Cleaning up the neighborhood: A full classification for adversarial partial monitoring

Tor Lattimore
DeepMind

Csaba Szepesvári
DeepMind

Abstract

Partial monitoring is a generalization of the well-known multi-armed bandit framework where the loss is not directly observed by the learner. We complete the classification of finite adversarial partial monitoring to include *all* games, solving an open problem posed by Bartók et al. [2014]. Along the way we simplify and improve existing algorithms and correct errors in previous analyses. Our second contribution is a new algorithm for the class of games studied by Bartók [2013] where we prove upper and lower regret bounds that shed more light on the dependence of the regret on the game structure.

1 Introduction

Partial monitoring is a generalization of the bandit framework that relaxes the relationship between the feedback and the loss, which makes the framework applicable to a wider range of practical problems such as spam filtering and product testing. Equally importantly, it offers a rich and elegant framework to study the exploration-exploitation dilemma beyond bandits [Rustichini, 1999].

We consider the finite adversarial version of the problem where a learner and adversary interact over n rounds. At the start of the game the adversary secretly chooses a sequence of n outcomes from a finite set. In each round the learner chooses one of finitely many actions and receives a feedback that depends on its action and the choice of the adversary for that round. The loss is also determined by the action/outcome pair, but is not directly observed by the learner. Although the learner does not know the choices of the adversary, the feedback/loss functions are known in advance and the learner must use this infer a good policy. The learner's goal is to minimize the regret, which is the difference between the total loss suffered and the loss that would have been suffered by playing the best single action given knowledge of the adversaries choices.

The study of partial monitoring games started with the work by Rustichini [1999] where the definition of regret differed slightly from what is used here and the results have an asymptotic flavor. These results have been strengthened in an interesting line of work by Mannor and Shimkin [2003], Perchet [2011], Mannor et al. [2014], the last of which gives non-asymptotic rates for this more general definition of regret that unfortunately do not reduce to the optimal rate in our setting. The regret we consider was first considered by Piccolboni and Schindelhauer [2001], who showed that a variant of exponential weights achieves $O(n^{3/4})$ regret in nontrivial games. This was improved to $O(n^{2/3})$ by Cesa-Bianchi et al. [2006], who also showed that in general this result is not improvable, but that there exist many types of game for which the regret is $O(n^{1/2})$. They posed the question of classifying finite adversarial partial monitoring games in terms of the achievable minimax regret. An effort started around 2010 to achieve this goal, which eventually led to the paper by Bartók et al. [2014] who made significant progress towards solving this problem. In particular, they gave an almost complete characterization of partial monitoring games by identifying four regimes: trivial, easy, hard and hopeless games. The characterization, however, left out the set of games with actions that are only optimal on low-dimensional subspaces of the adversaries choices. Although these actions are never uniquely optimal, they can be informative and until now it was not known how to use these actions when balancing exploration and exploitation. Games in this tricky regime

have been called 'degenerate', but there is no particular reason to believe these games should not appear in practice. This problem is understood in the stochastic variant of partial monitoring where the adversary chooses the outcomes independently at random [Antos et al., 2013], but a complete understanding of the adversarial setup has remained elusive.

Contributions

- We develop an improved version of NEIGHBOURHOODWATCH by Foster and Rakhlin [2012] that correctly deals with degenerate games and completes the classification for *all* finite partial monitoring games, closing an open question posed by Bartók et al. [2014].[1] Another benefit is that Foster and Rakhlin [2012] and Bartók et al. [2014] inadvertently exchanged an expectation and maximum during the localisation argument of their analysis. A correction is presumably possible, but this would add another level of complexity to an already intricate proof. Our algorithm also enjoys a regret guarantee that holds with high probability.
- Bartók [2013] introduced a class of partial monitoring games and suggested a complicated algorithm with improved regret relative to NEIGHBOURHOODWATCH. We propose a novel algorithm and prove that for these games its regret satisfies $O(F\sqrt{nK_{\text{loc}}\log(K)})$, where K is the number of actions and F is the number of feedback symbols. The quantity K_{loc} depends on the game and satisfies $K_{\text{loc}} \leq K$. This bound improves on the result of Bartók [2013] in several ways: *(a)* we eliminate the dependence on arbitrarily large game-dependent constants, *(b)* the new algorithm is simpler, *(c)* our bound is better by logarithmic factors of the horizon and *(d)* the analysis by Bartók mistakenly combines bounds that hold in expectation in 'local games' into a bound for the whole game as if they were high probability bounds. We expect this could be corrected by modifying the algorithm and analysis, but the resulting algorithm would be even more complicated and the regret would not improve.
- We prove a variety of lower bounds. First correcting a minor error in the proof by Bartók et al. [2014] and second showing the linear dependence on the number of feedbacks is unavoidable in general.
- The new algorithms and analysis simplify existing results, which think is a contribution in its own right and we hope encourages more research into this fascinating topic with many open questions.

Problem setup Given a natural number n let $[n] = \{1, 2, \ldots, n\}$. We use $\langle x, y \rangle$ to denote the usual inner product in Euclidean space. The d-simplex is $\mathcal{P}_d = \{x \in [0,1]^{d+1} : \|x\|_1 = 1\}$, where for $p \geq 1$, $\|x\|_p$ is the p-norm of x. The relative interior of $\mathcal{P}_d$ is $\text{ri}(\mathcal{P}_d) = \{x \in (0,1)^{d+1} : \|x\|_1 = 1\}$. The dimension of a set $A \subset \mathbb{R}^{d+1}$ is the dimension of its affine hull. For any set A the indicator function is $\mathbb{1}_A(\cdot)$ and for function $f : A \to \mathbb{R}$ the supremum norm of f is $\|f\|_\infty = \sup_{a \in A} |f(a)|$. A partial monitoring problem $G = (\mathcal{L}, \Phi)$ is a game between a learner and an adversary over n rounds and is specified by a *loss matrix* $\mathcal{L} \in [0,1]^{K \times E}$ and a *feedback matrix* $\Phi \in [F]^{K \times E}$ for natural numbers E, F and K. At the beginning of the game the learner is given $\mathcal{L}$ and Φ and the adversary secretly chooses a sequence of *outcomes* $i_{1:n} = (i_1, \ldots, i_n)$ where $i_t \in [E]$ for each $t \in [n]$. In each round t the learner chooses an action $A_t \in [K]$ and observes feedback $\Phi_t = \Phi_{A_t i_t}$. The loss incurred by playing action a in round t is $y_{ta} = \mathcal{L}_{a i_t}$. In contrast to bandit and full information problems the loss in partial monitoring is *not* observed by the learner, even for the action played. A policy π is a map from sequences of action/observation pairs to a distribution over the action-set $[K]$. The performance of a policy π is measured by its *regret*, $R_n(\pi, i_{1:n}) = \max_{a \in [K]} \sum_{t=1}^n (y_{tA_t} - y_{ta})$. When the outcome sequence and policy are fixed we abbreviate $R_n = R_n(\pi, i_{1:n})$. The minimax expected regret associated with partial monitoring game G is the worst-case expected regret of the best policy. $R_n^*(G) = \inf_\pi \max_{i_{1:n}} \mathbb{E}[R_n(\pi, i_{1:n})]$ where the inf is taken over all policies, the max over all outcome sequences of length n and the expectation with respect to the randomness in the actions. We let $\mathcal{F}_t = \sigma(A_1, A_2, \ldots, A_t)$ be the σ-algebra generated by the information available after round t and abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. A core question in partial monitoring is to understand how $\mathcal{L}$ and Φ affect the growth of $R_n^*(G)$ in terms of the horizon. The

Game type	$R_n^*(G)$
Trivial	0
Easy	$\tilde{\Theta}(n^{1/2})$
Hard	$\Theta(n^{2/3})$
Hopeless	$\Omega(n)$

[1] Historical note: Foster and Rakhlin [2012] claim a modification of their argument would handle degenerate games but give no details. The followup paper explicitly mentions the difficulties and poses the open problem [Bartók et al., 2014, Remark 4 and §8].

main theorem of Bartók et al. [2014] shows that for all 'nondegenerate' games the minimax regret falls into one of four categories as illustrated in the table. The colloquial meaning of the adjective degenerate suggests that only nondegenerate games are interesting, but this is not the case. The term is used in a technical sense (to be clarified soon) referring to a subclass of games that we have no reason to believe should be less important than the nondegenerate ones.

Preliminaries To illustrate some of the difficulties of partial monitoring relative to bandits we formalize a simplistic version of the spam filtering problem.

Example 1 Let $c \geq 0$ and define partial monitoring game $G = (\mathcal{L}, \Phi)$ by

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ c & c \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

Loss ($\mathcal{L}$)	Spam	Not spam
Spam	0	1
Not spam	1	0
Don't know	c	c

Feedback (Φ)	Spam	Not spam
Spam	1	1
Not spam	1	1
Don't know	1	2

The idea is also illustrated in the tables on the right. Rows correspond to actions of the learner and columns to outcomes selected by the adversary. The learner has three actions in this game corresponding to 'spam', 'not spam' and 'don't know' while the adversary chooses between 'spam' and 'not spam'. The learner suffers a loss of 1 if it guesses incorrectly. Alternatively the learner can say they don't know in which case they suffer a loss of c and observe some meaningful feedback. The minimax regret for this game depends on the price of information. If $c > 1/2$, then the minimax regret is $\Theta(n^{2/3})$. On the other hand, if $c \in (0, 1/2]$ the minimax regret is $\tilde{\Theta}(n^{1/2})$ where $\tilde{\Theta}(\cdot)$ indicates growth up to logarithmic factors. Finally, when $c = 0$ a policy can suffer no regret by playing just the third action.

Example 2 The game on the right is hopeless because the learner cannot gain information about her loss and the adversary can always force the expected regret to be $\Omega(n)$.

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Cell decomposition In order to understand what makes a partial monitoring game hard, easy or hopeless, it helps to introduce a linear structure. Let $u_t = e_{i_t} \in \mathcal{P}_{E-1}$ be the standard basis vector that is nonzero in the coordinate of the outcome i_t chosen by the adversary in round t. For action a let $\ell_a \in [0,1]^E$ be the ath row of matrix $\mathcal{L}$. The *cell* C_a of action a is the subset of $\mathcal{P}_{E-1}$ on which action a is optimal: $C_a = \{u \in \mathcal{P}_{E-1} : \max_{b \in [K]} \langle \ell_a - \ell_b, u \rangle = 0\}$. Action a is optimal in hindsight if and only if $\frac{1}{n}\sum_{t=1}^{n} u_t \in C_a$. Each nonempty C_a is a polytope and the collection $\{C_a : a \in [K]\}$ is called the *cell decomposition* of G. An action is called *dominated* if it is never optimal: $C_a = \emptyset$. We define the *dimension* of nondominated action a to be the dimension of C_a, which ranges between 0 and $E-1$. Nondominated actions with dimension less than $E-1$ are called *degenerate* while actions with dimension $E-1$ are called *Pareto optimal*. A partial monitoring game is *degenerate* if it has at least one degenerate action. For each $u \in \mathcal{P}_{E-1}$ let $a_u^* \in \arg\min_a \langle \ell_a, u \rangle$ and $a_t^* \in \arg\min_a \sum_{s=1}^{t} \langle \ell_a, u_s \rangle$, which means that a_u^* is an optimal action if the adversary is playing u on average and a_t^* is the optimal action in hindsight when the adversary plays the sequence $(u_1, \ldots, u_t)$. Without loss of generality we assume that a_u^* and a_t^* are nondegenerate. A pair of nondegenerate actions a, b are *neighbors* if $C_a \cap C_b$ has dimension $E-2$. They are *weak neighbors* if $C_a \cap C_b \neq \emptyset$. Actions a and b are called *duplicates* if $\ell_a = \ell_b$. We let $\mathcal{N}_a$ be the set of actions consisting of a and its neighbors (but *not* the duplicates of a). For any pair of neighbors (a, b) let $\mathcal{N}_{ab} = \{c \in [K] : C_a \cap C_b \subseteq C_c\}$. Although a is not a neighbor of itself we define $\mathcal{N}_{aa} = \emptyset$.

Lemma 1 (Bartók et al. 2014, Lem. 11)**.** *Let a and b be neighbors. Then for all $d \in \mathcal{N}_{ab}$ there exists a unique $\alpha \in [0,1]$ such that $\ell_d = \alpha \ell_a + (1-\alpha)\ell_b$.*

A corollary is that for $d \in \mathcal{N}_{ab}$ and if α from the lemma lies in $(0,1)$, then $C_d = C_a \cap C_b$. Degenerate and dominated actions can never be uniquely optimal in hindsight, but they can provide information to the learner that proves the difference between a hard and hopeless game (or easy and hard). This is also true for duplicate actions, which have the same loss, but not necessarily the same feedback.

Observability The neighborhood structure determines which actions can be uniquely optimal and when. This is only half of the story. The other half is the relationship between the feedback and

loss matrices that defines the difficulty of identifying the optimal action. A natural first attempt towards designing an algorithm would be to construct an unbiased estimator of y_{ta} for each Pareto optimal action a. A moments thought produces easy games where this is impossible (Exhibit 1 in Appendix I). A more fruitful idea is to estimate the loss differences $y_{ta} - y_{tb}$ for Pareto optimal actions a and b, which is sufficient (and essentially necessary) to discover the optimal action. Suppose in round t the learner has chosen to sample $A_t \sim P_t$ where $P_t \in \mathrm{ri}(\mathcal{P}_{K-1})$. A conditionally unbiased estimator of $y_{ta} - y_{tb}$ is a function $g : [K] \times [F] \to \mathbb{R}$ such that $\mathbb{E}_{t-1}[g(A_t, \Phi_t)] = \sum_a P_{ta} g(a, \Phi_{ai_t}) = y_{ta} - y_{tb}$. Whether or not such an estimator exists and its structure determines the difficulty of a partial monitoring game. A pair of actions (a, b) are called *globally observable* if there exists a function $v : [K] \times [F] \to \mathbb{R}$ such that

$$\sum_{c=1}^{K} v(c, \Phi_{ci}) = \ell_{ai} - \ell_{bi} \qquad \text{for all } i \in [E]\,. \tag{1}$$

They are *locally observable* if in addition to the above a and b are neighbors and $v(c, f) = 0$ whenever $c \notin \mathcal{N}_{ab}$. Finally, they are *pairwise observable* if $v(c, f) = 0$ whenever $c \notin \{a, b\}$. If the learner is sampling action A_t from distribution $P_t \in \mathrm{ri}(\mathcal{P}_{K-1})$, then the existence of a function satisfying Eq. (1) means that $v(A_t, \Phi_t)/P_{tA_t}$ is an unbiased estimator of $\langle \ell_a - \ell_b, u_t \rangle = y_{ta} - y_{tb}$. A game G is called *globally/locally observable* if all pairs of neighbors are globally/locally observable. A game is called *point-locally observable* if all pairs of weak neighbors are pairwise observable. The cell decomposition and observability structure for the spam game is described in detail in Exhibit 2. Note that in globally observable games it is easy to see that any pair of Pareto optimal actions are globally observable, not just the neighbors.

2 Classification theorem

The following theorem classifies partial monitoring games into four categories depending on the observability structure.

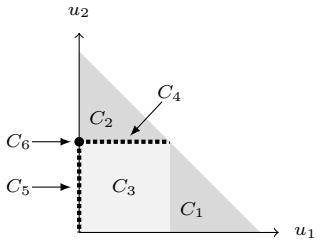
Theorem 1. *The minimax regret of partial monitoring game $G = (\mathcal{L}, \Phi)$ satisfies*

$$R_n^*(G) = \begin{cases} 0, & \text{if } G \text{ has no pairs of neighboring actions}\,; \\ \tilde{\Theta}(\sqrt{n}), & \text{if } G \text{ is locally observable and has neighboring actions}\,; \\ \Theta(n^{2/3}), & \text{if } G \text{ is globally observable, but not locally observable}\,; \\ \Omega(n), & \text{otherwise}\,. \end{cases}$$

The theorem follows by proving upper and lower bounds for each class of games. Most of the pieces already exist in the literature. The upper bound for globally observable games is by Cesa-Bianchi et al. [2006]. The upper bound for games with no pairs of neighboring actions is trivial, since in this case there exists an action a with $C_a = \mathcal{P}_{E-1}$ and playing this action alone ensures zero regret. The lower bound for easy games is by Antos et al. [2013, §6] and for hard games by Bartók et al. [2014, §4]. All that remains is to prove an upper bound for locally observable games with at least one pair of neighboring actions.

3 Algorithm for locally observable games

Fix a locally observable game $G = (\mathcal{L}, \Phi)$ with at least one pair of neighboring actions. We introduce a policy called NEIGHBORHOODWATCH2 (Algorithm 1).



Preprocessing The new algorithm always chooses its action $A_t \in \cup_{a,b} \mathcal{N}_{ab}$ where the union is over pairs of neighboring actions. For example, in the game with cell decomposition shown in the figure the policy only plays actions 1, 2, 3 and 4. Removing (some) degenerate actions can only increase the minimax regret so from now on we assume that all actions in $[K]$ are in $\mathcal{N}_{ab}$ for some neighbors a and b. Let $\mathcal{A}$ be an arbitrary largest subset of Pareto optimal actions such that $\mathcal{A}$ does not contain actions that are duplicates of each other and $\mathcal{D} = [K] \setminus \mathcal{A}$ be the remaining actions.

Estimating loss differences The definition of local observability means that for each pair of neighboring actions a, b there exists a function $v^{ab} : [K] \times [F] \to \mathbb{R}$ satisfying Eq. (1) and with $v^{ab}(c, f) = 0$ whenever $c \notin \mathcal{N}_{ab}$. Even though a is not a neighbor of itself, for notational convenience we define $v^{aa}(c, f) = 0$ for all c and f. The policy works for any such of v^{ab}, but the analysis suggests minimizing $V = \max_{a,b} \|v^{ab}\|_\infty$ with the maximum over all pairs of neighbors.

Algorithm 1 NEIGHBORHOODWATCH2

Input $\mathcal{L}$, Φ, η, γ
for $t \in 1, \ldots, n$ **do**
  For $a, k \in [K]$ let $Q_{tka} = \mathbb{1}_{\mathcal{A}}(k) \dfrac{\mathbb{1}_{\mathcal{N}_k \cap \mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)}{\sum_{b \in \mathcal{N}_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)} + \mathbb{1}_{\mathcal{D}}(k) \dfrac{\mathbb{1}_{\mathcal{A}}(a)}{|\mathcal{A}|}$
  Find distribution $\tilde{P}_t$ such that $\tilde{P}_t^\top = \tilde{P}_t^\top Q_t$
  Compute $P_t = (1-\gamma)$REDISTRIBUTE$(\tilde{P}_t) + \frac{\gamma}{K}\mathbf{1}$ and sample $A_t \sim P_t$
  Compute loss-difference estimators for each $k \in \mathcal{A}$ and $a \in \mathcal{N}_k \cap \mathcal{A}$.

$$\hat{Z}_{tka} = \frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}} \quad \text{and} \quad \beta_{tka} = \eta V^2 \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}} \quad \text{and} \quad \tilde{Z}_{tka} = \hat{Z}_{tka} - \beta_{tka} \tag{2}$$

end for
function REDISTRIBUTE(p)
  $q \leftarrow p$
  for $d \in \mathcal{D}$ **do**
    Find a, b such that $d \in \mathcal{N}_{ab}$ and $\alpha \in [0,1]$ such that $\ell_d = \alpha \ell_a + (1-\alpha)\ell_b$ (Lemma 1)
    $c_a \leftarrow \frac{\alpha q_b}{\alpha q_b + (1-\alpha) q_a}$ and $c_b \leftarrow 1 - c_a$ and $\rho \leftarrow \frac{1}{2K} \min\left\{ \frac{p_a}{q_a c_a}, \frac{p_b}{q_b c_b} \right\}$
    $q_d \leftarrow \rho c_a q_a + \rho c_b q_b$ and $q_a \leftarrow (1 - \rho c_a) q_a$ and $q_b \leftarrow (1 - \rho c_b) q_b$
  end for
  return q
end function

Description In each round the algorithm first computes a collection of exponential weights distribution $Q_{tk} \in \mathcal{P}_{K-1}$, one for each $k \in \mathcal{A}$. The distribution Q_{tk} is supported on the $\mathcal{N}_k \cap \mathcal{A}$ when $k \in \mathcal{A}$ and for $k \in \mathcal{D}$ it is uniform on $\mathcal{A}$. These local distributions are then combined into a global distribution $\tilde{P}_t$, which is taken to be the stationary distribution of right-stochastic matrix Q_t, which means that

$$\tilde{P}_{ta} = \sum_{k \in \mathcal{A}} \tilde{P}_{tk} Q_{tka} \text{ for any } a, k \in \mathcal{A}\,. \tag{3}$$

These steps are the same as the original NEIGHBORHOODWATCH, which samples its action from $(1-\gamma)\tilde{P}_t + \gamma \mathbf{1}/K$. This does not work when there are degenerate actions because $Q_{tkd} = 0$ when $d \in \mathcal{D}$, which by the above display means that $P_{td} = \gamma/K$ for actions $d \in \mathcal{D}$ and non-adaptive forced exploration is not sufficient for $O(\sqrt{n})$ regret in partial monitoring. This is the role of the redistribution function, which is analyzed formally in Appendix A. The final part of the algorithm is to estimate the loss differences for each $k \in \mathcal{A}$ and $a \in \mathcal{N}_k \cap \mathcal{A}$. Our choice of loss estimators are another departure from the original algorithm, which only updated the estimators for one local game in each round and then used a complicated aggregation strategy. This is one source of significant simplification in the new algorithm.

Remark 1. The special treatment of degenerate actions using the redistribution function seems like a big hassle. You might wonder why we did not simply include the degenerate actions in the local games and then play the stationary distribution, possibly with a little exploration. Unfortunately this idea does not work. Let d be a degenerate action in $\mathcal{N}_{ak}$ where a and k are neighbors. Then Lemma 1 shows that the loss-difference between k and d can be estimated by $\hat{Z}_{skd} = \alpha \hat{Z}_{skk} + (1-\alpha)\hat{Z}_{ska}$ with α such that $\ell_d = \alpha \ell_k + (1-\alpha)\ell_a$. Intuitively, a degenerate action d in $\mathcal{N}_{ak}$ is only useful for learning about the loss differences between actions a and k, which suggests the algorithm should not assign much more probability to d than the minimum probability of playing a and k. At a technical level the proof does not go through because the predictable variation of the estimator

above is roughly $\Omega(\max(1/P_{tk}, 1/P_{ta}))$ and yet P_{td} can be $\Omega(\max(P_{tk}, P_{ta}))$ and in the analysis of exponential weights these terms are required to cancel.

Remark 2. The estimators $\tilde{Z}_{tka}$ are negatively biased by β_{tka} in order to prove high probability bounds, which is reminiscent of the Exp3.P algorithm for finite-armed adversarial bandits [Auer et al., 2002]. As a minor contribution, we generalize their analysis to the case where the loss estimators satisfy certain constraints, rather than taking the specific importance-weighted form used for adversarial bandits. Choosing $\beta_{tka} = 0$ in the algorithm leads to a bound on the expected regret as we soon show.

Theorem 2. *Suppose Algorithm 1 is run on locally observable $G = (\mathcal{L}, \Phi)$ with parameters $\delta \in (0,1)$ and $\eta = \frac{1}{V}\sqrt{\log(K/\delta)/(nK)}$ and $\gamma = VK\eta$. Then with probability at least $1-\delta$ the regret is bounded by $R_n \leq C_G\sqrt{n\log(e/\delta)}$, where C_G is a constant that depends on the game G, but not the horizon n or confidence level δ.*

The complete proof of Theorem 2 given in Appendix B. Here we prove a bound on the expected regret in the simple case where there are no degenerate actions and $\beta_{tka} = 0$. Although this proof does not highlight one of our main contributions (how to deal with degenerate actions), it does emphasize the enormous simplification of the new algorithm. The first step is a localization argument to bound the regret in terms of the 'local regret' in each neighborhood. We need a simple lemma, which for completeness we prove in the the appendix.

Lemma 2 (Bartók et al. 2014)**.** *There exists a constant $\varepsilon_G > 0$ depending only on G such that for all pairs of actions $a, \tilde{a} \in \mathcal{A}$ and $u \in C_{\tilde{a}}$ there exists an action $b \in \mathcal{N}_a \cap \mathcal{A}$ such that $\langle \ell_a - \ell_{\tilde{a}}, u\rangle \leq \langle \ell_a - \ell_b, u\rangle/\varepsilon_G$.*

Since there are no degenerate actions, the REDISTRIBUTE function has no effect and $P_t = (1-\gamma)\tilde{P}_t + \gamma\mathbf{1}/K$. Let $B_1, \ldots, B_n$ be a sequence of random variables with $B_t \sim \tilde{P}_t$ that is conditionally independent of A_t given the observations up to time t. Then by Hoeffding-Azuma's inequality

$$R_n = \sum_{t=1}^{n}\langle \ell_{A_t} - \ell_{a_n^*}, u_t\rangle \leq n\gamma + \sqrt{8\log(1/\delta)} + \sum_{t=1}^{n}\langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle\,. \tag{4}$$

Next we apply Lemma 2 to localize the second term,

$$\text{(A)} = \sum_{t=1}^{n}\langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle = \sum_{a\in[K]}\left\langle \ell_a - \ell_{a_n^*}, \textstyle\sum_{t:B_t=a} u_t\right\rangle \leq \frac{1}{\varepsilon_G}\max_{\phi\in\mathcal{H}}\sum_{t=1}^{n}\langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\,,$$

where $\mathcal{H}$ is the set of functions $\phi : [K] \to [K]$ with $\phi(a) \in \mathcal{N}_a$ for all a. Then using Hoeffding-Azuma's inequality and a union bound over all $\phi \in \mathcal{H}$ shows that with probability at least $1-\delta$,

$$\begin{aligned}
\text{(A)} &\leq \frac{1}{\varepsilon_G}\max_{\phi\in\mathcal{H}}\sum_{t=1}^{n}\sum_{a\in[K]}\tilde{P}_{ta}\langle \ell_a - \ell_{\phi(a)}, u_t\rangle + \sqrt{8n\log\left(\frac{|\mathcal{H}|}{\delta}\right)} \\
&= \frac{1}{\varepsilon_G}\max_{\phi\in\mathcal{H}}\sum_{t=1}^{n}\sum_{k\in[K]}\tilde{P}_{tk}\sum_{a\in\mathcal{N}_k}Q_{tka}\langle \ell_a - \ell_{\phi(k)}, u_t\rangle + \sqrt{8n\log\left(\frac{|\mathcal{H}|}{\delta}\right)} \\
&= \frac{1}{\varepsilon_G}\sum_{k\in[K]}\underbrace{\max_{b\in\mathcal{N}_k}\sum_{t=1}^{n}\sum_{a\in\mathcal{N}_k}Q_{tka}(\tilde{P}_{tk}y_{ta} - \tilde{P}_{tk}y_{tb})}_{\text{local regret}} + \sqrt{8n\log\left(\frac{|\mathcal{H}|}{\delta}\right)}\,,
\end{aligned} \tag{5}$$

where the first equality uses the fact that $\tilde{P}_t$ is the stationary distribution of Q_t (see (3)). The local regret is bounded using the tools from online convex optimization. Of course the losses are never actually observed and must be replaced with the loss difference estimators. Then it remains to control the variance of these estimators. The 'standard' analysis of Exp3 [Auer et al., 1995, Cesa-Bianchi and Lugosi, 2006] shows that

$$\max_{b\in\mathcal{N}_k}\sum_{t=1}^{n}Q_{tka}\left(\tilde{P}_{tk}y_{ta} - \tilde{P}_{tk}y_{tb}\right) \leq \frac{\log(K)}{\eta} + \eta\sum_{t=1}^{n}\sum_{a\in\mathcal{N}_k}Q_{tka}\hat{Z}_{tka}^2\,. \tag{6}$$

In order to bound the second term we substitute the definition of $\hat{Z}_{tka}$, which shows that

$$\sum_{a\in\mathcal{N}_k} Q_{tka}\hat{Z}_{tka}^2 = \sum_{a\in\mathcal{N}_k} \frac{\tilde{P}_{tk}^2 Q_{tka} v^{ak}(A_t,\Phi_t)^2}{P_{tA_t}^2} \le \frac{\tilde{P}_{tk}V^2}{P_{tA_t}} \sum_{a\in\mathcal{N}_k} \frac{\tilde{P}_{tk}Q_{tka}\mathbb{1}_{\{a,k\}}(A_t)}{P_{tA_t}} \le \frac{2\tilde{P}_{tk}V^2}{P_{tA_t}}\,.$$

where in the first inequality we used the fact that $\|v^{ak}\|_\infty \le V$. The second inequality follows by considering two cases. First, if $A_t = k$, then all entries of the sum are non-zero and $\sum_{a\in\mathcal{N}_k} \tilde{P}_{tk}Q_{tka} = \tilde{P}_{tk} \le 2P_{tA_t}$, which is true by choosing $\gamma \le 1/2$. For the second case $A_t = a$ for $a \in \mathcal{N}_k$ and $a \neq k$, which means that only one term of the sum is non-zero. Then the definition of $\tilde{P}_t$ as the stationary distribution of Q_t means that $\tilde{P}_{tk}Q_{tka} \le \tilde{P}_{ta} \le 2P_{tA_t}$. Combining this with Eqs. (4) to (6) and a union bound shows that with probability at least $1-\delta$.

$$R_n \le n\gamma + \frac{1}{\varepsilon_G}\left(\frac{K\log(K)}{\eta} + 2\eta V^2 \sum_{t=1}^n \frac{1}{P_{tA_t}}\right) + \sqrt{8n\log(2/\delta)} + \sqrt{8n\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}\,.$$

Now $\mathbb{E}[\sum_{t=1}^n P_{tA_t}^{-1}] = nK$, which means that

$$\mathbb{E}[R_n] \le n\gamma + \frac{1}{\varepsilon_G}\left(\frac{K\log(K)}{\eta} + 2\eta nKV^2\right) + 2\sqrt{8n(1+\log(2|\mathcal{H}|))} = O\left(\frac{KV}{\varepsilon_G}\sqrt{n\log(K)}\right)\,,$$

where we first used Lemma 3 below along with naive bounding and the fact that $|\mathcal{H}| \le K^K$. The Big-O follows by choosing $\eta = \frac{1}{V}\sqrt{\log(K)/n}$ and $\gamma = \eta KV$. The choice of γ ensures that the loss-difference estimate satisfies $\eta|\hat{Z}_{tka}| \le \eta V/P_{tA_t} \le \eta VK/\gamma = 1$ on which the proof of Eq. (6) relies. We prove in Appendix G that for games without degenerate actions the loss-difference estimators can always be chosen so that $V \le 1+F$.

Lemma 3. *Suppose $a \ge 0$ and $b \ge 1$ are constants and X, Y are random variables such that $\mathbb{P}(X \ge Y + \sqrt{a\log(b/\delta)}) \le \delta$ for all $\delta \in (0,1)$. Then $\mathbb{E}[X] \le \mathbb{E}[Y] + \sqrt{a(1+\log(b))}$.*

Dealing with degenerate actions The presence of degenerate actions makes the calculation significantly more fiddly. The first step is to show that the redistribution process guarantees that the expected loss accumulated by playing P_t rather than $\tilde{P}_t$ is not too great. The localization argument is then repeated and the remaining question is how to control the variance of the loss difference estimates. The redistribution process guarantees that the degenerate actions have sufficient mass that the variance is at most $O(K)$ larger than what we saw in the above calculation. The process is complicated slightly by the desire to have a high probability bound.

4 Algorithm for point-locally observable games

The weakened neighbor definition and pairwise observability makes the analysis of point-locally observable games less delicate than locally observable games and the results are correspondingly stronger. Perhaps the most striking improvement is that asymptotically the bound does not dependent on arbitrarily large game-dependent constants. Here we present a simple new algorithm based on EXP3 called RELEXP3 ('Relative Exp3'). The name is derived from the fact that the algorithm does not estimate losses directly, but rather the loss differences relative to an 'anchor' arm that varies over time and is the arm to which the algorithm assigns the largest probability. As we shall see, this reduces the variance of the loss difference estimates.

Preprocessing The definition of pairwise observability means that degenerate and dominated actions are not needed to estimate the loss differences. Since removing these actions can only increase the minimax regret, for the remainder of this section we fix a point-locally observable game $G = (\mathcal{L}, \Phi)$ for which there are no dominated or degenerate actions. A *point-local game* is a largest subset of actions $A \subseteq [K]$ with $\bigcap_{a\in A} C_a \neq \emptyset$ (a maximal clique of the graph over actions with edges representing weak neighbors). We let K_{loc} be the size of the largest point-local game.

Estimation functions For each pair of actions a, b let v^{ab} be an estimation function satisfying Eq. (1) and furthermore assume that $v^{aa} = 0$ and $v^{ab}(c,f) = 0$ if a, b are weak neighbors and $c \notin \{a, b\}$. The existence of these functions is guaranteed by the definition of a point-locally observable game. Given pair of actions a, b let S^{ab} be the set of actions needed to estimate the loss difference

between a and b, which is $S^{ab} = \{a, b\} \cup \{c \in [K] : \text{exists } f \in [F] \text{ such that } v^{ab}(c, f) \neq 0\}$. Our assumptions ensure that $S^{ab} = \{a, b\}$ if a and b are weak neighbors. Define $V^{ab} = \|v^{ab}\|_\infty$ and $V = \max_{a,b \in [K]} V^{ab}$ and $V_{\text{loc}} = \max_{a,b:C_a \cap C_b \neq \emptyset} V^{ab}$. We show in Appendix G that v^{ab} can be chosen so that $V_{\text{loc}} \leq 1 + F$.

Decreasing learning rates The algorithm makes use of a sequence of decreasing learning rates $(\eta_t)_{t=1}^\infty$ and exploration parameters $(\alpha_t)_{t=1}^\infty$. On top of this the algorithm also has a dynamic exploration component that ensures the loss difference estimates are not too large. The decreasing learning rate is one of the essential innovations that allows us to prove an asymptotic bound that is independent of arbitrarily large game-dependent quantities. As an added bonus, it also means the algorithm does not require advance knowledge of the horizon.

Algorithm 2 RELEXP3

$\hat{L}_{0a} = 0$ for all $a \in [K]$
for $t = 1, \ldots, n$ **do**
  For each $a \in [K]$ let $\tilde{P}_{ta} = \dfrac{\exp(-\eta_t \hat{L}_{t-1,a})}{\sum_{b=1}^K \exp(-\eta_t \hat{L}_{t-1,b})}$
  Let $B_t = \arg\max_a \tilde{P}_{ta}$ and $M_t = \left\{ a : \tilde{P}_{ta} \exp\left(\dfrac{\eta_t V^{aB_t}}{\alpha_t} \right) > \dfrac{\eta_t}{t} \right\}$
  Let $S_t = \bigcup_{a \in M_t} S^{aB_t}$ and $\gamma_{ta} = \mathbb{1}_{S_t}(a) \eta_t \max_{a \in M_t} V^{aB_t} + \dfrac{\alpha_t}{K}$ and $P_t = (1 - \|\gamma_t\|_1) \tilde{P}_{ta} + \gamma_t$
  Sample $A_t \sim P_t$ and observe feedback Φ_t
  For each $a \in [K]$ compute estimates $\hat{Z}_{ta} = \dfrac{v^{aB_t}(A_t, \Phi_t)}{P_{tA_t}}$ and update $\hat{L}_{ta} = \hat{L}_{t-1,a} + \hat{Z}_{ta}$
end for

Theorem 3. *Let $G = (\mathcal{L}, \Phi)$ be point-locally observable, then with appropriately tuned parameters* RELEXP3 *satisfies* $\limsup_{n \to \infty} \mathbb{E}[R_n] / \sqrt{n} \leq 8 \sqrt{2 K_{loc} (1 + F)(2 + F) \log(K)}$. *Furthermore, the linear dependence on F is unavoidable (see Appendix E).*

Note that the constant hidden by the asymptotics *does* depend on arbitrarily large game-dependent constants. The proof of Theorem 3 may be found in the Appendix C, but the general idea is to show the forced exploration ensures for sufficiently large t that the algorithm is almost always playing in a point-local game that contains the optimal action and at this point the variance of the importance-weighted estimators is well behaved.

5 Summary and open problems

We completed the classification of all finite partial monitoring games. Along the way we greatly simplified existing algorithms and analysis and proved that for a large class of games the asymptotic regret does not depend on arbitrarily large game-dependent constants, which is the first time this has been demonstrated in the adversarial setting. There are many fascinating open problems. One of the most interesting is to understand to what extent it is possible to adapt to 'easy data'. For example, globally observable games may have locally observable subgames and one might hope for an algorithm with $O(\sqrt{n})$ regret if the adversary is playing in this subgame and $O(n^{2/3})$ regret otherwise. Another question is to refine the definition of the regret to differentiate between algorithms in hopeless games where linear regret is unavoidable, but the coefficient can depend on the algorithm [Rustichini, 1999]. Yet another question is to understand to what extent V is a fundamental quantity in the regret for easy games and whether or not the arbitrarily large game-dependent constants are real for large n as we have shown they are not for point-locally observable games.

References

András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *Conference on Learning Theory*, pages 696–710, 2013.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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

Dean Foster and Alexander Rakhlin. No internal regret via neighborhood watch. In *Artificial Intelligence and Statistics*, pages 382–390, 2012.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 02 1975.

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

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

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

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

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

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

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

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science*, pages 222–227, 1977.

A Redistribution properties

Here we collect a number of properties of the REDISTRIBUTE function in Algorithm 1.

Lemma 4. *Assume $\gamma \in [0, 1/2]$ and let $u \in \mathcal{P}_{E-1}$, and $k, a \in \mathcal{A}$ arbitrary neighbors. Then $P_t \in \mathcal{P}_{K-1}$ is a probability vector and the following hold:*

(a) $P_{ta} \geq \tilde{P}_{ta}/4$; *(b)* $\left|\sum_{a=1}^{K}(P_{ta} - \tilde{P}_{ta})\langle \ell_a, u\rangle\right| \leq \gamma$;

(c) $P_{tb} \geq \dfrac{\tilde{P}_{tk}Q_{tka}}{4K}$ *for any non-duplicate* $b \in \mathcal{N}_{ka}$; *(d)* $P_{ta} \geq \gamma/K$;

(e) $P_{td} \geq \dfrac{\tilde{P}_{tk}}{4K}$ *for any* $d \in [K]$ *such that* $\ell_d = \ell_k$.

Proof. First we show that P_t is indeed a probability vector. By assumption $\tilde{P}_t$ is the stationary distribution, which is a probability distribution. Let $\bar{P}_t = \textsc{Redistribute}(\tilde{P}_t)$ so that

$$P_t = (1-\gamma)\bar{P}_t + \frac{\gamma}{K}\mathbf{1}\,,$$

which means we need to show that $\bar{P}_t$ is a probability distribution. Since $\bar{P}_t$ is obtained by the iterative procedure given in the REDISTRIBUTE function it is sufficient to show that the vector q tracked by this algorithm is indeed a distribution. The claim is that each loop of the REDISTRIBUTE function does not break this property. The first observation is that the algorithm always moves mass from actions in $\mathcal{A}$ to actions in $\mathcal{D}$. All that must be shown is that $\bar{P}_{ta} \geq 0$ for all $a \in \mathcal{A}$. To see this note first that if $a \in \mathcal{A}$ is one of the choices of the algorithm in Line 11, then $\rho c_a q_a \leq p_a/(2K)$ and so

$$\bar{P}_{ta} \geq \tilde{P}_{ta}/2 \qquad \text{for all } a \in \mathcal{A} \geq 0\,. \tag{7}$$

Part (a): Since $\gamma \leq 1/2$ this follows from Eq. (7).

Part (b): First we show that $\sum_{a\in[K]}(\bar{P}_{ta} - \tilde{P}_{ta})\ell_a = 0$. It suffices to show that the redistribution in each inner loop of the algorithm does not change this value, which is true because

$$\begin{aligned}(c_a q_a + c_b q_b)\ell_d &= (c_a q_a + c_b q_b)(\alpha\ell_a + (1-\alpha)\ell_b) \\ &= \frac{q_a q_b}{\alpha q_b + (1-\alpha)q_a}(\alpha\ell_a + (1-\alpha)\ell_b) \\ &= \rho c_a q_a \ell_a + \rho c_b q_b \ell_b\,.\end{aligned}$$

Then using the definition of P_t we have

$$\left|\sum_{a\in[K]}(P_{ta} - \tilde{P}_{ta})\langle \ell_a, u\rangle\right| = \left|\sum_{a\in[K]}(P_{ta} - \bar{P}_{ta})\langle \ell_a, u\rangle\right| = \gamma\left|\sum_{a\in[K]}\left(\frac{1}{K} - \bar{P}_{ta}\right)\langle \ell_a, u\rangle\right| \leq \gamma\,,$$

where we used the assumption that $\ell_a \in [0,1]^E$ for all actions and $u \in \mathcal{P}_{E-1}$ so that $\langle \ell_a, u\rangle \in [0,1]$.

Part (c): There are three cases: Either $b = k$ or $b = a$ or b is degenerate. If $b = k$, then the result is immediate from Part (a). If $b = a$, then, Part (a) combined with (3) implies that $P_{tb} = P_{ta} \geq \tilde{P}_{ta}/4 \geq \tilde{P}_{tk}Q_{tka}/4 \geq \tilde{P}_{tk}Q_{tka}/(4K)$. Finally, if b is degenerate, then by the definition of the rebalancing algorithm we have

$$\bar{P}_{tb} \geq \frac{\min(\tilde{P}_{tk}, \tilde{P}_{ta})}{2K} \geq \frac{\min(\tilde{P}_{tk}, \tilde{P}_{tk}Q_{tka})}{2K} = \frac{\tilde{P}_{tk}Q_{tka}}{2K}$$

and the result follows from Eq. (7).

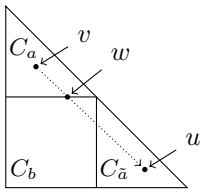
Part (d): This is trivial from the definition of P_t.

Part (e): Let $b \in \mathcal{A}$ be the Pareto optimal action chosen by the rebalancing algorithm when d is given weight. Since $\ell_d = \ell_a$ it follows that $\alpha = 1$ and so $c_a = 1$ and $c_b = 1$, which means that $\bar{P}_{td} = \tilde{P}_{ta}/2$ and using Eq. (7) again yields the result. □

B Proof of Theorem 2

We start by proving Lemma 2.

Proof of Lemma 2. Since $u \in C_{\tilde{a}}$, $0 \le \langle \ell_a - \ell_{\tilde{a}}, u\rangle$. The result is trivial if $a, \tilde{a}$ are neighbors or $\langle \ell_a - \ell_{\tilde{a}}, u\rangle = 0$. From now on assume that $\langle \ell_a - \ell_{\tilde{a}}, u\rangle > 0$ and that $a, \tilde{a}$ are not neighbors. Let v be the centroid of C_a and consider the line segment connecting v and u. Then let w be the first point on this line segment for where there exists a $b \in \mathcal{N}_a \cap \mathcal{A}$ with $w \in C_b$ (see figure). Note that w is well-defined by the Jordan-Brouwer separation theorem and b is well-defined because $\mathcal{A}$ is a maximal duplicate-free subset of the Pareto optimal actions. Using twice that $\langle \ell_a - \ell_b, w\rangle = 0$, we calculate



$$\langle \ell_a - \ell_b, u\rangle = \langle \ell_a - \ell_b, u - w\rangle = \frac{\|u-w\|_2}{\|v-w\|_2}\langle \ell_a - \ell_b, w - v\rangle = \frac{\|u-w\|_2}{\|v-w\|_2}\langle \ell_b - \ell_a, v\rangle > 0\,, \tag{8}$$

where the second equality used that $w \neq v$ is a point of the line segment connecting v and u, hence $w - v$ and $u - w$ are parallel and share the same direction and $\|v - w\|_2 > 0$. The last inequality follows because v is the centroid of C_a and a, b are distinct Pareto optimal actions. Let v_c be the centroid of C_c for any $c \in \mathcal{A}$. Then,

$$\begin{aligned}
\frac{\langle \ell_a - \ell_{\tilde{a}}, u\rangle}{\langle \ell_a - \ell_b, u\rangle} &= \frac{\langle \ell_a - \ell_{\tilde{a}}, w + u - w\rangle}{\langle \ell_a - \ell_b, u\rangle} \overset{\text{(a)}}{\le} \frac{\langle \ell_a - \ell_b, w\rangle + \langle \ell_a - \ell_{\tilde{a}}, u - w\rangle}{\langle \ell_a - \ell_b, u\rangle} \\
&\overset{\text{(b)}}{=} \frac{\langle \ell_a - \ell_{\tilde{a}}, u - w\rangle}{\langle \ell_a - \ell_b, u\rangle} \overset{\text{(c)}}{=} \frac{\|v-w\|_2\langle \ell_a - \ell_{\tilde{a}}, u - w\rangle}{\|u-w\|_2\langle \ell_b - \ell_a, v\rangle} \\
&\overset{\text{(d)}}{\le} \frac{\|v-w\|_2\|\ell_a - \ell_{\tilde{a}}\|_2}{\langle \ell_b - \ell_a, v\rangle} \overset{\text{(e)}}{\le} \frac{\sqrt{2E}}{\min_{c\in\mathcal{A}}\min_{d\in\mathcal{N}_c}\langle \ell_d - \ell_c, v_c\rangle} = \frac{1}{\varepsilon_G}\,,
\end{aligned}$$

where (a) follows since by (8), $\langle \ell_a - \ell_b, u\rangle > 0$ and also because $w \in C_b$ implies that $\langle \ell_a - \ell_{\tilde{a}}, w\rangle \le \langle \ell_a - \ell_b, w\rangle$, (b) follows since $\langle \ell_a - \ell_b, w\rangle = 0$ (which is used in other steps, too), (c) uses (8), (d) is by Cauchy-Schwartz and in (e) we bounded $\|w - v\|_2 \le \sqrt{2}$ and used that $\|\ell_a - \ell_{\tilde{a}}\|_2 \le \sqrt{E}$ and $\langle \ell_b - \ell_a, v\rangle = \langle \ell_b - \ell_a, v_a\rangle \ge \min_{c\in\mathcal{A}}\min_{d\in\mathcal{N}_c}\langle \ell_d - \ell_c, v_c\rangle > 0$. The final equality serves as the definition of $1/\varepsilon_G$. □

Lemma 5. *Let $\mathcal{H}$ be the set of functions $\phi : \mathcal{A} \to \mathcal{A}$ with $\phi(a) \in \mathcal{N}_a$ for all $a \in \mathcal{A}$ and define $a_n^* = \arg\min_{a\in[K]}\sum_{t=1}^n \langle \ell_a, u_t\rangle$. Then, for any $(B_t)_{1\le t\le n}$ sequence of actions in $\mathcal{A}$,*

$$\sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle \le \frac{1}{\varepsilon_G}\max_{\phi\in\mathcal{H}}\sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\,.$$

Proof. With no loss of generality we assume that $a_n^* \in \mathcal{A}$ because $\mathcal{A}$ is a maximal duplicate-free subset of Pareto optimal actions. Apply the previous lemma on subsequences of rounds where $B_t = a$ for each $a \in \mathcal{A}$. □

Lemma 6. *Let $\delta \in (0,1)$. Then with probability at least $1 - 2\delta$ it holds that*

$$R_n \le \gamma n + \frac{1}{\varepsilon_G}\sum_{k\in\mathcal{A}}\max_{b\in\mathcal{N}_k\cap\mathcal{A}}\sum_{t=1}^n \tilde{P}_{tk}\sum_{a\in\mathcal{A}} Q_{tka}\,(y_{ta} - y_{tb}) + \sqrt{8n\log(|\mathcal{H}|/\delta)}\,.$$

Proof. For $t \in [n]$, let $B_t \sim \tilde{P}_t$. Define the surrogate regret $R'_n = \sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle$. By the definition of A_t and B_t and part (b) of Lemma 4 we have $\mathbb{E}_{t-1}[\langle \ell_{A_t} - \ell_{B_t}, u_t\rangle] \le \gamma$. Furthermore, $|\langle \ell_a - \ell_b, u_t\rangle| \le 1$ for all a, b. Therefore, by Hoeffding-Azuma, with probability at least $1 - \delta$,

$$R_n \le R'_n + \gamma n + \sqrt{2n\log(1/\delta)}\,. \tag{9}$$

By Lemma 5, the surrogate regret is bounded in terms of the local regret:

$$R'_n = \sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle \le \frac{1}{\varepsilon_G}\max_{\phi\in\mathcal{H}}\sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\,. \tag{10}$$

We prepare to use Hoeffding-Azuma again. Fix $\phi \in \mathcal{H}$ arbitrarily. Then,

$$\begin{aligned}\mathbb{E}_{t-1}\left[\langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\right] &= \sum_{k\in\mathcal{A}} \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka} \langle \ell_a - \ell_{\phi(k)}, u_t\rangle \\ &= \sum_{k\in\mathcal{A}} \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka} (y_{ta} - y_{t\phi(k)})\,,\end{aligned} \tag{11}$$

where we used the fact that $\tilde{P}_{ta} = \sum_k \tilde{P}_{tk} Q_{tka}$. Hoeffding-Azuma's inequality now shows that with probability at least $1 - \delta/|\mathcal{H}|$,

$$\sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle \le \sum_{k\in\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)}) + \sqrt{2n\log(|\mathcal{H}|/\delta)}\,.$$

The result is completed via a union bound over all $\phi \in \mathcal{H}$ and chaining with Eqs. (9) and (10), and noting that

$$\begin{aligned}\max_\phi \sum_{k\in\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)}) &\le \sum_{k\in\mathcal{A}} \max_\phi \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)}) \\ &= \sum_{k\in\mathcal{A}} \underbrace{\max_{b\in\mathcal{N}_k\cap\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{tb})}_{R_{nk}}\,. \qquad \square\end{aligned}$$

Proof of Theorem 2. The proof has two steps. First bounding the local regret R_{nk} for each $k \in \mathcal{A}$ and then merging the bounds using the previous lemma.

Step 1: Bounding the local regret For the remainder of this step we fix $k \in \mathcal{A}$ and bound the local regret R_{nk}. First, we need to massage the local regret into a form in which we can apply Theorem 6, which is a generic version of the Exp3.P analysis by Auer et al. [2002]. Let $Z_{tka} = \tilde{P}_{tk}(y_{ta} - y_{tk})$ and $\mathcal{G}_t$ be the σ-algebra generated by $(A_1, \dots, A_t)$ and $\mathcal{G} = (\mathcal{G}_t)_{t=0}^n$ be the associated filtration. A simple rewriting shows that

$$R_{nk} = \max_{b\in\mathcal{N}_k\cap\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}\,(y_{ta} - y_{tb}) = \max_{b\in\mathcal{N}_k\cap\mathcal{A}} \sum_{t=1}^n \sum_{a\in\mathcal{A}} Q_{tka}\,(Z_{tka} - Z_{tkb})\,.$$

In order to apply the result in Theorem 6 we need to check the conditions. Since $(P_t)_t$ and $(\tilde{P}_t)_t$ are $\mathcal{G}$-predictable it follows that $(\beta_t)_t$ and $(Z_t)_t$ are also $\mathcal{G}$-predictable. Similarly, $(\hat{Z}_t)_t$ is $\mathcal{G}$-adapted because $(A_t)_t$ and $(\Phi_t)_t$ are $\mathcal{G}$-adapted. It remains to show that assumptions *(a–d)* are satisfied. For *(a)* let $a \in \mathcal{N}_k \cap \mathcal{A}$. By Lemma 4.(d) we have $P_{tb} \ge \gamma/K$ for all t and $b \in [K]$. Furthermore, $|v^{ak}(A_t, \Phi_t)| \le V$ so that $\eta|\hat{Z}_{tka}| = |\eta \tilde{P}_{tk} v^{ak}(A_t, \Phi_t)/P_{tA_t}| \le \eta V K/\gamma = 1$, where the equality follows from the choice of γ. Assumption *(b)* is satisfied in a similar way with $\eta\beta_{tka} = \eta^2 V^2 \sum_{b\in\mathcal{N}_{ak}} \tilde{P}_{tk}^2/P_{tb} \le \eta^2 K^2 V^2/\gamma = \eta K V \le 1$, where in the last inequality we used the definition of η and assumed that $n \ge K\log(K/\delta)$. To make sure that the regret bound holds even for smaller values of n, we require $C_G \ge K\sqrt{\log(eK)}$ so that when $n < K^2\log(K/\delta)$, the regret bound is trivial. For assumption *(c)*, we have

$$\mathbb{E}_{t-1}[\hat{Z}_{tka}^2] = \mathbb{E}_{t-1}\left[\left(\frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}}\right)^2\right] \le V^2 \tilde{P}_{tk}^2 \mathbb{E}_{t-1}\left[\frac{\mathbb{1}_{\mathcal{N}_{ak}}(A_t)}{P_{tA_t}^2}\right] = V^2 \sum_{b\in\mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}} = \frac{\beta_{tka}}{\eta}\,.$$

Finally *(d)* is satisfied by the definition of v^{ak} and the fact that $P_t \in \mathrm{ri}(\mathcal{P}_{K-1})$. The result of Theorem 6 shows that with probability at least $1 - (K+1)\delta$,

$$R_{nk} \le \frac{3\log(1/\delta)}{\eta} + 5\sum_{t=1}^n \sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\beta_{tka} + \eta \sum_{t=1}^n \sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\hat{Z}_{tka}^2\,.$$

Step 2: Aggregating the local regret Using the result from the previous step in combination with a union bound over $k \in \mathcal{A}$ we have that with probability at least $1 - K(K+1)\delta$,

$$\sum_{k \in \mathcal{A}} R_{nk} \leq \frac{3K \log(1/\delta)}{\eta} + 5 \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} + \eta \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_a \cap \mathcal{A}} Q_{tka} \hat{Z}_{tka}^2 \,. \tag{12}$$

For bounding the second term we use the definition of β_{tka} from (2) and write

$$\sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} = \eta V^2 \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}} = \eta V^2 \tilde{P}_{tk} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}}{P_{tb}} \,.$$

The sum over $b \in \mathcal{N}_{ak}$ is split into two, separating duplicates of k and the rest:

$$\begin{aligned}
\sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}}{P_{tb}} &= \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b : \ell_b = \ell_k} \frac{\tilde{P}_{tk}}{P_{tb}} + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak} : \ell_b \neq \ell_k} \frac{\tilde{P}_{tk}}{P_{tb}} \\
&= \sum_{b : \ell_b = \ell_k} \frac{\tilde{P}_{tk}}{P_{tb}} + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \sum_{b \in \mathcal{N}_{ak} : \ell_b \neq \ell_k} \frac{Q_{tka} \tilde{P}_{tk}}{P_{tb}} \\
&\leq 4K \left(\sum_{b : \ell_b = \ell_k} 1 + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \sum_{b \in \mathcal{N}_{ak} : \ell_b \neq \ell_k} 1 \right) \leq 4K^2 \,,
\end{aligned}$$

where the first equality used that $\sum_a Q_{tka} = 1$, the second to last inequality follows using parts (c) and (e) of Lemma 4, and the last inequality uses the reasoning above. Summing over all rounds and $k \in \mathcal{A}$ yields

$$5 \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} \leq 20 \eta n K^2 V^2 \,.$$

For the last term in Eq. (12) we use the definition of $\hat{Z}_{tka}$ and Lemma 4.(c) to show that

$$\begin{aligned}
\eta \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \hat{Z}_{tka}^2 &= \eta \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \frac{Q_{tka} \tilde{P}_{tk}^2 v^{ak}(A_t, \Phi_t)^2}{P_{tA_t}^2} \\
&\leq \eta V^2 \sum_{t=1}^{n} \frac{1}{P_{tA_t}} \sum_{k \in \mathcal{A}} \tilde{P}_{tk} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \frac{Q_{tka} \tilde{P}_{tk} \mathbb{1}_{\mathcal{N}_{ak}}(A_t)}{P_{tA_t}} \\
&\leq 4 \eta K V^2 \sum_{t=1}^{n} \frac{1}{P_{tA_t}} \,.
\end{aligned}$$

Now, from Lemma 4 (d), $\gamma / K \, (1/P_{ta}) \leq 1$ for all a, and in particular, holds for $a = A_t$. Furthermore, $\mathbb{E}_{t-1}[1/P_{tA_t}] = K$ and $\mathbb{E}_{t-1}[1/P_{tA_t}^2] = \sum_a 1/P_{ta} \leq K^2/\gamma$. By the result in Lemma 10 it holds that with probability at least $1 - \delta$ that

$$\sum_{t=1}^{n} \frac{1}{P_{tA_t}} \leq 2nK + \frac{K \log(1/\delta)}{\gamma} \,.$$

Another union bound shows that with probability at least $1 - (1 + K(K+1))\delta$,

$$\sum_{k \in \mathcal{A}} R_{nk} \leq \frac{3K \log(1/\delta)}{\eta} + 28 \eta n V^2 K^2 + 4 V K \log(1/\delta) \,.$$

The result follows from the definition of η, Lemma 6 and the definition of R_{nk}. □

C Proof of Theorem 3

Before the proof we need a simple lemma showing that if actions a and b are not weak neighbours, then the regret of either a or b grows linearly in t.

Lemma 7. *There exists a game-dependent constant $\varepsilon_G > 0$ such that:*

(a) If a and b are not weak neighbours, then $\inf_{u \in \mathcal{P}_{E-1}} \langle \ell_a + \ell_b - 2\ell_{a_u^*}, u \rangle \geq \varepsilon_G$.

(b) If $u \in \mathcal{P}_{E-1}$ and $M_u = \{a : \langle \ell_a - \ell_{a_u^*}, u \rangle < \varepsilon_G\}$, then $|M_u| \leq K_{loc}$.

Proof. For (a) let $c \in [K]$ be arbitrary. Since $C_a \cap C_b = \emptyset$ it follows that $\langle \ell_a + \ell_b - 2\ell_c, u \rangle > 0$ for all $u \in C_c$. By compactness of C_c and the continuity of the inner product in u we conclude that $\inf_{u \in C_c} \langle \ell_a + \ell_b - 2\ell_c, u \rangle > 0$. Taking the minimum over all c shows that $2\varepsilon_G = \inf_{u \in \mathcal{P}_{E-1}} \langle \ell_a + \ell_b - 2\ell_{a_u^*}, u \rangle > 0$. For (b), let $a, b \in M_u$. Then $\langle \ell_a + \ell_b - 2\ell_{a_u^*}, u \rangle < 2\varepsilon_G$, which by (a) means that a and b are weak neighbours. Therefore all actions in M_u are weak neighbours of each other so $|M_u| \leq K_{\text{loc}}$. □

The next lemma uses the concentration of the loss estimators to show that with high probability the distribution $\tilde{P}_t$ calculated by RELEXP3 assigns negligible probability to actions that are either not neighbours of B_t or for which the loss is large relative to the optimal action.

Lemma 8. *Let $Z_{ta} = \langle \ell_a - \ell_{B_t}, u_t \rangle$ and $L_{ta} = \sum_{s=1}^t Z_{sa}$. Then there exists an event* FAIL *with $\mathbb{P}(\text{FAIL}) \leq 1/n$ and function $g : \mathbb{N} \to [0, \infty)$ such that if* FAIL *does not hold, then*

(a) $\tilde{P}_{ta} \leq \exp(-\eta_t g(t))$ for all a that are not neighbours of B_t.

(b) $\tilde{P}_{ta} \leq \exp(-\eta_t g(t))$ for all a with $\langle \ell_a - \ell_{a_t^*}, \bar{u}_t \rangle \geq \varepsilon_G$.

(c) There exist constants $c_1, c_2 \geq 0$ depending on $G = (\mathcal{L}, \Phi)$ and the choice of ε in the definition of α_t such that for all $t \geq c_1 \log^{c_2}(n)$ it holds that $g(t) \geq \frac{1}{2}\varepsilon_G t$.

Proof. Define random variable $\phi_t = \max_{a,b} |\sum_{s=1}^t (\hat{Z}_{sb} - Z_{sb} + Z_{sa} - \hat{Z}_{sa})|$. Given an arbitrary pair of arms (a, b), from the triangle inequality we have

$$\left|\hat{L}_{ta} - \hat{L}_{tb}\right| = \left|\sum_{s=1}^t \left(\hat{Z}_{sa} - \hat{Z}_{sb}\right)\right| \geq \left|\sum_{s=1}^t (Z_{sa} - Z_{sb})\right| - \left|\sum_{s=1}^t \left(\hat{Z}_{sa} - Z_{sa} + Z_{sb} - \hat{Z}_{sb}\right)\right|$$
$$= |L_{ta} - L_{tb}| - \left|\sum_{s=1}^t \left(\hat{Z}_{sa} - Z_{sa} + Z_{sb} - \hat{Z}_{sb}\right)\right| \geq |L_{ta} - L_{tb}| - \phi_t \,.$$

The quantity ϕ_t is bounded with high probability via a union bound over all pairs of arms and a martingale version of Bernstein's bound [Freedman, 1975], which shows there exists a game-dependent constant $C_G > 0$ such that

$$\mathbb{P}\Big(\underbrace{\text{exists } t \in [n] : \phi_t \geq C_G t^{3/4+\varepsilon/2} \log^{\frac{1}{2}}(n)}_{\text{FAIL}}\Big) \leq \frac{1}{n} \,.$$

Choose $g(t) = \max\{0, (t-1)\varepsilon_G - C_G t^{3/4+\varepsilon/2} \log^{\frac{1}{2}}(n)\}$, which clearly satisfies the condition in *(c)*. First suppose that a is not a weak neighbour of B_{t+1}, which by the definition of B_{t+1}, ϕ_t and Lemma 7 ensures that

$$\hat{L}_{ta} - \hat{L}_{tB_{t+1}} \geq \hat{L}_{ta} + \hat{L}_{tB_{t+1}} - 2\hat{L}_{ta_t^*} \geq L_{ta} + L_{tB_{t+1}} - 2L_{ta_t^*} - 2\phi_t \geq t\varepsilon_G - 2\phi_t \,.$$

On the other hand if $\langle \ell_a - \ell_{a_t^*}, \bar{u}_t \rangle \geq \varepsilon_G$, then

$$\hat{L}_{ta} - \hat{L}_{tB_{t+1}} \geq \hat{L}_{ta} - \hat{L}_{ta_t^*} \geq L_{ta} - L_{ta_t^*} - 2\phi_t \geq t\varepsilon_G - 2\phi_t \,.$$

The result follows from the fact that $\tilde{P}_{ta} \leq \exp(\eta_t(\hat{L}_{t-1,B_t} - \hat{L}_{t-1,a}))$ for any action. □

Proof of Theorem 3. Choose $\varepsilon \in (0, 1/2)$ and

$$\eta_t = \min\left\{\frac{1}{4KV}, \frac{1}{2V_{\text{loc}}}\sqrt{\frac{\log(K)}{2tK_{\text{loc}}}}\right\} \qquad \text{and} \qquad \alpha_t = \min\left\{\frac{1}{4K}, t^{-1/2-\varepsilon}\right\} .$$

First note that the choices of η_t and α_t ensures that $\|\gamma_t\|_1 \leq 1/2$ and so P_t is indeed a probability distribution and $P_{ta} \geq \tilde{P}_{ta}/2$ for all t and a. Let N_t be the set of weak neighbours of B_t. Since $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$ we have for $p = e_{a_n^*}$ that

$$\begin{aligned}
\mathbb{E}[R_n] &= \mathbb{E}\left[\sum_{t=1}^n Z_{tA_t} - Z_{ta_n^*}\right] = \mathbb{E}\left[\sum_{t=1}^n \langle \tilde{P}_t - p, Z_t\rangle\right] + \mathbb{E}\left[\sum_{t=1}^n \langle \gamma_t - \|\gamma_t\|_1 \tilde{P}_t, Z_t\rangle\right] \\
&\leq \mathbb{E}\underbrace{\left[\sum_{t=1}^n \langle \tilde{P}_t - p, \hat{Z}_t\rangle\right]}_{\tilde{R}_n} + 2\mathbb{E}\left[\sum_{t=1}^n \|\gamma_t\|_1\right],
\end{aligned}$$

where in the inequality we used Cauchy-Schwartz and the fact that $\|Z_t\|_\infty \leq 1$. The analysis of exponential weights given in Theorem 2.3 of the book by Cesa-Bianchi and Lugosi [2006] yields

$$\mathbb{E}[\tilde{R}_n] \leq \log(K)\left(\frac{2}{\eta_n} + \frac{1}{\eta_1}\right) + \sum_{t=1}^n \mathbb{E}\underbrace{\left[\sum_{a=1}^K \tilde{P}_{ta}\hat{Z}_{ta} + \frac{1}{\eta_t}\log\left(\sum_{a=1}^K \tilde{P}_{ta}\exp\left(-\eta_t \hat{Z}_{ta}\right)\right)\right]}_{(\mathrm{A})_t}.$$

Note that we have stopped the proof before the application of Hoeffding's lemma, which is not appropriate for bandits due to the large range of the loss estimates. Suppose that $a \in M_t$, then for any $b \in S^{aB_t} \subseteq S_t$ we have $P_{tb} \geq \gamma_{tb} \geq \eta_t V^{aB_t}$, which means that

$$\eta_t |\hat{Z}_{ta}| = \frac{\eta_t v^{aB_t}(A_t, \Phi_t)}{P_{tA_t}} \leq \frac{\eta_t V^{aB_t} \mathbb{1}_{S^{aB_t}}(A_t)}{P_{tA_t}} \leq 1.$$

Then using $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ leads to

$$\tilde{P}_{ta}\exp\left(-\eta_t \hat{Z}_{ta}\right) \leq \tilde{P}_{ta} - \eta_t \tilde{P}_{ta}\hat{Z}_{ta} + \eta_t^2 \tilde{P}_{ta}\hat{Z}_{ta}^2.$$

On the other hand, if $a \notin M_t$ then by the definitions of M_t, $\hat{Z}_{ta}$ and P_t,

$$\tilde{P}_{ta}\exp\left(-\eta_t \hat{Z}_{ta}\right) \leq \tilde{P}_{ta}\exp\left(\eta_t |\hat{Z}_{ta}|\right) \leq \tilde{P}_{ta}\exp\left(\frac{\eta_t V^{aB_t}}{\alpha_t}\right) \leq \frac{\eta_t}{t},$$

which by the fact that $x \leq 1 + x \leq \exp(x)$ for all x also implies that $\tilde{P}_{ta}\hat{Z}_{ta} \leq 1/t$. Using $\log(1 + x) \leq x$,

$$\begin{aligned}
(\mathrm{A})_t &= \sum_{a=1}^K \tilde{P}_{ta}\hat{Z}_{ta} + \frac{1}{\eta_t}\log\left(\sum_{a=1}^K \tilde{P}_{ta}\exp\left(-\eta_t \hat{Z}_{ta}\right)\right) \\
&\leq \sum_{a=1}^K \tilde{P}_{ta}\hat{Z}_{ta} + \frac{1}{\eta_t}\log\left(\frac{\eta_t K}{t} + 1 - \eta_t \sum_{a \in M_t} \tilde{P}_{ta}\hat{Z}_{ta} + \eta_t^2 \sum_{a \in M_t} \tilde{P}_{ta}\hat{Z}_{ta}^2\right) \\
&\leq \frac{2K}{t} + \eta_t \sum_{a \in M_t} \tilde{P}_{ta}\hat{Z}_{ta}^2,
\end{aligned}$$

Next we bound the conditional second moment of $\hat{Z}_{ta}$. If a and B_t are weak neighbours, then

$$\begin{aligned}
\tilde{P}_{ta}\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] &= \tilde{P}_{ta}\mathbb{E}_{t-1}\left[\frac{v^{aB_t}(A_t, \Phi_t)^2}{P_{tA_t}^2}\right] \\
&\leq \tilde{P}_{ta}V_{\mathrm{loc}}^2\mathbb{E}_{t-1}\left[\frac{\mathbb{1}\{A_t \in \{a, B_t\}\}}{P_{tA_t}^2}\right] \leq 2V_{\mathrm{loc}}^2 + o(1),
\end{aligned}$$

where in the last line we used the fact that $P_{tA_t} \geq P_{ta}$ whenever $A_t \in \{a, B_t\}$ and $\mathbb{E}_{t-1}[\mathbb{1}\{A_t \in \{a, B_t\}\}] = 2$ and $P_{tA_t} \geq (1 - \|\gamma_t\|_1)\tilde{P}_{tA_t} = \tilde{P}_{tA_t}(1 - o(1))$. On the other hand, if

a and B_t are not weak neighbours, then $\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq K^2V^2/\alpha_t$ and so

$$\sum_{t=1}^{n} \mathbb{E}\left[(\mathrm{A})_t\right] \leq \sum_{t=1}^{n} \frac{2K}{t} + \sum_{t=1}^{n} \eta_t \mathbb{E}\left[\sum_{a \in M_t \cap N_t} \tilde{P}_{ta} \hat{Z}_{ta}^2\right] + \sum_{t=1}^{n} \eta_t \mathbb{E}\left[\sum_{a \in M_t \cap N_t^c} \tilde{P}_{ta} \hat{Z}_{ta}^2\right]$$
$$\leq 2V_{\mathrm{loc}}^2 \sum_{t=1}^{n} \eta_t \mathbb{E}\left[|M_t|\right] + K^2V^2 \sum_{t=1}^{n} \frac{\eta_t}{\alpha_t} \mathbb{E}\left[\sum_{a \in N_t^c} \tilde{P}_{ta}\right] + o(\sqrt{n})\,. \tag{13}$$

The second sum is bounded using parts *(a)* and *(c)* of Lemma 8, which shows that

$$\sum_{t=1}^{n} \frac{\eta_t}{\alpha_t} \mathbb{E}\left[\sum_{a \in N_t^c} \tilde{P}_{ta}\right] \leq \mathbb{P}\left(\textsc{Fail}\right) \sum_{t=1}^{n} \frac{\eta_t}{\alpha_t} + \sum_{t=1}^{n} \frac{\eta_t}{\alpha_t} \sum_{a \in N_t^c} \exp\left(-\eta_t g(t)\right) = o(\sqrt{n})\,. \tag{14}$$

Suppose that FAIL does not hold and define t_0 by

$$t_0 = \min\left\{t : \text{for all } s \geq t,\, \exp\left(\frac{\eta_s V}{\alpha_s} - \eta_s g(s)\right) \leq \frac{\eta_s}{s}\right\},$$

which by part *(c)* of Lemma 8 and rearrangement satisfies $t_0 = O(\mathrm{polylog}(n))$. The definition of t_0 ensures that if $t \geq t_0$ and a is an action with $\langle \ell_a - \ell_{a_{t-1}^*}, \bar{u}_{t-1} \rangle > \varepsilon_G$, then

$$\tilde{P}_{ta} \exp\left(\frac{\eta_t V^{aB_t}}{\alpha_t}\right) \leq \exp\left(\frac{\eta_t V^{aB_t}}{\alpha_t} - \eta_t g(t)\right) \leq \exp\left(\frac{\eta_t V}{\alpha_t} - \eta_t g(t)\right) \leq \frac{\eta_t}{t}$$

and so $a \notin M_t$. Therefore when FAIL does not hold and $t \geq t_0$,

$$M_t \subseteq \left\{a : \langle \ell_a - \ell_{a_{t-1}^*}, \bar{u}_{t-1} \rangle \leq \varepsilon_G\right\}.$$

But by Lemma 7 the number of arms in this set is at most K_{loc} and so in this case $|M_t| \leq K_{\mathrm{loc}}$. Since $|M_t| \leq K$ regardless of t or the failure event,

$$\sum_{t=1}^{n} \eta_t \mathbb{E}[|M_t|] \leq \mathbb{P}\left(\textsc{Fail}\right) \sum_{t=1}^{n} \eta_t K + \sum_{t=1}^{t_0} \eta_t K + K_{\mathrm{loc}} \sum_{t=t_0+1}^{n} \eta_t = K_{\mathrm{loc}} \sum_{t=1}^{n} \eta_t + o(\sqrt{n})\,.$$

Combining the above display with Eqs. (13) and (14) shows that

$$\sum_{t=1}^{n} \mathbb{E}[(\mathrm{A})_t] = 2V_{\mathrm{loc}}^2 K_{\mathrm{loc}} \sum_{t=1}^{n} \eta_t + o(\sqrt{n})\,.$$

Next we bound the sum of the expectations of $\|\gamma_t\|_1$. To begin notice that if M_t contains only neighbours of B_t, then $S_t = M_t$ and $\max_{a \in M_t} V^{aB_t} \leq V_{\mathrm{loc}}$. The definitions of γ_t and $\alpha_t = o(\sqrt{1/t})$ means that $\|\gamma_t\|_1 = \eta_t |S_t| \max_{a \in M_t} V^{aB_t} + o(\sqrt{1/t})$ and so the same argument as above shows that

$$2\mathbb{E}\left[\sum_{t=1}^{n} \|\gamma_t\|_1\right] = 2V_{\mathrm{loc}} K_{\mathrm{loc}} \sum_{t=1}^{n} \eta_t + o(\sqrt{n})\,.$$

Putting the pieces together and using the fact that $\sum_{t=1}^{n} \sqrt{1/t} \leq 2\sqrt{n}$,

$$\limsup_{n \to \infty} \frac{\mathbb{E}[R_n]}{\sqrt{n}} \leq \limsup_{n \to \infty} \frac{1}{\sqrt{n}} \left(\frac{2\log(K)}{\eta_n} + 2K_{\mathrm{loc}} V_{\mathrm{loc}}(V_{\mathrm{loc}} + 1) \sum_{t=1}^{n} \eta_t\right)$$
$$= 8\sqrt{2V_{\mathrm{loc}}(1 + V_{\mathrm{loc}}) K_{\mathrm{loc}} n \log(K)}\,.$$

The result is completed by recalling that $v^{ab}(\cdot)$ were chosen so that $V_{\mathrm{loc}} \leq 1 + F$. □

D Lower Bounds for Hard Games

In this section we prove a $\Omega(n^{2/3})$ lower bound on the minimax regret in hard partial monitoring games. Like for bandits, by Yao's minimax principle [Yao, 1977], the lower bounds are most easily proven using a stochastic adversary. In stochastic partial monitoring we assume that $u_1, \dots, u_n$ are chosen independently at random from the same distribution. To emphasise the randomness we switch to capital letters. Given a partial monitoring problem $G = (\mathcal{L}, \Phi)$ and a probability vector $u \in \mathcal{P}_{E-1}$ the stochastic partial monitoring environment associated with u samples a sequence of independently and identically distribution random variables $U_1, \dots, U_n$ with $U_t \in \{e_1, \dots, e_E\}$ with $\mathbb{P}(U_t = e_i) = u_i$. In each round t a policy chooses action A_t and receives feedback $\Phi_t = \Phi(A_t, U_t)$. The regret is

$$R_n(u, \pi, G) = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{n} \langle \ell_{A_t} - \ell_a, U_t \rangle\right] = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{n} \langle \ell_{A_t} - \ell_a, u \rangle\right] .$$

The mentioned minimax principle implies that $R_n^*(G) \geq \inf_\pi \sup_u R_n(u, \pi, G)$. Hence, in what follows, we lower bound $\sup_u R_n(u, \pi, G)$ for fixed π.

Given $u, v \in \mathcal{P}_{E-1}$, let $\mathrm{KL}(u, v)$ be the relative entropy between categorical distributions with parameters u and v respectively:

$$\mathrm{KL}(u, v) = \sum_{i=1}^{K} u_i \log\left(\frac{u_i}{v_i}\right) \leq \sum_{i=1}^{K} \frac{(u_i - v_i)^2}{v_i} , \tag{15}$$

where the second inequality follows from the fact that for measures $P \ll Q$ we have $\mathrm{KL}(P, Q) \leq \chi^2(P, Q)$. We need one more simple result that is frequently used in lower bound proofs. Given measures P and Q on the same probability space, Lemma 2.6 in the book by Tsybakov [2008] says that for any event A,

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-\mathrm{KL}(P, Q)\right) . \tag{16}$$

Theorem 4. *Let $G = (\mathcal{L}, \Phi)$ be a globally observable partial monitoring problem with that is not locally observable. Then there exists a constant $c_G > 0$ such that $R_n^*(G) \geq c_G n^{2/3}$.*

Proof. The proof involves several steps. Roughly, we need to define two alternative stochastic partial monitoring problems. We then show these environments are hard to distinguish without playing an action associated with a large loss. Finally we balance the cost of distinguishing the environments against the linear cost of playing randomly.

Fix a policy π and a partial monitoring game G with the required properties. For $u \in \mathcal{P}_{E-1}$ let $\mathbb{P}_u$ denote the measure on sequences of outcomes $(A_1, \Phi_1, \dots, A_n, \Phi_n)$ induced by the interaction of a fixed policy and the stochastic partial monitoring problem determined by u and G and denote by $\mathbb{E}_u$ the corresponding expectation. Note that $R_n(\pi, u, G) = \max_a \mathbb{E}_u\left[\sum_{t=1}^{n} \langle \ell_{A_t} - \ell_a, u \rangle\right]$.

Step 1: Defining the alternatives Let a, b be a pair neighbouring actions that are not locally observable. Then by definition $C_a \cap C_b$ is a polytope of dimension $E - 2$. Let u be the centroid of $C_a \cap C_b$ and

$$\varepsilon = \min_{c \notin \mathcal{N}_{ab}} \langle \ell_c - \ell_a, u \rangle . \tag{17}$$

The value of ε is well-defined, since by global observability of G, but nonlocal observability of (a, b) there must exist some action $c \notin \mathcal{N}_{ab}$. Furthermore, since $c \notin \mathcal{N}_{ab}$ it follows that $\varepsilon > 0$. We also have $u \in \mathrm{ri}(\mathcal{P}_{E-1})$. We now define two stochastic partial monitoring problems. Since (a, b) are not locally observable, there is no function $v : [K] \times [F] \to \mathbb{R}$ such that for all $i \in [E]$,

$$\sum_{c \in \mathcal{N}_{ab}} v(c, \Phi_{ci}) = \ell_{ai} - \ell_{bi} . \tag{18}$$

To facilitate the next step we rewrite this using a linear structure. For action $c \in [K]$ let $S_c \in \{0, 1\}^{F \times E}$ be the matrix with $(S_c)_{fi} = \mathbb{1}\{\Phi(c, i) = f\}$, which is chosen so that $S_c e_i = e_{\Phi_{ci}}$.

Define the linear map $S : \mathbb{R}^E \to \mathbb{R}^{|\mathcal{N}_{ab}|F}$ by

$$S = \begin{pmatrix} S_a \\ S_b \\ \vdots \\ S_c \end{pmatrix},$$

which is the matrix formed by stacking the matrices $\{S_c : c \in \mathcal{N}_{ab}\}$. Then there exists a v satisfying Eq. (18) if and only if there exists a vector $w \in \mathbb{R}^{|\mathcal{N}_{ab}|F}$ such that

$$\ell_a - \ell_b = w^\top S\,.$$

In other words, actions (a, b) are locally observable if and only if $\ell_a - \ell_b \in \operatorname{img}(S^\top)$. Since we have assumed that (a, b) are not locally observable, it follows that $\ell_a - \ell_b \notin \operatorname{img}(S^\top)$. Let $z \in \operatorname{img}(S^\top)$ and $w \in \ker(S)$ be such that $\ell_a - \ell_b = z + w$, which is possible since $\operatorname{img}(S^\top) \oplus \ker(S) = \mathbb{R}^E$. Since $\ell_a - \ell_b \notin \operatorname{img}(S^\top)$ it holds that $w \neq 0$ and $\langle w, \ell_a - \ell_b\rangle = \langle w, z + w\rangle = \langle w, w\rangle \neq 0$. Finally let $v = w/\langle w, \ell_a - \ell_b\rangle$. It follows that $Sv = 0$ and $\langle v, \ell_a - \ell_b\rangle = 1$.[2] Let $\Delta > 0$ be some small constant to be tuned subsequently and define $u_a = u - \Delta v$ and $u_b = u + \Delta v$ so that

$$\langle \ell_b - \ell_a, u_a\rangle = \Delta \qquad \text{and} \qquad \langle \ell_a - \ell_b, u_b\rangle = \Delta\,.$$

We note that if Δ is sufficiently small, then $u_a \in C_a \cap \operatorname{ri}(\mathcal{P}_{E-1})$ and $u_b \in C_b \cap \operatorname{ri}(\mathcal{P}_{E-1})$. This means that action a is optimal if the environment plays u_a on average and b is optimal if the environment plays u_b on average and that u_a and u_b are in the relative interior of the $(E-1)$-simplex (see Fig. 1).

Step 2: Calculating the relative entropy Given action c and $w \in \mathcal{P}_{E-1}$ let $\mathbb{P}_{cw}$ be the distribution on the feedback observed by the learner when playing action c in stochastic partial monitoring environment determined by w. That is $\mathbb{P}_{cw}(f) = \mathbb{P}_w(\Phi_t = f | A_t = c) = (S_c w)_f$. Let $T_c(n)$ be the number of times action c is played over all n rounds. The chain rule for relative entropy shows that

$$\operatorname{KL}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) = \sum_{c \in [K]} \mathbb{E}_{u_a}[T_c(n)] \operatorname{KL}(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b})\,. \tag{19}$$

By definition of u_a and u_b we have $S_c u_a = S_c u_b$ for all $c \in \mathcal{N}_{ab}$. Therefore $\mathbb{P}_{cu_a} = \mathbb{P}_{cu_b}$ and so $\operatorname{KL}(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) = 0$ for all $c \in \mathcal{N}_{ab}$. On the other hand, if $c \notin \mathcal{N}_{ab}$, then thanks to $u_a, u_b, u \in \operatorname{ri}(\mathcal{P}_{E-1})$ and Eq. (15),

$$\operatorname{KL}(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) \leq \operatorname{KL}(u_a, u_b) \leq \sum_{i=1}^{E} \frac{(u_{ai} - u_{bi})^2}{u_{bi}} = 4\Delta^2 \sum_{i=1}^{K} \frac{v_i^2}{u_i - \Delta v_i} \leq C_u \Delta^2\,,$$

where C_u is a suitably large constant and we assume that Δ is chosen sufficiently small that $u_i - \Delta v_i \geq u_i/2$. Therefore

$$\operatorname{KL}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) \leq C_u \mathbb{E}_{u_a}[\tilde{T}(n)] \Delta^2\,, \tag{20}$$

where $\tilde{T}(n)$ is the number of times an arm not in $\mathcal{N}_{ab}$ is played:

$$\tilde{T}(n) = \sum_{c \notin \mathcal{N}_{ab}} T_c(n)\,.$$

Step 3: Comparing the regret By Eq. (17) and the Cauchy-Schwartz inequality for $c \notin \mathcal{N}_{ab}$ we have $\langle \ell_c - \ell_a, u_a\rangle = \varepsilon + \langle \ell_c - \ell_a, \Delta v\rangle \geq \varepsilon - 2\Delta \|v\|_\infty$ and $\langle \ell_c - \ell_b, u_b\rangle \geq \varepsilon - 2\Delta \|v\|_\infty$. By Lemma 1, for each action $c \in \mathcal{N}_{ab}$ there exists an $\alpha \in [0, 1]$ such that $\ell_c = \alpha \ell_a + (1 - \alpha)\ell_b$. Therefore

$$\langle \ell_c - \ell_a, u_a\rangle + \langle \ell_c - \ell_b, u_b\rangle = (1 - \alpha)\langle \ell_b - \ell_a, u_a\rangle + \alpha\langle \ell_a - \ell_b, u_b\rangle = \Delta\,, \tag{21}$$

which means that $\max(\langle \ell_c - \ell_a, u_a\rangle, \langle \ell_c - \ell_b, u_b\rangle) \geq \Delta/2$. Define $\bar{T}(n)$ as the number of times some arm in $\mathcal{N}_{ab}$ is played that is at least $\Delta/2$ suboptimal in u_a:

$$\bar{T}(n) = \sum_{c \in \mathcal{N}_{ab}} \mathbb{1}\left\{\langle \ell_c - \ell_a, u_a\rangle \geq \frac{\Delta}{2}\right\} T_c(n)\,.$$

[2] The minor error in Bartók et al. [2014] appears in their definition of v, which is in the kernel of a *different* S constructed by stacking just S_a and S_b and not the degenerate/duplicate actions in between.

Assume that Δ is chosen sufficiently small so that $2\Delta \left\|v\right\|_\infty \leq \varepsilon/2$. Then

$$\begin{aligned}
R_n(\pi, u_a, G) + R_n(\pi, u_b, G) &= \mathbb{E}_{u_a}\left[\sum_{c\in[K]} T_c(n)\langle \ell_c - \ell_a, u_a\rangle\right] + \mathbb{E}_{u_b}\left[\sum_{c\in[K]} T_c(n)\langle \ell_c - \ell_b, u_b\rangle\right] \\
&\geq \frac{\varepsilon}{2}\mathbb{E}_{u_a}\left[\tilde{T}(n)\right] + \frac{n\Delta}{4}\left(\mathbb{P}_{u_a}(\bar{T}(n) \geq n/2) + \mathbb{P}_{u_b}(\bar{T}(n) < n/2)\right) \\
&\geq \frac{\varepsilon}{2}\mathbb{E}_{u_a}\left[\tilde{T}(n)\right] + \frac{n\Delta}{8}\exp\left(-\operatorname{KL}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b})\right) \\
&\geq \frac{\varepsilon}{2}\mathbb{E}_{u_a}\left[\tilde{T}(n)\right] + \frac{n\Delta}{8}\exp\left(-C_u\Delta^2\mathbb{E}_{u_a}\left[\tilde{T}(n)\right]\right) ,
\end{aligned}$$

In the above display we used (21) and

$$\sum_c T_c(n)\mathbb{1}\left\{\langle \ell_c - \ell_b, u_b\rangle > \Delta/2\right\} = \sum_c T_c(n)\mathbb{1}\left\{\langle \ell_c - \ell_a, u_a\rangle \leq \Delta/2\right\} = n - \bar{T}(n) ,$$

where the second inequality follows from the high probability version of Pinsker's inequality Eq. (16) and the third from Eq. (20). The bound is completed by a simple case analysis. If $\mathbb{E}_{u_a}[\tilde{T}(n)] > n^{2/3}$, the result holds for any value of Δ. Otherwise choosing $\Delta = (c/n)^{1/3}$ for appropriate positive game-dependent constant c establishes the bound. □

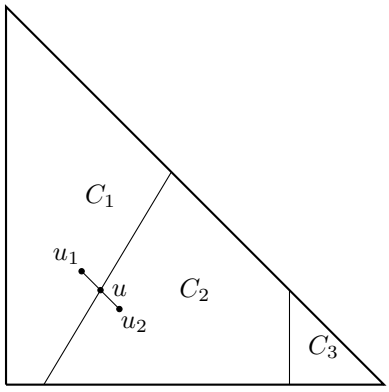


Figure 1: Lower bound construction for hard partial monitoring problems

E Lower Bound for Theorem 3

We consider the following game with $K = 2$ and $E = 2F - 2$ and $G = (\mathcal{L}, \Phi)$ given by

$$\mathcal{L} = \begin{pmatrix} 1 & 0 & 1 & 0 & \cdots & 1 & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 & 1 \end{pmatrix} , \quad \Phi = \begin{pmatrix} 1 & 2 & 2 & 3 & 3 & 4 & \cdots & F-1 & F-1 & F \\ 1 & 1 & 2 & 2 & 3 & 3 & \cdots & F-2 & F-1 & F-1 \end{pmatrix} .$$

Theorem 5. *For n sufficiently large the minimax regret of G is at least $R_n^*(G) \geq \frac{F-1}{45}\sqrt{n}$.*

Proof. Let $\Delta = \sqrt{1/(17n)}$ and $u \in \mathcal{P}_{E-1}$ be constants to be tuned subsequently and $u' \in \mathcal{P}_{E-1}$ by $u'_i = u_i + 2(-1)^i\Delta$. Using the notation of the previous section we have

$$\operatorname{KL}(\mathbb{P}_{1u}, \mathbb{P}_{1u'}) \leq \chi^2(\mathbb{P}_{1u}, \mathbb{P}_{2u'}) = 4\Delta^2\left(\frac{1}{u'_1} + \frac{1}{u'_E}\right) \qquad \text{and} \qquad \operatorname{KL}(P_{2u}, P_{2u'}) = 0 .$$

Let $p = 1/2 - (E-2)\Delta/4$ and choose $u_1 = p + \Delta$ and $u_E = p - \Delta$ and for $i \in \{2, \ldots, E-1\}$ let $u_i = \Delta(1-(-1)^i)$. For sufficiently large horizon, Δ is small enough so that $\operatorname{KL}(\mathbb{P}_{1u}, \mathbb{P}_{1u'}) \leq 17\Delta^2$ and

$$\operatorname{KL}(\mathbb{P}_u, \mathbb{P}_{u'}) = \mathbb{E}_u[T_1(n)]\operatorname{KL}(\mathbb{P}_{1u}, \mathbb{P}_{1u'}) \leq 17n\Delta^2 .$$

Using the fact that $R_n(u) = 2\Delta(F-1)\mathbb{E}_u[T_1(n)]$ and $R_n(u') = 2\Delta(F-1)\mathbb{E}_{u'}[T_2(n)]$ leads to

$$R_n(u) + R_n(u') \geq \frac{n(F-1)\Delta}{2}\exp\left(-n17\Delta^2\right) = \frac{(F-1)\sqrt{\frac{n}{17}}}{2e} .$$

The result follows since $\max(a, b) \geq (a+b)/2$ and by naive simplification. □

F Generic Bound for Exponential Weights

The proof of Theorem 2 depends on a generic regret analysis for a variant of the EXP3.P bandit algorithm by Auer et al. [1995]. The main difference is that loss estimators are assumed to be god-given and satisfy certain properties, rather than being explicitly defined as biased importance-weighted estimators. Nothing here would startle an expert, but we do not know where an equivalent result is written in the literature. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t=0}^n, \mathbb{P})$ be a filtered probability space and abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$. To reduce clutter we assume for the remainder that t ranges in $[n]$ and $a \in [K]$. Recall that a sequence of random elements (X_t) is called adapted if X_t is $\mathcal{F}_t$-measurable for all t, while (X_t) is called predictable if X_t is $\mathcal{F}_{t-1}$-measurable for all t. Let (Z_t) and $(\tilde{Z}_t)$ be sequences of random elements in $\mathbb{R}^K$. Given nonempty $\mathcal{A} \subseteq [K]$ and positive constant η define the probability vector $Q_t \in \mathcal{P}_{K-1}$ by

$$Q_{ta} = \frac{\mathbb{1}_{\mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{sa}\right)}{\sum_{b \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{sb}\right)} .$$

Theorem 6. *Assume that the $\mathbb{R}^K$-valued process $(Z_t)_t$ is predictable, the $\mathbb{R}^K$-valued process $(\tilde{Z}_t)_t$ is adapted and that $\tilde{Z}_t = \hat{Z}_t - \beta_t$, where $(\hat{Z}_t)_t$ is adapted and $(\beta_t)_t$ is predictable. Assume the following hold for all $a \in \mathcal{A}$:*

(a) $\eta|\hat{Z}_{ta}| \leq 1$,
(b) $\eta\beta_{ta} \leq 1$,
(c) $\eta\mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \beta_{ta}$ *almost surely*,
(d) $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$ *almost surely*.

Let $A^* = \arg\min_{a \in \mathcal{A}} \sum_{t=1}^n Z_{ta}$. Then, for any $0 \leq \delta \leq 1/(K+1)$, with probability at least $1 - (K+1)\delta$,

$$\sum_{t=1}^n \sum_{a=1}^K Q_{ta}(Z_{ta} - Z_{tA^*}) \leq \frac{3\log(1/\delta)}{\eta} + \eta \sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}\hat{Z}_{ta}^2 + 5\sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}\beta_{ta} .$$

Proof. We proceed in five steps.

Step 1: Decomposition

$$\begin{aligned}
&\sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}(Z_{ta} - Z_{tA^*}) \\
&\quad = \underbrace{\sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}(\tilde{Z}_{ta} - \tilde{Z}_{tA^*})}_{\text{(A)}} + \underbrace{\sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}(Z_{ta} - \tilde{Z}_{ta})}_{\text{(B)}} + \underbrace{\sum_{t=1}^n (\tilde{Z}_{tA^*} - Z_{tA^*})}_{\text{(C)}} .
\end{aligned}$$

Step 2: Bounding (A) By assumption *(c)* we have $\beta_{ta} \geq 0$, which by assumption *(a)* means that $\eta\tilde{Z}_{ta} \leq \eta\hat{Z}_{ta} \leq \eta|\hat{Z}_{ta}| \leq 1$ for all $a \in \mathcal{A}$. Then the standard mirror descent analysis with negentropy regularisation [Hazan, 2016] shows that (A) is bounded by

$$\begin{aligned}
\text{(A)} &\leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}\tilde{Z}_{ta}^2 \\
&= \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}(\hat{Z}_{ta}^2 + \beta_{ta}^2) - \frac{2\eta}{1-\gamma} \sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}\hat{Z}_{ta}\beta_{ta} \\
&\leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}\hat{Z}_{ta}^2 + 3\sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{ta}\beta_{ta} ,
\end{aligned}$$

where in the last two line we used the assumptions that $\eta\beta_{ta} \leq 1$ and $\eta|\hat{Z}_{ta}| \leq 1$.

Step 3: Bounding (B) For (B) we have

$$\text{(B)} = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} Q_{ta}(Z_{ta} - \tilde{Z}_{ta}) = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} Q_{ta}(Z_{ta} - \hat{Z}_{ta} + \beta_{ta}) \,.$$

We prepare to use Lemma 10. By assumptions *(c)* and *(d)* respectively we have $\eta \mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \beta_{ta}$ and $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$. By Jensen's inequality,

$$\eta \mathbb{E}_{t-1}\left[\left(\sum_{a \in \mathcal{A}} Q_{ta}(Z_{ta} - \hat{Z}_{ta})\right)^2\right] \leq \eta \sum_{a \in \mathcal{A}} Q_{ta} \mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \sum_{a \in \mathcal{A}} Q_{ta} \beta_{ta} \,.$$

Therefore by Lemma 10, with probability at least $1 - \delta$

$$\text{(B)} \leq 2 \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} Q_{ta} \beta_{ta} + \frac{\log(1/\delta)}{\eta} \,.$$

Step 4: Bounding (C) For (C) we have

$$\text{(C)} = \sum_{t=1}^{n} (\tilde{Z}_{tA^*} - Z_{tA^*}) = \sum_{t=1}^{n} \left(\hat{Z}_{tA^*} - Z_{tA^*} - \beta_{tA^*}\right) \,.$$

Because A^* is random we cannot directly apply Lemma 10, but need a union bound over all actions. Let $a \in \mathcal{A}$ be fixed. Then by Lemma 10 and the assumption that $\eta |\hat{Z}_{ta}| \leq 1$ and $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$ and $\eta \mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \beta_{ta}$, with probability at least $1 - \delta$.

$$\sum_{t=1}^{n} \left(\hat{Z}_{ta} - Z_{ta} - \beta_{ta}\right) \leq \frac{\log(1/\delta)}{\eta} \,.$$

Therefore by a union bound we have with probability at most $1 - K\delta$,

$$\text{(C)} \leq \frac{\log(1/\delta)}{\eta} \,.$$

Step 5: Putting it together Combining the bounds on (A), (B) and (C) in the last three steps with the decomposition in the first step shows that with probability at least $1 - (K+1)\delta$,

$$R_n \leq \frac{3 \log(1/\delta)}{\eta} + \eta \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} Q_{ta} \hat{Z}_{ta}^2 + 5 \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} Q_{ta} \beta_{ta} \,.$$

where we used the assumption that $\delta \leq 1/K$. □

G Bounding the Norm of the Estimators

Lemma 9. *Let a and b be pairwise observable and $\Delta = \ell_a - \ell_b \in [-1, 1]^E$, then there exists a function $v : \{a, b\} \times [F] \to \mathbb{R}$ such that:*

(a) $\|v\|_\infty \leq 1 + F$.

(b) $v(a, \Phi_{ai}) + v(b, \Phi_{bi}) = \Delta_i$ *for all* $i \in [E]$.

Proof. By the definition of pairwise observable there exists a function $v : \{a, b\} \times [F] \to \mathbb{R}$ such that $v(a, \Phi_{ai}) + v(b, \Phi_{bi}) = \Delta_i$ for all $i \in [E]$. Define bipartite graph with vertices $\mathcal{V} = \{a, b\} \times [F]$ and edges $\mathcal{E} = \{((a, \Phi_{ai}), (b, \Phi_{bi})) : i \in [E]\}$. Assume without loss of generality this graph is fully connected. If not then apply the following procedure to each connected component. For any $f, f' \in [F]$ let $(a, f_1), (b, f_2), (a, f_3), \ldots, (a, f_k)$ be a loop free path with $f_1 = f$ and $f_k = f'$ and for $j \in [k-1]$ let $i_j \in [E]$ correspond to the edge connecting $(\cdot, f_j)$ and $(\cdot, f_{j+1})$. Then

$$|v(a, f) - v(a, f')| = \left|\sum_{j=1}^{k-1} (-1)^{j-1} \Delta_{i_j}\right| \leq 2F \,.$$

We may assume that $\max_{f \in [F]} |v(a, f)| \leq \frac{1}{2} \max_{f, f' \in [F]} |v(a, f) - v(a, f')|$, which is always possible by translating $v(a, \cdot)$ by a constant α and $v(b, \cdot)$ by $-\alpha$. Finally for each $f \in [F]$ there exists an $f' \in [F]$ and $i \in [E]$ such that $v(b, f) + v(a, f') = \Delta_i$, which ensures that $|v(b, f)| \leq |v(a, f')| + |\Delta_i| \leq F + 1$. □

H Concentration

Lemma 10. *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables adapted to filtration $(\mathcal{F}_t)_t$ and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$ and $\mu_t = \mathbb{E}_{t-1}[X_t]$. Suppose that $\eta > 0$ satisfies $\eta X_t \leq 1$ almost surely. Then*

$$\mathbb{P}\left(\sum_{t=1}^{n}(X_t - \mu_t) \geq \eta \sum_{t=1}^{n} \mathbb{E}_{t-1}[X_t^2] + \frac{1}{\eta}\log\left(\frac{1}{\delta}\right)\right) \leq \delta\,.$$

Proof. Let $\sigma_t^2 = \mathbb{E}_{t-1}[X_t^2]$. By Chernoff's method we have

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{n}\left(X_t - \mu_t - \eta\sigma_t^2\right) \geq \frac{\log(1/\delta)}{\eta}\right) &= \mathbb{P}\left(\exp\left(\eta\sum_{t=1}^{n}\left(X_t - \mu_t - \eta\sigma_t^2\right)\right) \geq \frac{1}{\delta}\right) \\ &\leq \delta\mathbb{E}\left[\exp\left(\eta\sum_{t=1}^{n}\left(X_t - \mu_t - \eta\sigma_t^2\right)\right)\right].\end{aligned}$$

The result is completed by showing the term inside the expectation is a supermartingale. For this, we have

$$\begin{aligned}\mathbb{E}_{t-1}\left[\exp\left(\eta\left(X_t - \mu_t - \eta\sigma_t^2\right)\right)\right] &= \exp\left(-\eta\mu_t - \eta^2\sigma_t^2\right)\mathbb{E}_{t-1}\left[\exp\left(\eta X_t\right)\right] \\ &\leq \exp\left(-\eta\mu_t - \eta^2\sigma_t^2\right)\left(1 + \eta\mu_t + \eta^2\sigma_t^2\right) \leq 1\,,\end{aligned}$$

where we used the inequalities $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$. Chaining the above inequality completes the proof. □

Proof of Lemma 3. Let Λ be the smallest value such that $X \leq Y + \sqrt{2a\log(b/\Lambda)}$, which is almost surely positive. Taking expectations of both sides shows that

$$\mathbb{E}[X] \leq \mathbb{E}[Y] + \sqrt{2a}\mathbb{E}[\sqrt{\log(b/\Lambda)}]$$

The second expectation is bounded by

$$\begin{aligned}\mathbb{E}[\sqrt{\log(b/\Lambda)}] &= \int_0^\infty \mathbb{P}\left(\sqrt{\log(b/\Lambda)} \geq x\right)dx \\ &\leq \inf_{y>0} y + \int_y^\infty \mathbb{P}\left(\sqrt{\log(b/\Lambda)} \geq x\right)dx \\ &\leq \inf_{y>0} y + \int_y^\infty \mathbb{P}\left(\Lambda \leq b\exp(-x^2)\right)dx \\ &\leq \inf_{y>0} y + \int_y^\infty b\exp(-x^2)dx \\ &= \inf_{y>0} y + \frac{b\sqrt{\pi}}{2}\operatorname{erfc}(y) \\ &= \sqrt{\log(b)} + \frac{b\sqrt{\pi}}{2}\operatorname{erfc}(\sqrt{\log(b)}) \\ &\leq \sqrt{1 + \log(b)}\,.\end{aligned}$$

□

I Gallery

Exhibit 1. In the following game the learner cannot estimate the actual losses, but the loss differences can be calculated from the feedback directly.

$$\mathcal{L} = \begin{pmatrix} 1 & 1/2 & 1/2 & 0 \\ 1/2 & 1 & 0 & 1/2 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}.$$

Exhibit 2. A useful way to think about the cell decomposition is to assume that $\mathcal{L}$ has positive entries and consider the intersection of the hypograph of concave function $f(u) = \min_a \langle \ell_a, u \rangle$ with domain $u = \mathcal{P}_{E-1}$ and the epigraph of $\mathcal{P}_{E-1}$. To illustrate the idea let $G = (\mathcal{L}, \Phi)$ be the variant of the spam game where $c = 1/3$, which is defined by

$$\mathcal{L} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

In this case $\mathcal{P}_{E-1} = \mathcal{P}_1$ is 1-dimensional, which means the intersection of the epigraph of $\mathcal{P}_{E-1}$ and the hypograph of f is 2-dimensional and is shown in the left figure below. The intersection is itself a polytope and the faces (1-dimensional in this case) pointing upwards correspond to cells of nondegenerate actions. If c is increased to $1/2$, then the third action becomes degenerate, which is observable from the right-hand figure below by noting that the dimension of its intersection with the polytope is now zero. Increasing c any further would make this action dominated.

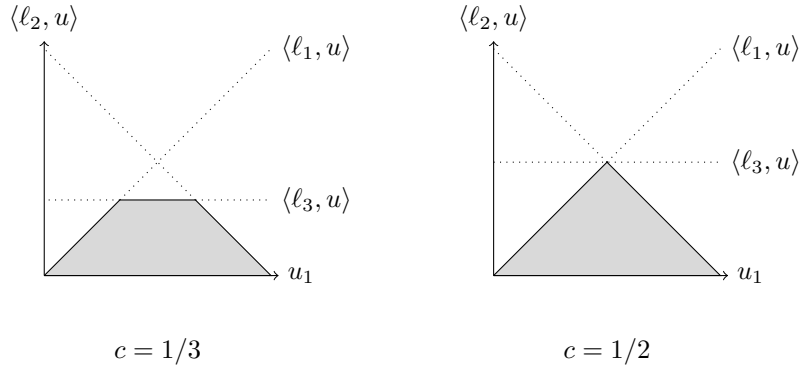
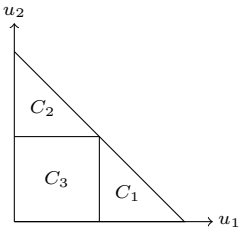


Figure 2: Alternative view of cell decomposition for spam game with $c = 1/3$ and $c = 1/2$.

Exhibit 3. The following game demonstrates that not all locally observable games are point-locally observable.

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1/2 & 1/2 & 1/2 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}.$$



The cell decomposition for this game is shown on above-right. Notice that $(1, 2)$ are not neighbours, but are weak neighbours. And yet $(1, 2)$ are not pairwise observable. Therefore the game is not point-locally observable. On the other hand, both sets of neighbours $(1, 2)$ and $(1, 3)$ are locally observable.

Exhibit 4. This game produces the cell decomposition depicted at the start of Section 3. The only neighbours are $(2, 3)$ and $(1, 3)$, which are locally observable. Therefore the game is locally observable. Actions 4,5 and 6 are degenerate. NEIGHBOURHOODWATCH2 will only play actions 1, 2, 3 and 4 with actions 5 and 6 ruled out because their cells are not equal to the intersection of any

neighbours cells. Notice that $\ell_4 = \ell_2/2 + \ell_3/2$ is a convex combination of ℓ_2 and ℓ_3.

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1/2 & 1/2 & 1/2 \\ 3/4 & 1/4 & 3/4 \\ 1 & 1/2 & 1/2 \\ 1 & 1/4 & 3/4 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$