# Exploration by Optimisation in Partial Monitoring

Tor Lattimore and Csaba Szepesvári

DeepMind

**Abstract**

We provide a simple and efficient algorithm for adversarial $k$-action $d$-outcome non-degenerate locally observable partial monitoring games for which the $n$-round minimax regret is bounded by $6(d+1)k^{3/2}\sqrt{n\log(k)}$, matching the best known information-theoretic upper bound. The same algorithm also achieves near-optimal regret for full information, bandit and globally observable games.

## 1 Introduction

Partial monitoring is a generalisation of the bandit framework that decouples the loss and the observations. The framework is sufficiently rich to model bandits, linear bandits, full information games, dynamic pricing, bandits with graph feedback and many problems between and beyond these examples. For positive integer $m$ let $[m] = \{1, \ldots, m\}$. A finite adversarial partial monitoring game is determined by a signal matrix $\Phi \in \Sigma^{k\times d}$ and loss matrix $\mathcal{L} \in [0,1]^{k\times d}$ where $\Sigma$ is an arbitrary finite set. Both $\Phi$ and $\mathcal{L}$ are known to the learner. The game proceeds over $n$ rounds. First the adversary chooses a sequence $(x_t)_{t=1}^n$ with $x_t \in [d]$. In each round $t \in [n]$ the learner chooses an action $A_t \in [k]$, suffers loss $\mathcal{L}_{A_t x_t}$, but only observes the signal $\sigma_t = \Phi_{A_t x_t}$. The regret is

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

where the expectation is with respect to the randomness in the actions and $\pi$ is the policy of the learner mapping action/observation sequences to distributions over the actions. When they are clear from the context, we omit $\pi$ and $(x_t)_{t=1}^n$ from the expression for the regret. The minimax regret is

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

which only depends on $n$, $\Phi$ and $\mathcal{L}$ and is our main interest. Our main contribution is a simple and efficient algorithm for finite non-degenerate locally observable partial monitoring games for which

$$\mathfrak{R}_n^* \leq 6k^{3/2}(d+1)\sqrt{n\log(k)}. \tag{1}$$

The same algorithm is adaptive to other types of game, achieving near-optimal regret for globally observable games, a regret of $\sqrt{2nk\log(k)}$ for bandits and $\sqrt{2n\log(k)}$ for full information games.

**Related work** Partial monitoring goes back to the work by Rustichini [1999], who derived Hannan consistent policies. There has been significant effort in understanding the dependence of the regret on the horizon. The key result is the classification theorem, showing that all finite partial monitoring games lie in one of four categories as illustrated in Table 1. The classification theorem also gives a procedure to decide into which category a given game belongs. Since the game is known in advance, there is no need to learn the classification of the game. This result has been pieced together over about a decade by a number of authors [Cesa-Bianchi et al., 2006; Foster and Rakhlin, 2012; Antos et al., 2013; Bartók et al., 2014; Lattimore and Szepesvári, 2019a]. Ironically, the 'easy' games present the greatest challenge for algorithm design and analysis.

| | |
|---|---|
| Trivial | $\mathfrak{R}_n^* = 0$ |
| Easy | $\mathfrak{R}_n^* = \Theta(n^{1/2})$ |
| Hard | $\mathfrak{R}_n^* = \Theta(n^{2/3})$ |
| Hopeless | $\mathfrak{R}_n^* = \Omega(n)$ |

**Table 1:** Classification of finite partial monitoring

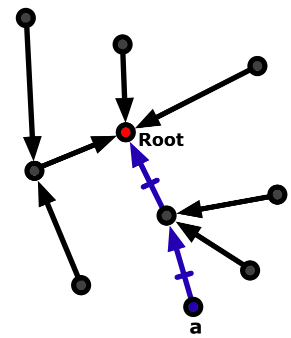
The best known bound for an efficient algorithm for 'easy' games is $\mathfrak{R}_n \leq C(\Phi, \mathcal{L})\sqrt{n \log(n)}$, where the constant $C(\Phi, \mathcal{L})$ can be arbitrarily large, even for fixed $k$ and $d$ [Foster and Rakhlin, 2012; Lattimore and Szepesvári, 2019a]. Furthermore, the algorithms achieving this bound are complicated to analyse and the proofs yield little insight into the structure of partial monitoring. Recently we proved that for the 'non-degenerate' (defined later) subset of easy games, the minimax regret is at most $\mathfrak{R}_n^* \leq (d+1)k^{3/2}\sqrt{8n\log(k)}$ [Lattimore and Szepesvári, 2019b]. Unfortunately, however, our proof non-constructively appealed to minimax duality and the Bayesian regret analysis techniques by Russo and Van Roy [2016]. No algorithm was provided, a deficiency we now resolve.

Partial monitoring has been studied in a variety of contexts. For example, bandits with graph feedback [Alon et al., 2015] and a linear feedback setting [Lin et al., 2014]. Some authors also consider a variant of the regret that refines the notion of optimality in hopeless games [Rustichini, 1999; Mannor and Shimkin, 2003; Perchet, 2011; Mannor et al., 2014]. Our focus is on the adversarial setting, but the stochastic setup is also interesting and is better understood [Bartók et al., 2011; Vanchinathan et al., 2014; Komiyama et al., 2015].

**Approach** Our algorithms are based on exponential weights with unbiased importance-weighted loss difference estimators [Freund and Schapire, 1997]. Crucially, the algorithms do not sample from the distribution proposed by exponential weights. Instead, they solve a convex optimisation problem to find an unbiased loss difference estimator and new distribution over actions for which the loss cannot be much larger than the proposal distribution and the 'stability' term in the bound of exponential weights is minimised. We then prove that the value of the optimisation problem appears in the resulting regret guarantee and provide upper bounds for different classes of games. The most challenging aspect is to prove the existence of a suitable exploration distribution for locally observable non-degenerate games, which follows by combining a minimax theorem with insights from the Bayesian setting. The idea to modify the distribution proposed by exponential weights is reminiscent of the work by McMahan and Streeter [2009] for bandits with expert advice, though the situation here is rather different.

## 2 Notation and concepts

We write $\mathbf{0}$ and $\mathbf{1}$ for the column vectors of all zeros and all ones respectively. For a positive semidefinite matrix $A$ and vector $x$, we let $\|x\|_A^2 = x^\top A x$ and $\operatorname{diag}(x)$ be the diagonal matrix with $x$ on the diagonal. We use $\|A\|_\infty = \max_{ij} |A_{ij}|$ for the (entrywise) maximum norm of $A$, which we also use for the special case that $A$ is a vector. The minimum entry of a matrix is $\min(A) = \min_{ij} A_{ij}$. The standard basis vectors are $e_1, \ldots, e_d$; we use the same symbols regardless of the dimension, which should be clear from the context in all cases.




**In-trees** An in-tree with vertex set $[k]$ is a set $\mathcal{T} \subseteq [k] \times [k]$ representing the edges of a directed tree with vertices $[k]$. Furthermore, we assume there is a root vertex denoted by $\operatorname{root}_\mathcal{T} \in [k]$ such that for all $a \in [k]$ there is a directed path $\operatorname{path}_\mathcal{T}(a) \subseteq \mathcal{T}$ from $a$ to the root. The path from the root is the empty set: $\operatorname{path}_\mathcal{T}(\operatorname{root}_\mathcal{T}) = \emptyset$. The figure depicts an in-tree over $k = 10$ vertices. The blue (barred) path is $\operatorname{path}_\mathcal{T}(a)$.

**Partial monitoring** Throughout we fix a partial monitoring game $\mathcal{G} = (\Phi, \mathcal{L})$ with loss matrix $\mathcal{L} \in [0,1]^{k \times d}$ and signal matrix $\Phi \in \Sigma^{k \times d}$. Let $\mathcal{D} = \{\nu \in [0,1]^d : \|\nu\|_1 = 1\}$ and $\mathcal{P} = \{p \in [0,1]^k : \|p\|_1 = 1\}$ be the probability simplices of dimension $d-1$ and $k-1$ respectively. It is helpful to notice that if $p \in \mathcal{P}$ and $\nu \in \mathcal{D}$, then $p^\top \mathcal{L} \nu$ is the expected loss suffered by a learner sampling an action from $p$ while the adversary samples its output from $\nu$. Given an action $a \in [k]$ let $C_a = \{\nu \in \mathcal{D} : e_a^\top \mathcal{L} \nu \leq \min_{b \in [k]} e_b^\top \mathcal{L} \nu\}$ be the set of probability vectors in $\mathcal{D}$ where action $a$ is optimal to play in expectation if the adversary plays randomly according to $\nu$. We call $C_a$ the cell of action $a$. Cells are convex polytopes because they are bounded and are determined by finitely many non-strict linear constraints. The collection $\{C_a : a \in [k]\}$, illustrated in Fig. 1, is called the cell decomposition.

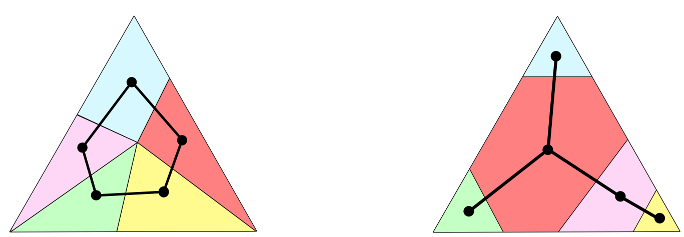


**Figure 1:** Cell decompositions and neighbourhood graphs for two games with $d = 3$ and $k = 5$.

**Remark 1.** A generalisation of the framework allows $(x_t)_{t=1}^n$ to be chosen in an arbitrary outcome space $\mathcal{X}$ and $\mathcal{L} : [k] \times \mathcal{X} \to [0,1]$ and $\Phi : [k] \times \mathcal{X} \to \Sigma$ are arbitrary functions. Our mathematical results continue to hold in this case with $d = |\mathcal{X}|$, but the proposed algorithms may not be computationally efficient when $|\mathcal{X}| = \infty$. A short discussion of infinite games appears in Section 7.

**Neighbourhood graph** A key concept in partial monitoring is the neighbourhood relation, which gives those pairs of potentially optimal actions that can be simultaneously optimal. An action $a$ is called Pareto optimal if $\dim(C_a) = d - 1$ where the dimension of a polytope is defined as the dimension of its affine hull as an affine subspace. Pareto optimal actions $a$ and $b$ are neighbours if $\dim(C_a \cap C_b) = d - 2$. More informally, actions are neighbours if their cells share a boundary of dimension $d - 2$. Note that $\dim(C_a \cap C_b) = d - 1$ is only possible when the rows in the loss matrix associated with actions $a$ and $b$ are the same: $e_a^\top \mathcal{L} = e_b^\top \mathcal{L}$. The neighbourhood relation defines a graph over $[k]$. We let $\mathcal{E} = \{(a, b) : a \text{ and } b \text{ are neighbours}\}$ be the set of edges in this graph. A game is called non-degenerate if all actions are Pareto optimal and $e_a^\top \mathcal{L} \neq e_b^\top \mathcal{L}$ for any actions $a \neq b$. Of course, $\dim(\mathcal{D}) = d - 1$, so actions $a$ with $\dim(C_a) < d - 1$ are optimal on a 'negligible' subset of $\mathcal{D}$, where they cannot be uniquely optimal. For the remainder we make the following simplifying assumption.

**Assumption 2.** $\mathcal{G}$ is globally observable and non-degenerate.

There is no particular reason to discard degenerate games except their analysis requires careful handling of certain edge cases, as we discuss briefly in the discussion and extensively in other work [Lattimore and Szepesvári, 2019a].

**Observability** The classification of a partial monitoring game depends on both the loss and signal matrices. What is important to make a non-degenerate game 'easy' is that the learner should have some way to estimate the loss differences between neighbouring actions by playing only those actions, a property known as local observability. To begin we introduce a linear structure on the signal matrix. For action $a$, let $S_a \in \{0, 1\}^{|\Sigma| \times d}$ be the matrix with $(S_a)_{\sigma x} = \mathbb{1}(\Phi_{ax} = \sigma)$, which means that $S_a e_x = e_{\Phi_{ax}}$. Here we have abused notation by indexing the rows of $S_a$ using signals, which we also do for vectors in $\mathbb{R}^{|\Sigma|}$. An important point is that $S_{A_t} e_{x_t} = e_{\sigma_t}$, which can be computed by the learner at the end of round $t$. A game is globally observable if for all edges $e = (a, b)$ in the neighbourhood graph there exists a tuple of 'estimation' vectors $(w_{ec})_{c=1}^k$ with $w_{ec} \in \mathbb{R}^{|\Sigma|}$ so that

$$\mathcal{L}_{ax} - \mathcal{L}_{bx} = \sum_{c=1}^{k} w_{ec}^\top S_c e_x = \sum_{c=1}^{k} (w_{ec})_{\Phi_{cx}} \text{ for all } x \in [d]\,. \tag{2}$$

A non-degenerate game is locally observable if Eq. (2) holds and additionally $w_{ec}$ can be chosen so that $w_{ec} = \mathbf{0}$ for all $c \notin \{a, b\}$. Of course, all locally observable games are globally observable.

**Remark 3.** The reader should be aware that for arbitrary (possibly degenerate) games the definition of local observability is that there exist estimation vectors such that Eq. (2) holds and $w_{ec} = \mathbf{0}$ unless $e_c^\top \mathcal{L} = \alpha e_a^\top \mathcal{L} + (1 - \alpha) e_b^\top \mathcal{L}$ for some $\alpha \in [0, 1]$. For non-degenerate games the definitions are equivalent by [Bartók et al., 2014, Lemma 11].

The classification theorem we mentioned in the introduction says that

$$\mathfrak{R}_n^* = \begin{cases} 0\,, & \text{if there is only one Pareto optimal action}\,; \\ \Theta(n^{1/2})\,, & \text{if the game is locally observable}\,; \\ \Theta(n^{2/3})\,, & \text{if the game is globally observable}\,; \\ \Omega(n)\,, & \text{otherwise}\,, \end{cases}$$

where the Big-Oh notation hides game-dependent constants.

**Estimation** The following lemma and discussion afterwards shows that for globally observable games Eq. (2) can be chained along paths in the neighbourhood graph to estimate the loss differences between any pair of actions, not just neighbours.

**Lemma 4.** *If $\mathcal{G}$ is non-degenerate and globally observable, then there exists a collection of matrices $(G_a)_{a=1}^k$ with $G_a \in \mathbb{R}^{k \times |\Sigma|}$ and $c \in \mathbb{R}^d$ such that $\sum_{a=1}^k G_a S_a = \mathcal{L} - [c_1 \mathbf{1}, \ldots, c_d \mathbf{1}]$.*

*Proof.* Let $\mathcal{T} \subseteq \mathcal{E}$ be any in-tree over $[k]$ and $(G_a)_{b\sigma} = \sum_{e \in \text{path}_{\mathcal{T}}(b)} (w_{ea})_\sigma$. Then

$$e_b^\top \sum_{a=1}^k G_a S_a e_x = \sum_{a=1}^k (G_a)_{b\Phi_{ax}} = \sum_{a=1}^k \sum_{e \in \text{path}_{\mathcal{T}}(b)} (w_{ea})_{\Phi_{ax}} = \mathcal{L}_{bx} - \mathcal{L}_{\text{root}_{\mathcal{T}}\, x}\,.$$

The result follows by setting $c_x = \mathcal{L}_{\text{root}_{\mathcal{T}}\, x}$. □

Given a distribution $p \in \mathcal{P} \cap (0,1)^k$ and $(G_a)_{a=1}^k$ satisfying the conclusion of Lemma 4, it follows that if $A$ is sampled from $p$ and $x \in [d]$ is arbitrary, then for actions $a, b$,

$$\mathbb{E}\left[\frac{(e_a - e_b)^\top G_A S_A e_x}{p_A}\right] = \sum_{c=1}^k (e_a - e_b)^\top G_c e_{\Phi_{cx}} = \mathcal{L}_{ax} - \mathcal{L}_{bx}\,. \tag{3}$$

The point is that $S_{A_t} e_{x_t} = e_{\sigma_t}$ is observed by the learner, which means the matrices $(G_a)_{a=1}^k$ can be used with importance-weighting to estimate the loss differences.

**Bandit and full information games** Bandit and full information games with finitely many possible losses can be modelled by finite partial monitoring games, and serve as useful examples. Bandit games are those with $\mathcal{L} = \Phi$ and full information games have $\Phi_{ax} = (\mathcal{L}_{1x}, \dots, \mathcal{L}_{kx})$. Estimation matrices witnessing the conclusion of Lemma 4 are easily constructed. The obvious choice for bandit games is $(G_a)_{b\sigma} = \mathbb{1}(a = b)\sigma$ while for full information games $(G_a)_{b\sigma} = p_a \sigma_b$ where $p \in \mathcal{P}$ is any probability distribution over the actions.

**Exponential weights** We briefly summarise a well-known bound on the regret of exponential weights. For $q \in \mathcal{P}$ define $\Psi_q : \mathbb{R}^k \to \mathbb{R}$ by

$$\Psi_q(z) = \langle q, \exp(-z) + z - 1 \rangle\,, \tag{4}$$

where the exponential function is applied coordinate-wise. Suppose that $(\hat{y}_t)_{t=1}^n$ is an arbitrary sequence of (loss) vectors with $\hat{y}_t \in \mathbb{R}^k$ and $(\eta_t)_{t=1}^n$ is a non-increasing sequence of positive learning rates. Define a sequence of probability vectors $(q_t)_{t=1}^n$ by

$$q_{ta} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{y}_{sa}\right)}{\sum_{b=1}^k \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{y}_{sb}\right)}\,.$$

Then the following bound on the regret holds for any $a^* \in [k]$ [Lattimore and Szepesvári, 2019, Chapter 28, for example],

$$\sum_{t=1}^n \sum_{a=1}^k q_{ta}(\hat{y}_{ta} - \hat{y}_{ta^*}) \leq \frac{\log(k)}{\eta_n} + \sum_{t=1}^n \frac{\Psi_{q_t}(\eta_t \hat{y}_t)}{\eta_t}\,. \tag{5}$$

Note, there is no randomness here. The term involving $\Psi$ is sometimes called the stability term. The following inequality is useful:

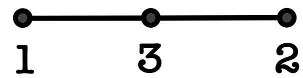
$$\Psi_q(\eta y) \leq \begin{cases} \eta^2 \|y\|^2_{\text{diag}(q)}\,, & \text{if } \eta y \geq -\mathbf{1}\,; \\ \frac{1}{2}\eta^2 \|y\|^2_{\text{diag}(q)}\,, & \text{if } \eta y \geq \mathbf{0}\,, \end{cases} \tag{6}$$

which follows from the inequalities $\exp(-x) \leq x^2 - x + 1$ for all $x \geq -1$ and $\exp(-x) \leq x^2/2 - x + 1$ for $x \geq 0$. We will use the fact that the perspective $(p, z) \mapsto p\Psi_q(z/p)$ is convex for $p > 0$.

## 3 Exploration by optimisation

Our algorithm is a combination of exponential weights and a careful exploration strategy. The following example game is helpful to gain some intuition:

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1/4 & 1/4 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} \perp & \perp \\ \perp & \perp \\ \text{H} & \text{T} \end{pmatrix}. \tag{7}$$




The figure on the right is the neighbourhood graph, with the first two actions separated by the third. The structure of the feedback matrix means that the learner only gains information by playing the third action. Suppose that $q \in \mathcal{P}$ is a distribution with $q_3$ close to zero and both $q_1$ and $q_2$ reasonably large. Sampling an action from $q$ leads to a low probability of gaining information and a correspondingly high variance when estimating the difference between the losses of the first and second actions. Consider the transformation of $q$ defined by $p = q - \min(q_1, q_2)(e_1 + e_2) + 2\min(q_1, q_2)e_3$, which is illustrated in Fig. 2. Then $p_3 \geq q_3$ and

$$(p - q)^\top \mathcal{L} = -\frac{1}{2}\min(q_1, q_2)\mathbf{1}\,. \tag{8}$$

Hence, any algorithm proposing to play distribution $q \in \mathcal{P}$ with $\min(q_1, q_2) > 0$ could improve its decision by playing $p$, which decreases the expected loss and increases the amount of information. Our new algorithm solves an optimisation problem to find a sampling distribution and estimation matrices that minimise the sum of the loss relative to a distribution proposed by exponential weights and the stability term in Eq. (5). In the example above the solution always results in a distribution $p$ with $\min(p_1, p_2) = 0$. By contrast, previous algorithms for adversarial locally observable partial monitoring games do not exhibit this behaviour [Foster and Rakhlin, 2012; Lattimore and Szepesvári, 2019a].

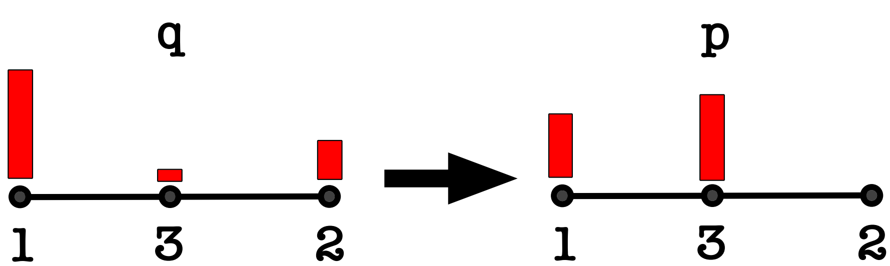



**Figure 2:** An exploration distribution $p$ derived from $q$ for the game in Eq. (7). The expected loss when playing $p$ is smaller than playing $q$ and simultaneously more information is gained because the third action is revealing.

**Optimisation problem** Suppose that exponential weights proposes a distribution $q \in \mathcal{P}$. Our algorithm solves an optimisation problem to find an exploration distribution and estimation matrices that determine the loss estimators. For $q \in \mathcal{P}$ and $\eta > 0$ let $\text{opt}_q(\eta)$ be the value of the following convex optimisation problem:

$$
\begin{aligned}
&\underset{(G_a)_{a=1}^k, c\in\mathbb{R}^d, p\in\mathcal{P}}{\text{minimise}} && \max_{x\in[d]} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L} e_x + \frac{1}{\eta^2}\sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G_a S_a e_x}{p_a}\right)\right] \\
&\text{subject to} && \sum_{a=1}^k G_a S_a = \mathcal{L} - [c_1 \mathbf{1}, \ldots, c_d \mathbf{1}]\,.
\end{aligned} \tag{9}
$$

Lemma 4 implies feasibility for non-degenerate globally observable games. We assume that Eq. (9) can be solved exactly to obtain minimising values for $(G_a)_{a=1}^k$ and $p \in \mathcal{P}$. Our algorithm, however, is robust to small perturbations of these quantities. Let

$$
\text{opt}^*(\eta) = \sup_{q\in\mathcal{P}} \text{opt}_q(\eta)\,.
$$

Note that both $\text{opt}_q(\eta)$ and $\text{opt}^*(\eta)$ depend on $\mathcal{G}$; this dependence is not shown to minimize clutter. At least for reasonably sized problems, the optimisation problem can be solved using off-the-shelf solvers and disciplined convex programming. The algorithm is a simple combination of exponential weights using the exploration distribution and estimation matrices provided by solving Eq. (9).

**input:** $\eta$

**for** $t = 1, \ldots, n$:

Compute $Q_{ta} = \dfrac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{y}_{sa}\right)}{\sum_{b=1}^k \exp\left(-\eta \sum_{s=1}^{t-1} \hat{y}_{sb}\right)}$

Solve (9) with $q = Q_t$ to find $P_t \in \mathcal{P}$ and $(G_{ta})_{a=1}^k$

Sample $A_t \sim P_t$ and observe $\sigma_t$

Compute $\hat{y}_t = \dfrac{G_{tA_t} e_{\sigma_t}}{P_{tA_t}}$

**Algorithm 1:** Exponential weights for partial monitoring

The regret of Algorithm 1 depends on the learning rate and the value of the optimisation problem, which depends on the structure of the game. Bounds on $\text{opt}^*(\eta)$ are provided subsequently.

**Theorem 5.** *For any $\eta > 0$, the regret of Algorithm 1 is bounded by $\mathfrak{R}_n \leq \dfrac{\log(k)}{\eta} + \eta n\, \text{opt}^*(\eta)$.*

*Proof.* Let $a^* = \arg\min_{a\in[k]} \sum_{t=1}^n \mathcal{L}(a, x_t)$ be the optimal action in hindsight. Then

$$
\begin{aligned}
\mathfrak{R}_n &= \mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k P_{ta}(\mathcal{L}_{ax_t} - \mathcal{L}_{a^* x_t})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k Q_{ta}(\mathcal{L}_{ax_t} - \mathcal{L}_{a^* x_t})\right] + \mathbb{E}\left[\sum_{t=1}^n (P_t - Q_t)^\top \mathcal{L} e_{x_t}\right]\,.
\end{aligned} \tag{10}
$$

Eq. (3) and the constraint in Eq. (9) shows that $\mathbb{E}[\hat{y}_{ta} - \hat{y}_{ta^*} \,|\, A_1, \sigma_1, \ldots, A_{t-1}, \sigma_{t-1}] = \mathcal{L}_{ax_t} - \mathcal{L}_{a^*x_t}$. Then

$$\mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k Q_{ta}(\mathcal{L}_{ax_t} - \mathcal{L}_{a^*x_t})\right] = \mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k Q_{ta}(\hat{y}_{ta} - \hat{y}_{ta^*})\right]$$
$$\leq \frac{\log(k)}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \Psi_{Q_t}(\eta \hat{y}_t)\right] \tag{11}$$
$$= \frac{\log(k)}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k P_{ta}\Psi_{Q_t}\left(\frac{\eta G_{ta} S_a e_{x_t}}{P_{ta}}\right)\right],$$

where Eq. (11) follows from Eq. (5) and the definitions of $Q_t$ and $\hat{y}_t$. Combining with Eq. (10) and the definition of $\mathrm{opt}^*(\eta)$,

$$\mathfrak{R}_n \leq \frac{\log(k)}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \left(\eta(P_t - Q_t)^\top \mathcal{L} e_{x_t} + \sum_{a=1}^k P_{ta}\Psi_{Q_t}\left(\frac{\eta G_{ta} S_a e_{x_t}}{P_{ta}}\right)\right)\right]$$
$$\leq \frac{\log(k)}{\eta} + \eta n \, \mathrm{opt}^*(\eta) .$$

□

**Applications** Table 2 provides bounds on $\mathrm{opt}^*(\eta)$ for different games and the regret bound that results from optimising the learning rate. The proofs are provided in Sections 5 and 6. Except for locally observable games, they mirror existing proofs bounding the stability of exponential weights. In this way many other results could be added to this table, including bandits with graph feedback [Alon et al., 2015] and linear bandits with finitely many arms [Bubeck et al., 2012].

| Game type | $\mathrm{opt}^*(\eta)$ bound | Conditions | Ref. | Regret |
|---|---|---|---|---|
| BANDIT | $k/2$ | | Prop. 8 | $\sqrt{2nk\log(k)}$ |
| FULL INFORMATION | $1/2$ | | Prop. 9 | $\sqrt{2n\log(k)}$ |
| GLOBALLY OBSERVABLE NON-DEGENERATE | $c_\mathcal{G}/\sqrt{\eta}$ | $\eta \leq 1/c_\mathcal{G}^2$ | Prop. 10 | $3(c_\mathcal{G} n/2)^{2/3}(\log(k))^{1/3}$ |
| LOCALLY OBSERVABLE NON-DEGENERATE | $9k^3(d+1)^2$ | $\eta \leq 1/(2k^2(d+1))$ | Prop. 11 | $6k^{3/2}(d+1)\sqrt{n\log(k)}$ |

**Table 2:** Upper bounds on $\mathrm{opt}^*(\eta)$ and the regret of Algorithm 1 for different games. The constant $c_\mathcal{G}$ is game-dependent and can be exponentially large in $d$, which we believe is unavoidable.

## 4 Online learning rate tuning

Tuning the learning rate used by Algorithm 1 is delicate. First, it is not clear that $\mathrm{opt}^*(\eta)$ can be computed efficiently in general. Second, the learning rate that minimises the bound in Theorem 5 may be overly conservative. Algorithm 2 mitigates these issue by using an adaptive learning rate. The algorithm is parameterised by a constant $B$ that determines the initialisation of the learning rate. $B$

should be chosen large enough that $\eta = 1/B$ satisfies the conditions for the relevant game in Table 2, but the additional regret from choosing $B$ too large is only additive.

```
input:  B
for t = 1, 2, ..., :
```

Set $\eta_t = \min\left\{\frac{1}{B}, \sqrt{\frac{\log(k)}{1 + \sum_{s=1}^{t-1} V_s}}\right\}$

Compute $Q_{ta} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{y}_{sa}\right)}{\sum_{b=1}^{k} \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{y}_{sb}\right)}$

Solve (9) with $\eta = \eta_t$ and $q = Q_t$ to find $V_t = \max\{0, \operatorname{opt}_{Q_t}(\eta_t)\}$ and corresponding $P_t$ and $(G_{ta})_{a=1}^k$

Sample $A_t \sim P_t$ and observe $\sigma_t$

Compute $\hat{y}_t = \frac{G_{tA_t} e_{\sigma_t}}{P_{tA_t}}$

**Algorithm 2:** Adaptive exponential weights for partial monitoring

We now present a general theorem that bounds the regret as a function of $(V_t)_t$, which is computed by the algorithm. This theorem implies that Algorithm 2 recovers all regret bounds in Table 2 up to small constant factors and additive terms.

**Theorem 6.** *There exists a universal constant $c > 0$ such that the regret of Algorithm 2 is bounded by*

$$\mathfrak{R}_n \leq 5\mathbb{E}\left[\sqrt{\left(1 + \sum_{t=1}^{n} V_t\right) \log(k)}\right] + \mathbb{E}\left[\max_{t \in [n]} V_t\right] \sqrt{\log(k)} + B \log(k)\,.$$

A corollary using the definition of $V_t$ is that the regret of Algorithm 2 is bounded by

$$\mathfrak{R}_n = O\left(\sqrt{n \sup\{\max\{0, \operatorname{opt}^*(\eta)\} : \eta \leq 1/B\} \log(k)} + B \log(k)\right)\,. \tag{12}$$

This bound is most useful for full information, bandit and locally observable non-degenerate games when $B$ can be chosen so that $\eta_1 \leq 1/B$ satisfies the conditions in the second column of Table 2. As a consequence, for games of this category Theorem 6 recovers the bounds in the last column Table 2 up to small constant factors and additive terms.

For games that are globally observable but not locally observable $\operatorname{opt}^*(\eta) \to 0$ as $\eta \to 0$ and the supremum in Eq. (12) is infinite. Soon, we will argue that the learning rate used by Algorithm 2 does not decrease too fast and that the algorithm still achieves the regret bound shown in Table 2 for globally observable games.

*Proof sketch of Theorem 6.* We explain only the differences relative to the proof of Theorem 5. Recall that $V_t = \max\{0, \operatorname{opt}_{Q_t}(\eta_t)\}$. Note, the learning rate $\eta_t$ is non-increasing. Hence, by Eq. (5),

$$\begin{aligned}
\mathfrak{R}_n &\leq \mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k Q_{ta}(\hat{y}_{ta} - \hat{y}_{ta^*}) + \sum_{t=1}^n (P_t - Q_t)^\top \mathcal{L} e_{x_t}\right] \\
&\leq \mathbb{E}\left[\frac{\log(k)}{\eta_n} + \sum_{t=1}^n \frac{\Psi_{Q_t}(\eta_t \hat{y}_t)}{\eta_t} + \sum_{t=1}^n (P_t - Q_t)^\top \mathcal{L} e_{x_t}\right] \qquad (13) \\
&\leq \mathbb{E}\left[\frac{\log(k)}{\eta_n} + \sum_{t=1}^n \eta_t V_t\right], \qquad (14)
\end{aligned}$$

where Eq. (14) follows from the same argument as the proof of Theorem 5 and the definition of $V_t$. The second term is bounded using Lemma 14 in the appendix by

$$\sum_{t=1}^n \eta_t V_t \leq 4\sqrt{\left(1 + \frac{1}{2}\sum_{t=1}^n V_t\right)\log(k)} + \max_{t \in [k]} V_t \sqrt{\log(k)}\,.$$

The definition of $(\eta_t)_{t=1}^n$ means that

$$\frac{\log(k)}{\eta_n} \leq B\log(k) + \sqrt{\left(1 + \sum_{t=1}^n V_t\right)\log(k)}\,.$$

The bound follows by combining the parts and naive algebra. □

As promised, we now show that for sufficiently large $B$ the algorithm achieves the best known regret for any globally observable, non-degenerate game.

**Proposition 7.** *Fix a globally observable non-degenerate game $\mathcal{G}$. Suppose that $\alpha > 0$ and $\operatorname{opt}^*(\eta) \leq \alpha/\sqrt{\eta}$ for all $\eta \leq 1/B$. Then, the regret of Algorithm 2 on $\mathcal{G}$ is at most*

$$\mathfrak{R}_n = O\left((n\alpha)^{2/3}(\log(k))^{1/3} + B\log(k)\right),$$

*where the Big-Oh hides only universal constants.*

Note that the conditions of this result will be satisfied with $\alpha = c_{\mathcal{G}}$ once $B \geq c_{\mathcal{G}}^2$ with (cf. Table 2).

*Proof.* The result follows from Theorem 6 and an almost sure bound on $\sum_{t=1}^n V_t$. Clearly, $\eta_t \leq 1/B$ and so by assumption $V_t = \max\{0, \operatorname{opt}_{Q_t}(\eta_t)\} \leq \alpha/\sqrt{\eta_t}$. Then, using the definition of $(\eta_t)_{t=1}^n$,

$$\frac{\log(k)}{\eta_{t+1}^2} \leq \frac{\log(k)}{\eta_t^2} + \frac{\alpha}{\eta_t^{1/2}} = \frac{\log(k)}{\eta_t^2} + \frac{\alpha}{\log(k)^{1/4}}\left(\frac{\log(k)}{\eta_t^2}\right)^{\frac{1}{4}}.$$

Hence, using the definition of $\eta_n$ and Lemma 15 in the appendix,

$$1 + \sum_{t=1}^n V_t \leq \frac{\log(k)}{\eta_n^2} \leq \left(\frac{3\alpha(n-1)}{4\log(k)^{1/4}} + \max\{1, B^2\log(k)\}^{3/4}\right)^{4/3}.$$

Substituting the above bound into the dominant term of Theorem 6 shows that

$$\sqrt{\left(1+\sum_{t=1}^n V_t\right)\log(k)} = O\Big((kn\alpha)^{2/3}(\log(k))^{1/3} + B\log(k)\Big)\,.$$

The result is completed by noting that $\max_{t\in[n]} V_t \leq \alpha\eta_n^{-1/2}$ is lower-order. □

## 5 Bandit, full information and globally observable games

We now bound $\mathrm{opt}^*(\eta)$ for bandit, full information and globally observation games. All results follow from the usual arguments for bounding the stability term in the regret guarantee for exponential weights in Eq. (5).

**Proposition 8.** *For bandit games,* $\mathrm{opt}^*(\eta) \leq k/2$ *for all* $\eta > 0$*.*

*Proof.* For bandit games the constraints of Eq. (9) are satisfied by $p = q$ and non-negative $(G_a)_{a=1}^k$, which may be chosen so that $G_a S_a e_x = \mathcal{L}_{ax} e_a$. This corresponds to the usual importance-weighted estimators used for $k$-armed bandits. Then

$$\begin{aligned}
\mathrm{opt}_q(\eta) &\leq \max_{x\in[d]} \frac{1}{\eta^2}\sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G_a S_a e_x}{p_a}\right) \\
&= \max_{x\in[d]} \frac{1}{\eta^2}\sum_{a=1}^k p_a \Psi_q\left(\frac{\eta \mathcal{L}_{ax} e_a}{p_a}\right) \\
&\leq \frac{1}{2}\max_{x\in[d]} \sum_{a=1}^k \mathcal{L}_{ax}^2 \leq \frac{k}{2}\,,
\end{aligned}$$

where in the second last inequality we used Eq. (6) and the fact that $p_a = q_a$. □

**Proposition 9.** *For full information games,* $\mathrm{opt}^*(\eta) \leq 1/2$ *for all* $\eta > 0$*.*

*Proof.* For these games we again let $p = q$ and $(G_a)_{a=1}^k$ can be chosen so that $G_a S_a = p_a \mathcal{L}$, which satisfies the constraints. The argument then follows along the same lines as the proof of Proposition 8. □

**Proposition 10.** *For non-degenerate globally observable games there exists a constant $c_{\mathcal{G}}$ depending only on $\Phi$ and $\mathcal{L}$ such that for all $\eta \leq 1/c_{\mathcal{G}}^2$,*

$$\mathrm{opt}^*(\eta) \leq \frac{c_{\mathcal{G}}}{\sqrt{\eta}}\,.$$

*Proof.* By the definition of a globally observable game there exist matrices $(G_a)_{a=1}^k$ satisfying the constraints in Eq. (9). Let $\beta = \max_a \|G_a\|_\infty$ and $c_{\mathcal{G}} = \max\{1, 2k\beta\}$. Then let $\gamma = k\beta\sqrt{\eta}$ and $p = (1-\gamma)q + \gamma\mathbf{1}/k$, which is a probability distribution since $\gamma \in [0,1]$ for $\eta \leq 1/c_{\mathcal{G}}^2$. We now claim that $\eta\langle e_b, G_a S_a e_x\rangle/p_a \geq -1$, which follows from the definitions of $\gamma$ and $\beta$ so that

$$p_a \geq \frac{\gamma}{k} = \beta\sqrt{\eta} \geq \beta\eta \geq \eta\max(-G_a S_a)\,,$$

where the second inequality uses the fact that $c_{\mathcal{G}} \geq 1$ and $\eta \leq 1/c_{\mathcal{G}}^2 \leq 1$. To bound the objective notice that for any $x \in [d]$ it holds that

$$\frac{1}{\eta}(p-q)^\top \mathcal{L} e_x = \frac{\gamma}{\eta}(\mathbf{1}/k - q)^\top \mathcal{L} e_x \leq \frac{\gamma}{\eta} = \frac{k\beta}{\sqrt{\eta}}\,.$$

For the second term in the objective, by Eq. (6),

$$\begin{aligned}\frac{1}{\eta^2}\max_{x\in[d]}\sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G_a S_a e_x}{p_a}\right) &\leq \max_{x\in[d]}\sum_{a=1}^k \frac{\|G_a S_a e_x\|^2_{\mathrm{diag}(q)}}{p_a} \\ &\leq \frac{k}{\gamma}\max_{x\in[d]}\sum_{a=1}^k\sum_{b=1}^k q_b\langle e_b, G_a S_a e_x\rangle^2 \\ &\leq \frac{k^2\beta^2}{\gamma} = \frac{k\beta}{\sqrt{\eta}}\,.\end{aligned}$$

The result follows by combining the previous two displays and the definition of $c_{\mathcal{G}}$. □

# 6 Locally observable games

Controlling $\mathrm{opt}^*(\eta)$ for locally observable games is more involved. The main result of this section is a proof of the following proposition.

**Proposition 11.** *For locally observable non-degenerate games and $\eta \leq 1/(2k^2(d+1))$,*

$$\mathrm{opt}^*(\eta) \leq 9k^3(d+1)^2\,.$$

We make use of the water transfer operator, which is a construction from our earlier paper that provides an exploration distribution suitable for locally observable games in the Bayesian setting. The challenge in partial monitoring is that the observability structure only allows for pairwise comparison between neighbours. This is problematic when two non-neighbouring actions are played with high probability and the actions separating them are played with low probability. Given a distributions $q \in \mathcal{P}$ and $\nu \in \mathcal{D}$, the water transfer operator 'flows' probability in $q$ towards the greedy action $a$ for which $\nu \in C_a$ (Fig. 3). Then all loss differences can be estimated relative to the greedy action. This decreases the variance of estimation without increasing the expected loss when the adversary samples its action from $\nu$. Shortly we use a minimax theorem to transfer a statement that for all $\nu$ there exists an operator with the desired to properties to show there exists an operator such that for all $\nu$ the desired properties hold.

**Lemma 12** (Lattimore and Szepesvári 2019b)**.** *Suppose that $\mathcal{G}$ is non-degenerate and locally observable and $\nu \in \mathcal{D}$. Then there exists a function $W_\nu : \mathcal{P} \to \mathcal{P}$ such that the following hold for all $q \in \mathcal{P}$:*

*(a) The expected loss does not increase: $(W_\nu(q) - q)^\top \mathcal{L}\nu \leq 0$.*

*(b) Action probabilities are not too small: $W_\nu(q)_a \geq q_a/k$ for all $a \in [k]$.*

*(c) Probabilities increase towards the root of some in-tree: there exists an in-tree $\mathcal{T} \subseteq \mathcal{E}$ over $[k]$ such that $W_\nu(q)_a \leq W_\nu(q)_b$ for all $(a,b) \in \mathcal{T}$.*

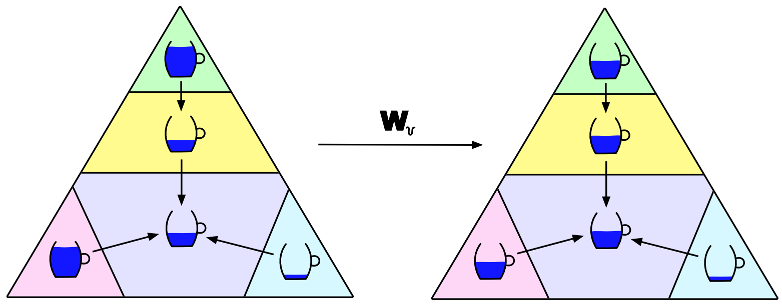



**Figure 3:** The water transfer operator. Imagine the mugs are on a flat table and arrows are one-way pipes connecting the mugs. Water flows from mugs with more water than their neighbours in a carefully specified way until stabilisation. Full details may be found in our previous paper [Lattimore and Szepesvári, 2019b].

The core result for proving Proposition 11 is the following lemma.

**Lemma 13.** *Consider a locally observable, non-degenerate game. For all $q \in \mathcal{P}$ there exists a $p \in \mathcal{P}$ and $(G_a)_{a=1}^k$ with $G_a \in \mathbb{R}^{k \times |\Sigma|}$ such that*

*(a) Losses can be estimated up to additive constants: there exists a vector $c \in [0,1]^d$ such that $\sum_{a=1}^k G_a S_a = \mathcal{L} - [c_1 \mathbf{1}, \ldots, c_d \mathbf{1}]$.*

*(b) The expected loss does not increase: $(p-q)^\top \mathcal{L} \leq \mathbf{0}$.*

*(c) The variance is controlled: $\max_{x \in [d]} \sum_{a=1}^k \frac{\|G_a S_a e_x\|^2_{\mathrm{diag}(q)}}{p_a} \leq 4k^3(d+1)^2$.*

*(d) The range is controlled: $\max_{a \in [k]} \|G_a S_a\|_\infty \leq k(d+1)$.*

*Proof.* Define a convex set of tuples of matrices

$$\mathcal{H} = \left\{ (G_a)_{a=1}^k : G_a \in [-k(d+1), k(d+1)]^{k \times |\Sigma|} \right\} ,$$

where the linear structure on $\mathcal{H}$ is defined in the obvious way by identifying $\mathcal{H}$ with a subset of $\mathbb{R}^{k^2|\Sigma|}$. Let

$$\mathcal{C}^{(1)} = \left\{ (p, G) \in \mathcal{P} \times \mathcal{H} : \text{ exists } c \in [0,1]^d \text{ such that } \sum_{a=1}^k G_a S_a = \mathcal{L} - [c_1 \mathbf{1}, \ldots, c_d \mathbf{1}] \right\} ,$$

$$\mathcal{C}_q^{(2)} = \left\{ (p, G) \in \mathcal{P} \times \mathcal{H} : \max_{x \in [d]} \sum_{a=1}^k \frac{\|G_a S_a e_x\|^2_{\mathrm{diag}(q)}}{p_a} \leq 4k^3(d+1)^2 \right\} .$$

Both sets $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$ are convex. Define $\mathcal{C}_q = \mathcal{C}^{(1)} \cap \mathcal{C}_q^{(2)}$ and $J_q : \mathcal{C}_q \times \mathcal{D} \to \mathbb{R}$ by $J_q((p,G),\nu) = (p-q)^\top \mathcal{L}\nu$. We now prove that

$$\max_{\nu \in \mathcal{D}} \min_{(p,G) \in \mathcal{C}_q} J_q((p,G),\nu) \leq 0 , \tag{15}$$

where the existence of the $\max$ and $\min$ are justified by the compactness of $\mathcal{D}$ and $\mathcal{C}_q$. The lemma will follow from Eq. (15) and Sion's minimax theorem to exchange the $\max$ and $\min$. Fix $\nu \in \mathcal{D}$ and let $p = W_\nu(q)$. Let $\mathcal{T} \subseteq \mathcal{E}$ be the in-tree provided by Lemma 12 and for any edge $e$ let $(w_{ea})_{a=1}^k$ be a tuple of vectors satisfying Eq. (2) and with $\|w_{ea}\|_\infty \leq d+1$ for all edges $e$ and actions $a$, which is possible by [Lattimore and Szepesvári, 2019a, Lemma 16]. Next, for each action $a$ define $G_a \in \mathbb{R}^{k \times |\Sigma|}$ element-wise by

$$(G_a)_{b\sigma} = \sum_{e \in \mathrm{path}_\mathcal{T}(b)} (w_{ea})_\sigma \,.$$

That $(G_a)_{a=1}^k \in \mathcal{H}$ follows from the bound $\|w_{ea}\|_\infty \leq d+1$ and the fact that paths have length at most $k$. We now show that $(p, G) \in \mathcal{C}_q$. By the proof of Lemma 4,

$$\sum_{a=1}^k G_a S_a = \mathcal{L} - [\mathcal{L}_{\mathrm{root}_\mathcal{T}\, 1}\mathbf{1}, \dots, \mathcal{L}_{\mathrm{root}_\mathcal{T}\, d}\mathbf{1}] \,.$$

Then using Parts (b) and (c) of Lemma 12,

$$\begin{aligned}
\sum_{a=1}^k \frac{\|G_a S_a e_x\|^2_{\mathrm{diag}(q)}}{p_a} &= \sum_{b=1}^k \sum_{a=1}^k \frac{q_b}{p_a} \left( \sum_{e \in \mathrm{path}_\mathcal{T}(b)} (w_{ea})_{\Phi_{ax}} \right)^2 \\
&\leq \sum_{b=1}^k (d+1)^2 \sum_{a=1}^k \frac{q_b}{p_a} \left( \sum_{e \in \mathrm{path}_\mathcal{T}(b)} \mathbb{1}(a \in e) \right)^2 \\
&\leq 4k^3(d+1)^2 \,,
\end{aligned}$$

where we used Part (b) of Lemma 12 to show that $q_b \leq kp_b$ and Part (c) to show that $p_a \geq p_b$ for $a \in \mathrm{path}_\mathcal{T}(b)$. Thus $(p, G) \in \mathcal{C}_q$. Finally, by Part (a) of Lemma 12, $J_q((p,G),\nu) = (p-q)^\top \mathcal{L}\nu \leq 0$, which means that for all $\nu \in \mathcal{D}$ there exists a $(p, G) \in \mathcal{C}_q$ with

$$J_q((p,G),\nu) \leq 0 \,,$$

which proves Eq. (15). Clearly $J_q$ is linear and continuous in both arguments and $\mathcal{D}$ is compact. Hence, by Sion's minimax theorem [Sion, 1958],

$$\min_{(p,G) \in \mathcal{C}_q} \max_{\nu \in \mathcal{D}} J_q((p,G),\nu) = \max_{\nu \in \mathcal{D}} \min_{(p,G) \in \mathcal{C}_q} J_q((p,G),\nu) \,. \tag{16}$$

By substituting Eq. (15) into the above display it follows that there exists a $(p, G) \in \mathcal{C}_q$ such that $J_q((p,G),\nu) \leq 0$ for all $\nu \in \mathcal{D}$, which completes the proof. □

*Proof of Proposition 11.* By Lemma 13, for any $q \in \mathcal{P}$ there exists an $r \in \mathcal{P}$ and $c \in [0,1]^d$ and $(G_a)_{a=1}^k$ satisfying conditions (a–d) in Lemma 13. Let $p = (1-\gamma)r + \gamma\mathbf{1}/k$ with $\gamma = \eta k^2(d+1) \leq 1/2$. The constraint in Eq. (9) only depends on $(G_a)_{a=1}^k$ and $c$ and is satisfied without further calculation. Using Part (d) of Lemma 13:

$$p_a \geq \frac{\gamma}{k} = \eta k(d+1) \geq \eta \max(-G_a S_a) \,,$$

which shows that for all $b \in [k]$ and $x \in [d]$,

$$\frac{\eta e_b^\top G_a S_a e_x}{p_a} \geq -1 \,. \tag{17}$$

We now control each term in the objective:

$$\begin{aligned}
\frac{1}{\eta^2} \max_{x \in [d]} \sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G_a S_a e_x}{p_a}\right) &\leq \max_{x \in [d]} \sum_{a=1}^k \frac{\|G_a S_a e_x\|^2_{\operatorname{diag}(q)}}{p_a} \\
&\leq \max_{x \in [d]} \sum_{a=1}^k \frac{\|G_a S_a e_x\|^2_{\operatorname{diag}(q)}}{(1-\gamma) r_a} \\
&\leq \frac{4k^3(d+1)^2}{1-\gamma} \leq 8k^3(d+1)^2 \,,
\end{aligned}$$

where in the first inequality we used Eq. (6) with Eq. (17), in the second inequality we used Part (c) of Lemma 13 and in the third that $\gamma \leq 1/2$. The other term in the objective is bounded by

$$\frac{(p-q)^\top \mathcal{L} e_x}{\eta} = \frac{((1-\gamma)r + \gamma \mathbf{1}/k - q)^\top \mathcal{L} e_x}{\eta} \leq \frac{\gamma(\mathbf{1}/k - q)^\top \mathcal{L} e_x}{\eta} \leq \frac{\gamma}{\eta} \leq k^3(d+1)^2 \,.$$

Hence, $\operatorname{opt}^*(\eta) \leq 9k^3(d+1)^2$. □

# 7 Discussion

We introduced a new algorithm for finite partial monitoring that is efficient, nearly parameter free and enjoys roughly the best known regret in all classes of games. Notably, this is the first efficient algorithm for which the regret is independent of arbitrarily large game-dependent constants for locally observable non-degenerate games. A natural criticism of previous algorithms for partial monitoring is that the algorithms are generally quite conservative and not practical for normal problems. As far as we can tell, the proposed algorithm does not suffer from this problem, at least recovering standard bounds in bandit and full information settings. In certain cases the algorithm may also adapt to the choices of the adversary. The principle for finding an exploration distribution and estimation procedure is generic and may work well in other problems.

**Lower bounds** The best known lower bound for locally observable partial monitoring games is either $\Omega(\sqrt{kn})$ or $\Omega(d\sqrt{n})$, which are witnessed by a standard Bernoulli bandit [Auer et al., 1995] and a result by the authors [Lattimore and Szepesvári, 2019a]. If pressed, we would speculate that $\Theta(kd\sqrt{n})$ is the correct worst-case regret over all $d$-outcome $k$-action non-degenerate locally observable partial monitoring games.

**Infinite outcome spaces** Finiteness of the outcome space was not used in the proofs of Theorem 5 or Theorem 6 and in particular the results in Table 2 continue to hold in this case (the result for locally observable games becomes vacuous). The main cost of infinite outcome spaces is that the optimisation problem Eq. (9) is unlikely to be tractable without additional structure. Classic examples of infinite games for which the regret can be well controlled are bandit and full information games. In both games the outcomes $(x_t)_{t=1}^n$ are chosen in $\mathcal{X} = [0,1]^k$ and $\mathcal{L}(a,x) = x_a$ (using the notation of Remark 1). The signal function is $\Phi(a,x) = x_a$ for bandits and $\Phi(a,x) = x$ for the full information games. Exploring the existence of a simple classification theorem for infinite-outcome games is an interesting future direction. Understanding when Eq. (9) is tractable is also intriguing.

**Game-dependent bounds** One of the objectives of this work was to design an efficient algorithm for which the regret does not depend on arbitrarily large game-dependent constants. Naturally it is desirable to have small game-dependent constants and adaptivity to the choices of the adversary. Table 2 provides upper bounds on $\operatorname{opt}^*(\eta)$ for various classes, but the actual values depends on the game. Understanding the dependence of this optimisation problem on the structure of the loss and signal matrices is an interesting open direction. Also interesting is whether or not $\operatorname{opt}^*(\eta)$ is a fundamental quantity for the difficulty of the game and/or the regret of our algorithms.

**Adaptivity** Algorithm 2 already exhibits some adaptivity in the lucky situation that $V_t$ is small. This is not entirely satisfactory, however, since $V_t$ is a random variable that depends on the choices of both the learner and the adversary. We anticipate that all the usual enhancements for adaptivity – log barrier, biased estimates and optimism – can be applied here [Rakhlin and Sridharan, 2013; Bubeck et al., 2018; Wei and Luo, 2018; Bubeck et al., 2019, for example]. A related challenge would be to seek a best-of-both-worlds result using the INF potential [Zimmert et al., 2019].

**Beyond exponential weights** The objective in Eq. (9) is chosen so that the terms in Eq. (11) are well controlled, which corresponds to bounding the stability term in the regret analysis of exponential weights. Other algorithms can be obtained by replacing exponential weights with follow the regularized leader and Legendre potential $F$. A standard regret bound (holding under certain technical conditions) is

$$\mathfrak{R}_n \leq \frac{\operatorname{diam}_F(\mathcal{P})}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k P_{ta} D_{F^*}\left(\nabla F(Q_t) - \frac{\eta G_{ta} S_a e_{x_t}}{P_{ta}}, \nabla F(Q_t)\right)\right] \tag{18}$$
$$+ \mathbb{E}\left[\sum_{t=1}^n (Q_t - P_t)^\top \mathcal{L} e_{x_t}\right].$$

where $\operatorname{diam}_F(\mathcal{P}) = \max_{x,y \in \mathcal{P}} F(x) - F(y)$ is the diameter and $D_{F^*}(x,y)$ is the Bregman divergence between $x$ and $y$ with respect to the Fenchel conjugate of $F$. Let

$$\Psi_q(z) = D_{F^*}(\nabla F(q) - z, \nabla F(q)).$$

Then convexity of $F^*$ implies that the perspective $(p,z) \mapsto p\Psi_q(z/p)$ is also convex for $p > 0$. When $F$ is the unnormalised negentropy, the definition above reduces to Eq. (4). All this means that the same approach holds more broadly for other potentials, which carry certain advantages in some settings [Audibert and Bubeck, 2009; Bubeck et al., 2018; Wei and Luo, 2018; Bubeck et al., 2019, and others]. For more details on follow the regularised leader and bounds of the form in Eq. (18), see [Lattimore and Szepesvári, 2019, Chapter 28] and [Hazan, 2016]. We leave a deeper exploration of these ideas for the future.

**Degenerate games** The non-degeneracy assumption is for simplicity only: for degenerate games the only change is that the exponential weights distribution should be computed over the Pareto optimal actions ($\dim(C_a) = d-1$) while the optimisation problem that determines the exploration distribution must include all actions. As in our previous work, even for locally observable games the constants can be exponential in $d$, which we believe is unavoidable [Lattimore and Szepesvári, 2019b].

**Connections between stability and the information ratio** Zimmert and Lattimore [2019] have shown that the generalised information ratio can be bounded by a worst-case bound on the stability term of mirror descent, which makes a connection between the information-theoretic tools and those from online convex optimisation. Here we work in the other direction, using duality and the techniques for bounding the information ratio to bound the stability term. The argument does not provide an equivalence between stability and the information ratio, but perhaps reinforces the feeling that there is an interesting connection here.

**Acknowledgements** We are grateful to András György for many insightful comments.

# References

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

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

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE.

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

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

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

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

Bubeck, S., Li, Y., Luo, H., and Wei, C.-Y. (2019). Improved path-length regret bounds for bandits. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 508–528, Phoenix, USA. PMLR.

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

Foster, D. and Rakhlin, A. (2012). No internal regret via neighborhood watch. In Lawrence, N. D. and Girolami, M., 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. PMLR.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

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

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

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

Lattimore, T. and Szepesvári, C. (2019a). Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In Garivier, A. and Kale, S., editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 529–556, Chicago, Illinois. PMLR.

Lattimore, T. and Szepesvári, C. (2019b). An information-theoretic approach to minimax regret in partial monitoring. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2111–2139, Phoenix, USA. PMLR.

Lin, T., Abrahao, B., Kleinberg, R., Lui, J., and Chen, W. (2014). Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pages 901–909.

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

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

McMahan, H. B. and Streeter, M. J. (2009). Tighter bounds for multi-armed bandits with expert advice. In *COLT*.

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

Pogodin, R. and Lattimore, T. (2019). On first-order bounds, variance and gap-dependent bounds for adversarial bandits.

Rakhlin, S. and Sridharan, K. (2013). Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074.

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

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

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

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

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

Zimmert, J. and Lattimore, T. (2019). Connections between mirror descent, Thompson sampling and the information ratio. *arXiv preprint arXiv:1905.11817*.

Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7683–7692, Long Beach, California, USA. PMLR.

# A Technical lemmas

**Lemma 14** (Pogodin and Lattimore 2019)**.** *Let $(a_t)_{t=1}^n$ be a sequence of non-negative reals. Then*

$$\sum_{t=1}^n \frac{a_t}{\sqrt{1+\sum_{s=1}^{t-1} a_s}} \leq 4\sqrt{1+\frac{1}{2}\sum_{t=1}^n a_t} + \max_{t\in[n]} a_t\,.$$

**Lemma 15.** *Let $\alpha > 0$ and $(a_t)_{t=1}^n$ be a sequence of non-negative reals with $a_{t+1} \leq a_t + \alpha a_t^{1/4}$. Then*

$$a_n \leq \left(\frac{3\alpha(n-1)}{4} + a_1^{3/4}\right)^{4/3}\,.$$

*Proof.* Consider the differential equation $y(0) = a_1$ and $y'(t) = \alpha y(t)^{1/4}$, which has solution

$$y(t) = \left(\frac{3\alpha t}{4} + a_1^{3/4}\right)^{4/3}\,.$$

By comparison, $a_n \leq y(n-1)$ and the result follows. □