \left\langle q_t - p, \frac{\hat{v}_t(a_t, \sigma_t)}{\eta}\right\rangle &\le \frac{\mathrm{diam}(\mathcal{D})}{\eta} + \frac{1}{\eta}\sum_{t=1}^n \mathrm{D}_\star\left(\nabla F(q_t) - \frac{\eta \hat{v}(a_t, \sigma_t)}{\pi_t(a)}, \nabla F(q_t)\right) \\
&\le \frac{m\log(n/m)}{\eta} + \frac{1}{\eta}\sum_{t=1}^n \mathrm{D}_\star\left(\nabla F(q_t) - \frac{\eta \hat{v}(a_t, \sigma_t)}{\pi_t(a)}, \nabla F(q_t)\right)\,.
\end{aligned}
$$

Therefore, by the definition of q_t, $\hat{v}_t$, π_t and $\Lambda_{q_t\eta}^\star$, the regret of Algorithm 1 is bounded by

$$
\mathfrak{R}_n \le n\epsilon + n\epsilon_{\mathcal{D}} + \frac{m\log(n/m)}{\eta} + \mathbb{E}\left[\sum_{t=1}^n \Lambda_{q_t\eta}^\star\right]\,. \tag{8}
$$

Step 2: Minimax duality   On the path to relating the information ratio and $\Lambda_{q\eta}^\star$ we need to apply minimax duality to the optimisation problem that defines $\Lambda_{q\eta}^\star$. Let $\mathcal{W} = \mathscr{P}(\mathcal{K} \times \mathcal{D})$. Given a distribution $\xi \in \mathcal{W}$ and $\pi \in \mathrm{ri}(\Delta_k)$, let $\mathbb{P}_\xi^\pi$ be a probability measure on some measurable space carrying random elements $X, P, A, Z, M, S \in \mathcal{K} \times \mathcal{D} \times [k] \times \mathcal{Z} \times [m] \times \Sigma$ where $(\mathbb{P}_\xi^\pi)_{X,P,A} = \xi \otimes \pi$ and $(\mathbb{P}_\xi^\pi)_{M|X,P} = P$ a.s. and $(\mathbb{P}_\xi^\pi)_{Z|X,P} = X$ a.s. and $S = \mathcal{S}(A, Z)$ and $M \perp A$ and $Z \perp A$. The expectation operator with respect to $\mathbb{P}_\xi^\pi$ is $\mathbb{E}_\xi^\pi$. Let $q \in \mathcal{D}$ and $\eta > 0$ and $\Delta_{k,\delta} = \Delta_k \cap [\delta, 1]^k$. We claim that by repeating the proof of Theorem 9 in the paper by Lattimore and György (2021),

$$
\Lambda_{q\eta}^\star \le \inf_{\delta > 0}\sup_{\xi \in \mathcal{W}}\min_{\pi \in \Delta_{k,\delta}} \mathbb{E}_\xi^\pi\left[\mathcal{V}(\pi, X) - \langle P, \mathcal{V}^m(X)\rangle - \frac{1}{\eta}\mathrm{D}\left(\mathbb{E}[P|A, S], \mathbb{E}[P]\right)\right]\,. \tag{9}
$$

For completeness we summarise the argument here, which is slightly easier when adapted to our setting because F and its gradients are bounded on $\mathcal{D}$. Remember that $\mathcal{E}$ is the space of functions from $[k] \times \Sigma$ to $\mathbb{R}^m$. Let $\mathcal{E}_b \subset \mathcal{E}$ be the set of all $\hat{v} \in \mathcal{E}$ with $\|\hat{v}(a, s)\|_\infty \le 2n/\eta$ for all $a \in [k]$ and $s \in \Sigma$. The set $\mathcal{E}_b$ is compact with respect to the product topology by Tychonoff's theorem. Abbreviate $\mathcal{H}_q(x) = \mathrm{D}_\star(\nabla F(q) - x, \nabla F(q))$, which is convex. Then by Sion's minimax theorem

(see aforementioned paper for topoligical details), for any $\delta > 0$,

$$
\begin{aligned}
\Lambda_{q\eta} &= \inf_{\substack{\pi \in \mathrm{ri}(\Delta_k) \\ \hat{v} \in \mathcal{E}}} \sup_{\xi \in \mathcal{W}} \mathbb{E}_\xi^\pi \left[\mathcal{V}(\pi, X) - \langle P, \mathcal{V}^m(X) \rangle + \left\langle P - q, \frac{\hat{v}(A, S)}{\pi(A)} \right\rangle + \frac{1}{\eta} \mathcal{H}_q \left(\frac{\eta \hat{v}(A, S)}{\pi(A)} \right) \right] \\
&\leq \min_{\substack{\pi \in \Delta_{k,\delta} \\ \hat{v} \in \mathcal{E}_b}} \sup_{\xi \in \mathcal{W}} \mathbb{E}_\xi^\pi \left[\mathcal{V}(\pi, X) - \langle P, \mathcal{V}^m(X) \rangle + \left\langle P - q, \frac{\hat{v}(A, S)}{\pi(A)} \right\rangle + \frac{1}{\eta} \mathcal{H}_q \left(\frac{\eta \hat{v}(A, S)}{\pi(A)} \right) \right] \\
&= \sup_{\xi \in \mathcal{W}} \min_{\substack{\pi \in \Delta_{k,\delta} \\ \hat{v} \in \mathcal{E}_b}} \mathbb{E}_\xi^\pi \left[\mathcal{V}(\pi, X) - \langle P, \mathcal{V}^m(X) \rangle + \left\langle P - q, \frac{\hat{v}(A, S)}{\pi(A)} \right\rangle + \frac{1}{\eta} \mathcal{H}_q \left(\frac{\eta \hat{v}(A, S)}{\pi(A)} \right) \right] .
\end{aligned}
$$
(10)

Note that in the application of Sion's we are using the fact that $\pi \mapsto \mathcal{V}(\pi, x)$ is convex for all x, which is the only distinction between the argument above and those of Lattimore and György (2021), where the corresponding term was linear. Letting $\xi \in \mathcal{W}$ and $\pi \in \Delta_{k,\delta}$ be fixed and

$$
\hat{v}(a, s) = \begin{cases} \frac{\pi(a)}{\eta} \left(\nabla F(q) - \nabla F(\mathbb{E}_\xi^\pi[P | S = s, A = a]) \right) & \text{if } \mathbb{P}_\xi^\pi(S = s, A = a) > 0 \\ 0 & \text{otherwise} , \end{cases}
$$

which is chosen to minimise the bound on the right-hand side Eq. (10). Note that $\hat{v} \in \mathcal{E}_b$ follows from the fact that q and $\mathbb{E}[P | S, A]$ are in $\mathcal{D}$ and hence when $\mathbb{P}_\xi^\pi(S = s, A = a) > 0$,

$$
\|\hat{v}(a, s)\|_\infty \leq \frac{1}{\eta} \left(\|\nabla F(q)\|_\infty + \|\nabla F(\mathbb{E}[P | S = s, A = a])\|_\infty \right) \leq \frac{2n}{\eta} ,
$$

where we used also that $\nabla F(q) = -1/q$. When $\mathbb{P}_\xi^\pi(S = s, A = a) = 0$, then $\|\hat{v}(a, s)\|_\infty = 0$ by definition. Substituting the definition of $\hat{v}$ into the second two terms of Eq. (10) and repeating the calculation in Equation (6) of Lattimore and György (2021),

$$
\mathbb{E}_\xi^\pi \left[\left\langle P - q, \frac{\hat{v}(A, S)}{\pi(A)} \right\rangle + \frac{1}{\eta} \mathcal{H}_q \left(\frac{\eta \hat{v}(A, S)}{\pi(A)} \right) \right] \leq -\frac{1}{\eta} \mathbb{E}_\xi^\pi \left[D(\mathbb{E}_\xi^\pi[P | S, A], \mathbb{E}_\xi^\pi[P]) \right] .
$$

Since $\delta > 0$ was arbitrary, Eq. (9) follows by combining the above with Eq. (10).

Step 3: From Bregman divergence to information gain For $\xi \in \mathcal{W}$, let $\mu_\xi \in \mathscr{P}_m(\mathcal{K})$ be the distribution with the same law as $\mathbb{E}_\xi^\pi[X | M]$. We claim that the expected Bregman divergence in

Eq. (9) is $\mathrm{I}(\pi, \mu_\xi)$, which follows from Bayes' law:

$$
\begin{aligned}
\mathbb{E}_\xi^\pi[\mathrm{D}(\mathbb{E}_\xi^\pi[P|A,S], \mathbb{E}_\xi^\pi[P])] &= \mathbb{E}_\xi^\pi[F(\mathbb{E}[P|A,S]) - F(\mathbb{E}_\xi^\pi[P])] \\
&= \mathbb{E}\left[\sum_{\alpha=1}^{m} \log\left(\frac{\mathbb{P}_\xi^\pi(M=\alpha)}{\mathbb{P}_\xi^\pi(M=\alpha|A,S)} \right) \right] \\
&= \sum_{\alpha=1}^{m} \sum_{a=1}^{k} \pi(a) \sum_{s\in\Sigma} \mathbb{P}_\xi^\pi(S=s|A=a) \log\left(\frac{\mathbb{P}_\xi^\pi(M=\alpha)}{\mathbb{P}_\xi^\pi(M=\alpha|S=s,A=a)} \right) \\
&= \sum_{\alpha=1}^{m} \sum_{a=1}^{k} \pi(a) \sum_{s\in\Sigma} \mathbb{P}_\xi^\pi(S=s|A=a) \log\left(\frac{\mathbb{P}_\xi^\pi(M=\alpha)\mathbb{P}_\xi^\pi(S=s|A=a)}{\mathbb{P}_\xi^\pi(S=s|M=\alpha,A=a)\mathbb{P}_\xi^\pi(M=\alpha|A=a)} \right) \\
&= \sum_{\alpha=1}^{m} \sum_{a=1}^{k} \pi(a) \sum_{s\in\Sigma} \mathbb{P}_\xi^\pi(S=s|A=a) \log\left(\frac{\mathbb{P}_\xi^\pi(S=s|A=a)}{\mathbb{P}_\xi^\pi(S=s|M=\alpha,A=a)} \right) \\
&= \sum_{\alpha=1}^{m} \sum_{a=1}^{k} \pi(a) \, \mathrm{KL}(\mathbb{P}_\xi^\pi(S=\cdot|A=a), \mathbb{P}_\xi^\pi(S=\cdot|A=a,M=\alpha)) \\
&= \sum_{\alpha=1}^{m} \sum_{a=1}^{k} \pi(a) \, \mathrm{KL}(\mathcal{S}(a, \mathbb{E}_\xi^\pi[X]), \mathcal{S}(a, \mathbb{E}_\xi^\pi[X|M=\alpha])) \\
&= \mathrm{I}(\pi, \mu_\xi) \,,
\end{aligned}
$$

where in this calculation we have adopted the measure-theoretic convention that $0 \times \infty = 0$. Note that because $P \in \mathcal{D}$, $\mathbb{P}(M=\alpha) > 0$ for all $\alpha \in [m]$.

Step 4: Introducing the information ratio   Combining the pieces from the previous steps,

$$
\begin{aligned}
\Lambda_{q\eta}^\star &\leq \sup_{\xi\in\mathcal{W}} \min_{\pi\in\Delta_{k,\delta}} \mathbb{E}_\xi^\pi\left[\mathcal{V}(\pi, X) - \langle P, \mathcal{V}^m(X)\rangle - \frac{1}{\eta}\mathrm{I}(\pi, \mu_\xi) \right] \\
&= \sup_{\xi\in\mathcal{W}} \min_{\pi\in\Delta_{k,\delta}} \mathbb{E}_\xi^\pi\left[\mathbb{E}_\xi^\pi[\mathcal{V}(\pi, X)|M] - \langle P, \mathcal{V}^m(X)\rangle - \frac{1}{\eta}\mathrm{I}(\pi, \mu_\xi) \right] \\
&\leq \sup_{\xi\in\mathcal{W}} \min_{\pi\in\Delta_{k,\delta}} \mathbb{E}_\xi^\pi\left[\mathcal{V}(\pi, \mathbb{E}_\xi^\pi[X|M]) - \langle P, \mathcal{V}^m(X)\rangle - \frac{1}{\eta}\mathrm{I}(\pi, \mu_\xi) \right] \\
&\leq \sup_{\xi\in\mathcal{W}} \min_{\pi\in\Delta_{k,\delta}} \mathbb{E}_\xi^\pi\left[\mathcal{V}(\pi, \mathbb{E}_\xi^\pi[X|M]) - \mathcal{V}_\star(\mathbb{E}_\xi^\pi[X|M]) - \frac{1}{\eta}\mathrm{I}(\pi, \mu_\xi) \right] \\
&= \sup_{\xi\in\mathcal{W}} \min_{\pi\in\Delta_{k,\delta}} \left(\Delta(\pi, \mu_\xi) - \frac{1}{\eta}\mathrm{I}(\pi, \mu_\xi) \right) \,,
\end{aligned}
$$

where in the first inequality we used the concavity of $x \mapsto \mathcal{V}(\pi, x)$ and the second follows because

$$
\begin{aligned}
\mathbb{E}_\xi^\pi[\langle P, \mathcal{V}(X)\rangle] &= \mathbb{E}_\xi^\pi[\mathbb{E}_\xi^\pi[\langle P, \mathcal{V}^m(X)\rangle|X, P]] = \mathbb{E}_\xi^\pi[\mathbb{E}_\xi^\pi[\mathcal{V}^m(X)_M|X, P]] \\
&= \mathbb{E}_\xi^\pi[\mathbb{E}_\xi^\pi[\mathcal{V}^m(X)_M|M]] = \mathbb{E}_\xi^\pi[\mathcal{V}^m(\mathbb{E}_\xi^\pi[X|M])_M] \geq \mathbb{E}_\xi^\pi[\mathcal{V}_\star(\mathbb{E}_\xi^\pi[X|M])] \,.
\end{aligned}
$$

Since $\delta > 0$ is arbitrary,

$$\Lambda_{q\eta}^\star \leq \inf_{\delta>0} \sup_{\xi \in \mathcal{W}} \min_{\pi \in \Delta_{k,\delta}} \left(\Delta(\pi, \mu_\xi) - \frac{1}{\eta} I(\pi, \mu_\xi) \right)$$

$$\leq \inf_{\delta>0} \sup_{\xi \in \mathcal{W}} \min_{\pi \in \Delta_k} \left(\Delta((1-k\delta)\pi + \delta\mathbf{1}, \mu_\xi) - \frac{1}{\eta} I((1-k\delta)\pi + \delta\mathbf{1}, \mu_\xi) \right)$$

$$\leq \inf_{\delta>0} \left[k\delta + (1-k\delta) \sup_{\xi \in \mathcal{W}} \min_{\pi \in \Delta_k} \left(\Delta(\pi, \mu_\xi) - \frac{1}{\eta} I(\pi, \mu_\xi) \right) \right] ,$$

where we used convexity of $\pi \mapsto \Delta(\pi, \mu_\xi)$ and that $\Delta(\pi, x) \in [0, 1]$ and $\pi \mapsto I(\pi, \mu_\xi)$ is linear and non-negative. By the definition of the information ratio and Lemma 14, for any ξ there exists a policy $\pi \in \Delta_k$ such that

$$\Delta(\pi, \mu_\xi) \leq (\Psi_\star^\lambda)^{1/\lambda} I(\pi, \mu_\xi)^{1/\lambda}$$

Therefore,

$$\Lambda_{q\eta}^\star \leq \inf_{\delta>0} \left[k\delta + (1-k\delta) \sup_{\xi \in \mathcal{W}} \left((\Psi_\star^\lambda)^{1/\lambda} I(\pi, \mu_\xi)^{1/\lambda} - \frac{I(\pi, \mu_\xi)}{\eta} \right) \right] \leq \frac{\lambda-1}{\lambda} \left(\frac{\eta \Psi_\star^\lambda}{\lambda} \right)^{1/(\lambda-1)} ,$$

where the second inequality follows by optimising the worst possible value of $I(\pi, \mu_\xi)$ and because of the $\inf$ over $\delta > 0$. Combining the above with Eq. (8) completes the proof.

Appendix F. Comparing the notions of regret

There is a subtle connection between the Rustichini and the standard regret. By Theorem 3, the minimax asymptotic regret for both settings is determined by the minimax information ratio and the resulting value of $\lambda_\star$. Given a game $\mathcal{G} = (\mathcal{L}, \mathcal{S})$, let $\lambda_\star^s(\mathcal{G})$ be the value of $\lambda_\star$ in game $\mathcal{G}$ when the standard regret is used and $\lambda_\star^r(\mathcal{G})$ be its value when the Rustichini regret is used. Throughout this section, assume that $\mathcal{Z}$ is finite and let $\mathcal{V}$ and $\mathcal{V}_\star$ be the Rustichini versions as defined in Eq. (2).

By our main theorem and the upper bounds by (Kwon and Perchet, 2017) we know that $\lambda_\star^r(\mathcal{G}) \leq 3$ for all finite games. The classification theorem for finite partial monitoring (Bartók et al., 2014) and our Theorem 3 show that $\lambda_\star^s(\mathcal{G}) \in \{1, 2, 3, \infty\}$. Games with $\lambda_\star^s(\mathcal{G}) = \infty$ are called hopeless and the learner suffers linear minimax standard regret. So the Rustichini and standard regret are not the same on these games. An example is the game in Section 5, for which $\lambda_\star^s(\mathcal{G}) = \infty$ and $\lambda_\star^r(\mathcal{G}) = 7/3$. A natural question is whether or not $\lambda_\star^s(\mathcal{G}) = \lambda_\star^r(\mathcal{G})$ when $\lambda_\star^s(\mathcal{G}) \in \{1, 2, 3\}$. The answer is obviously yes for all games where $x\mathrm{R}y$ if and only if $x = y$. The answer also turns out to be yes for the large class of non-degenerate games. For degenerate games there can be subtle differences.

To explain, we need to remind you about the classification of actions and games in the standard setting. We assume some familiarity with the classification theorem for finite games (Bartók et al., 2014) and the subtle definitions of the cell decomposition and observability. Detailed intuition to complement the definitions below can be found in the literature (Lattimore and Szepesvári, 2020). The cell decomposition for the standard regret is defined as follows. The cell associated with action a is

$$\mathcal{C}_a = \left\{ x \in \mathcal{K} : \mathcal{L}(a, x) = \min_{b \in [k]} \mathcal{L}(b, x) \right\} .$$

The dimension of a cell is the dimension of its affine hull. Actions a with $\dim(\mathcal{C}_a) = \dim(\mathcal{K})$ are called Pareto optimal. Actions a with $\mathcal{C}_a = \emptyset$ are called dominated and the remaining actions are called degenerate. A pair of actions a and b are called globally observable if there exists a function $v : [k] \times \Sigma \to \mathbb{R}$ such that

$$\mathcal{L}(a, z) - \mathcal{L}(b, z) = \sum_{c \in [k]} v(c, \mathcal{S}(c, z)) = \mathcal{L}(a, z) - \mathcal{L}(b, z) \text{ for all } z \in \mathcal{Z}, \qquad (11)$$

which just means that the loss differences between a and b can be estimated from the signals when playing all actions. The game is called globally observable if all pairs of Pareto optimal actions are globally observable. Note, there is no need to estimate the losses of dominated or degenerate actions because they cannot be uniquely optimal. This is the subtlety that causes problems.

Proposition 19 *Suppose that all pairs of actions (not just Pareto optimal) are globally observable. Then $\lambda_\star^s(\mathcal{G}) = \lambda_\star^r(\mathcal{G})$.*

Proof By the definition of global observability, there exists a function $\omega : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ such that for all actions a,

$$\mathcal{L}(a, y) = \mathcal{L}(a, x) + \omega(x, y).$$

Hence the definitions of the instantaneous regret (Eq. (3)) for both settings coincide. Since the information gain is always the same, the minimax information ratios also coincide and so $\lambda_\star^s(\mathcal{G}) = \lambda_\star^r(\mathcal{G})$. ∎

With a small modification we can extend this result to games with dominated actions.

Proposition 20 *For all globally observable games $\mathcal{G}$ with no degenerate actions, $\lambda_\star^s(\mathcal{G}) = \lambda_\star^r(\mathcal{G})$.*

Proof Let $\mathcal{G} = (\mathcal{L}, \mathcal{S})$ be a globally observable game with no degenerate actions, but possibly some dominated actions. Assume without loss of generality that actions $1, \ldots, p$ are Pareto optimal and actions $p + 1, \ldots, k$ are dominated. Suppose that a is a dominated action, which means that $\mathcal{L}(a, x) < \min_{b \in [k]} \mathcal{L}(b, x)$ for all $x \in \mathcal{K}$. By von Neummann's minimax theorem and compactness of $\mathcal{K}$ there exists an $\epsilon > 0$ such that for all dominated actions a,

$$\max_{\pi \in \Delta_k : \mathrm{spt}(\pi) \subset [p]} \min_{x \in \mathcal{K}} \mathcal{L}(a, x) - \mathcal{L}(\pi, x) = \min_{x \in \mathcal{K}} \max_{\pi \in \Delta_k : \mathrm{spt}(\pi) \subset [p]} \mathcal{L}(a, x) - \mathcal{L}(\pi, x) \geq \epsilon.$$

Hence for each dominated action $a \in [k]$ there exists a distribution ρ_a over Pareto optimal actions such that for all $x \in \mathcal{K}$,

$$\mathcal{L}(\rho_a, x) + \epsilon \leq \mathcal{L}(a, x).$$

Next, define a new loss function by

$$\tilde{\mathcal{L}}(a, x) = \begin{cases} \mathcal{L}(a, x) & \text{if } a \leq p \\ \mathcal{L}(\rho_a, x) + \epsilon \leq \mathcal{L}(a, x) & \text{if } a > p. \end{cases}$$

The game defined by $\tilde{\mathcal{G}} = (\tilde{\mathcal{L}}, \mathcal{S})$ has the same dominated actions as the original game and the same loss function for the non-dominated actions. Since the signal function is the same, the modified game has the same classification as the original game. On the other hand,

$$\max_{y:y\mathrm{R}x} \tilde{\mathcal{L}}(\pi, y) - \min_{\rho\in\Delta_k} \max_{y:y\mathrm{R}x} \tilde{\mathcal{L}}(\rho, y) = \max_{y:y\mathrm{R}x} \tilde{\mathcal{L}}(\pi, y) - \min_{\rho\in\Delta_k} \max_{y:y\mathrm{R}x} \mathcal{L}(\rho, y)$$
$$\leq \max_{y:y\mathrm{R}x} \mathcal{L}(\pi, y) - \min_{\rho\in\Delta_k} \max_{y:y\mathrm{R}x} \mathcal{L}(\rho, y).$$

Hence the instantaneous regret function using the Rustichini definition is smaller in the modified game than in the original. And since the signal structure is the same, the minimax information ratio is also smaller in the modified game than the original. Furthermore, by the definition of the modified game all pairs of actions are now globally observable, so by Proposition 19,

$$\lambda_\star^{\mathrm{r}}(\mathcal{G}) \geq \lambda_\star^{\mathrm{r}}(\tilde{\mathcal{G}}) = \lambda_\star^{\mathrm{s}}(\tilde{\mathcal{G}}) = \lambda_\star^{\mathrm{s}}(\mathcal{G})$$

For the upper bound we need to consider the cases. We already know that $\lambda_\star^{\mathrm{r}}(\mathcal{G}) \leq 3$, so we are done when $\lambda_\star^{\mathrm{s}}(\mathcal{G}) = 3$. Games where $\lambda_\star^{\mathrm{s}}(\mathcal{G}) = 1$ have a single action that is optimal everywhere, so in this case $\lambda_\star^{\mathrm{r}}(\mathcal{G}) = 1$ is trivial as well. We need to upper bound $\lambda_\star^{\mathrm{r}}(\mathcal{G})$ for games where $\lambda_\star^{\mathrm{s}}(\mathcal{G}) = 2$, which are the so-called non-trivial locally observable games. For these games it is known that there exists a policy with optimal regret (up to constant factors) that only plays actions a that are Pareto optimal or where $\mathcal{L}(a, \cdot) = (1 - \alpha)\mathcal{L}(b, \cdot) + \alpha\mathcal{L}(c, \cdot)$ for some $\alpha \in (0, 1)$ and with b and c Pareto optimal (Lattimore and Szepesvári, 2020). Let $\tilde{\mathcal{G}}$ be the game obtained from $\mathcal{G}$ by eliminating the actions that do not satisfy the above criteria. Since all the eliminated actions are degenerate, this can only increase the Rustichini instantaneous regret and hence $\lambda_\star^{\mathrm{r}}(\mathcal{G}) \leq \lambda_\star^{\mathrm{r}}(\tilde{\mathcal{G}})$. But in $\tilde{\mathcal{G}}$ all pairs of actions are globally observable and hence

$$\lambda_\star^{\mathrm{r}}(\mathcal{G}) \leq \lambda_\star^{\mathrm{r}}(\tilde{\mathcal{G}}) = \lambda_\star^{\mathrm{s}}(\tilde{\mathcal{G}}) = \lambda_\star^{\mathrm{s}}(\mathcal{G}),$$

where the final equality follows because removing the actions not satisfying the above criteria cannot change the classification of the game. ∎

When there are degenerate actions it is possible that $\lambda_\star^{\mathrm{r}}(\mathcal{G}) \neq \lambda_\star^{\mathrm{s}}(\mathcal{G})$. Consider the following example:

$$\mathcal{L} = \begin{bmatrix} 1 & 0 & 2 & 2 \\ 0 & 1 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \end{bmatrix} \qquad \mathcal{S} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

The adversary has four choices, so $\mathcal{K}$ is the probability simplex with four corners. Actions 3 and 4 are degenerate. They are only optimal if $x_1 = x_2 = 0$. They are the only actions that reveal information. Action 1 is optimal if $x_2 \geq x_1$ and action 2 is optimal otherwise. Equivalently, action 1 is optimal if $x_2 + x_3 \geq x_1 + x_3$. Both sides of this inequality can be estimated by playing actions 3 and 4. You can check that this game is globally observable but not locally observable, so $\lambda_\star^{\mathrm{s}}(\mathcal{G}) = 3$. Let $p(x) = x_1 + x_3$ and $q(x) = x_2 + x_3$. The adversary's optimal decision y with $y\mathrm{R}x$ is to play $y_3 = \min(p(x), q(x))$ and $y_1 = p(x) - y_3$ and $y_2 = q(x) - y_3$ with the remaining mass on y_4. In

other words, the adversary chooses the y with $y\mathrm{R}x$ with as much mass on x_3 and x_4 as possible. Calculating, we have

$$\mathcal{V}_\star(x) = 2 - 2(p(x) + q(x)) + 4\min(p(x), q(x))\,.$$

On the other hand, if π is a Dirac on any convex combination of actions 3 and 4, then obviously $\mathcal{V}(\pi, x) = 2$ and $\Delta(\pi, x) = 2|p(x) - q(x)|$. By the proof of our lower bound in Theorem 3, it suffices to prove an upper bound on $\lambda_\star^{\mathrm{r}}(\mathcal{G})$ by constructing a policy with small regret against a stochastic adversary. In this case the simple explore-then-commit policy that plays actions 3 and 4 uniformly until it identifies statistically the sign of $p(x) - q(x)$ and then plays optimally has regret of order $n^{1/2}$ up to logarithmic factors. Hence, $\lambda_\star^{\mathrm{r}}(\mathcal{G}) \le 2$. Since $\lambda_\star^{\mathrm{s}}(\mathcal{G}) = 3$, this gives an example where the two quantities are different. At an intuitive level, the difference arises because in the Rustichini definition of the regret the adversary is compelled to play the worst case over the indistinguishable outcomes. When degenerate actions are involved, this can be the difference between a degenerate action being near-optimal or not.