JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–15

A near-optimal algorithm for finite partial-monitoring games against adversarial opponents

Gábor Bartók BARTOK@INF.ETHZ.CH
ETH Zürich

Abstract

Partial monitoring is an online learning model where in every time step, after a learner and an opponent choose their actions, the loss and the feedback for the learner is calculated based on a loss and a feedback function, both of which are known to the learner ahead of time. As in other online learning scenarios, the goal of the learner is to minimize his cumulative loss. In this paper we present and analyze a new algorithm for locally observable partial monitoring games. We prove that the expected regret of our algorithm is of $\widetilde{O}(\sqrt{N'T})$, where T is the time horizon and N' is the size of the largest point-local game. The most important improvement of this bound compared to previous results is that it does not depend directly on the number of actions, but rather on the structure of the game.

Keywords: partial monitoring, online learning, limited feedback, regret analysis

1. Introduction

Consider the sequential game where a learner chooses actions while, at the same time, an opponent chooses an outcome, then the learner receives some feedback and suffers some (unobserved) loss. Both the loss and the feedback are deterministic functions of the action and the outcome and these actions are revealed to both party before the game begins. The goal of the learner is to minimize his cumulative loss.

A classical example is *dynamic pricing*. In this problem, a vendor wants to sell his product to customers. Every time a customer appears, the vendor (learner) puts a price tag on the product while the customer (opponent) secretly thinks of a maximum price for which he is willing to buy the product. Then, the transaction happens or not depending on the price tag and the maximum price. If the product was not sold then the vendor suffers some constant loss for lost sales. Even if the customer bought the product, the vendor suffers some loss if the price tag was too low (opportunity loss). This value of this latter kind of loss will never be revealed to the learner. His feedback is only whether the product was bought or not.

Partial monitoring models online learning scenarios with limited feedback, such as the example above. Under what conditions can the learner have good performance? What are the theoretical limitations of learning? How does the feedback structure influence the ability and speed of learning? These are some of the questions to investigate in partial monitoring.

While the goal of the learner is to minimize his cumulative loss, the performance is measured in terms of the *regret*, defined as the excess cumulative loss compared to that of the best fixed action in hindsight. If the regret grows sublinearly with the time horizon, we

can say that the learner learns the optimal action. Then the question becomes how quickly the learning happens, *i.e.,* what is the growth rate of the regret?

Regret analysis of partial monitoring problems started with investigating special cases. In the case of *full feedback*, the Weighted Majority algorithm due to Vovk (1990) and Littlestone and Warmuth (1994) achieves $\Theta(\sqrt{T \log N})$ expected regret against an oblivious adversarial opponent, where T is the time horizon and N is the number of actions. The other thoroughly investigated case is the so-called *bandit feedback* model where the learner receives his own loss as feedback. In this case, the algorithm Exp3 by Auer et al. (2002) achieves $O(\sqrt{TN \log N})$ expected regret.[1]

The general setting of partial monitoring was first considered by Piccolboni and Schindelhauer (2001). Their algorithm, FeedExp was proven to achieve $O(T^{3/4})$ expected regret in terms of the time horizon whenever the problem is learnable at all. Later, Cesa-Bianchi et al. (2006) tightened this bound to $O(T^{2/3})$. They also showed a problem instance where the above bound is optimal. The question remained, however, which problem instances allow regret smaller than $cT^{2/3}$. This question was partially answered by Bartók et al. (2011) who showed, against stochastic opponents, that some class of partial monitoring games can have $O(\sqrt{T})$ regret. The key condition that distinguishes games with smaller regret is the so called *local observability condition* (see Definition 4). The same answer was later given for oblivious adversarial opponents by Foster and Rakhlin (2012).

The above recent works on partial monitoring focus mainly on the regret growth rate in terms of the time horizon. It leaves the question open what bound we can have in terms of the number of actions? Most of the algorithms and bounds prove to be suboptimal when compared to the special cases of full feedback and bandit feedback. The work of Mannor and Shamir (2011) considers the case "between" bandit and full feedback games where the learner gets his own loss and possibly losses of other actions in every time step. They derive a bound that is, in some sense, unimprovable.

In our work, we investigate how the feedback and loss structure influences the worst-case regret of a game in terms of both the time horizon and the number of actions. In particular, we show that any non-degenerate locally observable finite partial monitoring game has worst-case expected regret of $\widetilde{O}(\sqrt{N'T})$, where N' is the size of the largest *point-local game* (see Definition 6).[2] This bound is optimal at least for the case of bandit feedback and substantially improves the bound of $\widetilde{O}(N\sqrt{T})$ given by Foster and Rakhlin (2012). Our algorithm, while similar to the algorithm NeighborhoodWatch of Foster and Rakhlin (2012), differs in two important points. First, the point-local games we define slightly differ from the local games in their work. Second, when deciding which (point-)local game to play, our algorithm *does not randomize*. We believe that this second property is the main insight that allows us to obtain the improved upper bound.

2. Preliminaries

An instance $\mathcal{G}$ of finite partial monitoring is defined by the matrices $\mathcal{L} \in \mathbb{R}^{N \times M}$ and $\mathcal{H} \in \Sigma^{N \times M}$ for some alphabet Σ of symbols. Before the game starts, these matrices are revealed

1. The algorithm INF by Audibert and Bubeck (2009) achieves the optimal bound of $O(\sqrt{TN})$, removing the logarithmic term.
2. The notation $\widetilde{O}(\cdot)$ hides logarithmic factors.

to both the learner and the opponent. In every round t, the learner chooses an action $I_t \in \{1, \dots, N\}$ and simultaneously the opponent chooses an outcome $J_t \in \{1, \dots, M\}$. Then the learner receives feedback $\mathcal{H}(I_t, J_t)$ and also suffers loss $\mathcal{L}(I_t, J_t)$, which is not revealed to him. The goal of the learner is to minimize his cumulative loss over a time horizon T:

$$L_T = \sum_{t=1}^{T} \mathcal{L}(I_t, J_t)\,.$$

In this paper, we assume that the opponent does not have access to the actions chosen by the learner. That is, we assume an *oblivious adversarial* opponent. Equivalently, we can assume that the opponent chooses an outcome sequence $(J_1, \dots, J_T)$ before the game begins. We measure the performance of the learner by comparing his cumulative loss to the cumulative loss of the best fixed action in hindsight. To this end, the regret is defined:

$$R_T = L_T - \min_{i \in \{1,\dots,N\}} \sum_{t=1}^{T} \mathcal{L}(i, J_t)\,.$$

2.1. Properties of a game

Most of the definitions of this section are taken from Bartók et al. (2011). Consider the game $\mathcal{G} = (\mathcal{L}, \mathcal{H})$ with N actions and M outcomes. For $1 \le i \le N$, let $\ell_i \in \mathbb{R}^M$ denote the column vector consisting of the i^{th} row of $\mathcal{L}$. Let Δ_M denote the ($M-1$-dimensional) probability simplex, that is, $\Delta_M = \{q \in \mathbb{R}^M \ : \ \|q\|_1 = 1, q \ge 0\}$. For any outcome sequence of length T, the vector q consisting of the relative frequencies with which each outcome occurs is in Δ_M. With this notation we can describe the cumulative loss of action i as

$$\sum_{t=1}^{T} \mathcal{L}(i, J_t) = T \cdot \ell_i^\top q\,.$$

We can say that the term $\ell_i^\top q$ is the average loss of action i. Thus, a relative frequency $q \in \Delta_M$ determines which action is optimal: $\operatorname{argmin}_i \ell_i^\top q$. This induces a *cell decomposition* of Δ_M:

Definition 1 (Cells) *Given a game $\mathcal{G} = (\mathcal{L}, \mathcal{H})$, the cell of action i is defined as*

$$\mathcal{C}_i = \{q \in \Delta_M \ : \ \ell_i^\top q = \min_j \ell_j^\top q\}\,.$$

In other words, the cell of an action is the set of outcome frequencies for which the action is optimal. It is easy to see that $\mathcal{C}_i$ is the solution of the linear inequality system $(\ell_i - \ell_j)^\top q \le 0, j = 1, \dots, N$ and thus every cell is an intersection of halfspaces (convex polytope). It is also obvious that the cells together cover Δ_M.

Now we are equipped to define *neighbors*:

Definition 2 *Two actions i and j are neighbors if $\mathcal{C}_i \cap \mathcal{C}_j \neq \emptyset$.*

Note that this definition slightly differs from that of Bartók et al. (2011) where two actions are neighbors only if their cells' intersection is an $M-2$-dimensional polytope. As a consequence, our result is weaker than that of Bartók et al. (2011) in the sense that our algorithm works for a smaller class of games. Also note that in this paper we restrict ourselves to games that have neither *duplicate* actions, *i.e.,* actions i, j with $\ell_i = \ell_j$, nor *degenerate* actions, whose cells are lower than $M-1$-dimensional. Games with actions of either of the above kinds are considered degenerate games. Nonetheless, any degenerate game can be made non-degenerate by slightly perturbing its loss matrix.

So far we have not dealt with the feedback structure of the game. To this end, the *signal matrix* of an action is defined.

Definition 3 (Bartók et al. (2011)) *For an action i, let $\sigma_1, \ldots, \sigma_{s_i} \in \Sigma$ be the symbols appearing in row i of $\mathcal{H}$. The* signal matrix *S_i of action i is defined as the incidence matrix of symbols and outcomes:*

$$S_i(k,l) = \mathbb{I}\{\mathcal{H}(i,l) = \sigma_k\} \qquad k = 1, \ldots, s_i\,; \qquad l = 1, \ldots, M\,.$$

For ease of reading, we illustrate the above definition with an example. Let the row of $\mathcal{H}$ corresponding to action i be $\begin{pmatrix} a & b & a & c & c \end{pmatrix}$. Then S_i is a 3×5 matrix:

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

A useful property of the signal matrix is that if we treat the outcomes as the standard basis vectors in the M-dimensional Euclidean space then, denoting by e_t the outcome at time step t, we get that the vector $S_i e_t$ is the basis vector for the corresponding symbol the learner receives as feedback, had he chosen action i. Also, for any relative frequency $q \in \Delta_M$, $S_i q$ gives a relative frequency over symbols observed by the learner.

Now we recite the *local observability condition*, the key condition for achieving $\widetilde{O}(\sqrt{T})$ regret.

Definition 4 (Bartók et al. (2011)) *Two neighboring actions i and j are* locally observable *if $\ell_i - \ell_j \in \operatorname{Im} S_i^\top \oplus \operatorname{Im} S_j^\top$.*

Here $\operatorname{Im}(\cdot)$ denotes the image space (or column space) of a matrix while $\oplus$ denotes the standard direct sum.

The above definition implies that two actions i and j are locally observable if $\ell_i - \ell_j$ can be expressed as $S_i^\top v_{i,j} - S_j^\top v_{j,i}$ with some vectors $v_{i,j} \in \mathbb{R}^{\sigma_i}$ and $v_{j,i} \in \mathbb{R}^{\sigma_j}$. The reason why this property is important becomes clear when we try to calculate the difference between the average loss of action i and j:

$$(\ell_i - \ell_j)^\top q = v_{i,j}^\top S_i q - v_{j,i}^\top S_j q\,.$$

Here the fact to note is that the vectors $S_i q$ and $S_j q$ are symbol frequencies and can be observed by choosing actions i and j. We will heavily use this property of locally observable neighbors in our algorithm.

The last definition of this section is *local observability* of a game.

Definition 5 *A game is* locally observable *if all of its neighboring action pairs are locally observable.*

In the following, we describe our algorithm that achieves near-optimal regret for locally observable games.

3. Algorithm

Our algorithm is similar in spirit to that of Foster and Rakhlin (2012). It also borrows ideas from Bartók et al. (2010). The skeleton of Foster and Rakhlin's algorithm NeighborhoodWatch is that the actions are split into "local games", and the algorithm plays a two-level hierarchical exponential-weights algorithm. On the top level the algorithm decides which local game to play, then the bottom level decides which action to play within the chosen local game. This way their algorithm achieves $\widetilde{O}(N\sqrt{T})$ expected regret against any outcome sequence.

The difference in our approach compared to that of Foster and Rakhlin is twofold: first, our "local games" are defined differently and second, to decide which local game to play at a time step is not decided by a randomized strategy but chosen deterministically. This way we are able to reduce the expected worst-case regret to be of $\widetilde{O}(\sqrt{N'T})$, where N' is the size of the largest *point-local* game, defined shortly.

In Foster and Rakhlin (2012), every action has a corresponding local game, namely, $\mathcal{G}_i$ consists of action i and all of its neighbors. For our algorithm, we define *point-local* games:

Definition 6 (Point-local games) *A subset of actions $i_1, \dots, i_s$ in $\mathcal{G}$ is called* point-local *if $\mathcal{C}_{i_1} \cap \dots \cap \mathcal{C}_{i_s} \neq \emptyset$ and adding an extra action i_{s+1} gives $\mathcal{C}_{i_1} \cap \dots \cap \mathcal{C}_{i_{s+1}} = \emptyset$.*

In words, a subset of actions is a point-local game if their cells have at least one point in common and if the subset is not extendable. For an illustration of cell decomposition and local versus point-local games, see Figure 1.

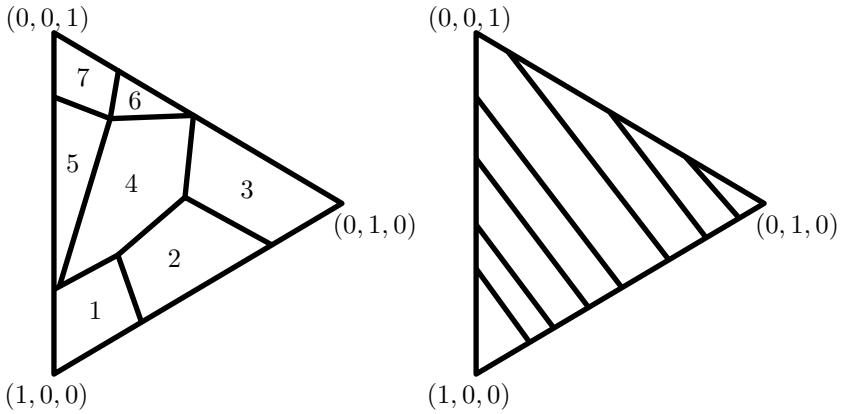
3.1. Playing in point-local games

In this section we describe how one plays in a single point-local game. This algorithm, called LocalExp3, will be used as a subroutine in our main algorithm.

For now, let $\mathcal{G}$ be a locally observable point-local game. That is, every two actions of the game are locally observable neighbors. For every action pair i, j, we define the vectors $v_{i,j}$ and $v_{j,i}$ according to local observability:

$$\ell_i - \ell_j = S_i^\top v_{i,j} - S_j^\top v_{j,i}\,.$$

Remark 7 *The vectors $v_{i,j}$ that satisfy the above equation are not unique. Actually, the choice of these vectors can influence the behaviour of the algorithm and thus also the regret. We do not exactly know what is the best choice but, as we will see from the regret bound, a choice that minimizes the max norm of the vectors will be helpful for us.*



(a) The largest local game is the whole game: 4 and all of its neighbors. On the other hand, the largest point-local game is $\{4, 5, 6, 7\}$.

(b) This game has many actions but the size of the largest point-local game N' is 2.

Figure 1: Examples of cell decomposition on the outcome frequency space Δ_3.

To see that $v_{i,j}$ are well defined, notice that $\ell_j - \ell_i = S_j^\top v_{j,i} - S_i^\top v_{i,j} = -(\ell_i - \ell_j)$.

The algorithm, as its name suggests, is similar to Exp3 due to Auer et al. (2002). The main difference is how the weights are updated; since we do not have access to losses of chosen actions, we need to do something else. To have low regret, we do not actually need unbiased estimates of the individual losses, it is enough to ensure that the expected difference of the updates of two actions equals to the difference of the losses at each time step.

The above goal is achieved by randomly choosing two actions: one that will be chosen by the learner (I_t) and one that it will be compared against (I'_t). We choose these two actions independently based on the same distribution $p_i(t)$ given by the algorithm the usual way. Then, the update after observing the feedback is given as[3]:

$$\hat{\ell}_i(t) = \left(\frac{\mathbb{I}\{I_t = i\}}{p_i(t)} v_{i,I'_t} - v_{I_t,i} \right)^\top S_{I_t} e_t \,, \tag{1}$$

where $v_{i,i} = 0$. Recall that e_t is defined as the basis vector corresponding to the outcome J_t chosen by the opponent at time step t. Note that while e_t is not observed, $S_{I_t} e_t$ is precisely the information that is observed by the learner at time step t. Pseudocode of the algorithm is shown in Algorithm 1.

To see why the above update makes sense we prove the following statement:

Lemma 8 *For every time step t and actions i and j,*

$$\mathbb{E}_t[\hat{\ell}_i(t) - \hat{\ell}_j(t)] = \mathcal{L}(i, J_t) - \mathcal{L}(j, J_t) \,,$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | I_1, \ldots, I_{t-1}]$ is the conditional expectation given all the choices of the algorithm up to time step $t-1$.

3. This update is very similar to that of Foster and Rakhlin (2012).

Algorithm 1: LocalExp3

Parameters: γ, η;
Initialization: Construct $v_{i,j}$, S_i from $(\mathcal{L}, \mathcal{H})$, $w_i \leftarrow 1$;
for *t=1:T* **do**
  $p_i \leftarrow (1-\gamma)\frac{w_i}{\sum_j w_j} + \frac{\gamma}{N} \quad i = 1, \ldots, N$;
  Draw actions I and I' with distribution p_i;
  Choose action I;
  Receive feedback vector g;
  for *i=1:N* **do**
    $\hat{\ell}_i \leftarrow \left(\frac{\mathbb{I}\{I=i\}}{p_i} v_{i,I'} - v_{I,i}\right)^\top g$;
    Update weights $w_i \leftarrow w_i \exp(-\eta \hat{\ell}_i)$;
  end
end

Proof

$$\begin{aligned}
\mathbb{E}_t[\hat{\ell}_i(t)] &= \frac{p_i(t)}{p_i(t)} \sum_{k=1}^{N} p_k(t) v_{i,k}^\top S_i e_t - \sum_{k=1}^{N} p_k(t) v_{k,i}^\top S_k e_t \\
&= \sum_{k=1}^{N} p_k(t) \left(v_{i,k}^\top S_i - v_{k,i}^\top S_k\right) e_t \\
&= \sum_{k=1}^{N} p_k(t) (\ell_i - \ell_k)^\top e_t \\
&= \mathcal{L}(i, J_t) - \sum_{k=1}^{N} p_k(t) \mathcal{L}(k, J_t)\,. \qquad (2)
\end{aligned}$$

Hence,

$$\mathbb{E}_t[\hat{\ell}_i(t) - \hat{\ell}_j(t)] = \mathcal{L}(i, J_t) - \mathcal{L}(j, J_t)\,.$$

■

Now we can prove that, with appropriately set parameters, LocalExp3 gives $\widetilde{O}(\sqrt{NT})$ expected regret.

Theorem 9 *Let $\mathcal{G} = (\mathcal{L}, \mathcal{H})$ be a point-local $N \times M$ partial-monitoring game. Let $V_{\max} = \max_{i,j} \|v_{i,j}\|_\infty$ and $L_{\max} = \max_{i,j} \mathcal{L}(i,j)$. For any outcome sequence of length T chosen by an opponent, algorithm* LocalExp3 *run with parameters*

$$\eta = \frac{1}{2V_{\max}} \sqrt{\frac{\log N}{TN}} \qquad\qquad \gamma = \sqrt{\frac{\log N}{TN}}\,,$$

achieves expected regret

$$\mathbb{E}[R_T] \le (2L_{\max} + 4V_{\max}) \sqrt{NT \log N}\,.$$

The proof of this theorem is rather technical and is almost identical to that of Exp3 and thus it is written in the appendix.

3.2. Playing in general games

Suppose now we have a game $\mathcal{G}$ that satisfies the local observability condition. Its actions are covered by K point-local games $\mathcal{G}_i$, with N_i actions ($i = 1, \dots, K$). Let N' denote $\max_i N_i$. In this section we present our algorithm GlobalExp3 and prove that it achieves $\widetilde{O}(\sqrt{N'T})$ expected regret against any outcome sequence. The algorithm plays a local game chosen based on the whereabouts of the outcome frequency vector. Thus, the algorithm has two main components: (1) explore and track the outcome frequency vector $q_t \in \Delta_M$, and (2) play in a local game that is believed to contain the optimal action.

To track the outcome frequency, in the preprocessing phase the algorithm chooses a subset of neighboring action pairs $\mathcal{M} = \{\{i_1, j_1\}, \dots, \{i_{|\mathcal{M}|}, j_{|\mathcal{M}|}\}\}$. The only constraint on $\mathcal{M}$ is that the set of loss difference vectors $\{\ell_i - \ell_j \ : \ \{i,j\} \in \mathcal{M}\}$ should be linearly independent and span the subspace generated by *all* loss differences. This way the size of $\mathcal{M}$ is at most $M - 1$. If it is less than $M - 1$, it means that some directions in Δ_M are irrelevant in terms of deciding which action is optimal.

Let the exploring parameter β be chosen later. In every round, the algorithm decides to explore with probability $2|\mathcal{M}|\beta$. Then, it chooses a pair P_t from $\mathcal{M}$ uniformly and finally one of the actions $I_t \in P_t$ from the chosen pair, also uniformly. Then, the difference estimates for every $\{i,j\} \in \mathcal{M}$ are updated as:

$$\hat{d}_{i,j}(t) = \frac{\mathbb{I}\{P_t = \{i,j\}\}}{\beta} \left(\mathbb{I}\{I_t = i\} v_{i,j}^\top S_i - \mathbb{I}\{I_t = j\} v_{j,i}^\top S_j \right) e_t \,,$$
$$\hat{f}_{i,j}(t) = \frac{(t-1)\hat{f}_{i,j}(t-1) + \hat{d}_{i,j}(t)}{t} \,.$$

Finally, our estimate $\hat{q}(t)$ for the outcome frequency vector is

$$\hat{q}(t) = \begin{pmatrix} \ell_{i_1}^\top - \ell_{j_1}^\top \\ \vdots \\ \ell_{i_{|\mathcal{M}|}}^\top - \ell_{j_{|\mathcal{M}|}}^\top \\ \mathbf{1}^\top \end{pmatrix}^\dagger \begin{pmatrix} \hat{f}_{i_1,j_1}(t) \\ \vdots \\ \hat{f}_{i_{|\mathcal{M}|},j_{|\mathcal{M}|}}(t) \\ 1 \end{pmatrix}$$

where $\dagger$ denotes the pseudo-inverse. In the following, first we prove that $\hat{q}(t)$ is an unbiased estimate of the true outcome frequency q_t. Then, we prove that the true frequency is within a constant confidence interval from the estimate, given that t is large enough.

Lemma 10 *For any pair $\{i,j\} \in \mathcal{M}$,*

$$\mathbb{E}[(\ell_i - \ell_j)^\top \hat{q}(t)] = (\ell_i - \ell_j)^\top q_t \,.$$

Proof First we derive $\mathbb{E}[\hat{d}_{i,j}(t)]$:

$$\mathbb{E}[\hat{d}_{i,j}(t)] = \frac{2\beta}{\beta} \left(\frac{1}{2} v_{i,j}^\top S_i - \frac{1}{2} v_{j,i}^\top S_j \right) e_t = (\ell_i - \ell_j)^\top e_t \,.$$

Now it easily follows that

$$\mathbb{E}[\hat{f}_{i,j}(t)] = \frac{1}{t}\sum_{s=1}^{t}\hat{d}_{i,j}(s) = (\ell_i - \ell_j)^\top \frac{1}{t}\sum_{s=1}^{t} e_s = (\ell_i - \ell_j)^\top q_t\,.$$

Finally,

$$\begin{aligned}
\mathbb{E}[(\ell_i - \ell_j)^\top \hat{q}(t)] &= (\ell_i - \ell_j)^\top \begin{pmatrix} \ell_{i_1}^\top - \ell_{j_1}^\top \\ \vdots \\ \ell_{i_{|\mathcal{M}|}}^\top - \ell_{j_{|\mathcal{M}|}}^\top \\ \mathbf{1}^\top \end{pmatrix}^\dagger \mathbb{E}\left[\begin{pmatrix} \hat{f}_{i_1,j_1}(t) \\ \vdots \\ \hat{f}_{i_{|\mathcal{M}|},j_{|\mathcal{M}|}}(t) \\ 1 \end{pmatrix}\right] \\
&= (\ell_i - \ell_j)^\top \begin{pmatrix} \ell_{i_1}^\top - \ell_{j_1}^\top \\ \vdots \\ \ell_{i_{|\mathcal{M}|}}^\top - \ell_{j_{|\mathcal{M}|}}^\top \\ \mathbf{1}^\top \end{pmatrix}^\dagger \begin{pmatrix} \ell_{i_1}^\top - \ell_{j_1}^\top \\ \vdots \\ \ell_{i_{|\mathcal{M}|}}^\top - \ell_{j_{|\mathcal{M}|}}^\top \\ \mathbf{1}^\top \end{pmatrix} q_t \\
&= (\ell_i - \ell_j)^\top q_t\,,
\end{aligned}$$

where the last equality follows from the fact that $A^\dagger A$ is the orthogonal projection to the rowspace of A. ■

In the next lemma we show that the real outcome frequency q_t is in a parallelotope around $\hat{q}(t)$ with high probability.[4]

Lemma 11 *Given confidence parameter δ and width ϵ, with probability at least $1-\delta$, for all $t \geq \frac{4V_{\max}^2 \log(2T|\mathcal{M}|/\delta)}{3\epsilon^2\beta}$, q_t lies within the parallelotope defined by*

$$|(\ell_i - \ell_j)^\top(\hat{q}(t) - q_t)| \leq \epsilon \qquad\qquad \{i,j\} \in \mathcal{M}\,.$$

Proof We start as follows:

$$(\ell_i - \ell_j)^\top(\hat{q}(t) - q_t) = \hat{f}_{i,j}(t) - \mathbb{E}[\hat{f}_{i,j}(t)]\,.$$

With the help of Bernstein's inequality (see *e.g.*, Boucheron et al. (2003, Theorem 3)), we write

$$\begin{aligned}
P(|\hat{f}_{i,j}(t) - \mathbb{E}[\hat{f}_{i,j}(t)]| > \epsilon) &\leq 2\exp\left(-\frac{\epsilon^2/2}{\frac{\mathrm{Var}(\hat{d}_{i,j}(s))}{t} + \frac{V_{\max}\epsilon}{3\beta t}}\right) \\
&\leq \exp\left(-\frac{3\epsilon^2\beta t}{4V_{\max}^2}\right)\,.
\end{aligned}$$

Letting $t \geq \frac{4V_{\max}^2 \log(2T|\mathcal{M}|/\delta)}{3\epsilon^2\beta}$ and rearranging gives

$$P(|\hat{f}_{i,j}(t) - \mathbb{E}[\hat{f}_{i,j}(t)]| > \epsilon) \leq \frac{\delta}{|\mathcal{M}|T}\,.$$

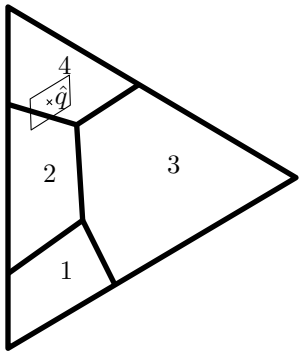


Figure 2: A game with two point-local games. The two point-local games are $\{1, 2, 3\}$ and $\{2, 3, 4\}$. The current estimate for q_t is $\hat{q}$ with the confidence parallelotope around it. The elements of $\mathcal{M}$ are $\{\{1, 2\}, \{2, 3\}\}$ and thus the edges of the parallelotope are parallel to the corresponding boundaries of $\mathcal{C}_1 \cap \mathcal{C}_2$ and $\mathcal{C}_2 \cap \mathcal{C}_3$. The picture suggests that the true outcome frequency must lie, with high confidence, in the second point-local game.

We get the statement of the lemma by using the union bound. ■

For an illustration of a global game and a confidence parallelotope, refer to Figure 2.

Now we are ready to explain the algorithm. In every round, first it decides if the round will be an exploratory round. If yes then it chooses a pair from $\mathcal{M}$ then an element of the chosen pair. It updates $\hat{q}$ and the next round begins. If the round is not an exploratory round then it plays in one of the local games. To decide which local game to play, the algorithm considers the following:

1. If the parallelotope is fully contained in the point-local game[5] played in the previous round, then play that local game again.

2. If not, then take any local game that contains the parallelotope. In this case, the local game's weights are reset.

Pseudocode of the algorithm is written in Algorithm 2. To prove a regret bound for the algorithm, we first show that, with high probability, GlobalExp3 does not change local games very often. To move toward this statement, we show the following:

Lemma 12 *For any game $\mathcal{G}$ there exists an $\epsilon_{\mathcal{G}} > 0$ such that for any $q \in \Delta_M$, the ball*

$$\{x \in \Delta_M \ : \ \|q - x\| \leq \epsilon_{\mathcal{G}}\}$$

is completely contained in at least one of the point local games.

4. Note that this parallelotope is "open" if $|\mathcal{M}| < M - 1$ but this fact does not affect our analysis.
5. We slightly abuse our notion of point-local games when we say that the parallelotope is contained in a point-local game. What we mean is that the parallelotope is contained in the union of the cells of the actions in the point-local game. We denote this union of cells by $\mathcal{C}_{\mathcal{G}_i} = \bigcup_{j \in \mathcal{G}_i} \mathcal{C}_j$.

Algorithm 2: GlobalExp3

Initialization: Construct local games $\mathcal{G}_i$, set $\epsilon_{\mathcal{G}}, \beta$, construct $\mathcal{M}$, $\hat{q} = 0$, $\hat{f}_{i,j} = 0$;
Initialization of local games: weights $w_{1,1}, \ldots, w_{1,N_1}; \ldots; w_{K,1}, \ldots, w_{K,N_K}$ to 1, set current local game $k = 1$, parameters η_i, γ_i;
for *t=1,...,T* **do**
  Draw X by Bernoulli with parameter $2|\mathcal{M}|\beta$;
  if *X=1* **then**
    Draw an action pair A from $\mathcal{M}$ uniformly;
    Chose action I from A uniformly;
    Receive feedback vector g;
    for $\{i,j\} \in \mathcal{M}$ **do**
      $\hat{d}_{i,j} \leftarrow \frac{\mathbb{I}\{A=\{i,j\}\}}{\beta}\Big(\mathbb{I}\{I=i\}v_{i,j}^\top g - \mathbb{I}\{I=j\}v_{j,i}^\top g\Big)$;
    end
  else
    if parallelotope$(\hat{q}, \epsilon_{\mathcal{G}}/6) \not\subset \mathcal{C}_{\mathcal{G}_k}$ **then**
      $k \leftarrow$ any local game that contains parallelotope$(\hat{q}, \epsilon_{\mathcal{G}})$;
      $w_{k,1} = \cdots = w_{k,N_k} \leftarrow 1$;
    end
    Play one round of LocalExp3 in $\mathcal{G}_k$;
  end
  for $\{i,j\} \in \mathcal{M}$ **do**
    $\hat{f}_{i,j} \leftarrow \frac{(t-1)\hat{f}_{i,j} + \hat{d}_{i,j}}{t}$;
  end
  $\hat{q} \leftarrow D^\dagger \bar{f}$ where $D = (\ell_{i_1} - \ell_{j_1}, \ldots, \ell_{i_{|\mathcal{M}|}} - \ell_{j_{|\mathcal{M}|}}, \mathbf{1})^\top$ and $\bar{f} = (\hat{f}_{i_1,j_1}, \ldots, \hat{f}_{i_{|\mathcal{M}|},j_{|\mathcal{M}|}})^\top$;
end

Proof The statement is trivial with $\epsilon_{\mathcal{G}}$ being the minimum distance between any two non-neighboring cells. ■

Now we show that the outcome frequency can not change too fast.

Lemma 13 *For any outcome sequence of length T, the total variation of $(q_t)_t$ is of $O(\log T)$.*

Proof For any time step t we have $\|q_{t+1} - q_t\| = \|e_{t+1} - q_t\|/(t+1) \le 1/(t+1)$. Thus,

$$\sum_{s=1}^{T} \|q_s - q_{s-1}\| \le \sum_{s=1}^{T} \frac{1}{s} \le \log T + 1\,. \qquad \blacksquare$$

Now we are ready to show the key lemma:

Lemma 14 *The number of times* GlobalExp3 *changes between local games is logarithmic in T.*

Proof We change local games when the parallelotope of width $\epsilon_{\mathcal{G}}/6$ goes out of the current local game. In this case we switch to a local game that contains the bigger, $\epsilon_{\mathcal{G}}$-wide

parallelotope. Thus, the next switch to happen, $\hat{q}$ needs to travel $\epsilon_{\mathcal{G}}/6$ far. But since we make sure that q_t and $\hat{q}(t)$ are, if confidence intervals do not fail, at most $\epsilon_{\mathcal{G}}/6$ far away from each other for high enough t, this also means that q_t has to travel at least $\epsilon_{\mathcal{G}}/6$ before the next switch. By Lemma 13 we get that the number of switches is at most $\frac{6}{\epsilon_{\mathcal{G}}}\log T$. ■

Now we state and prove the main theorem of the paper.

Theorem 15 *Given a finite partial monitoring game $\mathcal{G} = (\mathcal{L}, \mathcal{H})$, algorithm* GlobalExp3 *with appropriately set parameters achieves expected regret*

$$\mathbb{E}[R_T] \leq 1 + 24\frac{V_{\max}}{\epsilon_{\mathcal{G}}}\sqrt{|\mathcal{M}|T\log(2T^2/|\mathcal{M}|)} + \sqrt{\frac{6}{\epsilon_{\mathcal{G}}}}(2L_{\max} + 4V_{\max})\sqrt{N'T\log N'\log T}\,.$$

Proof Using $\epsilon = \epsilon_{\mathcal{G}}/6$ in Lemma 11 we let $\mathcal{E}$ denote the event that any confidence interval fails after time step $t_0 = \frac{72V_{\max}^2\log(2TM/\delta)}{\epsilon_{\mathcal{G}}^2\beta}$. From now on we assume the event $\mathcal{E}^c$.

For every local game $1 \leq i \leq K$, let T_i be the last time step when $q_t \in \mathcal{C}_{\mathcal{G}_i}$. Due to the construction of the confidence parallelotopes, the local game $\mathcal{G}_i$ is never played again from time step $T_i + 1$ on. Thus, we can investigate how much regret local game i produces by looking at its regret up to time step T_i. The optimal action(s) up to time step T_i must be in $\mathcal{G}_i$ and thus the regret bound of Theorem 9 can be applied to upper bound the regret of GlobalExp3.

Let Z_i denote the (random) number of times the algorithm switches to local game i after time step t_0. For every "epoch" $1 \leq l \leq Z_i$, let $R^{(i,l)}$ denote the regret of local game i in epoch l. Then, the regret accumulated from all the local games can be written as $\sum_{i=1}^{K}\sum_{l=1}^{Z_i} R^{(i,l)}$. By Lemma 14 we know that the number of terms in the above expression is bounded. Furthermore, the terms themselves can be bounded with the help of Theorem 9. The expected regret of GlobalExp3 comes from (1) event $\mathcal{E}$, (2) the first t_0 time steps, (3) number of times an exploratory action is taken, and (4) the regret of the local games:

$$\mathbb{E}[R_T] \leq \delta T + \frac{72V_{\max}^2\log\frac{2T|\mathcal{M}|}{\delta}}{\epsilon_{\mathcal{G}}^2\beta} + 2|\mathcal{M}|\beta T + \sqrt{\frac{6}{\epsilon_{\mathcal{G}}}}(2L_{\max} + 4V_{\max})\sqrt{N'T\log N'\log T}\,.$$

Setting $\delta = 1/T$ and $\beta = C_1 V_{\max}\sqrt{\log 2T^2|\mathcal{M}|}/(\epsilon_{\mathcal{G}}\sqrt{T|\mathcal{M}|})$ we get the desired result. ■

4. Discussion

In this paper we presented and analyzed an algorithm that achieves $\widetilde{O}(\sqrt{N'T})$ expected regret on locally observable non-degenerate partial monitoring games. This bound is substantially better than that of the previous state-of-the art algorithm. The main benefit of this new bound is that now we know that the expected regret does not directly depend on the number of actions, but rather, through the structure of the game, on the size of the largest point-local game. As an extreme example, one can think of a game with a large number of actions but $N' = 2$. In our solution we used the intuitive idea that the algorithm should not randomize when choosing which local game to play in. Indeed, randomization is usually needed when multiple actions are close to being optimal by a small margin. Due

to the construction of point-local games, this situation is avoided and thus we have the opportunity to play a point-local game of choice without randomizing.

One may notice that our bound contains the value $\epsilon_{\mathcal{G}}$ in the denominator. This value depends on the structure of the game and can get very small in some cases. One might also think that with more actions, ϵ_G decreases and thus N gets in the bound implicitly. However, there exist game instances with many actions and large ϵ_G. For an example consider cell decomposition in which every corner of the probability simplex has a local game with many actions, but these local games are far away from each other.

We would also like to note that a value related to ϵ_G naturally must appear in the regret upper bound. To understand why, consider a game where two non-neighboring actions do not satisfy the local observability condition, but their cells are very close to each other (and thus ϵ_G is small). Imagine these cells moving towards each other. When the gap becomes zero, the actions become neighbors and the game becomes non-locally observable; thus the regret will scale with $T^{2/3}$. Hence, as ϵ_G shrinks, the regret must go up.

The bound also shows a dependence on the number of outcomes (M). We conjecture that this dependence can be lifted with a more sophisticated way of tracking q_t. Our method of devoting some rounds to exploration seems suboptimal. Improving the bound in this aspect remains future work.

As a final remark we note a fact that we found interesting. If we use the algorithm LocalExp3 on a bandit game (we can because it is a point-local locally observable game), the algorithm does not reduce to Exp3. This is due to the fact that the expectation of the updates are offset by the value $\sum_{k=1}^{N} p_k(t)\mathcal{L}(k, J_t)$, which is in turn the expected loss of the algorithm at time step t. This "centralized" update might even improve upon the performance of Exp3 because it makes the absolute values of the updates smaller.

Acknowledgments

The author thanks the anonymous reviewers for their insightful comments and constructive suggestions. This research was supported in part by DARPA grant MSEE FA8650-11-1-7156.

References

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, 2009.

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

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. In *ALT*, pages 224–238, 2010.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. *Journal of Machine Learning Research - Proceedings Track (COLT)*, 19:133–154, 2011.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240, 2003.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Math. Oper. Res.*, 31(3):562–580, 2006.

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

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, 1994.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *NIPS*, pages 684–692, 2011.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *COLT/EuroCOLT*, pages 208–223, 2001.

V. G. Vovk. Aggregating strategies. In *COLT*, pages 371–386, 1990.

Appendix A. Proof of Theorem 9

Proof [Theorem 9] For every variable x used by the algorithm, we denote by $x(t)$ the value of the variable at the end of time step t. The proof of this regret bound is almost identical to that of Auer et al. (2002). We begin by lower bounding $W(T) = \sum_{i=1}^{N} w_i(T)$.

$$W(T) = \sum_{i=1}^{N} w_i(T) \geq w_{i^*}(T) = \exp\left(-\eta \sum_{t=1}^{T} \hat{\ell}_{i^*}(t)\right)$$

for any action i^*. We continue by upper bounding the term $W(t)/W(t-1)$ for any $= 1, \ldots, T$.

$$\begin{aligned}
\frac{W(t)}{W(t-1)} &= \frac{\sum_{i=1}^{N} w_i(t-1)\exp(-\eta\hat{\ell}_i(t))}{W(t-1)} \\
&\leq 1 - \eta \sum_{i=1}^{N} \frac{w_i(t-1)}{W(t-1)}\hat{\ell}_i(t) + \eta^2 \sum_{i=1}^{N} \frac{w_i(t-1)}{W(t-1)}\hat{\ell}_i(t)^2 \qquad (3) \\
&= 1 - \eta \sum_{i=1}^{N} \frac{p_i(t) - \gamma/N}{1-\gamma}\hat{\ell}_i(t) + \eta^2 \sum_{i=1}^{N} \frac{p_i(t) - \gamma/N}{1-\gamma}\hat{\ell}_i(t)^2 \\
&\leq \exp\left(-\eta \sum_{i=1}^{N} \frac{p_i(t) - \gamma/N}{1-\gamma}\hat{\ell}_i(t) + \eta^2 \sum_{i=1}^{N} \frac{p_i(t) - \gamma/N}{1-\gamma}\hat{\ell}_i(t)^2\right).
\end{aligned}$$

In (3) we used that $e^x \leq 1 + x + x^2$ if $|x| \leq 1$, and thus we must ensure this condition. For this, we need to set $\gamma \geq 2\eta V_{\max}$, where $V_{\max} = \max_{i,j} \|v_{i,j}\|_\infty$. We will make sure to

satisfy this condition when rendering values to the parameters. Now we merge the above lower and upper bound using telescopic sum:

$$\begin{aligned}\frac{\exp\left(-\eta\sum_{t=1}^{T}\hat{\ell}_{i^*}(t)\right)}{N} &\leq \frac{W(T)}{W(0)} \\ &\leq \exp\left(-\eta\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{p_i(t)-\gamma/N}{1-\gamma}\hat{\ell}_i(t)+\eta^2\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{p_i(t)-\gamma/N}{1-\gamma}\hat{\ell}_i(t)^2\right)\end{aligned}$$

Taking logarithm and expectation of both sides we get

$$-\eta\mathbb{E}\left[\sum_{t=1}^{T}\hat{\ell}_{i^*}(t)\right]-\log N \leq \mathbb{E}\left[-\eta\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{p_i(t)-\gamma/N}{1-\gamma}\hat{\ell}_i(t)+\eta^2\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{p_i(t)-\gamma/N}{1-\gamma}\hat{\ell}_i(t)^2\right]$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}p_i(t)\mathbb{E}_t\left[\hat{\ell}_i(t)-\hat{\ell}_{i^*}(t)\right]\right] \leq \frac{\log N}{\eta}+\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t\left[\sum_{i=1}^{N}\frac{\gamma}{N}\hat{\ell}_i(t)+\eta p_i(t)\hat{\ell}_i(t)^2\right]\right] \tag{4}$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}p_i(t)\left(\mathcal{L}(i,J_t)-\mathcal{L}(i^*,J_t)\right)\right] \leq \frac{\log N}{\eta}+\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t\left[\sum_{i=1}^{N}\frac{\gamma}{N}\hat{\ell}_i(t)+\eta p_i(t)\hat{\ell}_i(t)^2\right]\right] \tag{5}$$

$$\mathbb{E}[R(T)] \leq \frac{\log N}{\eta}+2L_{\max}\gamma TN+4V_{\max}^2\eta TN\,. \tag{6}$$

In (4) we used the tower rule for conditional expectation, in (5) we used Lemma 8, and in (6) we used the following two bounds:

$$\begin{aligned}\mathbb{E}_t[\hat{\ell}_i(t)] &\leq 2L_{\max} && \text{see (2)} \\ \mathbb{E}_t[\hat{\ell}_i(t)^2] &\leq \frac{4V_{\max}^2}{p_i}\,.\end{aligned}$$

Now, setting $\eta=\frac{1}{2V_{\max}}\sqrt{\frac{\log N}{TN}}$ and $\gamma=2V_{\max}\eta=\sqrt{\frac{\log N}{TN}}$ we get the desired result. ■