Partial monitoring – classification, regret bounds, and algorithms*

Gábor Bartók
Department of Computing Science
University of Alberta

Dean Foster
Department of Statistics
University of Pennsylvania

Dávid Pál
Department of Computing Science
University of Alberta

Alexander Rakhlin
Department of Statistics
University of Pennsylvania

Csaba Szepesvári
Department of Computing Science
University of Alberta

March 27, 2013

Abstract

In a partial monitoring game, the learner repeatedly chooses an action, the environment responds with an outcome, and then the learner suffers a loss and receives a feedback signal, both of which are fixed functions of the action and the outcome. The goal of the learner is to minimize his regret, which is the difference between his total cumulative loss and the total loss of the best fixed action in hindsight. In this paper we characterize the minimax regret of any partial monitoring game with finitely many actions and outcomes. It turns out that the minimax regret of any such game is either zero, $\widetilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$, or $\Theta(T)$. We provide computationally efficient learning algorithms that achieve the minimax regret within logarithmic factor for any game. In addition to the bounds on the minimax regret, if we assume that the outcomes are generated in an i.i.d. fashion, we prove individual upper bounds on the expected regret.

1 Introduction

Partial monitoring provides a mathematical framework for sequential decision making problems with imperfect feedback. Various problems of interest can be modeled as partial monitoring instances, such as learning with expert advice [Littlestone and Warmuth, 1994], the multi-armed bandit problem [Auer et al., 2002], dynamic pricing [Kleinberg and Leighton, 2003], the dark pool problem [Agarwal et al., 2010], label efficient prediction [Cesa-Bianchi et al., 2005], and linear and convex optimization with full or bandit feedback [Zinkevich, 2003, Abernethy et al., 2008, Flaxman et al., 2005].

In this paper we restrict ourselves to finite games, *i.e.*, games where both the set of actions available to the learner and the set of possible outcomes generated by the environment are finite. A finite partial monitoring game G is described by a pair of $N \times M$ matrices: the *loss matrix* L and the *feedback matrix* H. The entries $L_{i,j}$ of L are real numbers lying in, say, the interval $[0, 1]$. The entries $H_{i,j}$ of H belong to an alphabet Σ on which we do not impose any structure and we only assume that learner is able to distinguish distinct elements of the alphabet.

The game proceeds in T rounds according to the following protocol. First, $G = (L, H)$ is announced for both players. In each round $t = 1, 2, \ldots, T$, the learner chooses an action $I_t \in \{1, 2, \ldots, N\}$ and simultaneously, the environment, or *opponent*, chooses an outcome $J_t \in \{1, 2, \ldots, M\}$. Then, the learner receives

[*] This article is an extended version of Bartók, Pál, and Szepesvári [2011], Bartók, Zolghadr, and Szepesvári [2012], and Foster and Rakhlin [2012].

as a feedback the entry H_{I_t,J_t}. The learner incurs *instantaneous loss* L_{I_t,J_t}, which is *not revealed* to him. The feedback can be thought of as a masked information about the outcome J_t. In some cases H_{I_t,J_t} might uniquely determine the outcome, in other cases the feedback might give only partial or no information about the outcome.

The learner is scored according to the loss matrix L. In round t the learner incurs an *instantaneous loss* of L_{I_t,J_t}. The goal of the learner is to keep low his *total loss* $\sum_{t=1}^{T} L_{I_t,J_t}$. Equivalently, the learner's performance can also be measured in terms of his regret, *i.e.*, the total loss of the learner is compared with the loss of best fixed action in hindsight. Since no non-trivial bound can be given on the learner's total loss, we resort to regret analysis in which the total loss of the learner is compared with the loss of best fixed action in hindsight. The regret is defined as the difference of these two losses.

In general, the regret grows with the number of rounds T. If the regret is sublinear in T, the learner is said to be Hannan consistent, and this means that the learner's average per-round loss approaches the average per-round loss of the best action in hindsight.

Piccolboni and Schindelhauer [2001] were one of the first to study the regret of these games. They proved that for any finite game (L, H), either for any algorithm the regret can be $\Omega(T)$ in the worst case, or there exists an algorithm which has regret $\widetilde{O}(T^{3/4})$ on any outcome sequence[1]. This result was later improved by Cesa-Bianchi et al. [2006] who showed that the algorithm of Piccolboni and Schindelhauer has regret $O(T^{2/3})$. Furthermore, they provided an example of a finite game, a variant of label-efficient prediction, for which any algorithm has regret $\Theta(T^{2/3})$ in the worst case.

However, for many games $O(T^{2/3})$ is not optimal. For example, games with full feedback (*i.e.*, when the feedback uniquely determines the outcome) can be viewed as a special instance of the problem of learning with expert advice and in this case it is known that the "EWA forecaster" has regret $O(\sqrt{T})$; see *e.g.* Lugosi and Cesa-Bianchi [2006, Chapter 3]. Similarly, for games with "bandit feedback" (*i.e.*, when the feedback determines the instantaneous loss) the INF algorithm [Audibert and Bubeck, 2009] and the Exp3 algorithm [Auer et al., 2002] achieve $O(\sqrt{T})$ regret as well.[2]

This leaves open the problem of determining the minimax regret (*i.e.*, optimal worst-case regret) of any given game (L, H). A partial progress was made in this direction by Bartók et al. [2010] who characterized (almost) all finite games with $M = 2$ outcomes. They showed that the minimax regret of any "non-degenerate" finite game with two outcomes falls into one of four categories: zero, $\widetilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$ or $\Theta(T)$. They gave a combinatoric-geometric condition on the matrices L, H that determines the category a game belongs to. Additionally, they constructed an efficient algorithm that, for any game, achieves the minimax regret rate associated to the game within poly-logarithmic factor.

In this paper, we consider the general problem of classifying partial-monitoring games with any finite number of actions and outcomes. We investigate the problem under two different opponent models: the *oblivious adversarial* and the *stochastic* opponent. In the oblivious adversarial model, the outcomes are arbitrarily generated by an adversary with the constraint that they cannot depend on the actions chosen by the learner. Equivalently, an oblivious adversary can be thought of as an oracle that chooses a sequence of outcomes before the game begins. In the stochastic model, the outcomes are generated by a sequence of i.i.d. random variables.

In the stochastic model, an alternative definition of regret is used; instead of comparing the cumulative loss of the learner of that of the best fixed action in hindsight, the base of the comparison is the expected cumulative loss of the action with the smallest expected loss, given the distribution the outcomes are generated from. More formally, the regret of an algorithm $\mathcal{A}$ under outcome distribution p is defined as

$$R_T(\mathcal{A}, p) = \sum_{t=1}^{T} L_{I_t,J_t} - \min_{1\le i\le N} \mathbb{E}_p\left[\sum_{t=1}^{T} L_{i,J_t}\right] .$$

This paper is based on the results of Bartók, Pál, and Szepesvári [2011], Bartók, Zolghadr, and Szepesvári [2012], and Foster and Rakhlin [2012]. We summarize the results of these works to create a complete and

[1] The notations $\widetilde{O}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ hide polylogarithmic factors.

[2] We ignore the dependence of regret on the number of actions or any other parameters.

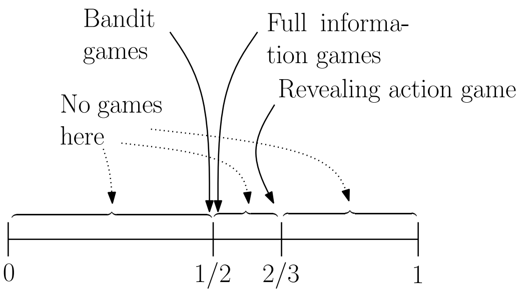


Figure 1: Diagram of the classification result. Points on the line segment represent the exponent on the time horizon T in the minimax regret of games. The gap between 0 and 1/2 was proven by Antos et al. [2012], while the gap between 2/3 and 1 was shown by Piccolboni and Schindelhauer [2001]. The minimax regret of the "revealing action game" was proven to be of $\Theta(T^{2/3})$ by Cesa-Bianchi et al. [2006]. The gap between 1/2 and 2/3 is the result of this work, completing the characterization.

self-contained reference for the recent advancements on finite partial monitoring. The results include a characterization of non-degenerate games against adversarial opponents, a full characterization of games as well as individual regret bounds against stochastic opponents.

The characterization result, in both cases, shows that there are only four classes of games in terms of the minimax regret:

- Trivial games with zero minimax regret,
- "Easy" games with $\widetilde{\Theta}(\sqrt{T})$ minimax regret,
- "Hard" games with $\Theta(T^{2/3})$ minimax regret, and
- Hopeless games with $\Omega(T)$ minimax regret.

A visualization of the classification is depicted in Figure 1.

2 Definitions and notations

Let $\underline{N}$ denote the set $\{1, \ldots, N\}$. For a subset $S \subset \underline{N}$ we use $1_S \in \{0,1\}^N$ to denote the vector with ones on the coordinates in S and zeros outside. A vector $a \in \mathbb{R}^N$ indexed by j is sometimes denoted by $[a_j]_{j\in[N]}$. Standard basis vectors are denoted by $\{e_i\}$.

Recall from the introduction that an instance of partial monitoring with N actions and M outcomes is defined by the pair of matrices $L \in \mathbb{R}^{N\times M}$ and $H \in \Sigma^{N\times M}$, where Σ is an arbitrary set of symbols. In each round t, the opponent chooses an outcome $j_t \in \underline{M}$ and simultaneously the learner chooses an action $i_t \in \underline{N}$. Then, the feedback H_{I_t,J_t} is revealed and the learner suffers the loss L_{i_t,j_t}. It is important to note that the loss is not revealed to the learner, whereas L and H are revealed before the game begins.

The following definitions are essential for understanding how the structure of L and H determines the "hardness" of a game. Let Δ_M denote the probability simplex in $\mathbb{R}^M$. That is, $\Delta_M = \{p \in \mathbb{R}^M \ : \ \forall 1 \le i \le M, p_i \ge 0, \sum_{i=1}^M p_i = 1\}$. Elements of Δ_M will also be called *opponent strategies* as $p \in \Delta_M$ represents an outcome distribution that a stochastic opponent can use to generate outcomes. Let ℓ_i denote the column vector consisting of the i^{th} row of L. Action i is called *optimal* under strategy p if its expected loss is not greater than that of any other action $i' \in \underline{N}$. That is, $\ell_i^\top p \le \ell_{i'}^\top p$. Determining which action is optimal under opponent strategies yields the *cell decomposition*[3] of the probability simplex Δ_M:

[3] The concept of cell decomposition also appears in Piccolboni and Schindelhauer [2001].

Definition 1 (Cell decomposition)**.** *For every action $i \in \underline{N}$, let $C_i = \{p \in \Delta_M \ : \ \text{action } i \text{ is optimal under } p\}$. The sets $C_1, \dots, C_N$ constitute the* cell decomposition *of Δ_M.*

Now we can define the following important properties of actions:

Definition 2 (Properties of actions)**.**

- *Action i is called* dominated *if $C_i = \emptyset$. If an action is not dominated then it is called* non-dominated.
- *Action i is called* degenerate *if it is non-dominated and there exists an action i' such that $C_i \subsetneq C_{i'}$.*
- *If an action is neither dominated nor degenerate then it is called* Pareto-optimal. *The set of Pareto-optimal actions is denoted by $\mathcal{P}$.*
- *Action i is called* duplicate *if there exists another action $j \neq i$ such that $\ell_i = \ell_j$.*

From the definition of cells we see that a cell is either empty or it is a closed polytope. Furthermore, Pareto-optimal actions have $(M-1)$-dimensional cells. The following definition, important for our analyses, also uses the dimensionality of polytopes:

Definition 3 (Neighbors)**.** *Two Pareto-optimal actions i and j are* neighbors *if $C_i \cap C_j$ is an $(M-2)$-dimensional polytope. Let $\mathcal{N}$ be the set of unordered pairs over $\underline{N}$ that contains neighboring action-pairs. The* neighborhood action set *of two neighboring actions i, j is defined as $N^+_{i,j} = \{k \in \underline{N} \ : \ C_i \cap C_j \subseteq C_k\}$.*

Note that the neighborhood action set $N^+_{i,j}$ naturally contains i and j. If $N^+_{i,j}$ contains some other action k then either $C_k = C_i$, $C_k = C_j$, or $C_k = C_i \cap C_j$.

Now we turn our attention to how the feedback matrix H is used. In general, the elements of the feedback matrix H can be arbitrary symbols. Nevertheless, the nature of the symbols themselves does not matter in terms of the structure of the game. What determines the feedback structure of a game is the occurrence of identical symbols in each row of H. To "standardize" the feedback structure, the *signal matrix* is defined for each action:

Definition 4. *Let s_i be the number of distinct symbols in the i^{th} row of H and let $\sigma_1, \dots, \sigma_{s_i} \in \Sigma$ be an enumeration of those symbols. Then the* signal matrix *$S_i \in \{0,1\}^{s_i \times M}$ of action i is defined as $(S_i)_{k,l} = \mathbb{I}\{H_{i,l} = \sigma_k\}$.*

Note that the signal matrix of action i is just the incidence matrix of symbols and outcomes, assuming action i is chosen. Furthermore, if $p \in \Delta_M$ is the opponent's strategy (or in the adversarial setting, the relative frequency of outcomes in time steps when action i is chosen), then $S_i p$ gives the distribution (or relative frequency) of the symbols underlying action i. In fact, it is also true that observing H_{I_t,J_t} is equivalent to observing the vector $S_{I_t} e_{J_t}$, where e_k is the k^{th} unit vector in the standard basis of $\mathbb{R}^M$. From now on we assume without loss of generality that the learner's observation at time step t is the random vector $Y_t = S_{I_t} e_{J_t}$. Note that the dimensionality of this vector depends on the action chosen by the learner, namely $Y_t \in \mathbb{R}^{s_{I_t}}$.

Let $\operatorname{Im} M$ denote the image space (or column space) of a matrix M. The following two definitions play a key role in classifying partial-monitoring games.

Definition 5 (Global observability [Piccolboni and Schindelhauer, 2001])**.** *A partial-monitoring game (L, H) admits the* global observability *condition, if for all pairs i, j of actions, $\ell_i - \ell_j \in \oplus_{k \in \underline{N}} \operatorname{Im} S_k^\top$.*

Definition 6 (Local observability)**.** *A pair of neighboring actions i, j is said to be* locally observable *if $\ell_i - \ell_j \in \oplus_{k \in N^+_{i,j}} \operatorname{Im} S_k^\top$. We denote by $\mathcal{L} \subset \mathcal{N}$ the set of locally observable pairs of actions (the pairs are unordered). A game satisfies the* local observability *condition if every pair of neighboring actions is locally observable,* i.e., *if $\mathcal{L} = \mathcal{N}$.*

The intuition behind these definitions is that if $\ell_i - \ell_j \in \oplus_{k\in D} \operatorname{Im} S_k^\top$ for some subset D of actions then the expected difference of the losses of actions i and j can be estimated with observations from actions in D. We will later see that the above condition necessary for haveing unbiased estimates for the loss differences.

It is easy to see that local observability implies global observability. Also, from Piccolboni and Schindelhauer [2001] we know that if global observability does not hold then the game has linear minimax regret. From now on, we only deal with games that admit the global observability condition.

2.1 Examples

To illustrate the concepts of global and local observability, we present some examples of partial-monitoring games.

Full-information games Consider a game $G = (L, H)$, where every row of the feedback matrix consists of pairwise different symbols. Without loss of generality we may assume that

$$H = \begin{pmatrix} 1 & 2 & \cdots & M \\ 1 & 2 & \cdots & M \\ \vdots & \vdots & & \vdots \\ 1 & 2 & \cdots & M \end{pmatrix} .$$

In this case the learner receives the outcome as feedback at the end of every time step, hence we call it the *full-information* case. It is easy to see that the signal matrix of any action i is the identity matrix of dimension M. Consequently, for any $\ell \in \mathbb{R}^M$, $\ell \in \operatorname{Im} S_i^\top$ and thus any full-information game is locally observable.

Bandit games The next games we consider are games $G = (L, H)$ with $L = H$. In this case the feedback the learner receives is identical to the loss he suffers at every time step. For this reason, we call these types of games *bandit* games.

For an action i, let the the rows of S_i correspond to the symbols $\sigma_1, \sigma_2, \ldots, \sigma_{s_i}$, where s_i is the number of different symbols in the i^{th} row of H. Since we assumed that $L = H$, we know that these symbols are real numbers (losses). It follows from the construction of the signal matrix that

$$\ell_i = S_i^\top \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_{s_i} \end{pmatrix}$$

for all $i \in \underline{N}$. It follows that all bandit games are locally observable.

A hopeless game We define the following game $G = (L, H)$ by

$$L = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}, \qquad H = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & * & * \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & * & * \end{pmatrix} .$$

We make the following observations:

1. Neither of actions 1 and 2 are dominated. Thus the game is not trivial.
2. The difference of the loss vectors $\ell_2 - \ell_1 = \begin{pmatrix} 5 & 3 & 1 & -1 & -3 & -5 \end{pmatrix}^\top$.
3. The image space of the signal matrices $\operatorname{Im} S_1 = \operatorname{Im} S_2 = \{\ell \in \mathbb{R}^6 \ : \ \ell[5] = \ell[6]\}$.

The three points together imply that the game is not globally observable.

Dynamic pricing In dynamic pricing, a vendor (learner) tries to sell his product to a buyer (opponent). The buyer secretly chooses a maximum price (outcome) while the seller tries to sell it at some price (action). If the outcome is lower than the action then no transaction happens and the seller suffers some constant loss. Otherwise the buyer buys the product and the seller's loss is the difference between the seller's price and the buyer's price. The feedback for the seller is, however, only the binary observation if the transaction happened (y for yes and n for no). The finite version of the game can be described with the following matrices:

$$L = \begin{pmatrix} 0 & 1 & 2 & \cdots & N-1 \\ c & 0 & 1 & \cdots & N-2 \\ c & c & 0 & \cdots & N-3 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c & \cdots & \cdots & c & 0 \end{pmatrix}; \qquad H = \begin{pmatrix} y & y & \cdots & y \\ n & y & \cdots & y \\ \vdots & \ddots & \ddots & \vdots \\ n & \cdots & n & y \end{pmatrix}.$$

Simple algebra gives that all action pairs are neighbors. In fact, there is a single point on the probability simplex that is common to all of the cells, namely

$$p = \begin{pmatrix} \frac{1}{c+1} & \frac{c}{(c+1)^2} & \cdots & \frac{c^{i-1}}{(c+1)^i} & \cdots & \frac{c^{N-2}}{(c+1)^{N-1}} & \frac{c^{N-1}}{(c+1)^{N-1}} \end{pmatrix}^\top .$$

We show that the locally observable action pairs are the "consecutive" actions ($\{i, i+1\}$). The difference $\ell_{i+1} - \ell_i$ is

$$\ell_{i+1} - \ell_i = \begin{pmatrix} 0 & \cdots & 0 & c & -1 & \cdots & -1 \end{pmatrix}$$

with $i-1$ zeros at the beginning. The signal matrix S_i is

$$S_i = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}$$

where the "switch" is after $i-1$ columns. Thus,

$$\ell_{i+1} - \ell_i = S_i^\top \begin{pmatrix} -c \\ 0 \end{pmatrix} + S_{i+1}^\top \begin{pmatrix} c \\ -1 \end{pmatrix} .$$

On the other hand, action pairs that are not consecutive are not locally observable. For example,

$$\ell_3 - \ell_1 = \begin{pmatrix} c & c-1 & -2 & \cdots & -2 \end{pmatrix}^\top ,$$

while both $\operatorname{Im} S_1^\top$ and $\operatorname{Im} S_3^\top$ contain only vectors whose first two coordinates are identical. Thus, dynamic pricing is not a locally observable game. Nevertheless, it is easy to see that global observability holds.

3 Summary of results

In this paper we present new algorithms for finite partial-monitoring games—NeigborhoodWatch for the adversarial case and CBP for the stochastic case—and provide regret bounds. Our results on the minimax regret can be summarized in the following two classification theorems.

Theorem 1 (Classification for games against stochastic opponents)**.** *Let $G = (L, H)$ be a finite partial-monitoring game. Let K be the number of non-dominated actions in G. The minimax expected regret of G against stochastic opponents is*

$$\mathbb{E}[R_T(G)] = \begin{cases} 0, & K = 1; \\ \widetilde{\Theta}(\sqrt{T}), & K > 1,\ G \text{ is locally observable;} \\ \Theta(T^{2/3}), & G \text{ is globally observable but not locally observable;} \\ \Theta(T), & G \text{ is not globally observable.} \end{cases}$$

To state our classification theorem for the case of adversarial opponents, we need a definition.

Definition 7 (Degenerate games)**.** *A partial-monitoring game G is called* degenerate *if it has degenerate or duplicate actions. A game is called* non-degenerate *if it is not degenerate.*

Theorem 2 (Classification for games against adversarial opponents)**.** *Let $G = (L, H)$ be a non-degenerate finite partial-monitoring game. Let K be the number of non-dominated actions in G. The minimax expected regret of G against adversarial opponents is*

$$\mathbb{E}[R_T(G)] = \begin{cases} 0, & K = 1; \\ \Theta(\sqrt{T}), & K > 1, \ G \text{ is locally observable;} \\ \Theta(T^{2/3}), & G \text{ is globally observable but not locally observable;} \\ \Theta(T), & G \text{ is not globally observable.} \end{cases}$$

For the stochastic case, we additionally present individual bounds on the regret of any finite partial-monitoring game, *i.e.*, bounds that depend on the strategy of the opponent.[4]

Theorem 3. *Let (L, H) be an N by M partial-monitoring game. For a fixed opponent strategy $p^* \in \Delta_M$, let δ_i denote the difference between the expected loss of action i and an optimal action. For any time horizon T, algorithm* CBP *with parameters $\alpha > 1$, $\nu_k = W_k^{2/3}$, $f(t) = \alpha^{1/3} t^{2/3} \log^{1/3} t$ has expected regret*

$$\begin{aligned}
\mathbb{E}[R_T] \le & \sum_{\{i,j\} \in \mathcal{N}} 2|V_{i,j}| \left(1 + \frac{1}{2\alpha - 1}\right) + \sum_{k=1}^{N} \delta_k \\
& + \sum_{\substack{k=1 \\ \delta_k > 0}}^{N} 4 W_k^2 \frac{d_k^2}{\delta_k} \alpha \log T \\
& + \sum_{k \in \mathcal{V} \setminus N^+} \delta_k \min\left(4 W_k^2 \frac{d_{l(k)}^2}{\delta_{l(k)}^2} \alpha \log T, \ \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3} T \right) \\
& + \sum_{k \in \mathcal{V} \setminus N^+} \delta_k \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3} T \\
& \qquad\qquad + 2 d_k \alpha^{1/3} W^{2/3} T^{2/3} \log^{1/3} T \,,
\end{aligned}$$

where $W = \max_{k \in \underline{N}} W_k$, $\mathcal{V} = \cup_{\{i,j\} \in \mathcal{N}} V_{i,j}$, $N^+ = \cup_{\{i,j\} \in \mathcal{N}} N_{i,j}^+$, and $d_1, \ldots, d_N$ are game-dependent constants.

Theorem 3 gives a very general bound on the regret. This bound will be used to derive all the upper bounds that concern the regret of games against stochastic environments: It translates to a logarithmic individual upper bound on the regret of locally observable games (Corollary 1); it gives the minimax upper bound of $\widetilde{O}(\sqrt{T})$ for locally observable games (Corollary 3), the minimax upper bound of $\widetilde{O}(T^{2/3})$ for globally observable games (Corollary 2). Additionally and quite surprisingly, it also follows from the above bound that even for not locally observable games, if we assume that the opponent is "benign" in some sense, the minimax regret of $\widetilde{O}(\sqrt{T})$ still holds. For the precise statement, see Theorem 5.

In the next section we give a lower bound on the minimax regret for games that are not locally observable. This bound is valid for both the stochastic and the adversarial settings and is necessary for proving the classification theorems. Then, in Sections 5 and 6, we describe and analyze the algorithms CBP and NeigborhoodWatch. The first algorithm, CBP for Confidence Bound Partial monitoring, is shown to achieve the desired regret upper bounds for any finite partial-monitoring game against stochastic opponents. The second algorithm, NeigborhoodWatch, works for locally observable non-degenerate games. We show that for these games, the algorithm achieves the desired $O(\sqrt{T})$ regret bound against adversarial opponents.

[4] Some of the notations used by the theorem is defined in the next section.

4 A lower bound for not locally observable games

In this section we prove that for any game that does not satisfy the local observability condition has expected minimax regret of $\Omega(T^{2/3})$.

Theorem 4. *Let $G = (L, H)$ be an N by M partial-monitoring game. Assume that there exist two neighboring actions i and j that are not locally observable. Then there exists a problem dependent constant $c(G)$ such that for any algorithm $\mathcal{A}$ and time horizon T there exists an opponent strategy p such that the expected regret satisfies*

$$\mathbb{E}\left[R_T\left(\mathcal{A}, p\right)\right] \geq c(G) T^{2/3} .$$

Proof. Without loss of generality we can assume that the two neighbor cells in the condition are C_1 and C_2. Let $C_3 = C_1 \cap C_2$. For $i = 1, 2, 3$, let $\mathcal{N}_i$ be the set of actions associated with cell C_i. Note that $\mathcal{N}_3$ may be the empty set. Let $\mathcal{N}_4 = \mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2 \cup \mathcal{N}_3)$. By our convention for naming loss vectors, ℓ_1 and ℓ_2 are the loss vectors for C_1 and C_2, respectively. Let $\mathcal{L}_3$ collect the loss vectors of actions which lie on the open segment connecting ℓ_1 and ℓ_2. It is easy to see that $\mathcal{L}_3$ is the set of loss vectors that correspond to the cell C_3. We define $\mathcal{L}_4$ as the set of all the other loss vectors. For $i = 1, 2, 3, 4$, let $k_i = |\mathcal{N}_i|$.

According to the lack of local observability, $\ell_2 - \ell_1 \not\in \operatorname{Im} S_1^\top \oplus \operatorname{Im} S_2^\top$. Thus, $\{\rho(\ell_2 - \ell_1) \ : \ \rho \in \mathbb{R}\} \not\subset \operatorname{Im} S_1^\top \oplus \operatorname{Im} S_2^\top$, or equivalently, $(\ell_2 - \ell_1)^\perp \not\supset \operatorname{Ker} S_1 \cap \operatorname{Ker} S_2$, where we used that $(\operatorname{Im} M)^\perp = \operatorname{Ker}(M^\top)$. Thus, there exists a vector v such that $v \in \operatorname{Ker} S_1 \cap \operatorname{Ker} S_2$ and $(\ell_2 - \ell_1)^\top v \neq 0$. By scaling we can assume that $(\ell_2 - \ell_1)^\top v = 1$. Note that since $v \in \operatorname{Ker} S_1 \cap \operatorname{Ker} S_2$ and the rowspaces of both S_1 and S_2 contain the vector $(1, 1, \ldots, 1)$, the coordinates of v sum up to zero.

Let p_0 be an arbitrary probability vector in the relative interior of C_3. It is easy to see that for any $\varepsilon > 0$ small enough, $p_1 = p_0 + \varepsilon v \in C_1 \setminus C_2$ and $p_2 = p_0 - \varepsilon v \in C_2 \setminus C_1$.

Let us fix a deterministic algorithm $\mathcal{A}$ and a time horizon T. For $i = 1, 2$, let $R_T^{(i)}$ denote the expected regret of the algorithm under opponent strategy p_i. For $i = 1, 2$ and $j = 1, \ldots, 4$, let N_j^i denote the expected number of times the algorithm chooses an action from $\mathcal{N}_j$, assuming the opponent plays strategy p_i.

From the definition of $\mathcal{L}_3$ we know that for any $\ell \in \mathcal{L}_3$, $\ell - \ell_1 = \eta_\ell(\ell_2 - \ell_1)$ and $\ell - \ell_2 = (1 - \eta_\ell)(\ell_1 - \ell_2)$ for some $0 < \eta_\ell < 1$. Let $\lambda_1 = \min_{\ell \in \mathcal{L}_3} \eta_\ell$ and $\lambda_2 = \min_{\ell \in \mathcal{L}_3}(1 - \eta_\ell)$ and $\lambda = \min(\lambda_1, \lambda_2)$ if $\mathcal{L}_3 \neq \emptyset$ and let $\lambda = 1/2$, otherwise. Finally, let $\beta_i = \min_{\ell \in \mathcal{L}_4} (\ell - \ell_i)^\top p_i$ and $\beta = \min(\beta_1, \beta_2)$. Note that $\lambda, \beta > 0$.

As the first step of the proof, we lower bound the expected regret $R_T^{(1)}$ and $R_T^{(2)}$ in terms of the values $N_j^i, \varepsilon, \lambda$ and β:

$$\begin{aligned} R_T^{(1)} &\geq N_2^1 \overbrace{(\ell_2 - \ell_1)^\top p_1}^{\varepsilon} + N_3^1 \lambda (\ell_2 - \ell_1)^\top p_1 + N_4^1 \beta \geq \lambda (N_2^1 + N_3^1)\varepsilon + N_4^1 \beta \,, \\ R_T^{(2)} &\geq N_1^2 \underbrace{(\ell_1 - \ell_2)^\top p_2}_{\varepsilon} + N_3^2 \lambda (\ell_1 - \ell_2)^\top p_2 + N_4^2 \beta \geq \lambda (N_1^2 + N_3^2)\varepsilon + N_4^2 \beta \,. \end{aligned} \tag{1}$$

For the next step, we need the following lemma.

Lemma 1. *There exists a (problem dependent) constant c such that the following inequalities hold:*

$$N_1^2 \geq N_1^1 - cT\varepsilon\sqrt{N_4^1} \,, \qquad N_3^2 \geq N_3^1 - cT\varepsilon\sqrt{N_4^1} \,,$$

$$N_2^1 \geq N_2^2 - cT\varepsilon\sqrt{N_4^2} \,, \qquad N_3^1 \geq N_3^2 - cT\varepsilon\sqrt{N_4^2} \,.$$

Using the above lemma we can lower bound the expected regret. Let $r = \operatorname{argmin}_{i \in \{1,2\}} N_4^i$. It is easy to see that for $i = 1, 2$ and $j = 1, 2, 3$,

$$N_j^i \geq N_j^r - c_2 T\varepsilon\sqrt{N_4^r} \,.$$

If $i \neq r$ then this inequality is one of the inequalities from Lemma 1. If $i = r$ then it is a trivial lower bounding by subtracting a positive value. From (1) we have

$$\begin{aligned}
R_T^{(i)} &\geq \lambda(N_{3-i}^i + N_3^i)\varepsilon + N_4^i\beta \\
&\geq \lambda(N_{3-i}^r - c_2T\varepsilon\sqrt{N_4^r} + N_3^r - c_2T\varepsilon\sqrt{N_4^r})\varepsilon + N_4^r\beta \\
&= \lambda(N_{3-i}^r + N_3^r - 2c_2T\varepsilon\sqrt{N_4^r})\varepsilon + N_4^r\beta\,.
\end{aligned}$$

Now assume that, at the beginning of the game, the opponent randomly chooses between strategies p_1 and p_2 with equal probability. The the expected regret of the algorithm is lower bounded by

$$\begin{aligned}
R_T &= \frac{1}{2}\left(R_T^{(1)} + R_T^{(2)}\right) \\
&\geq \frac{1}{2}\lambda(N_1^r + N_2^r + 2N_3^r - 4c_2T\varepsilon\sqrt{N_4^r})\varepsilon + N_4^r\beta \\
&\geq \frac{1}{2}\lambda(N_1^r + N_2^r + N_3^r - 4c_2T\varepsilon\sqrt{N_4^r})\varepsilon + N_4^r\beta \\
&= \frac{1}{2}\lambda(T - N_4^r - 4c_2T\varepsilon\sqrt{N_4^r})\varepsilon + N_4^r\beta\,.
\end{aligned}$$

Choosing $\varepsilon = c_3T^{-1/3}$ we get

$$\begin{aligned}
R_T &\geq \frac{1}{2}\lambda c_3T^{2/3} - \frac{1}{2}\lambda N_4^r c_3T^{-1/3} - 2\lambda c_2c_3^2T^{1/3}\sqrt{N_4^r} + N_4^r\beta \\
&\geq T^{2/3}\left(\left(\beta - \frac{1}{2}\lambda c_3\right)\frac{N_4^r}{T^{2/3}} - 2\lambda c_2c_3^2\sqrt{\frac{N_4^r}{T^{2/3}}} + \frac{1}{2}\lambda c_3\right) \\
&= T^{2/3}\left(\left(\beta - \frac{1}{2}\lambda c_3\right)x^2 - 2\lambda c_2c_3^2x + \frac{1}{2}\lambda c_3\right)\,,
\end{aligned}$$

where $x = \sqrt{N_4^r/T^{2/3}}$. Now we see that $c_3 > 0$ can be chosen to be small enough, independently of T so that, for any choice of x, the quadratic expression in the parenthesis is bounded away from zero, and simultaneously, ε is small enough so that the threshold condition in Lemma 10 is satisfied, completing the proof of Theorem 4. □

5 The stochastic case

In this section we present and analyze our algorithm CBP for *Confidence Bound Partial monitoring* that achieves near optimal regret for any finite partial-monitoring game against stochastic opponents. In particular, we show that CBP achieves $\widetilde{O}(\sqrt{T})$ regret for locally observable games and $O(T^{2/3})$ regret for globally observable games.

5.1 The proposed algorithm

In the core of the algorithm lie the concepts of *observer action sets* and *observer vectors*:

Definition 8 (Observer sets and observer vectors)**.** *The* observer set $V_{i,j} \subset \underline{N}$ *underlying a pair of neighboring actions* $\{i,j\} \in \mathcal{N}$ *is a set of actions such that*

$$\ell_i - \ell_j \in \oplus_{k \in V_{i,j}} \operatorname{Im} S_k^\top\,.$$

The observer vectors $(v_{i,j,k})_{k \in V_{i,j}}$ *underlying* $V_{i,j}$ *are defined to satisfy the equation* $\ell_i - \ell_j = \sum_{k \in V_{i,j}} S_k^\top v_{i,j,k}$*. In particular,* $v_{i,j,k} \in \mathbb{R}^{s_k}$*. In what follows, the choice of the observer sets and vectors is restricted so that* $V_{i,j} = V_{j,i}$ *and* $v_{i,j,k} = -v_{j,i,k}$*. Furthermore, the observer set* $V_{i,j}$ *is constrained to be a superset of* $N_{i,j}^+$ *and, in particular, when a pair* $\{i,j\}$ *is locally observable,* $V_{i,j} = N_{i,j}^+$ *must hold. Finally, for any action* $k \in \bigcup_{\{i,j\} \in \mathcal{N}} V_{i,j}$*, let* $W_k = \max_{i,j:k \in V_{i,j}} \|v_{i,j,k}\|_\infty$*.*

In a nutshell, CBP works as follows. For every neighboring action pair it maintains an unbiased estimate of the expected difference of their losses. It also keeps a confidence width for these estimates. If at time step t an estimate is "confident enough" to determine which action is better, the algorithm excludes some actions from the set of potentially optimal actions.

For two actions i, j, let $\delta_{i,j}$ denote the expected difference of their losses. That is, $\delta_{i,j} = (\ell_i - \ell_j)^\top p^*$ where p^* is the opponent strategy. At any time step t, the estimate of the loss difference of actions i and j is calculated as

$$\tilde{\delta}_{i,j}(t) = \sum_{k \in V_{i,j}} v_{i,j,k}^\top \frac{\sum_{s=1}^{t-1} \mathbb{I}\{I_s = k\} Y_s}{\sum_{s=1}^{t-1} \mathbb{I}\{I_s = k\}} .$$

The confidence bound of the loss difference estimate is defined as

$$c_{i,j}(t) = \sum_{k \in V_{i,j}} \|v_{i,j,k}\|_\infty \sqrt{\frac{\alpha \log t}{\sum_{s=1}^{t-1} \mathbb{I}\{I_s = k\}}}$$

with some preset parameter α. We call the estimate $\tilde{\delta}_{i,j}(t)$ confident if $|\tilde{\delta}_{i,j}(t)| \geq c_{i,j}(t)$.

In every time step t, the algorithm uses the estimates and the widths to select a set of candidate actions. If an estimate $\tilde{\delta}_{i,j}(t)$ is confident then the algorithm assumes that the opponent strategy p^* lies in the halfspace defined as $\{p \in \Delta_M \,:\, \operatorname{sgn}(\tilde{\delta}_{i,j}(t))(\ell_i - \ell_j)^\top p > 0\}$. Taking the intersection of these halfspaces for all the action pairs with confident estimates, we arrive at a polyitope that contains the opponent strategy with high probability. Then, the set of potentially optimal actions $\mathcal{P}(t)$ is defined as the actions whose cells intersect with the above polytope. We also need to maintain the set $\mathcal{N}(t)$ of neighboring actions, since it may happen that action pairs that are originally neighbors do not share an $M-2$ dimensional facet in this polytope. Then, the actions candidate for choosing by the algorithm is defined as the union of observer action sets of current neighboring pairs: $Q(t) = \cup_{\{i,j\} \in \mathcal{N}(t)} V_{i,j}$. Finally, the action is chosen to be the one that potentially reduces the remaining uncertainty the most:

$$I_t = \operatorname{argmax}_{k \in Q(t)} \frac{W_k^2}{\sum_{s=1}^{t-1} \mathbb{I}\{I_s = k\}} ,$$

where $W_k = \max\{\|v_{i,j,k}\|_\infty \,:\, k \in N_{i,j}^+\}$ with fixed $v_{i,j,k}$ precomputed and used by the algorithm.

Decaying exploration. The algorithm depicted above could be shown to achieve low regret for locally observable games. However, for a game that is only globally observable, the opponent can choose a strategy that causes the algorithm to suffer linear regret: Let action 1 and 2 be a neighboring action pair that is not locally observable. It follows that their observer action set must contain a third action 3 with $C_3 \not\supseteq C_1 \cap C_2$. If the opponent chooses a strategy $p \in C_1 \cap C_2$ then actions 1 and 2 are optimal while action 3 is not. Unfortunately, the algorithm will choose action 3 linearly many times in its effort to (futilely) estimate the loss difference of actions 1 and 2.

To prevent the algorithm from falling in the above trap, we introduce the *decaying exploration rule*. This rule, described below, upper bounds the number of times an action can be chosen for only information seeking purposes. For this, we introduce the set of rarely chosen actions,

$$\mathcal{R}(t) = \{k \in \underline{N} \,:\, n_k(t) \leq \eta_k f(t)\} ,$$

where $\eta_k \in \mathbb{R}$, $f : \mathbb{N} \mapsto \mathbb{R}$ are tuning parameters to be chosen later. Then, the set of actions available at time t is restricted to

$$Q(t) = \bigcup_{\{i,j\} \in \mathcal{N}(t)} N_{i,j}^+ \cup \left(\bigcup_{\{i,j\} \in \mathcal{N}(t)} V_{i,j} \cap \mathcal{R}(t) \right) .$$

Symbol	*Definition*	*Found in/at*
$N, M \in \mathbb{N}$	number of actions and outcomes	
$\underline{N}$	$\{1, \ldots, N\}$, set of actions	
$\Delta_M \subset^M$	M-dim. simplex, set of opponent strategies	
$p^* \in \Delta_M$	opponent strategy	
$L \in \mathbb{R}^{N \times M}$	loss matrix	
$H \in \Sigma^{N \times M}$	feedback matrix	
$\ell_i \in \mathbb{R}^M$	$\ell_i = L_{i,:}$, loss vector underlying action i	
$C_i \subseteq \Delta_M$	cell of action i	Definition 1
$\mathcal{P} \subseteq \underline{N}$	set of Pareto-optimal actions	Definition 2
$\mathcal{N} \subseteq \underline{N}^2$	set of unordered neighboring action-pairs	Definition 3
$N^+_{i,j} \subseteq \underline{N}$	neighborhood action set of $\{i, j\} \in \mathcal{N}$	Definition 3
$S_i \in \{0,1\}^{s_i \times M}$	signal matrix of action i	Definition 4
$\mathcal{L} \subseteq \mathcal{N}$	set of locally observable action pairs	Definition 6
$V_{i,j} \subseteq \underline{N}$	observer actions underlying $\{i, j\} \in \mathcal{N}$	Definition 8
$v_{i,j,k} \in^{s_k}$, $k \in V_{i,j}$	observer vectors	Definition 8
$W_i \in \mathbb{R}$	confidence width for action $i \in \underline{N}$	Definition 8

Table 1: List of basic symbols

We will show that with these modifications, the algorithm achieves $O(T^{2/3})$ regret on globally observable games, while it will also be shown to achieve an $O(\sqrt{T})$ regret when the opponent uses a benign strategy. Pseudocode for the algorithm is given in Algorithm 1.

It remains to specify the function getPolytope. It gets the array *halfSpace* as input. The array *halfSpace* stores which neighboring action pairs have a confident estimate on the difference of their expected losses, along with the sign of the difference (if confident). Each of these confident pairs define an open halfspace, namely

$$\Delta_{\{i,j\}} = \left\{ p \in \Delta_M \,:\, \textit{halfSpace}(i,j)(\ell_i - \ell_j)^\top p > 0 \right\} .$$

The function getPolytope calculates the open polytope defined as the intersection of the above halfspaces. Then for all $i \in \mathcal{P}$ it checks if C_i intersects with the open polytope. If so, then i will be an element of $\mathcal{P}(t)$. Similarly, for every $\{i, j\} \in \mathcal{N}$, it checks if $C_i \cap C_j$ intersects with the open polytope and puts the pair in $\mathcal{N}(t)$ if it does.

For the convenience of the reader, we include a list of symbols used in this Chapter in Table 1. The list of symbols used in the algorithm is shown in Table 2.

Computational complexity The computationally heavy parts of the algorithm are the initial calculation of the cell decomposition and the function getPolytope. All of these require linear programming. In the preprocessing phase we need to solve $N + N^2$ linear programs to determine cells and neighboring pairs of cells. Then in every round, at most N^2 linear programs are needed. The algorithm can be sped up by "caching" previously solved linear programs.

5.2 Analysis of the algorithm

The first theorem in this section is an individual upper bound on the regret of CBP.

Theorem 3. *Let (L, H) be an N by M partial-monitoring game. For a fixed opponent strategy $p^* \in \Delta_M$, let δ_i denote the difference between the expected loss of action i and an optimal action. For any time horizon*

Algorithm 1 CBP

```
Input: $L$, $H$, $\alpha$, $\eta_1, \ldots, \eta_N$, $f = f(\cdot)$
Calculate $\mathcal{P}$, $\mathcal{N}$, $V_{i,j}$, $v_{i,j,k}$, $W_k$
for $t = 1$ to $N$ do
  Choose $I_t = t$ and observe $Y_t$                                  {Initialization}
  $n_{I_t} \leftarrow 1$                                              {# times the action is chosen}
  $\nu_{I_t} \leftarrow Y_t$                                          {Cumulative observations}
end for
for $t = N+1, N+2, \ldots$ do
  for each $\{i,j\} \in \mathcal{N}$ do
    $\tilde{\delta}_{i,j} \leftarrow \sum_{k \in V_{i,j}} v_{i,j,k}^\top \frac{\nu_k}{n_k}$                     {Loss diff. estimate}
    $c_{i,j} \leftarrow \sum_{k \in V_{i,j}} \|v_{i,j,k}\|_\infty \sqrt{\frac{\alpha \log t}{n_k}}$        {Confidence}
    if $|\tilde{\delta}_{i,j}| \geq c_{i,j}$ then
      $\mathit{halfSpace}(i,j) \leftarrow \operatorname{sgn} \tilde{\delta}_{i,j}$
    else
      $\mathit{halfSpace}(i,j) \leftarrow 0$
    end if
  end for
  $[\mathcal{P}(t), \mathcal{N}(t)] \leftarrow \textsc{getPolytope}(\mathcal{P}, \mathcal{N}, \mathit{halfSpace})$
  $N^+(t) = \cup_{\{i,j\} \in \mathcal{N}(t)} N_{ij}^+$
  $\mathcal{V}(t) = \cup_{\{i,j\} \in \mathcal{N}(t)} V_{ij}$
  $\mathcal{R}(t) = \{k \in \underline{N} : n_k(t) \leq \eta_k f(t)\}$
  $\mathcal{S}(t) = \mathcal{P}(t) \cup N^+(t) \cup (\mathcal{V}(t) \cap \mathcal{R}(t))$
  Choose $I_t = \operatorname{argmax}_{i \in \mathcal{S}(t)} \frac{W_i^2}{n_i}$ and observe $Y_t$
  $\nu_{I_t} \leftarrow \nu_{I_t} + Y_t$
  $n_{I_t} \leftarrow n_{I_t} + 1$
end for
```

T, algorithm CBP *with parameters $\alpha > 1$, $\nu_k = W_k^{2/3}$, $f(t) = \alpha^{1/3} t^{2/3} \log^{1/3} t$ has expected regret*

$$
\begin{aligned}
\mathbb{E}[R_T] \leq & \sum_{\{i,j\} \in \mathcal{N}} 2|V_{i,j}| \left(1 + \frac{1}{2\alpha - 1}\right) + \sum_{k=1}^{N} \delta_k \\
& + \sum_{\substack{k=1 \\ \delta_k > 0}}^{N} 4W_k^2 \frac{d_k^2}{\delta_k} \alpha \log T \\
& + \sum_{k \in \mathcal{V} \setminus N^+} \delta_k \min\left(4W_k^2 \frac{d_{l(k)}^2}{\delta_{l(k)}^2} \alpha \log T,\ \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3} T\right) \\
& + \sum_{k \in \mathcal{V} \setminus N^+} \delta_k \alpha^{1/3} W_k^{2/3} T^{2/3} \log^{1/3} T \\
& \qquad + 2 d_k \alpha^{1/3} W^{2/3} T^{2/3} \log^{1/3} T \,,
\end{aligned}
$$

where $W = \max_{k \in \underline{N}} W_k$, $\mathcal{V} = \cup_{\{i,j\} \in \mathcal{N}} V_{i,j}$, $N^+ = \cup_{\{i,j\} \in \mathcal{N}} N_{i,j}^+$, and $d_1, \ldots, d_N$ are game-dependent constants.

Proof. We use the convention that, for any variable x used by the algorithm, $x(t)$ denotes the value of x at the *end* of time step t. For example, $n_i(t)$ is the number of times action i is chosen up to and including time step t.

Symbol	*Definition*
$I_t \in \underline{N}$	action chosen at time t
$Y_t \in \{0,1\}^{s_{I_t}}$	observation at time t
$\tilde{\delta}_{i,j}(t) \in$	estimate of $(\ell_i - \ell_j)^\top p$ ($\{i,j\} \in \mathcal{N}$)
$c_{i,j}(t) \in$	confidence width for pair $\{i,j\}$ ($\{i,j\} \in \mathcal{N}$)
$\mathcal{P}(t) \subseteq \underline{N}$	plausible actions
$\mathcal{N}(t) \subseteq \underline{N}^2$	set of admissible neighbors
$N^+(t) \subseteq \underline{N}$	$\cup_{\{i,j\}\in\mathcal{N}(t)} N^+_{i,j}$; admissible neighborhood actions
$\mathcal{V}(t) \subseteq \underline{N}$	$\cup_{\{i,j\}\in\mathcal{N}(t)} V_{i,j}$; admissible information seeking actions
$\mathcal{R}(t) \subseteq \underline{N}$	rarely sampled actions
$\mathcal{S}(t)$	$\mathcal{P}(t) \cup N^+(t) \cup (\mathcal{V}(t) \cap \mathcal{R}(t))$; admissible actions

Table 2: List of symbols used in the algorithm

The proof is based on three lemmas. The first lemma shows that the estimate $\tilde{\delta}_{i,j}(t)$ is in the vicinity of $\delta_{i,j}$ with high probability.[5]

Lemma 2. *For any $\{i,j\} \in \mathcal{N}$, $t \geq 1$,*

$$\mathbb{P}\left(|\tilde{\delta}_{i,j}(t) - \delta_{i,j}| \geq c_{i,j}(t)\right) \leq 2|V_{i,j}|t^{1-2\alpha}\,.$$

If for some t, i, j, the event whose probability is upper-bounded in Lemma 2 happens, we say that a confidence interval fails. Let $\mathcal{G}_t$ be the event that no confidence intervals fail in time step t and let $\mathcal{B}_t$ be its complement event. An immediate corollary of Lemma 2 is that the sum of the probabilities that some confidence interval fails is small:

$$\sum_{t=1}^{T} \mathbb{P}\left(\mathcal{B}_t\right) \leq \sum_{t=1}^{T} \sum_{\{i,j\}\in\mathcal{N}} 2|V_{i,j}|t^{-2\alpha} \leq \sum_{\{i,j\}\in\mathcal{N}} 2|V_{i,j}| \left(1 + \frac{1}{2\alpha - 2}\right)\,. \tag{2}$$

To prepare for the next lemma, we need some new notations. For the next definition we need to denote the dependence of the random sets $\mathcal{P}(t)$, $\mathcal{N}(t)$ on the outcomes ω from the underlying sample space Ω. For this, we will use the notation $\mathcal{P}_\omega(t)$ and $\mathcal{N}_\omega(t)$. With this, we define the *set of plausible configurations* to be

$$\Psi = \cup_{t\geq 1} \left\{ (\mathcal{P}_\omega(t), \mathcal{N}_\omega(t)) \,:\, \omega \in \mathcal{G}_t \right\}\,.$$

Call $\pi = (i_0, i_1, \ldots, i_r)$ ($r \geq 0$) a *path* in $\mathcal{N}' \subseteq \underline{N}^2$ if $\{i_s, i_{s+1}\} \in \mathcal{N}'$ for all $0 \leq s \leq r-1$ (when $r = 0$ there is no restriction on π). The path is said to start at i_0 and end at i_r. In what follows we denote by i^* an optimal action under p^* (i.e., $\ell_{i^*}^\top p^* \leq \ell_i^\top p^*$ holds for all actions i).

The set of paths that connect i to i^* and lie in $\mathcal{N}'$ will be denoted by $B_i(\mathcal{N}')$. The next lemma shows that $B_i(\mathcal{N}')$ is non-empty whenever $\mathcal{N}'$ is such that for some $\mathcal{P}'$, $(\mathcal{P}', \mathcal{N}') \in \Psi$:

Lemma 3. *Take an action i and a plausible pair $(\mathcal{P}', \mathcal{N}') \in \Psi$ such that $i \in \mathcal{P}'$. Then there exists a path π that starts at i and ends at i^* that lies in $\mathcal{N}'$.*

For $i \in \mathcal{P}$ define

$$d_i = \max_{\substack{(\mathcal{P}',\mathcal{N}')\in\Psi \\ i\in\mathcal{P}'}} \; \min_{\substack{\pi\in B_i(\mathcal{N}') \\ \pi=(i_0,\ldots,i_r)}} \sum_{s=1}^{r} |V_{i_{s-1},i_s}|\,.$$

[5] The proofs of technical lemmas can be found in the appendix.

According to the previous lemma, for each Pareto-optimal action i, the quantity d_i is well-defined and finite. The definition is extended to degenerate actions by defining d_i to be $\max(d_l, d_k)$, where k, l are such that $i \in N^+_{k,l}$.

Let $k(t) = \operatorname{argmax}_{i \in \mathcal{P}(t) \cup V(t)} W_i^2 / n_i(t-1)$. When $k(t) \neq I_t$ this happens because $k(t) \notin N^+(t)$ and $k(t) \notin \mathcal{R}(t)$, *i.e.*, the action $k(t)$ is a "purely" information seeking action which has been sampled frequently. When this holds we say that the *"decaying exploration rule is in effect at time step t"*. The corresponding event is denoted by $\mathcal{D}_t = \{k(t) \neq I_t\}$. Let δ_i be defined as $\max_{j \in \underline{N}} \delta_{i,j}$, *i.e.*, δ_i is the excess expected loss of action i compared to an optimal action.

Lemma 4. *Fix any* $t \geq 1$.

1. *Take any action* i. *On the event* $\mathcal{G}_t \cap \mathcal{D}_t$,[6] *from* $i \in \mathcal{P}(t) \cup N^+(t)$ *it follows that*

$$\delta_i \leq 2 d_i \sqrt{\frac{\alpha \log t}{f(t)}} \max_{k \in \underline{N}} \frac{W_k}{\sqrt{\eta_k}} \,.$$

2. *Take any action* k. *On the event* $\mathcal{G}_t \cap \mathcal{D}_t^c$, *from* $I_t = k$ *it follows that*

$$n_k(t-1) \leq \min_{j \in \mathcal{P}(t) \cup N^+(t)} 4 W_k^2 \frac{d_j^2}{\delta_j^2} \alpha \log t \,.$$

We are now ready to start the proof. By Wald's identity, we can rewrite the expected regret as follows:

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\left[\sum_{t=1}^T L_{I_t, J_t}\right] - \sum_{t=1}^T \mathbb{E}\left[L_{i^*, J_1}\right] = \sum_{k=1}^N \mathbb{E}[n_k(T)] \delta_i \\
&= \sum_{k=1}^N \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\left\{I_t = k\right\}\right] \delta_k \\
&= \sum_{k=1}^N \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\left\{I_t = k, \mathcal{B}_t\right\}\right] \delta_k + \sum_{k=1}^N \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\left\{I_t = k, \mathcal{G}_t\right\}\right] \delta_k \,.
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_{k=1}^N \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\left\{I_t = k, \mathcal{B}_t\right\}\right] \delta_k &\leq \sum_{k=1}^N \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\left\{I_t = k, \mathcal{B}_t\right\}\right] && \text{(because } \delta_k \leq 1\text{)} \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{k=1}^N \mathbb{I}\left\{I_t = k, \mathcal{B}_t\right\}\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\left\{\mathcal{B}_t\right\}\right] = \sum_{t=1}^T \mathbb{P}\left(\mathcal{B}_t\right) \,.
\end{aligned}$$

Hence,

$$\mathbb{E}[R_T] \leq \sum_{t=1}^T \mathbb{P}\left(\mathcal{B}_t\right) + \sum_{k=1}^N \mathbb{E}[\sum_{t=1}^T \mathbb{I}\left\{I_t = k, \mathcal{G}_t\right\}] \delta_k \,.$$

[6] Here and in what follows all statements that start with "On event X" should be understood to hold almost surely on the event. However, to minimize clutter we will not add the qualifier "almost surely".

Here, the first term can be bounded using (2). Let us thus consider the elements of the second sum:

$$\mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\left\{I_t = k, \mathcal{G}_t\right\}]\delta_k \leq \delta_k +$$

$$\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k \tag{3}$$

$$+\,\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k \tag{4}$$

$$+\,\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k \tag{5}$$

$$+\,\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k\,. \tag{6}$$

The first δ_k corresponds to the initialization phase of the algorithm when every action gets chosen once. The next paragraphs are devoted to upper bounding the above four expressions (3)-(6). Note that, if action k is optimal, that is, if $\delta_k = 0$ then all the terms are zero. Thus, we can assume from now on that $\delta_k > 0$.

Term (3): Consider the event $\mathcal{G}_t \cap D_t^c \cap \{k \in \mathcal{P}(t) \cup N^+(t)\}$. We use case 2 of Lemma 4 with the choice $i = k$. Thus, from $I_t = k$, we get that $i = k \in \mathcal{P}(t) \cup N^+(t)$ and so the conclusion of the lemma gives

$$n_k(t-1) \leq A_k(t) \stackrel{\text{def}}{=} 4W_k^2 \frac{d_k^2}{\delta_k^2} \alpha \log t\,.$$

Therefore, we have

$$\begin{aligned}
&\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\right\} \\
&\quad \leq \sum_{t=N+1}^{T} \mathbb{I}\left\{I_t = k, n_k(t-1) \leq A_k(t)\right\} \\
&\qquad + \sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k, n_k(t-1) > A_k(t)\right\} \\
&\quad = \sum_{t=N+1}^{T} \mathbb{I}\left\{I_t = k, n_k(t-1) \leq A_k(t)\right\} \\
&\quad \leq A_k(T) = 4W_k^2 \frac{d_k^2}{\delta_k^2} \alpha \log T
\end{aligned}$$

yielding

$$(3) \leq 4W_k^2 \frac{d_k^2}{\delta_k} \alpha \log T\,.$$

Term (4): Consider the event $\mathcal{G}_t \cap D_t^c \cap \{k \notin \mathcal{P}(t) \cup N^+(t)\}$. We use case 2 of Lemma 4. The lemma gives that that

$$n_k(t-1) \leq \min_{j \in \mathcal{P}(t) \cup N^+(t)} 4W_k^2 \frac{d_j^2}{\delta_j^2} \alpha \log t\,.$$

We know that $k \in \mathcal{V}(t) = \cup_{\{i,j\} \in \mathcal{N}(t)} V_{i,j}$. Let Φ_t be the set of pairs $\{i,j\}$ in $\mathcal{N}(t) \subseteq \mathcal{N}$ such that $k \in V_{i,j}$. For any $\{i,j\} \in \Phi_t$, we also have that $i, j \in \mathcal{P}(t)$ and thus if $l'_{\{i,j\}} = \operatorname{argmax}_{l \in \{i,j\}} \delta_l$ then

$$n_k(t-1) \leq 4W_k^2 \frac{d^2_{l'_{\{i,j\}}}}{\delta^2_{l'_{\{i,j\}}}} \alpha \log t\,.$$

Therefore, if we define $l(k)$ as the action with

$$\delta_{l(k)} = \min\left\{ \delta_{l'_{\{i,j\}}} \ : \ \{i,j\} \in \mathcal{N},\, k \in V_{i,j} \right\}$$

then it follows that

$$n_k(t-1) \leq 4W_k^2 \frac{d^2_{l(k)}}{\delta^2_{l(k)}} \alpha \log t\,.$$

Note that $\delta_{l(k)}$ can be zero and thus we use the convention $c/0 = \infty$. Also, since k is not in $\mathcal{P}(t) \cup N^+(t)$, we have that $n_k(t-1) \leq \eta_k f(t)$. Define $A_k(t)$ as

$$A_k(t) = \min\left(4W_k^2 \frac{d^2_{l(k)}}{\delta^2_{l(k)}} \alpha \log t, \eta_k f(t) \right)\,.$$

Then, with the same argument as in the previous case (and recalling that $f(t)$ is increasing), we get

$$(4) \leq \delta_k \min\left(4W_k^2 \frac{d^2_{l(k)}}{\delta^2_{l(k)}} \alpha \log T, \eta_k f(T) \right)\,.$$

We remark that without the concept of "rarely sampled actions", the above term would scale with $1/\delta^2_{l(k)}$, causing high regret. This is why the "vanilla version" of the algorithm fails on hard games.

Term (5): Consider the event $\mathcal{G}_t \cap D_t \cap \{k \in \mathcal{P}(t) \cup N^+(t)\}$. From case 1 of Lemma 4 we have that $\delta_k \leq 2d_k \sqrt{\frac{\alpha \log t}{f(t)}} \max_{j \in \underline{N}} \frac{W_j}{\sqrt{\eta_j}}$.

Thus,

$$(5) \leq d_k T \sqrt{\frac{\alpha \log T}{f(T)}} \max_{l \in \underline{N}} \frac{W_l}{\sqrt{\eta_l}}\,.$$

Term (6): Consider the event $\mathcal{G}_t \cap D_t \cap \{k \notin \mathcal{P}(t) \cup N^+(t)\}$. Since $k \notin \mathcal{P}(t) \cup N^+(t)$ we know that $k \in \mathcal{V}(t) \cap \mathcal{R}(t) \subseteq \mathcal{R}(t)$ and hence $n_k(t-1) \leq \eta_k f(t)$. With the same argument as in the cases (3) and (4) we get that

$$(6) \leq \delta_k \eta_k f(T)\,.$$

To conclude the proof of Theorem 3, we set $\eta_k = W_k^{2/3}$, $f(t) = \alpha^{1/3} t^{2/3} \log^{1/3} t$ and, with the notation

$W = \max_{k\in\underline{N}} W_k$, $\mathcal{V} = \cup_{\{i,j\}\in\mathcal{N}} V_{i,j}$, $N^+ = \cup_{\{i,j\}\in\mathcal{N}} N^+_{i,j}$, we write

$$\begin{aligned}
\mathbb{E}[R_T] \le & \sum_{\{i,j\}\in\mathcal{N}} 2|V_{i,j}|\left(1+\frac{1}{2\alpha-2}\right) + \sum_{k=1}^{N}\delta_k \\
& + \sum_{\substack{k=1\\ \delta_k>0}}^{N} 4W_k^2\frac{d_k^2}{\delta_k}\alpha\log T \\
& + \sum_{k\in\mathcal{V}\setminus N^+} \delta_k \min\left(4W_k^2\frac{d_{l(k)}^2}{\delta_{l(k)}^2}\alpha\log T,\ \alpha^{1/3}W_k^{2/3}T^{2/3}\log^{1/3}T\right) \\
& + \sum_{k\in\mathcal{V}\setminus N^+} \delta_k\alpha^{1/3}W_k^{2/3}T^{2/3}\log^{1/3}T \\
& \qquad + 2d_k\alpha^{1/3}W^{2/3}T^{2/3}\log^{1/3}T\,.
\end{aligned}$$

□

An implication of Theorem 3 is an upper bound on the individual regret of locally observable games:

Corollary 1. *If G is locally observable then*

$$\mathbb{E}[R_T] \le \sum_{\{i,j\}\in\mathcal{N}} 2|V_{i,j}|\left(1+\frac{1}{2\alpha-1}\right) + \sum_{k=1}^{N}\delta_k + 4W_k^2\frac{d_k^2}{\delta_k}\alpha\log T\,.$$

Proof. If a game is locally observable then $\mathcal{V}\setminus N^+ = \emptyset$, leaving the last two sums of the statement of Theorem 3 zero. □

The following corollary is an upper bound on the minimax regret of any globally observable game.

Corollary 2. *Let G be a globally observable game. Then there exists a constant c such that the expected regret can be upper bounded independently of the choice of p^* as*

$$\mathbb{E}[R_T] \le cT^{2/3}\log^{1/3}T\,.$$

The following theorem is an upper bound on the minimax regret of any globally observable game against "benign" opponents. To state the theorem, we need a new definition. Let A be some subset of actions in G. We call A a *point-local game* in G if $\bigcap_{i\in A}\mathcal{C}_i \neq \emptyset$.

Theorem 5. *Let G be a globally observable game. Let $\Delta' \subseteq \Delta_M$ be some subset of the probability simplex such that its topological closure $\overline{\Delta'}$ has $\overline{\Delta'}\cap\mathcal{C}_i\cap\mathcal{C}_j = \emptyset$ for every $\{i,j\}\in\mathcal{N}\setminus\mathcal{L}$. Then there exists a constant c such that for every $p^*\in\Delta'$, algorithm* CBP *with parameters $\alpha > 1$, $\nu_k = W_k^{2/3}$, $f(t) = \alpha^{1/3}t^{2/3}\log^{1/3}t$ achieves*

$$\mathbb{E}[R_T] \le cd_{pmax}\sqrt{bT\log T}\,,$$

where b is the size of the largest point-local game, and d_{pmax} is a game-dependent constant.

Proof. To prove this theorem, we use a scheme similar to the proof of Theorem 3. Repeating that proof, we arrive at the same expression

$$
\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T} \mathbb{I}\left\{I_t = k, \mathcal{G}_t\right\}]\delta_k \leq \delta_k +& \\
&\mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k && (3)\\
&+ \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k && (4)\\
&+ \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k && (5)\\
&+ \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t, k \notin \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k\,, && (6)
\end{aligned}
$$

where $\mathcal{G}_t$ and $\mathcal{D}_t$ denote the events that no confidence intervals fail, and the decaying exploration rule is in effect at time step t, respectively.

From the condition of Δ' we have that there exists a positive constant ρ_1 such that for every neighboring action pair $\{i, j\} \in \mathcal{N} \setminus \mathcal{L}$, $\max(\delta_i, \delta_j) \geq \rho_1$. We know from Lemma 4 that if $\mathcal{D}_t$ happens then for any pair $\{i, j\} \in \mathcal{N} \setminus \mathcal{L}$ it holds that $\max(\delta_i, \delta_j) \leq 4N\sqrt{\frac{\alpha \log t}{f(t)}} \max(W_{k'}/\sqrt{\eta_{k'}}) \overset{\text{def}}{=} g(t)$. It follows that if $t > g^{-1}(\rho_1)$ then the decaying exploration rule can not be in effect. Therefore, terms (5) and (6) can be upper bounded by $g^{-1}(\rho_1)$.

With the value ρ_1 defined in the previous paragraph we have that for any action $k \in \mathcal{V} \setminus N^+$, $l(k) \geq \rho_1$ holds, and therefore term (4) can be upper bounded by

$$(4) \leq 4W^2 \frac{4N^2}{\rho_1^2} \alpha \log T\,,$$

using that d_k, defined in the proof of Theorem 3, is at most $2N$. It remains to carefully upper bound term (3). For that, we first need a definition and a lemma. Let $A_\rho = \{i \in \underline{N} \;:\; \delta_i \leq \rho\}$.

Lemma 5. *Let $G = (L, H)$ be a finite partial-monitoring game and $p \in \Delta_M$ an opponent strategy. There exists a $\rho_2 > 0$ such that A_{ρ_2} is a point-local game in G.*

To upper bound term (3), with ρ_2 introduced in the above lemma and $\gamma > 0$ specified later, we write

$$
\begin{aligned}
(3) &= \mathbb{E}[\sum_{t=N+1}^{T} \mathbb{I}\left\{\mathcal{G}_t, \mathcal{D}_t^c, k \in \mathcal{P}(t) \cup N^+(t), I_t = k\right\}]\, \delta_k \\
&\leq \mathbb{I}\left\{\delta_k < \gamma\right\} n_k(T)\delta_k + \mathbb{I}\left\{k \in A_{\rho_2}, \delta_k \geq \gamma\right\} 4W_k^2 \frac{d_k^2}{\delta_k} \alpha \log T + \mathbb{I}\left\{k \notin A_{\rho_2}\right\} 4W^2 \frac{8N^2}{\rho_2} \alpha \log T \\
&\leq \mathbb{I}\left\{\delta_k < \gamma\right\} n_k(T)\gamma + |A_{\rho_2}| 4W^2 \frac{d_{pmax}^2}{\gamma} \alpha \log T + 4NW^2 \frac{8N^2}{\rho_2} \alpha \log T\,,
\end{aligned}
$$

where d_{pmax} is defined as the maximum d_k value within point-local games.

Let b be the number of actions in the largest point-local game. Putting everything together we have

$$\begin{aligned}\mathbb{E}[R_T] \le & \sum_{\{i,j\}\in\mathcal{N}} 2|V_{i,j}|\left(1+\frac{1}{2\alpha-2}\right)+g^{-1}(\rho_1)+\sum_{k=1}^{N}\delta_k \\ & +16W^2\frac{N^3}{\rho_1^2}\alpha\log T+32W^2\frac{N^3}{\rho_2}\alpha\log T \\ & +\gamma T+4bW^2\frac{d_{pmax}^2}{\gamma}\alpha\log T\,.\end{aligned}$$

Now we choose γ to be

$$\gamma = 2Wd_{pmax}\sqrt{\frac{b\alpha\log T}{T}}$$

and we get

$$\mathbb{E}[R_T]\le c_1+c_2\log T+4Wd_{pmax}\sqrt{b\alpha T\log T}\,.$$

□

Remark 1. *Note that the above theorem implies that* CBP *does not need to have any prior knowledge about* Δ' *to achieve* $\sqrt{T}$ *regret. This is why we say our algorithm is "adaptive".*

An immediate implication of Theorem 5 is the following minimax bound for locally observable games:

Corollary 3. *Let G be a locally observable finite partial monitoring game. Then there exists a constant c such that for every $p\in\Delta_M$,*

$$\mathbb{E}[R_T]\le c\sqrt{T\log T}\,.$$

5.3 Example

In this section we demonstrate the results of the previous section through the example of Dynamic Pricing. From Section 2.1 we know that dynamic pricing is not a locally observable game. That is, the minimax regret of the game is $\Theta(T^{2/3})$.

Now, we introduce a restriction on the space of opponent strategies such that the condition of Theorem 5 is satisfied. We need to prevent non-consecutive actions from being simultaneously optimal. A somewhat stronger condition is that out of three actions $i<j<k$, the loss of j should not be more than that of *both* i and k. We can prevent this from happening by preventing it for every triple $i-1, i, i+1$. Hence, a "bad" opponent strategy would satisfy

$$\ell_{i-1}^\top p\le \ell_i^\top p \qquad\text{and}\qquad \ell_{i+1}^\top p\le \ell_i^\top p\,.$$

After rearranging, the above two inequalities yield the constraints

$$p_i\le\frac{c}{c+1}p_{i-1}$$

for every $i=2,\dots,N-1$. Note that there is no constraint on p_N. If we want to avoid by a margin these inequalities to be satisfied, we arrive at the constraints

$$p_i\ge\frac{c}{c+1}p_{i-1}+\rho$$

for some $\rho>0$, for every $i=2,\dots,N-1$.

In conclusion, we define the restricted opponent set to

$$\Delta' = \left\{ p \in \Delta_M \ : \ \forall i = 2, \ldots, N-2, \ p_i \geq \frac{c}{c+1} p_{i-1} + \rho \right\} .$$

The intuitive interpretation of this constraint is that the probability of the higher maximum price of the costumer should not decrease too fast. This constraint does not allow to have zero probabilities, and thus it is too restrictive.

Another way to construct a subset of Δ_M that is isolated from "dangerous" boundaries is to include only "hilly" distributions. We call a distribution $p \in \Delta_M$ hilly if it has a peak point $i^* \in \underline{N}$, and there exist $\xi_1, \ldots, \xi_{i^*-1} < 1$ and $\xi_{i^*+1}, \ldots, \xi_N < 1$ such that

$$\begin{aligned} p_{i-1} &\leq \xi_{i-1} p_i && \text{for } 2 \leq i \leq i^*, \text{ and} \\ p_{i+1} &\leq \xi_{i+1} p_i && \text{for } i^* \leq i \leq N-1. \end{aligned}$$

We now show that with the right choice of ξ_i, under a hilly distribution with peak i^*, only action i^* and maybe action i^*-1 can be optimal.

1. If $i \leq i^*$ then

$$\begin{aligned} (\ell_i - \ell_{i-1})^\top p &= c p_{i-1} - (p_i + \cdots + p_N) \\ &\leq c \xi_{i-1} p_i - p_i - (p_{i+1} + \cdots + p_N) , \end{aligned}$$

   thus, if $\xi_{i-1} \leq 1/c$ then the expected loss of action i is less than or equal to that of action $i-1$.

2. If $i \geq i^*$ then

$$\begin{aligned} (\ell_{i+1} - \ell_i)^\top p &= c p_i - (p_{i+1} + \cdots + p_N) \\ &\geq p_i \left\{ c - (\xi_{i+1} + \xi_{i+1}\xi_{i+2} + \cdots + \prod_{j=i+1}^{N} \xi_j) \right\} . \end{aligned}$$

   Now if we let $\xi_{i^*+1} = \cdots = \xi_N = \xi$ then we get

$$\begin{aligned} (\ell_{i+1} - \ell_i)^\top p &\geq p_i \left(c - \xi \frac{1 - \xi^{N-1}}{1 - \xi} \right) \\ &\geq p_i \left(c - \frac{\xi}{1 - \xi} \right) , \end{aligned}$$

   and thus if we choose $\xi \leq \frac{c}{c+1}$ then the expected loss of action i is less than or equal to that of action $i+1$.

   So far in all the calculations we allowed equalities. If we want to achieve that only action i^* and possibly action i^*-1 are optimal, we use

$$\xi_i \begin{cases} < 1/c, & \text{if } 2 \leq i \leq i^* - 2; \\ = 1/c, & \text{if } i = i^* - 1; \\ < c/(c+1), & \text{if } i^* + 1 \leq i \leq N. \end{cases}$$

If an opponent strategy is hilly with ξ_i satisfying all the above criteria, we call that strategy *sufficiently hilly.* Now we are ready to state the corollary of Theorem 5:

Corollary 4. *Consider the dynamic pricing game with N actions and M outcomes. If we restrict the set of opponent strategies Δ' to the set of all sufficiently hilly distributions then the minimax regret of the game is upper bounded by*

$$\mathbb{E}[R_T] \leq C\sqrt{T}$$

for some constant $C > 0$ that depends on the game $G = (L, H)$ and the choice of Δ'.

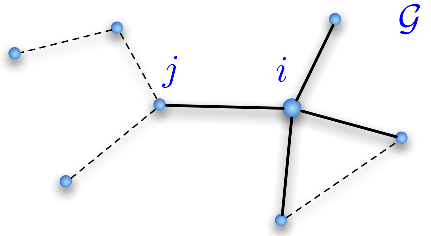


Figure 2: To each vertex i in the graph $\mathcal{G}$ we associate an algorithm $\mathcal{A}_i$. The algorithm plays an action from the distribution q_i^t over its neighborhood set N_i and receives partial information about relative loss between the node i and its neighbor. The other piece of the partial information comes from the times when a neighboring algorithm $\mathcal{A}_j$ is run and the action i is picked.

Remark 2. *Note that the number of actions and outcomes $N = M$ does not appear in the bound because the size of the largest point local game with the restricted strategy set is always* 2*, irrespectively of the number of actions.*

6 The adversarial case

Now we turn our attention to playing against adversarial opponents. We propose and analyze the algorithm NeigborhoodWatch. We show that the algorithm achieves $O(\sqrt{T})$ regret on locally observable games.

6.1 Method

The method is a two-level procedure motivated by Foster and Vohra [1997], and Blum and Mansour [2007]. The intuition stems from the following observation. Consider the graph whose vertices are the actions, and two vertices are connected with an edge if the corresponding actions are neighbors. Suppose for each vertex i we have a distribution $q_i \in \Delta_N$ supported on the neighbor set N_i. Let $p \in \Delta_N$ be defined by $p = Qp$ where Q is the matrix $[q_1, \ldots, q_N]$. Then there are two equivalent ways of sampling an action from p. The first way is to directly sample the vertex according to p. The second is to sample a vertex i according to p and then choose a vertex j within the neighbor set N_i according to q_i. Because of the stationarity (or *flow*) condition $p = Qp$, the two ways are equivalent. This idea of finding a fixed point is implicit in Foster and Vohra [1997], and Blum and Mansour [2007], who show how stationarity can be used to convert external regret guarantees into an internal regret statement.[7] We show here that, in fact, this conversion can be done "locally" and only with "comparison" information between neighboring actions.

Our procedure is as follows. We run N different algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_N$, each corresponding to a vertex and its neighbor set. Within this neighbor set we obtain small regret because we can construct estimates of loss differences among the actions, thanks to the local observability condition. Each algorithm $\mathcal{A}_i$ produces a distribution $q_i^t \in \Delta_N$ at round t, reflecting the relative performance of the vertex i and its neighbors. Since $\mathcal{A}_i$ is only concerned with its local neighborhood, we require that q_i^t has support on N_i and is zero everywhere else. The meta algorithm NeigborhoodWatch combines the distributions $Q^t = [q_1^t, \ldots, q_N^t]$ and computes p^t as a fixed point

$$p^t = Q^t p^t \ . \tag{7}$$

How do we choose our actions? At each round, we draw $K_t \sim p_t$ and then $I_t \sim q_{K_t}^t$ according to our two-level scheme. The action I_t is the action we play in the partial monitoring game against the adversary.

[7] For the definition of internal regret, see the next section. The external regret is just the regret, the word "external" is used as "not internal".

Algorithm 2 NeigborhoodWatch Algorithm

```
For all $i = \{1, \dots, N\}$, initialize algorithm $\mathcal{A}_i$ with $q_i^1 = x_i^1 = \mathbf{1}_{N_i}/|N_i|$
for t=1,..., T do
    Let $Q^t = [q_1^t, \dots, q_N^t]$, where $q_i^t$ is furnished by $\mathcal{A}_i$
    Find $p^t$ satisfying $p^t = Q^t p^t$
    Draw $k_t$ from $p^t$
    Play $I_t$ drawn from $q_{k_t}^t$ and obtain signal $S_{I_t} e_{j_t}$
    Run local algorithm $\mathcal{A}_{k_t}$ with the received signal
    For any $i \neq k_t$, $q_i^{t+1} \leftarrow q_i^t$
end for
```

Algorithm 3 Local Algorithm $\mathcal{A}_i$

```
If $t = 1$, initialize $s = 1$
For $r \in \{\tau_i(s-1)+1, \dots, \tau_i(s)\}$ (i.e. for all $r$ since the last time $\mathcal{A}_i$ was run) construct
    $$b_{(i,j)}^r = v_{i,j}^\intercal \begin{bmatrix} \mathbb{I}\{I_r = i\} S_i \\ \mathbb{I}\{k_r = i\} \mathbb{I}\{I_r = j\} S_j / q_i^r(j) \end{bmatrix} e_{j_r}$$
    for all $j \in N_i$
Define for all $j \in N_i$,
    $$h_{(i,j)}^s = \sum_{r=\tau_i(s-1)+1}^{\tau_i(s)} b_{(i,j)}^r$$
    and let
    $$\tilde{f}_i^s = \left[ h_{(i,j)}^s \cdot \mathbb{I}\{j \in N_i\} \right]_{j \in [N]}$$
Pass the cost $\tilde{f}_i^s$ to a full-information online convex optimization algorithm over the simplex (e.g. Exponential Weights Algorithm) and receive the next distribution $x^{s+1}$ supported on $N_i$
Define
    $$q_i^{t+1} \leftarrow (1-\gamma) x^{s+1} + (\gamma/|N_i|) 1_{N_i}$$
Increase the count $s \leftarrow s + 1$
```

Let the action played by the adversary at time t be denoted by J_t. Then the feedback we obtain is $S_{I_t} e_{J_t}$. This information is passed to $\mathcal{A}_{K_t}$ which updates the distributions $q_{K_t}^t$. In Section 6.2.2 we detail how this is done.

The advantage of the above two-level method is that while the actions are still chosen with respect to the distribution q^t, the loss difference estimations are only needed locally. The local observability condition ensures that these local estimations can be done without using "non-local" actions.

6.2 Analysis of NeigborhoodWatch

Before presenting the main result of this section, we need the concept of *local internal regret.*

6.2.1 Local internal regret

Let $\phi : \{1, \dots, N\} \mapsto \{1, \dots, N\}$ be a *departure function* [Cesa-Bianchi et al., 2006], and let I_t and J_t denote the moves at time t of the player and the opponent, respectively. At the end of the game, regret with respect to ϕ is calculated as the difference of the incurred cumulative cost and the cost that would have been incurred had we played action $\phi(I_t)$ instead of I_t, for all t. Let Φ be a set of departure functions. The Φ-regret is defined as

$$\frac{1}{T}\sum_{t=1}^{T} c(I_t, J_t) - \inf_{\phi\in\Phi}\frac{1}{T}\sum_{t=1}^{T} c(\phi(I_t), J_t)$$

where the cost function considered in this paper is simply $c(i,j) = L_{i,j}$. If $\Phi = \{\phi_k : k \in [N]\}$ consists of constant mappings $\phi_k(i) = k$, the regret is called *external*, or just simply regret: this definition is equivalent to the regret definition in the introduction. For (global) internal regret, the set Φ consists of all departure functions $\phi_{i\to j}$ such that $\phi_{i\to j}(i) = j$ and $\phi_{i\to j}(h) = h$ for $h \neq i$.

Definition 9. *For a game G, let the graph $\mathcal{G}$ be its* neighborhood graph*: its vertices are the actions of the game, and two vertices are connected with an edge if the corresponding actions are neighbors. A departure function $\phi_{i\to j}$ is called* local *if j is a neighbor of i in the neighborhood graph $\mathcal{G}$. Let Φ_L be the set of all local departure functions. The Φ_L-regret defined with respect to the set of all local departure functions is called* local internal regret*.*

The main result of the paper is the following internal regret guarantee.

Theorem 6. *The local internal regret of Algorithm 2 is bounded as*

$$\sup_{\phi\in\Phi_L}\mathbb{E}\left\{\sum_{t=1}^{T}(e_{I_t} - e_{\phi(I_t)})^\intercal L e_{j_t}\right\} \le 4N\bar{v}\sqrt{6(\log N)T}$$

where $\bar{v} = \max_{(i,j)}\|v_{(i,j)}\|_\infty$.

To prove that the same bound holds for the external regret we need two observations. The fist observation is that the local internal regret is equal to the the internal regret:

Lemma 6. *There exists a problem dependent constant K such that the internal regret is at most K times the local internal regret.*

The second (well-known) observation is that the internal regret is always greater than or equal to the external regret.

Corollary 5. *External regret of Algorithm 2 is bounded as*

$$\mathbb{E}\{R_T\} \le 4KN\bar{v}\sqrt{6(\log N)T}$$

where K is the upper bound from Lemma 6.

We remark that high probability bounds can also be obtained in a rather straightforward manner, using, for instance, the approach of Abernethy and Rakhlin [2009]. Another extension, the case of random signals, is discussed in Section 6.3.

The rest of this section is devoted to prove Theorem 6.

6.2.2 Estimating loss differences

The random variable k_t drawn from p^t at time t determines which algorithm is active on the given round. Let

$$\tau_i(s) = \min\{t \ : \ s = \sum_{r=1}^{t}\mathbb{I}\{k_t = i\}\}$$

denote the (random) time when the algorithm $\mathcal{A}_i$ is invoked for the s-th time. By convention, $\tau_i(0) = 0$. Further, define

$$\pi_i(t) = \min\{t' \ge t \ : \ k_{t'} = i\}$$

to denote the next time the algorithm is run on or after time t. When invoked for the s-th time, the algorithm $\mathcal{A}_i$ constructs estimates

$$b^r_{(i,j)} \triangleq v_{i,j}^\intercal \left[\begin{array}{c} \mathbb{I}\left\{I_r = i\right\} S_i \\ \mathbb{I}\left\{k_r = i\right\} \mathbb{I}\left\{I_r = j\right\} S_j / q_i^r(j) \end{array} \right] e_{j_r} \qquad (r \in \{\tau_i(s-1)+1, \ldots, \tau_i(s)\},\ j \in N_i)$$

for all the rounds after it has been run the last time, until (and including) the current time $r = \tau_i(s)$. We can assume $b^t_{(i,j)} = 0$ for any $j \notin N_i$. The estimates $b^t_{(i,j)}$ can be constructed by the algorithm because $S_{I_r} e_{j_r}$ is precisely the feedback given to the algorithm.

Let $\mathcal{F}_t$ be the σ-algebra generated by the random variables $\{k_1, I_1, \ldots, k_t, I_t\}$. For any t, the (conditional) expectation satisfies

$$\begin{aligned}
\mathbb{E}\left[b^t_{(i,j)} | \mathcal{F}_{t-1}\right] &= \sum_{k=1}^N p_k^t q_k^t(i) v_{i,j}^\intercal \left[\begin{array}{c} S_i \\ 0 \end{array} \right] e_{j_t} + p_i^t q_i^t(j) v_{i,j}^\intercal \left[\begin{array}{c} 0 \\ S_j / q_i^t(j)) \end{array} \right] e_{j_t} \\
&= p_i^t v_{i,j}^\intercal S_{(i,j)} e_{j_t} \\
&= p_i^t (\ell_j - \ell_i)^\intercal e_{j_t} \\
&= p_i^t (e_j - e_i)^\intercal L e_{j_t} \qquad (8)
\end{aligned}$$

where in the second equality we used the fact that $\sum_{k=1}^N p_k^t q_k^t(i) = p_i^t$ by stationarity (7). Thus each algorithm $\mathcal{A}_i$, on average, has access to unbiased estimates of the loss differences within its neighborhood set.

Recall that algorithm $\mathcal{A}_i$ is only aware of its neighborhood, and therefore we peg coordinates of q_i^t to zero outside of N_i. However, for convenience, our notation below still employs full N-dimensional vectors, and we keep in mind that only coordinates indexed by N_i are considered and modified by $\mathcal{A}_i$.

When invoked for the s-th time (that is, $t = \tau_i(s)$), $\mathcal{A}_i$ constructs linear functions (cost estimates) $\tilde{f}_i^s \in \mathbb{R}^N$ defined by

$$\tilde{f}_i^s = \left[h^s_{(i,j)} \cdot \mathbb{I}\left\{j \in N_i\right\} \right]_{j \in [N]} ,$$

where

$$h^s_{(i,j)} = \sum_{r = \tau_i(s-1)+1}^{\tau_i(s)} b^r_{(i,j)} .$$

We now show that $\tilde{f}_i^s \cdot q_i^{\tau(s)}$ has the same conditional expectation as the actual loss of the meta algorithm NeigborhoodWatch at time $t = \tau_i(s)$. That is, by bounding expected regret of the black-box algorithm operating on $\{\tilde{f}_i^s\}$, we bound the actual regret suffered by the meta algorithm on the rounds when $\mathcal{A}_i$ was invoked.

Lemma 7. *Consider algorithm $\mathcal{A}_i$. It holds that*

$$\mathbb{E}\left\{ (q_i^{\tau_i(s+1)} - e_u)^\intercal L e_{j_{\tau_i(s+1)}} \;\middle|\; \mathcal{F}_{\tau_i(s)} \right\} = \mathbb{E}\left\{ \tilde{f}_i^{s+1} \cdot (q_i^{\tau_i(s+1)} - e_u) \;\middle|\; \mathcal{F}_{\tau_i(s)} \right\}$$

for any $u \in N_i$.

Proof. Throughout the proof, we drop the subscript i on τ_i to ease the notation. Note that $q_i^{\tau(s+1)} = q_i^{\tau(s)+1}$ since the distribution is not updated when algorithm $\mathcal{A}_i$ is not invoked. Hence, conditioned on $\mathcal{F}_{\tau(s)}$, the variable $(q_i^{\tau(s+1)} - e_u)$ can be taken out of the expectation. We therefore need to show that

$$(q_i^{\tau(s+1)} - e_u) \cdot \mathbb{E}\left\{ L e_{j_{\tau(s+1)}} | \mathcal{F}_{\tau(s)} \right\} = (q_i^{\tau(s+1)} - e_u) \cdot \mathbb{E}\left\{ \tilde{f}_i^{s+1} | \mathcal{F}_{\tau(s)} \right\} \qquad (9)$$

First, we can write

$$\begin{aligned}
\mathbb{E}\left\{h^{s+1}_{(i,j)} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} &= \mathbb{E}\left\{\sum_{t=\tau(s)+1}^{\tau(s+1)} b^t_{(i,j)} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \mathbb{E}\left\{\sum_{t=\tau(s)+1}^{\infty} b^t_{(i,j)} \mathbb{I}\left\{t \le \tau(s+1)\right\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{\mathbb{E}\left[b^t_{(i,j)} \mathbb{I}\left\{t \le \tau(s+1)\right\} \;\middle|\; \mathcal{F}_{t-1}\right] \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{\mathbb{I}\left\{t \le \tau(s+1)\right\} \mathbb{E}\left[b^t_{(i,j)} \;\middle|\; \mathcal{F}_{t-1}\right] \;\middle|\; \mathcal{F}_{\tau(s)}\right\} .
\end{aligned}$$

The last step follows because the event $\{t \le \tau(s+1)\}$ is $\mathcal{F}_{t-1}$-measurable (that is, variables $k_1, \ldots, k_{t-1}$ determine the value of the indicator). By (8), we conclude

$$\mathbb{E}\left\{h^{s+1}_{(i,j)} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} = \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{\mathbb{I}\left\{t \le \tau(s+1)\right\} p^t_i (e_j - e_i)^\intercal L e_{j_t} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} . \tag{10}$$

Since $\mathbb{I}\left\{t = \tau(s+1)\right\} = \mathbb{I}\left\{k_t = i\right\} \mathbb{I}\left\{t \le \tau(s+1)\right\}$, we have

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{I}\left\{t = \tau(s+1)\right\} e_{j_t} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\mathbb{I}\left\{k_t = i\right\} \mathbb{I}\left\{t \le \tau(s+1)\right\} e_{j_t} \;\middle|\; \mathcal{F}_{t-1}\right\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \mathbb{E}\left\{\mathbb{I}\left\{t \le \tau(s+1)\right\} e_{j_t} \mathbb{E}\left\{\mathbb{I}\left\{k_t = i\right\} \;\middle|\; \mathcal{F}_{t-1}\right\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \mathbb{E}\left\{\mathbb{I}\left\{t \le \tau(s+1)\right\} \mathcal{P}(k_t = i \mid \mathcal{F}_{t-1}) e_{j_t} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \mathbb{E}\left\{\mathbb{I}\left\{t \le \tau(s+1)\right\} p^t_i e_{j_t} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} .
\end{aligned}$$

Combining with (10),

$$\begin{aligned}
\mathbb{E}\left\{h^{s+1}_{(i,j)} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} &= \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{\mathbb{I}\left\{t \le \tau(s+1)\right\} p^t_i (e_j - e_i)^\intercal L e_{j_t} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{\mathbb{I}\left\{t = \tau(s+1)\right\} (e_j - e_i)^\intercal L e_{j_t} \;\middle|\; \mathcal{F}_{\tau(s)}\right\}
\end{aligned}$$

Observe that coordinates of $\tilde{f}^{s+1}_i$, $q^{\tau(s+1)}_i$, and e_u are zero outside of N_i. We then have that

$$\begin{aligned}
\mathbb{E}\left\{\tilde{f}^{s+1}_i \;\middle|\; \mathcal{F}_{\tau(s)}\right\} &= \left[\mathbb{I}\left\{j \in N_i\right\} \mathbb{E}\left\{h^{s+1}_{(i,j)} \;\middle|\; \mathcal{F}_{\tau(s)}\right\}\right]_{j \in N} \\
&= \left[\mathbb{I}\left\{j \in N_i\right\} \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{(e_j - e_i)^\intercal L e_{j_t} \mathbb{I}\left\{t = \tau(s+1)\right\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\}\right]_{j \in N} \\
&= \left[\mathbb{I}\left\{j \in N_i\right\} \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{e_j L e_{j_t} \mathbb{I}\left\{t = \tau(s+1)\right\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\}\right]_{j \in N} - c \cdot 1_{N_i}
\end{aligned}$$

where

$$c = \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{e_i L e_{j_t} \mathbb{I}\left\{t = \tau(s+1)\right\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\}$$

is a scalar. When multiplying the above expression by $q_i^{\tau(s+1)} - e_u$, the term $c \cdot 1_{N_i}$ vanishes. Thus, minimizing regret with relative costs (with respect to the ith action) is the same as minimizing regret with the absolute costs. We conclude that

$$\begin{aligned}
(q_i^{\tau(s+1)} - e_u)\mathbb{E}\left\{\tilde{f}_i^{s+1} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} &= (q_i^{\tau(s+1)} - e_u) \cdot \left[\sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{e_j L e_{j_t} \mathbb{I}\{t = \tau(s+1)\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\}\right]_{j \in N_i} \\
&= (q_i^{\tau(s+1)} - e_u) \cdot \sum_{t=\tau(s)+1}^{\infty} \mathbb{E}\left\{L e_{j_t} \mathbb{I}\{t = \tau(s+1)\} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} \\
&= (q_i^{\tau(s+1)} - e_u) \cdot \mathbb{E}\left\{L e_{j_{\tau(s+1)}} \;\middle|\; \mathcal{F}_{\tau(s)}\right\} . \qquad \square
\end{aligned}$$

6.2.3 Regret Analysis

For each algorithm $\mathcal{A}_i$, the estimates $\tilde{f}_i^s$ are passed to a full-information black box algorithm which works only on the coordinates N_i. From the point of view of the full-information black box, the game has length $T_i = \max\{s : \tau_i(s) \le T\}$, the (random) number of times action i has been played within T rounds.

We proceed similarly to Abernethy and Rakhlin [2009]: we use a full-information online convex optimization procedure with an entropy regularizer (also known as the Exponential Weights Algorithm) which receives the vector $\tilde{f}_i^s$ and returns the next mixed strategy $x^{s+1} \in \Delta_N$ (in fact, effectively in $\Delta_{|N_i|}$). We then define

$$q_i^{t+1} = (1-\gamma)x^{s+1} + (\gamma/|N_i|)1_{N_i}$$

where γ is to be specified later. Since $\mathcal{A}_i$ is run at time t, we have $\tau_i(s) = t$ by definition. The next time $\mathcal{A}_i$ is active (that is, at time $\tau_i(s+1)$), the action $I_{\tau_i(s+1)}$ will be played as a random draw from $q_i^{t+1} = q_i^{\tau_i(s+1)}$; that is, the distribution is not modified on the interval $\{\tau_i(s)+1, \ldots, \tau_i(s+1)\}$.

We prove Theorem 6 by a series of lemmas. The first one is a direct consequence of an external regret bound for a Follow the Regularized Leader (FTRL) algorithm in terms of local norms [Abernethy and Rakhlin, 2009]. For a strictly convex "regularizer" F, the local norm $\|\cdot\|_x$ is defined by $\|z\|_x = \sqrt{z^\intercal \nabla^2 F(x) z}$ and its dual is $\|z\|_x^* = \sqrt{z^\intercal \nabla^2 F(x)^{-1} z}$.

Lemma 8. *The full-information algorithm utilized by $\mathcal{A}_i$ has an upper bound*

$$\mathbb{E}\left\{\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - e_{\phi(i)})\right\} \le \eta \mathbb{E}\left\{\sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2\right\} + \eta^{-1} \log N + T\gamma\bar{\ell}$$

on its external regret, where $\phi(i) \in N_i$ is any neighbor of i, $\bar{\ell} = \max_{i,j} L_{i,j}$, and η is a learning rate parameter to be tuned later.

Proof. Since our decision space is a simplex, it is natural to use the (negative) entropy regularizer, in which case FTRL is the same as the Exponential Weights Algorithm. From Abernethy and Rakhlin [2009, Thm 2.1], for any comparator u with zero support outside $|N_i|$, the following regret guarantee holds:

$$\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (x^s - u) \le \eta \sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2 + \eta^{-1} \log(|N_i|) .$$

An easy calculation shows that in the case of entropy regularizer F, the Hessian $\nabla^2 F(x) = \mathrm{diag}(x_1^{-1}, x_2^{-1}, \ldots, x_N^{-1})$ and $\nabla^2 F(x)^{-1} = \mathrm{diag}(x_1, x_2, \ldots, x_N)$. We refer to Abernethy and Rakhlin [2009] for more details.

Let $\phi : \{1, \ldots, N\} \mapsto \{1, \ldots, N\}$ be a local departure function (see Definition 9). We can then write a regret guarantee

$$\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (x^s - e_{\phi(i)}) \le \eta \sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2 + \eta^{-1} \log(|N_i|) .$$

Since, in fact, we play according to a slightly modified version $q_i^{\tau_i(s)}$ of x^s, it holds that

$$\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - e_{\phi(i)}) \leq \eta \sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2 + \eta^{-1} \log(|N_i|) + \sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - x^s) \ .$$

Taking expectations of both sides and upper bounding $|N_i|$ by N, we get

$$\mathbb{E}\left\{\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - e_{\phi(i)})\right\} \leq \eta \mathbb{E}\left\{\sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2\right\} + \eta^{-1} \log N + \mathbb{E}\left\{\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - x^s)\right\} \ .$$

A proof identical to that of Lemma 7 gives

$$\begin{aligned} \mathbb{E}\left\{\tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - x^s) \;\middle|\; \mathcal{F}_{\tau_i(s-1)}\right\} &= \mathbb{E}\left\{(q_i^{\tau_i(s)} - x^s)^\intercal L e_{j_{\tau_i(s)}} | \mathcal{F}_{\tau_i(s-1)}\right\} \\ &\leq \mathbb{E}\left\{\|q_i^{\tau_i(s)} - x^s\|_1 \cdot \|L e_{j_{\tau_i(s)}}\|_\infty \;\middle|\; \mathcal{F}_{\tau_i(s-1)}\right\} \\ &\leq \gamma \bar{\ell} \end{aligned}$$

for the last term, where $\bar{\ell}$ is the upper bound on the magnitude of entries of L. Putting everything together,

$$\mathbb{E}\left\{\sum_{s=1}^{T_i} \tilde{f}_i^s \cdot (q_i^{\tau_i(s)} - e_{\phi(i)})\right\} \leq \eta \mathbb{E}\left\{\sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2\right\} + \eta^{-1} \log N + T\gamma\bar{\ell}$$

where we have upper bounded T_i by T. □

As with many bandit-type problems, effort is required to show that the variance term is controlled. This is the subject of the next lemma.

Lemma 9. *The variance term in the bound of Lemma 8 is upper bounded as*

$$\sum_{i=1}^{N} \mathbb{E}\left\{\sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2\right\} \leq 24 \bar{v}^2 NT \,.$$

Proof. First, fix an $i \in [N]$ and consider the term $\mathbb{E}\left\{\sum_{s=1}^{T_i} (\|\tilde{f}_i^s\|_{x^s}^*)^2\right\}$. Until the last step of the proof, we will sometimes omit i from the notation.

We start by observing that $\tilde{f}_i^s$ is a sum of $\tau(s) - \tau(s-1) - 1$ terms of the type $v_{i,j}^\intercal S_i e_{j_r}$ (that is, of constant magnitude) and one term of the type $v_{i,j}^\intercal S_j e_{j_r} / q_i^r(j)$. In controlling $\|\tilde{f}_i^s\|_{x^s}^*$, we therefore have two difficulties: controlling the number of constant-size terms and making sure the last term does not explode due to division by a small probability $q_i^r(j)$. The former is solved below by a careful argument below, while the latter problem is solved according to usual bandit-style arguments.

More precisely, we can write $\tilde{f}_i^s = g_{\tau_i(s)}^{\tau_i(s-1)} + h^{\tau_i(s)}$ where the vectors $g_{\tau_i(s)}^{\tau_i(s-1)}, h^{\tau_i(s)} \in \mathbb{R}^N$ are defined as

$$g_{\tau_i(s)}^{\tau_i(s-1)}(j) \triangleq g^{\tau_i(s-1)}(j) \triangleq \sum_{r=\tau_i(s-1)}^{\tau_i(s)-1} \mathbb{I}\left\{I_r = i\right\} v_{i,j}^\intercal S_i e_{j_r} \mathbb{I}\left\{j \in N_i\right\}$$

and

$$h^{\tau_i(s)}(j) = \mathbb{I}\left\{I_{\tau_i(s)} = j\right\} v_{i,I_{\tau_i(s)}}^\intercal S_{I_{\tau_i(s)}} e_{j_{\tau_i(s)}} / q_i^{\tau_i(s)}(I_{\tau_i(s)}) \ .$$

Then

$$(\|\tilde{f}_i^s\|_{x^s}^*)^2 = (\|g^{\tau_i(s-1)} + h^{\tau_i(s)}\|_{x^s}^*)^2 \leq 2(\|g^{\tau_i(s-1)}\|_{x^s}^*)^2 + 2(\|h^{\tau_i(s)}\|_{x^s}^*)^2$$

We will bound each of the two terms separately, in expectation. For the second term,

$$(\|h^{\tau_i(s)}\|^*_{x^s})^2 = x^s(I_\tau)(v^\intercal_{i,I_\tau} S_{I_\tau} e_{j_\tau}/q_i^\tau(I_\tau))^2 \le x^s(I_\tau)(\bar{v}/q_i^\tau(I_\tau))^2$$

where $\tau = \tau_i(s)$. Since $q_i^{\tau_i(s)} = (1-\gamma)x^s + (\gamma/|N_i|)1_{N_i}$, it is easy to verify that $x^s(I_\tau)/q_i^\tau(I_\tau) \le 2$ (whenever $\gamma < 1/2$) and thus

$$(\|h^{\tau_i(s)}\|^*_{x^s})^2 \le 2\bar{v}^2/q_i^\tau(I_\tau) \ .$$

The remaining division by the probability disappears under the expectation:

$$\mathbb{E}\left\{(\|h^{\tau_i(s)}\|^*_{x^s})^2 \;\middle|\; \sigma(k_1, I_1, \dots, k_{\tau_i(s)})\right\} \le 2\bar{v}^2 \sum_{j=1}^{N} q_i^{\tau_i(s)}(j)/q_i^{\tau_i(s)}(j) = 2N\bar{v}^2 \ . \tag{11}$$

Consider now the second term. As discussed in the proof of Lemma 8, the inverse Hessian of the entropy function shrinks each coordinate i precisely by $x^s(i) \le 1$, implying that the local norm is dominated by the Euclidean norm :

$$\|g^{\tau_i(s-1)}\|^*_{x^s} \le \|g^{\tau_i(s-1)}\|_2.$$

It is therefore enough to upper bound $\mathbb{E}\left\{\sum_{s=1}^{T_i} \|g^{\tau_i(s)}\|_2^2\right\}$. The idea of the proof is the following. Observe that $P(k_t = i|\mathcal{F}_{t-1}) = P(I_t = i|\mathcal{F}_{t-1})$. Conditioned on the event that either $k_t = i$ or $I_t = i$, each of the two possibilities has probability 1/2 of occurring. Note that $g^{\tau_i(s-1)}$ inflates every time $k_t \neq i$, yet $I_t = i$ occurs. It is then easy to see that magnitude of $g^{\tau_i(s-1)}$ is unlikely to get large before algorithm $\mathcal{A}_i$ is run again. We now make this intuition precise.

The function g^t is presently defined only for those time steps when $t = \tau_i(s)$ for some s (that is, when the algorithm $\mathcal{A}_i$ is invoked). We extend this definition as follows. Let the jth coordinate of g^t be defined as

$$g^t_{\pi(t+1)}(j) \triangleq g^t(j) \triangleq \sum_{r=t}^{\pi(t+1)-1} \mathbb{I}\{I_r = i\}\, v_{(i,j)} S_i e_{j_r}$$

for $j \in N_i$ and 0 otherwise. The function g^t can be thought of as accumulating partial pieces on rounds when $I_t = i$ until $k_t = i$ occurs. Let us now define an analogue of τ and π for the event that *either* $I_t = i$ or $k_t = i$:

$$\gamma_i(s) = \min\left\{t \ : \ s = \sum_{r=1}^{t} \mathbb{I}\{k_t = i \text{ or } I_t = i\}\right\}$$

Further, for any t, let

$$\nu_i(t) = \min\{t' \ge t : k_t = i \text{ or } I_t = i\},$$

the next time occurrence of the event $\{k_\tau = i \text{ or } I_\tau = i\}$ on or after t. Let

$$\mathcal{I} = \mathbb{I}\{\nu_i(t) \neq \pi_i(t)\}$$

be the indicator of the event that the first time after t that $\{k_\tau = i \text{ or } I_\tau = i\}$ occurred it was also the case that the algorithm was not run (i.e. $k_\tau \neq i$). Note that $g^t(j)$ can now be written recursively as

$$g^t(j) = \mathcal{I} \cdot \left[v_{(i,j)} S_i e_{j_{\nu(t)}} + g^{\nu(t)+1}_{\pi(\nu(t)+1)}(j)\right].$$

As argued before, $\mathcal{P}(\mathcal{I} = 1|\mathcal{F}_{t-1}) = 1/2$. We will now show that $\mathbb{E}\{g^t(j) \mid \mathcal{F}_{t-1}\} \le 2\bar{v}$ by the following inductive argument, whose base case trivially holds for $t = T$:

$$\begin{aligned}
\mathbb{E}\left\{g^t(j) \mid \mathcal{F}_{t-1}\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\mathcal{I}\cdot\left[v_{(i,j)}S_i e_{j_{\nu(t)}} + g^{\nu(t)+1}(j)\right] \,\Big|\, \mathcal{F}_{\nu(t)}\right\} \,\Big|\, \mathcal{F}_{t-1}\right\} \\
&= \mathbb{E}\left\{\mathcal{I}v_{(i,j)}S_i e_{j_{\nu(t)}} + \mathcal{I}\mathbb{E}\left\{g^{\nu(t)+1}(j) \,\Big|\, \mathcal{F}_{\nu(t)}\right\} \,\Big|\, \mathcal{F}_{t-1}\right\} \\
&\le \bar{v} + \mathbb{E}\left\{\mathcal{I}g^{\nu(t)+1}(j) \,\Big|\, \mathcal{F}_{t-1}\right\} \\
&= \bar{v} + \mathbb{E}\Big\{\mathcal{I}\underbrace{\mathbb{E}\left[g^{\nu(t)+1}(j) \,\Big|\, \mathcal{F}_{\nu(t)}\right]}_{\le 2\bar{v} \text{ by induction}} \,\Big|\, \mathcal{F}_{t-1}\Big\} \\
&\le \bar{v} + \mathbb{E}\left\{\mathcal{I} \mid \mathcal{F}_{t-1}\right\}2\bar{v} \\
&\le \bar{v} + (1/2)2\bar{v} = 2\bar{v}
\end{aligned}$$

The expected value of $(g^t(j))^2$ can be controlled in a similar manner. To ease the notation, let $z = v_{(i,j)}S_i e_{j_{\nu(t)}}$. Using the upper bound for the conditional expectation of $g^t(j)$ calculated above,

$$\begin{aligned}
\mathbb{E}\left\{(g^t(j))^2 \mid \mathcal{F}_{t-1}\right\} &= \mathbb{E}\left\{\mathcal{I}\cdot\left(z^2 + (g^{\nu(t)+1}(j))^2 + 2zg^{\nu(t)+1}(j)\right) \,\Big|\, \mathcal{F}_{t-1}\right\} \\
&= \mathbb{E}\left\{\mathcal{I}z^2 + \mathcal{I}\mathbb{E}\left\{(g^{\nu(t)+1}(j))^2 \,\Big|\, \mathcal{F}_{\nu(t)}\right\} + 2\mathcal{I}z\mathbb{E}\left\{g^{\nu(t)+1}(j) \,\Big|\, \mathcal{F}_{\nu(t)}\right\} \,\Big|\, \mathcal{F}_{t-1}\right\} \\
&\le 5\bar{v}^2 + \mathbb{E}\left\{\mathcal{I}\mathbb{E}\left\{(g^{\nu(t)+1}(j))^2 \,\Big|\, \mathcal{F}_{\nu(t)}\right\} \,\Big|\, \mathcal{F}_{t-1}\right\}
\end{aligned}$$

The argument now proceeds with backward induction exactly as above. We conclude that

$$\mathbb{E}\left\{(g^t(j))^2 \mid \mathcal{F}_{t-1}\right\} \le 10\bar{v}^2$$

and, hence,

$$\mathbb{E}\left\{\|g^{\tau_i(s-1)}\|_2^2\right\} \le 10N\bar{v}^2$$

Together with (11), we conclude that

$$\mathbb{E}\left\{(\|\tilde{f}_i^s\|_{x^s}^*)^2\right\} \le 2(2N\bar{v}^2 + 10N\bar{v}^2) = 24\bar{v}^2 N.$$

Summing over $t = 1, \ldots, T$ and observing that only one algorithm is run at any time t proves the statement. □

Proof of Theorem 6. The flow condition $p^t = Q^t p^t$ comes in crucially in several places throughout the proofs, and the next argument is one of them. Observe that

$$\mathbb{E}\left\{e_{\phi(I_t)} \,\middle|\mathcal{F}_{t-1}\right\} = \sum_{k=1}^N\sum_{i=1}^N p_k^t q_k^t(i)e_{\phi(i)} = \sum_{i=1}^N e_{\phi(i)}\sum_{k=1}^N p_k^t q_k^t(i) = \sum_{i=1}^N e_{\phi(i)}p_i^t = \mathbb{E}\left\{e_{\phi(k_t)} \,\middle|\mathcal{F}_{t-1}\right\}$$

and thus

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^T e_{\phi(I_t)}^\intercal L e_{j_t}\right\} &= \mathbb{E}\left\{\sum_{t=1}^T \mathbb{E}\left\{e_{\phi(I_t)} \,\middle|\mathcal{F}_{t-1}\right\}^\intercal L e_{j_t}\right\} \\
&= \mathbb{E}\left\{\sum_{t=1}^T \mathbb{E}\left\{e_{\phi(k_t)} \,\middle|\mathcal{F}_{t-1}\right\}^\intercal L e_{j_t}\right\} \\
&= \mathbb{E}\left\{\sum_{t=1}^T e_{\phi(k_t)}^\intercal L e_{j_t}\right\}
\end{aligned}$$

It is because of this equality that external regret with respect to the local neighborhood can be turned into local internal regret. We have that

$$\begin{aligned}\mathbb{E}\left\{\sum_{t=1}^{T}(e_{I_t}-e_{\phi(I_t)})^{\intercal}Le_{j_t}\right\} &= \mathbb{E}\left\{\sum_{t=1}^{T}(e_{I_t}-e_{\phi(k_t)})^{\intercal}Le_{j_t}\right\}\\ &= \mathbb{E}\left\{\sum_{t=1}^{T}(q^t_{k_t}-e_{\phi(k_t)})^{\intercal}Le_{j_t}\right\}\\ &= \sum_{i=1}^{N}\mathbb{E}\left\{\sum_{t=1}^{T}\mathbb{I}\left\{k_t=i\right\}(q^t_i-e_{\phi(i)})^{\intercal}Le_{j_t}\right\}\end{aligned}$$

By Lemma 7,

$$\mathbb{E}\left\{(q_i^{\tau_i(s)}-e_{\phi(i)})^{\intercal}Le_{j_{\tau_i(s)}}|\mathcal{F}_{\tau_i(s-1)}\right\}=\mathbb{E}\left\{\tilde{f}_i^s\cdot(q_i^{\tau_i(s)}-e_{\phi(i)})\;\middle|\;\mathcal{F}_{\tau_i(s-1)}\right\}$$

and so by Lemma 8

$$\begin{aligned}E\left\{\sum_{t=1}^{T}(e_{I_t}-e_{\phi(I_t)})^{\intercal}Le_{j_t}\right\} &= \sum_{i=1}^{N}\mathbb{E}\left\{\sum_{s=1}^{T_i}\tilde{f}_i^s\cdot(q_i^{\tau_i(s)}-e_{\phi(i)})\right\}\\ &\le \eta\sum_{i=1}^{N}\mathbb{E}\left\{\sum_{s=1}^{T_i}(\|\tilde{f}_i^s\|^*_{x^s})^2\right\}+N(\eta^{-1}\log N+T\gamma\bar{\ell})\end{aligned}$$

With the help of Lemma 9,

$$\mathbb{E}\left\{\sum_{t=1}^{T}(e_{I_t}-e_{\phi(I_t)})^{\intercal}Le_{j_t}\right\}\le\eta 24\bar{v}^2NT+N(\eta^{-1}\log N+T\gamma\bar{\ell})=4N\bar{v}\sqrt{6(\log N)T}+TN\gamma\bar{\ell}$$

with $\eta=\sqrt{\frac{\log N}{24\bar{v}^2T}}$.

We remark that for the purposes of "in expectation" bounds, we can simply set $\gamma=0$ and still get $O(\sqrt{T})$ guarantees (see Abernethy and Rakhlin [2009]). This point is obscured by the fact that the original algorithm of Auer et al. [2003] uses the same parameter for the learning rate η and exploration γ. If these are separated, the "in expectation" analysis of Auer et al. [2003] can be also done with $\gamma=0$. However, to prove high probability bounds on regret, a setting of $\gamma\propto T^{-1/2}$ is required. Using the techniques in Abernethy and Rakhlin [2009], the high-probability extension of results in this paper is straightforward (tails for the terms $\|g^{\tau_i(s-1)}\|_2^2$ in Lemma 9 can be controlled without much difficulty). □

6.3 Random Signals

We now briefly consider the setting of partial monitoring with random signals, studied by Rustichini [1999], Lugosi, Mannor, and Stoltz [2008], and Perchet [2011]. Without much modification of the above arguments, the local observability condition yet again yields $O(\sqrt{T})$ internal regret.

Suppose that instead of receiving deterministic feedback $H_{i,j}$, the decision maker now receives a random signal $d_{i,j}$ drawn according to the distribution $H_{i,j}\in\Delta(\Sigma)$ over the signals. In the problem of deterministic feedback studied in the paper so far, the signal $H_{i,j}=\sigma$ was identified with the Dirac distribution δ_σ.

Given the matrix H of distributions on Σ, we can construct, for each row i, a matrix $\Xi_i\in\mathbb{R}^{s_i\times M}$ as

$$\Xi_i(k,j)\triangleq H_{i,j}(\sigma_k)$$

where the set $\sigma_1,\ldots,\sigma_{s_i}$ is the union of supports of $H_{i,1},\ldots,H_{i,M}$. Columns of Ξ_i are now distributions over signals. Given the actions I_t and j_t of the player and the opponent, the feedback provided to the player can be equivalently written as $S^t_{I_t} e_{j_t}$ where each column r of the random matrix $S^t_{I_t} \in \mathbb{R}^{s_i \times M}$ is a standard unit vector drawn independently according to the distribution given by the column r of Ξ_i. Hence, $\mathbb{E} S^t_i = \Xi_i$.

As before, the matrix $\Xi_{(i,j)}$ is constructed by stacking Ξ_i on top of Ξ_j. The local observability condition, adapted to the case of random signals, can now be stated as:

$$\ell_i - \ell_j \in \operatorname{Im} \Xi^\top_{(i,j)}$$

for all neighboring actions i, j.

Let us specify the few places where the analysis slightly differs from the arguments of the paper. Since we now have an extra (independent) source of randomness, we define $\mathcal{F}_t$ to be the σ-algebra generated by the random variables $\{k_1, I_1, S^1 \ldots, k_t, I_t, S^t\}$ where S^t is the random matrix obtained by stacking all S^t_i. We now define the estimates

$$b^r_{(i,j)} \triangleq v^\top_{i,j} \begin{bmatrix} \mathbb{I}\{I_r = i\}\, S^t_i \\ \mathbb{I}\{k_r = i\}\, \mathbb{I}\{I_r = j\}\, S^t_j / q^r_i(j) \end{bmatrix} e_{j_r} \ , \qquad \forall r \in \{\tau_i(s-1)+1, \ldots, \tau_i(s)\},\ \forall j \in N_i$$

with the only modification that S^t_i and S^t_j are now random variables. Equation (8) now reads

$$\begin{aligned} \mathbb{E}\left[b^t_{(i,j)} | \mathcal{F}_{t-1} \right] &= \sum_{k=1}^{N} p^t_k q^t_k(i) \cdot v^\top_{i,j} \begin{bmatrix} \Xi_i \\ 0 \end{bmatrix} e_{j_t} + p^t_i q^t_i(j) \cdot v^\top_{i,j} \begin{bmatrix} 0 \\ \Xi_j / q^t_i(j)) \end{bmatrix} e_{j_t} \\ &= p^t_i v^\top_{i,j} \Xi_{(i,j)} e_{j_t} \\ &= p^t_i (e_j - e_i)^\top L e_{j_t} \ . \end{aligned} \tag{12}$$

The rest of the analysis follows as in Section 6.2.3, with Ξ in place of S.

7 Classification – putting everything together

In this section we use the results of this paper, along with some previous results, to prove the classification theorems 1 and 2. For the convenience of the reader, we recite these theorems:

Theorem 1 (Classification for games against stochastic opponents)**.** *Let $G = (L, H)$ be a finite partial-monitoring game. Let K be the number of non-dominated actions in G. The minimax expected regret of G against stochastic opponents is*

$$\mathbb{E}[R_T(G)] = \begin{cases} 0, & K = 1; \\ \widetilde{\Theta}(\sqrt{T}), & K > 1,\ G \textit{ is locally observable;} \\ \Theta(T^{2/3}), & G \textit{ is globally observable but not locally observable;} \\ \Theta(T), & G \textit{ is not globally observable.} \end{cases}$$

Theorem 2 (Classification for games against adversarial opponents)**.** *Let $G = (L, H)$ be a non-degenerate finite partial-monitoring game. Let K be the number of non-dominated actions in G. The minimax expected regret of G against adversarial opponents is*

$$\mathbb{E}[R_T(G)] = \begin{cases} 0, & K = 1; \\ \Theta(\sqrt{T}), & K > 1,\ G \textit{ is locally observable;} \\ \Theta(T^{2/3}), & G \textit{ is globally observable but not locally observable;} \\ \Theta(T), & G \textit{ is not globally observable.} \end{cases}$$

Proof of Theroems 1 and 2. The following lower bound results are sufficient for both theorems:

- If a game is not globally observable then its minimax regret is $\Theta(T)$ [Piccolboni and Schindelhauer, 2001].

- If a game has more than one Pareto-optimal actions then its minimax regret is $\Omega(\sqrt{T})$ [Antos et al., 2012].

- If a game is not locally observable then its minimax regret is $\Omega(T^{2/3})$ (Theorem 4).

On the other hand, the following upper bounds completes the proofs of the theorems:

- If a game has only one Pareto-optimal action then the minimax regret is 0 (trivial: an optimal algorithm chooses the Pareto-optimal action in every time step).

- If a game is globally observable then the algorithm FeedExp by Piccolboni and Schindelhauer [2001] achieves $O(T^{2/3})$ expected regret [Cesa-Bianchi et al., 2006].

- For locally observable games,

  1. the algorithm CBP achieves $\widetilde{O}(\sqrt{T})$ expected regret against stochastic opponents (Corollary 3);
  2. if the game is non-degenerate then NeigborhoodWatch achieves $O(\sqrt{T})$ expected regret against adversarial opponents (Corollary 5); □

8 Discussion

This paper presents the recent advances made in understanding finite partial-monitoring games. The main achievement of this work is a classification of games based on their minimax regret. Algorithms are presented that achieve the minimax regret within logarithmic factors for any given game.

The immediate open problem is to include the degenerate games in the classification under the adversarial model. We conjecture that the classification extends to degenerate games the same way as under the stochastic model. From a more practical point of view, more computationally efficient algorithms would be helpful, especially in the stochastic case. If the number of actions or outcomes is high, the running time of the CBP algorithm dramatically increases. This is due to the fact that the algorithm runs LP solvers in every time step.

Another important extension is partial monitoring with side information. In the model investigated in this paper, the learner does not receive any additional information about the outcome, or how the outcome is generated, before taking an action. In many practical applications it is not the case. In dynamic pricing, for example, the vendor might (and should) have additional information about the customer, *e.g.*, how much he needs the product or his financial situation. This leads to the model of partial monitoring with side information. A recent work by Bartók and Szepesvári [2012] investigates this setting. They prove that local observability remains the key condition to achieve root-T regret under partial monitoring with side information.

Acknowledgements

We thank Vianney Perchet and Gilles Stoltz for their helpful comments on the first draft of this paper.

References

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

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 263–273. Citeseer, 2008.

Alekh Agarwal, Peter Bartlett, and Max Dama. Optimal allocation strategies for the dark pool problem. In *13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), May 12-15, 2010, Chia Laguna Resort, Sardinia, Italy*, 2010.

A. Antos, G. Bartók, D. Pál, and Cs. Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 2012. to appear.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

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

G. Bartók and Cs. Szepesvári. Partial monitoring with side information. In *ALT*, 2012. To appear.

G. Bartók, D. Pál, and C. Szepesvári. Toward a classification of finite partial-monitoring games. In *Algorithmic Learning Theory*, pages 224–238. Springer, 2010.

G. Bartók, D. Pál, and C. Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Conference on Learning Theory*, 2011.

G. Bartók, N. Zolghadr, and Cs. Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *ICML*, 2012. submitted.

A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8 (1307-1324):3–8, 2007.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, June 2005.

T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley, New York, second edition, 2006.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, page 394. Society for Industrial and Applied Mathematics, 2005.

D.P. Foster and A. Rakhlin. No internal regret via neighborhood watch. *Journal of Machine Learning Research - Proceedings Track (AISTATS)*, 22:382–390, 2012.

D.P. Foster and R.V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, 1997.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of 44th Annual IEEE Symposium on Foundations of Computer Science 2003 (FOCS 2003)*, pages 594–605. IEEE, 2003.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

G. Lugosi, S. Mannor, and G. Stoltz. Strategies for prediction under imperfect monitoring. *Math. Oper. Res*, 33:513–528, 2008.

Gábor Lugosi and Nicolò Cesa-Bianchi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

V. Perchet. Internal regret with partial monitoring: Calibration-based optimal algorithms. *Journal of Machine Learning Research*, 12:1893–1921, 2011.

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

A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1-2):224–243, 1999.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of Twentieth International Conference on Machine Learning (ICML 2003)*, 2003.

Appendix

Here we give the proofs of the lemmas used in the main proof. For the convenience of the reader, we restate the lemmas.

Lemma 1. *There exists a (problem dependent) constant c such that the following inequalities hold:*

$$N_1^2 \geq N_1^1 - cT\varepsilon\sqrt{N_4^1}\,, \qquad N_3^2 \geq N_3^1 - cT\varepsilon\sqrt{N_4^1}\,,$$
$$N_2^1 \geq N_2^2 - cT\varepsilon\sqrt{N_4^2}\,, \qquad N_3^1 \geq N_3^2 - cT\varepsilon\sqrt{N_4^2}\,.$$

Proof. For any $1 \leq t \leq T$, let $f^t = (f_1, \ldots, f_t) \in \Sigma^t$ be a feedback sequence up to time step t. For $i = 1, 2$, let p_i^* be the probability mass function of feedback sequences of length $T-1$ under opponent strategy p_i and algorithm $\mathcal{A}$. We start by upper bounding the difference between values under the two opponent strategies. For $i \neq j \in \{1, 2\}$ and $k \in \{1, 2, 3\}$,

$$\begin{aligned}
N_k^i - N_k^j &= \sum_{f^{T-1}} \left(p_i^*(f^{T-1}) - p_j^*(f^{T-1})\right) \sum_{t=0}^{T-1} 1\mathcal{A}(f^t) \in \mathcal{N}_k \\
&\leq \sum_{\substack{f^{T-1}: \\ p_i^*(f^{T-1}) - p_j^*(f^{T-1}) \geq 0}} \left(p_i^*(f^{T-1}) - p_j^*(f^{T-1})\right) \sum_{t=0}^{T-1} 1\mathcal{A}(f^t) \in \mathcal{N}_k \\
&\leq T \sum_{\substack{f^{T-1}: \\ p_i^*(f^{T-1}) - p_j^*(f^{T-1}) \geq 0}} p_i^*(f^{T-1}) - p_j^*(f^{T-1}) = \frac{T}{2}\|p_1^* - p_2^*\|_1 \\
&\leq T\sqrt{\mathrm{KL}(p_1^*||p_2^*)/2}\,, \qquad (13)
\end{aligned}$$

where $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence and $\|\cdot\|_1$ is the L_1-norm. The last inequality follows from Pinsker's inequality [Cover and Thomas, 2006]. To upper bound $\mathrm{KL}(p_1^*||p_2^*)$ we use the chain rule for KL-divergence. By overloading p_i^* so that $p_i^*(f^{t-1})$ denotes the probability of feedback sequence f^{t-1} under opponent strategy p_i and algorithm $\mathcal{A}$, and $p_i^*(f_t|f^{t-1})$ denotes the conditional probability of feedback $f_t \in \Sigma$ given that the past feedback sequence was f^{t-1}, again under p_i and $\mathcal{A}$. With this notation we have

$$\begin{aligned}
\mathrm{KL}(p_1^*||p_2^*) &= \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) \sum_{f_t} p_1^*(f_t|f^{t-1}) \log \frac{p_1^*(f_t|f^{t-1})}{p_2^*(f_t|f^{t-1})} \\
&= \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) \sum_{i=1}^{4} 1\mathcal{A}(f^{t-1}) \in \mathcal{N}_i \sum_{f_t} p_1^*(f_t|f^{t-1}) \log \frac{p_1^*(f_t|f^{t-1})}{p_2^*(f_t|f^{t-1})} \qquad (14)
\end{aligned}$$

Let $a_{f_t}^\top$ be the row of S that corresponds to the feedback symbol f_t.[8] Assume $k = \mathcal{A}(f^{t-1})$. If the feedback set of action k does not contain f_t then trivially $p_i^*(f_t|f^{t-1}) = 0$ for $i = 1, 2$. Otherwise $p_i^*(f_t|f^{t-1}) = a_{f_t}^\top p_i$. Since $p_1 - p_2 = 2\varepsilon v$ and $v \in \operatorname{Ker} S$, we have $a_{f_t}^\top v = 0$ and thus, if the choice of the algorithm is in either $\mathcal{N}_1, \mathcal{N}_2$ or $\mathcal{N}_3$, then $p_1^*(f_t|f^{t-1}) = p_2^*(f_t|f^{t-1})$. It follows that the inequality chain can be continued from (14) by writing

$$\begin{aligned}
\mathrm{KL}(p_1^*||p_2^*) &\le \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) 1\mathcal{A}(f^{t-1}) \in \mathcal{N}_4 \sum_{f_t} p_1^*(f_t|f^{t-1}) \log \frac{p_1^*(f_t|f^{t-1})}{p_2^*(f_t|f^{t-1})} \\
&\le c_1 \varepsilon^2 \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) 1\mathcal{A}(f^{t-1}) \in \mathcal{N}_4 \qquad (15) \\
&\le c_1 \varepsilon^2 N_4^1 \,.
\end{aligned}$$

In (15) we used Lemma 10 (see below) to upper bound the KL-divergence of p_1 and p_2. Flipping p_1^* and p_2^* in (13) we get the same result with N_4^2. Reading together with the bound in (13) we get all the desired inequalities. □

Lemma 10. *Fix a probability vector $p \in \Delta_M$, and let $\epsilon \in \mathcal{R}^M$ such that $p - \epsilon, p + \epsilon \in \Delta_M$ also holds. Then*

$$\mathrm{KL}(p - \epsilon || p + \epsilon) = O(\|\epsilon\|_2^2) \qquad \text{as } \epsilon \to 0.$$

The constant and the threshold in the $O(\cdot)$ notation depends on p.

Proof. Since p, $p + \epsilon$, and $p - \epsilon$ are all probability vectors, notice that $|\epsilon(i)| \le p(i)$ for $1 \le i \le M$. So if a coordinate of p is zero then the corresponding coordinate of ϵ has to be zero as well. As zero coordinates do not modify the KL divergence, we can assume without loss of generality that all coordinates of p are positive. Since we are interested only in the case when $\epsilon \to 0$, we can also assume without loss of generality that $|\epsilon(i)| \le p(i)/2$. Also note that the coordinates of $\epsilon = (p + \epsilon) - \epsilon$ have to sum up to zero. By definition,

$$\mathrm{KL}(p - \epsilon || p + \epsilon) = \sum_{i=1}^{M} (p(i) - \epsilon(i)) \log \frac{p(i) - \epsilon(i)}{p(i) + \epsilon(i)}.$$

We write the term with the logarithm

$$\log \frac{p(i) - \epsilon(i)}{p(i) + \epsilon(i)} = \log\left(1 - \frac{\epsilon(i)}{p(i)}\right) - \log\left(1 + \frac{\epsilon(i)}{p(i)}\right),$$

so that we can use that, by second order Taylor expansion around 0, $\log(1 - x) - \log(1 + x) = -2x + r(x)$, where $|r(x)| \le c|x|^3$ for $|x| \le 1/2$ and some $c > 0$. Combining these equations, we get

$$\begin{aligned}
\mathrm{KL}(p - \epsilon || p + \epsilon) &= \sum_{i=1}^{M} (p(i) - \epsilon(i)) \left[-2\frac{\epsilon(i)}{p(i)} + r\left(\frac{\epsilon(i)}{p(i)}\right)\right] \\
&= \sum_{i=1}^{M} -2\epsilon(i) + \sum_{i=1}^{M} 2\frac{\epsilon^2(i)}{p(i)} + \sum_{i=1}^{M} (p(i) - \epsilon(i)) r\left(\frac{\epsilon(i)}{p(i)}\right).
\end{aligned}$$

Here the first term is 0, letting $\underline{p} = \min_{i \in \{1,\ldots,M\}} p(i)$ the second term is bounded by

[8] Here without loss of generality we assume that different actions have difference feedback symbols, and thus a row of S corresponding to a symbol is unique.

$2\sum_{i=1}^{M} \epsilon^2(i)/\underline{p} = (2/\underline{p})\|\epsilon\|_2^2$, and the third term is bounded by

$$\begin{aligned}\sum_{i=1}^{M}(p(i)-\epsilon(i))\left| r\left(\frac{\epsilon(i)}{p(i)}\right)\right| &\le c\sum_{i=1}^{M}\frac{p(i)-\epsilon(i)}{p^3(i)}|\epsilon(i)|^3 \\ &\le c\sum_{i=1}^{M}\frac{|\epsilon(i)|}{p^2(i)}\epsilon^2(i) \\ &\le \frac{c}{2}\sum_{i=1}^{M}\frac{1}{\underline{p}}\epsilon^2(i) = \frac{c}{2\underline{p}}\|\epsilon\|_2^2 .\end{aligned}$$

Hence, $\mathrm{KL}(p-\epsilon||p+\epsilon) \le \frac{4+c}{2\underline{p}}\|\epsilon\|_2^2 = O(\|\epsilon\|_2^2)$. □

Lemma 2. *For any $\{i,j\} \in \mathcal{N}$, $t \ge 1$,*

$$\mathbb{P}\left(|\tilde{\delta}_{i,j}(t) - \delta_{i,j}| \ge c_{i,j}(t)\right) \le 2|V_{i,j}|t^{1-2\alpha} .$$

Proof.

$$\begin{aligned}&\mathbb{P}\left(|\tilde{\delta}_{i,j}(t) - \delta_{i,j}| \ge c_{i,j}(t)\right) \\ &\le \sum_{k\in N_{i,j}^+} \mathbb{P}\left(|v_{i,j,k}^\top \frac{\nu_k(t-1)}{n_k(t-1)} - v_{i,j,k}^\top S_k p^*| \ge \|v_{i,j,k}\|_\infty \sqrt{\frac{\alpha \log t}{n_k(t-1)}}\right) \qquad (16) \\ &= \sum_{k\in N_{i,j}^+}\sum_{s=1}^{t-1} \mathbb{I}\left\{n_k(t-1)=s\right\} \mathbb{P}\left(|v_{i,j,k}^\top \frac{\nu_k(t-1)}{s} - v_{i,j,k}^\top S_k p^*| \ge \|v_{i,j,k}\|_\infty \sqrt{\frac{\alpha \log t}{s}}\right) \qquad (17) \\ &\le \sum_{k\in N_{i,j}^+} 2t^{1-2\alpha} \qquad (18) \\ &= 2|N_{i,j}^+|t^{1-2\alpha} ,\end{aligned}$$

where in (16) we used the triangle inequality and the union bound and in (18) we used Hoeffding's inequality. □

Lemma 3. *Take an action i and a plausible pair $(\mathcal{P}', \mathcal{N}') \in \Psi$ such that $i \in \mathcal{P}'$. Then there exists a path π that starts at i and ends at i^* that lies in $\mathcal{N}'$.*

Proof. If $(\mathcal{P}', \mathcal{N}')$ is a valid configuration, then there is a convex polytope $\Pi \subseteq \Delta_M$ such that $p^* \in \Pi$, $\mathcal{P}' = \{i \ : \ \dim \mathcal{C}_i \cap \Pi = M-1\}$ and $\mathcal{N}' = \{\{i,j\} \ : \ \dim \mathcal{C}_i \cap \mathcal{C}_j \cap \Pi = M-2\}$.

Let p' be an arbitrary point in $\mathcal{C}_i \cap \Pi$. We enumerate the actions whose cells intersect with the line segment $\overline{p'p^*}$, in the order as they appear on the line segment. We show that this sequence of actions $i_0, \ldots, i_r$ is a feasible path.

- It trivially holds that $i_0 = i$, and i_r is optimal.
- It is also obvious that consecutive actions on the sequence are in $\mathcal{N}'$.

For an illustration we refer the reader to Figure 3 □

Next, we want to prove lemma 4. For this, we need the following auxiliary result:

Lemma 11. *Let action i be a degenerate action in the neighborhood action set $N_{k,l}^+$ of neighboring actions k and l. Then ℓ_i is a convex combination of ℓ_k and ℓ_l.*

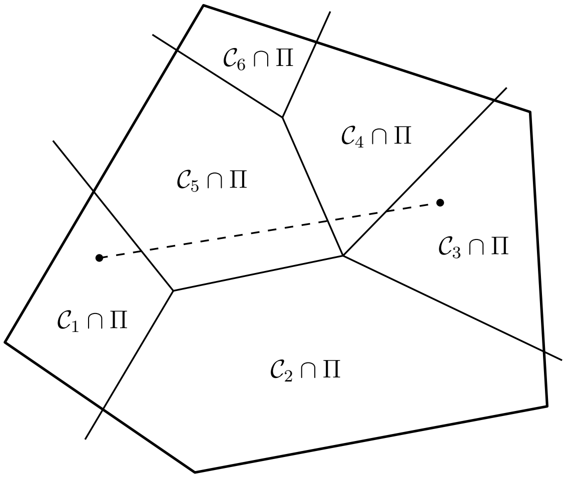


Figure 3: The dashed line defines the feasible path $1, 5, 4, 3$.

Proof. For simplicity, we rename the degenerate action i to action 1, while the other actions k, l will be called actions 2 and 3, respectively. Since action 1 is a degenerate action between actions 2 an 3, we have that

$$(p \in \Delta_M \text{ and } p \perp (\ell_1 - \ell_2)) \to (p \perp (\ell_1 - \ell_3) \text{ and } p \perp (\ell_2 - \ell_3))$$

implying

$$(\ell_1 - \ell_2)^{\perp} \subseteq (\ell_1 - \ell_3)^{\perp} \cap (\ell_2 - \ell_3)^{\perp} \ .$$

Using de Morgan's law we get

$$\langle \ell_1 - \ell_2 \rangle \supseteq \langle \ell_1 - \ell_3 \rangle \oplus \langle \ell_2 - \ell_3 \rangle \, .$$

This implies that for any $c_1, c_2 \in$ there exists a $c_3 \in$ such that

$$\begin{aligned} c_3(\ell_1 - \ell_2) &= c_1(\ell_1 - \ell_3) + c_2(\ell_2 - \ell_3) \\ \ell_3 &= \frac{c_1 - c_3}{c_1 + c_2}\ell_1 + \frac{c_2 + c_3}{c_1 + c_2}\ell_2 \, , \end{aligned}$$

suggesting that ℓ_3 is an affine combination of (or collinear with) ℓ_1 and ℓ_2.

We know that there exists $p_1 \in \Delta$ such that $\ell_1^\top p_1 < \ell_2^\top p_1$ and $\ell_1^\top p_1 < \ell_3^\top p_1$. Also, there exists $p_2 \in \Delta_M$ such that $\ell_2^\top p_2 < \ell_1^\top p_2$ and $\ell_2^\top p_2 < \ell_3^\top p_2$. Using these and linearity of the dot product we get that ℓ_3 must be the middle point on the line, which means that ℓ_3 is indeed a convex combination of ℓ_1 and ℓ_2. □

Lemma 4. *Fix any* $t \geq 1$.

1. *Take any action* i. *On the event* $\mathcal{G}_t \cap \mathcal{D}_t$,[9] *from* $i \in \mathcal{P}(t) \cup N^+(t)$ *it follows that*

$$\delta_i \leq 2 d_i \sqrt{\frac{\alpha \log t}{f(t)}} \max_{k \in \underline{N}} \frac{W_k}{\sqrt{\eta_k}} \, .$$

[9] Here and in what follows all statements that start with "On event X" should be understood to hold almost surely on the event. However, to minimize clutter we will not add the qualifier "almost surely".

2. Take any action k. On the event $\mathcal{G}_t \cap \mathcal{D}_t^c$, from $I_t = k$ it follows that

$$n_k(t-1) \le \min_{j \in \mathcal{P}(t) \cup N^+(t)} 4W_k^2 \frac{d_j^2}{\delta_j^2} \alpha \log t \,.$$

Proof. First we observe that for any neighboring action pair $\{i,j\} \in \mathcal{N}(t)$, on $\mathcal{G}_t$ it holds that $\delta_{i,j} \le 2c_{i,j}(t)$. Indeed, from $\{i,j\} \in \mathcal{N}(t)$ it follows that $\tilde{\delta}_{i,j}(t) \le c_{i,j}(t)$. Now, on $\mathcal{G}_t$, $\delta_{i,j} \le \tilde{\delta}_{i,j}(t) + c_{i,j}(t)$. Putting together the two inequalities we get $\delta_{i,j} \le 2c_{i,j}(t)$.

Now, fix some action i that is not dominated. We define the "parent action" i' of i as follows: If i is not degenerate then $i' = i$. If i is degenerate then we define i' to be the Pareto-optimal action such that $\delta_{i'} \ge \delta_i$ and i is in the neighborhood action set of i' and some other Pareto-optimal action. It follows from Lemma 11 that i' is well-defined.

Consider case 1. Thus, $I_t \ne k(t) = \operatorname{argmax}_{j \in \mathcal{P}(t) \cup \mathcal{V}(t)} W_j^2 / n_j(t-1)$. Therefore, $k(t) \notin \mathcal{R}(t)$, *i.e.*, $n_{k(t)}(t-1) > \eta_{k(t)} f(t)$. Assume now that $i \in \mathcal{P}(t) \cup N^+(t)$. If i is degenerate then i' as defined in the previous paragraph is in $\mathcal{P}(t)$ (because the rejected regions in the algorithm are closed). In any case, by Lemma 3, there is a path $(i_0, \ldots, i_r)$ in $\mathcal{N}(t)$ that connects i' to i^* ($i^* \in \mathcal{P}(t)$ holds on $\mathcal{G}_t$). We have that

$$\begin{aligned}
\delta_i \le \delta_{i'} &= \sum_{s=1}^{r} \delta_{i_{s-1}, i_s} \\
&\le 2 \sum_{s=1}^{r} c_{i_{s-1}, i_s} \\
&= 2 \sum_{s=1}^{r} \sum_{j \in V_{i_{s-1}, i_s}} \|v_{i_{s-1}, i_s, j}\|_\infty \sqrt{\frac{\alpha \log t}{n_j(t-1)}} \\
&\le 2 \sum_{s=1}^{r} \sum_{j \in V_{i_{s-1}, i_s}} W_j \sqrt{\frac{\alpha \log t}{n_j(t-1)}} \\
&\le 2 d_i W_{k(t)} \sqrt{\frac{\alpha \log t}{n_{k(t)}(t-1)}} \\
&\le 2 d_i W_{k(t)} \sqrt{\frac{\alpha \log t}{\eta_{k(t)} f(t)}} \,.
\end{aligned}$$

Upper bounding $W_{k(t)} / \sqrt{\eta_{k(t)}}$ by $\max_{k \in \underline{N}} W_k / \sqrt{\eta_k}$ we obtain the desired bound.

Now, for case 2 take an action k, consider $\mathcal{G} \cap \mathcal{D}_t^c$, and assume that $I_t = k$. On D_t^c, $I_t = k(t)$. Thus, from $I_t = k$ it follows that $W_k / \sqrt{n_k(t-1)} \ge W_j / \sqrt{n_j(t-1)}$ holds for all $j \in \mathcal{P}(t)$. Let $J_t = \operatorname{argmin}_{j \in \mathcal{P}(t) \cup N^+(t)} \frac{d_j^2}{\delta_j^2}$. Now, similarly to the previous case, there exists a path $(i_0, \ldots, i_r)$ from the parent action $J_t' \in \mathcal{P}(t)$ of J_t to i^* in $\mathcal{N}(t)$. Hence,

$$\begin{aligned}
\delta_{J_t} \le \delta_{J_t'} &= \sum_{s=1}^{r} \delta_{i_{s-1}, s} \\
&\le 2 \sum_{s=1}^{r} \sum_{j \in V_{i_{s-1}, i_s}} W_j \sqrt{\frac{\alpha \log t}{n_j(t-1)}} \\
&\le 2 d_{J_t} W_k \sqrt{\frac{\alpha \log t}{n_k(t-1)}} \,,
\end{aligned}$$

implying

$$\begin{aligned} n_k(t-1) &\le 4W_k^2 \frac{d_{J_t}^2}{\delta_{J_t}^2}\alpha \log t \\ &= \min_{j\in\mathcal{P}(t)\cup N^+(t)} 4W_k^2 \frac{d_j^2}{\delta_j^2}\,\alpha \log t\,. \end{aligned}$$

This concludes the proof of Lemma 4. □

Lemma 5. *Let $G = (L, H)$ be a finite partial-monitoring game and $p \in \Delta_M$ an opponent strategy. There exists a $\rho_2 > 0$ such that A_{ρ_2} is a point-local game in G.*

Proof. For any (not necessarily neighboring) pair of actions $\{i, j\}$, the boundary between them is defined by the set $B_{i,j} = \{p \in \Delta_M \ : \ (\ell_i - \ell_j)^\top p = 0\}$. We generalize this notion by introducing the *margin*: for any $\xi \ge 0$, let the margin be the set $B_{i,j}^{\xi} = \{p \in \Delta_M \ : \ |(\ell_i - \ell_j)^\top p| \le \xi\}$. It follows from finiteness of the action set that there exists a $\xi^* > 0$ such that for any set K of neighboring action pairs,

$$\bigcap_{\{i,j\}\in K} B_{i,j} \neq \emptyset \qquad \Longleftrightarrow \qquad \bigcap_{\{i,j\}\in K} B_{i,j}^{\xi^*} \neq \emptyset\,. \tag{19}$$

Let $\rho_2 = \xi^*/2$. Let $A = A_{\rho_2}$. Then for every pair i, j in A, $(\ell_i - \ell_j)^\top p^* = \delta_{i,j} \le \delta_i + \delta_j \le \rho_2$. That is, $p^* \in B_{i,j}^{\xi^*}$. It follows that $p^* \in \bigcap_{i,j\in A\times A} B_{i,j}^{\xi^*}$. This, together with (19), implies that A is a point-local game. □

Lemma 6. *There exists a problem dependent constant K such that the internal regret is at most K times the local internal regret.*

Let Φ be a set of transformations $\underline{N} \mapsto \underline{N}$. Φ-regret is defined as

$$\sup_{\phi\in\Phi} \mathbb{E}\left\{\sum_{t=1}^T \ell_{I_t}^\top e_{j_t} - \sum_{t=1}^T \ell_{\phi(I_t)}^\top e_{j_t}\right\}$$

Let Φ_L be the set of local transformations. The claim is that there exists a problem-dependent constant K (independent of T) such that Φ-regret is upper bounded by K times Φ_L-regret.

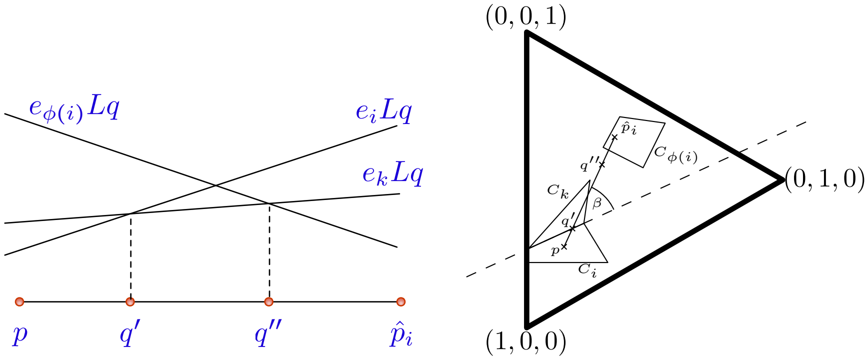
Let us first write

$$\sum_{t=1}^T \ell_{I_t}^\top e_{j_t} - \sum_{t=1}^T \ell_{\phi(I_t)}^\top e_{j_t} = \sum_{i=1}^N \sum_{t\in\underline{T}:I_t=i} (\ell_i - \ell_{\phi(i)})^\top e_{j_t} = \sum_{i=1}^N s_i(\ell_i - \ell_{\phi(i)})^\top \hat{p}_i$$

where $s_i = |\{t : I_t = i\}|$ and $\hat{p}_i = \frac{1}{s_i}\sum_{t\in[T]:I_t=i} e_{j_t}$, the empirical frequency of adversarial actions on the rounds when our choice is action i. To prove the claim it is enough to show that for any $i \in \underline{N}$ there exists a $K > 0$ (that does not depend on T) and a neighboring action $k \in \mathcal{N}_i$ such that

$$(\ell_i - \ell_{\phi(i)})^\top \hat{p}_i \le K(\ell_i - \ell_k)^\top \hat{p}_i\ .$$

We may assume that $\phi(i)$ is the best response action to $\hat{p}_i$ (in other words, $\hat{p}_i \in C_{\phi(i)}$) since this makes the above requirement harder to satisfy. If $\phi(i)$ is a neighbor of i, the claim is trivially satisfied with $K = 1$. Otherwise, pick $p \in C_i$ to be the centroid of C_i and consider the segment $[p, \hat{p}_i] \subset \Delta_M$. Note that on this segment the function $f(q) = \min_i e_i L q$ is concave and piece-wise linear. Since i and $\phi(i)$ are not neighbors, there exists an action $k \neq \phi(i)$ such that k is a neighbor of i and there exists $q' \in [p, \hat{p}_i]$ such that $\ell_i^\top q' = \ell_k^\top q'$. It then follows that $\ell_{\phi(i)}^\top \hat{p}_i \le \ell_k^\top \hat{p}_i < \ell_i^\top \hat{p}_i$. The first inequality is an equality when $\hat{p}_i = q''$, in which case we



(a) There must be a neighbor of i, denoted by k between i and $\phi(i)$.

(b) The dashed line denotes the intersection of the probability simplex and the hyperplane $(\ell_i - \ell_k)^\top x = 0$. The minimum angle β must be isolated from zero because p is the centroid of C_i.

Figure 4: Illustrations for Lemma 6.

may simply choose $K = 1$. Otherwise, let u denote a unit vector in the direction $q'' - p$. We may express $\hat{p}_i$ as $q'' + \alpha u$ for a constant $\alpha > 0$. Then we are seeking an upper bound on the ratio

$$\frac{(\ell_i - \ell_k)^\top q'' + \alpha(\ell_i - \ell_{\phi(i)})^\top u}{(\ell_i - \ell_k)^\top q'' + \alpha(\ell_i - \ell_k)^\top u} \leq \frac{(\ell_i - \ell_{\phi(i)})^\top u}{(\ell_i - \ell_k)^\top u} \,.$$

Obviously, the enumerator of the above fraction can be upper bounded by $\|\ell_i - \ell_{\phi(i)}\|$. Now what is left is to lower bound the denominator. The lower bound depends on the angle between orthogonal of $(\ell_i - \ell_k)$ and the direction $\hat{p}_i - p$. Since p was chosen as the centroid of C_i, this angle (β on Figure 4(b)) is isolated from zero.