# Partial monitoring with side information

Gábor Bartók and Csaba Szepesvári

University of Alberta
Edmonton, Canada

**Abstract.** In a partial-monitoring problem in every round a learner chooses an action, simultaneously an opponent chooses an outcome, then the learner suffers some loss and receives some feedback. The goal of the learner is to minimize his (unobserved) cumulative loss. In this paper we explore a variant of this problem where in every round, before the learner makes his decision, he receives some side-information. We assume that the outcomes are generated randomly from a distribution that is influenced by the side-information. We present a "meta" algorithm scheme that reduces the problem to that of the construction of an algorithm that is able to estimate the distributions of observations while producing confidence bounds for these estimates. Two specific examples are shown for such estimators: One uses linear estimates, the other uses multinomial logistic regression. In both cases the resulting algorithm is shown to achieve $\widetilde{O}(\sqrt{T})$ minimax regret for locally observable partial-monitoring games.

## 1 introduction

Partial monitoring is a framework to model online learning games with arbitrary feedback structure. In every time step, a learner chooses an *action* and simultaneously an opponent chooses an *outcome*. Then, the learner suffers some loss and receives some feedback, both of which are deterministic functions of the action and the outcome. The loss and feedback functions are both known to the learner and the opponent and together they define the *partial monitoring game*. The goal of the learner is to keep his cumulative loss as low as possible. His performance is measured in terms of the *regret*: the learner's excess cumulative loss compared to that of the best fixed action in hindsight.

Canonical examples of partial-monitoring include *product testing* and *dynamic pricing*. In the case of product testing, the learner has to decide to test or not test products arriving on a production line. The learner receives feedback about the quality of the product only if he decided to test the product. On the other hand, he suffers a constant loss in every time step when either a good product was tested (unable to sell, e.g., when the test means the destruction of the product) or a bad product was not tested (complaining costumers). In the case of dynamic pricing, a vendor (learner/he) sets the price of a product while the consumer (opponent/she) secretly chooses a maximum price she is willing to buy the product for. In case the sale price is below the consumer-chosen price, the product is sold. The information received by the learner is the single bit whether this happens. The loss suffered in a round when the product is sold is the difference between the consumer-chosen prices and the sale price, while in a round when the product is not sold a fixed storage cost is incurred.

In this paper we extend the basic partial monitoring problem to allow the learner to use some *side information* to make a more informed decision. For example, in product testing, before deciding about whether to use a potentially destructive testing procedure the learner can take a look at the product. Similarly, in dynamic pricing, the learner may use information available about the customer (gender, age, etc.) for determining a more competitive price. Formally, the assumption is that in each round the learner receives the so-called side information (sometimes also called "a context") before making a decision. The side information is not subject to any restrictions, but in this paper we assume that the outcome for the given round is a *stochastically* function of the side information shown to the learner. Then, instead of competing with the single best action, the learner competes with the oracle that knows the mapping that maps the side information to the outcome distributions and who makes optimal decisions given this knowledge.

### 1.1 Related work

The model of partial monitoring was introduced by Piccolboni and Schindelhauer [2001]. They designed the algorithm FeedExp and showed for any game, either the worst-case expected regret is linear in the time horizon $T$, or the algorithm achieves expected regret of $O(T^{3/4})$ for any outcome sequence. This upper bound was later improved to $O(T^{2/3})$ by Cesa-Bianchi et al. [2006]. In the same paper, Cesa-Bianchi et al. show that there exists a game whose *minimax regret*—the worst case regret of the best possible algorithm—scales as $\Omega(T^{2/3})$. However, they noted that some games enjoy minimax regret growth rate of $\Theta(\sqrt{T})$, and posed the problem of determining exactly which games have minimax regret rate better than $\Theta(T^{2/3})$. This problem was solved in the works of Bartók et al. [2011] against stochastic opponents, while by providing a new algorithm Foster and Rakhlin [2011] showed that the classification of games worked out by Bartók et al. [2011] continues to hold even against adversarial opponents. According to the solution, partial-monitoring games with a finite number of actions and outcomes can be classified into four categories based on the growth rate of the minimax regret: *trivial* games with minimax regret 0, *easy* games with minimax regret[1] of $\widetilde{\Theta}(\sqrt{T})$, *hard* games with minimax regret $\Theta(T^{2/3})$, and *hopeless* games with linear minimax regret. The condition that separates easy games from hard games is the *local observability condition* (see Definition 2). In the bandit literature learning with side-information has been considered before under various conditions, see Auer [2003], Dudík et al. [2011] and references therein, while Helmbold et al. [2000] considered a special case of our framework when both the number of actions and outcomes is two, with one action revealing the actual

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

outcome, while the other action not yielding any information about the outcome, the hidden relationship between the side information and hidden information is deterministic and the loss is the zero-one loss.

## 2 Problem definition

An instance of a partial-monitoring game with side-information is described by the tuple $\mathbf{G} = (\mathbf{L}, \mathbf{H}, \mathcal{F})$, where $\mathbf{L} \in \mathbb{R}^{N \times M}$ is the *loss matrix*, $\mathbf{H} \in \Sigma^{N \times M}$ is the *feedback matrix* ($\Sigma$ is the set of feedback symbols), and $\mathcal{F} \subseteq \{f \mid f : \mathcal{X} \to \Delta_M\}$ is a subset of all functions that map elements from some side-information set $\mathcal{X}$ to the set of outcome distributions. For convenience, we assume that $\max_{i \in \underline{N}, j \in \underline{M}}(\mathbf{L}_{i,j}) - \min_{i \in \underline{N}, j \in \underline{M}}(\mathbf{L}_{i,j}) \leq 1$, where for a natural number $n \in \mathbb{N}$ we used $\underline{n}$ to denote the set $\{1, 2, \ldots, n\}$. The partial-monitoring game proceeds in turns. Before the first turn, both the learner and the opponent is given $\mathbf{G}$ and the opponent secretly chooses a function $f \in \mathcal{F}$. In turn $t$ ($t = 1, 2, \ldots$), first the learner receives the side-information $x_t \in \mathcal{X}$. Then, the learner chooses an action $I_t \in \underline{N}$, while at the same time the opponent draws an outcome $J_t$ from the distribution $f(x_t)$. No stochastic assumption is made about the side information sequence, $\{x_t\}$ and, in fact, we also allow $x_t$ to be chosen based on the history $\mathcal{H}_{t-1} = (x_1, I_1, J_1, \ldots, x_{t-1}, I_{t-1}, J_{t-1})$. After the learner and the opponent made their decisions, the learner receives the feedback $\mathbf{H}_{I_t, J_t}$ and suffers the loss $\mathbf{L}_{I_t, J_t}$. It is important to emphasize that the loss is not revealed to the learner.

The goal of the learner is to minimize his cumulative loss given the knowledge of the game $\mathbf{G}$. His performance is measured in terms of the regret, defined as the excess cumulative loss he suffers as compared to the expected cumulative loss of the oracle that knows $f$ and chooses the action with the smallest expected loss as a function of the side-information in every round. In other words,

$$R_T = \sum_{t=1}^{T} \mathbf{L}_{I_t, J_t} - \min_{g \in \underline{N}^{\mathcal{X}}} \sum_{t=1}^{T} \mathbb{E}[\mathbf{L}_{g(x_t), J_t} | \mathcal{H}_{t-1}, x_t] \, .$$

## 3 Preliminaries

In this section we introduce the necessary notations and definitions that we will need. Most of the definitions presented here are taken from Bartók et al. [2011].

Let $\mathbf{G} = (\mathbf{L}, \mathbf{H}, \mathcal{F})$ be a partial-monitoring game. For an action $i$, the column vector $\ell_i$ consisting of the elements of the $i^{\text{th}}$ row of $\mathbf{L}$ is called the *loss vector* of action $i$. Let the probability simplex of dimension $n$ be denoted by $\mathcal{K}_n \subseteq \mathbb{R}^n$. Thus, the set of all outcome distributions is $\mathcal{K}_M$. It is easy to see that the expected loss of action $i$ at time step $t$ given the past and $x_t$ equals $\mathbb{E}[\mathbf{L}_{i,J_t} | H_{t-1}, x_t] = \ell_i^\top f(x_t)$.

For an action $i$, let the cell of $i$ be the set of outcome distributions under which action $i$ is optimal:

$$\mathcal{C}_i = \{p \in \mathcal{K}_M \mid \forall j \in \underline{N} : (\ell_i - \ell_j)^\top p \leq 0\} \, .$$

It is easy to see that for every $i \in \underline{N}$, $\mathcal{C}_i$ is either empty or a closed convex polytope, with $\bigcup_{i \in \underline{N}} \mathcal{C}_i = \mathcal{K}_M$. We call $\mathbb{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_N\}$ the *cell decomposition* of $\mathcal{K}_M$. For clarity of presentation, in this paper we only deal with games that are non-degenerate: for every action $i$, $\mathcal{C}_i$ is $M-1$ dimensional and for $i \neq j$, $\mathcal{C}_i \neq \mathcal{C}_j$. We remark that our results generalize to degenerate games, but the algorithm and its analysis are somewhat more involved.

For an action $i \in \underline{N}$, we define the *signal matrix* of $i$ as follows:

**Definition 1.** *For an action $i$, let $\alpha_1, \alpha_2, \dots, \alpha_{\sigma_i} \in \Sigma$ be the distinct symbols in the $i^{\text{th}}$ row of the feedback matrix $\mathbf{H}$. The signal matrix $S_i \in \{0,1\}^{\sigma_i \times M}$ is defined as the incidence matrix of the $i^{\text{th}}$ row of the feedback matrix $\mathbf{H}$:*

$$(S_i)_{k,l} = \mathbb{I}_{\{H_{i,l} = \alpha_k\}} \, .$$

An important property of the signal matrix $S_i$ is that if $p \in \mathcal{K}_M$ is the outcome distribution chosen by the opponent then $S_i p$ is the probability distribution over the set of observations $\{\alpha_1, \dots, \alpha_k\}$ induced by $p$ under action $i$. From now on, without loss of generality, we assume that the feedback at time step $t$ is presented as the unit vector corresponding to the received symbol $\mathbf{H}_{I_t,J_t}$. We shall denote this unit vector by $Y_t$.

If for two actions $i$ and $j$, $\dim(\mathcal{C}_i \cap \mathcal{C}_j) = M-2$ we say that $i$ and $j$ are *neighbors*. The set of neighboring action pairs is denoted by $\mathcal{N}$. Now we are ready to recall the *local observability condition* from Bartók et al. [2011]:

**Definition 2.** *Let $\{i,j\} \in \mathcal{N}$ be two neighboring actions. We say that $\{i,j\}$ is* locally observable *if $\ell_i - \ell_j \in \operatorname{Im}(S_i^\top) \oplus \operatorname{Im}(S_j^\top)$. The game is called* locally observable *(or we say that it satisfies the local observability condition) if every neighboring action pair is locally observable. For a pair of distinct action $\{i,j\} \in \mathcal{N}$, a pair of vectors, $v_{i,j}$, $v_{j,i}$ is called* observer vectors *for $\{i,j\}$ if*

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

If a neighboring action pair is locally observable then the local observability condition yields the existence of these observer vectors. From now on, for locally observable neighboring action pairs we shall fix such observer vectors. Note that the observer vectors are *not* uniquely defined. We will discuss good choices of the observer vectors later on.

## 4 The algorithm

Bartók et al. [2011] proved that if a game is locally observable then a minimax regret of $\widetilde{O}(\sqrt{T})$ is achievable against a stochastic opponent. Now we extend their result to partial monitoring with side-information. In particular, we show that the $\widetilde{O}(\sqrt{T})$ regret bound remains true in this richer model.

In this section we describe the algorithm scheme CBP-Side for "Confidence Bound Partial monitoring with Side-information" that when fed with a method that estimates the outcome distributions and their uncertainty defines a learning strategy. In Section 5 we give a bound on the expected regret as a function of how fast the uncertainty of the outcome distribution estimates decays. Then, in Section 6 we present two examples that illustrate how this general bound translates into actual regret bounds for two different classes of functions $\mathcal{F}$.

The algorithm is a generalization of the algorithm "Confidence Bound Partial monitoring" (CBP) from Bartók et al. [2012]. Pseudocode for the algorithm is given in Algorithm 1.

Throughout the algorithm, some statistics $\mathcal{S}$ is maintained that is used by the functions getObsEst and getConfWidth (which are left generic for now). The statistics might be the whole sequence of observations and actions up to time step $t-1$, or just some average of the observations and maybe the number of times each action was chosen. After receiving the side-information for time step $t$, estimates for the observation probabilities and their confidence widths are obtained by calling the functions getObsEst and getConfWidth. Then the algorithm calculates estimates of the loss differences (denoted by $\tilde{\Delta}_{i,j}$) for neighboring action pairs, along with their confidence widths $c_{i,j}$. If, for some pair $i, j \in \mathcal{N}$ the absolute value of the loss-difference estimate is greater than its confidence width, we know that, with high probability, $p_t = f(x_t)$ lies in the half space $\{p \in \mathbb{R}^M \mid \operatorname{sgn}(\tilde{\Delta}_{i,j})(\ell_i - \ell_j)^\top p \geq 0\}$. Thus, the intersection of all these half spaces and the probability simplex determines the convex polytope $\mathcal{K}_t$ that $p_t$ belongs to (with high probability), giving rise to the set of admissible actions $Q$. To compute this set the method getNeighbors computes $\mathcal{N}(t) = \{\{i,j\} \in \mathcal{N} : \mathcal{C}_i \cap \mathcal{C}_j \cap \operatorname{int}(K_t) \neq \emptyset\}$. Then, $Q = \cup\mathcal{N}(t)$. Finally, the action $I_t$ from $Q$ that has the greatest potential of reducing the confidence width for the next rounds is chosen and based on the information received the statistics $\mathcal{S}$ is updated.

## 5 Analysis of CBP-Side

In this section we provide an upper bound on the expected regret suffered by the algorithm on any given game with any plugged-in estimate and confidence width functions. Note that the upper bound contains the expectation of some random values that depend on the outcomes drawn randomly at every time step. In the next sections, we will see how these can be upper-bounded by some (small) deterministic quantities in some specific cases.

From now on, we use the convention that for any variable $v$, we denote by $v(t)$ the value assigned to $v$ in time step $t$.

**Theorem 1.** *Assume that there exist numbers $\delta_1, \delta_2, \ldots, \delta_T \in [0,1]$ and a norm $\|\cdot\|$ such that for every time step $t$ it holds that*

$$\mathbb{P}\left(\|\hat{q}_i(t) - S_i f(x_t)\| > w_i(t)\right) \leq \delta_t \tag{1}$$

*for every $i \in \underline{N}$. Then, the expected regret of* CBP-Side *on game $\mathbf{G} = (\mathbf{L}, \mathbf{H}, \mathcal{F})$ can be upper bounded as*

$$\mathbb{E}[R_T] \leq \sum_{t=1}^{T} N\delta_t + \sum_{t=1}^{T} \mathbb{E}\left[\min\left\{4NW_{I_t} w_{I_t}(t), 1\right\}\right],$$

**Algorithm 1** The algorithm CBP-Side

```
Input: $\mathbf{L}$, $\mathbf{H}$, $\alpha$
Calculate $\mathcal{P}$, $\mathcal{N}$, $v_{i,j}$, $W_k$
$\mathcal{S} \leftarrow$initStatistic()                                  {Some statistics as needed}
for $t = 1$ to $T$ do
    Receive side information $x_t$
    for each $i \in \underline{N}$ do
        $\tilde{q}_i \leftarrow$getObsEst($\mathcal{S}, x_t$)               {Observation distribution estimate}
        $w_i \leftarrow$getConfWidth($\mathcal{S}, x_t$)                                    {Confidence}
    end for
    for each $\{i, j\} \in \mathcal{N}$ do
        $\tilde{\Delta}_{i,j} \leftarrow v_{i,j}^\top \tilde{q}_i - v_{j,i}^\top \tilde{q}_j$                            {Loss diff. estimate}
        $c_{i,j} \leftarrow \|v_{i,j}\|_* w_i + \|v_{j,i}\|_* w_j$                               {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
    $\mathcal{N}(t) \leftarrow$ getNeighbors($\mathcal{P}, \mathcal{N}, \mathit{halfSpace}$)
    $Q \leftarrow \bigcup \mathcal{N}(t)$                                            {Admissible actions}
    Choose $I_t = \operatorname{argmax}_{i \in Q}(W_i w_i)$                        {$W_i = \max_j \|v_{i,j}\|_*$}
    Observe $Y_t$
    $\mathcal{S} \leftarrow$updateStatistic($\mathcal{S}, x_t, I_t, Y_t$)
end for
```

*where* $W_i = \max_j \|v_{i,j}\|_*$ *with* $\|\cdot\|_*$ *being the dual norm of* $\|\cdot\|$*.*

*Proof.* For any $i, j \in \underline{N}$ and $x \in \mathcal{X}$, let $\Delta_{i,j}(x)$ denote the expected loss difference of actions $i$ and $j$ given side-information $x$, written as $\Delta_{i,j}(x) \overset{\triangle}{=} (\ell_i - \ell_j)^\top f(x)$. Further, let $\Delta_i(x) \overset{\triangle}{=} \max_j \Delta_{i,j}(x)$ be the "gap" between the expected loss of action $i$ and that of an optimal action given side-information $x$. It is easy to see that the expected regret of an algorithm can be rewritten as $\mathbb{E}[R_T] = \sum_{t=1}^T \mathbb{E}[\Delta_{I_t}(x_t)]$. Let $\mathcal{E}_t$ be the event that some confidence width fails at time step $t$. Then, $\mathbb{E}[R_T] = \sum_{t=1}^T \mathbb{E}[\Delta_{I_t}(x_t)] \leq \sum_{t=1}^T N\delta_t + \sum_{t=1}^T \mathbb{E}[\Delta_{I_t}(x_t)\mathbb{I}_{\{\mathcal{E}_t^c\}}]$, where we used that $\Delta_i(x) \leq 1$. Thus, it remains to bound $\Delta_{I_t}(x_t)$ assuming that for all $i \in \underline{N}$, $\|\tilde{q}_i(t) - S_i f(x_t)\| \leq w_i(t)$ holds.

If $i$ and $j$ are in $\mathcal{N}(t)$ (that is, they are neighbors at time step $t$), then $\tilde{\Delta}_{i,j}(t)$ is a "good" approximation of $\Delta_{i,j}(x_t)$:

$$\begin{aligned}
|\Delta_{i,j}(x_t) - \tilde{\Delta}_{i,j}(t)| &= \left|(\ell_i - \ell_j)^\top f(x_t) - \left(v_{i,j}^\top \tilde{q}_i(t) - v_{j,i}^\top \tilde{q}_j(t)\right)\right| \\
&\leq \|v_{i,j}\|_* \, \|S_i f(x_t) - \tilde{q}_i(t)\| + \|v_{j,i}\|_* \|S_j f(x_t) - \tilde{q}_j(t)\| \\
&\leq \|v_{i,j}\|_* \, w_i(t) + \|v_{j,i}\|_* \, w_j(t) \\
&= c_{i,j}(t)\,.
\end{aligned} \tag{2}$$

We know from line 12 of the algorithm that if $\{i,j\} \in \mathcal{N}(t)$ then $\tilde{\varDelta}_{i,j}(t) \leq c_{i,j}$. This together with Equation (2) gives

$$\varDelta_{i,j}(x_t) \leq 2c_{i,j}\,. \tag{3}$$

Let $i^*$ be an optimal action at time step $t$ (that is, $\min_i \ell_i^\top f(x_t) = \ell_{i^*}^\top f(x_t)$). Then

$$\varDelta_{I_t,i^*}(t) = \sum_{s=1}^{r} \varDelta_{k_{s-1},k_s}(t)\,,$$

where $I_t = k_0, k_1, \ldots, k_r = i^*$ is a sequence of actions such that $\{k_{s-1}, k_s\} \in \mathcal{N}(t)$ for all $1 \leq s \leq r$. This sequence always exists thanks to how the algorithm constructs the set of admissible actions[2]. With the help of Equation (3) we get

$$\begin{aligned}\varDelta_{I_t,i^*}(t) &\leq 2\sum_{s=1}^{r} c_{k_{s-1},k_s}(t) = 2\sum_{s=1}^{r}\Big(\left\|v_{k_{s-1},k_s}\right\|_q w_{k_{s-1}}(t) + \left\|v_{k_s,k_{s-1}}\right\|_q w_{k_s}(t)\Big)\\ &\leq 4NW_{I_t}w_{I_t}(t)\,,\end{aligned}$$

where in the last line we used line 20 of the algorithm and the fact that $r \leq N$, thus finishing the proof. □

*Remark 1 (On the choice of the observer vectors $v_{i,j}$.).* We mentioned earlier that the choice of the observer vectors is not unique and thus we have some freedom in choosing them. Theorem 1 indicates that for different estimators, the best choice of the observer vectors might differ. In particular, it depends on the norm the estimate uses: to optimize the bound of Theorem 1, we should choose the vectors that minimize $\|v_{i,j}\|_*$. If the norm used is the 2-norm then there is a closed form solution for the best $v_{i,j}$:

$$\begin{pmatrix} v_{i,j} \\ -v_{j,i} \end{pmatrix} = \left(S_i^\top\ S_j^\top\right)^+ (\ell_i - \ell_j)\,,$$

where $A^+$ denotes the pseudo-inverse of the matrix $A$.

## 6 Examples

In this section we demonstrate the power of Theorem 1 through specific examples.

### 6.1 Linear side-information, least-squares estimate

In the first example, the side-information set is the probability simplex $\mathcal{K}_d$ of some dimension $d > 0$ while the function set $\mathcal{F}$ is the set of all linear maps where

[2] For a thorough proof of this statement, we refer the reader to Bartók et al. [2012].

the underlying matrix is a stochastic matrix of size $M \times d$. The estimator we use is regularized least squares. We introduce the following notations. For every action $i$, let $\theta_i^* = S_i K \in \mathbb{R}^{\sigma_i \times d}$, where $K$ is the matrix underlying the the linear map $f$ chosen by the opponent (thus, $f(x) = Kx$). Let $t_i(s)$ be the time step when action $i$ is chosen by the algorithm the $s^{\text{th}}$ time. Let $n_i(t)$ be the number of times action $i$ is chosen up to time step $t$. Then the regularized least squares estimator is defined by the equation

$$\tilde{\theta}_i(t) = \min_{\theta \in \mathbb{R}^{M \times d}} \sum_{s=1}^{n_i(t-1)} \left( Y_{t_i(s)} - \theta x_{t_i(s)} \right)^2 + \lambda_i \|\theta\|_2^2 \,.$$

For the closed form solution we define the matrices

$$X_{i,t} = \left( x_{t_i(1)} \; x_{t_i(2)} \; \cdots \; x_{t_i(n_i(t-1))} \right) , \quad \mathcal{Y}_{i,t} = \left( Y_{t_i(1)} \; Y_{t_i(2)} \; \cdots \; Y_{t_i(n_i(t-1))} \right) \,.$$

Then,

$$\tilde{\theta}_i(t) = \mathcal{Y}_{i,t} X_{i,t}^\top \left( \lambda_i I_d + X_{i,t} X_{i,t}^\top \right)^{-1} ,$$

where $I_d$ is the $d \times d$ identity matrix. Let $V_{i,t} = \lambda_i I_d + X_{i,t} X_{i,t}^\top$.

For some positive definite matrix $S$, let $\| \cdot \|_S$ denote the $S$-weighted 2-norm: $\|v\|_S^2 = v^\top S v$. In the rest of the paper, we will need a number of results, which, for the sake of completeness, we recite here.

**Theorem 2 (Abbasi-Yadkori et al. [2011, Theorem 1]).** *Let $\{F_t\}_{t=1}^\infty$ a filtration. Let $\{\eta_t\}_{t=1}^T$ be a real-valued stochastic process such that $\eta_t$ is $F_t$-measurable and $\eta_t$ is conditionally $R$-sub-Gaussian for some $R \geq 0$. Let $\{x_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that $x_t$ is $F_{t-1}$-measurable. Let $\lambda > 0$. For any $t \geq 0$, define*

$$V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top , \qquad\qquad S_t = \sum_{s=1}^t \eta_s x_s \,.$$

*Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,*

$$\|S_t\|_{V_t^{-1}}^2 \leq 2R^2 \log \left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right) \,.$$

**Theorem 3 (Abbasi-Yadkori and Szepesvári [2011, Theorem 1]).** *Let $(x_0, Y_1), \ldots, (x_t, Y_{t+1}), x_i \in \mathbb{R}^d, Y_i \in \mathbb{R}^n$ satisfy the linear model Assumption*[3] *A1 with some $L > 0$, $\Theta_* \in \mathbb{R}^{d \times n}$, $\operatorname{tr}(\Theta_*^\top \Theta_*) \leq S^2$ and let $\mathcal{F} = (\mathcal{F}_t)$ be the associated filtration. Consider the $\ell^2$-regularized least-squares parameter estimate $\hat{\Theta}_t$ with regularization coefficient $\lambda > 0$. Let*

$$V_t = \lambda I + \sum_{i=0}^{t-1} x_i x_i^\top$$

[3] Reciting this assumption is beyond the scope of this paper. In a nutshell, it says that $x_t$ and $Y_t$ are $\mathcal{F}_t$-measurable, $\mathbb{E}[Y_{t+1} | \mathcal{F}_t] = \Theta^\top x_t$ for some matrix $\Theta$, the noise $Y_{t+1} - \mathbb{E}[Y_{t+1} | \mathcal{F}_t]$ is componentwise sub-Gaussian with parameter $L$.

*be the regularized design matrix underlying the covariates. Define*

$$\beta_t(\delta) = \left( nL\sqrt{2\log\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}} + \lambda^{1/2}S \right)^2 .$$

*Then, for any $0 < \delta < 1$ and stopping time $N$, with probability at least $1-\delta$,*

$$\operatorname{tr}\left( (\hat{\Theta}_N - \Theta_*)^\top V_N (\hat{\Theta}_N - \Theta_*) \right) \le \beta_N(\delta) .$$

**Lemma 1 (Abbasi-Yadkori et al. [2011, Lemma 10]).** *Let $x_1, \ldots, x_t \in \mathbb{R}^d$ be such that for any $1 \le s \le t$, $\|x_s\|_2 \le L$. Let $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$ for some $\lambda > 0$. Then,*

$$\det(V_t) \le (\lambda + tL^2/d)^d .$$

In the following lemma $z_1, z_2, \ldots \in \mathbb{R}^d$ is an arbitrary sequence of $d$-dimensional vectors and $V_t = \lambda I + \sum_{s=1}^t z_s z_s^\top$ for some $\lambda > 0$.

**Lemma 2 (Abbasi-Yadkori and Szepesvári [2011, Lemma 10]).** *The following holds for any $t \ge 1$:*

$$\sum_{k=0}^{t-1} \min\left( \|z_k\|^2_{V_k^{-1}}, 1) \right) \le 2\log\frac{\det(V_t)}{\det(\lambda I)} .$$

*Further, when the covariates satisfy $\|z_t\| \le c_m, t \ge 0$ with some $c_m > 0$ w.p.1 then*

$$\log\frac{\det(V_t)}{\det(\lambda I)} \le (n+d)\log\frac{\lambda(n+d) + tc_m^2}{\lambda(n+d)} .$$

With the help of Theorem 1 of Abbasi-Yadkori and Szepesvári [2011] we get that for any $0 < \delta_t < 1$,

$$\operatorname{tr}((\tilde{\theta}_i(t) - \theta_i^*) V_{i,t} (\tilde{\theta}_i(t) - \theta_i^*)^\top) \le d^2 \left( \sqrt{2\log\frac{\det(V_{i,t})^{1/2}}{\delta_t \lambda_i^{d/2}}} + \sigma_i \lambda_i^{1/2} \right)^2$$

with probability at least $1 - \delta_t$. Lemma 10 of Abbasi-Yadkori et al. [2011] gives

$$\det(V_{i,t}) \le (\lambda_i + n_i(t-1))^d .$$

Using the above two inequalities together with $\operatorname{tr}(A^\top A) \ge \|A\|_2^2$ and plugging in $\lambda_i = 1$ we arrive at

$$\|(\tilde{\theta}_i(t) - \theta_i^*) V_{i,t}^{1/2}\|_2 \le d\left( \sqrt{d\log t + 2\log(1/\delta_t)} + \sigma_i \right) .$$

Now, we are ready to derive the confidence width for the estimate $\tilde{q}_i(t)$:

$$\begin{aligned} \|\tilde{q}_i(t) - q_i(t)\|_2 &= \|(\tilde{\theta}_i(t) - \theta_i^*) x_t\|_2 \\ &\le \|(\tilde{\theta}_i(t) - \theta_i^*) V_{i,t}^{1/2}\|_2 \|V_{i,t}^{-1/2} x_t\|_2 \\ &\le d\left( \sqrt{d\log t + 2\log(1/\delta_t)} + \sigma_i \right) \|x_t\|_{V_{i,t}^{-1}} \triangleq w_i(t) . \end{aligned} \qquad (4)$$

With these definitions we get the following result from Theorem 1:

**Theorem 4.** *Let $\mathbf{G} = (\mathbf{L}, \mathbf{H}, \mathcal{F})$ be a partial-monitoring game with $\mathcal{X} = \mathcal{K}_d$ and $\mathcal{F} = \{x \mapsto Kx : K \in \mathbb{R}^{M \times d}, K \text{ stochastic}\}$. Then, the regret of* CBP-Side *run with the least-squares estimator and confidence widths defined above satisfies*

$$\mathbb{E}[R_T] \le C_1 N + C_2 N^{3/2} d^2 \sqrt{T} \log T$$

*with some $\mathbf{G}$-dependent constants $C_1, C_2 > 0$.*

*Proof.* Plugging in the confidence widths from Equation (4) gives

$$\begin{aligned}
&\sum_{t=1}^{T} \min\{4NW_{I_t} w_{I_t}(t), 1\} \\
&\le 4N \sum_{i=1}^{N} V_i \sum_{s=1}^{n_i(T)} \min\{w_i(t_i(s)), 1\} \qquad (5)\\
&\le 4N \max_{i \in \underline{N}} W_i \\
&\qquad \sum_{i=1}^{N} \sqrt{n_i(T) \sum_{s=1}^{n_i(T)} d\left(\sqrt{d \log t_i(s) + 2\log(1/\delta_{t_i(s)})} + \sigma_i\right) \min\left\{\|x_t\|^2_{V^{-1}_{i,t_i(s)}}, 1\right\}} \\
&\le 4Nd \max_{i \in \underline{N}} W_i \left(\sqrt{d \log T + 2\log(1/\delta_T)} + \sum_{i=1}^{N} \sigma_i\right) \sum_{i=1}^{N} \sqrt{n_i(T) 2d \log T} \qquad (6)\\
&\le 4N^{3/2} d^{3/2} \max_{i \in \underline{N}} W_i \left(\sqrt{d \log T + 2\log(1/\delta_T)} + \sum_{i=1}^{N} \sigma_i\right) \sqrt{T 2 \log T}\,,
\end{aligned}$$

where in (6) we used Lemma 10 from Abbasi-Yadkori and Szepesvári [2011]. Setting $\delta_t = 1/t^2$ gives the regret bound $\mathbb{E}[R_T] \le C_1 N + C_2 N^{3/2} d^2 \sqrt{T} \log T$. □

### 6.2 Multinomial logistic regression

In this section we will consider the case when for any given action the observations follow a multinomial logit model. A $\sigma$-dimensional multinomial logit model $q^\theta : \mathcal{X} \to \mathcal{K}_\sigma$ is defined using a feature map $\Phi : \mathcal{X} \to \mathbb{R}^{\sigma \times D}$. Here, $\theta \in \mathbb{R}^D$ is the parameter vector of the model and the dependence of $q_k^\theta$ on $x$ is given by

$$q_k^\theta(x) = \frac{\exp(\eta_k^\theta(x))}{N^\theta(x)}, \quad \eta_k^\theta(x) = \phi_k(x)^\top \theta\,, \quad \text{where } N^\theta(x) = \sum_{k=1}^{\sigma} \exp(\eta_k^\theta(x))\,,$$

and the feature-vectors $(\phi_k^\top(x))_{k=1,\dots,K}$ are the rows of matrix $\Phi(x)$:

$$\Phi(x) = \begin{pmatrix} \phi_1^\top(x) \\ \vdots \\ \phi_K^\top(x) \end{pmatrix}.$$

The set $\mathcal{F}$ is implicitly defined as the set of maps such that the observations, for all actions, follow some multinomial logit model. More precisely, let $\mathcal{Q}_i$ be the set of admissible symbol-distribution models; in this section these will be some subset of all $\sigma_i$-dimensional multinomial logit models with some feature maps $\Phi_i : \mathcal{X} \to \mathbb{R}^{\sigma \times D_i}$. Define $\mathcal{F}_i = \{f : \mathcal{X} \to \mathcal{K}_M \,:\, S_i f \in \mathcal{Q}_i\}$, where $S_i f : X \to \mathcal{K}_{\sigma_i}$ is given by $(S_i f)(x) = S_i f(x)$, $x \in \mathcal{X}$. Then, $\mathcal{F} = \cap_{i \in \underline{N}} \mathcal{F}_i$. In what follows we shall assume that $\mathcal{F}$ is non-empty. This holds, for example, when the features underlying all actions correspond to a common underlying discretization of the side-information set.

As in the previous section, for each action $i$, the parameters $\theta_i$ of the $i^{\text{th}}$ model are estimated using (constrained) maximum likelihood based on the observation available for that action. To simplify the presentation of the following developments, from here on we fix an action $i$ and we will suppress the indexing of the features, parameters, etc. by the action $i$. Thus, $\Phi$ will denote the feature map for action $i$, $\theta$ will denote the underlying parameter to be tuned, etc. Thus, the set of admissible models is $\mathcal{Q} = \{q^\theta \,:\, \theta \in \Theta\}$, where $q^\theta = (q_k^\theta)_{1 \le k \le sn}$ and $\Theta$ is the set of admissible parameters.

The log-likelihood of the data available for the selected action is given by

$$\ell_t(\theta) = \sum_{s=1}^{n_i(t-1)} \sum_{k=1}^{\sigma} Z_{t_i(s),k} \log q_k^\theta(x_{t_i(s)}) \,, \quad \text{where } Z_{t,k} = \mathbb{I}_{\{Y_t = k\}}$$

and $n_i(\cdot)$, $t_i(\cdot)$ are as in the previous section. To simplify the presentation we will reindex the variables $(Z_{t_i(s),k}, x_{t_i(s)}, Y_{t_i(s)})$ as $(Z_\tau, x_\tau, Y_\tau; \tau = 1, 2, \ldots)$ (*e.g.*, $Z_{t_i(1)}$ is identified with $Z_\tau$ with $\tau = 1$). Note that the reindexing does not impact the dependence structure of the variables. In particular, by our assumption, for any $\tau > 0$ we have $Y_\tau \sim q_k^{\theta^*}(x_\tau)$ for some $\theta^* \in \Theta$. We will also drop the $i$ subindex of $n_i(t)$.

Let us first derive the estimator that we wish to use. A simple calculation shows that

$$\frac{\partial}{\partial \theta} q_k^\theta(x) = \sum_{j=1}^{\sigma} \left\{ \mathbb{I}_{\{k=j\}} - p_j^\theta(x) \right\} \phi_j^\top(x).$$

Using $\sum_{k=1}^{\sigma} Z_{\tau,k} = 1$, from this we get that

$$\frac{\partial}{\partial \theta} \ell_t(\theta) = D_t - g_t(\theta), \text{ where}$$

$$D_t = \sum_{k=1}^{\sigma} \sum_{\tau=1}^{n(t-1)} Z_{\tau,k} \phi_k(x_\tau), \quad g_t(\theta) = \sum_{k=1}^{\sigma} \sum_{\tau=1}^{n(t-1)} q_k^\theta(x_\tau) \phi_k(x_\tau) \,.$$

Let $\hat\theta_t$ be the maximum likelihood solution: $D_t = g_t(\hat\theta_t)$. We will show below that $\hat\theta_t$, the maximizer of the likelihood $\ell_t(\theta)$ is uniquely defined. Since $\hat\theta_t$ might be outside of the set of admissible parameters $\Theta$, we "project it back" to $\Theta$. Our final estimator $\hat\theta_t$ is defined as the

$$\tilde\theta_t = \operatorname{argmin}_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat\theta_t) \right\|^2_{V_t^{-1}} \,.$$

Here and in what follows, for a positive definite matrix $S \succ 0$. Further,

$$V_t = \sum_{\tau=1}^{n(t-1)} \sum_{k=1}^{\sigma} \phi_k(x_\tau)\phi_k(x_\tau)^\top \,.$$

The role of $V_t$ will become clear in the analysis. Note that in a practical implementation first one should check $\hat\theta_t \in \Theta$ because if this holds then $\tilde\theta_t = \hat\theta_t$.

To ensure that $V_t$ is invertible we assume that the algorithm generates $D\sigma$ "virtual data points" $(x_\tau)_{\tau=1,\dots,D\sigma}$ such that

$$V_{D\sigma,k} \triangleq \sum_{\tau=1}^{D\sigma} \phi_k(x_\tau)\phi_k(x_\tau)^\top \succeq \lambda_0 I \succ 0, \quad 1 \le k \le \sigma \,. \tag{7}$$

Note that this must be done for each action, independently of each other. The corresponding observations $(Y_\tau)_{\tau=1,\dots,D\sigma}$ are arbitrarily assigned to one of the available features. (This initialization allows one to encode prior information about the models, too.)

In what follows we shall assume that the following holds:

**Assumption A1** The following are assumed to hold:

(i) The set $\Theta$ is such that for all $1 \le k \le \sigma$ it holds that $0 < \inf_{\theta\in\Theta, x\in X} q_k^\theta(x) \le \sup_{\theta\in\Theta, x\in X} q_k^\theta(x) < 1$.

(ii) The constant $C_L > 0$ is known such that for any $x \in X$, $\theta, \theta' \in \Theta$, $1 \le k \le \sigma$, $|p_k^\theta(x) - p_k^{\theta'}(x)| \le C_L \|\Phi(x)(\theta - \theta')\|$, *i.e.*, $p_k^{\cdot}(x)$ is $C_L$-Lipschitzian.

Now, we are ready to state our first result:

**Lemma 3.** *Let Assumption A1 hold. Define*

$$\varepsilon_{\tau,k} = Z_{\tau,k} - q_k^{\theta_*}(x_\tau), \quad \xi_t = \sum_{\tau=1}^{n(t-1)} \sum_{k=1}^{\sigma} \varepsilon_{\tau,k}\phi_k(x_\tau) \,.$$

*Then, if* (7) *holds for some* $\lambda_0 > 0$ *then there exists some constant* $C > 0$ *such that for any* $1 \le j \le \sigma$, $x \in X$, $t \ge 1$,

$$|q_j^{\theta_*}(x) - q_j^{\tilde\theta_t}(x)| \le C \|\xi_t\|_{V_t^{-1}} \sqrt{\sum_{k=1}^{\sigma} \|\phi_k(x)\|^2_{V_t^{-1}}} \,.$$

Note that the constant can be computed as a function of the upper and lower bounds for the logit model values in Assumption A1(i) and $\lambda_0$.

*Proof.* We follow the constructions from Filippi et al. [2010]. The Hessian of the log-likelihood takes the form

$$H_t(\theta) \triangleq \frac{\partial}{\partial\theta} g_t(\theta) = \sum_{j,k=1}^{\sigma} \sum_{\tau=1}^{n(t-1)} \left[ (\mathbb{I}_{\{k=j\}} - q_j^\theta(x_\tau)) q_k^\theta(x_\tau) \right] \phi_k(x_\tau)\phi_j^\top(x_\tau) \,.$$

Using (A1)(i), one can prove that there exists some constant $C_H > 0$ such that for any $\theta \in \Theta$, $H_t(\theta) \succeq C_H V_t \succeq C_H V_D \succeq C_H \lambda_0 I \succ 0$ holds. Now define

$$\hat{H}_t = \int_0^1 \frac{\partial}{\partial \theta} g_t\big(u\theta_* + (1-u)\tilde{\theta}_t\big)\, du\,.$$

Since $g_t$ is continuous, by the Fundamental Theorem of Calculus,

$$g_t(\theta_*) - g_t(\tilde{\theta}_t) = \hat{H}_t(\theta_* - \tilde{\theta}_t). \tag{8}$$

Now, since $H_t(\theta) \succeq C_H V_t \succ 0$, $\hat{H}_t$ is non-singular and in particular

$$\hat{H}_t^{-1} \preceq \frac{1}{C_H} V_t^{-1}\,. \tag{9}$$

By Assumption A1(ii) and (8),

$$\begin{aligned}
|q_j^{\theta_*}(x) - q_j^{\tilde{\theta}_t}(x)|^2 &\le C_L^2 \sum_{k=1}^{\sigma} \left|\langle \phi_k(x), \theta_* - \tilde{\theta}_t\rangle\right|^2 \\
&\le C_L^2 \sum_{k=1}^{\sigma} \left|\langle \phi_k(x), \hat{H}_t^{-1}(g_t(\theta_*) - g_t(\tilde{\theta}_t))\rangle\right|^2\,.
\end{aligned}$$

Applying Cauchy-Schwartz and (9) gives

$$\begin{aligned}
\langle \phi_k(x), \hat{H}_t^{-1}(g_t(\theta_*) - g_t(\tilde{\theta}_t))\rangle &\le \|\phi_k(x)\|_{\hat{H}_t^{-1}} \|g_t(\theta_*) - g_t(\tilde{\theta}_t)\|_{\hat{H}_t^{-1}} \\
&\le \frac{1}{C_H} \|\phi_k(x)\|_{V_t^{-1}} \|g_t(\theta_*) - g_t(\tilde{\theta}_t)\|_{V_t^{-1}}\,.
\end{aligned}$$

Let us now bound the second term on the right-hand side:

$$\begin{aligned}
\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\|_{V_t^{-1}} &\le \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{V_t^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\tilde{\theta}_t)\|_{V_t^{-1}} \\
&\le 2\|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{V_t^{-1}}
\end{aligned}$$

Here, the second inequality follows from the optimizer property of $\tilde{\theta}_t$ and because $\theta_* \in \Theta$ by assumption. Now, it remains to put together the inequalities and to notice that $\xi_t = g_t(\hat{\theta}_t) - g_t(\theta_*)$. □

Now, we use the result of Lemma 3 to construct the confidence widths $w_i(t)$. First, we upper bound the term $\|\xi_t\|_{V_t^{-1}}$. Define $V_{t,k} = \sum_{\tau=1}^{n(t-1)} \phi_k(x_\tau)\phi_k(x_\tau)^\top$ to get

$$\begin{aligned}
\|\xi_t\|_{V_t^{-1}} &\le \sum_{k=1}^{\sigma} \left\| \sum_{\tau=1}^{n(t-1)} \varepsilon_{\tau,k}\phi_k(x_\tau) \right\|_{V_t^{-1}} \\
&\le \sum_{\tau=1}^{D\sigma} \sum_{k=1}^{\sigma} \|\phi_k(x_\tau)\|_{V_{D\sigma,k}^{-1}} + \sum_{k=1}^{\sigma} \left\| \sum_{\tau=D\sigma+1}^{n(t-1)} \varepsilon_{\tau,k}\phi_k(x_\tau) \right\|_{V_{t,k}^{-1}}\,.
\end{aligned}$$

Here we separated the terms that are obtained during the initialization as for those terms $\varepsilon_{\tau,k}$ are arbitrary (they do not posses the martingale property possesed by $\varepsilon_{\tau,k}$ coming after the initialization phase). Assuming $\lambda_0 = 1$ and that the 2-norm of $\phi_k(x_\tau)$ for any $k$ and $\tau$ is upper bounded by the $R > 0$, we get

$$\|\xi_t\|_{V_t^{-1}} \le RD\sigma^2 + \sum_{k=1}^{\sigma} \left\| \sum_{\tau=D\sigma+1}^{n(t-1)} \varepsilon_{\tau,k}\phi_k(x_\tau) \right\|_{V_{t,k}^{-1}} .$$

Now, Theorem 1 of Abbasi-Yadkori et al. [2011] gives

$$\begin{aligned}\|\xi_t\|_{V_t^{-1}} &\le RD\sigma^2 + \sum_{k=1}^{\sigma} \sqrt{2\log\frac{\det(V_{t,k})^{1/2}}{\delta_{n(t-1)}}} \\ &\le RD\sigma^2 + \sigma\sqrt{2D(1+n(t-1)R^2/D) + 2\log(1/\delta_{n(t-1)})}\,,\end{aligned}$$

Thus the confidence width $w(t)$ is defined as

$$\begin{aligned}w(t) \triangleq C\left(\sqrt{\left(2D(1+n(t-1)R^2/D) + 2\log(1/\delta_{n(t-1)})\right)} + RD\sigma^2\right) \cdot \\ \sqrt{\sum_{k=1}^{\sigma} \|\phi_k(x)\|_{V_t^{-1}}^2}\,.\end{aligned}$$

Note that this confidence bound must be computed for each action.

Now we state the regret bound result using Theorem 1.

**Theorem 5.** *With the estimate and confidence function described above,* CBP-Side *achieves expected regret*

$$\mathbb{E}\left[R_T\right] \le C_3 N + C_4 N^{3/2} D^2 \sqrt{T}\log T\,,$$

*where $C_3, C_4 > 0$ are some $\mathbf{G}$-dependent constants.*

*Proof.* The proof follows the same steps as that of Theorem 4 and thus it is omitted. □

## 7 Conclusions

In this paper we have considered partial-monitoring problems when the learner receives side information before he has to make a decision. Our solution shows that the strategy of Bartók et al. [2012] can be successfully generalized to this setting. The main idea is to use estimators that estimate the distributions of the observable symbols for each action given the side information. We have shown how the knowledge of these distributions (and confidence bounds for these distributions) can be used to make inferences about the losses of the individual actions, and thus eliminate suboptimal actions. As this approach does not attempt to directly estimate the outcome distribution, building suitable, computationally efficient estimators with good confidence bounds is expected to be less of a problem than if we attempted to estimate the distribution of the (unobserved) outcomes. However, estimating this distribution might allow the better use of information and thus may improve the dependence on the number of arms. It remains for future work to see if constructing such an estimator is feasible. In general, the dependence on the various problem dependent constants in our bounds is expected to be improvable, too. An interesting (and probably challenging) problem is to derive an estimator that matches existing lower bounds known for the bandit case such as given by Auer [2003]. Finally, we note that our results apply even when the side information is generated in a non-oblivious adversarial fashion. This is due to the strong pointwise bounds used in the construction of the confidence bounds.

## Bibliography

Y. Abbasi-Yadkori and Cs. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. *Journal of Machine Learning Research - Proceedings Track (COLT'11)*, 19:1–26, 2011.

Y. Abbasi-Yadkori, D. Pál, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits (extended version). In *NIPS*, pages 2312–2320, 2011. URL http://www.ualberta.ca/ szepesva/papers/linear-bandits-NIPS2011.pdf.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:422, 2003.

G. Bartók, D. Pál, and Cs. Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. *Journal of Machine Learning Research - Proceedings Track (COLT'11)*, 19:133–154, 2011.

G. Bartók, N. Zolghadr, and Cs. Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. To appear in ICML, 2012.

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

M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *UAI*, pages 169–178, 2011.

S. Filippi, O. Cappé, A. Garivier, and Cs. Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, pages 586–594, 2010.

D.P. Foster and A. Rakhlin. No internal regret via neighborhood watch. *CoRR*, abs/1108.6088, 2011.

D.P. Helmbold, N. Littlestone, and P.M. Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.

A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. *Lecture notes in computer science*, pages 208–223, 2001.