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 partial monitoring games. For non-degenerate locally observable games the n-round minimax regret is bounded by $2mk^{3/2}\sqrt{3n\log(k)}$, matching the best known information-theoretic upper bound. The same algorithm also achieves near-optimal regret for full information, bandit and globally observable games. High probability bounds and simple experiments are also provided.

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 $[m] = \{1, \dots, m\}$. 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 Σ is an arbitrary finite set. Both Φ 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

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

The minimax regret is

$$\mathfrak{R}_n^* = \inf_\pi \sup_{(x_t)_{t=1}^n} \mathbb{E}\left[\mathfrak{R}_n\right] ,$$

where the expectation is with respect to the randomness in the actions and π is the policy of the learner mapping action/observation sequences to distributions over the actions. Our main contribution is a simple and efficient algorithm for finite non-degenerate locally observable partial monitoring games for which

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

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

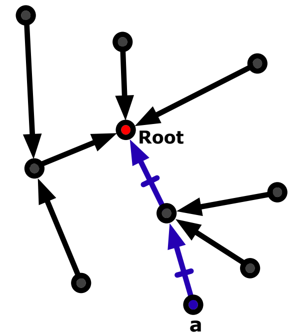
The best known bound for an efficient algorithm for 'easy' games is $\mathbb{E}[\mathfrak{R}_n] \leq C(\Phi, \mathcal{L})\sqrt{n\log(n)}$, where the constant $C(\Phi, \mathcal{L})$ can be arbitrarily large, even for fixed k and d [Foster and Rakhlin, 2012; Lattimore and Szepesvári, 2019a]. Furthermore, 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 $\mathfrak{R}_n^* \leq mk^{3/2}\sqrt{2n\log(k)}$ [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 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 a 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 $\|x\|_A^2 = x^\top Ax$ 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 $\mathrm{root}_{\mathcal{T}} \in [k]$ such that for all $a \in [k]$ there is a directed path $\mathrm{path}_{\mathcal{T}}(a) \subseteq \mathcal{T}$ from a to the root. The path from the root is the empty set: $\mathrm{path}_{\mathcal{T}}(\mathrm{root}_{\mathcal{T}}) = \emptyset$. The figure depicts an in-tree over $k = 10$ vertices. The blue (barred) path is $\mathrm{path}_{\mathcal{T}}(a)$.

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 ν. 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 ν. 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.

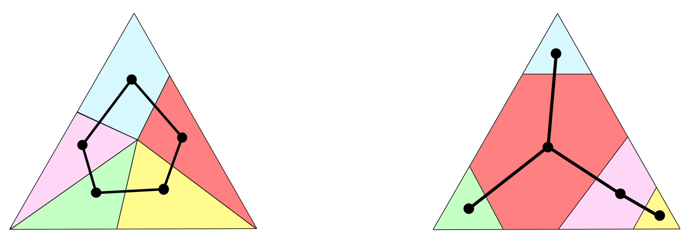


Figure 1: Cell decompositions and neighbourhood graphs for two games with $d = 3$ and $k = 5$.

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 optimal simultaneously. 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. The set of Pareto optimal actions is denoted by $\Pi = \{a : \dim(C_a) = d - 1\}$. An action a with $C_a \neq \emptyset$ and $\dim(C_a) \leq d - 2$ is called degenerate while actions with $C_a = \emptyset$ are dominated. Distinct actions a and b are duplicates if $(e_a - e_b)^\top \mathcal{L} = \mathbf{0}$. 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 a and b are duplicates. 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 it has no degenerate actions. 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, non-degenerate and contains no duplicate actions.

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]. No modifications to the algorithm are required.

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. A game is globally observable if for all edges $e = (a, b) \in \mathcal{E}$ in the neighbourhood graph there exists a function $w_e : [k] \times \Sigma \to \mathbb{R}$ such that

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

A non-degenerate game is locally observable if Eq. (2) holds and additionally w_e can be chosen so that $w_e(c, \sigma) = 0$ for all $c \notin \{a, b\}$ and all σ. 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 $w_e(c, \sigma) = 0$ 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. Let $\mathcal{H}$ be the set of all functions $G : [k] \times \Sigma \to \mathbb{R}^k$.

Lemma 4. *If $\mathcal{G}$ is globally observable, then there exists a function $G \in \mathcal{H}$ such that for all $b, c \in \Pi$,*

$$\sum_{a=1}^{k} (G(a, \Phi_{ax})_b - G(a, \Phi_{ax})_c) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \,.$$

Proof. Let $\mathcal{T} \subseteq \mathcal{E}$ be any in-tree over Π and for $b \in \Pi$ let $G(a, \sigma)_b = \sum_{e \in \text{path}_{\mathcal{T}}(b)} w_e(a, \sigma)$. Then

$$\sum_{a=1}^{k} G(a, \Phi_{ax})_b = \sum_{a=1}^{k} \sum_{e \in \text{path}_{\mathcal{T}}(b)} w_e(a, \Phi_{ax}) = \mathcal{L}_{bx} - \mathcal{L}_{\text{root}_{\mathcal{T}}\, x}\,.$$

The result follows by repeating the argument for $c \in \Pi$ and taking the difference. $\square$

Given a distribution $p \in \mathcal{P} \cap (0,1)^k$ and G 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,

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

In other words, the function G can be used with importance-weighting to estimate the loss differences. The set of functions that satisfy the consequences of Lemma 4 are denoted by

$$\mathcal{H}_\circ = \left\{ G : (e_b - e_c)^\top \sum_{a=1}^{k} G(a, \Phi_{ax}) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \text{ for all } b, c \in \Pi \text{ and } x \in [d] \right\}.$$

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 $\Phi_{ax} = (\mathcal{L}_{1x}, \ldots, \mathcal{L}_{kx})$. Estimation functions witnessing the conclusion of Lemma 4 are easily constructed. The obvious choice for bandit games is $G(a, \sigma) = e_a \sigma$ while for full information games $G(a, \sigma) = p_a \sigma$ 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

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

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 Ψ is sometimes called the stability term. The following inequality is useful:

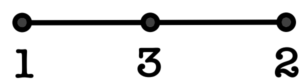
$$\Psi_q(\eta y) \leq \begin{cases} \eta^2 \left\|y\right\|^2_{\mathrm{diag}(q)} \,, & \text{if } \eta y \geq -\mathbf{1}\,; \\ \frac{1}{2}\eta^2 \left\|y\right\|^2_{\mathrm{diag}(q)} \,, & \text{if } \eta y \geq \mathbf{0}\,, \end{cases} \tag{6}$$

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, called costly matching pennies, is helpful to gain some intuition:

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ c & c \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} \bot & \bot \\ \bot & \bot \\ \mathrm{H} & \mathrm{T} \end{pmatrix}. \tag{7}$$



The figure on the right is the neighbourhood graph when $c = 1/4$, which shows the first two actions are 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 function 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].

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 function that determine the loss estimators. Given an estimation function $G \in \mathcal{H}$ and outcome $x \in [d]$, define a 'bias' function that measures the degree of bias when using importance-weighting to estimate loss differences:

$$\mathrm{bias}_q(G; x) = \left\langle q,\, \mathcal{L}e_x - \sum_{a=1}^k G(a, \Phi_{ax}) \right\rangle + \max_{c \in \Pi} \left(\sum_{a=1}^k G(a, \Phi_{ax})_c - \mathcal{L}_{cx} \right).$$

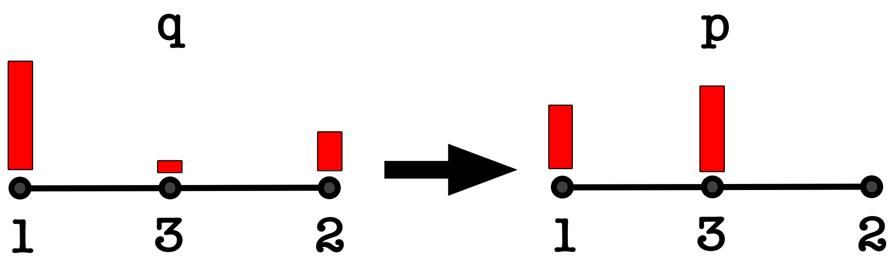


Figure 2: An exploration distribution p derived from q for the game in Eq. (7). The expected loss when playing p is smaller than playing q and simultaneously more information is gained because the third action is revealing.

As a function of G the bias is max-affine and hence convex. It is always non-negative and vanishes when the estimation function $G \in \mathcal{H}_\circ$ is unbiased. For $q \in \mathcal{P}$ and $\eta > 0$ let $\operatorname{opt}_q(\eta)$ be the value of the following convex optimisation problem:

$$\operatorname*{minimise}_{G \in \mathcal{H}, p \in \mathcal{P}} \quad \max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x + \operatorname{bias}_q(G; x)}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^k p_a \Psi_q \left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right] . \tag{9}$$

We assume that Eq. (9) can be solved exactly to obtain minimising values for $G \in \mathcal{H}$ and $p \in \mathcal{P}$. Our algorithm, however, is robust to small perturbations of these quantities. Numerical issues and a practical approximation are discussed in Appendix F. Let

$$\operatorname{opt}_*(\eta) = \sup_{q \in \mathcal{P}} \operatorname{opt}_q(\eta) .$$

Note that both $\operatorname{opt}_q(\eta)$ and $\operatorname{opt}_*(\eta)$ depend on $\mathcal{G}$; this dependence is not shown to minimize clutter. The optimisation problem can be formulated as an exponential cone problem and solved using off-the-shelf solvers. The algorithm is a simple combination of exponential weights using the exploration distribution and estimation function provided by solving Eq. (9).

input: η

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

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

Solve (9) with $q = Q_t$ to find $P_t \in \mathcal{P}$ and $G_t \in \mathcal{H}$

Sample $A_t \sim P_t$, observe σ_t and compute $\hat{y}_t = \dfrac{G_t(A_t, \sigma_t)}{P_{tA_t}}$

Algorithm 1: Exponential weights for partial monitoring with confidence

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 $\operatorname{opt}_*(\eta)$ are provided subsequently.

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

Proof. Let $a^* = \arg\min_{a \in [k]} \sum_{t=1}^n \mathcal{L}(a, x_t)$ be the optimal action in hindsight, where ties are broken so that $a^* \in \Pi$ is Pareto optimal. Then

$$
\begin{aligned}
\mathbb{E}[\mathfrak{R}_n] &= \mathbb{E}\left[\sum_{t=1}^n \sum_{b=1}^k P_{tb}(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t})\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^n \Big((P_t - Q_t)^\top \mathcal{L} e_{x_t} + \operatorname{bias}_{Q_t}(G_t; x_t)\Big)\right] + \mathbb{E}\left[\sum_{t=1}^n \sum_{b=1}^k Q_{tb}(\hat{y}_{tb} - \hat{y}_{ta^*})\right]. \qquad (10)
\end{aligned}
$$

Next,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^n \sum_{b=1}^k Q_{tb}(\hat{y}_{tb} - \hat{y}_{ta^*})\right] &\leq \frac{\log(k)}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \Psi_{Q_t}(\eta \hat{y}_t)\right] \qquad (11) \\
&= \frac{\log(k)}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k P_{ta} \Psi_{Q_t}\left(\frac{\eta G_t(a, \Phi_{ax_t})}{P_{ta}}\right)\right],
\end{aligned}
$$

where Eq. (11) follows from Eq. (5) and the definitions of Q_t and $\hat{y}_t$. The result follows by combining Eq. (10) and the definition of $\operatorname{opt}_*(\eta)$. □

Applications Table 2 provides bounds on $\operatorname{opt}_*(\eta)$ 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	$\operatorname{opt}^*(\eta)$ 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	$c_{\mathcal{G}}/\sqrt{\eta}$	$\eta \leq 1/c_{\mathcal{G}}^2$	Prop. 11	$3(c_{\mathcal{G}}n/2)^{2/3}(\log(k))^{1/3}$
LOCALLY OBSERVABLE NON-DEGENERATE	$3k^3m^2$	$\eta \leq 1/(mk^2)$	Prop. 12	$2k^{3/2}m\sqrt{3n\log(k)}$

Table 2: Upper bounds on $\operatorname{opt}_*(\eta)$ 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 issues 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 $t = 1, 2, \dots,$:

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

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

Solve (9) with $\eta = \eta_t$ and $q = Q_t$ to find $V_t = \max\{0, \mathrm{opt}_{Q_t}(\eta_t)\}$ and corresponding P_t and G_t

Sample $A_t \sim P_t$, observe σ_t and compute $\hat{y}_t = \frac{G_t(A_t, \sigma_t)}{P_{tA_t}}$

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*

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

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

$$\mathbb{E}[\mathfrak{R}_n] = O\left(\sqrt{n \sup\{\max\{0, \mathrm{opt}_*(\eta)\} : \eta \le 1/B\}\log(k)} + B\log(k)\right)\,. \qquad (12)$$

This bound is most useful for full information, bandit and locally observable non-degenerate games when B can be chosen so that $\eta_1 \le 1/B$ 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 $\mathrm{opt}_*(\eta) \to \infty$ 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 $V_t = \max\{0, \mathrm{opt}_{Q_t}(\eta_t)\}$. Note, the learning rate η_t is non-increasing. Hence, by Eq. (5),

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

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 22 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 game.

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

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

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 $V_t = \max\{0, \mathrm{opt}_{Q_t}(\eta_t)\} \leq \alpha/\sqrt{\eta_t}$. 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 η_n and Lemma 23 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 $\operatorname{opt}_*(\eta)$ for bandit, full information and globally observable 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,* $\operatorname{opt}_*(\eta) \leq k/2$ *for all* $\eta > 0$.

Proof. Let $q \in \mathcal{P}$ be arbitrary and let $p = q$ and $G(a, \sigma) = e_a\sigma$. This corresponds to the usual importance-weighted estimators used for k-armed bandits. Then

$$\operatorname{opt}_q(\eta) \leq \max_{x\in[d]} \frac{1}{\eta^2}\sum_{a=1}^{k} p_a \Psi_q\left(\frac{\eta G(a,\Phi_{ax})}{p_a}\right) = \max_{x\in[d]} \frac{1}{\eta^2}\sum_{a=1}^{k} p_a \Psi_q\left(\frac{\eta \mathcal{L}_{ax} e_a}{p_a}\right) \leq \frac{1}{2}\max_{x\in[d]}\sum_{a=1}^{k}\mathcal{L}_{ax}^2 \leq \frac{k}{2} ,$$

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

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

Proof. As in the previous proof let $p = q$, but now choose $G(a, \sigma) = p_a\sigma$, which is unbiased. The argument then follows along the same lines as the proof of Proposition 8. □

Remark 10. You might wonder whether these choices of p and G actually minimise Eq. (9), in which case the algorithm would reduce to Hedge for full information games. As we show in Appendix D, however, they do not. The minimisers of Eq. (9) shift the loss estimates and play a distribution that is close to q, but not exactly the same. A similar story holds for bandits.

Proposition 11. *For non-degenerate globally observable games there exists a constant* $c_{\mathcal{G}}$ *depending only on* Φ *and* $\mathcal{L}$ *such that for all* $\eta \leq 1/c_{\mathcal{G}}^2$*,*

$$\operatorname{opt}_*(\eta) \leq \frac{c_{\mathcal{G}}}{\sqrt{\eta}} .$$

Proof. By the definition of a globally observable game there exists an unbiased estimation function $G \in \mathcal{H}_\circ$. Let $\beta = \|G\|_\infty$ 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 claim that $\eta G(a,\Phi_{ax})/p_a \geq -\mathbf{1}$, which follows from the definitions of γ and β so that

$$p_a\mathbf{1} \geq \frac{\gamma}{k}\mathbf{1} = \beta\sqrt{\eta}\mathbf{1} \geq \beta\eta\mathbf{1} \geq \eta G(a,\Phi_{ax}) ,$$

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),

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

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 12. *For locally observable non-degenerate games and $\eta \leq 1/(2mk^2)$,*

$$\mathrm{opt}_*(\eta) \leq 3m^2k^3\,.$$

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 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 $\nu \in C_a$. 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 ν.

Lemma 13 (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 $(a, b) \in \mathcal{T}$.

A simplified proof of the above lemma is provided for completeness in Appendix B.

Proof of Proposition 12. Let $q \in \mathcal{P}$. By Sion's minimax theorem

$$\begin{aligned}\operatorname{opt}_q(\eta) &\leq \min_{G \in \mathcal{H}_\circ, p \in \mathcal{P}} \max_{\nu \in \mathcal{D}} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu + \frac{1}{\eta^2}\sum_{x=1}^d \nu_x \sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G(a, \Phi_{ax})}{p_a}\right)\right] \\ &= \max_{\nu \in \mathcal{D}} \min_{G \in \mathcal{H}_\circ, p \in \mathcal{P}} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu + \frac{1}{\eta^2}\sum_{x=1}^d \nu_x \sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G(a, \Phi_{ax})}{p_a}\right)\right],\end{aligned}$$

where in the first inequality we added the constraint that $G \in \mathcal{H}_\circ$, which zeros the bias term. Let $\nu \in \mathcal{D}$ and let $\mathcal{T}$ and $r = W_\nu(q)$ be the in-tree over $[k]$ and distribution in $\mathcal{P}$ provided by the water transfer operator (Lemma 13). Define $G \in \mathcal{H}$ by

$$G(a, \sigma)_b = \sum_{e \in \operatorname{path}_\mathcal{T}(b)} w_e(a, \sigma)\,.$$

By Lemma 20 and the assumption that $\mathcal{G}$ is non-degenerate, w_e can be chosen so that $\|w_e\|_\infty \leq m/2$. Since paths in $\mathcal{T}$ have length at most k it follows that

$$\|G\|_\infty \leq km/2\,.$$

Furthermore, $G \in \mathcal{H}_\circ$ by the proof of Lemma 4. Then let $\gamma = \eta m k^2/2$ and $p = (1-\gamma)r + \gamma \mathbf{1}/k$, which means that for any $x \in [d]$,

$$\frac{\eta G(a, \Phi_{ax})}{p_a} \geq -\frac{\eta m k^2}{2\gamma} = -1\,.$$

Additionally, the assumption that $\eta \leq 1/(mk^2)$ means that $\gamma \leq 1/2$ so that $r \geq p/2$. Hence, by Eq. (6) and using Parts (b) and (c) of Lemma 13 with the definition of r,

$$\begin{aligned}\frac{1}{\eta^2}\sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G(a, \Phi_{ax})}{p_a}\right) &\leq \sum_{a=1}^k \frac{\|G(a, \Phi_{ax})\|^2_{\operatorname{diag}(q)}}{p_a} \\ &\leq 2\sum_{a=1}^k \frac{\|G(a, \Phi_{ax})\|^2_{\operatorname{diag}(q)}}{r_a} \\ &= 2\sum_{b=1}^k \sum_{a=1}^k \frac{q_b}{r_a}\left(\sum_{e \in \operatorname{path}_\mathcal{T}(b)} w_e(a, \Phi_{ax})\right)^2 \\ &\leq \frac{m^2}{2}\sum_{b=1}^k \sum_{a=1}^k \frac{q_b}{r_a}\left(\sum_{e \in \operatorname{path}_\mathcal{T}(b)} \mathbb{1}(a \in e)\right)^2 \\ &\leq 2k^3 m^2\,,\end{aligned}$$

where we used Part (b) of Lemma 13 to show that $q_b \leq k r_b$ and Part (c) to show that $r_a \geq r_b$ for $a \in \operatorname{path}_\mathcal{T}(b)$. Finally,

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

Hence $\operatorname{opt}_q(\eta) \leq 3k^3 m^2$. □

Remark 14. The bound can be improved to $\operatorname{opt}_q(\eta) \leq 3km^2 \operatorname{diam}(\mathcal{E})^2$, where $\operatorname{diam}(\mathcal{E})$ is the diameter of the neighbourhood graph.

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 $\Theta(d\sqrt{kn})$ is the correct worst-case regret over all d-outcome k-action non-degenerate locally observable partial monitoring games, at least as n tends to infinity.

High probability bounds By replacing the bias term in Eq. (9) with a constraint on a certain moment-generating function the algorithm can be adapted to prove high probability bounds. Details are provided in Appendix A.

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 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 $\mathrm{opt}_*(\eta)$ 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 $\mathrm{opt}_*(\eta)$ 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, perhaps 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

$$\mathbb{E}[\mathfrak{R}_n] \leq \frac{\operatorname{diam}_F(\mathcal{P})}{\eta} + \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^n \sum_{a=1}^k P_{ta} D_{F^*}\left(\nabla F(Q_t) - \frac{\eta G_t(a, \Phi_{ax_t})}{P_{ta}}, \nabla F(Q_t)\right)\right] \tag{15}$$
$$+ \mathbb{E}\left[\sum_{t=1}^n (P_t - Q_t)^\top \mathcal{L} e_{x_t} + \operatorname{bias}_{Q_t}(G_t; x_t)\right].$$

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. (15), 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 purely for simplicity. Only the proof of Proposition 12 and its dependents need to be modified in minor ways. The notable difference is that the magnitude of the estimation vectors is no longer guaranteed to be small. More specifically, Lemma 20 does not hold when estimating loss differences between actions (a, b) for which there are degenerate actions c with $C_c = C_a \cap C_b$. As in our previous work, using Proposition 21 instead introduces constants that may be exponential in d, which we believe is unavoidable [Lattimore and Szepesvári, 2019b]. Duplicate actions can be handled similarly and have the same affect.

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

Abernethy, J. D. and Rakhlin, A. (2009). Beating the adaptive bandit with high probability. In *COLT*.

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.

Alon, N. and Vũ, V. H. (1997). Anti-hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs. *Journal of Combinatorial Theory, Series A*, 79(1):133–160.

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.

Gerchinovitz, S. and Lattimore, T. (2016). Refined lower bounds for adversarial bandits. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, NIPS, pages 1198–1206. Curran Associates, Inc.

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*.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2017). SCS: Splitting conic solver, version 2.1.1. `https://github.com/cvxgrp/scs`.

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.

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 High probability bounds

The same design principle can be used to construct algorithms for which the regret is controlled with high probability. The idea is to replace the constraint in the optimisation problem that the estimators are unbiased with constraints on the range of the loss estimators and on an appropriately chosen moment-generating function. Given $q \in \mathcal{P}$ and $\eta > 0$, let $\operatorname{opthp}_q(\eta)$ be the solution to the following optimisation problem:

$$\begin{aligned}
&\underset{G \in \mathcal{H}, p \in \mathcal{P}, \lambda \geq 0}{\text{minimise}} && \lambda + \frac{2}{\eta^2} \max_{x \in [d]} \sum_{a=1}^k p_a \Psi_q\left(\frac{\eta G(a, \Phi_{ax})}{p_a}\right) \\
&\text{subject to} && \sum_{a=1}^k p_a \exp\left(\eta\left(\mathcal{L}_{ax} - \mathcal{L}_{cx} - \frac{\langle q - e_c, G(a, \Phi_{ax})\rangle}{p_a}\right)\right) \leq \exp(\lambda \eta^2) \\
&\text{and} && \eta \left\|G(a, \sigma)\right\|_\infty \leq p_a \text{ for all } a \text{ and } \sigma\,.
\end{aligned} \tag{16}$$

The optimisation problem in Eq. (16) is not convex, but the solution can be approximated efficiently within a factor of two. Let $\operatorname{opthp}_q(\eta, \lambda)$ be the optimal value of Eq. (16) with a fixed value of λ, which is convex. A larger value of λ leads to a larger constraint set and hence $\lambda \mapsto \operatorname{opthp}_q(\eta, \lambda) - \lambda$ is decreasing. Then the bisection method can be used to find (approximately) the value of λ such that $\operatorname{opthp}_q(\eta, \lambda) = 2\lambda$ and you can check that for this choice $\operatorname{opthp}_q(\eta, \lambda) \leq 2 \operatorname{opthp}_q(\eta)$. We also define

$$\operatorname{opthp}_*(\eta) = \sup_{q \in \mathcal{P}} \operatorname{opthp}_q(\eta)\,.$$

The algorithm is exactly the same as Algorithm 1 except that the optimisation problem in Eq. (16) is used instead of Eq. (9).

input: η

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

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

Solve (16) with $q = Q_t$ to find $\lambda_t \in \mathbb{R}$ and $P_t \in \mathcal{P}$ and $G_t \in \mathcal{H}$

Sample $A_t \sim P_t$ and observe σ_t

Compute $\hat{y}_t = \dfrac{G_t(A_t, \sigma_t)}{P_{tA_t}}$

Algorithm 3: Exponential weights for partial monitoring

Theorem 15. *With probability at least $1 - 2\delta$ the regret of Algorithm 3 is bounded by*

$$\mathfrak{R}_n \leq \frac{\log(k) + 2\log(1/\delta)}{\eta} + \eta n \operatorname{opthp}_*(\eta)\,.$$

Proof. Let $(\lambda_t)_{t=1}^n$ be the sequence of real values as defined in the algorithm. Using Lemma 26, the regret is bounded with probability at least $1 - \delta$ by

$$\mathfrak{R}_n = \sum_{t=1}^n (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t}) \leq \frac{\log(1/\delta)}{\eta} + \eta \sum_{t=1}^n \lambda_t + \sum_{t=1}^n \sum_{b=1}^k Q_{tb} \left(\hat{y}_{tb} - \hat{y}_{ta^*}\right)\,.$$

The second sum is bounded as in the proof of Theorem 5 using Eq. (5) by

$$\sum_{t=1}^n \sum_{b=1}^k Q_{tb} \left(\hat{y}_{tb} - \hat{y}_{ta^*}\right) \leq \frac{\log(k)}{\eta} + \frac{1}{\eta} \sum_{t=1}^n \Psi_q\left(\eta \hat{y}_t\right)\,.$$

Let $\mathbb{E}_{t-1}[\cdot]$ denote the expectation conditioned on the history observed after $t-1$ rounds. The last constraint in Eq. (16) ensure that $|\eta \hat{y}_t| \leq 1$. Therefore $\Psi_q(\eta \hat{y}_t) \in [0, 1]$ and

$$\mathbb{E}_{t-1}\left[\exp\left(\Psi_{Q_t}(\eta \hat{y}_t) - \mathbb{E}_{t-1}[\Psi_{Q_t}(\eta \hat{y}_t)]\right)\right] \leq 1 + \mathbb{E}_{t-1}\left[(\Psi_{Q_t}(\eta \hat{y}_t))^2\right] \leq \exp\left(\mathbb{E}_{t-1}\left[\Psi_{Q_t}(\eta \hat{y}_t)\right]\right)\,,$$

where we used that $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and that $\mathbb{E}[X^2] \leq \mathbb{E}[X]$ for random variables $X \in [0, 1]$. Hence, another application of Lemma 26 shows that with probability at least $1 - \delta$,

$$\begin{aligned}\frac{1}{\eta} \sum_{t=1}^{n} \Psi_{Q_t}\left(\eta \hat{y}_t\right) &\leq \frac{2}{\eta} \sum_{t=1}^{n} \mathbb{E}_{t-1}\left[\Psi_{Q_t}\left(\eta \hat{y}_t\right)\right] + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right) \\ &\leq \eta \sum_{t=1}^{n} \left(\operatorname{opthp}_{Q_t}(\eta) - \lambda_t\right) + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right).\end{aligned}$$

Combining the pieces shows that the regret is bounded with probability at least $1 - 2\delta$ by

$$\mathfrak{R}_n \leq \frac{\log(k) + 2\log(1/\delta)}{\eta} + \eta \sum_{t=1}^{n} \operatorname{opthp}_{Q_t}(\eta) \leq \frac{\log(k) + 2\log(1/\delta)}{\eta} + \eta n \operatorname{opthp}_*(\eta). \qquad \square$$

Remark 16. Algorithm 3 can be modified with a little effort to adapt the learning rate in a similar manner as Algorithm 2. The analysis remains more-or-less the same except a version of Lemma 26 must be proven for decreasing sequences of learning rates.

Applications Like $\operatorname{opt}_q(\eta)$, the quantity $\operatorname{opthp}_q(\eta)$ is game-dependent. In all the applications that we know of the stability component of the optimisation problem in Eq. (9) can be bounding by choosing p and G so that the loss estimators do not have magnitude larger than $1/\eta$ and then using the bounds on Ψ_q in Eq. (6). The following lemma extracts the core assumptions needed for this argument. Afterwards we give applications for full information, bandit and partial monitoring games.

Lemma 17. *Let $q \in \mathcal{P}$ and $\eta \in (0, 1/2)$ and suppose there exists a $p \in \mathcal{P}$ and $G \in \mathcal{H}_\circ$ and $\varphi \in [0, \infty)^k$ such that for all actions a and b and outcomes x,*

$$\left|\frac{\eta G(a, \Phi_{ax})_b}{p_a}\right| \leq \frac{1}{2} \qquad \text{and} \qquad \sum_{a=1}^{k} \frac{G(a, \Phi_{ax})_b^2}{p_a} \leq \varphi_b \qquad \text{and} \qquad \eta^2 \varphi_b \leq \frac{1}{2}. \tag{17}$$

Then $\operatorname{opthp}_q(\eta/2) \leq 3\langle q, \varphi\rangle + \frac{1}{\eta} \max_{x \in [d]} (p - q)^\top \mathcal{L} e_x$.

Proof. Let $\lambda = 2\langle q, \lambda\rangle$ and G' be given by $G'(a, \sigma) = G(a, \sigma) - \eta p_a \varphi$. At heart this is the same biased loss estimator used by Auer et al. [1995] and generalised by Abernethy and Rakhlin [2009]. We now get to work bounding the moment generating functions that appear in the constraint of Eq. (16). For any action c,

$$\begin{aligned}\sum_{a=1}^{k} p_a \exp\left(\eta\left(\frac{G'(a, \Phi_{ax})_c}{p_a} - \mathcal{L}_{cx}\right)\right) &= \sum_{a=1}^{k} p_a \exp\left(\eta\left(\frac{G(a, \Phi_{ax})_c}{p_a} - \mathcal{L}_{cx} - \eta\varphi_b\right)\right) \\ &\leq \left(1 + \varphi_b \eta^2 + \eta \mathcal{L}_{cx}\right) \exp\left(-\eta \mathcal{L}_{cx} - \varphi_b \eta^2\right) \\ &\leq 1,\end{aligned}$$

where we used the inequality $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and the conditions of the lemma and then the fact that $1 + x \leq \exp(x)$. Similarly, using $\exp(x) \leq 1 + x + x^2$ and the first and second conditions in Eq. (17),

$$\begin{aligned}
&\sum_{a=1}^{k} p_a \exp\left(\eta\left(\mathcal{L}_{ax} - \frac{\langle q, G'(a, \Phi_{ax})\rangle}{p_a}\right)\right) \\
&\qquad = \sum_{a=1}^{k} p_a \exp\left(\eta\left(\mathcal{L}_{ax} - \frac{\langle q, G(a, \Phi_{ax})\rangle}{p_a} + \eta\langle q, \varphi\rangle\right)\right) \\
&\qquad \leq \left(1 + \eta(p-q)^\top \mathcal{L} e_x + \eta^2 \sum_{a=1}^{k} p_a \left(\mathcal{L}_{ax} - \frac{\langle q, G(a, \Phi_{ax})\rangle}{p_a}\right)^2\right) \exp\left(\eta^2 \langle q, \varphi\rangle\right) \\
&\qquad \leq \exp\left(\eta(p-q)^\top \mathcal{L} e_x + 2\eta^2 \langle q, \varphi\rangle\right) .
\end{aligned}$$

Combining the previous two displays and the inequality $\exp(\eta(x+y)/2) \leq \exp(\eta x)/2 + \exp(\eta y)/2$ shows that

$$\sum_{a=1}^{k} p_a \exp\left(\frac{\eta}{2}\left(\mathcal{L}_{ax} - \mathcal{L}_{cx} - \sum_{b=1}^{k} \frac{\langle q - e_c, G(a, \Phi_{ax})\rangle}{p_a}\right)\right) \leq \exp\left(2\eta^2 \langle q, \varphi\rangle\right) .$$

The second part of the objective is bounded using the same idea:

$$\begin{aligned}
\sum_{a=1}^{k} p_a \Psi_q\left(\frac{\eta G'(a, \Phi_{ax})}{2p_a}\right) &= \sum_{a=1}^{k} p_a \left\langle q,\, \exp\left(-\frac{\eta G(a, \Phi_{ax})}{2p_a} + \frac{\eta^2 \varphi}{2}\right) - 1 + \frac{\eta G(a, \Phi_{ax})}{2p_a} - \frac{\eta^2 \varphi}{2}\right\rangle \\
&\leq \frac{1}{4} \sum_{a=1}^{k} p_a \sum_{b=1}^{k} q_b \left(-\frac{\eta G(a, \Phi_{ax})_b}{p_a} + \frac{\eta^2 \varphi_b}{2}\right)^2 \\
&\leq \frac{\eta^2 \langle q, \varphi\rangle}{2} + \frac{\eta^2}{2} \sum_{b=1}^{k} q_b \eta^2 \varphi_b^2 \\
&\leq \eta^2 \langle q, \varphi\rangle ,
\end{aligned}$$

where in the first inequality we again used that $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and the first and third conditions in Eq. (17). In the second inequality we used that $(x+y)^2 \leq 2x^2 + 2y^2$ and finally we used the assumption that $\eta^2 \varphi_b \leq 1$. Finally, by the first and third conditions in Eq. (17),

$$\eta \left\|G'(a, \Phi_{ax})\right\|_\infty = \eta \left\|G(a, \Phi_{ax}) - \eta p_a \varphi\right\|_\infty \leq \eta \left\|G(a, \Phi_{ax})\right\|_\infty + p_a \eta^2 \left\|\varphi\right\|_\infty \leq p_a .$$

Hence, $\mathrm{opt}_q(\eta/2) \leq 3\langle q, \varphi\rangle + \frac{1}{\eta} \max_{x \in [d]} (p-q)^\top \mathcal{L} e_x$. $\square$

Using the same analysis as in the proofs of Propositions 8, 9, 11 and 12 you can prove all the bounds in Table 3. For example, Lemma 17 can be applied to full information games by defining $G(a, \sigma) = p_a \sigma$ and $p = q$ and $\varphi = \mathbf{1}$. Then $\langle q, \varphi\rangle = 1$ and for $\eta \leq 1/2$ it follows that $\mathrm{opthp}^*(\eta) \leq 3$. And hence the familiar bound of $O(\sqrt{n \log(k/\delta)})$ is recovered using Theorem 15.

Remark 18. The bounds in Table 3 are obtained by tuning the learning rate in a manner that depends on δ. The learning rate can be tuned without the knowledge of δ, but then the dependence on $\log(1/\delta)$ moves outside the square root, a price that is known to be unavoidable [Gerchinovitz and Lattimore, 2016].

Game type	**Regret**
FULL INFORMATION	$O\left(\sqrt{n \log(k/\delta)}\right)$
BANDIT	$O\left(\sqrt{nk \log(k/\delta)}\right)$
GLOBALLY OBSERVABLE	$O\left((c_{\mathcal{G}} n)^{2/3} \log(k/\delta)^{1/3}\right)$
LOCALLY OBSERVABLE NON-DEGENERATE	$O\left(mk^{3/2}\sqrt{n \log(k/\delta)}\right)$

Table 3: High probability regret upper bounds that hold for a given $\delta \in (0,1)$. The constant $c_{\mathcal{G}}$ is game-dependent and can be exponentially large in d, which we believe is unavoidable.

B Water transfer operator

Here we provide a simple proof of Lemma 13. Let $\mathcal{T}$ be an in-tree over $[k]$. A vector $y \in \mathbb{R}^k$ is called $\mathcal{T}$-increasing if $y_a \leq y_b$ for all $(a,b) \in \mathcal{T}$, which means the function $a \mapsto y_a$ is increasing towards the root of $\mathcal{T}$. Similarly, y is $\mathcal{T}$-decreasing if $y_a \geq y_b$ for all $(a,b) \in \mathcal{T}$.

Lemma 19. *Given a tree $\mathcal{T}$ over $[k]$ and $q \in \mathcal{P}$, there exists an $r \in \mathcal{P}$ such that:*

(a) $r \geq q/k$.

(b) r *is* $\mathcal{T}$*-increasing.*

(c) $\langle r - q, y\rangle \leq 0$ *for all* $\mathcal{T}$*-decreasing* $y \in \mathbb{R}^k$.

Proof. Let $\operatorname{desc}_{\mathcal{T}}(a)$ be the descendants of a in $\mathcal{T}$ with the convention that $a \in \operatorname{desc}_{\mathcal{T}}(a)$. Define $d_{\mathcal{T}}(a)$ as the depth of a in $\mathcal{T}$ with $d_{\mathcal{T}}(\operatorname{root}_{\mathcal{T}}) = 1$. Define $r_a = \sum_{b \in \operatorname{desc}_{\mathcal{T}}(a)} q_b / d_{\mathcal{T}}(b)$, which is illustrated in Fig. 3. That $r \in \mathcal{P}$ follows since

$$\sum_{a=1}^k \sum_{b \in \operatorname{desc}_{\mathcal{T}}(a)} \frac{q_b}{d_{\mathcal{T}}(b)} = \sum_{b=1}^k \frac{q_b}{d_{\mathcal{T}}(b)} \sum_{a=1}^k \mathbb{1}(b \in \operatorname{desc}_{\mathcal{T}}(a)) = \sum_{b=1}^k q_b = 1\,.$$

Part (a) follows because $d_{\mathcal{T}}(b) \leq k$. Part (b) follows because if $(a,b) \in \mathcal{T}$, then $\operatorname{desc}_{\mathcal{T}}(a) \subset \operatorname{desc}_{\mathcal{T}}(b)$. For the last part, the fact that y is $\mathcal{T}$-decreasing means that

$$\langle r, y\rangle = \sum_{a=1}^k y_a \sum_{b \in \operatorname{desc}_{\mathcal{T}}(a)} \frac{q_b}{d_{\mathcal{T}}(b)} \leq \sum_{a=1}^k \sum_{b \in \operatorname{desc}_{\mathcal{T}}(a)} \frac{y_b q_b}{d_{\mathcal{T}}(b)} = \sum_{b=1}^k \frac{y_b q_b}{d_{\mathcal{T}(b)}} \sum_{a=1}^k \mathbb{1}(b \in \operatorname{desc}_{\mathcal{T}}(a)) = \langle q, y\rangle\,.$$

Rearranging completes the proof. □

Proof of Lemma 13. The result follows from Lemma 19 and by proving there exists an in-tree $\mathcal{T}$ over $[k]$ such that $\mathcal{L}\nu$ is $\mathcal{T}$-decreasing. We start by proving the existence of $\mathcal{T}$ when $\nu \in \operatorname{ri}(C_{a^*_\nu})$ for some Parent optimal action a^*_ν. Define a function $\operatorname{par} : [k] \to [k]$ by

$$\operatorname{par}(a) = \underset{b:(a,b)\in\mathcal{E}}{\arg\min}\, e_b^\top \mathcal{L}\nu\,,$$

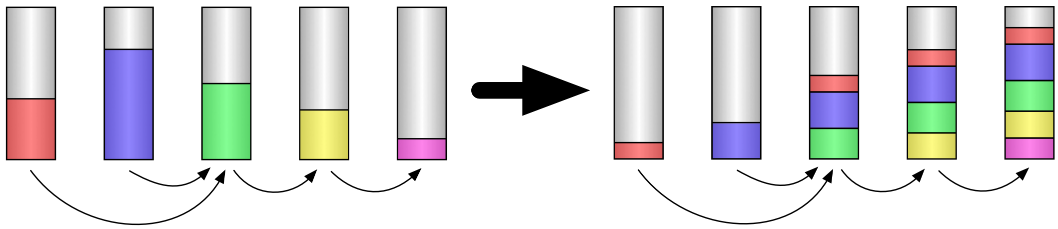


Figure 3: Illustration of r as defined in the proof of Lemma 13.

where the ties in the $\arg\min$ are broken arbitrarily. We will shortly show that $(e_a - e_{\mathrm{par}(a)})^\top \mathcal{L}\nu > 0$ for all $a \neq a^*_\nu$, which means that $\mathcal{T} = \{(a, \mathrm{par}(a)) : a \neq a^*_\nu\}$ is an in-tree over $[k]$ on which $\mathcal{L}^\top \nu$ is $\mathcal{T}$-decreasing. Let $D = \{(a, b, c) : \dim(C_a \cap C_b \cap C_c) \leq d - 3\}$ and

$$A = \bigcup_{a,b,c \in D} C_a \cap C_b \cap C_c\,.$$

Let $a \neq a^*_\nu$ and $\mu \in \mathrm{ri}(C_a)$ be such that the chord connecting μ and ν does not intersect A. Next, let $\rho \in \partial C_a$ be such that $\rho - \mu$ is proportional to $\nu - \mu$ and $b \neq a$ be an action with $\rho \in C_b$. Since $\mu \in \mathrm{ri}(C_a)$ we have $e_a^\top \mathcal{L}\mu < e_b^\top \mathcal{L}\mu$ and since $\rho \in C_a \cap C_b$ we have $e_a^\top \mathcal{L}\rho = e_b^\top \mathcal{L}\rho$. Hence $e_b^\top \mathcal{L}\nu < e_a^\top \mathcal{L}\nu$. The choice of μ ensures that $\rho \notin A$ and hence $(a, b) \in \mathcal{E}$, which means that $\mathrm{par}(a)$ is well defined and satisfies the claimed monotonicity conditions. Suppose now that ν is arbitrary and $a^*_\nu \in C_\nu$. Then take a sequence $(\nu_t)_{t=1}^\infty$ converging to ν and with $\nu_t \in \mathrm{ri}(a^*_\nu)$. By the previous argument there exists a sequence of in-trees $(\mathcal{T}_t)_{t=1}^\infty$ such that $\mathcal{L}\nu_t$ is $\mathcal{T}_t$-decreasing. Since the space of trees is finite, the sequence $(\mathcal{T}_t)_{t=1}^\infty$ has a cluster point $\mathcal{T}$ and it is easy to see that $\mathcal{L}\nu$ is $\mathcal{T}$-decreasing. □

C Bounds on the estimation functions

The polynomial dependence on k and m in locally observable non-degenerate games follows from the simple combinatorial structure when loss differences are estimated by playing two actions only. We provide the following lemma, which strengthens slightly our previous result [Lattimore and Szepesvári, 2019a].

Lemma 20. *If $\mathcal{G} = (\Phi, \mathcal{L})$ is locally observable and non-degenerate and actions $(a, b) \in \mathcal{E}$ are neighbours, then there exist functions $w_a, w_b : \Sigma \to \mathbb{R}$ such that $\|w_a\|_\infty \leq m/2$ and $\|w_b\|_\infty \leq m/2$ and*

$$\mathcal{L}_{ax} - \mathcal{L}_{bx} = w_a(\Phi_{ax}) + w_b(\Phi_{bx}) \text{ for all } x \in [d]\,. \tag{18}$$

Proof of Lemma 20. By the definition of local observability and non-degeneracy there exists w_a, w_b satisfying Eq. (18). Consider the bipartite graph over $V = \{(a, 1), \ldots, (a, m), (b, 1), \ldots, (b, m)\}$ and edges between vertices (a, σ) and (b, σ') if there exists an $x \in [d]$ such that $\Phi_{ax} = \sigma$ and $\Phi_{bx} = \sigma'$. Define a function $f : V \to \mathbb{R}$ by $f((a, \sigma)) = w_a(\sigma)$ and $f((b, \sigma)) = w_b(\sigma)$. Since entries in the loss matrix are bounded in $[0, 1]$ it holds that $f(w) + f(v) \in [0, 1]$ for all edges (w, v). The result follows from Lemma 25. □

For degenerate games the learner may need more than two actions to produce unbiased loss estimates, which unfortunately introduces the potential for an unpleasant combinatorial structure that makes learning much harder. Nevertheless, the norm of the estimation vectors can be uniformly bounded in terms of d and k.

Proposition 21. *Suppose that $(\mathcal{L}, \Phi)$ is globally observable and a and b are neighbours. Then there exists a function $w : [k] \times \Sigma \to \mathbb{R}$ such that for all $x \in [d]$,*

$$\sum_{c=1}^{k} w(c, \Phi_{cx}) = \mathcal{L}_{ax} - \mathcal{L}_{bx} .$$

Furthermore, w can be chosen so that $\|w\|_\infty \leq d^{1/2} k^{d/2}$.

Proof. 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. Let $S = (S_1^\top, \ldots, S_k^\top)$, which means that $S \in \mathbb{R}^{d \times mk}$. Then let $y = (e_a - e_b)^\top \mathcal{L} \in [-1,1]^k$. We identify w with a vector in $\mathbb{R}^{km}$. By the assumption of global observability there exists a $w \in \mathbb{R}^{km}$ such that $Sw = y$. Hence we may take $w = S^+ y$ with S^+ the Moore-Penrose pseudo-inverse and for which

$$\|w\|_\infty \leq \|w\|_2 \leq \|S^+\|_2 \, \|y\|_2 \leq d^{1/2} \|S^+\|_2 \leq d^{1/2} k^{d/2} ,$$

where $\|S^+\|_2$ is the spectral norm of S^+ and the final inequality follows from Lemma 24. □

D Non-equivalence to Hedge

Even in the full information setting Algorithm 1 does not reduce to Hedge, even when an additional constraint is added to Eq. (9) that the estimation function is unbiased: $G \in \mathcal{H}_\circ$. The full information game with binary losses and k actions has $d = 2^k$ outcomes, which we associate with $\{0,1\}^k$ via some arbitrary bijection and then view the outcomes as being in $\{0,1\}^k$ instead of $[d]$. The signal matrix is $\Phi_{ax} = x \in \{0,1\}^k$ and the loss matrix is $\mathcal{L}_{ax} = x_a$. Given distribution $q \in \mathcal{P}$, the estimation function $G \in \mathcal{H}_\circ$ that minimise the objective in Eq. (9) for the full information game can be calculated analytically as

$$G(a, \sigma) = p_a(\sigma + c(\sigma)) ,$$

where the shifting constant $c(\sigma)$ is given by

$$c(\sigma) = \frac{1}{\eta} \log\left(\langle q, \exp(-\eta\sigma)\rangle\right) = -\langle q, \sigma\rangle + O(\eta) .$$

The sampling distribution p should be the minimiser of

$$\frac{1}{\eta} \min_{p \in \mathcal{P}} \max_{x \in \{0,1\}^k} \left(\langle p - q, x\rangle + \frac{1}{\eta}\left(\eta\langle q, x\rangle + \log\left(\langle q, \exp(-\eta x)\rangle\right)\right)\right)$$
$$\approx \frac{1}{\eta} \min_{p \in \mathcal{P}} \max_{x \in \{0,1\}^k} \left(\langle p - q, x\rangle + \frac{\eta}{2}\langle q, x^2\rangle\right) .$$

The inner optimisation problem is not especially pleasant, but as η tends to zero the linear term dominates and the optimal p tends to q. Generally speaking, however, the optimal p is not equal to q. A numerical calculation shows that when $k = 2$ and $\eta = 0.5$ and $q = (0.9, 0.1)$, then the optimal p is approximately $p = (0.897, 0.103) \neq q$.

E Technical lemmas

Lemma 22 (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 23. *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. □

The next lemma provides a lower bound on the smallest non-zero eigenvalue of a positive semi-definite matrix with integer entries. Such results are somehow the reverse of the more well-known Hadamard problem of finding the maximum determinant [Alon and Vũ, 1997]. Presumably the naive bound below is known to experts, but a source seems hard to find.

Lemma 24. *Let $k \geq 3$ and $A \in \{0, \ldots, k\}^{d\times d}$ be non-zero and positive semi-definite with eigenvalues $\lambda_1, \ldots, \lambda_d$. Then $\min\{\lambda_i : \lambda_i > 0\} \geq k^{-d}$.*

Proof. Assume without loss of generality that $(\lambda_j)_{j=1}^d$ is decreasing and i is the index of the smallest non-zero eigenvalue. If $i = 1$, then $\lambda_i \geq 1$ and the result is immediate. Suppose now that $i > 1$. Since A has integer coefficients, the product of its non-zero eigenvalues is a positive integer, which means that $\prod_{i:\lambda_i>0} \lambda_i \geq 1$. Hence, by the arithmetic-geometric mean inequality,

$$\frac{1}{\lambda_i} \leq \prod_{j=1}^{i-1} \lambda_j \leq \left(\frac{1}{i-1}\sum_{j=1}^{i-1}\lambda_j\right)^{i-1} \leq \left(\frac{\operatorname{tr}(A)}{i-1}\right)^{i-1} \leq \left(\frac{dk}{i-1}\right)^{i-1} \leq k^d \,. \qquad \square$$

Lemma 25. *Let V_1 and V_2 be disjoint sets with $|V_1| = |V_2| = m$ and $V = V_1 \cup V_2$. Suppose that (V, E) is a bipartite graph with $E \subseteq V_1 \times V_2$ and $f : V \to \mathbb{R}$ is a function such that $f(u) + f(v) \in [0, 1]$ for all $(u, v) \in E$. Then there exists a function $g : V \to \mathbb{R}$ such that*

(a) $\|g\|_\infty \leq \frac{m}{2}$.

(b) $g(u) + g(v) = f(u) + f(v)$ *for all* $u, v \in E$.

Proof. We define g on each connected component of (V, E). For edge $(u, v) \in E$ we abuse notation by writing $f(e) = f(u) + f(v)$. Let $U \subseteq V$ be a connected component and $u = \arg\min_{v\in U\cap V_1} f(v)$.

$$g(v) = \begin{cases} f(v) - f(u) - m/2 + 1, & \text{if } v \in V_1; \\ f(v) + f(u) + m/2 - 1, & \text{if } v \in V_2 \,. \end{cases}$$

Then for any $v \in U \cap V_1$ there exists a path $(e_t)_{t=1}^n$ from v to u with $n \leq 2(m-1)$ and

$$g(v) + m/2 - 1 = g(v) - g(u) = \sum_{t=1}^n (-1)^{t+1} f(e_t) \leq m - 1\,.$$

Hence $g(v) \in [-m/2+1, m/2]$ for all $v \in U \cap V_1$ and so $g(v) \in [-m/2, m/2]$ for all $v \in U \cap V_2$. □

The following lemma has been seen before in many forms [Auer et al., 1995, for example] and follows immediately from the Chernoff method.

Lemma 26. *Suppose that $(X_t)_{t=1}^n$ is a sequence of random variables adapted to filtration $(\mathcal{F}_t)_{t=1}^n$ and $(\lambda_t)_{t=1}^n$ is $(\mathcal{F}_t)$-predictable and for $\eta > 0$,*

$$\mathbb{E}[\exp(\eta X_t - \lambda_t^2) \mid \mathcal{F}_{t-1}] \leq 1 \;\; a.s.\,.$$

Then for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\sum_{t=1}^n X_t \geq \sum_{t=1}^n \frac{\lambda_t^2}{\eta} + \frac{\log(1/\delta)}{\eta}\right) \leq \delta\,.$$

Proof. By Markov's inequality and the tower rule for conditional expectation,

$$\mathbb{P}\left(\exp\left(\sum_{t=1}^n \eta X_t - \lambda_t^2\right) \geq \frac{1}{\delta}\right) \leq \delta\,.$$

Re-arranging completes the proof. □

F A second-order cone approximation

The optimisation problem in Eq. (9) is convex and can be written as an exponential cone program. For small problems and reasonably large η it is amenable to standard methods. Numerical instability seems to be a problem when η is small, however. A practical resolution is to move some of the analysis into the optimisation problem by adding constraints on the magnitude of the estimation function and then approximating Ψ_q by an upper bound as in Eq. (6). This leads to the following formulation of the approximation of Eq. (9) as a second-order cone program:

$$\begin{array}{ll} \underset{G \in \mathcal{H}, p \in \mathcal{P}}{\text{minimise}} & \max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x + \text{bias}_q(G; x_t)}{\eta} + \sum_{a=1}^k \frac{\langle q, G(a, \Phi_{ax})^2 \rangle}{p_a}\right] \\ \text{subject to} & G(a,\sigma) + \frac{p_a}{\eta}\mathbf{1} \geq \mathbf{0} \text{ for all } a \text{ and } \sigma \\ \text{and} & p_a \geq \varepsilon \text{ for all } a\,. \end{array} \tag{19}$$

The first constraint justifies using the bound in Eq. (6) to approximate Ψ_q. The parameter $\varepsilon \geq 0$ in the second constraint is present to improve numerical stability and should be chosen so that its impact on the regret is negligible. For example, $\varepsilon = \eta^2$.

Let $\mathrm{opt}_q^{\sim}(\eta)$ be the optimal value of the above optimisation problem and

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

It is straightforward to show that the value of Eq. (9) at the optimiser of Eq. (19) is at most $\mathrm{opt}_q^{\sim}(\eta)$. Indeed, the upper bounds on $\mathrm{opt}_q(\eta)$ were all proven in this manner. We are not aware of a situation where $\mathrm{opt}_q(\eta) \ll \mathrm{opt}_q^{\sim}(\eta)$.

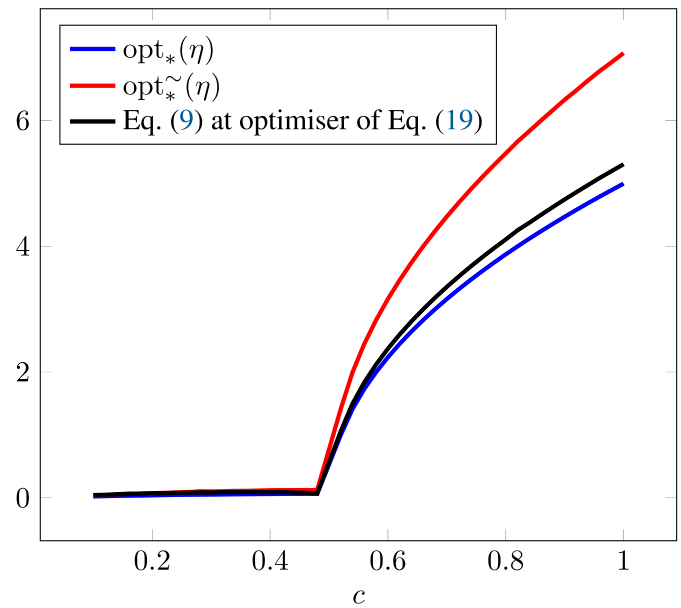


Figure 4: The plot illustrates the quality of the approximation in Eq. (19) for the matching pennies game with the cost varying on the x-axis and a fixed learning rate: $\eta = 0.01$. The blue line shows $\mathrm{opt}_*(\eta)$. The red line shows $\mathrm{opt}_*^{\sim}(\eta)$ and the black line is the value of Eq. (9) evaluated at the optimiser of Eq. (19). At least for this game the approximation is quite reasonable. The abrupt increase when $c > 1/2$ occurs because this is where the game transitions from being locally observable to only globally observable. Both Eq. (9) and Eq. (19) were solved using the Splitting Cone Solver [O'Donoghue et al., 2016, 2017].

G Experiments

In our simple experiments we use the Splitting Cone Solver [O'Donoghue et al., 2016, 2017] to solve the optimisation problem in Eq. (19). The performance of the algorithm is illustrated on the costly matching pennies game (Eq. (7)), which is locally observable and non-degenerate for $c \in (0, 1/2)$ and degenerate and globally observable for $c > 1/2$. When $c = 1/2$ it is degenerate and locally observable. When $c = 0$ it is trivial. The next figure shows the regret of ExpPM in costly matching pennies for two different values of c.

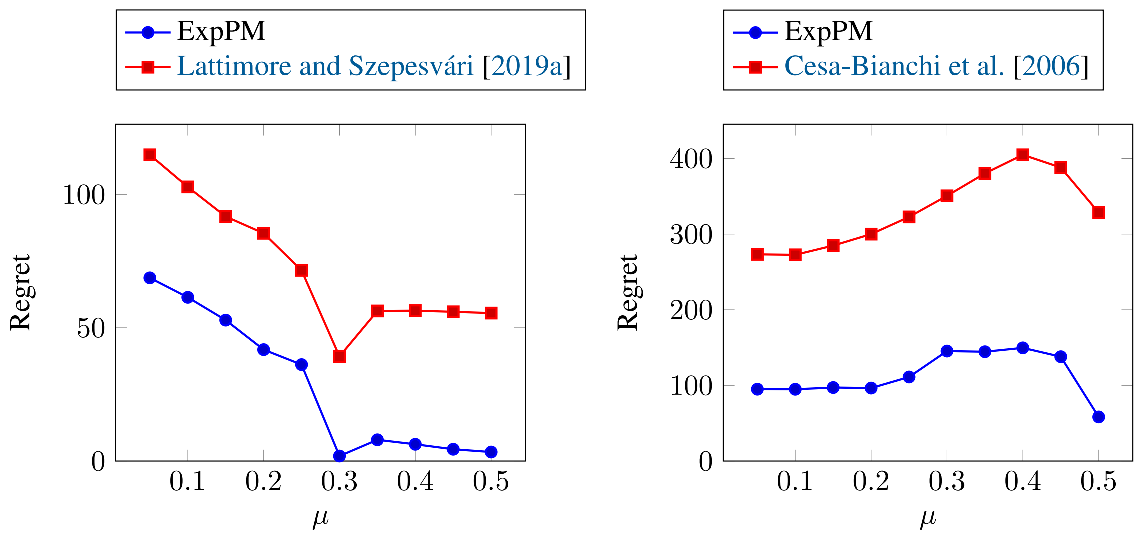


Figure 5: Costly matching pennies where the adversary is stochastic and samples from the outcomes i.i.d. from distribution $(\mu, 1-\mu)$. The horizon is $n = 2000$. On the left plot the cost is $c = 3/10$ and the algorithm is compared to Neighbourhood Watch 2 [Lattimore and Szepesvári, 2019a]. On the right plot the cost is $c = 1$ and the algorithm is compared to the algorithm by Cesa-Bianchi et al. [2006].