An Information-Theoretic Approach to Minimax Regret in Partial Monitoring

Tor Lattimore LATTIMORE@GOOGLE.COM
DeepMind, London

Csaba Szepesvári SZEPI@GOOGLE.COM
DeepMind, London

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We prove a new minimax theorem connecting the worst-case Bayesian regret and minimax regret under finite-action partial monitoring with no assumptions on the space of signals or decisions of the adversary. We then generalise the information-theoretic tools of Russo and Van Roy (2016) for proving Bayesian regret bounds and combine them with the minimax theorem to derive minimax regret bounds for various partial monitoring settings. The highlight is a clean analysis of 'easy' and 'hard' finite partial monitoring, with new regret bounds that are independent of arbitrarily large game-dependent constants and eliminate the logarithmic dependence on the horizon for easy games that appeared in earlier work. The power of the generalised machinery is further demonstrated by proving that the minimax regret for k-armed adversarial bandits is at most $\sqrt{2kn}$, improving on existing results by a factor of 2. Finally, we provide a simple analysis of the cops and robbers game, also improving best known constants.

Keywords: Online learning, partial monitoring, minimax theorems, bandits.

1. Introduction

Partial monitoring is a generalisation of the multi-armed bandit framework with an interestingly richer structure. In this paper we are concerned with the finite-action version. Let k be the number of actions. A finite-action partial monitoring game is described by two functions, the loss function $\mathcal{L} : [k] \times \mathcal{X} \to [0,1]$ and a signal function $\Phi : [k] \times \mathcal{X} \to \Sigma$, where $[k] = \{1, 2, \ldots, k\}$ and $\mathcal{X}$ and Σ are topological spaces. At the start of the game the adversary secretly chooses a sequence of outcomes $(x_t)_{t=1}^n$ with $x_t \in \mathcal{X}$, where n is the horizon. The learner knows $\mathcal{L}$, Φ and n and sequentially chooses actions $(A_t)_{t=1}^n$ from $[k]$. In round t, after the learner chooses A_t they suffer a loss of $\mathcal{L}(A_t, x_t)$ and observe only $\Phi(A_t, x_t)$ as a way of indirectly learning about the loss. A policy π is a function mapping action/signal sequences to probability distributions over actions (the learner is allowed to randomise) and the regret of policy π in environment $x = (x_t)_{t=1}^n$ is

$$\mathfrak{R}_n(\pi, x) = \max_{a \in [k]} \mathbb{E}\left[\sum_{t=1}^n \mathcal{L}(A_t, x_t) - \mathcal{L}(a, x_t)\right],$$

where the expectation is taken with respect to the randomness in the learner's choices which follow π. The minimax regret of a partial monitoring game is

$$\mathfrak{R}_n^* = \inf_{\pi \in \Pi} \sup_{x \in \mathcal{X}^n} \mathfrak{R}_n(\pi, x),$$

where Π is the space of all policies. Our objective is to understand how the minimax regret depends on the horizon n and the structure of $\mathcal{L}$ and Φ. Note, this is the oblivious setting because the adversary chooses all the losses at the start of the game. Some classical examples of partial monitoring games are given in Table 1 and Fig. 5 in the appendix.

Setting	$\boldsymbol{\mathcal{X}}$	$\boldsymbol{\Sigma}$	$\boldsymbol{\Phi(a,x)}$	$\boldsymbol{\mathcal{L}(a,x)}$
Full information	$[0,1]^k$	$[0,1]^k$	x	x_a
Bandit	$[0,1]^k$	$[0,1]$	x_a	x_a
Cops and robbers	$[0,1]^k$	$[0,1]^{k-1}$	$x_1,\dots,x_{a-1},x_{a+1},\dots,x_k$	x_a
Finite partial monitoring	$[d]$	arbitrary	arbitrary	arbitrary

Table 1: Example environment classes. In the last row, d is a natural number.

Bayesian viewpoint Although our primary objective is to shed light on the minimax adversarial regret, we establish our results by first proving uniform bounds on the Bayesian regret that hold for any prior. Then a new minimax theorem demonstrates the existence of an algorithm with the same minimax regret. While these methods are not constructive, we demonstrate that they lead to elegant analysis of various partial monitoring problems, and better control of the constants in the bounds.

Let $\mathscr{Q}$ be a space of probability measures on $\mathcal{X}^n$ with the Borel σ-algebra. The Bayesian regret of a policy π with respect to prior $\nu \in \mathscr{Q}$ is

$$\mathfrak{BR}_n(\pi,\nu) = \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi,x) d\nu(x)\,.$$

The minimax Bayesian optimal regret is

$$\mathfrak{BR}_n^*(\mathscr{Q}) = \sup_{\nu\in\mathscr{Q}} \inf_{\pi\in\Pi_{\mathrm{M}}} \mathfrak{BR}_n(\pi,\nu)\,,$$

where Π_{M} is a space of policies so that $x \mapsto \mathfrak{R}_n(\pi,x)$ is measurable, which we define formally in Section 3. When $\mathscr{Q}$ is clear from the context, we write $\mathfrak{BR}_n^*$ in place of $\mathfrak{BR}_n^*(\mathscr{Q})$.

Contributions Our first contribution is to generalise the machinery developed by Russo and Van Roy (2016, 2017) and Bubeck et al. (2015). In particular, we prove a minimax theorem for finite-action partial monitoring games with no restriction on either the loss or the feedback function. The theorem establishes that the Bayesian optimal regret and minimax regret are equal: $\mathfrak{BR}_n^* = \mathfrak{R}_n^*$. Next, the information-theoretic machinery of Russo and Van Roy (2017) is generalised by replacing the mutual information with an expected Bregman divergence. The power of the generalisation is demonstrated by showing that $\mathfrak{R}_n^* \le \sqrt{2kn}$ for k-armed adversarial bandits, which improves on the best known bounds by a factor of 2. The rest of the paper is focussed on applying these ideas to finite partial monitoring games. The results enormously simplify existing analysis by sidestepping the complex localisation arguments. At the same time, our bounds for the class of 'easy non-degenerate' games do not depend on arbitrarily large game-dependent constants, which was true of all prior analysis. Finally, for a special class of bandits with graph feedback called cops and robbers, we show that $\mathfrak{R}_n^* \le \sqrt{2n\log(k)}$, improving on prior work by a factor of $5/\sqrt{2}$.

2. Related work

Since partial monitoring is so generic, the related literature is vast, with most work focussing on the full information setting (see Cesa-Bianchi and Lugosi (2006)) or the bandit setting (Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2019)). The information-theoretic machinery that we build on was introduced by Russo and Van Roy (2016, 2017) in the context of minimizing the Bayesian regret for stationary stochastic bandits (with varying structural assumptions). Bubeck et al. (2015) noticed the results also applied to the 'adversarial' Bayesian setting and applied minimax theory to prove worst-case bounds for convex bandits. Minimax theory has also been used to transfer Bayesian regret bounds to adversarial bounds. For example, Abernethy et al. (2009) explores this in the context of online convex optimisation in the full-information setting and Gravin et al. (2016) for prediction with expert advice. The finite version of partial monitoring was introduced by Rustichini (1999), who developed Hannan consistent algorithms. The main challenge since then has been characterizing the dependence of the regret on the horizon in terms of the structure of the loss and signal functions. It is now known that all games can be classified into one of exactly four types. Trivial and hopeless, for which $\mathfrak{R}_n^* = 0$ and $\mathfrak{R}_n^* = \Omega(n)$ respectively. Between these extremes there are 'easy' games where $\mathfrak{R}_n^* = \Theta(n^{1/2})$ and 'hard' games for which $\mathfrak{R}_n^* = \Theta(n^{2/3})$. The classification result is proven by piecing together upper and lower bounds from various papers (Cesa-Bianchi et al., 2006; Foster and Rakhlin, 2012; Antos et al., 2013; Bartók et al., 2014; Lattimore and Szepesvári, 2019). A caveat of the classification theorem is that the focus is entirely on the dependence of the minimax regret on the horizon. The leading constant is game-dependent and poorly understood. Existing bounds for easy games depend on a constant that can be arbitrarily large, even for fixed d and k. One of the contributions of this paper is to resolve this issue. Another disadvantage of the current partial monitoring literature, especially in the adversarial setting, is that the algorithms and analysis tend to be rather complicated. Although our results only prove the existence of an algorithm witnessing a claimed minimax bound, the Bayesian algorithm and analysis are intuitive and natural. There is also a literature on stochastic partial monitoring, with early analysis by Bartók et al. (2011). A quite practical algorithm was proposed by Vanchinathan et al. (2014). The asymptotics have also been worked out (Komiyama et al., 2015). Although a frequentist regret bound in a stochastic setting normally implies a Bayesian regret bound, in our Bayesian setup the environments are not stationary, while all the algorithms for the stochastic case rely heavily that the distribution of the adversary is stationary. Generalising these algorithms to the non-stationary case does not seem straightforward. Finally, we should mention there is an alternative definition of the regret that is less harsh on the learner. For trivial, easy and hard games it is the same, but for hopeless games the regret captures the hopelessness of the task and measures the performance of the learner relative to an achievable objective. We do not consider this definition here. Readers interested in this variation can consult the papers by Rustichini (1999); Mannor and Shimkin (2003); Perchet (2011) and Mannor et al. (2014).

3. Notation and conventions

The maximum/supremum of the empty set is negative infinity. The standard basis vectors in $\mathbb{R}^d$ are $e_1, \dots, e_d$. The column vector of all ones is $\mathbf{1} = (1, 1, \dots, 1)^\top$. The standard inner product is $\langle \cdot, \cdot \rangle$. The ith coordinate of vector $x \in \mathbb{R}^d$ is x_i. The $(d-1)$-dimensional probability simplex is $\Delta^{d-1} = \{x \in [0,1]^d : \|x\|_1 = 1\}$. The interior of a topological space Z is $\operatorname{int}(Z)$ and its boundary is ∂Z. The relative entropy between probability measures μ and ν over the same measurable space is $\mathrm{D}(\nu \,\|\, \mu) = \int \log(\frac{d\nu}{d\mu}) d\nu$ if $\nu \ll \mu$ and $\mathrm{D}(\nu \,\|\, \mu) = \infty$ otherwise, where $\log$ is the natural logarithm. When X is a random variable with $X \in [a, b]$ almost surely, then Pinsker's inequality combined with straightforward inequalities shows that

$$\int X(d\mu - d\nu) \leq (b-a) \|\mu - \nu\|_{\mathrm{TV}} \leq (b-a)\sqrt{\frac{1}{2}\mathrm{D}(\mu \,\|\, \nu)}\,, \tag{1}$$

where $\|\mu - \nu\|_{\mathrm{TV}}$ is the total variation distance. When $\nu \ll \mu$, the squared Hellinger distance can be written as $h(\nu, \mu)^2 = \int (1 - \sqrt{d\nu/d\mu})^2 d\mu$. Given a measure $\mathbb{P}$ and jointly distributed random elements X and Y we let $\mathbb{P}_X$ denote the law of X and (unconventionally) we let $\mathbb{P}_{X|Y}$ be the conditional law of X given Y, which satisfies $\mathbb{P}_{X|Y}(\cdot) = \mathbb{P}(X \in \cdot \,|\, Y)$. One can think of $\mathbb{P}_{X|Y}$ as a random probability measure over the range of X that depends on Y. In none of our analysis do we rely on exotic spaces where such regular versions do not exist. When $Y \in [k]$ is discrete we let $\mathbb{P}_{X|Y=i}$ denote $\mathbb{P}(X \in \cdot \,|\, Y = i)$ for $i \in [k]$. With this notation the mutual information between X and Y is $I(X;Y) = \mathbb{E}[\mathrm{D}(\mathbb{P}_{X|Y} \,\|\, \mathbb{P}_X)]$. The domain of a convex function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is $\mathrm{dom}(F) = \{x : F(x) < \infty\}$. The Bregman divergence with respect to convex/differentiable F is $D_F : \mathrm{dom}(F) \times \mathrm{dom}(F) \to [0, \infty]$. For $x, y \in \mathrm{dom}(F)$ this is defined by $D_F(x, y) = F(x) - F(y) - \nabla_{x-y} F(y)$, where $\nabla_v F(y)$ is the directional derivative of F in direction v at y. The relative entropy between categorical distributions $p, q \in \Delta^{k-1}$ is the Bregman divergence between p and q where F is the unnormalised negentropy: $F(p) = \sum_{i=1}^k (p_i \log(p_i) - p_i)$ with domain $[0, \infty)^k$. The diameter of a convex set $\mathcal{K}$ with respect to F is $\mathrm{diam}_F(\mathcal{K}) = \sup_{x,y \in \mathcal{K}} F(x) - F(y)$.

Probability spaces, policies and environments The Borel σ-algebra on topological space Z is $\mathfrak{B}(Z)$. Recall that $\mathcal{X}$ and Σ are assumed to carry a topology, which we will use for ensuring measurability of the regret. More about the choices of these topologies later. We assume the signal function $\Phi(a, \cdot)$ is $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\Sigma)$-measurable and the loss function $\mathcal{L}(a, \cdot)$ is $\mathfrak{B}(\mathcal{X})$-measurable. A policy is a function $\pi : \cup_{t=1}^n ([k] \times \Sigma)^{t-1} \to \Delta^{k-1}$ and the space of all policies is Π. A policy is measurable if $h_{t-1} \mapsto \pi(h_{t-1})$ is $\mathfrak{B}(([k] \times \Sigma)^t)$-measurable for all $h_{t-1} = a_1, \sigma_1, \ldots, a_{t-1}, \sigma_{t-1}$, which coincides with the usual definition of a probability kernel. The space of all measurable policies is Π_{M}. In general Π_{M} is a strict subset of Π. For most of the paper we work in the Bayesian framework where there is a prior probability measure ν on $(\mathcal{X}^n, \mathfrak{B}(\mathcal{X}^n))$. Given a prior ν and a measurable policy $\pi \in \Pi_{\mathrm{M}}$, random elements $X \in \mathcal{X}^n$ and $A \in [k]^n$ are defined on common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We let $\Phi_t(a) = \Phi(a, X_t)$ and $\mathcal{L}_t(a) = \mathcal{L}(a, X_t)$. Expectations $\mathbb{E}$ are with respect to $\mathbb{P}$. For $t \in [n+1]$ we let $\mathcal{F}_t = \sigma(A_1, \Phi(A_1, X_1), \ldots, A_{t-1}, \Phi(A_{t-1}, X_{t-1})) \subseteq \mathcal{F}$, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \,|\, \mathcal{F}_t]$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \,|\, \mathcal{F}_t)$. Note that $\mathcal{F}_1 = \{\emptyset, \Omega\}$ is the trivial σ algebra. The σ-algebra $\mathcal{F}$ and the measure $\mathbb{P}$ are such that

1. The law of the adversaries choices satisfies $\mathbb{P}(X \in \cdot\,) = \nu(\cdot)$.

2. For any $t \in [n]$, the law of the actions almost surely satisfies

$$\mathbb{P}_t(A_t \in \cdot\,) = \mathbb{P}_t(A_t \in \cdot \,|\, X) = \pi(A_1, \Phi_1(A_1), \ldots, A_{t-1}, \Phi_{t-1}(A_{t-1}))(\cdot)\,. \tag{2}$$

The existence of a probability space satisfying these properties is guaranteed by Ionescu-Tulcea (Kallenberg, 2002, Theorem 6.17). The last condition captures the important assumption that, conditioned on the observed history, A_t is sampled independently from X. In particular, it implies that

X_t and A_t are independent under $\mathbb{P}_t$. The optimal action is $A^* = \arg\min_{a \in [k]} \sum_{t=1}^n \mathcal{L}_t(a)$. It is not hard to see that the Bayesian regret is well defined and satisfies

$$\mathfrak{BR}_n(\pi, \nu) = \mathbb{E}\left[\sum_{t=1}^n \mathcal{L}_t(A_t) - \mathcal{L}_t(A^*)\right] = \mathbb{E}\left[\sum_{t=1}^n \Delta_t\right],$$

where $\Delta_t = \mathcal{L}_t(A_t) - \mathcal{L}_t(A^*)$. To minimise clutter, when the policy π and prior ν are clear from the context, we abbreviate $\mathfrak{BR}_n(\pi, \nu)$ to $\mathfrak{BR}_n$. We let $P_{ta} = \mathbb{P}_t(A_t = a)$, which means that $P_t \in \Delta^{k-1}$ is a probability vector.

4. Minimax theorem

Our first main result is a theorem that connects the minimax regret to the worst-case Bayesian regret over all finitely supported priors. The regret $\mathfrak{R}_n(\pi, x)$ is well defined for any x and any policy $\pi \in \Pi$, but the Bayesian regret depends on measurability of $x \mapsto \mathfrak{R}_n(\pi, x)$. If ν is supported on a finite set $x_1, \ldots, x_m \in \mathcal{X}^n$, however, we can write

$$\mathfrak{BR}_n(\pi, \nu) = \sum_{i=1}^m \nu(\{x_i\}) \mathfrak{R}_n(\pi, x_i),$$

which does not rely on measurability. By considering finitely supported priors we free ourselves from any concern that $x \mapsto \mathfrak{R}_n(\pi, x)$ might not be measurable. This also means that if Σ (or $\mathcal{X}$) came with some topologies, we simply replace them with the discrete topology (which makes all maps continuous and measurable, implying $\Pi = \Pi_{\mathrm{M}}$).

Theorem 1 *Let $\mathcal{Q}$ be the space of all finitely supported probability measures on $\mathcal{X}^n$. Then*

$$\inf_{\pi \in \Pi} \sup_{x \in \mathcal{X}^n} \mathfrak{R}_n(\pi, x) = \sup_{\nu \in \mathcal{Q}} \min_{\pi \in \Pi} \mathfrak{BR}_n(\pi, \nu).$$

An equivalent statement of this theorem is that if $\mathcal{X}$ and Σ carry the discrete topology then $\mathfrak{R}_n^* = \mathfrak{BR}_n^*(\mathcal{Q})$, which is the form we prove in Appendix A. The strength of this result is that it depends on no assumptions except that the action set is finite.

Our proof borrows techniques from a related result by Bubeck et al. (2015). The main idea is to replace the policy space Π with a simpler space of 'mixtures' over deterministic policies, which is related to Kuhn's celebrated result on the equivalence of behavioral and mixed strategies (Kuhn, 1953). We then establish that this space is compact and use Sion's theorem to exchange the minimum and maximum. While we borrowed the ideas from Bubeck et al. (2015), our proof relies heavily on the finiteness of the action space, which allowed us to avoid any assumptions on Σ and $\mathcal{X}$, which also necessitated our choice of $\mathcal{Q}$. Neither of the two results imply each other.

Theorem 1 is a minimax theorem for a special kind of two-player multistage zero-sum deterministic partial information game. Minimax theorems for this case are nontrivial because of challenges related to measurability and the use of Sion's theorem. Although there is a rich and sophisticated literature on this topic, we are not aware of any result implying our theorem. Tools include the approach we took using the weak topology (Bernhard, 1992), or the so-called weak-strong topology (Leao et al., 2000) and reduction to completely observable games and then using dynamic programming (Ghosh et al., 2004). An interesting challenge is to extend our result to compact action spaces.

One may hope to generalise the proof by Bubeck et al. (2015), but some important details are missing (for example, the measurable space on which the priors live is undefined, the measurability of the regret is unclear as is the compactness of distributions induced by *measurable* policies). We believe that the approach of Ghosh et al. (2004) can complete this result.

5. The regret information tradeoff

Unless otherwise mentioned, all expectations $\mathbb{E}$ are with respect to the probability measure over interactions between a fixed policy $\pi \in \Pi_{\mathrm{M}}$ and an environment sampled from a prior ν on $(\mathcal{X}^n, \mathfrak{B}(\mathcal{X}^n))$. Before our generalisation we present a restatement of the core theorem in the analysis by Russo and Van Roy (2016). Let $I_t(X;Y)$ be the mutual information between X and Y under $\mathbb{P}_t$. Although the proof is identical, the setup here is different because the prior ν is arbitrary.

Theorem 2 (Russo and Van Roy (2016)) *Suppose there exists a constant $\beta \geq 0$ such that $\mathbb{E}_t[\Delta_t] \leq \sqrt{\beta I_t(A^*; \Phi_t(A_t), A_t)}$ almost surely for all t. Then $\mathfrak{BR}_n \leq \sqrt{n\beta \log(k)}$.*

This elegant result provides a bound on the regret in terms of the information gain about the optimal arm. Our generalisation replaces the information gain with an expected Bregman divergence.

Theorem 3 *Let $(M_t)_{t=1}^{n+1}$ be an $\mathbb{R}^d$-valued martingale adapted to $(\mathcal{F}_t)_{t=1}^{n+1}$ and $M_t \in \mathcal{D} \subset \mathbb{R}^d$ almost surely for all t. Then let F be a convex function with $\operatorname{diam}_F(\mathcal{D}) < \infty$. Suppose there exist constants $\alpha, \beta \geq 0$ such that $\mathbb{E}_t[\Delta_t] \leq \alpha + \sqrt{\beta \mathbb{E}_t[D_F(M_{t+1}, M_t)]}$ almost surely for all t. Then $\mathfrak{BR}_n \leq \alpha n + \sqrt{n\beta \operatorname{diam}_F(\mathcal{D})}$.*

Proof We calculate

$$\begin{aligned}
\mathbb{E}_t[D_F(M_{t+1}, M_t)] &= \mathbb{E}_t\left[\liminf_{h\to 0+}\left(F(M_{t+1}) - F(M_t) - \frac{F(M_t + h(M_{t+1} - M_t)) - F(M_t)}{h}\right)\right] \\
&\leq \liminf_{h\to 0+}\left(\mathbb{E}_t\left[F(M_{t+1}) - F(M_t) - \frac{F((1-h)M_t + hM_{t+1}) - F(M_t)}{h}\right]\right) \\
&= \mathbb{E}_t\left[F(M_{t+1})\right] - F(M_t) + \liminf_{h\to 0+}\frac{F(M_t) - \mathbb{E}_t[F((1-h)M_t + hM_{t+1})]}{h} \\
&\leq \mathbb{E}_t\left[F(M_{t+1})\right] - F(M_t) + \liminf_{h\to 0+}\frac{F(M_t) - F(\mathbb{E}_t[(1-h)M_t + hM_{t+1}])}{h} \\
&= \mathbb{E}_t\left[F(M_{t+1})\right] - F(M_t)\,,
\end{aligned} \tag{3}$$

where the first inequality follows from Fatou's lemma and the second from convexity of F. The last equality is because $\mathbb{E}_t[M_{t+1}] = M_t$. Hence

$$\begin{aligned}
\mathfrak{BR}_n = \mathbb{E}\left[\sum_{t=1}^n \Delta_t\right] &\leq \alpha n + \mathbb{E}\left[\sum_{t=1}^n \sqrt{\beta \mathbb{E}_t[D_F(M_{t+1}, M_t)]}\right] \\
&\leq \alpha n + \sqrt{\beta n \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}_t[D_F(M_{t+1}, M_t)]\right]} \leq \alpha n + \sqrt{\beta n \operatorname{diam}_F(\mathcal{D})}\,,
\end{aligned}$$

where the first inequality follows from the assumption in the theorem, the second by Cauchy-Schwarz, while the third follows by Eq. (3), telescoping and the definition of the diameter. ■

A natural choice for M_t is the posterior distribution of the optimal action. Let $P^*_{ta} = \mathbb{P}_t(A^* = a)$, which is the posterior probability that $A^* = a$ based on the information available at the start of round t. By the tower rule, we have $\mathbb{E}_t[P^*_{t+1}] = P^*_t$ so that $(P^*_t)_{t=1}^{n+1}$ is a martingale adapted to $(\mathcal{F}_t)_{t=1}^{n+1}$.

Corollary 4 *Let $F : \mathbb{R}^k \to \mathbb{R}$ be a convex function with $\operatorname{diam}_F(\Delta^{k-1}) < \infty$. Suppose there exist constants $\alpha, \beta \geq 0$ such that $\mathbb{E}_t[\Delta_t] \leq \alpha + \sqrt{\beta \mathbb{E}_t[D_F(P^*_{t+1}, P^*_t)]}$ almost surely for all t. Then $\mathfrak{BR}_n \leq \alpha n + \sqrt{n\beta \operatorname{diam}_F(\Delta^{k-1})}$.*

Remark 5 *That Theorem 3 generalises Theorem 2 follows by choosing F as the unnormalised negentropy for which $\operatorname{diam}_F(\Delta^{k-1}) \leq \log(k)$ and $\mathbb{E}_t[D_F(P^*_{t+1}, P^*_t)] = I_t(A^*; \Phi_t(A_t), A_t)$. The assumption that $M_t \in \mathbb{R}^d$ can be relaxed. The result continues to hold when M_t takes values in a bounded subset of a Banach space, where the martingale is defined using the Bochner integral. The Bregman divergence generalises naturally via the Gateoux derivative.*

6. Finite-armed bandits

In the bandit setting the learner observes the loss of the action they play, which is modelled by choosing $\Sigma = [0, 1]$, $\mathcal{X} = [0, 1]^k$ and $\Phi(a, x) = \mathcal{L}(a, x) = x_a$. The best known bound is by Bubeck and Cesa-Bianchi (2012), who prove that online mirror descent with an appropriate potential satisfies $\mathfrak{R}^*_n \leq \sqrt{8kn}$. Using the same potential in combination with Theorem 3 allows us to improve this result to $\mathfrak{R}^*_n \leq \sqrt{2kn}$.

Theorem 6 *The minimax regret for k-armed adversarial bandits satisfies $\mathfrak{R}^*_n \leq \sqrt{2kn}$.*

Proof Let $F(p) = -2\sum_{a=1}^k \sqrt{p_a}$, which has domain $[0, \infty)^k$ and $\operatorname{diam}_F(\Delta^{k-1}) \leq 2\sqrt{k}$. Combine Corollary 4 and Theorem 1 and Lemma 7 below for Thompson sampling, which is the policy that samples A_t from $P_t = P^*_t$. ■

Lemma 7 *Let F be as above and $P_t = P^*_t$. Then $\mathbb{E}_t[\Delta_t] \leq \sqrt{k^{1/2}\mathbb{E}_t[D_F(P^*_{t+1}, P^*_t)]}$ a.s..*

Remark 8 *Potentials other than the negentropy have been used in many applications in bandits and online convex optimisation. The log barrier, for example, leads to first order bounds for k-armed bandits (Wei and Luo, 2018). Alternative potentials also appear in the context of adversarial linear bandits (Bubeck et al., 2012, 2018) and follow the perturbed leader (Abernethy et al., 2014). Investigating the extent to which these applications transfer to the Bayesian setting is an interesting direction for the future.*

7. Finite partial monitoring games

Recall from Table 1 that a finite partial monitoring game is characterised by functions $\mathcal{L} : [k] \times [d] \to [0, 1]$ and $\Phi : [k] \times [d] \to \Sigma$ where d is a natural number and Σ is arbitrary. Finite partial monitoring enjoys a rich linear structure, which we now summarise. A picture can help absorbing these concepts, and is provided with an example at the beginning of Appendix I. For $a \in [k]$, let $\ell_a \in [0, 1]^d$ be the vector with $\ell_{ax} = \mathcal{L}(a, x)$. Actions a and b are duplicates if $\ell_a = \ell_b$.

The cell associated with action a is $C_a = \{u \in \Delta^{d-1} : \langle \ell_a, u\rangle \leq \min_{b \neq a}\langle \ell_b, u\rangle\}$, which is the subset of distributions $u \in \Delta^{d-1}$ where action a minimises $\mathbb{E}_{x\sim u}[\mathcal{L}(a, x)]$. Note that $C_a \subset \mathbb{R}^d$ is a closed convex polytope and its dimension $\dim(C_a)$ is defined as the dimension of the affine space it generates. An action a is called Pareto optimal if C_a has dimension $d-1$ and degenerate otherwise. Of course $\cup_a C_a = \Delta^{d-1}$, but cells may have nonempty intersection. When a and b are not duplicates, the intersection $C_a \cap C_b$ is a (possibly empty) polytope of dimension at most $d-2$. A pair of Pareto optimal actions a and b are called neighbours if $C_a \cap C_b$ has dimension $d-2$. A game is called non-degenerate if there are no degenerate actions and no duplicate actions. So far none of the concepts have depended on the signal function. Local observability is a property of the signal and loss functions that allows the learner to estimate loss differences between actions a and b by playing only those actions. For neighbours a and b let $\mathcal{N}_{ab} = \{c : C_c \subseteq C_a \cap C_b\}$, which contains a and its duplicates, b and its duplicates, and degenerate actions c with $C_c = C_a \cap C_b$. A game is globally observable if for all pairs of neighbours there exists a function $f : [k] \times \Sigma \to \mathbb{R}$ such that

$$\mathcal{L}(a, x) - \mathcal{L}(b, x) = \sum_{c=1}^{k} f(c, \Phi(c, x))\,. \tag{4}$$

The game is locally observable if for all pairs of neighbours a and b the function f can be chosen satisfying Eq. (4) and additionally that $f(c, \Phi(c, x)) = 0$ for all $c \notin \mathcal{N}_{ab}$. In the standard analysis of partial monitoring the function f is used to derive importance-weighted estimators of the loss differences. In the following f is used more directly. A quantity that appears naturally in the analysis is the supremum norm of the estimation functions f. Given a globally observable game, we let $v \geq 0$ be the smallest value such that for all pairs of neighbours a and b there exists a function satisfying Eq. (4) with $\|f\|_\infty \leq v$. For locally observable games v is defined in the same way, but with the additional restriction that f is supported on $\mathcal{N}_{ab}$. The neighbourhood of a is $\mathcal{N}_a = \{b : \dim(C_a \cap C_b) \geq d-2\}$. The neighbourhood graph over $[k]$ has edges $\{(a, b) : a, b \text{ are neighbours}\}$. For non-degenerate games, the neighbourhood graph is connected.

The following theorem classifies all partial monitoring games into one of four categories. All results were known previously except that previous upper bounds for locally observable games were $\mathfrak{R}_n^* = O((n\log(n))^{1/2})$.

Theorem 9 *The minimax regret for finite partial monitoring game G satisfies the following:*

$$\mathfrak{R}_n^* = \begin{cases} 0 & \text{if there are no neighbouring actions} \\ \Theta(n^{1/2}) & \text{if there are neighbouring actions and } G \text{ is locally observable} \\ \Theta(n^{2/3}) & \text{if } G \text{ is globally observable and not locally observable} \\ \Omega(n) & \text{otherwise}\,. \end{cases}$$

Summary of new results The main theorem is the following, which improves on previous bounds that all depend on arbitrarily large game-dependent constants, even when k and d are fixed.

Theorem 10 *For any locally observable non-degenerate game:* $\mathfrak{R}_n^* \leq k^{3/2}(d+1)\sqrt{8n\log(k)}$.

For degenerate locally observable games the bound differs only due to the increased norm of the estimation functions. In particular, we have the following theorem, which improves on prior work in terms of constants and logarithmic factors (Lattimore and Szepesvári, 2019).

Theorem 11 *For any locally observable game:* $\mathfrak{R}_n^* \leq vk^{3/2}\sqrt{8n\log(k)}$*, where* v *is a bound on the supremum norm of the estimation functions.*

The bound for globally observable games has the same order as the prior work, but with slightly improved constants (Cesa-Bianchi et al., 2006).

Theorem 12 *For any globally observable game:* $\mathfrak{R}_n^* \leq 3(nkv)^{2/3}(\log(k)/2)^{1/3}$*, where* v *is a bound on the supremum norm of the estimation functions.*

Finally, for any locally/globally observable game, Lemma 25 in the appendix shows that the norm of the estimators is bounded by at most $v \leq d^{1/2}(1+k)^{d/2}$, which provides an explicit upper bound that is independent of the loss and signal matrix. We believe the exponential dependence on the dimension is unavoidable in general.

8. Proof of Theorem 10

For this section we assume the game is non-degenerate and locally observable. Before the proof of Theorem 10 we provide the algorithm, which seems to be novel among previous algorithms for partial monitoring. Note that Thompson sampling can suffer linear regret in partial monitoring (Appendix G). Let $G_t = \arg\min_{a\in[k]} \mathbb{E}_t[\mathcal{L}_t(a)]$ be the greedy action that minimises the 1-step Bayesian expected loss. The idea is to define a directed tree with vertex set $[k]$ and root G_t and where all paths lead to G_t. A little notation is needed. Define an undirected graph with vertices V_t and edges E_t by $V_t = \{a \in [k] : \mathbb{E}_t[\mathcal{L}_t(a)] = \mathbb{E}_t[\mathcal{L}_t(G_t)]\}$ and $E_t = \{a, b \in V_t : a \text{ and } b \text{ are neighbours}\}$, which is connected by Lemma 23. Note that $V_t = \{G_t\}$ when G_t is unique, but this is not always the case. For $a \in V_t$ let $\rho_t(a)$ be the length of the shortest path from a to G_t in (V_t, E_t) with $\rho_t(G_t) = 0$ by definition. Let $\mathcal{P}_t : [k] \to [k]$ be the 'parent' function:

$$\mathcal{P}_t(a) = \begin{cases} \arg\min_{b\in\mathcal{N}_a} \mathbb{E}_t[\mathcal{L}_t(b)] & \text{if } a \notin V_t \\ \arg\min_{b\in\mathcal{N}_a\cap V_t} \rho_t(b) & \text{otherwise}\,. \end{cases}$$

The following lemma is proven in Appendix H.

Lemma 13 *The directed graph over vertex set* $[k]$ *with an edge from* a *to* b *if* $a \neq G_t$ *and* $b = \mathcal{P}(a)$ *is a directed tree with root* G_t*.*

Let $\mathcal{A}_t(a)$ be the set of ancestors of action a in the tree defined in Lemma 13. We adopt the convention that $a \in \mathcal{A}_t(a)$. By the previous lemma, $G_t \in \mathcal{A}_t(a)$ for all a. Let $\mathcal{D}_t(a)$ be the set of descendants of a, which does not include a (Fig. 7). The depth of an action a in round t is the distance between a and the root G_t. An action a is called anomalous for $P \in \Delta^{k-1}$ in round t if $P_a < \max_{b\in\mathcal{D}_t(a)} P_b$. Algorithm 2 defines the 'water transfer' operator $W_t : \Delta^{k-1} \to \Delta^{k-1}$ that corrects this deficiency by transferring mass towards the root of the tree defined in Lemma 13 while ensuring that (a) the loss suffered when playing the according to the transformed distribution does not increase and (b) the distribution is not changed too much. The process is illustrated in Fig. 1 in Appendix F, where you will also find the proof of the next lemma.

Lemma 14 *Let* $P \in \Delta^{k-1}$ *and* $Q = W_t^k P = W_t \cdots W_t P$*. Then:*

1. $\sum_{a=1}^k Q_a \mathbb{E}_t[\mathcal{L}_t(a)] \leq \sum_{a=1}^k P_a \mathbb{E}_t[\mathcal{L}_t(a)]$.
2. $Q_a \leq Q_{\mathcal{P}_t(a)}$ *for all* $a \in [k]$.
3. $Q_a \geq P_a/k$ *for all* $a \in [k]$.

Our new algorithm samples A_t from $P_t = W_t^k P_t^*$. Because of the plumbing and randomisation, the new algorithm is called Mario sampling (Algorithm 1). The proof of Theorem 10 follows immediately from Theorems 1 and 2 and the following lemma.

```
input:  partial monitoring game (Σ, ℒ, Φ) and prior ν
for t = 1, ..., n
    compute P_t^* and P_t = W_t^k P_t^*.  Then sample A_t ~ P_t.
```

Algorithm 1: Mario sampling

Lemma 15 *For Mario sampling:* $\mathbb{E}_t[\Delta_t] \leq (d+1)k^{3/2}\sqrt{8I_t(A^*; \Phi_t(A_t), A_t)}$ *a.s..*

Proof We assume an appropriate zero measure set is discarded so that we can omit the qualification 'almost surely' for the rest of the proof. By the first part of Lemma 14,

$$\mathbb{E}_t[\Delta_t] \leq \sum_{a=1}^k P_{ta}^* \left(\mathbb{E}_t[\mathcal{L}_t(a)] - \mathbb{E}_t[\mathcal{L}_t(a) \mid A^* = a]\right) . \tag{5}$$

For $b \neq G_t$ let $f_b, g_b : \Sigma \to \mathbb{R}$ be a pair of functions such that $\max\{\|f_b\|_\infty, \|g_b\|_\infty\} \leq d+1$ and $f_b(\Phi(b,x)) + g_b(\Phi(\mathcal{P}_t(b), x)) = \mathcal{L}(b,x) - \mathcal{L}(\mathcal{P}_t(b), x)$ for all $x \in [d]$. The existence of such functions is guaranteed by Lemma 24 and the fact that $\mathcal{P}_t(b) \in \mathcal{N}(b)$ and because we assumed the game is non-degenerate, locally observable. The expected loss of a can be decomposed in terms of the sum of differences to the root,

$$\begin{aligned}
\mathbb{E}_t[\mathcal{L}_t(a)] &= \mathbb{E}_t\left[\mathcal{L}_t(G_t) + \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} (\mathcal{L}_t(b) - \mathcal{L}_t(\mathcal{P}_t(b)))\right] \\
&= \mathbb{E}_t\left[\mathcal{L}_t(G_t) + \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} f_b(\Phi_t(b)) + g_b(\Phi_t(\mathcal{P}_t(b)))\right] .
\end{aligned} \tag{6}$$

In the same way,

$$\mathbb{E}_t[\mathcal{L}_t(a) \mid A^* = a] = \mathbb{E}_t\left[\mathcal{L}_t(G_t) + \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} f_b(\Phi_t(b)) + g_b(\Phi_t(\mathcal{P}_t(b))) \,\middle|\, A^* = a\right] . \tag{7}$$

Then, because $\mathcal{A}_t(a)$ and G_t are $\mathcal{F}_t$-measurable,

$$
\begin{aligned}
\mathbb{E}_t[\Delta_t] &\leq \sum_{a=1}^k P_{ta}^* \left(\mathbb{E}_t[\mathcal{L}_t(a)] - \mathbb{E}_t[\mathcal{L}_t(a) \,|\, A^* = a]\right) && \text{(Eq. (5))} \\
&= \sum_{a=1}^k P_{ta}^* \Bigg[\sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} \left(\mathbb{E}_t[f_b(\Phi_t(b))] - \mathbb{E}_t[f_b(\Phi_t(b)) \,|\, A^* = a]\right) \\
&\quad + \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} \left(\mathbb{E}_t[g_b(\Phi_t(\mathcal{P}_t(b)))] - \mathbb{E}_t[g_b(\Phi_t(\mathcal{P}_t(b))) \,|\, A^* = a]\right) \Bigg] && \text{(Eqs. (6) and (7))} \\
&\leq (d+1) \sum_{a=1}^k P_{ta}^* \sum_{b \in \mathcal{A}_t(a)} \sqrt{8\, \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)}\right)} && \text{(Eq. (1), D} \geq 0) \\
&\leq k(d+1) \sqrt{8 \sum_{a=1}^k P_{ta}^* \sum_{b \in \mathcal{A}_t(a)} P_{ta}^*\, \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)}\right)} && \text{(Cauchy-Schwarz)} \\
&\leq k^{3/2}(d+1) \sqrt{8 \sum_{a=1}^k P_{ta}^* \sum_{b \in \mathcal{A}_t(a)} P_{tb}\, \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)}\right)} && \text{(Lemma 14, Part 3)} \\
&\leq k^{3/2}(d+1) \sqrt{8 \sum_{a=1}^k P_{ta}^* \sum_{b=1}^k P_{tb}\, \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)}\right)} && \text{(D} \geq 0) \\
&= k^{3/2}(d+1) \sqrt{8 I_t(A^*; \Phi_t(A_t), A_t)}\,. && \text{(Lemma 26)}
\end{aligned}
$$

■

Remark 16 *In many games there exists a constant m such that $|\mathcal{A}_t(a)| \leq m$ almost surely for all a and t. In this case Part 3 of Lemma 14 improves to $P_{ta} \geq P_{ta}^*/m$ and the application of Cauchy-Schwarz in Lemma 15 can be strengthened. This means the bound in Theorem 10 becomes $m(d+1)\sqrt{8kn\log(k)}$. For the game illustrated in Fig. 7, $m = 5$ while $k = 7$, but more extreme examples are easily constructed.*

9. Discussion and future directions

One of the main benefits of the information-theoretic approach is the simplicity and naturality of the arguments, which is particularly striking in partial monitoring. Even for the k-armed bandit analysis there is no tuning of learning rates or careful bounding of dual norms. In exchange, our results are existential, though we emphasise that the Bayesian setting is interesting in its own right. We anticipate that Theorem 3 will have many other applications and there is clearly more to understand about this generalisation. Is it a coincidence that the same potential leads to minimax bounds using both online stochastic mirror descent and Thompson sampling?

```
input:  P ∈ Δ^{k−1} and tree determined by 𝒫_t
find action a at the greatest depth such that P_a < max_{b∈𝒟_t(a)} P_b.
if no such action is found, let W_t P = P and return.
for α ∈ [0,1] let 𝒟_t(a;α) = {b ∈ 𝒟_t(a) : P_b ≥ α}.
let α* be the largest α ∈ {P_b : b ∈ 𝒟_t(a)} such that
```

$$p_\alpha = \frac{1}{1 + |\mathcal{D}_t(a;\alpha)|} \sum_{b \in \mathcal{D}_t(a;\alpha) \cup \{a\}} P_b > q_\alpha = \max\{P_b : b \in \mathcal{D}_t(a) \setminus \mathcal{D}_t(a;\alpha)\}\,.$$

```
let (W_t P)_b = p_{α*} if b ∈ 𝒟_t(a;α*) ∪ {a} and (W_t P)_b = P_b otherwise.
```

Algorithm 2: The water transfer operator $W_t : \Delta^{k-1} \to \Delta^{k-1}$.

Information-directed sampling Thompson sampling depends on the prior, but not the potential that appears in Theorem 3. Russo and Van Roy (2014) noted that the information-theoretic analysis is tightest when the algorithm is chosen to minimize $\mathbb{E}_t[\Delta_t]^2/\mathbb{E}_t[D_F(P^*_{t+1}, P^*_t)]$, where F is the unnormalised negentropy. Our generalisation provides a means of constructing new algorithms by changing the potential.

Open problems An obvious next step is stress test the applicability of Theorem 3. Bandits with graph feedback beyond cops and robbers might be a good place to start (Alon et al., 2015). One may also ask whether in adversarial linear bandits the results by Bubeck et al. (2018) can be replicated or improved using Theorem 3. There are many open problems in partial monitoring, a few of which we now describe. We hope some readers will be inspired to work on them!

Adaptivity There exist games where for 'nice' adversaries the regret should be $O(n^{1/2})$ while for truly adversarial data the regret is as large as $\Theta(n^{2/3})$. Designing algorithms that adapt to a broad range of adversaries is an interesting challenge. Some work on this topic in the stochastic setting is by Bartók et al. (2012). A related question is understanding how to use the information-theoretic machinery to provide adaptive bounds.

Constants Our results have eliminated arbitrarily large constants from the analysis of easy non-degenerate games. Still, we do not yet understand how the regret should depend on the structure of $\mathcal{L}$ or Φ except in special cases. The result in Remark 16 is a small step in this direction, but there is much to do. The best place to start is probably lower bounds. Currently generic lower bounds for finite partial monitoring focus on the dependence on the horizon. One concrete question is whether or not the minimum supremum norm of the estimation functions that appears in Theorem 12 is a fundamental quantity.

Stochastic analysis of Mario sampling Theorem 2 and Lemma 15 show that for any prior Mario sampling satisfies $\mathfrak{BR}_n \leq k^{3/2}(d+1)\sqrt{8n\log(k)}$. In the stationary stochastic setting we expect that for a suitable prior it should be possible to prove a bound on the frequentist regret of this algorithm. Perhaps the techniques developed by Agrawal and Goyal (2013) or Kaufmann et al. (2012) generalise to this setting.

References

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

J. D. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. In M. F. Balcan, V. Feldman, and Cs. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 807–823, Barcelona, Spain, 13–15 Jun 2014. PMLR.

S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

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

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. Pál, and Cs. Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 133–154, 2011.

G. Bartók, N. Zolghadr, and Cs. Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML, pages 1779–1786, USA, 2012. Omnipress.

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.

P. Bernhard. Information and strategies in dynamic games. *SIAM Journal on Control and Optimization*, 30(1):212–228, 1992.

V. I. Bogachev. *Measure theory*, volume 2. Springer Science & Business Media, 2007.

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

S. Bubeck, N. Cesa-Bianchi, and S. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Annual Conference on Learning Theory*, volume 23, pages 41–1. Microtome, 2012.

S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 266–278, Paris, France, 03–06 Jul 2015. PMLR.

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

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

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

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

S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

M. K. Ghosh, D. McDonald, and S. Sinha. Zero-sum stochastic games with partial information. *Journal of optimization theory and applications*, 121(1):99–118, 2004.

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

O. Kallenberg. *Foundations of modern probability*. Springer-Verlag, 2002.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In NaderH. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 199–213. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34105-2.

J. Komiyama, J. Honda, and H. Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 1792–1800. Curran Associates, Inc., 2015.

H.W. Kuhn. Extensive games and the problem of information, contributions to the theory of games II. *Annals of Mathematics Studies*, 28:193–216, 1953.

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

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

D. Leao, J. B. R. do Val, and M. D. Fragoso. Nonstationary zero sum stochastic games with incomplete observation. In *39th IEEE Conference on Decision and Control*, pages 2278–2283, 2000.

S. Mannor and N. Shimkin. On-line learning with imperfect monitoring. In *Learning Theory and Kernel Machines*, pages 552–566. Springer, 2003.

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

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

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

D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2017.

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

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

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

H. P. Vanchinathan, G. Bartók, and A. Krause. Efficient partial monitoring with prior information. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, NIPS, pages 1691–1699. Curran Associates, Inc., 2014.

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

Appendix A. Proof of Theorem 1

The proof depends on a little functional analysis. The important point is that the space of policies written as probability measures over deterministic policies is compact and the Bayesian regret is linear and continuous as a function of the measures over policies and priors over environments. Then minimax theorems can be used to exchange the $\min$ and $\sup$. Guaranteeing compactness and continuity and avoiding any kind of measurability issues requires careful choice of topologies.

For a topological space Z, let $\mathscr{P}_r(Z)$ be the space of Radon probability measures when Z is equipped with the Borel σ-algebra. The weak* topology on $\mathscr{P}_r(Z)$ is the coarsest topology such that $\mu \mapsto \int f d\mu$ is continuous for all bounded continuous functions $f : Z \to \mathbb{R}$.

Recall that $\mathcal{X}$ is the space of outcomes and Σ is the space of feedbacks and these are arbitrary sets. A deterministic policy can be represented as a function $\pi : \cup_{t=1}^n \Sigma^{t-1} \to [k]$. By the choice of topology on Σ, these are all continuous, hence, measurable. Let Π_{D} be the space of all such policies, Π_{DM} be the space of the measurable policies amongst these. By Tychonoff's theorem, Π_{D} is compact with the product topology, where $[k]$ has the discrete topology. Π_{D} is Hausdorff because the product of Hausdorff spaces is Hausdorff. By Theorem 8.9.3 in the two volume book by Bogachev (2007), the space $\mathscr{P}_r(\Pi_{\mathrm{D}})$ is weak*-compact. Clearly $\mathscr{P}_r(\Pi_{\mathrm{D}})$ is also convex.

Let $\mathscr{Q}$ be the space of finitely supported probability measures on $\mathcal{X}^n$, which is a convex subset of $\mathscr{P}_r(\mathcal{X}^n)$ where $\mathcal{X}^n$ is taken to have the discrete topology. Equip with $\mathscr{Q}$ with the weak* topology. If $f = f(\mu, \nu)$ with $f : \mathscr{P}_r(\Pi_{\mathrm{D}}) \times \mathscr{Q} \to \mathbb{R}$ is linear and continuous in both μ and ν individually. Since $\mathscr{P}_r(\Pi_{\mathrm{D}})$ is compact, by Sion's minimax theorem (Sion, 1958),[1]

$$\min_{\mu \in \mathscr{P}_r(\Pi_{\mathrm{D}})} \sup_{\nu \in \mathscr{Q}} f(\mu, \nu) = \sup_{\nu \in \mathscr{Q}} \min_{\mu \in \mathscr{P}_r(\Pi_{\mathrm{D}})} f(\mu, \nu) \,.$$

We are going to choose f to be the Bayesian regret and argue that Π can be identified with $\mathscr{P}_r(\Pi_{\mathrm{D}})$. First, we need to check some continuity conditions for the regret. Since $\mathcal{X}^n$ has the discrete topology the map $x \mapsto \mathfrak{R}_n(\pi, x)$ is continuous for fixed π. Now we check that $\pi \mapsto \mathfrak{R}_n(\pi, x)$, $\pi \in \Pi_{\mathrm{D}}$, is continuous for fixed x. Let $\Phi_t(a) = \Phi(a, x_t)$ and $\mathcal{L}_t(a) = \mathcal{L}(a, x_t)$, which are both continuous since $[k]$ has the discrete topology. Then let $\sigma_t : \Pi_{\mathrm{D}} \to \Sigma$ and $a_t : \Pi_{\mathrm{D}} \to [k]$ be defined inductively by

$$a_t(\pi) = \pi(\sigma_1(\pi), \dots, \sigma_{t-1}(\pi)) \quad \text{and} \quad \sigma_t(\pi) = \Phi_t(a_t(\pi)) \,.$$

Writing the definition of the regret,

$$\mathfrak{R}_n(\pi, x) = \sum_{t=1}^n \mathcal{L}_t(a_t(\pi)) - \min_{a \in [k]} \sum_{t=1}^n \mathcal{L}_t(a) \,.$$

The second term is constant and, as we mentioned already, $a \mapsto \mathcal{L}_t(a)$ is continuous. So it remains to check that a_t is continuous for each t. This follows by induction. The definition of the product topology means that for any fixed $\sigma_1, \dots, \sigma_{t-1}$ and $b \in [k]$, the set

$$U_b(\sigma_1, \dots, \sigma_{t-1}) = \{\pi : \pi(\sigma_1, \dots, \sigma_{t-1}) = b\}$$

is open in Π_{D}. Let ϵ denote the empty tuple. Then $a_1^{-1}(b) = U_b(\epsilon)$ is open in Π_{D}. We confirm that a_2 is continuous and leave the rest to the reader. That a_2 is continuous follows by writing

$$a_2^{-1}(c) = \bigcup_{b=1}^k \left(U_b(\epsilon) \cap U_c(\Phi_1(b))\right) \,.$$

Hence $\pi \mapsto \mathfrak{R}_n(\pi, x)$ is continuous and also measurable with respect to the Borel σ-algebra on Π_{D}. Then let $f(\mu, \nu)$ be given by

$$f(\mu, \nu) = \int_{\Pi_{\mathrm{D}}} \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x) d\mu(\pi) = \int_{\mathcal{X}^n} \int_{\Pi_{\mathrm{D}}} \mathfrak{R}_n(\pi, x) d\mu(\pi) d\nu(x) \,,$$

1. Sion's theorem is more general, it only assumes that f is quasiconvex/quasiconcave in each argument and upper/lower semicontinuous respectively.

where the exchange of integrals is justified by Fubini's theorem, which is applicable because the regret is bounded in $[-n, n]$. Clearly f is linear in both arguments. We now claim that both $\mu \mapsto f(\mu, \nu)$ and $\nu \mapsto f(\mu, \nu)$ are continuous. To see that $\nu \mapsto f(\mu, \nu)$ is continuous, note that $x \mapsto \int_{\Pi_{\rm D}} d\mu(\pi)\mathfrak{R}_n(\pi, x)$ is a $\mathcal{X}^n \to [-n, n]$ continuous map owning to the choice of the discrete topology on $\mathcal{X}^n$. Since $\mathscr{Q} \subset \mathscr{P}_r(\mathcal{X}^n)$ is equipped with the weak*-topology, this implies the continuity of $\nu \mapsto f(\mu, \nu)$. The argument for the continuity of $\mu \mapsto f(\mu, \nu)$ is similar: In particular, first note that $\pi \mapsto \int_{\mathcal{X}^n} d\nu(x)\mathfrak{R}_n(\pi, x)$ is a $\Pi_{\rm D} \to [-n, n]$ continuous map, since owning to the choice of $\mathscr{Q}$, the integral with respect to ν is a finite sum, and we have already established that for $x \in \mathcal{X}^n$ fixed, $\pi \mapsto \mathfrak{R}_n(\pi, x)$ is a $\Pi_{\rm D} \to [-n, n]$ continuous map. Again, the choice of the weak*-topology on $\mathscr{P}_r(\Pi_{\rm D})$ implies the desired continuity.

The final step is to note that for each policy $\mu \in \mathscr{P}_r(\Pi_{\rm D})$ there exists a policy $\pi \in \Pi$ such that for all $x \in \mathcal{X}^n$,

$$\mathfrak{R}_n(\pi, x) = \int_{\Pi_{\rm D}} \mathfrak{R}_n(\pi_d, x) d\mu(\pi_d)\,.$$

In particular, it is not hard to show that π can be defined through $\pi(a_1, \phi(a_1, x), \ldots, a_t, \phi(a_t, x))_a = \mathbb{P}_{\mu,x}(A_{t+1} = a | A_1 = a_1, \ldots, A_t = a_t)$, where $\mathbb{P}_{\mu,x}$ is the distribution resulting from using μ on the environment x. Here, the right-hand side is well defined (as a completely regular measure) because of the choice of $\mathcal{A}$. That π is well defined and is suitable follows from the definitions. Putting things together,

$$\begin{aligned}
\mathfrak{R}_n^* &= \inf_{\pi\in\Pi} \sup_{x\in\mathcal{X}^n} \mathfrak{R}_n(\pi, x) \le \min_{\mu\in\mathscr{P}_r(\Pi_{\rm D})} \sup_{x\in\mathcal{X}^n} \int_{\Pi_{\rm D}} \mathfrak{R}_n(\pi, x) d\mu(\pi) \overset{(a)}{\le} \min_{\mu\in\mathscr{P}_r(\Pi_{\rm D})} \sup_{\nu\in\mathscr{Q}} f(\mu, \nu) \\
&\overset{(b)}{=} \sup_{\nu\in\mathscr{Q}} \min_{\mu\in\mathscr{P}_r(\Pi_{\rm D})} f(\mu, \nu) \overset{(c)}{=} \sup_{\nu\in\mathscr{Q}} \min_{\pi\in\Pi_{\rm D}} \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x) \\
&\overset{(d)}{=} \sup_{\nu\in\mathscr{Q}} \min_{\pi\in\Pi} \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x) \overset{(e)}{=} \mathfrak{BR}_n^*(\mathscr{Q})\,,
\end{aligned} \tag{8}$$

where in (a) we used the fact that the Dirac measures are in $\mathscr{Q}$, (b) follows from Sion's theorem. In (c) we used the fact that the Dirac measures in $\mathscr{P}_r(\Pi_{\rm D})$ are minimisers of $f(\cdot, \nu)$ for any ν, in (d) we used that $\Pi_{\rm D} \subset \Pi$ and, via a dynamic programming argument, that the deterministic policies from $\Pi_{\rm D}$ minimise the Bayesian regret. For (e), let $\mathbb{P}_{\pi\nu}$ be the joint induced by π and ν over $[k]^n \times \mathcal{X}^n$, $\mathbb{E}_{\pi\nu}$ the corresponding expectation and define $\mathfrak{r}_n(a, x) = \sum_{t=1}^n \mathcal{L}(a_t, x_t) - \min_{b\in[k]} \sum_{t=1}^n \mathcal{L}(b, x_t)$. Then, note that $\mathbb{P}_{\pi\nu}$ almost surely, $\mathbb{E}_{\pi\nu}[\mathfrak{r}_n(A, X)|X] = \mathfrak{R}_n(\pi, X)$, and thus, by the tower rule and because $\mathbb{P}_{\pi\nu,X} = \nu$ by assumption, $\int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x) = \mathfrak{BR}_n(\pi, \nu)$. That $\mathfrak{BR}_n^*(\mathscr{Q}) \le \mathfrak{R}_n^*$ follows from

$$\begin{aligned}
\mathfrak{R}_n^* &= \inf_{\pi\in\Pi} \sup_{x\in\mathcal{X}^n} \mathfrak{R}_n(\pi, x) = \inf_{\pi\in\Pi} \sup_{\nu\in\mathscr{Q}} \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x) \ge \sup_{\nu\in\mathscr{Q}} \inf_{\pi\in\Pi} \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x) \\
&= \mathfrak{BR}_n^*(\mathscr{Q})\,,
\end{aligned} \tag{9}$$

where the second equality used that for any fixed $\pi \in \Pi$, $\nu \mapsto \int_{\mathcal{X}^n} \mathfrak{R}_n(\pi, x) d\nu(x)$ is a linear functional on $\mathscr{P}_r(\mathcal{X}^n)$, which is thus maximised in the extreme points of $\mathscr{P}_r(\mathcal{X}^n)$, which are all the Dirac measures over $\mathcal{X}^n$. Combining Eqs. (8) and (9) gives the desired result.

Appendix B. Proof of Theorem 7

Using the fact that the total variation distance is upper bounded by the Hellinger distance (Tsybakov, 2008, Lemma 2.3) and the first inequality in Eq. (1),

$$\begin{aligned}\mathbb{E}_t[\Delta_t] &= \sum_{a:P^*_{ta}>0} P^*_{ta}\left(\mathbb{E}_t[X_{ta}] - \mathbb{E}_t[X_{ta} \mid A^* = a]\right) \\ &\leq \sum_{a:P^*_{ta}>0} P^*_{ta} \sqrt{\int_{[0,1]} \left(1 - \sqrt{\frac{d\mathbb{P}_{t,X_{ta}|A^*=a}}{d\mathbb{P}_{t,X_{ta}}}}\right)^2 d\mathbb{P}_{t,X_{ta}}} \qquad (10) \\ &\leq \sqrt{k^{1/2} \sum_{a:P^*_{ta}>0} (P^*_{ta})^{3/2} \int_{[0,1]} \left(1 - \sqrt{\frac{d\mathbb{P}_{t,X_{ta}|A^*=a}}{d\mathbb{P}_{t,X_{ta}}}}\right)^2 d\mathbb{P}_{t,X_{ta}}}\,. \qquad (11)\end{aligned}$$

Eq. (10) is true because the total variation distance is upper bounded by the Hellinger distance. Eq. (11) uses Cauchy-Schwarz and the fact that $\sum_{a=1}^k (P^*_{ta})^{1/2} \leq k^{1/2}$, which also follows from Cauchy-Schwarz. The next step is to apply Bayes law to the square root term. There are no measurability problems because both X_{ta} and A^* live in Polish spaces (Ghosal and van der Vaart, 2017).

$$\begin{aligned}\int_{[0,1]} \left(1 - \sqrt{\frac{d\mathbb{P}_{t,X_{ta}|A^*=a}}{d\mathbb{P}_{t,X_{ta}}}}\right)^2 d\mathbb{P}_{t,X_{ta}} &= \int_{[0,1]} \left(1 - \sqrt{\frac{\mathbb{P}_t(A^*=a \mid X_{ta})(x)}{\mathbb{P}_t(A^*=a)}}\right)^2 d\mathbb{P}_{t,X_{ta}}(x) \\ &= \mathbb{E}_t\left[\left(1 - \sqrt{\frac{\mathbb{P}_t(A^*=a \mid X_{ta})}{\mathbb{P}_t(A^*=a)}}\right)^2\right] \\ &= \frac{1}{\sqrt{\mathbb{P}_t(A^*=a)}} \mathbb{E}_t\left[\frac{\left(\sqrt{\mathbb{P}_t(A^*=a)} - \sqrt{\mathbb{P}_t(A^*=a \mid X_{ta})}\right)^2}{\sqrt{\mathbb{P}_t(A^*=a)}}\right].\end{aligned}$$

Substituting the above into Eq. (11) and using the fact that $P_{ta} = P^*_{ta} = \mathbb{P}_t(A^* = a)$ yields

$$\begin{aligned}\mathbb{E}_t[\Delta_t] &\leq \sqrt{k^{1/2} \sum_{a:P_{ta}>0} P_{ta}\mathbb{E}_t\left[\frac{\left(\sqrt{\mathbb{P}_t(A^*=a)} - \sqrt{\mathbb{P}_t(A^*=a \mid X_{ta})}\right)^2}{\sqrt{\mathbb{P}_t(A^*=a)}}\right]} \\ &\leq \sqrt{k^{1/2} \sum_{a:P_{ta}>0} P_{ta}\mathbb{E}_t\left[\sum_{c:P^*_{tc}>0} \frac{\left(\sqrt{\mathbb{P}_t(A^*=c)} - \sqrt{\mathbb{P}_t(A^*=c \mid X_{ta})}\right)^2}{\sqrt{\mathbb{P}_t(A^*=c)}}\right]}\,,\end{aligned}$$

where the second inequality follows by introducing the sum over c. Finally, note that

$$D_F(p,q) = \sum_{c:p_c \neq q_c} \frac{\left(\sqrt{p_c} - \sqrt{q_c}\right)^2}{\sqrt{q_c}}\,.$$

The result follows from a direct computation using the independence of A_t and X_t under $\mathbb{P}_t$ (Lemma 26).

Appendix C. Cops and robbers

To further demonstrate the flexibility of the approach we consider this special case of bandits with graph feedback. In cops and robbers the learner observes the losses associated with all actions except the played action. Except for constant factors, this problem is no harder than the full information setting where all losses are observed. Cops and robbers is formalised in the partial monitoring framework by choosing $\Sigma = [0,1]^{k-1}$, $\mathcal{X} = [0,1]^k$, $\mathcal{L}(a,x) = x_a$ and

$$\Phi(a,x) = (x_1, \dots, x_{a-1}, x_{a+1}, \dots, x_k)\,.$$

Theorem 17 *The minimax regret of cops and robbers satisfies* $\mathfrak{R}_n^* \leq \sqrt{2n\log(k)}$.

This improves on the result by Alon et al. (2015) that $\mathfrak{R}_n^* \leq 5\sqrt{n\log(k)}$. We leave for the future the interesting question of whether or not this method recovers other known results for bandits with graph feedback. Theorem 17 follows immediately from Theorems 2 and 1, and the following lemma.

Lemma 18 *Thompson sampling for cops and robbers satisfies* $\mathbb{E}_t[\Delta_t] \leq \sqrt{2I_t(A^*;\Phi_t(A_t),A_t)}$ *almost surely for all* t.

Proof Fix $t \in [n]$ and let $G_t = \arg\max_a P_{ta}$. Here, we assume that we have already discarded a suitable set of measure zero, so that we do not need to keep repeating the qualification 'almost surely'. Then, subtracting and adding $\mathcal{L}_t(G_t)$, expanding the definitions and using that $P_t^* = P_t$,

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &= \sum_{a\neq G_t} P_{ta}\mathbb{E}_t[\mathcal{L}_t(a) - \mathcal{L}_t(G_t)] + \sum_{a \neq G_t} P_{ta}\mathbb{E}_t[\mathcal{L}_t(G_t) - \mathcal{L}_t(a) \mid A^* = a] \\
&\leq \sum_{a\neq G_t} P_{ta}\left(\sqrt{\frac{1}{2}\,\mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(G_t)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(G_t)}\right)} + \sqrt{\frac{1}{2}\,\mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(a)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(a)}\right)}\right) \\
&\leq \sqrt{\text{(A)}} + \sqrt{\text{(B)}}\,,
\end{aligned}$$

where the first inequality follows from grouping the terms that involve $\mathcal{L}_t(G_t)$ and those that involve $\mathcal{L}_t(a)$ and then using Pinsker's inequality (1), while the second follows from Cauchy-Schwarz and the definitions,

$$\begin{aligned}
\text{(A)} &= \frac{1-P_{tG_t}}{2}\sum_{a\neq G_t} P_{ta}\,\mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(G_t)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(G_t)}\right)\,, \\
\text{(B)} &= \frac{1-P_{tG_t}}{2}\sum_{a\neq G_t} P_{ta}\,\mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(a)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(a)}\right)\,.
\end{aligned}$$

The result is completed by bounding each term separately. Using that $1 - P_{tG_t} = \sum_{b \neq G_t} P_{tb}$,

$$
\begin{aligned}
(\mathrm{A}) &= \frac{1 - P_{tG_t}}{2} \sum_{a \neq G_t} P_{ta} \, \mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(G_t)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(G_t)}\right) \\
&= \frac{1}{2} \sum_{a \neq G_t} P_{ta} \sum_{b \neq G_t} P_{tb} \, \mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(G_t)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(G_t)}\right) \\
&\leq \frac{1}{2} \sum_{a \neq G_t} P_{ta} \sum_{b \neq G_t} P_{tb} \, \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)}\right) \\
&\leq \frac{1}{2} I_t(A^*; \Phi_t(A_t), A_t) \,,
\end{aligned}
$$

where the first inequality follows from the data processing inequality (for $b \neq G_t$, $\mathcal{L}_t(G_t)$ is a deterministic function of $\Phi_t(b)$) and the last from Lemma 26. The second term is bounded in almost the same way. Here we use the fact that $1 - P_{tG_t} \leq 1 - P_{ta}$ for all $a \in [k]$:

$$
\begin{aligned}
(\mathrm{B}) &= \frac{1 - P_{tG_t}}{2} \sum_{a \neq G_t} P_{ta} \, \mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(a)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(a)}\right) \\
&\leq \frac{1}{2} \sum_{a \neq G_t} (1 - P_{ta}) P_{ta} \, \mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(a)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(a)}\right) \\
&= \frac{1}{2} \sum_{a \neq G_t} P_{ta} \sum_{b \neq a} P_{tb} \, \mathrm{D}\left(\mathbb{P}_{t,\mathcal{L}_t(a)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\mathcal{L}_t(a)}\right) \\
&\leq \frac{1}{2} \sum_{a \neq G_t} P_{ta} \sum_{b \neq a} P_{tb} \, \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)}\right) \\
&\leq \frac{1}{2} I_t(A^*; \Phi_t(A_t), A_t) \,.
\end{aligned}
$$

Combining the previous displays and rearranging completes the proof. ■

Appendix D. Proof of Theorem 11

We need the following lemma, which characterises actions $c \in \mathcal{N}_{ab}$ as having loss vectors ℓ_c that are convex combinations of ℓ_a and ℓ_b.

Lemma 19 (Bartók et al. 2014) *For all actions $c \in \mathcal{N}_{ab}$ there exists an $\alpha \in [0, 1]$ such that $\ell_c = \alpha \ell_a + (1 - \alpha)\ell_b$.*

Proof [Theorem 11] In order to define the algorithm we first choose a subset $\mathcal{C} \subseteq [k]$ such that $\mathcal{C}$ contains no duplicate or degenerate actions and $\cup_{c \in \mathcal{C}} C_c = \Delta^{k-1}$. We assume additionally that P_t^* is constant on duplicate actions. Construct the parent function $\mathcal{P}_t$ on actions in $\mathcal{C}$ in the same way as Mario sampling. For $a \neq b$ let $\mathcal{T}_{ab} = (c_1, \ldots, c_m)$ be an ordering of

$$
\{c \in ([k] \setminus \mathcal{C}) \cup \{b\} : c = b \text{ or exists } \alpha \in (0, 1] \text{ with } \ell_c = \alpha \ell_a + (1 - \alpha)\ell_b\}
$$

ordered by decreasing α values and with $c_m = b$. In other words $\mathcal{T}_{ab}$ is a sequence of actions starting with duplicates of a, then actions c for which ℓ_c is a strict convex combination of ℓ_a and ℓ_b, with actions that are 'closer' to a sorted first. The last element of $\mathcal{T}_{ab}$ is b itself. Duplicates of b are not included in $\mathcal{T}_{ab}$. Let $\mathcal{T}_{aa}$ be the duplicates of a, excluding a, in an arbitrary order. Then define

$$\mathcal{P}'_t(c) = \begin{cases} \mathcal{T}_{c\mathcal{P}(c)}[1] & \text{if } c \in \mathcal{C} \setminus \{G_t\} \\ \mathcal{T}_{cc}[1] & \text{if } c = G_t \text{ and } \mathcal{T}_{cc} \neq \emptyset \\ \mathcal{T}_{ab}[i+1] & \text{if } c = \mathcal{T}_{ab}[i]\,. \end{cases}$$

Let W'_t be the water transfer operator using the tree generated by $\mathcal{P}'_t$ instead of $\mathcal{P}_t$ and $P_t = (W'_t)^k P^*_t$. Now we follow the proof of Theorem 10. Let $t \in [n]$ be fixed. We start by bounding $\mathbb{E}_t[\Delta_t]$ in terms of the expected information gain. Given $b \in \mathcal{C} \setminus \{G_t\}$ let $f_b : \mathcal{N}_{a\mathcal{P}_t(a)} \to \mathbb{R}$ be a function with

$$\sum_{c \in \mathcal{N}_{b\mathcal{P}_t(b)}} f_b(c, \Phi_t(c)) = \mathcal{L}_t(b) - \mathcal{L}_t(\mathcal{P}_t(b))\,,$$

which exists by the definition of local observability. By definition we may assume that $\|f_b\|_\infty \leq v$. Then

$$\mathbb{E}_t[\mathcal{L}_t(a)] = \mathbb{E}_t\left[\mathcal{L}_t(G_t) + \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} \sum_{c \in \mathcal{N}_{b\mathcal{P}_t(b)}} f_b(c, \Phi_t(c))\right]\,.$$

Therefore

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &\leq v \sum_{a=1}^k P^*_{ta} \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} \sum_{c \in \mathcal{N}_{b\mathcal{P}_t(b)}} \sqrt{2\,\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&\leq vk \sqrt{4 \sum_{a=1}^k P^*_{ta} \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} \sum_{c \in \mathcal{N}_{b\mathcal{P}_t(b)}} P^*_{ta}\,\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&\leq vk^{3/2} \sqrt{4 \sum_{a=1}^k P^*_{ta} \sum_{b \in \mathcal{A}_t(a) \setminus \{G_t\}} \sum_{c \in \mathcal{N}_{b\mathcal{P}_t(b)}} P_{tc}\,\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&\leq vk^{3/2} \sqrt{8 \sum_{a=1}^k P^*_{ta} \sum_{c=1}^k P_{tc}\,\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&= vk^{3/2} \sqrt{8 I_t(A^*; \Phi_t(A_t), A_t)}\,.
\end{aligned}$$

And the result follows from Theorem 2 and Theorem 1. ■

Appendix E. Proof of Theorem 12

Again Thompson sampling does not explore sufficiently often. The most straightforward correction is to simply add a small amount of forced exploration, which was also used in combination with

Exp3 in prior analysis of these games (Cesa-Bianchi et al., 2006). We let

$$P_t = (1-\gamma)P_t^* + \gamma \mathbf{1}/k\,, \tag{12}$$

where ties in the $\arg\max$ that defines P_t^* are broken by prioritising Pareto optimal actions, which means that $P_{ta}^* = 0$ for all degenerate actions. As usual, the crucial step is to bound the expected 1-step regret in terms of the information gain.

Lemma 20 *For the policy playing according to Eq. (12) it holds almost surely that*

$$\mathbb{E}_t[\Delta_t] \leq \gamma + kv\sqrt{\frac{2I_t(A^*;\Phi_t(A_t),A_t)}{\gamma}}\,.$$

Proof Let $a_\circ$ be an arbitrary fixed Pareto optimal action and for each Pareto optimal action a let $f_a : [k] \times \Sigma \to \mathbb{R}$ be a function with $\|f_a\|_\infty \leq v$ such that

$$\sum_{c=1}^k f_a(c, \Phi(c,x)) = \mathcal{L}(a,x) - \mathcal{L}(a_\circ, x) \quad \text{for all } x \in [d]\,.$$

The next step is to decompose the expected loss in terms of f:

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &= \sum_{a=1}^k P_{ta}\mathbb{E}_t[\mathcal{L}_t(a)] - \sum_{a=1}^k P_{ta}^*\mathbb{E}_t[\mathcal{L}_t(a) \mid A^* = a] \\
&\leq \gamma + \sum_{a=1}^k P_{ta}^* \left(\mathbb{E}_t[\mathcal{L}_t(a)] - \mathbb{E}_t[\mathcal{L}_t(a) \mid A^* = a]\right) \\
&= \gamma + \sum_{a=1}^k P_{ta}^* \left(\mathbb{E}_t[\mathcal{L}_t(a) - \mathcal{L}_t(a_\circ)] - \mathbb{E}_t[\mathcal{L}_t(a) - \mathcal{L}_t(a_\circ) \mid A^* = a]\right) \\
&= \gamma + \sum_{a=1}^k P_{ta}^* \left(\mathbb{E}_t\left[\sum_{c=1}^k f_a(c, \Phi_t(c))\right] - \mathbb{E}_t\left[\sum_{c=1}^k f_a(c,\Phi_t(c)) \,\middle|\, A^* = a\right]\right)\,,
\end{aligned}$$

where the inequality follows from the definition of P_t and the fact that losses are bounded in $[0,1]$. Then, by Pinsker's inequality (1),

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &\leq \gamma + v\sum_{c=1}^k\sum_{a=1}^k P_{ta}^*\sqrt{2\,\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&\leq \gamma + v\sqrt{2k\sum_{a=1}^k P_{ta}^* \sum_{c=1}^k \mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&\leq \gamma + kv\sqrt{\frac{2}{\gamma}\sum_{a=1}^k P_{ta}^* \sum_{c=1}^k P_{tc}\,\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(c)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(c)}\right)} \\
&= \gamma + kv\sqrt{\frac{2I_t(A^*;\Phi_t(A_t),A_t)}{\gamma}}\,,
\end{aligned}$$

where the first inequality follows from Pinsker's inequality (1), the second from Cauchy-Schwarz, the third because $1 \leq kP_{tc}/\gamma$ for all c. The last term follows from Lemma 26. ■

Proof [Theorem 12] By the previous lemma and Corollary 4,

$$\mathfrak{BR}_n \leq n\gamma + kv\sqrt{\frac{2n\log(k)}{\gamma}} \leq 3(nkv)^{2/3}(\log(k)/2)^{1/3}\,,$$

where we choose $\gamma = n^{-1/3}(kv)^{2/3}(\log(k)/2)^{1/3}$ and note that when $\gamma > 1$ the claim in the theorem is immediate. ■

Appendix F. The water transfer operator

Here we explain in more detail the water transfer operator defined by Algorithm 2 and provide the proof of Lemma 14. An example with $k = 6$ is illustrated below.

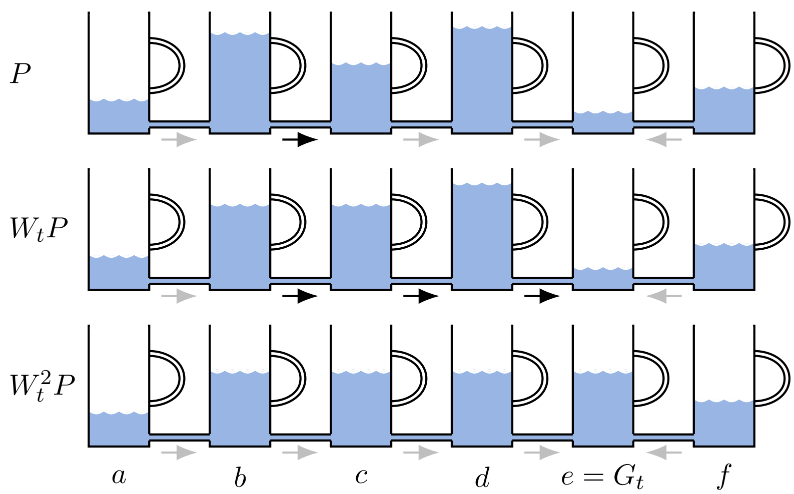

Figure 1: Water transfer process

The mugs correspond to actions and are connected at the bottom with valves that default to being closed. The total volume of water sums to 1. Arrows correspond to edges in the tree. The dark arrows indicate which valves are open in each iteration and show the direction of flow. In the first application of W_t, mug c is anomalous and the water in mugs b and c is averaged. Imagine opening the valve connecting b and c. The water in a is too low to be included in the average. In the second application, the water in mugs b, c, d and e is averaged. Further applications of W_t have no effect because there are no anomalous actions.

Remark 21 *Another way to think about the application of W_t to P is as follows. First the anomalous action a is identified, if it exists. Then water flows continuously into a from the set of descendants of a that contain more water than a until a is no longer anomalous.*

Proof [Lemma 14] To begin, notice that every application of the water transfer operator reduces the number of anomalous actions by at least one because: (1) If a is selected by Algorithm 2 then a is not anomalous in $W_t P$ and (2) only actions that were anomalous in P can be anomalous in $W_t P$. Since there are at most k anomalous actions in any P, the water transfer operator ceases to have any affect after more than k operations. Hence $Q_a \geq \max_{b \in \mathcal{D}_t(a)} Q_b$ for all a and the second part follows. For the first part we show that the loss of $W_t P$ is always smaller than P. Let $\bar{L}(b) = \mathbb{E}_t[\mathcal{L}_t(b)]$ for $b \in [k]$ and $a \in [k]$ be the anomalous action in P selected by the algorithm. Then let $\mathcal{C} = \{b \in [k] : (W_t P)_b \neq P_b\}$ be the set of actions for which the distribution is changed. By the definition of the tree, $\bar{L}(b) \geq \bar{L}(a)$ for any $b \in \mathcal{C}$,

$$\sum_{b=1}^{k} (P_b - (W_t P)_b)\bar{L}(b) = (P_a - (W_t P)_a)\bar{L}(a) + \sum_{b \in \mathcal{C}, b \neq a} (P_b - (W_t P)_b)\bar{L}(b)$$
$$= \sum_{b \in \mathcal{C}, b \neq a} (P_b - (W_t P)_b)(\bar{L}(b) - \bar{L}(a)) \geq 0\,,$$

which shows that W_t decreases the expected loss. For the last part, notice that during each iteration of the water transfer operator the update occurs by averaging the contents of a number of mugs so that all have the same level (Fig. 1). Once a group of mugs have been averaged together, subsequently they are always averaged together. It follows that after every iteration the actions $[k]$ can be partitioned so that the level in each partition is the average of P_a. Suppose that a is in partition $S \subseteq [k]$. Then $Q_a = \frac{1}{|S|} \sum_{b \in S} P_b \geq P_a / k$. ∎

Appendix G. Failure of Thompson sampling for partial monitoring

The following example with $k = 3$ and $d = 2$ illustrates the failure of Thompson sampling for locally observable non-degenerate partial monitoring games. The game is a toy 'spam filtering' problem where the learner can either classify an email as spam/not spam or pay a small cost for the true label. The functions Φ and $\mathcal{L}$ are represented by the tables below, with the learner choosing the rows and adversary the columns.

Losses $\mathcal{L}$	NOT SPAM	SPAM
SPAM	1	0
NOT SPAM	0	1
UNKNOWN	c	c

Signals Φ	NOT SPAM	SPAM
SPAM	$\perp$	$\perp$
NOT SPAM	$\perp$	$\perp$
UNKNOWN	NOT SPAM	SPAM

Figure 2: The 'spam' partial monitoring game. For $c < 1/2$ the game is locally observable and non-degenerate. For $c = 1/2$ the game is locally observable, but degenerate. For $c > 1/2$ the game is not locally observable, but is globally observable. For $c = 0$ the game is trivial.

The learner only elicits meaningful feedback in the spam game by paying a cost of c to observe the true label. For appropriately chosen c and prior, we will see that Thompson sampling never chooses the revealing action, cannot learn, and hence suffers linear regret. Let $c > 0$ and ν be the mixture of two Dirac's: $\nu = \frac{1}{2}\delta^n_{\text{SPAM}} + \frac{1}{2}\delta^n_{\text{NOT SPAM}}$, where δ^n_i is the Dirac measure on $(i, i, \dots, i)$. With these choices the optimal action is almost surely either SPAM or NOT SPAM. Since choosing these actions does not reveal any information, the posterior is equal to the prior and Thompson sampling plays these two actions uniformly at random. Clearly this leads to linear regret relative to the optimal policy that plays the exploratory action once to identify the adversary and plays optimally for the remainder. Since this result holds for any strictly positive cost, it also shows that Thompson sampling does not work for globally observable games.

Appendix H. Structural lemmas for partial monitoring

Lemma 22 *Let $a, b \in [k]$ be distinct actions in a non-degenerate game and $u \in C_a$. Then there exists an action $c \in \mathcal{N}_b \setminus \{b\}$ such that $\langle \ell_b - \ell_c, u\rangle \geq 0$. Furthermore, if $u \notin C_b$, then $\langle \ell_b - \ell_c, u\rangle > 0$.*

Proof Let w be a point in the relative interior of C_b, which means that $\langle \ell_b, w\rangle < \min_{c \neq b}\langle \ell_c, w\rangle$. Now let $c \in \mathcal{N}_b \setminus \{b\}$ be an action such that $v = u + \alpha(w - u) \in C_b \cap C_c$ for some $\alpha \in [0, 1)$, which exist because C_b is closed convex set and hence $\{u + \alpha(w - u) : \alpha \in \mathbb{R}\} \cap C_b$, which is nonempty, must be a closed segment. Let $f(x) = \langle \ell_b - \ell_c, u + x(w - u)\rangle$. By definition, $f(\alpha) = 0$ and $f(1) < 0$. Since f is linear it follows that $f(0) = \langle \ell_b - \ell_c, u\rangle \geq 0$. The second part follows because if $u \notin C_b$, then $\alpha > 0$, which means that $f(0) > f(\alpha) = 0$. ■

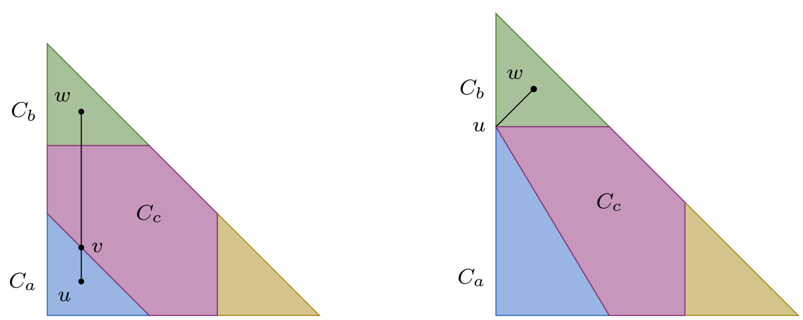

Figure 3: Illustration for the proof of Lemma 22. The bottom left region is C_a and $u \in C_a$ so that a minimises $\mathbb{E}_{x\sim u}[\mathcal{L}(a, x)]$. The lemma proves that for the situation in the left figure: $\mathbb{E}_{x\sim u}[\mathcal{L}(c, x)] < \mathbb{E}_{x\sim u}[\mathcal{L}(b, x)]$. The strict inequality is replaced by an equality if $u \in C_a \cap C_b$ as in the right figure, when $u = v$.

Lemma 23 *Consider a non-degenerate game and let $u \in \Delta^{k-1}$ and $V = \{a : u \in C_a\}$ and $E = \{(a, b) \in V : a \text{ and } b \text{ are neighbours}\}$. Then the graph (V, E) is connected.*

Proof This must be a known result about the facet graph of convex polytopes. We give a dimension argument. You may find Fig. 4 useful. Let $B_\varepsilon(x) = \{y \in \Delta^{d-1} : \|y - x\|_2 \leq \varepsilon\}$, $\mathcal{H}_d$ be the d-dimensional Hausdorff measure and ri be the relative interior operator. Since the cells are closed, there exists an $\varepsilon > 0$ such that $B_\varepsilon(u) \cap C_c = \emptyset$ for all $c \notin V$. Then let $a^*(v) = \{a \in [k] : v \in C_a\}$ be the set of actions that are optimal at $v \in \Delta^{d-1}$. It is easy to see that if $a^*(v) = \{a, b\}$ for some $v \in \Delta^{d-1}$, then a and b are neighbours. Let $N = \{v \in \Delta^{d-1} : |a^*(v)| > 2\}$, which by the assumption that there are no duplicate/degenerate actions has dimension at most $d-3$ and hence $\mathcal{H}_{d-2}(N) = 0$. Let $a, b \in V$ be distinct and $v, w \in B_\varepsilon(u)$ be such that $B_\delta(v) \subset C_a$ and $B_\delta(w) \subset C_b$ for some $\delta > 0$, which by definition means that the interval $[v, w] \cap C_c = \emptyset$ for all $c \notin V$. Let A be the affine space containing v with normal $v - w$ and $P = \{\arg\min_{x \in A} \|x - y\|_2 : y \in N\}$ be the projection of N onto A. Since projection onto a plane cannot increase the Hausdorff measure, $\mathcal{H}_{d-2}(P) = 0$. On the other hand, the fact that $A \cap B_\delta(v)$ has dimension $d-2$ means that $\mathcal{H}_{d-2}(A \cap B_\delta(v)) > 0$. Therefore $\mathcal{H}_{d-2}(B_\delta(v) \cap (A \setminus P)) > 0$ and hence there exists an $x \in B_\delta(v) \cap A$ and $y = x + w - v \in B_\delta(w)$ such that $[x, y] \cap N = \emptyset$. Then the set $\cup_{z \in [x,y]} a^*(u)$ forms a connected path in V between a and b. ∎

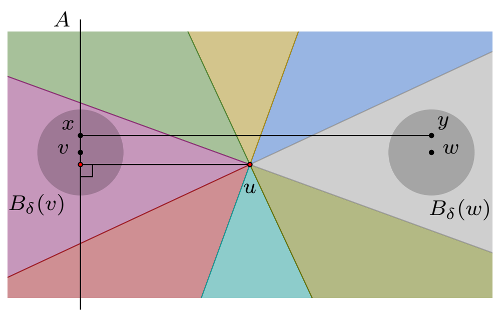

Figure 4: Illustration for the proof of Lemma 23 when $d = 3$. The whole region shown is a subset of $B_\varepsilon(u)$. The set N in this case consists only of u, which has 1-dimensional Hausdorff measure zero. The cells crossed by the interval $[x, y]$ form the path between a and b in V.

Proof [Lemma 13] By definition there are no edges starting from G_t. By Lemma 22, for all $a \notin V_t$ there is a neighbour $b \in \mathcal{N}_a$ with strictly smaller loss, $\mathbb{E}_t[\mathcal{L}_t(b)] < \mathbb{E}_t[\mathcal{L}_t(a)]$. Hence the definition of $\mathcal{P}_t(a)$ ensures there are no cycles and that every path starting from $a \notin V_t$ eventually leads to V_t. Then by Lemma 22 the graph (V_t, E_t) is connected, which means that for $a \in V_t$ the parent $\mathcal{P}_t(a)$ is a vertex $b \in V_t$ that is closest to G_t. Hence all paths lead to G_t. ∎

The next two lemmas bound on the supremum norms of the estimation functions. The first is restricted to the non-degenerate case where the result was already known and the second holds for all globally observable games.

Lemma 24 (Lattimore and Szepesvári (2019), Lemma 9) *For locally observable non-degenerate games, the function f in Eq. (4) can be chosen so that $\|f\|_\infty \leq d + 1$.*

Lemma 25 *If $(\Phi, \mathcal{L})$ is globally observable, then for each pair of neighbours a and b there exists a function f satisfying Eq. (4) such that $\|f\|_\infty \leq d^{1/2}(1+k)^{d/2}$. If $(\Phi, \mathcal{L})$ is also locally observable, then f can be chosen so that $f(c, \sigma) = 0$ for all $c \notin \mathcal{N}_{ab}$.*

Proof We prove only the first part. The proof for locally observable games is the same, but the signal matrices defined below are restricted to $c \in \mathcal{N}_{ab}$. Assume without loss of generality that $\Sigma = [d]$ and $d \geq 2$. For $c \in [k]$ define $S_c \in \{0,1\}^{d \times d}$ to be the matrix with $(S_c)_{\sigma x} = 1$ if $\Phi(c, x) = \sigma$.

Then let $S \in \{0,1\}^{d \times dk}$ be formed by horizontally stacking the matrices $\{S_c : c \in [k]\}$. By the definition of local observability it holds that $\ell_a - \ell_b \in \mathrm{im}(S)$. Let S^+ be the Moore-Penrose pseudo-inverse of S and let $w = S^+(\ell_a - \ell_b)$, which satisfies $Sw = \ell_a - \ell_b$. Then f can be chosen so that $\|f\|_\infty = \|w\|_\infty \leq \|w\|_2$. Since losses are bounded in $[0,1]$ we have $\|w\|_2 \leq \|S^+\|_2 \|\ell_a - \ell_b\|_2 \leq d^{1/2}\sigma_{\min}^{-1}$, where $\sigma_{\min}$ is the smallest nonzero singular value of S. Hence we need to lower bound the smallest nonzero eigenvalue of $B = SS^\top$, which is a $d \times d$ matrix with entries in $\{0, 1, \ldots, k\}$. The characteristic polynomial of B is $\chi(\lambda) = \det(\lambda I - B) = \sum_{i=0}^d a_i \lambda^d$, where $a_d = 1$ and, up to a sign, a_i is the sum principle minors of B of size $d - i$. Since the geometric mean is smaller than the arithmetic mean, for matrix $A \in [0,k]^{i \times i}$ it holds that $\det(A) \leq (\mathrm{tr}(A)/i)^i \leq k^i$. Hence,

$$|a_{d-i}| \leq \binom{d}{i} k^i \,.$$

By the binomial theorem,

$$\sum_{i=0}^d |a_i| \leq \sum_{i=0}^d \binom{d}{i} k^i = (1+k)^d \,.$$

Let $i_{\min} = \min\{i : a_i \neq 0\}$ and suppose that $\lambda > 0$ is the smallest nonzero root of χ, which must be positive. Then

$$0 = |\chi(\lambda)| = \left|\sum_{i=i_{\min}}^d a_i \lambda^i\right| = \lambda^{i_{\min}} \left|\sum_{i=i_{\min}}^d a_i \lambda^{i - i_{\min}}\right| \geq \lambda^{i_{\min}} \left|1 - (1+k)^d \lambda\right| \,,$$

where we used the fact that (a_i) are integer-valued. Therefore $\lambda \geq (1+k)^{-d}$, which means that $\|S^+\|_2 \leq (1+k)^{d/2}$ and hence $\|f\|_\infty \leq d^{1/2}(1+k)^{d/2}$. ■

Appendix I. Figures and examples

Finite partial monitoring example Below is a 4-action finite-outcome, finite-action partial monitoring game with feedback set $\Sigma = \{\perp, ☂, ☃, ☀\}$. The left table shows the loss function and the right shows the signal function. By staying indoors you cannot evaluate the quality of the snow, but climbing or skiing in poor conditions is no fun.

Losses $\mathcal{L}$	SUN	SNOW	RAIN
SKI	3/4	0	1
CLIMB	0	3/4	1
MATH	1/2	1/2	1/4
RAINDANCE	1	1	0

Signals Φ	SUN	SNOW	RAIN
SKI	☀	☃	☂
CLIMB	☀	☃	☂
MATH	$\perp$	$\perp$	$\perp$
RAINDANCE	☀	☃	☂

Figure 5: Example finite partial monitoring game

The following figure shows the cell decomposition for the above game, Δ^{d-1} is parameterised by $(p, q, 1-p-q)$. In this game all actions a Pareto optimal. All actions are neighbours of MATH and otherwise CLIMB and SKI are neighbours and MATH and RAINDANCE. The game is locally observable because the loss of all actions can be identified by playing that action, except for MATH, the losses of which can be identified by playing any of its neighbours.

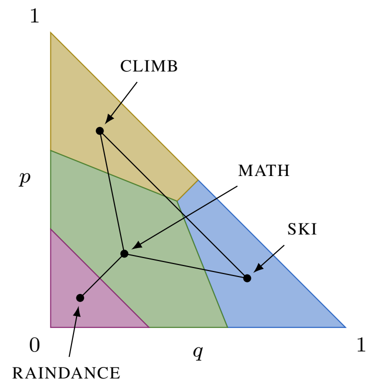

Figure 6: Cell decomposition for the game described above where $d = 3$. The figure shows Δ^{d-1} projected onto the plane by the parameterisation $(p, q, 1 - p - q)$. All actions are Pareto optimal, so their cells have dimension $d - 1 = 2$. The intersection of the cells of neighbouring actions are the lines shared by the cells, which have dimension 1.

Tree construction The figure depicts the cell decomposition for a partial monitoring game with seven actions and the tree structure defined in Lemma 13. Arrows indicate the parent relationship. All paths leading towards G_t. Red nodes are descendants of a. Blue nodes are ancestors. Dotted lines indicate connections in the neighbourhood graph that are not part of the tree.

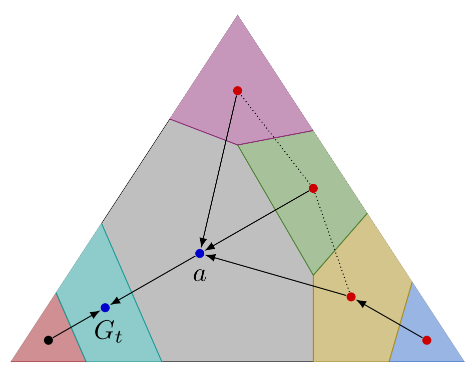

Figure 7: Tree construction

Appendix J. Technical calculation

Lemma 26 *Let $P_{ta} = \mathbb{P}_t(A_t = a)$. Then the following hold almost surely:*

$$\mathbb{E}_t[D_F(P^*_{t+1}, P^*_t)] = \sum_{a=1}^k P_{ta} \mathbb{E}_t \left[D_F(\mathbb{P}_{t,A^*|\Phi_t(a)}, \mathbb{P}_{t,A^*}) \right] ,$$

$$I_t(A^*; \Phi_t(A_t), A_t) = \sum_{a=1}^k P^*_{ta} \sum_{b=1}^k P_{tb}\, \mathbb{E}_t \left[\mathrm{D} \left(\mathbb{P}_{t,\Phi_t(b)|A^*=a} \,\middle\|\, \mathbb{P}_{t,\Phi_t(b)} \right) \right] .$$

Proof Recall that $P^*_{t+1} = \mathbb{P}_{t+1}(A^* \in \cdot) = \mathbb{P}_t(A^* \in \cdot \mid A_t, \Phi_t(A_t))$. Then

$$\begin{aligned}
\mathbb{E}_t[D_F(\mathbb{P}_{t,A^*|A_t,\Phi_t(A_t)}, \mathbb{P}_{t,A^*})] &= \mathbb{E}_t \left[\mathbb{E}_t[D_F(\mathbb{P}_{t,A^*|A_t,\Phi_t(A_t)}, \mathbb{P}_{t,A^*}) \mid A_t] \right] \\
&= \mathbb{E}_t \left[\mathbb{E}_t[D_F(\mathbb{P}_{t,A^*|\Phi_t(A_t)}, \mathbb{P}_{t,A^*}) \mid A_t] \right] \\
&= \sum_{a=1}^k P_{ta} \mathbb{E}_t[D_F(\mathbb{P}_{t,A^*|\Phi_t(a)}, \mathbb{P}_{t,A^*}) \mid A_t = a] \\
&= \sum_{a=1}^k P_{ta} \mathbb{E}_t[D_F(\mathbb{P}_{t,A^*|\Phi_t(a)}, \mathbb{P}_{t,A^*})] ,
\end{aligned}$$

where in the second and fourth inequalities we used the independence of A_t and X under $\mathbb{P}_t$. The second part of the lemma follows from an identical argument. ∎